STATISTICS II

Bachelor's degrees in Economics, Finance and Management

2nd year/2nd Semester
2025/2026

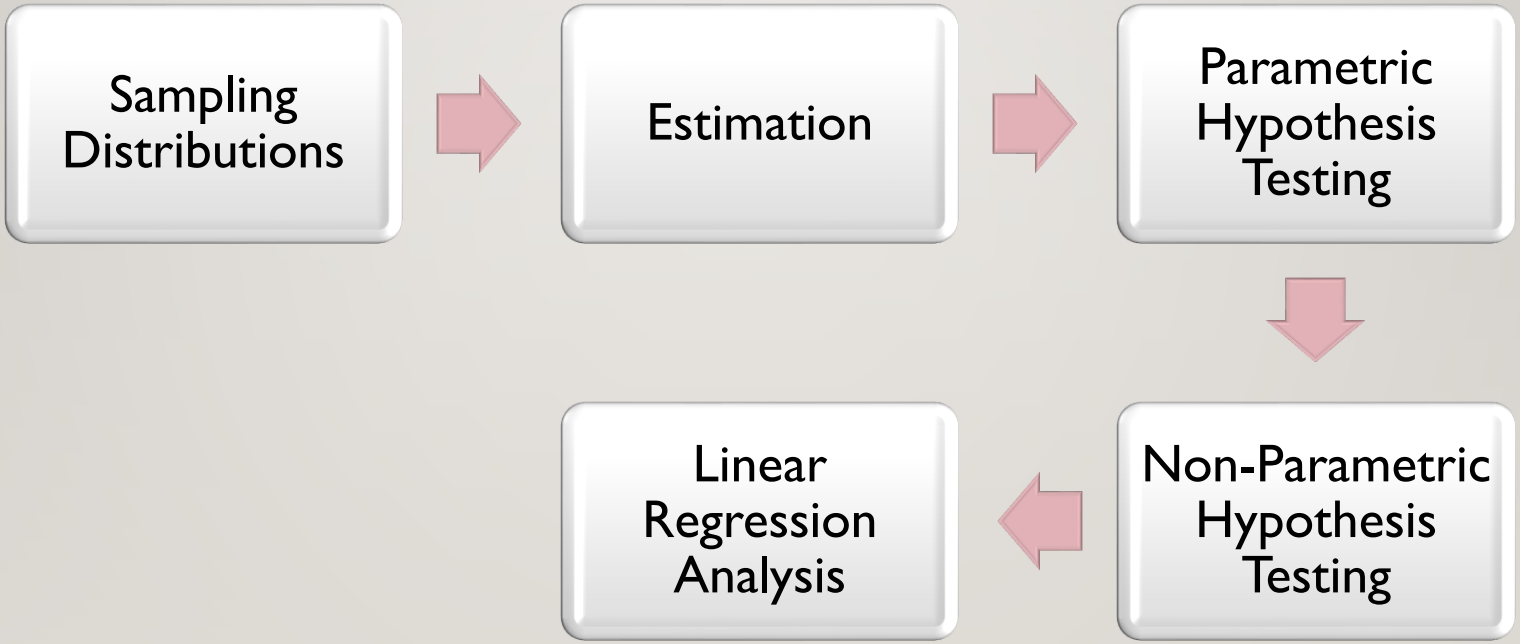
CONTACT

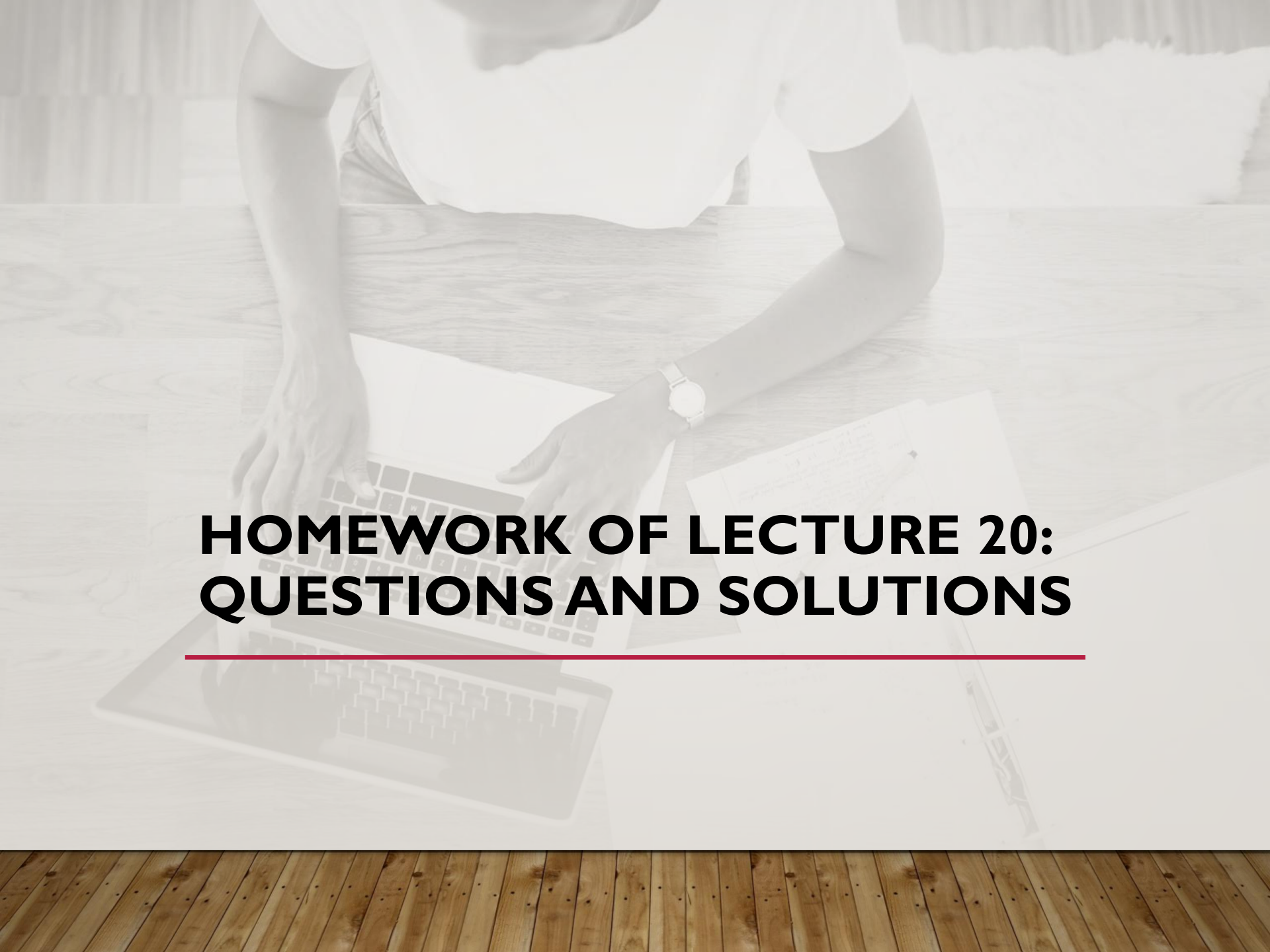
Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

