Uma introdução ao uso do gretl

Alexandre Loures
Rodrigo Nobre Fernandez
Universidade Federal de Pelotas

16 de outubro de 2023

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Loures, Alexandre
Uma introdução ao uso do gretl [livro eletrônico] / Alexandre Loures, Rodrigo Nobre Fernandez. – 1. ed. – Pelotas, RS: Ed. dos autores, 2023.
PDF

Bibliografia.
ISBN 978-65-00-82283-0

1. Econometria 2. Estatística 3. Estatística – Métodos 4. Linguagem de programação (Computadores) 5. Software I. Fernandez, Rodrigo Nobre. II. Título.

23–176338 CDD–330.015195

Índice para catálogo sistemático:

1. Econometria 330.015195
Aline Graziele Benitez – Bibliotecária – CRB-1/3129

Prefácio

A motivação para a elaboração deste material se deu na dificuldade apresentada por muitos alunos do Curso de Ciências Econômicas da UFPel no desenvolvimento de trabalhos aplicados nas disciplinas relacionadas a elaboração do Trabalho de Conclusão de Curso. Mesmo que possa parecer surpreendente, alguns acadêmicos ainda não sabem como utilizar planilhas eletrônicas, um tema que é fundamental para análise e manipulação de dados.

Dessa forma, o `gretl` foi o software escolhido para podermos introduzir nossos alunos à Econometria Aplicada. Esta ferramenta, é bastante amigável, não sendo necessário o conhecimento prévio em programação. Adicionalmente, o software possibilita o uso de diversas técnicas estatísticas e econométricas, o que possibilita a realização de uma gama de análises.

Descubra o fascinante mundo da econometria e análise estatística com a apostila "Uma introdução ao uso do `gretl`". Projetada para iniciantes e entusiastas que desejam mergulhar no universo da modelagem econômica, esta apostila oferece uma abordagem abrangente e prática para a utilização do `gretl`, um poderoso software estatístico de código aberto.

Através de uma narrativa didática e exemplos elucidativos, os leitores serão guiados desde os conceitos básicos até a aplicação avançada do `gretl`. Aprenda a manipular dados, realizar análises de regressão, testar hipóteses e interpretar resultados, tudo isso utilizando uma ferramenta eficiente e amigável.

Os capítulos apresentam exercícios práticos que ajudam a consolidar o conhecimento adquirido, permitindo que os leitores desenvolvam habilidades prontamente aplicáveis em suas pesquisas, estudos acadêmicos ou projetos profissionais.

Seja você um estudante de economia, pesquisador em ciências sociais ou profissional que busca aprimorar suas habilidades estatísticas, "Uma introdução ao uso do `gretl`" é o guia essencial para desbravar o vasto terreno da análise econômica com confiança e destreza. Transforme dados em *insights* valiosos e leve sua compreensão estatística para o próximo nível com esta apostila abrangente e acessível.

Devemos agradecer ao professor Lee Adkins que publicou a quinta edição do texto "Using Gretl for Principles of Econometrics" em 2018. Em muitas partes, nosso material é uma tradução para a língua portuguesa deste manual. No entanto, fizemos algumas adaptações e utilizamos principalmente o ambiente gráfico do `gretl` (GUI) porque a nossa abordagem não está relacionada ao uso de programação.

Por fim, esperamos que esse livro possa servir como um instrumento para um primeiro contato com a Econometria. Recomendamos que, após alguma familiaridade com o software e com as técnicas estatísticas e econométricas, o leitor se aventure no uso do `R` e do `Python` que são linguagens usualmente mais solicitadas no mercado de trabalho.

Rodrigo Nobre Fernandez e Alexandre Loures

Sumário

Capítulo 1

Regressão linear simples

O modelo de regressão linear simples, que é estimado usando o princípio dos mínimos quadrados, será apresentado através de um modelo simples de gastos com alimentos. Mais precisamente, será calculada uma elasticidade – uma vez que se trata de um modelo simples, previsões serão feitas, os dados serão apresentados graficamente e algumas outras estatísticas calculadas usando resultados de mínimos quadrados ordinários.

O modelo de regressão simples é dado por:

$$food\_exp_i = \beta_1 + \beta_2\, income_i + e_i \quad i = 1,\, 2,\, \ldots,\, n \tag{1.1}$$

em que $food\_exp_i$ caracteriza-se como sendo a variável dependente, $income_i$ por sua vez representa a variável independente, e_i é denominado o termo de erro e β_1 e β_2 são os parâmetros a serem estimados.

Para iniciar o modelo simples de gastos com alimentos deve-se carregar os dados contendo as informações sobre despesas com alimentos e receitas (renda familiar) no **gretl**.[1]

[1] O arquivo de dados `food.gdt` está disponível em: http://www.learneconometrics.com/gretl/poe5/POE5Data.zip

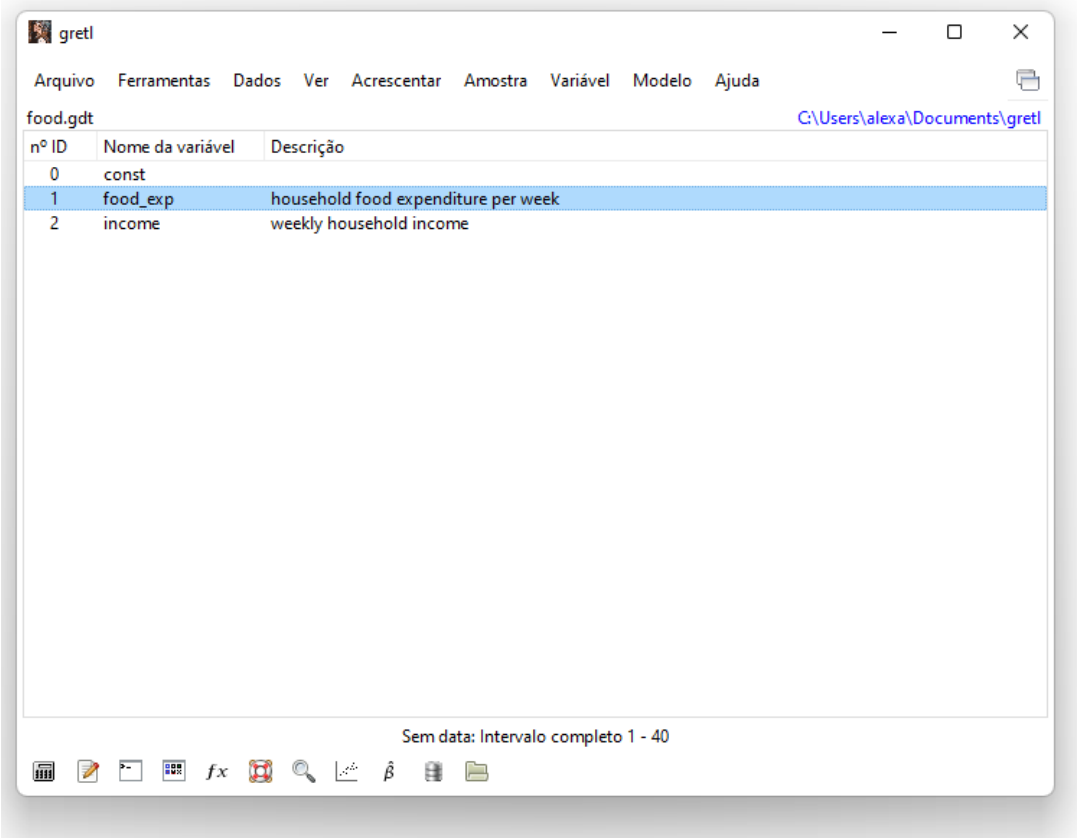


Figura 1.1: Janela principal do **gretl**.

Os dados com gastos dos alimentos são carregados através do comando **Arquivo>Abrir dados>Arquivo do usuário**,[2] na barra de menu, e escolhendo o conjunto de dados de alimentos – `food.gdt` – disponível no arquivo `POE5Data`. A primeira observação que se faz é que, a coluna **Descrição** contém algumas informações sobre as variáveis que estão na memória do programa. Importante destacar que nem sempre essas informações estão disponíveis, entretanto, é possível rotular manualmente uma variável. Por exemplo, a Figura 1.2 mostra que se deve destacar (i.e., sombreado de azul claro) a variável para qual se irá acrescentar o rótulo e, então, clica-se com o botão direito do mouse para abrir um menu que irá conter algumas opções, entre essas está **Editar características**. Selecione essa opção para que se possa abrir uma janela, Figura 1.3, em que será possível escrever uma descrição para a variável selecionada – no presente caso $food\_exp$.

[2]Ou simplesmente clique duas vezes sobre o ícone da base de dados.

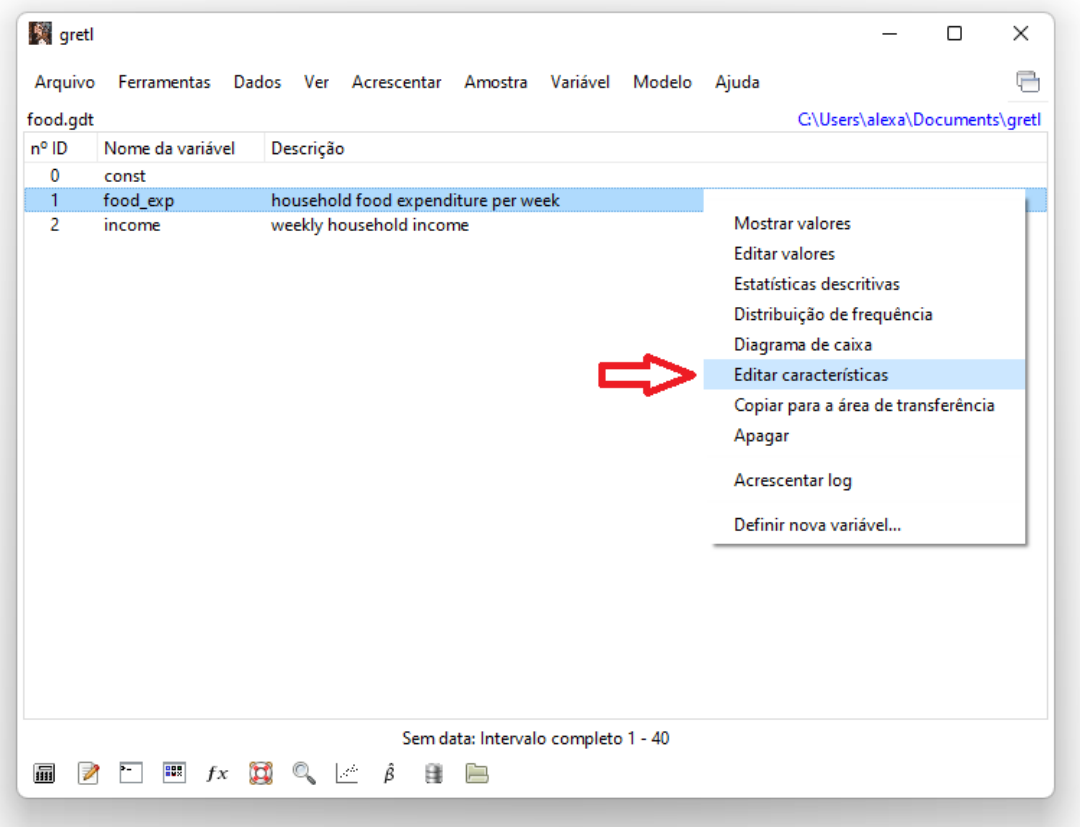


Figura 1.2: Destacando a variável de interesse.

Note que nessa janela que se abre será possível alterar o nome da variável, rotular a variável bem como adicionar um nome que será apresentado nos gráficos. Para exemplificar, na opção **Nome a apresentar (mostrado nos gráficos)** coloca-se **Despesas alimentação/semana** para a variável $food\_exp$ e **Renda semanal** ($ 100) para a variável *income*. Essas manipulações nas variáveis da base de dados se justificam para tornarem as saídas mais fácil de entender.

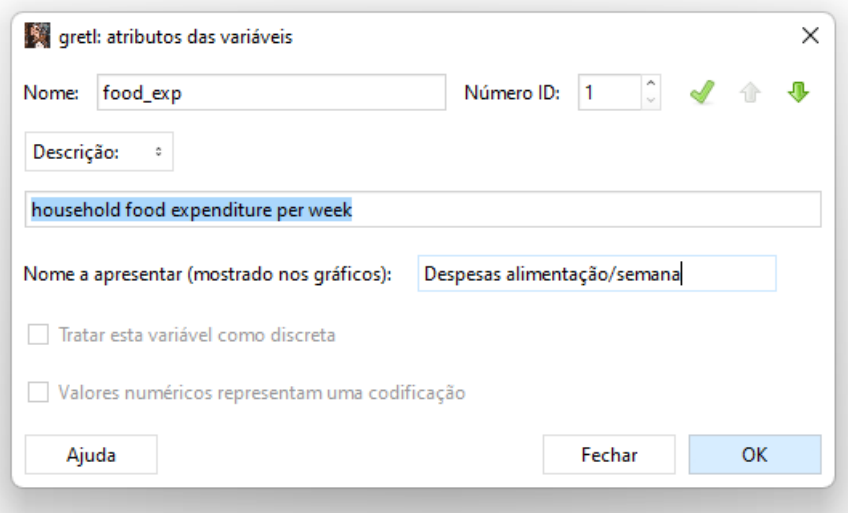


Figura 1.3: Caixa de diálogo de edição de variável.

1.1 Representando graficamente os dados

Para gerar um gráfico de dispersão entre as variáveis *food_exp* e *income*, na barra de menu, deve-se seguir o seguinte comando **Ver>Gráfico das variáveis>X-Y em dispersão**. Essa sequência de passos abrirá a janela mostra na Figura 1.4. Outra forma seria usar o quarto ícone da direita para a esquerda, , na barra de ferramentas do **gretl**, parte inferior da janela principal. Note que os rótulos aplicados na Figura 1.4 aparecem nos eixos do gráfico, Figura 1.5.

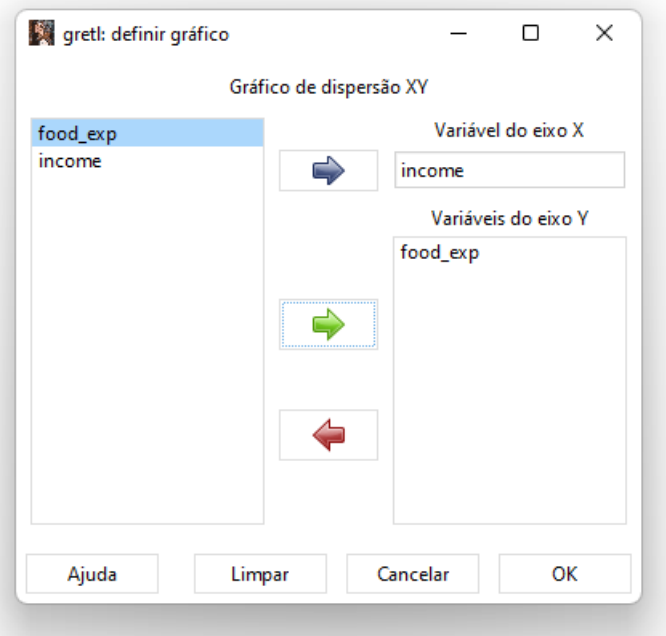


Figura 1.4: Caixa de diálogo para o gráfico de dispersão.

A Figura 1.5 mostra os gastos semanais com alimentação no eixo y enquanto no eixo x tem-se a renda semanal. Por padrão, o **gretl** também traça a linha de regressão ajustada. Agora torna-se mais fácil compreender a utilidade em se rotular as variáveis por meio da caixa de diálogo da Figura 1.3. A saída do gráfico mostra ambos os eixos x e y rotulados de uma forma intuitiva bem como o título do gráfico.

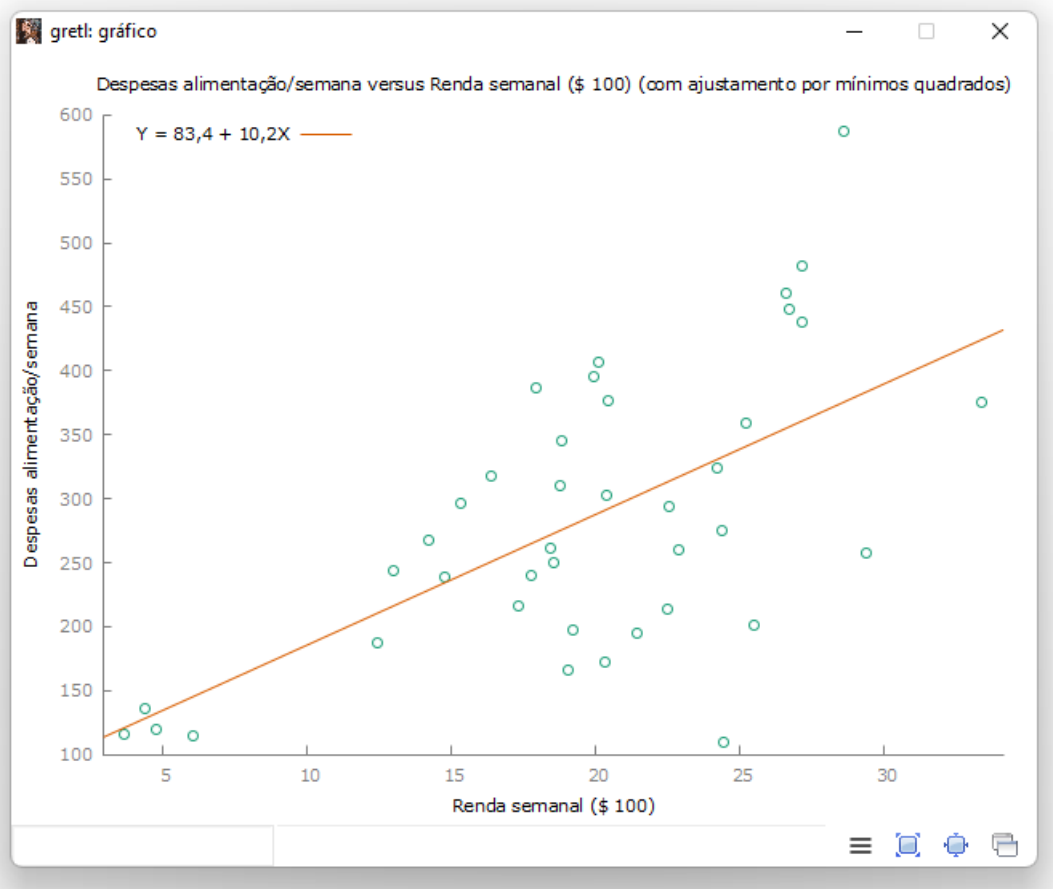


Figura 1.5: Gráfico de dispersão dos dados de gastos com alimentos.

1.2 Estimando o modelo de gastos com alimentação

Nesta seção, será demonstrado como usar o **gretl** para estimar os parâmetros da equação de gastos com alimentação:

$$food\_exp_i = \beta_1 + \beta_2\, income_i + e_i \quad i = 1, 2, \ldots n \tag{1.2}$$

Na barra de menus, selecione **Modelo>Mínimos Quadrados Ordinários** no menu suspenso, Figura 1.6, para abrir a caixa de diálogo mostrada na Figura 1.7.

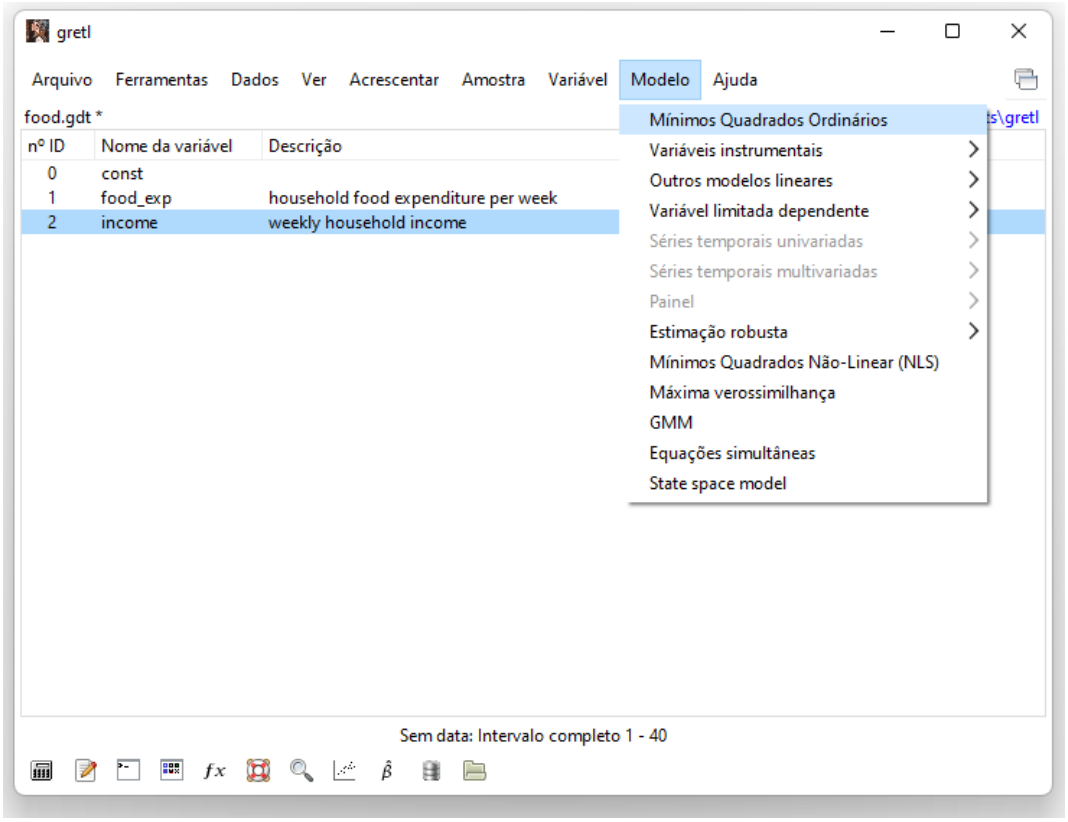


Figura 1.6: Caixa de diálogo para os mínimos quadrados ordinários.

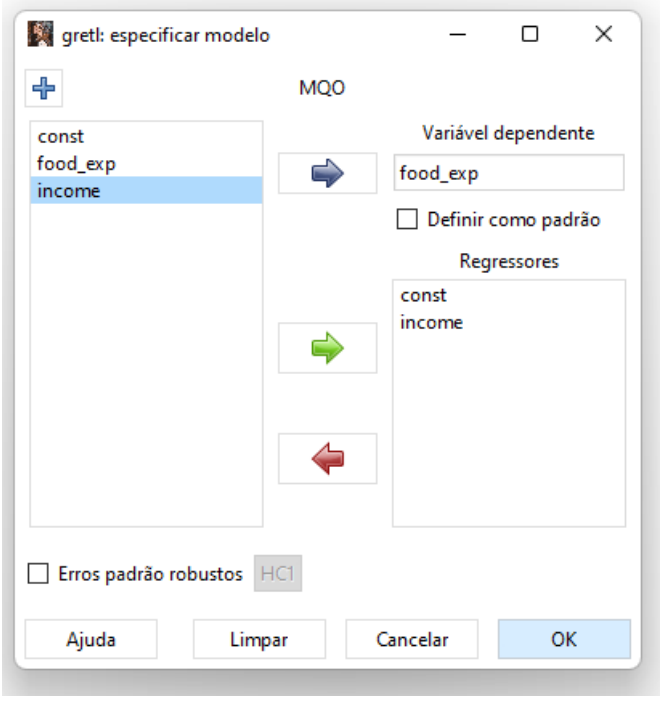


Figura 1.7: Caixa de diálogo para especificação do modelo.

Nessa caixa de diálogo, Figura 1.7, deve-se informar ao **gretl** qual variável usar como variável dependente e qual será a variável independente. Observe que, por padrão, o **gretl** assume que se deseja estimar um intercepto (β_1) e, assim, inclui uma constante como variável independente – colocando a variável `const` na lista de regressores por padrão. Por outro lado, para colocar x, no presente caso *income*, como uma variável independente, destaque-a com o cursor (i.e., sombreado azul claro), Figura 1.7, e clique no botão de seta verde,  , para adicioná-la. Para adicionar a variável dependente destaque-a (i.e., sombreado azul claro) com o cursor e clique no botão de seta azul,  e, por sua vez, para retirar um regressor da lista basta destacá-lo (i.e., sombreado azul claro) e clicar no botão de seta vermelha,  , para excluí-lo. Uma vez especificado o modelo clique no botão `OK` da caixa de diálogo da Figura 1.7. Isso reportará a janela mostrada na Figura 1.8.

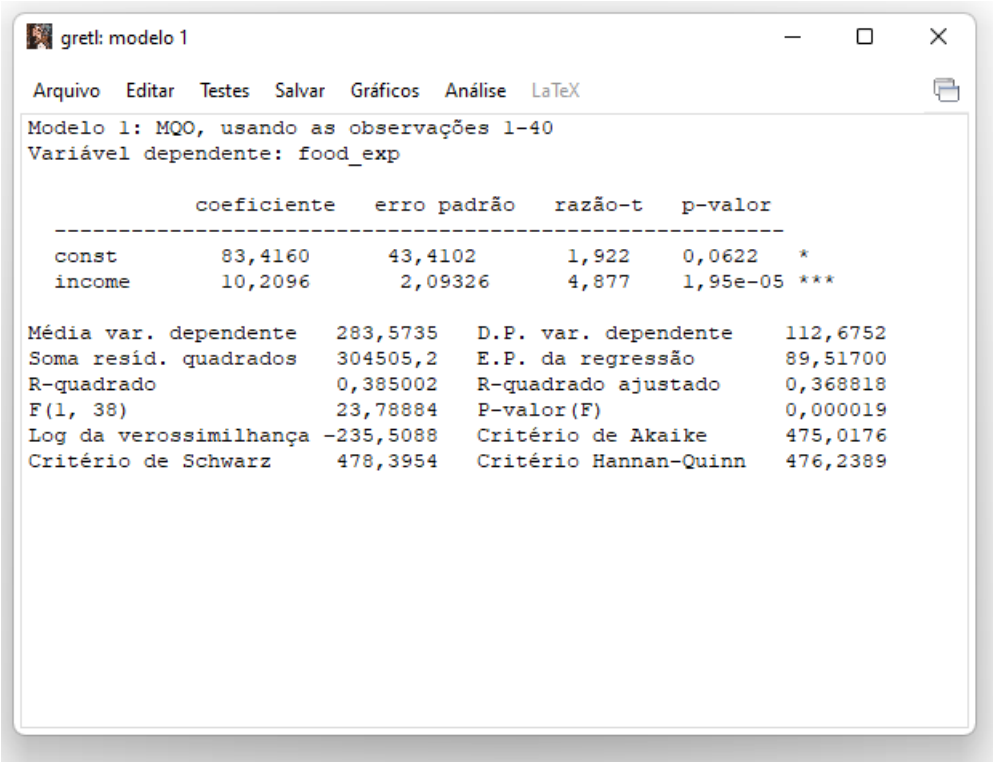


Figura 1.8: Resultados da regressão.

Destaca-se que, uma vez estimado o modelo, pode-se realizar operações subsequentes (gráficos, testes, análises, etc.) sobre o modelo. Uma forma mais elegante para apresentar os resultados, especialmente em modelos muito pequenos como a regressão linear simples, é usar a forma de equação. Neste formato, os resultados para o modelo de gastos com alimentação podem ser apresentados como:

$$\widehat{food\_exp} = \underset{(43,4102)}{83,4160} + \underset{(2,09326)}{10,2096}\, income$$

$$n = 40 \quad \bar{R}^2 = 0,3688 \quad F = (1,\ 38) = 23,789 \quad \hat{\sigma} = 89,517$$

(erros padrão entre parênteses)

1.3 Elasticidade

A elasticidade é um conceito importante em economia e caracteriza-se como sendo o percentual de variação em uma determinada variável, dada uma variação percentual em outra variável. Pode ser relacionada com sensibilidade ou reação da variável em questão em relação a outras variáveis.

$$\epsilon = \frac{\text{mudança percentual em } y}{\text{mudança percentual em } x} = \frac{\Delta y}{\Delta x}\,. \tag{1.3}$$

Em termos do modelo de gastos com alimentação, está interessado na elasticidade dos gastos médios com alimentos em relação às mudanças da renda:

$$\epsilon = \frac{\Delta\,(y)\,/\,E\,(y)}{\Delta\,x\,/\,x} = \beta_2\,\frac{x}{E\,(y)}\,, \tag{1.4}$$

em que $E\,(y)$ e x são usualmente substituídos por suas médias amostrais e β_2 por sua estimativa. Note que a média para $food\_exp$ e renda (x) pode ser obtidas através do comando **Ver>Estatísticas descritivas**. Na caixa de diálogo que abrir, Figura 1.9 use o cursor para destacar (i.e., sombreado azul claro) ambas as variáveis e, em seguida, clique no botão de seta verde,  , e clique no botão `OK`.

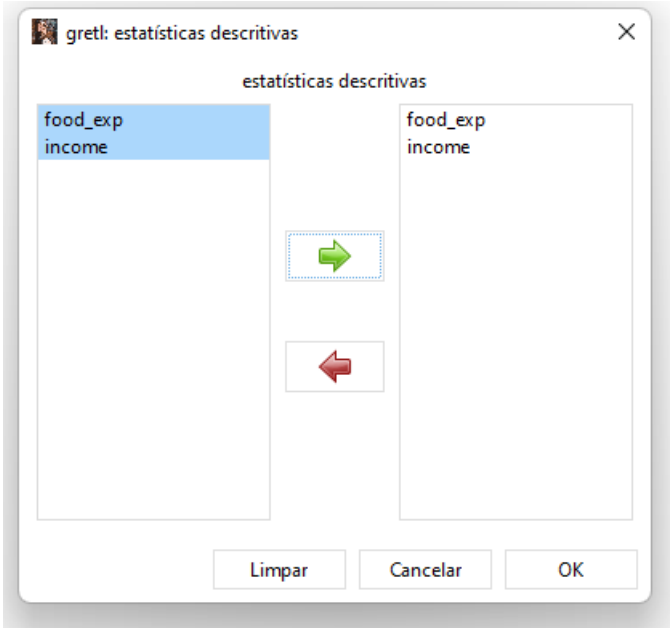


Figura 1.9: Caixa de diálogo para estatísticas descritivas.

Isso irá produzir a saída mostrada na Figura 1.10. Assim, a Equação 1.4 pode ser calculada manualmente. Então, usando o parâmetro da regressão e as estatísticas descritivas tem-se que: $\hat{\beta}_2 \times (income \,/\, E\,(food\_exp)) = 10,2096 \times (19,605 \,/\, 283,54) = 0,705855$. Assim, como o valor para a elasticidade ficou abaixo de 1, os gastos com alimentação são inelástico a variações na renda. Mais precisamente, a variação nos gasto com alimentação é proporcionalmente menor que a variação na renda.

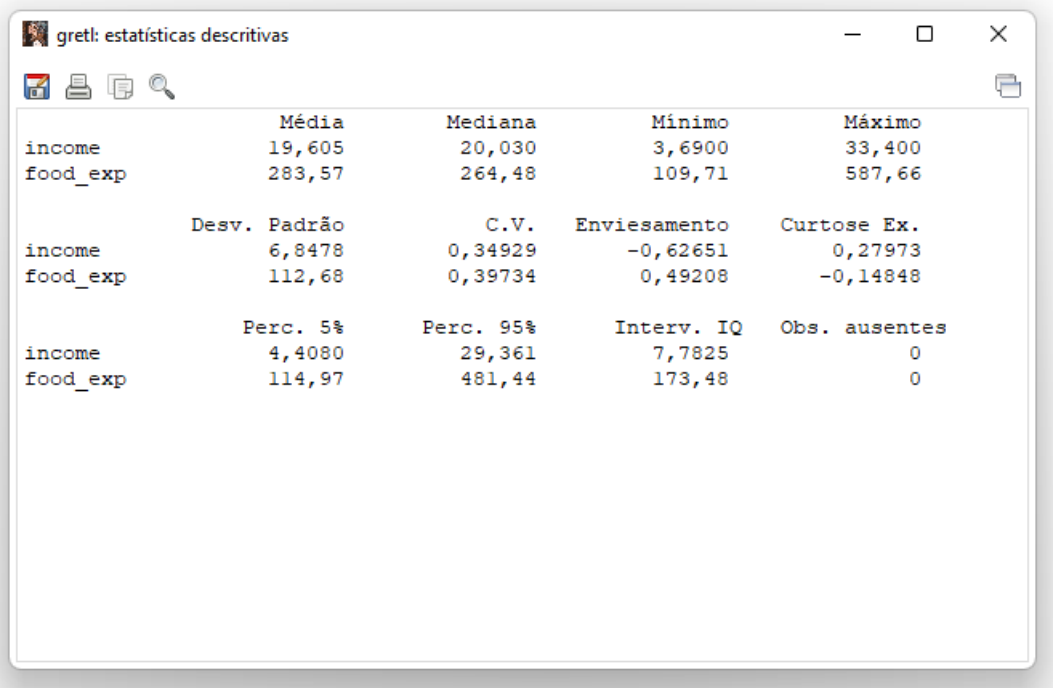


Figura 1.10: Estatísticas descritivas.

1.4 Predição

Uma vez de posse dos resultados da estimação, pode-se fazer previsões sobre os gastos com alimentação para uma dada renda x qualquer. Por exemplo, suponha que se queira saber qual o gasto com alimentação para uma família cuja renda média semanal familiar é de $\$\,2.000$. Como a renda é medida em $\$\,100$, então, $\dfrac{\$\,2.000}{\$\,100} = 20$. Logo,

$$\widehat{food\_exp}_i = 83,42 + 10,21\, income_i = 83,42 + (10,21 \times 20) = 287,61 \qquad (1.5)$$

Ou seja, uma família cuja renda média semanal é de $\$\,2.000$ terá um gasto semanal com alimentação de $\$\,287,61$.

1.4.1 Estimando a variância

Uma vez que o modelo é estimado empregando Mínimos Quadrados Ordinários, as variâncias e covariância estimadas podem ser obtidas selecionando o comando **Análise>Matriz de covariâncias dos coeficientes**, Figura 1.11.

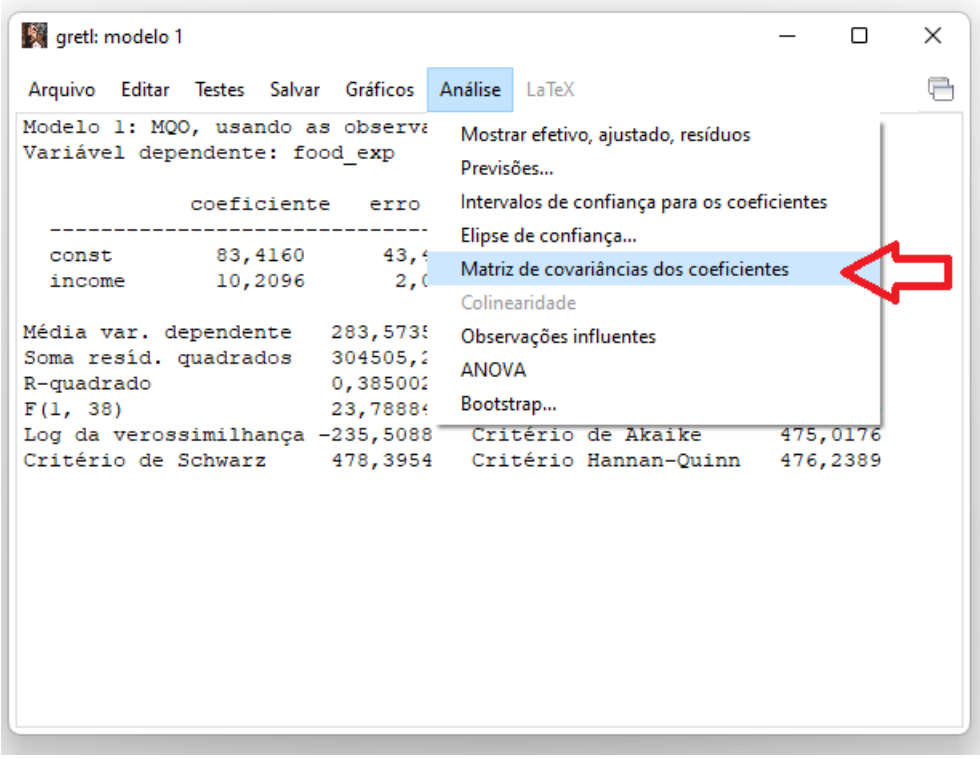


Figura 1.11: Obtendo a matriz das variâncias e covariância.

Na Figura 1.12 apresenta as variâncias estimadas do estimador de Mínimos Quadrados Ordinários para o intercepto (β_1) e para a inclinação (β_2) que são, respectivamente, $1.884,44$ e $4,38175$. Note que os erros padrão, na Figura 1.8, são simplesmente as raízes quadradas desses valores. Por sua vez, a covariância estimada entre o intercepto e a inclinação é $-85,9032$.

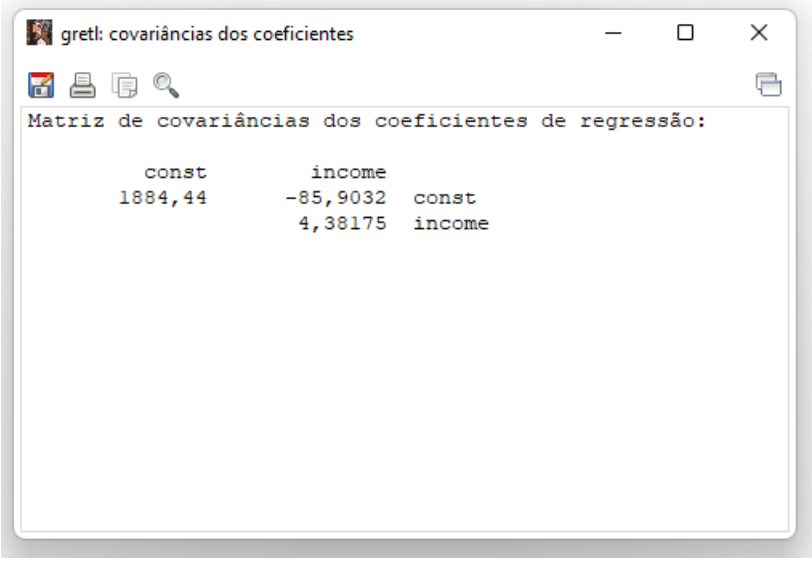


Figura 1.12: Matriz de variância-covariância.

Capítulo 2

Estimação de intervalo e teste de hipóteses

Discutiremos como gerar intervalos de confiança e testar hipóteses usando **gretl**. O software inclui vários utilitários úteis que o ajudarão a obter valores críticos e valores `p` de várias distribuições de probabilidade importantes. Uma maneira de fazer isso é observar a estimativa do parâmetro dos Mínimos Quadrados Ordinários (MQO) juntamente com uma medida de sua precisão, ou seja, seu erro padrão estimado.

O intervalo de confiança serve a um propósito semelhante, embora seja muito mais simples de interpretar porque fornece limites superiores e inferiores entre os quais o parâmetro desconhecido ficará com uma determinada frequência em amostras repetidas.

No **gretl**, você pode obter intervalos de confiança por meio de uma caixa de diálogo ou construindo-os manualmente usando resultados de regressão salvos. Você pode procurar o valor crítico apropriado em uma tabela ou usar a função crítica do **gretl**. Considere a equação de um intervalo de confiança:

$$P\left[b_k - t_c\, se\left(b_k\right) \;\leq\; \beta_k \;\leq\; b_k + t_c\, se\left(b_k\right)\right] = 1 - \alpha \tag{2.1}$$

Lembre-se de que b_k é o estimador de MQO de β_k e que $se\left(b_k\right)$ é seu erro padrão estimado. A constante t_c é o valor crítico de $\alpha\,/\,2$ da distribuição `t` e α é a probabilidade total desejada associada à área de "rejeição" (a área fora do intervalo de confiança). Você precisará saber o valor crítico t_c, que pode ser obtido de uma tabela estatística, da caixa de diálogo **Ferramentas>Tabelas estatísticas** contidas no programa.

Primeiro, tente usar a caixa de diálogo mostrada na Figura 2.1. Escolha a guia para a distribuição `t` e diga ao **gretl** quanto peso colocar na cauda direita da distribuição de probabilidade e quantos graus de liberdade sua estatística `t` tem, no nosso caso, 38. Depois de fazer isso, clique em `OK`. Você obterá o resultado mostrado na Figura 2.2. Ele mostra que para o t_{38} com $\alpha\,/\,2$ probabilidade de cauda direita de 0.025 e $\alpha = 0.05$, o valor crítico é 2.02439.

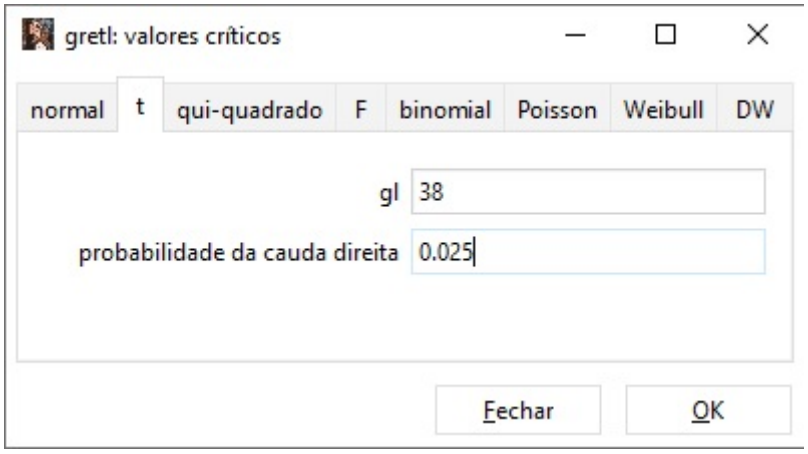


Figura 2.1: Obtenção dos valores críticos Ferramentas>Tabelas estatísticas.

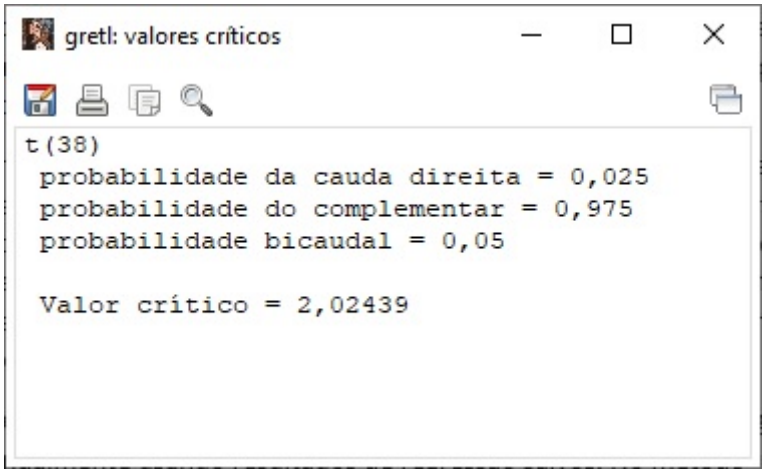


Figura 2.2: O valor crítico obtido na caixa de diálogo Ferramentas>Tabelas estatísticas.

Exemplo: com arquivo food.gdt

Este exemplo é baseado no modelo de gastos com alimentos:

$$food\_exp_i = \beta_1 + \beta_2\, income + e_i \qquad\qquad i = 1, \ldots, n \tag{2.2}$$

O objetivo é estimar um intervalo de confiança de 95% para a inclinação, β_2. Estime o modelo usando os mínimos quadrados da maneira usual. Clique em **Modelo>Mínimos quadrados ordinários** no menu principal, preencha as variáveis dependentes e independentes na caixa de diálogo do MQO e clique em `OK`.

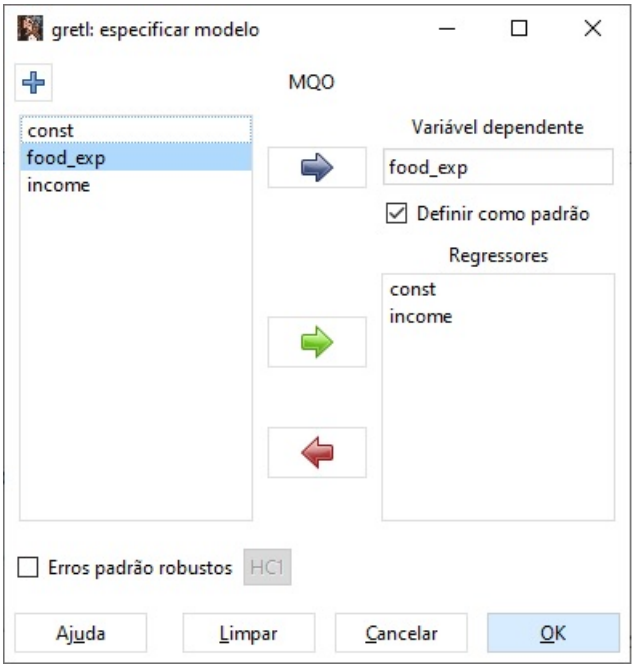


Figura 2.3: Configuração usual do modelo de MQO.

Agora escolha **Análise>Intervalos de confiança** para coeficientes no menu suspenso da janela de modelos para gerar o resultado mostrado na Figura 2.3. O ícone α em caixa pode ser usado para alterar o tamanho do intervalo de confiança, que:

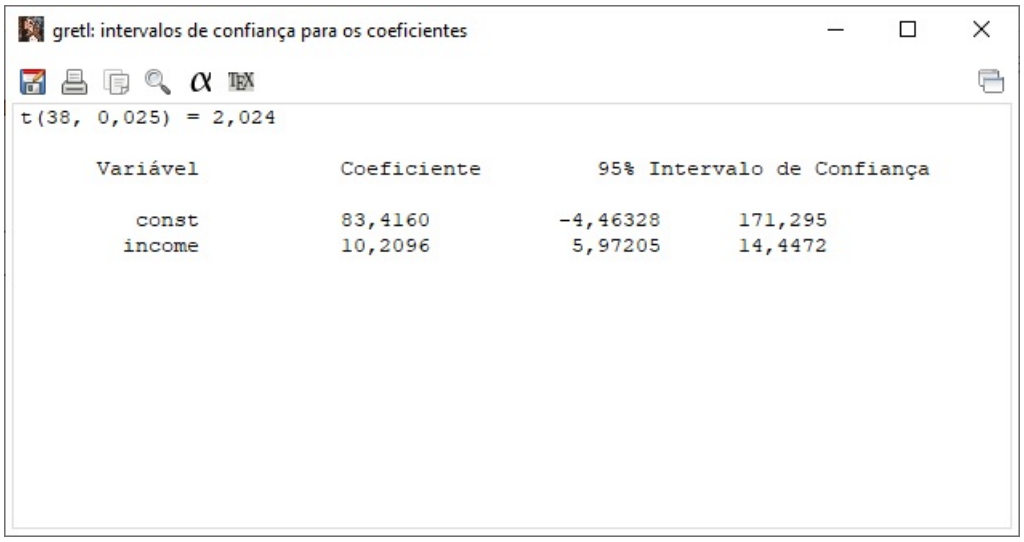


Figura 2.4: O intervalo de confiança de 95% para o coeficiente de renda no exemplo de gasto com alimentação usando o diálogo.

2.1 Teste de hipóteses

Testes de hipóteses permitem comparar o que supomos ser verdade com o que observamos por meio de dados. Suponha que eu acredite que o gasto autônomo semanal com comida não seja inferior a $ 40, eu extraio uma amostra, calculo uma estatística que mede o gasto com comida e então comparo minha estimativa com minha conjectura usando um teste de hipóteses. A hipótese nula é que $\beta_2 = 0$ contra a alternativa de que é positivo (ou seja, $\beta_2 > 0$). A estatística de teste é:

$$t = \frac{(\beta_2 - 0)}{s_e\ (\beta_2)} \sim t_{38} \tag{2.3}$$

se $\beta_2 = 0$ (a hipótese nula é verdadeira). Selecione $\alpha = 0.05$ o que torna o valor crítico para a alternativa unilateral ($\beta_2 > 0$) igual a $1,686$. A regra de decisão é rejeitar H_0 em favor da alternativa se o valor calculado da estatística `t` estiver dentro da região de rejeição do teste; isto é, se for maior que $1,686$. A informação necessária para calcular `t` está contida nos resultados de estimativa de mínimos quadrados produzidos por **gretl**:

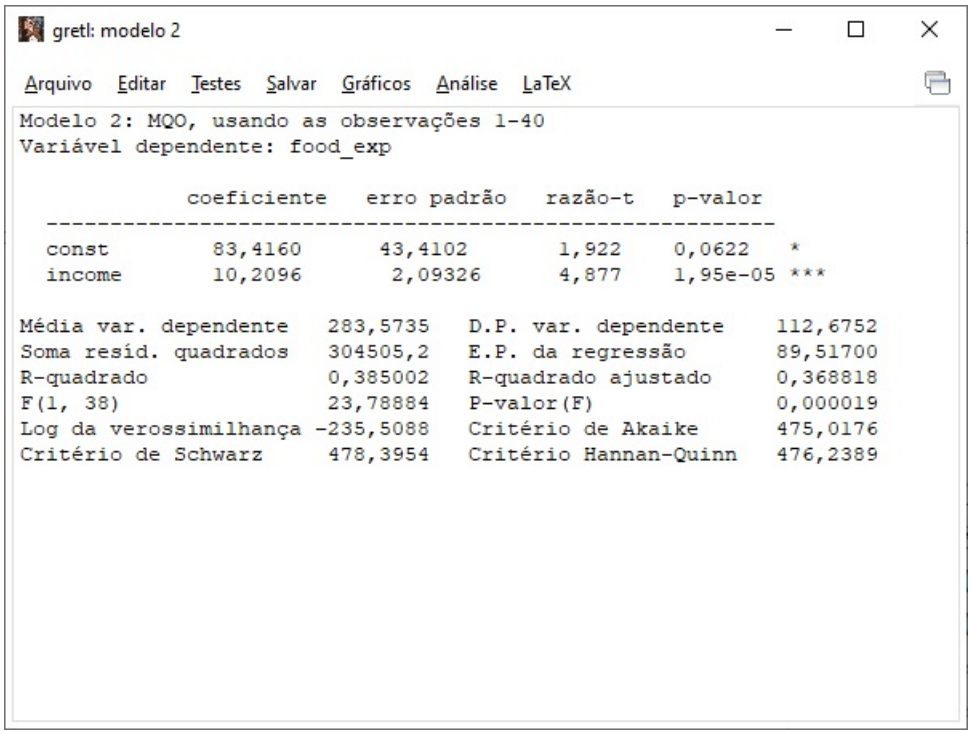


Figura 2.5: Resultados do modelo de MQO

Os cálculos:

$$t = \frac{(\beta_2 - 0)}{s_e\ (\beta_2)} = (10.21 - 0) \,/\, 2.09 = 4.889 \tag{2.4}$$

Como esse valor está dentro da região de rejeição, há evidências suficientes no nível de significância de 5% para nos convencer de que a hipótese nula está incorreta; a

hipótese nula é rejeitada neste nível de significância. **gretl** é usado para obter o valor `p` para este teste usando o menu superior Ferramentas (Figura 2.5). Nesta caixa de diálogo, você insere os graus de liberdade desejados para sua distribuição t_{38}, o valor de:

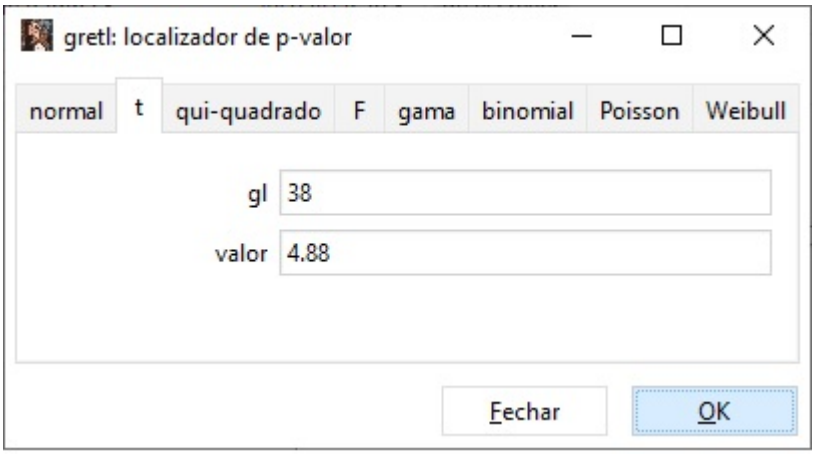


Figura 2.6: Ferramentaas>Localizador de p-valor

Substituindo na Equação 2.4 β_2 (10.21), seu valor sob a hipótese nula - algo que **gretl** se refere como "média" (0) e o erro padrão estimado da impressão (2, 09). Isso produz as informações da Figura 2.6:

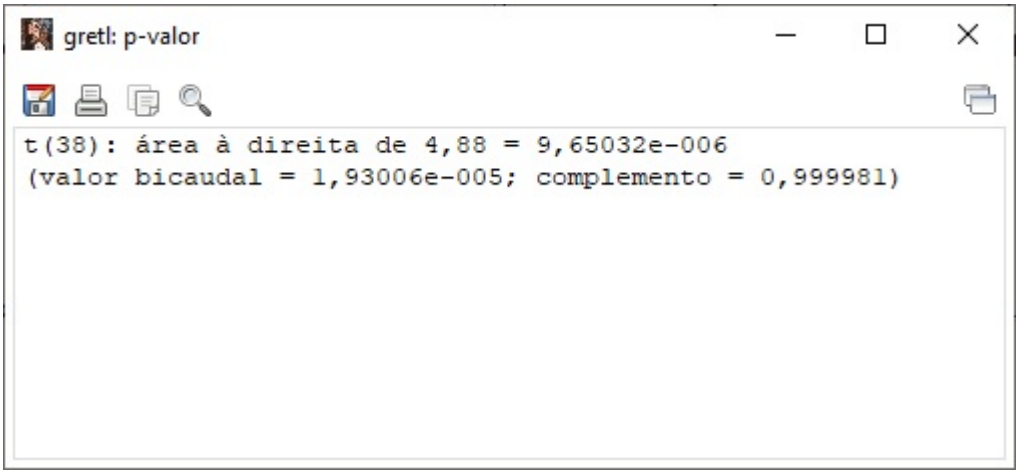


Figura 2.7: Ferramentas>Localizador de p-valor

Assim, a área de uma variável aleatória t_{38} à direita de 4, 88, ou seja, o valor `p` do teste, é quase zero. Como o valor de `p` está bem abaixo de $\alpha = 0.05$, a hipótese é rejeitada.

Capítulo 3

Previsão, qualidade do ajuste e problemas de especificação

Neste capítulo serão apresentadas diversas extensões do modelo de regressão linear simples. Primeiramente, predições condicionais serão geradas usando os resultados armazenados na memória do **gretl** após estimar um modelo. Logo após se discute um teste estatístico comumente utilizado para checar a qualidade do ajuste do modelo fornecida pela regressão. Mais precisamente, este teste estatístico determinará quão bem os dados da amostra se ajustam a uma distribuição de uma população com distribuição normal. Simplificando, este teste levanta a hipótese se uma amostra é distorcida ou representa os dados que se esperaria encontrar na população real.

Destaca-se que a escolha de uma forma funcional adequada para uma regressão linear é de suma importância. Sendo assim, este capítulo apresentará algumas formas funcionais para uma regressão linear, entre as seguintes especificações possíveis:

1. Polinomiais;
2. Logarítmicas;
3. linear-log – variável dependente em nível e variável(is) independente(s) em log;
4. log-linear – variável dependente em log e variável(is) independente(s) em nível ;
5. log-log – variável dependente em log e variável(is) independente(s) também em log.

3.1 Previsão no modelo de gastos com alimentação

A geração de valores previstos para os valores de gastos com alimentação para uma família com um dado nível de renda é muito simples no **gretl**. Isto já foi demonstrado na Seção 1.4 em que, para uma família que possui uma renda semanal igual a $income_0 = \$\,2000$, foi previsto que essa família gaste aproximadamente $\$\,287,61$ com alimentação por semana (lembre-se que a renda é medida em US$\$\,100$ no conjunto de dados).

Por outro lado, para obter o intervalo de confiança de 95% é um pouco mais difícil uma vez que não existem comandos no **gretl** para realizarem esse cálculo. No entanto, essa estatística pode ser obtida manualmente através da seguinte fórmula:

$$\widehat{var}\,(f) = \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{T} + (income_0 - \overline{income})^2 \times \widehat{var}(\beta_2) \tag{3.1}$$

Na Figura 1.8 nota-se que o erro padrão da regressão é igual a $89,517$, logo, tem-se que $\hat{\sigma}^2 = (89,517)^2 = 8013,29$. Por sua vez, da Figura 1.12 tem-se que $\widehat{var}(\beta_2) = 4,3818$. Já o comando para obter o valor médio da renda foi apresentado na Seção 1.3, Figura 1.10, sendo o valor igual a $19,605$. O valor crítico de t_{38} 5% é de $2,0244$, Figura 2.2. Assim, o cálculo do intervalo de confiança será:

$$\widehat{var}\,(f) = 8013,2941 + \frac{8013,2941}{40} + (20 - 19,605)^2 \times 4,3818 = 8214,31 \tag{3.2}$$

Então, o intervalo de confiança para os valores previstos é dado por:

$$\widehat{food\_exp}_0 = \pm t_x\, se\,(f) = 287,6069 \pm 2,0244\sqrt{8214,31} = [104,132;\ 471,086] \tag{3.3}$$

Isso implica que o intervalo de confiança de 95% centrado em $287,609$ é $(104,132;\ 471,086)$.

3.2 Qualidade do ajuste

O coeficiente de determinação é utilizado na teoria da regressão linear e expressa quão bem a equação de regressão se ajusta aos dados, i.e., qualidade do ajuste. Mais precisamente, qual a proporção da variação na variável dependente que é explicada pela variação da(s) variável(is) independente(s). R^2 é a razão entre a variação explicada e a variação total; assim, ele é interpretado como a *fração da variação amostral em y que é explicada por x*. É muito pouco provável que se tenha uma correlação perfeita ($R^2 = 1$) na prática, uma vez que existem muitos fatores que determinam as relações entre variáveis na vida real.

A forma mais simples de se obter o R^2 é diretamente da saída da regressão no **gretl**. Isso é mostrado na figura Figura 3.1 através da estatística `R-quadrado` igual a $0,385002$, sombreado com azul claro na janela **gretl modelo 1**.

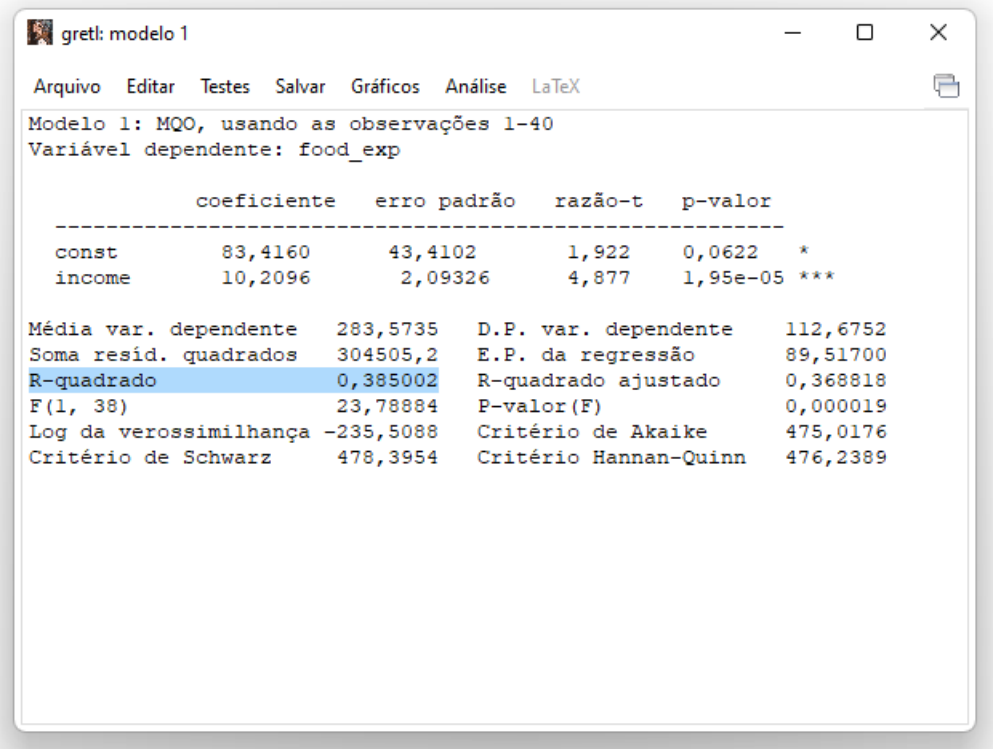


Figura 3.1: Coeficiente de determinação.

Manualmente o coeficiente de determinação pode ser calculado usando a tabela ANOVA obtida após uma regressão usando o comando **Analysis>ANOVA** no menu suspenso da janela do modelo conforme a Figura 3.2. Na tabela ANOVA apresenta na Figura 3.3 são encontrados os valores para Soma dos Quadrados dos Resíduos (SQR), Soma dos Quadrados Explicados (SQE) e Soma Total de Quadrados (STQ) bem como o **gretl** faz o cálculo para o coeficiente de determinação, R^2. Então, o R^2 é calculado da seguinte forma:

$$R^2 = \frac{\text{SQE}}{\text{STQ}} = 1 - \frac{\text{SQR}}{\text{STQ}} = \frac{190627}{495132} = 0,385002 \tag{3.4}$$

em que, conforme a Figura 3.3, SQE = 190627, SQT = 495132 e SQR = 304505.

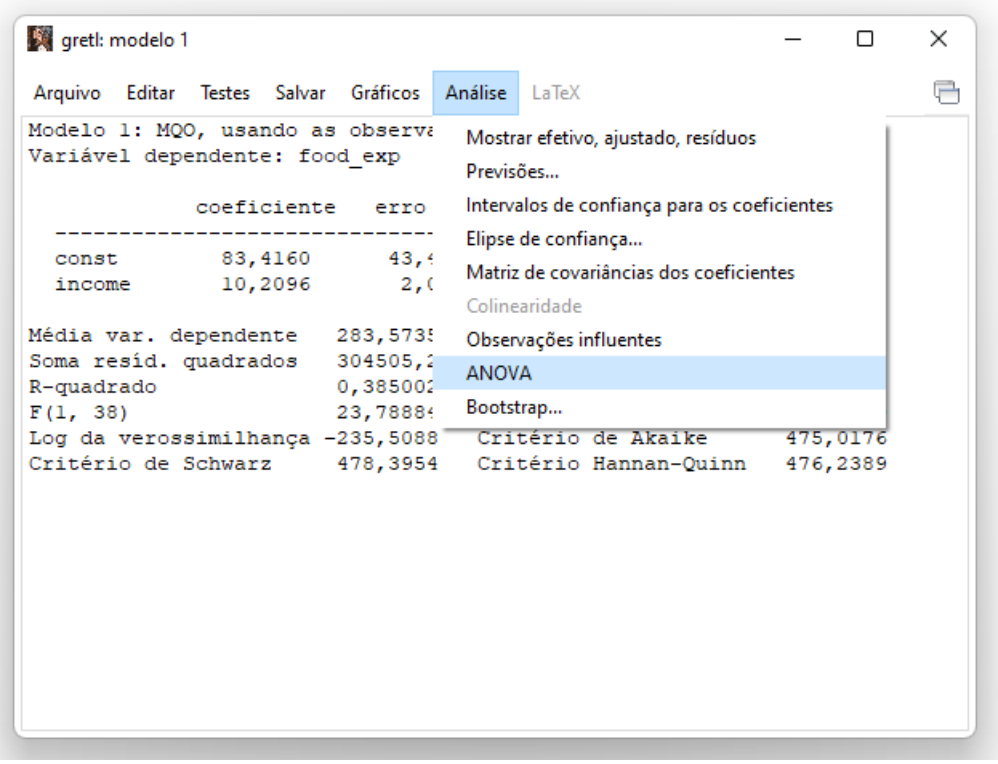


Figura 3.2: Tabela ANOVA.

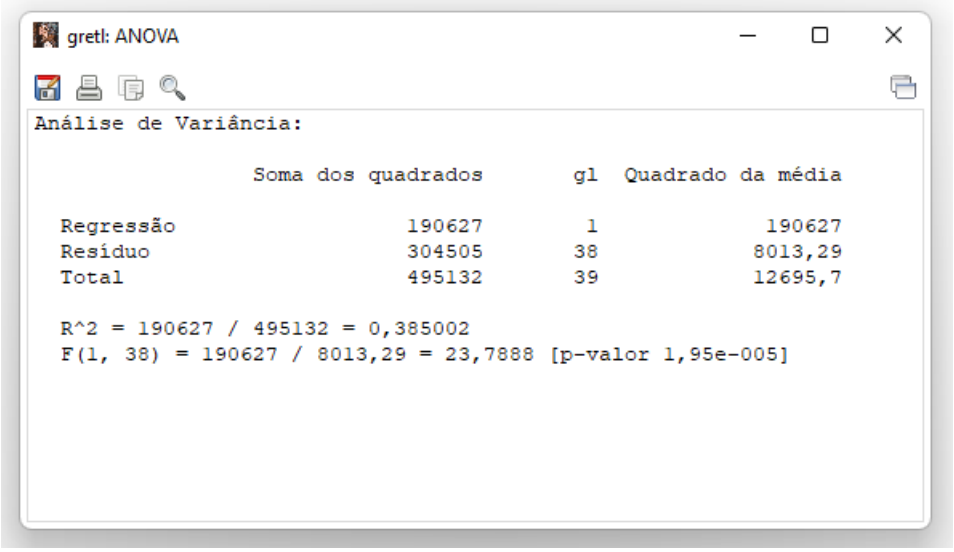


Figura 3.3: Saída da tabela ANOVA.

3.3 Escolhendo a forma funcional

Não há razão para considerar que gasto com alimentação e renda apresentem uma relação linear. Na verdade, é bem provável que essas duas variáveis apresentem uma relação não linear, pois um assalariado de baixa renda gastará todo Real (R$) adicional em comida enquanto um assalariado de alta renda gastará bem menos de cada Real (R$) adicional que recebe.

Entretanto, como se sabe, as não linearidades podem ser contornadas com a transformação da variável dependente (y) ou independente (x) ou de ambas. Outro exemplo é relação entre insumos e produto que é regida no curto prazo pela lei dos rendimentos decrescentes, sugerindo que uma curva convexa é mais apropriada. Mas como já dito, uma simples transformação das variáveis (y, x ou ambas) produz um modelo linear nos parâmetros (mas não necessariamente nas variáveis).

Importante destacar que a forma funcional escolhida deve ser consistente com a forma como os dados são realmente gerados. A escolha de uma forma funcional que, quando devidamente parametrizado, não consegue gerar seus dados, seu modelo está mal especificado, ou seja, especificado incorretamente. O modelo, na melhor das hipóteses, pode não ser útil e, na pior das hipóteses, ser totalmente enganoso.

A transformação de variáveis no **gretl** é bastante simples e é realizada na janela principal através do menu suspenso do comando `Acrescentar`, Figura 3.4. Esse menu suspenso fornece acesso a várias transformações. Uma vez escolhida um tipo de transformação, a variável transformada será adicionada automaticamente ao conjunto de dados, bem como sua descrição.

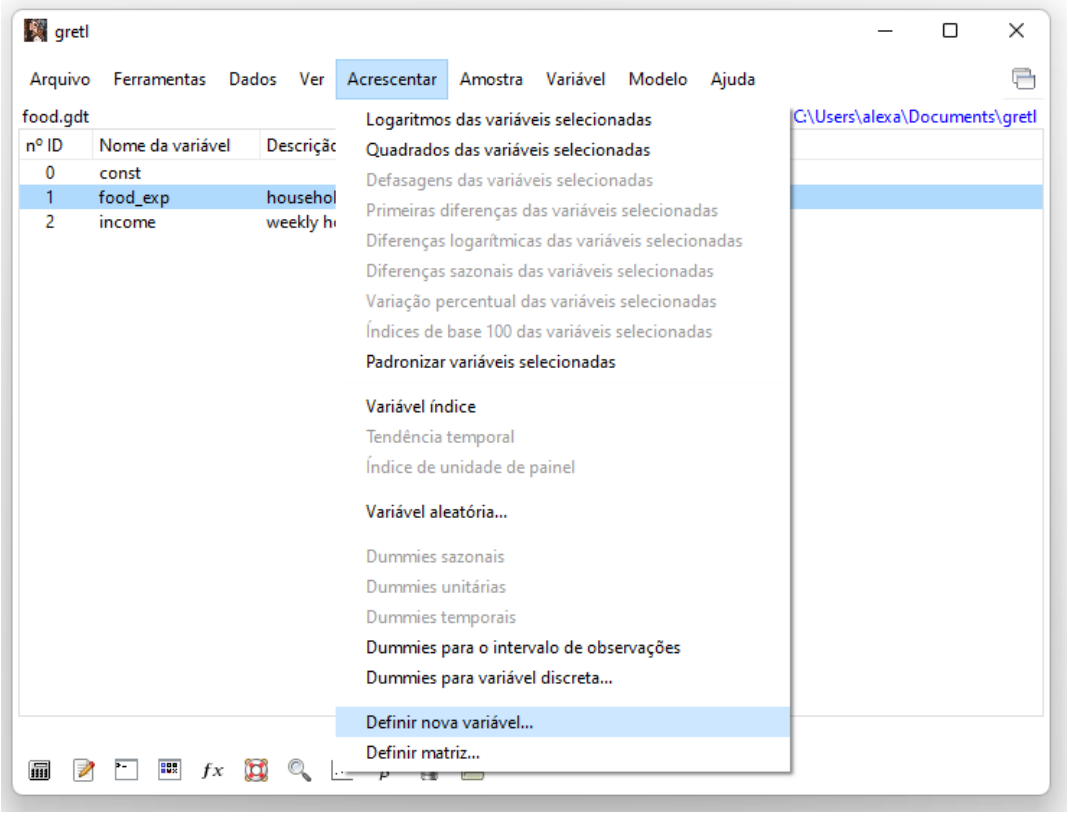


Figura 3.4: Menu para transformação de variáveis.

A penúltima opção, `Definir nova variável...`, (sombreada de azul claro) permite realizar transformações mais complicadas tais como: raiz quadrada, seno, cosseno, valor absoluto, exponencial, mínimo, máximo, etc..

3.3.1 Especificação linear-log

A especificação **linear-log** do modelo de gastos com alimentação usa o logaritmo neperiano (natural) da renda como variável independente:

$$food\_exp = \beta_1 + \beta_2 \ln(income) + e \tag{3.5}$$

Assim, para adicionar o logaritmo da variável *income* ao conjunto de dados executa-se o comando **Acrescentar>Logaritmos das variáveis selecionadas**. Porém, note que antes de executar tal comando a variável para qual se deseja o logaritmo deve estar destacada (sombreada de azul claro) na janela principal do **gretl**, conforme Figura 3.5. Após executar esse comando a janela principal do **gretl** passará a mostrar a nova variável criada (*l_income*), Figura 3.6.

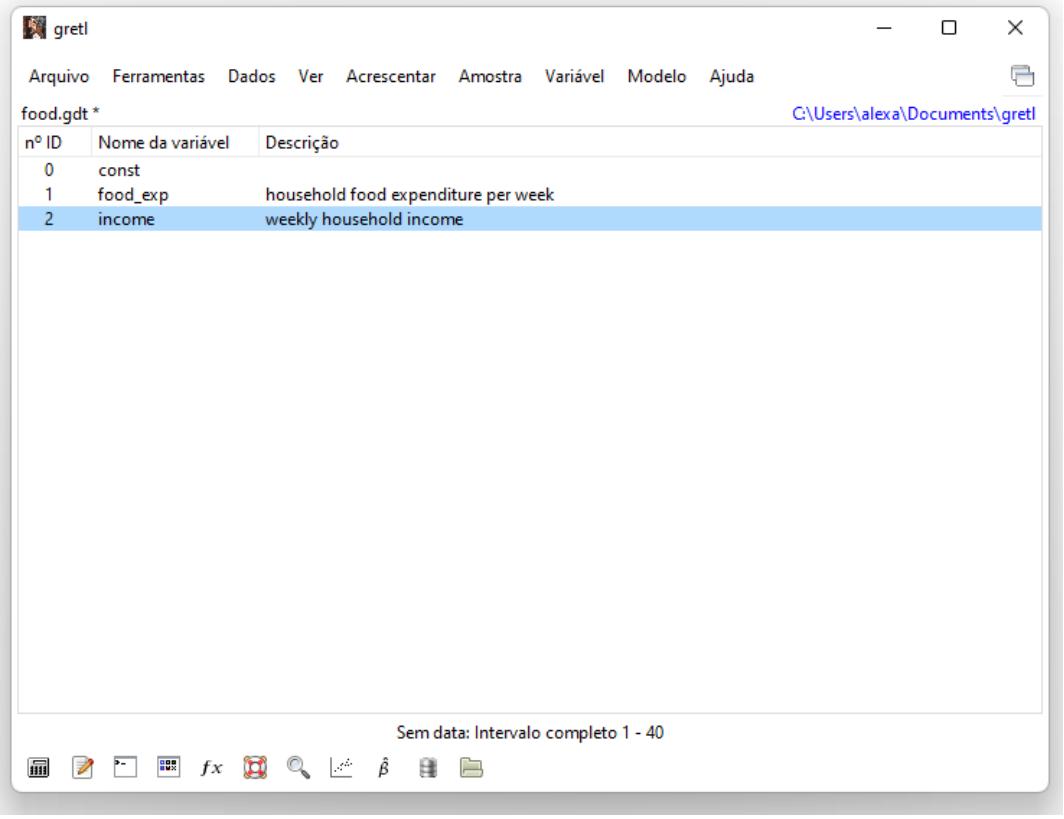


Figura 3.5: Selecionando a variável a ser transformada.

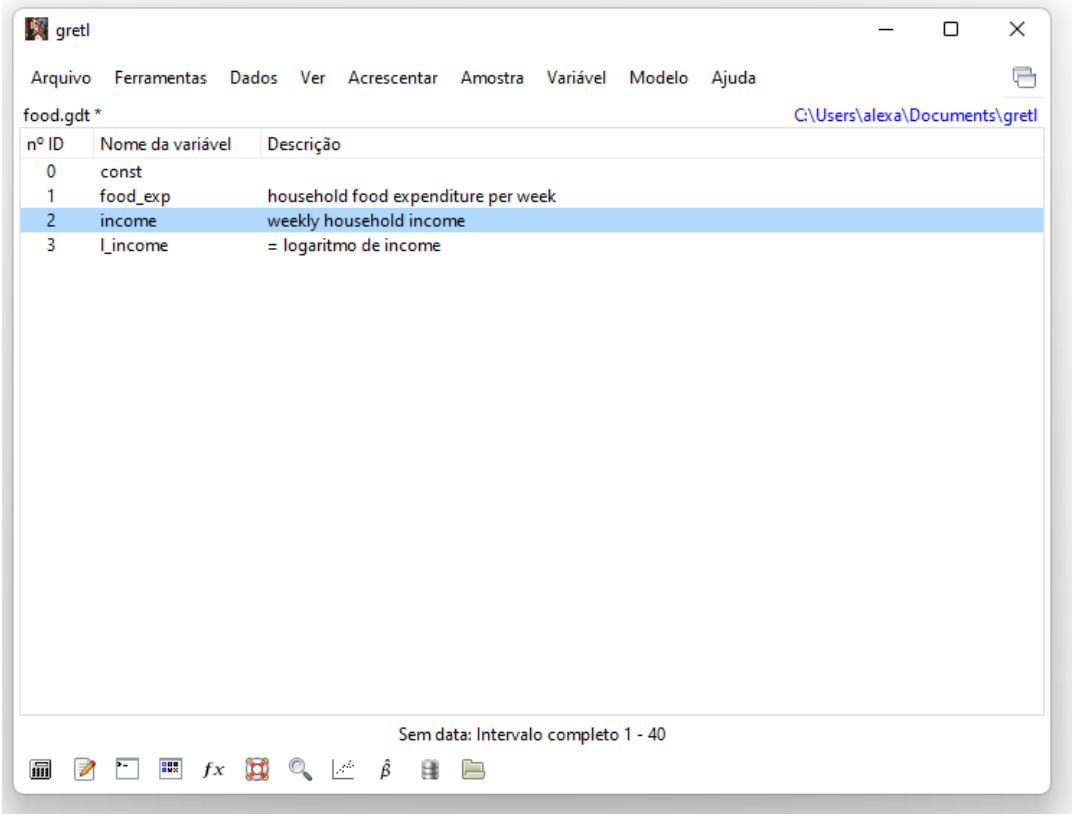


Figura 3.6: Janela principal com a nova variável.

Estimando o modelo produz

$$\widehat{food\_exp} = \underset{(84,2374)}{-97,1864} + \underset{(28,8046)}{132,166}\, l\_income$$

$$n = 40 \quad \bar{R}^2 = 0,3396 \quad F\,(1,\ 38) = 21,053 \quad \hat{\sigma} = 91,567$$

(erros padrão entre parênteses)

A seguir tem-se o gráfico de dispersão, Figura 3.7, da relação entre gastos com alimentação e renda. Uma vez que se estimou um modelo usando logaritmo neperiano (natural) da renda espera-se que uma relação positiva, i.e., não linear. Para gerar esse gráfico primeiramente estime a regressão para que seja aberta a janela de modelos. A seguir, execute o seguinte comando **Salvar>Valores ajustados**, Figura 3.8. Nomeie a variável valor ajustado como `yhat2` e clique em `Ok`.

Agora volte à janela principal e destaque (sombreado azul claro) as três variáveis ($food\_exp$, yhat2 e $income$) e, então, use o comando **Ver>Gráficos das variáveis>** X**-**Y **em dispersão**. Isso abrirá uma janela igual a da Figura 3.9. Escolha como `Variável do eixo` X $income$ e como `Variáveis do eixo` Y as variáveis $food\_exp$ e $yhat2$.

