BMMS2074 Statistics for Data Science

Contents:

Reference Books:

1. Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications*. (7th ed.). Thomson.
2. Ken Black (2013). *Applied Business Statistics: Making Better Business Decisions*. (7th ed.). John Wiley.
3. Chris Chatfield. 2004. *The Analysis of Time Series : An Introduction*. 6th Edition. Chapman & Hall.

Chapter 1 : Simple Linear Regression

1.1 Introduction

Suppose we are interested in estimating the average GPA of all students at TAR UC. How would we do this? (Assume we do not have access to any student records.)

- Define the population:
  All TAR UC students

- Define the parameter of interest:
  Let μ be the population mean of the GPA of all TAR UC students.

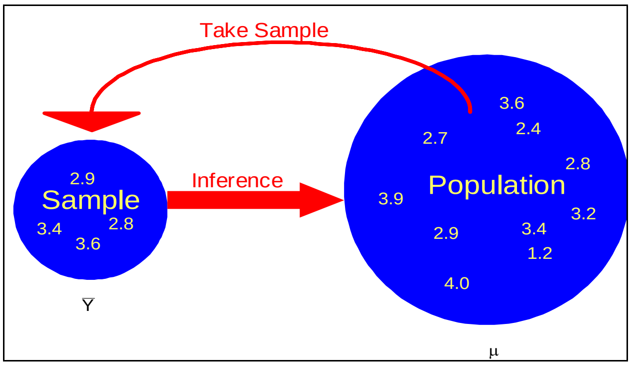
- Take a representative sample from the population:
  Suppose a random sample of 100 students is selected and the GPA of each student are recorded and let Y denote the GPA of all TAR UC students.

- Calculate the sample statistic that estimates the parameter:

  $$\bar{y} = \frac{\sum_{i=1}^{100} y_i}{100}$$ which is the sample mean computed to estimate μ.

- Make an inference about the value of the parameter using statistics:
  Construct confidence intervals or perform hypothesis tests using the sample mean and sample standard deviation computed.

The diagram below demonstrates these steps. Note that not all GPAs could be shown in the diagram.

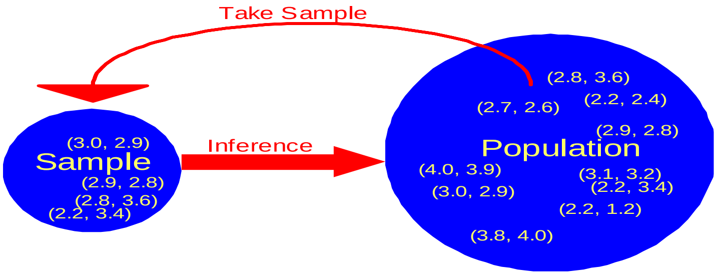


What factors may be related to GPA?

- High school GPA
- Rank in high school class
- Involvement in activities
- High school overall rating
- Etc.

Suppose we are interested in the relationship between college and HS GPA and we want to use HS GPA to predict college GPA. How could we do this?

Use similar steps as on page 1, but now with regression models.



Data shown as: (HS GPA, College GPA)

1.1.1 Scatterplot

The main objective of this chapter is to analyze a collection of paired sample data (or **bivariate data**) and determine whether there appears to be a **relationship** between the two variables.

A set of bivariate data is denoted as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

A ***correlation*** exists between two variables when one of them is related to the other in some way.

A ***scatterplot*** (or ***scatter diagram***) is a graph in which the paired (*x*, *y*) sample data are plotted with a horizontal *x*-axis and a vertical *y*-axis. Each individual (*x*, *y*) pair is plotted as a single point.

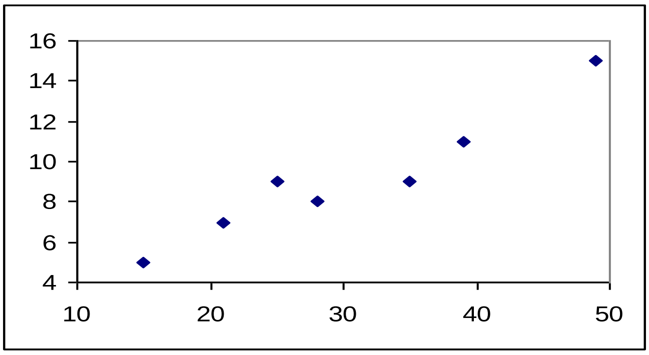
Example

Suppose we take a sample of seven households and collect information on their incomes and food expenditures for the past month. The information obtained (in hundreds of RM) is given below.

Income (hundreds)	35	49	21	39	15	28	25
Food expenditure (hundreds)	9	15	7	11	5	8	9

Solution

The scatter diagram for this set of data is



1.2 Simple Linear Regression Model

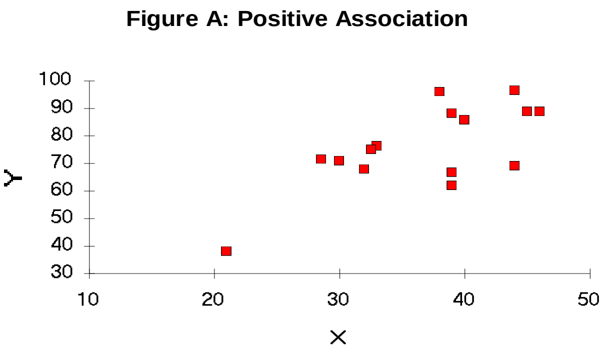
The **response** or **dependent** variable is the response of interest and is usually denoted by Y.

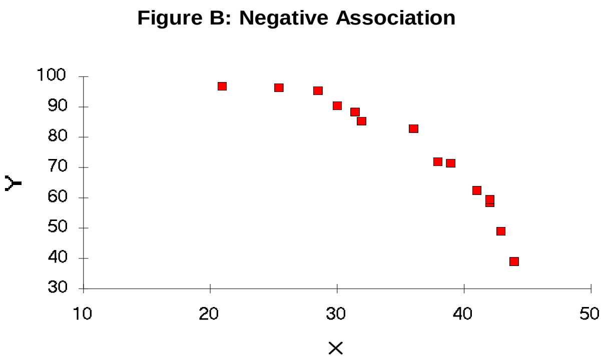
The **explanatory**, **independent or predictor** variable attempts to explain the response and is usually denoted by X.

