Chap 2: Regression Analysis

Ani Shabri

Department of Mathematical Sciences,
Faculty of Science, Universiti Teknologi
Malaysia,
81310 UTM Johor Bahru, Malaysia
ani@utm.my

Jun 8, 2014

Chap 2: Regression Analysis

Outline:

- Introduction to linear regression model
- Least squares estimation in linear regression models
- Estimating the parameter regression by matrix
- Test for significance of regression
- Tests on individual regression coefficients
- Multiple regression
- Test for significance of multiple regression
- Tests on individual multiple regression coefficients
- Non-linear regression
- Regression models for seasonal time series data

Introduction to linear regression model

Liner regression model is useful when the time series has a clear trend. This model can be used to estimate the relationship between a dependent variables and an independent variables. This model is defined as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where β_0 and β_1 are parameters. The estimated model linear regression model is

$$\hat{Y}_i = a + bX_i$$

Note: For linear trend, independent variable X for linear regression is usually as time period t. The linear trend is defined as

$$Y_t = a + bt$$

Least squares estimation in linear regression models

The parameters of sample regression or linear trend can be estimated by the method of ordinary least squares (OLS). OLS estimates are obtained by minimizing the sum of squared errors (SSE):

$$SSE = L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

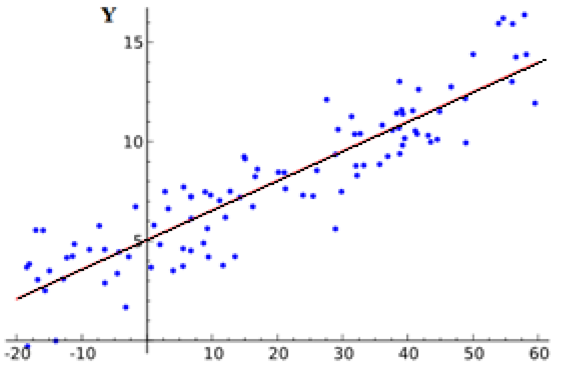


Figure 1: The least-squares line procedure

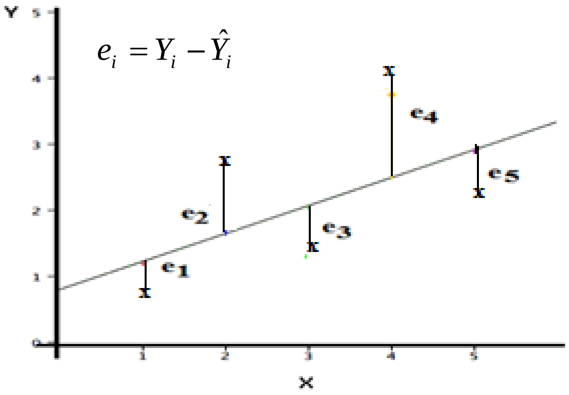


Figure 2: Least Square Estimates

Let

$$SSE = L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

Taking the partial derivatives of SSE with respect to *a* and *b*, and setting them equal to zero, we get

$$\frac{\partial L}{\partial a} = -\sum_{i=1}^{n} 2(Y_i - (a + bX_i)) = 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} 2(Y_i - (a + bX_i))X_i = 0$$

These equations can be written as

$$b = \frac{n\sum_{i=1}^{n} XY - \sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n\sum_{i=1}^{n} X^2 - \left(\sum_{i=1}^{n} X\right)^2} \qquad a = \bar{Y} - b\bar{X}$$

Estimating the parameters regression by matrix

The linear regression model can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{b}$$

To obtain the estimated regression coefficients from linear regression model by matrix methods, we pre-multiply $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ both sides by the inverse of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b}$

We then find, since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ and $\mathbf{I}\mathbf{b} = \mathbf{b}$, we get

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$$

Test for Significance of regression

F test is used to test whether the linear regression model as a whole is useful to explain Y, i.e., at least one X–variable in the regression model is useful to explain Y.

