Chapter 11
Simple Linear Regression and Correlation

11.1 Introduction to Linear Regression

Objectives:

1. **To introduce the linear relationship between two r.v.'s.**
2. **To define the regression model.**
3. **To define a bivariate random variable.**

- Linear regression studies the relationship (linear) between two or more dependent variables, and assumes that the relationship can be expressed by the formula for a straight line as follows: $Y = \alpha + \beta X$.
- Y denotes the **dependent** variable (response) and X denotes the **independent** variable (predictor or explanatory), where α and β represents the y-intercept and the slope of the **regression line** of the population, respectively.
- To estimate the best regression line that relates Y to X, we need to draw a bivariate random sample of size n from the bivariate population (X,Y), i.e. $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$.

11.2 The Simple Linear Regression Model

Objectives:

1. **To define the simple linear regression model.**
2. **To plot the scatter diagram.**

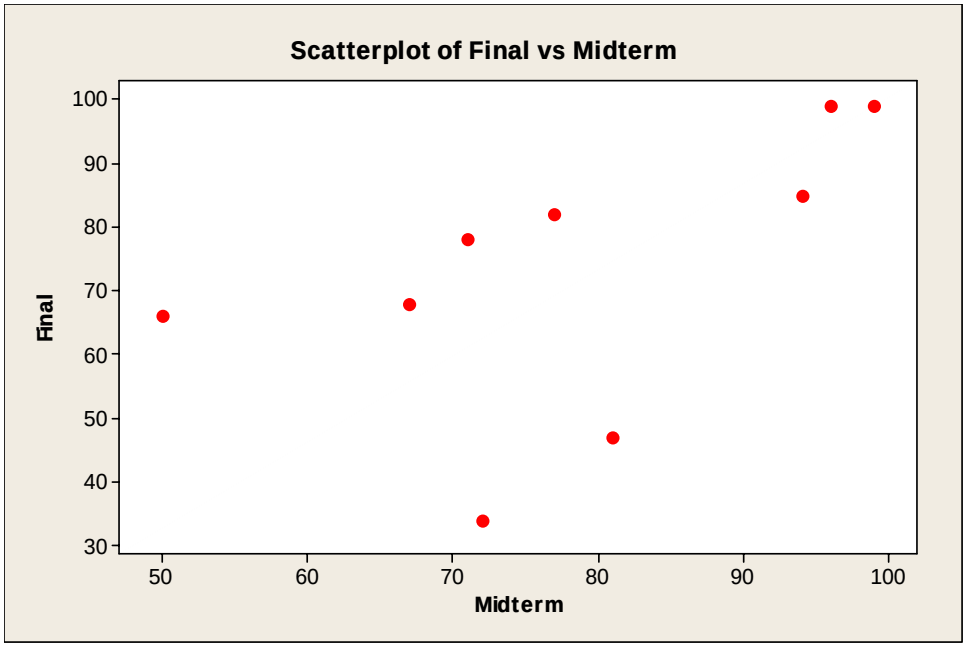
Definition: The **simple linear regression model** is given by $Y = \alpha + \beta X + \varepsilon$, where ε is called the random error.

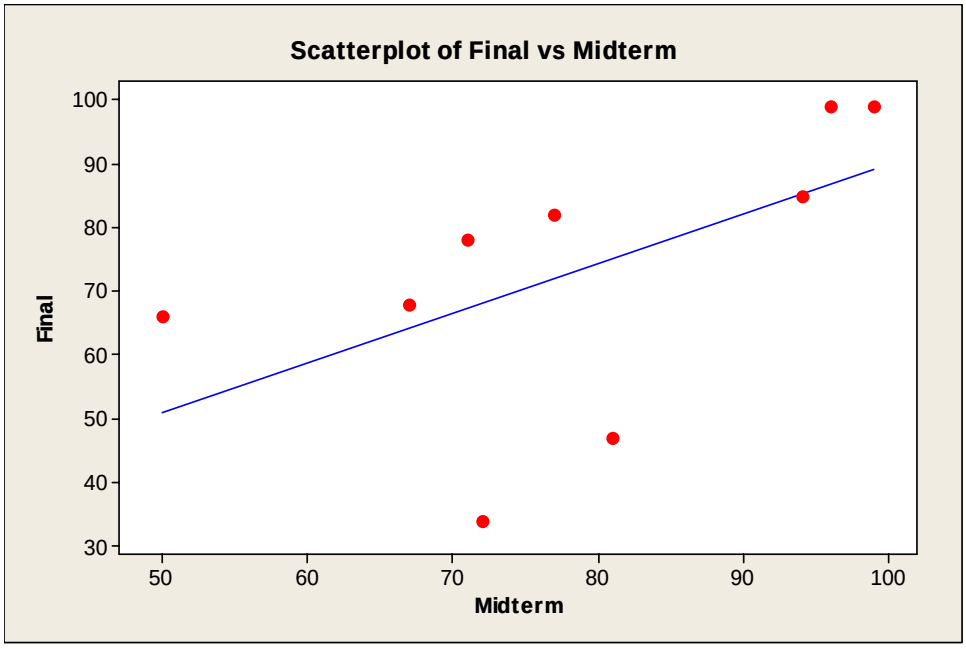
Assumptions: ε is a random variable which is **normally distributed** with mean $\boldsymbol{E(\varepsilon) = 0}$ and variance $\boldsymbol{V(\varepsilon) = \sigma^2}$ (constant). The errors from one x value to another are **independent**. The constant σ^2 is called the **error** variance or the **residual** variance.

