Lisbon School of Economics & Management
Universidade de Lisboa

Estatística II

Licenciatura em Gestão do Desporto
2.º Ano/2.º Semestre
2025/2026

Aulas Teórico-Práticas N.^{os} 21 e 22 (Semana 12)

Docente: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

Conteúdos Programáticos

Aulas Teórico-Práticas (Semanas 1 a 3)	Aulas Teórico-Práticas (Semanas 4 a 7)	Aulas Teórico-Práticas (Semanas 7 a 9)	Aulas Teórico-Práticas (Semanas 10 a 13)
• **Capítulo 1**: Revisões e Distribuições de Amostragem	• **Capítulo 2**: Estimação	• **Capítulo 3**: Testes de Hipóteses	• **Capítulo 4**: Modelo de Regressão Linear Múltipla

Material didático: Exercícios do Livro Murteira et al (2015), Formulário e Tabelas Estatísticas

Bibliografia: B. Murteira, C. Silva Ribeiro, J. Andrade e Silva, C. Pimenta e F. Pimenta; *Introdução à Estatística*, 2ª ed., Escolar Editora, 2015.

1 Teste de Hipóteses de Independência do Q-Q

Hipóteses, Estatística de Teste e Decisão

5. Um inspetor de qualidade recolheu uma amostra de 176 produtos alimentares num centro de distribuição. Sabendo que cada produto pode ser proveniente de uma de três fábricas e pode ou não estar contaminado; o inspetor avaliou todos os produtos e obteve os seguintes resultados:

	Fábrica A	**Fábrica B**	**Fábrica C**	**Total**
Contaminado	8	15	11	34
Não contaminado	55	67	20	142
Total	63	82	31	176

Pode-se afirmar que o facto de um produto estar contaminado é independente da sua fábrica de origem, considerando $\alpha = 0{,}01$?

[*Adaptado da fonte: https://www.ime.unicamp.br/~veronica/Coordenadas1s/aula8pr.pdf*]

Formulário

Teste de Independência: $Q = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(N_{ij}-fe_{ij})^2}{fe_{ij}} \overset{a}{\sim} \chi^2_{((r-1)(s-1))}$

Exercício: Teste de Independência do Qui-Quadrado

Dados
N = 176
VOE = 7,024
Valor-p = 0,03
$\alpha = 1\% \Rightarrow 1-\alpha = 99\%$
df = g.l.´s = 2

Hipóteses

H_0: As variáveis são independentes

Versus

H_1: As variáveis não são independentes

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Estatística de teste

$$\sum_i\sum_j\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \overset{apr}{\cap} \chi^2_{(l-1)(c-1)}$$

Tabela da distribuição do Qui-Quadrado

Decisão

Pelo valor crítico: Valor da Estatística de Teste = 7,024 não pertence a RR = $[\chi^2_{0,99;2}; +\infty[= [9,21; +\infty[$

Pelo valor-p: valor-p = 0,030 > 0,01

Não se rejeita H_0 para $\alpha = 1\%$. Assim, não existe evidência estatística para afirmar que as variáveis não são independentes para $\alpha = 1\%$.

Exercício: Teste de Independência do Qui-Quadrado

Estatística de Teste:

$$\sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \overset{apr}{\cap} \chi^2_{(l-1)(c-1)}$$

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Contaminado * Fábrica Tabulação cruzada

			Fábrica			
			Fábrica A	Fábrica B	Fábrica C	Total
Contaminado	Sim	Contagem	8	15	11	34
		Expected Count	12,2	15,8	6,0	34,0
	Não	Contagem	55	67	20	142
		Expected Count	50,8	66,2	25,0	142,0
Total		Contagem	63	82	31	176
		Expected Count	63,0	82,0	31,0	176,0

Valor da Estatística de Teste

Testes de chi-quadrado

	Valor	df	Sig. Assint. (2 lados)
Chi-quadrado de Pearson	7,024[a]	2	,030
Razão de probabilidade	6,450	2	,040
Associação Linear por Linear	6,099	1	,014
N de Casos Válidos	176		

a. 0 células (0,0%) esperam contagem menor do que 5. A contagem mínima esperada é 5,90.

A Condição de Aplicabilidade dos Testes do Qui-Quadrado é satisfeita:

- Todas as frequências esperadas $E_i \geq 5$.

Teste de Independência do Qui-Quadrado

Região de rejeição ou crítica:
7,024 não pertence a RR = $[\chi^2_{0,99;2}; +\infty[$ = $[9,21; +\infty[$

Decisão (para $\alpha = 0,01$)
Pelo valor crítico: VOE = 7,024 > $\chi^2_{0,99;2} = 9,21$

Tabela da Distribuição do Qui-Quadrado

Regra de decisão pelo valor crítico ou região de rejeição (RR):
$\begin{cases} VOE \geq \chi^2_{1-\alpha} \\ VOE \in RR = [\chi^2_{1-\alpha}; +\infty[\end{cases} \Rightarrow$ Rejeita-se H_0 para α

Quantil de probabilidade 1-α da distribuição do Qui-Quadrado

Pelo valor-p: valor-p = 0,03 > 0,01

Regra de decisão pelo valor-p:
Valor-p = $P(X^2 \geq VOE) < \alpha \Rightarrow$ Rejeita-se H_0 para α

Não se rejeita-se H_0 para $\alpha = 0,01$. Assim, não existe evidência estatística para afirmar que as variáveis não são independentes para $\alpha = 1\%$.

Cálculo do Quantil da Distribuição Qui-Quadrado de Probabilidade 1-α e com (l-1)×(c-1) g.l.´s

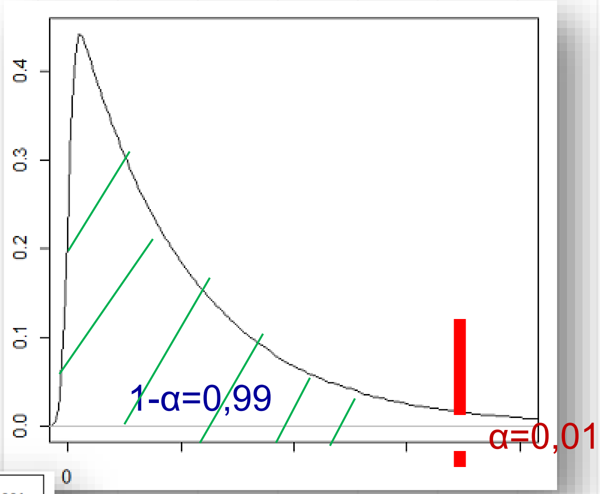
Nível de confiança (1-α=0,99)

Nível de significância (α=0,01)

Área total é igual a 1

O nível de significância é igual a α = 0,01, então tem-se 1-α = 0,99

$\chi^2_{0,99;2}$ = 9,21 (ver tabela)