H_0 : all slope coefficients are equal to zero

(i.e. $\beta_1 = 0$)

H_a : not all slope coefficients are equal to zero

Test statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2 / p}{\sum(Y_i - \hat{Y}_i)^2 / (n-p-1)}$$

Decision rule: reject null hypothesis if $F > F_{a,\, n\text{-}p\text{-}1}$ or p-value < a

Tests on individual regression coefficients

One way to evaluate the regression model is to test the hypothesis that the population slope (β) equal to zero (indicating no linear relationship):

$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0$$

The t-test statistic is computed in the following manner

$$t = \frac{b - \beta}{s_b}$$

where $s_b = \sqrt{\dfrac{s_e^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$ and $s_e = \sqrt{\dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^{\ 2}}{n-2}}$

Reject H_0 if $t > \left| t_{\alpha/2, v=n-2} \right|$

Coefficient of determination

Coefficient of determination (R^2) is used to summarize how well a linear regression model fits the data. R^2 is defined as

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$\hat{y}_i$ is the predicted value

R^2 shows the proportion of variation in the forecast variable that is explained by the regression model.

- R^2 lies between 0 and 1.
- If $R^2 \approx 1$ means the predictions are close to actual values.
- If $R^2 \approx 0$ shows the predictions are unrelated to the actual values.

Correlation

Correlation is denoted by *R* is used to measure of the strength of the relationship between independent and dependent variables.

$$R = \sqrt{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}}$$

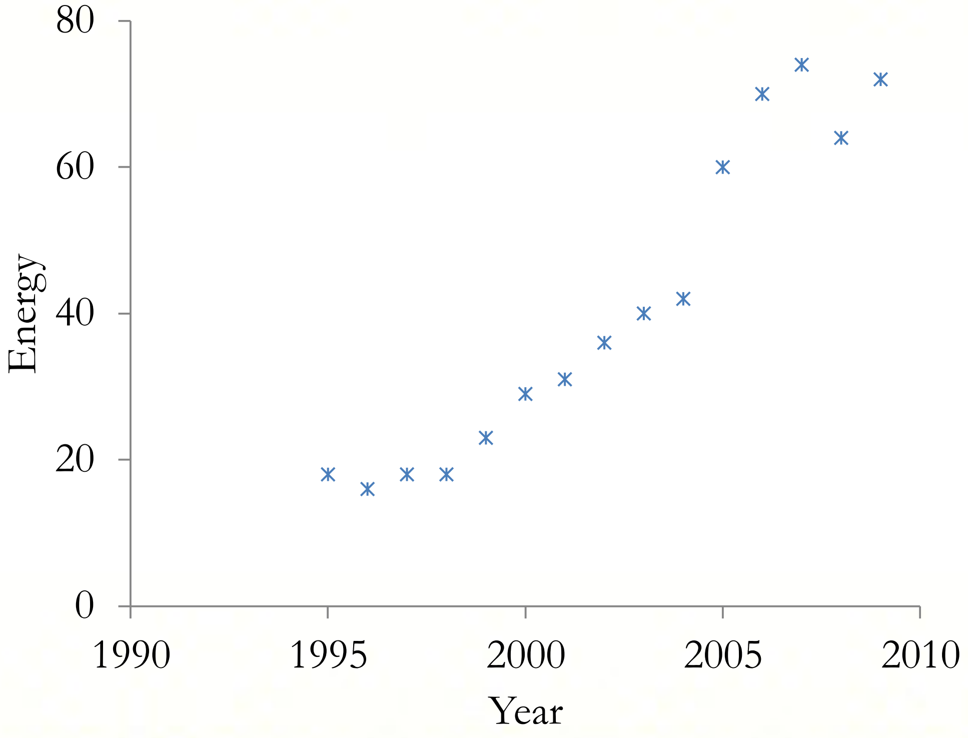
- Values *R* lies between +1 and -1.
- Value of *R* near zero indicates little or no relationship between variables.
- Values of *R* near 1.00 indicate a strong positive linear relationship
- Values of *R* near -1.00 indicate a strong negative linear relationship

Example: Linear regression analysis

The number of energy consumption

(in quardrillion BTUs)

Year	Energy
1995	18
1996	16
1997	18
1998	18
1999	23
2000	29
2001	31
2002	36
2003	40
2004	42
2005	60
2006	70
2007	74
2008	64
2009	72



Example: Estimating the parameters model

The data displays an obvious trend over time and a *linear regression method*, can be used to forecast the data. The number of producing energy.

