Chapter 06.05

Adequacy of Models for Regression

Quality of Fitted Model

In the application of regression models, one objective is to obtain an equation $y=f(x)$ that best describes the n response data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Consequently, we are faced with answering two basic questions.

1. Does the model $y=f(x)$ describes the data adequately, that is, is there an adequate fit?
2. How well does the model predict the response variable (predictability)?

To answer the above questions, let us start from the examination of some measures of discrepancies between the whole data and some key central tendency. Look at the two equations given below.

$$S_r = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad (1)$$

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (2)$$

where S_r is the sum of the square of the residuals (residual is the difference between the observed value, y_i and the predicted value, $\hat{y}_i$), and S_t is the sum of the square of the difference between the observed value and the average value.

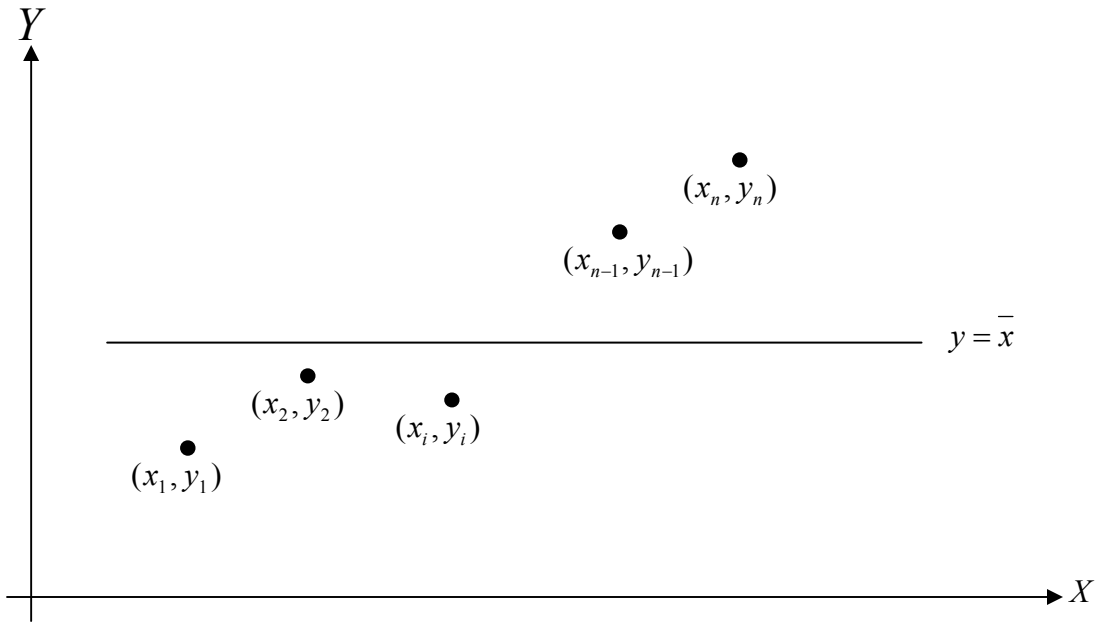


Figure 1. Spread of data about the mean value of **y**

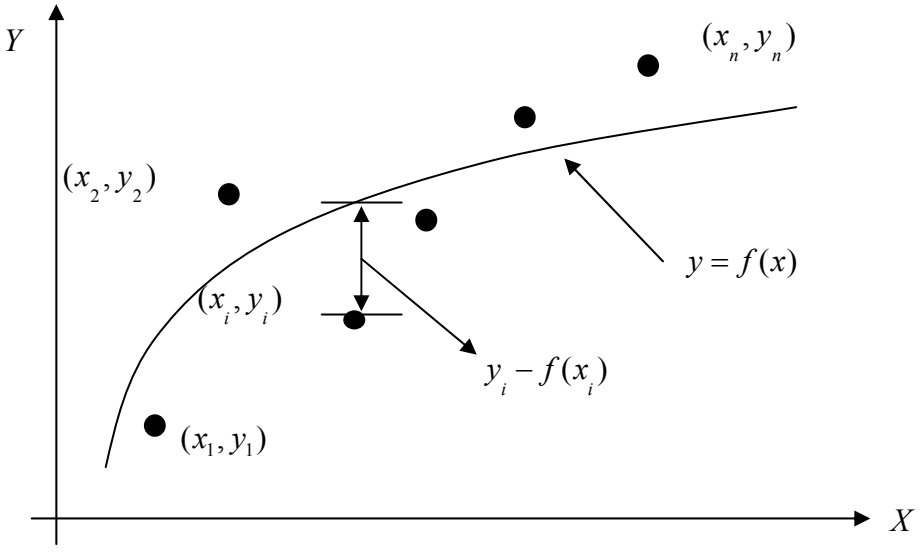


Figure 2. Spread of data about the regression line.

