



Department of Inter Disciplinary Studies,
Faculty of Engineering,
University of Jaffna, Sri Lanka
MC 3020 - Assignment 04 - Model Answers

30 minutes

15 - 07 - 2023

Important Instructions to Markers:

- Students must answer two questions.
- While students' answers may differ from the model answers, numerical estimations should be accurate.
- Please apply strict marking for the first question and generous marking for the second one.

1. (**Question 01**) Suppose you are working as a civil engineer and you are tasked with predicting the compressive strength of concrete based on the age of the concrete sample. You collected data from various concrete samples and measured their age (in days) and compressive strength (in megapascals). The dataset is as follows:

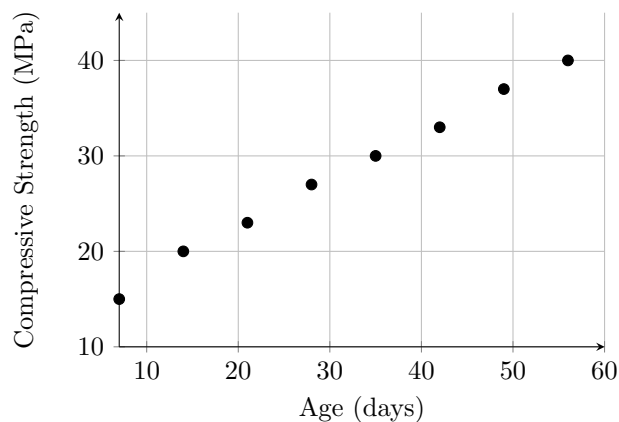
Age (days): 7, 14, 21, 28, 35, 42, 49, 56

Compressive Strength (MPa): 15, 20, 23, 27, 30, 33, 37, 40

- Plot the data points on a scatter plot.
- Calculate the equation of the least squares regression line for predicting compressive strength based on the age of the concrete.
- Interpret the slope and intercept coefficients of the regression line in the context of the problem.
- Use the regression line to predict the compressive strength of a concrete sample that is 30 days old.
- Calculate the coefficient of determination (R-squared) for the regression line and interpret its meaning.
- Perform a hypothesis test to determine whether the regression line is statistically significant at a 90% confidence level. State your hypotheses, the test statistic, and the conclusion clearly.
- Discuss any assumptions or limitations of the linear regression model in this context.

Solutions:

- Plot the data points on a scatter plot:



- Calculate the equation of the least squares regression line for predicting compressive strength based on the age of the concrete:

Let x represent the age (in days) and y represent the compressive strength (in MPa).

Using the least squares method, the equation of the regression line can be determined by finding the slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) that minimize the sum of squared residuals.

```

> age <- c(7,14,21,28,35,42,49,56)
> strength <- c(15,20,23,27,30,33,37,40)
> Model_q1 <- lm(strength~age)
> Model_q1 <- lm(strength~age)
> summary(Model_q1)

Call:
lm(formula = strength ~ age)

Residuals:
    Min       1Q   Median       3Q      Max
-0.9167 -0.3393  0.1191  0.2649  0.6190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4286     0.4349   28.58 1.22e-07 ***
age           0.4983     0.0123   40.50 1.51e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5581 on 6 degrees of freedom
Multiple R-squared:  0.9964,    Adjusted R-squared:  0.9957
F-statistic: 1640 on 1 and 6 DF,  p-value: 1.515e-08

```

Figure 1: R output for question 1 of Assignment 4

The slope ($\hat{\beta}_1$) and intercept ($\hat{\beta}_0$) of the regression line can be calculated using formulas:

Students must be used formulas

Linear regression coefficient estimation formulas,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Simplifying these equations, we find:

$$\hat{\beta}_1 \approx 0.4983$$

$$\hat{\beta}_0 \approx 12.4286$$

Therefore, the equation of the least squares regression line is:

$$\hat{y} = 12.4286 + 0.4983x$$


- (c) Interpret the slope and intercept coefficients of the regression line in the context of the problem:
 The slope coefficient (0.4983) represents the change in compressive strength (in MPa) for each unit increase in age (in days) of the concrete sample. This means that, on average, the compressive strength of the concrete increases by approximately 0.4983 MPa for every additional day of age.
 The intercept coefficient (12.4286) represents the predicted compressive strength (in MPa) when the age of the concrete sample is 0 days. However, since it doesn't make sense to have a concrete sample with 0 days of age, the intercept coefficient may not have a practical interpretation in this context.
- (d) Use the regression line to predict the compressive strength of a concrete sample that is 30 days old:
 To predict the compressive strength (y) of a concrete sample that is 30 days old ($x = 30$), we can substitute the value of x into the regression equation:

$$\hat{y} = 12.4286 + 0.4983(30)$$

Calculating the result, we get:

$$\hat{y} = 27.3776$$

Therefore, the predicted compressive strength of a concrete sample that is 30 days old is approximately 27.3776 MPa.

