STATISTICAL LABORATORY

Lecture 6: Introduction of Correlation and Regression (Theory and Practice)

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

allenlui@iseg.ulisboa.pt

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Program

1. Fundamental Concepts of Statistics
2. Exploratory Data Analysis
3. Organizing and Summarizing Data
4. **Association and Relationships Between Variables**
5. Index Numbers
6. Time Series Analysis

Exercise book

An exercise book containing practice questions and answers has been uploaded to the Fenix portal. If you need more practice, please use it.

Correlation and Regression

- Purpose
  - Analyse the relationship between two quantitative variables X and Y.
- Scatterplot
  - First step in studying the relationship between variables.
  - Plots each pair (X, Y) as a point in a coordinate system.
  - Allows visual detection of:
    - **Direction** (positive or negative)
    - **Form** (linear or non-linear)
    - **Strength** of the relationship
- Correlation
  - Measures the degree and direction of a linear relationship.
  - Expressed by the correlation coefficient (r) ranges from -1 to $+1$.
  - $r > 0$: variables increase together.
  - $r < 0$: one increases as the other decreases.
- Regression
  - Describes or predicts Y based on X.
  - **Simple Linear Regression Model:** $Y = \beta_0 + \beta_1 X + \varepsilon$
  - Parameters estimated by the least squares method.

Correlation and Regression

Key ideas:

- The scatterplot shows the relationship.
- Correlation quantifies the strength of association.
- Regression explains and predicts the relationship.

Measures of relationships between variables

Two measures of the relationship between variables are:

- **Covariance**
  - a measure of the direction of a linear relationship between two variables
- **Correlation Coefficient**
  - a measure of both the direction and the strength of a linear relationship between two variables

- Only concerned with the strength of the relationship
- No causal effect is implied

Newbold et al. (2013)

Covariance

Definition:

Consider two variables X and Y, for which we have n paired observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

The sample covariance between X and Y is given by

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Exercise 4.6

Show that the covariance can also be calculated using the following formula:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Interpreting covariance

Covariance between two variables:

$$\text{Cov}(x, y) > 0 \rightarrow x \text{ and } y \text{ tend to move in the same direction}$$

$$\text{Cov}(x, y) < 0 \rightarrow x \text{ and } y \text{ tend to move in opposite directions}$$

$$\text{Cov}(x, y) = 0 \rightarrow x \text{ and } y \text{ are uncorrealted}$$

Notes:

- The sign of the covariance indicates the direction of the relationship between the variables.

Correlation Coefficient

Definition: The sample correlation coefficient measures the degree and direction of the linear relationship between two variables X and Y. It is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

or equivalently,

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

Features of correlation coefficient

- Unit free
- Ranges between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any linear relationship

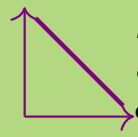
Properties:

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} > 0$: positive linear relationship
- $r_{xy} < 0$: negative linear relationship
- $r_{xy} = 0$: no linear relationship

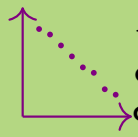
Interpreting correlation coefficient (Silvestre, 2007)

Value of r_{xy}	**Interpretation**
$r_{xy} = -1$	Perfect negative linear correlation (exact linear relationship)
$-1 < r_{xy} < 0$	Negative linear relationship (not perfect)
$r_{xy} = 0$	No linear correlation (absence of linear relationship)
$0 < r_{xy} < 1$	Positive linear relationship (not perfect)
$r_{xy} = 1$	Perfect positive linear correlation (exact linear relationship)

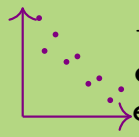
Strength and direction of correlation



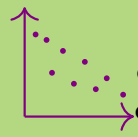
$r = -1$
- Perfect negative linear correlation



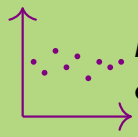
$-1 < r \leq -0.8$
- Strong negative linear correlation



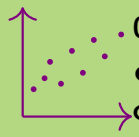
$-0.8 < r \leq -0.5$
- Moderate negative linear correlation



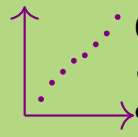
$-0.5 < r < 0$
- Weak negative linear correlation