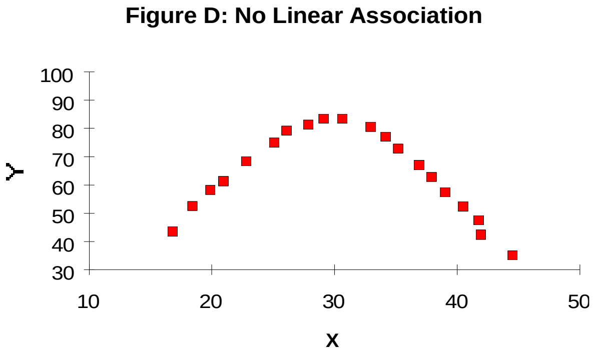
A **scatter plot** shows the relationship between two quantitative variables X and Y. The values of the X variable are marked on the horizontal axis, and the values of the Y variable are marked on the vertical axis. Each pair of observations (x_i, y_i) is represented as a point in the plot.

Two variables are said to be **positively associated** if, as X increases, the value of Y tends to increase. Two variables are said to be **negatively associated** if, as X increases, the value of Y tends to decrease.

Interpretation:
Form, Direction, Strength, Any Deviations







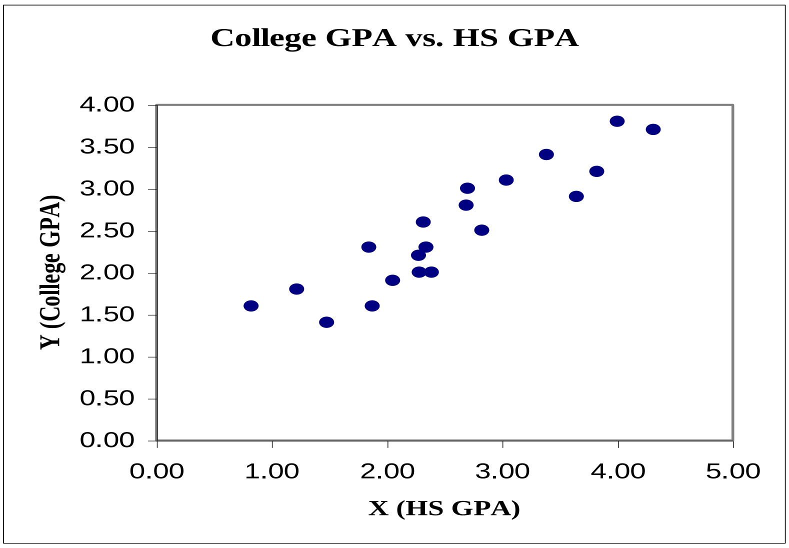
- Figure A: shows a moderately strong, positive, linear relationship.
- Figure B: shows a strong, negative, slightly curved to linear relationship.
- Figure C: shows no association between the two variables.
- Figure D: shows a very strong association, not a linear one, but rather more quadratic (curvilinear).

Example 1.1

A random sample of UTARUC students is taken producing the data set below.

Student	X (HS GPA)	Y (College GPA)
1	x_1=3.04	y_1=3.10
2	x_2=2.35	y_2=2.30
3	2.70	3.00
. . .	. . .	. . .
18	4.00	3.80
19	2.28	2.20
20	1.88	1.60

Scatter plot of the data:



It shows fairly strong positive linear association between College GPA and HS GPA.

A **functional relation** between two variables is expressed by a *mathematical formula*. If X denotes the independent variable and Y the dependent variable, a functional relation is of the form: $Y = f(X)$. Given a particular value of X, the function f indicates the corresponding value of Y. All values fall directly on the line of functional relationship.

Statistical relation (Regression) between two variables is *not* a perfect fit. In general, the observations do not fall directly on the curve of relationship.

Example:

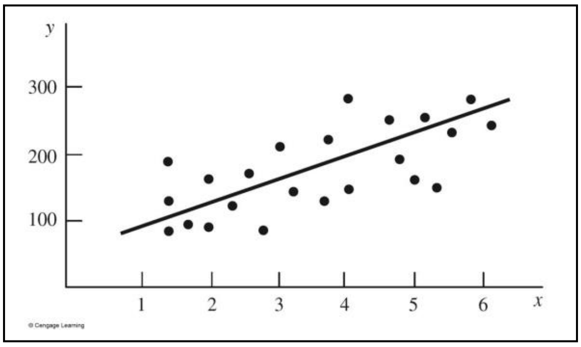
Consider the relation between dollar sales (Y) of a product sold at a fixed price and number of units sold (X). If the selling price is RM2 per unit, the relation is expressed by the equation:

$$Y = 2X$$

Number of Units Sold, X	Sales, Y (RM)
75	150
25	50
130	260

Example:

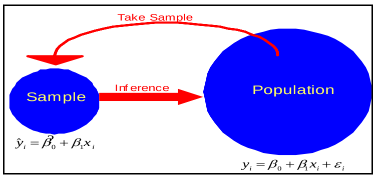
Performance evaluations for 23 employees were obtained at midyear (0 – 10 scale) and at year-end (0 – 400 points). These data are plotted in the following figure.



The figure clearly suggests that there is a positive linear relation between midyear and year–end evaluation. However, the relation is not a perfect fit. The scattering of the points suggesting that some of the evaluations is not accounted for by midyear performance assessments. For instance, two employees had midyear evaluation of $x = 4$, and yet they received different year–end evaluation.

1.2.1 Formal Statement of Model

Suppose you are interested in studying the relationship between two variables X and Y.


Take Sample
Sample
Inference
Population
$\hat{y}_i = \hat{\beta}_0 + \beta_1 x_i$
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

The population model can be stated as follows:

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

where:

(i) y_i is the value of the response (dependant) variable in the i^{th} trial/observation.

(ii) The regressor x_i is a known constant (fix). (i.e. the value in the predictor (independent) variable in the i^{th} trial).

(iii) The intercept β_0 and the slope β_1 are unknown constants (parameters).