The fitted regression line

- We want to estimate the unknown parameters α and β which are called the **regression coefficients**.
- Let $\hat{\alpha} = a$ and $\hat{\beta} = b$, then $\hat{Y} = a + bx$ is called the **fitted (estimated or predicted) regression line** which should be as close as possible to the true regression line for the population.
- The **scatter diagram** is a graphical representation, of the relationship between the two dependent variables X & Y, which can be obtained by plotting the ordered pairs (x_i, y_i) from the sample to see whether a linear relationship between X & Y exists.

Ex.1: problem (2/358). Let X: midterm grade, Y: final grade.





Note: The errors that are above the line are positive and the ones below are negative and they sum to 0.

11.3 The Least Squares Estimate and the Fitted Model

Objectives:

1. **To define the residuals.**
2. **To find the LSE's of α and β.**

Definition: Given a **set of regression data** $\{(x_i,y_i),\ i = 1, 2, \ldots, n\}$ and a fitted model $\hat{Y}_i = a + bx_i$, then the i^{th} **residual** (error or difference) e_i is defined as $e_i = y_i - \hat{y}_i = y_i - a - bx_i$, $i = 1, 2, \ldots, n$, where $e_i = \hat{\varepsilon}_i$. Note that $\sum_{i=1}^{n} e_i = 0$.

The Method of Least Squares

The two estimates a and b will be found so that the sum of the squared residuals, denoted by *SSE*, is minimum, i.e. the function $SSE = Q(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$ is to be minimized for a and b simultaneously.

Definition: Given $\{(x_i, yi),\ i = 1, 2\ldots\ n\}$, then the LSE's of α and β are:

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

and $\hat{\alpha} = a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$, and so, $\hat{y} = a + bx$ is called the **best** fitted (predicted) regression line.

Ex.1 (11.1/357): Estimate the regression line for table 11.1, find the predicted value for x = 31%, and find the error in prediction (estimate error).

Ex.2: Problem (5/359).

11.12 The Correlation

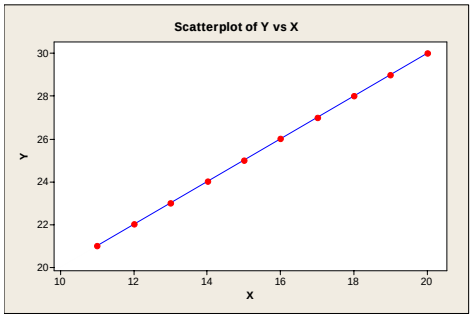
Objectives:

1. **To define the Pearson (simple) correlation coefficient.**
2. **To define the coefficient of determination.**
2. **To test hypotheses about the correlation coefficient.**

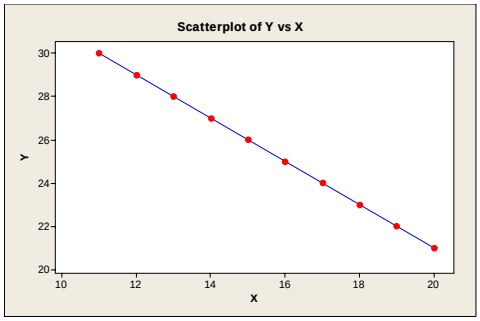
Definition: The **Pearson product-moment correlation coefficient** (simple correlation coefficient) is a numerical measure of the linear association between two dependent r.v.'s X and Y, denoted by ρ, which can be estimated by the sample correlation coefficient (***r***), such that $r = b\sqrt{\frac{s_{xx}}{s_{yy}}} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$.