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

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

PROGRAM


Sampling Distributions
Estimation
Parametric Hypothesis Testing
Non-Parametric Hypothesis Testing
Linear Regression Analysis



HOMEWORK OF LECTURE 20: QUESTIONS AND SOLUTIONS

EXERCISE 15.2

15.2 Given the following analysis of variance table, compute mean squares for between groups and within groups. Compute the F ratio and test the hypothesis that the group means are equal.

Source of Variation	Sum of Squares	Degrees of Freedom
Between groups	879	3
Within groups	798	16
Total	1,677	19

Newbold et al (2013)

EXERCISE 15.2: SOLUTION

✅ Answer:

Source of Variation	Sum of Squares	Degrees of Freedom
Between groups	879	3
Within groups	798	16
Total	1,677	19

1. Compute the Mean Squares

Mean Square Between Groups (MSB)

$$MS_B = \frac{SS_B}{df_B} = \frac{879}{3} = 293$$

Mean Square Within Groups (MSW)

$$MS_W = \frac{SS_W}{df_W} = \frac{798}{16} = 49.875$$

2. Compute the F ratio

$$F = \frac{MS_B}{MS_W} = \frac{293}{49.875} \approx 5.88$$

EXERCISE 15.2: SOLUTION

✅ Answer:

3. Hypothesis Test

Hypotheses

- H_0: All group means are equal
- H_1: At least one group mean is different

Test Statistic

$$F \approx 5.88$$

with:

- numerator degrees of freedom: $df_1 = 3$
- denominator degrees of freedom: $df_2 = 16$

Decision Rule

At the 5% significance level ($\alpha = 0.05$):

$F_{0.95;\,3,16} \approx 3.24$

Conclusion

RR = [3.24; +∞[

Since:

$$5.88 > 3.24,$$

we **reject** H_0.

4. Final Conclusion

There is **statistically significant evidence** to conclude that **the group means are not all equal.**

EXERCISE 14.2

14.2 A 2008 survey investigated favorite water sports in Australia, and it found out that 45% of the interviewees voted for surfing, 40% voted for scuba diving, and the rest voted for other water sports. In 2011, a similar survey was conducted; out of a sample of 200 respondents, 102 declared they prefer surfing, 82 chose scuba diving, and the remaining 16 selected other water sports. Is it possible to conclude at the 5% level that in 2011 these preferences remained the same?

Newbold et al (2013)

EXERCISE 14.2: SOLUTION

✅ Answer:

Step 1: Hypotheses + Observed vs. Expected Frequencies

$$H_0 : p_{\text{surfing}} = 0.45,\ p_{\text{scuba}} = 0.40,\ p_{\text{other}} = 0.15$$

$$H_a : \text{Preferences in 2011 differ from 2008}$$

Sample size: $n = 200$

Water Sport	Observed O_i	Expected $E_i = n \cdot p_i$
Surfing	102	200 × 0.45 = 90
Scuba	82	200 × 0.40 = 80
Other	16	200 × 0.15 = 30

EXERCISE 14.2: SOLUTION

✅ Answer:

Step 2: Test Statistic

$$q_0 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(102-90)^2}{90} + \frac{(82-80)^2}{80} + \frac{(16-30)^2}{30}$$

Compute each term:

$$\frac{(102-90)^2}{90} = \frac{12^2}{90} = \frac{144}{90} \approx 1.60$$

$$\frac{(82-80)^2}{80} = \frac{2^2}{80} = \frac{4}{80} = 0.05$$

$$\frac{(16-30)^2}{30} = \frac{(-14)^2}{30} = \frac{196}{30} \approx 6.53$$

$$q_0 \approx 1.60 + 0.05 + 6.53 = 8.18$$

EXERCISE 14.2: SOLUTION

✅ Answer:

Step 3: Rejection Region

- Degrees of freedom: $df = k - 1 = 3 - 1 = 2$
- Significance level: $\alpha = 0.05$
- Critical value: $\chi^2_{0.05,2} \approx 5.991$
- **Reject H_0 if** $q_0 > 5.991$

Step 4: P-value

$$p = P(Q_0 > 8.18) \approx 0.017$$

Step 5: Conclusion

- $q_0 = 8.18 > 5.991 \Rightarrow$ in the rejection region
- P-value $p = 0.017 < 0.05$

RR = [5.991; +∞[

P-value = 0.017

Decision: Reject H_0

Interpretation (slide-ready): There is **statistically significant evidence** that water sports preferences in 2011 **differ from those in 2008**.