(iv) ε_i is the random error

Assumptions

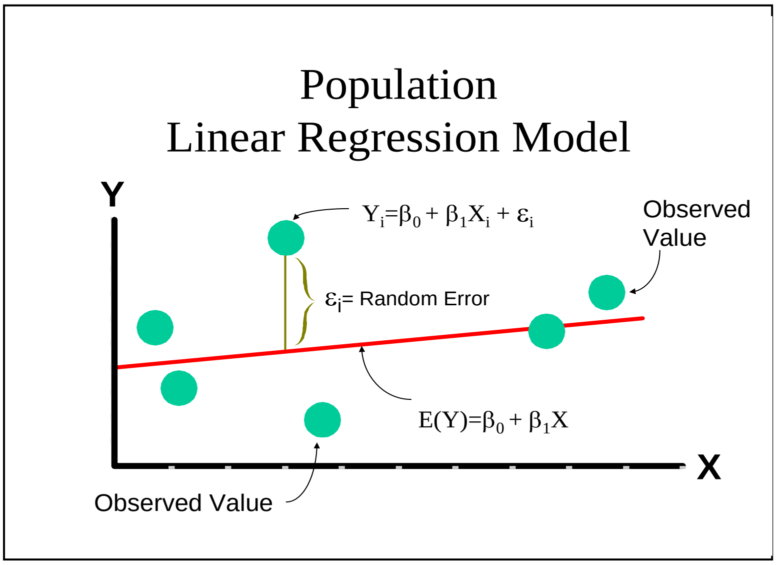
1. The error terms ε_i are normally and independently distributed with $E(\varepsilon_i) = 0$ and constant variance $Var(\varepsilon_i) = \sigma^2$.
2. The error (thus, the y_i, also) are uncorrelated with each other.

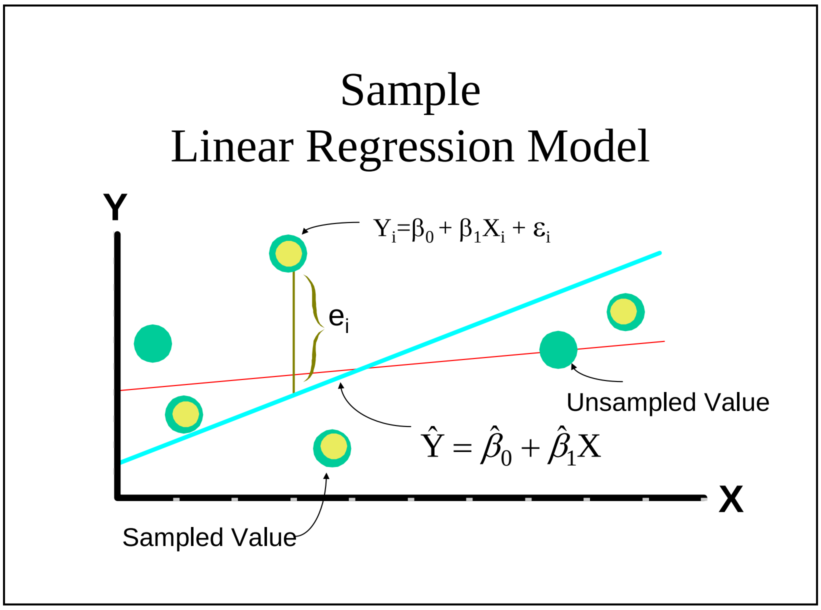
   $\therefore \varepsilon_i \sim NID(0, \sigma^2)$
3. $E(Y \mid x) = \beta_0 + \beta_1 x$; $Var(Y \mid x) = \sigma^2$

Note:

- The above model is said to be *simple, linear in the parameters*, and *linear in the predictor variables*.
- It is "simple" in that there is only one predictor variable, "linear in the parameters" because no parameter appears as an exponent or is multiplied or divided by another parameter, and "linear in the predictor variable," because this predictor variables appears only in the first power.
- A model that is linear in the parameters and in the predictor variable is also called a *first–order model*.

The parameters β_0 and β_1 are unknown and can be estimated using n pairs of sample data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.




Sample
Linear Regression Model
Y
$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$
e_i
Unsampled Value
$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$
X
Sampled Value

1.3 Least Square Estimation of the parameters

The line that minimizes the sum of squares of the deviations of observed values of y_i from those predicted is the best–fitting line.

The least–squares (LS) criterion is defined as

$$S\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} e_i^2 = \sum \left(y_i - \hat{y}\right) = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

which is the error sum of squares.

As discussed, the best–fitted line is that one which minimized LS, that is

$$\frac{\partial S}{\partial \hat{\beta}_0} = 0 \quad \text{and} \quad \frac{\partial S}{\partial \hat{\beta}_1} = 0$$

The solution for the LS estimators, $\hat{\beta}_0$ and $\hat{\beta}_1$, is:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ and } \hat{\beta}_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{S_{XY}}{S_{XX}}$$

where $S_{XX} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$ and $S_{XY} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$.

Note that $S_{YY} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$ (will be used later.)

A test of the second partial derivatives will show that a minimum is obtained with the least squares estimators of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Thus, the fitted simple linear regression model (estimated regression equation or line) is

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

Properties of LSE:

(a) It can be shown that $\hat{\beta}_1 = \dfrac{\sum\limits_{\text{i=1}}^{\text{n}}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sum\limits_{\text{i=1}}^{\text{n}}(x_i - \bar{x})y_i - \sum\limits_{\text{i=1}}^{\text{n}}(x_i - \bar{x})\bar{y}}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sum\limits_{\text{i=1}}^{\text{n}}(x_i - \bar{x})y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \sum k_i y_i$

where $k_i = \dfrac{x_i - \bar{x}}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$. Since the k_i are known constants, is $\hat{\beta}_1$ a linear combination of the y_i and hence is a linear estimator. In the similar fashion, it can be shown that $\hat{\beta}_0$ is a linear estimator as well.

(b) Some interesting properties of coefficient k_i:

(i) $\sum k_i = 0$

(ii) $\sum k_i x_i = 1$

(iii) $\sum k_i^2 = \dfrac{1}{\sum (x_i - \bar{x})^2} = \dfrac{1}{S_{XX}}$

(c) The LSEs are unbiased estimators of the parameters β_0 and β_1.

$E(\hat{\beta}_0) = \beta_0$ $\qquad$ $E(\hat{\beta}_1) = \beta_1$

(d) The variance of the LSEs are

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)$$

(e) **Gauss–Markov Theorem** stated that:

Under the conditions of regression, the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators. i.e. BLUE –Best Linear unbiased Estimators.

i. $\hat{\beta}_0$ is the BLUE of β_0.

ii. $\hat{\beta}_1$ is the BLUE of β_1.

iii. $c_1\hat{\beta}_0 + c_2\hat{\beta}_1$ is the BLUE of $c_1\beta_0 + c_2\beta_1$.