To normalize with respect to the number of data points, we calculate standard deviation, σ as

$$\sigma = \sqrt{\frac{S_t}{n-1}} \tag{3}$$

However, why is S_t divided by *(n-1)* and not *n* as we have *n* data points? This is because with the use of the mean in calculating S_t, we lose the independence of one of the degrees of freedom. That is, if you know the mean of *n* data points, then the value of one of the *n* data points can be calculated by knowing the other *n-1* data points. The standard deviation is an estimate of the spread of the data about its average.

Similarly, to normalize the sum of the square of the residuals with respect to the number of data points, the standard error of estimate is calculated as

$$s_{y/x} = \sqrt{\frac{S_r}{n-m}} \tag{4}$$

where m is the number of constants of the model (a straight line model $y = a_0 + a_1 x$ has two constants, a_0 and a_1; an exponential model $y = a_0 e^{a_1 x}$ has two constants, a_0 and a_1; a polynomial model $y = a_0 + a_1 x + a_2 x^2$ has three constants, a_0, a_1 and a_2). The subscript *y/x* stands for that the error in the predicted value of *y* for a chosen value of *x*.

Why is S_r divided by *(n-m)* and not *n* as we have *n* data points? This is because with the use of the mean in calculating S_r, we lose the independence of *m* degrees of freedom.

What inferences can we make about the two equations? Equation (2) measures the discrepancy between the data and the mean. Recall that the mean of the data is a measure of a single point that measures the central tendency of the whole data. Equation (2) contrasts with Equation (1) as Equation (1) measures the discrepancy between the vertical distance of the point from the regression line (another measure of central tendency). This line obtained by the least squares method gives the best estimate of a line with least sum of deviation. S_R as calculated quantifies the spread around the regression line.

The objective of least squares method is to obtain a compact equation that best describes all the data points. The mean can also be used to describe all the data points. The magnitude of the sum of squares of deviation from the mean or from the least squares line is therefore **a good indicator of how well the mean or least squares**

characterizes the whole data. We can liken the sum of squares deviation around the mean, S_T as the error or variability in y without considering the regressor variable, x, while S_R, the sum of squares deviation around the least square regression line is error or variability in y remaining after the dependent variable x has been considered.

The difference between these two parameters measures the error due to describing or characterizing the data in one form instead of the other. A relative comparison of this difference $(S_t - S_r)$, with the sum of squares deviation associated with the mean (S_t) describes a quantity called **coefficient of determination**, r^2

$$r^2 = \frac{S_t - S_r}{S_t} \tag{5}$$

$$r = \sqrt{r^2} = \sqrt{\frac{S_t - S_r}{S_t}} \tag{6}$$

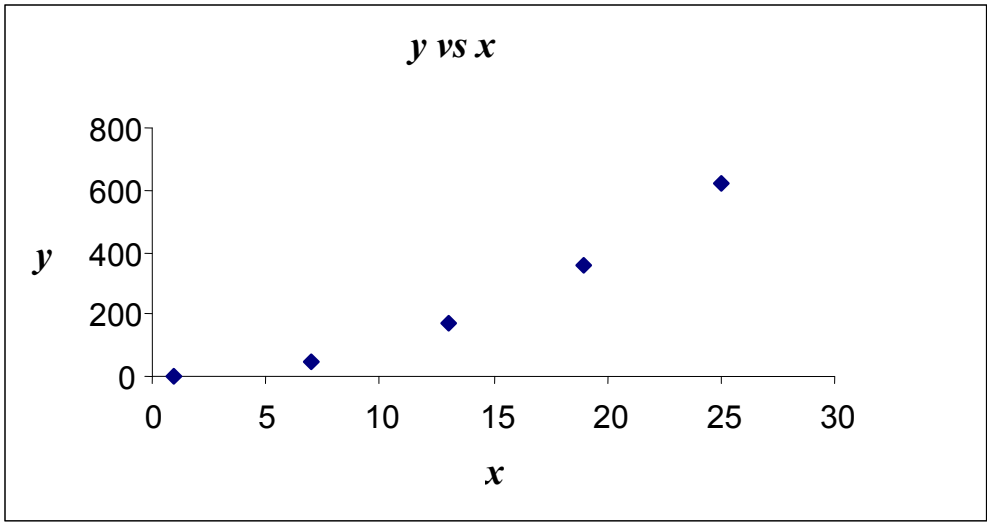
Where r, called Pearson's product moment correlation coefficient (PPMCC)