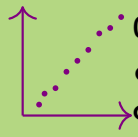
$r = 0$
- No linear correlation



$0 < r < 0.5$
- Weak positive linear correlation



$0.5 \leq r < 0.8$
- Moderate positive linear correlation



$0.8 \leq r < 1$
- Strong positive linear correlation

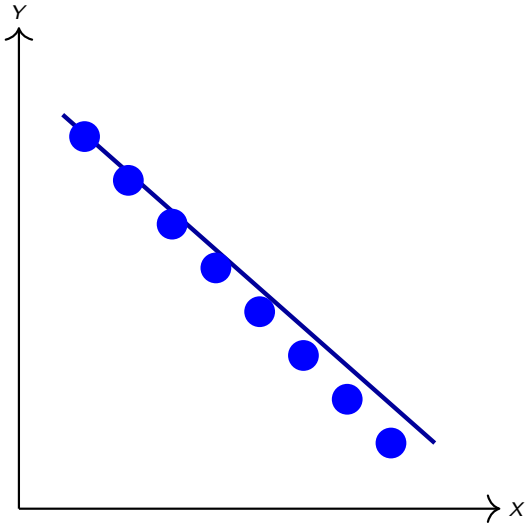


$r = 1$
- Perfect positive linear correlation

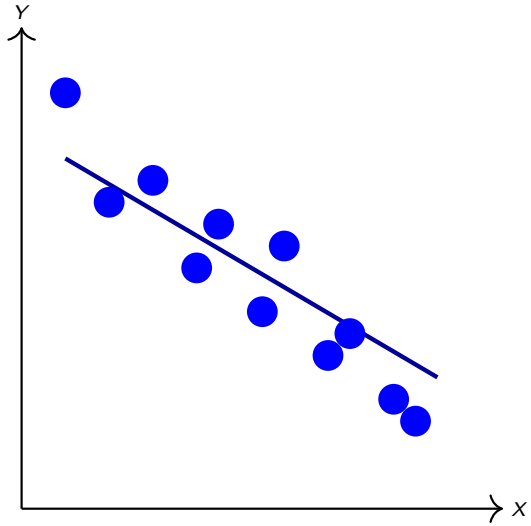
Note:

- The sign of r indicates the direction of the association.
- The absolute value $|r|$ indicates the strength of the linear relationship.

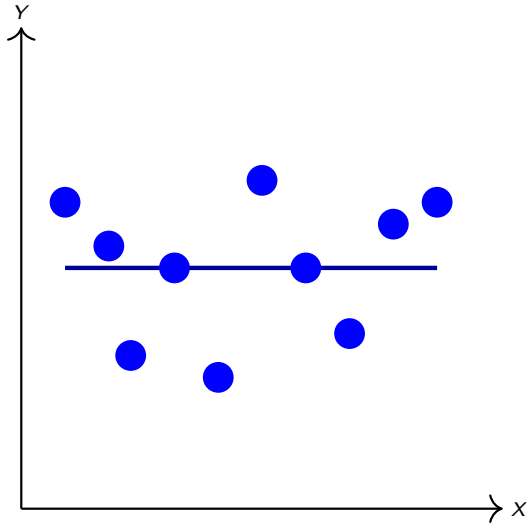
Scatterplots of data with various correlation coefficients



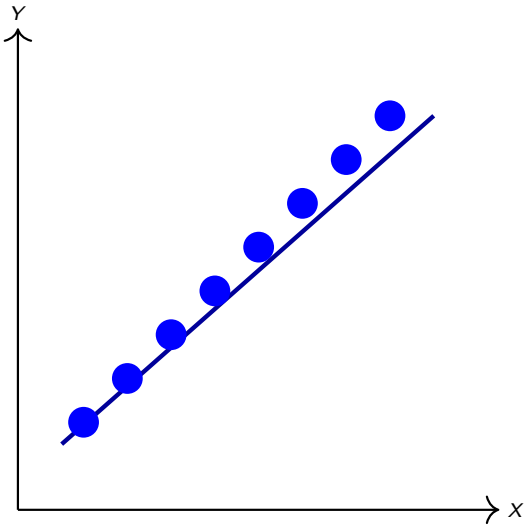
$r = -1$



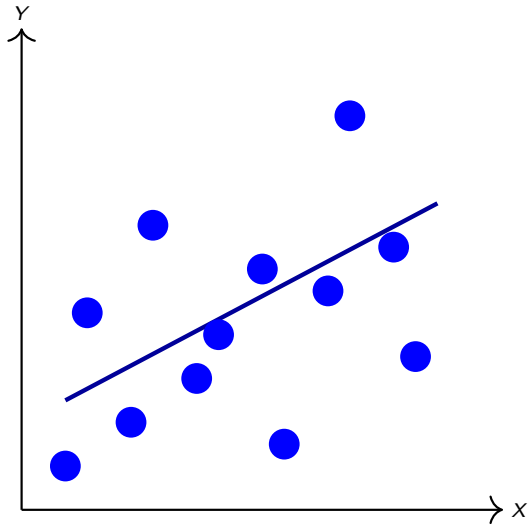
$r = -.6$



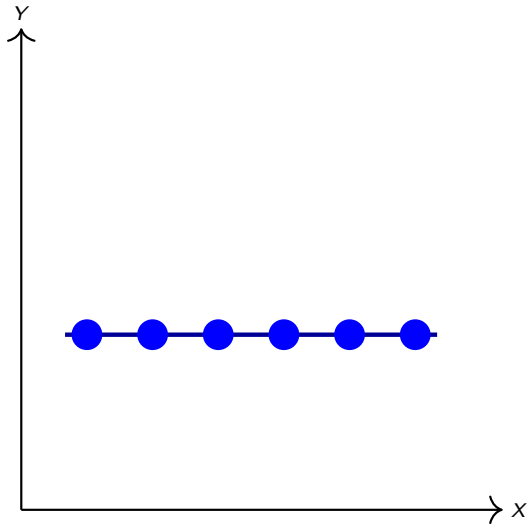
$r = 0$



$r = +1$



$r = +.3$



$r = 0$

Exercise 2.55

A random sample for five exam scores produced the following (hours of study, grade) data values:

Hours Studied (x)	**Test Grade (y)**
3.5	88
2.4	76
4.0	92
5.0	85
1.1	60

a) Compute the covariance.

b) Compute the correlation coefficient.