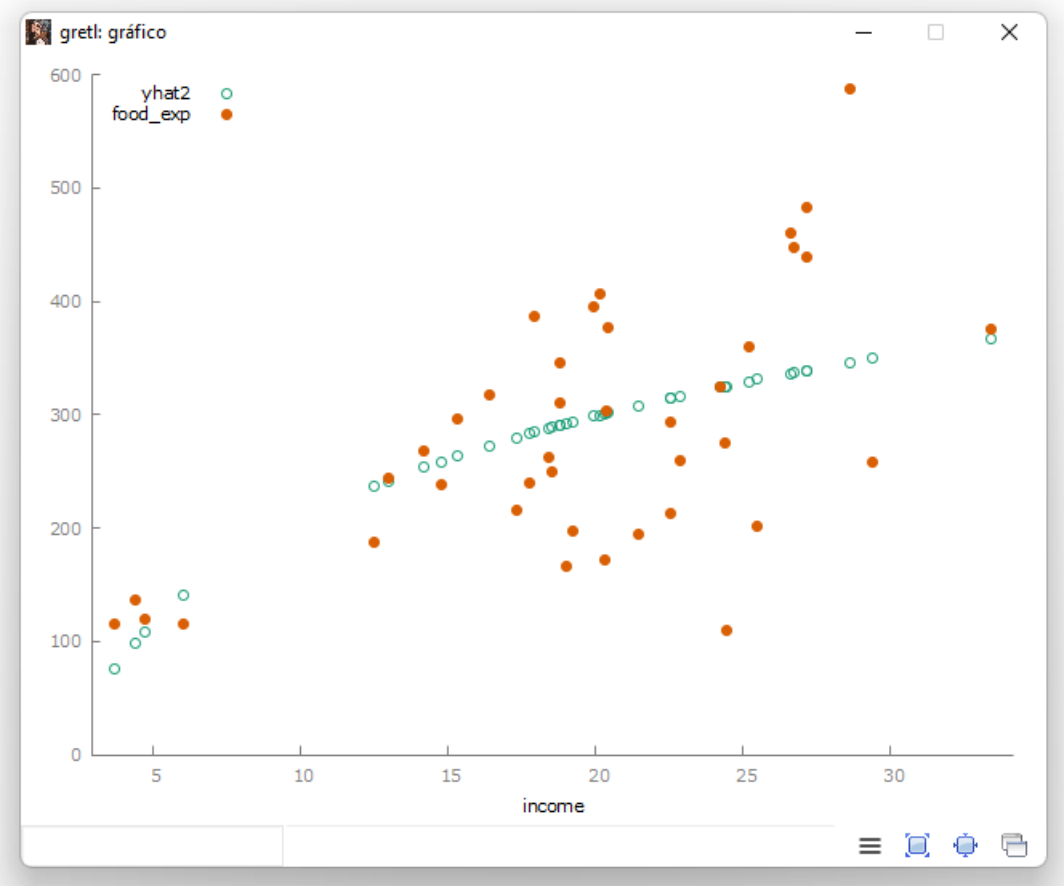


Figura 3.7: Menu suspenso para salvar os `Valores ajustados`.

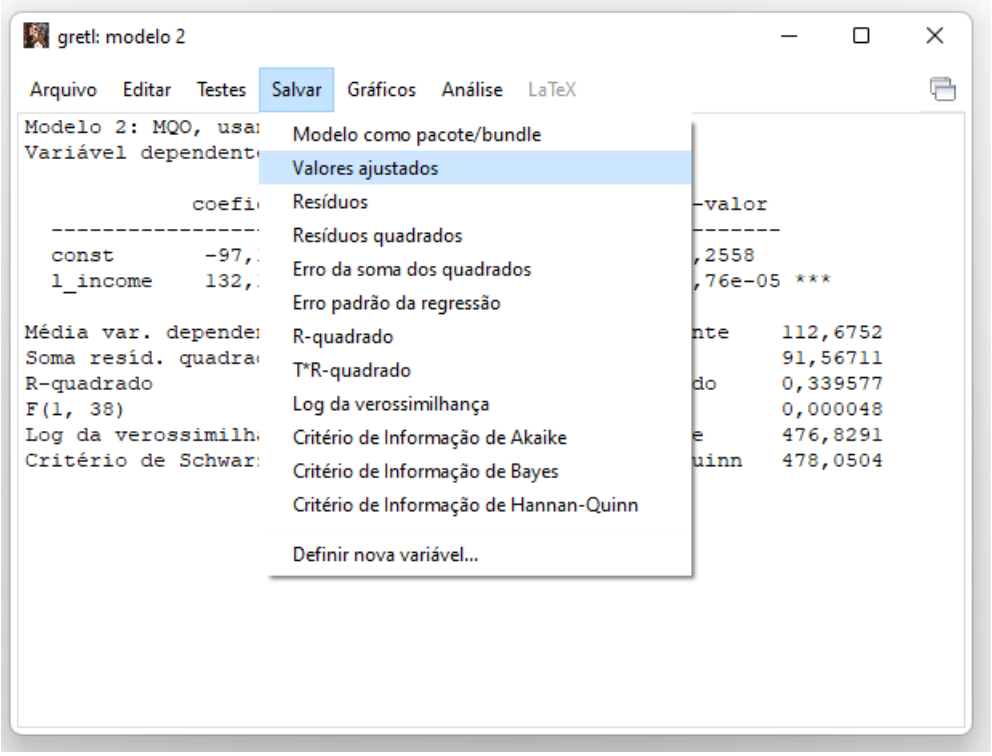


Figura 3.8: Menu para definir as variáveis.

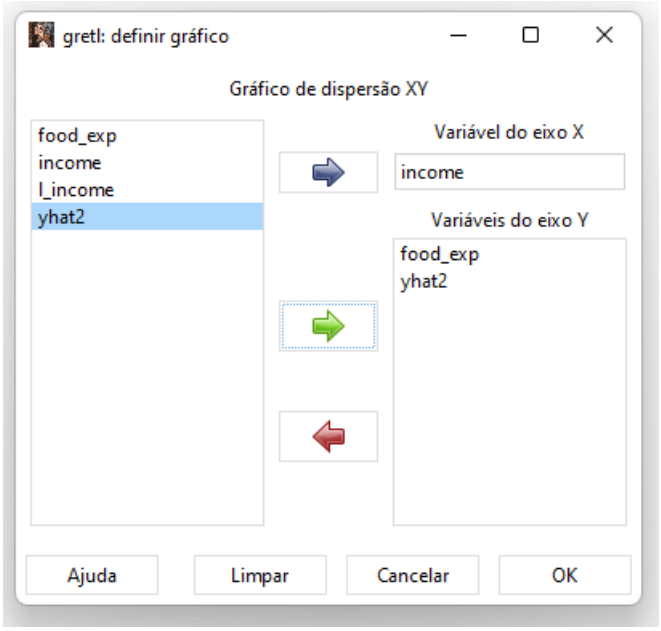


Figura 3.9: Gráfico de dispersão.

3.3.2 Teste para má especificação – gráfico dos resíduos

A tomada de decisões com base nos resultados de uma regressão pode levar a sérios problemas se a forma funcional estiver mal especificada. Por isso, após uma estimação deve-se realizar alguns testes estatísticos para confirmar a robustez dos resultados. Um dos primeiros teste a ser realizado é o diagnóstico de problemas de especificação. Destaca-se que existem diversos testes para identificar uma má especificação, entretanto, os pesquisadores geralmente começam examinando o gráfico dos resíduos da regressão em busca de evidências de qualquer erro de especificação.

Gráficos da distribuição dos resíduos de uma regressão semelhantes ao apresentado na Figura 3.10 garantem que as suposições do modelo de regressão linear normal clássico se mantêm e, assim, garantindo que os mínimos quadrados sejam a variância mínima não viesada.

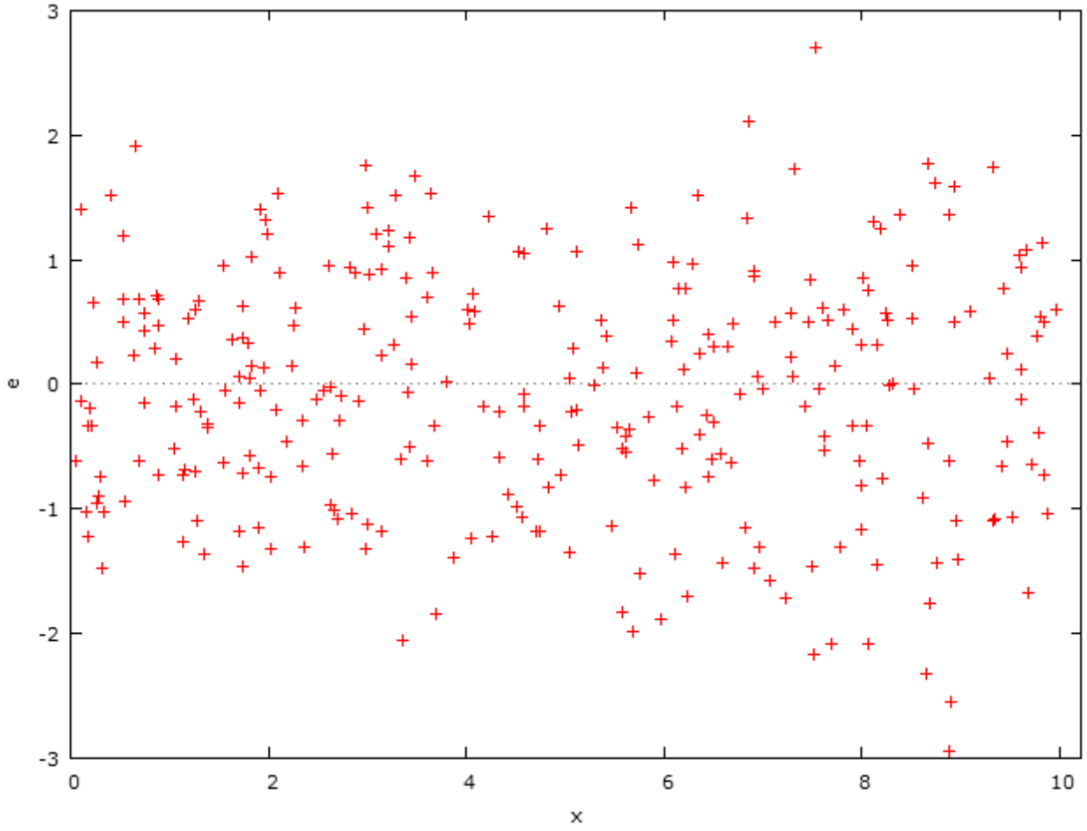


Figura 3.10: Resíduos distribuídos aleatoriamente.

Modelo linear-log

A Figura 3.11 refere-se ao gráfico dos resíduos de mínimos quadrados do modelo de regressão linear-log dos gastos com alimentação. Note que esses não parecem ser estritamente aleatórios, mas, pelo contrário, parecem ser heterocedásticos. Significando que para alguns níveis de renda o gasto com alimentação varia mais do que para outros níveis – nota-se que rendas mais altas a variação é maior.

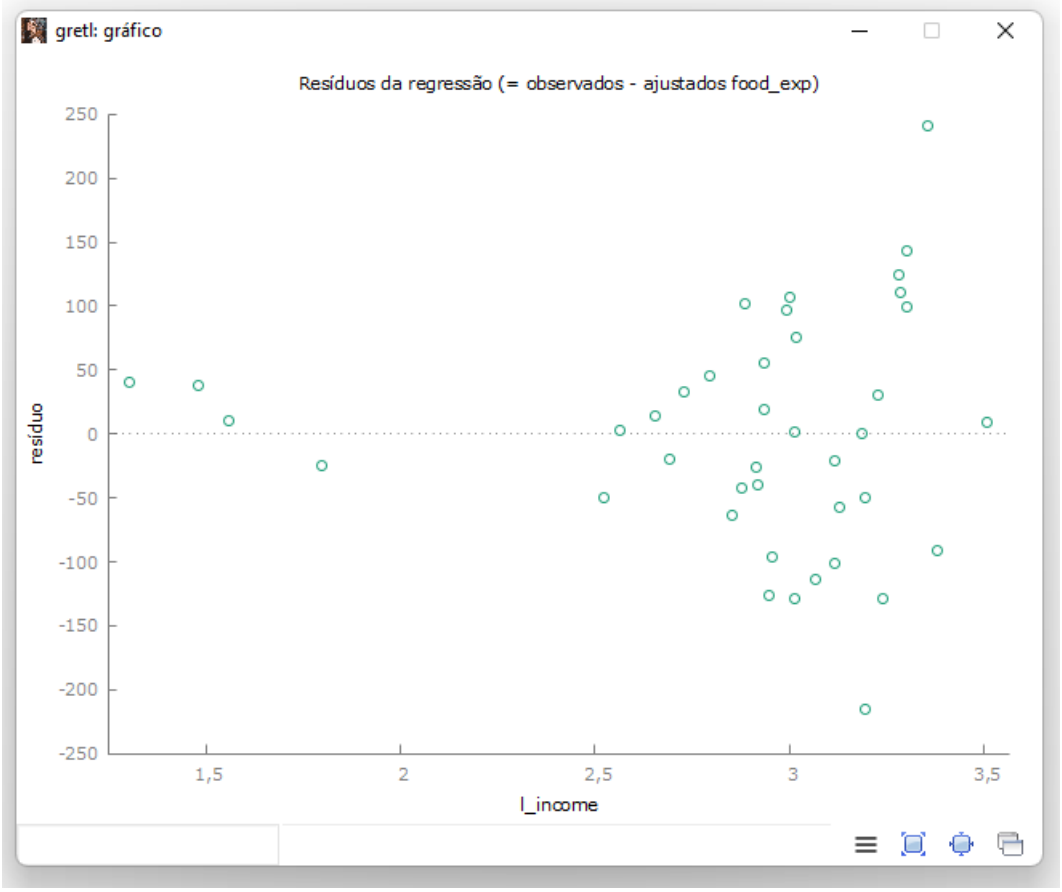


Figura 3.11: Distribuição dos resíduos do modelo linear-log.

Em função disso, os mínimos quadrados podem até ser imparciais nesse caso, porém, não é eficiente. Assim, a validade dos testes de hipóteses e intervalos é afetada e alguns cuidados devem ser tomados para garantir que sejam feitas inferências estatísticas adequadas.

Modelo log-linear

Agora, o modelo dos gastos com alimentação é estimado adotando a estrutura log-linear. Mais uma vez, os resíduos não apresentam uma distribuição aleatório, mas, pelo contrário, continuam sendo heterocedásticos. Porém, quando comparados ao modelo linear-log pode-se dizer que são levemente heterocedástico, Figura 3.12.

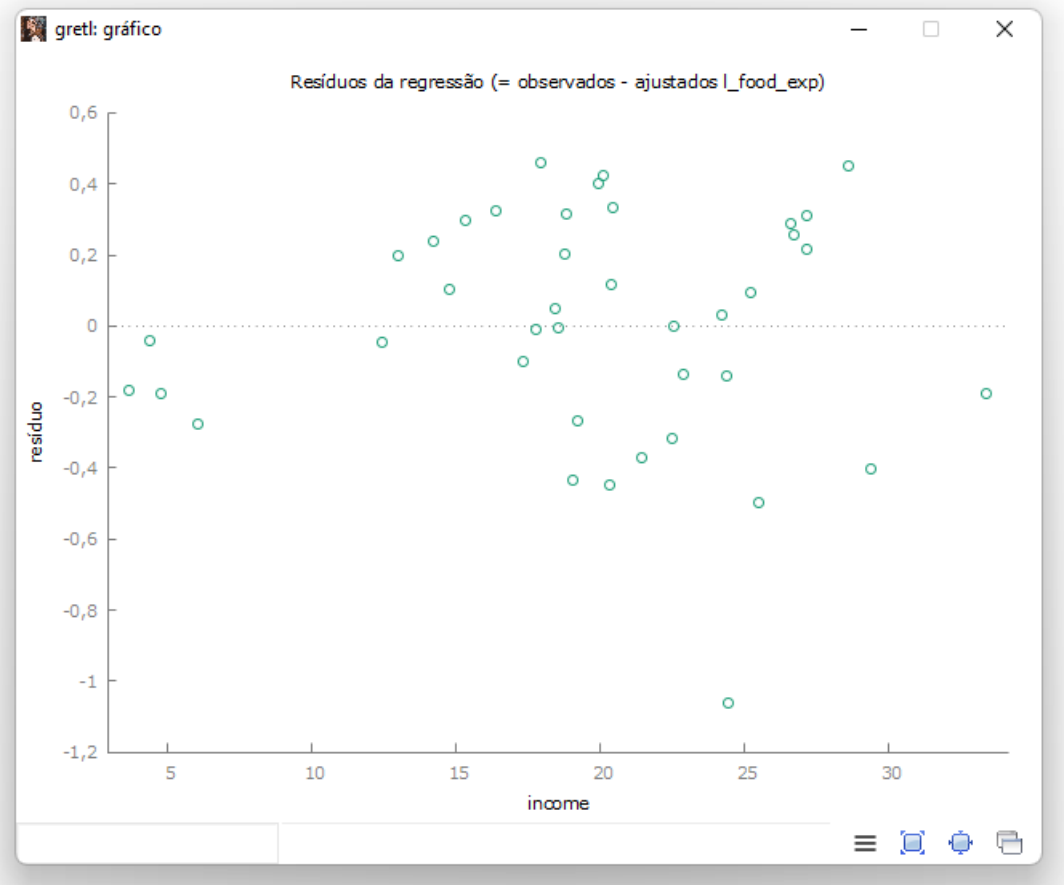


Figura 3.12: Distribuição dos resíduos do modelo log-linear.

3.3.3 Teste de normalidade

O teste de normalidade Jarque-Bera – JB – é calculado usando a assimetria e a curtose dos resíduos de mínimos quadrados. Primeiramente, é necessário estimar o modelo usando Mínimos Quadrados Ordinários e salvar os resíduos no conjunto de dados. Assim, para o modelo de gastos com alimentação, após a estimação salva-se os resíduos aplicando o comando **Salvar>Resíduos**, Figura 3.13.

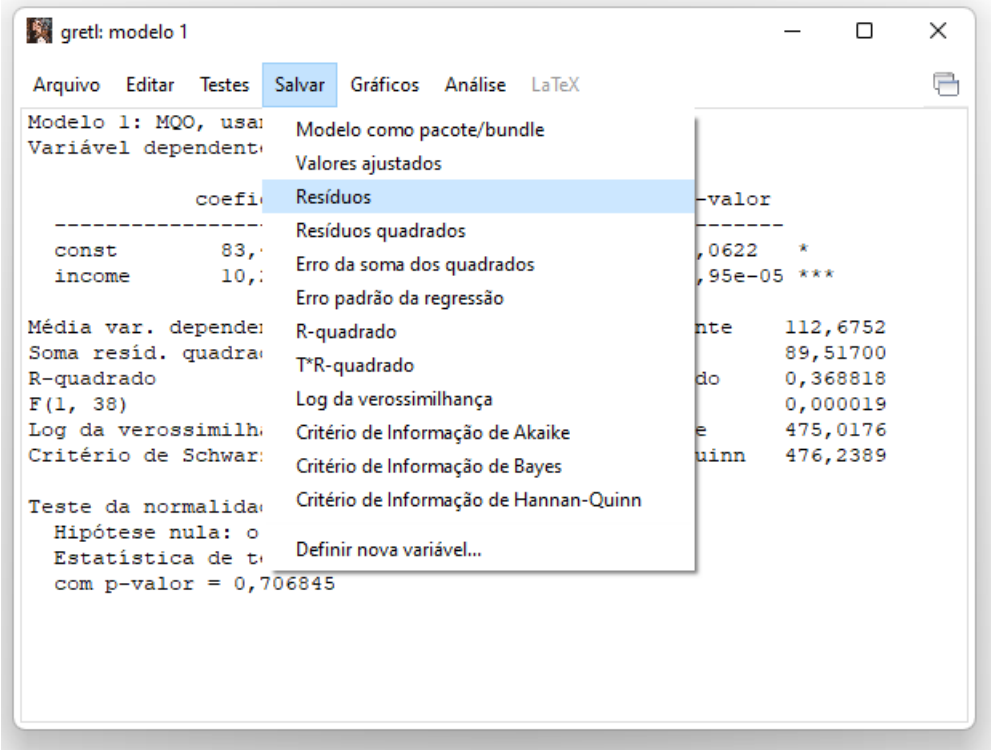


Figura 3.13: Salvando os resíduos.

Importante mencionar que o **gretl** reporta o excesso de curtose em vez da curtose e, assim, o cálculo é dado por:

$$JB = \frac{T}{6}\left(\text{assimetria}^2 + \frac{(\text{excesso de curtose})^2}{4}\right) \tag{3.6}$$

Variáveis aleatoriamente normalmente distribuídas não possuem nem assimetria nem curtose e, portanto, a estatística JB é igual a zero. Entretanto, essa estatística fica maior quanto maior a assimetria e quanto maior o grau de excesso de curtose exibido pelos dados. Agora, uma vez salvado os resíduos no conjunto de dados, usa-se a janela de comandos para realizar o cálculo da estatística Jarque-Bera. Para acessar a janela de comandos, clique no terceiro ícone da esquerda, , na parte inferior da janela principal do **gretl**. Na janela que abrir, nomeada de `console`, digite o comando `normtest uhat1 --jbera`, Figura 3.14.

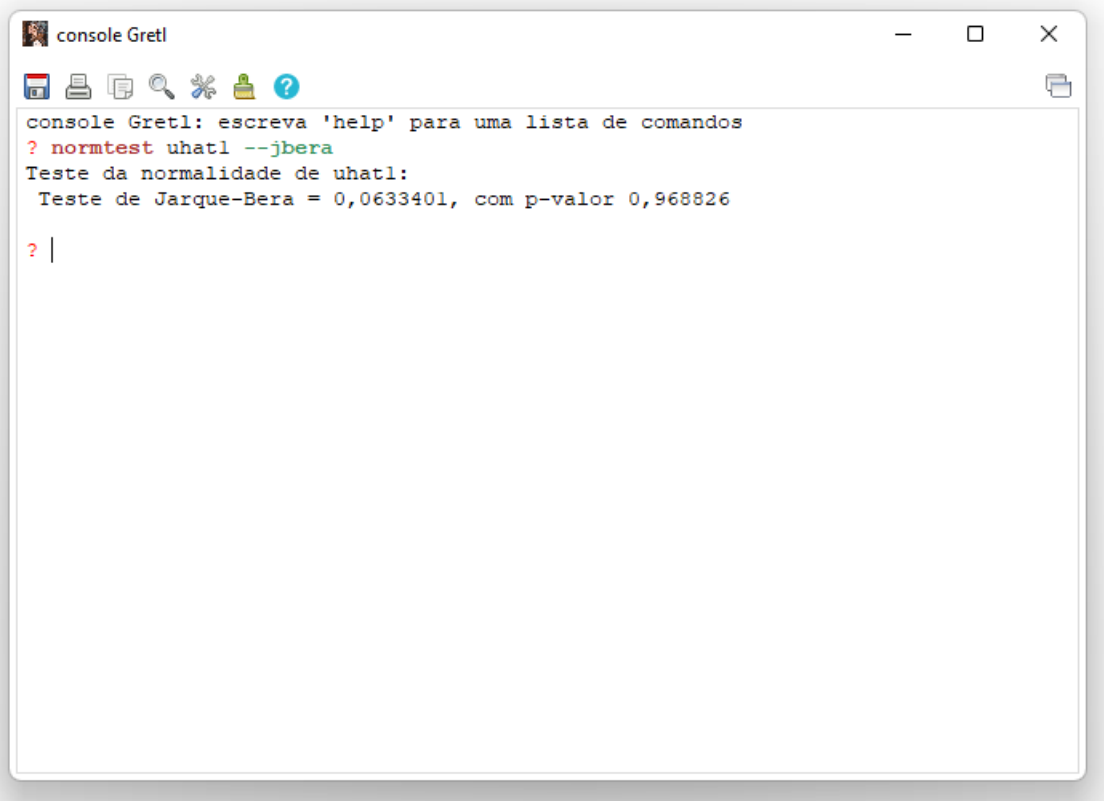


Figura 3.14: Saída do teste Jarque-Bera.

Outros testes para a normalidade dos resíduos podem ser obtidos digitando na janela `console` do **gretl** o seguinte comando: `normtest uhat1 --all`. Um dos testes reportados é o teste de Doornik-Hansen – DH – que é computacionalmente mais complexo que o teste de Jarque-Bera. Ademais, para plotar um gráfico básico da distribuição dos resíduos pode-se executar o comando **Testes>Normalidade dos resíduos** na janela da regressão do modelo, Figura 3.15. Uma vantagem de se usar o `normtest` é que se pode testar a normalidade para qualquer variável, não apenas dos resíduos.

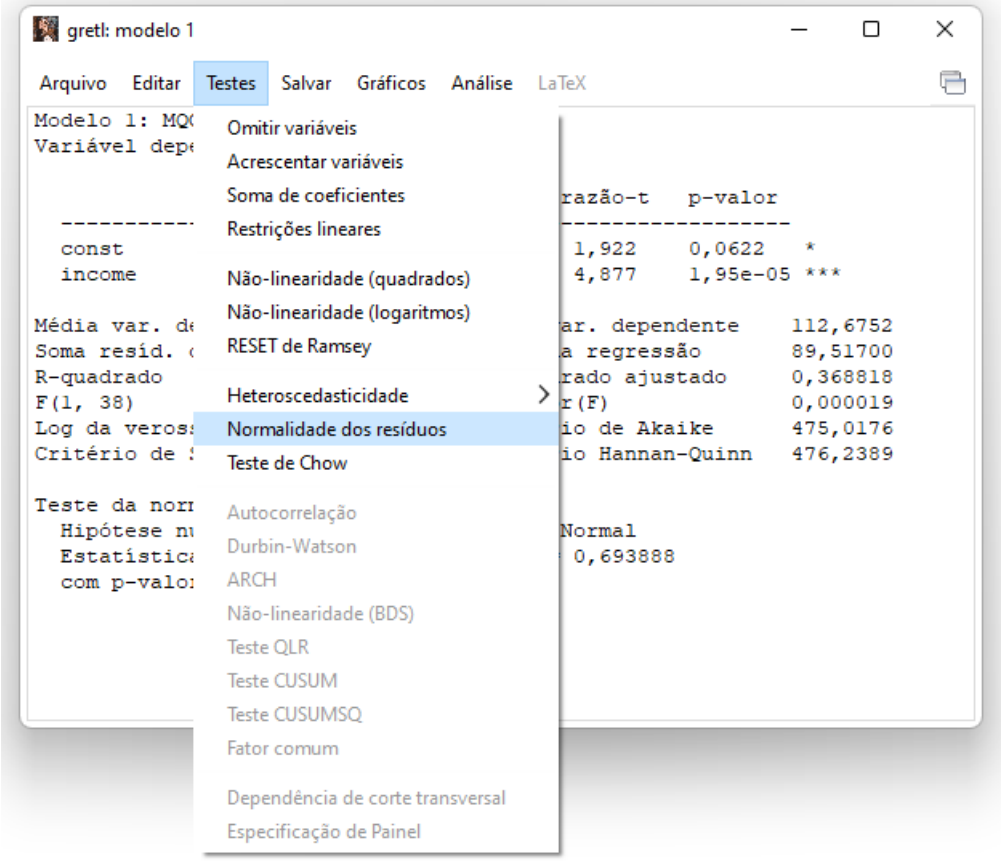


Figura 3.15: Teste DH de normalidade dos resíduos.

Um histograma dos resíduos é gerado com uma densidade normal sobreposta à distribuição dos resíduos, Figura 3.16.

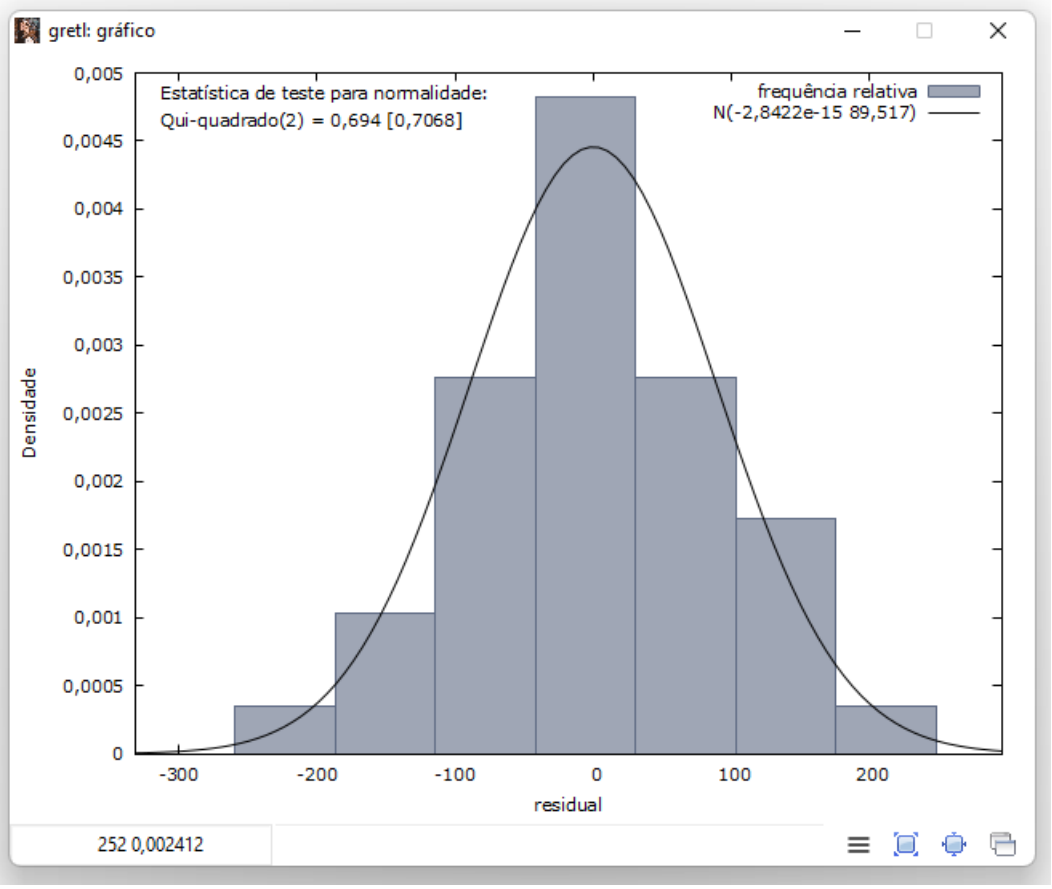


Figura 3.16: Histograma da distribuição dos resíduos.

Capítulo 4

Modelo de regressão múltipla

O modelo de regressão múltipla é uma extensão do modelo de regressão simples. A principal diferença é que o modelo linear de regressão múltipla contém mais do que uma variável explicativa. Essa condição muda ligeiramente a interpretação dos coeficientes e impõe uma condição especial aos dados. A forma geral do modelo é mostrada na Equação 4.1 abaixo:

$$y_i = \beta_1 + \beta_2\, x_{i2} + \cdots + \beta_k\, x_{ik} + e_i \qquad i = 1,\, 2,\, \ldots,\, n \tag{4.1}$$

em que y_i é variável dependente, x_{ij} é a i^{th} observação da j^{th} variável independente, $j = 2,\, 3,\, \ldots,\, k$; e_i é o erro aleatório e $\beta_1,\, \beta_2,\, \ldots,\, \beta_k$ são os parâmetros que se deseja estimar. Assim, como o modelo de regressão linear simples, cada erro $e_i \mid x_{ij}$ tem um valor zero para cada valor das j's variáveis independentes. Cada variável possui a mesma variável σ^2 e são correlacionados com qualquer um dos outros termos de erros.

Para estimar cada um dos β_s, nenhuma das variáveis independentes pode ser exatamente uma combinação linear das demais variáveis independentes. Essa condição serve como um requisito para que a variável independente assuma pelo menos dois valores diferentes na amostra. As suposições sobre o termo de erro podem ser resumidas como: $e_i \mid x_{i2},\, x_{i3},\, \ldots,\, x_{ik}\ i.i.d\ (0,\, \sigma^2)$. Lembre-se que a expressão $i.i.d$ significa que os erros são estatisticamente independentes uns dos outros (e, portanto, não são correlacionados) e cada um dos resíduos tem a mesma distribuição de probabilidade.

Os parâmetros $\beta_1,\, \beta_2,\, \ldots,\, \beta_k$ são considerados como inclinações e cada inclinação mede o efeito de a mudança de uma unidade de x_{ij} na média do valor de y_i, mantendo todas as outras variáveis na equação constantes. A interpretação condicional do coeficiente é importante para lembrar quando se utiliza a regressão linear múltipla.

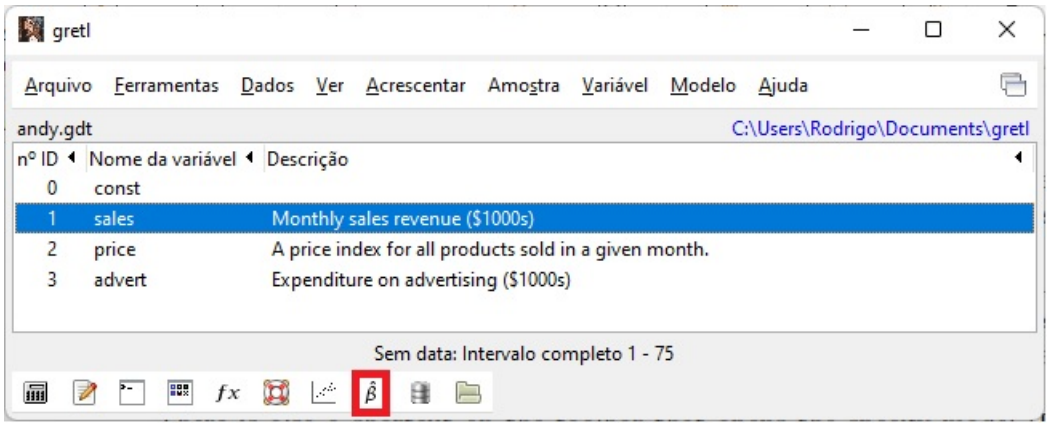
O primeiro exemplo usado é o modelo de vendas do Big Andy's Burger Barn. O modelo inclui duas variáveis explicativas e uma constante:

$$sales_i = \beta_1 + \beta_2\, price_i + \beta_3\, advert_i + e_i \qquad i = 1,\, 2, \cdots,\, n \tag{4.2}$$

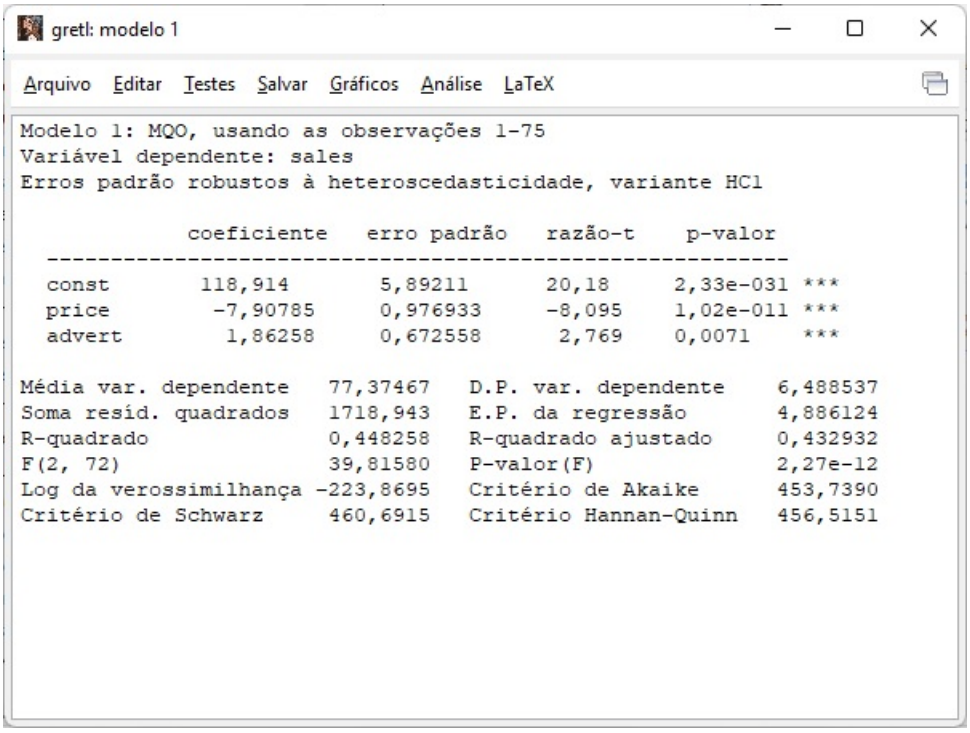
em que $sales_i$ são as vendas mensais em uma dada cidade sendo medida em \$1.000 incrementos, $price_i$ é o preço do hambúrguer medido em dólares e $advert_i$ são os gastos em propaganda também medidas em milhares de dólares.

4.1 Regressão linear

Para estimar-se a regressão linear múltipla, deve-se clicar em **Modelo>Mínimos Quadrados Ordinários**. Também há um atalho na barra de ferramentas que abre o modelo a ser especificado. Lembre que a barra de ferramentas está localizada na parte inferior da janela principal do **gretl**. Lá encontra-se um botão rotulado como $\hat{\beta}$:



Clicando no botão $\hat{\beta}$ pode-se especificar o modelo, obtendo os seguintes resultados.



4.2 Qualidade do ajuste

Uma importante estatística incluída na saída do **modelo 1** é a Soma dos Quadrados dos Resíduos (SQR) a qual o **gretl** se refere como `Soma dos quadrados resíduo`. Nesse modelo o $SQR = 1718,943$. Para obter a variância estimada, $\hat{\sigma}^2$, dividi-se a SQR pelos graus de liberdade disponíveis para obter:

$$\hat{\sigma}^2 = \frac{\text{SQR}}{n-k} = \frac{1718,94}{75-3} = 23,873 \tag{4.3}$$

em que n corresponde ao número de observações e k é o grau de liberdade.

A raiz quadrada desse número é $4,88612$ que é referida pelo **gretl** como `E.P da regressão` (Erro Padrão da Regressão). Se o economista empírico deseja computar suas próprias versões dessas estatísticas usando a soma dos quadrados do modelo, poderá utilizar o menu gerado pela própria janela do modelo **Análise>ANOVA**. Para computar o R^2 mostrado na saída padrão do **gretl** deve-se lembrar que:

$$\hat{\sigma}_y = \sqrt{\frac{\text{STQ}}{n-1}} \tag{4.4}$$

em que STQ é a Soma Total dos Quadrados e n o número de observações.

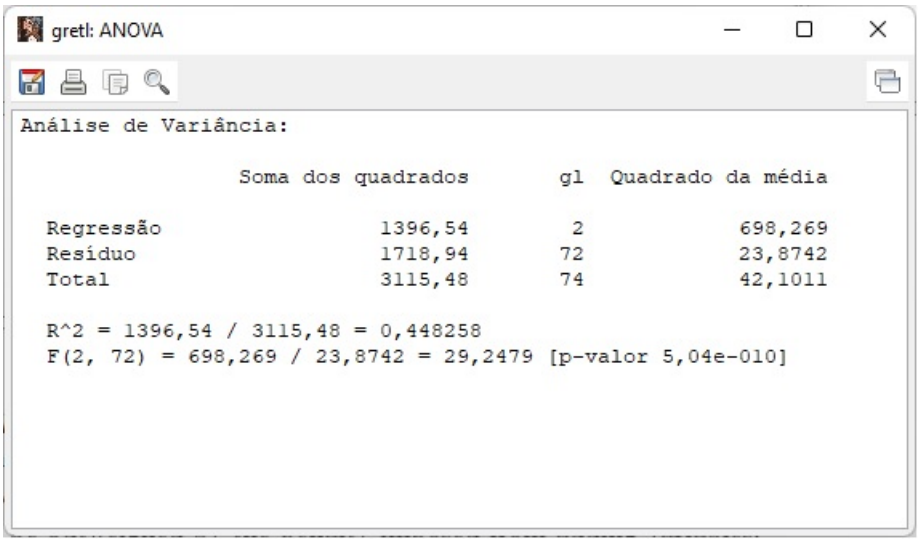
A esteatítica $\hat{\sigma}_y$ é mostrada pelo **grelt** como `D.P da var. dependente` que é $6,48854$. Com um pouco de álgebra tem-se que:

$$\text{STQ} = (n-1)\hat{\sigma}_y = 74 \times 6,48854 = 3115,785 \tag{4.5}$$

em que STQ é a Soma Total dos Quadrados e n o número de observações. Então:

$$R^2 = 1 - \frac{\text{SQE}}{\text{STQ}} = 1 - \frac{1718,94}{3115,485} = 0,448 \tag{4.6}$$

em que SQE é a Soma dos Quadrados Explicados e STQ a Soma Total dos Quadrados. Dessa forma, as estatísticas de qualidade de ajuste impressas na saída da regressão **gretl** ou na tabela `ANOVA` são perfeitamente aceitáveis.



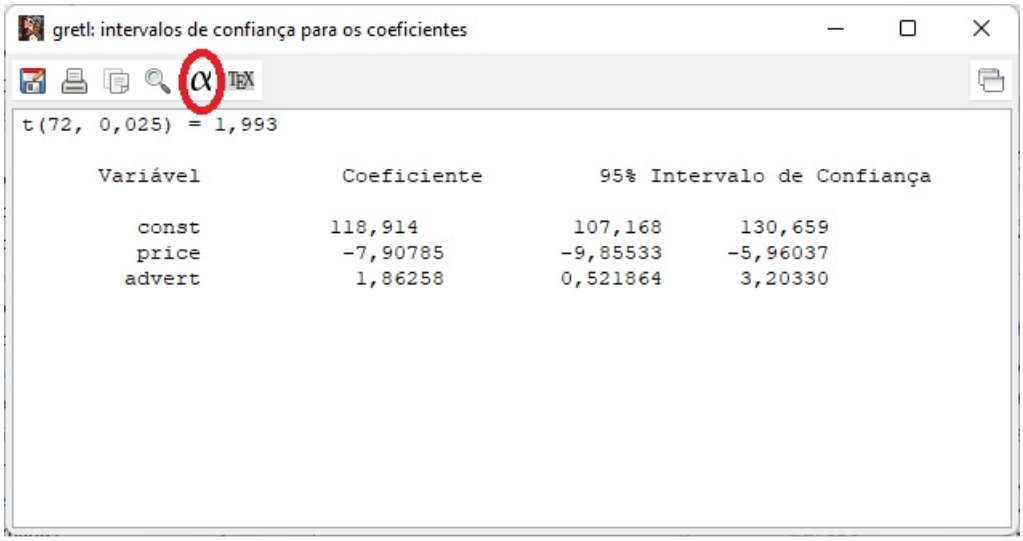
O **gretl** também reporta o R^2 – ajustado na saída padrão da regressão. O R^2 – ajustado impõe uma pequena penalização para o R^2 padrão quando uma nova variável é inserida no modelo. Adicionando uma nova variável qualquer a correlação com y sempre reduz a SQE e aumenta o tamanho do R^2. Por sua vez, o R^2 – ajustado pode se tornar menor à medida que novas variáveis são adicionadas. A fórmula é:

$$\bar{R}^2 = 1 - \frac{SQE\,(n-k)}{SQT\,(n-1)} = 1 \tag{4.7}$$

O **gretl** refere-se a essa medida como R-quadrado ajustado. Para o exemplo do Big Andy's Burger Barn o R^2 – ajustado é igual a $0,4329$.

4.3 Intervalos de confiança

Os intervalos de confiança pode ser obtidos usando o menu **Análise>Intervalos de confiança para os coeficientes**.



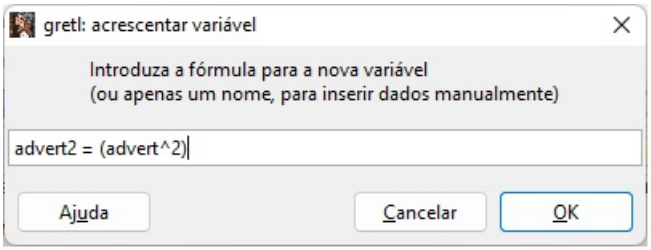
Clicando em α pode-se selecionar o nível de confiança desejado.

4.4 Polinômios

Uma forma de permitir um relacionamento não linear entre a variável dependente e a independente é introduzir polinômios ao modelo de regressão. No exemplo espera-se que o efeito marginal de um dólar adicional investido em propaganda reduza ao aumentar os gastos em propaganda.

$$sales_i = \beta_1 + \beta_2\, price_i + \beta_3\, advert_i + \beta_4\, advert_i^2 + e_i \qquad i = 1,\, 2,\, \ldots,\, n \tag{4.8}$$

Para poder estimar os parâmetros desse modelo, deve-se criar uma nova variável $advert_i^2$ e adicioná-la ao modelo de mínimos quadrados. Para isso basta clicar no menu **Acrescentar>Definir nova variável**.



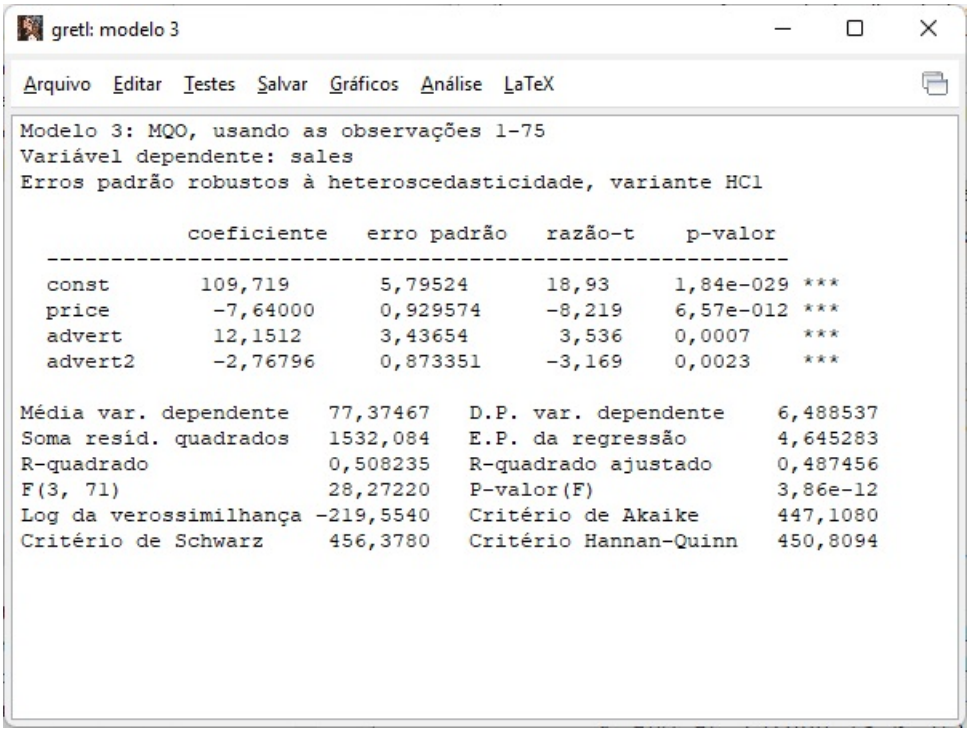
A criação dessa variável *advert*2 é um exemplo simples do que pode ser chamado de variável de interação. A forma mais simples de pensar sobre uma variável de interação é que a magnitude de seu efeito sobre a variável dependente depende de outra variável, ou seja, as duas variáveis interagem para determinar o valor médio da variável dependente. Neste exemplo, o efeito da publicidade nas vendas médias depende do nível da própria publicidade.

4.5 Efeitos marginais

Quando as variáveis interagem o efeito marginal de uma variável na média de outra deve ser computado baseando-se em cálculo. Ao tomar a derivada parcial das vendas médias em relação ao nível de propaganda obtém-se o efeito marginal médio das vendas sobre o aumento de uma unidade na propaganda:

$$\frac{\partial E\,(sales)}{\partial\, advert} = \beta_3 + 2\,\beta_4 \tag{4.9}$$

A magnitude do efeito marginal depende dos parâmetros bem como do nível de propaganda. Veja os resultados das estimativas para poder calcular o efeito marginal:



O efeito marginal de um acrescimento de $ 1.000 dólares em propaganda pode ser calculado da seguinte forma:

$$\beta_3 + 2\,\beta_4 = 12,15 + 2 \times (-2,76) \times 1 = 6,63$$

4.6 Efeitos de interação

Nesse exemplo fez-se a interação entre a variável experiência e a variável salário. Para isso, utiliza-se o arquivo `cps5_small.gdt`. A ideia é que o nível de experiência afeta o retorno de um ano a mais de escolaridade (ou, outro ano de educação afeta o retorno de um ano a mais de experiência). O modelo a ser estimado se torna:

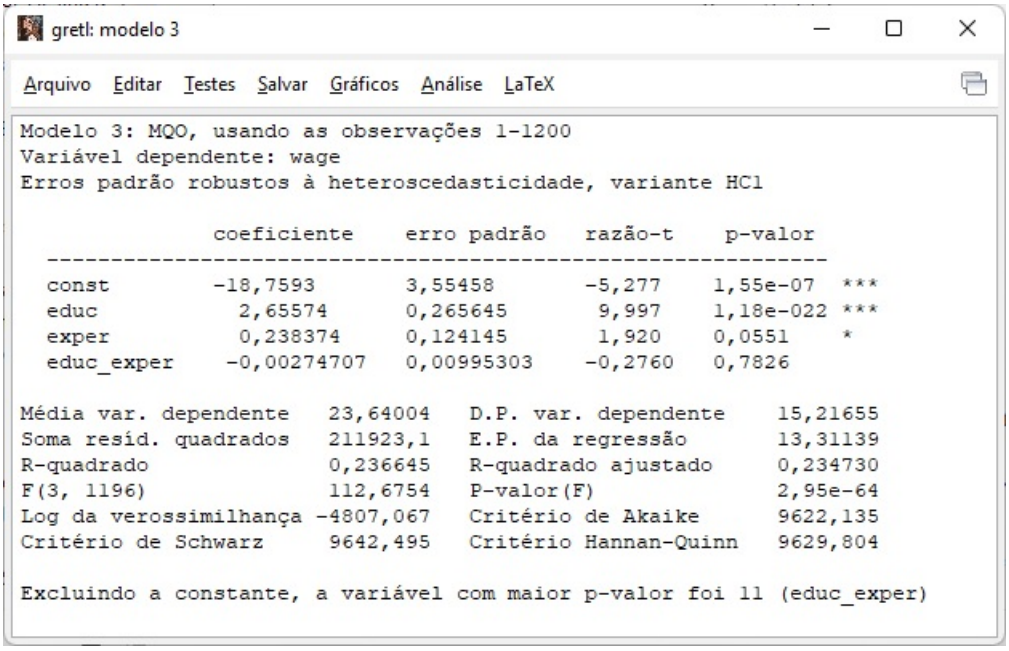
$$wage = \beta_1 + \beta_2\, educ + \beta_3\, exper + \beta_4\, educ \times exper + e \tag{4.10}$$

O efeito marginal depende dos níveis de educação e da experiência. Eles são medidos pelos trabalhadores que possuem entre 8 e 16 anos de escolaridade e para aqueles trabalhadores que possuem 20 anos de experiência:

$$\frac{\partial\, E\,(wage \mid educ,\, exper)}{\partial\, exper} = \beta_1 + \beta_4\, educ \tag{4.11}$$

$$\frac{\partial\, E\,(wage \mid educ,\, exper)}{\partial\, educ} = \beta_1 + \beta_4\, exper \tag{4.12}$$

Abaixo seguem as estimativas do modelo:



Os efeitos marginais da experiência são os seguintes:

Quando a experiência é $0 = 2,65$

$$\text{Quando a experiência é } 20 = 2,65 - (0,00275) \times 20 \cong 2,6$$

Os efeitos marginais da educação:

$$\text{Quando a educação é } 8 = 0,24 - (0,00275) \times 8 \cong 2,18$$

$$\text{Quando a educação é } 16 = 0,24 - (0,00275) \times 8 \cong 0,196$$

$$\text{Quando a educação é } 20 = 0,24 - (0,00275) \times 20 \cong 0,185$$

Pode-se expandir esse exemplo utilizando um termo quadrático:

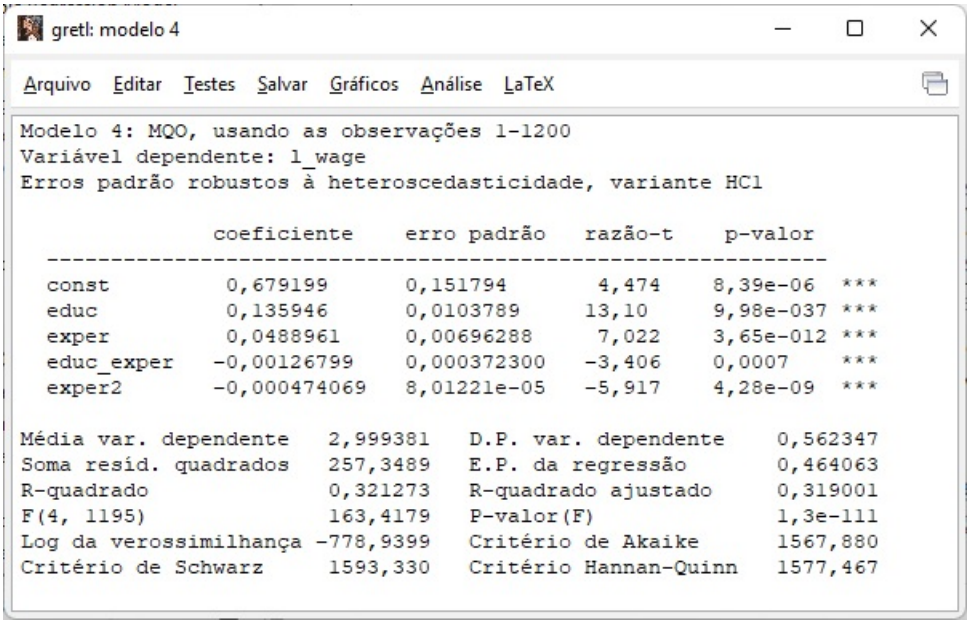
$$\ln(wage) = \beta_1 + \beta_2\, educ + \beta_3\, exper + \beta_4\, educ \times exper + \beta_5\, exper^2 + e \tag{4.13}$$

Os efeitos marginais são:

$$\frac{\partial\, E(\ln(wage) \mid educ,\, exper)}{\partial\, exper} = \beta_1 + \beta_4\, educ + 2\beta_5\, exper \tag{4.14}$$

$$\frac{\partial\, E(\ln(wage) \mid educ,\, exper)}{\partial\, educ} = \beta_1 + \beta_4\, exper \tag{4.15}$$

As estimativas do modelo podem ser vistas na figura abaixo:



Efeitos marginais da experiência no salário de 8 anos de educação e 20 anos de experiência:

$$0,05 + (-0,00127 \times 8) + (2 \times -0.0005 \times 20) = 0,0198 = 1,98\%$$

Efeitos marginais da educação no salário de 8 anos de educação e 20 anos de experiência:

$$0,136 + (-0,00127 \times 20) = 0,116 = 11,6\%$$

Capítulo 5

Inferência adicional no modelo de regressão múltipla

Neste capítulo aprofunda-se a análise dos modelos de regressão linear múltipla apresentando novas estatísticas auxiliares para checar a qualidade do ajuste do modelo. Primeiramente testa-se a hipóteses conjuntas sobre os parâmetros em um modelo e, a seguir, aprende-se a como impor restrições lineares aos parâmetros. Ademais, a especificação do modelo será determinada usando regras de seleção do modelo, previsão fora da amostra e um teste formal funcional. A colinearidade e a detecção de *outliers* – observações influentes – são discutidas e os mínimos quadrados não lineares são apresentados.

5.1 Teste F

A estatística t associada a qualquer coeficiente de MQO pode ser usada para testar se o parâmetro desconhecido correspondente na população é igual a qualquer constante dada, geralmente, mas nem sempre, zero – $\beta_k = 0$. Observe que essa hipótese envolve uma *única* restrição. No entanto, frequentemente, deseja-se testar hipóteses *múltiplas* sobre os parâmetros subjacentes β_0, β_1, ..., β_k. Logo, inicia-se com o procedimento principal de testar se um conjunto de variáveis independentes não tem efeito parcial sobre uma variável dependente.

5.1.1 Teste de restrições de exclusão

Sabe-se como testar se uma variável determinada não tem efeito parcial sobre a variável dependente: use a estatística t. Agora, o que se quer é testar se um *grupo* de variáveis não tem efeito sobre a variável dependente. Mais precisamente, a hipótese nula é que um conjunto de variáveis não tem efeito sobre y, já que outro conjunto de variáveis foi controlado.

Como uma ilustração do porquê testar a significância de um grupo é útil, considere o seguinte modelo do Big Andy's Burger Ban (conjunto de dados `andy.gdt`):

$$sales = \beta_1 + \beta_2\, price + \beta_3\, advert + \beta_4\, advert^2 + e \tag{5.1}$$

Suponha que se deseja testar a hipótese de que a propaganda ($advert$) não tem efeito sobre as vendas médias ($sales$) contra a hipótese alternativa de que tem. Assim,

tem-se que:

$$\begin{cases} H_0 : \beta_3 = \beta_4 = 0 \\ H_1 : \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0 \end{cases} \tag{5.2}$$

O modelo sob H_0 é restrito em comparação com o modelo sob H_1, pois nele $\beta_3 = 0$ e $\beta_4 = 0$. Ou seja, a hipótese nula constitui duas **restrições de exclusão**: se H_0 é verdadeiro, então, *advert* e $advert^2$ não têm efeito sobre *sales* após *price* ter sido controlado e, portanto, deveriam ser excluídos do modelo. Esse é um exemplo de conjunto de **restrições múltiplas** porque são colocadas mais de uma restrição sobre os parâmetros do Modelo 5.1; posteriormente, serão vistos mais exemplos gerais de restrições múltiplas. Um teste de restrições múltiplas é chamado **teste de hipóteses múltiplas** ou o **teste de hipóteses conjuntas**.

A estatística F usada para testar H_0 contra H_1 estima cada modelo por mínimos quadrados e compara sua respectiva soma de erros quadrados usando a estatística:

$$F = \frac{(SQR_r - SQR_{ir}) \,/\, J}{SQR_{ir} \,/\, (n-k)} \sim F_{J,\,n-k} \quad \text{se } H_0 \text{ é verdadeiro} \tag{5.3}$$

em que SQR_r é a Soma dos Quadrados dos Resíduos do modelo restrito enquanto SQR_{ir} caracteriza-se como sendo a Soma dos Quadrados dos Resíduos do modelo irrestrito. Por sua vez, J indica o número de hipóteses sendo testadas, no presente exemplo duas ($\beta_3 = 0$ e $\beta_4 = 0$). Já o denominador é dividido pelo número total de graus de liberdade na regressão irrestrita, $n - k$, em que n é o tamanho da amostra e k é o número de parâmetros na regressão irrestrita.

A seguir são apresentados os passos para calcular a estatística F no **gretl** usando o Modelo 5.1. Assim, inicialmente cria-se a variável $advert^2$ conforme a Figura 5.1.

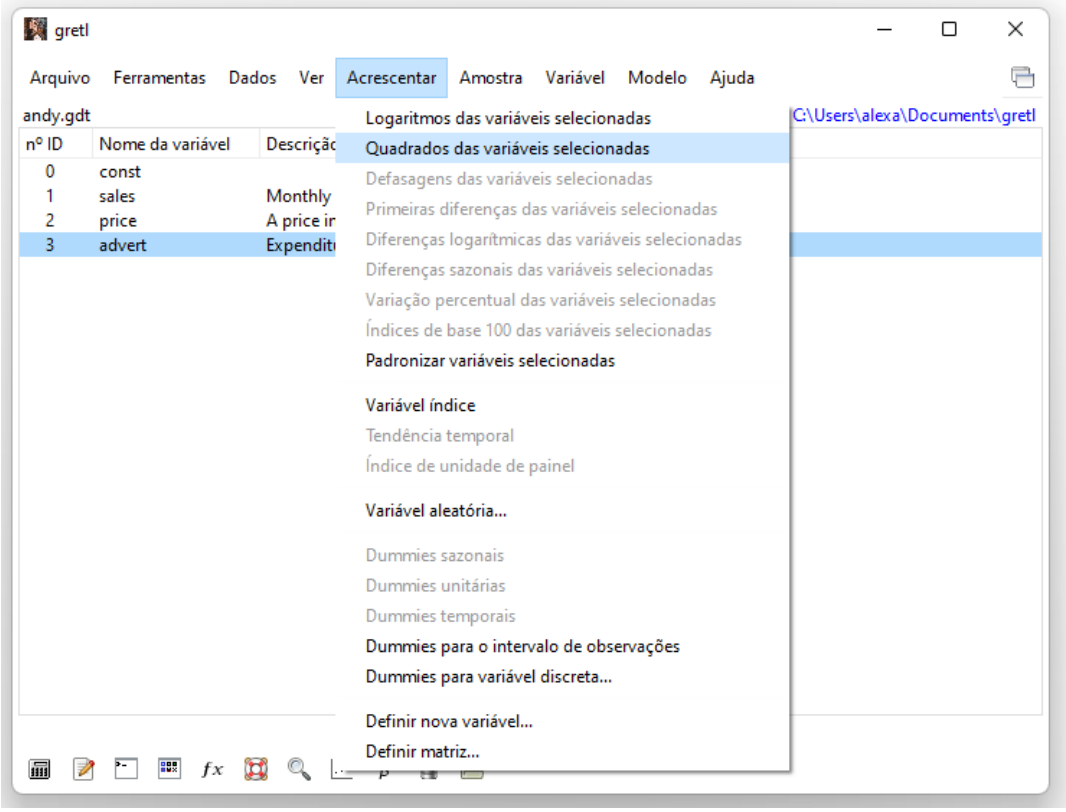


Figura 5.1: Caixa de diálogo para adicionar uma nova variável.

Uma vez criada essa variável a janela principal do **gretl** terá a seguinte aparência (Figura 5.2):

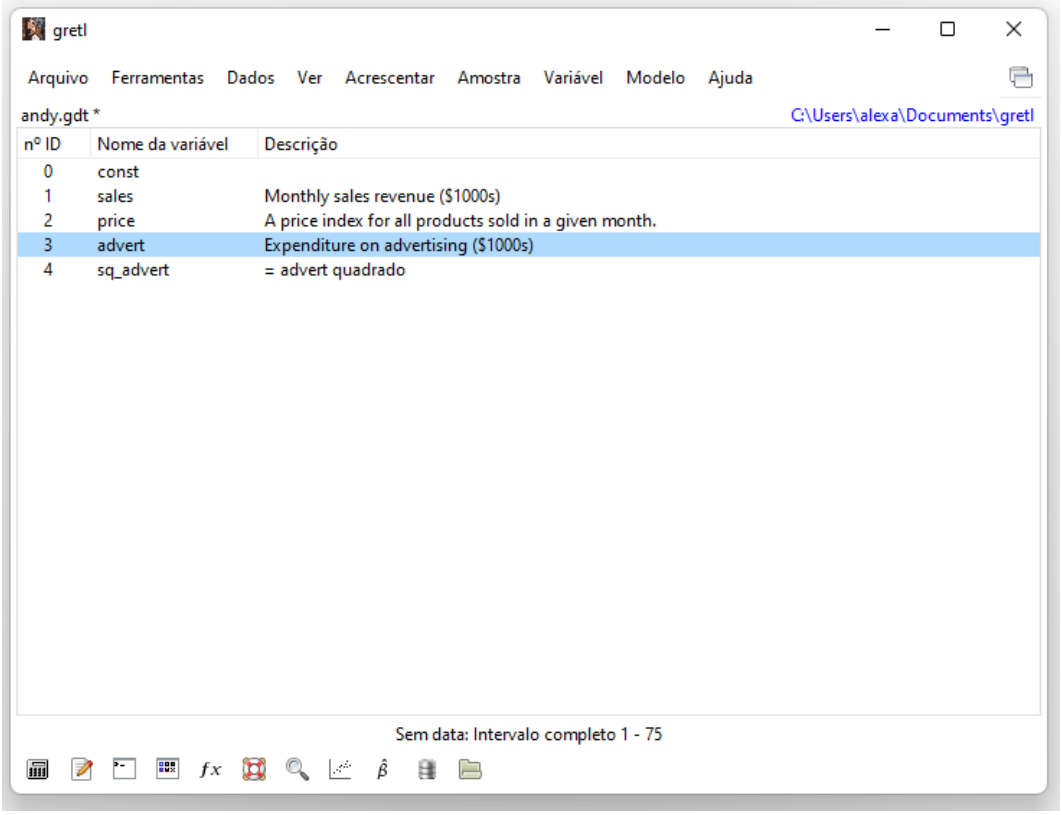


Figura 5.2: Janela principal do **gretl**.

Após definir a especificação a ser estimada, conforme a Figura 5.3, será aberta a janela com os resultados da estimação, Figura 5.4. Uma vez que o teste que se deseja executar envolve a imposição de restrições zero nos coeficientes de *advert* (publicidade) e $advert^2$ (publicidade ao quadrado), então, pode-se usar a opção `Omitir variáveis`. Sendo assim, na janela da Figura 5.4 execute o seguinte comando **Testes>Omitir Variáveis**. Isso abrirá a janela da Figura 5.5. Nessa janela, selecione as variáveis a serem testadas, no presente caso *advert* e $advert^2$ e marque a opção `Estimar modelo reduzido`, destacado com uma seta vermelha. Feito isso, clique em `Ok` e será apresentada a janela da Figura 5.6.

Com base no `p-valor` reportado nos resultados do `teste` F, Figura 5.6, rejeita-se a hipótese nula (H_0) de que os parâmetros β_3 e β_4, respectivamente das variáveis *advert* e $advert^2$, são iguais a zero e, portanto, o modelo Big Andy's Burger Ban deve ser estimado incluindo essas duas variáveis independentes – regressores.

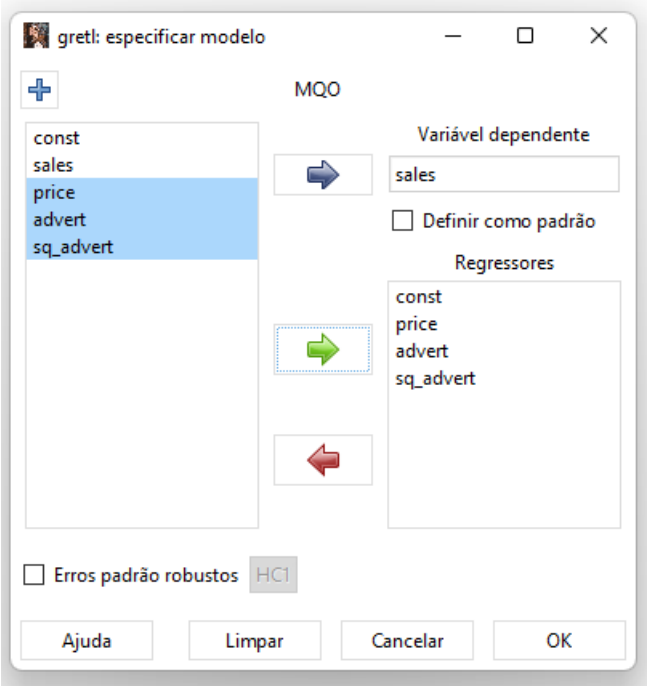


Figura 5.3: Definindo a especificação do modelo.

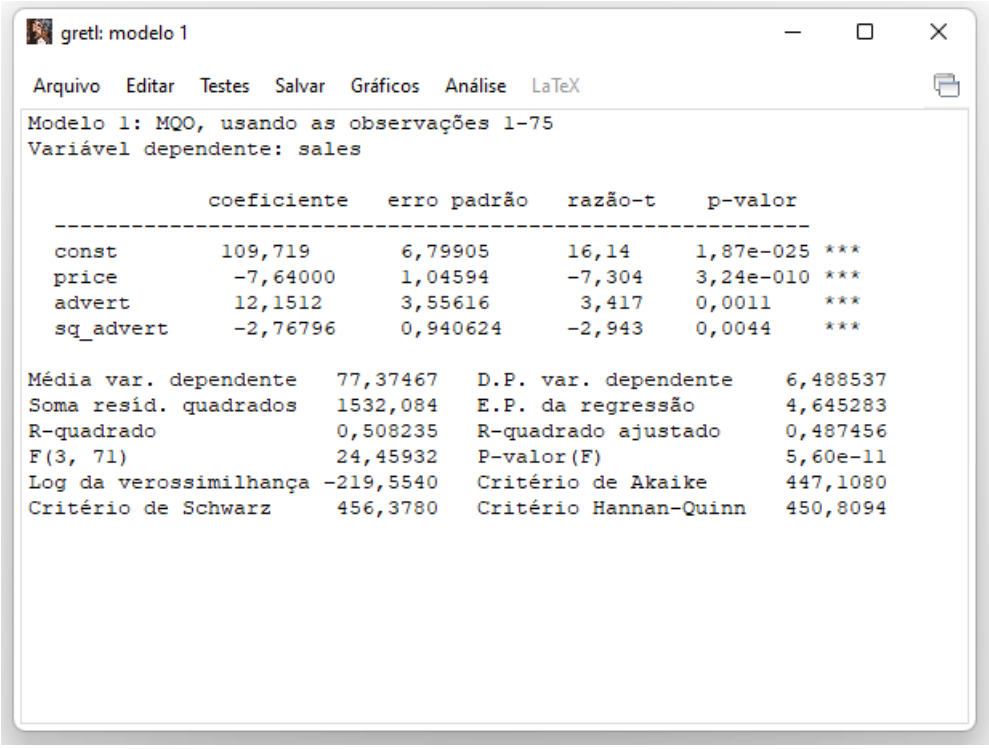


Figura 5.4: Resultados do modelo Big Andy's Burger Ban.

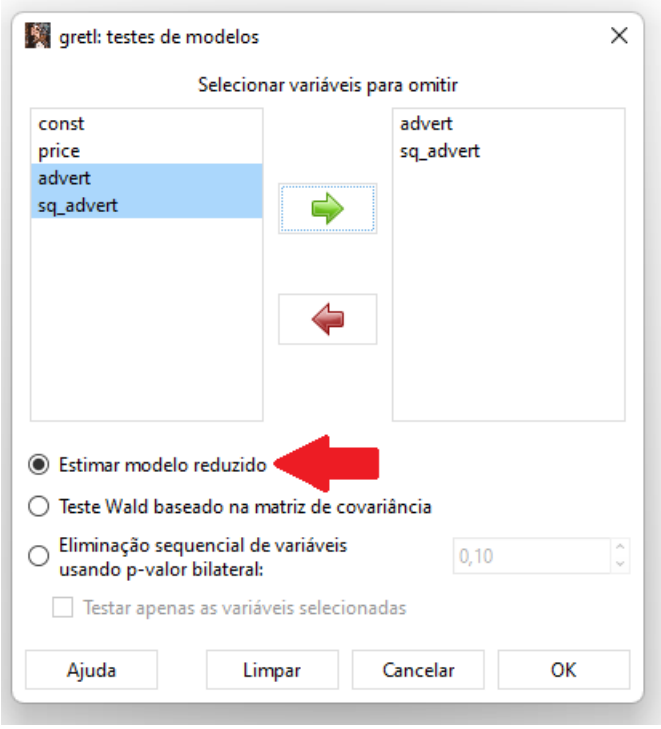


Figura 5.5: Definindo as variáveis a serem testadas.

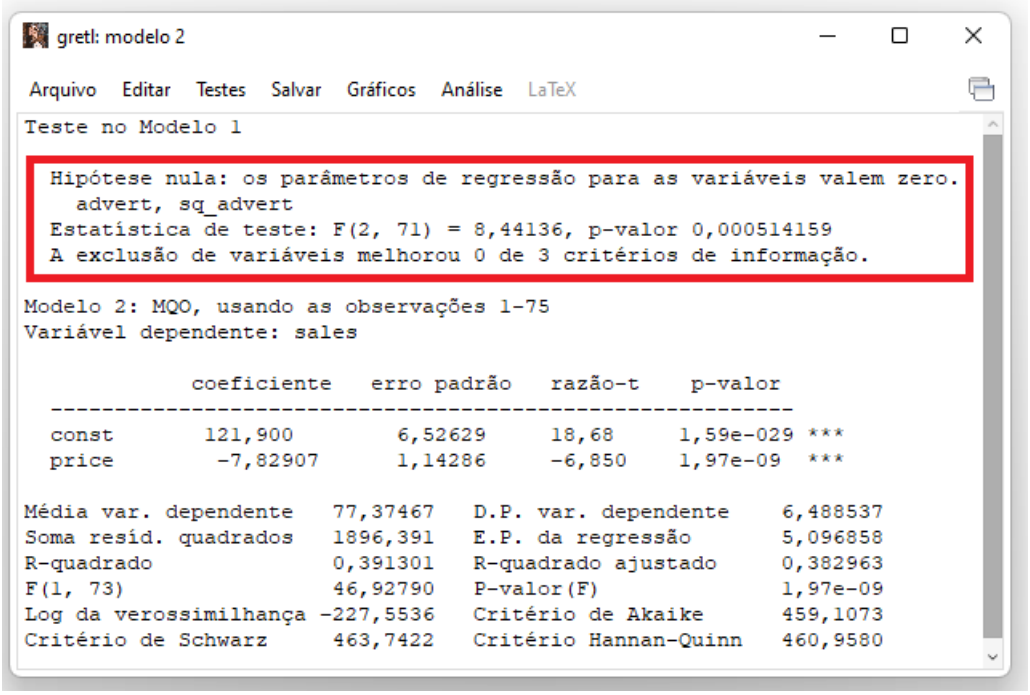


Figura 5.6: Resultado para o teste F.

5.1.2 Significância da regressão

A estatística F – teste-F – é usada para determinar se as variáveis em um modelo têm algum efeito sobre o valor médio da variável dependente y. Nesse caso, a hipótese nula, H_0, é a proposição de que y não depende de nenhuma das variáveis independentes enquanto a hipótese alternativa, H_1, é que y depende das variáveis independentes. Essa hipótese nula é, de certa maneira, muito pessimista. Note que que a hipótese nula trata-se de um conjunto de $k-1$ restrições lineares. Algebricamente, tem-se que (Equação 5.4):

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \beta_4 = \cdots = \beta_k = 0 \\ H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0 \text{ ou } \beta_4 \neq 0 \text{ ou } \ldots \text{ ou } \beta_k \neq 0 \end{cases} \tag{5.4}$$

O teste de significância geral da regressão é importante o suficiente para que todos os *softwares* econométricos e estatísticos reportem-o na saída padrão de cada regressão linear estimada. No **gretl** a estatística F $(24,45932)$ e seu `p-valor` $(5,60\text{e-}11)$, para o modelo Big Andy's Burger Ban, estão destacados na Figura 5.7. Ou seja, são reportados na janela principal do modelo. Uma vez que o `p-valor` é menor que $0,01$, então, rejeita-se a hipótese nula de que o modelo é insignificante no nível de significância de um por cento (1%).

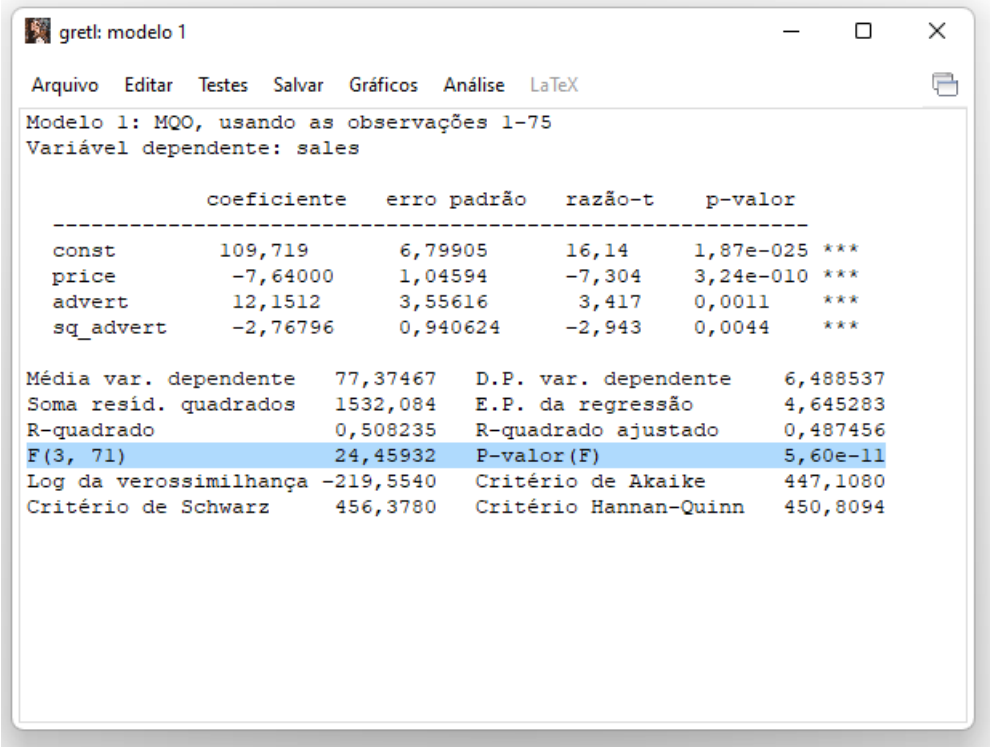


Figura 5.7: Estatística `F` de significância geral da regressão.

5.1.3 Relação entre o teste t e o teste F

Viu-se na Seção 5.1.2 como a estatística F pode ser usada para testar se um grupo de variáveis deve ou não ser incluído em um modelo. Entretanto, pode-se questionar o que aconteceria se aplicasse a estatística F ao caso de testar a significância de uma *única* variável independente? Ou seja, pode-se usar o a estatística F para testar uma *única* variável explicativa? Por exemplo, suponha que se descreva a hipótese nula como $H_0 : \beta_k = 0$ para testar a única restrição de exclusão, usando a estatística F, de que x_k pode ser excluído do modelo. Entretanto, sabe-se que a estatística t de β_k pode ser usada para testar essa hipótese.