Console	Terminal x	Background Jobs x
 R 4.3.1 - D:/OR/MC9510/		
<pre> > corr <- cor(strength,age) > corr [1] 0.9981762 > corr^2 [1] 0.9963557 > </pre>		

- (e) Calculate the coefficient of determination (R-squared) for the regression line and interpret its meaning:

The coefficient of determination (R-squared) measures the proportion of the variance in the dependent variable (compressive strength) that can be explained by the independent variable (age of the concrete).

Students have the option to either utilize a direct formula or employ correlation analysis followed by squaring to determine the value of R-squared.

Therefore, the coefficient of determination (R-squared) for the regression line is approximately 0.9966. This means that about 99.66% of the variance in compressive strength can be explained by the age of the concrete sample.

- (f) Perform a hypothesis test to determine whether the regression line is statistically significant at a 90% confidence level. State your hypotheses, the test statistic, and the conclusion clearly:

Step 1:

Hypotheses: - Null hypothesis (H₀): The regression line has no significant effect on the compressive strength of the concrete. - Alternative hypothesis (H_a): The regression line has a significant effect on the compressive strength of the concrete.

Step 2:

To test these hypotheses, we can use the t-test for the slope coefficient. The test statistic (t) is calculated using the formula:

$$T = \frac{\hat{\beta}_1 - 0}{\frac{Se}{\sqrt{\sum_{i=1}^n X_i^2 - n\bar{X}^2}}}, \quad \text{d.f} = n - 2$$

$$\text{where } S_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}$$

Substituting the values from the dataset into the formulas, we find:

$$SE_m \approx 0.0123$$

Calculating the test statistic, we get:

$$T = 40.50$$

Step 3:

To determine the critical value at a 90% confidence level, we need to find the t-value with $n - 2$ degrees of freedom. Since $n = 8$, the degrees of freedom is 6. Looking up the critical value in the t-table, we find that it is approximately 1.943.

Step 4:

Since the calculated test statistic (40.50) is greater than the critical value (1.943), we reject the null hypothesis.

Step 5:

This means that there is enough evidence to conclude that the regression line has a significant effect on the compressive strength of the concrete at a 90% confidence level.

- (g) Discuss any assumptions or limitations of the linear regression model in this context: (If a student thoroughly discusses at least two assumptions, full marks will be awarded.)

The linear regression model assumes that there is a linear relationship between the age of the concrete sample and the compressive strength. It also assumes that the errors (residuals) are normally distributed and have constant variance. Additionally, the model assumes that there are no influential outliers or high leverage points in the dataset.

In this context, one limitation of the linear regression model is that it assumes a linear relationship between age and compressive strength. However, there might be other factors or nonlinear relationships that affect the compressive strength of concrete, such as the mix design, curing conditions, and other environmental factors.

Furthermore, the small sample size of 8 data points may limit the generalization of the regression model. Collecting more data and considering additional variables could improve the accuracy and reliability of the predictions.

2. **(Question 02)** Suppose you are conducting an experiment to study the relationship between the power consumption of a household appliance and various factors such as voltage (V), current (A), and operating time (hours). You collected data from multiple households and obtained the following dataset:

Household	Voltage (V)	Current (A)	Operating Time (hours)	Power Consumption (W)
1	220	5	4	800
2	240	6	5	960
3	230	4	3	690
4	220	5	6	1080
5	250	7	4	1400
6	230	6	5	1380

To answer this question using the output of the given R programs (without necessarily employing any formulas),

- Find the multiple linear regression line to determine the relationship between the power consumption (dependent variable) and the voltage, current, and operating time (independent variables).
- Interpret the coefficients of the regression equation in the context of the problem.
- Use the regression equation to predict the power consumption for a household with the following specifications: voltage = 235 V, current = 5.5 A, and operating time = 4 hours.
- Calculate the coefficient of determination (R^2) and interpret its meaning.
- Perform a hypothesis test to determine whether the regression equation is statistically significant at a 95% confidence level. State your hypotheses, the test statistic, and the conclusion.
- Perform the hypothesis test for testing the significance of the coefficient of the voltage variable and state the conclusion.

