Topic 1
Simple Linear Regression

1. A sample of 6 pairs of values (x_i, y_i) is given in the table below.

x_i	8	10	12	14	16	18
y_i	9	14	15	13	19	18

(a) Draw a scatter diagram for X and Y.
(b) Fit a simple regression model to the data.
(c) Draw the regression line in the scatter diagram. Does the regression line pass through the centroid $(\bar{x}, \bar{y})$?
(d) Estimate the variance of the error terms.

2. A sample of 10 pairs of values (x_i, y_i) which will be represented by the following model:

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

The summaries of the data are:

$$\sum_{i=1}^{10} x_i = 100; \quad \sum_{i=1}^{10} y_i = 200; \quad \sum_{i=1}^{10} x_i y_i = 2000; \quad \sum_{i=1}^{10} x_i^2 = 2000; \quad \sum_{i=1}^{10} y_i^2 = 5000$$

Calculate the least squares estimate of β_0 and β_1.

3. The number of pounds of steam used per month at a plant is thought to be related to the average monthly ambient temperature. Last year's usage Y (in 1000 pounds) and temperatures X (in $^\circ F$) were studied, the summaries of the data is given below.

$$\sum_{i=1}^{12} x_i = 558; \quad \sum_{i=1}^{12} y_i = 5062.34; \quad \sum_{i=1}^{12} x_i y_i = 265869.63; \quad \sum_{i=1}^{12} x_i^2 = 29256; \quad \sum_{i=1}^{12} y_i^2 = 2416234.61;$$

$$\sum_{i=1}^{12} (x_i - \bar{x})^2 = 3309; \quad \sum_{i=1}^{12} (y_i - \bar{y})^2 = 280627.4204; \quad \sum_{i=1}^{12} (y_i - \hat{y}_i)^2 = 37.8547.$$

(a) (i) Fit a simple regression model to the data. State one of the assumptions made in this model. Interpret the meaning of $\hat{\beta}_1$ found.
(ii) State the mean and variance for the least square estimator of β_1.
(iii) State the relationship between the correlation coefficient and the slope, and hence find the correlation coefficient and interpret its meaning.
(b) Plant management believes that an increase in average ambient temperature of $1^\circ F$ will increase average monthly steam usage by $9100\,lb$. Do the data support this statement, use the significance level, $\alpha = 0.01$?
(c) Construct a 99% prediction interval on y_h with $x_h = 58^\circ F$. Interpret the interval obtained.

4. A sample of 20 pairs of values (x_i, y_i) which will be represented by the following model:

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

The summaries of the data are:

$$\hat{\beta}_0 = 138.561; \quad \hat{\beta}_1 = -1.104; \quad \sum (x_i - \bar{x})^2 = 10.668; \quad \sum (y_i - \bar{y})^2 = 20.838;$$

$$\sum e_i^2 = \sum (y_i - \hat{y})^2 = 7.832; \quad \bar{x} = 2$$

a) Construct a 99% confidence interval for the mean per capita consumption of natural gas ($E(Y_h)$) for all gas stations with the prices of natural gas at $2.11.

b) Construct a 99% prediction interval for the per capita consumption of natural gas Y_h for a gas stations with the prices of natural gas at $2.11.

5. A sample of 10 pairs of values (x_i, y_i) which will be represented by the following model:

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

The summaries of the data are:

$$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 400; \quad \sum_{i=1}^{10}(y_i - \bar{y})^2 = 425; \quad \sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2 = 225$$

Calculate the t-statistic used for testing the hypothesis $H_0 : \beta_1 = 0$.

6. A sample of 10 pairs of values (x_i, y_i) which will be represented by the following model:

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

The summaries of the data are:

$$\sum_{i=1}^{10} y_i = 30; \quad \sum_{i=1}^{10} y_i^2 = 690; \quad R^2 = 0.6.$$ Calculate MS_E .

7. A sample of 20 pairs of values (x_i, y_i) which will be represented by the following model:

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

F-ratio for testing the hypothesis $H_0 : \beta_1 = 0$ is equal to 12. Calculate R^2.

8. A sample of 8 pairs of values (x_i, y_i) which will be represented by the following model:

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

The summaries of the data are:

$$\hat{\beta}_1 = -35.69 ; \quad \sum(x_i - \bar{x})^2 = 1.62; \quad \sum(y_i - \bar{y})^2 = 2394$$

Determine the 90% confidence interval for β_1.

9. In a medical survey, the weight X (in pounds) and systolic blood pressure Y of 26 randomly selected males in the age group 25 – 30 are recorded and the summaries are given below.

$$\sum_{i=1}^{26} x_i = 4743; \sum_{i=1}^{26} y_i = 3786; \sum_{i=1}^{26} x_i y_i = 697076; \sum_{i=1}^{26} x_i^2 = 880545; \sum_{i=1}^{26} y_i^2 = 555802;$$

$$\sum_{i=1}^{26}(x_i - \bar{x})^2 = 15312.35; \sum_{i=1}^{26}(y_i - \bar{y})^2 = 4502.15; \sum_{i=1}^{26}(y_i - \hat{y}_i)^2 = 1808.5726$$

(a) Test if there is a linear correlation between X and Y with $\alpha = 0.05$.

(b) Fit a simple regression model to the data.

(c) Find a 99% confidence interval on the slope. Interpret the interval obtained.

(d) Construct the ANOVA table and test for significance of regression by using the significance level, $\alpha = 0.05$.

(e) Construct a 95% confidence interval on the mean response with $x_h = 180\ lb$. Interpret the interval obtained.

(f) Find the coefficient of determination and explain what it means.