$\chi^2_{0,99;2} = 9,21$

$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

Regra de decisão pelo valor-p:
Valor-p = $P(X^2 \geq VOE) < \alpha \Rightarrow$ Rejeita-se H_0 para α

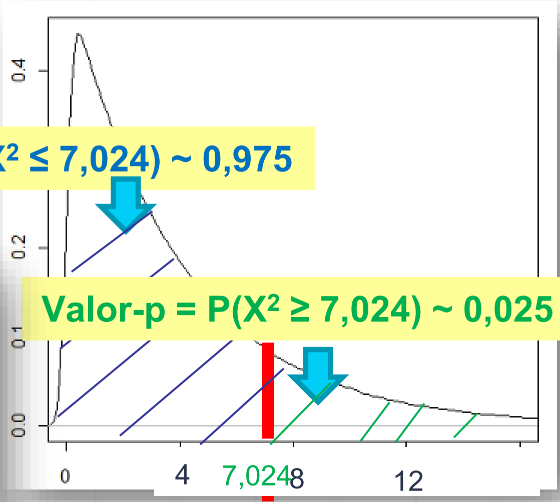
Cálculo do Valor-p quando a Estatística de Teste tem Distribuição Qui-Quadrado

valor-p = $P(X^2 \geq 7{,}024) \sim P(X^2 \geq 7{,}378) = 0{,}025$

$$\chi^2_{n.\varepsilon} : P(X > \chi^2_{n.\varepsilon}) = \varepsilon$$

n \ ε	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005	.001
1	.000	.000	.001	.004	.016	.102	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.827
2	.010	.020	.051	.103	.211	.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.815
3	.072	.115	.216	.352	.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	.207	.297	.484	.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.466
5	.412	.554	.831	1.145	1.610	2.675	4.351	6.626	9.236	11.070	12.832	15.086	16.750	20.515
6	.676	.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.457
7	.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.321
8	1.344	1.647	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588

Área total é igual a 1



$P(X^2 \leq 7{,}024) \sim 0{,}975$

Valor-p = $P(X^2 \geq 7{,}024) \sim 0{,}025$

Exercício: Teste de Independência do Qui-Quadrado

Condições de Aplicabilidade dos Testes do Qui-Quadrado:
- As frequências esperadas devem ser ≥ 5.
- No caso de tal não se verificar, então pelo menos 80% das frequências esperadas ≥ 5 e todas > 1 (não é válido para tabelas 2x2).

Frequências esperadas

$$E_{ij} = \frac{L_i \times C_j}{N}$$

L = total linhas e C = total colunas
N = nº total de elementos

Verificação das condições de aplicabilidade

Neste caso, todas as células têm frequências esperadas superiores a 5.

O teste do Qui-Quadrado apenas informa sobre a independência entre variáveis, mas nada diz sobre o grau de associação existente. Para esse efeito calculam-se **medidas de associação** tais como o coeficiente Phi, o coeficiente V de Cramer e o coeficiente de contingência.

Contaminado * Fábrica Tabulação cruzada

			Fábrica			
			Fábrica A	Fábrica B	Fábrica C	Total
Contaminado	Sim	Contagem	8	15	11	34
		Expected Count	12,2	15,8	6,0	34,0
	Não	Contagem	55	67	20	142
		Expected Count	50,8	66,2	25,0	142,0
Total		Contagem	63	82	31	176
		Expected Count	63,0	82,0	31,0	176,0

Murteira et al (2015)
Capítulo 9

16. Uma companhia de seguros está interessada em saber se existe independência entre a realização de seguros de dois tipos: " Seguro de Vida" e "Seguro de Saúde". Para tal analisou as apólices detidas por uma amostra casual de 200 clientes tendo-se verificado que: 90 tinham simultaneamente um seguro de vida e um seguro de saúde; 35 tinham apenas seguro de vida; 42 tinham apenas seguro de saúde; os restantes não tinham qualquer destes dois seguros. Com base num teste a 0.05, que pode concluir?

Exercício 16

$H_0: \quad P_{ij} = P_{i\bullet} P_{\bullet j} \quad (i, j = 1, 2)$

$H_1: \exists (i, j): P_{ij} \neq P_{i\bullet} P_{\bullet j}$

$$Q = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} \left(N_{ij} - fe_{ij}\right)^2}{fe_{ij}} \overset{a}{\sim} \chi^2_{\{(\overset{\#\text{ linhas}}{2}-1)(\overset{\#\text{ colunas}}{2}-1)\}} = \chi^2(1)$$

Total | 132 68 | 200

$\hat{P}_{\bullet 1} = \frac{132}{200} = 0.66 \qquad \frac{68}{200} = 0.34 = P_{\bullet 2}$

Exercício 16

$$H_0: \quad P_{ij} = P_{i\bullet} P_{\bullet j} \quad (i, j = 1, 2)$$

$$H_1: \exists (i, j): P_{ij} \neq P_{i\bullet} P_{\bullet j}$$

$$Q = \frac{\sum_{i=1}^{2} \sum_{j=1}^{2} (N_{ij} - fe_{ij})^2}{fe_{ij}} \overset{a}{\sim} \chi^2\{(\overset{\#\,\text{linhas}}{2}-1)\,(\overset{\#\,\text{colunas}}{2}-1)\} = \chi^2(1)$$

Exercício 16

Frequencias esperadas sob H_0 (fe_{ij}):

$$fe_{ij} = n\,\hat{P}_{i\bullet}\,\hat{P}_{\bullet j} \quad (i, j = 1, 2)$$

		seguro saúde	
		sim	não
seguro vida	sim	$200 \times 0.625 \times 0.66 = 82.5$	$200 \times 0.625 \times 0.34 = 42.5$
	não	$200 \times 0.375 \times 0.66 = 49.5$	$200 \times 0.375 \times 0.34 = 25.5$

Exercício 16

$$\chi^2_{1,0.05} = 3.841 \qquad W_Q = \{Q_{obs}: Q_{obs} > 3.841\}$$

$$Q_{obs} = \frac{(90-82.5)^2}{82.5} + \frac{(35-42.5)^2}{42.5} + \frac{(42-49.5)^2}{49.5} + \frac{(33-25.5)^2}{25.5} =$$

$$= 5.3476 \in W_Q \rightarrow \text{Rejeitamos } H_0.$$

2

Modelo de Regressão Linear Simples

Interpretação dos Coeficientes da Reta de Regressão

Overview of Linear Models

- An equation can be fit to show the best linear relationship between two variables:

$$\boxed{Y = \beta_0 + \beta_1 X}$$

Where Y is the dependent variable and

X is the independent variable

β_0 is the Y-intercept

β_1 is the slope

Newbold et al (2013)

Least Squares Regression

- Estimates for coefficients β_0 and β_1 are found using a Least Squares Regression technique
- The least-squares regression line, based on sample data, is

$$\boxed{\hat{y} = b_0 + b_1 x}$$

- Where b_1 is the slope of the line and b_0 is the *y*-intercept:

$$\boxed{b_1 = \frac{Cov(x, y)}{s_x^2} = r\left(\frac{s_y}{s_x}\right)} \qquad \boxed{b_0 = \overline{y} - b_1\overline{x}}$$

Newbold et al (2013)

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain
(also called the endogenous variable)

Independent variable: the variable used to explain the dependent variable
(also called the exogenous variable)

Newbold et al (2013)

Linear Regression Model

- The relationship between X and Y is described by a linear function
- Changes in Y are assumed to be influenced by changes in X
- Linear regression population equation model

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

- Where β_0 and β_1 are the population model coefficients and ε is a random error term.