Então, surge a dúvida: existem duas formas para testar hipóteses sobre um único coeficiente? A resposta é não. Embora as duas abordagens levem exatamente ao mesmo resultado,[1] desde que a hipótese alternativa seja bilateral, a estatística t é mais flexível para testar uma única hipótese, uma vez que essa pode ser usada para testar alternativas unilaterais. Usando o comando `Omitir` da Subsecção 5.1.1 para o modelo Big Andy's Burger Ban, Equação 5.1, obtém-se a Figura 5.8. Lembre-se de deixar a caixa `Estimar modelo reduzido` marcada.

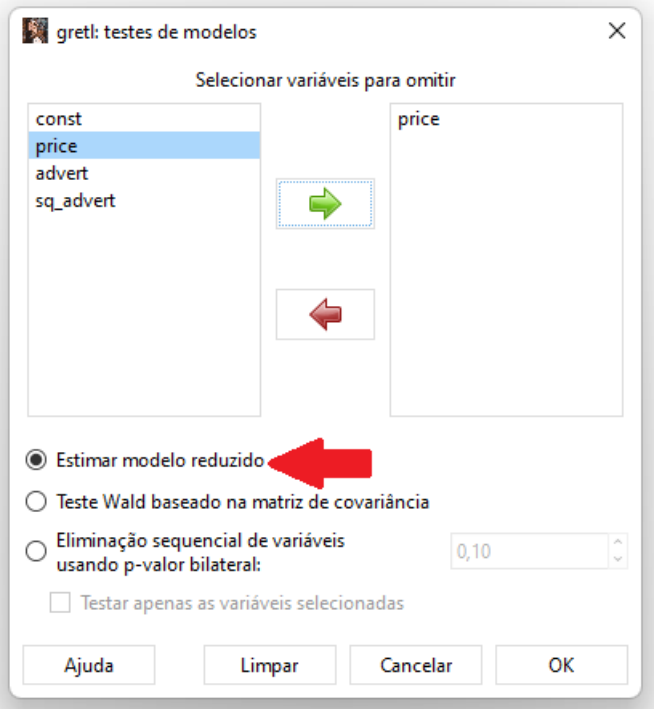


Figura 5.8: Definindo a variável a ser testada.

Ao clicar em `Ok`, na caixa de diálogo da Figura 5.8, abrirá a janela da Figura 5.9. Note que a estatística $F\,(1,\,71)$ é igual a $53,3549$ com um `p-valor` de $3,23648\text{e-}010$, que é muito menor do que $0,01$, logo, o coeficiente é significante a um nível de 1% de significância. Agora note que o quadrado da estatística `t` para a variável *price*, Figura

[1] A estatística F para testar a exclusão de uma única variável é igual ao *quadrado* da estatística t correspondente.

5.7, resultará, aproximadamente, no mesmo valor: $(-7,304)^2 = 53,348416$. Ademais, os `p-valores` também serão iguais: para a estatística F; 3, 23648e-010, (Figura 5.9) enquanto para a estatística t; 3, 24e-010 (Figura 5.7).

Destaca que o que se espera da estatística F é que essa revele se qualquer combinação de um conjunto de coeficientes $(\beta_1, \beta_2, \ldots, \beta_k)$ seja diferente de zero. Mas, entretanto, essa estatística nunca será o melhor teste para determinar se um *único* coeficiente é diferente de zero. Na verdade, a estatística t se apresenta como o teste mais adequado para testar uma *única* hipótese. Ademais, dado que as estatísticas t também são mais fáceis de serem obtidas do que as estatísticas F, uma vez que, por padrão, em todos os *softwares* econométricos e estatísticos, essas são reportadas juntamente com as demais estatísticas nas saídas da estimação, não há razão para usar uma estatística F para testar hipóteses sobre um *único* parâmetro.

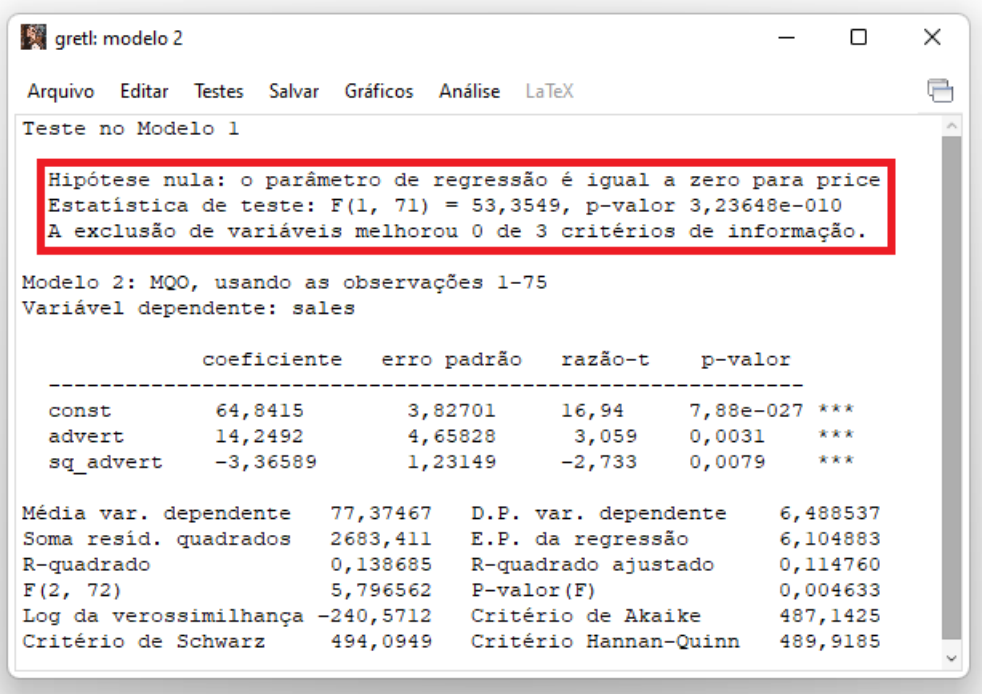


Figura 5.9: Resultado para o teste F.

5.2 Modelos restrito e irrestrito

Nesta seção, um modelo restrito[2] `log-log` de demanda por cerveja será estimado. Os dados estão disponíveis no arquivo `beer.gdt` cujas variáveis estão armazenadas em nível. O modelo é dado por:

$$\ln(q) = \beta_1 + \beta_2 \ln(pb) + \beta_3 \ln(pl) + \beta_4 \ln(pr) + \beta_5 \ln(i) + e \tag{5.5}$$

Assim, uma vez que as variáveis encontram-se na forma de nível, a primeira

[2]Importante destacar que essa abordagem é de suma importância para as funções Cobb-Douglas uma vez que o somatório dos parâmetros devem ser igual a um, i.e., $\alpha + \beta = 1$. Portanto, modelos empregando funções Cobb-Douglas caracterizam-se como sendo um modelo restrito.

coisa a se fazer é transformar cada uma das variáveis para logaritmo natural ou logaritmo neperiano. Para isso, bastar usar o comando **Acrescentar>Logaritmos das variáveis selecionadas**, Figura 5.10. Logo após a criação dos logaritmos neperiano das variáveis selecionadas a janela principal do **gretl** terá a aparência da Figura 5.11.

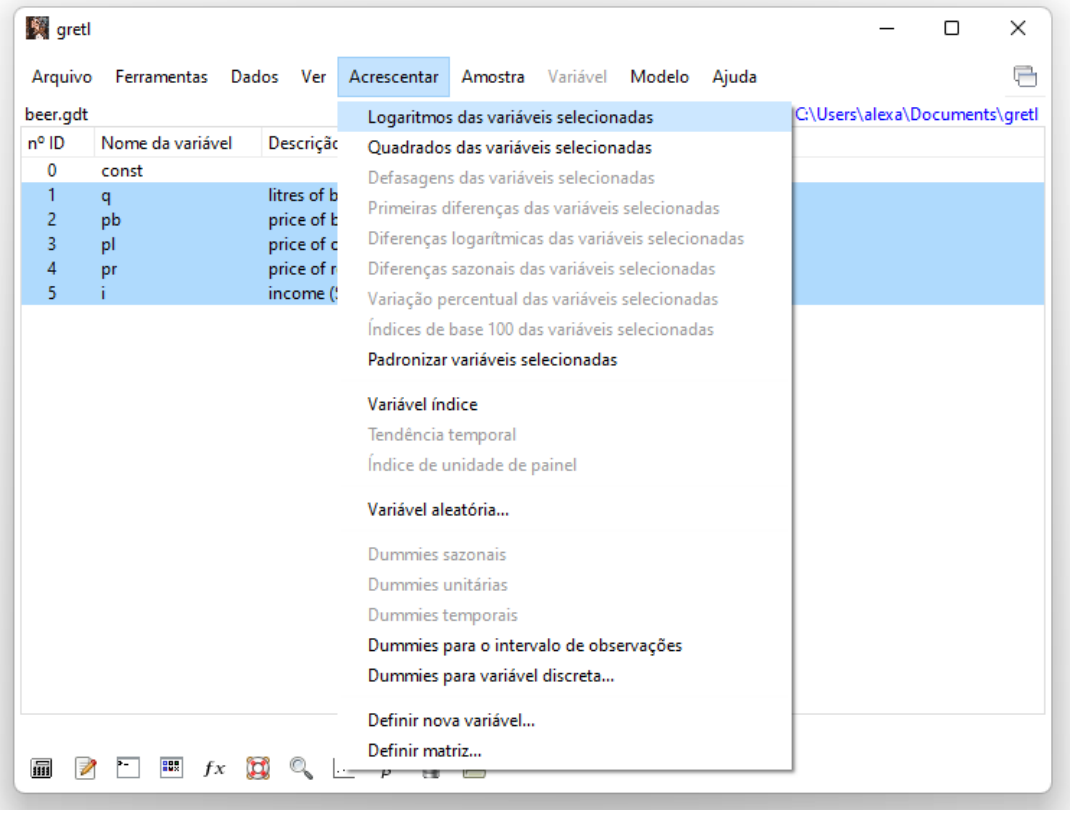


Figura 5.10: Obtendo o logaritmo das variáveis de interesse.

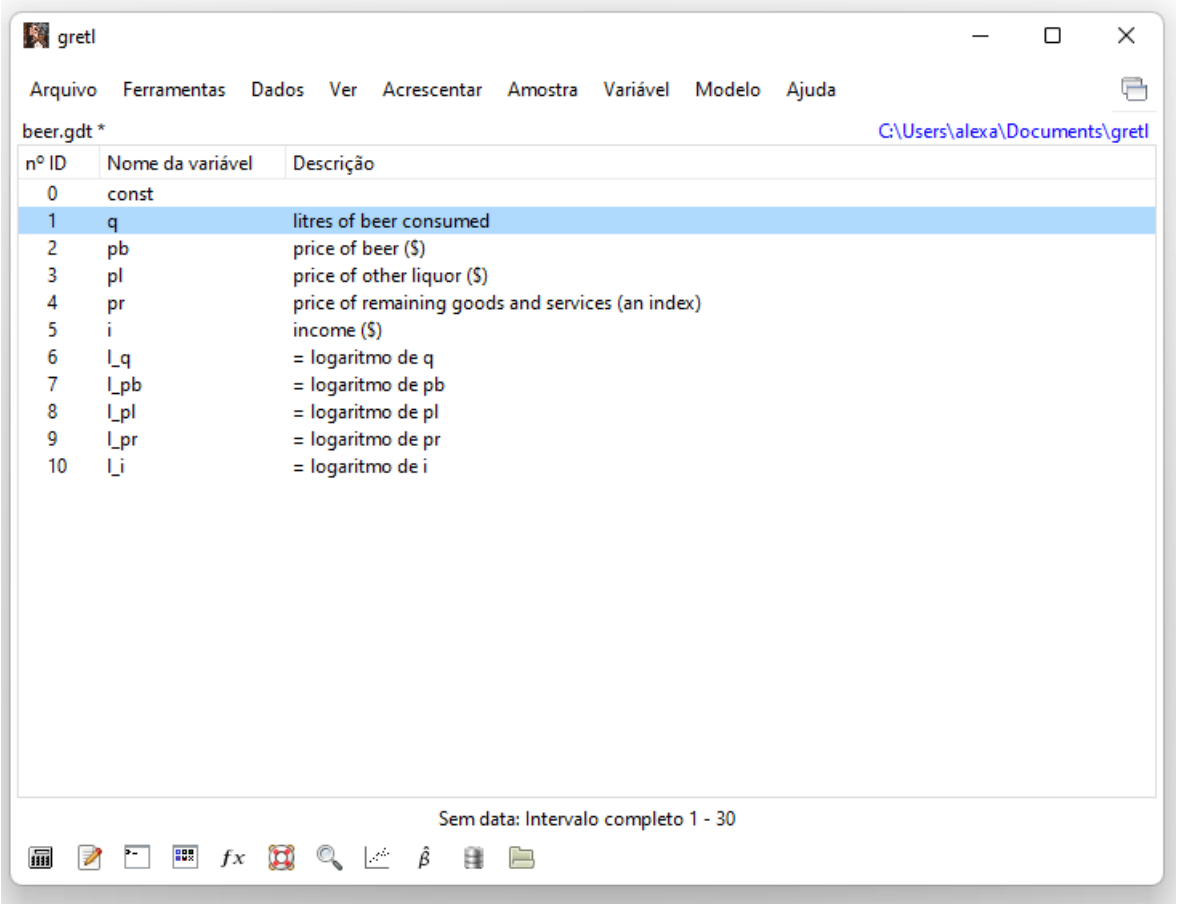


Figura 5.11: Janela principal com os logaritmos neperiano das variáveis selecionadas.

Agora se está interessado em estimar um modelo mas com a restrição de que o somatório dos parâmetros β_2, β_3, β_4 e β_5 seja igual a zero, ou seja, $\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$. Nesse caso, inicialmente estima um modelo irrestrito usando o comando **Modelo>Mínimos Quadrados Ordinários**, Figura 1.12. Posteriormente, usa-se o comando **Testes>Restrições lineares** para informar ao **gretl** que a estimação tem como restrição que o somatório dos parâmetros β_{2-5} deve ser igual a zero, ou seja, estima-se um modelo restrito – Figura 5.13.

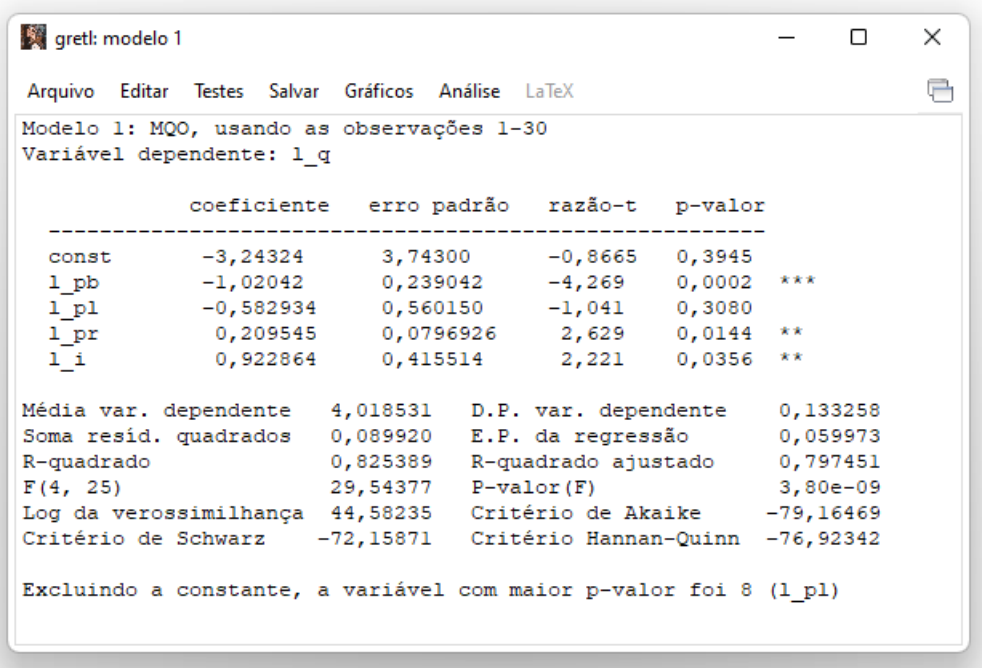


Figura 5.12: Resultados do modelo irrestrito de demanda por cerveja.

As restrições para o modelo restrito devem ser informadas manualmente com a seguinte relação: $\beta_2 =$ b [l_pb], $\beta_3 =$ b [l_pl], $\beta_4 =$ b [l_pr] e $\beta_5 =$ b [l_i]. Os resultados para o modelo restrito são apresentados na Figura 5.14. Note que o somatório dos coeficientes β_{2-5} totaliza zero $(-1,29939 + 0,186816 + 0,166742 + 0,945829 = 0)$.

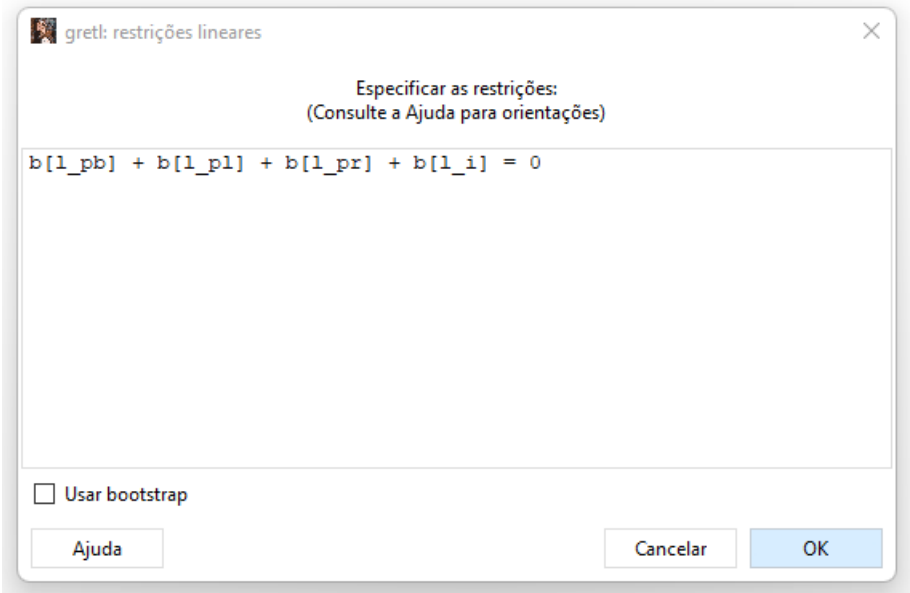


Figura 5.13: Restrições para o modelo restrito de demanda por cerveja.

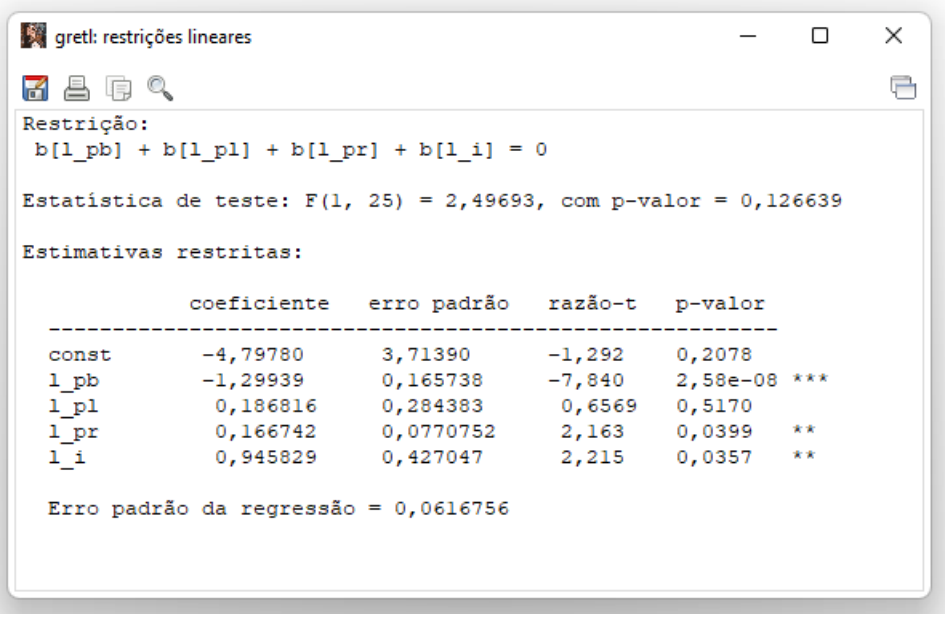


Figura 5.14: Resultados para o modelo restrito de demanda por cerveja.

5.3 Especificação do modelo

Diversas questões relacionadas à especificação de um modelo serão abordadas nesta seção. Inicialmente será considerado o problema de viés de variável omitida. Isso ocorre quando se omite variáveis independentes relevantes para o modelo. Uma variável independente é dita relevante quando essa afeta a média da variável dependente. Mais precisamente, quando se omite uma variável relevante que está correlacionada com qualquer um dos outros regressores, o estimador de Mínimos Quadrados sofre de viés de variável omitida.

Por outro lado, incluir variáveis irrelevantes ao modelo também gera problema para a estimação. Ou seja, incluir regressores que não afetam y (a variável dependente) ou, se afetam, não correlacionados com os demais regressores. A inclusão de variáveis independentes irrelevantes no modelo torna os Mínimos Quadrados menos precisos do que seriam – isso aumenta os erros-padrão, reduz o poder dos testes de hipóteses do modelo bem com aumenta o tamanho dos intervalos de confiança do modelo.

Nesta seção, os exemplos irão usar o conjunto de dados `edu_inc.gdt`. O primeiro modelo é dado por:

$$l\_faminc_i = \beta_1 + \beta_2\, he_i + \beta_3\, we_i + e_i \tag{5.6}$$

em que l_faminc é o logaritmo neperiano da renda familiar, he são os anos de escolaridade do marido e we são os anos de escolaridade da esposa. São estimadas diversas variações desse modelo que incluem o número de crianças menores de 6 anos no domicílio ($kl6$) e duas variáveis irrelevantes – x_5 e x_6.

Os dados são carregados no **gretl**, o logaritmo neperiano da renda familiar é obtido e, então, estima-se a Equação 5.6, considerada a equação "*baseline*". Serão estimados duas especificações, i) uma especificação completa, ou seja, incluindo tanto a escolaridade do marido quanto a escolaridade da esposa e; ii) uma especificação em

que a escolaridade da esposa é omitida. Uma vez estimada as duas especificações coloca-se os resultados das duas estimações em uma única janela (Figura 5.15).

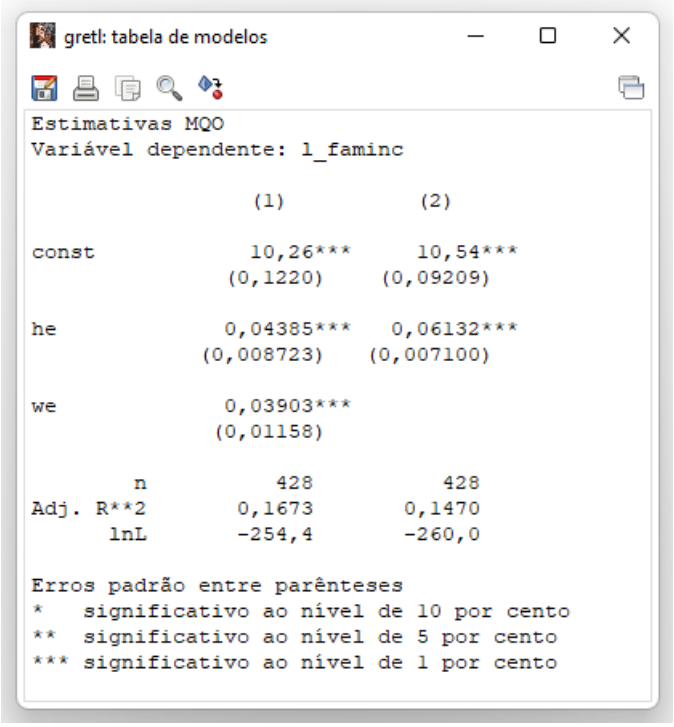


Figura 5.15: Tabela de modelos.

Para conseguir a tabela da Figura 5.15 estima o modelo irrestrito, denominado **modelo 1**. Na janela dos resultados do modelo execute o comando **Arquivo>Salvar para sessão como ícone** (Figura 5.16. Isso abrirá a janela `gretl: visualização de ícones`, Figura 5.17, que conterá um ícone denominado `Modelo 1`. Siga os mesmos passos para o modelo restrito e, assim, na janela `gretl: visualização de ícones` existirão dois ícones – `Modelo 1` e `Modelo 2`. Então, para obter a Figura 5.15 arraste o ícone `Modelo 1` para o ícone `Tabela de modelos` bem como arraste o ícone `Modelo 2` para o ícone `Tabela de modelos`. Observação, arraste um ícone por vez. Feito isso, basta dar um duplo clique no ícone `Tabela de modelos` para que a tabela da Figura 5.15 abra.

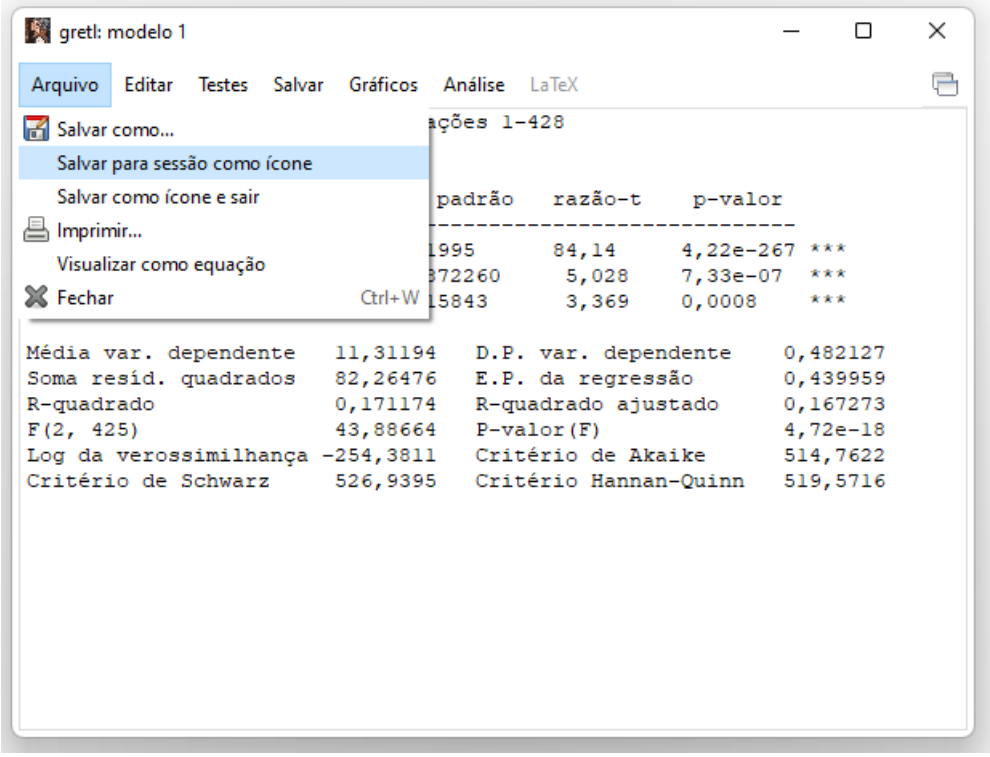


Figura 5.16: Salvar para sessão como ícone.

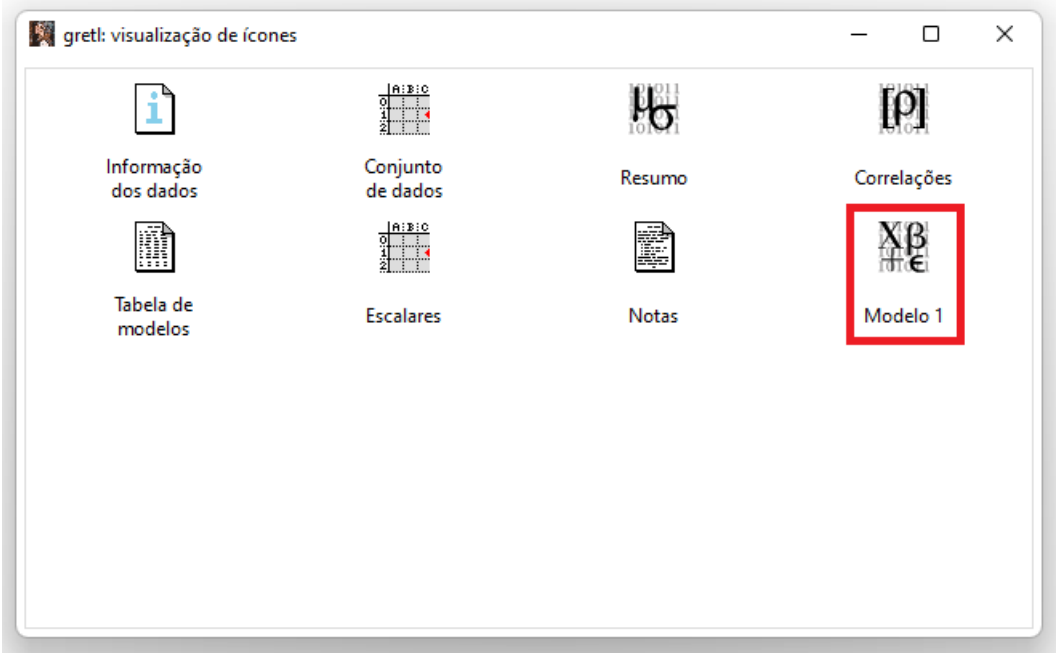


Figura 5.17: Visualização de ícones.

A seguir apresenta-se a tabela comparativa para a estimação das 5 diferentes

especificações, Figura 5.18.

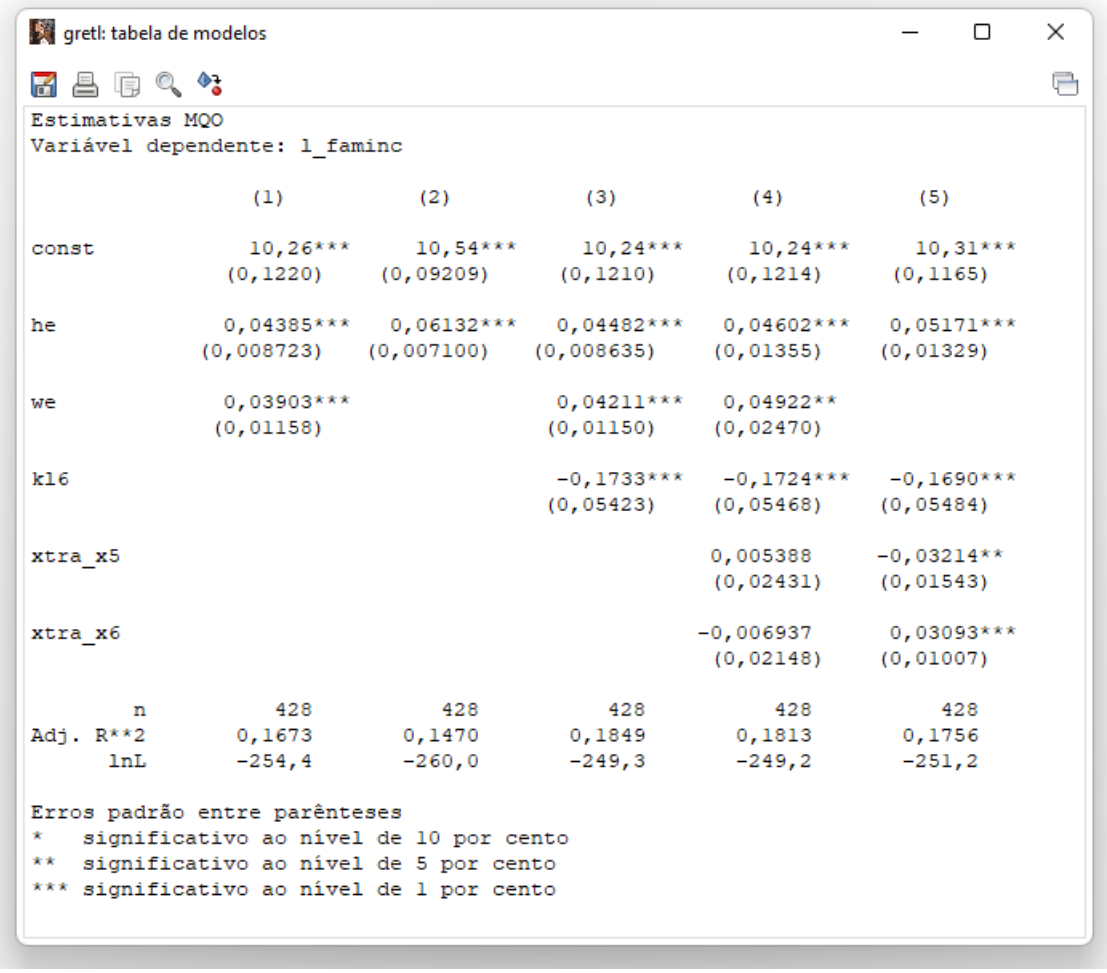


Figura 5.18: Tabela de modelos.

Note que, do **Modelo** 1 para o **Modelo** 2, houve a exclusão de uma variável relevante da especificação, a variável `we`. Assim, o R^2 ajustado ficou menor (passou de $0,1673$ para $0,1470$). Ou seja, o poder de explicação do modelo ficou menor ao se excluir uma variável independente relevante para o modelo. Por outro lado, comparando o **Modelo** 1 com o **Modelo** 3 percebe-se que a inclusão de uma variável independente relevante para o modelo eleva o poder de explicação. Pois o R^2 ajustado passou de $0,1673$ para $0,1849$.

Ao contrário, a inclusão de variáveis independentes irrelevantes para o modelo irão, como supracitado, aumentar os erros-padrão, reduzir o poder dos testes de hipóteses do modelo, aumentar o tamanho dos intervalos de confiança do modelo bem como reduzir o poder de explicação do modelo. Comparando o **Modelo** 3 com o **Modelo** 4 percebe-se que a inclusão dos regressores $xtra\_x5$ e $xtra\_x6$ não afetam a variável dependente, mas aumenta os erros-padrão. Por outro lado, o comparativo entre o **Modelo** 3 e **Modelo** 5 nota-se que a exclusão do regressor `we` e a inclusão dos regressores $xtra\_x5$ e $xtra\_x6$ torna esses dois regressores significativos, entretanto, como são irrelevantes para o

modelo, provocam o aumento dos erros-padrão das demais variáveis do modelo.

5.4 Seleção do modelo

Um desafio para todo estudo empírico é a escolha de um modelo apropriado. A omissão de variáveis relevantes que estão correlacionadas com as demais variáveis faz com que os Mínimos Quadrados sejam tendenciosos e inconsistentes.[3] A inclusão de variáveis irrelevantes reduz a precisão dos Mínimos Quadrados. Assim, do ponto de vista puramente técnico, é importante estimar um modelo que contenha todas as variáveis relevantes necessárias e nenhuma irrelevante. Além disso, é de suma importância a adoção de uma forma funcional (uma especificação) adequada. Entretanto, destaca-se que não existe nenhum conjunto de regras mecênicas que se possa seguir para garantir que o modelo seja especificado corretamente, mas há algumas coisas que se pode fazer para aumentar as chances de ter um modelo adequado para usar nas tomadas de decisões.

A seguir têm-se algumas regras de ouro que podem auxiliar estudos empíricos:

1. Use a literatura pregressa bem como a teoria econômica para selecionar uma forma funcional. Por exemplo, se estiver estimando uma função de produção de curto prazo, a teoria econômica sugere que os retornos de produção diminuem. Portanto, deve-se escolher uma forma funcional que permita retornos de produção decrescente e, nesse caso, adota-se uma forma funcional do tipo `log-log`;

2. Se os parâmetros estimados tiverem sinais opostos ou magnitudes não razoáveis ao esperado pela literatura pregressa, é prudente reavaliar a forma funcional ou se uma ou mais variáveis relevantes foram omitidas;

3. Pode-se realizar testes de hipóteses conjuntas para detectar a inclusão de conjuntos de variáveis irrelevantes. O teste não é infalível, pois sempre há a probabilidade positiva de que o erro do tipo 1 ou do tipo 2 esteja sendo cometido;

4. Pode-se usar as regras de seleção de modelo para encontrar conjuntos de regressores que são "ótimos" em termos de um *trade-off* estimado de viés/precisão e;

5. Pode-se usar um `teste RESET` para detectar possível especificação incorreta da forma funcional.

Nesta seção, serão apresentado alguns comandos do **gretl** para ajudar com as duas últimas regras de ouro: seleção de modelo e `teste RESET`. Ademais, considera-se três regras para seleção de modelo: $\bar{R}^2$, AIC e SC. Porém, destaca-se que não se está recomendando a aplicação dessas três regras, pois há muitos problemas estatísticos causados pelo uso da amostra para estimar, especificar e testar hipóteses em um modelo, mas as vezes se têm poucas opções.

[3] Dada a hipótese de que u_i segue a distribuição normal, os estimadores de Mínimos Quadrados Ordinários têm, entre outras, a seguinte propriedade: São consistentes; à medida que o tamanho da amostra aumenta indefinidamente, os estimadores convergem para os verdadeiros valores da população.

5.4.1 R^2 ajustado

O coeficiente de determinação R^2 usual é "ajustado" – $\bar{R}^2$ – para impor uma penalidade quando uma variável independente é adicionada ao modelo. Adicionar uma variável independente com qualquer correlação com a variável dependente y sempre reduz a Soma dos Quadrados Explicados (SQE) e aumenta o valor do R^2 usual. Por sua vez, com a versão "ajustada", i.e., $\bar{R}^2$, a melhoria no ajuste pode ser penalizada e pode ser menor à medida que variáveis independentes são adicionadas ao modelo. A fórmula é:

$$\bar{R}^2 = 1 - \frac{\text{SQE} \,/\, (n \,-\, k)}{\text{STQ} \,/\, (n \,-\, 1)} \tag{5.7}$$

em que SQE é a Soma dos Quadrados Explicados, STQ é a Soma Total dos Quadrados, n caracteriza-se como sendo o número de observações e k corresponde ao grau de liberdade.

Destaca-se que uma desvantagem em usar o $\bar{R}^2$ – R^2 ajustado ou R^2 barra – como regra de seleção de modelo é que a penalidade imposta por essa regra a cada regressor adicionado é muito pequena em média. Assim, esse critério de seleção de modelo tende a levar a modelos que contêm variáveis independentes irrelevantes.

5.4.2 Critério de informação

Por padrão, o **gretl** calcula o Critério de Informação Akaike (AIC) e o Critério de Schwarz (SC), esse último é também conhecido como Bayesian Information Criterion (BIC), e os inclui na saída da regressão padrão. Os valores que o **gretl** reporta são baseados na maximização de uma função de verossimilhança logarítmica (erros normais). Esses dois critérios são utilizados como regras para a seleção de modelo. As fórmulas desses critérios são:

$$\text{AIC} = \ln\left(\text{SQE} \,/\, n\right) + 2\,k \,/\, n \tag{5.8}$$

$$\text{SC} = \text{BIC} = \ln\left(\text{SQE} \,/\, n\right) + k\,\ln\left(n\right) / \, n \tag{5.9}$$

em que SQE corresponde a Soma dos Quadrados Explicados, n caracteriza-se como sendo o número de observações e, por sua vez, k representa o grau de liberdade.

Para proceder a seleção de modelo deve-se calcular AIC ou SC para cada modelo em consideração e escolher o modelo que minimiza o critério desejado. Lembre-se que os modelos devem ser estimados utilizando-se o mesmo número de observações, i.e., n. Assim, uma vez que o tamanho da amostra deve ser mantido constante ao usar regras de seleção de modelo, percebe-se que os dois critérios (AIC ou BIC) levarão exatamente a mesma escolha do modelo.

5.4.3 teste RESET

O `teste RESET` é utilizado para checar se a forma funcional empregada é adequada. A hipótese nula (H_0) é que a forma funcional é adequada enquanto a hipótese alternativa (H_1 ou H_a) implica que a forma funcional não é adequada. O `teste RESET` envolve calcular algumas regressões e calcular uma estatística `F`.
Considere o seguinte modelo:

$$y_i = \beta_1 + \beta_2\, x_{i\,2} + \beta_3\, x_{i\,3} + e_i \tag{5.10}$$

E as seguintes hipóteses:

$$\begin{aligned} &\mathrm{H}_0 : E\left[y \,|\, x_{i\,2},\, x_{i\,3}\right] = \beta_1 + \beta_2\, x_{i\,2} + \beta_3\, x_{i\,3} \\ &\mathrm{H}_1 : \text{não } \mathrm{H}_0 \end{aligned}$$

Se H_0 for rejeitado implica que a forma funcional empregada não é suportada pelos dados. Para proceder este teste, primeiramente estime a Equação 5.10 usando Mínimos Quadrados Ordinários (MQO) e salve os valores previstos, $\hat{y}_i$. Então, eleve os valores previstos $\hat{y}_i$ ao quadrado e ao cubo e os adicionem ao modelo:

$$\begin{aligned} y_i &= \beta_1 + \beta_2\, x_{i\,2} + \beta_3\, x_{i\,3} + \gamma_1\, \hat{y}_i^2 + e_i \\ y_i &= \beta_1 + \beta_2\, x_{i\,2} + \beta_3\, x_{i\,3} + \gamma_1\, \hat{y}_i^2 + \gamma_2\, \hat{y}_i^3 + e_i \end{aligned}$$

As hipótese nulas a testar contra a hipótese alternativa (H_1 : não H_0) são:

$$\begin{aligned} &\mathrm{H}_0 : \gamma_1 = 0 \\ &\mathrm{H}_0 : \gamma_1 = \gamma_2 = 0 \end{aligned}$$

Para realizar o **teste RESET** use o comando **Testes>RESET de Ramsey** na janela com os resultados da regressão após a estimação do modelo por Mínimos Quadrados Ordinários (MQO), conforme a Figura 5.19.

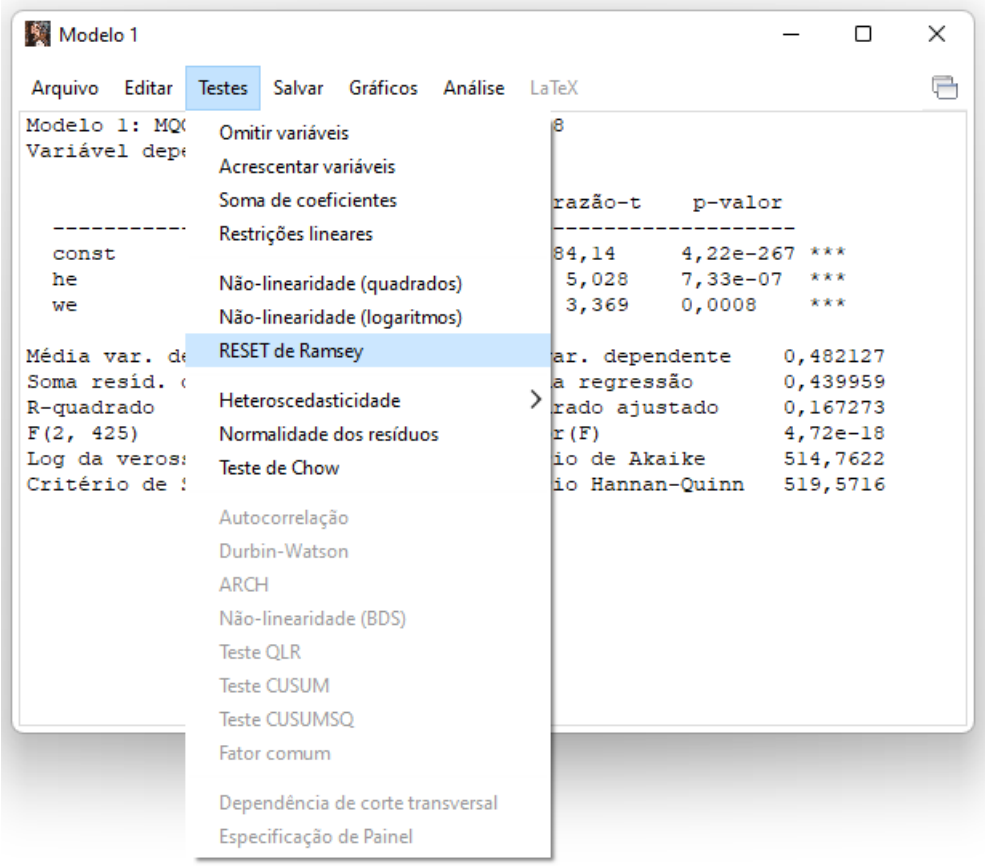


Figura 5.19: Teste RESET de Ramsey.

Ao clicar em **RESET de Ramsey** abrirá uma janela igual a da Figura 5.20. Observe que nessa janela estão disponíveis as seguintes opções: i) `quadrados e cubos`; ii) `apenas quadrados`; iii) `apenas cubos` e; iv) `todas as variantes`. Inicialmente realiza-se um teste `apenas quadrados` e, a seguir, um teste para `quadrados e cubos`.

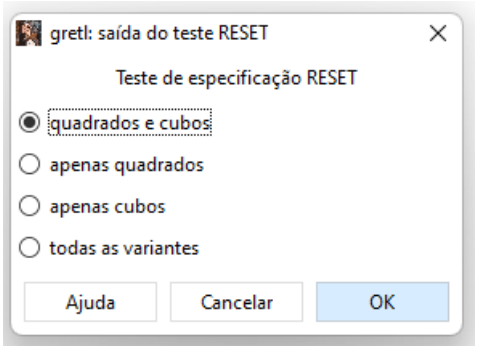


Figura 5.20: Janela para o teste de especificação RESET.

Os resultados do `teste RESET` para a Equação 5.6 são os seguintes (Figuras 5.21 e 5.22):

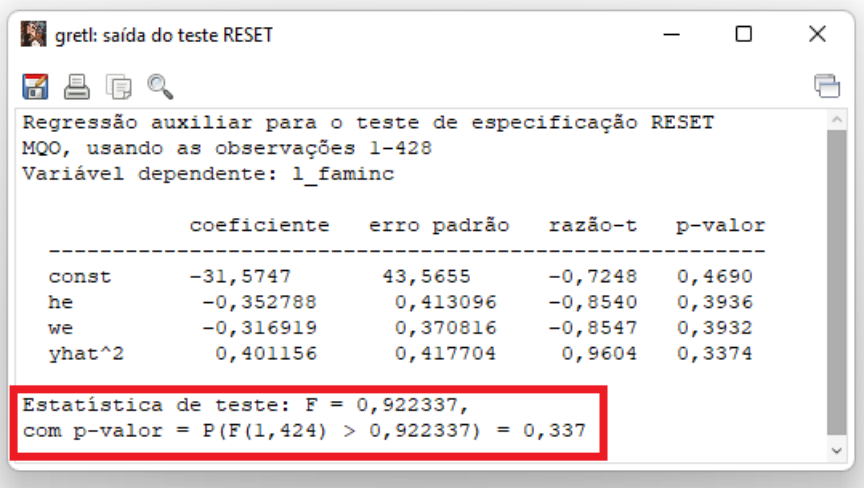


Figura 5.21: Teste RESET apenas quadrados.

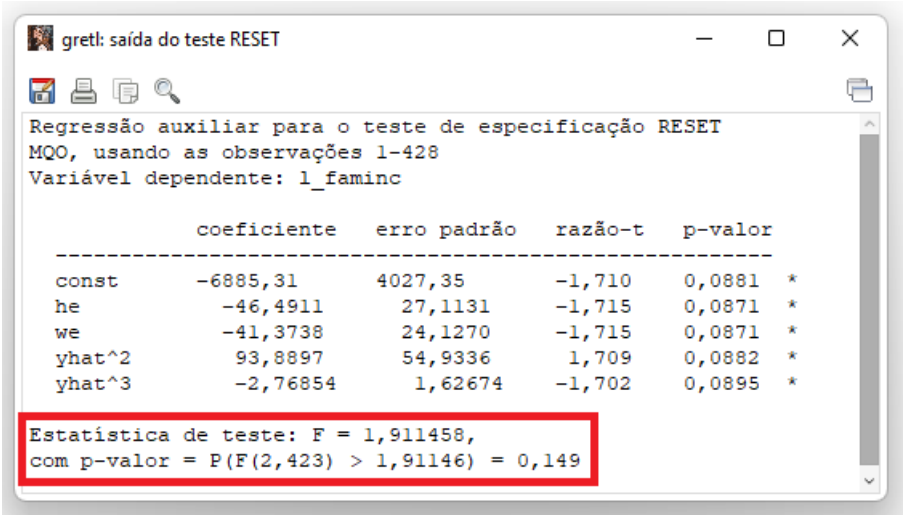


Figura 5.22: Teste RESET quadrados e cubos.

Pelas Figuras 5.21 e 5.22 nota-se que a adequação da forma funcional não é rejeitada ao nível de significância de 5% para ambos os testes. Uma vez que os `p-valores` foram, respectivamente, 0,337 e 0,149.

5.4.4 Colinearidade

As estatísticas descritivas de um conjunto de dados podem fornecer informações úteis sobre os dados, servindo a vários propósitos. Por exemplo, se houver algum problema com o conjunto de dados, as estatísticas descritivas podem fornecer alguma indicação. O tamanho da amostra é o esperado? A média, o mínimo e o máximo são razoáveis? Caso contrário, precisa-se fazer algum trabalho investigativo. Além disso, ao observar as estatísticas descritivas se tem uma ideia de como as variáveis foram dimensionadas.

Isso é de suma importância quando se trata de extrair sentido econômico dos resultados. A magnitude dos coeficientes faz sentido? Por meio das estatísticas descritivas também é possível identificar variáveis discretas, que requerem algum cuidado na interpretação.

O comando **Ver>Estatísticas descritivas** incluem as seguintes estatísticas:

1. Média;
2. Mediana;
3. Mínimo (Min);
4. Máximo (Max);
5. Desvio padrão (D.P.);
6. Coeficiente de variação (CV);
7. Assimetria e;
8. Excesso de curtose.

O comando **Ver>Matriz de correlação** calcula a correlação simples entre as variáveis. Isso pode ser útil para obter uma compreensão inicial se as variáveis são altamente colineares ou não. Embora outras medidas sejam mais úteis, nunca é demais olhar para as correlações. Qualquer um desses dois comandos podem ser usado com uma lista de variáveis selecionadas para limitar a quantidade de variáveis resumidas ou correlacionadas. Por exemplo, usando a base de dados `rice5.gdt`, na Figura 5.23 foram selecionadas previamente apenas as variáveis `firm`, `area`, `fert`, `labor`, `prod` e `year`, sombreadas de azul claro, para a obtenção das estatísticas descritivas e correlação.

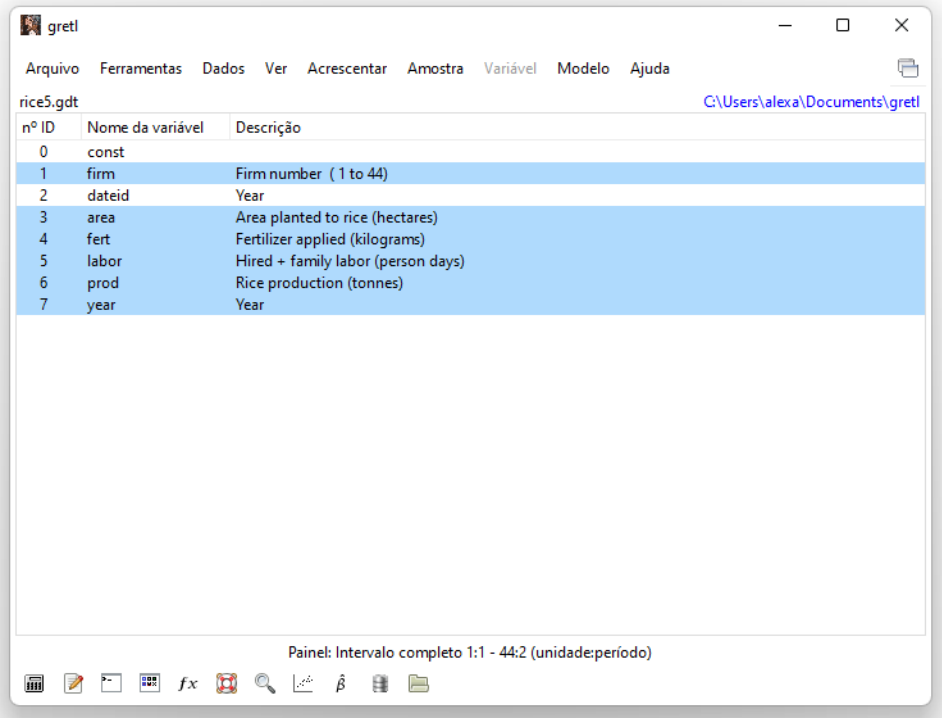
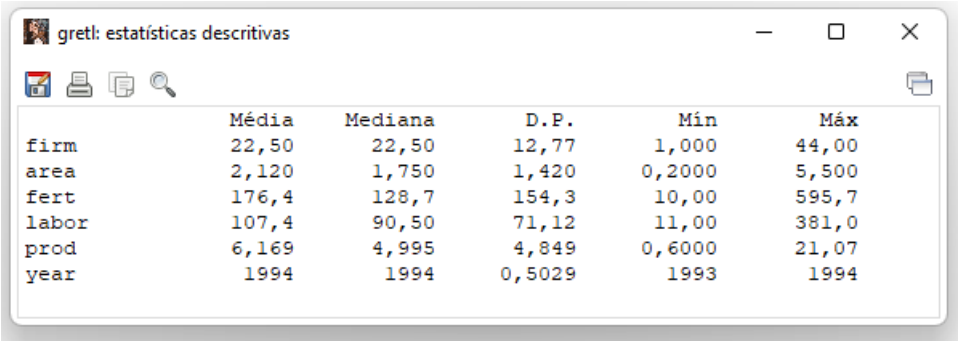


Figura 5.23: Janela principal com as variáveis de interesse selecionadas.

Considere o exemplo de produção de arroz (dados contidos no arquivo `rice5.gdt`). Esse é um modelo `log-log` de produção (toneladas de arroz) que depende da área cultivada (hectares), mão de obra (pessoa-dia) e fertilizante (quilogramas).

$$\ln(prod) = \beta_1 + \beta_2 \ln(area) + \beta_3 \ln(labor) + \beta_4 \ln(fert) + e \tag{5.11}$$

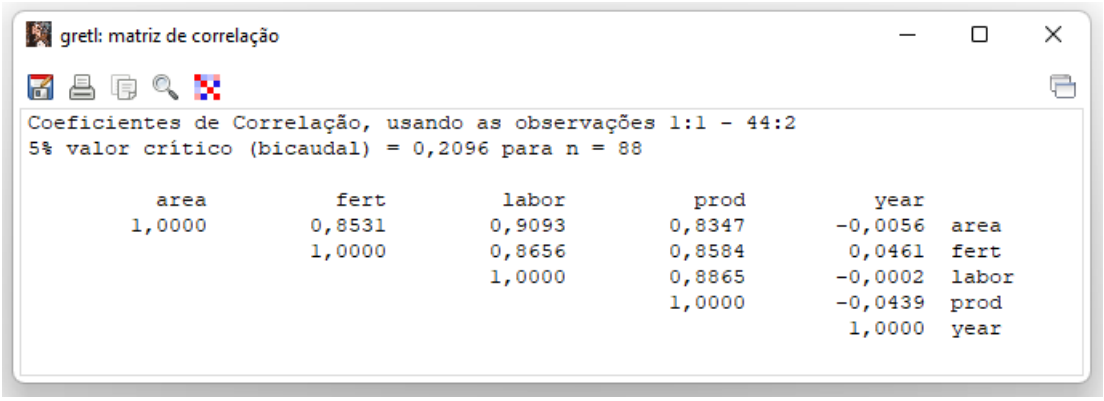
A Figura 5.24 apresenta as principais estatísticas descritivas (média, mediana, desvio padrão (D.P.), Mínimo e Máximo) para as variáveis em nível, ou seja, antes da transformação logarítmica:



	Média	Mediana	D.P.	Mín	Máx
firm	22,50	22,50	12,77	1,000	44,00
area	2,120	1,750	1,420	0,2000	5,500
fert	176,4	128,7	154,3	10,00	595,7
labor	107,4	90,50	71,12	11,00	381,0
prod	6,169	4,995	4,849	0,6000	21,07
year	1994	1994	0,5029	1993	1994

Figura 5.24: Tabela de estatísticas descritivas.

Por sua vez, a matriz de correlação para o mesmo conjunto de variáveis (menos a variável `firm`) está demonstrada na Figura 5.25. Nota-se por essa matriz que as variáveis na amostra são altamente correlacionadas. Por exemplo, a correlação entre `area` e `labor` é de $0,9093$. Quanto maior a área da fazenda maior o emprego de mão de obra. Nenhuma surpresa!



Coeficientes de Correlação, usando as observações 1:1 - 44:2
5% valor crítico (bicaudal) = 0,2096 para n = 88

area	fert	labor	prod	year	
1,0000	0,8531	0,9093	0,8347	-0,0056	area
	1,0000	0,8656	0,8584	0,0461	fert
		1,0000	0,8865	-0,0002	labor
			1,0000	-0,0439	prod
				1,0000	year

Figura 5.25: Matriz de correlação para as variáveis em nível.

Tomar o logaritmo das variáveis não provocará grandes mudanças nas correlações. As correlações entre os logaritmos das variáveis são apresentados na Figura 5.26. A correlação entre $\ln(area)$ e $\ln(labor)$ na verdade aumenta ligeiramente de $0,9093$ para

0, 9320.

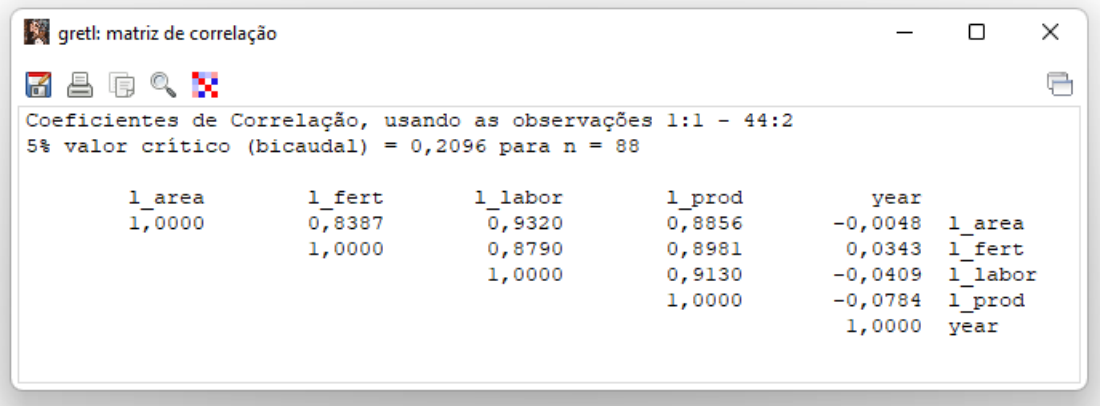


Figura 5.26: Matriz de correlação para o logaritmo das variáveis.

O modelo de produção de arroz, Equação 5.11, é estimado para o ano de 1994 e os resultados são apresentados na Figura 5.27. Para estimar o modelo apenas para o ano de 1994 utiliza-se os seguintes comando no console do **gretl**.

smpl (year == 1994) −−restrict

m_1994 < − ols l_prod const l_area l_labor l_fert

omit l_area l_labor −−test−only

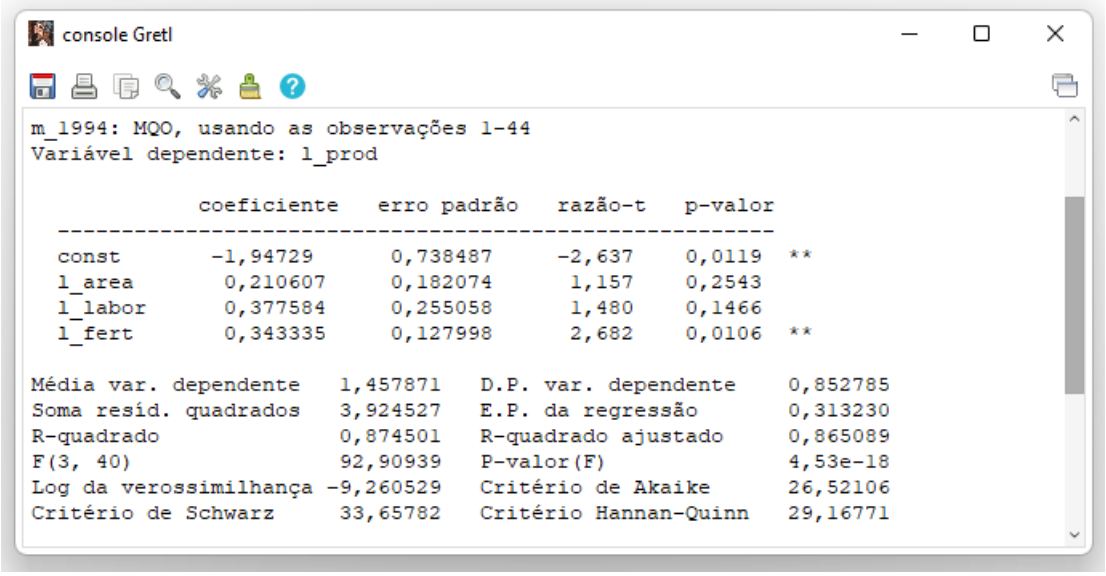


Figura 5.27: Resultados do modelo de produção de arroz.

Nota-se da Figura 5.27 que além da constante a única variável significativa foi l_fert, ao nível de 5%. A estatística `F` é de 92, 90939 com `p-valor` de 4, 53e-18, bem abaixo de 1%. O coeficiente de determinação R^2 é de 0, 874501, que parece bastante

grande. A significância conjunta de β_2 e β_3 é testada usando o comando `omit`, Figura 5.28. Os coeficientes são conjuntamente diferentes de zero uma vez que o `p-valor` para este teste foi $0,00214705$. Assim, pode-se rejeitar a hipótese nula de $\beta_2 = \beta_3 = 0$ ao nível de significância de 1%, pois $0,00214705 < 0,01$.

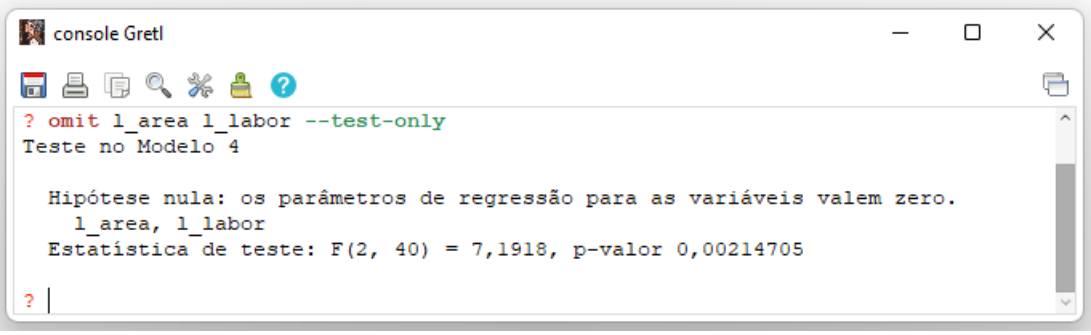


Figura 5.28: Significância conjunta de β_2 e β_3.

Finalmente, a colinearidade é examinada usando a função `vif` após a regressão. `vif` significa Variance Inflation Factor (Fator de Inflação de Variância) e é usado como um diagnóstico de colinearidade por muitos *softwares*, incluindo o **gretl**. A função `vif` está relacionada com a recomendação de Hill *et al.* (2018) p.(91) que sugere usar o coeficiente de determinação R^2 de regressões auxiliares para determinar até que ponto cada variável independente pode ser explicada como funções lineares das outras variáveis independentes. A função `vif` regride x_j contra todas as outras variáveis independentes e compara o R_j^2 da regressão auxiliar com 10. Se R_j^2 exceder 10 haverá evidência de um problema de colinearidade.

O vif_j relata as mesmas informações, mas de uma forma menos direta. O `vif` associado ao j-ésimo regressor é calculado da seguinte forma:

$$vif_j = \frac{1}{1 - R_j^2} \tag{5.12}$$

que é uma função apenas de R_j^2 da j-ésima regressão auxiliar. Ademais, observe que quando $R_j^2 > 0,9$, o $vif_j > 10$. Portanto, a regra prática para as duas regras é, na verdade, a mesma. Um vif_j maior que 10 é equivalente a um R_j^2 maior que $0,9$ da regressão auxiliar. Para realizar o teste de colinearidade, estime o modelo e, na janela do modelo, use o comando **Análise>Colinearidade**, Figura 5.29, e os resultados aparecerão na saída do **gretl**.

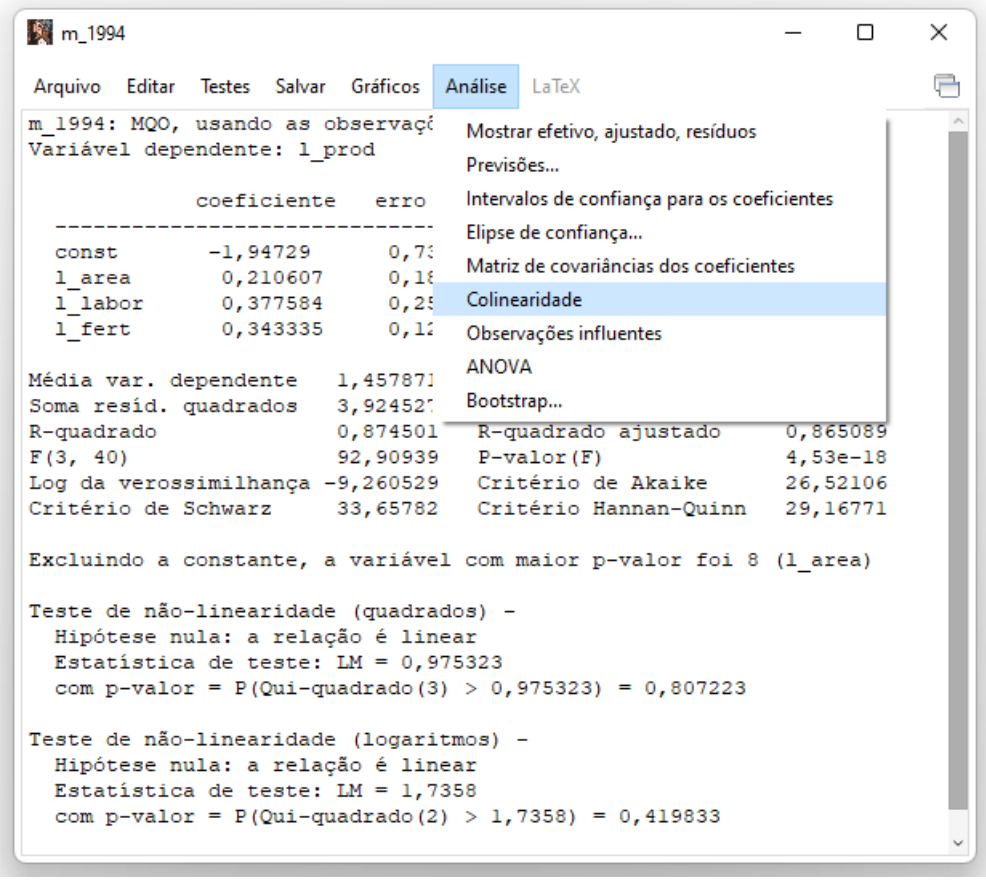


Figura 5.29: Janela do modelo de regressão.

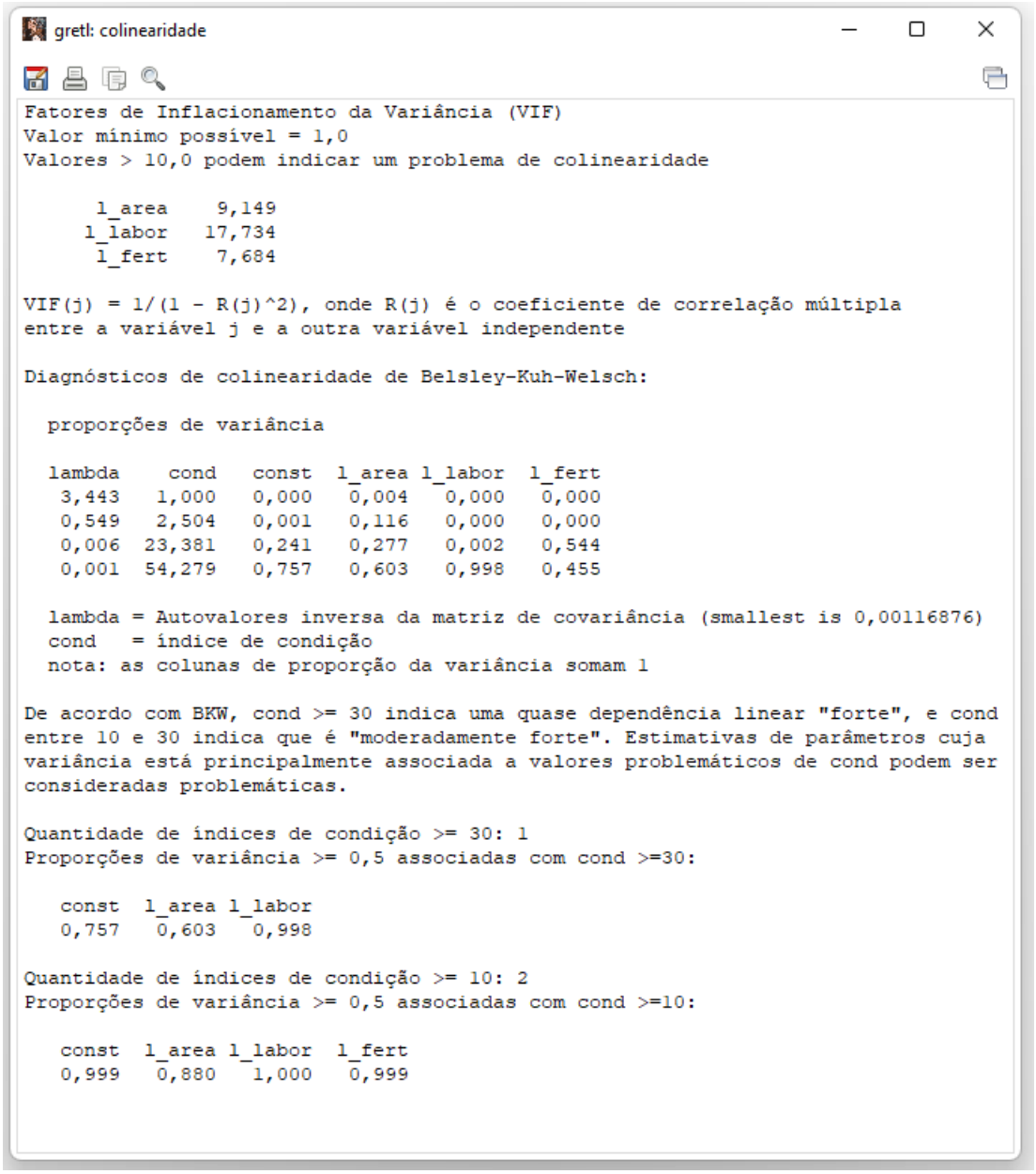


Figura 5.30: Resultados para o teste de colinearidade.

Mais uma vez, a saída do **gretl** é bastante informativa, fornece o limite para alta colinearidade ($vif_j > 10$) e a relação entre vif_j e R_j^2. Pela Figura 5.30 nota-se que esses dados são altamente colineares com o fator de inflação de variância – `vif` – para a variável independente *l_labor* acima do limite.

Para obter as estimativas dos intervalos de confiança para cada uma das inclinações, ou seja, para cada um dos coeficientes, use o comando **Análise>Intervalos de confiança para os coeficientes**, na janela do modelo (Figura 5.31). Isso abrirá a janela da Figura 5.32.

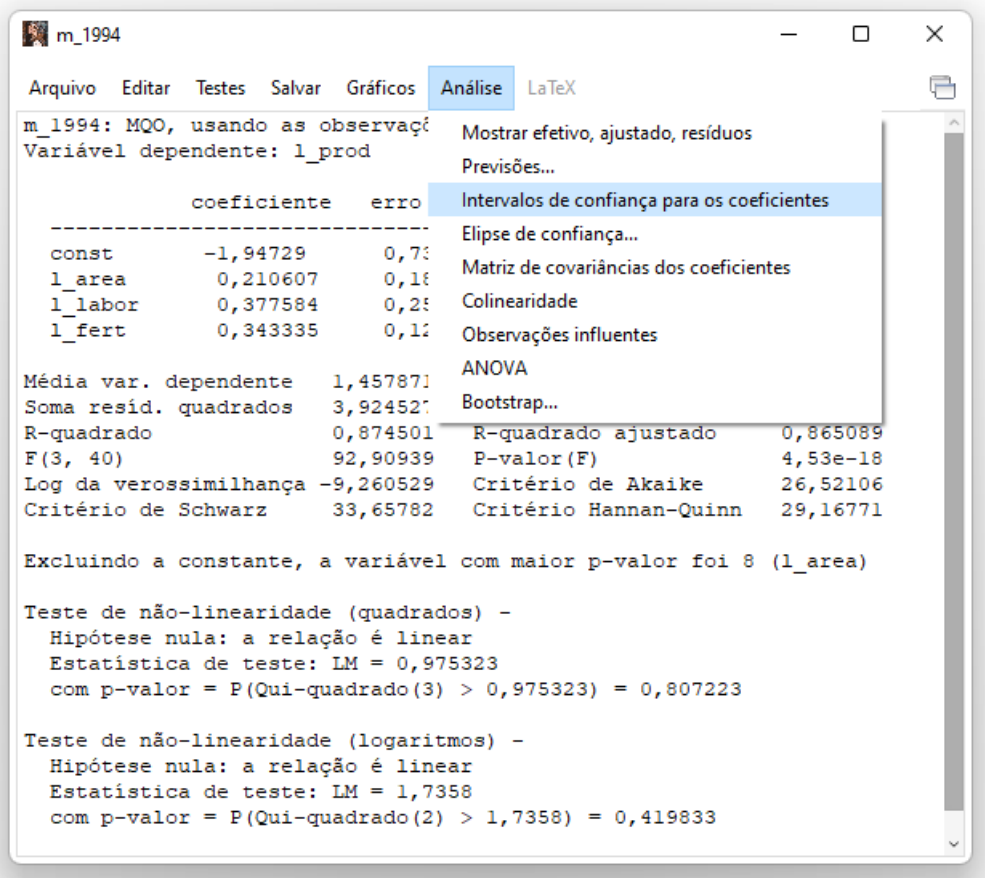


Figura 5.31: Janela do modelo de regressão.

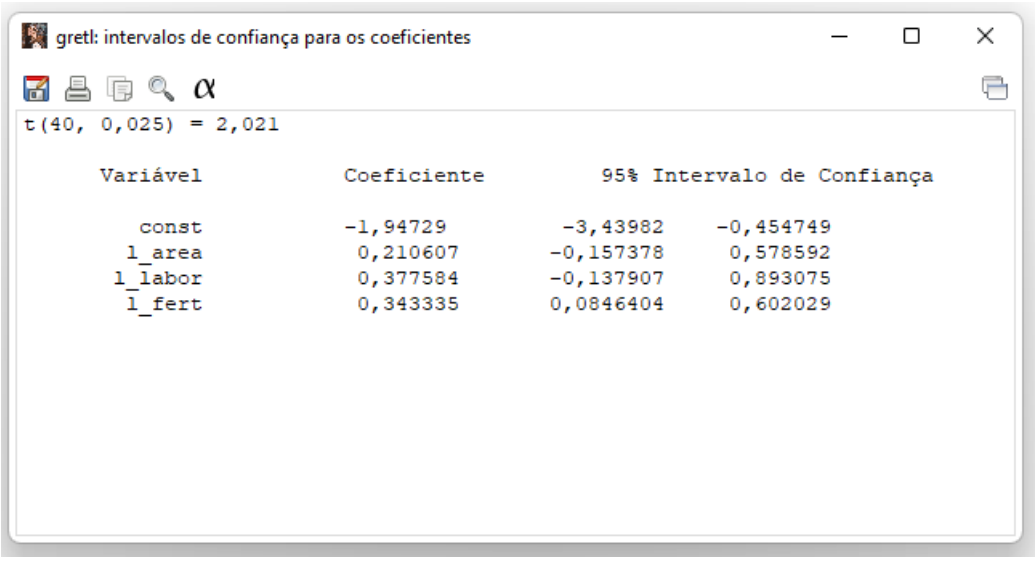


Figura 5.32: Intervalos de confiança para os coeficientes.

Uma sugestão para contornar o problema da colinearidade é impor restrições aos parâmetros do modelo. Por exemplo, suponha que se saiba que os retornos da produção de arroz sejam constantes. Isso implica então, a seguinte restrição sobre os parâmetros do modelo: $\beta_2+\beta_3+\beta_4 = 1$. Ou seja, o somatório de β_{2-4} é igual a unidade (1), Figura 5.33. Para estimar um modelo restrito veja a Seção 5.2. Note da Figura 5.33 que o somatório dos coeficientes de $\beta_{2-4} = 1$, pois $0,226228 + 0,483419 + 0,290253 = 1$.