Standardization of a variable

Let x be a variable with n observations $(x_1, x_2, \ldots, x_n)$.

The standardized values z_i are defined as:

$$z_i = \frac{x_i - \bar{x}}{s_x}, \quad i = 1, 2, \ldots, n$$

where:

- $\bar{x}$ is the mean of x
- s_x is the standard deviation of x
- z_i is a relative number
- The standardized variable z has **mean 0** and **variance 1**

Silvestre (2007)

Least Square Regression Line

Least Square Regression Line

- Let x and y be variables with n observations $(x_i, y_i), i = 1, 2, \ldots, n$, which appear to be correlated.
- The regression line is estimated as:

$$\hat{y}_i = a + bx_i$$

  The Least Squares Method is used to determine the coefficients a and b by minimizing **the sum of squared** estimation errors.
- The estimation error (also called regression residual) is defined as

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

Least squares method and estimators

The Least Squares Method minimizes the sum of squared errors

$$\min_{a,b} Q(a,b) = \min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

- The slope estimator b is

$$\hat{b} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{s_{xy}}{s_x^2}$$

- The intercept estimator is

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

- The regression line is

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Exercise 4.2

The following data refer to the expenses (in monetary units) of an individual on food (F) and leisure (L) over 10 vacation days:

F	20	22	18	10	15	12	20	18	21	11
L	15	17	20	15	14	10	18	20	15	15

a) Compute the mean and variance of each variable.

b) Assess the existence of a possible linear relationship between the variables using an appropriate measure. Justify your answer.

What is the intuition of OLS?

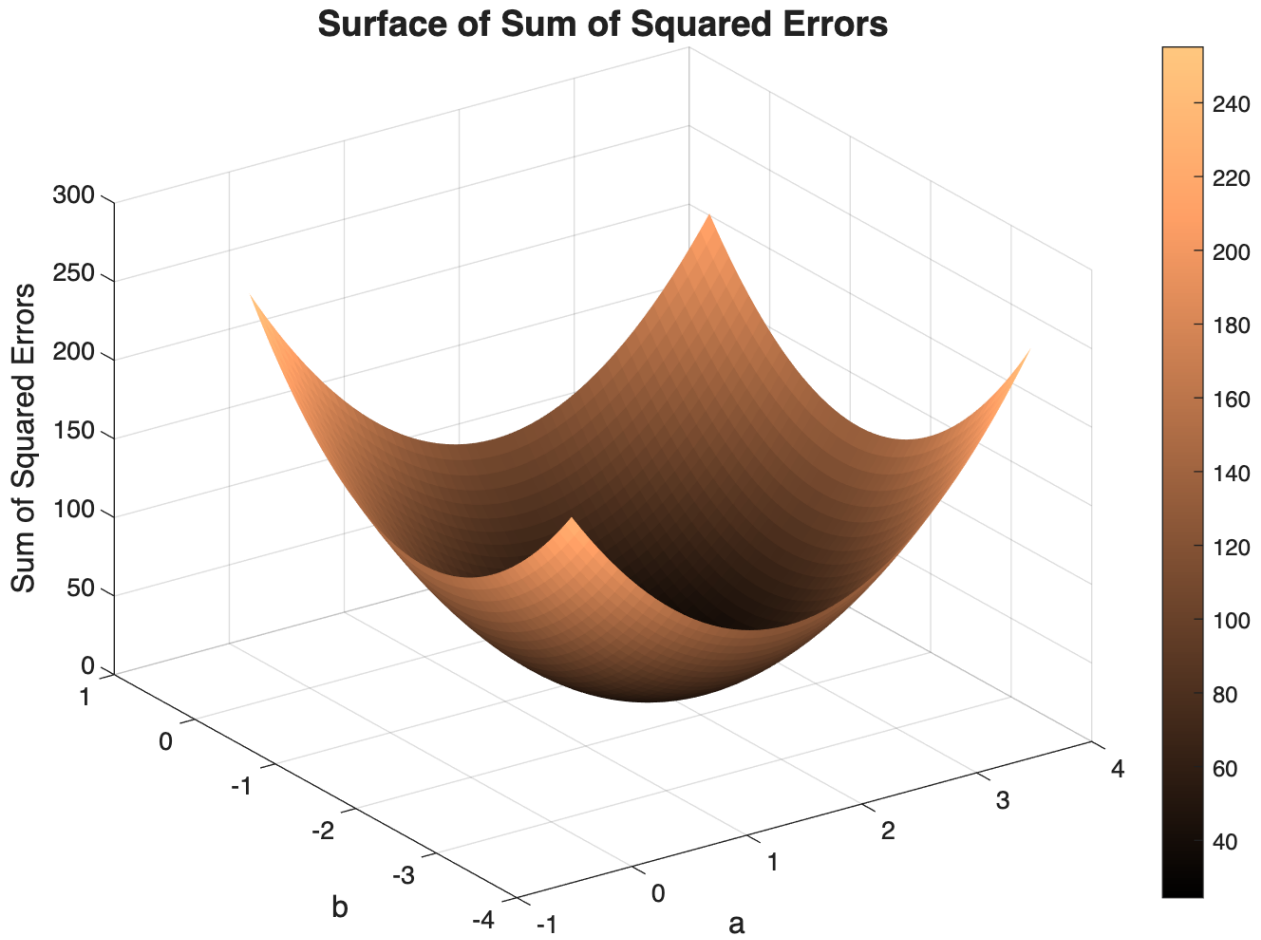
The Ordinary Least Squares (OLS) estimators are defined by:

$$\hat{a}, \hat{b} := \arg\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

In words:

- We look at the rhs as a function in a and b
- That function happens to be quadratic
- We find the values of a and b that minimize that function
- The values that minimize that function are called the solution
- We give the solution a specific name: $\hat{a}$ and $\hat{b}$

Geometry of the minimization problem



The single point at the very bottom (the unique minimum) is denoted $(\hat{a}, \hat{b})$

A numerical example of OLS minimization

Consider $n = 4$ observations: $(1, 2)$, $(2, 3)$, $(3, 5)$, and $(4, 6)$.

Observation (i)	**Dependent (y_i)**	**Independent (x_i)**
1	2	1
2	3	2
3	5	3
4	6	4

Comparing different parameter choices for $y_i = a + bx_i$:

A numerical example of OLS minimization

			OLS ($a = 0.5, b = 1.4$)			**Guess 1** ($a = 0, b = 1.5$)			**Guess 2** ($a = 1, b = 1$)		
i	y_i	x_i	$\hat{y}_i$	e_i	e_i^2	$\hat{y}_i$	e_i	e_i^2	$\hat{y}_i$	e_i	e_i^2
1	2	1	1.9	0.1	0.01	1.5	0.5	0.25	2.0	0.0	0.00
2	3	2	3.3	-0.3	0.09	3.0	0.0	0.00	3.0	0.0	0.00
3	5	3	4.7	0.3	0.09	4.5	0.5	0.25	4.0	1.0	1.00
4	6	4	6.1	-0.1	0.01	6.0	0.0	0.00	5.0	1.0	1.00
SSE ($\sum_{i=1}^{4} e_i^2$):				**0.20**			**0.50**			**2.00**	
$\sqrt{\textbf{SSE}}$ ($\sqrt{\sum_{i=1}^{4} e_i^2}$):				**0.45**			**0.71**			**1.41**	

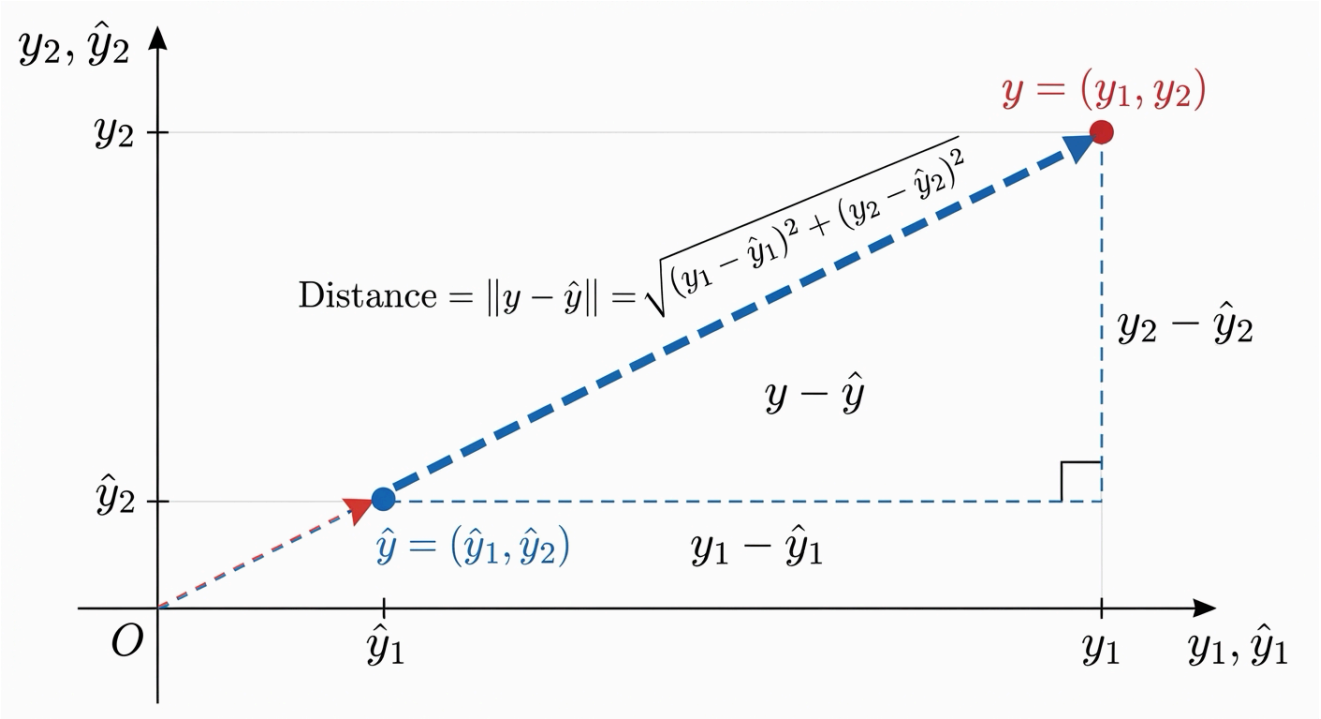
No alternative parameter pair (a, b) can achieve a smaller SSE than the OLS minimizers.