Newbold et al (2013)

Simple Linear Regression Model

The population regression model:

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

$$\boxed{y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}}$$

Newbold et al (2013)

Linear Regression Assumptions

- The true relationship form is linear (*Y* is a linear function of *X*, plus random error)
- The error terms, ε_i are independent of the *x* values
- The error terms are random variables with mean 0 and constant variance, σ^2

  (the uniform variance property is called homoscedasticity)

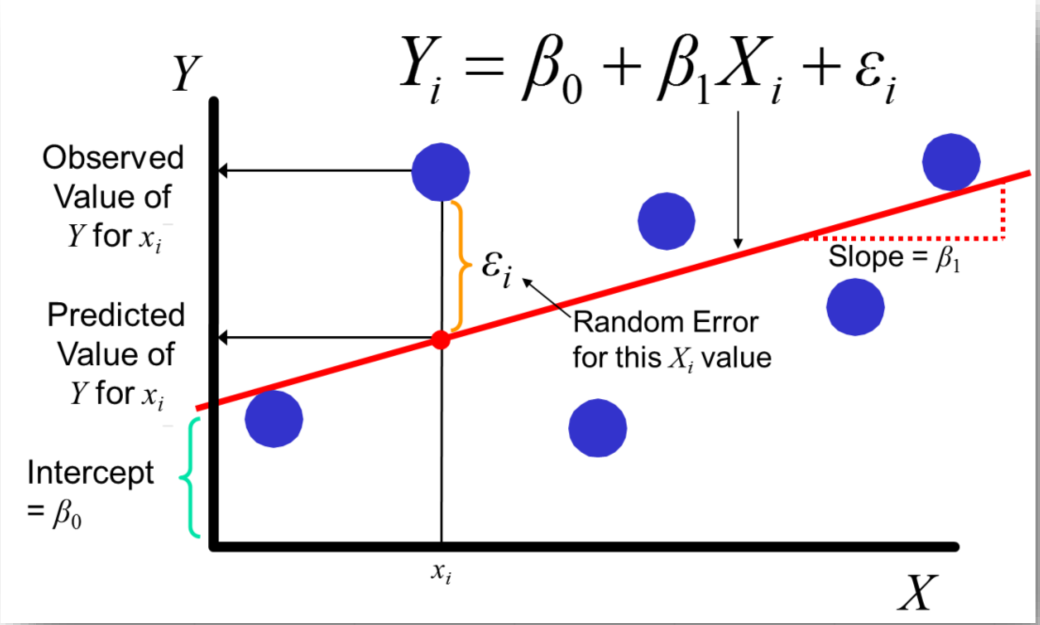
$$E[\varepsilon_i] = 0 \text{ and } E\left[\varepsilon_i^2\right] = \sigma^2 \text{ for } (i = 1, \ldots, n)$$

- The random error terms ε_i, are not correlated with one another, so that

$$E\left[\varepsilon_i \varepsilon_j\right] = 0 \text{ for all } i \neq j$$

Newbold et al (2013)

Simple Linear Regression Model



Newbold et al (2013)

Simple Linear Regression Equation

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) *y* value for observation *i*

Estimate of the regression intercept

Estimate of the regression slope

Value of *x* for observation *i*

$$\boxed{\hat{y}_i = b_0 + b_1 x_i}$$

The individual random error terms e_i have a mean of zero

$$e_i = (y_i - \hat{y}_i) = y_i - (b_0 + b_1 x_i)$$

Newbold et al (2013)

Coefficient Estimators

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals (errors), SSE:

$$\boxed{\begin{aligned}\min \ \text{SSE} &= \min \sum_{i=1}^{n} e_i^2 \\ &= \min \sum (y_i - \hat{y}_i)^2 \\ &= \min \sum [y_i - (b_0 + b_1 x_i)]^2\end{aligned}}$$

Differential calculus is used to obtain the coefficient estimators b_0 and b_1 that minimize SSE

Newbold et al (2013)

Least Squares Coefficient Estimators

- The slope coefficient estimator is

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2} = r\frac{s_y}{s_x}}$$

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

- And the constant or y-intercept is

$$\boxed{b_0 = \bar{y} - b_1\bar{x}}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$

- The regression line always goes through the mean $\bar{x}, \bar{y}$

Newbold et al (2013)

Computer Computation of Regression Coefficients

- The coefficients b_0 and b_1, and other regression results in this chapter, will be found using a computer
  - Hand calculations are tedious
  - Statistical routines are built into Excel
  - Other statistical analysis software can be used

Newbold et al (2013)

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of *y* when the value of *x* is zero (if *x* = 0 is in the range of observed *x* values)

- b_1 is the estimated change in the average value of *y* as a result of a one-unit change in *x*

Newbold et al (2013)

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Newbold et al (2013)

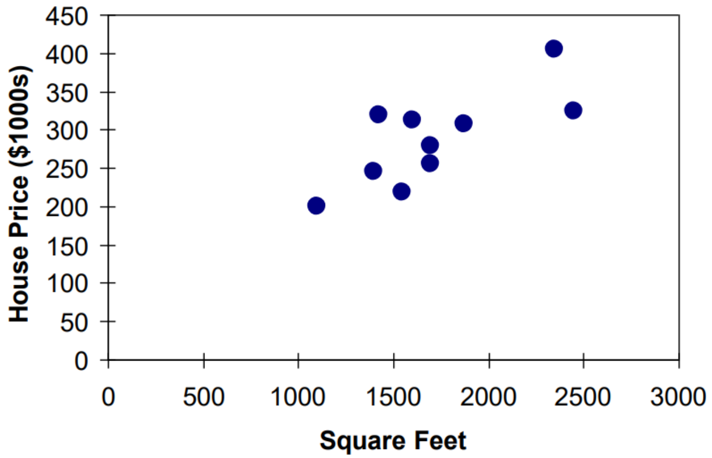
Sample Data for House Price Model

House Price in $1000s ($Y$)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Newbold et al (2013)