It can also be shown that $\hat{y}$ is the BLUE of $E(Y)$

Properties of the Fitted Regression Model

1. The difference between the observed value y_i and the corresponding fitted value $\hat{y}$ is a **residual**:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

and the sum of the residual is always zero:

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

2. The LS regression line always passes through the centroid $(\bar{x}, \bar{y})$ of the data.

3. The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

4. The sum of the residuals weighted by the corresponding value of the regressor variable always equals zero.

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

5. The sum of the residuals weighted by the corresponding fitted values always equals zero.

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

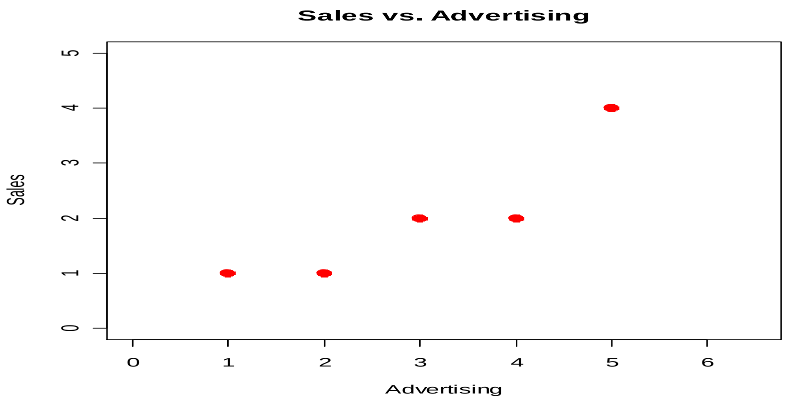
Example 1.2

What is the relationship between sales and advertising costs for a company?

Let X be the advertising costs in RM100,000.
Let Y be the sales units (in 10,000 units)
Assume monthly data below and independence between monthly sales.

	x	y	x^2	y^2	xy
	1	1	1	1	1
	2	1	4	1	2
	3	2	9	4	6
	4	2	16	4	8
	5	4			
Σ					

$\therefore$ The regression line is
The corresponding scatterplot:



Example 1.3

(a) What do the estimated parameters in *Ex. 1.2* mean?

(b) What are the estimated sales when the advertising cost is RM100,000 and RM250,000, respectively?

Extrapolation is using the regression line to predict the value of a response corresponding to a x value that is outside the range of the data used to determine the regression line. Extrapolation can lead to unreliable predications

Example 1.4: Childhood Growth

The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female during childhood, from four to nine years old, were compiled and the least squares regression line was obtained as $\hat{y} = 31.496 + 2.3622x$, where Y denotes height in inches and X denotes age in years.

(a) Interpret the value of the estimated slope $\hat{\beta}_1 = 2.3622$.

(b) Would interpretation of the value of the estimated Y-intercept, $\hat{\beta}_0 = 31.496$, make sense here? If yes, interpret it. If no, explain why not.

(c) What would you predict the height to be for a female at 8 years old?

(d) What would you predict the height to be for a female at 25 years old?

(e) Why do you think your answer to part (d) was so inaccurate?

1.4 Estimating σ^2

Population simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{where } \varepsilon_i \sim NID(0, \sigma^2)$$

σ^2 measures the variability of the ε_i.

σ^2 can be estimated based on the residual or error sum of squares where

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

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

This unbiased estimator of σ^2 is called the **residual mean square** and its square root is called **standard error of regression**.

Note:

(a) SS_E has $n-2$ degrees of freedom associated with it. Two degrees of freedom are lost due to the estimation of $\hat{\beta}_0$ and $\hat{\beta}_1$ (remember that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$).

(b) $SS_E = S_{YY} - \hat{\beta}_1 S_{XY}$

Example 1.5: Sales and Advertising

	x_i	y_i	$\hat{y}_i = -0.1 + 0.7x$	$e_i = (y_i - \hat{y}_i)$	$e_i = (y_i - \hat{y}_i)^2$
	1	1	0.6	0.4	0.16
	2	1	1.3	–0.3	0.09
	3	2	2.0	0	0
	4	2	2.7	–0.7	0.49
	5	4			

1.5 Correlation

The *linear correlation coefficient*, r (or R), (is also called the *Pearson product moment correlation coefficient*) measures the strength of the **linear** relationship between the paired *x*- and *y*-quantitative values in a sample. It describes the direction of the linear association and indicates how closely the points in a scatter plot are to the least squares regression line

1.5.1 The formula

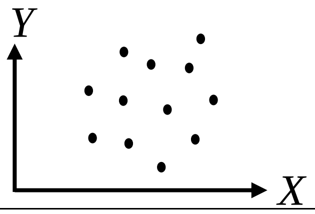
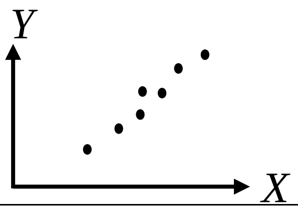
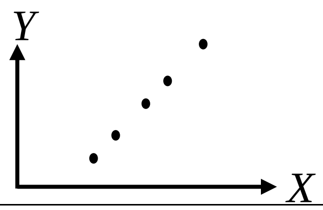
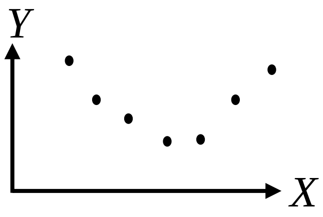
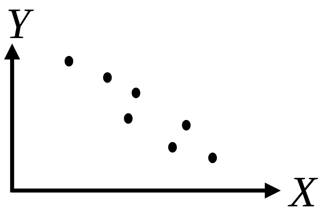
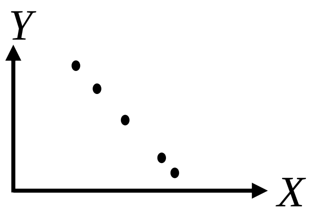
$$\boxed{r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\,\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}}$$

Properties of the linear correlation coefficient, r,

1. The value of r is always between –1 and 1 inclusive. That is $-1 \le r \le 1$.
2. r measures the strength of a *linear relationship*. It is not designed to measure the strength of a relationship that is not linear.
3. If $r = 0$, then there is no **linear relationship** between the two variables
4. $0 \le r^2 \le 1$