EXERCISE 14.18

14.18 University administrators have collected the following information concerning student grade point average and the school of the student's major.

Determine if there is any association between GPA and major.

School	GPA < 3.0	GPA 3.0 or Higher
Arts and Sciences	50	35
Business	45	30
Music	15	25

Newbold et al (2013)

EXERCISE 14.18: SOLUTION

✅ Answer:

Chi-Square Test of Independence

Step 1: Hypotheses + Observed vs. Expected Frequencies

H_0 : GPA is independent of school/major $vs.$ H_a : GPA depends on school/major

Observed table O_{ij}:

School	GPA < 3.0	GPA ≥ 3.0	Row Total
Arts and Sciences	50	35	85
Business	45	30	75
Music	15	25	40
Column Total	110	90	200

EXERCISE 14.18: SOLUTION

✅ Answer:

Expected frequencies $E_{ij} = \frac{(\text{row total})(\text{column total})}{n}$:

$$E_{Arts,<3} = \frac{85 \cdot 110}{200} = 46.75, \quad E_{Arts,\geq 3} = \frac{85 \cdot 90}{200} = 38.25$$

$$E_{Business,<3} = \frac{75 \cdot 110}{200} = 41.25, \quad E_{Business,\geq 3} = \frac{75 \cdot 90}{200} = 33.75$$

$$E_{Music,<3} = \frac{40 \cdot 110}{200} = 22, \quad E_{Music,\geq 3} = \frac{40 \cdot 90}{200} = 18$$

School	Observed O_{ij}	Expected E_{ij}
Arts and Sciences	50, 35	46.75, 38.25
Business	45, 30	41.25, 33.75
Music	15, 25	22, 18

EXERCISE 14.18: SOLUTION

✅ Answer:

Step 2: Test Statistic

$$Q_0 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Compute each term:

- Arts, <3: $(50 - 46.75)^2/46.75 = 10.56/46.75 \approx 0.226$
- Arts, ≥3: $(35 - 38.25)^2/38.25 = 10.56/38.25 \approx 0.276$
- Business, <3: $(45 - 41.25)^2/41.25 = 14.06/41.25 \approx 0.341$
- Business, ≥3: $(30 - 33.75)^2/33.75 = 14.06/33.75 \approx 0.417$
- Music, <3: $(15 - 22)^2/22 = 49/22 \approx 2.227$
- Music, ≥3: $(25 - 18)^2/18 = 49/18 \approx 2.722$

$$q_0 \approx 0.226 + 0.276 + 0.341 + 0.417 + 2.227 + 2.722 = 6.209$$

EXERCISE 14.18: SOLUTION

Answer:

Step 3: Rejection Region

- Degrees of freedom: $df = (r-1)(c-1) = (3-1)(2-1) = 2$
- Significance level: $\alpha = 0.05$
- Critical value: $\chi^2_{0.05,2} \approx 5.991$
- **Reject** H_0 i $q_0 > 5.991$

Step 4: P-value

$$p = P(\ Q_0 > 6.209\ ) \approx 0.045$$

Step 5: Conclusion

- $q_0 = 6.209 > 5.991 \Rightarrow$ in rejection region
- P-value $p = 0.045 < 0.05$

RR = [5.991; +∞[

P-value = 0.045

Decision: Reject H_0

Interpretation: There is **statistically significant evidence** of an association between GPA and school/major.

LECTURE 21: SIMPLE REGRESSION

OVERVIEW OF LINEAR MODELS

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

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

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)

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

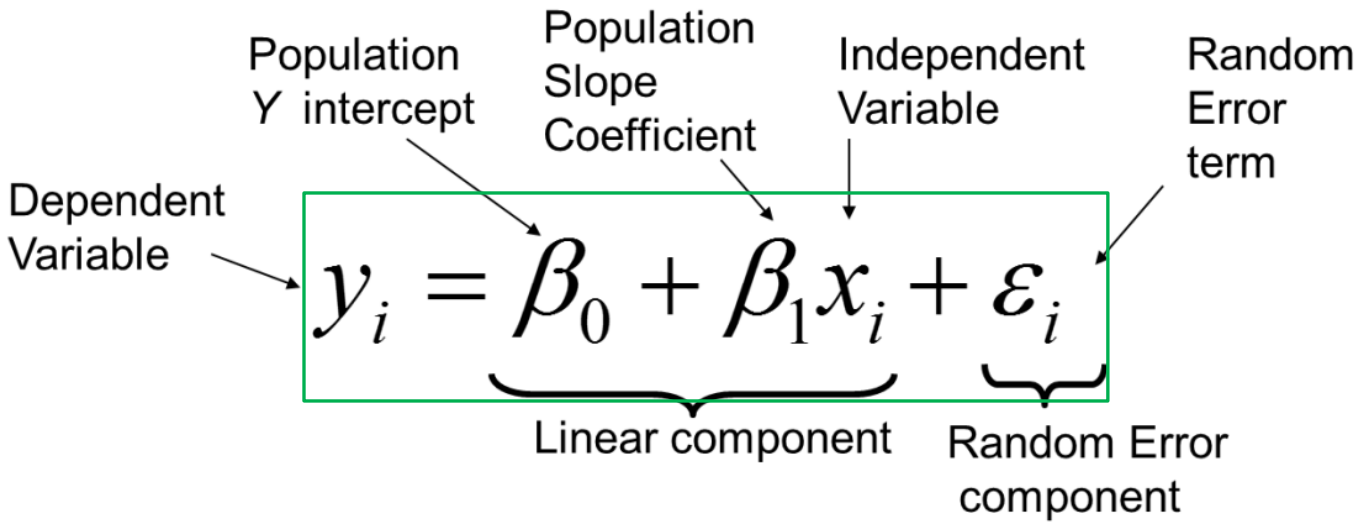
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 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:



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)

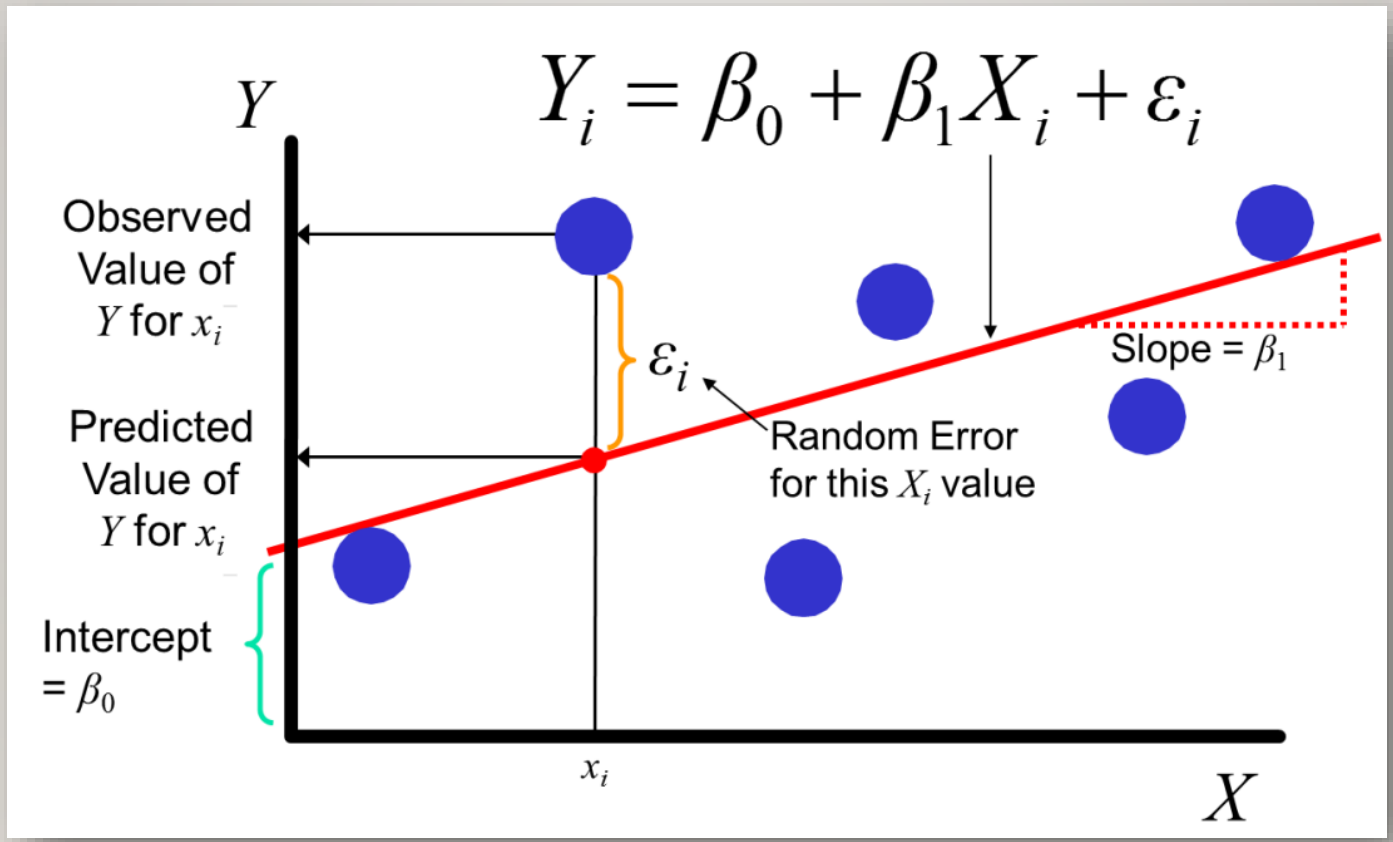
$$E[\varepsilon_i] = 0 \text{ and } E[\varepsilon_i^2] = \sigma^2 \text{ for } (i = 1, \ldots, n)$$

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

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

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*

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

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:

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

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

Newbold et al (2013)

LEAST SQUARES 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

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

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

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

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

Newbold et al (2013)

LEAST SQUARES COEFFICIENT ESTIMATORS

- The slope coefficient estimator is

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

r is the Pearson Correlation Coefficient.

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

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

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

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

- And the constant or *y*-intercept is

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

- 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)