Graphical Presentation

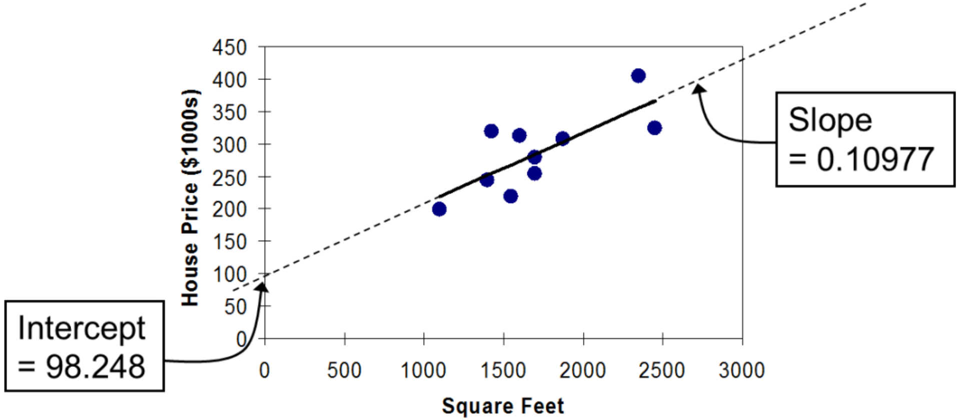
- House price model: scatter plot



Newbold et al (2013)

Graphical Presentation

- House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Newbold et al (2013)

Interpretation of the Intercept, b Sub 0

$$\widehat{\text{house price}} = \boxed{98.24833} + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Newbold et al (2013)

Interpretation of the Slope Coefficient, b Sub 1

$$\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by $.10977(\$1000) = \$109.77$, on average, for each additional one square foot of size

Newbold et al (2013)

Exercise 11.18 A)

11.18 Compute the coefficients for a least squares regression equation and write the equation, given the following sample statistics.

a. $\overline{x} = 50, \overline{y} = 100, s_x = 25, s_y = 75, r_{xy} = 0.6, n = 60$

Newbold et al (2013)

Exercise 11.18 A): Solution

Answer:

Given data

- $\bar{x} = 50$
- $\bar{y} = 100$
- $s_x = 25$
- $s_y = 75$
- $r_{xy} = 0.6$
- $n = 60$

Step 1: Compute the slope coefficient

For a simple linear regression, the slope is:

$$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\beta}_1 = 0.6 \times \frac{75}{25} = 0.6 \times 3 = 1.8$$

Exercise 11.18 A): Solution

✅ Answer:

Step 2: Compute the intercept

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

$$\hat{\beta}_0 = 100 - 1.8 \times 50 = 100 - 90 = 10$$

Step 3: Write the regression equation

$$\hat{Y} = 10 + 1.8X$$

✅ **Final answer**

The least squares regression equation is:

$$\boxed{\hat{Y} = 10 + 1.8X}$$

Exercise 11.20

11.20 For a sample of 20 monthly observations, a financial analyst wants to regress the percentage rate of return (Y) of the common stock of a corporation on the percentage rate of return (X) of the Standard & Poor's 500 index. The following information is available:

$$\sum_{i=1}^{20} y_i = 22.6 \quad \sum_{i=1}^{20} x_i = 25.4 \quad \sum_{i=1}^{20} x_i^2 = 145.7 \quad \sum_{i=1}^{20} x_i y_i = 150.5$$

a. Estimate the linear regression of Y on X.
b. Interpret the slope of the sample regression line.
c. Interpret the intercept of the sample regression line.

Newbold et al (2013)

Exercise 11.20 A): Solution

Answer:

a) Estimation of the linear regression of Y on X

We consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Given:

- $n = 20$
- $\sum y_i = 22.6 \Rightarrow \bar{y} = \frac{22.6}{20} = 1.13$
- $\sum x_i = 25.4 \Rightarrow \bar{x} = \frac{25.4}{20} = 1.27$
- $\sum x_i^2 = 145.7$
- $\sum x_i y_i = 150.5$

Step 1: Compute S_{xy} and S_{xx}

$$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$$

$$S_{xy} = 150.5 - 20(1.27)(1.13) = 150.5 - 28.702 = 121.798$$

$$S_{xx} = \sum x_i^2 - n\bar{x}^2$$

$$S_{xx} = 145.7 - 20(1.27^2) = 145.7 - 32.258 = 113.442$$

Exercise 11.20 A): Solution

Answer:

Step 2: Estimate the regression coefficients

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{121.798}{113.442} \approx 1.074$$

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

$$\hat{\beta}_0 = 1.13 - 1.074(1.27) \approx -0.234$$

Estimated regression line:

$$\hat{Y} = -0.234 + 1.074X$$

Exercise 11.20 B): Solution

Answer:

b) Interpretation of the slope

The slope estimate $\hat{\beta}_1 = 1.074$ indicates that **for each additional 1 percentage point increase in the return of the S&P 500 index, the stock's return is expected to increase on average by approximately 1.07 percentage points**.

This suggests that the stock is more volatile than the market (beta greater than 1).

Exercise 11.20 C): Solution

✅ Answer:

c) Interpretation of the intercept

The intercept estimate $\hat{\beta}_0 = -0.234$ represents the **expected return of the stock when the market return is zero**.

Although this interpretation has limited practical relevance in finance, it serves as a baseline adjustment ensuring the best linear fit of the model.

Exercise 11.21

11.21 A corporation administers an aptitude test to all new sales representatives. Management is interested in the extent to which this test is able to predict sales representatives' eventual success. The accompanying table records average weekly sales (in thousands of dollars) and aptitude test scores for a random sample of eight representatives.

Weekly sales	10	12	28	24	18	16	15	12
Test score	55	60	85	75	80	85	65	60

a. Estimate the linear regression of weekly sales on aptitude test scores.
b. Interpret the estimated slope of the regression line.

Newbold et al (2013)

Exercise 11.21 A): Solution

Answer:

a) Estimation of the linear regression

Step 1: Compute sample means

$$\bar{x} = \frac{55 + 60 + 85 + 75 + 80 + 85 + 65 + 60}{8} = \frac{565}{8} = 70.625$$

$$\bar{y} = \frac{10 + 12 + 28 + 24 + 18 + 16 + 15 + 12}{8} = \frac{135}{8} = 16.875$$

Step 2: Compute S_{xx} and S_{xy}

$$S_{xx} = \sum (x_i - \bar{x})^2 = 906.88$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 301.88$$

Exercise 11.21 A): Solution

✅ Answer:

Step 3: Estimate the regression coefficients

Slope:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{301.88}{906.88} \approx 0.333$$

Intercept:

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

$$\hat{\beta}_0 = 16.875 - 0.333(70.625) \approx -6.66$$

Estimated regression equation

$$\boxed{\hat{Y} = -6.66 + 0.333X}$$

where:

- Y = weekly sales (thousands of dollars)
- X = aptitude test score

Exercise 11.21 B): Solution

Answer:

b) Interpretation of the slope

The estimated slope $\hat{\beta}_1 = 0.333$ indicates that:

> **For each additional one-point increase in the aptitude test score, average weekly sales are expected to increase by approximately 0.33 thousand dollars (about $330), on average.**

This suggests a positive relationship between aptitude test performance and sales success.