Another way of defining r^2 (see Equation 3) is to describe it as the proportion of variation in the response data that is explained by the regression model. We note that $0 \le r^2 \le 1$. When all points in a data set lie on the regression model, the largest value of $r^2=1$ is obtained, while a minimum value of $r^2=0$ is obtained when there is only one data point or if the regression model is a constant line given by the average of the y data values.

Example 1

The following y vs. x data is given

x	y
1	1
7	49
13	169
19	361
25	625



. **Figure 3**. Data points of the y vs x data

Although $y = x^2$ is an exact fit to the data, a scientist thinks that $y = a_0 + a_1 x$ can explain the data. Find

a) constants of the model, a_0, and a_1,

b) standard deviation of the data points,

c) standard error of estimate of the straight line model,

d) the coefficient of determination for the straight-line model?

Solution

a) First find the constants of the assumed model

$$y = a_0 + a_1 x$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$n = 5$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{5} x_i y_i = 1 \times 1 + 7 \times 49 + 13 \times 169 + 19 \times 361 + 25 \times 625 = 25025$$

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{5} x_i^2 = 1^2 + 7^2 + 13^2 + 19^2 + 25^2 = 1205$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{5} y_i = 1 + 49 + 169 + 361 + 625 = 1205$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{5} x_i = 1 + 7 + 13 + 19 + 25 = 65$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$a_1 = \frac{5(25025) - (65)(1205)}{5(1205) - (65)^2}$$

$$= 26$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$= \frac{1205}{5} - 26\frac{65}{5}$$

$$= (241) - 26(13)$$

$$= -97$$

This gives

$$y = a_0 + a_1 x$$

$$y = -97 + 26x$$

is the regression formula.

b) The sum of the squares of the difference between observed value and average value, S_t is given by

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= \sum_{i=1}^{5} (y_i - \bar{y})^2$$

$$= (1-241)^2 + (49-241)^2 + (169-241)^2 + (361-241)^2 + (625-241)^2$$

$$= 261504$$

The standard deviation of the observed values is

$$\sigma = \sqrt{\frac{S_t}{n-1}}$$

$$= \sqrt{\frac{261504}{5-1}}$$

$$= 255.68$$

c) The sum of the squares of the residuals, that is the sum of the square of differences between the observed values and the predicted values is

$$S_r = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i \right)^2$$

x_i	y_i	$a_0 + a_1 x_i$	$y_i - a_0 - a_1 x_i$
1	1	-71	72
7	49	85	-36
13	169	241	-72
19	361	397	-36
25	625	553	72

$$\begin{aligned} S_r &= \sum_{i=1}^{5} \left(y_i - a_0 - a_1 x_i \right)^2 \\ &= (72)^2 + (-36)^2 + (-72)^2 + (-36)^2 + (72)^2 \\ &= 18144 \end{aligned}$$

The standard error of estimate is

$$\begin{aligned} s_{y/x} &= \sqrt{\frac{S_r}{n-2}} \\ &= \sqrt{\frac{18144}{5-2}} \\ &= 77.77 \end{aligned}$$

Since there is an improvement from a standard deviation of 255.68 to a standard error of estimate of 77.77, there is merit to explaining the data by the straight line

$$y = -97 + 26x.$$

d) Then using equation (5), we get

$$\begin{aligned} r^2 &= \frac{S_t - S_r}{S_t} \\ &= \frac{261504 - 18144}{261504} \\ &= 0.9306 \end{aligned}$$

This implies that *93.06%* of the original uncertainty in the data is explained by the straight line $y = -97 + 26x$.

Caution in the use of r^2

1. r^2 can be made larger (assumes no collinear points) by adding more terms to the model. For instance, $n-1$ terms in a regression equation for which n data points are used will give an r^2 value of 1 if there are no collinear points.