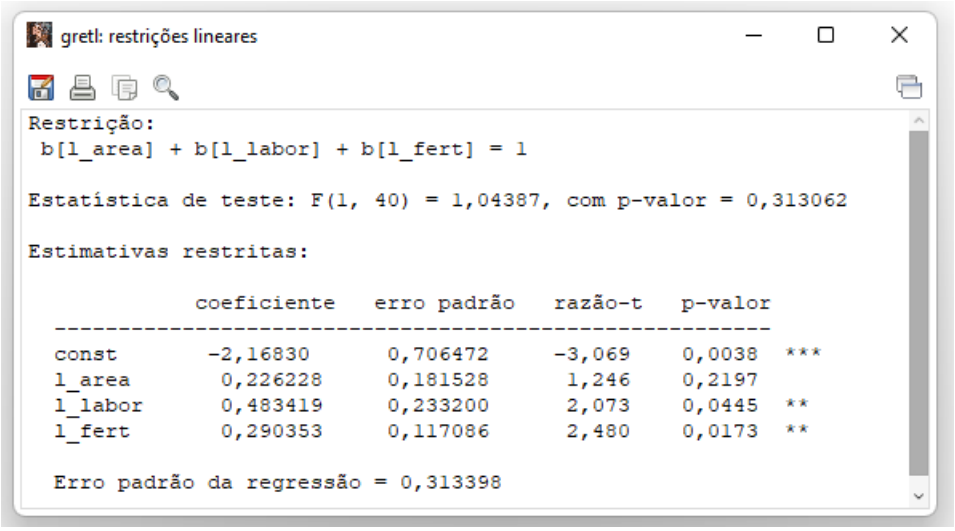


Figura 5.33: Estimativas do modelo restrito.

A restrição como hipótese nula (H_0) não é rejeita ao nível de 5%, uma vez que reportou um `p-valor` igual a $0,313062$. Ademais, no modelo restrito a variável independente `l_labor` passou a ser significativa.

Por fim, repete-se a estimativa do modelo de produção de arroz usando a amostra completa, ou seja, usando os dados para os anos de 1993 e 1994. Além disso, calcula-se o fator de inflação de variância `vif` bem como os intervalos de confiança de 95% para esse novo modelo. Os resultados para esta nova regressão são apresentados na Figura 5.34.

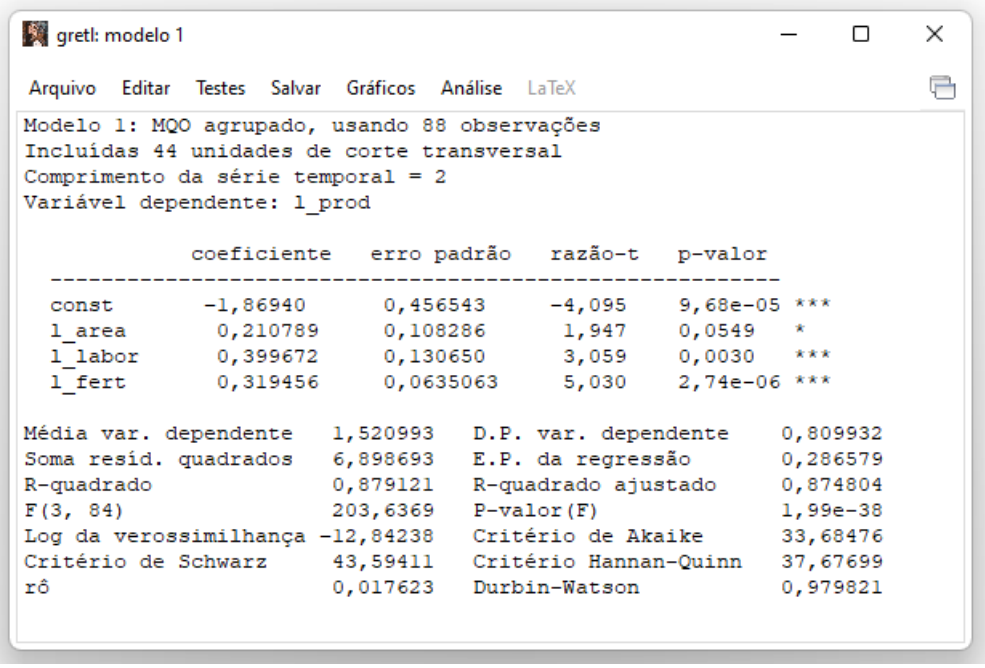


Figura 5.34: Resultados para o modelo de produção de arroz *full*.

Por sua vez, a Figura 5.35 apresenta os novos intervalos de confiança a 95% para os coeficientes. Enquanto a saída para o teste de colinearidade é apresentado na Figura 5.36. Destaca-se que o `vif` da variável `l_labor` caiu de 17,734 para 10,051, ou seja, é melhor do que o modelo para o ano de 1994. Todavia, ainda sinaliza um problema de colinearidade uma vez que é maior do que 10.

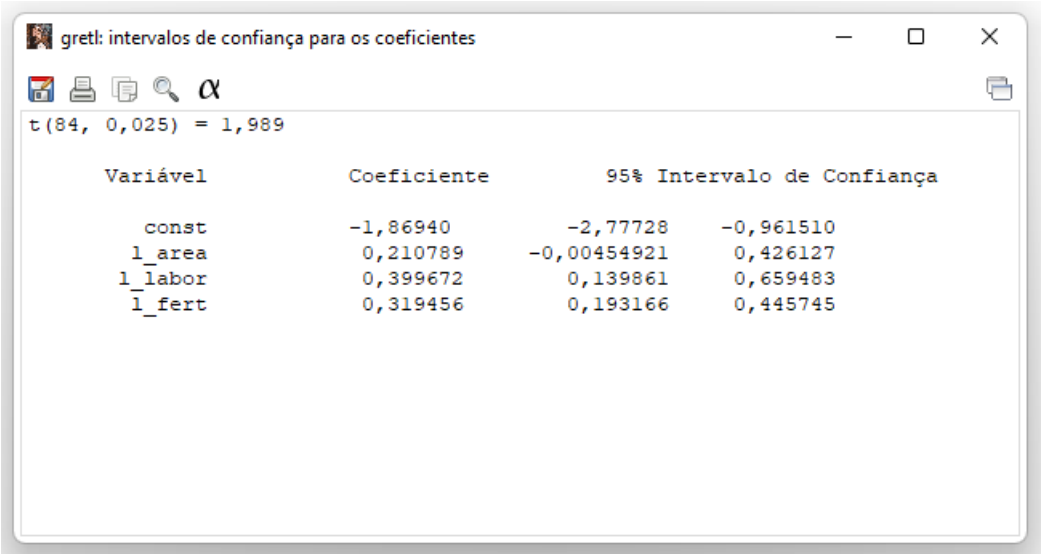


Figura 5.35: Intervalos de confiança para o modelo de produção de arroz *full*.

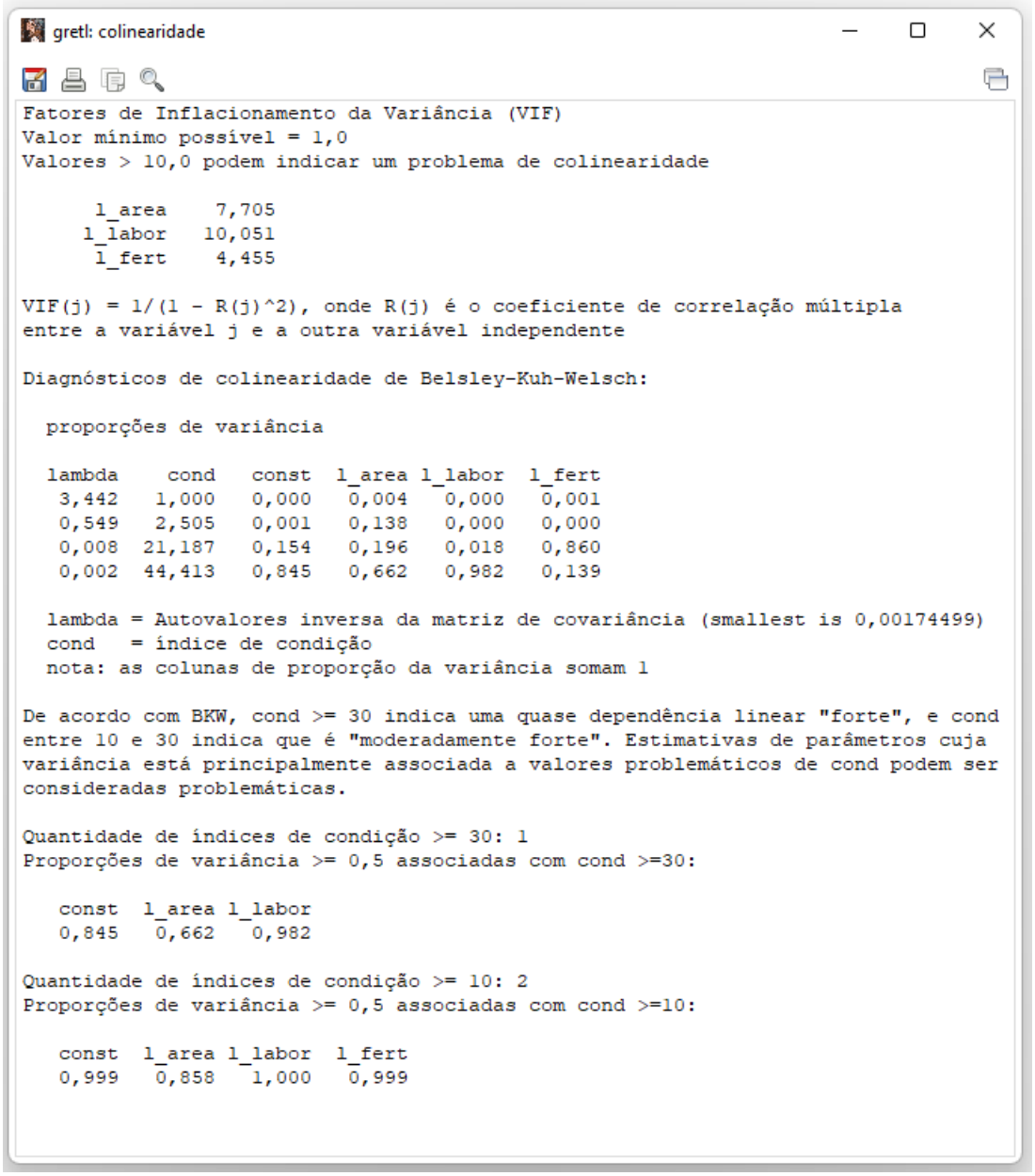


Figura 5.36: Teste de colinearidade do modelo de produção de arroz *full*.

5.4.5 Mínimos quadrados não-linear

A não linearidade nos parâmetros bem como um termo de erro aditivo implica que o modelo não pode ser estimado por Mínimos Quadrados Ordinários mas, na realidade, esses dois problemas sinalizam para estimativas de Mínimos Quadrados Não-Linear. A seguir, estima-se um modelo usando o estimador de Mínimos Quadrados Não-Linear.

$$y_t = \beta\, x_{t1} + \beta^2\, x_{t2} + e_t \tag{5.13}$$

Uma vez que o parâmetro é elevado ao quadrado (β^2) e o termo de erro é aditivo, este modelo é um candidato para estimação não-linear de mínimos quadrados, pois o mínimo da função da soma dos erros quadrados não pode ser resolvido analiticamente

para β em termos dos dados. Assim, uma solução numérica para as equações normais de mínimos quadrados deve ser encontrada.

Destaca-se que os Mínimos Quadrados Não-Linear, bem como outros estimadores não-linear, usam métodos numéricos, em vez de métodos analíticos, para minimizar a função objetivo da soma dos erros quadrados. Assim, os Mínimos Quadrados Não-Lineares requerem mais poder computacional do que a estimativa linear, entretanto, atualmente isso não é uma grande restrição devido ao avanço computacional.

No **gretl**, para estimar um modelo de Mínimos Quadrados Não-Linear o usuário deve especificar a função de regressão. Essa conterá variáveis nomeadas no conjunto de dados e um conjunto de parâmetros nomeados pelo usuário. Esses parâmetros devem ser declarados e informado seus valores (os palpites do usuário quanto ao valor que os parâmetros devam assumir). Opcionalmente, pode-se fornecer as derivadas analíticas da função de regressão em relação a cada um dos parâmetros que determinam a direção da próxima etapa. Porém, se essas derivadas não forem fornecidas, deve-se fornecer uma lista dos parâmetros a serem estimados (separados por espaço ou vírgula) e precedidos da palavra-chave `params`. Já a tolerância, o critério para o encerramento do procedimento de estimativa iterativa, pode ser ajustada usando o comando `set`.

A Equação 5.13 será estimada usando o conjunto de dados `nlls.gdt`. Com essa base carregada no **gretl**, use o comando **Modelo>Mínimos Quadrados Não-Linear (NLS)**, Figura 5.37. Isso abrirá uma janela igual a da Figura 5.38 onde será passada a estrutura do modelo a ser estimado. Ou seja:

1. A primeira linha fornece o valor inicial (o palpite do usuário) do parâmetro `b` como 1;

2. A segunda linha define a estrutura do modelo a ser estimado e;

3. A terceira linha fornece a lista dos parâmetros, que no presente caso é apenas um, `b`.

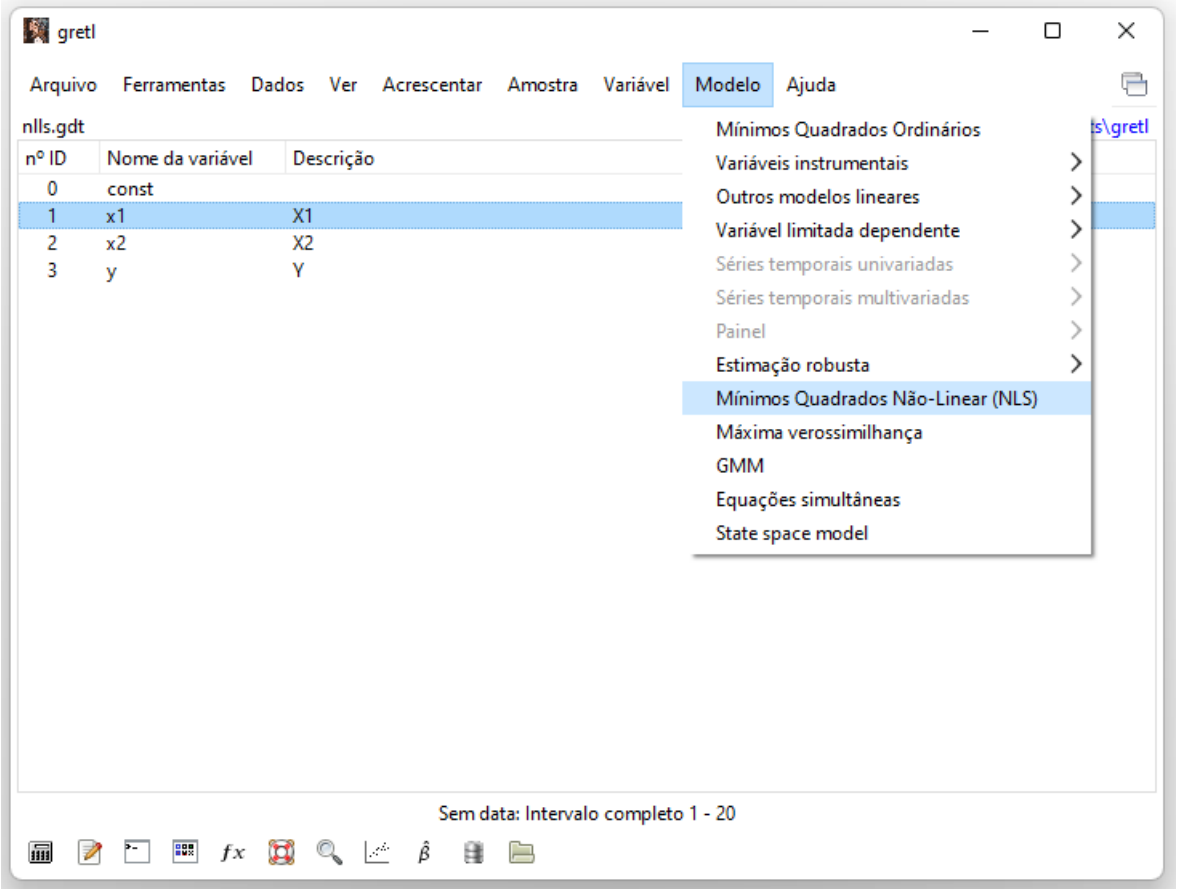


Figura 5.37: Mínimos Quadrados Não-Linear (NLS).

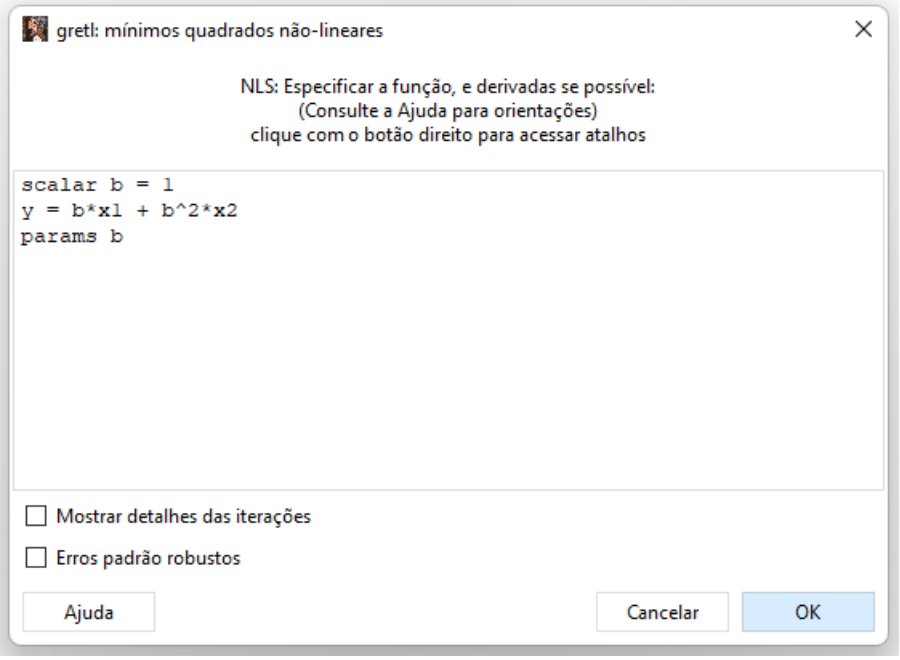


Figura 5.38: Definindo a estrutura do modelo.

Uma vez que foram repassada todas as informações necessárias clica-se no botão `OK` da Figura 5.38 que abrirá a janela com a saída do modelo de regressão, Figura 5.39.

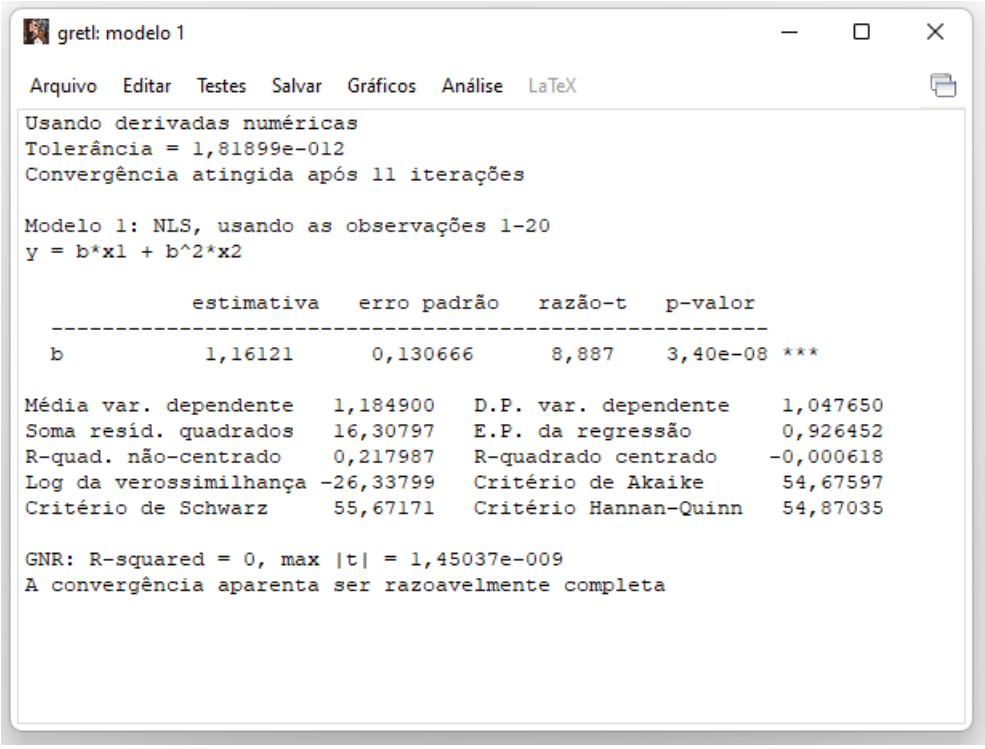


Figura 5.39: Resultado dos Mínimos Quadrados Não-Linear (NLS).

Nota-se da Figura 5.39 que a estimativa para β é de $1,6121$ enquanto o erro-padrão estimado é de aproximadamente $0,131$. Ademais, importante destacar que o R^2 centrado é negativo. Contudo, isso não deve gerar nenhuma surpresa uma vez que em modelos não-linear essa estatística não é limita entre 0 e 1.

Para uma melhor compreensão, a seguir estima-se mais um exemplo de um modelo não-linear simples, porém, esse novo modelo possui três parâmetros. Na verdade, estima-se uma curva de crescimento logístico usando dados sobre a parcela de produção total de aço bruto dos EUA que é produzida por fornos elétrico a arco disponível no conjunto de dados `steel.gdt`. O modelo é dado por:

$$y_t = \frac{\alpha}{1 + \exp\left(-\beta - \delta\, t\right)} + e_t \tag{5.14}$$

A estruturação para esse modelo de produção total de aço bruto é apresentado na Figura 5.40 enquanto a saída para o estimador de Mínimos Quadrados Não-Linear encontra-se na Figura 5.41.

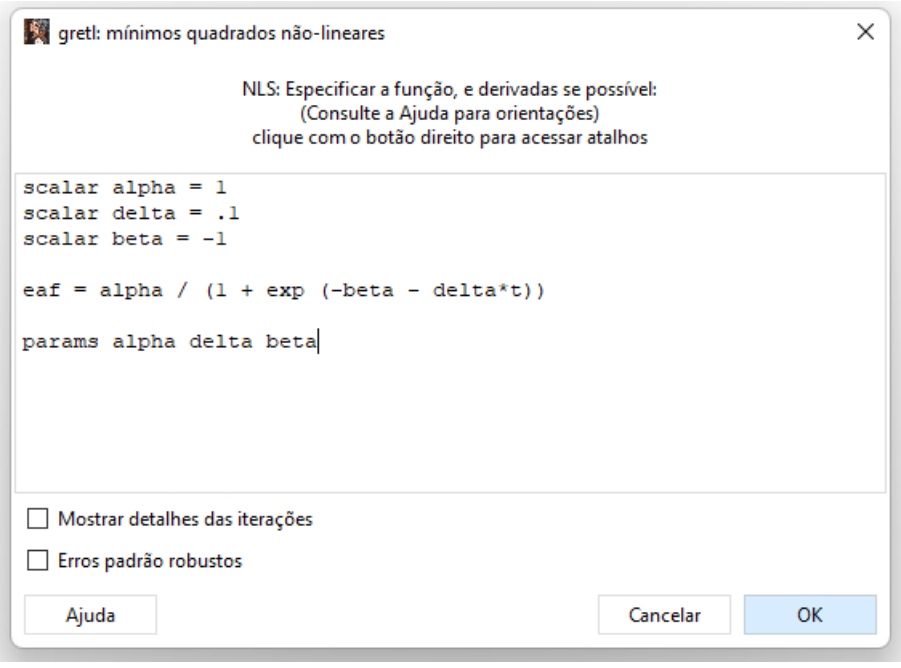


Figura 5.40: Estrutura do modelo de produção de aço.

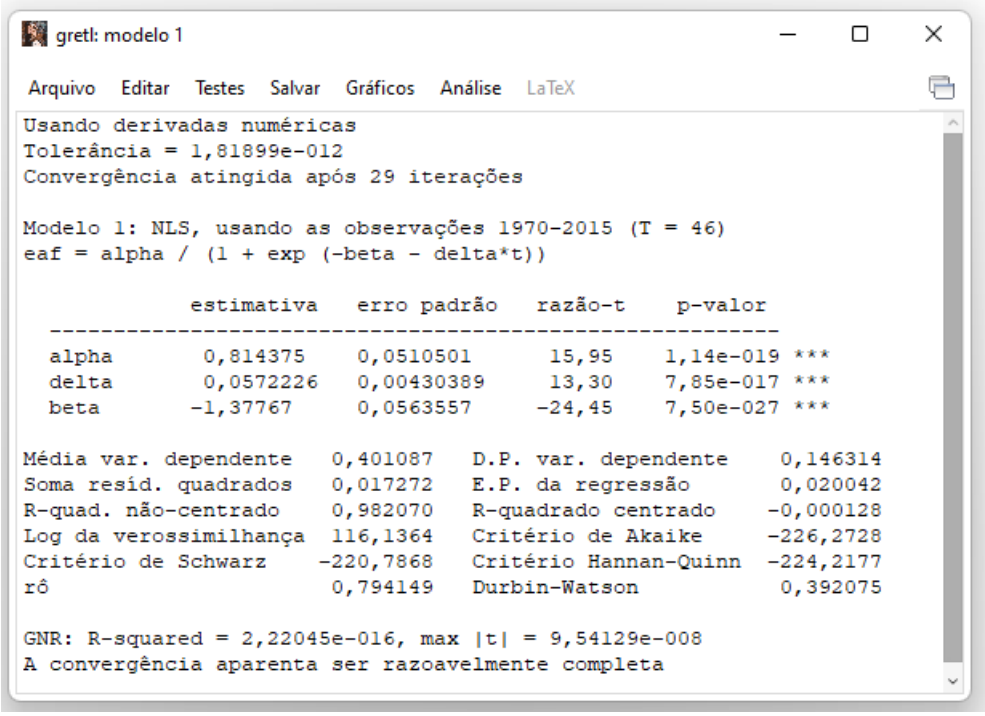


Figura 5.41: Saída do modelo de produção de aço.

Capítulo 6

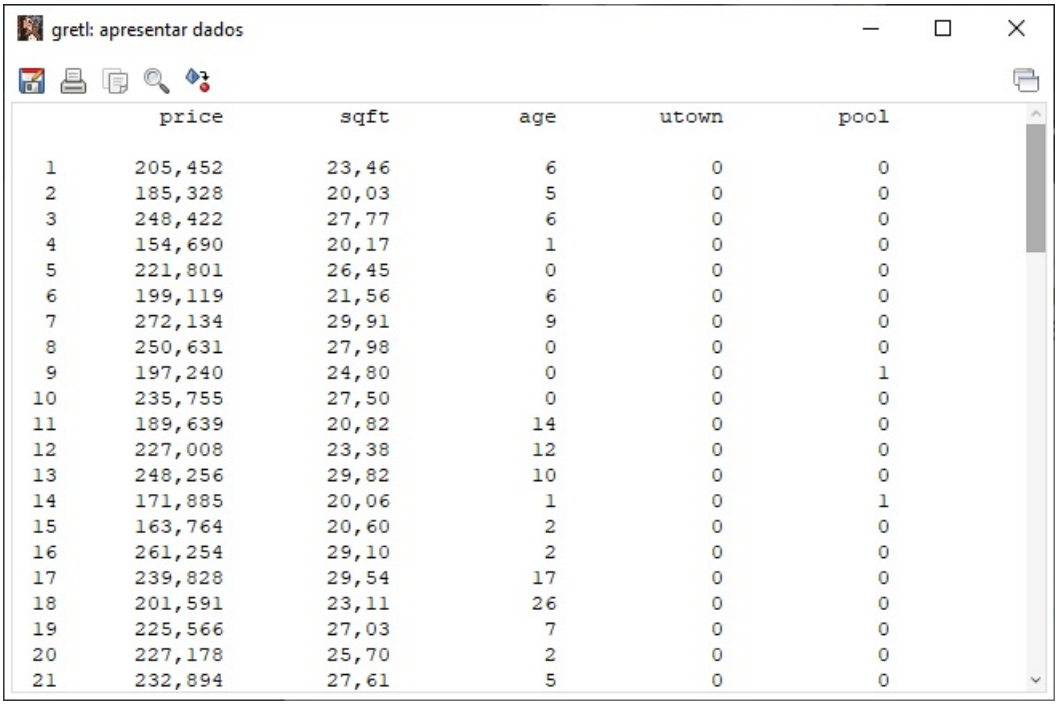
Usando variáveis indicadoras

Neste capítulo, explora-se o uso de variáveis indicadoras na análise de regressão. A discussão incluirá como criá-las, estimar modelos usando-as e como interpretar os resultados desses modelos. Também se discuti várias aplicações, as quais incluem o uso de indicadores para criar interações, indicadores regionais e realizar testes Chow de equivalência de regressão em diferentes categorias. Por fim, a utilização dessas variáveis na estimativas de modelos de probabilidade linear e na avaliação dos efeitos do tratamento e nos estimadores de diferenças em diferenças.

6.1 Variáveis indicadoras

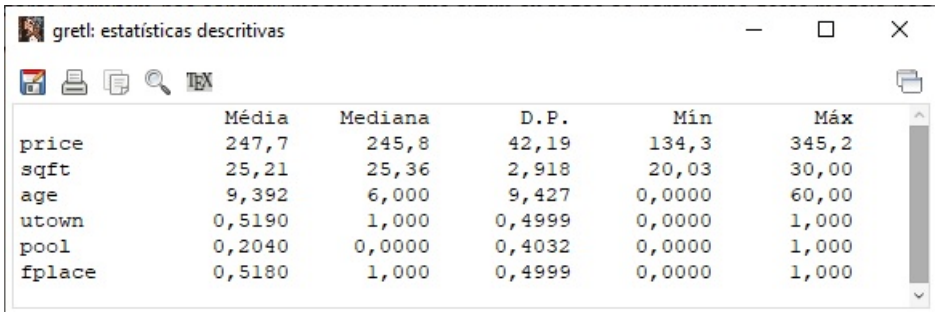
Variáveis indicadoras permitem construir modelos em que algum ou todos os parâmetros desse modelo podem mudar para um subconjunto da amostra. Uma variável indicador indica se uma determinada condição é satisfeita. Se isso é verdade a variável é igual a 1 e se não é igual a 0. Pode-se referir a elas como variáveis *dummies* e o **gretl** usa esse termo para a criação de variáveis indicadoras.

O exemplo usado nesta seção é novamente baseado nos dados imobiliários `utown.gdt`. Primeiro deve-se abrir o conjunto de dados e examiná-los. Pode-se selecionar todas as variáveis e então clicar com o botão direito do mouse na opção `Mostrar Valores`:



	price	sqft	age	utown	pool
1	205,452	23,46	6	0	0
2	185,328	20,03	5	0	0
3	248,422	27,77	6	0	0
4	154,690	20,17	1	0	0
5	221,801	26,45	0	0	0
6	199,119	21,56	6	0	0
7	272,134	29,91	9	0	0
8	250,631	27,98	0	0	0
9	197,240	24,80	0	0	1
10	235,755	27,50	0	0	0
11	189,639	20,82	14	0	0
12	227,008	23,38	12	0	0
13	248,256	29,82	10	0	0
14	171,885	20,06	1	0	1
15	163,764	20,60	2	0	0
16	261,254	29,10	2	0	0
17	239,828	29,54	17	0	0
18	201,591	23,11	26	0	0
19	225,566	27,03	7	0	0
20	227,178	25,70	2	0	0
21	232,894	27,61	5	0	0

No caso atual, seis observações são suficientes para ver que *price* e $sqft$ são contínuos, que a idade é discreta e que *utown*, *pool* e $fplace$ provavelmente são variáveis indicadoras. As estatísticas descritivas simples para toda a amostra dão uma ideia do alcance e variabilidade de *price*, $sqft$ e *idade*. As médias informam sobre as proporções de residências próximas à Universidade e que possuem piscinas ou lareiras. Para isso selecione todas as variáveis clique com o botão direito e selecione a opção **Estatísticas Descritivas>Mostrar Estatísticas Principais**.

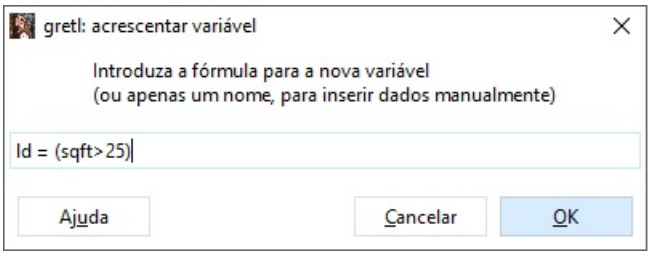


	Média	Mediana	D.P.	Mín	Máx
price	247,7	245,8	42,19	134,3	345,2
sqft	25,21	25,36	2,918	20,03	30,00
age	9,392	6,000	9,427	0,0000	60,00
utown	0,5190	1,000	0,4999	0,0000	1,000
pool	0,2040	0,0000	0,4032	0,0000	1,000
fplace	0,5180	1,000	0,4999	0,0000	1,000

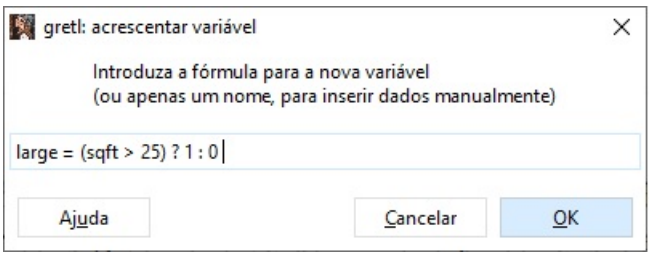
Pode-se ver que metade das casas da amostra está perto da Universidade (519 / 1000). Também é bastante claro que os preços são medidos em unidades de $ 1.000 e metros quadrados em unidades de 100. A casa mais antiga tem 60 anos e há algumas novas na amostra (idade = 0). Mínimos e máximos de 0 e 1, respectivamente, geralmente significam que se tem variáveis indicadoras na amostra. Isso confirma o que se conclui observando as primeiras observações da amostra.

6.2 Criando variáveis indicadoras

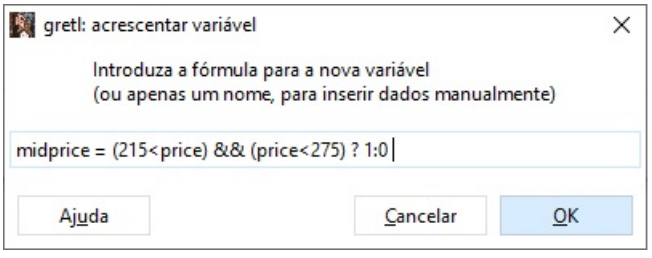
É fácil criar variáveis indicadoras utilizando o **gretl**. Suponha que se deseja criar uma variável *dummy* para indicar que uma casa é grande. Grande nesse caso significa ser maior do que 250 pés quadrados (1 pé quadrado equivale a 0,093 metros quadrados). Para isso precisa-se ir no menu **Acrescentar>Definir nova variável**:



A variável `ld` assumirá o valor 1 para todos os valores de $sqft$ maiores que 25 e será zero caso contrário. Pode-se também usar um operador condicional para criar variáveis indicadoras:

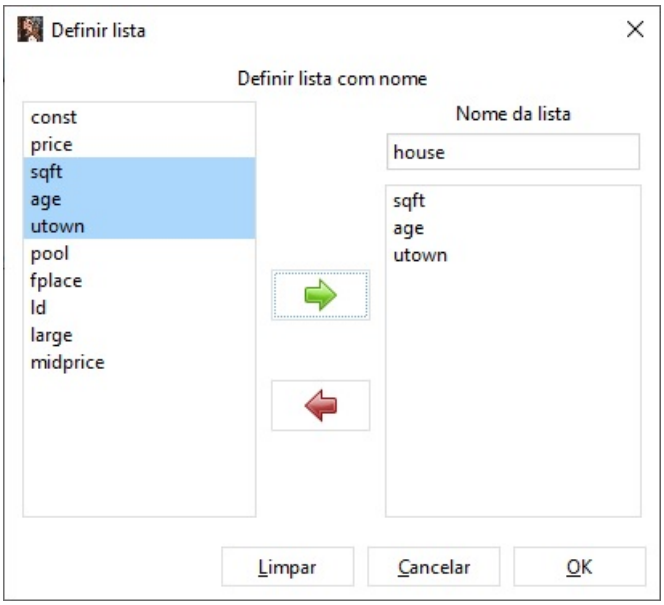


A série seria chamada de *large* e se a expressão entre parênteses for verdadeira (ou seja, a casa tiver mais de 2.500 pés quadrados), então assume o valor que segue o ponto de interrogação (?), que é 1. Se a afirmação não for verdadeira , é atribuído o valor que segue os dois pontos (ou seja, 0). O operador de atribuição condicional, também pode ser usado com lógica composta. No próximo exemplo, uma série chamada preço médio recebe o valor 1 se o preço estiver entre 215 e 275:

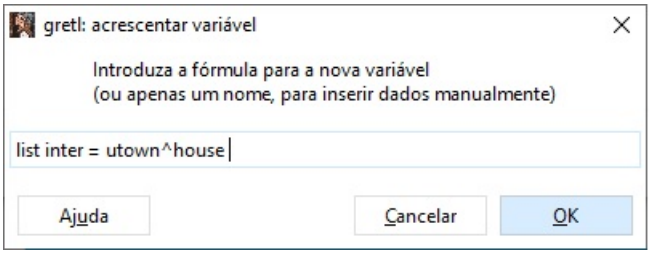


Nesse caso, a variável *midprice* receberá o valor 1 se as duas condições entre

parênteses forem verdadeiras. Finalmente, os indicadores podem interagir com outros indicadores ou variáveis contínuas usando listas. Suponha que foram criada duas listas. A primeira contém um indicador, *utown*, que é 0 se a casa não estiver localizada no bairro Cidade Universitária. A segunda lista contém indicadores contínuos e indicadores ($sqft$, *age* e *pool*). Para isso deve-se ir no menu **Dados>Criar ou editar lista:**



Para criar uma interação entre a lista *utown* e *house*, deve-se acrescentar uma nova variável e usar o seguinte comando:



Após executar esse comando, perceberá que o **gretl** criará variáveis com o final 0 e outra com o final 1. Por exemplo, *age_utown_0* repete os valores de *age* quanto *utown* é igual a zero. Já *age_utown_1* é o produto $age \times utown$, ou seja, repete os valores de *age* quando *utown* é igual a 1.

6.2.1 Estimando uma regressão

A seguinte regressão será efetuada usando como plataforma o mesmo conjunto de dados. O modelo a ser estimado é o seguinte:

$$price = \beta_1 + \beta_2\, sqft + \beta_3\, age + \delta_1\, utown + \delta_2\, pool + \delta_3\, fplace + \gamma\,(sqft \times utown) + \varepsilon$$

A saída dessa regressão é a seguinte:

gretl: modelo 1

Arquivo Editar Testes Salvar Gráficos Análise LaTeX

```
Modelo 1: MQO, usando as observações 1-1000
Variável dependente: price
Erros padrão robustos à heteroscedasticidade, variante HC1

                coeficiente   erro padrão   razão-t    p-valor
  ------------------------------------------------------------
  const          24,5000       6,35096      3,858    0,0001    ***
  sqft            7,61218      0,247540    30,75     2,00e-146 ***
  age            -0,190086     0,0524720   -3,623    0,0003    ***
  utown          27,4530       8,47187      3,240    0,0012    ***
  pool            4,37716      1,11489      3,926    9,23e-05  ***
  fplace          1,64918      0,967905     1,704    0,0887    *
  sqft_utown_1    1,29940      0,332953     3,903    0,0001    ***

Média var. dependente   247,6557   D.P. var. dependente    42,19273
Soma resid. quadrados   230184,4   E.P. da regressão       15,22521
R-quadrado              0,870570   R-quadrado ajustado     0,869788
F(6, 993)               1053,477   P-valor(F)              0,000000
Log da verossimilhança -4138,379   Critério de Akaike      8290,758
Critério de Schwarz     8325,112   Critério Hannan-Quinn   8303,815
```

O coeficiente na variável indicadora de inclinação $sqft \times utown$ é significativamente diferente de zero no nível de 5%. Isso significa que o tamanho de uma casa perto da universidade tem um impacto diferente no preço médio da casa. Com base no modelo estimado, pode-se tirar as seguintes conclusões:

- O prêmio de localização para lotes próximos à universidade é de $\$\,27.453$;
- A mudança no preço esperado por metro quadrado adicional é de US$\$\,89,12$ $(10 \times (\beta_2 + \gamma))$ perto da universidade e US$\$\,76,12$ $(10 \times \beta_2)$ em outros lugares;
- Casas depreciam $\$\,190,10$/ano $(1000 \times \beta_3)$;
- Uma piscina vale $\$\,4.377,30$ $(1000 \times \delta_2)$ e;
- Uma lareira vale $\$\,1.649,20$ $(1000 \times \delta_3)$.

6.3 Aplicando variáveis indicadoras

Nessa seção serão dados exemplos sobre a estimação e a interpretação de regressões que incluem variáveis indicadoras.

6.3.1 Interações

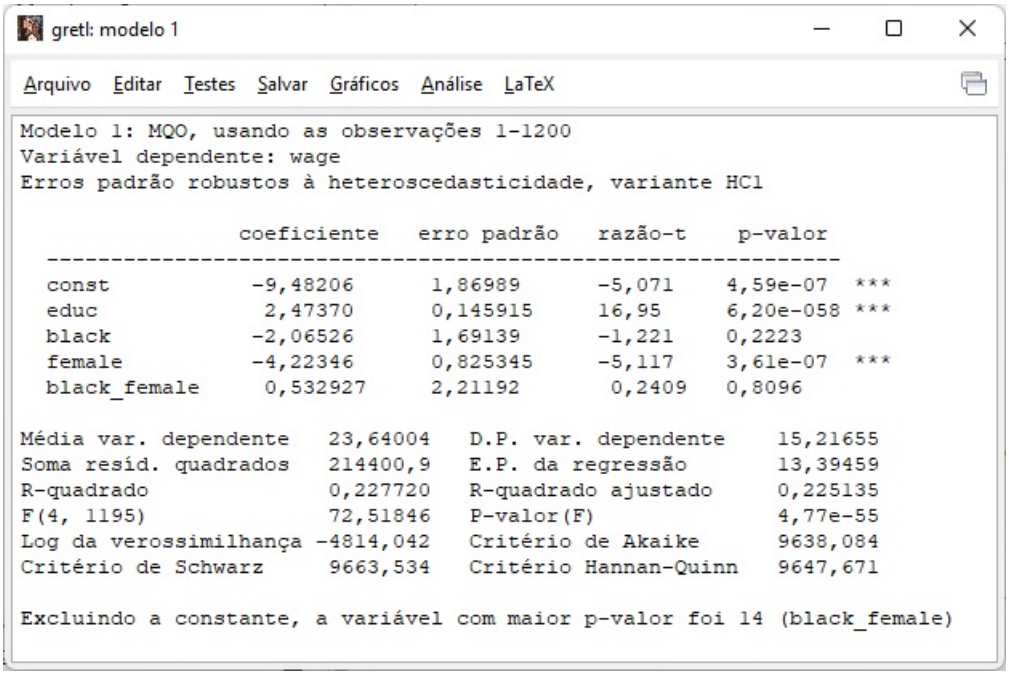
Considere a simples equação de salário:

$$wage = \beta_1 + \beta_2\, educ + \delta_1\, black + \delta_2\, female + \gamma\,(female \times black) + \varepsilon$$

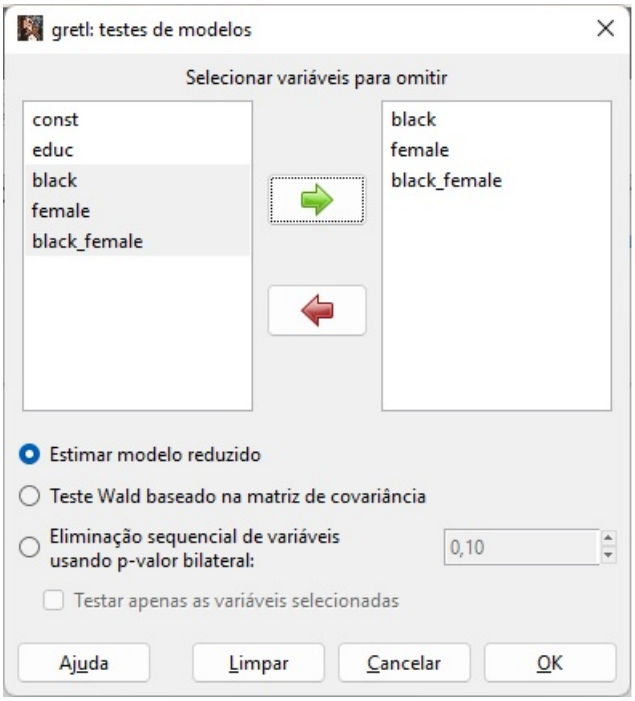
Em que *black* e *female* são variáveis indicadoras. Tomando o valor esperado do $ln\,(wage)$ tem-se os seguintes casos considerados na regressão:

$$E\,[wage \mid educ] = \begin{cases} \beta_1 + \beta_2\, educ & \text{Homens Brancos} \\ \beta_1 + \delta_1 + \beta_2\, educ & \text{Homens Negros} \\ \beta_1 + \delta_2 + \beta_2\, educ & \text{Mulheres Brancas} \\ \beta_1 + \delta_1 + \delta_2 + \gamma + \beta_2\, educ & \text{Mulheres Negras} \end{cases}$$

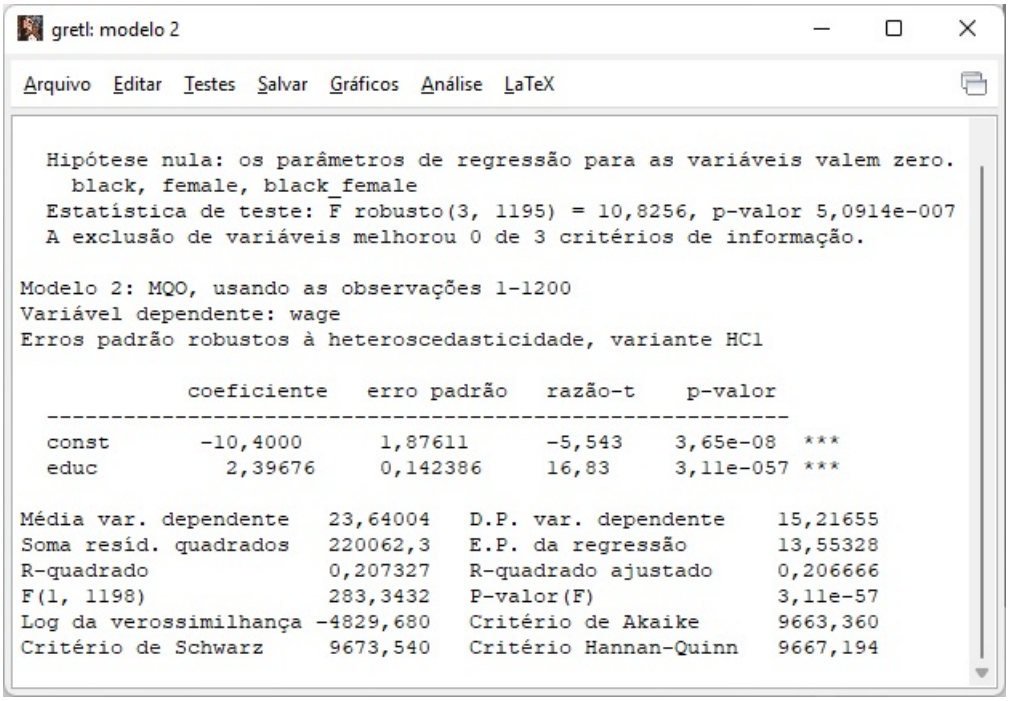
O grupo de referência é aquele em que todas as variáveis indicadoras são zero, ou seja, homens brancos. O parâmetro δ_1 mede o efeito de ser negro, em relação ao grupo de referência; δ_2 mede o efeito de ser mulher em relação ao grupo de referência, e γ mede o efeito de possuir as duas características ser mulher e ser negra. O modelo é estimado usando o arquivo `cps5_small.gdt` como segue:



Mantendo os anos de escolaridade constantes, os homens negros ganham US$\,2,07/hora a menos que os homens brancos. Para a mesma escolaridade, as mulheres brancas ganham US$\,4,22 a menos e as negras ganham US$\,0,53 a mais. No entanto, o coeficiente no termo de interação não é significativo ao nível de 5%. Pode-se testar a significância conjunta de $\delta_1 = \delta_2 = \gamma = 0$. Para isso, na tela anterior clique em **Testes>Omitir Variáveis**:



Após clicar em `ok` a seguinte saída será mostrada:



A estatística de teste é $10,82$ e o valor `p-valor` da distribuição `F` $(3, 1195)$ está bem abaixo de 5%, na verdade é praticamente zero. Dessa forma, pode-se rejeitar a hipótese nula que os três coeficientes são iguais a zero.

6.3.2 Indicadores regionais

Nesse exemplo, um conjunto de variáveis indicadoras regionais serão adicionadas ao modelo. Há quatro regiões mutuamente exclusivas a serem consideradas. O grupo de referência deve ser escolhido, nesse caso será a região nordeste. O modelo se torna:

$$wage = \begin{array}{c} \beta_1 + \beta_2\, educ + \delta_1\, black + \delta_2\, female + \gamma\,(female \times black) + \\ \theta_1\, south + \theta_2\, midwest + \theta_3\, west + \varepsilon \end{array}$$

Note que o grupo de referência é composto por homens brancos que residem na região nordeste. Todas as variáveis regionais são variáveis *dummy* (indicadoras). Tomando o valor esperado do $\ln(wage)$ tem-se os seguintes casos:

$$E\left[wage \mid educ\right] = \begin{cases} \beta_1 + \beta_2\, educ & \text{nordeste} \\ \beta_1 + \theta_1 + \beta_2\, educ & \text{sul} \\ \beta_1 + \theta_2 + \beta_2\, educ & \text{centro oeste} \\ \beta_1 + \theta_3 + \beta_2\, educ & \text{oeste} \end{cases}$$

As estimativas para o modelo completo são as seguintes:

gretl: modelo 3

Arquivo Editar Testes Salvar Gráficos Análise LaTeX

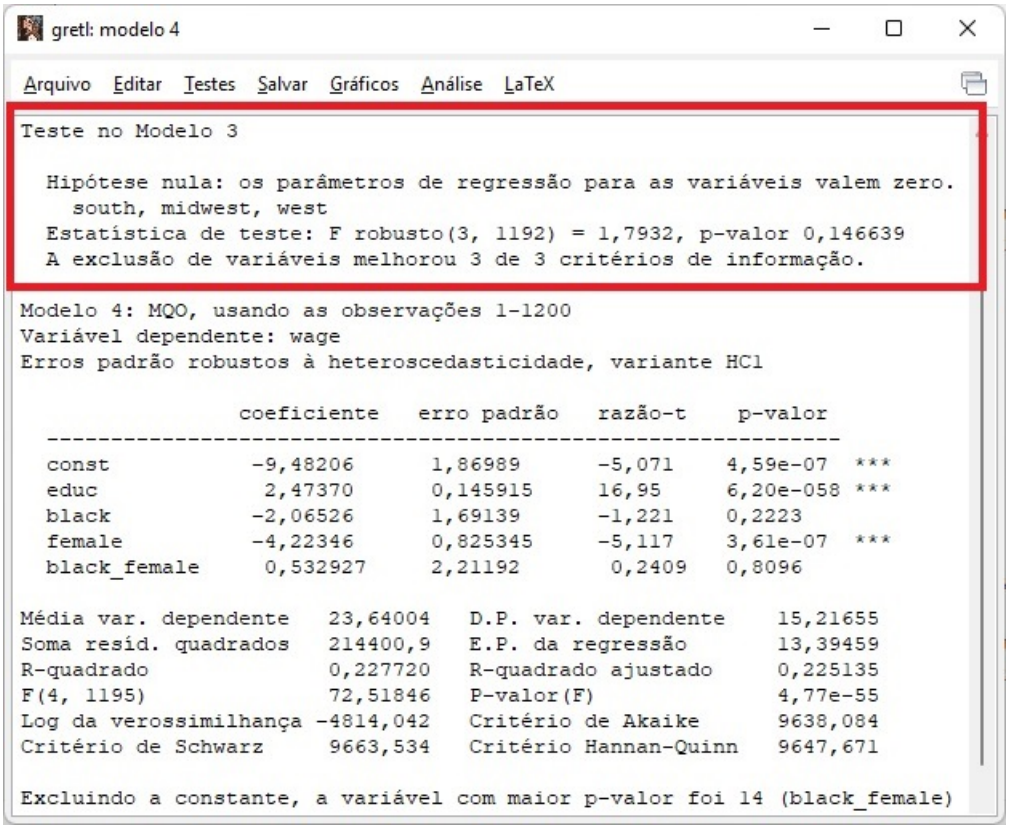
```
Modelo 3: MQO, usando as observações 1-1200
Variável dependente: wage
Erros padrão robustos à heteroscedasticidade, variante HC1

                coeficiente   erro padrão   razão-t    p-valor
  ------------------------------------------------------------
  const          -8,37082      2,01716     -4,150    3,56e-05  ***
  educ            2,46700      0,145020    17,01     2,87e-058 ***
  black          -1,87772      1,74214     -1,078    0,2813
  female         -4,18605      0,826252    -5,066    4,70e-07  ***
  black_female    0,618998     2,19979      0,2814   0,7785
  south          -1,65226      1,14902     -1,438    0,1507
  midwest        -1,93920      1,03560     -1,873    0,0614    *
  west           -0,145190     1,15163     -0,1261   0,8997

Média var. dependente   23,64004   D.P. var. dependente    15,21655
Soma resid. quadrados   213552,1   E.P. da regressão       13,38486
R-quadrado              0,230777   R-quadrado ajustado     0,226260
F(7, 1192)              44,15414   P-valor(F)              1,08e-55
Log da verossimilhança -4811,662   Critério de Akaike      9639,324
Critério de Schwarz     9680,044   Critério Hannan-Quinn   9654,663

Excluindo a constante, a variável com maior p-valor foi 10 (west)
```

Espera-se que os trabalhadores do sul ganhem US$ 1,65 a menos por hora do que os do nordeste mantendo outras variáveis constantes. No entanto, nenhum dos indicadores regionais é individualmente significativo a 5%. Os resultados do teste conjunto são:



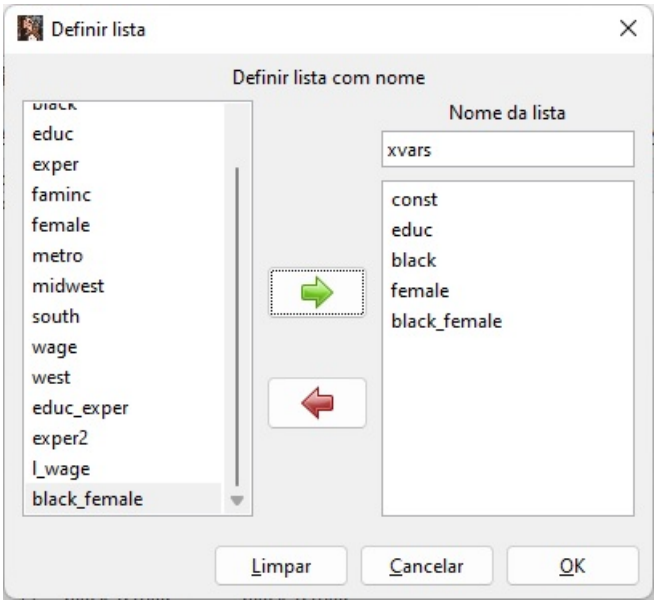
A estatística de teste tem uma distribuição `F` $(3, 1192)$ e é igual a $1,79$. O `p-valor` é superior a 5% e, assim, conclui-se que os indicadores não são conjuntamente significativos. Dessa forma, não foi possível concluir que os trabalhadores com mesma escolaridade, raça e gênero recebem salários por hora diferentes entre entre as regiões analisadas.

6.3.3 Testando a equivalência entre duas regiões

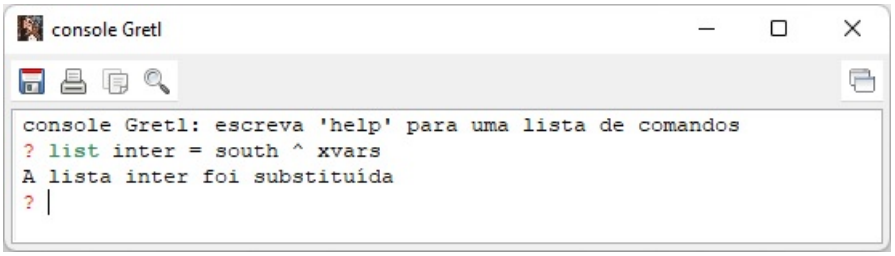
Pode-se levantar o seguinte questionamento: os salários recebidos no sul são diferentes para as demais regiões do país? Há várias formas de verificar isso no **gretl**. Pode-se utilizar a interação enter variáveis indicadoras ou estimar diferentes modelos com subamostras. Ainda, pode-se realizar o `teste de Chow` que permite testar a equivalência de regressões de subamostras com base em uma variável indicadora. Para ilustrar isso, considere o seguinte modelo de salários:

$$wage = \beta_1 + \beta_2\, educ + \delta_1\, black + \delta_2\, female + \gamma\,(female \times black) + \varepsilon$$

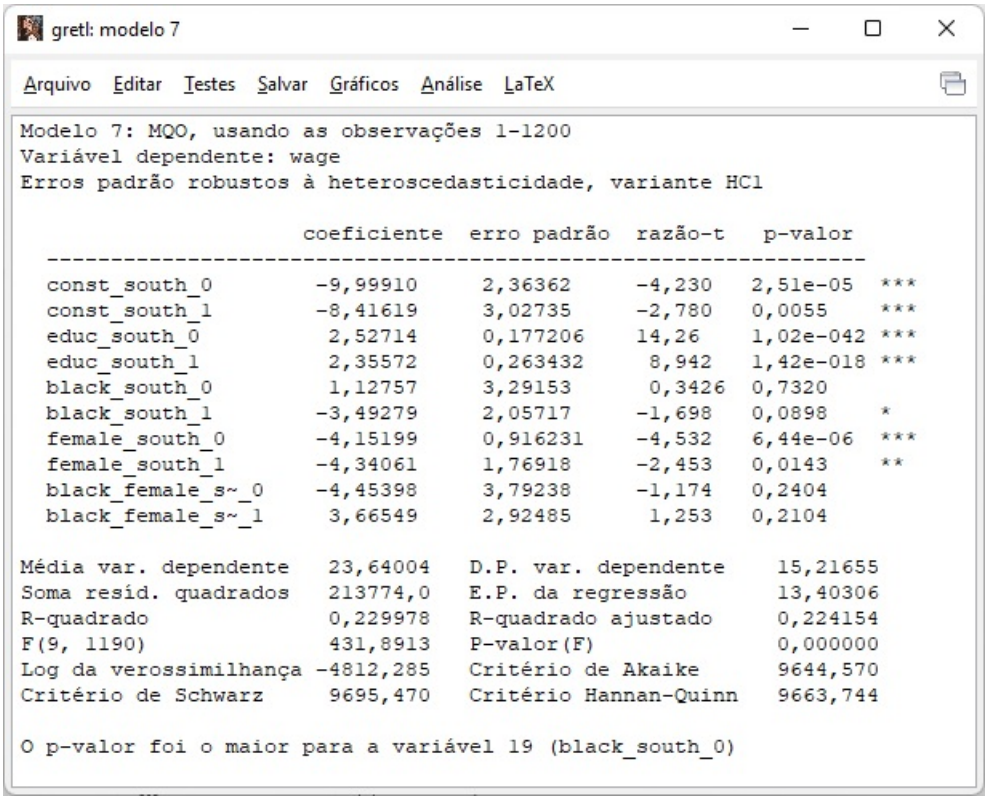
Se os salários são determinados de forma diferente na região sul, então as inclinações e os interceptos devem ser diferentes. Primeiro cria-se uma lista chamada `xvars`:



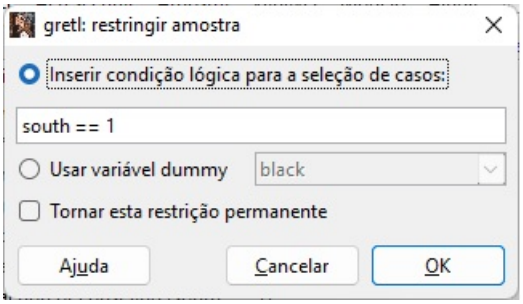
Depois faz a interação dessa lista com a variável *south*. Para isso pode acrescentar uma nova variável e digitar o comando abaixo ou utilizar o próprio console do **gretl**:



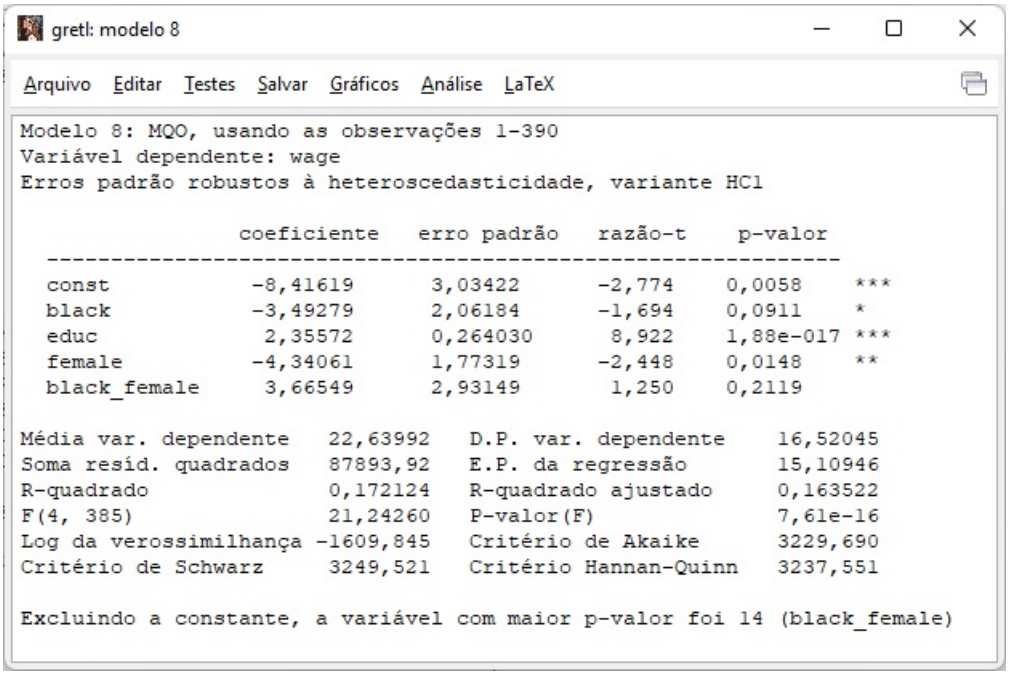
Posteriormente deve-se estimar uma regressão utilizando essas variáveis de interação:



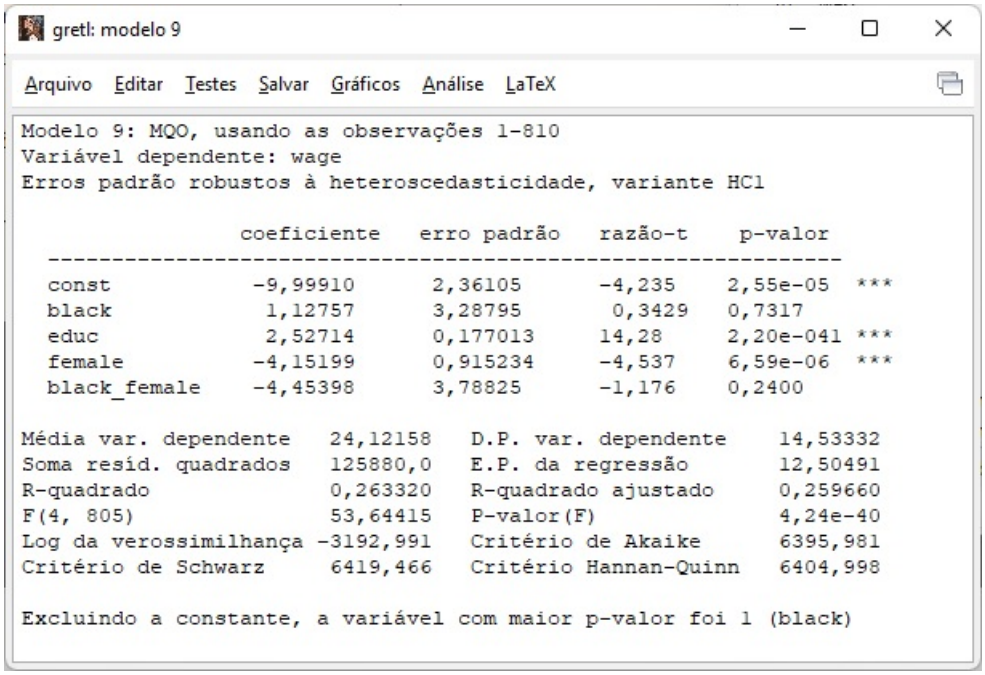
Ao interagir cada uma das variáveis, incluindo a constante, com o indicador, estimamos essencialmente duas regressões separadas em um único modelo. Observe que os erros padrão são calculados com base na suposição de que as duas subamostras têm a mesma variância geral, σ^2. Agora deve-se estimar duas equações separadamente, uma para amostra restrita aos salários recebidos pelos trabalhadores que residem na região sul e uma mostra para os trabalhadores das outras regiões. Para isso, deve-se clicar no menu **Amostra>Restringir baseado em critérios:**



A seguir estima-se o modelo para a amostra restrita a região sul:



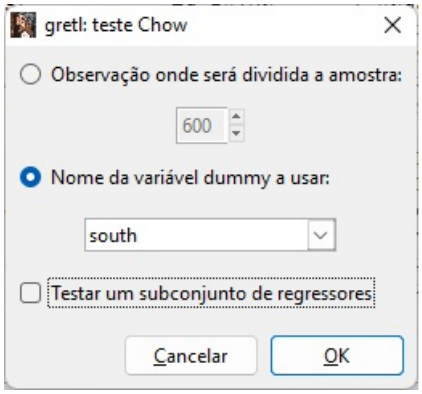
Para as outras regiões, deve-se clicar no menu **Amostra>Restaurar intervalo completo**. Depois, repeti-se o procedimento anterior restringindo o intervalo para $south == 0$ e reestima-se o modelo:



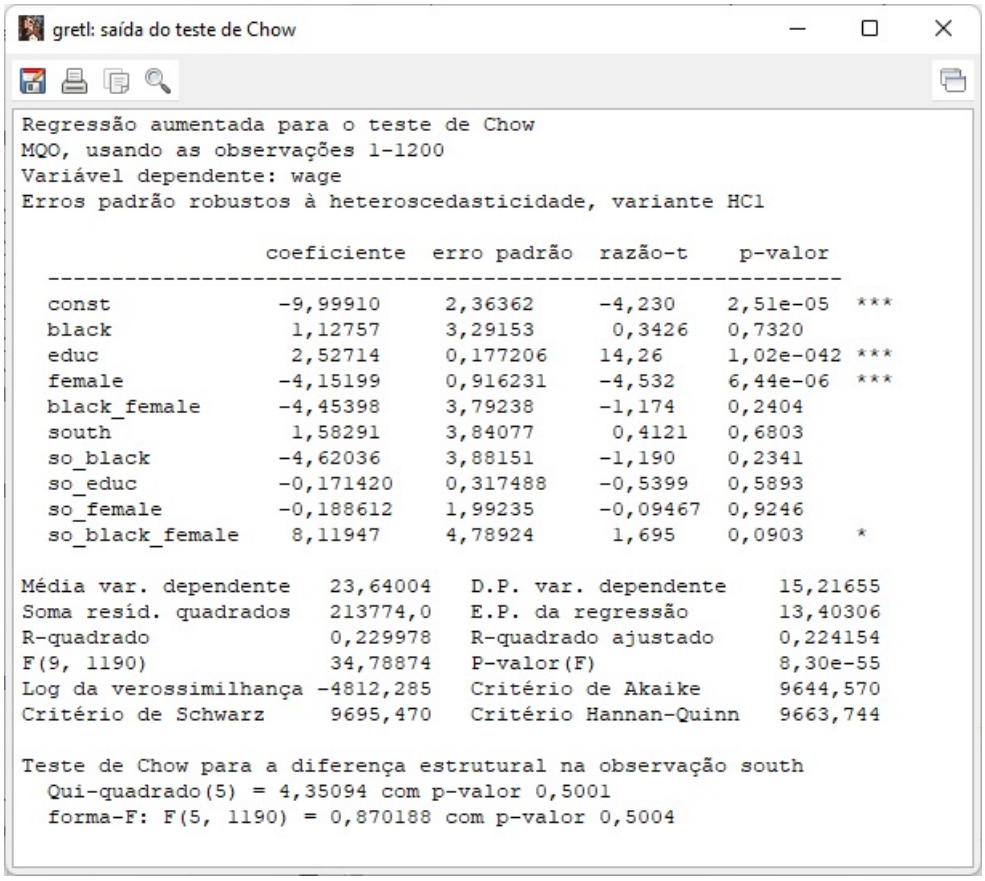
As estimativas dos coeficientes coincidem com aquelas obtidas por meio dos indicadores. Como esperado, os erros padrão são diferentes.

Um `teste de Chow` é usado para verificar a presença de quebras estruturais ou alterações em uma regressão. Em outras palavras, esse procedimento testa se uma

subamostra possuí um intercepto e uma inclinação diferentes de outra. Ele pode ser usado para detectar quebras estruturais em modelos de séries temporais ou para determinar se, no exemplo em questão, os salários do sul são determinados de forma diferente dos do resto do país. Para realizar o teste estime o modelo por Mínimos Quadrados e clique no menu **Teste>Teste de Chow.**



Após isso tem-se a seguinte saída:



Observe que `p-valor` associado ao teste é $0,625$, fornecendo evidências insuficientes para convencer de que os salários são estruturalmente diferentes no sul.

6.3.4 Modelos log-lineares com variáveis indicadores

Nesse exemplo, uma variável indicadora é incluída num modelo log linear. Para tanto, basea-se no modelo do exemplo anterior:

$$\ln(wage) = \beta_1 + \beta_2\, educ + \delta_1\, female + \varepsilon$$

A estimação do modelo por mínimos quadrados permite computar a diferença percentual entre os salários entre homens e mulheres. Com um pouco de álgebra pode-se verificar que essa diferença percentual é:

$$100\left(e^{\widehat{\delta}-1}\right)\%$$

Para isso suponha que $female = 0$:

$$\ln(wage) = \beta_1 + \beta_2\, educ + \varepsilon$$

Subtraia as duas equações:

$$\begin{aligned} \ln(wage_f) &= \beta_1 + \beta_2\, educ + \delta_1 + \varepsilon \\ - & \\ \ln(wage_{sf}) &= \beta_1 + \beta_2\, educ + \varepsilon \end{aligned}$$

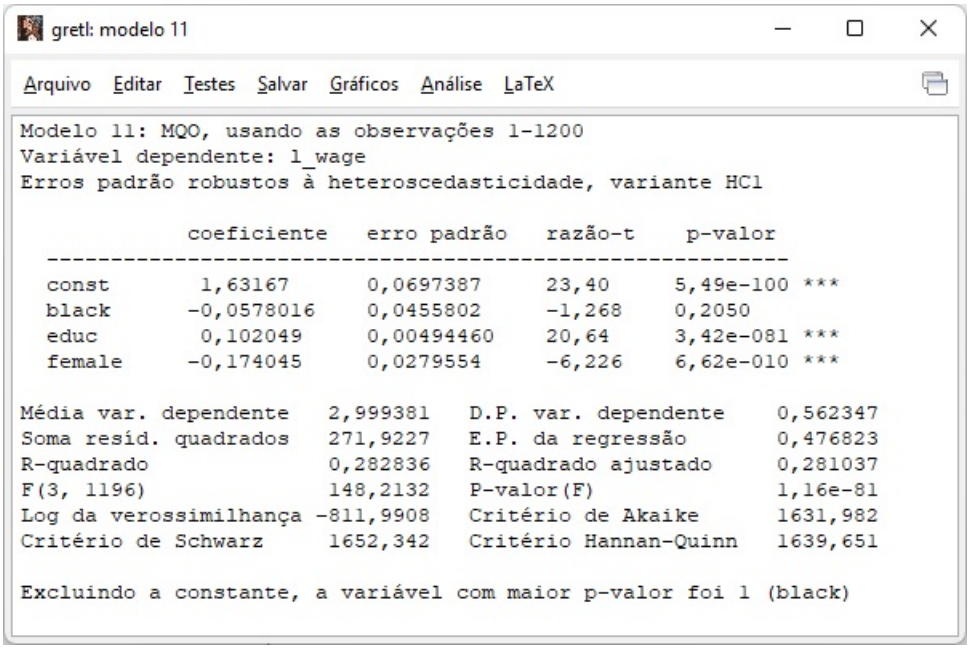
O que resulta em:

$$\ln\left(\frac{wage_f}{wage_{sf}}\right) = \delta_1$$

Subtraindo 1 dos dois lados, aplicando o exponencial e multiplicando por 100:

$$\Delta\, wage = 100 \times \exp(\delta_1 - 1)$$

Assim pode-se estimar o modelo:



O coeficiente de escolaridade sugere que um ano adicional de escolaridade aumenta o salário médio em $10,24\%$, mantendo o sexo constante. O diferencial salarial estimado entre homens e mulheres de escolaridade semelhante é de $17,78\%$. Usando a equação para computar a diferença percentual obtem-se o valor de -16.29. Esse número sugere que as mulheres ganham cerca de $16,29\%$ menos do que os homens que têm níveis de educação semelhantes.

6.4 Modelo de probabilidade linear

O modelo de probabilidade linear é uma regressão que a variável dependente é uma indicadora. Esse modelo pode ser estimado por mínimos quadrados. Suponha que:

$$y_i = \begin{cases} 1 \text{ se a alternativa é escolhida} \\ 0 \text{ se a alternativa não é escolhida} \end{cases}$$

Adicionalmente, suponha que $Pr\left(y_i = 1\right) = \pi_i$. Para uma variável discreta:

$$E\left[y_i\right] = 1 \times Pr\left(y_i = 1\right) + 0 \times Pr\left(y_i = 0\right) = \pi_i$$

Dessa forma, a média de uma variável aleatória binária pode ser interpretada como uma probabilidade, isto é, a probabilidade que $y = 1$.

Quando a regressão: $E\left[y_i \mid x_{i2}, x_{i3}, \ldots, x_{iK}\right]$ é linear então:

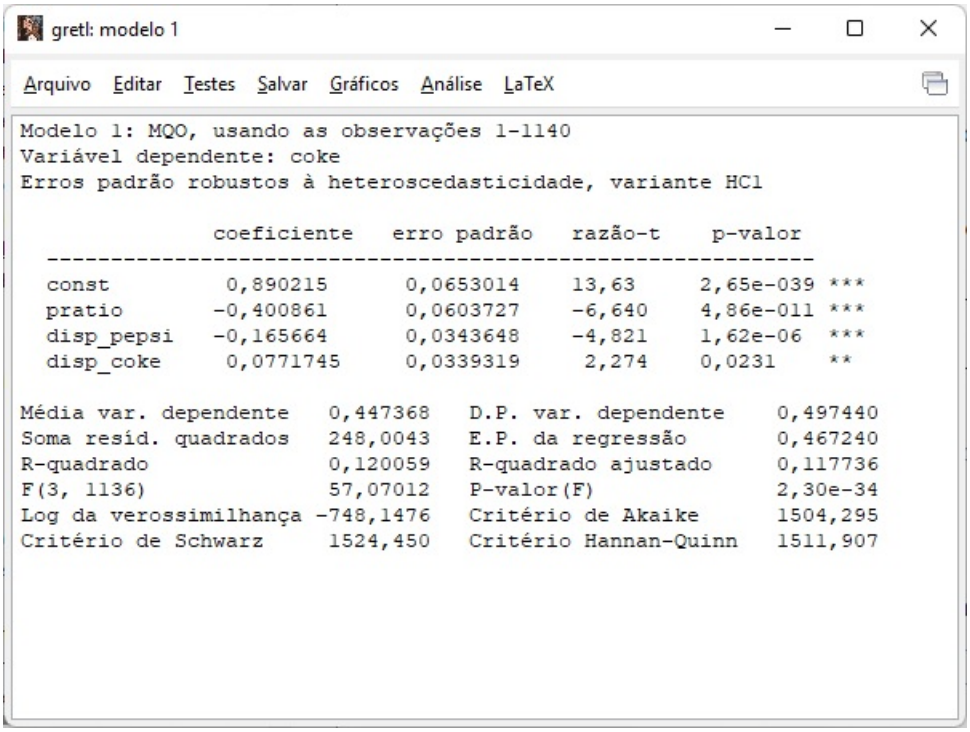
$$E\left[y_i\right] = \beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \cdots + \beta_k\, x_{iK}$$

A variância de uma variável binária é:

$$var\left[y_i\right] = \pi_i\left(1 - \pi_i\right)$$

O que significa que será diferente para cada indivíduo. Substituindo a probabilidade não observada $E(y_i)$, com a variável indicadora observar isso requer adicionar um termo de erro ao modelo que pode ser estimado via mínimos quadrados ordinários.

No exemplo a seguir, utiliza-se o arquivo `coke.gdt`, que contém 1.140 observações de indivíduos que compraram Coca-Cola ou Pepsi. A variável dependente assume o valor 1 se a pessoa comprar Coca-Cola e 0 se Pepsi. Estes dependem da relação dos *prices*, *pratio*, e duas variáveis indicadoras, *disp_coke* e *disp_pepsi*. Estas variáveis indicam se a loja que vende as bebidas tinha *folders* promocionais de Coca-Cola ou Pepsi no momento da compra. As estimativas são mostradas a seguir:



O modelo foi estimado usando um estimador de matriz de variância-covariância que é consistente quando os termos de erro do modelo possuem variâncias que dependem da observação. Esse é o caso aqui.