Connection to High-School Mathematics

Consider two 2×1 vectors, y and $\hat{y}$:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad \text{and} \quad \hat{y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \end{pmatrix}$$

If we treat them as points in 2-D coordinate space, the distance between the two vectors y and $\hat{y}$ can be calculated using the Pythagorean theorem as:



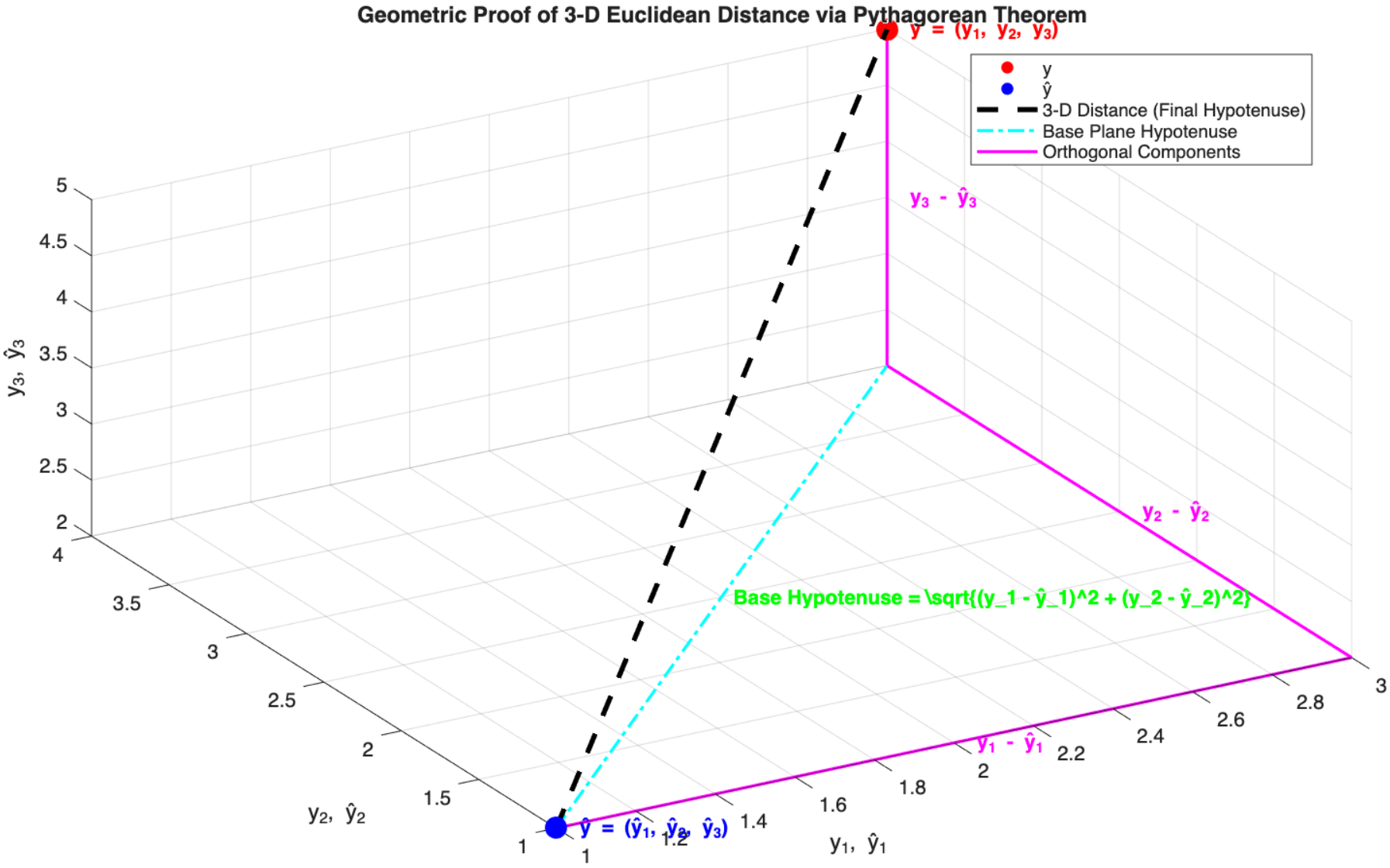
From 2 Dimensions to n Dimensions

- In an n-dimensional space, the actual data vector y and the predicted vector $\hat{y}$ are points with n coordinates.
- The straight-line distance between them generalizes the Pythagorean theorem to n dimensions:

$$\begin{aligned} \|y - \hat{y}\| &= \sqrt{(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \end{aligned}$$

- This is called **Euclidean distances**.
- OLS seeks the parameter pair (a, b) that generates a fitted vector $\hat{y}$ with the minimum Euclidean distance to y.