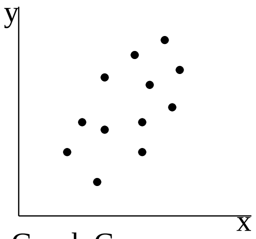
Degree of correlation	Positive correlation	Negative correlation
Perfect	+1	–1
Strong	$0.8 \le r < 1.0$	$-1.0 < r \le -0.8$
Moderate	$0.4 \le r < 0.8$	$-0.8 < r \le -0.4$
Weak	$0 < r < 0.4$	$-0.4 < r < 0$
Absent	0	0

Scatter diagrams and correlation

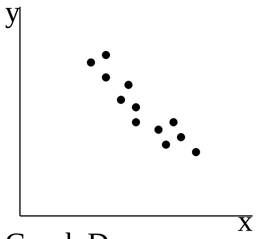
		
No relationship	positive linear correlation	perfect positive linear correlation
		
Non–linear relationship	negative linear correlation	perfect negative linear correlation

Example 1.6:

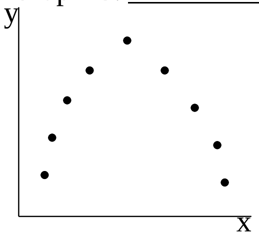
Graph A: ____________



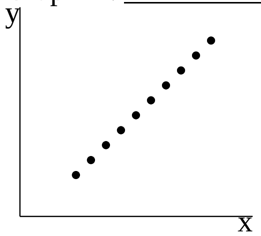
Graph B: ____________



Graph C: ____________



Graph D: ____________



Example 1.7:

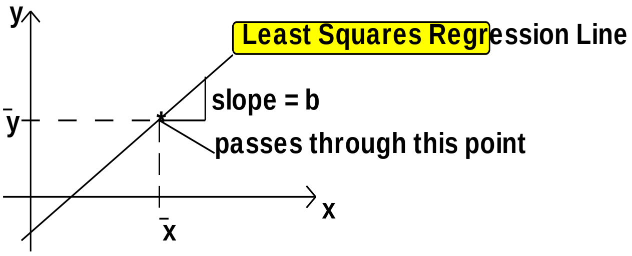
Compute the correlation coefficient r for Test 1 versus Test 2

	x	y	x^2	y^2	xy
	8	9	64	81	72
	10	13			
	12	14	144	196	168
	14	15	196	225	210
	16	19	256	361	304
Σ	60	70		1032	884

1.5.3 Relationship between r and the slope

$$\hat{\beta}_1 = r\left(\frac{s_Y}{s_X}\right)$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}$$



Note: The least squares regression line always passes through the point $(\bar{x}, \bar{y})$.

Example 1.8:

The scores on the midterm and final exam for 500 students were obtained. The possible values for each exam are between 0 and 100. The least squares regression line for predicting the final exam from the midterm exam was obtained for these data. Suppose the correlation coefficient is 0.5 for these data, $r = 0.5$.

Susan, a student in this class, received a midterm score that was one standard deviation above the average midterm score. Suppose the average and standard deviation for the midterm scores were 80 and 10, respectively. Also suppose that the average and standard deviation for the final exam scores were 60 and 20, respectively.

Predict Susan’s final exam score

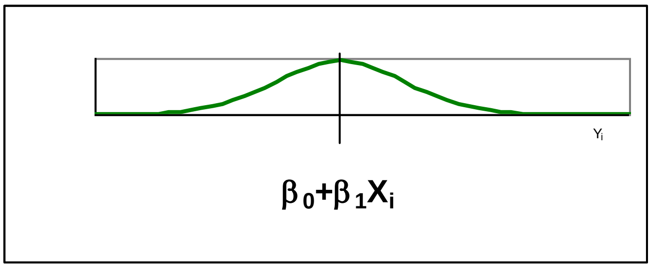
1.6 Assumptions of the error

From $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_i \sim NID(0, \sigma^2)$

This implies that

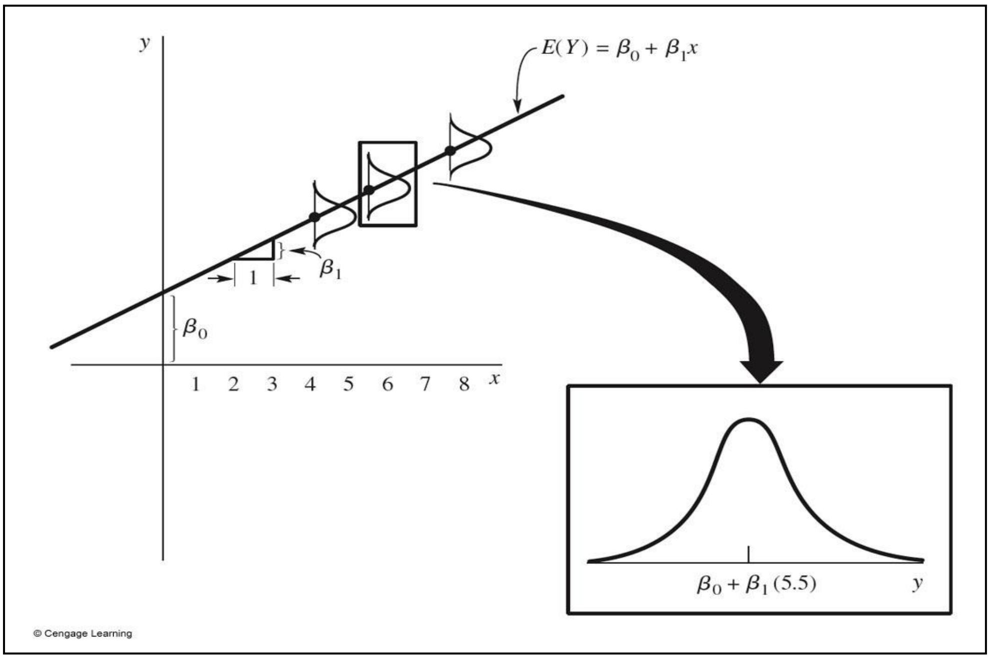
1. The y_i's are independent
2. $E(y_i) = \beta_0 + \beta_1 x_i$
3. $Var(y_i) = \sigma^2$
4. y_i follow a normal distribution

Since x_i β_0, and β_1 are assumed to be constant in the regression model.
Thus for a particular x_i value:



Note:

1) There is a probability distribution of Y for each level of X .
2) The means of these probability distributions vary in some systematic fashion with X .
3) All the probability distributions of y_i exhibit the same variability, σ^2, in conformance with the assumptions of simple regression model .