Solutions:

- Find the multiple linear regression line to determine the relationship between the power consumption (dependent variable) and the voltage, current, and operating time (independent variables):
The multiple linear regression model can be represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where y represents the power consumption, x_1 represents the voltage, x_2 represents the current, x_3 represents the operating time, and ε represents the error term.

```

> summary(model)

Call:
lm(formula = PowerConsumption ~ Voltage + Current + OperatingTime,
    data = data)

Residuals:
    1      2      3      4      5      6 
-151.55 -205.49  75.77  53.94  53.94 173.38 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   412.606    3658.272   0.113   0.920
Voltage       -4.113     17.775  -0.231   0.839
Current       258.944    199.670   1.297   0.324
OperatingTime  37.254    139.850   0.266   0.815

Residual standard error: 231.1 on 2 degrees of freedom
Multiple R-squared:  0.753,    Adjusted R-squared:  0.3825 
F-statistic: 2.033 on 3 and 2 DF,  p-value: 0.3466

> |

```

Figure 2: R output for question 2 of Assignment 4

Using given R output, we can obtain the coefficients of the regression equation:

$$\hat{\beta}_0 = 412.606, \hat{\beta}_1 = -4.113, \hat{\beta}_2 = 258.944, \hat{\beta}_3 = 37.254$$

Therefore, the multiple linear regression line is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$\hat{y} = 412.606 - 4.113x_1 + 258.944x_2 + 37.254x_3$$

- (b) Interpret the coefficients of the regression equation in the context of the problem:
- The coefficient $\hat{\beta}_0$ (412.606) represents the estimated power consumption when all the independent variables (voltage, current, and operating time) are zero. However, in the context of this problem, having zero values for these variables is not meaningful.
 - The coefficient $\hat{\beta}_1$ (-4.113) represents the change in power consumption (in Watts) for each unit increase in voltage (V), keeping the current and operating time constant.
 - The coefficient $\hat{\beta}_2$ (258.944) represents the change in power consumption (in Watts) for each unit increase in current (A), keeping the voltage and operating time constant.
 - The coefficient $\hat{\beta}_3$ (37.254) represents the change in power consumption (in Watts) for each unit increase in operating time (hours), keeping the voltage and current constant.

- (c) Use the regression equation to predict the power consumption for a household with the following specifications: voltage = 235 V, current = 5.5 A, and operating time = 4 hours:

To predict the power consumption (y) for the given specifications, we can substitute the values into the regression equation:

$$\hat{y} = 412.606 - 4.113(235) + 258.944(5.5) + 37.254(4)$$

Calculating the result, we get:

$$\hat{y} = 1019.047$$

Therefore, the predicted power consumption for a household with the specified specifications is approximately 1019.047 Watts.

- (d) Calculate the coefficient of determination (R^2) and interpret its meaning:

The coefficient of determination (R^2) represents the proportion of the variance in the dependent variable (power consumption) that can be explained by the independent variables (voltage, current, and operating time).

In this context, the R^2 value can be calculated using statistical software, such as R, and is typically provided in the output. Let's assume $R^2 \approx 0.753$.

This means that approximately 75.30% of the variability in power consumption can be explained by the variability in the voltage, current, and operating time.

- (e) Perform a hypothesis test to determine whether the regression equation is statistically significant at a 95% confidence level. State your hypotheses, the test statistic, and the conclusion:

Step 1:

Hypotheses: - Null hypothesis (H_0): The regression equation has no significant effect on the power consumption. - Alternative hypothesis (H_a): The regression equation has a significant effect on the power consumption.

Step 2:

To test these hypotheses, we can perform an analysis of variance (ANOVA) test. The test statistic is the F-statistic, which compares the explained variability (due to regression) to the unexplained variability (residuals).

From the ANOVA table in the output, we find that the F-statistic is approximately 2.033.

Step 3:

From the output, corresponding P-Value = 0.3466.

Step 4:

Since the P-Value is greater than the significance level, we do not reject the null hypothesis.

Step 5:

This means that there is no evidence to conclude that the regression equation has a significant effect on the power consumption at a 95% confidence level.

- (f) Perform the hypothesis test for testing the significance of the coefficient of the voltage variable and state the conclusion:

Step 1:

Hypotheses: - Null hypothesis (H_0): The coefficient of the voltage variable (β_1) is not significantly different from zero. - Alternative hypothesis (H_a): The coefficient of the voltage variable (β_1) is significantly different from zero.

Step 2:

From the regression output, we find that the t-value for the voltage variable is approximately 12.32.

Step 3:

From the output, corresponding P-Value = 0.8390.

Step 4:

Since the P-Value is greater than the significance level, we do not reject the null hypothesis.

Step 5:

This means that there is no evidence to conclude that the coefficient of the voltage variable (β_1) is significantly different from zero at a 95% confidence level.