$$\bar{x}=\frac{120}{5}=24 \quad \bar{y}=\frac{611}{15}=122.2$$

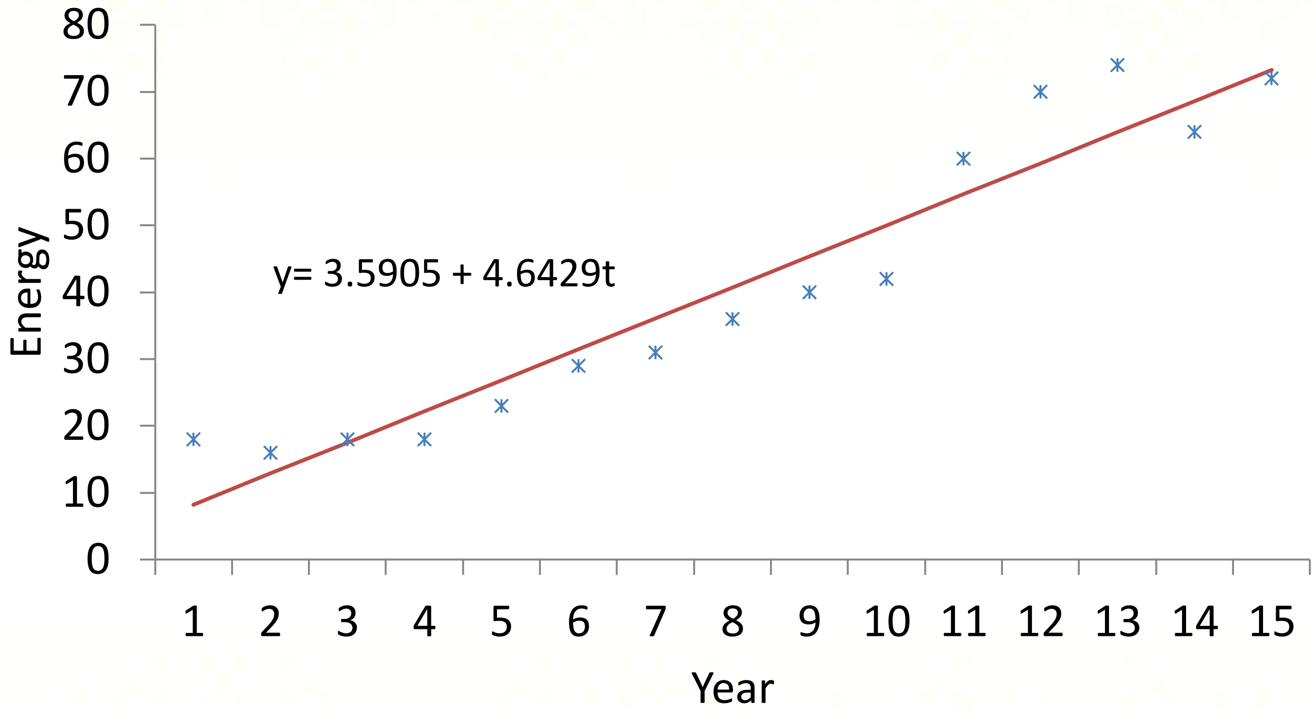
$$b=\frac{\sum xy-n\bar{x}\bar{y}}{\sum x^2-n\bar{x}^2}=\frac{6188-(15)(8)(40.7333)}{1240-15(8)^2}=4.6429$$

$$a=\bar{y}-b\bar{x}=40.7333-(4.6429)(8)=3.5905$$

$$y=3.5905+4.6429t \quad \textit{linear regression line}$$

for period 16, $t=16$, $y=3.5905+4.6429(16)=77.87$

Example : The linear regression plot



Example: Another way to determine trend: use the excel regression function

- Run linear regression to test b_1 in the model $Y_t = b_0 + b_1 t + e_t$
- Excel results:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.59047619	3.538153165	1.014788	0.328722
X Variable 1	4.642857143	0.389144966	11.93092	2.24E-08

This smaller P-value indicates
that there is strong evidence that trend exists

- Conclusion: A trend regression model is appropriate.

Example: Output Excel for linear regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9572
R Square	0.9163
Adjusted R Square	0.9098
Standard Error	6.5116
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6035.71	6035.71	142.34	2.24E-08
Residual	13	551.22	42.40		
Total	14	6586.93			

The computed of F is larger than the critical value, thus we reject the null hypothesis of no linear relationship between Y and independent variables. The F test value shows that the trend model fit the data very well.

Example: Coefficient of determination

- The coefficient of determination is the percentage of the variation in the dependent variable that results from the independent variable.
- Computed by squaring the correlation coefficient, r.