Notes:

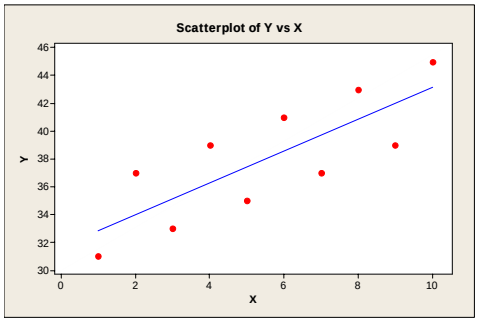
1. $-1 \le r \le +1 \rightarrow 0 \le r^2 \le +1$.
2. $r = +1 \rightarrow$ the correlation is **perfect** (complete) **positive** (direct).
3. $r = -1 \rightarrow$ the correlation is **perfect** (complete) **negative** (inverse).
4. $r = 0 \rightarrow$ there is no **linear** correlation between the r.v.'s X & Y.
5. $r^2 = b^2\frac{s_{xx}}{s_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = R^2 =$ **sample coefficient of determination** = the percentage of variation in the response variable Y that is explained by the variation in the predictor X.



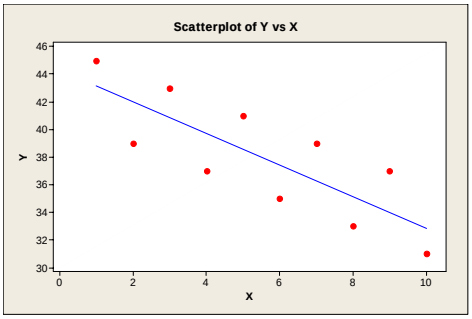
$r = +1$



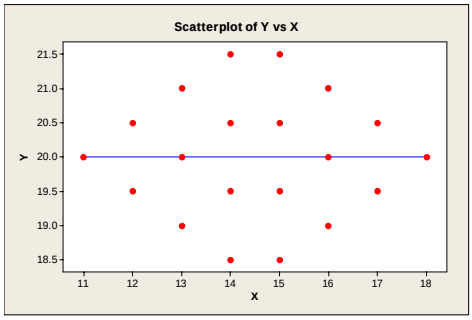
$r = -1$



$r = +0.802$

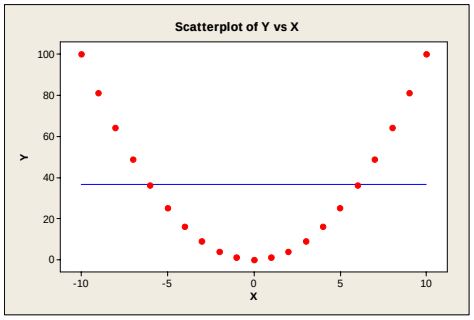


$r = -0.802$



$r = 0$

(No correlation at all)



$r = 0$

(No linear but quadratic correlation)

Ex.1: Problem (1/396). Find the correlation coefficient between X & Y, the coefficient of determination, and interpret its value.

Testing about ρ

For testing H_0: $\rho = 0$ vs. H_1: ρ (>, <, or ≠) 0, the following test statistic $T_0 = r\sqrt{\frac{n-2}{1-r^2}}$: t_{n-2} is used which is equivalent to another test statistic.

Ex.2: Problem (2/396). Refer to ex.1 and Test H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$.

11.4 Properties of the Least Squares Estimates

Objectives:

1. **To introduce the properties of the LSE's.**
2. **To define the mean square error.**

If the regression model is $Y_i = \alpha + \beta X_i + \varepsilon_i$ such that ε is a r.v. with mean 0 and constant variance σ^2, where $\hat{\alpha} = a$ and $\hat{\beta} = b$, then the values of a and b may change from sample to sample. So, the point estimators for α and β

are denoted by A and B respectively. In addition, $Y_1, Y_2, \ldots, Y_n \sim N(\mu_{Y|X_i}, \sigma^2_{Y|X_i})$, where $\mu_{Y|X_i} = \alpha + \beta X_i$ and $\sigma^2_{Y|X_i} = \sigma^2$.