2. The magnitude of r^2 also depends on the range of variability of the regressor (x) variable. Increase in the spread of x increases r^2 while a decrease in the spread of x decreases r^2.
3. Large regression slope will also yield artificially high r^2.
4. r^2 does not measure the appropriateness of the linear model. r^2 may be large for nonlinearly related x and y values.
5. Large r^2 value does not necessarily imply the regression will predict accurately.
6. r^2 does not measure the magnitude of the regression slope.

These statements above imply that one should not choose a regression model solely based on consideration of r^2.

Other checks for adequacy

a) Plot the graph and see if the regression model visually explains the data.
b) Plot the residuals as a function of x to check for increasing variance, outliers or nonlinearity.
c) Check if 95% of the values of scaled residuals are within [-2, 2]. The scaled residuals, *SR* are given by

$$SR = \frac{y_i - f(x_i)}{\sqrt{\frac{S_r}{n-m}}} = \frac{y_i - f(x_i)}{s_{y/x}} \tag{7}$$

Example 2: Make the other checks for the adequacy of the model in Example 1.

Solution

a) Plot the graph as given below (Figure 4). See if the straight-line regression model visually explains the data. Although you may see a nonlinear trend in how the data points are around the straight line, this trend gets visually less exaggerated by extending the axis (Figure 5).

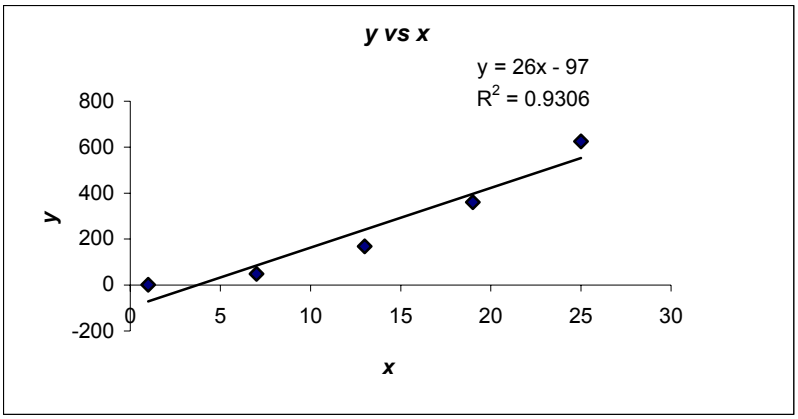


Figure 4. Linear regression model for data

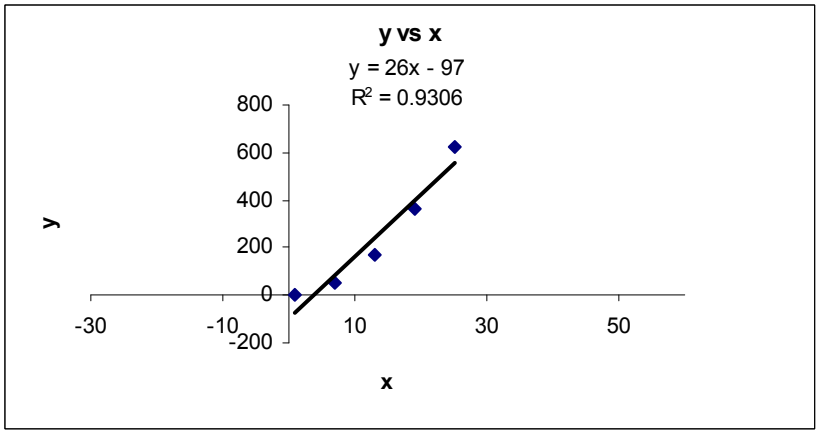


Figure 5. Linear regression model for data in Figure 4 with extended *x*-axis

b) Plot the residuals as a function of *x* to check for increasing variance, outliers or nonlinearity. As seen from the residual plot, the residuals $(y_i - a_0 - a_1 x_i)$ do show nonlinearity. This may be the first indication so far, that the model is not adequate.