For the example:

$$R = .9572, R^2 = .9163$$

- This value indicates that 91.63% of the amount of variation in wells can be attributed to the number of times, with the remaining 10.1% due to other, unexplained, factors.

Multiple regression

Multiple regression analysis is one of the most widely used for all statistical methods. Is used to determine the relationship between one dependent (y) and more than one independent variables (x's). The estimated multiple linear regression model is given by

$$y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

The parameters of the model can be calculated using the least square estimates

where $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$$

$$\hat{\mathbf{Y}} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12}\ldots & x_{1p} \\ 1 & x_{21} & x_{22}\ldots & x_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{n1} & x_{n2}\ldots & x_{np} \end{pmatrix}$$

Test for significance of multiple regression

F test is used to test whether the multiple regression model as a whole is useful to explain Y, i.e., at least one X–variable in the regression model is useful to explain Y.

H_0 : all slope coefficients are equal to zero

(i.e. $\beta_1 = \beta_2 = \ldots = \beta_p = 0$)

H_a : not all slope coefficients are equal to zero

Test statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{\sum(\hat{Y}_i - \bar{Y})^2 / p}{\sum(Y_i - \hat{Y}_i)^2 / (n-p-1)}$$

Decision rule: reject null hypothesis if $F > F_{a,\ n\text{-}p\text{-}1}$ or p-value $< a$

Tests on individual multiple regression coefficients

Tests on Individual Regression Coefficient to test suitable the parameters of multiple regression model

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

The test statistic $t_0 = \dfrac{\beta_j}{\sqrt{s\,C_{jj}}}$

$$\mathbf{C} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \qquad \text{and} \qquad s^2 = \frac{SSE}{n-p} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-p}$$

Decision rule: reject the null hypothesis at α level of significance if p-value < α or $\left| t_0 \right| > t_{\alpha/2,n-p}$

Non-linear regression

- Linear regression model is suitable when the data shows a linear trend or displays an obvious trend over time.
- If the data shows non-linear trend, the non-linear regression models are suitable model to fit to this data.
- The parameters of non-linear regression models can be estimated by linear regression by using Log-Transform.
- Examples of non-linear regression models are

Non-linear Regression	Linear Regression
$y = ax^b$	$ln\ y = ln\ a + b\ lnx$
$y = ab^x$	$ln\ y = ln\ a + x\ lnb$
$y = ae^{bx}$	$ln\ y = ln\ a + bx$

- For data shows curvilinear trend, the high-order polynomial model can be use

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_p x^p$$

Regression Models For Seasonal Time Series Data

If a variable Y exhibits both trend and seasonality, the trend model with the seasonal model are used. The model is given by

$$Y_t = \alpha + \beta t + \sum_{i=1}^{s} \gamma_i S_{it} + a_t$$

where

$$S_{it} = \begin{cases} 1 & \text{if seasonal occur at time t} \\ 0 & \text{others} \end{cases}$$

If the series shows no trend and pure seasonal, the dummy model is given by:

$$Y_t = \sum_{i=1}^{s} \gamma_i S_{it} + a_t$$

If the series shows curvilinear trend and pure seasonal, the dummy model is given by:

$$Y_t = \alpha + \beta t + \gamma t^2 + \sum_{i=1}^{s} \gamma_i S_{it} + a_t$$

Exercise: Regression for Seasonal time series data

Data below shows the sales of car for 16 periods.

Period	Sales
1	7
2	11
3	12
4	17
5	8
6	11
7	13
8	19
9	8
10	12
11	13
12	20
13	8
14	13
15	14
16	22