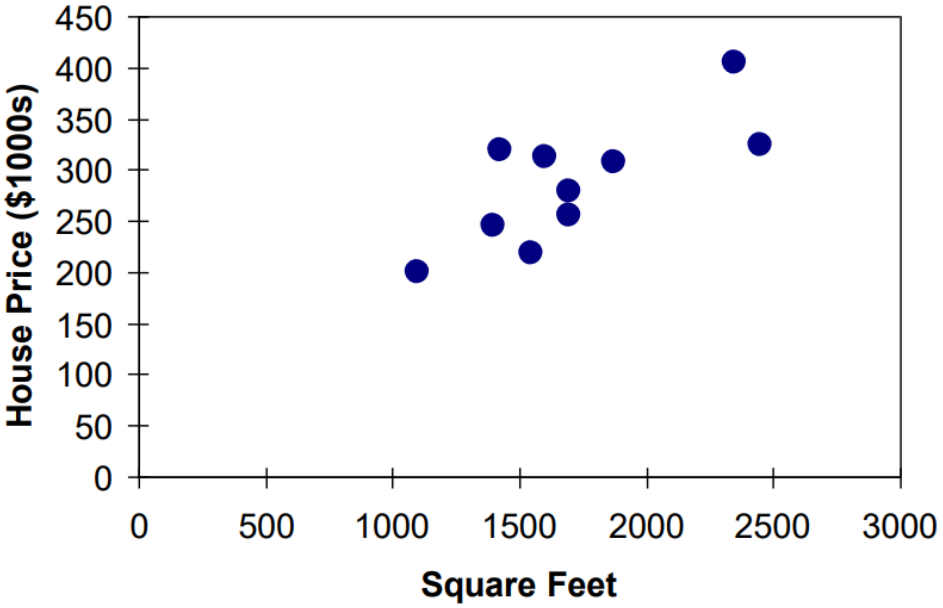
SAMPLE DATA FOR HOUSE PRICE MODEL

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

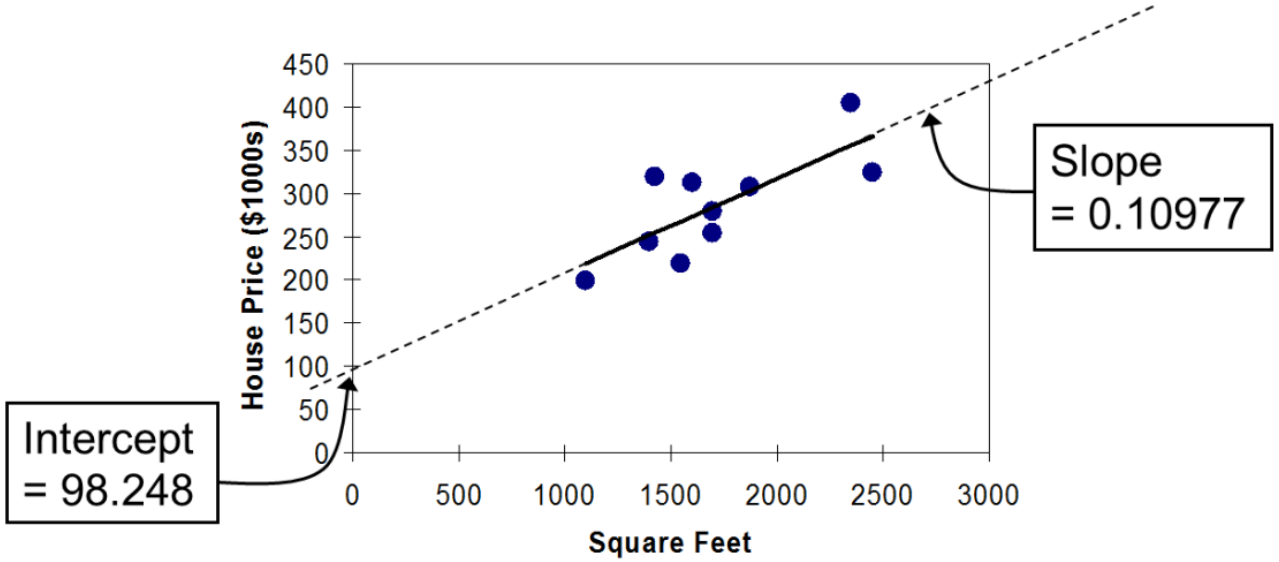
- House price model: scatter plot



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

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.

EXPLANATORY POWER OF A LINEAR REGRESSION EQUATION

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

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

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

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

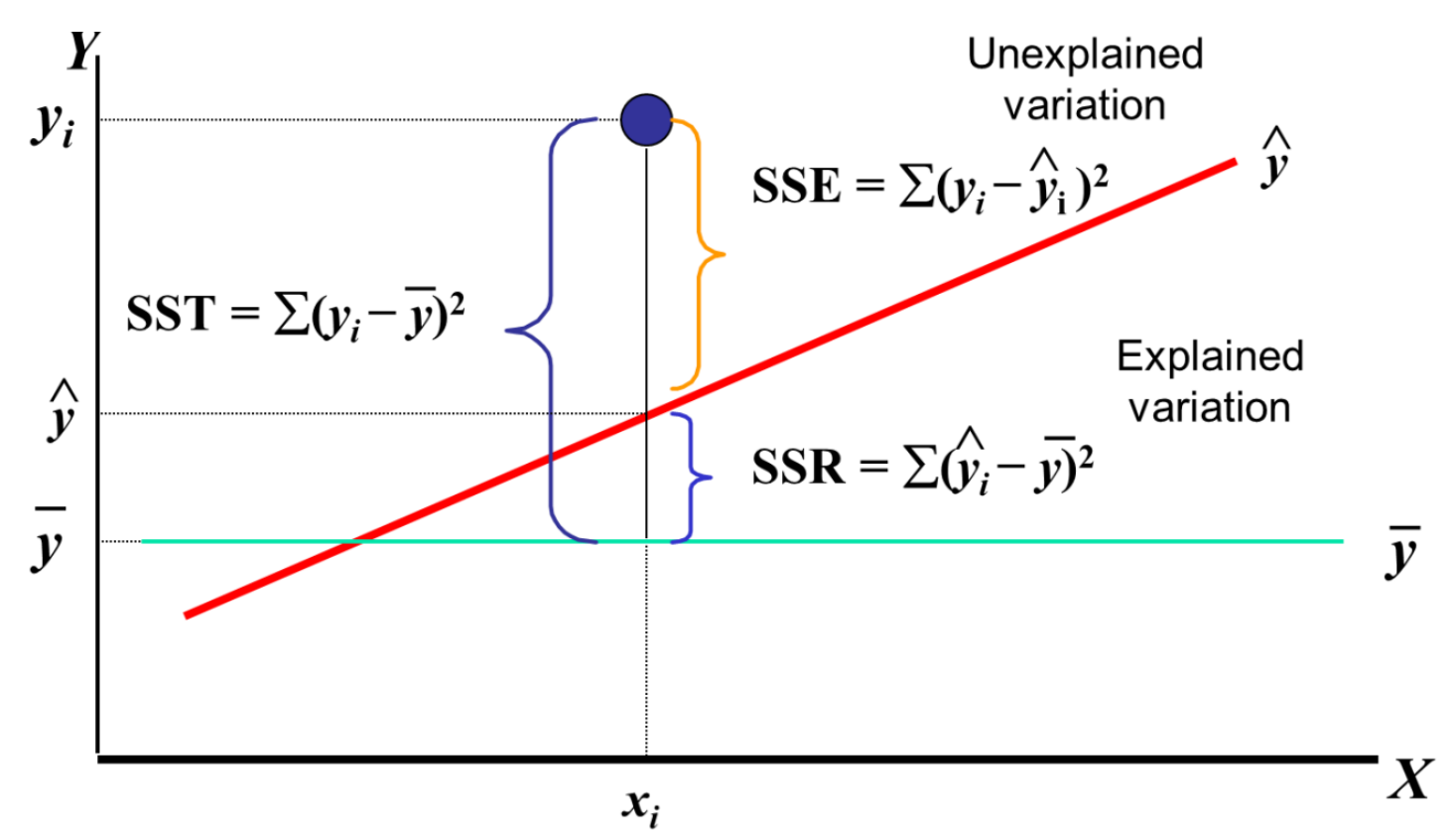
Newbold et al (2013)