Thus, the response Y_i, when the level of X in the ith trial is X_i, comes from a probability distribution whose mean is:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

1.7 Inferences concerning β_1 and β_0

1.7.1 The sampling distribution for $\hat{\beta}_1$

Population regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

From the assumption that $\varepsilon_i \sim NID(0, \sigma^2)$, we have $y_i \sim NID(\beta_0 + \beta_1 x_i, \sigma^2)$

Since $\hat{\beta}_1 = \sum k_i y_i \sim NID(\beta_1, \dfrac{\sigma^2}{S_{XX}})$

To test the hypothesis that the slope equals a constant, we have $H_0 : \beta_1 = \beta_{10}$ and the test statistic is $z = \dfrac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{XX}}} \sim N(0,1)$.

If σ^2 is unknown and the unbiased estimator MS_E and the test statistic becomes

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_E / S_{XX}}} \sim t(n-2)$$

The term $se(\hat{\beta}_1) = \sqrt{MS_E / S_{XX}}$ is the (estimated) standard error of $\hat{\beta}_1$.

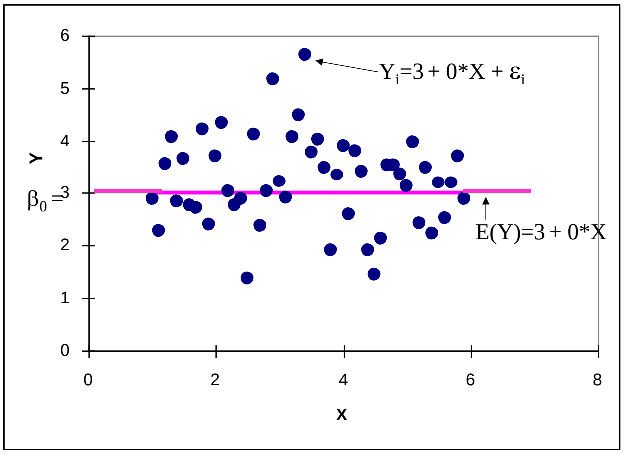
1.7.2 Special Case: Testing Significance of Regression

Test for

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Suppose $\beta_1 = 0$, then $y_i = \beta_0 + 0.x_i + \varepsilon_i = \beta_0 + \varepsilon_i$

Example plot: Suppose $\beta_0 = 3$.



Note:

1. The hypothesis testing can be done via (i) $t-$test or $z-$test, or (ii) analysis of variance (ANOVA)

2. Accept null hypothesis H_0 indicate that there is no linear relationship between X and Y which means
   (i) X is of little value in explaining the variation of Y
   (ii) The true relationship between X and Y is not linear.
3. Reject null hypothesis H_0 indicate that
   (i) X is of value in explaining the variation of Y
   (ii) The straight-line model is adequate or better results could be obtained with addition of higher order polynomial terms in X .

Example 1.9:
For *Ex. 1.2,* Is advertising linearly related to sales? Use $\alpha = 0.05$.

1.7.3 Hypothesis test on β_0

To test the hypothesis that the intercept equals a constant, we have $H_0 : \beta_0 = \beta_{00}$ and the test statistic is $t = \dfrac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_E\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{XX}}\right)}} \sim t(n-2)$,

The term $se(\hat{\beta}_0) = \sqrt{MS_E\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{XX}}\right)}$ is the (estimated) standard error of $\hat{\beta}_0$.

Example 1.10:
Refer to *Ex.1.7,* Test 1 vs Test 2.

(a) What is the regression line that relate Test 1 to Test 2.

(b) Is there sufficient evidence to conclude that a linear relationship exists between Test 1 and Test 2? Use $\alpha = 0.05$.

(c) Test whether there is a direct(positive) relationship between Test 1 and Test 2. Use $\alpha = 0.05$.

(d) The following test has no practical significance in this problem. Test whether the intercept is zero.

1.7.4 Interval Estimation in Simple Linear Regression

$(1-\alpha)100\%$ confidence interval (C.I.) for β_1 is

$$\hat{\beta}_1 - t_{\alpha/2;n-2}se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2;n-2}se(\hat{\beta}_1)$$

$(1-\alpha)100\%$ confidence interval (C.I.) for β_0 is

$$\hat{\beta}_0 - t_{\alpha/2;n-2}se(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2;n-2}se(\hat{\beta}_0)$$

$(1-\alpha)100\%$ confidence interval (C.I.) for σ^2 is

$$\frac{(n-2)MS_E}{\chi^2_{\alpha/2;n-2}} \le \sigma^2 \le \frac{(n-2)MS_E}{\chi^2_{1-\alpha/2;n-2}}$$

Example 1.11:
At a used car dealership, let X be an independent variable representing the age in years of a motorcycle and Y be the dependent variable representing the selling price of a motorcycle. Find a 95% confidence interval for β_1.

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$(x_i-\bar{x})^2$	$(y_i-\hat{y})^2$
	5 10 12 14 15	500 400 300 200 100	25 100 144 196 225	250000 160000 90000 40000 10000	2500 4000 3600 2800 1500	38.44 1.44 0.64 7.84 14.44	1367.52 2923.56 929.64 47.75 3011.81
$\sum$	56	1500	690	550000	14400	62.8	8280.28

With 95% confidence, we estimate that the change in the mean of the selling price (decrease) of a motorcycle when the age in years of a motorcycle increase by one unit, is somewhere between \$17.12 and \$59.32.

Note:
The resulting 95% confidence interval is -59.32 to -17.12. Since the interval does not contain 0, you can conclude that the true value of β_1 is not 0, and you can reject the null hypothesis $H_0:\beta_1=0$ in favor of $H_1:\beta_1\neq 0$. Furthermore, the confidence interval estimate indicates that there is a decrease of \$17.12 to \$59.32 in selling price for each year increase in the age of the motorcycle.

1.7.5 Some considerations on making inferences concerning β_1

The sampling distributions for $\hat{\beta}_1$ hold true ONLY if the assumption that ε_i ~ $NID(0,\sigma^2)$ holds; however, if the normality assumption does not hold, these sampling distributions are usually still good approximations.

$\hat{\beta}_1$ are somewhat "robust" against normality.