Illustration in 3 Dimension



Properties of the least squares regression

1. The mean of the estimation errors is zero:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

2. The regression line passes through the point of means:

$$(\bar{x}, \bar{y})$$

3. The mean of the predicted values equals the mean of the observed values:

$$\bar{\hat{y}} = \bar{y}$$

4. The estimation errors are uncorrelated with the variable x:

$$\sum_{i=1}^{n} x_i e_i = 0$$

Exercise 11.19

A company sets different prices for a particular DVD system in eight different regions of the country. The accompanying table shows the numbers of units sold and the corresponding prices (in dollars).

Sales	420	380	350	400	440	380	450	420
Price	104	195	148	204	96	256	141	109

a. Graph these data, and estimate the linear regression of sales on price.

b. What effect would you expect a \$50 increase in price to have on sales?

Newbold et al. (2013)

Assessment of fit quality: MSE and residual variance

Mean Squared Error (MSE):

- Average of the squared prediction errors

$$\text{MSE} = \frac{1}{m}\sum_{k=1}^{m}(\hat{y}_{n+k} - y_{n+k})^2$$

- Assuming m forecasts and corresponding observed values $(\hat{y}_{n+k}, y_{n+k}),\ k = 1, 2, \ldots, m$.
- $m =$ number of **predictions** you are evaluating.
- $y_{n+k} =$ **observed value** at step $n + k$.
- $\hat{y}_{n+k} =$ **predicted value** for step $n + k$.
- $k = 1, 2, \ldots, m$ is just an **index for each prediction** after the initial sample of size n.
- **MSE** is used to assess the quality of out-of-sample or future predictions. The MSE is computed using data points outside the original dataset.

Assessment of fit quality: MSE and residual variance

Variance of estimation errors:

- Measures the dispersion of the estimation (or residual) errors

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Interpretation:

- Smaller values of both **MSE** and residual variance indicate better model fit.
- Both measure the magnitude of prediction errors: the closer predicted values are to observed values, the smaller these quantities. Large values suggest poor predictive performance or high variability.

Assessment of Fit Quality: Coefficient of Determination

Coefficient of Determination (R^2)

- Starts from the following decomposition:

$$\text{Total Variation} = \text{Residual Variation} + \text{Explained Variation}$$

By dividing by n, we obtain:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Using variance notation:

$$s_y^2 = s_e^2 + \hat{s}_y^2 \tag{1}$$

where:

- s_y^2 = total variance of y
- s_e^2 = variance of the residuals (errors)
- $\hat{s}_y^2$ = variance of the fitted values (explained)

Assessment of Fit Quality: Coefficient of Determination

Divided both side of Equation (1) by s_y^2, we get

$$1 = \frac{s_e^2}{s_y^2} + \frac{\hat{s}_y^2}{s_y^2} \tag{2}$$

From this, the Coefficient of Determination is defined as:

$$R^2 = \frac{\hat{s}_y^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

Interpretation

- $0 \leq R^2 \leq 1$
- The closer R^2 is to 1, the better the model fits the data.

Assessment of Fit Quality: Coefficient of Determination

- Total Sum of Square (SST): The overall spread of the actual data points around their mean ($\sum_{i=1}^{n}(y_i - \bar{y})^2$).
- Residual Sum of Square (SSE): The spread that is left unexplained by your model ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$). These are the errors or residuals–the distance between what the model predicted and what actually happened.
- Regression Sum of Squares (SSR): The variation captured by the model's predictions ($\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$). Measures the variability explained by the regression model.

Assessment of Fit Quality: Coefficient of Determination

So

$$SST = SSR + SSE,$$

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

- Represents the proportion of the total variation in y explained by the model.
- Higher R^2 means better fit.

One thing to watch out for when using this formula

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

is that, if there is not a constant in the regression, the R^2 could be negative, this happens when the regression without intercept performs worse than a sample mean (regression with intercept only).

Assessment of Fit Quality: Coefficient of Determination

- There is no fixed cutoff for MSE or residual variance (s_e^2) to define a "bad" fit, because their values depend on the scale of the data and context.
- However, for the coefficient of determination (R^2), general guidelines can be used:

- $R^2 > 0.9 \rightarrow$ excellent fit
- $0.7 < R^2 \leq 0.9 \rightarrow$ good fit
- $0.5 < R^2 \leq 0.7 \rightarrow$ moderate fit
- $R^2 < 0.5 \rightarrow$ weak fit, the model explains little of the variance

Note: In macroeconomics, R^2 can be as high as 0.99 but in microeconomics or finance, R^2 can be as low as 0.2 and the model is still considered to be fine.

Exercise 11.75

The following table shows, for eight vintages of select wine, purchases per buyer (y) and the wine buyer's rating in a year (x):

x	3.6	3.3	2.8	2.6	2.7	2.9	2.0	2.6
y	24	21	22	22	18	13	9	6

a. Estimate the regression of purchases per buyer on the buyer's rating.

b. Interpret the slope of the estimated regression line.

c. Find and interpret the coefficient of determination.