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

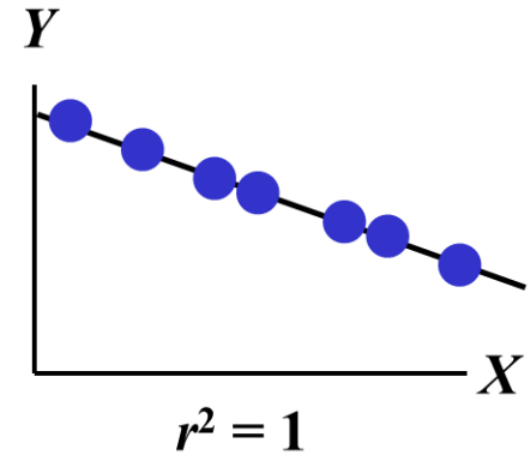
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

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

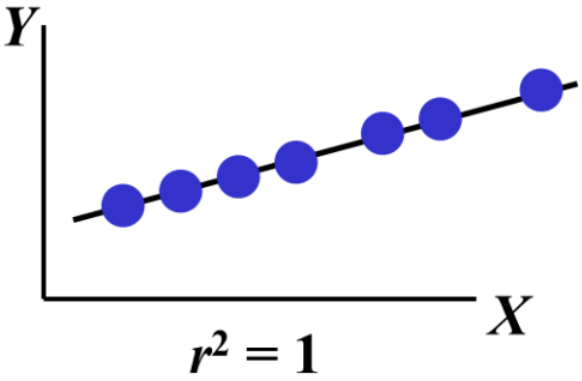
note: $0 \leq R^2 \leq 1$

Newbold et al (2013)

EXAMPLES OF APPROXIMATE R SQUARED VALUES



$r^2 = 1$



$r^2 = 1$

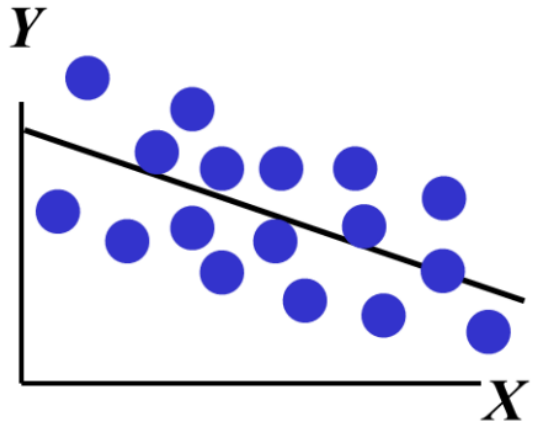
$r^2 = 1$

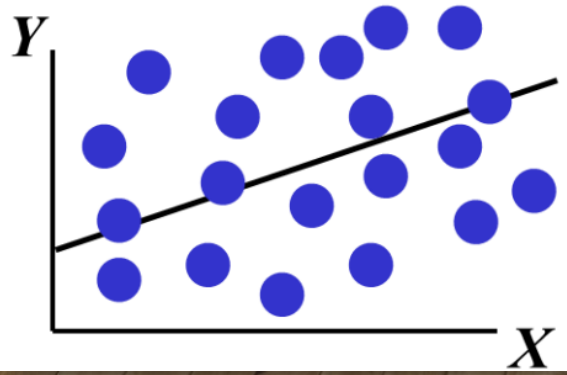
Perfect linear relationship between *X* and *Y*:

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

Newbold et al (2013)