1.8 Estimating the Mean Response

Let x_h be any value of the regressor variable within the range of the original data X used to fit the model. (Note that x_h may or may not be one of the values in the sample.) The mean response $E(Y|\,x_h) = \mu_{Y|\,x_h} = E(Y_h)$ can be estimated by

$$\hat{E}(Y|\,x_h) = \hat{\mu}_{Y|\,x_h} = \hat{E}(Y_h) = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

What is the difference between $\hat{y}_h$ and $\hat{E}(Y_h)$ for a given value x_h?

Example 1.12:
Let X be the score for Quiz 1;
Let Y be the score for Quiz 2;
The data collected are as follows:

Quiz 1	0	2	4	6	8
Quiz 2	6	5	8	7	9

We obtained:
$\hat{\beta}_0$ = 5.4; $\hat{\beta}_1$=0.4

If we want to estimate the mean Quiz 2 score for all students in the population who score a 6 (i.e. $x_h = 6$) on Quiz 1, then the estimate will be

$$\hat{E}(Y_h) = \hat{\beta}_0 + \hat{\beta}_1 x_h = 5.4 + 0.4(6) = 7.8$$

On the other hand, we may want to predict the Quiz 2 score for a student who score a 6 (i.e. $x_h = 6$) on Quiz 1, then the estimate will be

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = 5.4 + 0.4(6) = 7.8 \text{ .}$$

1.8.1 Confidence Interval for a Mean Response

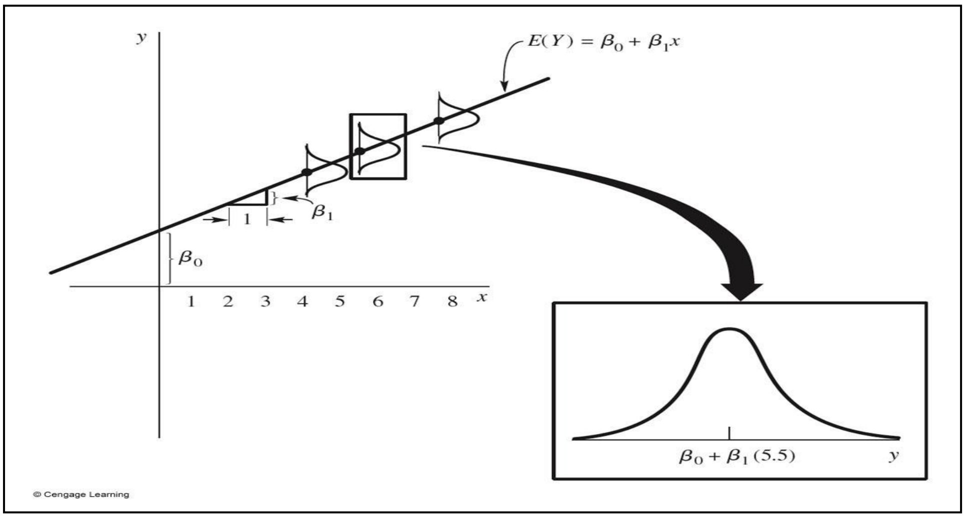
Note that the variance of $\hat{E}(Y_h)$ is

$$Var[\hat{E}(Y_h)] = \sigma^2\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]$$

$(1-\alpha)100\%$ CI on the mean response at the point for $x = x_h, \mu_{Y|\,x_h}$ is

$$\hat{E}(Y_h) - t_{\alpha/2;n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]} \le \mu_{Y|x_h} \le \hat{E}(Y_h) + t_{\alpha/2;n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]}$$

Example 1.13:
Consider the data in *Ex. 1.12.* Construct a 95% confidence interval for the mean Quiz 2 score for all students who scored 6 on Quiz 1.



1.9 Prediction of a new observation

If x_h is the value of the regressor of interest, the point estimate of the new value of the response, y_h is $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$

Note that the random variable

$$\Psi = Y_h - \hat{Y}_h \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]\right)$$

1.9.1 Prediction Interval for an Individual Response

$(1-\alpha)100\%$ prediction interval (PI) on the future observation Y_h at a specified value of $x = x_h$ is

$$\hat{Y}_h - t_{\alpha/2;n-2}\sqrt{MS_E\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]} \le Y_h \le \hat{Y}_h + t_{\alpha/2;n-2}\sqrt{MS_E\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{S_{XX}}\right]}$$

Example 1.14:

Consider the data in *Ex. 1.12*. Compute a 95% prediction interval for an individual student who scores 6 on Quiz 1.

With 95% confidence, a student with a score of 6 on Quiz 1 should expect between a 3.83 and 11.77 score on Quiz 2.

Note:

Prediction intervals resemble confidence intervals. However, they differ conceptually:

(i) A CI represents an inference on a parameter and is an interval that is intended to cover the value of the parameter.

(ii) A PI is a statement about the value to be taken by a random variable, the new observation Y_h.

1.10 Analysis of Variance Approach to Regression Analysis

The *total sum of squares*, denoted by SS_T is given by,

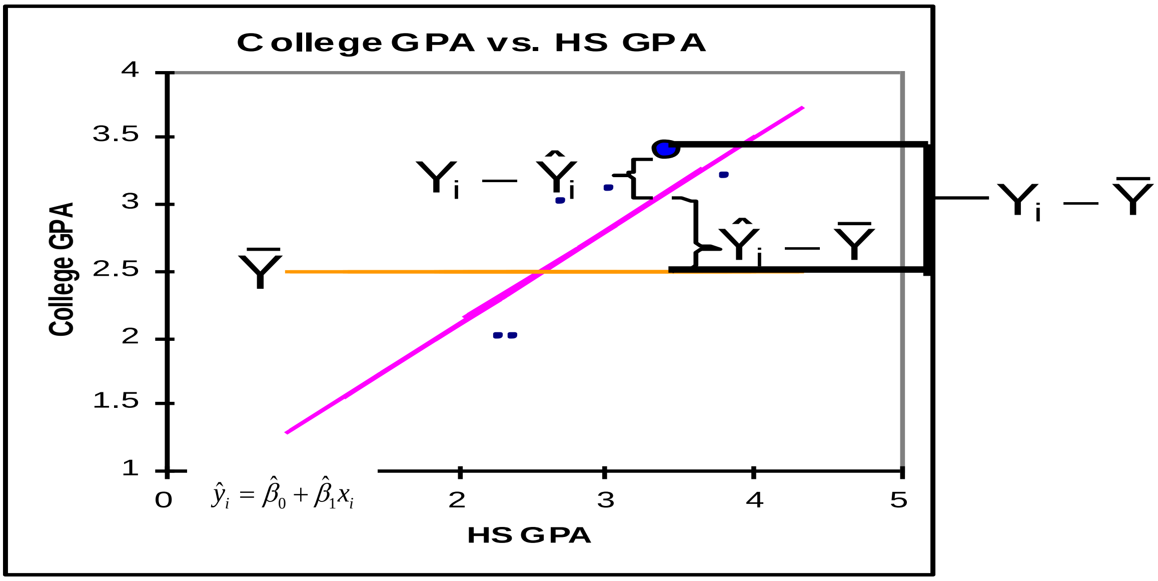
$$SS_T = \Sigma(y-\bar{y})^2$$
$$= S_{YY} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