6.5 Efeito do tratamento

Com o propósito de entender o impacto dos efeitos do tratamento, considere um simples modelo de regressão no qual a variável explicativa é uma *dummy*, indicando quando um indivíduo em particular está no grupo de tratamento ou de controle. Seja y a variável de resultado, que mede a característica que deve ser afetada pelo tratamento. Defina a variável indicadora d como:

$$d_i = \begin{cases} 1 & \text{se é tratado} \\ 0 & \text{se não é tratado} \end{cases}$$

O efeito do tratamento na variável de resultado pode ser modelado como:

$$y_i = \beta_1 + \beta_2 d_i + e_i \qquad \text{para } i = 1, 2, \ldots, N$$

sendo e_i a coleção de outros fatores que afetam a variável de resultado. As funções de tratamento para os grupos de tratamento e de controle são:

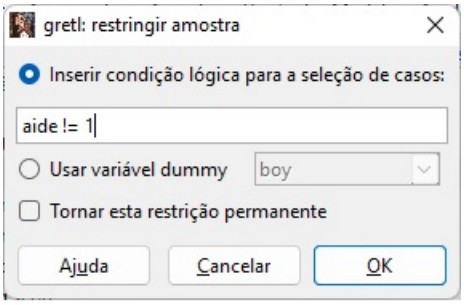
$$E(y_i) = \begin{cases} \beta_1 + \beta_2 & \text{se o indivíduo é tratado} \\ \beta_1 & \text{se não é tratado} \end{cases}$$

O efeito do tratamento que se deseja medir é β_2. O estimador de mínimos quadrados de β_2 é:

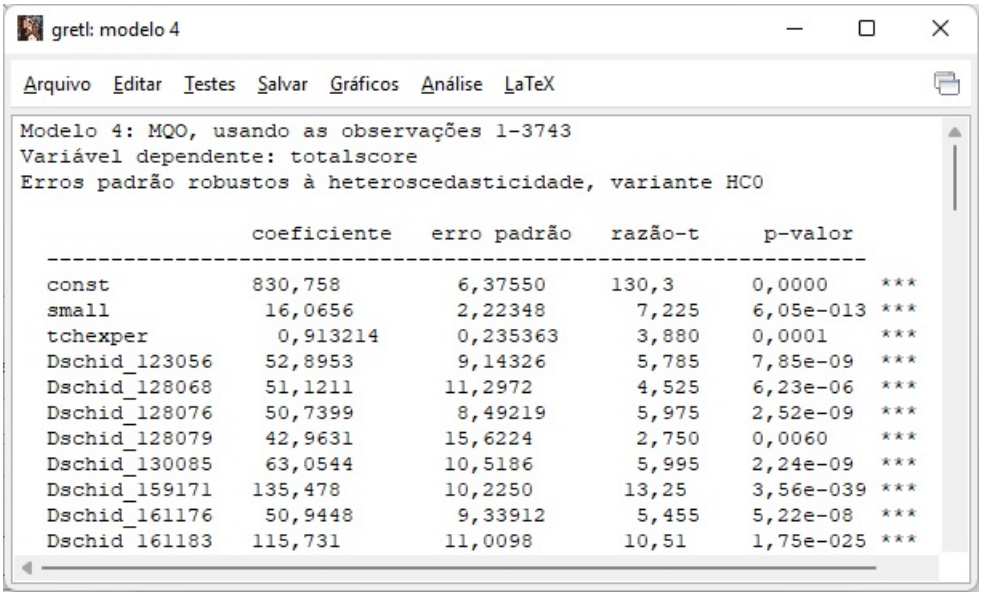
$$b_2 = \frac{\sum_{i=1}^{N}(d_i - \bar{d})(y_i - \bar{y})}{\sum_{i=1}^{N}(d_i - \bar{d})^2} = \bar{y}_1 - \bar{y}_0$$

em que $\bar{y}_1$ é a média das observações de y para o grupo de tratamento e $\bar{y}_0$ é a média amostral para as observações do grupo não tratamento. Nessa abordagem de tratamento/controle o estimado b_2 é chamado de estimador de diferença por causa da diferença entre as médias amostrais dos grupos de controle e de tratamento.

Para exemplificar esse modelo, utiliza-se o arquivo `star.gdt`. Primeiramente, deseja-se descartar as observações para as salas de aula que possuem professor auxiliares. Para isso deve-se restringir a amostra da seguinte forma:



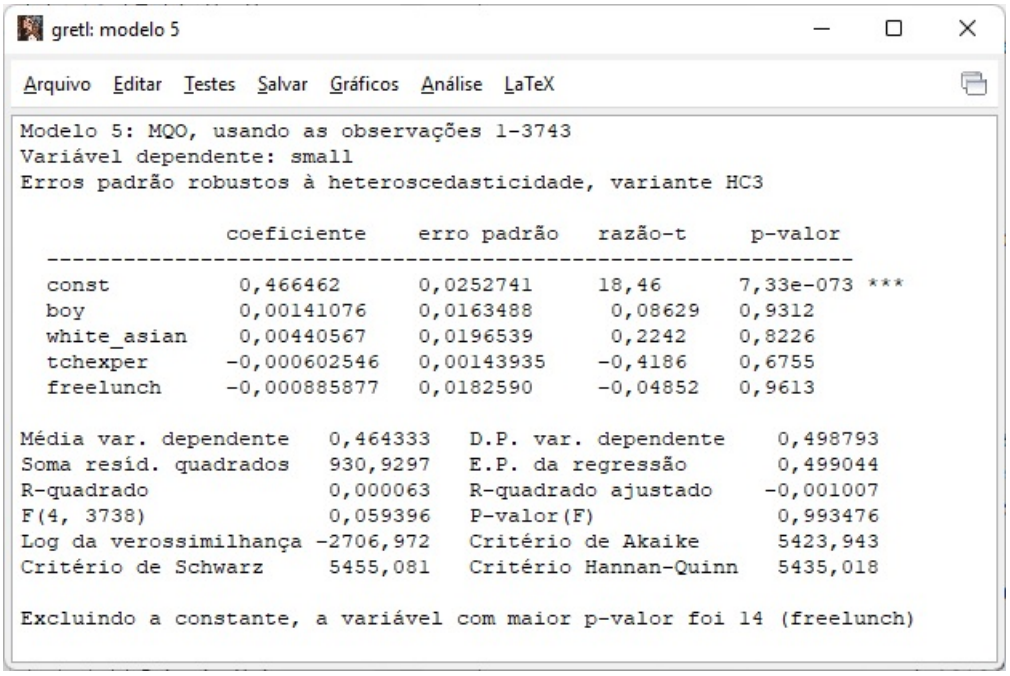
Além disso, pode ser que a atribuição de grupos de tratamento esteja relacionada a uma ou mais das características observáveis (tamanho da escola ou experiência do professor). Uma maneira de controlar esses efeitos omitidos é usar a estimativa de efeitos fixos. Aborda-se esse ponto com mais detalhes posteriormente. Os efeitos fixos de escola, nada mais são do que variáveis *dummy* que identificam cada escola. Para isso, clique com o botão direito do mouse na variável *schid* e selecione a opção `Transformar em dummy`. Em seguida escolha a primeira opção, `Codificar todos os valores` e aperte no botão `Ok`. Posteriormente estime um modelo de mínimos quadrados, com a seguinte configuração. Não esqueça de retirar a primeira *dummy* criada para identificar a escola, pois ela será utilizada como grupo de referência.



Com essa estimativa verifica-se o impacto do efeito de uma turma pequena (*small*) no escore total do aluno (*totalscore*). Na regressão utiliza-se como controle a experiência do professor e também adiciona-se os efeitos fixos de escola. Observe que esses efeitos fixos são significativos. Em média, pode-se dizer que os escores de leitura e de matemática dos alunos que estudam em turmas pequenas são 16.06 pontos mais altos do que aqueles que estudam em turmas "grandes".

6.5.1 Usando um modelo de probabilidade linear para verificar a atribuição aleatória

No modelo estimado para medir o efeito do tratamento das turmas pequenas, foi omitido muitas variáveis do modelo. Esse procedimento é seguro fazê-lo considerando que essas variáveis não estejam correlacionadas com regressores. Caso fossem correlacionadas, isso seria uma evidência que as atribuições ao grupo de controle são sistemáticas. Para verificar esse fato, pode-se usar uma regressão. Como *small* é uma variável *dummy*, usa-se uma regressão de probabilidade linear. As variáveis independentes são *boy*, *white_asian*, *tchexper* e $freelunch$.



Pode-se observar que a estatística `F` não é significativa a 10%. Nenhuma das razões `t-individuais` é significativa. Esses resultados sugerem que a atribuição das crianças em turmas pequenas ou grandes é totalmente aleatório, algo como jogar uma moeda. Dessa forma, pode-se considerar seguro omitir essas variáveis explicativas do modelo de regressão.

6.6 Diferenças em diferenças

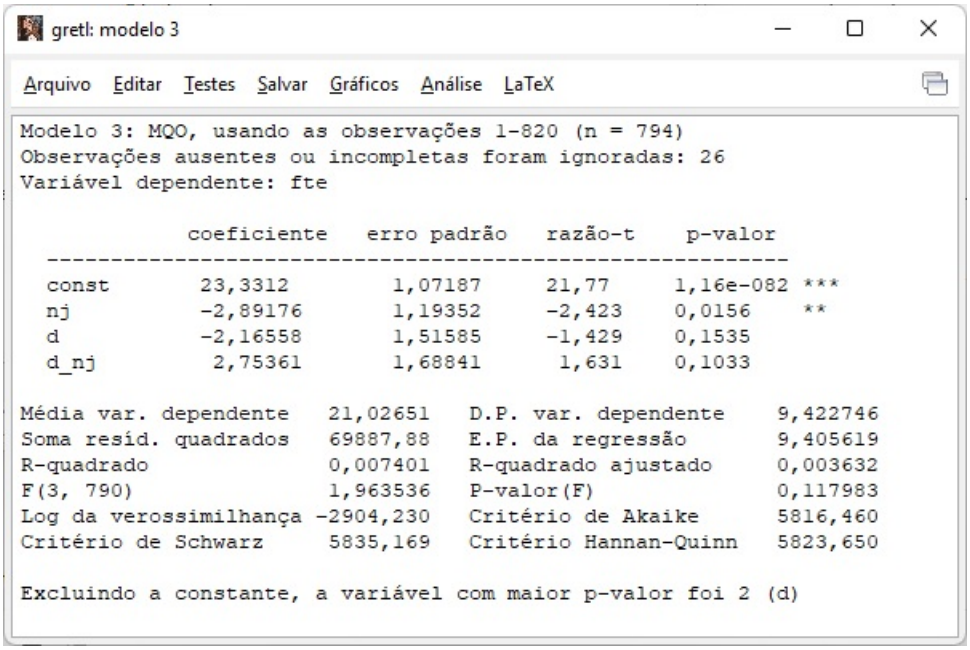
Se deseja saber como uma mudança na política afeta os resultados, nada supera um experimento aleatório controlado. Infelizmente, eles são raros em economia porque são muito caros ou moralmente inaceitáveis. Ninguém quer determinar qual é o retorno à escolaridade atribuindo aleatoriamente pessoas a um determinado número de anos de escolaridade. Essa escolha deve ser individual e não de um formulador de políticas públicas. Mas, a avaliação de políticas públicas não é impossível quando experimentos controlados randomizados são possíveis.

A vida oferece situações que acontecem a diferentes grupos de indivíduos em diferentes pontos no tempo. Esses eventos não são realmente aleatórios, mas, do ponto de vista estatístico, o tratamento pode parecer atribuído aleatoriamente. É disso que tratam os chamados experimentos naturais. Você tem dois grupos de pessoas semelhantes. Por qualquer motivo, um grupo é tratado com a política e o outro não. Diferenças comparativas são atribuídas à política.

No exemplo, será visto os efeitos de uma mudança no salário mínimo. Isso é possível porque o salário mínimo foi aumentado em um estado e não em outro. A semelhança dos estados é importante porque o estado não tratado será usado como grupo de comparação. Os dados são de Card e Krueger e estão no arquivo `njmin3.gdt`.

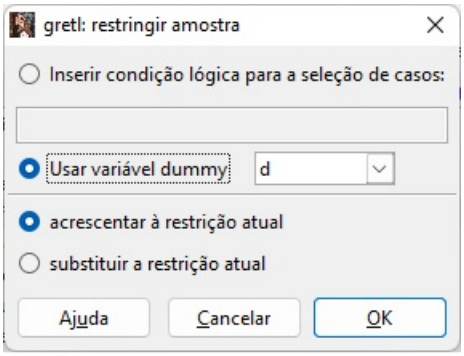
Como se quer ter uma ideia do que aconteceu em NJ e PA antes e depois do aumento do salário mínimo em NJ, pode-se restringir a amostra para antes do aumento e verificar

as estatísticas descritivas. Restaure a amostra completa e, em seguida, restrinja-a após a política $d = 1$. Repita as estatísticas de resumo para `fte`. Os resultados não irão indicar muita diferença.

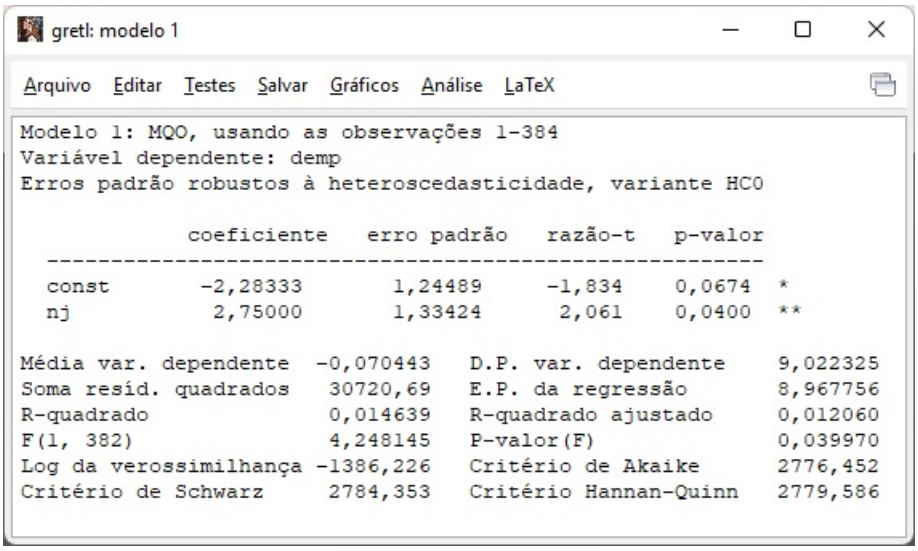


O coeficiente de *dn_j* é o estimador de diferenças em diferenças da mudança no emprego devido a uma mudança no salário mínimo. Não é significativamente diferente de zero neste caso e, sendo assim, pode-se concluir que o aumento do salário mínimo em Nova Jersey não afetou negativamente o emprego.

Na análise anterior não foi explorado uma característica importante dos dados de Card e Krueger. Os mesmos restaurantes foram observados antes e depois em ambos os estados em 384 das 410 observações. Parece razoável limitar a comparação antes e depois às mesmas unidades. Isso requer a adição de um efeito fixo individual ao modelo e a eliminação de observações que não tenham antes ou depois com as quais comparar. Além disso, será preciso limitar a amostra às observações únicas (no original, cada uma é duplicada). Para isso clique na variável *demp* e selecione a opção no menu **Amostra>Descartar obervações com valores ausentes**. Depois selecione a variável *d* clique no menu **Amostra>Restringir baseado em critérios**.



Feito isso estime o seguinte modelo:



O coeficiente de nj não é significativamente menor que zero ao nível de 5% e, portanto, conclui-se que o aumento do salário mínimo não reduziu o emprego.

Capítulo 7

Heterocedasticidade

Uma hipótese importante do modelo clássico de regressão linear é que os termos de erro e_i que aparecem na função de regressão populacional são homocedásticos, ou seja, todos têm a mesma variância. Contudo, em uma regressão qualquer, não há a garantia de que o termo estocástico do modelo, o termo de erro e_i, tenha a mesma variabilidade. Ou seja, algumas observações podem ter uma variância maior ou menor do que outras. Essa condição é conhecida como heterocedasticidade. A seguir tem-se um modelo de regressão linear geral:

$$y_i = \beta_1 + \beta_2\, x_{i\,2} + \cdots + \beta_k\, x_{i\,k} + e_i \qquad i = 1,\, 2,\, \ldots,\, N \tag{7.1}$$

em que y_i caracteriza-se como sendo a variável dependente; $x_{i\,j}$ corresponde à i-ésima observação sobre a j-ésima variável independente (com $j = 2,\, 3,\, \ldots,\, k$); e_i representa o termo de erro e $\beta_1,\, \beta_2,\, \ldots,\, \beta_k$ são os parâmetros a serem estimados. Note que neste modelo de regressão múltipla (Equação 7.1) a variância de e_i agora depende de i, ou seja, da observação a que pertence. Indexar a variância com o subscrito i é apenas uma forma de indicar que as observações podem ter diferentes quantidades de variabilidade associadas a elas. As suposições de erro podem ser resumidas como $e_i \,|\, x_{i\,2},\, x_{i\,3},\, \ldots,\, x_{i\,k} \;\; idd \;\; N\,(0,\, \sigma^2)$.

O intercepto e as inclinações $(\beta_1,\, \beta_2,\, \ldots,\, \beta_k)$ são consistentemente estimados por mínimos quadrados mesmo se os dados forem heterocedásticos. Infelizmente, os estimadores usuais dos erros padrão dos mínimos quadrados e os testes baseados neles são inconsistentes e inválidos. Neste capítulo, várias maneiras de detectar a heterocedasticidade são consideradas bem como são exploradas formas estatisticamente válidas de estimar os parâmetros da Equação 7.1 e testar hipóteses sobre os β's quando os dados são heterocedáticos.

7.1 Exemplo despesa com alimentação

O modelo de de regressão linear simples de gastos com alimentação é estimado usando mínimos quadrados. O modelo é:

$$food\_exp_i = \beta_1 + \beta_2\, income_i + e_i \qquad i = 1,\, 2,\, \ldots,\, n \tag{7.2}$$

em que $food\_exp_i$ caracteriza-se como sendo gastos com alimentação e $income_i$ é a renda do i-ésimo indivíduo. Quando os erros do modelo são heterocedásticos

o estimador de mínimos quadrados dos coeficientes são consistentes.[1] Significando que as estimativas pontuais de mínimos quadrados do intercepto bem como da(s) inclinação(ões) são úteis. No entanto, quando os erros são heterocedásticos, os erros padrão de mínimos quadrados usuais são inconsistentes e, portanto, não devem ser usados para formar intervalos de confiança ou testar hipóteses.

Para usar estimativas de mínimos quadrados com dados heterocedásticos deve-se, no mínimo, usar um estimador consistente de seus erros padrão para construir testes e intervalos de confiança válidos. Um cálculo simples foi proposto por White. Os erros padrão calculados usando a técnica de White são referidos como robustos, mas é preciso tomar cuidado ao usar esse termo. Pois os erros padrão são robustos à presença de heterocedasticidade nos erros do modelo, mas não necessariamente a outras formas de especificação incorreta do modelo.

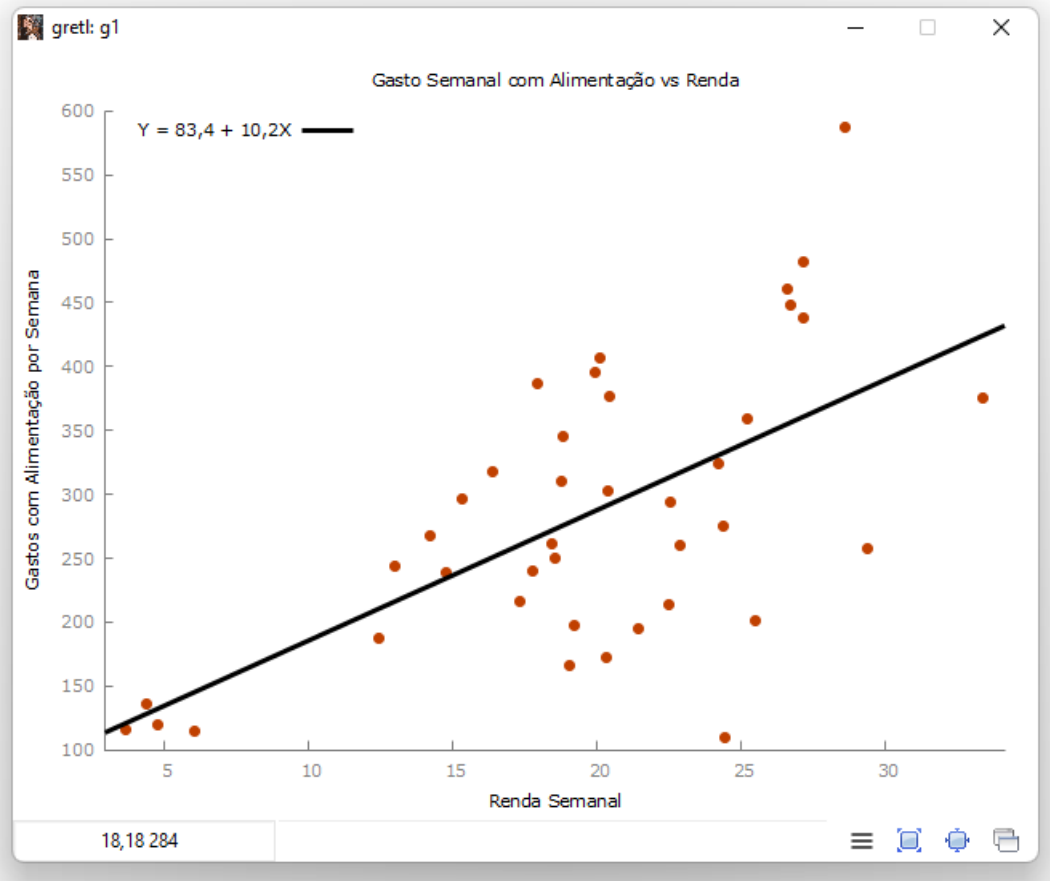


Figura 7.1: Regressão dos gastos com alimentação.

Abra o conjunto de dados `food.gdt` no **gretl** e estime o modelo usando mínimos quadrados. Se os dados forem heterocedástico isso produzirá as estimativas usuais dos parâmetros, contudo, os erros padrão não são confiáveis para construir intervalo de

[1]Dada a hipótese de que e_i segue a distribuição normal, os estimadores de mínimos quadrados são consistentes, ou seja, à medida que o tamanho da amostra aumenta indefinidamente, os estimadores convergem para os verdadeiros valores da população.

confiança, realizar testes de hipóteses e outros procedimentos. Uma inspeção visual do gráfico de regressão do modelo pode sinalizar se os dados são heterocedásticos. No caso do modelo de gastos com alimentação se os dados forem heterocedástico em relação à renda, haverá mais variação em torno da linha de regressão para alguns níveis de renda. Observando o gráfico da Figura 7.1 parece que esse é o caso para o modelo de gastos com alimentação, pois há uma variação significativamente maior nos dados para rendas altas do que para rendas baixas.

7.2 Estimativa robusto de covariância

Para obter os erros padrão robustos à heterocedasticidade execute o comando **Modelo>Mínimos Quadrados Ordinários**, para abrir a caixa de diálogo `especificar modelo`, nessa caixa de diálogo marque a opção `Erros padrão robustos`, conforme Figura 7.2. Note que há um botão à direita chamado `HC1`. Clicando nesse botão é aberta uma caixa de diálogo na qual uma, das duas opções, podem ser selecionadas: i) `Selecione a partir das opções do HCCME Regular` e ii) `Agrupar por`. Marcando a primeira opção, abrirá uma caixa de diálogo de preferências, Figura 7.3. Note que nessa caixa de diálogo foi selecionado a aba `HCCME`, na opção `Para dados de corte` optou-se por `HC3` e marcou a caixa `Usar por padrão a matriz de covariância robusta`.

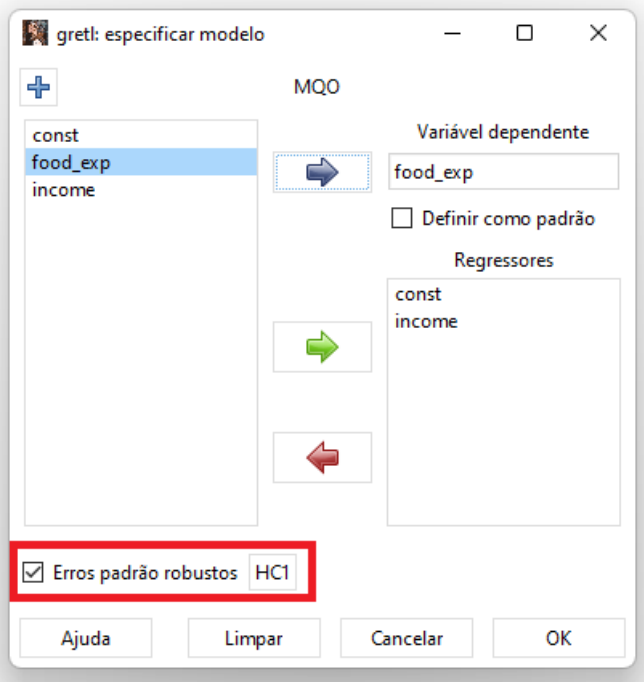


Figura 7.2: Caixa para erros padrão robustos à heterocedasticidade.

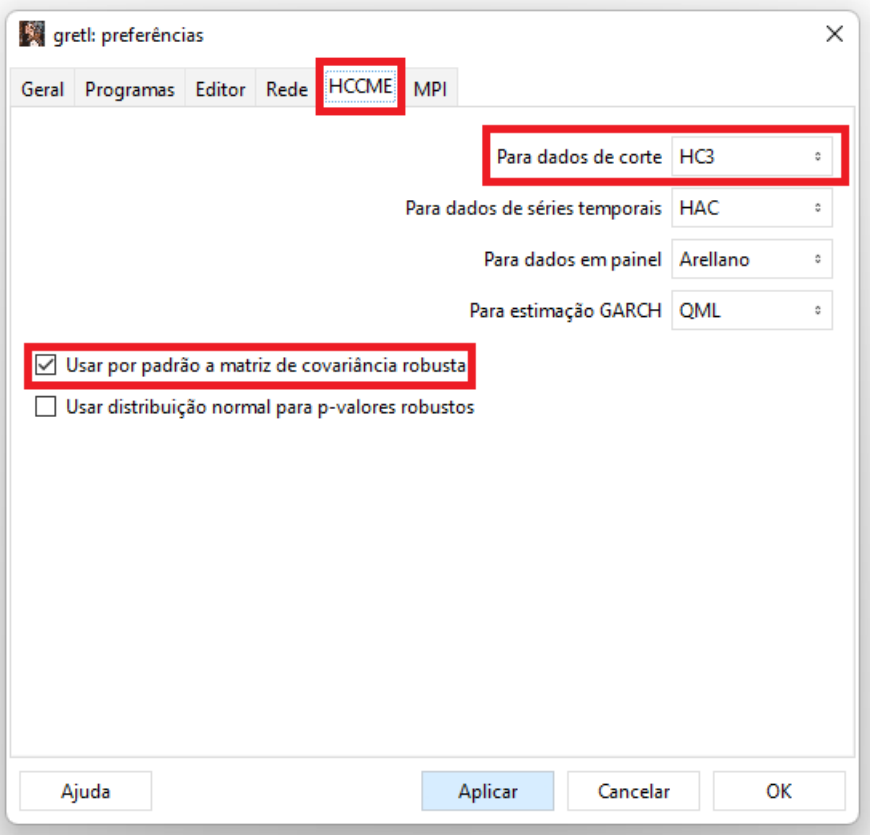


Figura 7.3: Defina o método para calcular erros padrão robustos.

Os resultados do modelo de gastos com alimentação aparecem na Figura 7.4. Objetivando uma análise do intervalo de confiança, execute o comando **Análise>Intervalos de confiança para os coeficientes** na janela principal do modelo, Figura 7.4. Uma vez que esse modelo foi estimado utilizando a opção de erros robustos, os erros do modelo serão baseados na variante dos erros padrão de `White` uma vez que foi escolhido a opção `HC3`, como se pode observar na Figura 7.3. O resultado para o intervalo de confiança é apresentado na Figura 7.5.

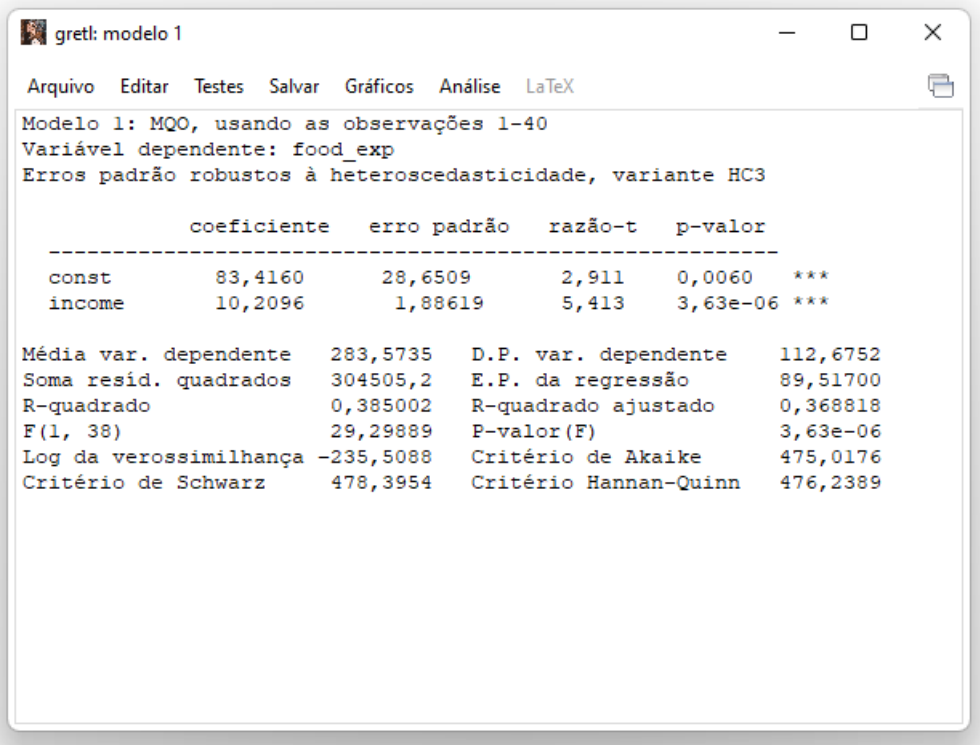


Figura 7.4: Saída do modelo de gastos com alimentação.

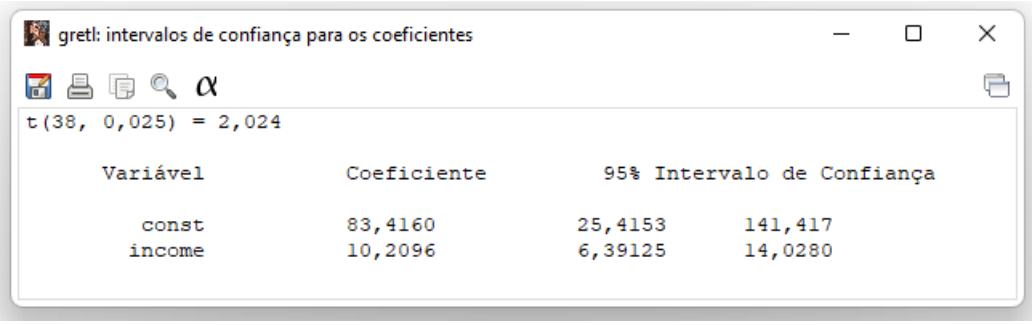


Figura 7.5: Intervalo de confiança para os coeficientes.

7.3 Detecção de heterocedasticidade usando gráficos dos resíduos

Na Seção 7.1 utilizou-se o gráfico da regressão (Figura 7.1) para se ter uma ideia inicial se os dados são heterocedásticos. Agora, porém, utiliza-se os gráficos dos resíduos para tentar identificar se há heterocedasticidades nos dados. Entretanto, chama-se a atenção para o fato de que se deve ter cuidado ao gerar os gráficos dos resíduos bem como ao interpretá-los. Pois, por sua própria natureza, os gráficos dos resíduos só permitem que se análise as relações de uma única variável por vez. Mas, todavia, se a

heterocedasticidade envolver mais de uma variável, os gráficos dos resíduos podem não ser muito reveladores.

A Figura 7.6 caracteriza-se como sendo o gráfico dos Mínimos Quadrados em relação à renda. Analisando visualmente o gráfico da Figura 7.6 parece que para maiores níveis de renda há uma variação muito maior nos resíduos. Esse gráfico pode ser gerado executando o comando **Gráficos>Gráfico dos resíduos>Comparado com income** a partir da janela do modelo, Figura 7.7. Importante destacar que a aparência desse gráfico foi alterada clicando com o botão direito do mouse sobre o gráfico e escolhendo a opção `Editar`.

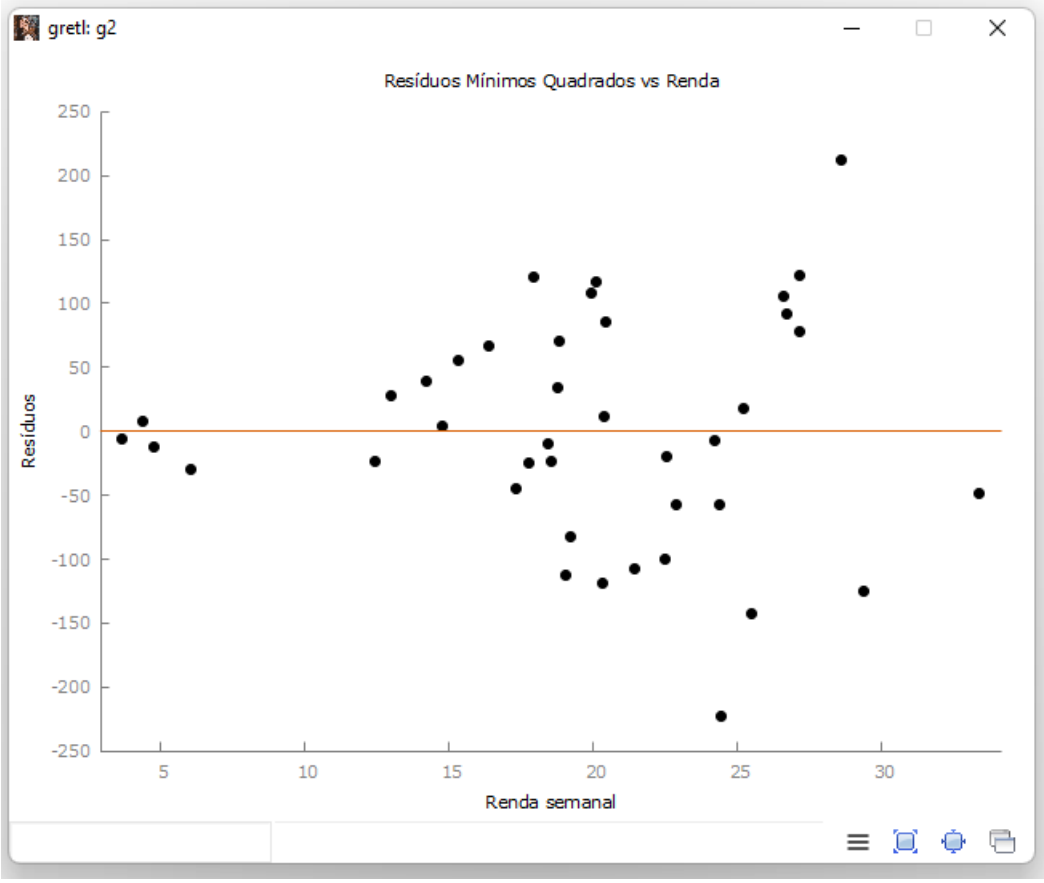


Figura 7.6: Gráfico dos resíduos dos Mínimos Quadrados.

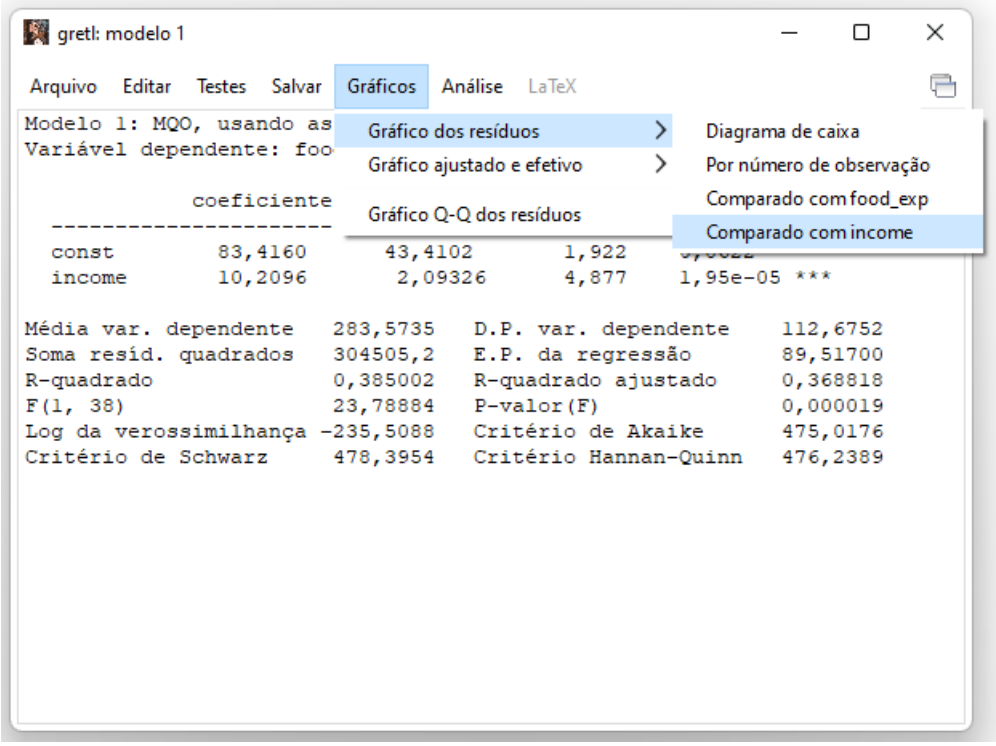


Figura 7.7: Caixa de diálogo para o gráfico dos resíduos.

Outro método gráfico que mostra a relação entre a magnitude dos resíduos e a variável independente é mostrado na Figura 7.8. O primeiro passo para gerar esse gráfico é salvar o valor absoluto dos resíduos dos Mínimos Quadrados em uma nova variável denominada `abs_e`, representada na Figura 7.8 por `|e|`. A seguir, plota-se essa variável (`|e|`) contra a renda como um gráfico de dispersão e como um gráfico de dispersão suavizado e ponderado localmente, estimado pelo processo chamado `loess`. `loess` é considerado um suavizador desejável pois tende a seguir os dados. Diferentemente dos métodos de suavização polinomial que são globais e, assim, o que acontece na extrema direita de um gráfico de dispersão pode afetar os valores ajustados na extrema esquerda. O gráfico da Figura 7.8 foi criado executando os comandos da figura Figura 7.9. Já para a abrir a janela de console para executar os comandos clique no terceiro ícone da esquerda para direita na janela principal do **gretl**, Figura 7.10.

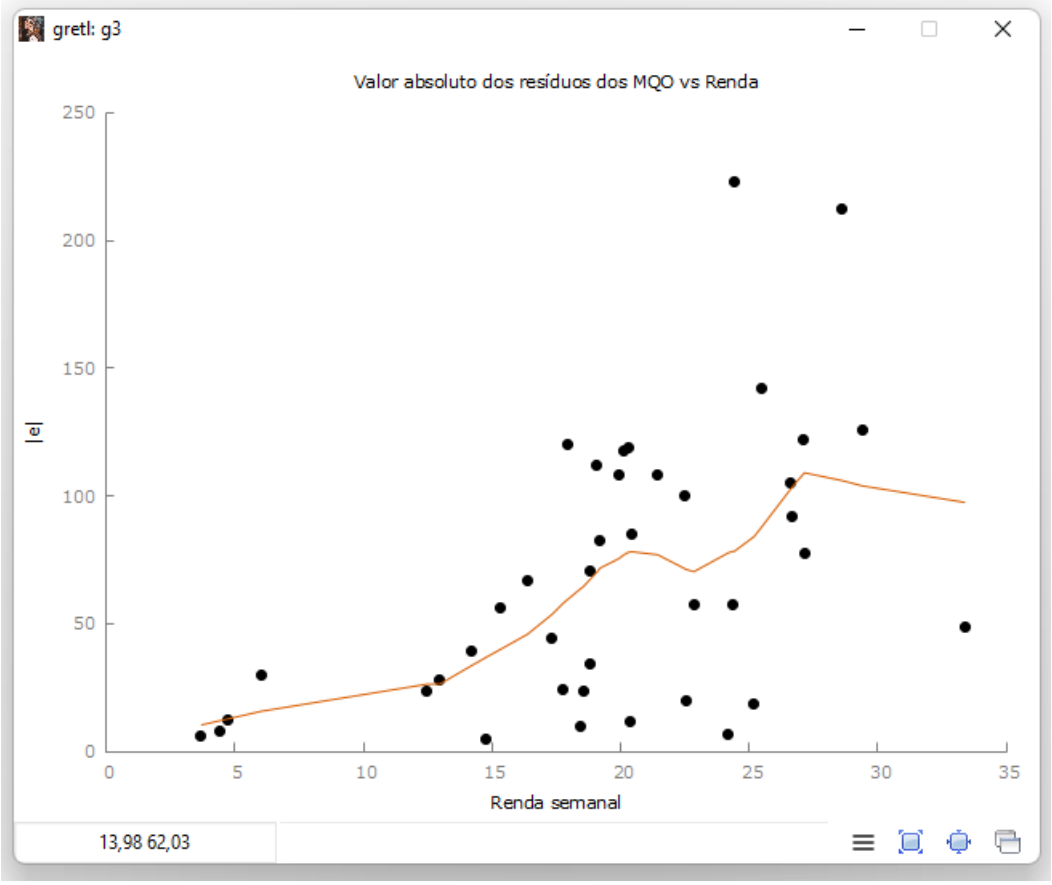


Figura 7.8: Gráfico do valor absoluto dos resíduos com *fit loess*.

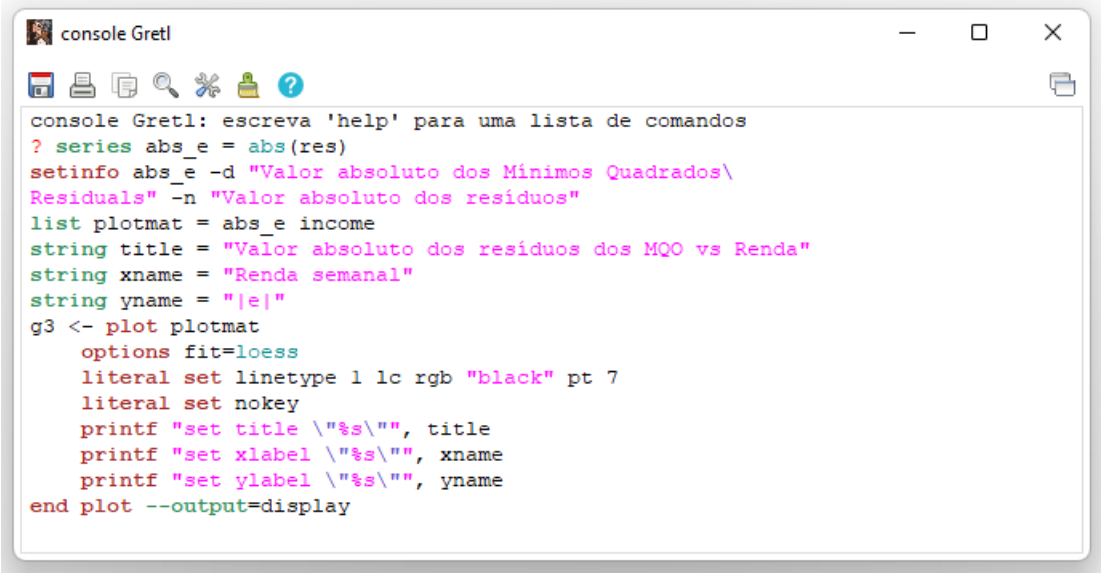


Figura 7.9: Console do **gretl** com as linhas de comando do gráfico com *fit loess*.

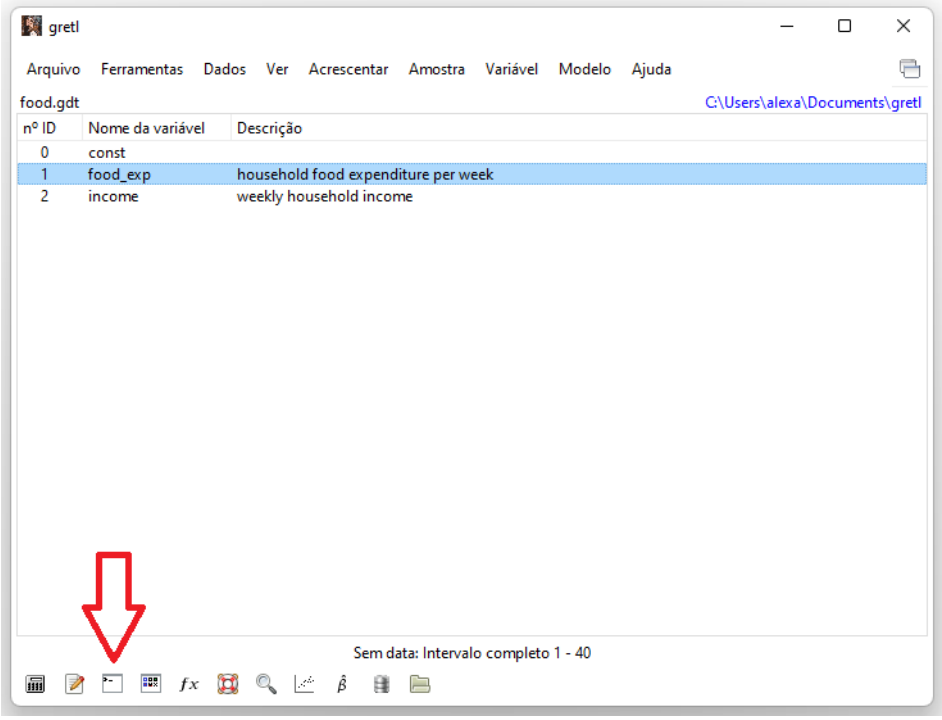


Figura 7.10: Janela principal do **gretl**.

7.4 Mínimos quadrados ponderados

Modelos em que os dados são heterocedásticos as observações com uma variância alta não possuem muita informação sobre a linha de regressão quanto as observações com baixa variância. Nesse caso, uma forma de contornar o problema da heterocedasticidade é a adoção do estimador de Mínimos Quadrados Ponderados (MQP). Isso é possível uma vez que o MQP irá reponderar os dados para que todas as observações contenham o mesmo nível de informação, ou seja, mesma variância, sobre a localização da linha de regressão. Na prática, as observações que contêm mais (menos) ruído recebem menos (mais) peso. Reponderar os dados dessa maneira é conhecido como Mínimos Quadrados Ponderados (MQP).

Suponha que os erros variem proporcionalmente com x_i de acordo com:

$$var\left(e_i\right) = \sigma^2 x_i \tag{7.3}$$

Os erros serão heterocedásticos pois cada erro terá uma variância diferente, cujo valor depende de x_i. Entretanto, como descrito acima o Mínimos Quadrados Ponderados (MQP) reponderá cada uma das observações no modelo de modo que cada observação transformada tenha a mesma variância que as outras. Algebricamente,

$$\frac{1}{\sqrt{x_i}} var\left(e_i\right) = \sigma^2 \tag{7.4}$$

Então, multiplique a Equação 7.1 por $\frac{1}{\sqrt{x_i}}$ para completar a transformação. Assim, o modelo resultante, o modelo transformado, é homocedástico e tanto os Mínimos Quadrados quanto os erros padrão dos Mínimos Quadrados são estatisticamente válidos

e eficientes. Para estimar um modelo de MQP, com a base de dados `food.gdt` carregada no **gretl** clique com o botão direito do mouse em qualquer área da janela principal do **gretl**. Isso abrirá uma janela cuja última opção é `Definir nova variável...`. Clicando nessa opção abrirá uma janela igual a da Figura 7.11. Nessa janela digite `genr peso = 1 / income` para criar a variável `peso` $1/income$ que será usada para reponderar o modelo e, assim, contornar o problema da heterocedasticidade. Uma vez criada a variável `peso` execute o comando **Modelo>Outros modelos lineares>Mínimos Quadrados Ponderados**. Isso abrirá a caixa de diálogo para a especificação do modelo, Figura 7.12.

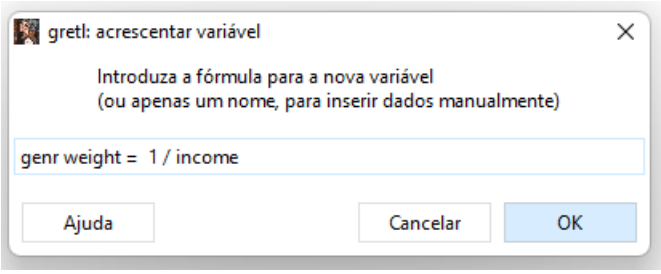


Figura 7.11: Caixa de diálogo para criar uma nova variável.

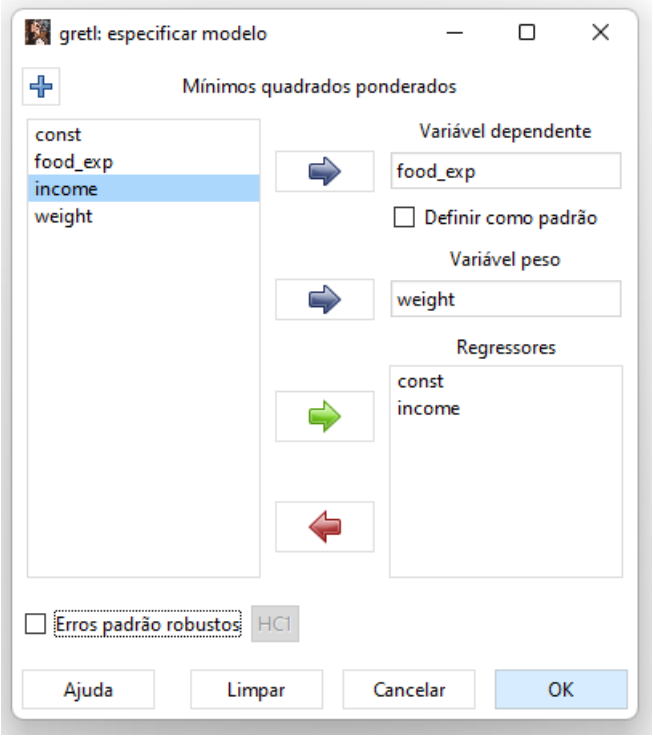


Figura 7.12: Caixa de diálogo de especificação do modelo.

Uma vez que a caixa de diálogo para especificação do modelo for aberta (Figura 7.12) defina como Variável dependete `food_exp`, como Variável peso `weight` e como Regressores `const` e `income` e clique no botão `OK`. A saída do modelo de gastos com alimentação utilizando o estimador de Mínimos Quadrados Ponderados é apresentada na Figura 7.13.

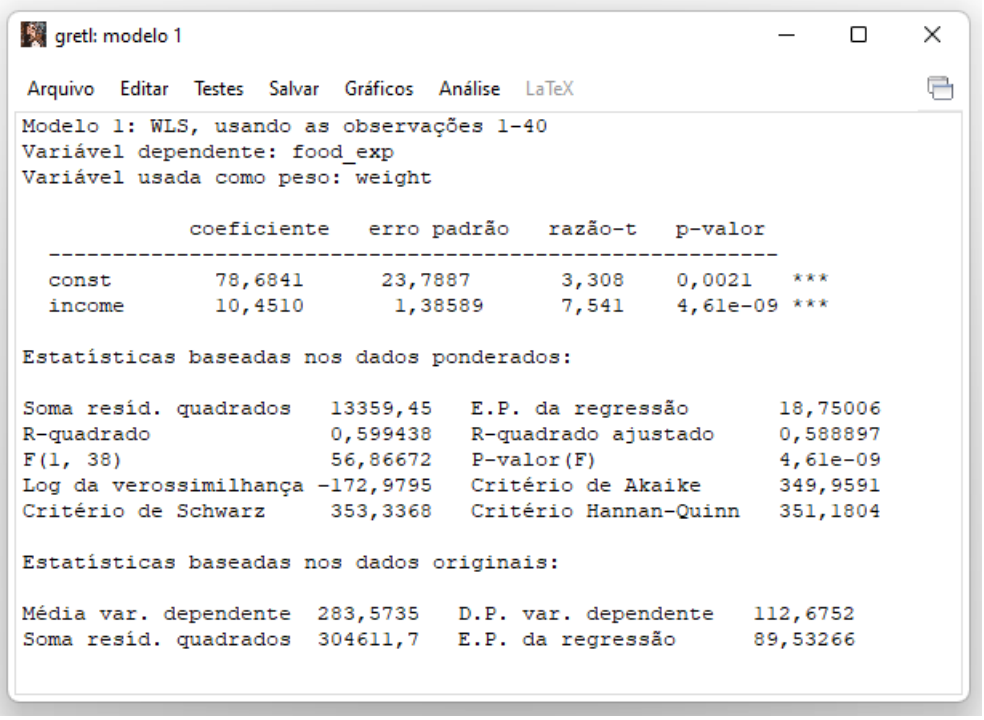


Figura 7.13: Saída do modelo de gasto com alimentação.

Para checar a performance do estimador de Mínimos Quadrados Ponderados a Figura 7.14 plota os resíduos para a estimação empregando MQP, `ehat_wls`, bem como os resíduos para a estimação utilizando o estimador de MQO, `ehat`. Visualmente os resíduos dos MQP, `ehat_wls` parecem ser homocedásticos quando comparados aos resíduos do estimador de MQO, `ehat`. O *script* para a geração do gráfico da Figura 7.14 é apresentado na Figura 7.15, não esqueça de digitar cada uma das linhas do *script* por vez.

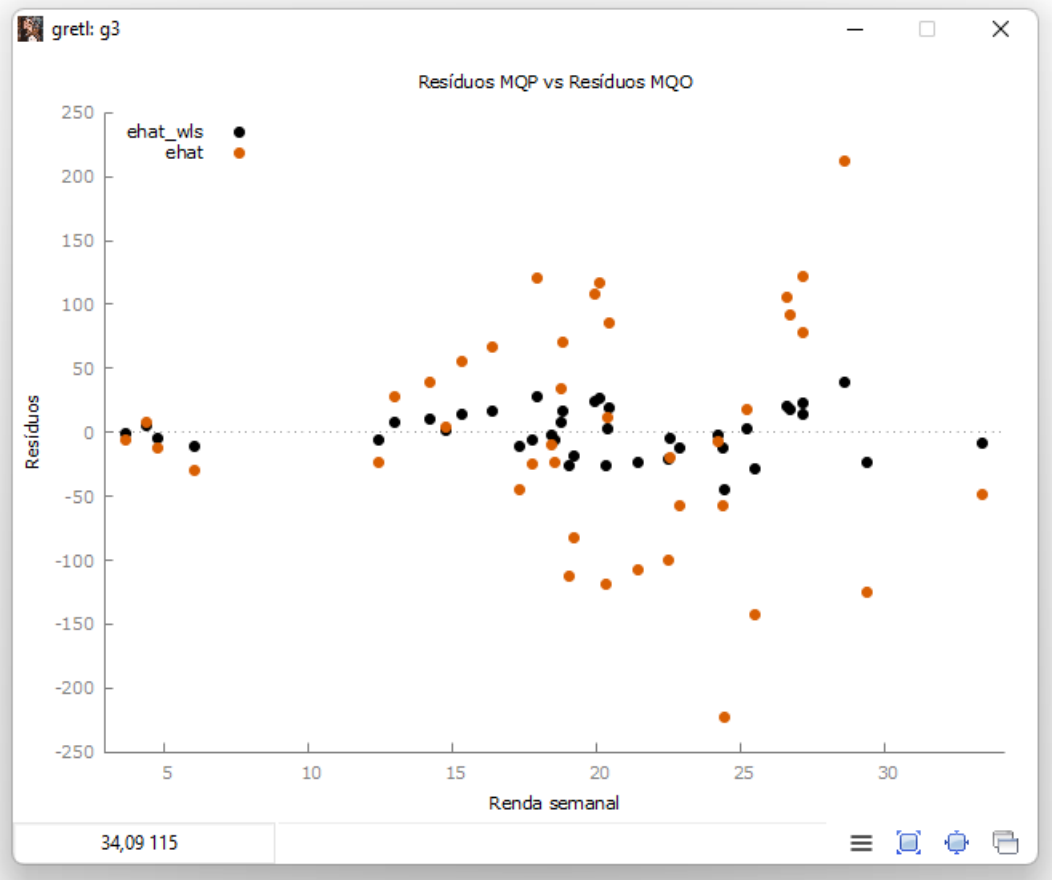


Figura 7.14: Resíduos MQP vs resíduos MQO.

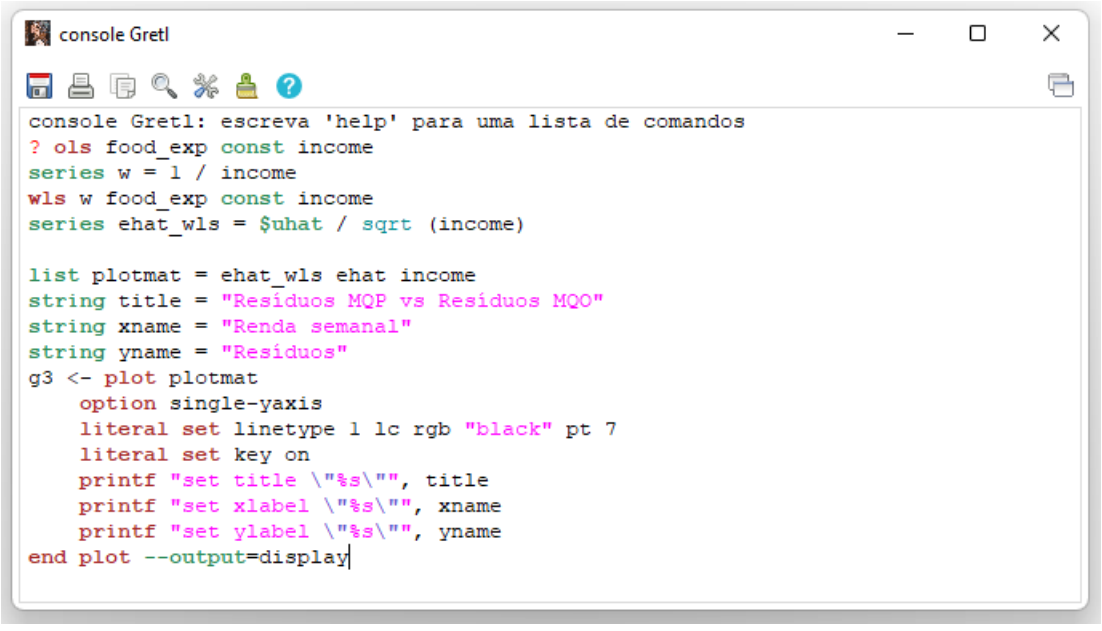


Figura 7.15: Linhas de comando do gráfico dos Resíduos MQP vs Resíduos MQO.

7.5 Detectando heterocedasticidade usando testes de hipótese

7.5.1 Testes do multiplicador de Lagrange

Existem muitos testes de hipótese nula para a homocedasticidade, dois deles são baseados nos `multiplicadores de Lagrange`. Esses são testes particularmente simples de fazer e úteis. O primeiro é algumas vezes denominado de `teste de Breusch-Pagan` (BP). Por sua vez, o segundo é conhecido como `teste de White` e é creditado a White. As hipóteses nula (H_0) e alternativa (H_1) para o `teste de Breusch-Pagan` são:

$$\begin{aligned} H_0 &: \sigma_i^2 = \sigma^2 \\ H_1 &: \sigma_i^2 = h\left(\alpha_1 + \alpha_2\, z_{i2} + \cdots + \alpha_s\, z_{is}\right) \end{aligned} \tag{7.5}$$

A hipótese nula, H_0, é que os dados são homocedásticos enquanto a hipótese alternativa, H_1 ou H_A, é de que os dados são heterocedásticos de uma forma que dependa das variáveis z_{is}, $s = 2, 3, \ldots, S$. Essas variáveis são exógenas e correlacionadas com as variáveis do modelo. Destaca-se que a função $h(.)$ é uma função linear das variáveis z. No caso do modelo de gastos com alimentação, Equação 1.1, para realizar o `teste de Breusch-Pagan` de heterocedasticidade deve-se executar o comando **Testes**>**Heterocedasticidade**>**Breusch-Pagan** na janela da regressão do modelo, conforme Figura 7.16.

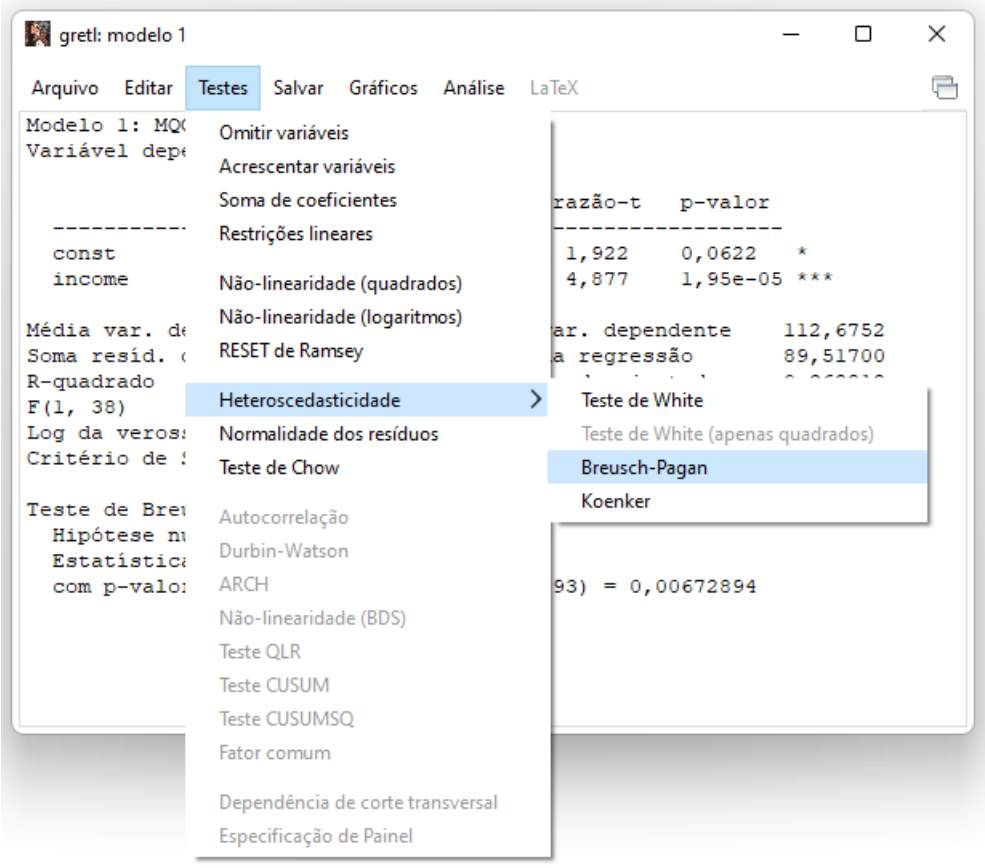


Figura 7.16: Teste de Breusch-Pagan.

Nota-se pela Figura 7.17 que o `teste de Breusch-Pagan` rejeita a hipótese nula, H_0, de homocedasticidade, `p-valor` inferior à 1%

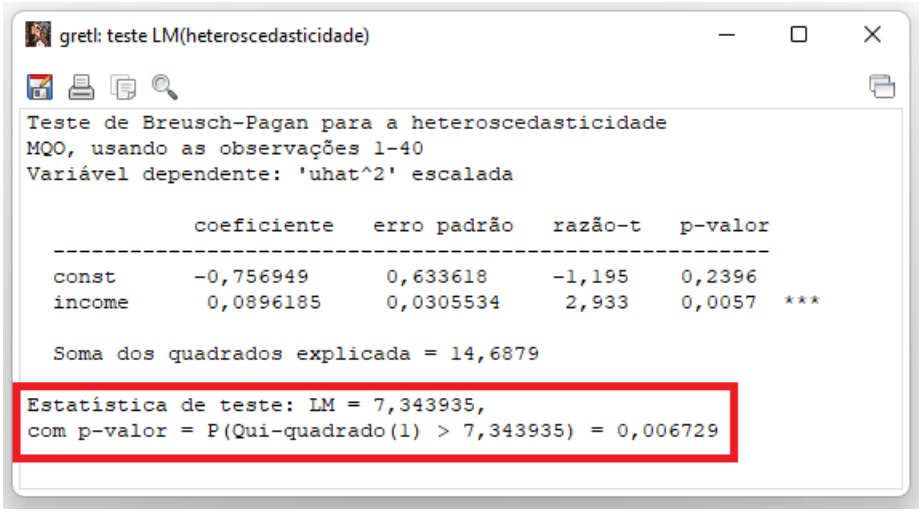


Figura 7.17: Resultado do teste de Breusch-Pagan.

7.5.2 O teste de White

Destaca-se que o `teste de White` caracteriza-se como sendo uma pequena variação do `teste de Breusch-Pagan` em que as hipóteses nula, H_0, e alternativa, H_1 ou H_A, são dados por:

$$\begin{aligned} H_0 &: \sigma_i^2 = \sigma^2 \qquad \text{para todo } i \\ H_1 &: \sigma_i^2 \neq \sigma_j^2 \qquad \text{para pelo menos 1 } i \neq j \end{aligned} \tag{7.6}$$

Esta é uma alternativa composta que captura todas as possibilidades exceto aquela coberta pelo nulo. Se o pesquisador não sabe nada sobre a natureza da heterocedasticidade em seus dados, este é um bom teste para se começar. O teste é muito semelhante ao `teste de Breusch-Pagan`. Porém, no `teste de White` as variáveis relacionadas à heterocedasticidade (z_{is}, $s = 2, 3, \ldots, S$) incluem cada regressor não redundante, seu quadrado e todos os produtos cruzados entre os regressores. No caso do modelo de gastos com alimentação há apenas o intercepto e um regressor contínuo (a renda). Portanto, a constante ao quadrado e o produto cruzado entre a constante e a renda são redundantes. Dessa forma, existe apenas um variável para adicionar ao modelo, renda ao quadrado. Note que, assim como no `teste de Breusch-Pagan`, a hipótese nula de homocedasticidade dos dados foi rejeitada, mas, agora, ao nível de 5%, Figura 7.18.

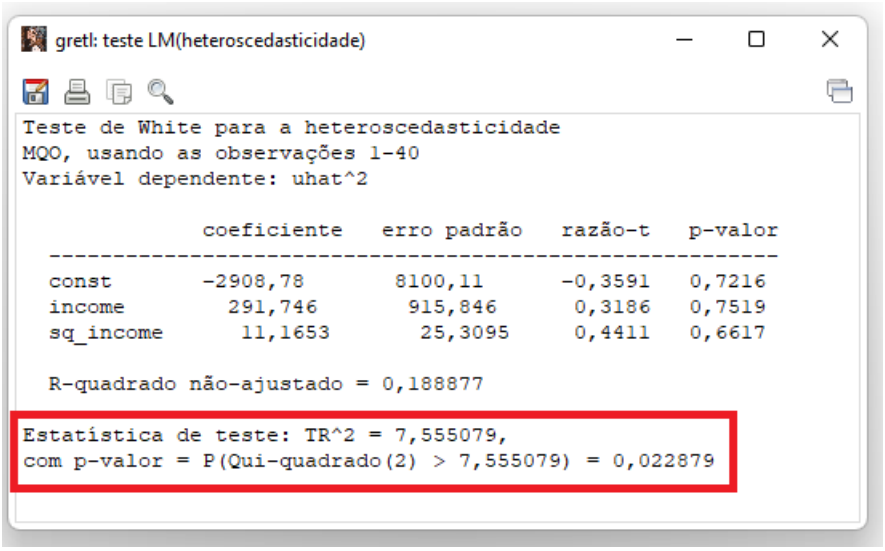


Figura 7.18: Resultado do teste de White.

7.6 Erros padrão consistentes com heterocedasticidade

Lembre-se que na Seção 7.2 foi demonstrado que o estimador de Mínimos Quadrados Ordinários – MQO – pode ser usado para estimar o modelo linear mesmo quando os erros são heterocedásticos, e isso, destaca-se, com bom resultado. Pois o problema

com o uso de MQO em um modelo heterocedástico é que o estimador usual de precisão (matriz de variância-covariância estimada) não é consistente. Assim, a forma mais simples de contornar esse problema é usar MQO para estimar o intercepto e as inclinações (regressores) e usar um estimador de covariância de MQO que seja consistente, sejam os erros heterocedásticos ou não. Esse é o chamado estimador robusto de heterocedasticidade de covariância que o **gretl** usa, Figura 7.2.

A seguir, o modelo de gastos com alimentação é usado para estimar o modelo usando MQO padrão (Ou seja, sem considerar erros padrão robustos) bem como três conjuntos robustos de erros padrão – HC1, HC2 e HC3. Observe, Figura 7.19, que as estimativas dos coeficientes são as mesmas nas quatro colunas (83, 42), contudo, os erros padrão estimados são diferentes. O erro padrão robusto para a inclinação é menor do que o habitual, quando o modelo é estimado sem marcar a caixa `Erros padrão robustos`. Chama-se ainda a atenção para o fato de que vários comandos se comportam de maneira diferente quando são usados após o uso de um modelo que emprega `Erros padrão robustos`. O uso dessa opção força os `testes de Wald` subsequentes com base nas estimativas de MQO a usar o `HCCME` para computação. Isso irá garantir que os resultados de omitir ou restringir serão estatisticamente válidos sob heterocedasticidade quando a regressão for estimada utilizando a opção `Erros padrão robustos`, Figura 7.20. Para mais detalhe sobre como selecionar qual estimador de covariância empregar (HC1, HC2, entre outros) veja a Seção 7.2.

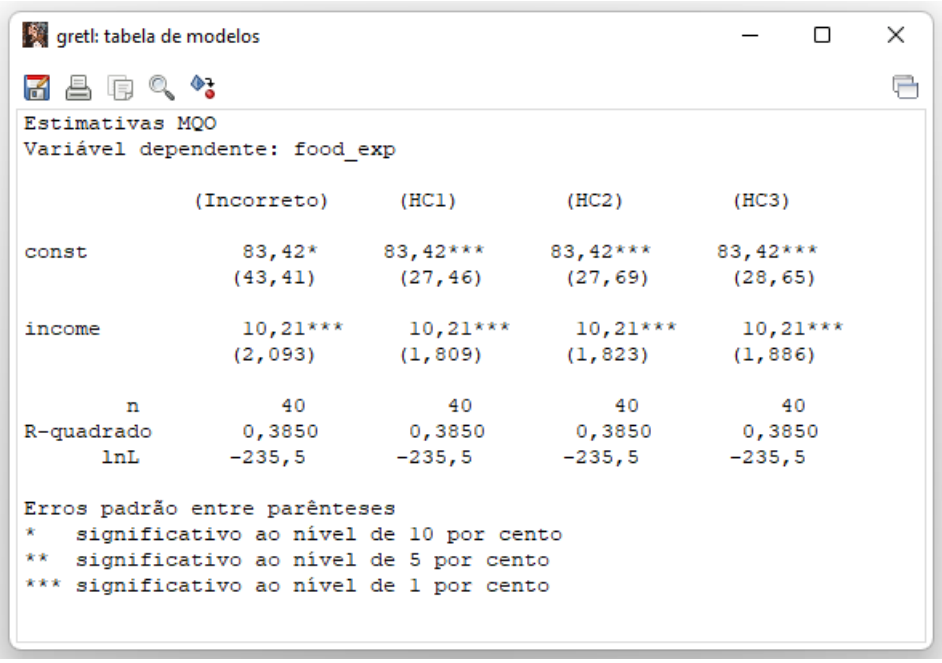


Figura 7.19: Erros padrão robustos *vs* não-robustos.

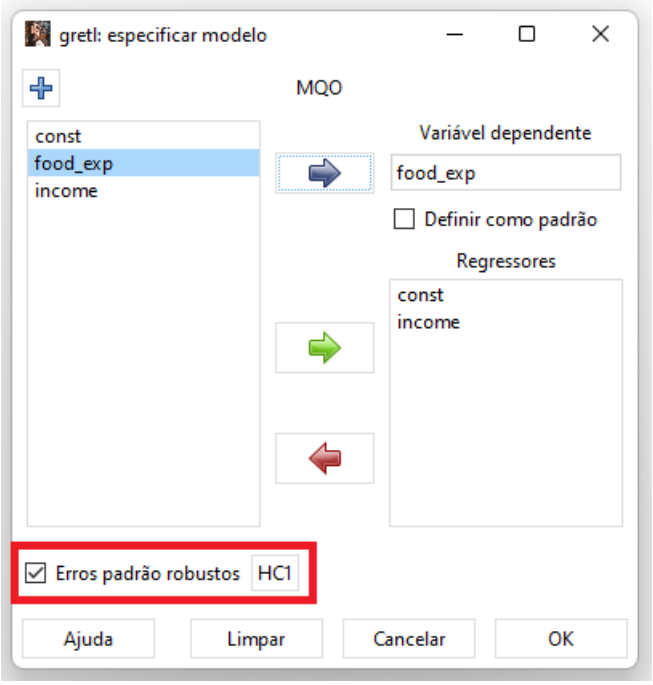


Figura 7.20: Opção para erros padrão robustos à heterocedasticidade.

Capítulo 8

Séries estacionárias

O objetivo principal deste capítulo é explorar as propriedades de séries temporais dos dados usando **gretl**. Um dos pontos básicos em econometria é que as propriedades dos estimadores e sua utilidade para estimativas pontuais e testes de hipóteses dependem de como os dados se comportam. Por exemplo, em um modelo de regressão linear em que os erros são correlacionados com os regressores, os mínimos quadrados não serão consistentes e, consequentemente, não devem ser usados para estimativas ou testes subsequentes.

Na maioria das regressões de séries temporais, os dados devem ser estacionários para que os estimadores tenham propriedades desejáveis. Isso requer que as médias, variâncias e covariâncias das séries de dados sejam independentes do período de tempo em que são observadas. Por exemplo, a média e a variância da distribuição de probabilidade que gerou o PIB no terceiro trimestre de 1973 não pode ser diferente daquela que gerou o PIB do 4º trimestre de 2006. Observações sobre séries temporais estacionárias podem ser correlacionadas entre si, mas a natureza dessa correlação não pode mudar ao longo do tempo. O PIB está crescendo ao longo do tempo (não significa estacionário) e pode ter se tornado menos volátil (não a variação estacionária). Mudanças na tecnologia da informação e nas instituições podem ter encurtado a persistência dos choques na economia (não a covariância estacionária).

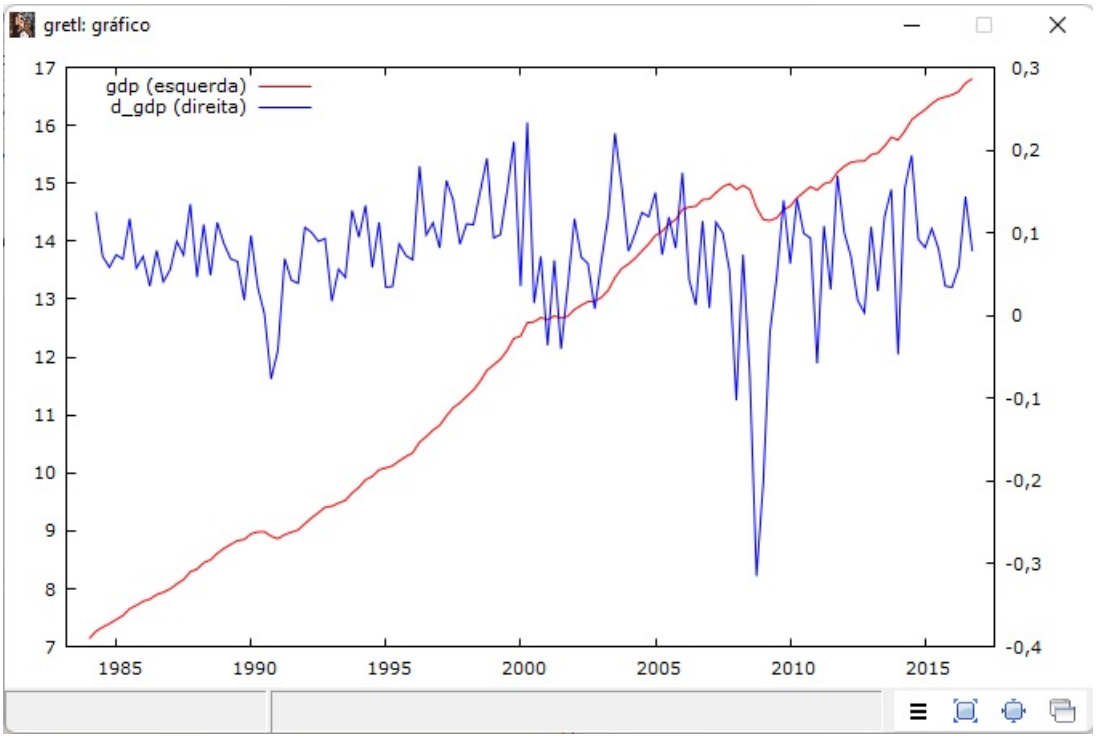
As séries temporais não estacionárias devem ser usadas com cuidado na análise de regressão. Métodos para lidar efetivamente com esse problema forneceram um rico campo de pesquisa para econometristas nos últimos anos.

