Resumo do conteúdo - Prova 1

Análise de correlação linear

Covariância

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{SPXY}{n-1}$$

Coeficiente de correlação linear de Pearson

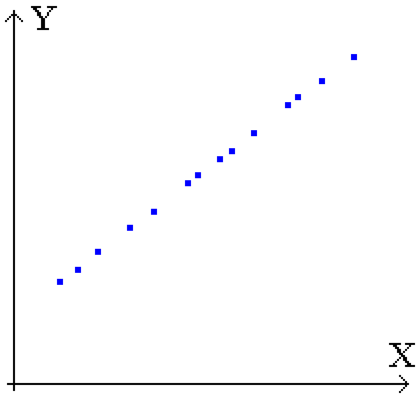
$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{SPXY}{\sqrt{SQX.SQY}}$$

$$SPXY = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

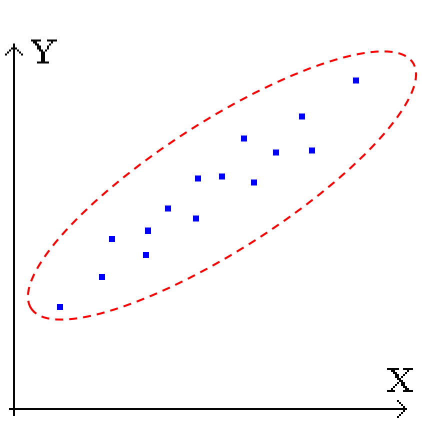
$$SQX = \sum(x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$SQY = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

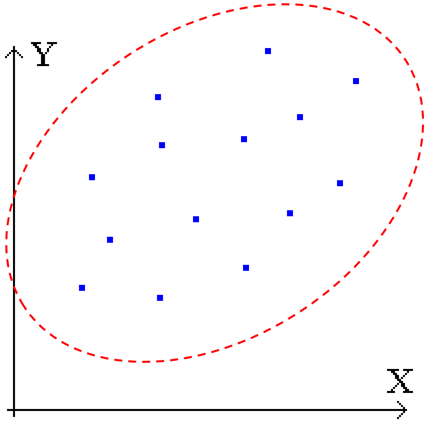
Interpretação do coeficiente de correlação



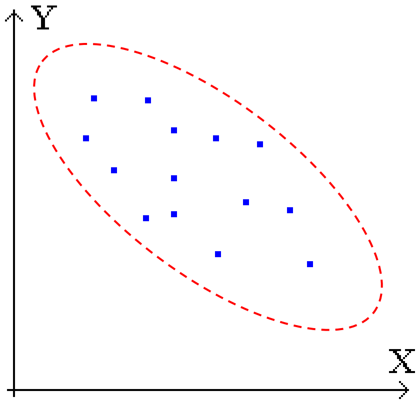
Positiva perfeita
r = 1



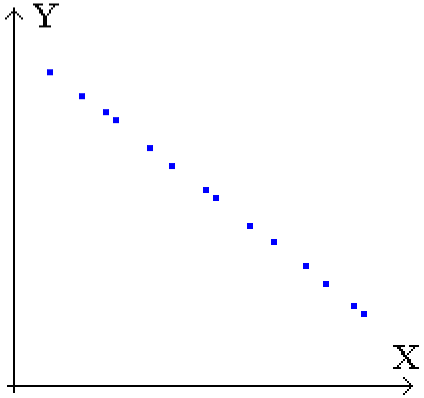
Positiva elevada
r = 0,8



Positiva baixa
r = 0,1



Negativa média
r = -0,5



Negativa perfeita
r = -1

Teste de hipótese sobre o coeficiente de correlação ρ

Pressuposição:

- Distribuição da variável (X,Y) é normal bivariada;

Hipótese estatística:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Estatística do teste:

$$T = \frac{R_{xy}}{\sqrt{\dfrac{1 - R_{xy}^2}{n-2}}}$$

Intervalo de confiança para o coeficiente de correlação ρ

Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;
- As variáveis são correlacionadas, ou seja, $\rho \neq 0$.