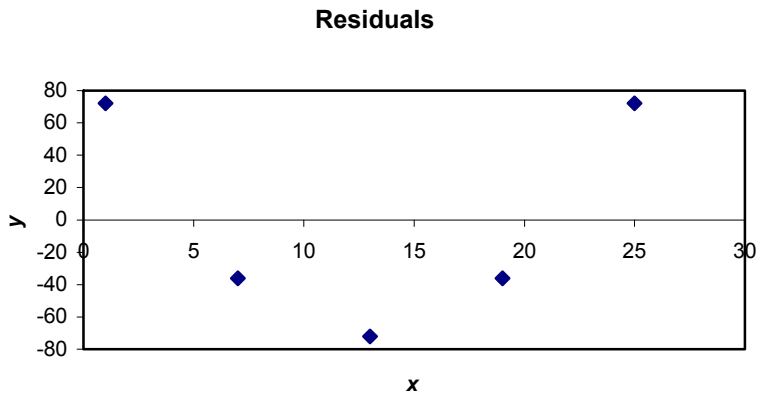


Figure 6. Residuals for data points

c) Check if 95% of the values of scaled residuals are within [-2, 2]. The scaled residuals, *SR* are given by

$$SR = \frac{y_i - f(x_i)}{\sqrt{\frac{S_r}{n-m}}} = \frac{y_i - f(x_i)}{s_{y/x}}$$

where $f(x)$ the regression function, n is the number of data points and *m* is the number of degrees of freedom lost (constants of the model).

x_i	y_i	$f(x_i) = a_0 + a_1 x_i$	SR
1	1	-71	$\frac{1-(-71)}{77.77} = 0.9258$
7	49	85	-0.4629
13	169	241	-0.9258
19	361	397	-0.4629
25	625	553	0.9258

All the scaled residuals are between [-2, 2], that is, more than 95% of the scaled residuals are in between [-2,2].

Adequacy of Coefficient of Regression

A key consideration in any model is the adequacy of the model. One must always ask the question, does the fitted model adequately approximate the response variable? A

negative answer to this question requires reevaluation of the assumed model. Prior to the interpretation of the prediction equation, one needs to consider the adequacy of the fit. This is done by the evaluation of the coefficient of determination. Having answered the adequacy question based on coefficient of determination, one might think that the regression coefficient estimates must be close to the true parameter values. There is a fallacy in this belief because wrongly specified model can provide acceptable residuals, r^2 and σ^2 even with poorly estimated model parameters. This is the reason why we must examine the adequacy of the model parameters estimators.

Hypothesis Testing in Linear Regression

The test for significance if regression is to check if a linear relationship exists between *y* and *x*. The hypothesis is that

$$H_0 : a_1 = 0$$

$$H_1 : a_1 \neq 0$$

If we are unable to reject the hypothesis $H_0 : a_1 = 0$, it would mean that there is no linear relationship between *x* and *y*. This implies whether the relationship between *x* and *y* is a constant line or that a linear relationship between *y* and *x* does not exist. Assuming normal distribution and using test statistics, a standard normal random valuable is given by

$$Z = \frac{a_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

We would reject H_0 if $|Z| > Z_{\alpha/2}$, where α denotes the size (probability of Type I error) of the test. If σ^2 is not linear, a t-static can be replaced by $s_{y/x}$

$$t = \frac{a_1}{\sqrt{\frac{s_{y/x}}{S_{xx}}}}$$

$$= \frac{a_1}{Var(a_1)} \qquad (8)$$

Which is distributed a student's t with $(n-2)$ degrees of freedom. Hence the null hypothesis is rejected if $|t| > t\left(\frac{\alpha}{2}, n-2\right)$ where $t\left(\frac{\alpha}{2}, n-2\right)$ denotes the value of the t-distribution such that $prob\left(t > t\left(\frac{\alpha}{2}, n-2\right)\right) = \frac{\alpha}{2}$

Example 3: From the data of Example 1, determine if the assumption of linear relationship is reasonable?

Solution:

We to test the regression variable influence the response, that is hypothesis

$$t = \frac{a_1 - a_p}{Var(a_1)}$$
$$= \frac{26-0}{0.2160}$$
$$= 120.37$$

Since $t_{0.025,3} = 7.453$ and since 120.37>7.453, we reject the null hypothesis and accept the fact that x influences the response of variable y.

Model Estimators

The theoretical properties of the model parameter estimators are tied to the model assumptions. The model assumptions include

- Model is correctly specified.
- Predictor variables are non-random and are measured without error.
- Model error terms have constant variances, zero means and are uncorrelated.
- Model error terms are normally, independently distributed with mean zero and constant variance.