The *regression* or *model sum of squares*, denoted by SS_R is given as

$$SS_R = \Sigma(\hat{y}_i - \bar{y})^2$$

and $$SS_T = SS_R + SS_{\text{Res}}$$

Example 1.15: College and HS GPA



Notes:

1. SS_T has $n-1$ degrees of freedom (1 is lost through the estimation of μ by $\bar{y}$).
2. SS_R's degrees of freedom corresponds to the number of independent variables in the model.
3. "Mean squares" are formed from dividing the sum of squares by their corresponding degrees of freedom.

$$MS_E = \frac{SS_E}{n-2}, \qquad MS_R = \frac{SS_R}{1}, \qquad \frac{SS_{TO}}{n-1} \neq MS_R + MS_E$$

4. Analysis of variance (ANOVA) table

Source of variation	df	SS	MS	F
Regression	1	SS_R	MS_R	$F = \dfrac{MS_R}{MS_E}$
Error	$n-2$	SS_E	MS_E	
Total	$n-1$	SS_T		

Note that F follows an F – distribution with 1 degree of freedom for the numerator and $n-2$ degrees of freedom for the denominator.

1.10.1 $F-$Test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$

Note:

For a given significance level α, $F-$test of $\beta_1 = 0$ vs. $\beta_1 \neq 0$ is equivalent algebraically to the two-tailed $t-$test.

(i) The test statistic $F^* = \dfrac{SS_R \div 1}{SS_E \div (n-2)} = \dfrac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{MS_E} = \dfrac{\hat{\beta}_1^2}{se^2(\hat{\beta}_1)} = \left(\dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2 = (t)^2$

(ii) The required percentiles of the t and F distributions for the tests: $[t(1-\alpha/2; n-2)]^2 = F(1-\alpha; 1, n-2)$. Remember that $t-$ test is two-tailed test whereas the $F-$ test is right-tailed test.
Eg: $[t(0.975;23]^2 = (2.069)^2 = 4.28 = F(0.95;1,23)$

The $t-$test is more flexible since it can be used for one–sided alternatives involving $H_0 : \beta_1 \leq 0$ or $H_0 : \beta_1 \geq 0$, while the $F-$test cannot have such tests.

Example 1.16:

Reconsider *Ex. 1.12*, by using $\alpha = 0.05$, is Quiz 1 linearly related to Quiz2?

1.11 Coefficient of Determination

The *coefficient of determination*, denoted by R^2, represents the proportion of SS_T that is explained by the use of the linear regression model, is defined as:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\text{Res}}}{SS_T}$$

It is a measure of the variability in Y without considering the effect of the regressor variable X.

The computational formula for R^2 is

$$R^2 = \frac{S_{XY}^2}{S_{XX}\ S_{YY}} \qquad \text{and} \qquad 0 \leq R^2 \leq 1.$$

Notes:

1. R^2 measures the proportion of variation in Y that explained by the regressor variable X. (i.e. $R^2 100\%$ of the variation in Y can be "explained" by using X to predict Y); or The error in predicting Y can be reduced by $R^2 100\%$ when the regression model is used instead of just $\bar{y}$.

2. R^2 is a measure of "fit" for the regression line

   0 0.25 0.5 0.75 1.0

   Bad Fit Good Fit

3. $r = \sqrt{R^2}$ is the coefficient of correlation. The square of this is the coefficient of determination in simple linear regression.

4. From the relationship $\hat{\beta}_1 = r\sqrt{\dfrac{S_{YY}}{S_{XX}}}$, we obtain

$$\hat{\beta}_1^2 = R^2\left(\frac{\sum(y_i-\bar{y})^2}{\sum(x_i-\bar{x})^2}\right) = \frac{SS_R}{SS_T}\left(\frac{\sum(y_i-\bar{y})^2}{\sum(x_i-\bar{x})^2}\right) \text{ and } SS_R = \hat{\beta}_1^2\sum(x_i-\bar{x})^2$$

5. $$F = \frac{R^2/1}{(1-R^2)/(n-2)} = \frac{(n-2)R^2}{1-R^2}$$

Example 1.17:

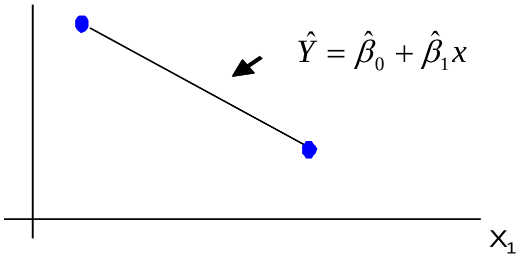
Reconsider *Ex.1.2*. Find R^2 and give an interpretation for this quantity.

Warning

1. Use R^2 as a measure of fit when the sample size is substantially larger than the number of variables in the model; otherwise, R^2 may be artificially high.

   For example:

   Suppose the estimated model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, and a random sample of size 2 is used to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$. Then a scatter plot with the estimated regression line plotted upon it would look something like:



   and $R^2 = 1$. In this case the sample size is not substantially larger than the number of variables in the model causing R^2 to be artificially high.

2. R^2 is only measuring the linear relationship.

3. R^2 is a measure of how the estimated regression line fits in the sample only.

9.2 Hypothesis Testing for Population Correlation ρ

⇨ A significance test can be conducted to test whether the correlation between two variables X and Y is significant or not.

- The Null and Alternative Hypotheses.

	H_0	H_1	Type of test
(i)	$\rho = 0$	$\rho \neq 0$	Two-tailed test
(ii)	$\rho = 0$ or $\rho \geq 0$	$\rho < 0$	Left-tailed test
(iii)	$\rho = 0$ or $\rho \leq 0$	$\rho > 0$	Right-tailed test

- Test Statistic.

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$