Newbold et al (2013)

Relationship Between Correlation Coefficient and Regression Coefficient

Given: Simple linear regression:

$$\hat{y} = \hat{a} + \hat{b}x$$

Slope:

$$\hat{b} = \frac{s_{xy}}{s_x^2} \quad \text{where } s_{xy} = \text{cov}(X, Y)$$

Pearson correlation:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Relationship Between Correlation Coefficient and Regression Coefficient

- Step 1: Express r_{xy} in terms of $\hat{b}$

  Substitute $s_{xy} = \hat{b}s_x^2$ into $r_{xy} = \frac{s_{xy}}{s_x s_y}$:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\hat{b}s_x^2}{s_x s_y} = \hat{b}\frac{s_x}{s_y}$$

  This proves:

$$\boxed{r_{xy} = \hat{b}\frac{s_x}{s_y}}$$

- Step 2: Express r_{xy}^2

$$r_{xy}^2 = \left(\hat{b}\frac{s_x}{s_y}\right)^2 = \hat{b}^2\frac{s_x^2}{s_y^2}$$

  This proves:

$$\boxed{r_{xy}^2 = \frac{\hat{b}^2 s_x^2}{s_y^2}}$$

Relationship Between Correlation Coefficient and Regression Coefficient

Conclusion/Interpretation

- The square of the correlation coefficient is directly related to the **squared slope** of the regression line, scaled by the ratio of the variances of X and Y.
- This shows mathematically that **strong correlation** $\rightarrow$ **steep slope** (if s_x, s_y fixed).

Exercise 4.1

The advertising expenses and sales (in monetary units) of a company are presented below:

Year	1	2	3	4	5	6	7	8	9	10
Sales	120	150	160	200	210	250	300	360	450	550
Expenses	10	12	14	20	22	25	30	38	50	60

a) Assess the relationship between sales and advertising expenses using an appropriate graphical representation.

b) Estimate the regression line of advertising expenses as a function of sales. Interpret the result.

c) Evaluate the quality of the estimated model using an appropriate measure.

Relationship Between Correlation Coefficient and Regression Coefficient

Simple Linear Regression Model

$$Y_i = \hat{a} + \hat{b}X_i + e_i$$

Definitions:

$$r = \frac{s_{xy}}{s_x s_y} \quad \text{and} \quad R^2 = \frac{SSR}{SST}$$

Derivation:

$$\hat{b} = \frac{s_{xy}}{s_x^2}$$

$$SSR = \hat{b}^2 \sum (X_i - \bar{X})^2 = n\frac{s_{xy}^2}{s_x^2}$$

$$SST = ns_y^2$$

$$\Rightarrow R^2 = \frac{SSR}{SST} = \frac{s_{xy}^2}{s_x^2 s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = r^2$$

Relationship Between Correlation Coefficient and Regression Coefficient

Model	**Relationship**
Simple Linear Regression	$R^2 = r^2$
Multiple Linear Regression	$R^2 \neq r^2$ (but can be showed that $R^2 = Corr(y, \hat{y})^2$)

Interpretation

In simple linear regression, the coefficient of determination (R^2) equals the square of the correlation coefficient (r), meaning the proportion of variance in Y explained by X is directly related to their linear association.

Derivation of OLS Estimates

The basic approach is *multivariate calculus* similar to your microeconomic course.
First step: differentiate $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ (with chain rule).

$$\frac{\partial SSR}{\partial a} = 2\sum_{i=1}^{n}(y_i - a - bx_i)\cdot(-1) = -2\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$\frac{\partial SSR}{\partial b} = 2\sum_{i=1}^{n}(y_i - a - bx_i)\cdot(-x_i) = -2\sum_{i=1}^{n}x_i(y_i - a - bx_i)$$

This is a set of two linear equations and two unknowns (a and b).

First-Order Conditions for OLS

Second step: set derivative to zero
(at this step, $a = \hat{a}$ and $b = \hat{b}$)

$$0 = -2 \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b} x_i)$$

$$0 = -2 \sum_{i=1}^{n} x_i \cdot (y_i - \hat{a} - \hat{b} x_i)$$

These two equations are the *first order necessary conditions* (*foc*) for a minimum.

Solving for the Intercept ($\hat{a}$)

Third step: using the first foc, solve for $\hat{a}$ as a function of $\hat{b}$

$$0 = -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)$$

$$0 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{a} - \sum_{i=1}^{n} \hat{b}x_i$$

$$= \sum_{i=1}^{n} y_i - n\hat{a} - \hat{b}\sum_{i=1}^{n} x_i$$

Which is equivalent to: $n\hat{a} = \sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i$, And:

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{b}\frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{b}\bar{x}$$

Result: $\hat{a}$ is the sample average of y minus $\hat{b}$ times the sample average of x.