Properties of LSE's:

1. Given a bivariate r.s. $\{(x_i,y_i),\ i = 1, 2, \ldots, n\}$, $\hat{\alpha}$ and $\hat{\beta}$ are both unbiased, i.e. $E(\hat{\alpha}) = E(A) = \alpha$ and $E(\hat{\beta}) = E(B) = \beta$.
2. $\sigma^2_{\hat{\alpha}} = \sigma^2_A = \dfrac{\sum x^2}{n\sum (x-\bar{x})^2}\sigma^2 = \dfrac{\sigma^2\sum x^2}{n\sum (x-\bar{x})^2} = \dfrac{\sigma^2\sum x^2}{ns_{xx}}$.
3. $\sigma^2_{\hat{\beta}} = \sigma^2_B = \dfrac{\sigma^2}{\sum (x-\bar{x})^2} = \dfrac{\sigma^2}{\sum x^2 - n\bar{x}^2} = \dfrac{\sigma^2}{s_{xx}}$.

Theorem: An unbiased estimator of σ^2, i.e. the **Mean Square Error**, is

$$MSE = S_e^2 = \frac{SST - SSR}{n-2} = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{s_{yy} - bs_{xy}}{n-2}.$$

Taking the square root of *MSE* gives $\boldsymbol{S_e}$ which is called the **standard erro**r of the regression model.

11.5 Inferences Concerning the Regression Coefficients

Objectives:

1. To construct CI's about the regression coefficients.
2. To test hypotheses about the regression coefficients.
3. To revisit the coefficient of determination.

Given that $\varepsilon_i \sim N(0, \sigma^2)$ and $Y_i \sim N(\alpha+\beta X_i, \sigma^2) \rightarrow A \sim N(\alpha, \sigma_A{}^2)$ and $B \sim N(\beta, \sigma_B{}^2)$.

I. Inference about β

Definition: A (1-α)100% C.I. for the slope β of the regression line is

$$b \pm t_{\alpha/2,n-2}\frac{s_e}{\sqrt{s_{xx}}} \text{ where } s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{s_{yy} - bs_{xy}}{n-2}} \text{ which}$$

is called the **standard error** of the regression line, and the quantity $SE(\hat{\beta}) = SE(B) = \dfrac{S_e}{\sqrt{s_{xx}}}$ is called the **standard error (of *B*)** for estimating the slope of the regression line β.

Ex.1 (11.2/364): Refer to Ex. 11.1 and find a 95% C.I. for β.

Definition: For testing H_0: $\beta = \beta_0$ vs. H_1: β (>, <, or ≠) β_0, the test statistic

$$t_0 = \frac{B - \beta_0}{S_e/\sqrt{s_{xx}}} \text{ has n-2 degrees of freedom.}$$

The Decision Rule:

First: Using the classic (critical-value) approach;

1. (Right-Tailed) For H_1: $\beta > \beta_0$, Reject H_0 if $t_0 > t_{\alpha,n-2}$.
2. (Left-Tailed) For H_1: $\beta < \beta_0$, Reject H_0 if $t_0 < -t_{\alpha,n-2}$.
3. (Two-Tailed) For H_1: $\beta \neq \beta_0$, Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$.

Second: Using the *p*-value approach; Reject H_0 if *p*-value < α, where;

1. (Right-Tailed) For H_1: $\beta > \beta_0$, *p*-value = $P(t_{n-2} > t_0)$.
2. (Left-Tailed) For H_1: $\beta < \beta_0$, *p*-value = $P(t_{n-2} < t_0)$.
3. (Two-Tailed) For H_1: $\beta \neq \beta_0$, *p*-value = $P(|t_{n-2}| > |t_0|)$.

Third: Using the C.I. approach, only for **two-tailed test**; Reject H_0 if β_0 is outside a $(1-\alpha)$ 100% C.I. for β.

Ex.2 (11.3/364): Refer to Ex. 11.1 and test H_0: $\beta = 1$ vs. H_1: $\beta < 1$.

Note: The most important test is H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, because the acceptance of H_0 means that there is NO significant linear correlation between X & Y. it is exactly equivalent to testing H_0: $\rho = 0$ vs. H_1: $\rho \neq 0$ in the sense that the two test statistics values are equal for the same sample.

II. Inference about α

Definition: A $(1-\alpha)100\%$ C.I. for the slope α of the regression line is $a \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{\sum x^2}{n s_{xx}}}$. So, $SE(\hat{\alpha}) = SE(A) = s_e \sqrt{\frac{\sum x^2}{n s_{xx}}}$ is called the **standard error (of *A*)** for estimating the y-intercept of the regression line α.