8.1 Gráficos das séries temporais

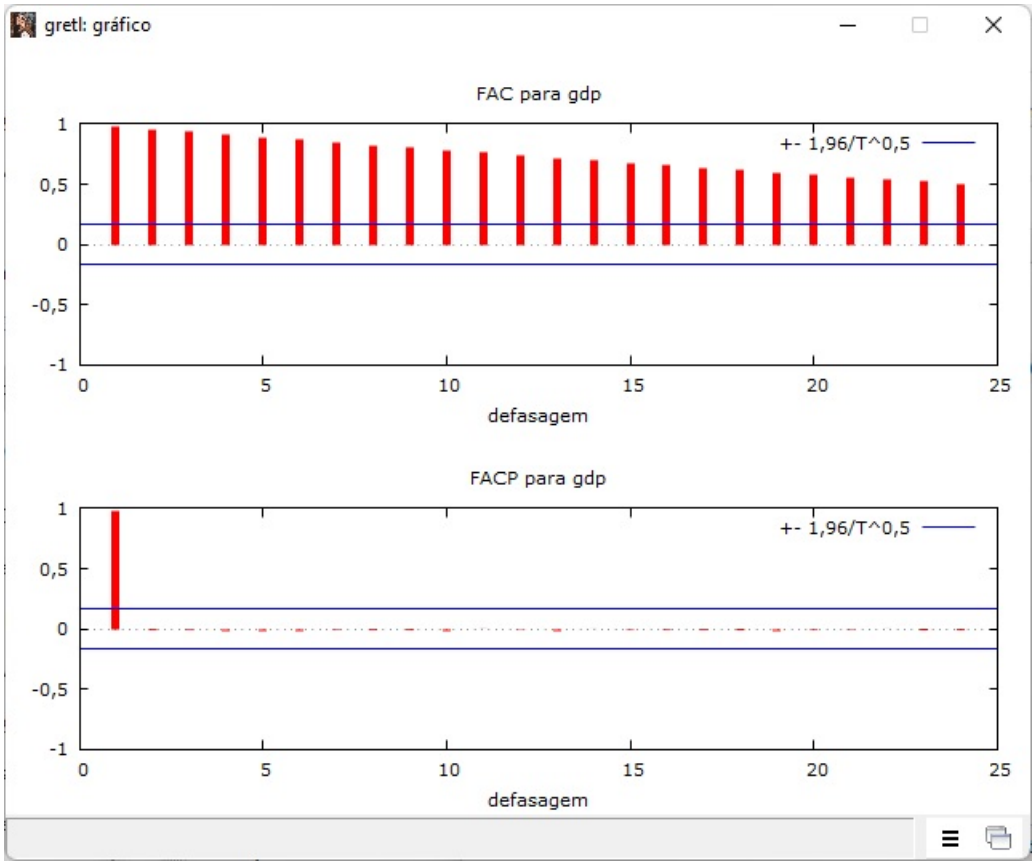
A primeira coisa a fazer ao trabalhar com séries temporais é observá-las graficamente. Um gráfico de série temporal revelará possíveis problemas com seus dados e sugerirá maneiras de proceder estatisticamente. Os gráficos de séries temporais são simples de serem gerados. Abra o arquivo de dados `gdp5.gdt` e crie as primeiras diferenças de GDP . A plotagem da série pode ser feita de várias maneiras. Por exemplo, pode-se clicar no menu **Ver**>**Gráfico das variáveis**>**Séries Temporais**. Alternativamente, pode-se clicar com o botão direito do mouse sobre a série e escolher a opção `Gráfico de Série Temporal`.

Antes de se fazer o gráfico, toma-se a primeira diferença da série do PIB (GDP). Clique no menu **Acrescentar**>**Primeiras diferenças das variáveis selecionadas**.

Também é possível obter o mesmo resultado clicando com o botão direito do mouse sobre a variável desejada e selecionar **Acrescentar diferença**. Selecione as duas variáveis e as coloque em um único gráfico:



Autocorrelações de amostra podem revelar uma potencial não estacionaridade em uma série. Séries não estacionárias tendem a ter grandes autocorrelações em defasagens longas. Isso é evidente para a série do PIB, conforme mostrado abaixo. As grandes autocorrelações para o PIB persistem além de 24 defasagens, um sinal claro de que a série não é estacionária. Apenas as duas primeiras autocorrelações são significativas para a série de mudanças.



Para produzir o gráfico acima é necessário clicar no menu **Variável>Correlograma**.

8.2 Tendências determinísticas

Variáveis não estacionárias que parecem vagar para cima e para baixo por um tempo são chamadas de tendências estocásticas. Por outro lado, algumas tendências são persistentes e são ditas ser determinista. Uma série temporal pode possuir ambos os tipos de tendência. Uma tendência determinística simples para uma série y_t pode ser modelada:

$$y_t = c_1 + c_2\, t + u_t$$

em que t é o índice temporal. Uma tendência quadrática poderia ser:

$$y_t = c_1 + c_2\, t + c_2\, t^2 + u_t$$

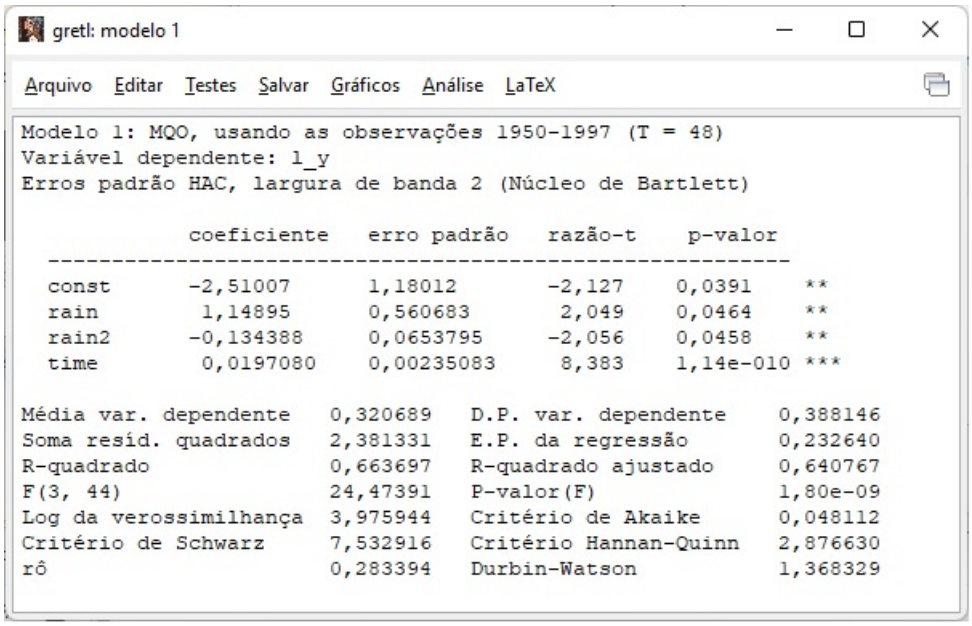
Adicionalmente, uma tendência em mudança percentual pode ser modelada como:

$$\ln(y_t) = c_1 + c_2\, t + u_t$$

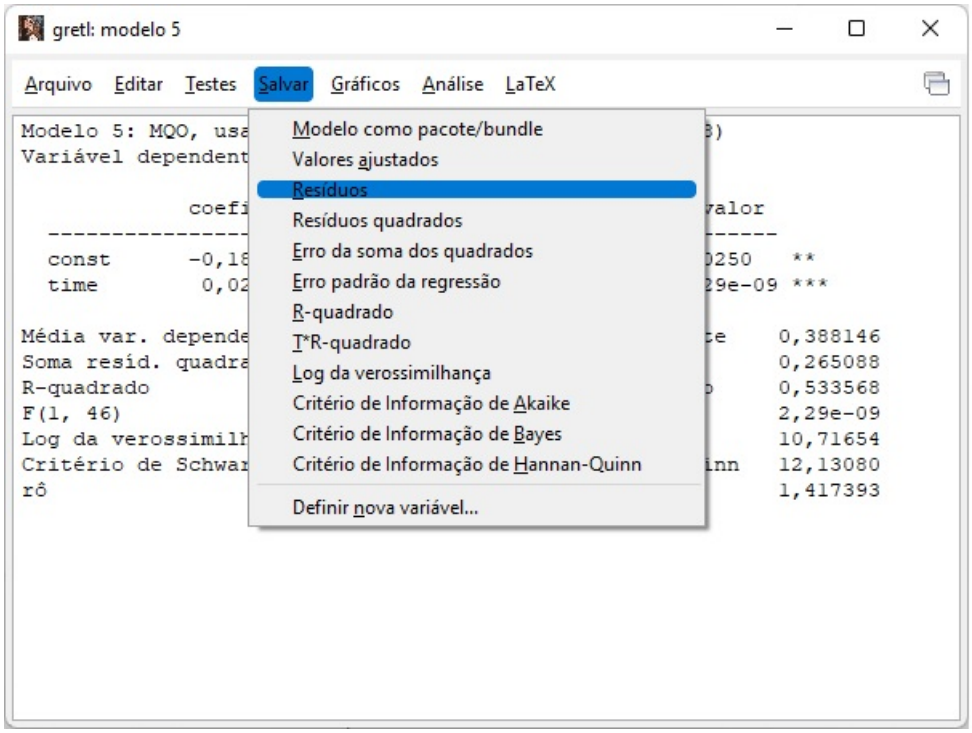
Em cada caso, o efeito temporal é parametrizado e pode ser estimado.

A seguir, será visto um exemplo em que se modela a produção de trigo em Toodyay Shire na Austrália . A produção de trigo depende das chuvas e da produtividade, que

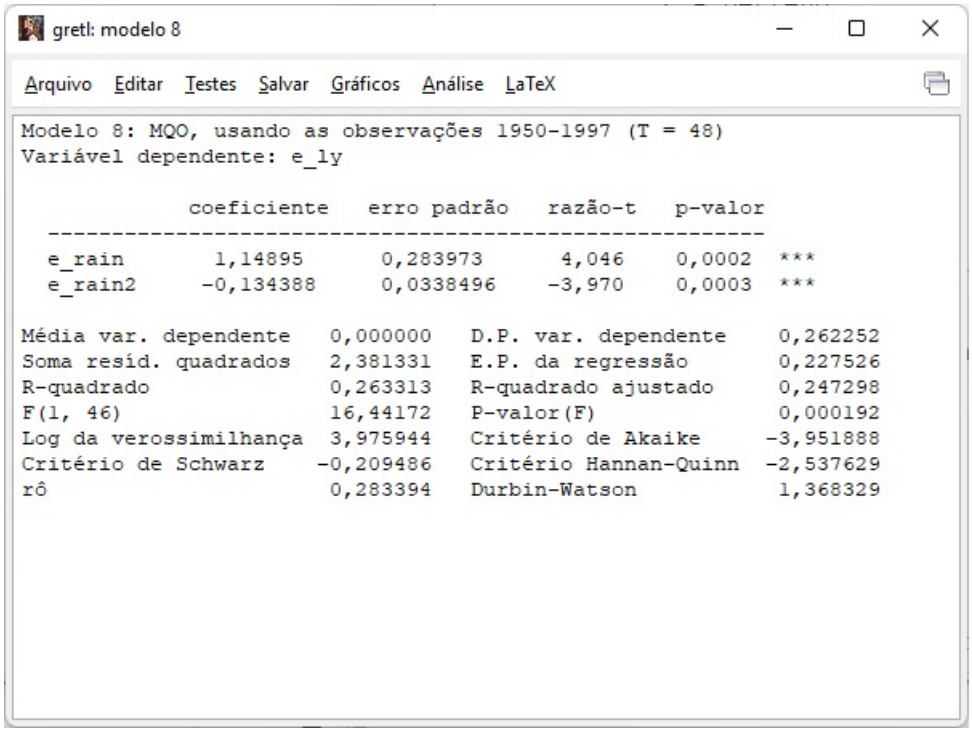
tende a melhorar com o tempo. Assim, é razoável que o rendimento possa apresentar uma tendência determinística. A precipitação também pode mudar ao longo do tempo, possivelmente devido as mudanças no clima global. Após carregar os dados, que estão em `toody5.gdt`, adicione o logaritmo natural da produtividade e o quadrado da precipitação ao conjunto de dados. Pode-se adicionar uma tendência linear clicando no menu **Acrescentar>Tendência Temporal** . A seguir, estima-se um modelo que incluí essa tendência e o quadrado da variável `rain`:



Pode-se observar que a tendência é estatisticamente significativa. Pode-se remover a tendência das séries e rodar um novo modelo sem a tendência temporal. Para isso, precisa-se estimar um modelo de Mínimos Quadrados Ordinários para cada variável contra a tendência e a constante e salvar os resíduos. Após estimar o modelo para l_y, clique em **Salvar>Resíduos**.



Escolha um nome para a nova variável, como por exemplo *e_ly*. Posteriormente repita esse procedimento para todas as variáveis usadas no modelo original e, por fim, estime o seguinte modelo sem constante:



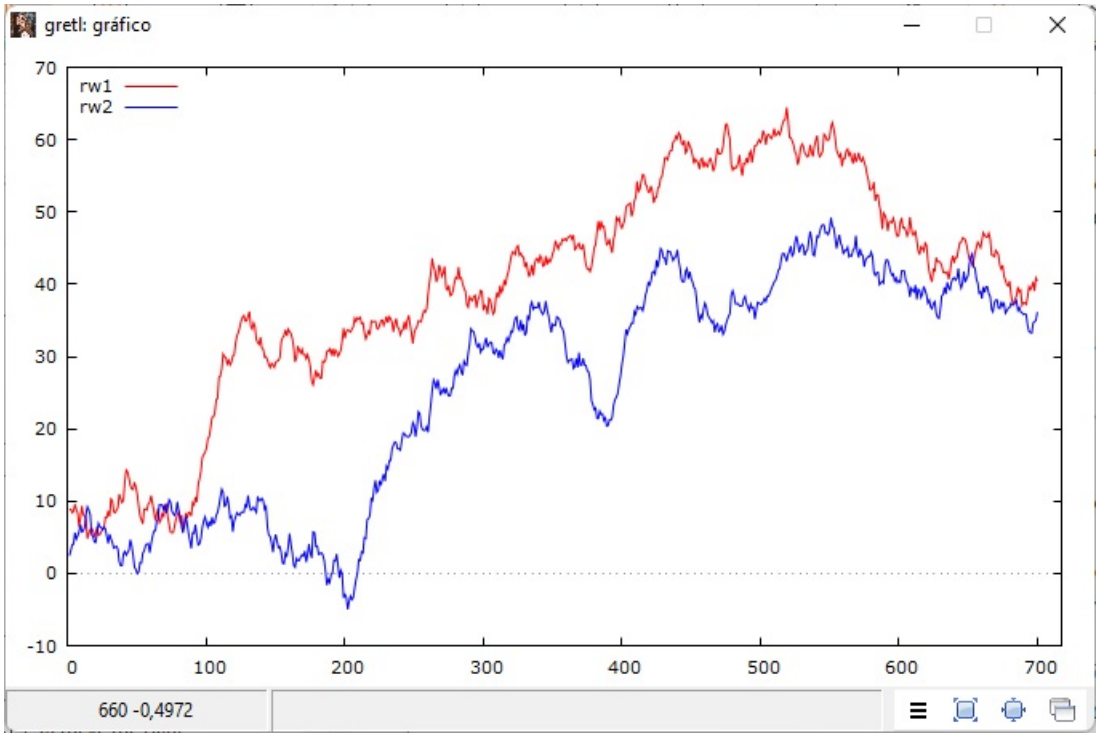
8.3 Regressão espúria

É possível estimar uma regressão e encontrar uma relação estatisticamente significativa mesmo que não exista nenhuma. Na análise de séries temporais, isso é realmente uma ocorrência comum quando os dados não são estacionários. Este exemplo usa duas séries de dados, `rw1` e `rw2`, que foram geradas como `caminhos aleatórios` (random walk) independentes:

$$rw_1 : y_t = y_{t-1} + v_{1t}$$
$$rw_2 : x_t = x_{t-1} + v_{2t}$$

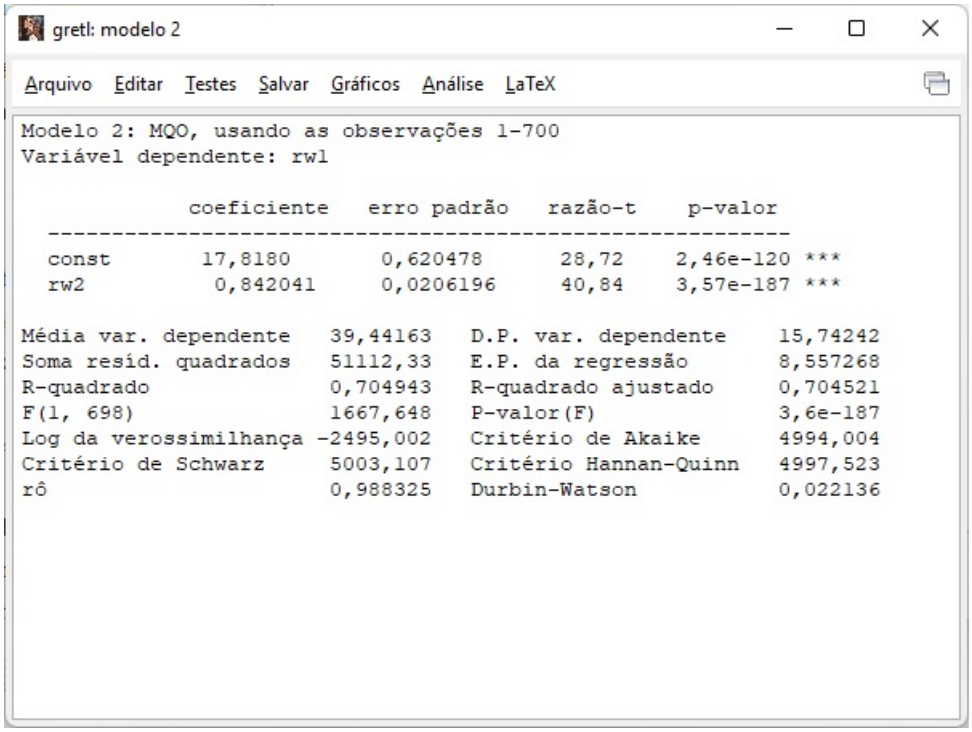
Os erros são desvios aleatórios normais padrão independentes gerados usando um gerador de números pseudoaleatórios. Como se pode ver, x_t e y_t não são relacionados. Para explorar a relação empírica entre essas séries não relacionadas, carregue os dados `spurious.gdt`. Em seguida defina os dados como séries temporais. Para isso clique no **Menu Dados>Estrutura do Conjunto de Dados>Séries Temporais**. Como as séries são fictícias, escolha a periodicidade **Outro**.

Depois plota-se os dados usando um gráfico de série temporal. Para colocar ambas as séries no mesmo gráfico de série temporal, selecione **Ver>Gráfico de variáveis>Série temporal**. Coloque ambas as séries na caixa do lado direito e clique em `OK`.

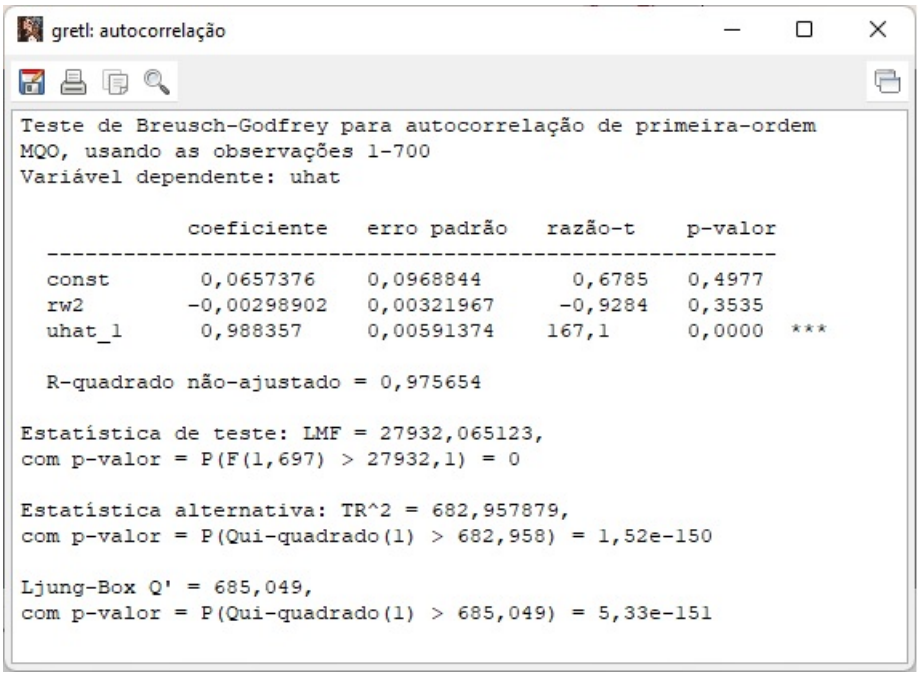


Depois estima-se um modelo de Mínimos Quadrados Ordinários. O coeficiente em `rw2` é positivo $(0,842)$ e significativo $(t = 40.84 > 1.96)$. No entanto, estas variáveis não estão relacionadas umas com as outras! A relação observada é puramente espúria. A causa do resultado espúrio é a não estacionariedade das duas séries. É por isso que

se deve verificar a estacionaridade de seus dados sempre que usar séries temporais em uma regressão.



Finalmente, os resíduos da regressão espúria são testados para autocorrelação de 1ª ordem usando o `teste LM`. No modelo estimado clique no menu **Testes>Autocorrelação** e escolha a ordem 1 para o teste, como segue:



A estatística do `teste LM` é $682,95$ e seu `valor-p` está bem abaixo do limite de 5%. As conclusões baseadas em evidências visuais são confirmadas estatisticamente, ou seja, os erros são autocorrelacionados.

8.4 Testes de estacionariedade

O `teste Dickey-Fuller (aumentado)` pode ser usado para testar a estacionariedade dos dados. O teste é baseado no seguinte modelo de regressão. A versão aumentada do `teste Dickey-Fuller` adiciona várias diferenças defasadas ao modelo. Para o modelo com uma tendência constante e sem determinística, isso seria:

$$\Delta y_t = \alpha + \gamma\, y_{t-1} + \sum_{s=1}^{m} a_s\, \Delta y_{t-s} + v_t$$

Para realizar o teste, algumas decisões devem ser tomadas em relação à série temporal. As decisões geralmente são tomadas com base na inspeção visual dos gráficos das séries temporais. Os gráficos são usados para identificar quaisquer tendências determinísticas na série. Se a tendência da série for quadrática, a versão diferenciada da série terá uma tendência linear.

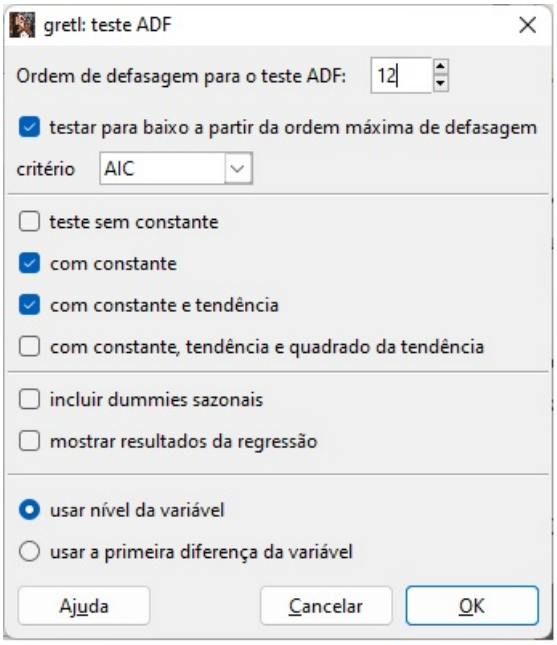
Deve-se determinar o número de termos defasados a serem incluídos nas regressões `ADF`. Há várias formas de fazer isso. Em princípio, os resíduos da regressão `ADF` devem ser isentos de qualquer autocorrelação. Inclua apenas os `lags` suficientes de $\Delta\, y_{t-s}$ para garantir que os resíduos não sejam correlacionados. O número de termos defasados também pode ser determinado examinando a função de autocorrelação (`ACF`) dos resíduos ou a significância dos coeficientes de defasagem estimados.

A hipótese nula do `teste ADF` é que a série temporal possui raiz unitária e não é estacionária. Se essa hipótese for rejeitada, concluirá que a série é estacionária. Não rejeitar a hipótese nula significa que a série em nível não é estacionária. Importante destacar uma característica sobre os resultados do `teste ADF`, **gretl** expressa o modelo de maneira ligeiramente diferente, como segue:

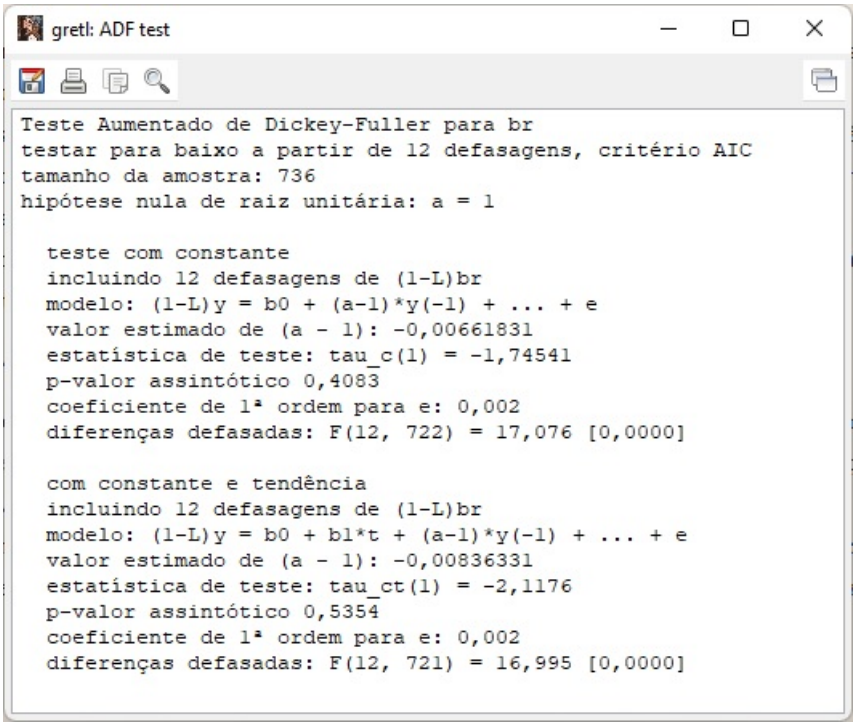
$$(1 - L)\, y_t = \beta_0 + (\alpha\, -\, 1)\, y_{t-1} + \alpha_1\, \Delta\, y_{t-1} + e_t$$

O coeficiente β_0 está incluído porque a série pode ter uma tendência, $(\alpha\, -\, 1) = \gamma$ é o coeficiente de interesse na regressão de `Dickey-Fuller` e α_1 é o parâmetro para o termo que "aumenta" a regressão de `Dickey-Fuller`. Ele é incluído para eliminar a autocorrelação nos erros do modelo, e_t, e mais defasagens podem ser incluídas, se necessário, para realizar isso. A notação no lado esquerdo da equação $(1\, -\, L)\, y_t$ faz uso do operador de `lag`, L. O operador `lag` realiza a mágica $Ly_t = y_{t-1}$. Assim, $(1\, -\, L)\, y_t = y_t - Ly_t = y_t - y_{t-1} = \Delta\, y_t$.

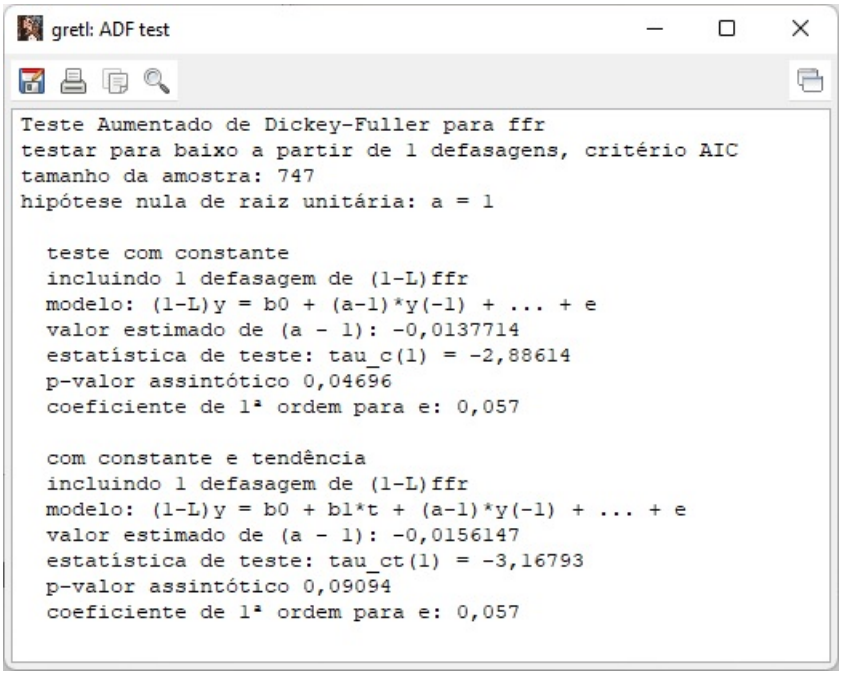
No exemplo a seguir, são consideradas a taxa dos fundos federais (`ffr`) e a taxa dos títulos de 3 anos (br). O arquivo a ser usado é o `usdata5.gdt`. Para realizar os testes `Dickey-Fuller`, primeiro decida se deve usar uma tendência constante e/ou determinística. Deve-se selecionar uma das séries, por exemplo `ffr` e clicar no menu **Variável>Testes de Raíz Unitária>Teste de Dickey-Fuller Aumentado**. As opções mostradas na figura abaixo são as padrões que o **gretl** dá para o `teste ADF`:



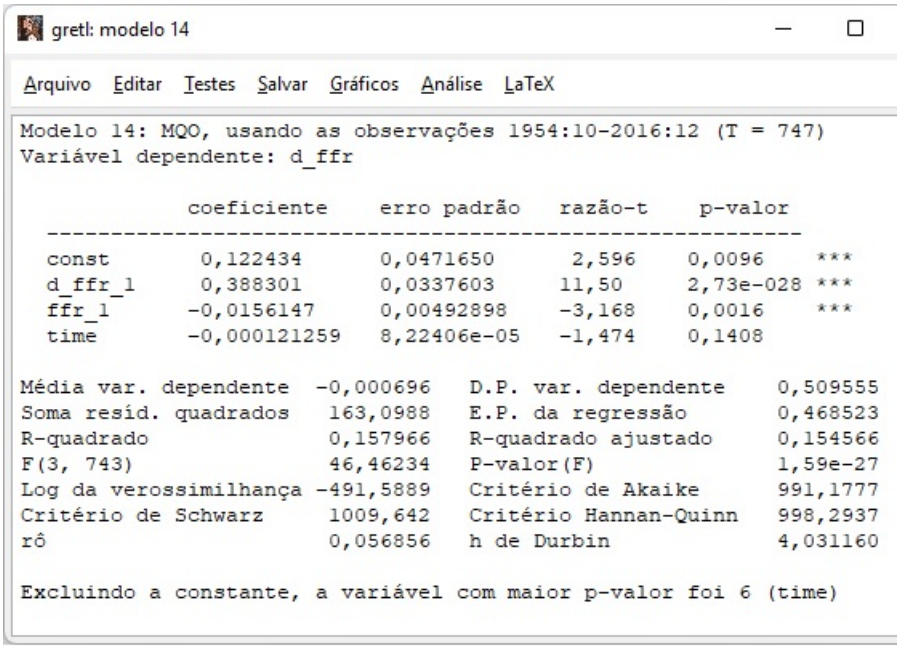
Após rodar o teste os seguintes resultados são mostrados:



Os resultados do teste são bastante informativos. Para os modelos com constante e constante e tendência, não se pode rejeitar a hipótese nula de raiz unitária. Em outras palavras a série dos títulos federais americanos não é estacionária em nível. Agora será utilizado apenas uma defasagem. Os resultados do teste são os seguintes:

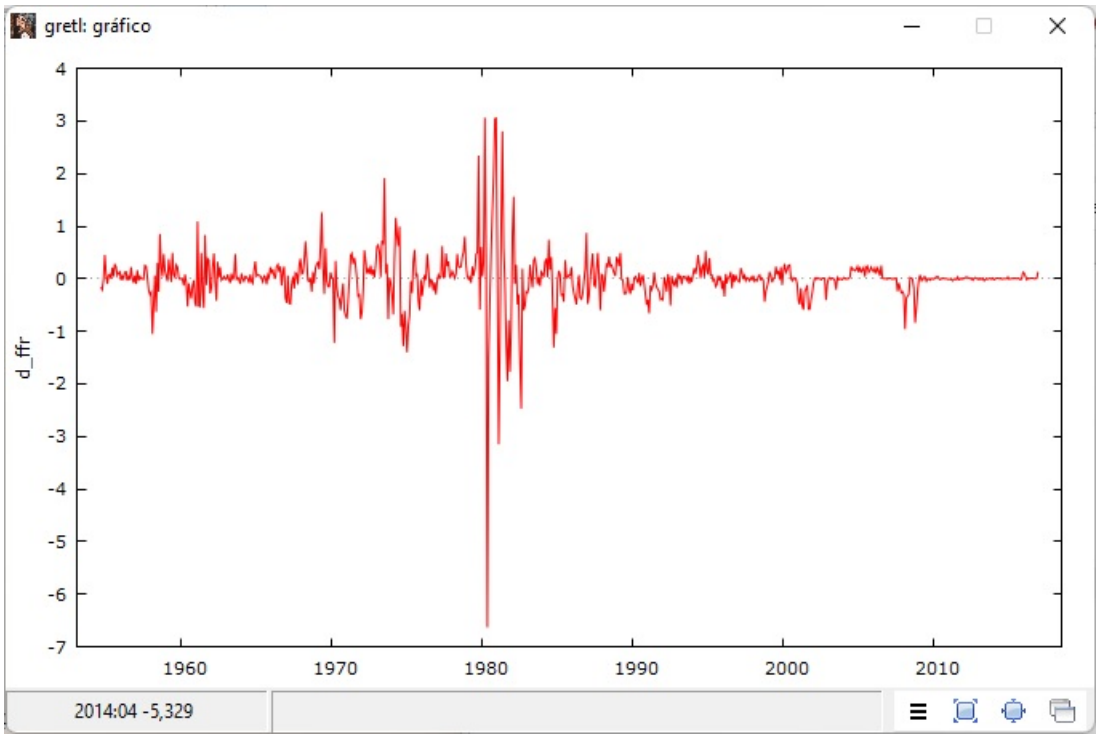


Os resultados do teste podem ser reproduzidos rodando um MQO tendo como variável dependente a primeira diferença `ffr` contra `ffr` defasada e a primeira diferença da variável defasada. Use o botão `Acrescentar` para criar as defasagens e tomar a primeira diferença. Veja que a estatística t da variável `ffr_1` é igual a do `tau_ct(1)` do modelo com constante e tendência.



Assim não se pode rejeitar a hipótese nula de raíz unitária. Em outras palavras, a série `ffr` não é estacionária em nível. Agora veja o gráfico dessa série quando se toma

a primeira diferença.

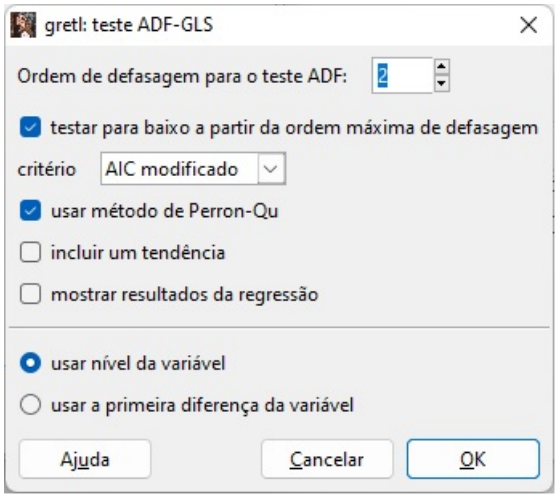


A série é estacionária. Faça o `teste ADF` para conferir.

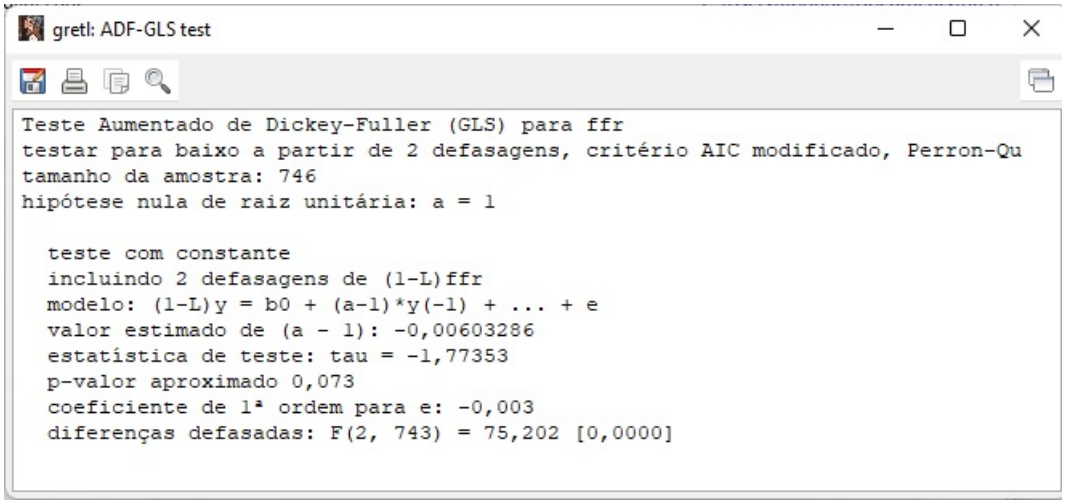
8.4.1 Outros testes para não estacionariedade

Há outros testes para não estacionariedade no **gretl**. O primeiro é o `teste DF-GLS`. Esse procedimento performa o `teste t` modificado de `Dickey-Fuller` (conhecido como `teste DF-GLS`) proposto por Elliott *et al.* (1996). Essencialmente, o teste é um `teste de Dickey-Fuller aumentado`, exceto que a série temporal é transformada por meio de uma regressão de Mínimos Quadrados Generalizados (GLS) antes de estimar o modelo. Elliott *et al.* (1996) mostraram que esse teste tem poder significativamente maior do que as versões anteriores do `teste Dickey-Fuller aumentado`. Consequentemente, não é incomum que este teste rejeite a hipótese nula da não estacionariedade quando o `teste de Dickey-Fuller aumentado` usual não o faz.

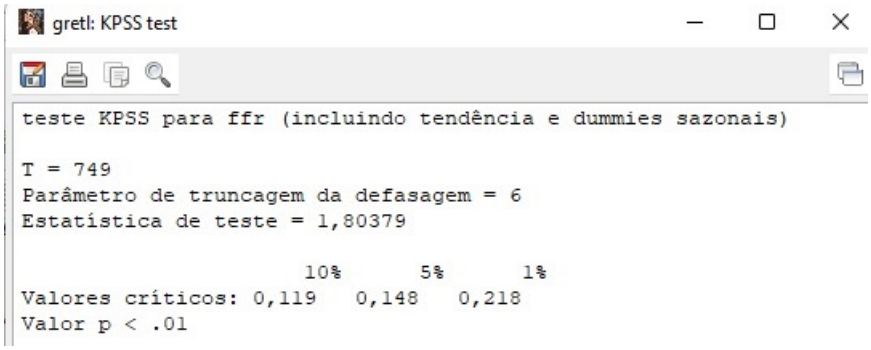
Para executar o `teste ADF-GLS` deve-se selecionar a variável desejada (`ffr`) e clicar no menu **Variável>Teste de raiz unitária>Teste ADF-GLS**. Para tanto selecione as seguintes opções:



Ao clicar em `OK`, tem-se os seguintes resultados:



A estatística do teste é -1, 7735 e tem um `p-valor` de 0, 0723, que está na zona de rejeição a 10% . Ao nível de significância de 10%, a série é estacionária. O **gretl** também pode realizar o `teste KPSS` proposto por Kwiatkowski *et al.* (1992). A hipótese nula desse teste é que a variável em questão é estacionária. Para executá-lo o caminho é o mesmo, basta selecionar a variável e clicar no menu **Variável>Teste de raiz unitária>Teste KPSS**. A seguir, tem-se o resultado do teste:



O `p-valor` é menor que 0.01, então pode-se rejeitar a hipótese nula que a série é estacionária.

8.5 Integração e cointegração

Duas séries temporais não estacionárias são cointegradas se tendem a se mover juntas ao longo do tempo. Por exemplo, estabelece que os níveis da taxa de fundos federais e do título de 3 anos são não estacionárias.

Na linguagem opaca usada na literatura de séries temporais, diz-se que cada série é integrada de ordem 1 ou `I (1)`. Se as duas séries não estacionárias se movem juntas ao longo do tempo, diz que são cointegradas. A teoria econômica sugeriria que eles deveriam ser vinculados por meio de arbitragem, mas isso não é garantido. Nesse contexto, o teste de cointegração equivale a um teste da substituibilidade desses ativos.

O teste básico é muito simples. Deve-se regredir uma variável `I (1)` contra outra usando mínimos quadrados. Se as séries forem cointegradas, os resíduos dessa regressão serão estacionários. Isso é verificado usando o `teste de Dickey-Fuller aumentado`, com um novo conjunto de valores críticos que levam em conta que a série de resíduos utilizada no teste é estimada a partir de dados. Engle e Granger usaram simulações para determinar os valores críticos corretos para o teste, o teste recebe o nome dos dois pesquisadores.

A hipótese nula é que os resíduos são não estacionários, o que implica que as séries não são cointegradas. Para obtê-lo, use **Modelo>Série temporal Multivariadas>Teste de cointegração (Engle-Granger)** na janela principal do **gretl**. Na caixa de diálogo, indique quantas defasagens devem ser incluídas nas regressões `Dickey-Fuller` iniciais em cada uma das variáveis, quais variáveis se deseja incluir na relação de cointegração e se uma constante, tendência ou tendência quadrática é necessária nas regressões.

gretl: teste de cointegração

```
  diferenças defasadas: F(6, 733) = 26,319 [0,0000]

Passo 3: regressão de cointegração

Regressão de cointegração -
MQO, usando as observações 1954:08-2016:12 (T = 749)
Variável dependente: br

             coeficiente    erro padrão   razão-t    p-valor
  ------------------------------------------------------------
  const       1,52824       0,0954937     16,00     8,90e-050 ***
  ffr         0,825232      0,00999627    82,55     0,0000    ***
  time       -0,000445478   0,000166732   -2,672    0,0077    ***

Média var. dependente   5,427810   D.P. var. dependente    3,151571
Soma resíd. quadrados   679,1410   E.P. da regressão       0,954137
R-quadrado              0,908588   R-quadrado ajustado     0,908343
Log da verossimilhança -1026,118   Critério de Akaike      2058,235
Critério de Schwarz     2072,091   Critério Hannan-Quinn   2063,575
rô                      0,916971   Durbin-Watson           0,164479

Passo 4: teste para uma raiz unitária em uhat

Teste Aumentado de Dickey-Fuller para uhat
incluindo 6 defasagens de (1-L)uhat
tamanho da amostra: 742
hipótese nula de raiz unitária: a = 1

  teste sem constante
  modelo: (1-L)y = (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,0768769
  estatística de teste: tau_ct(2) = -4,90933
  p-valor assintótico 0,001217
  coeficiente de 1ª ordem para e: 0,001
  diferenças defasadas: F(6, 735) = 13,823 [0,0000]

Existe evidência de uma relação de cointegração se:
(a) A hipótese de raiz unitária não é rejeitada para as variáveis individuais e;
(b) A hipótese de raiz unitária é rejeitada para os resíduos (uhat) da
    regressão de cointegração.
```

Pode-se rejeitar a hipótese nula que os resíduos possuem raiz unitária, ou seja, são não estacionários. Em outra palavras, as séries são cointegradas.

8.6 Correção de erro

A cointegração é uma relação entre duas variáveis não estacionárias, `I (1)`. Essas variáveis compartilham uma tendência comum e tendem a se mover juntas no longo prazo. Nesta seção, examina-se uma relação dinâmica de curto prazo entre variáveis `I (0)` que incorpora uma relação de cointegração conhecida como modelo de correção de erros.

Inicia-se com um modelo `ARDL (1, 1)`:

$$y_t = \delta + \theta_1 \, y_{t-1} + \delta_0 \, x_t + \delta_1 \, x_{t-1} + v_t$$

após alguma manipulação:

$$\Delta y_t = -(1 - \theta_1)(y_{t-1} - \beta_1 - \beta_2 x_{t-1}) + \delta_0 \Delta x_t + v_t$$

O termo no segundo conjunto de parênteses é uma relação de cointegração em que os níveis de y e x estão linearmente relacionados. Seja $\alpha = (1 - \theta_1)$ e os parâmetros da equação podem ser estimados por mínimos quadrados não lineares. É uma questão meramente empírica a opção de adicionar ou não as defasagens de Δx_t e Δy_t como regressores. Novamente, devemos incluir defasagens suficientes para remover a autocorrelação dos resíduos.

O modelo de correção de erro a ser estimado é:

$$\Delta br_t = -\alpha(br_{t-1} - \beta_1 - \beta_2 ffr_{t-1}) + \gamma_1 \Delta br_{t-1} + \gamma_2 \Delta br_{t-2}$$
$$+\delta_0 \Delta ffr_t + \delta_1 \Delta ffr_{t-1} + \delta_2 \Delta ffr_{t-2} + \delta_3 \Delta ffr_{t-3} + \delta_4 \Delta ffr_{t-4} + e_t$$

Os mínimos quadrados não lineares requerem valores iniciais. A regressão cointegrante é usada para inicializar β_1 e β_2. Os resíduos são obtidos e defasados para serem incluídos em uma regressão linear para inicializar os outros parâmetros. O parâmetro de correção de erros é inicializado em zero.

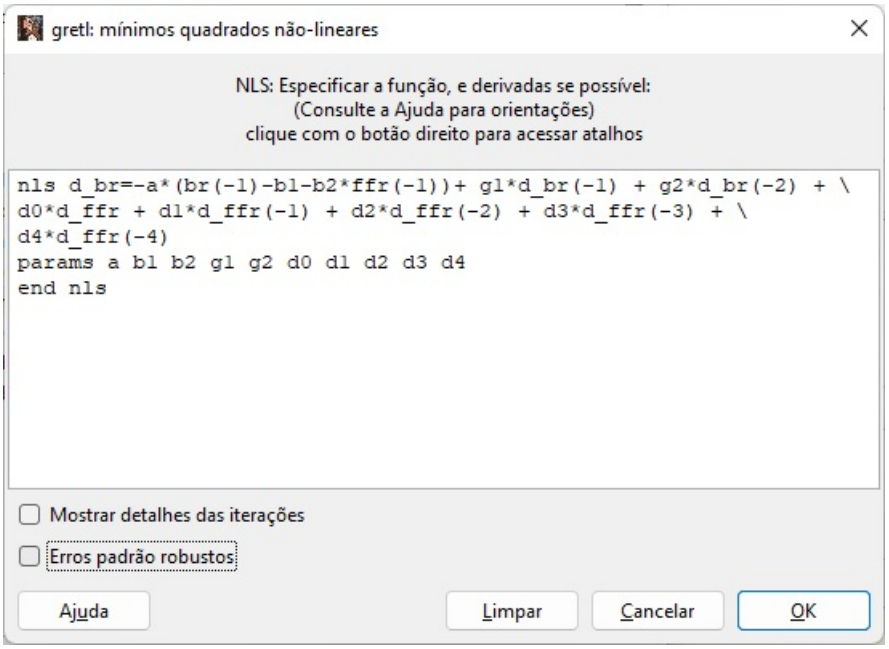
Deve-se estimar uma regressão de `br` contra `ffr` e uma constante. Depois armazena-se os resíduos. Estima-se outra regressão usando `br` em primeira diferença (*d_br*) contra os resíduos defasados, as defasagens de 1 até 2 e a primeira diferença de `ffr` até a sua quarta defasagem. Após rodar o modelo salve os valores dos coeficientes como variáveis:

- *g1 = $coeff(d_br_1)*
- *g2 = $coeff(d_br_2)*
- *d0 = $coeff(d_ffr)*
- *d1 = $coeff(d_ffr_1)*
- *d2 = $coeff(d_ffr_2)*
- *d3 = $coeff(d_ffr_3)*
- *d4 = $coeff(d_ffr_4)*

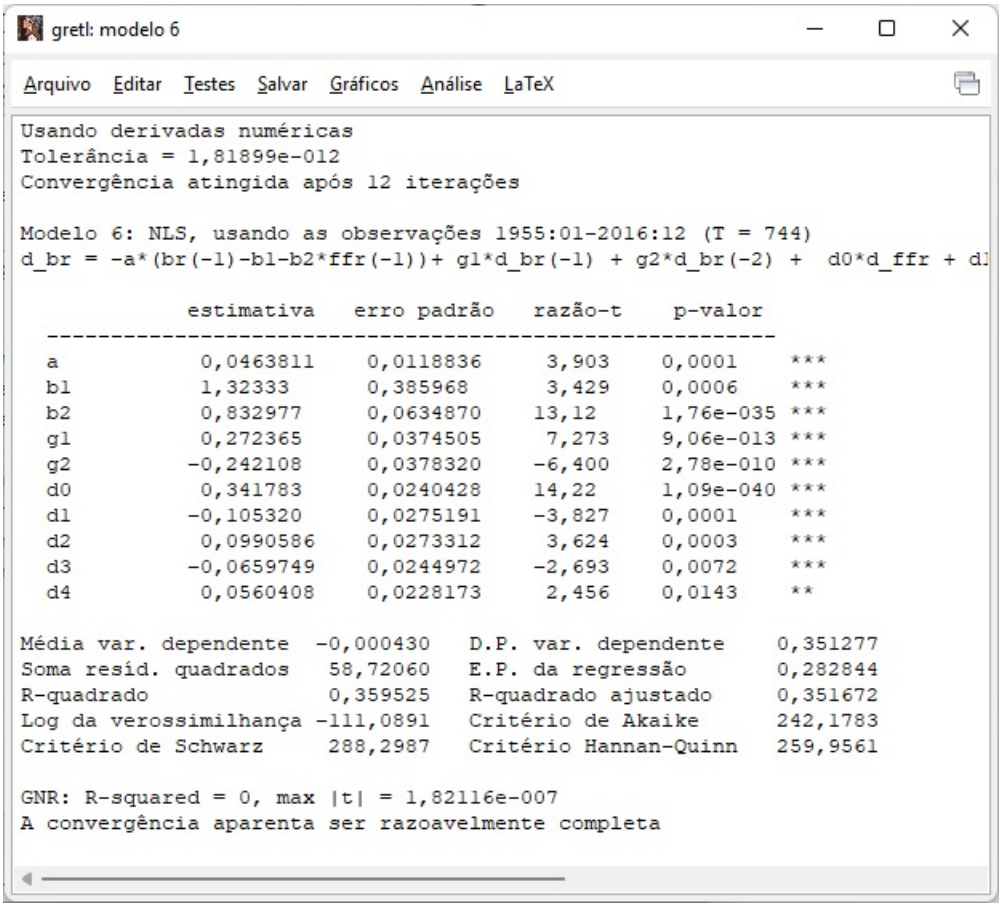
Em seguida rode uma regressão de `br` contra constante e `ffr` e salve os coeficientes da constante e de `ffr`

- *b1 = $coeff(const)*
- *b2 = $coeff(ffr)*
- *a = 0*

Uma vez que os valores declarados são obtidos, um bloco `nls` é construído para estimar o modelo acima. Para estimar esse modelo, clique no menu **Modelo>Mínimos Quadrados Não-Linear (NLS)**. Insira o seguinte código:



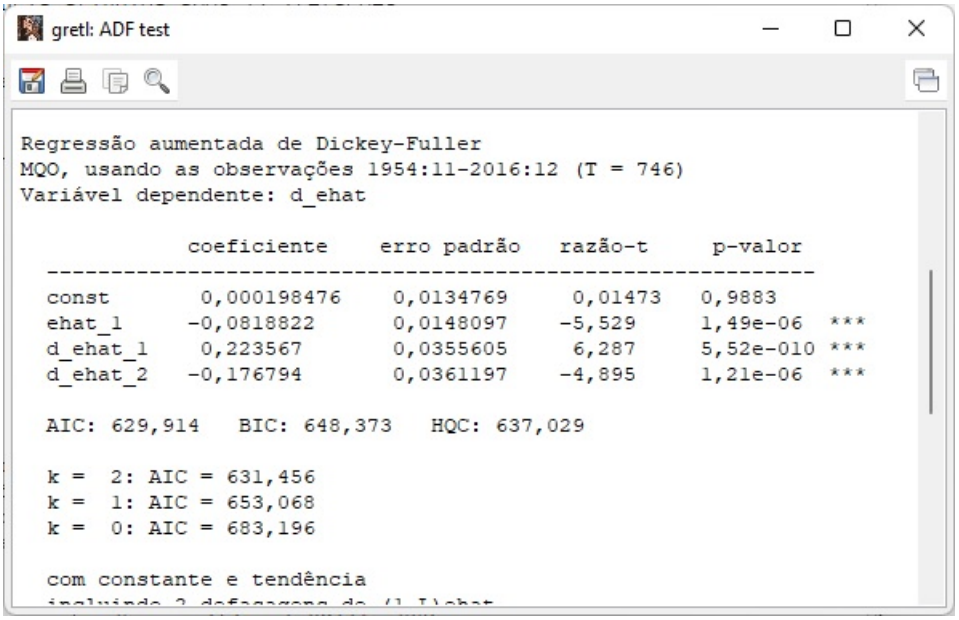
A estimativa pode ser vista na seguinte tela:



Estes correspondem aos resultados anteriores. As estimativas dos parâmetros de cointegração são muito próximas das obtidas por uma regressão simples de `br` sobre `ffr` e uma constante. Finalmente, os parâmetros de cointegração estimados `b1` e `b2` são usados para calcular os resíduos e estes são testados quanto à estacionaridade (também conhecido como `Engle-Granger`). Deve-se utilizar um `teste ADF` e a estatística de teste deve ser comparada com o valor crítico adequado. Para isso, clica-se no menu **Salvar>Definir nova variável**.

- *theta1 = 1-$coeff(a)*
- *ehat = br-$coeff(b1)-$coeff(b2)*ffr*

Depois executa-se um `teste ADF`:



A razão `t` do resíduo defasado é -5.52. Observe que o relacionamento de cointegração contém um intercepto. A conclusão é que a taxa de títulos e a taxa de fundos federais são cointegradas.

Capítulo 9

Vetor de Correção de Erro e Vetor Autorregressivo

O modelo de vetor autorregressivo é uma estrutura geral usada para descrever a interrelação dinâmica entre variáveis estacionárias. Portanto, o primeiro passo na análise deve ser determinar se os dados são estacionários em nível. Caso contrário, tome as primeiras diferenças de seus dados e tente novamente. Normalmente, se os níveis (ou níveis em logaritmo) de sua série temporal não forem estacionários, as primeiras diferenças serão. Se as séries temporais não forem estacionárias, a estrutura VAR precisa ser modificada para permitir uma estimativa consistente das relações entre as séries. O modelo vetorial de correção de erro (VECM) é apenas um caso especial do VAR para variáveis que são estacionárias em suas diferenças (ou seja, $I(1)$). O VECM também pode levar em conta quaisquer relações de cointegração entre as variáveis.

9.1 Modelos VAR e VEC

Considere duas séries temporais com as variáveis y_t e x_t. Generalizando a discussão sobre o relacionamento dinâmico dessas duas séries interrelacionadas em um sistema de equações:

$$y_t = \beta_{10} + \beta_{11}\, y_{t-1} + \beta_{12}\, x_{t-1} + v_t^y$$
$$x_t = \beta_{20} + \beta_{21}\, x_{t-1} + \beta_{22}\, y_{t-1} + v_t^x$$

As equações descrevem um sistema em que cada variável é uma função de sua própria defasagem e da defasagem da outra variável no sistema. Juntas, as equações constituem um sistema conhecido como vetor autorregressivo (VAR). Neste exemplo, como o `lag` máximo é de ordem um, temos um VAR(1).

Se y e x são estacionários, o sistema pode ser estimado usando mínimos quadrados ordinários aplicados a cada equação. Se y e x não são estacionários em seus níveis, mas estacionários em diferenças (ou seja, $I(1)$), então pegue as diferenças e estime:

$$\Delta\, y_t = \Delta\, \beta_{11}\, y_{t-1} + \beta_{12}\, \Delta\, x_{t-1} + v_t^{\Delta y}$$
$$\Delta\, x_t = \beta_{21}\, \Delta\, x_{t-1} + \beta_{22}\, \Delta\, y_{t-1} + v_t^{\Delta x}$$

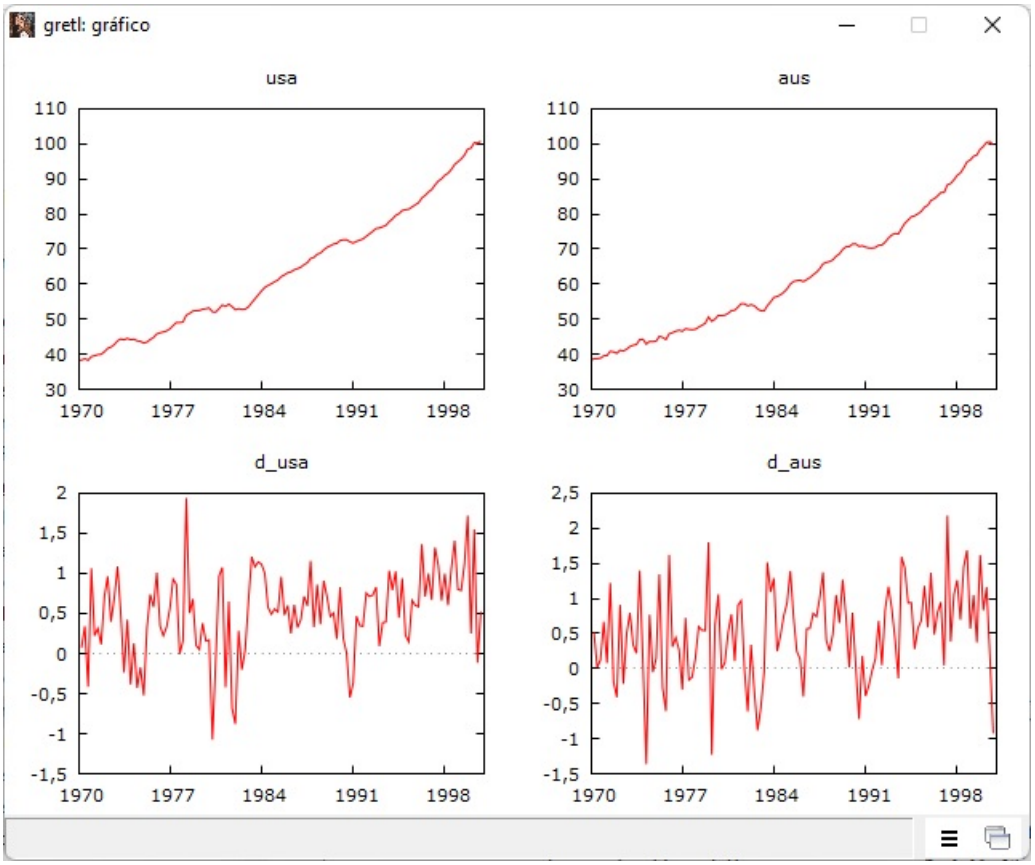
Se y e x são $I(1)$ e cointegrados, então o sistema de equações pode ser modificado para permitir a relação de cointegração entre as variáveis. A relação leva a um modelo conhecido como modelo de correção de erro vetorial (VEC). Serão utilizados dados

macroeconômicos sobre o PIB real para uma economia grande e pequena; `usa` é o PIB real trimestral para os Estados Unidos e `aus` é a série correspondente para a Austrália. Os dados podem ser obtidos no arquivo `gdp.gdt` e já foram dimensionados para que ambas as economias tenham PIB real de 100 no ano 2000, ou seja, ano base igual 2000.

Utiliza-se um modelo VEC porque as séries temporais não são estacionárias em nível, mas são em suas primeiras diferenças e as variáveis são cointegradas.

9.1.1 Gráficos de séries temporais

As impressões iniciais devem ser obtidas observando os gráficos das duas séries. Os gráficos de dados são obtidos da maneira usual após a importação do conjunto de dados. Os dados sobre o PIB dos EUA e da Austrália são encontrados no arquivo `gdp.gdt` e foram coletados de $1970:1 - 2000:4$. Toma-se a primeira diferença das variáveis e plota-se um gráfico de múltiplas séries temporais:

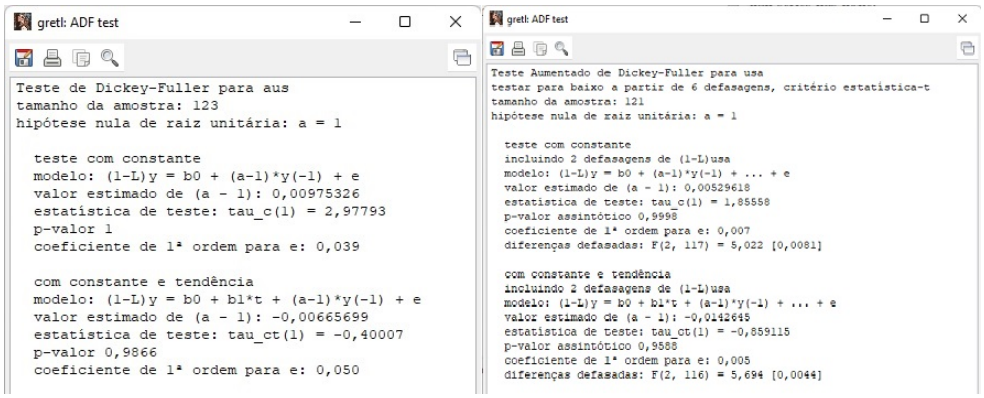


A partir dos gráficos de séries temporais, parece que as séries em nível possuem uma tendência de crescimento ao longo do tempo. As primeiras diferenças possuem uma pequena tendência ascendente. Isso significa que as variáveis em primeira diferença podem ser estacionárias. Pode-se verificar se isso é verdade realizando um `teste ADF`.

Deve-se levar em conta, quantas defasagens devem ser utilizadas no `teste ADF`. Há várias maneiras de selecionar lags e o **gretl** automatiza algumas delas. O conceito básico é incluir lags suficientes nas regressões do `teste ADF` para tornar os resíduos de

ruído branco.

A primeira estratégia é incluir defasagens suficientes para que a última delas seja estatisticamente significativa. O **gretl** automatiza esse processo quando se utiliza a opção do `teste ADF`: *testar para baixo a partir da ordem máxima de defasagem* . Inicie as regressões do ADF com um número bastante generoso de defasagens e o **gretl** reduz automaticamente esse número até que a proporção t na defasagem restante mais longo seja significativa no nível de 10%.



Os `p-valores` da estatística são muito altos para a séries indicando que ambas são não estacionárias em nível. Se esse teste for repetido com as primeiras diferenças das duas séries pode-se ver que elas são estacionárias.

A outra estratégia é testar os resíduos das regressões do `Teste de Dickey-Fuller Aumentado` para autocorrelação. Comece com um modelo pequeno e teste os resíduos da regressão para autocorrelação usando um `teste LM` (ou `LMF`). Se os resíduos forem autocorrelacionados, adicione outra diferença defasada da série à regressão ADF e teste os resíduos novamente. Uma vez que a estatística `LM` é insignificante, termine a testagem. É necessário começar com um número bastante razoável de defasagens no modelo ou os testes não possuirão propriedades desejáveis.

9.1.2 Teste de cointegração

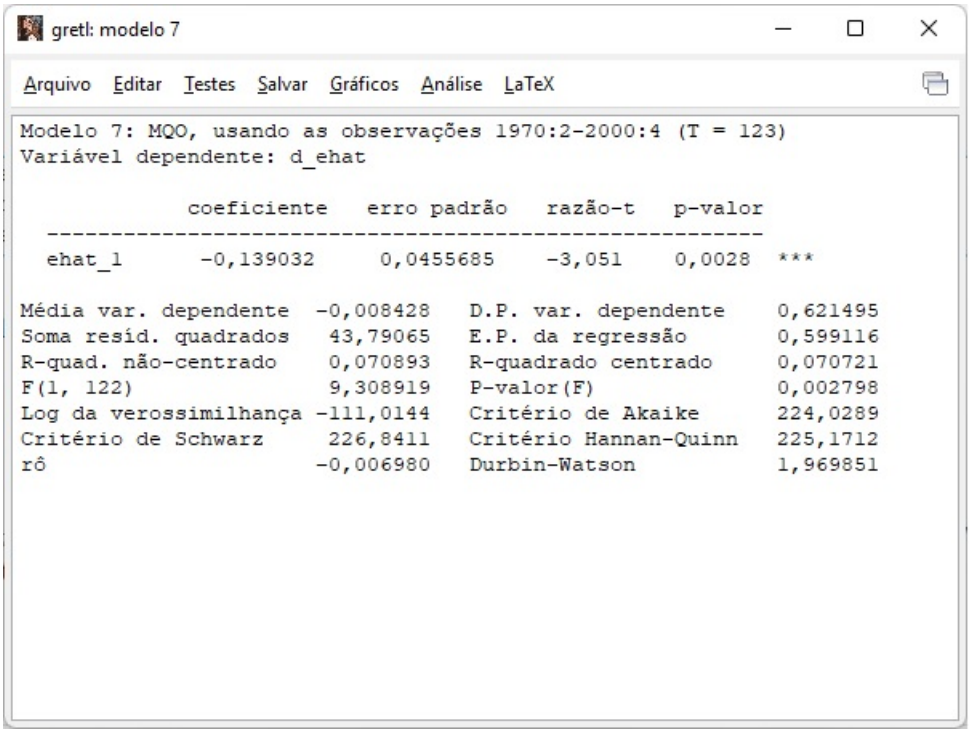
Dado que as duas séries são estacionárias em suas diferenças (ou seja, ambas são $I(1)$), o próximo passo é testar se elas são cointegradas. Para fazer isso, use os mínimos quadrados para estimar a regressão a seguir.

$$aus_t = \beta\, usa_t + e_t$$

Obtenha os resíduos, $\hat{e}_t$ e então estime o seguinte modelo:

$$\Delta\, \hat{e}_t = \gamma\, \hat{e}_{t-1} + u_t$$

Para isso estime a regressão de `aus` contra `usa` e salve os resíduos. A seguir, tome a primeira diferença dos resíduos e faça a regressão da primeira diferença dos resíduos contra os resíduos defasados (sem a inclusão da constante).



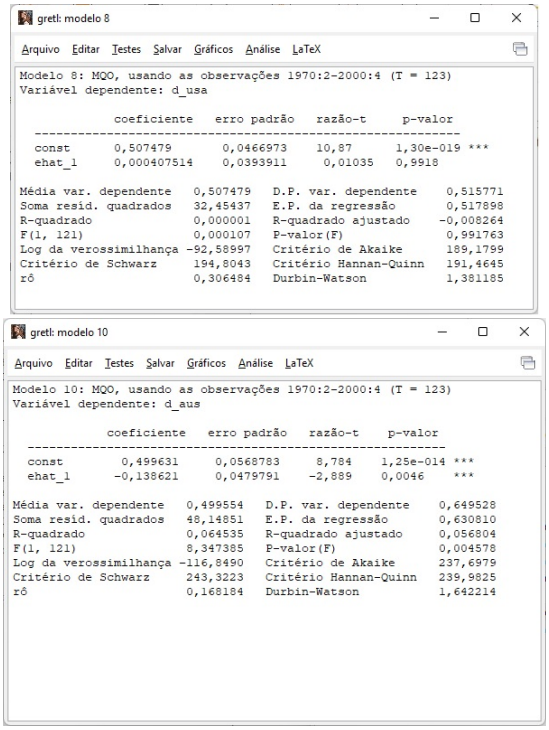
Veja que os resíduos defasados são significativos a 1%, o que permite rejeitar a hipótese nula de não cointegração.

9.1.3 VECM: PIB australiano e americano

Possui-se duas séries que são estacionárias em primeira diferença. Consequentemente, um modelo de correção de erros da dinâmica de curto prazo pode ser estimado usando mínimos quadrados. Um modelo simples de correção de erros é:

$$\Delta\, aus_t = \beta_{11} + \beta_{12}\, \hat{e}_{t-1} + v_{1t}$$
$$\Delta\, aus_t = \beta_{21} + \beta_{22}\, \hat{e}_{t-1} + v_{2t}$$

e as estimativas são dadas por:



O coeficiente negativo significativo em $\hat{e}_{t-1}$ indica que o PIB australiano responde a um desequilíbrio temporário entre os EUA e a Austrália. Os EUA não parecem responder a um desequilíbrio entre as duas economias. A razão t em $\hat{e}_{t-1}$ é insignificante. Esses resultados apoiam a ideia de que as condições econômicas na Austrália dependem daquelas nos EUA mais do que as condições nos EUA dependem da Austrália. Em um modelo simples de comércio de duas economias, os EUA são uma grande economia fechada e a Austrália é uma pequena economia aberta.

9.1.4 Usando o comando vecm

O exemplo do PIB da Austrália/EUA foi realizado manualmente em uma série de etapas para familiarizá-lo com a estrutura do modelo VEC. Na maioria das aplicações, o economista empírico provavelmente usará outros métodos para estimar o VECM. Eles fornecem informações adicionais úteis e geralmente mais eficientes.

Depois de algumas experimentações acaba-se usando um modelo de terceira ordem com apenas 1 vetor cointegrante. Como existem apenas 2 séries, o número máximo e único de vetores cointegrantes é 1. O padrão, "caso 3", que é uma constante irrestrita, é usado para modelar os componentes determinísticos do modelo. Escolher o caso correto é outra parte da arte de fazer um estudo VECM. Assim, cabe ao economista empírico aprofundar os estudos nesta questão para resolver esse problema complicado.

Para estimar o modelo VECM clique em **Modelo>Séries Temporais Multivariadas>VECM**. É possível adicionar variáveis endógenas ao VAR, variáveis exógenas (que devem ser $I(0)$), escolher defasagens, número de vetores cointegrantes e um modelo que incluí uma tendência determinística. A janela oferece acesso imediato a testes, gráficos e ferramentas adicionais para análise. Além disso, há também um

recurso prático que permite uma rápida reespecificarão do modelo. Na barra de menu da janela do modelo, escolha **Editar>Revisar especificação** para abrir a caixa de diálogo VECM novamente para alterar as configurações.

A seguir estão as estimativas da equação de cointegração. Os vetores de ajuste são, na verdade, os coeficientes dos resíduos defasados da relação de cointegração. Geralmente, estes devem ter sinais opostos em dois modelos de variáveis, caso contrário os ajustes aos choques podem não ser equilibrados. Finalmente, algumas estatísticas de seleção de modelo (não mostradas aqui) aparecem na parte inferior que podem ser úteis para determinar a ordem do VECM.

Equação 1: d_usa

	coeficiente	erro padrão	razão-t	p-valor	
const	0,331044	0,114549	2,890	0,0046	***
d_usa_1	0,239698	0,103381	2,319	0,0222	**
d_usa_2	0,289738	0,105083	2,757	0,0068	***
d_aus_1	0,0233260	0,0801741	0,2909	0,7716	
d_aus_2	-0,0866754	0,0788639	-1,099	0,2740	
EC1	-0,0213517	0,0397365	-0,5373	0,5921	

Média var. dependente	0,512457	D.P. var. dependente	0,518283
Soma resíd. quadrados	27,24164	E.P. da regressão	0,486707
R-quadrado	0,154880	R-quadrado ajustado	0,118136
rô	-0,006820	Durbin-Watson	1,990841

Equação 2: d_aus

	coeficiente	erro padrão	razão-t	p-valor	
const	-0,0295258	0,143083	-0,2064	0,8369	
d_usa_1	0,208285	0,129133	1,613	0,1095	
d_usa_2	0,224547	0,131259	1,711	0,0898	*
d_aus_1	4,99573e-05	0,100145	0,0004988	0,9996	
d_aus_2	-0,0431849	0,0985088	-0,4384	0,6619	
EC1	0,125124	0,0496348	2,521	0,0131	**

Média var. dependente	0,503389	D.P. var. dependente	0,653412
Soma resíd. quadrados	42,50365	E.P. da regressão	0,607945
R-quadrado	0,170396	R-quadrado ajustado	0,134326
rô	-0,024521	Durbin-Watson	1,997441

O coeficiente de correção de erro é negativo e diferente de zero para os EUA. A autocorrelação nos resíduos não é evidente. Para a Austrália, o termo de correção de erro não é significativamente diferente de zero e não há autocorrelação remanescente. Uma maneira de avaliar se foram feitas as escolhas de modelagem adequadas é examinar várias estatísticas na saída para verificar a significância dos atrasos, bem como as magnitudes e os sinais dos coeficientes. Verifique se defasagens desnecessárias foram incluídas no modelo (razões t insignificantes nas defasagens mais longas), verifique o valor da estatística de `Durbin-Watson` (deve ser próximo de 2) e verifique os sinais e a significância dos termos de correção de erros . Neste caso, os sinais são os esperados, e apenas a economia australiana se ajusta significativamente aos choques no curto prazo.

Mais uma coisa vale a pena conferir. Plote os termos de correção de erro. Este gráfico mostra que a maior parte do desequilíbrio é negativo. A Austrália está constantemente tentando alcançar os EUA. Note que o coeficiente na equação de cointegração é -1, 025. A estimativa simples dos mínimos quadrados foi -0, 985. Suspeitando que esse parâmetro deva ser igual a -1 (essas economias de mercado são

aproximadamente comparáveis), teste isso usando uma instrução restrita.

Equação 1: d_usa

	coeficiente	erro padrão	razão-t	p-valor	
const	0,331044	0,114549	2,890	0,0046	***
d_usa_1	0,239698	0,103381	2,319	0,0222	**
d_usa_2	0,289738	0,105083	2,757	0,0068	***
d_aus_1	0,0233260	0,0801741	0,2909	0,7716	
d_aus_2	-0,0866754	0,0788639	-1,099	0,2740	
EC1	-0,0213517	0,0397365	-0,5373	0,5921	

Média var. dependente	0,512457	D.P. var. dependente	0,518283
Soma resíd. quadrados	27,24164	E.P. da regressão	0,486707
R-quadrado	0,154880	R-quadrado ajustado	0,118136
rô	-0,006820	Durbin-Watson	1,990841

Equação 2: d_aus

	coeficiente	erro padrão	razão-t	p-valor	
const	-0,0295258	0,143083	-0,2064	0,8369	
d_usa_1	0,208285	0,129133	1,613	0,1095	
d_usa_2	0,224547	0,131259	1,711	0,0898	*
d_aus_1	4,99573e-05	0,100145	0,0004988	0,9996	
d_aus_2	-0,0431849	0,0985088	-0,4384	0,6619	
EC1	0,125124	0,0496348	2,521	0,0131	**

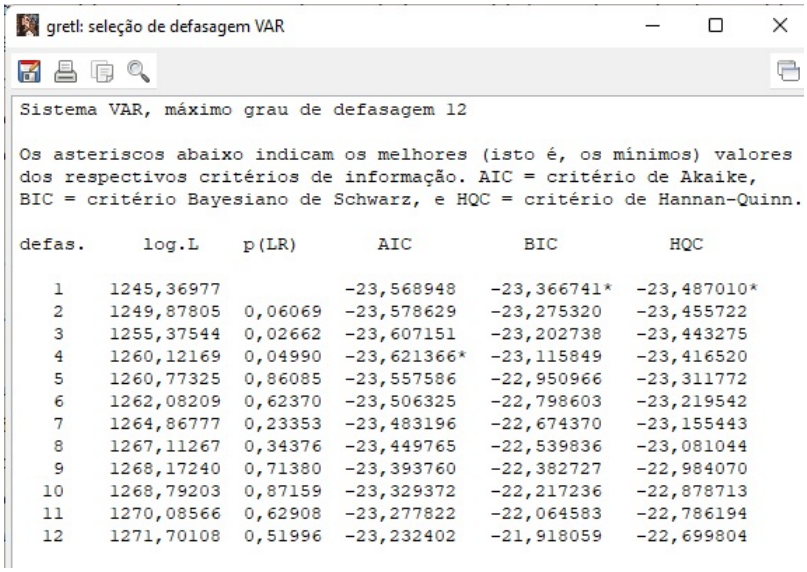
Média var. dependente	0,503389	D.P. var. dependente	0,653412
Soma resíd. quadrados	42,50365	E.P. da regressão	0,607945
R-quadrado	0,170396	R-quadrado ajustado	0,134326
rô	-0,024521	Durbin-Watson	1,997441

9.2 Vetor autoregressivo

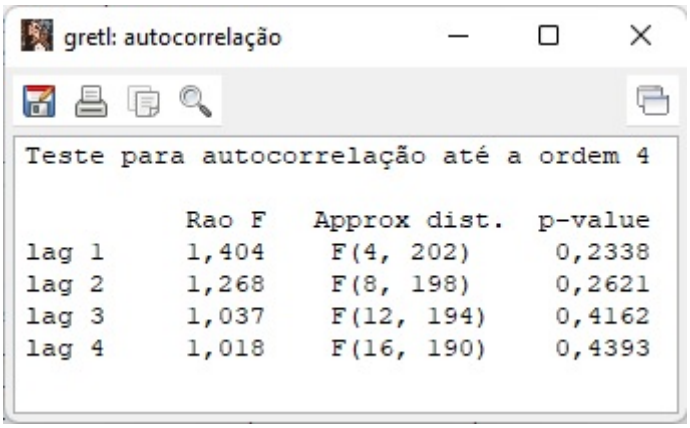
O modelo de vetor autoregressivo (VAR) é, na verdade, um pouco mais simples do que estimar o modelo VEC. É utilizado quando não há cointegração entre as variáveis e é estimado a partir de séries temporais estacionárias.