Explanatory Power of a Linear Regression Equation

- Total variation is made up of two parts:

$$\boxed{\text{SST} \quad = \quad \text{SSR} \quad + \quad \text{SSE}}$$

Total Sum of Squares | Regression Sum of Squares | Error (residual) Sum of Squares

$$\text{SST} = \sum (y_i - \bar{y})^2 \qquad \text{SSR} = \sum (\hat{y}_i - \bar{y})^2 \qquad \text{SSE} = \sum (y_i - \hat{y}_i)^2$$

where:

$\bar{y}$ = Average value of the dependent variable

y_i = Observed values of the dependent variable

$\hat{y}_i$ = Predicted value of *y* for the given x_i value

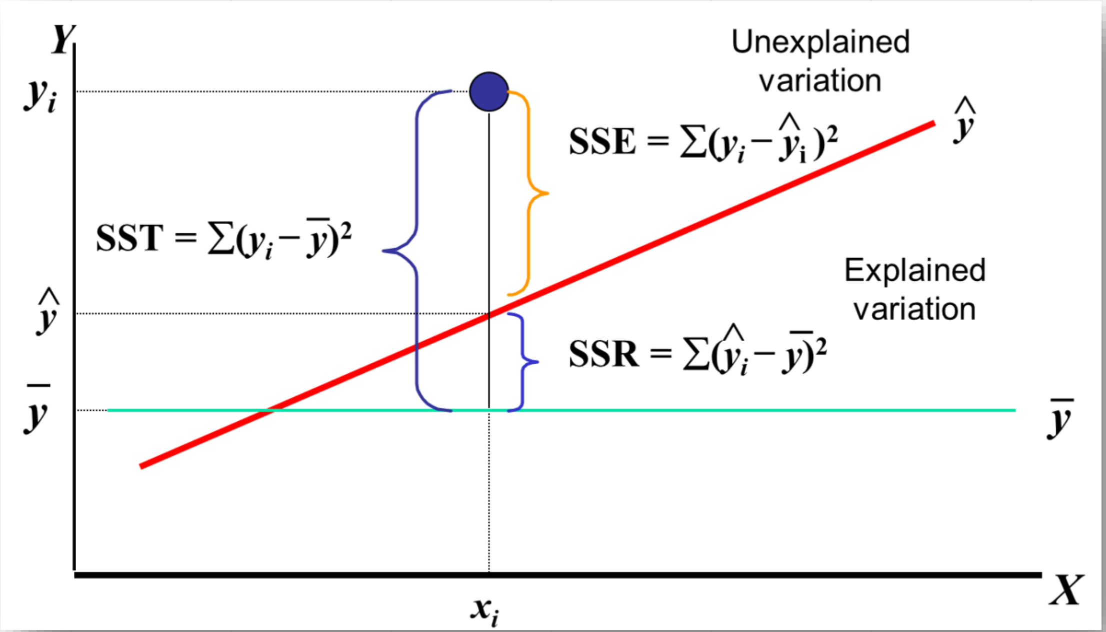
Newbold et al (2013)

Analysis of Variance

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean, $\bar{y}$
- SSR = regression sum of squares
  - Explained variation attributable to the linear relationship between *x* and *y*
- SSE = error sum of squares
  - Variation attributable to factors other than the linear relationship between *x* and *y*

Newbold et al (2013)

Analysis of Variance



Newbold et al (2013)

Coefficient of Determination, R Squared

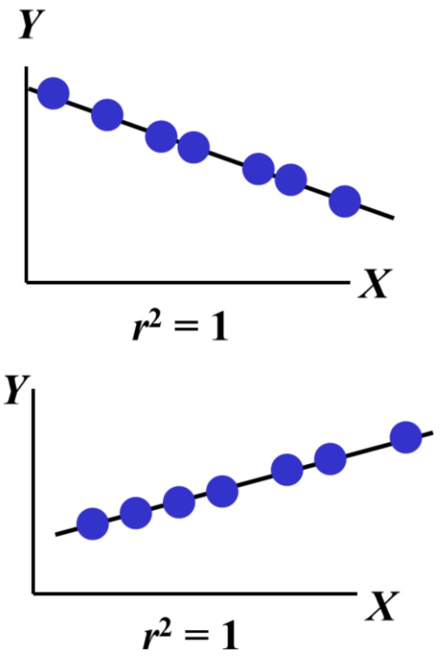
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called *R*-squared and is denoted as R^2

$$\boxed{R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

note: $0 \leq R^2 \leq 1$

Newbold et al (2013)

Examples of Approximate r Squared Values



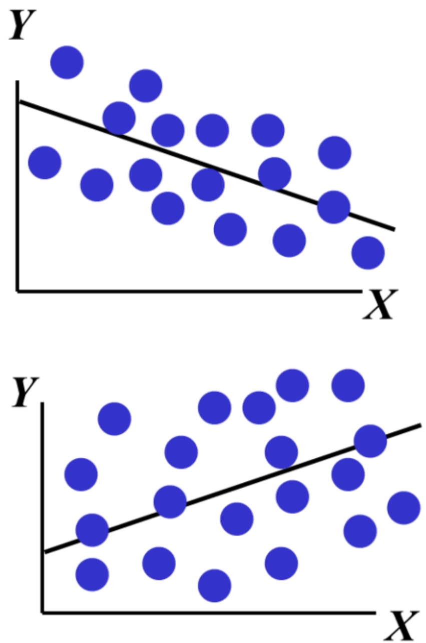
$$r^2 = 1$$

Perfect linear relationship between *X* and *Y*:

100% of the variation in *Y* is explained by variation in *X*

Newbold et al (2013)

Examples of Approximate r Squared Values



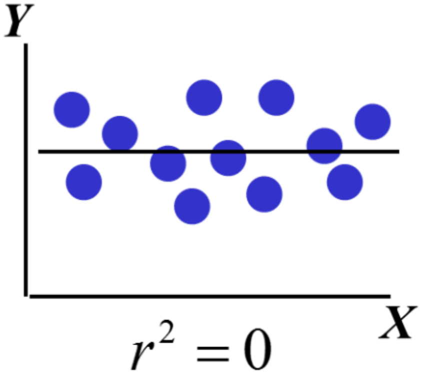
$$0 < r^2 < 1$$

Weaker linear relationships between *X* and *Y*:

Some but not all of the variation in *Y* is explained by variation in *X*

Newbold et al (2013)

Examples of Approximate r Squared Values



$$r^2 = 0$$

No linear relationship between *X* and *Y*:

The value of *Y* does not depend on *X*. (None of the variation in *Y* is explained by variation in *X*)

Newbold et al (2013)

Correlation and R Squared

- The coefficient of determination, R^2, for a simple regression is equal to the simple correlation squared

$$\boxed{R^2 = r^2}$$

Newbold et al (2013)

Exercise 11.22

11.22 In Wanchai Computer Centers in Hong Kong, there are dozens of computer shops selling multiple laptop brands. After a survey in one of them, 10 were selected. The ordered pairs show the speed of each computer's CPU in gigahertz and its price in Hong Kong dollars (1 USD = 7.78 HKD).