Solving for the Slope ($\hat{b}$)

Fourth step: substitute $\hat{a} = \bar{y} - \hat{b}\bar{x}$ into the second first-order condition and solve for $\hat{b}$

$$
\begin{aligned}
0 &= -2\sum_{i=1}^{n} x_i \cdot (y_i - \hat{a} - \hat{b}x_i) \\
0 &= \sum_{i=1}^{n} x_i \cdot (y_i - \hat{a} - \hat{b}x_i) \\
&= \sum_{i=1}^{n} x_i \cdot (y_i - \bar{y} + \hat{b}\bar{x} - \hat{b}x_i) \\
&= \sum_{i=1}^{n} x_i \cdot (y_i - \bar{y} - \hat{b}(x_i - \bar{x})) \\
&= \sum_{i=1}^{n} \left(x_i y_i - \bar{y}x_i - \hat{b}(x_i^2 - \bar{x}x_i) \right) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \hat{b}\sum_{i=1}^{n}(x_i^2 - \bar{x}x_i)
\end{aligned}
$$

Solving for the Slope ($\hat{b}$) - Continued

Repeating the last line (for convenience):

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - \hat{b} \sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)$$

This is equivalent to

$$\hat{b} \sum_{i=1}^{n} (x_i^2 - \bar{x} x_i) = \left(\sum_{i=1}^{n} x_i y_i \right) - n \bar{x} \bar{y}$$

Where the LHS can be simplified

$$\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i) = \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i = \left(\sum_{i=1}^{n} x_i^2 \right) - n \bar{x}^2$$

And isolating $\hat{b}$ on the left

$$\hat{b} = \frac{\left(\sum_{i=1}^{n} x_i y_i \right) - n \bar{x} \bar{y}}{\left(\sum_{i=1}^{n} x_i^2 \right) - n \bar{x}^2}$$

Solving for the Slope ($\hat{b}$) - Continued

Dividing the numerator and the denominator by n, and recalling that

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y},$$

we have

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \\ &= \frac{s_{xy}}{s_x^2}. \end{aligned} \tag{3}$$

Exercise 3.7

Show that the mean of the squares of the deviations of the values of a variable from a constant *ccc* is minimized when this constant is equal to the mean of the variable.

Exercise 2.55 (Suggested Answers)

a)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{3.5 + 2.4 + 4.0 + 5.0 + 1.1}{5} = \frac{16.0}{5} = 3.2$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{88 + 76 + 92 + 85 + 60}{5} = \frac{401}{5} = 80.2$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Exercise 2.55 (Suggested Answers)

Table: Cross-Moment Calculation Table

i	$(x_i - \bar{x})(y_i - \bar{y})$
1	$(3.5 - 3.2)(88 - 80.2) = 2.34$
2	$(2.4 - 3.2)(76 - 80.2) = 3.36$
3	$(4.0 - 3.2)(92 - 80.2) = 9.44$
4	$(5.0 - 3.2)(85 - 80.2) = 8.64$
5	$(1.1 - 3.2)(60 - 80.2) = 42.42$
Sum	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 66.2$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{66.2}{5} = 13.24$$

Exercise 2.55 (Suggested Answers)

b) Correlation coefficient: $r = \frac{s_{xy}}{s_x s_y}$.

- Using the sums of squares from our data:$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 9.02$
- $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 648.80$
- n =5.

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{9.02}{5} = 1.804$$

$$s_x = \sqrt{1.804} \approx 1.34313$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{648.80}{5} = 129.76$$

$$s_y = \sqrt{129.76} \approx 11.39122$$

$$r = \frac{13.24}{(1.34313)(11.39122)} = \frac{13.24}{15.30006} \approx 0.8654$$

Exercise 4.2 (Suggested Answer)

a) Mean for food:

$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i = \frac{20+22+18+10+15+12+20+18+21+11}{10} = \frac{167}{10}$$

Variance for food:

$$\sum_{i=1}^{n}(f_i - \bar{f})^2 = (20-16.7)^2 + (22-16.7)^2 + \cdots + (11-16.7)^2 = 174.1$$

$$s_f^2 = \frac{174.1}{10} = 17.41$$

Mean for leisure:

$$\bar{l} = \frac{1}{n}\sum_{i=1}^{n} l_i = \frac{15+17+20+15+14+10+18+20+15+15}{10} = \frac{159}{10}$$

Variance for leisure:

$$\sum_{i=1}^{n}(l_i - \bar{l})^2 = (15-15.9)^2 + (17-15.9)^2 + \cdots + (15-15.9)^2 = 80.9$$

Exercise 4.2 (Suggested Answer)

b) To assess the possible linear relationship, we calculate the correlation coefficient ($r = \frac{s_{fl}}{s_f s_l}$)

$$s_{fl} = \frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})(l_i - \bar{l})$$

$$\sum_{i=1}^{n}(f_i - \bar{f})(l_i - \bar{l}) = (20 - 16.7)(15 - 15.9) + \cdots + (11 - 16.7)(15 - 15.9) = 58.7$$

$$s_{fl} = \frac{58.7}{10} = 5.87, s_f = \sqrt{17.41} \approx 4.1725, s_l = \sqrt{8.09} \approx 2.8443$$

$$r = \frac{5.87}{(4.1725)(2.8443)} = \frac{5.87}{11.8872} \approx 0.4946$$

This indicates a weak-to-moderate positive linear relationship between an individual's food expenses and leisure expenses during vacation days.

Exercise 3.7 (Suggested Answer)

The mean of the squares of the deviations from a constant c is given by the function $f(c)$:

$$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

foc:

$$\frac{d}{dc}f(c) = \frac{d}{dc}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2\right] = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}2(x_i - \hat{c})(-1) = 0$$

$$\frac{1}{n}\left(\sum_{i=1}^{n}x_i - n\hat{c}\right) = 0$$

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$$

THANKS!

Questions?