Since the properties of the estimators are tied to the above assumptions, the adequacy of the estimators depend upon the correctness of the assumptions made in deriving the model. The estimators β_0 and β_1 are unbiased and their variances are given as

$$Var(a_0) = s^2_{y/x}\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \quad (9)$$

$$Var(a_1) = \frac{s^2_{y/x}}{S_{xx}} \quad (10)$$

where $s^2_{y/x}$ is the error variance. For tests of hypothesis, estimated standard errors for the slope and intercept are required and use of their variances becomes important. Estimation of the error variance is useful here.

In tests of hypothesis, the estimate of error variance is useful in the calculation of estimated errors of regression model coefficients. It is useful in assessing quality of fit and prediction capability of the regression model.

Example 4: Find the estimated standard error of the slope and the intercept for Example 1.

Solution:

The straight-line regression model calculated was

$y=-97+26x$

The estimated standard error of slope is

$$Var(a_1) = \sqrt{\frac{s_{y/x}}{S_{xx}}}$$

$s_{y/x} = 77.77$ from Example 2

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{1+7+13+19+25}{5}$$

$$= 13$$

$$S_{xx} = \sum_{i=1}^{5} (x_i - \bar{x})^2$$

$$= (1-13)^2 + (7-13)^2 + (13-13)^2 + (19-13)^2 + (25-13)^2$$

$$= 360$$

$$Var(a_1) = \sqrt{\frac{s_{y/x}}{S_{xx}}}$$

$$= \sqrt{\frac{77.77}{360}}$$

$$= 0.2160$$

$$Var(a_0) = \sqrt{s_{y/x}\left[\frac{1}{n} + \frac{\overline{x^2}}{S_{xx}}\right]}$$

$$= \sqrt{77.77\left[\frac{1}{5} + \frac{13^2}{360}\right]}$$

$$= 7.215$$

The estimated standard error of the slope is 0.2160 and the estimated standard error of the intercept is 7.215.

Confidence Intervals

To make inferences about the model coefficients, a more informative way could be the use of intervals with in which the parameters will lie. The probability statement associated with the student-t estimation is that for $(1-\alpha)$ confidence interval $(100(1-\alpha)\%)$ of slope a_1 in linear regression is

$$\alpha_1 - t_{\alpha/2,n-2}Var(a_1) \le a_1 \le \alpha_1 + t_{\alpha/2,n-2}Var(a_1) \qquad (11)$$

Similarly for the intercept a_0 in linear regression, the $(1-\alpha)$ confidence interval $(100(1-\alpha)\%)$

$$\alpha_0 - t_{\alpha/2,n-2}Var(a_0) \le a_0 \le \alpha_0 + t_{\alpha/2,n-2}Var(a_0) \qquad (12)$$

Example 5: Find the 95% confidence intervals for the slope and intercept of the regression model found in Example 1.

Solution: The confidence intervals on slope a_1 is

$$\alpha_1 - t_{\alpha/2,n-2}Var(a_1) \le a_1 \le \alpha_1 + t_{\alpha/2,n-2}Var(a_1)$$

And for a_0 is

$$\alpha_0 - t_{\alpha/2,n-2}Var(a_1) \le a_0 \le \alpha_0 + t_{\alpha/2,n-2}Var(a_0)$$

$\alpha = 0.05$

$n = 5$

$Var(a_1) = 0.2160$ (From Example 4)

$Var(a_0) = 7.215$ (From Example 4)

$\alpha_1 = 26$ (From Example 1)

$\alpha_0 = -97$ (From Example 1)

Hence for the slope a_1, the confidence interval is

$$26 - t_{0.025,5-2}(0.2160) \le a_1 \le 26 + t_{0.025,5-2}(0.2160)$$

$$26 - t_{0.025,3}(0.2160) \le a_1 \le 26 + t_{0.025,3}(0.2160)$$
$$26 - (3.182)(0.2160) \le a_1 \le 26 + (3.182)(0.2160)$$
$$25.31 \le a_1 \le 26.82$$

and the confidence interval for the intercept a_0 is

$$-97 - t_{0.025,5-2}(7.215) \le a_0 \le -97 + t_{0.025,5-2}(7.215)$$
$$-97 - t_{0.025,3}(7.215) \le a_0 \le -97 + t_{0.025,3}(7.215)$$
$$-97 - (3.812)(7.215) \le a_0 \le -97 + (3.812)(7.215)$$
$$-120.0 \le a_0 \le -69.50$$

Data Hazards in Regression

The quality or goodness of the relationship between a response variable and one or more predictor variables in regression analysis depends largely on the quality of the data used. Thus, whether accurate conclusions are made or quality fit is obtained is determined by the representativeness of the data used. It is theoretically possible to obtain a fit irrespective of the nature of the data, hence the saying that "garbage in and garbage out". Data that are not representative or inconsistent or even not properly compiled can result in poor fits and erroneous conclusions.