EXAMPLES OF APPROXIMATE R SQUARED VALUES





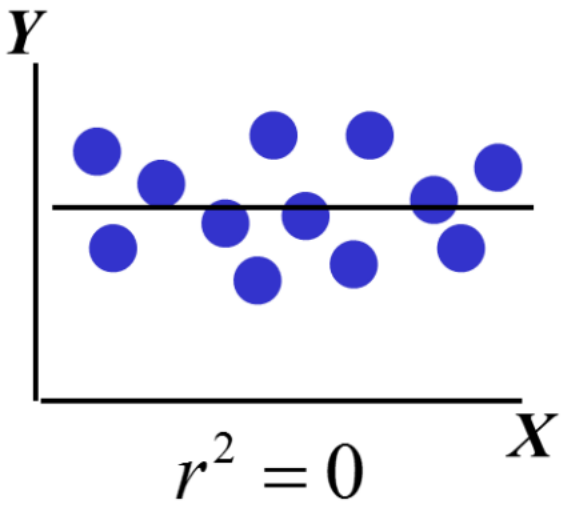
$$0 < r^2 < 1$$

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

$$R^2 = r^2$$

Newbold et al (2013)

Pearson correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Explanation (in English)

- r: Pearson correlation coefficient
- x_i, y_i: individual observations
- $\bar{x}, \bar{y}$: sample means of X and Y
- n: sample size

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:

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

where:

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

ESTIMATION OF MODEL ERROR VARIANCE

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

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

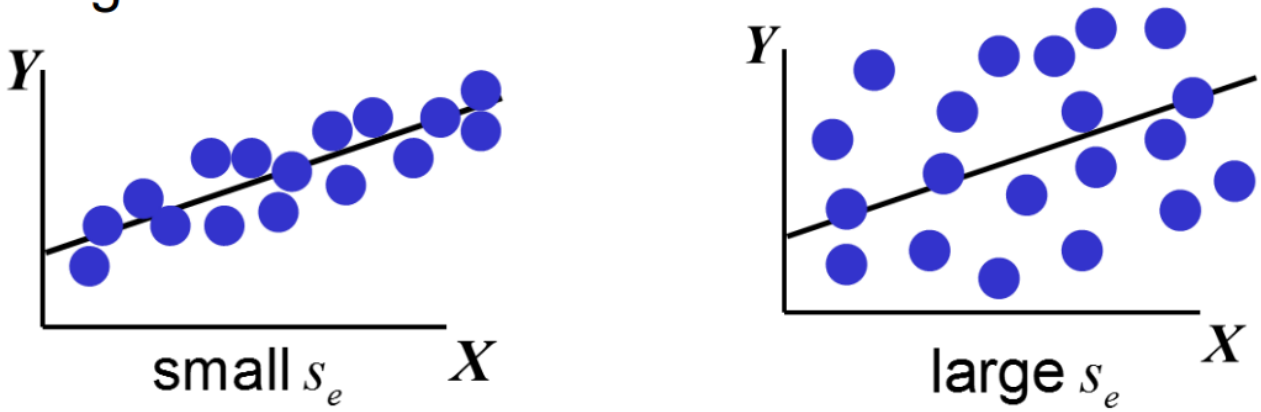
- Division by $n - 2$ instead of $n - 1$ is because the simple regression model uses two estimated parameters, b_0 and b_1, instead of one

$s_e = \sqrt{s_e^2}$ 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:

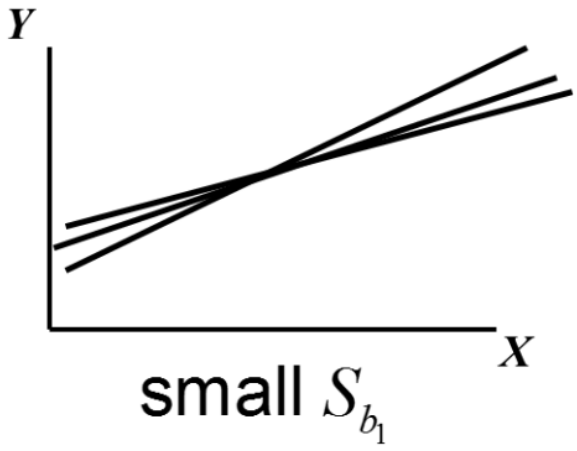
s_{b_1} = Estimate of the standard error of the least squares slope

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

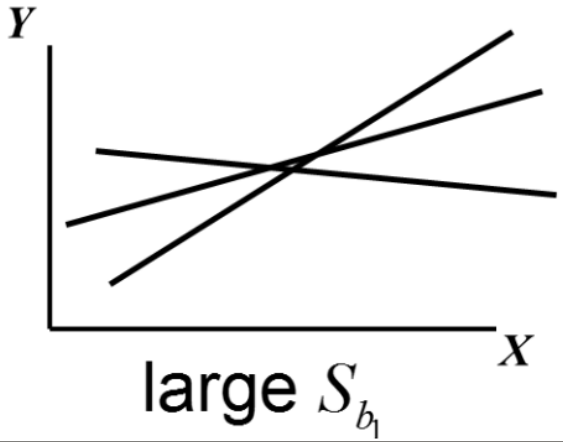
Newbold et al (2013)

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



small S_{b_1}



large S_{b_1}

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

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

- Test statistic

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

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

d.f. = 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: $t = 3.329$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

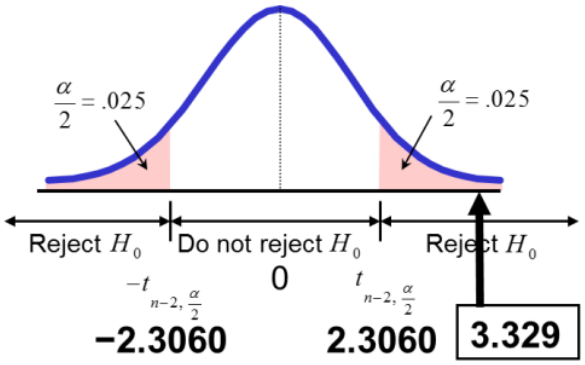
d.f. = 10 − 2 = 8

$t_{8,.025} = 2.3060$

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



RR =] -∞; -2.3060]U[2.3060; +∞[

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.)