Cálculo dos limites do intervalo:

$$z_{\alpha/2} = z_{0,475} = 1,96$$

$$IC(\rho; 0,95): \left[\frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0,9983}{1-0,9983}\right) - \frac{1,96}{\sqrt{22-3}}\right]\right\} - 1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0,9983}{1-0,9983}\right) - \frac{1,96}{\sqrt{22-3}}\right]\right\} + 1} ; \frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0,9983}{1-0,9983}\right) + \frac{1,96}{\sqrt{22-3}}\right]\right\} - 1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0,9983}{1-0,9983}\right) + \frac{1,96}{\sqrt{22-3}}\right]\right\} + 1} \right]$$

Coeficiente de correlação de postos de Spearman

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$ onde: d_i = posto de x_i – posto de y_i

Teste de hipótese sobre o coeficiente de correlação ρ_s

Hipótese estatística:

$$\begin{cases} H_0 : \rho_s = 0 \\ H_1 : \rho_s \neq 0 \end{cases}$$

Decisão:

Buscar na tabela o valor absoluto mínimo (crítico) para o r_s ser significativo, com base no tamanho da amostra (n) e no nível de significância (α) e compará-lo com o valor calculado.

$|r_s \text{ calculado}| \geq r_s \text{ crítico} \rightarrow$ Rejeita-se H_0

Análise de regressão simples

Análise de regressão simples

Modelo estatístico amostral: $y_i = \beta_0 + \beta_1 x_i + e_i$

Objetivo: determinar a equação que melhor representa a relação existente entre duas variáveis e, a partir desta equação, fazer predições para a variável resposta.

Para isso, uma sequência de passos deve ser seguida:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0 e β_1 para ajustar a equação da regressão.

2. Aplicação de testes de hipóteses para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

3. Construção de intervalos de confiança para os valores estimados pela equação de regressão.

1. Estimar os parâmetros do modelo

Estimativa do coeficiente de regressão

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{SPXY}{SQX}$$

Estimativa do intercepto

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Quantidades necessárias (obtidas da tabela auxiliar)

$$SQX = \sum (x_i - \bar{x})^2 = \boxed{\sum x_i^2 - n\bar{x}^2}$$

$$SQY = \sum (y_i - \bar{y})^2 = \boxed{\sum y_i^2 - n\bar{y}^2}$$

$$SPXY = \sum (x_i - \bar{x})(y_i - \bar{y}) = \boxed{\sum x_i y_i - n\bar{x}\,\bar{y}}$$

2. Ajustar a equação da reta

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$$

3. Estimar os valores esperados (médias) de Y

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x$$

4. Estimar os erros

$$\hat{e}_i = y_i - \hat{\mu}_i$$

Testar de hipóteses sobre β_1

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Metodologias para testar H_0 $\begin{cases} \text{Teste t} \\ \text{Análise da variância} \end{cases}$

Teste t

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Podemos testar H_0 utilizando a estatística T com distribuição de t de Student.

$$T = \frac{\hat{\beta}_1}{S(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{S^2(\hat{\beta}_1)}} \sim t(\nu = n - 2)$$

Variâncias das estimativas dos coeficientes

$$S^2(\hat{\beta}_1) = \frac{\dfrac{\sum \hat{e}_i^2}{n-2}}{\sum (x_i - \bar{x})^2}$$

$$S^2(\hat{\beta}_1) = \frac{S^2 Res}{SQX}$$

Quantidades utilizadas no teste t e nos intervalos de confiança para β_1 e β_2

$$S^2(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \frac{\sum \hat{e}_i^2}{n-2}$$

$$S^2(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}^2}{SQX} \right] S^2 Res$$

Análise da variância

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	ν_{Reg}=p-1	$\sum(\hat{\mu}_i - \bar{y})^2$	$S^2_{Reg} = \frac{SQ_{Reg}}{\nu_{Reg}}$	$\frac{S^2_{Reg}}{S^2_{Res}}$
Resíduo	ν=n-p	$\sum(y_i - \hat{\mu}_i)^2$	$S^2_{Res} = \frac{SQ_{Res}}{\nu}$	-
Total	ν_{Total}=n-1	$\sum(y_i - \bar{y})^2$	-	-

$$SQ_{Total} = \sum(y_i - \bar{y})^2 = SQY$$

$$SQ_{Reg} = \sum(\hat{\mu}_i - \bar{y})^2 = \beta_1^2 \sum(x_i - \bar{x})^2 = \beta_1^2 SQX$$

$$SQ_{Res} = \sum(y_i - \hat{\mu}_i)^2 = \sum \hat{e}_i^2 \text{ (por diferença)}$$