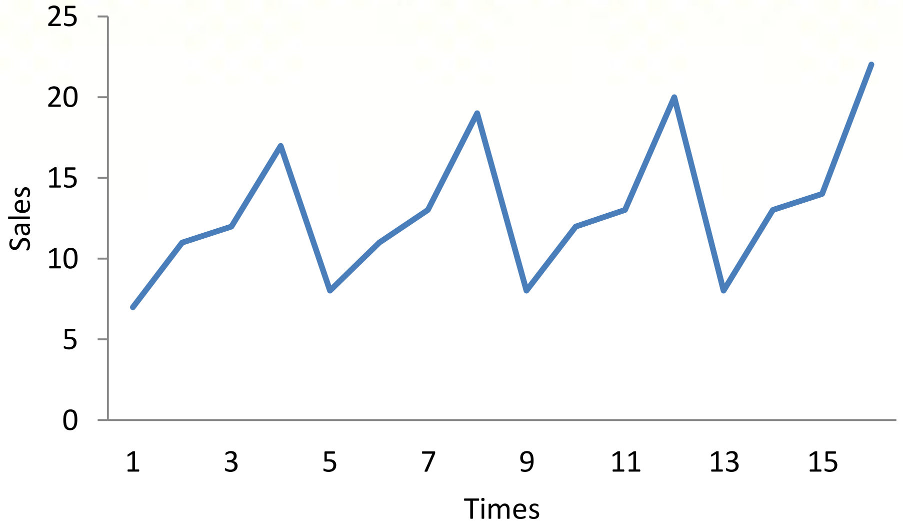


Fig: Sales of car for 16 periods

Using the seasonal dummy model to fit the data.

Example: Solution

The data containing linear trend and seasonality. The model is given by

$$Y_t = \alpha + \beta t + \sum_{i=1}^{s} \gamma_i S_{it} + a_t$$

$$= \alpha + \beta t + \gamma_1 S1_t + \gamma_2 S2_t + \gamma_3 S3_t$$

$\alpha + \beta t$ - is the trend in the data

$\sum_{i=1}^{s} \gamma_i S_{it}$ - is the additive seasonal component

Example: Solution

The following array presents an example of how quarterly data can be arranged:

Sales	Period	Dummy Variables		
y(t)	t	S1(t)	S2(t)	S3(t)
7	1	1	0	0
11	2	0	1	0
12	3	0	0	1
17	4	0	0	0
8	5	1	0	0
11	6	0	1	0
13	7	0	0	1
19	8	0	0	0
8	9	1	0	0
12	10	0	1	0
13	11	0	0	1
20	12	0	0	0
8	13	1	0	0
13	14	0	1	0
14	15	0	0	1
22	16	0	0	0

Example: Solution

Using the EXCEL, a multiple regression analysis was run with dummy variables representing the first quarters. The following result was obtained.

Table : Output Form Excel

		Standard		
	Coefficients	*Error*	*t Stat*	*P-value*
Intercept	17.5	0.5399	32.4142	0.0000
X Variable 1	0.2	0.0402	4.9701	0.0004
X Variable 2	-11.15	0.5231	-21.3140	0.0000
X Variable 3	-7.35	0.5153	-14.2626	0.0000
X Variable 4	-6.3	0.5106	-12.3385	0.0000

The seasonal dummy model is

$$\hat{Y}_t = 17.5 + 0.2t - 11.15S1_t - 7.35S2_t - 6.3S3_t$$

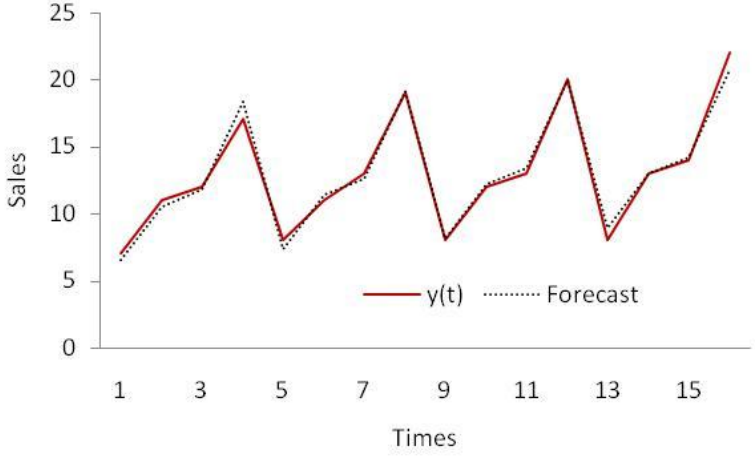
Example: Output form Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9906
R Square	0.9813
Adjusted R Square	0.9744
Standard Error	0.7198
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	298.3	74.575	143.917	2.0E-09
Residual	11	5.7	0.518		
Total	15	304			



Actual and forecast sales

As seen the table, R Square is 0.9813 indicate 98.13% of the variance in the actual data. The computed of F is larger than the critical value, thus we reject the null hypothesis of no linear relationship between Y and independent variables. The Figure and F test Show that the seasonal dummy model fit the data very well.