Serão utilizados os dados macroeconômicos de RPDI e RPCE para os Estados Unidos. Os dados são encontrados no conjunto de dados `fred.gdt` e já foram transformados em seus logaritmos naturais. Na base de dados, y é o logaritmo da renda disponível real e c é o logaritmo das despesas reais de consumo. O primeiro passo é determinar se as variáveis são estacionárias. Se não forem, deve-se transformá-las em séries temporais estacionárias e verificar se há cointegração. Os dados precisam ser analisados da mesma forma que a série do PIB no exemplo do VECM. Examine os gráficos para determinar possíveis tendências e use os `testes ADF` para determinar em quais formas os dados são estacionários. Esses dados são não estacionários em níveis, mas estacionários em diferenças. Em seguida, estime o vetor de cointegração e teste a estacionaridade de seus resíduos. Se os resíduos forem estacionários, as séries são cointegradas e, então, estima-se um VECM. Caso contrário, um tratamento VAR é suficiente.

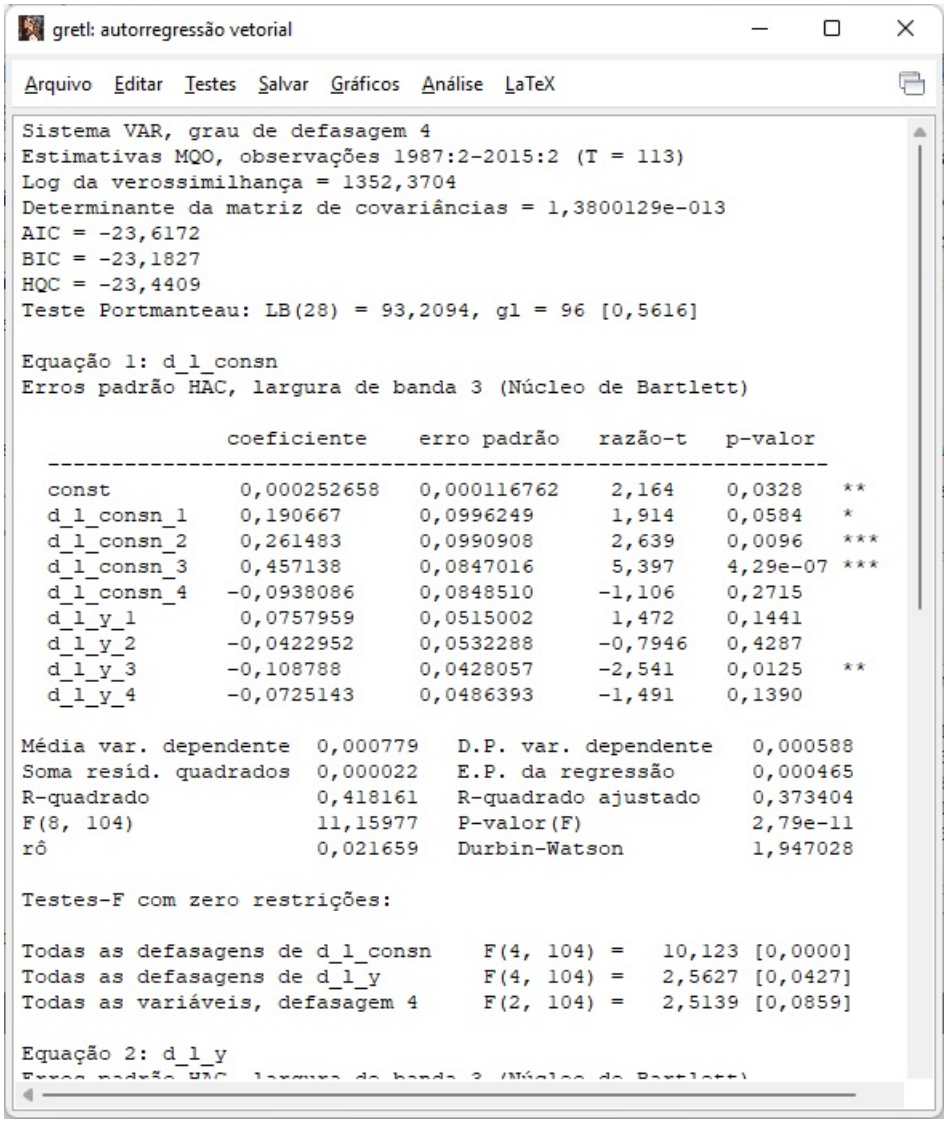
Para selecionar o número de defagens a serem incluídas no VAR, clique no menu **Modelo>Séries Temporais Multivariadas>Seleção de defasagnes do VAR**. Escolha um número suficientemente grande de defasagem para a testagem.



Pode-se observar que conforme os valores dos `testes BIC` e `HQC` deve-se escolher o modelo com apenas 1 defasagem. No entanto, deve-se verificar se há alguma correlação serial nos resíduos. Para isso, após estimarmos o modelo VAR clicar no menu **Modelo>Séries Temporais Multivariadas>Autoregressão Vetorial** com apenas 1 defagem e uma matriz de `variância-covariância HAC`, deve-se realizar um `teste de autocorrelação de Ljung-Box`. Observe que a autocorrelação some após inserirmos 4 defasagens. Em outras palavras, os `p-valores` são superiores a 0.10 o que permite não rejeitar a hipótese nula de não autocorrelação.



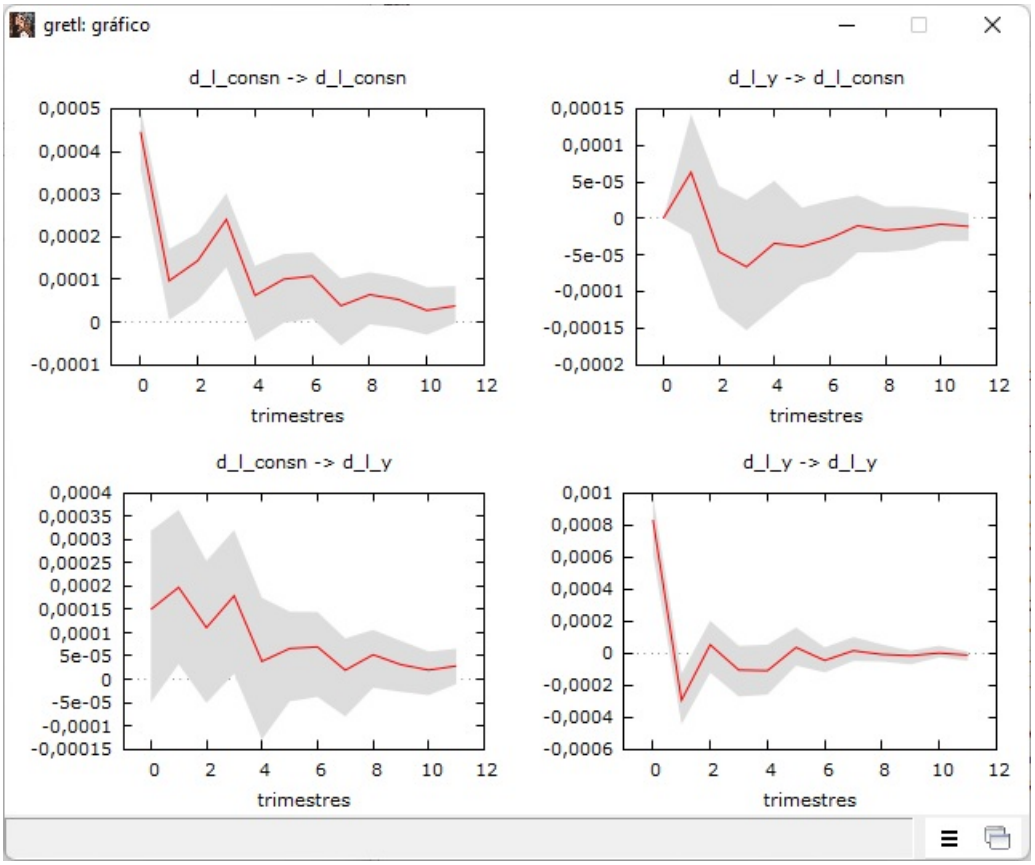
Dessa forma, deve-se estimar um modelo VAR com 4 defasagens:



9.2.1 Funções de impulso resposta e decomposição de variância

As funções de impulso resposta mostram os efeitos dos choques na trajetória de ajuste das variáveis. As decomposições da variância do erro de previsão medem a contribuição de cada tipo de choque para a variância do erro de previsão. Ambos os cálculos são úteis para avaliar como os choques nas variáveis econômicas reverberam em um sistema. Funções de impulso resposta e decomposições de variância de erro de previsão podem ser produzidas após a estimação dos modelos VAR e VECM. Os resultados podem ser apresentados em uma tabela ou gráfico.

Para gerar os gráficos das funções de impulso resposta, após estimar o modelo VAR, deve-se clicar no menu **Gráfico>Impulso Resposta**. Nesse menu, pode-se escolher se quer observar os efeitos dos choques em um único gráfico ou se deseja acompanhar o efeito exclusivo em alguma das variáveis do modelo. Abaixo apresenta-se um gráfico com os múltiplos choques:



O período escolhido para acompanhar os choques foi de 12 trimestres. Um exemplo de interpretação é que o efeito de um choque na variação da renda pessoal disponível faz com que a variação dos gastos em consumo aumente muito pouco nos 2 primeiros trimestres. Após isso, essa variação será negativa até se aproximar de zero no sexto até o oitavo trimestre.

No menu análise, pode-se ver os valores para as funções de impulso resposta, bem como, para a decomposição de variância do erro de previsão.

gretl: decomposição da variância

Decomposição da variância para d_l_consn

período	erro padrão	d_l_consn	d_l_y
1	0,000446184	100,0000	0,0000
2	0,00046083	98,1247	1,8753
3	0,000484852	97,4231	2,5769
4	0,000545186	96,4821	3,5179
5	0,000549787	96,1531	3,8469
6	0,000560198	95,8163	4,1837
7	0,000571087	95,7461	4,2539
8	0,000572431	95,7353	4,2647
9	0,00057622	95,7110	4,2890
10	0,000578778	95,6953	4,3047
11	0,000579463	95,6866	4,3134
12	0,000580787	95,6705	4,3295

Decomposição da variância para d_l_y

período	erro padrão	d_l_consn	d_l_y
1	0,000845999	3,1465	96,8535
2	0,000917498	7,3029	92,6971
3	0,000925642	8,6091	91,3909
4	0,000948832	11,7679	88,2321
5	0,00095599	11,7535	88,2465
6	0,000958916	12,1570	87,8430
7	0,000962532	12,5872	87,4128
8	0,000962837	12,6196	87,3804
9	0,00096431	12,8780	87,1220
10	0,000964959	12,9672	87,0328
11	0,000965161	13,0036	86,9964
12	0,000965665	13,0768	86,9232

Capítulo 10

Dados em Painel

Um painel de dados consiste em um grupo de unidades transversais (pessoas, empresas, estados ou países) que são observadas ao longo do tempo. Denota-se o número de unidades transversais por n e o número de períodos de tempo que são observados como T. Para usar os procedimentos predefinidos para estimar modelos usando dados de painel em **gretl**, deve-se ter certeza de que os dados foram estruturados corretamente no programa.

As caixas de diálogo para atribuir a estrutura do conjunto de dados do painel usando variáveis de índice. Para usar este método, os dados devem incluir variáveis que identifiquem cada indivíduo e período de tempo. O **gretl** fornece acesso fácil a vários conjuntos de dados de painel úteis por meio de seu servidor de banco de dados. Incluí a Penn World Table e os dados de Barro e Lee (1996) sobre desempenho educacional internacional. Esses dados podem ser instalados usando o menu **Arquivo>Base de Dados>No servidor de base de dados**.

10.1 Um modelo básico

A expressão mais geral dos modelos de regressão linear que possuem dimensões de tempo e unidade é vista na equação abaixo.

$$y_{it} = \beta_{1it} + \beta_{2it}\, x_{2it} + \beta_{3it}\, x_{3it} + e_{it} \tag{10.1}$$

sendo $i = 1, \ldots, n$ e $t = 1, \ldots, T$. Se tiver um conjunto completo de observações de tempo para cada indivíduo, haverá nT observações totais na amostra. Neste caso, diz que o painel está equilibrado. Não é incomum ter algumas observações de tempo perdido para um ou mais indivíduos. Quando isso acontece, o número total de observações é menor que nT e o painel fica desbalanceado.

O maior problema com a Equação 10.1 é que mesmo que o painel esteja balanceado, o modelo contém 3 vezes mais parâmetros do que observações (nT)! Para poder estimar o modelo, algumas suposições devem ser feitas a fim de reduzir o número de parâmetros. Uma das suposições mais comuns é que as inclinações são constantes para cada indivíduo e para cada período de tempo; além disso, as interceptações variam apenas por indivíduo. Este modelo é mostrado na Equação 10.2.

$$y_{it} = \beta_{1i} + \beta_2\, x_{2it} + \beta_3\, x_{3it} + e_{it} \tag{10.2}$$

Essa especificação incluí $n + 2$ parâmetros, incluí variáveis *dummy* que permitem separar cada intercepto para cada indivíduo. Tal modelo implica que não há mudanças substantivas na função de regressão em curtos períodos de tempo. Obviamente, quanto maior a dimensão do tempo, maior a probabilidade de essa suposição ser falsa.

10.2 Efeitos Fixos

Na Equação 10.2 os parâmetros que variam por indivíduo são chamados de efeitos fixos individuais e o modelo é referido como efeitos fixos unidirecionais. O modelo é adequado quando os indivíduos da amostra diferem uns dos outros de uma forma que não varia ao longo do tempo. É uma maneira útil de evitar diferenças não observadas entre os indivíduos da amostra que, de outra forma, teriam de ser omitidas. Lembre-se de que a omissão de variáveis relevantes pode fazer com que os mínimos quadrados sejam tendenciosos e inconsistentes; um modelo de efeitos fixos unidirecional, que requer o uso de dados de painel, pode ser muito útil para mitigar o viés associado a efeitos não observáveis invariantes no tempo.

Para painéis mais longos em que a função de regressão está mudando ao longo do tempo, variáveis fictícias de tempo $T - 1$ podem ser adicionadas ao modelo. O modelo torna-se:

$$y_{it} = \beta_{1i} + \beta_{1t} + \beta_2\, x_{2it} + \beta_3\, x_{3it} + e_{it} \tag{10.3}$$

em que β_{1i} ou β_{1t} devem ser omitidos para evitar colinearidade perfeita. Este modelo contém $n + (T - 1) + 2$ parâmetros que geralmente é menor que as nT observações na amostra. A Equação 10.3 é chamada de modelo de efeitos fixos bidirecionais porque contém parâmetros que serão estimados para cada indivíduo e cada período de tempo.

Ainda é possível reescrever a Equação 10.3 da seguinte forma:

$$y_{it} = \beta_2\, x_{2it} + \beta_3\, x_{3it} + c_i + e_{it} \tag{10.4}$$

sendo c_i o efeito fixo individual que está potencialmente correlacionado com os regressores x. Pode-se escrever Equação 10.4 tomando a média para cada unidade i:

$$\bar{y}_i = \beta_2\, \bar{x}_{2i} + \beta_3\, \bar{x}_{3i} + \bar{c}_i + \bar{e}_i \tag{10.5}$$

Subtraindo (10.4) de (10.5) tem-se que:

$$(y_{it} - \bar{y}_i) = \beta_2\, (x_{2it} - \bar{x}_{2i}) + \beta_3\, (x_{3it} - \bar{x}_{3i}) + (c_i - \bar{c}_i) + (e_{it} - \bar{e}_i)$$

$$y^*_{it} = \beta_2\, x^*_{2it} + \beta_3\, x^*_{3it} + e^*_{it} \tag{10.6}$$

Os termos com asterisco se referem aos termos entre parênteses que são diferenciados pela média. Observe que c_i e todos os demais termos que são constantes no tempo serão eliminados com esta transformação. Esse estimador é conhecido `Within` e pode ser estimado por MQO.

10.3 Primeira diferença

Antes de apresentar o modelo de primeira diferenças importa-se o conjunto de dados `nls_panel.csv` que inclui um subconjunto do *National Longitudinal Survey*, conduzido pelo Departamento de Trabalho dos EUA. A base de dados inclui observações sobre mulheres, em 1968, com idades compreendidas entre os 14 e os 24 anos. Em seguida, acompanha-as ao longo do tempo, registando vários aspectos das suas vidas anualmente até 1973 e semestralmente depois. A amostra é composta por 716 mulheres observadas em 5 anos (1982, 1983, 1985, 1987 e 1988). O painel é equilibrado e há um total de 3.580 observações.

O primeiro passo trata-se da importação desse conjunto de dados. Para tanto, clica-se no menu **Arquivo>Abri dados>Arquivos do usuário**. Não esqueça de selecionar a opção para leitura de arquivos `CSV` ou para a leitura de qualquer tipo de arquivo. Essa opção fica no menu localizado acima do botão `Abrir`. Ao abrir os dados o **gretl** perguntará se deseja interpretar a primeira apenas como uma coluna, você deve marcar que `Não`. Posteriormente, o software lhe oferecerá algumas opções para que se possa definir a base de dados no formato de dados em painel. Quando perguntado sobre a estrutura de dados, selecione a opção **dados em painel>usar variáveis índice**. Selecione o `id` como variável de unidade ou de grupo e o ano (year) como variável de índice de tempo. A dimensão temporal do painel é anual tendo começado em 1982.

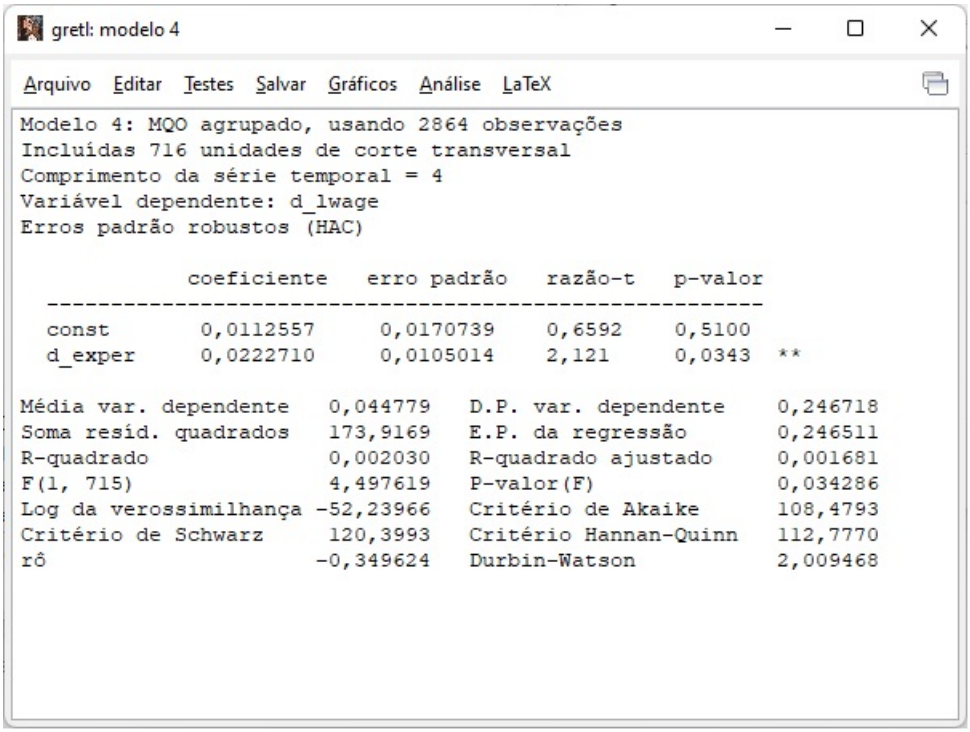
Para utilizar o estimador de primeiras diferenças são necessários pelo menos dois períodos de tempo, e se deve diferenciar as variáveis no tempo e estimar o modelo por MQO. As variáveis invariantes no tempo e a interceptação saem do modelo após a diferenciação. Por exemplo, se desejar estimar o seguinte modelo:

$$\ln(wage_{it}) = \beta_{1i} + \beta_2\, educ_{it} + \beta_3\, exper_{3it} + e_{it} \tag{10.7}$$

Tomando a primeira diferença, note que o termo *educ* desaparece da equação:

$$\Delta \ln(wage_{it}) = \Delta\, \beta_3\, exper_{it} + \Delta\, e_{it} \tag{10.8}$$

Para estimar esse modelo, clique no menu **Modelo>Mínimos Quadrados Ordinários** e selecione as variáveis em primeira diferença.



Embora o modelo seja simples, é possível observar que a variação da experiência influencia positivamente a variação no salário.

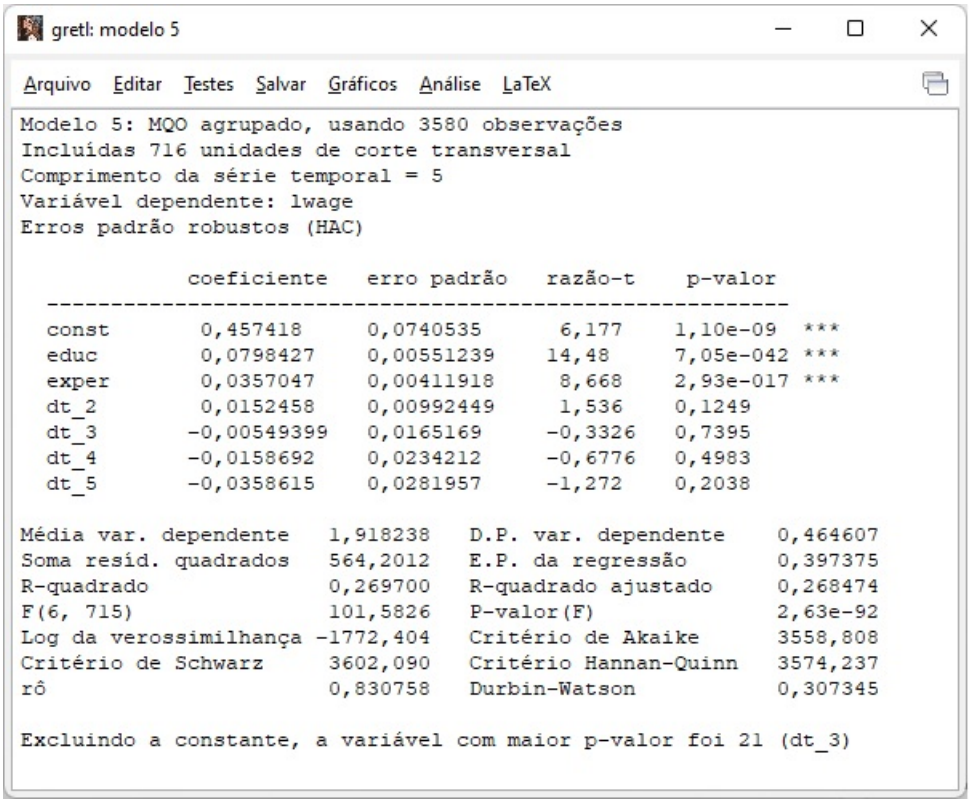
Por fim, qual estimador utilizar: efeitos fixos ou primeira diferença? O estimador de primeira diferença pode ser usado se $T > 2$. Se $T = 2$ ambos estimadores são idênticos. Para $T > 2$, o estimador de efeitos fixos é mais eficiente se os pressupostos clássicos são satisfeitos. O método de primeira diferença pode ser melhor caso os resíduos apresentem correlação serial e se `T` é muito grande e o número de unidades `N` não é tão grande. Nesse caso, o painel apresenta características de séries temporais e alguns problemas de dependência podem surgir, assim provavelmente o estimador de primeiras diferenças é mais apropriado. Caso contrário, é melhor realizar as duas estimativas e checar a robustez.

10.4 Painel Agrupado

Para estimar o modelo da Equação 10.7 deve-se fazer o mesmo procedimento com as variáveis em nível, sem estarem em primeira diferença. Para isso deve-se estimar a seguinte equação:

$$\ln(wage_{it}) = \beta_1 + \beta_2\, educ_{it} + \beta_3\, exper_{3it} + \gamma_t + e_{it} \tag{10.9}$$

Note que foram incluído efeitos fixos temporais (γ_t), isto é, *dummies* de ano. Em seguida realiza-se a estimação desse modelo por MQO.



Veja que a educação e a experiência possuem um efeito positivo sobre o salário. Observe que as *dummies* temporais não são significativas.

10.5 Efeitos Aleatórios

O estimador de efeitos aleatórios trata as diferenças individuais como sendo atribuídas aleatoriamente aos indivíduos. Ao invés de estimá-los como parâmetros como realizado no modelo de efeitos fixos, aqui eles são incorporados ao erro do modelo, que em um painel terá uma estrutura específica. O termo β_{1i} na Equação 10.3 é modelado:

$$\beta_{1i} = \bar{\beta}_1 + u_t \tag{10.10}$$

em que u_i são as diferenças individuais aleatórias que são as mesmas em cada período de tempo.

$$\begin{aligned} y_{it} &= \bar{\beta}_1 + \beta_2\, x_{2it} + \beta_3\, x_{3it} + (e_{it} + u_i) \\ &= \bar{\beta}_1 + \beta_2\, x_{2it} + \beta_3\, x_{3it} + v_{it} \end{aligned} \tag{10.11}$$

o termo combinado de erro é chamado de erro de composição:

$$v_{it} = e_{it} + u_i$$

A propriedade chave é que novo termos de erro é homocedástico e serialmente correlacionado:

$$\sigma_v^2 = var\left(v_{it}\right) = var\left(e_{it} + u_i\right) = \sigma_u^2 + \sigma_e^2$$

Para o indivíduo i a covariância entre os erros é:

$$cov\left(v_{it},\, v_{is}\right) = 0$$

para $t \neq s$. Além disso, a covariância entre quaisquer dois indivíduos é zero. Uma das principais vantagens do modelo de efeitos aleatórios é que os parâmetros dos regressores invariantes no tempo podem ser estimados. As estimativas dos parâmetros são realmente obtidas através de Mínimos Quadrados Generalizados Factível (MQGEF). A transformação que é usada nas variáveis do modelo é algumas vezes chamada de quase-degradação. É baseado no cálculo de:

$$\theta = 1 - \frac{\sigma_e}{\sqrt{T\sigma_u^2 + \sigma_e^2}}$$

Com $\theta \in [0,\, 1]$. Lembre-se do estimador `Within` de efeitos fixos. Deve-se fazer a diferenciação da média de cada unidade i multiplicada pelo parâmetro θ , como segue:

$$\left(y_{it} - \theta\, \bar{y}_i\right) = \left(\bar{\beta}_1 - \theta\, \bar{\beta}_1\right) + \beta_2\left(x_{2it} - \theta\, \bar{x}_{2it}\right) + \beta_3\left(x_{3it} - \theta\, \bar{x}_{3it}\right) + \left(v_{it} - \theta\, \bar{v}_{it}\right)$$

$$y_{it}^* = \beta_1 + \beta_2\, x_{2it}^* + \beta_3\, x_{3it}^* + v_{it}^*$$

As variáveis em asterisco referem-se aos termos em parênteses e a constante é definida como $\beta_1 = \left(\bar{\beta}_1 - \theta\, \bar{\beta}_1\right)$.

10.6 Testes de diagnóstico de painel

Há alguns testes de especificação chave que devem ser feitos antes de confiar nos efeitos fixos, aleatórios ou nos estimadores de mínimos quadrados agrupados. Para consistência, todos exigem que a heterogeneidade não observada não esteja correlacionada com os regressores do modelo. Isso é testado usando uma versão de um `teste de Hausman`. O outro teste é para a presença de efeitos aleatórios, esse teste é um `teste LM` que às vezes é referido como `Breusch-Pagan`.

10.6.1 Breusch-Pagan

O `teste de Breusch-Pagan` é baseado numa estatística teste de um `multiplicador de Lagrange` e é calculado da seguinte forma:

$$LM = \sqrt{\frac{nT}{2\left(T-1\right)}}\ \left\{\frac{\sum_{i=1}^{n}\left(\sum_{i=1}^{n}\hat{e}_{it}\right)^2}{\sum_{i=1}^{n}\sum_{i=1}^{T}\hat{e}_{it}^2} - 1\right\}$$

Com a hipótese nula $H_0 : \sigma_u^2 = 0$ contra a alternativa que $H_1 : \sigma_u^2 \geq 0$. Sob a hipótese nula $LM \sim N(0,\, 1)$ e a melhor ideia é realizar um teste unicaudal. Infelizmente o **gretl** e outros softwares relatam o LM^2 e usam uma distribuição $\chi^2\left(1\right)$ que faz com que a hipótese alternativa seja $H_1 : \sigma_u^2 \neq 0$.

A boa notícia é que pelo menos **gretl** calcula LM^2 por padrão sempre que se estima um modelo de efeitos aleatórios. A rejeição da hipótese nula significa que o indivíduo (e neste modelo, aleatório) as diferenças possuem variância. Se o economista empírico não conseguir rejeitar a hipótese nula, provavelmente desejará usar Mínimos Quadrados Agrupados. Se os efeitos individuais aleatórios estiverem correlacionados com regressores, então o estimador de efeitos aleatórios não será consistente. Um teste estatístico desta proposição deve ser feito sempre que este estimador for utilizado, a fim de reduzir a chance de erro de especificação do modelo.

10.6.2 Hausman

O `teste de Hausman` prova a consistência do estimador de efeitos aleatórios. A hipótese nula é que essas estimativas são consistentes, ou seja, exige que a hipótese de ortogonalidade dos resíduos seja satisfeita. O teste é baseado numa medida, H, que é uma "distância" entre os estimadores de efeitos fixos e efeitos aleatórios. Essa medida é construída de modo que sob o nulo segue a distribuição χ^2 com graus de liberdade iguais ao número de regressores, J, que variam no tempo. Se o valor de H for grande, isso sugere que o estimador de efeitos aleatórios não é consistente e o modelo de efeitos fixos é preferível.

Para calcular o teste, os seguintes procedimentos devem ser realizados:

1. Considere o modelo de efeitos aleatórios como o "modelo restrito", e salve a soma dos quadrados dos resíduos como (SQR_r);

2. Estime via MQO um modelo irrestrito em que a variável dependente é y (diferenciada da média) e os regressores incluem X (diferenciado na média) (como no modelo RE) e as variantes diminuídas de todas as variáveis variantes no tempo (ou seja, os regressores de efeitos fixos);

3. Registre a soma dos resíduos quadrados deste modelo como SQR_u e;

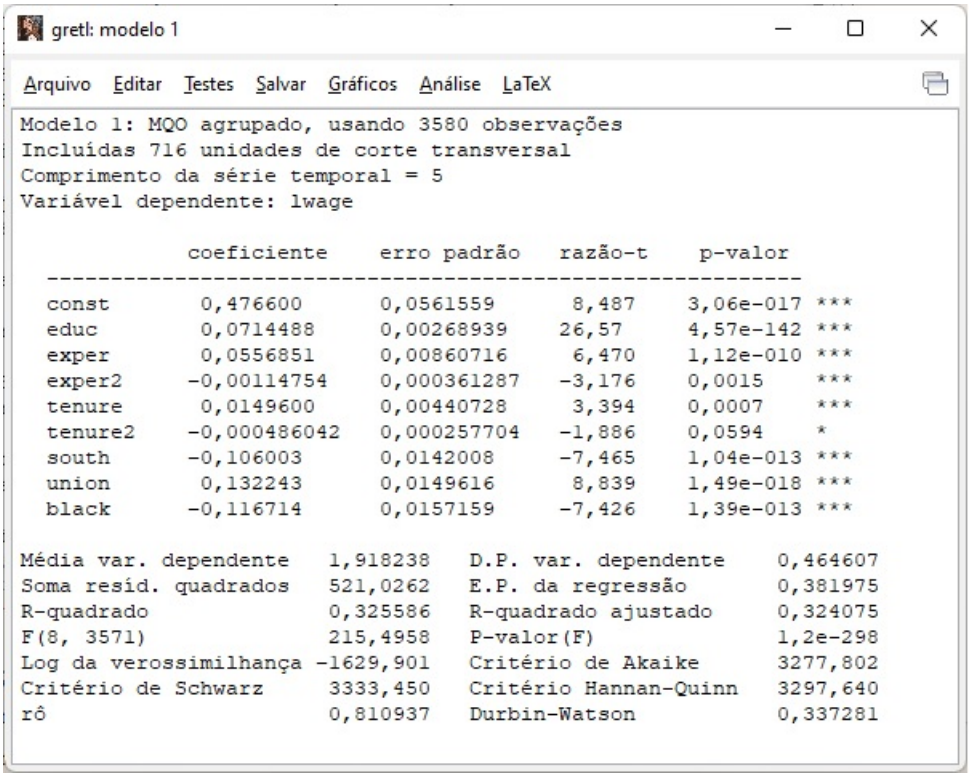
4. Calcule $H = {}^{n(SSR_r - SSR_u)}/_{SSR_u}$, em que n é o número total de observações usadas. Nesta variante, H não pode ser negativo, uma vez que adicionar regressores adicionais ao modelo efeitos aleatórios não pode aumentar o SQR.

10.7 Exemplo

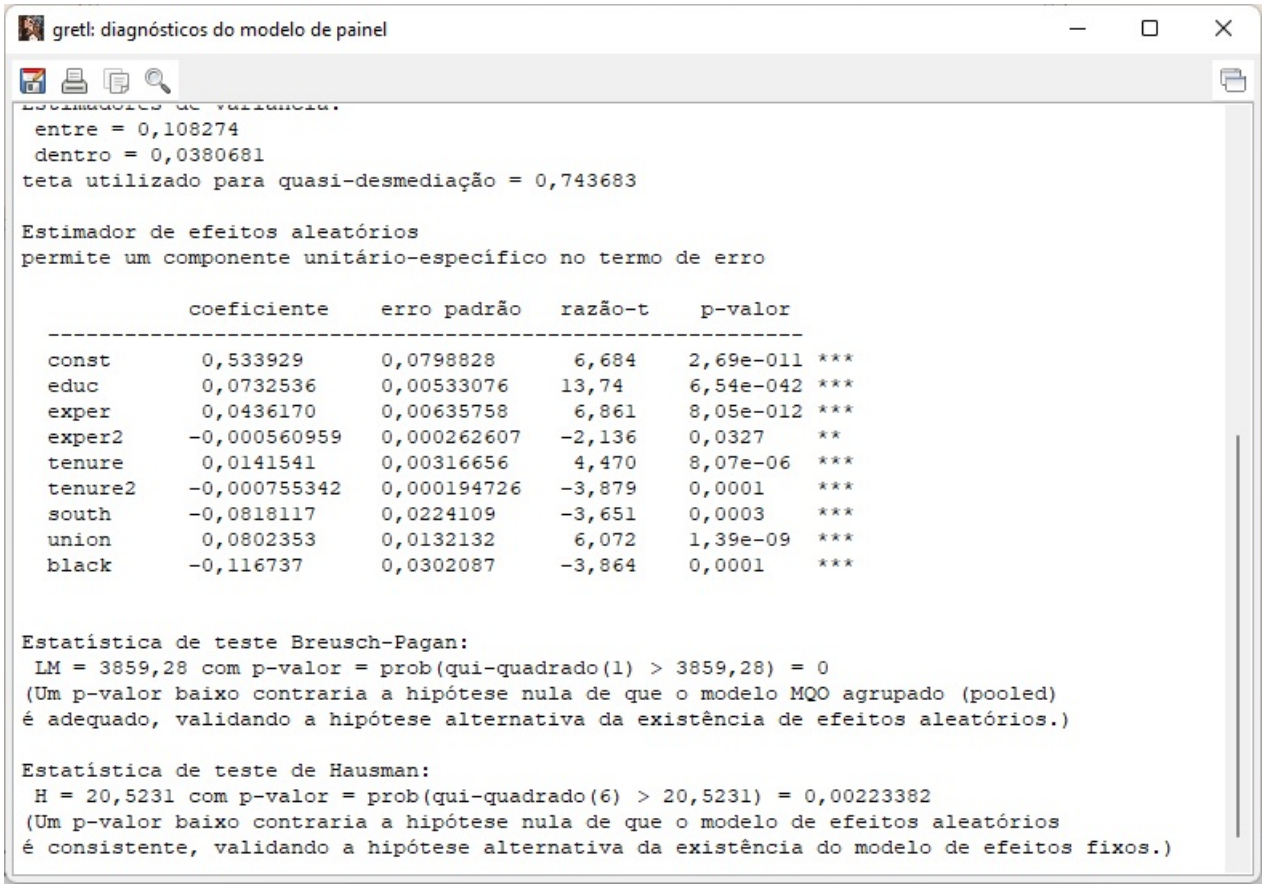
Com base no arquivo `nls_panel.gdt` estima-se o seguinte modelo:

$$\begin{aligned}\ln(wage_{it}) = {} & \beta_1 + \beta_2\, educ_{it} + \beta_3\, exper_{it} + \beta_4\, exper_{it}^2 + \beta_5\, tenure_{it} + \\ & + \beta_6\, tenure_{it}^2 + \gamma_1\, south + \gamma_2\, union + \gamma_3\, black + e_{it}\end{aligned} \tag{10.12}$$

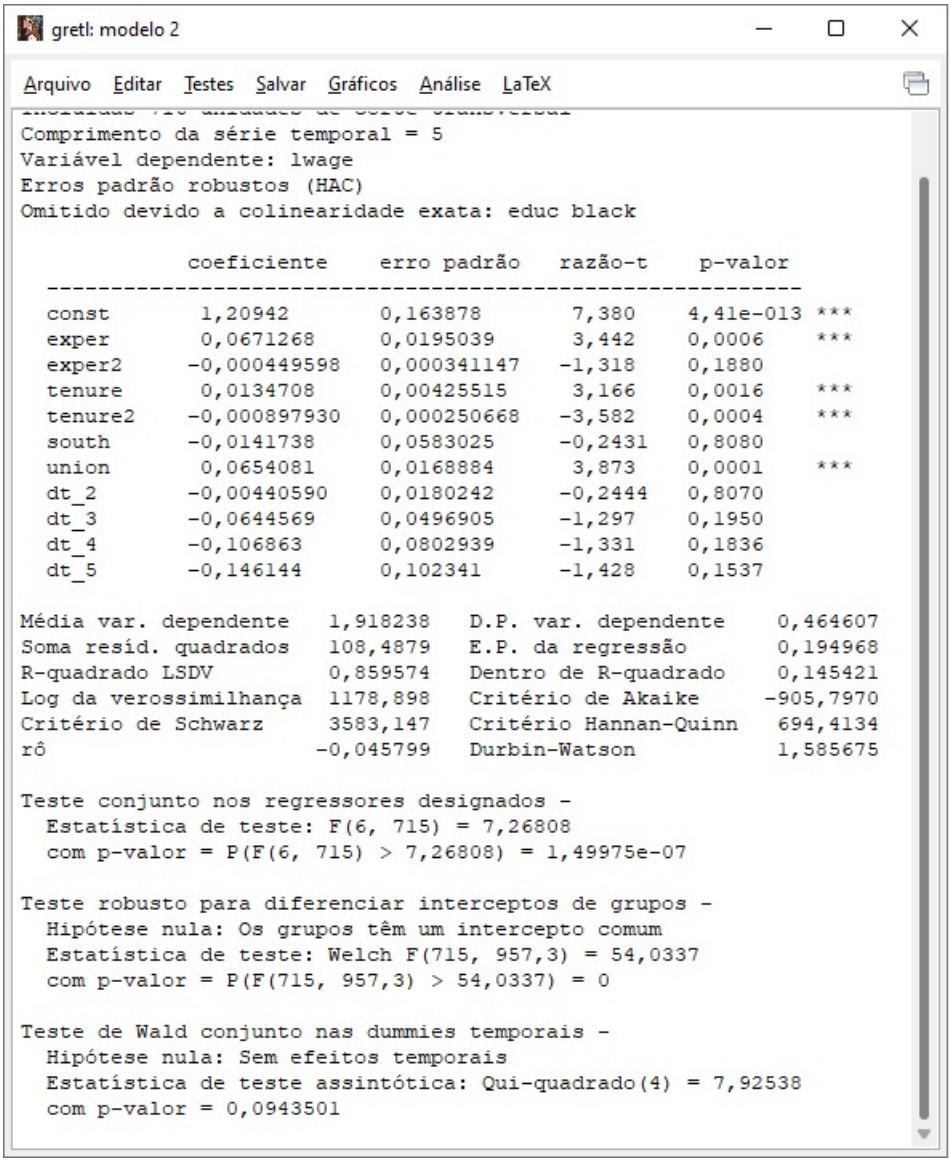
Para isso clique no menu **Modelo>Mínimos Quadrados Ordinários**. Esse é o modelo de painel agrupado:



Após a estimativa deve-se clicar no menu **Teste>Especificação de Painel**. Ao fazer isso o **gretl** nos mostrará a seguinte saída:



Veja que o **gretl** já faz os dois testes de especificação que foram discutidos anteriormente. De acordo com o `teste LM` o modelo de efeitos aleatórios é adequado em relação ao MQO. Conforme o `teste de Hausman`, verifica-se que o modelo de Efeitos Fixo é adequado em relação ao modelo de Efeitos Aleatórios. Dessa forma, deve-se realizar a estimação do modelo de efeitos fixos. Para isso, clique no menu **Modelo>Painel>Efeitos Fixos ou Aleatórios**. Escolha o modelo de efeitos fixos e marque as opções para inclusão de *dummies* temporais e erros padrões robustos.



Como as variáveis *educ* e *black* possuem pouca ou nenhuma varição temporal elas são removidas do modelo. Note que a inclusão das *dummies* temporais não foi importante para estimação do modelo.

Capítulo 11

Modelos com variável dependente qualitativa ou categórica

Há muitos eventos na economia que não podem ser quantificados de forma significativa. Como você vota em uma eleição, se você vai para a pós-graduação, se você possui o trabalho assalariado ou qual faculdade você escolhe não há uma forma natural de ser quantificado. Cada um deles expressa uma qualidade ou condição que você possui. Modelos de como essas decisões são determinadas por variáveis que são chamados de escolha qualitativa ou modelos de variáveis qualitativas.

As escolhas podem ser entre duas (binárias) ou mais (multinomiais) alternativas. Escolhas multinomiais podem ser feitas a partir de uma hierarquia (ordenadas) ou não. Por exemplo, uma escolha de uma escala de satisfação é ordenada e a escolha de ir a pé, de carro ou de ônibus para o trabalho não. Uma variável dependente limitada é contínua, mas sua faixa de valores é restrita de alguma forma. Alguns dos valores da variável dependente não são observados ou, se todos forem observados, alguns são restritos ao mesmo valor se o valor real exceder (ou cair abaixo) algum limite. Versões simples de ambos os tipos de modelo são consideradas abaixo.

Inicia-se com decisões binárias e depois passa-se para modelos de escolha multinomial. Modelos para dados de contagem são estimados e regressões censuradas e truncadas são consideradas.

11.1 Modelo de probabilidade linear

Em um modelo de escolha binária, a decisão de modelar tem apenas dois resultados possíveis. Um número artificial é atribuído a cada resultado antes que análises empíricas adicionais possam ser feitas. Em um modelo de escolha binária, é convencional atribuir "1" à variável se ela possuir uma qualidade específica ou se existir uma condição e "0" caso contrário. Assim, a variável dependente é:

$$y_i = \begin{cases} 1 & \text{se o indivíduo } i \text{ possui a característica} \\ 0 & \text{caso contrário} \end{cases}$$

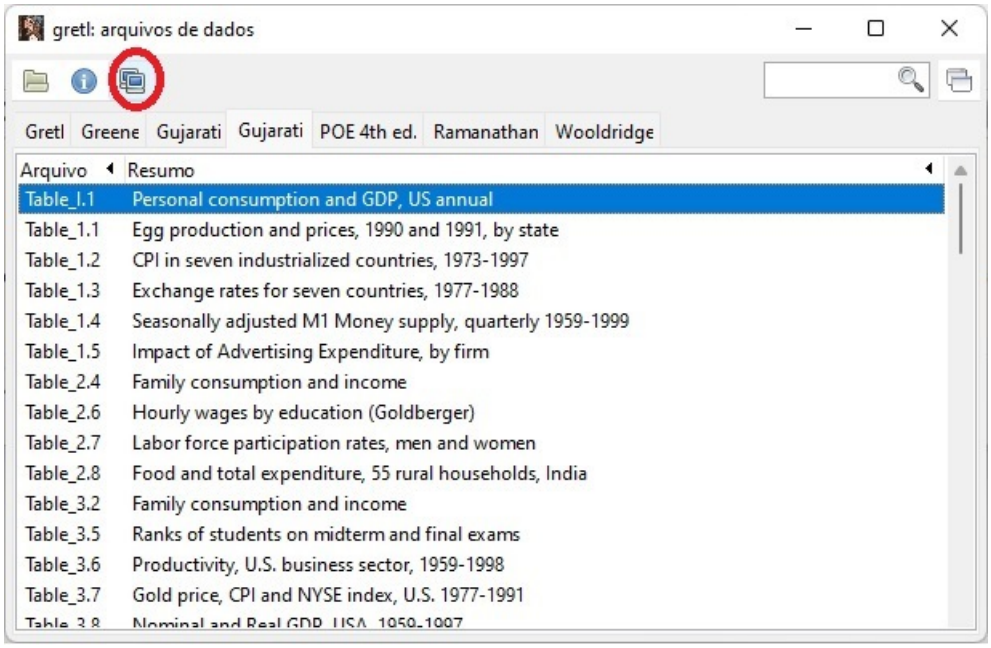
O modelo de probabilidade linear, modela a probabilidade de que $y_i = 1$ como uma função linear das variáveis independentes. Neste exemplo, é tomada uma decisão binária sobre dirigir de automóvel ou usar o transporte público.

$$auto_i = \begin{cases} 1 & \text{se o indivíduo } i \text{ escolhe o carro} \\ 0 & \text{se o transporte público é escolhido} \end{cases}$$

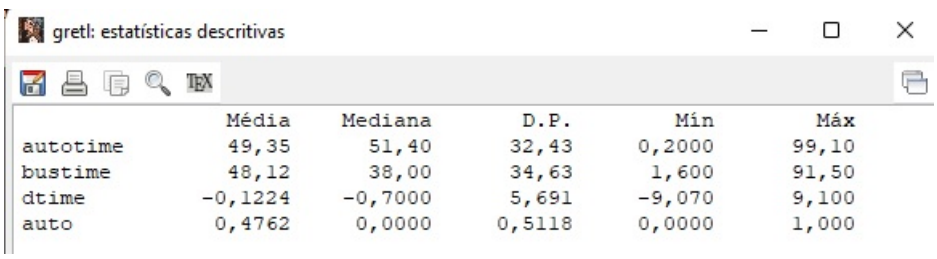
Isso é estimado em função do diferencial de tempo de deslocamento entre as duas alternativas. Isso é $dtime = {}^{(bustime - autotime)}/_{10}$. Em um modelo de probabilidade linear, isso se torna:

$$auto_i = \beta_1 + \beta_2\, dtime_i + e_i$$

Utiliza-se os dados da base `transport.gdt`. Esta base de dados pode ser baixada diretamente do servidor. Clique no menu **Arquivo>Arquivo de exemplos** e observe que há um pequeno computador (ver no servidor), selecione a opção `POE 4th`:



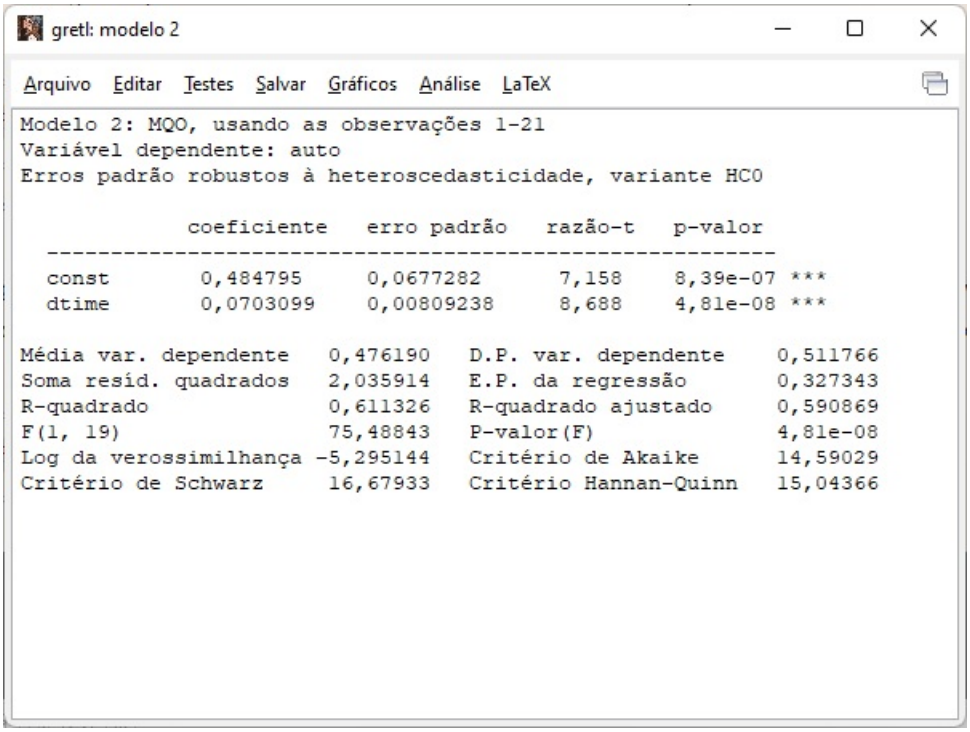
Ademais, também pode-se fazer o download de várias outras bases como as dos manuais de econometria de Wooldridge e Gujarati. Primeiramente obtém-se as estatísticas descritivas (**Ver>Estatísticas Descritivas**) dos dados:



A média da variável auto representa a proporção de indivíduos que escolhem o transporte por automóvel. Note que esse valor é o que corresponde ao número 1 da

variável de escolha binária. Em outras palavras, $47,62\%$ dos indivíduos da amostra preferem esse tipo de transporte.

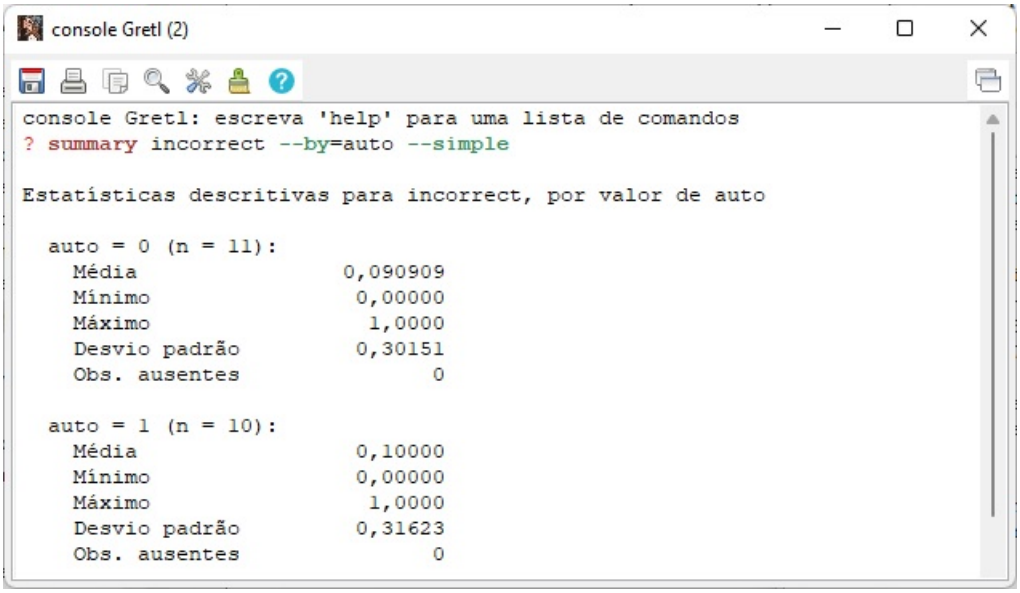
O modelo é estimado por mínimos quadrados usando erros padrões robustos, pois uma variável dependente binária é heterocedástica. Posteriormente calcula-se uma nova série que assume o valor e a probabilidade prevista estiver acima de 50%. Também calcula-se a previsão incorreta, quando o modelo prevê que o indivíduo escolherá o automóvel, mas ele de fato pega o ônibus. A média desta série mede a frequência relativa de previsões incorretas.



O coeficiente em `dtime` é positivo (significativamente a 5%), o que indica que quanto maior o diferencial de tempo, maior a probabilidade de uma pessoa fazer uma viagem de automóvel. Após estimar o modelo clica-se no menu **Salvar>valores ajustados**. Então salva-se os valores previstos como y_hat. Em seguida cria-se as seguintes variáveis usando o Menu da janela principal **Acrescentar>Definir nova variável**:

- `series y_pred = y_hat>0.5`
- `series incorrect = abs(auto-y_pred)`

Em seguida clique no menu **Ferramentas>Console** do **gretl** e digite o seguinte comando: **summary incorrect --by = auto --simple**. Esse comando mostra as principais estatísticas descritivas separadas por "auto":



A partir deles pode-se determinar que apenas 1 de 11 passageiros de ônibus $(1/11 = 0,091)$ e 1 de 10 passageiros de automóveis $(1/10 = 0,10)$ foram previstos incorretamente. O número total de previsões corretas é igual a $19/21 = 90\%$. Esse número pode ser calculado, definindo uma nova variável, da seguinte forma: **scalar correct = $nobs - sum (abs (auto - y_pred).** Lembre que 21 é o número de observações da nossa amostra.

11.2 Probit

O modelo estatístico Probit expressa a probabilidade p tal que $y_i = 1$ como uma função das variáveis independentes:

$$P\left[(y_i \mid x_{i2},\, x_{i3}) = 1\right] = \Phi(\beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3})$$

sendo Φ a função de distribuição cumulativa normal (cdf). O argumento dentro de Φ é linear nos parâmetros e chamado de função de índice. Φ mapeia os valores da função de índice no intervalo fechado $[0, 1]$. Estima-se este modelo usando uma função de máxima verossimilhança já disponível no **gretl**. Utiliza-se a mesma base de dados a qual foi usada para estimar o MPL (Modelo de Probabilidade Linear). A vantagem dos Probit e do Logit em relação a esse modelo é que todos os valores previstos estarão dentro do intervalo probabilístico entre zero e um. A seguir será estimada a seguinte equação:

$$P\left[auto_i = 1\right] = \Phi\left(\beta_1 + \beta_2\, dtime_i\right)$$

Para isso seleciona-se o seguinte menu **Modelo>Variável dependente limitada>Probit>Binário**. Escolha a seguinte configuração (Veja Figura 11.1):

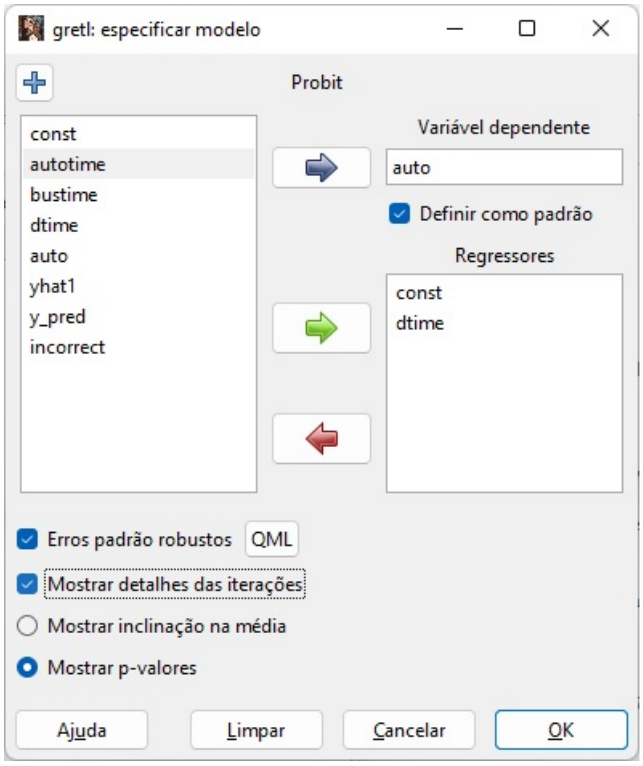


Figura 11.1: Especificar modelo.

A saída do modelo será a seguinte – Figura 11.2:

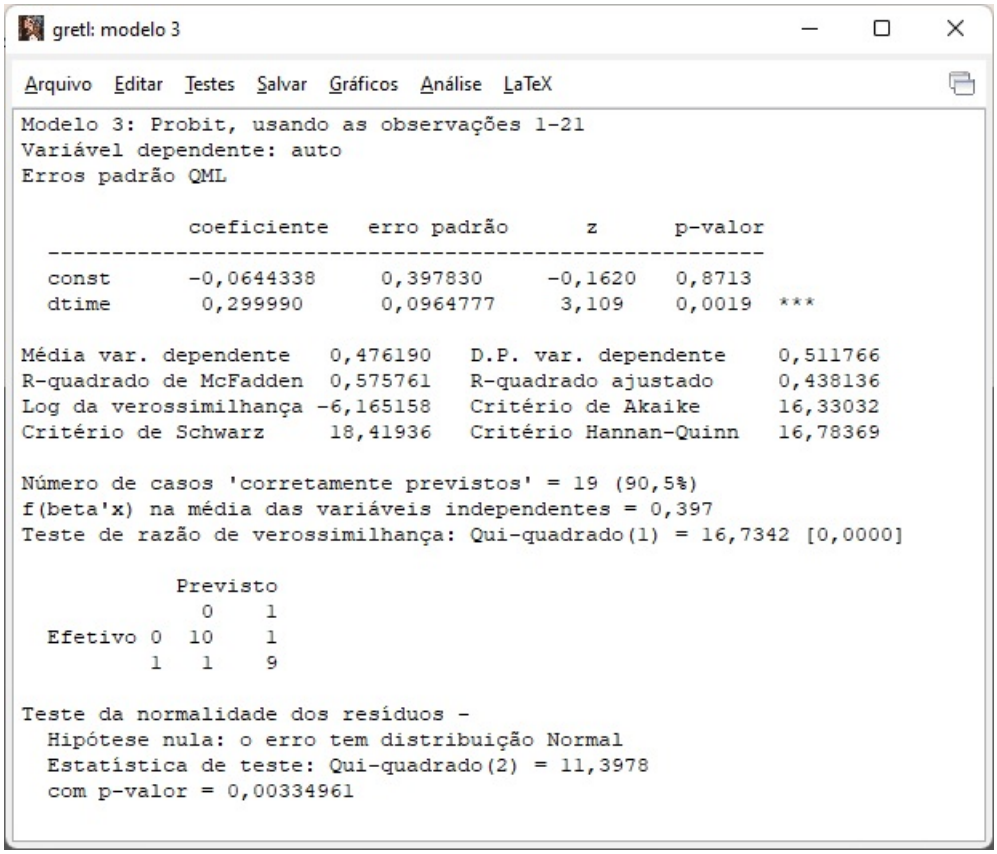


Figura 11.2: Saída do modelo Probit.

O diferencial de tempo aumenta em média as chances do indivíduo escolher o uso do automóvel. Agora será visto como interpretar mudanças pontuais e na média da variável independente e seus efeitos na variável dependente.

11.2.1 Efeitos marginais e efeitos marginais médios

O efeito marginal de uma mudança em x_{ij} na probabilidade de escolha P_i é:

$$\frac{\partial P_i}{\partial x_{ij}} = \Phi\left(\beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3}\right)\beta_j$$

em que $\Phi\left(\cdot\right)$ é a densidade da distribuição de probabilidade normal. Isso significa que os efeitos marginais dependem de todos os parâmetros do modelo bem como os valores das variáveis. Dado que a viagem por transporte público atualmente leva 20 ($dtime = 2$) minutos a mais do que o automóvel, o efeito marginal estimado foi:

$$\frac{\partial\, P_i}{\partial\, dtime_i} = \Phi\left(\hat{\beta}_1 + \hat{\beta}_2\, dtime_i\right)\hat{\beta}_2 = \Phi\left(-0.0644 + 0.3 \times 2\right) \times 0.3 = 0.1037$$

Os efeitos marginais para variáveis indicadoras necessitam de uma abordagem diferente. Para um regressor indicador, a probabilidade é calculada para cada um de seus estados (0 e 1), mantendo os valores das outras variáveis constantes nos valores

selecionados. As demais variáveis podem ser avaliadas em suas médias amostrais ou em pontos representativos.

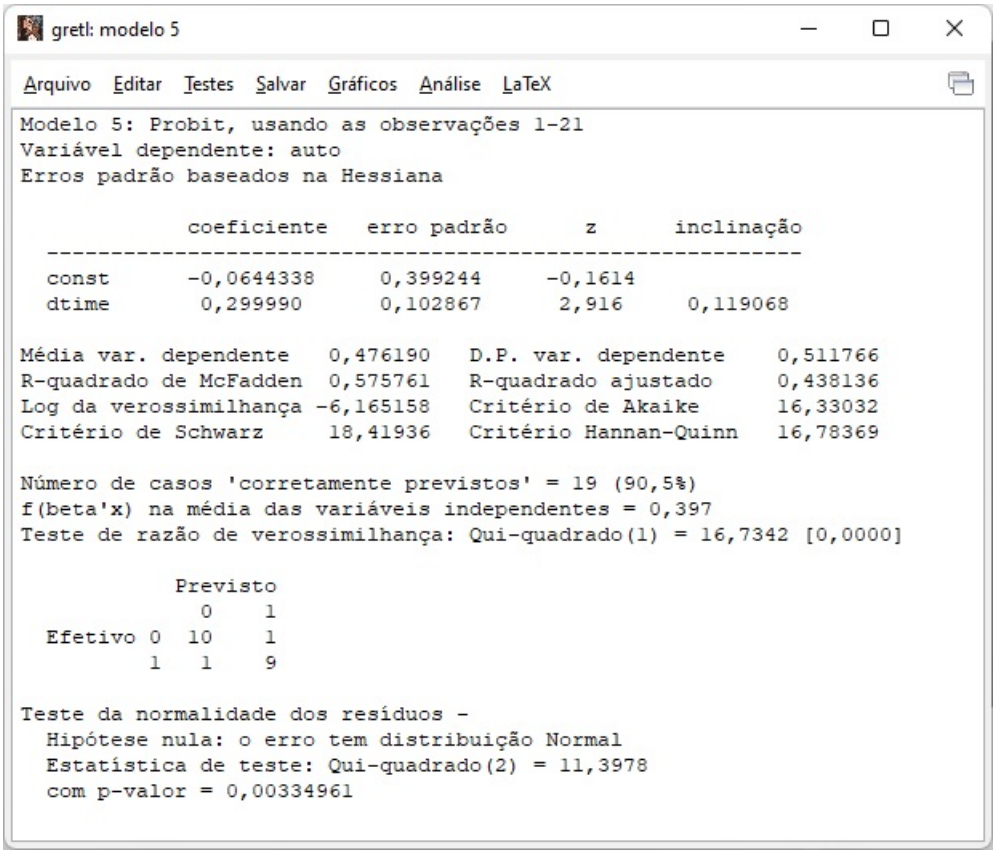
Uma abordagem bastante popular é calcular os efeitos marginais médios. O efeito marginal de uma mudança de x_{ij} em P_i é:

$$\widehat{AME_j} = \frac{1}{N} \sum_{i=1}^{N} \Phi(\hat{\beta}_1 + \hat{\beta}_2\, x_{i2} + \cdots + \hat{\beta}_N\, x_{iN})\, \hat{\beta}_j$$

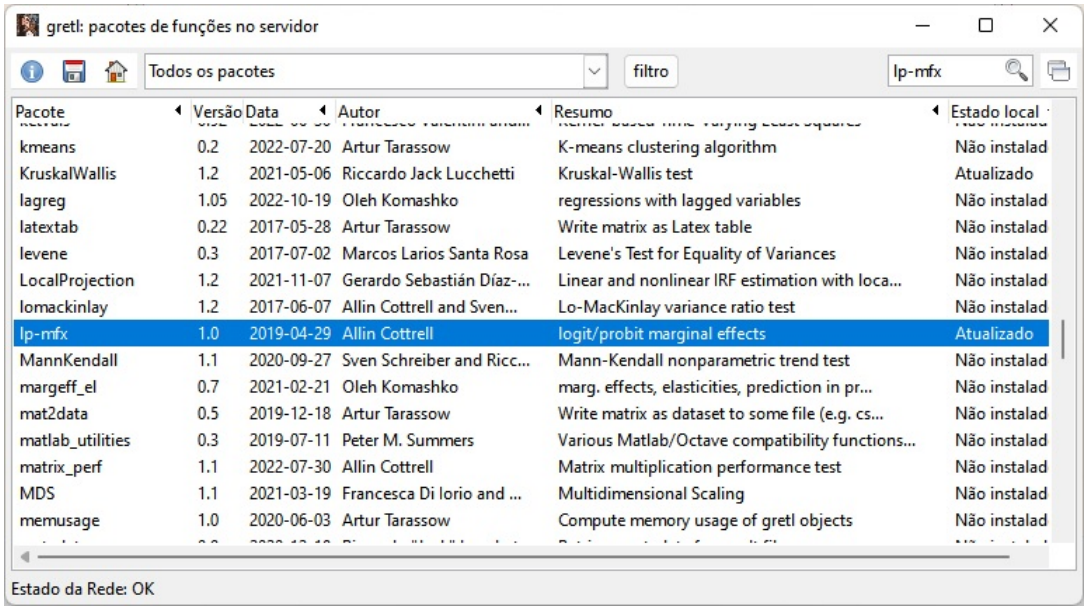
Também é comum avaliar os efeitos na média de cada variável independente. Isso é efeito do seguinte modo:

$$\widehat{ME_j} = \frac{1}{N} \sum_{i=1}^{N} \Phi\left(\hat{\beta}_1 + \hat{\beta}_2\, \bar{x}_2 + \cdots + \hat{\beta}_N\, \bar{x}_N\right) \hat{\beta}_j$$

Os efeitos de $\widehat{ME_j}$ são calculados e rotulados no **gretl** como inclinação. A maior desvantagem em usá-los é que os valores médios das variáveis podem não ser representativos. Isso ocorre com muita frequência se uma ou mais das variáveis independentes for um indicadora ou *dummy*. Por esse motivo, é indicado uso do AME, a menos que haja casos específicos a serem considerados. Pode-se ter uma boa ideia dos efeitos marginais (médios) observando as inclinações estimadas de um modelo de probabilidade linear. Para ver os efeitos marginais médios, deve-se selecionar a opção "mostrar a inclinação na média" quando for estimar o modelo Probit:



O efeito de uma mudança na média da diferença de tempo afeta em 0,11 a probabilidade do individuo optar pelo uso do automóvel. Para computar os efeitos marginais individuais, médios ou na média de todas as variáveis dependentes é possível usar a função (pacote) `lp-mfx`. Para instalar este pacote, clique no menu **Arquivo>Pacotes de Funções>No Servidor**.



Clique no disquete para instalar. Depois estime novamente o modelo probit. Na tela de estimação do modelo, clique no menu `Análise>Marginal effects` – Figura 11.3

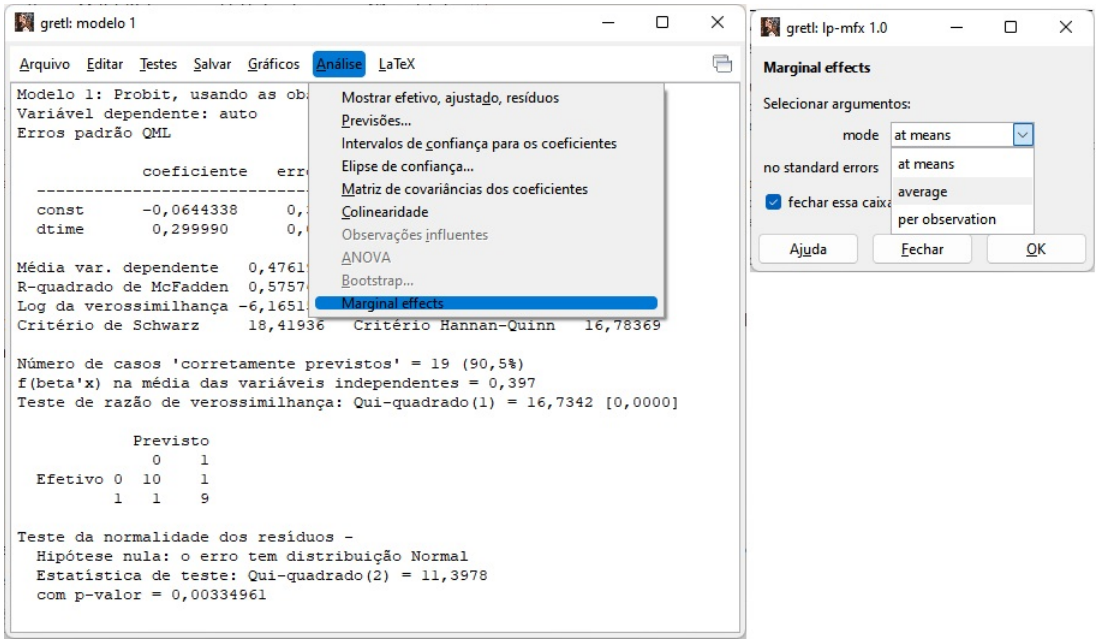


Figura 11.3: Selecionando a opão Marginal effects.

Serão obtido as seguintes saídas – Figura 11.4:

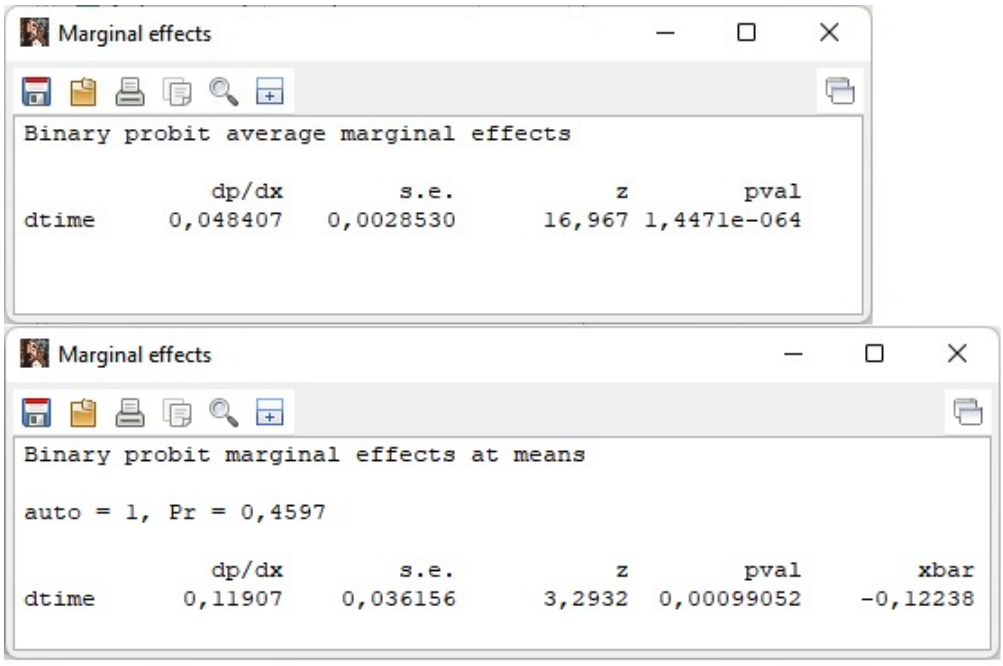


Figura 11.4: Marginal effects.

Note que quando se seleciona mostrar "inclinação" o **gretl** calcula o efeito marginal "at means", isto é, o $\widehat{AME}_j$ para a variação de uma unidade da variável avaliada. A função `lp-mfx` também pode ser utilizada para calcular os efeitos marginais do modelo

logit que será apresentado na próxima subseção.

11.3 Logit

O modelo logit é muito similar ao probit. No entanto, a probabilidade de um evento a ser descrito pelo evento por uma distribuição normal é modelada utilizando uma distribuição logística. As distribuições normal e logística possuem uma forma (curvatura) bastante similares, portanto a estimação desses modelos são muito próximas. A probabilidade que o indivíduo i escolha a alternativa é:

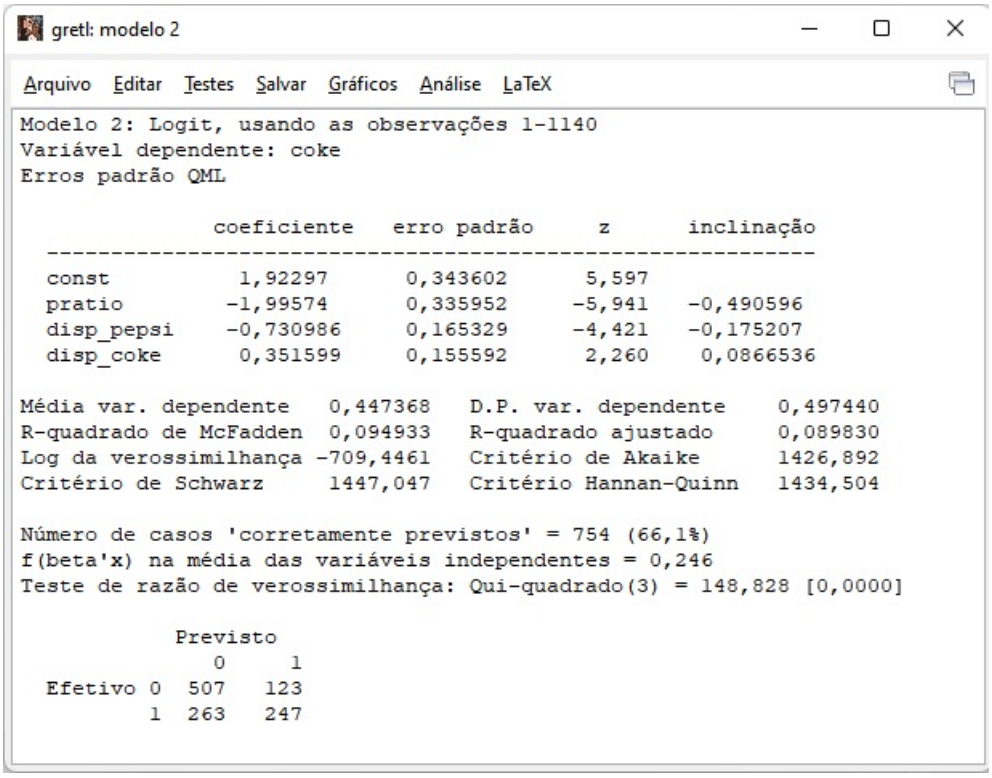
$$P_i = (F\, z_i) = \Lambda\,(z_i) = \frac{1}{1 + e^{-z_i}}$$

$$z_i = \sum_{j=1}^{k} x_{ij}\, \beta_j$$

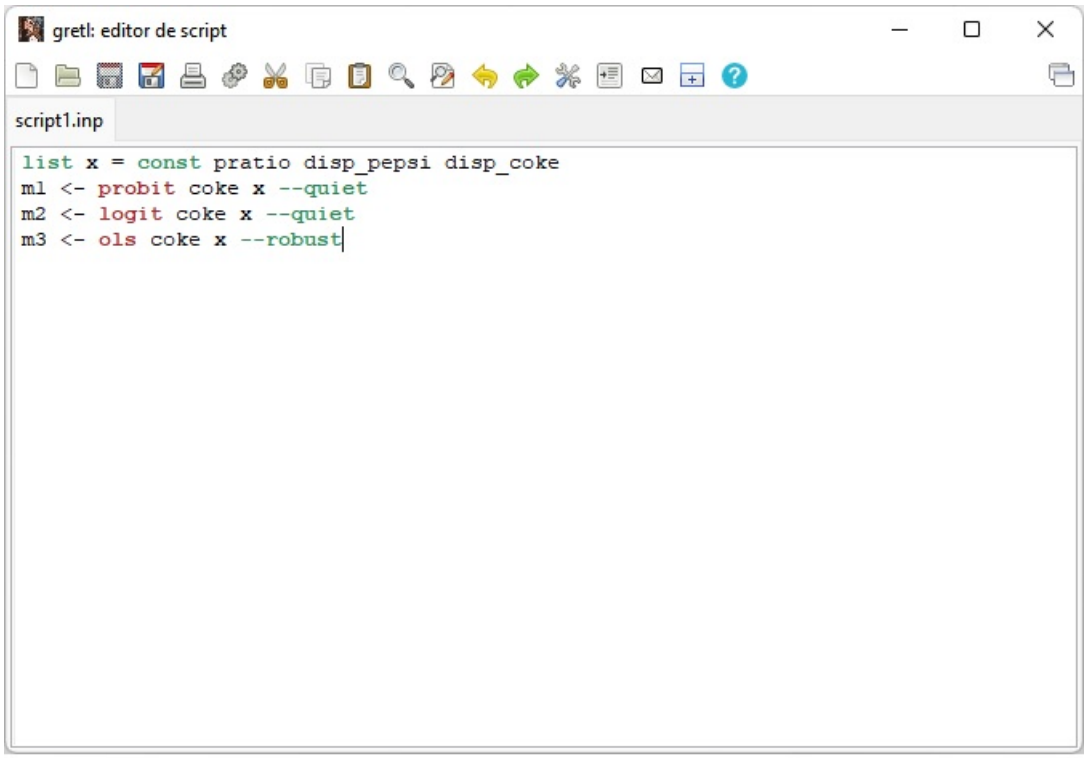
No logit, a probabilidade é modelada utilizando $\Lambda\,(z_i$ ao invés de $\Phi\,(z_i)$ como no modelo probit. Para exemplificar o uso do modelo logit, será utilizado a escolha pelo consumo de refrigerante, sendo a variável dependente igual a um se o consumidor comprar Coca-Cola e zero caso contrário. Modela-se essa relação como uma função da razão entre o preço da Coca-Cola (Coke, em inglês) e o preço da Pepsi. O modelo é:

$$Pr\,(Coke_i = 1) = \phi\,(\beta_1 + \beta_2\, pratio + \beta_3\, disp\_coke + \beta_4\, disp\_pepsi)$$

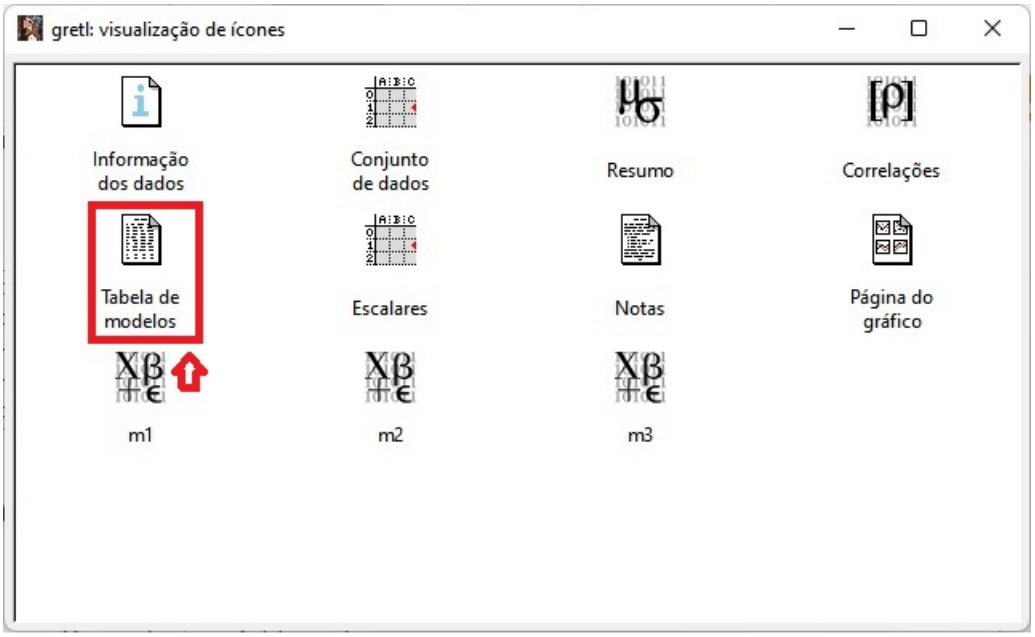
Para isso usa-se o arquivo `coke.gdt`. Para estimar esse modelo, clique no menu **Modelo>Variável dependente limitada>Logit>Binário**:



A tabela dos valores previstos revela que com logit, dos $(507 + 123) = 630$ consumidores que escolheram Pepsi (Pepsi = 0). O modelo previu 507 deles corretamente ($80,48\%$ correto para Pepsi). Para Coca-Cola o modelo previu $247/(263+247) = 247/510 = 48,43\%$. A porcentagem total que foi prevista corretamente é $754/1140 = 66,1\%$. Para compar as estimativas do logit, com as do probit e do mpl utiliza-se o `script`. Para isso clique no menu **Arquivo>Arquivos de script>Novo script>Script Gretl:**



Para executar o `script` clique nas engrenagens, que estão ao lado da impressora e da tesoura. Cada modelo ficará disponível na tela de ícones. Deve-se arrastar o ícone de cada modelo para a Tabela de modelos e posteriormente clicar duas vezes nesse ícone.