Coeficiente de determinação (r^2)

$$r^2 = \frac{SQ_{Reg}}{SQ_{Total}}$$

Intervalo de confiança para os coeficientes

$$IC(\beta_1;1-\alpha):\hat{\beta}_1 \pm t_{\alpha/2}\sqrt{S^2(\hat{\beta}_1)}$$

$$IC(\beta_0;1-\alpha):\hat{\beta}_0 \pm t_{\alpha/2}\sqrt{S^2(\hat{\beta}_0)}$$

Intervalo de confiança para médias (predição)

$$IC(\mu_i;1-\alpha):\hat{\mu}_i \pm t_{\alpha/2}\sqrt{S^2(\hat{\mu}_i)}$$

Intervalo de confiança para observações (previsão)

$$IC(y_i;1-\alpha):\hat{\mu}_i \pm t_{\alpha/2}\sqrt{S^2(y_i)}$$

Variância das estimativas de valores esperados

$$S^2(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)$$

$$S^2(\hat{\mu}_i) = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQX}\right)\left(S^2 Res\right)$$

Variância de uma observação

$$S^2(y_i) = \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)\left(\frac{\sum \hat{e}_i^2}{n-2}\right)$$

$$S^2(y_i) = \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQX}\right)\left(S^2 Res\right)$$

Quantidades utilizadas nos intervalos de confiança para μ_i e y_i

Análise de regressão múltipla com duas variáveis preditoras

Análise de regressão múltipla com duas variáveis preditoras

Modelo estatístico amostral: $y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_j$

Objetivo: determinar a equação que melhor representa a relação existente entre as três variáveis e, a partir desta equação, fazer predições para a variável resposta.

Para isso, uma sequência de passos deve ser seguida:

1. Obtenção das estimativas (pontuais) dos coeficientes β_0, β_1 e β_2 para ajustar a equação da regressão.

2. Aplicação de testes de hipóteses para as estimativas obtidas, a fim de verificar se a equação de regressão é adequada.

3. Construção de intervalos de confiança para os valores estimados pela equação de regressão.

1. Estimar os parâmetros do modelo

Estimadores dos coeficientes de regressão parciais

$$\begin{cases} \hat{\beta}_1 \sum (x_{1j} - \bar{x}_1)^2 + \hat{\beta}_2 \left[\sum (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)\right] = \sum (x_{1j} - \bar{x}_1)(y_j - \bar{y}) \\ \hat{\beta}_1 \left[\sum (x_{1j} - \bar{x}_1)(x_{2j} - \bar{x}_2)\right] + \hat{\beta}_2 \sum (x_{2j} - \bar{x}_2)^2 = \sum (x_{2j} - \bar{x}_2)(y_j - \bar{y}) \end{cases}$$

$$\begin{cases} \hat{\beta}_1 SQX_1 + \hat{\beta}_2 SPX_1X_2 = SPX_1Y \\ \hat{\beta}_1 SPX_1X_2 + \hat{\beta}_2 SQX_2 = SPX_2Y \end{cases}$$

Estimador do intercepto

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

Quantidades necessárias (obtidas da tabela auxiliar)

$$SQX_1 = \sum (x_{1j} - \bar{x}_1)^2 = \boxed{\sum x_{1j}^2 - n\bar{x}_1^2}$$

$$SQX_2 = \sum (x_{2j} - \bar{x}_2)^2 = \boxed{\sum x_{2j}^2 - n\bar{x}_2^2}$$

$$SQY = \sum (y_j - \bar{y})^2 = \boxed{\sum y_j^2 - n\bar{y}^2}$$

$$SPX_1X_2 = \sum \left[(x_{1j} - \bar{x})(x_{2j} - \bar{x}_2) \right] = \boxed{\sum x_{1j}x_{2j} - n\bar{x}_1\bar{x}_2}$$

$$SPX_1Y = \sum \left[(x_{1j} - \bar{x}_1)(y_j - \bar{y}) \right] = \boxed{\sum x_{1j}y_j - n\bar{x}_1\bar{y}}$$

$$SPX_2Y = \sum \left[(x_{2j} - \bar{x}_2)(y_j - \bar{y}) \right] = \boxed{\sum x_{2j}y_j - n\bar{x}_2\bar{y}}$$