For illustrative purposes, a study in which data obtained from racially homogeneous schools cannot be useful in making inferences about racial interactions in class and other sundry issues. Data inconsistency arises when there is no consistence in the data sampling. Multi collinearity arises when there is a near- linear relationship among the regressors. This is a serious problem that can impact the usefulness of the regression model since it affects ones ability to estimate regression coefficients. Four primary sources of multicollinearity include:

- Data collection method: When the analyst samples only a subspace of a region, the data collection method can lead to multicollinearity problems.
- Model or population constraints: Constraints on the model or in the population being sampled can cause multicollinearity. Example here can be data of two regressors that lie approximately along a straight line.
- Choice of model: The model specification can result in multicollinearity. Adding polynomial terms to a regression model may lead to multicollinear- since it gives rise to ill-conditioning of the matrix product X'X. Also, if the

range of the regressor variable is small, adding a squared regressor term can result in significant multicollinearity.

- Over-defined model: A model that has more regressor variables than observations is over defined and this may lead to multicollinearity. This is common in medical research.

Because of limited space, we will not discuss in details several techniques available for detecting multicollinearity. But the simplest method of measuring collinearity is the inspection of the off-diagonal elements r_{ij} in X'X matrix product for which $|r_{ij}|$ will be near unity if regressors x_i and x_j are linearly dependent. Also the determinant of X'X can be used as an index of multicollinearity since $0 \le |X'X| \le 1$ for matrix X'X which is in correlation form. For the regressors, orthogonality arises when $|X'X| = 1$ and linear dependence when $|X'X| = 0$

Outliers are data points that are not typical of the rest of the data and thus have considerably large residuals form the mean. The existence of outliers should be carefully investigated as to find out the reason why they occur. Reasons for their existence may be useful in rejecting or accepting them. Faculty measurement, lack of precision, incorrect recording of data, faulty instrument or analysis can all lead to outliers. Outliers may point out inadequacies in the model and thus a follow-up to ascertaining values of the regressor when the response was observed may be a useful exercise to improving the model. It is not recommended to just drop and outlier without first understanding the reason for its existence-is it a bad point or what? Various statistical tests exist for detecting and rejecting outliers. The easiest to apply involves the maximum normed residual test: $\frac{|E_i|}{\sqrt{\sum_{i=1}^{n} E_i^2}}$ which is very large if the response is an outlier. Effect of an outlier may be checked by dropping it in the regression model and re-fitting the regression equation. If the summary statistics are overtly very sensitive to an outlier, that may not be acceptable model.

In considering data hazards, the role of controlled and confounding variables in the regression equations must not be overlooked. Controlled variables are independent variables that the experimenter can manipulate in a systematic way.

The controlled variable contrasts with the confounding variable, which although and independent variable but for some reason is influence on the outcome of experimental results instead of the controlled variables only. The results are said to be confounded in this case. An example may suffice to illustrate the concept.

Consider a case of drug experiment in which two groups are compared. In one group a placebo is given and for the experimental group the active drug is prescribed. However, in the analysis of the data, it is discovered that the controlled group has a higher average age than the experimental group. The disease incidence for which the drug is prescribed is age related. It is possible that the observed difference in the treatment results between the two groups may be due to the age difference instead of the drug. The age difference is said to have confounded the findings.

The effects of confounding can include the outcome result appearing smaller (under-estimated) or appearing bigger than it is (over-estimated). The direction of the observed effect may change as a result of confounding, resulting in a harmful factor appearing to be protective or vice versa. An effective method of controlling potential confounding factors is through good experimental design and rigorous checking for confounding factors at all stages of the study.

ADEQUACY OF REGRESSION MODELS	
Topic	Adequacy of Regression Models
Summary	Textbook notes of Adequacy of Regression Models
Major	General Engineering
Authors	Egwu Kalu, Autar Kaw
Date	November 20, 2009
Web Site	http://numericalmethods.eng.usf.edu