P-value = 2 × P(T > 3.329) = 0.01039

Decision: *P*-value $< \alpha$ 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:

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

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

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

Note: The test statistic F follows an F distribution with 1 and $n - 2$ degrees of freedom.

$$F \sim F(1, n-2)$$

- The decision rule is

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

Newbold et al (2013)

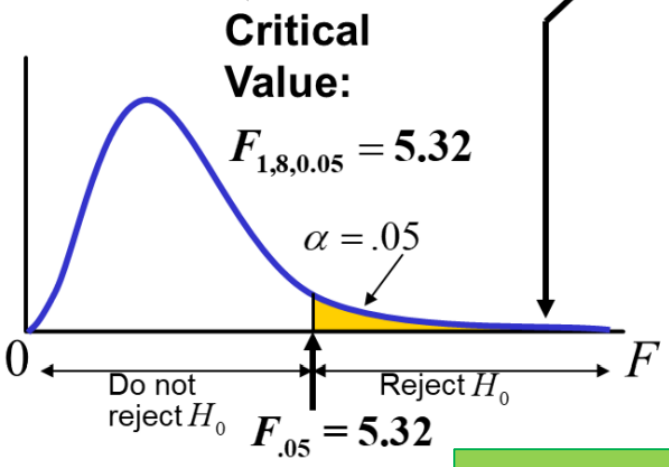
F-TEST FOR SIGNIFICANCE

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$\alpha = .05$

$df_1 = 1 \quad df_2 = 8$



Test Statistic:

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

Decision:

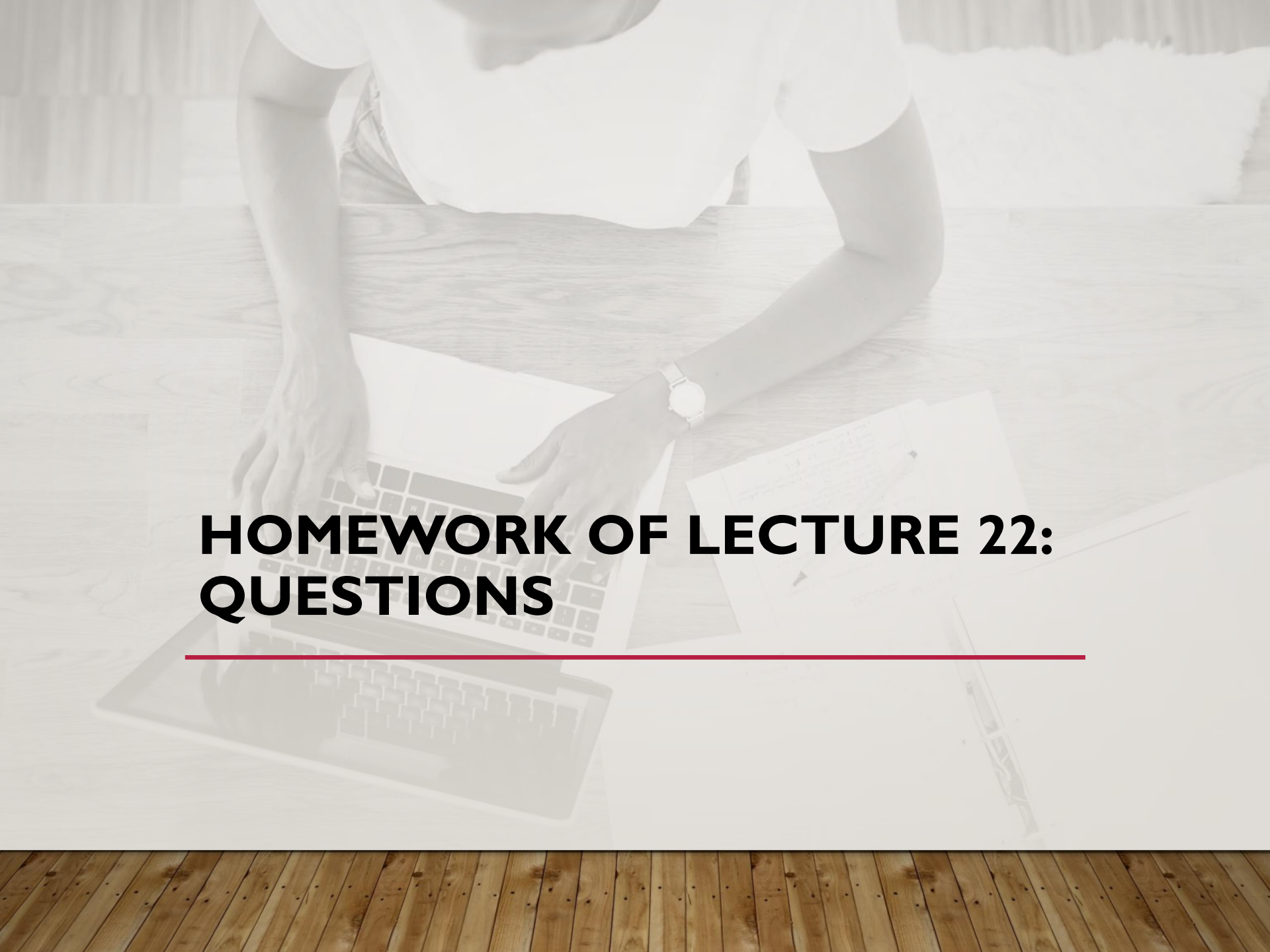
Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

RR = [5.32; +∞[

Newbold et al (2013)



HOMEWORK OF LECTURE 22: QUESTIONS

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)

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)

THANKS!

Questions?