Ex.3 (11.4/366): Refer to Ex. 11.1 and find a 95% C.I. for α.

Definition: For testing H_0: $\alpha = \alpha_0$ vs. H_1: α (>, <, or $\neq$) α_0, the test statistic $t_0 = \frac{A - \alpha_0}{S_e \sqrt{\frac{\sum x^2}{n s_{xx}}}}$ has n-2 degrees of freedom.

The Decision Rule:

First: Using the critical-value; the same as in (I) above.

Second: Using the *p*-value; the same as in (I) above.

Third: Using a C.I. only for **two-tailed test**; Reject H_0 if α_0 is outside a $(1-\alpha)$ 100% C.I. for α.

Ex.4 (11.5/366): Refer to Ex. 11.1 and test H_0: $\alpha = 0$ vs. H_1: $\alpha \neq 0$.

The Coefficient of Determination R^2

Recall that it is a measure for the percentage of variation in the response variable Y that is explained by the variation in the predictor variable X and is denoted by; $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r^2 = b^2 \frac{s_{xx}}{s_{yy}} = \frac{s_{xy}^2}{s_{xx} s_{yy}}$, where $SST = \sum (y - \bar{y})^2 = \sum y^2 - n\bar{y}^2 = s_{yy}$, $SSR = \sum (\hat{y} - \bar{y})^2 = b s_{xy} = \frac{s_{xy}^2}{s_{xx}}$, and consequently $SSE = \sum (y - \hat{y})^2 = SST - SSR = s_{yy} - b s_{xy} = s_{yy} - b^2 s_{xx}$.

Notes:

1. If $R^2 \approx 1$ then the linear fit is perfect and $SSE \approx 0$.
2. If $R^2 \approx 0$ then the linear fit is invalid and $SSE \approx SST$.

Ex.5: Refer to Ex. 11.1 and find R^2.

Ex.6: Problem (7/371). Refer to problem (5/359).

Ex.7: Refer to problem (5/359), find R^2 and interpret its value.

11.6 Prediction

Objectives:

1. **To construct a C.I. for the mean response of *Y* ($\mu_{Y|X_\circ}$).**
2. **To construct a P.I. for the mean response of *Y* ($\mu_{Y|X_\circ}$).**

Definition: The point estimate of $\mu_{Y|X_\circ} = \alpha + \beta X_\circ$ is $\hat{Y}_\circ = A + BX_\circ$, where $\hat{Y}_\circ \sim N(\mu_{\hat{Y}_\circ}, \sigma^2_{\hat{Y}_\circ})$ such that $\mu_{\hat{Y}_\circ} = E(\hat{Y}_\circ) = E(A + BX_\circ) = \alpha + \beta X_\circ = \mu_{Y|X_\circ}$ and $\sigma^2_{\hat{Y}_\circ} = \sigma^2_{A+BX_\circ} = \sigma^2\left(\frac{1}{n} + \frac{(X_\circ - \overline{X})^2}{S_{xx}}\right)$.

Definition: A(1 – α)100% C.I. for the **mean response** $\boldsymbol{\mu_{Y|X_\circ}}$ is given by $\hat{y}_\circ \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_\circ - \bar{x})^2}{s_{xx}}}$.

Ex.1 (11.6/368): Refer to Ex. (11.1/357) and construct a 95% C.I. for $\mu_{Y|X_\circ}$ where x_o = 20%.

Prediction Interval (*P.I.*): is a confidence interval for a future single observed response.

Definition: The r.v. $\hat{Y}_\circ - Y_\circ$ is normally distributed with mean: $\mu_{\hat{Y}_\circ - Y_\circ} = E(\hat{Y}_\circ - Y_\circ) = E(A + BX_\circ - \alpha + \beta X_\circ + \varepsilon)$, and variance $\sigma^2_{\hat{Y}_\circ - Y_\circ} = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_\circ - \overline{X})^2}{S_{xx}}\right)$.

Definition: A(1 – α)100% P.I. for a **single response $\boldsymbol{Y_o}$** is given by $\hat{y}_\circ \pm t_{\alpha/2,n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_\circ - \bar{x})^2}{s_{xx}}}$.

Ex.2 (11.7/368): Refer to Ex. (11.1/357) and construct a 95% P.I. for y_o $\mu_{Y|X_\circ}$ when x_o = 20%.

Ex.3: Problem (14/372). Refer to problem (5/359) and construct both a 95% C.I. for the mean y, and a 95% P.I. for y_o when x_o = 50 C°.