2. Ajustar a equação do plano

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

3. Estimar os valores esperados (médias) de Y

$$\hat{\mu}_{y(x_1,x_2)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

4. Estimar os erros

$$\hat{e}_j = y_j - \hat{\mu}_j$$

Testes de hipóteses sobre os parâmetros

1. Testes da hipótese geral

Hipótese estatística:

$$\begin{cases} H_0 : \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1 : \beta_i \neq 0, \text{ para pelo menos um } i \ \ (i = 1 \text{ e/ou } 2) \end{cases}$$

2. Testes das hipóteses parciais

Hipóteses estatísticas:

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases}$$

$$\begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

Hipótese geral: Análise da variância

Hipóteses estatísticas: $\begin{cases} H_0: \beta_i = 0, \text{ sendo } i = 1, 2 \\ H_1: \beta_i \neq 0, \text{ para pelo menos um } i \ (i = 1 \text{ e/ou } 2) \end{cases}$

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	ν_{Reg}=3-1	$\sum(\hat{\mu}_j - \bar{y})^2$	$\frac{SQ_{Reg}}{\nu_{Reg}}$	$\frac{S^2_{Reg}}{S^2}$
Resíduo	ν=n-3	$\sum(y_j - \hat{\mu}_j)^2$	$\frac{SQ}{\nu}$	-
Total	ν_{Total}=n-1	$\sum(y_j - \bar{y})^2$	-	-

Obtenção das somas de quadrados:

$$SQ_{Total} = \sum(y_j - \bar{y})^2 = SQY$$

$$SQ_{Reg} = \sum(\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1 \times SPX_1Y + \hat{\beta}_2 \times SPX_2Y$$

$$SQ_{Res} = \sum(y_j - \hat{\mu}_j)^2 = \sum \hat{e}_j^2 \text{ (por diferença)}$$

Critério de decisão

Se $f > f_\alpha$, rejeita-se H_0

Se $f < f_\alpha$, não se rejeita H_0

Ajustamento do modelo de regressão linear

Fonte de variação	GL (ν)	SQ	QM (S^2)	F
Regressão	ν_{Reg}=3-1	$\sum(\hat{\mu}_j - \bar{y})^2$	$\frac{SQ_{Reg}}{\nu_{Reg}}$	$\frac{S^2_{Reg}}{S^2}$
Resíduo	ν=n-3	$\sum(y_j - \hat{\mu}_j)^2$	$\frac{SQ}{\nu}$	-
Total	ν_{Total}=n-1	$\sum(y_j - \bar{y})^2$	-	-

Coeficiente de determinação (r^2)

$$r^2 = \frac{SQ_{Reg}}{SQ_{Total}}$$

Coeficiente de determinação corrigido

$$r_C^2 = r^2 - \frac{2}{n-3}\left(1 - r^2\right)$$

Testes de hipóteses sobre os parâmetros

2. Testes das hipóteses parciais

$$\begin{cases} H_0^1 : \beta_1 = 0 \\ H_1^1 : \beta_1 \neq 0 \end{cases} \quad \text{e} \quad \begin{cases} H_0^2 : \beta_2 = 0 \\ H_1^2 : \beta_2 \neq 0 \end{cases}$$

Testes das hipóteses parciais pela estatística t

$$T = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\sqrt{S^2(\hat{\beta}_i)}}$$

$$S^2(\hat{\beta}_1) = \frac{SQX_2}{SQX_1 \times SQX_2 - (SPX_1X_2)^2} \left(S^2 Res\right)$$

$$S^2(\hat{\beta}_2) = \frac{SQX_1}{SQX_1 \times SQX_2 - (SPX_1X_2)^2} \left(S^2 Res\right)$$

Dois critérios devem ser considerados na escolha do modelo:

(1) A significância do efeito linear de X_i sobre Y;

(2) O coeficiente de determinação (r^2).

Mesmo que o efeito linear de X_i sobre Y seja significativo, se o r^2 não for elevado, o modelo não deve ser adotado. Nestes casos, outros modelos não lineares podem ser testados.