(1.8, 14,500), (1.6, 12,290), (2.0, 17,500), (1.6, 16,500), (1.8, 19,650), (2.4, 21,000), (1.2, 7,500), (1.4, 12,500), (1.6, 14,650), (2.0, 18,350)

a. Determinate the regression equation of the sample.
b. Find the intercept and the slope of the equation.
c. Compute the coefficient of determination and interpret its meaning in this specific context.

Newbold et al (2013)

Exercise 11.22 A): Solution

Answer:

Given data

Let

- x = CPU speed (GHz)
- y = Laptop price (HKD)

Sample size: $n = 10$

Sample means:

$$\bar{x} = 11.74, \qquad \bar{y} = 15\,390$$

Formulas for simple linear regression

Slope

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

Exercise 11.22 A): Solution

Answer:

Computation of the coefficients

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 17\,498.6$$

$$\sum(x_i - \bar{x})^2 = 1.6836$$

Thus,

$$b_1 = \frac{17\,498.6}{1.6836} = 10\,391.19$$

$$b_0 = 15\,390 - 10\,391.19 \times 11.74 = -106\,548.54$$

a) Regression equation of the sample

$$\boxed{\hat{y} = -106\,548.54 + 10\,391.19\,x}$$

Exercise 11.22 B): Solution

Answer:

b) Intercept and slope with interpretation

From the regression equation

$$\hat{y} = -106\,548.54 + 10\,391.19\,x$$

we obtain:

Intercept (b_0)

$$b_0 = -106\,548.54$$

Interpretation:
The intercept represents the estimated price of a laptop when the CPU speed is **0 GHz**.
Although this value has **no practical meaning** in this context (since a CPU cannot have zero speed), it is a mathematical consequence of fitting a linear model to the data.

Slope (b_1)

$$b_1 = 10\,391.19$$

Interpretation:
For each additional **1 GHz** increase in CPU speed, the laptop price increases on average by approximately **HKD 10,391**, according to the sample data.

Exercise 11.22 C): Solution

✅ Answer:

c) Coefficient of determination

How R^2 is obtained

The coefficient of determination is defined as:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

where:

- $\mathrm{SST} = \sum(y_i - \bar{y})^2$ is the **total sum of squares**,
- $\mathrm{SSR} = \sum(\hat{y}_i - \bar{y})^2$ is the **regression sum of squares**,
- $\mathrm{SSE} = \sum(y_i - \hat{y}_i)^2$ is the **error sum of squares**.

Equivalently, in simple linear regression,

$$R^2 = r^2$$

where r is the sample correlation coefficient between x and y.

For this data set, the sample correlation is:

$$r \approx 0.880$$

Thus,

$$R^2 = (0.880)^2 = 0.7743$$

Exercise 11.22 C): Solution

✅ Answer:

Interpretation of R^2

$$R^2 = 0.7743$$

This means that approximately **77.4% of the variability in laptop prices** is explained by the linear relationship between CPU speed and price in this sample.

The remaining **22.6%** of the variability is due to other factors not included in the model, such as brand, RAM, storage capacity, graphics card, and design.

Exercise 11.25 A)

11.25 Compute SSR, SSE, s_e^2, and the coefficient of determination, given the following statistics computed from a random sample of pairs of X and Y observations.

a. $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 100{,}000$, $R^2 = 0.50$, $n = 52$

Newbold et al (2013)

Exercise 11.25 a): Solution

✅ Answer:

Given

- Total sum of squares:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 100{,}000$$

- Coefficient of determination:

$$R^2 = 0.50$$

- Sample size:

$$n = 52$$

1) Regression sum of squares (SSR)

By definition,

$$R^2 = \frac{SSR}{SST}$$

So,

$$SSR = R^2 \times SST = 0.50 \times 100{,}000 = 50{,}000$$

Exercise 11.25 A): Solution

Answer:

2) Error sum of squares (SSE)

$$SSE = SST - SSR = 100{,}000 - 50{,}000 = 50{,}000$$

3) Estimator of the error variance s_e^2

In simple linear regression:

$$s_e^2 = \frac{SSE}{n-2}$$

$$s_e^2 = \frac{50{,}000}{52-2} = \frac{50{,}000}{50} = 1{,}000$$

4) Coefficient of determination

This was already given, but for completeness:

$$R^2 = 0.50$$

Exercise 11.28

11.28 Find and interpret the coefficient of determination for the regression of DVD system sales on price, using the following data.

Sales	420	380	350	400	440	380	450	420
Price	98	194	244	207	89	261	149	198

Newbold et al (2013)

Exercise 11.28: Solution

Answer:

Step 1: Compute sample means

$$\bar{x} = \frac{98 + 194 + 244 + 207 + 89 + 261 + 149 + 198}{8} = 180$$

$$\bar{y} = \frac{420 + 380 + 350 + 400 + 440 + 380 + 450 + 420}{8} = 405$$

Step 2: Compute sums of squares

$$S_{xx} = \sum (x_i - \bar{x})^2 = 22{,}088$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = 8{,}200$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = -8{,}180$$

Step 3: Compute the correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-8{,}180}{\sqrt{22{,}088 \times 8{,}200}} \approx -0.608$$

Exercise 11.28: Solution

✅ Answer:

Step 4: Coefficient of determination

$$R^2 = r^2 = (-0.608)^2 \approx 0.37$$

Interpretation

The coefficient of determination $R^2 \approx 0.37$ means that:

> **Approximately 37% of the variation in DVD system sales can be explained by changes in price through the linear regression model.**

The remaining 63% of the variation in sales is due to other factors not included in the model, such as advertising, product features, competition, or random variation.