O resultado será o seguinte:

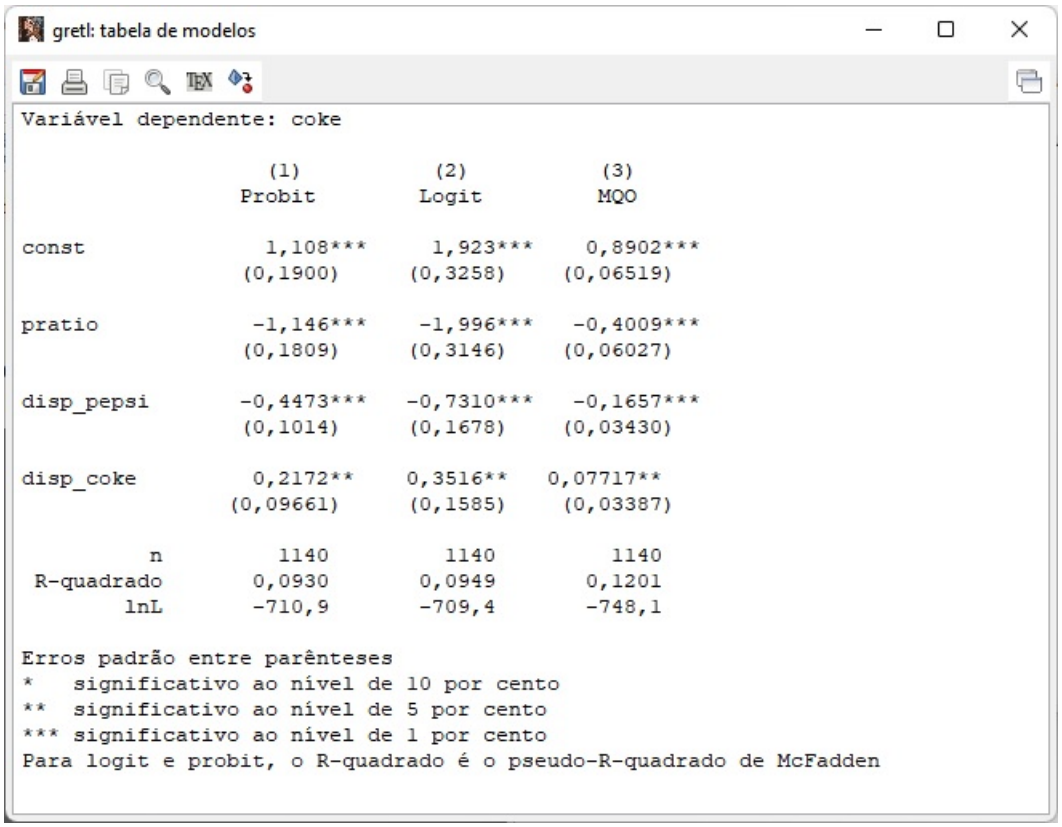


Figura 11.5: Tabela de modelos.

Os sinais e as `razões t` são aproximadamente iguais entre os estimadores. Nos

modelos logit e probit, os coeficientes e os sinais são consistentes com a direção dos efeitos marginais. As magnitudes dos coeficientes diferem apenas por causa das diferenças implícitas em como os coeficientes são normalizados. Embora, não seja óbvio, há uma relação aproximada entre os coeficientes de "inclinação" dos três conjuntos de estimativas.

$$\widetilde{\gamma}_{logit} \cong 4\,\hat{\beta}_{MPL}$$
$$\widetilde{\beta}_{probit} \cong 2.5\,\hat{\beta}_{MPL}$$
$$\widetilde{\gamma}_{logit} \cong 1.6\,\hat{\beta}_{probit}$$

Portanto, $4(-0,4009) = -1,6036$ é bastante próximo da estimativa de $-1,996$ para o coeficiente `pratio` na coluna logit. Mais importante ainda, existem semelhanças mais próximas entre os efeitos marginais implícitos por logit e probit. Suas médias (AME) são muito próximas do coeficiente correspondente no modelo de probabilidade linear. Pode-se esperar que eles se tornem mais próximos à medida que o tamanho da amostra aumenta. O primeiro conjunto de estatísticas computadas é o AME de cada um dos modelos. Isso é fácil para o MPL, pois os efeitos marginais são os mesmos, independentemente do valor de x. Para probit e logit requer o uso do método delta para obter estimadores consistentes do erros padrão.

11.3.1 Teste de Razão de Verossimilhança

Como os modelos probit e logit são estimados pelo método da verossimilhança máxima, também pode realizar um teste de razão de verossimilhança. A razão de verossimilhança é:

$$LR = 2\,(lnL_U - lnL_R) \sim \chi^2\,(J)$$

Se a hipótese nula for verdadeiro. O parâmetro J são os graus de liberdade para o $\chi^2\,(J)$ e é igual ao número de hipóteses que se está testando em conjunto, neste caso são duas. Os parâmetros L_U e L_R são as log verossimilhanças dos modelos irrestrito (U) e restrito (R), respectivamente. O procedimento é estimar modelos restritos e irrestritos, calcular a log-verossimilhança de cada um, compor a estatística LR e calcular seu `p-valor`.

Para isso volta-se ao exemplo anterior e estima-se o seguinte modelo:

$$P_{coke-U} = \phi\,(\beta_1 + \beta_2\,pratio + \beta_3\,disp\_coke + \beta_4\,disp\_pepsi)$$

Chamando esse modelo de irrestrito (U). Para o modelo restrito considera que $\beta_3 = 0$.

$$P_{coke-R} = \phi\,(\beta_1 + \beta_2\,pratio + \beta_4\,disp\_pepsi)$$

Desta forma, estima-se um modelo Probit irrestrito e clica-se no menu **Salvar>Log da verossimilhança**. Defina o nome da variável como lr_u. Depois clique no menu **Modificar modelo** e estime um novo modelo sem a variável $disp\_coke$. Novamente clique em **Salvar>Log da verossimilhança**. Defina essa variável como lr_r. Em seguida vá no menu **Acrescentar>Definir nova variável**. A fórmula é a mesma mostrada na equação $ll = scalar2*\,(lr\_u - lr\_r)$. No menu **Ferramentas**, selecione o console do **gretl** e digite os seguintes comandos:

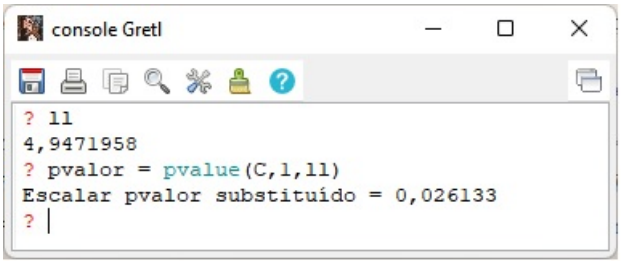
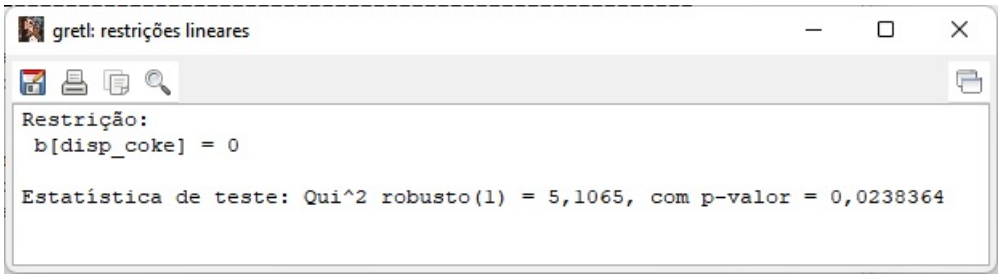


Figura 11.6: Console do **gretl**.

Este é quase o mesmo resultado obtido usando o `teste de Wald`. Para estimadores não lineares, essas estatísticas normalmente produzirão resultados (ligeiramente) diferentes. Pode-se rejeitar a H_0 que $\beta_3 = 0$ a um nível de 5%. Alternativamente, pode-se fazer um teste de restrições lineares! Estima-se o modelo completo e clica-se no menu **Testes>Restrições Lineares**. Deve-se ainda inserir a opção $b3 = 0$ e apertar `ok`.



Note que os resultados são muito próximos!!!

11.4 Regressores endógenos

Com um regressor contínuo e endógeno, há pelo menos duas abordagens que podem ser adotadas para estimar os parâmetros do modelo de forma consistente. A primeira é usar mínimos quadrados lineares de dois estágios. Esta é a contraparte do regressor endógeno para o modelo de probabilidade linear. A outra abordagem é usar uma variável instrumental probit (ou logit). Este NÃO é um estimador de dois estágios no mesmo sentido que o 2SLS linear. Requer alguns cuidados na prática.

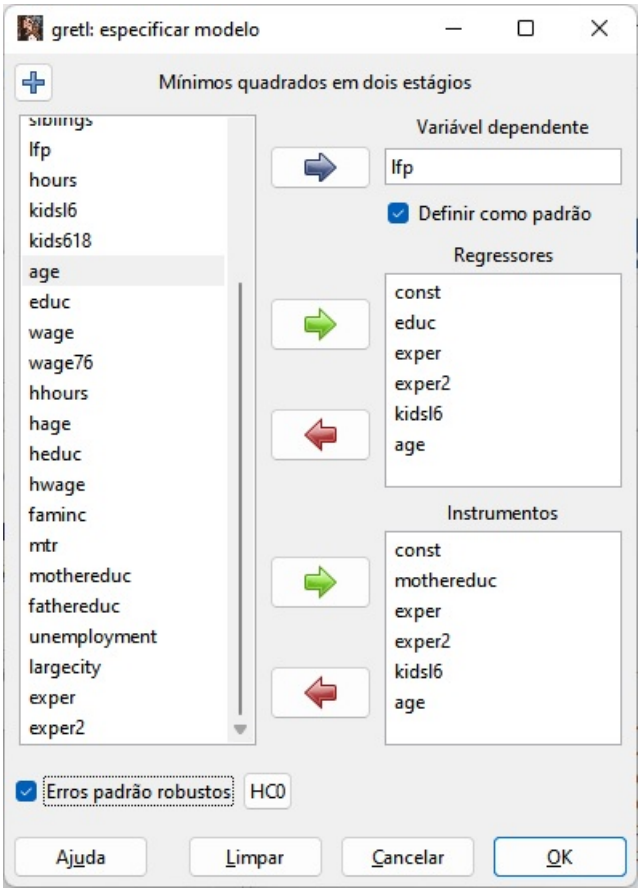
A seguir serão utilizados os dados contidos no arquivo `mroz.gdt` para estimar um modelo de participação feminina na força de trabalho (LFP). A variável LFP é binária, assumindo o valor 1 se uma mulher estiver na força de trabalho e 0 caso contrário. O modelo de probabilidade linear estimado é:

$$LFP = \beta_1 + \alpha_1\, educ + \beta_2\, exper + \beta_3\, exper^2 + \beta_4\, kidls6 + \beta_5\, age + e$$

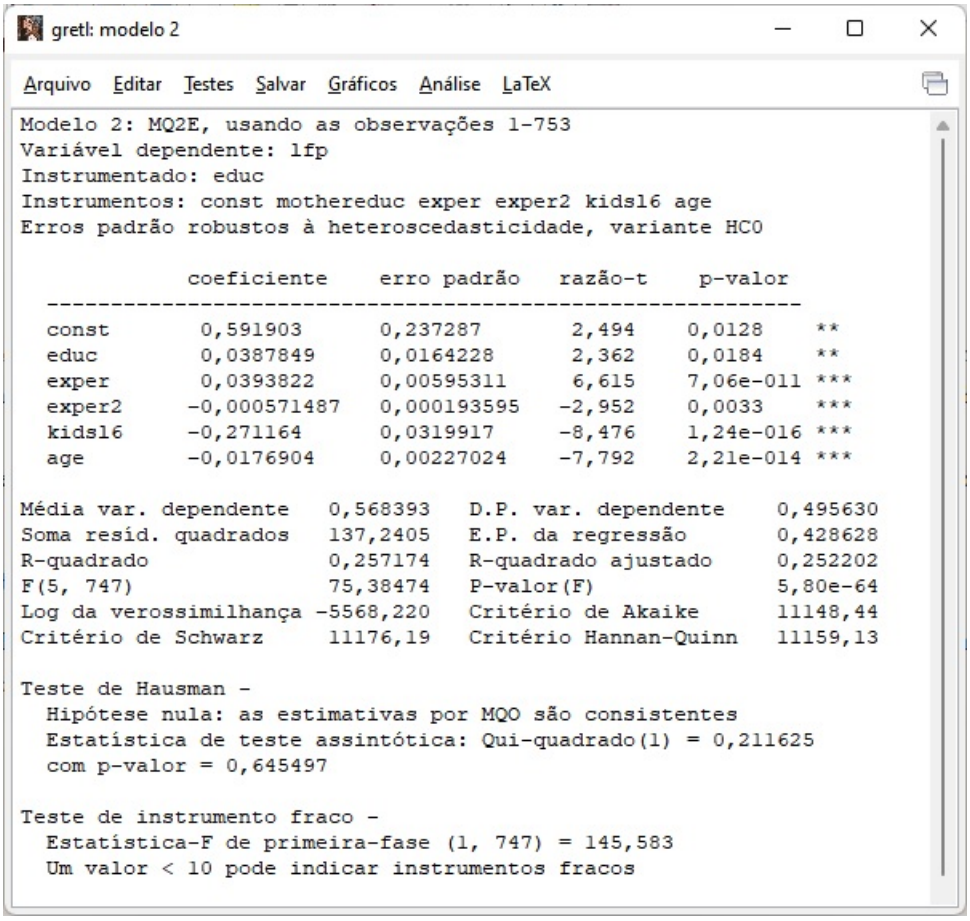
A escolaridade da mulher, *educ*, é considerada endógena. Para o modelo de Mínimos Quadrados em Dois Estágios (MQO2E), precisa-se de um instrumento. Neste caso, será utilizado a educação da mãe (*mothereduc*) como instrumento para *educ*. Para

isso clique no menu **Modelos>Variáveis Instrumentais>Mínimos Quadrados em Dois Estágios.**

Isso é proporcionado pela educação da mãe, *mothereduc*.

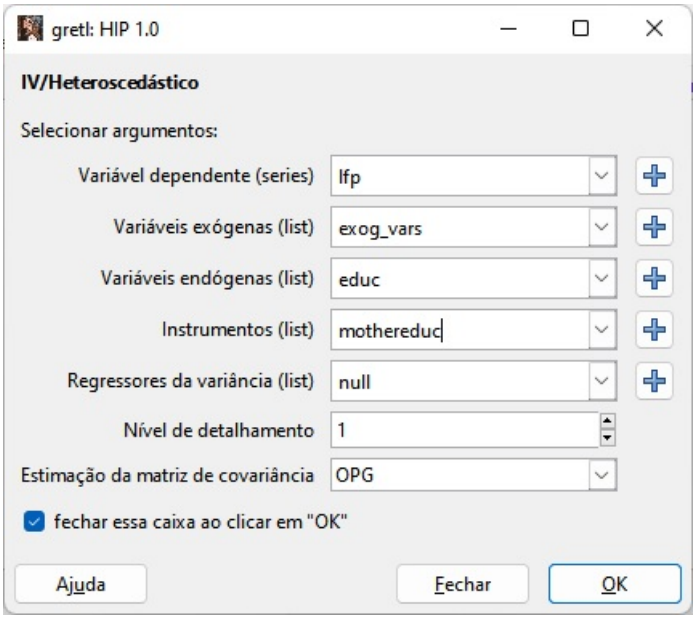


As estimativas do modelo MQO2E:



Embora o instrumento pareça forte ($F = 144,4$), o `teste de Hausman` para a exogeneidade da educação não é rejeitado a 5%. Uma outra possibilidade é estimar uma versão do modelo probit com variáveis instrumentais. Isso pode ser feito usando um pacote chamado `HIP`. O pacote `HIP` foi escrito por Riccardo Lucchetti e Claudia Pigini e apresenta uma coleção de `scripts` para estimar modelos probit heterocedásticos, que podem incluir regressores endógenos.

Primeiramente cria-se uma lista de variáveis exógenas e instrumentos. Para tanto, clique no menu **Dados>Criar ou editar lista**. Crie uma lista chamada *exog_vars* com as variáveis *const*, *exper*, *exper*2, *kidsl*6 e *age*. Em seguida deve-se clicar no menu **Modelo>Variável Limitada Dependente>Probit>IV/Heterocedástico**.



As estimativas do modelo podem ser vistas – Figura 11.7:

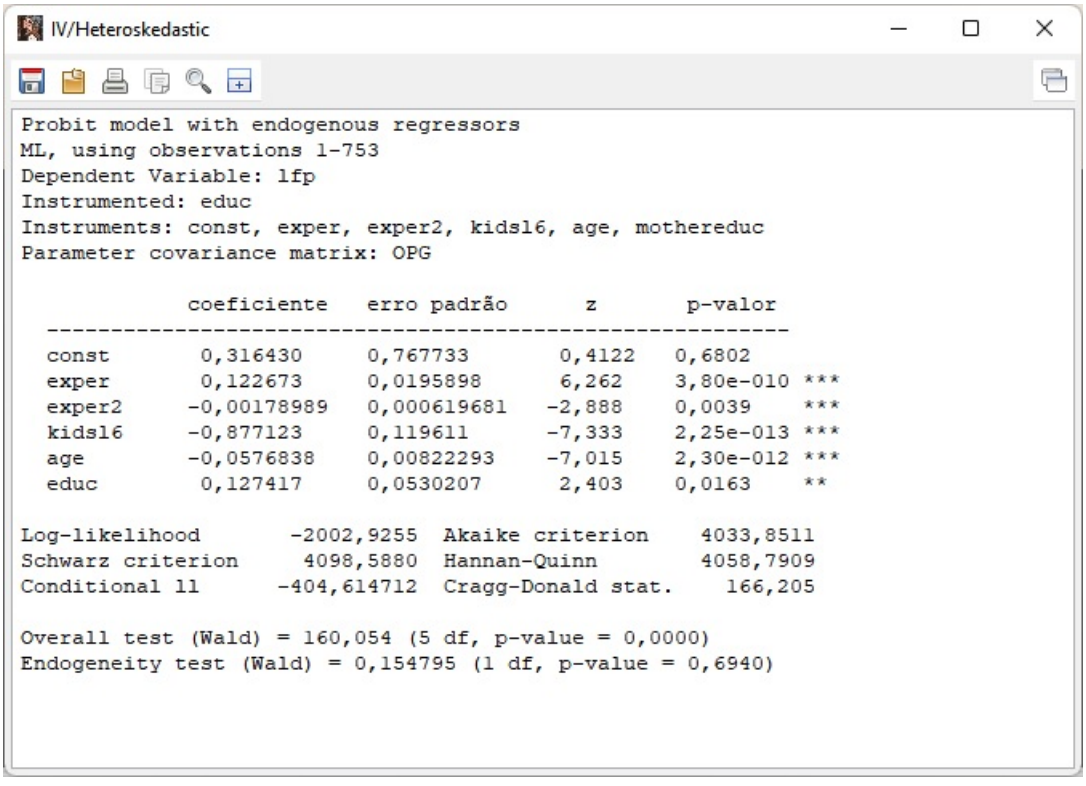


Figura 11.7: IV/Heteroskedastic

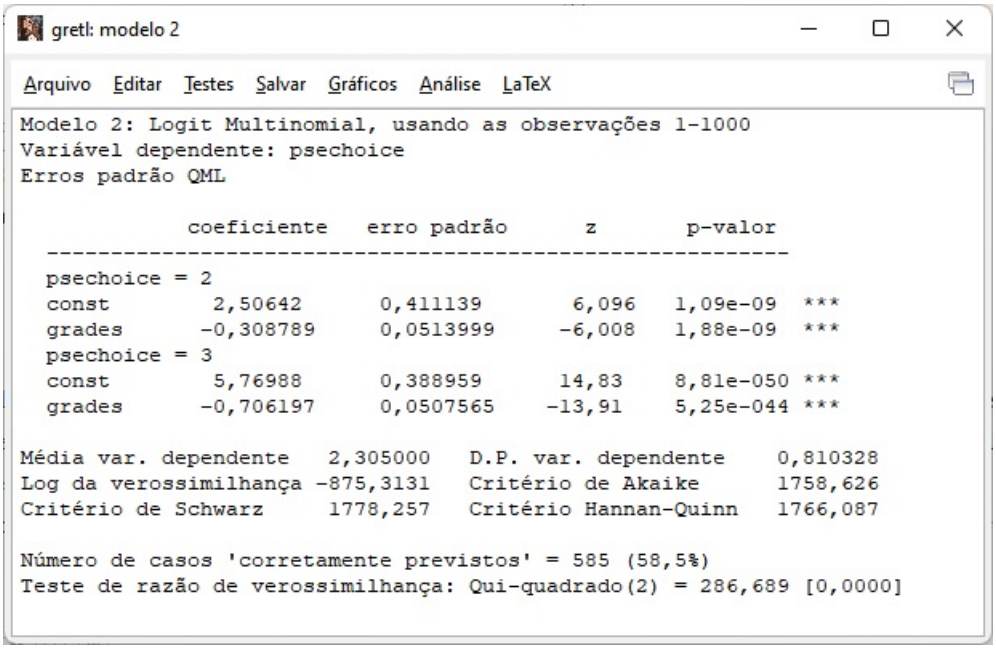
Os resultados do teste são bastante semelhantes aos do MPL/IV. A educação não é considerada endógena em 5%. A `razão t` em educação foi de $2,35$ na versão LPM e

é de 2, 4 na versão IV/probit. É claro que calcular os efeitos marginais no IV/probit é complicado pela não linearidade do modelo.

11.5 Logit Multinomial

No modelo Logit Multinomial, a variável dependente é categórica e codificada da seguinte maneira. Um estudante concluindo o ensino médio escolhe entre três alternativas: não frequentar a faculdade *psechoice* = 1, matricular-se em uma faculdade de 2 anos *psechoice* = 2 ou matricular-se em uma faculdade de 4 anos *psechoice* = 3. A variável explicativa são as notas, que é um índice que varia de 1, 0 (nível mais alto, nota A+) a 3, 0 (nível mais baixa, nota F) e representa o desempenho combinado em inglês, matemática e estudos sociais. Para este exemplo, as opções são tratadas como não ordenadas, há 1.000 observações.

Para estimar o modelo de escolha da escola em função das notas e uma constante, abra o conjunto de dados `nels_small.gdt` e clique no menu **Modelo>Variável Dependente Limitada>Logit>Multinomial**.

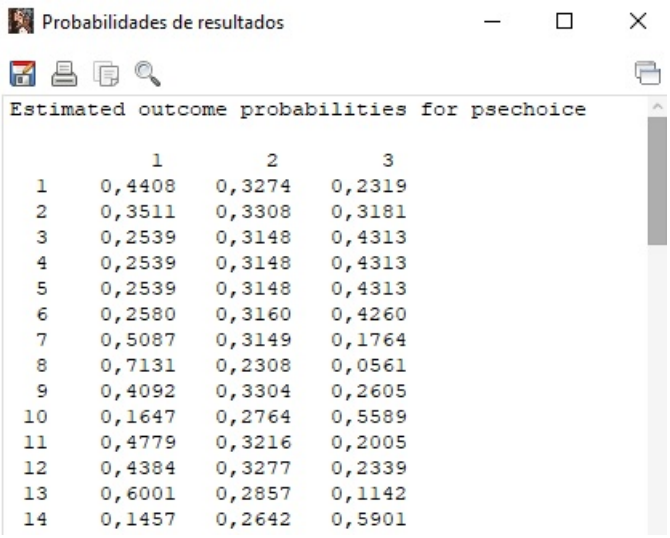


Os coeficientes aparecem agrupados. O primeiro grupo contém os coeficientes que estão associados a escolha de *psechoice* = 2 e o segundo grupo associa-se com *psechoice* = 3. Isso implica que o **gretl** escolheu *psechoice* = 1 como grupo de referência. A probabilidade de escolher uma alternativa em um modelo logit multinomial é:

$$p_{ij} = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\beta_{1j} + \beta_{2j}\, x_{i2} + \cdots + \beta_{kj}\, x_{ik}\right)} \quad \text{para} \quad j = 1$$

$$p_{ij} = \frac{\exp\left(\beta_{1j} + \beta_{2j}\, x_{i2} + \cdots + \beta_{kj}\, x_{ik}\right)}{1 + \sum_{j=2}^{J} \exp\left(\beta_{1j} + \beta_{2j}\, x_{i2} + \cdots + \beta_{kj}\, x_{ik}\right)} \quad \text{para} \quad j \neq 1$$

A obtenção das probabilidades estimadas pelo modelo é bastante simples. Estime o modelo **Modelo>Variável dependente limitada>Logit>Multinomial**. Na janela do modelo, selecione **Análise>Probabilidades de resultado** para produzir as probabilidades previstas para cada caso na amostra:



11.6 Probit Ordenado

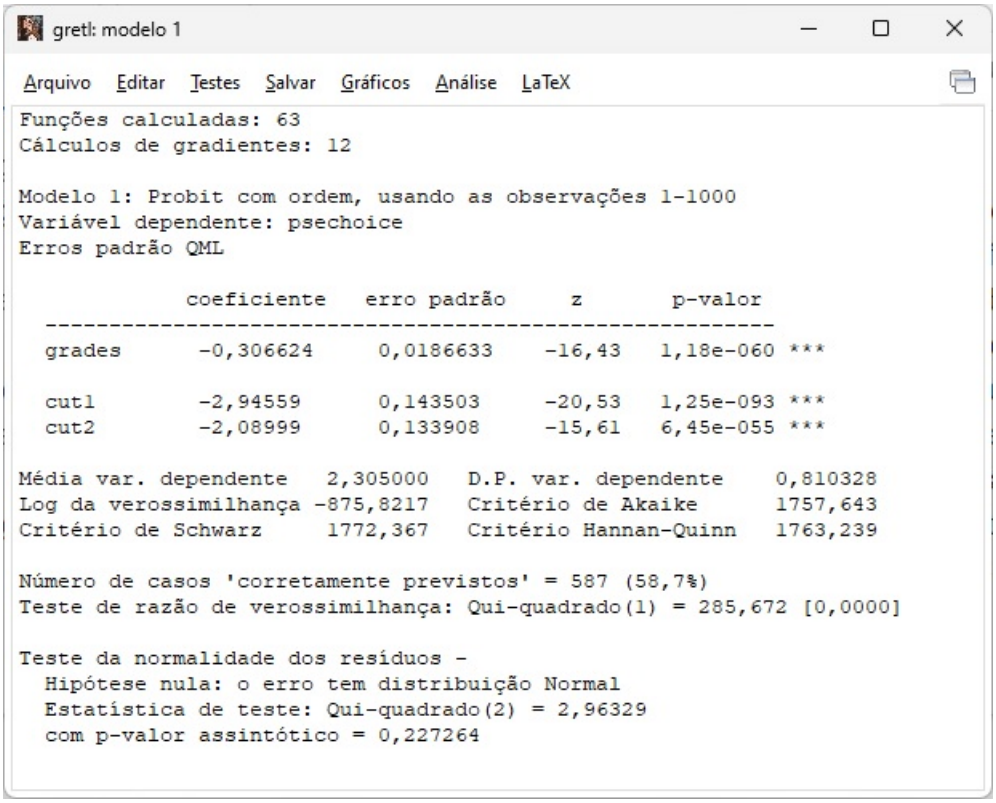
A seguir apresenta-se um exemplo em que as probabilidades de não frequentar a faculdade, de frequentar por 2 anos e por 4 anos, são modeladas como uma função das notas do aluno. Em princípio, espera-se que os estudantes com notas mais altas no ensino médio, possuem maior chance de frequentar uma faculdade por 4 anos e menos chances de pular o ensino superior. No conjunto de dados, as notas são medidas em uma escala de 1 a 13, sendo 1 a mais alta. Isso significa que se notas mais altas aumentam a probabilidade de ir para uma faculdade de 4 anos, o coeficiente nas notas será negativo. As probabilidades são modeladas usando a distribuição normal neste modelo onde os resultados representam níveis crescentes de dificuldade. O modelo é:

$$y_i^* = \beta\, grades_i + e_i$$

A variável y_i^* é uma variável latente, ou seja, o seu valor é não observado. Na verdade, observa-se as escolhas categóricas de entrada na faculdade:

$$y_i = \begin{cases} 3 & \text{Faculdade por 4 anos} \\ 2 & \text{Faculdade por 2 anos} \\ 1 & \text{não frequentou} \end{cases}$$

Os dados utilizados serão os de `nels_small.gdt`. Essa plataforma consiste em conjunto de 1.000 observações coletadas como parte do Estudo Longitudinal de Educação Nacional de 1988. As notas variáveis medem a nota média em matemática, inglês e estudos sociais na escala de 13 pontos, sendo 1 a mais alta. Para estimar o modelo vá no menu **Modelo>Variável dependente limitada>Probit>Ordenado**. Escolha uma variável dependente e um conjunto de regressores:



O coeficiente nas notas é negativo e significativo a 5%. Isso significa que, à medida que a variável de notas aumenta (as notas pioram), o índice fica menor e nas margens 2 anos os participantes da faculdade estão sendo empurrados para nenhuma faculdade e os participantes da faculdade de 4 anos estão sendo empurrados para a opção de 2 anos. Sabe-se que a probabilidade de estar na categoria mais baixa aumenta e de estar na categoria mais alta diminui. O que quer que aconteça no meio depende dos efeitos líquidos das pessoas sendo empurradas para fora da categoria 3 e puxadas para a categoria 1.

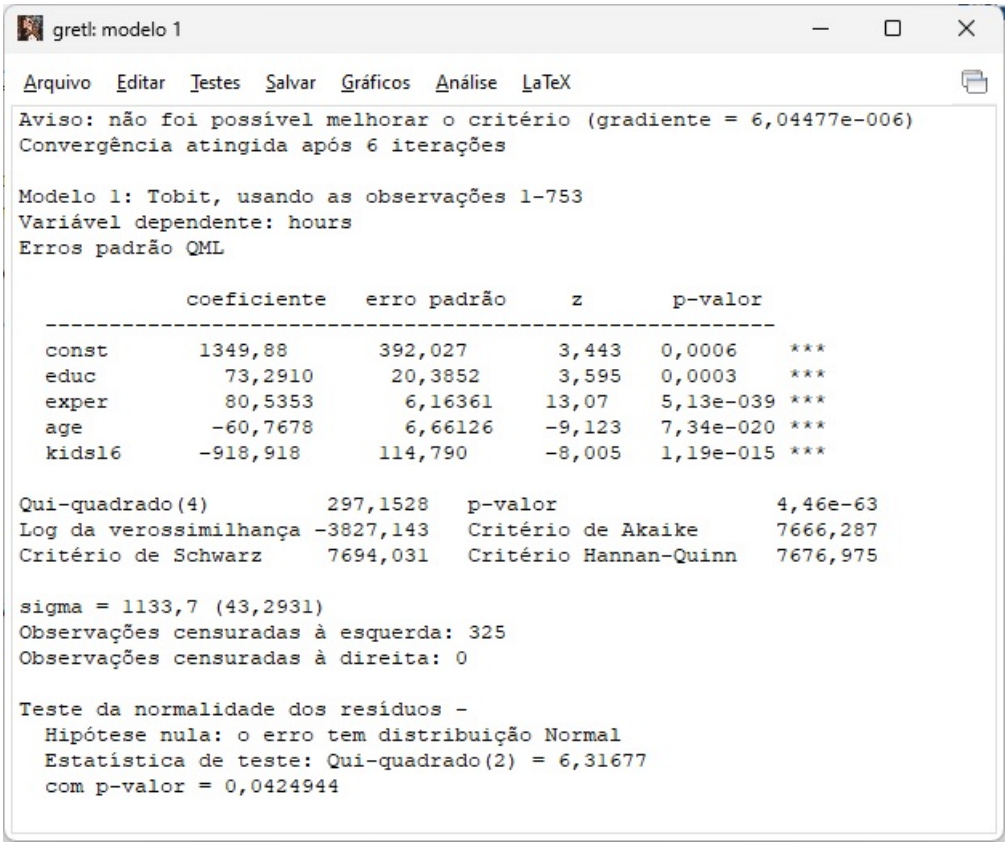
11.7 Tobit

O modelo Tobit é uma regressão linear em que algumas observações da variável dependente foram censuradas. Uma variável censurada é aquela que uma vez que atinge a um limite, esse valor limitador é registrado, não importa o valor de fato. Por exemplo, algum indivíduo com ganhos acima de 1 milhão de reais ou mais por ano poderia ser registrado no limite superior que seria o de ganhos acima de 1 milhão. Isso significa que indivíduos que ganham valores próximos ao limite superior, por exemplo, 1 milhão e 100 mil reais estão no mesmo grupo daqueles indivíduos que ganham 10 milhões de reais. Para dados desse tipo, o modelo de mínimos quadrados pode ser seriamente enviesado e então é aconselhável usar um modelo de regressão censurado (tobit) para estimar os parâmetros da regressão.

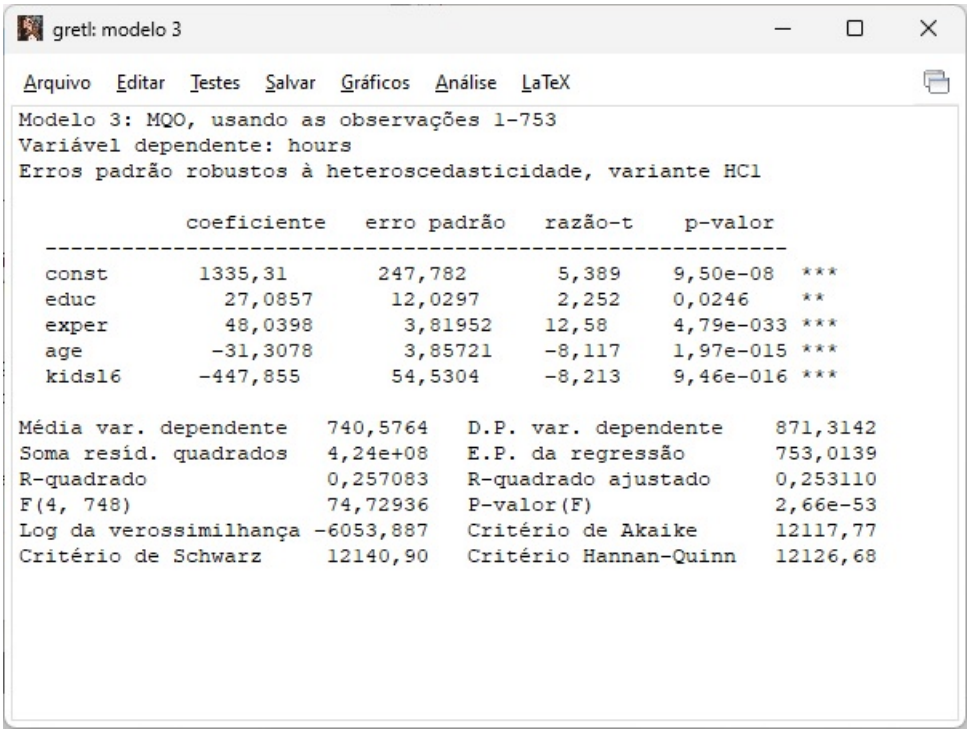
Considere o seguinte modelo de regressão, tendo como variável dependente o número de horas trabalhadas por uma amostra composta apenas por mulheres.

$$hours_i = \beta_1 + \beta_2\, educ_i + \beta_3\, exper_i + \beta_4\, age_i + \beta_5\, kidsl6_i + e_i$$

Pode-se estimar um modelo como uma regressão censurada, uma vez que várias mulheres na amostra trabalham zero horas, ou seja, não trabalham. Será utilizado a base `mroz.gdt`. Posteriormente, clique no menu **Modelo>Variável dependente limitada>Tobit**.



Ao estimar a regressão por Tobit observa-se um efeito positivo e significativo da educação, nas horas trabalhadas. Em outras palavras, um maior nível de escolaridade aumenta a chance da mulher trabalhar mais. Se estimar a equação acima por um modelo de MQO, percebe-se que o efeito da educação será subestimado, como segue:



11.8 Heckit

O viés de seleção ocorre quando em alguma das observações não se tem os dados para a variável dependente por alguma razão. Os problemas estatísticos ocorrem quando a causa da limitação da amostra está relacionada por alguma razão com a variável dependente. Ignorando a correlação, o modelo pode ser estimado usando Mínimos Quadrados, Tobit ou Mínimos Quadrados Censurados (regressão censurada). De qualquer forma, não é possível obter estimativas consistentes dos parâmetros de regressão quando a causa das observações faltantes está correlacionada com a variável dependente do modelo de regressão.

Considere um modelo que consiste em duas equações. A primeira será denominada de equação de seleção e pode ser definida como:

$$z_i^* = \gamma_1 + \gamma_2 w_i + u_i, \quad i = 1, \ldots, N$$

em que z_i^* é uma variável latente, γ_1 e γ_2 são os parâmetros, w_i é uma variável explicativa e u_i é o distúrbio aleatório. Uma variável latente é não observável, mas, por sua vez, uma variável dicotômica pode ser observada:

$$z_i = \begin{cases} 1 & z_i^* > 0 \\ 0 & \text{caso contrário} \end{cases}$$

A segunda equação é chamada de equação de regressão, e é o modelo de regressão linear de interesse.

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, n \quad ; N > n$$

em que y_i é uma variável aleatória observável, β_1 e β_2 são os parâmetros, x_i é uma variável exógena e e_i é um erro aleatório. Assumi-se que os erros aleatórios das duas equações são distribuídos como:

$$\begin{bmatrix} u_i \\ e_i \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma_e^2 \end{pmatrix}\right]$$

O problema de seleção surge quando y_i é observado somente quando $z_i = 1$ e $\rho \neq 0$. Nesse caso, os estimadores de mínimos quadrados de β é viesado e inconsistente. Um estimador consistente foi sugerido por Heckman (1979) e é comumente referenciado como o estimador de dois passos de Heckman ou simplesmente Heckit. Isso ocorre porque os erros são normalmente distribuídos e também os parâmetros são estimados por máxima verossimilhança. O estimador Heckit está baseado na média condicional de y_i quando essa variável pode ser observada:

$$E\left[y_i \mid z_i > 0\right] = \beta_1 + \beta_2\, x_i + \beta_\lambda \lambda_i$$

em que:

$$\lambda_i = \frac{\phi\left(\gamma_1 + \gamma_2\, w_i\right)}{\Phi\left(\gamma_1 + \gamma_2\, w_i\right)}$$

é a razão inversa de Mill. $\phi\left(\gamma_1 + \gamma_2 w_i\right)$ é uma função de densidade de probabilidade valorada ao índice i e; $\Phi\left(\gamma_1 + \gamma_2 w_i\right)$ é a função de densidade cumulativa da distribuição normal avaliada a esse índice. Adicionando um erro aleatório temos:

$$y_i = \beta_1 + \beta_2 x_i + \beta_\lambda \lambda_i + v_i$$

Pode-se mostrar que a equação acima é heterocedástica e se λ_i fosse conhecido (e não estocástico), então o modelo com a correção do viés de seleção poderia ser estimado por Mínimos Quadrados Generalizados. Como alternativa, pode ser estimado por mínimos quadrados ordinários, usando o estimador de covariância consistente de heterocedasticidade de `White (HCCME)` para teste de hipótese e construção de intervalos de confiança. Infelizmente, λ_i não é conhecido e deve ser estimado usando a amostra. A natureza estocástica de λ_i torna inapropriado o uso automático de HCCME neste contexto.

Os dois passos do estimador `Heckit` consistem em:

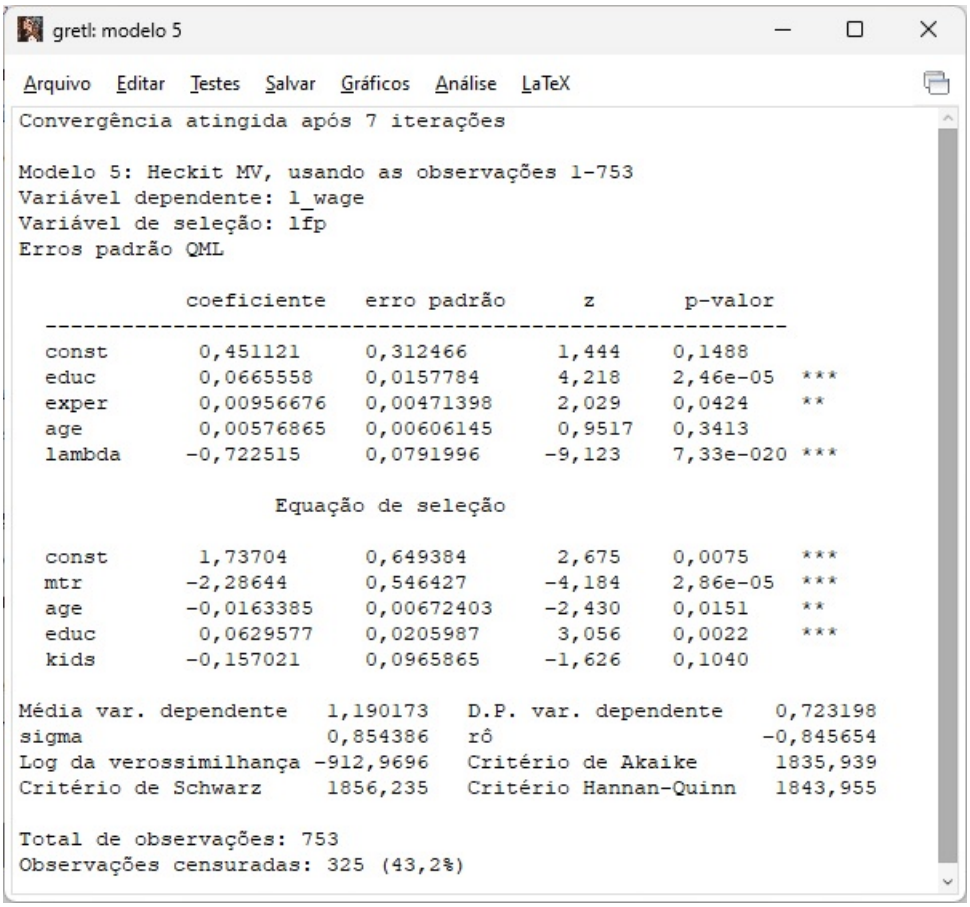
1. Estime a equação de seleção para obter $\hat{\gamma_1}$ e $\hat{\gamma_2}$. Use-os para estimar a `razão inversa de Mill`, $\hat{\lambda_i}$.

2. Adicione $\hat{\lambda_i}$ ao modelo de regressão como na equação e estime-o usando mínimos quadrados.

O procedimento `Heckit` leva em consideração que a decisão de trabalhar por remuneração pode estar correlacionada com o salário que uma pessoa ganha. Ele começa modelando a decisão de trabalhar e estimando a equação de seleção resultante usando um modelo probit. O modelo pode conter mais de uma variável explicativa, w_i, e neste exemplo há quatro: a idade de uma mulher, seus anos de escolaridade, uma variável *dummy* para saber se ela tem filhos e a alíquota marginal de imposto que ela pagaria sobre os ganhos se estivesse empregada.

A base de dados `mroz.gdt` continuará sendo utilizada. O primeiro passo é criar o logaritmo da variável salário (*wage*), selecionando-a e pressionando o botão direito do mouse. A seguir cria-se uma variável *dummy kids* para verificar se há a presença de crianças na residência da família. Para isso, utiliza o menu **Acrescentar>Definir nova variável** bem como a seguinte expressão: $serieskids = (kidsl6 + kids618 > 0)$. Em seguinda, selecione **Modelo>Variável dependente limitada>Heckit** na janela principal do **gretl**. Insira l_wage como a variável dependente e a variável indicadora lfp como a variável de seleção. Em seguida, insira as variáveis independentes desejadas para as equações de regressão e seleções.

Por fim, selecione o botão de estimativa em 2 etapas na parte inferior da caixa de diálogo e clique em `OK`. Então, será possível notar que as estimativas dos coeficientes são idênticas às produzidas manualmente acima. No entanto, os erros padrão, que agora são estimados de forma consistente, mudaram. O **gretl** também produz as estimativas da equação de seleção, que aparecem diretamente abaixo daquelas da regressão.



Capítulo 12

Modelos de equações simultâneas

Este capítulo apresenta um modelo de oferta e demanda. Sendo assim, importante destacar que esse modelo econométrico contém duas variáveis dependentes e duas equações. Uma característica dos modelos de equações simultâneas é que os valores de duas (ou mais) variáveis são determinados conjuntamente. Isso significa que uma mudança em uma das variáveis faz com que a outra variável mude e vice-versa. A estimativa de um modelo de equações simultâneas é demonstrada usando o exemplo da trufa. Para isso utilizará a base `truffles.gdt`.

12.1 Exemplo do modelo de equações simultâneas para trufa

Considere um modelo de oferta e de demanda para trufas:

$$q_i = \alpha_1 + \alpha_2\, p_i + \alpha_3\, ps_i + \alpha_4\, di_i + e_i^d \tag{12.1}$$

$$q_i = \beta_1 + \beta_2\, p_i + \beta_3\, pf_i + e_i^s \tag{12.2}$$

A Equação 12.1 é a demanda por trufas em que q representa a quantidade demandada em um determinado mercado, p é o preço de mercado da trufa, ps é o preço de um bem substituto e d_i é a renda disponível *per capita* do mercado local. Por sua vez, a Equação 12.2 caracteriza-se como sendo a equação de oferta. Essa equação contém a variável pf que representa o preço de um fator de produção. Cada observação é indexada por meio do índice $i = 1, 2, \ldots, N$. Como será visto, preços e quantidades em um mercado são determinados conjuntamente, portanto, neste modelo econométrico, p e q são ambos endógenos ao sistema.

12.2 As equações na forma reduzida

Destaca-se que as equações na forma reduzida expressam cada variável endógena como função linear de cada variável exógena em todo o sistema. Assim,

$$q_i = \pi_{11} + \pi_{21}\, ps_i + \pi_{31}\, di_i + \pi_{41}\, pf_i + v_{i1} \tag{12.3}$$

$$p_i = \pi_{12} + \pi_{22}\, ps_i + \pi_{32}\, di_i + \pi_{42}\, pf_i + v_{i2} \tag{12.4}$$

Uma vez que cada uma das covariáveis (variáveis independentes) é exógena em relação a q e p, as equações na forma reduzida (12.3) e (12.4) podem ser estimadas usando mínimos quadrados.

Os resultados do **gretl** aparecem abaixo. Cada uma das variáveis é individualmente diferente de zero a 5%. As estatísticas F gerais são $19,79$ e $69,19$; ambas, também, significantes a 5%.

$$\underset{(3,243)}{\hat{q} = 7,895} + \underset{(0,1425)}{0,6564\, ps} + \underset{(0,7005)}{2,167\, di} - \underset{(0,1213)}{0,5070\, pf}$$

$$n = 30 \quad \bar{R}^2 = 0,6625 \quad F\,(3,\ 26) = 19,973 \quad \hat{\sigma} = 2,6801$$

(erros padrão entre parênteses)

$$\underset{(7,984)}{\hat{p} = -32,51} + \underset{(0,3509)}{1,708\, ps} + \underset{(1,724)}{7,602\, di} + \underset{(0,2985)}{1,354\, pf}$$

$$n = 30 \quad \bar{R}^2 = 0,8758 \quad F\,(3,\ 26) = 69,189 \quad \hat{\sigma} = 6,5975$$

(erros padrão entre parênteses)

12.3 As equações estruturais

As equações estruturais são estimadas empregando o estimador de Mínimos Quadrados em Dois Estágios (MQ2E). Os instrumentos utilizados na estimação do MQ2E consistem em todas as variáveis exógenas, i.e., as mesmas empregadas para estimar as equações na forma reduzida (12.3) e (12.4).

A seguir apresenta-se os comandos, a serem passados no console do **gretl** – Figura 12.1, para abrir os dados da base `truffles.gdt` e estimar as equações estruturais empregado o estimador MQ2E no **gretl**.

1. list z = const ps di pf

2. tsls q const p ps di; z

3. tsls q const p pf; z

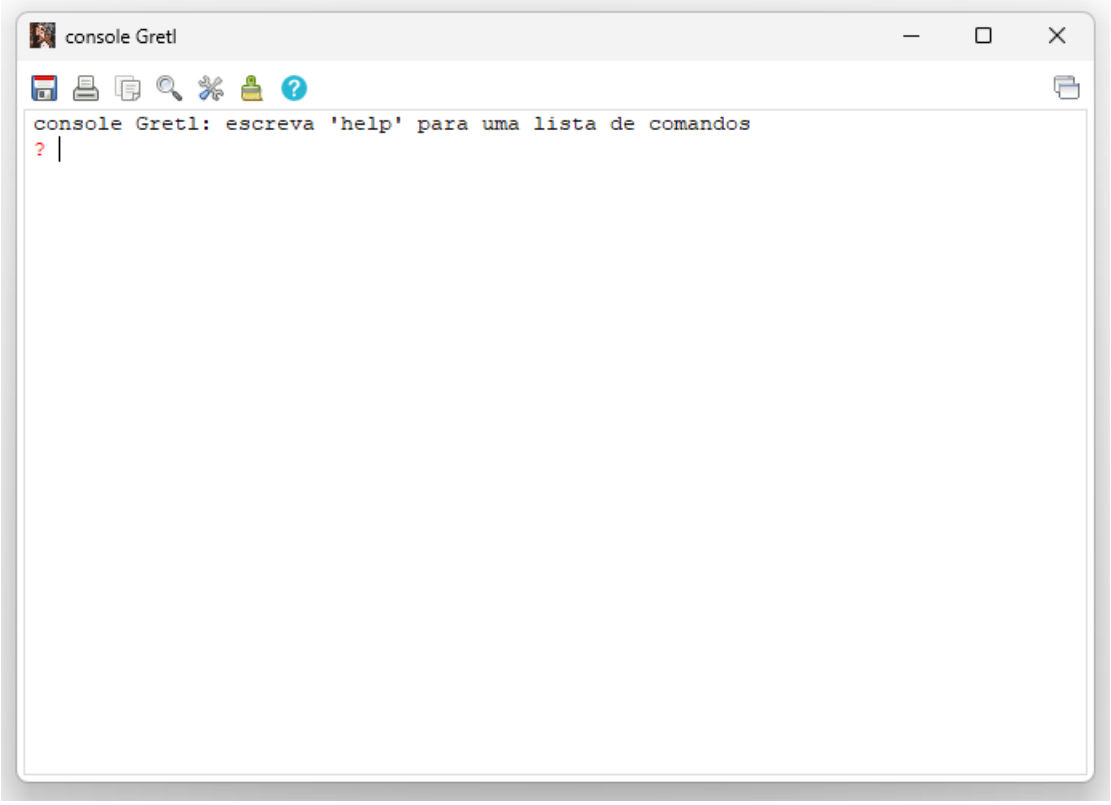


Figura 12.1: Console do **gretl**.

Observe que a primeira linha do *script* cria uma lista chamada de z e que contém todas as variáveis exógenas. Essas variáveis são usadas para calcular a regressão de primeiro estágio, ou seja, a lista de instrumentos. Por sua vez, a linha 2 estima os coeficientes da equação demanda por trufa empregando o estimador `TSLS`. Importante salientar que o comando `tsls` do **gretl** solicita o estimador MQ2E e é seguido pela especificação da equação estrutural que se deseja estimar – no presente exemplo, a variável dependente q e as variáveis independentes *const*, p, ps e di. Note que o ponto e vírgula separa o modelo que se deseja estimar da lista de instrumentos, agora contidos na lista z. Já a terceira linha segue o mesmo raciocínio da equação demanda, porém, agora para estimar os parâmetros da equação de oferta de trufa.

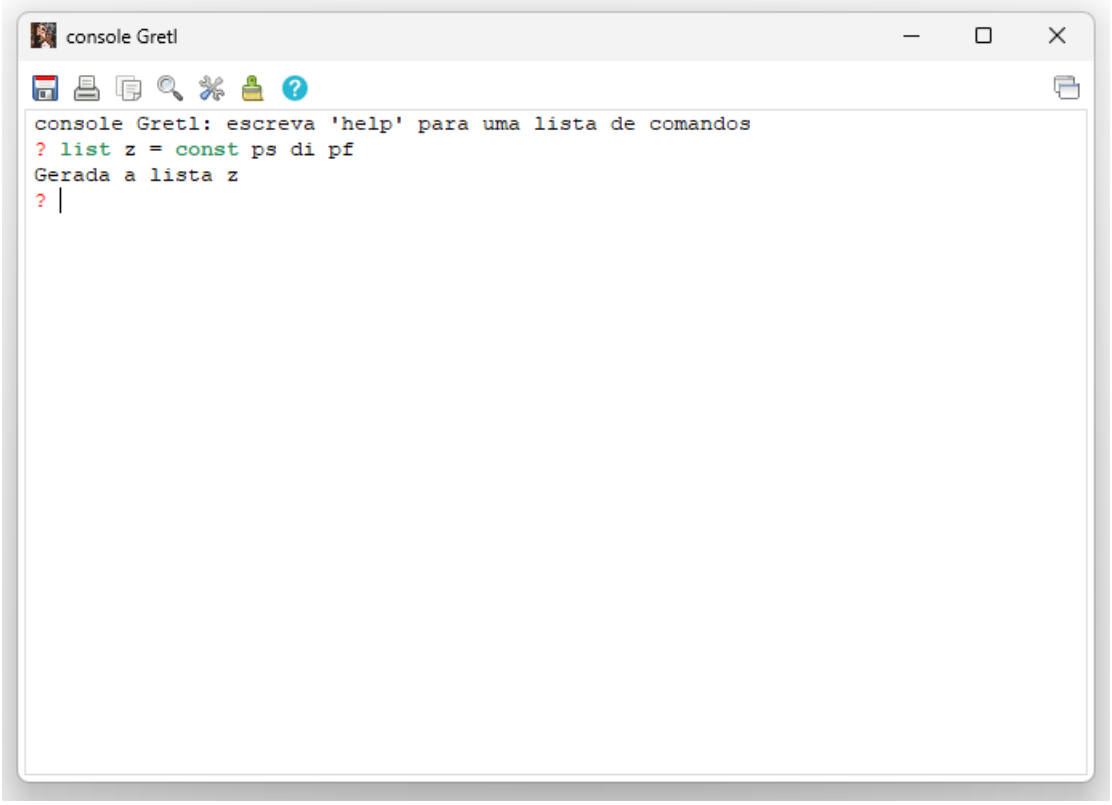


Figura 12.2: Criando uma lista com todas as variáveis exógenas.

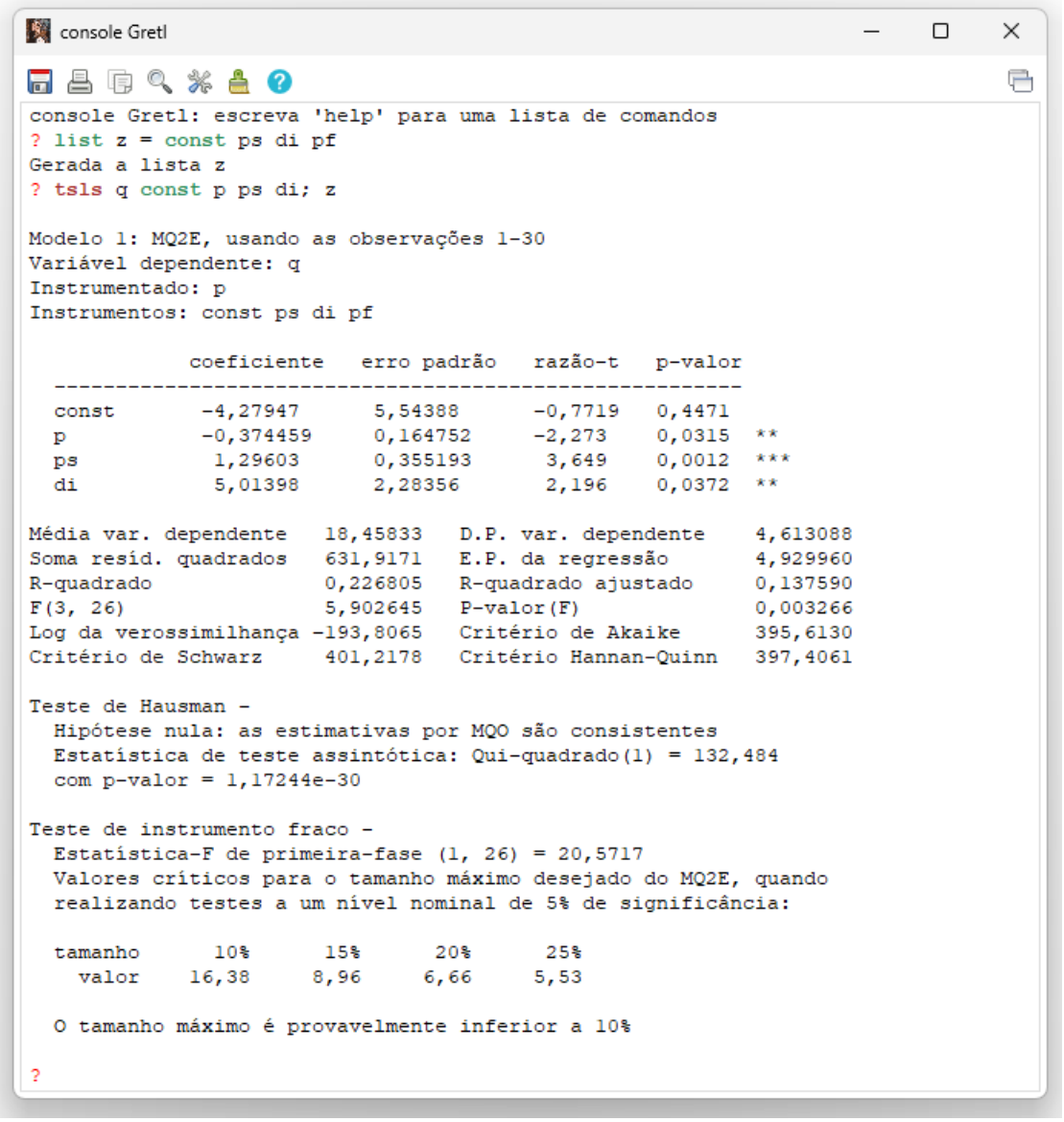


Figura 12.3: Estimando os coeficientes da equação demanda.

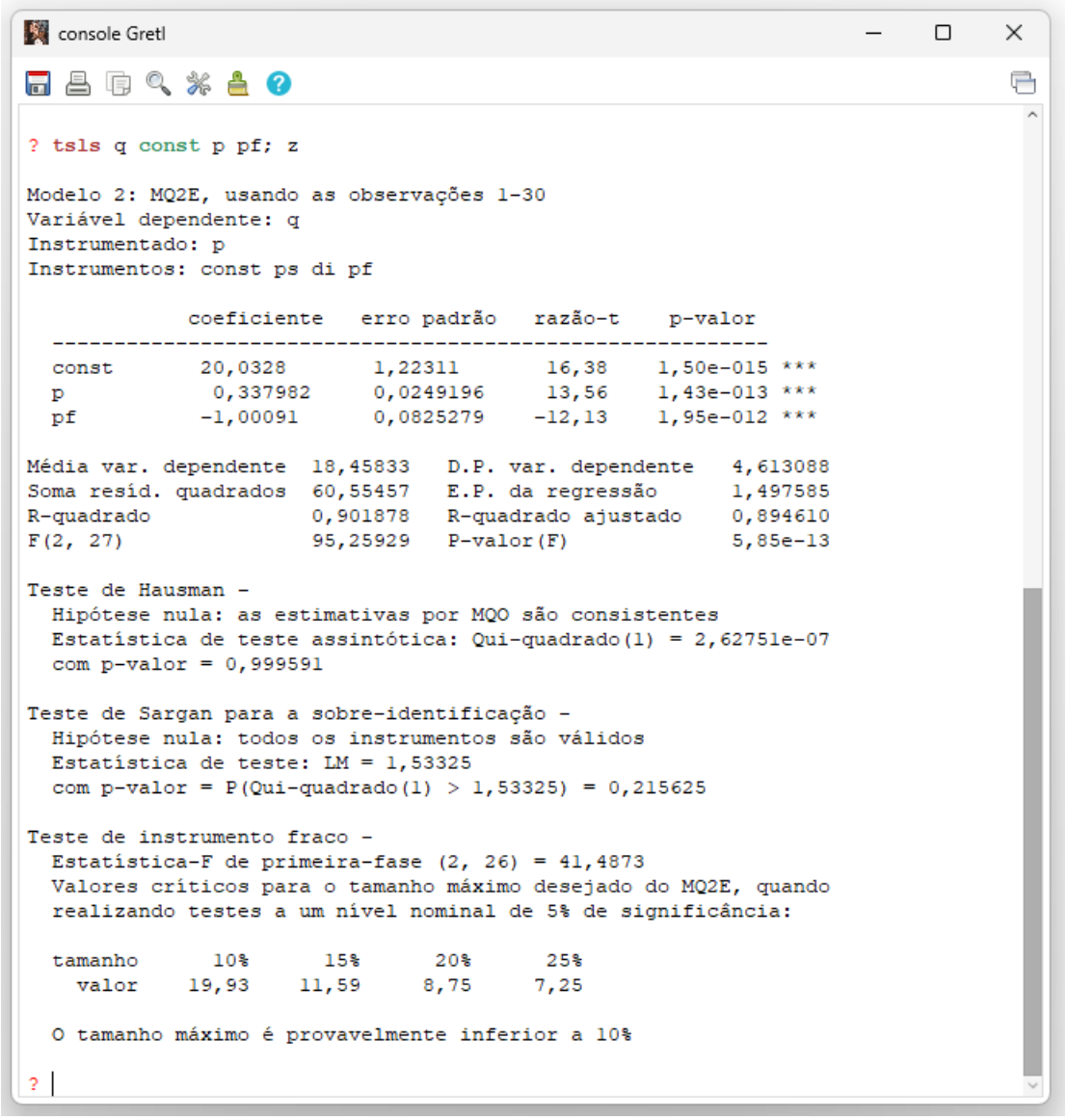


Figura 12.4: Estimando os parâmetros da equação oferta.

A Figura 12.3 mostra os resultados reportados pela estimativa de Mínimos Quadrados em Dois Estágios (MQ2E) da equação de demanda. O coeficiente do preço na equação de demanda é de $-0,374$ e é significativamente negativo a 5%. Lembre-se de que as curvas de demanda são negativamente inclinadas. Ademais, o `teste de Hausman` reportou um valor de $132,484$ com um `p-valor` próximo de zero e, assim, evidenciando que o preço não é uma variável exógeno. O teste de instrumentos fracos excede 10 e, portanto, o conjunto de instrumentos é bastante forte.

Os resultados para a estimação em dois estágios da especificação da oferta são apresentados na Figura 12.4. Como esperado, o coeficiente do preço é positivo. O resultado do `teste de Sargan`, `p-valor` $= 0,215625 > 0,05$, caracteriza que o modelo é adequadamente superidentificado. Além disso, o teste de instrumento fraco demonstra que os instrumentos utilizados na estimação são adequadamente fortes (`estatística-F` $(2, 26) = 41,4873$).

Capítulo 13

Modelos de contagem

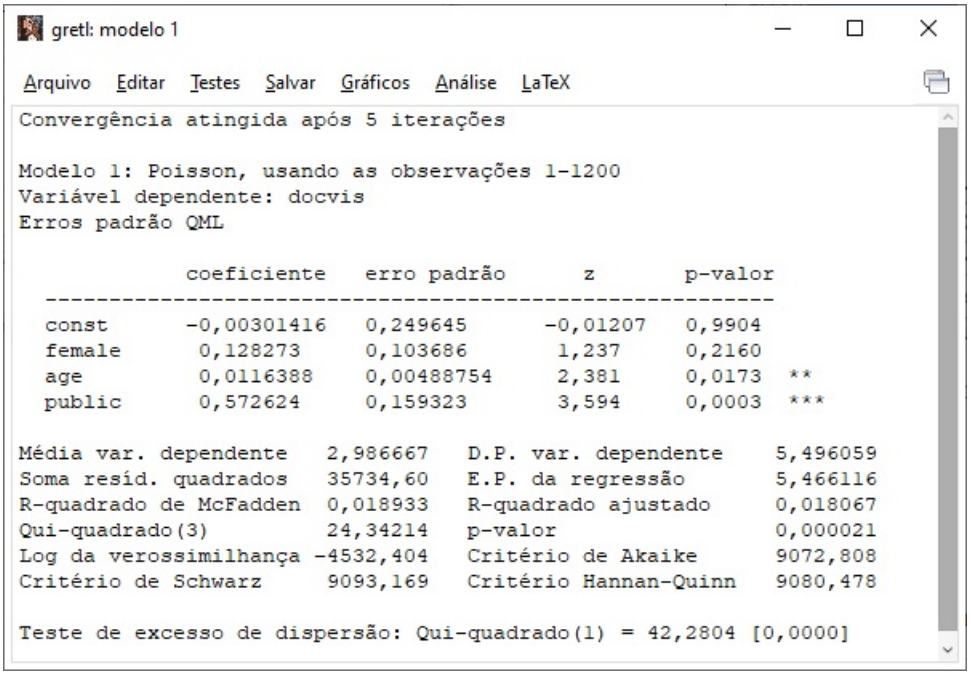
Quando a variável dependente em um modelo de regressão é uma "contagem" do número de ocorrências de um evento, pode-se querer usar o modelo de regressão de Poisson. Nestes modelos, a variável dependente é um número inteiro não negativo (ou um número natural), que representa o número de ocorrências de um determinado evento. Quando se está trabalhando com dados de contagem, inicia-se a estimação dos parâmetros por meio de um modelo de regressão Poisson, devido à sua simplicidade. Neste caso, a variável dependente de um modelo de regressão Poisson deve seguir uma distribuição Poisson com média igual à variância. Nestes casos, trabalha-se com a estimação de um modelo de regressão binomial negativo.

A probabilidade de um determinado número de ocorrências é modelada em função de variáveis independentes.

$$P\left(Y = y \,|\, x\right) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

em que $\lambda = \beta_1 + \beta_2\, x$ é a função de regressão.

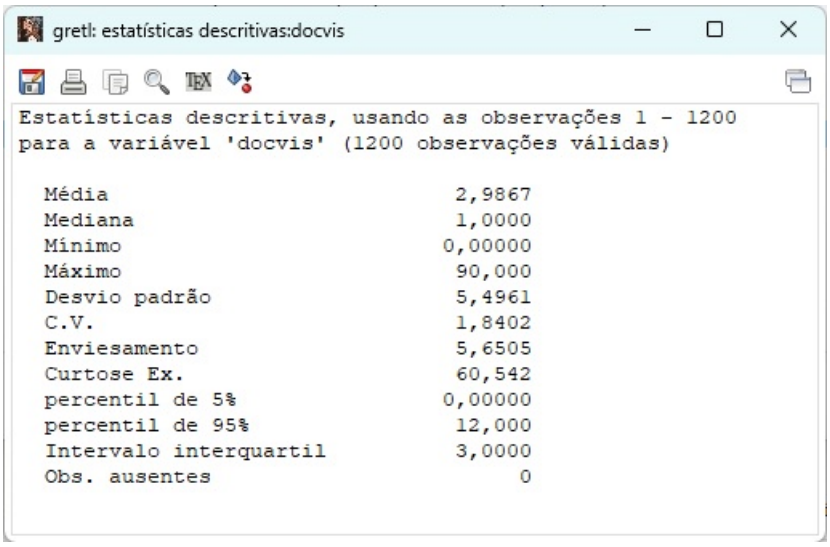
A estimação desse modelo, se dá por máxima verossimilhança. Como exemplo, será usado o número de consultas médicas nos últimos três anos. Este número será modelado em função da idade da pessoa, sexo e se ela tem seguro público ou privado. Os dados estão em `rwm88_small.gdt`, que são um subconjunto do German Socioeconomic Panel Survey de 1988. Depois que os dados são carregados, os modelos para dados de contagem podem ser acessados por meio do sistema de menu usando **Modelo>Variável dependente limitada>Contagem**.



As variáveis *age* e *public* são estatisticamente diferentes de zero.

13.1 Teste de superdispersão

Caso a variância da variável dependente seja consideravelmente maior do que a sua média, a estimação de um modelo Poisson poderá gerar parâmetros viesados, por conta do problema conhecido por superdispersão. É sempre recomendável, portanto, que, após a estimação de um modelo de regressão Poisson, seja elaborado um teste para verificação da existência de superdispersão e, caso sua presença seja detectada, será recomendada a estimação de um modelo de regressão binomial negativo. Seguindo o exemplo anterior, pode-se gerar as estatísticas descritivas da variável dependente:



Como observado, a média é diferente da variância. Cameron e Trivedi (1990) propõem um interessante procedimento para verificação da existência de superdispersão em modelos de regressão Poisson. Para tanto, é preciso que seja gerada uma variável Y^*, da seguinte maneira:

$$Y_i^* = \frac{[(Y_i - \hat{\mu}_i)^2 - Y_i]}{\hat{\mu}_i}$$

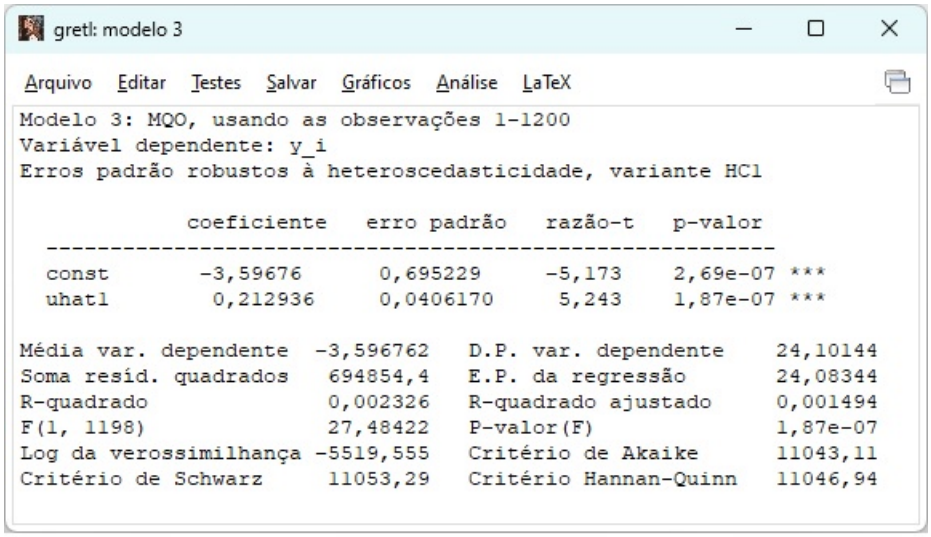
Sendo Y_i o número de ocorrências para cada observação da amostra. O termo $\hat{\mu}_i$ é o resíduo da regressão de Poisson. Após criar Y_i^* deve-se estimar o seguinte modelo de regressão:

$$\hat{Y_i^*} = \beta \hat{\mu}_i$$

Após a estimação do modelo de regressão se o parâmetro β for estatisticamente diferente de zero observa-se o fenômeno da superdispersão. Para isso, após rodar a regressão deve-se salvar os resíduos. Para isso clique no menu **Salvar>Resíduos**. Guarde a variável como $uhat1$. Posteriormente adiciona-se uma nova variável, clicando no menu **Acrescentar>definir nova variável**. A fórmula é a seguinte:

$$y_i = \frac{(docvis - uhat1)^2 - docvis)}{uhat1}$$

Em seguida estima-se o seguinte modelo de regressão por Mínimos Quadrados Ordinários:



Observe que o parâmetro β é estatisticamente diferente de zero, logo há o fenômeno da superdispersão e as estimativas devem ser executadas por meio da distribuição binomial negativa.

13.2 Binomial Negativa

Esta distribuição é também conhecida por distribuição Poisson-Gama por ser uma combinação de duas distribuições que foi desenvolvida para levar em consideração o fenômeno da superdispersão que é comumente observado em dados de contagem. Ainda segundo os autores, leva este nome por aplicar o teorema binomial com um expoente negativo. Se, por exemplo, a média do número de ocorrências de uma distribuição Poisson possuir uma parcela aleatória, a expressão (14.5) passará ser escrita da seguinte forma:

$$\lambda_i = e^{(\alpha+\beta_1 x_{1i}+\cdots+\beta_1 x_{ki}+\epsilon_i)}$$

que pode ser escrita como:

$$\lambda_i = u_i\, v_i$$

que possui uma distribuição binomial negativa, em que o primeiro termo (u_i) representa o valor esperado de ocorrências e possui uma distribuição Poisson e o segundo termo (v_i) corresponde à parcela aleatória do número de ocorrências da variável dependente e possui uma distribuição Gama. Para determinada observação i ($i = 1, 2, \ldots, n$ em que n é o tamanho da amostra), a função da distribuição de probabilidade da variável v_i:

$$p\,(v_i) = \frac{\delta^{\psi}\, v_i^{\psi-1}\, e^{-v_i\delta}}{\Gamma\,(\psi)}$$

O parâmetro de forma é $\psi > 0$ e o parâmetro de taxa $\delta > 0$. Pode-se combinar as expressões de modo a gerar a função da probabilidade de uma distribuição binomial negativa, o que nos permitirá calcular a probabilidade de ocorrência de uma contagem m, dada determinada exposição.

$$p\,(Y_i = m) = \left(\begin{array}{c} m+\psi-1 \\ \psi-1 \end{array}\right)\left(\frac{\psi}{u_i+\psi}\right)^{\psi}\left(\frac{u_i}{u_i+\psi}\right)^{m}, \qquad m = 0,\, 1,\, 2,\, \ldots$$

que representa a função de probabilidade da distribuição binomial negativa para a ocorrência de uma contagem m, com as seguintes estatísticas:

$$\text{Média: } E\,(Y) = u$$

$$\text{Variância: } Var\,(Y) = u + \alpha\, u^2$$

sendo $\alpha = \frac{1}{\psi}$.

O segundo termo da expressão de variância da distribuição binomial negativa representa a superdispersão. Se observar que $\alpha \to 0$, este fenômeno não estará presente nos dados. No entanto, quando ϕ é estatisticamente maior do que zero, deve-se estimar um modelo de regressão binomial negativo.

O **gretl** permite a estimação de dois modelos de regressão binomial negativo. O modelo apresentado acima é conhecido como NB2 (negative binomial 2 regression model). Uma versão alternativa, utiliza a seguinte expressão para a variância:

$$Var\,(Y) = u\,(1+\alpha)$$

e, é conhecido por modelo de regressão NB1 (negative binomial 1 regression model). Utiliza-se a mesma regressão aplicada no modelo de Poisson, utilizando a distribuição `NegBin2:`