Machine Learning Models 2024-2025

# Exercises Sheet #1

## Exercise 1

The Tri-City Office Equipment Corporation sells an imported desk calculator on a franchise basis and performs preventive maintenance and repair service on this calculator. The data below have been collected from 18 recent calls on users to perform routine preventive maintenance service. For each call, X is the number of machines serviced and Y is the total number of minutes spent by the service person. Assume that the simple linear regression model is appropriate.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 7 | 6 | 5 | I | 5 | 4 | 7 | 3 | 4 |
| $Y_i$ | 97 | 86 | 78 | 10 | 75 | 62 | 101 | 39 | 53 |

| $i$ | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 2 | 8 | 5 | 2 | 5 | 7 | I | 4 | 5 |
| $Y_i$ | 33 | 118 | 65 | 25 | 71 | 105 | 17 | 49 | 68 |

Summary calculation results are:

$\sum Y_i = 1{,}152$, $\sum X_i = 81$, $\sum(Y_i - \bar{Y})^2 = 16{,}504$, $\sum(X_i - \bar{X})^2 = 74.5$, $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 1{,}098$.

*Questions:*

1. Obtain the estimated regression function.
2. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?
3. Obtain a point estimate of the mean service time when X = 5 machines are serviced.

## Exercise 2

The director of admissions of a small college administered a newly designed entrance test to 20 students selected at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the entrance test score (X). The results of the study follow. Assume that the simple linear regression model is appropriate.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ | 5.5 | 4.8 | 4.7 | 3.9 | 4.5 | 6.2 | 6.0 | 5.2 | 4.7 | 4.3 | 4.9 | 5.4 | 5.0 | 6.3 | 4.6 | 4.3 | 5.0 | 5.9 | 4.1 | 4.7 |
| $Y_i$ | 3.1 | 2.3 | 3.0 | 1.9 | 2.5 | 3.7 | 3.4 | 2.6 | 2.8 | 1.6 | 2.0 | 2.9 | 2.3 | 3.2 | 1.8 | 1.4 | 2.0 | 3.8 | 2.2 | 1.5 |

Summary calculation results are: $\sum X_i = 100.0$, $\sum Y_i = 50.0$, $\sum X_i^2 = 509.12$, $\sum Y_i^2 = 134.84$, $\sum X_i Y_i = 257.66$.

*Questions*:

1. Obtain the least squares estimates of the paprameters , and state the estimated regression function.
2. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

3. Obtain a point estimate of the mean freshman GPA when the entrance test score is X = 5.0.
4. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

**Exercise 3**

Consider the following data :

| $x$ | $y$ |
|---|---|
| 0.86 | 2.49 |
| 0.09 | 0.83 |
| -0.85 | -0.25 |
| 0.87 | 3.10 |
| -0.44 | 0.87 |
| -0.43 | 0.02 |
| -1.10 | -0.12 |
| 0.40 | 1.81 |
| -0.96 | -0.83 |
| 0.17 | 0.43 |

*Questions* :

1) Compute the parameter values W of the best line that fit these data ?
2) Compute the determination coefficient $r^2$?
3) How does the model fit the data?

**Exercise 4**

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

| **Age** | **Number of Driver Deaths per 100,000** |
|---|---|
| 17.5 | 38 |
| 22 | 36 |
| 29.5 | 24 |
| 44.5 | 20 |
| 64.5 | 18 |
| 80 | 28 |

Based on the given data and using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable :

1) Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
2) What is the slope of the least squares (best-fit) line?
3) Find the determination coefficient $r^2$.
4) Is there a linear relationship between age of a driver and driver fatality rate?
5) What is the number of deaths for ages 40 and 60.

**Exercise 5**

The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

| Swim Time | Heart Rate |
|---|---|
| 34.12 | 144 |
| 35.72 | 152 |
| 34.72 | 124 |
| 34.05 | 140 |
| 34.13 | 152 |
| 35.73 | 146 |
| 36.17 | 128 |
| 35.57 | 136 |
| 35.37 | 144 |
| 35.57 | 148 |

1) Find the equation of the least-squares regression line.
2) How well does the regression line fit the data? Explain your response.
3) Which point has the largest residual (prediction error)? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

# Appendix: Least Squares (Simple Linear regression)

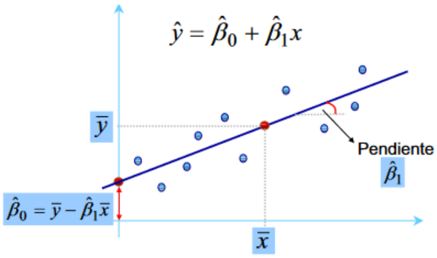
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

## Least squares estimators

The results are given by

$$\boxed{\hat{\beta}_1 = \frac{cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

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




Or

**Definition: least squares regression Line**

Given a collection of pairs $(x, y)$ of numbers (in which not all the $x$-values are the same), there is a line $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ that best fits the data in the sense of minimizing the sum of the squared errors. It is called the *least squares regression line*. Its slope $\hat{\beta}_1$ and $y$-intercept $\hat{\beta}_0$ are computed using the formulas

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

and

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

where

$$SS_{xx} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

and

$$SS_{xy} = \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right)$$

$\bar{x}$ is the mean of all the $x$-values, $\bar{y}$ is the mean of all the $y$-values, and $n$ is the number of pairs in the data set.

The equation

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

specifying the least squares regression line is called the least squares regression equation.