Estimation of Model Error Variance

- An estimator for the variance of the population model error is

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\text{SSE}}{n-2}$$

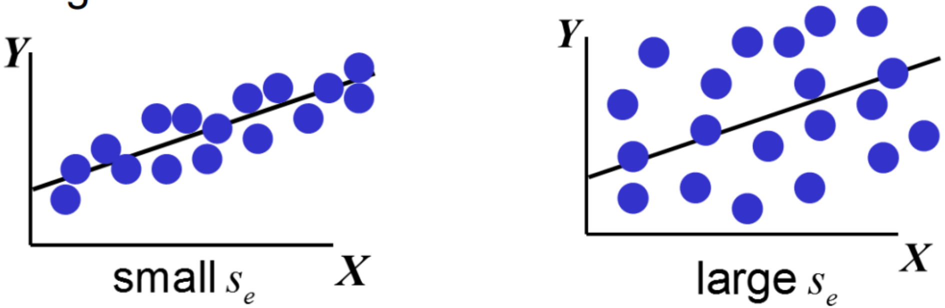
- Division by *n* − 2 instead of *n* − 1 is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$$s_e = \sqrt{s_e^2}$$ is called the standard error of the estimate

Newbold et al (2013)

Comparing Standard Errors

s_e is a measure of the variation of observed *y* values from the regression line



The magnitude of s_e should always be judged relative to the size of the *y* values in the sample data

i.e., $s_e = \$41.33K$ is moderately small relative to house prices in the $\$200 - \$300K$ range

Newbold et al (2013)

Statistical Inference: Hypothesis Tests and Confidence Intervals

- The variance of the regression slope coefficient (b_1) is estimated by

$$s_{b_1}^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2}$$

where:

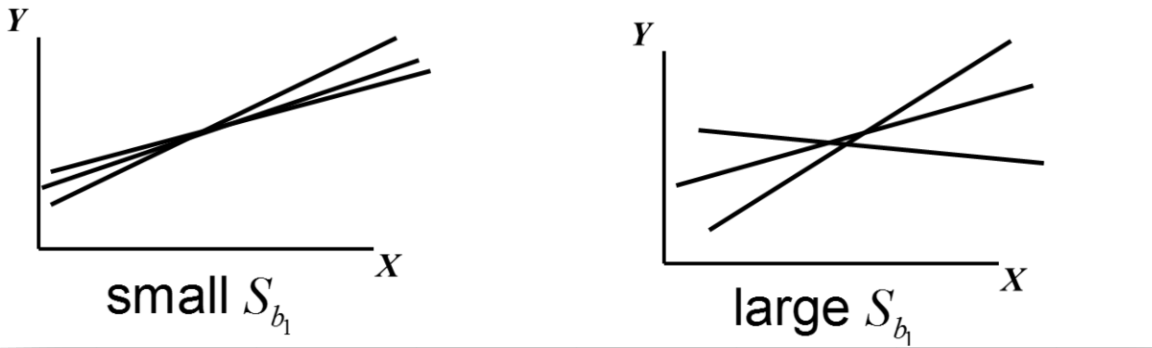
s_{b_1} = Estimate of the standard error of the least squares slope

$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$ = Standard error of the estimate

Newbold et al (2013)

Comparing Standard Errors of the Slope

S_{b_1} is a measure of the variation in the slope of regression lines from different possible samples



Newbold et al (2013)

Inference About the Slope: t Test

- *t* test for a population slope
  - Is there a linear relationship between *X* and *Y*?
- Null and alternative hypotheses

$$H_0 : \beta_1 = 0 \qquad \text{(no linear relationship)}$$
$$H_1 : \beta_1 \neq 0 \qquad \text{(linear relationship does exist)}$$

- Test statistic

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

$$\text{d.f.} = \text{n} - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

s_{b_1} = standard error of the slope

Newbold et al (2013)

Inference About the Slope: t Test

House Price in $1000s ($y$)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house significantly affect its sales price?

Newbold et al (2013)

Inferences About the Slope: t Test Example

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

From Excel output:

b_1 s_{b_1}

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Newbold et al (2013)

Inferences About the Slope: t Test Example

Test Statistic: ***t* = 3.329**

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

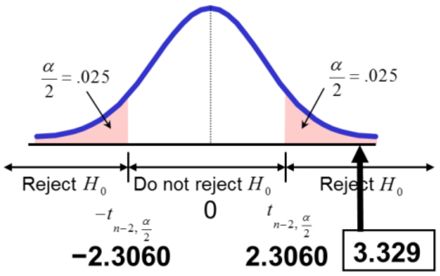
d.f. = 10 − 2 = 8

$t_{8,.025} = 2.3060$

From Excel output:

b_1 S_{b_1} t

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Newbold et al (2013)

Inferences About the Slope: t Test Example

P-value = **0.01039**

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

From Excel output:

P-value

	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept	98.24833	58.03348	1.69296	0.12892
Square Feet	0.10977	0.03297	3.32938	0.01039

This is a two-tail test, so the *p*-value is

$P(t > 3.329) + P(t < -3.329)$

$= 0.01039$

(for 8 d.f.)

Decision: *P*-value < α so Reject H_0

Conclusion:
There is sufficient evidence that square footage affects house price

Newbold et al (2013)

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 - t_{n-2,\frac{\alpha}{2}} s_{b_1} < \beta_1 < b_1 + t_{n-2,\frac{\alpha}{2}} s_{b_1}$$

$$\text{d.f.} = n - 2$$

Excel Printout for House Prices:

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Newbold et al (2013)

Confidence Interval Estimate for the Slope

	Coefficients	Standard Error	*t* Stat	*P*-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Newbold et al (2013)

Hypothesis Test for Population Slope Using the F Distribution

- *F* Test statistic:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where

$$\text{MSR} = \frac{\text{SSR}}{k}$$

$$\text{MSE} = \frac{\text{SSE}}{n-k-1}$$

where *F* follows an *F* distribution with *k* numerator and ($n - k - 1$) denominator degrees of freedom

(k = the number of independent variables in the regression model)

Newbold et al (2013)

Hypothesis Test for Population Slope Using the F Distribution

- An alternate test for the hypothesis that the slope is zero:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Use the *F* statistic

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}}{s_e^2}$$

- The decision rule is

reject H_0 if $F \geq F_{1,n-2,\alpha}$

Newbold et al (2013)

F-Test for Significance

$H_0 : \beta_1 = 0$

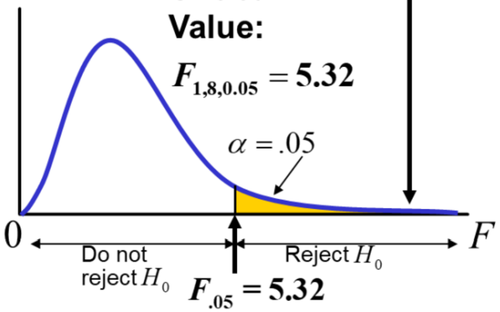
$H_1 : \beta_1 \neq 0$

$\alpha = .05$

$\underbrace{df_1 = 1 \quad df_2 = 8}$

Critical Value:

$F_{1,8,0.05} = 5.32$



Test Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Newbold et al (2013)

Exercise 11.32 A)

11.32 Given the simple regression model

$$Y = \beta_0 + \beta_1 X$$

and the regression results that follow, test the null hypothesis that the slope coefficient is 0 versus the alternative hypothesis of greater than zero using probability of Type I error equal to 0.05, and determine the two-sided 95% and 99% confidence intervals.

a. A random sample of size $n = 38$ with $b_1 = 5 \quad s_{b_1} = 2.1$

Newbold et al (2013)

Exercise 11.32 B)

11.32 Given the simple regression model

$$Y = \beta_0 + \beta_1 X$$

and the regression results that follow, test the null hypothesis that the slope coefficient is 0 versus the alternative hypothesis of greater than zero using probability of Type I error equal to 0.05, and determine the two-sided 95% and 99% confidence intervals.

b. A random sample of size $n = 46$ with $b_1 = 5.2 \quad s_{b_1} = 2.1$

Newbold et al (2013)

Exercise 11.34

11.34 Mumbai Electronics is planning to extend its marketing region from the western United States to include the midwestern states. In order to predict its sales in this new region, the company has asked you to develop a linear regression of DVD system sales on price, using the following data supplied by the marketing department:

Sales	418	384	343	407	432	386	444	427
Price	98	194	231	207	89	255	149	195

a. Use an unbiased estimation procedure to find an estimate of the variance of the error terms in the population regression.
b. Use an unbiased estimation procedure to find an estimate of the variance of the least squares estimator of the slope of the population regression line.
c. Find a 90% confidence interval for the slope of the population regression line.

Newbold et al (2013)

Exercise 11.34: Solution

Answer:

Step 1: Compute sample means

$$\bar{x} = \frac{98 + 194 + 231 + 207 + 89 + 255 + 149 + 195}{8} = 177.25$$

$$\bar{y} = \frac{418 + 384 + 343 + 407 + 432 + 386 + 444 + 427}{8} = 405.13$$

Step 2: Compute sums of squares

$$S_{xx} = \sum (x_i - \bar{x})^2 = 21{,}584$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = -7{,}503$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = 8{,}468$$

Step 3: Estimate the regression coefficients

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-7{,}503}{21{,}584} \approx -0.347$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 405.13 + 0.347(177.25) \approx 466.7$$

Regression line:

$$\hat{Y} = 466.7 - 0.347X$$

Exercise 11.34 A): Solution

Answer:

a) Unbiased estimate of the variance of the error terms

First compute:

$$SSR = \frac{S_{xy}^2}{S_{xx}} = \frac{(-7{,}503)^2}{21{,}584} \approx 2{,}609$$

$$SSE = S_{yy} - SSR = 8{,}468 - 2{,}609 = 5{,}859$$

The unbiased estimator of the error variance is:

$$\boxed{\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{5{,}859}{6} \approx 976.5}$$

Exercise 11.34 B): Solution

✅ Answer:

b) Unbiased estimate of the variance of the slope estimator

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}$$

$$\boxed{\mathrm{Var}(\hat{\beta}_1) = \frac{976.5}{21{,}584} \approx 0.0453}$$

Standard error of the slope:

$$s_{\hat{\beta}_1} = \sqrt{0.0453} \approx 0.213$$

Exercise 11.34 C): Solution

✅ Answer:

c) 90% confidence interval for the slope

Degrees of freedom:

$$df = n - 2 = 6$$

Critical value:

$$t_{0.95,6} \approx 1.943$$

Confidence interval:

$$\hat{\beta}_1 \pm t \cdot s_{\hat{\beta}_1} = -0.347 \pm 1.943(0.213)$$

$$= -0.347 \pm 0.414$$

$$\boxed{(-0.761,\ 0.067)}$$

Exercise 11.35

11.35 A fast-food chain decided to carry out an experiment to assess the influence of advertising expenditure on sales. Different relative changes in advertising expenditure, compared to the previous year, were made in eight regions of the country, and resulting changes in sales levels were observed. The accompanying table shows the results.

Increase in advertising expenditure (%)	0	4	14	10	9	8	6	1
Increase in sales (%)	2.4	7.2	10.3	9.1	10.2	4.1	7.6	3.5

a. Estimate by least squares the linear regression of increase in sales on increase in advertising expenditure.

Newbold et al (2013)

Exercise 11.35 A): Solution

✅ Answer:

(a) Least squares regression of sales on advertising

Step 1: Compute sample means

$$\bar{X} = \frac{0+4+14+10+9+8+6+1}{8} = \frac{52}{8} = 6.5$$

$$\bar{Y} = \frac{2.4+7.2+10.3+9.1+10.2+4.1+7.6+3.5}{8} = \frac{54.4}{8} = 6.8$$

Step 2: Compute sums of squares

$$S_{xx} = \sum (X_i - \bar{X})^2 = 146$$

$$S_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 80.6$$

Step 3: Estimate slope and intercept

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{80.6}{146} \approx 0.552$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 6.8 - 0.552(6.5) \approx 3.21$$

✅ Estimated regression line

$$\boxed{\hat{Y} = 3.21 + 0.552X}$$

Interpretation:

An increase of 1% in advertising expenditure is associated with an average increase of about **0.55% in sales**.

Exercise 11.35 B): Solution

Answer:

(b) 90% confidence interval for the slope

Step 1: Compute residual variance

$$S_{yy} = \sum (Y_i - \bar{Y})^2 = 64.64$$

$$\mathrm{SSE} = S_{yy} - \hat{\beta}_1 S_{xy} = 64.64 - (0.552)(80.6) \approx 20.15$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{20.15}{6} \approx 3.36$$

Step 2: Standard error of the slope

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}} = \sqrt{\frac{3.36}{146}} \approx 0.152$$

Step 3: Critical value

Degrees of freedom:

$$df = n - 2 = 6$$

For a **90% confidence interval**:

$$t_{0.05,6} \approx 1.943$$

Exercise 11.35 B): Solution

✅ Answer:

Step 4: Confidence interval

$$\hat{\beta}_1 \pm t \cdot \mathrm{SE} = 0.552 \pm 1.943(0.152)$$
$$= 0.552 \pm 0.295$$

✅ **90% confidence interval for the slope**

$$\boxed{(0.26,\ 0.85)}$$

Prediction

- The regression equation can be used to predict a value for *y*, given a particular *x*
- For a specified value, x_{n+1}, the predicted value is

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Newbold et al (2013)

Predictions Using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098\ (\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is $317.85(\$1{,}000\text{s}) = \$317{,}850$

Newbold et al (2013)

Obrigada!

Questões?