Chapter 06.05
Adequacy of Models for Regression

After reading this chapter, you should be able to

1. *determine if a linear regression model is adequate*
2. *determine how well the linear regression model predicts the response variable.*

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 $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Consequently, we are faced with answering two basic questions.

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

To answer these questions, let us limit our discussion to straight line models as nonlinear models require a different approach. Some authors [1] claim that nonlinear model parameters are not unbiased.

To exemplify our discussion, we will take example data to go through the process of model evaluation. Given below is the data for the coefficient of thermal expansion vs. temperature for steel. We assume a linear relationship between the data as

$$\alpha(T) = a_0 + a_1 T$$

Table 1 Values of coefficient of thermal expansion vs. temperature.

$T(^\circ\text{F})$	α (µin/in/$^\circ$F)
-340	2.45
-260	3.58
-180	4.52
-100	5.28
-20	5.86
60	6.36

Following the procedure for conducting linear regression as given in Chapter 06.03, we get

$$\alpha(T) = 6.0325 + 0.0096964T$$

Let us now look at how we can evaluate the adequacy of a linear regression model.

1. Plot the data and the regression model.

Figure 1 shows the data and the regression model. From a visual check, it looks like the model explains the data adequately.

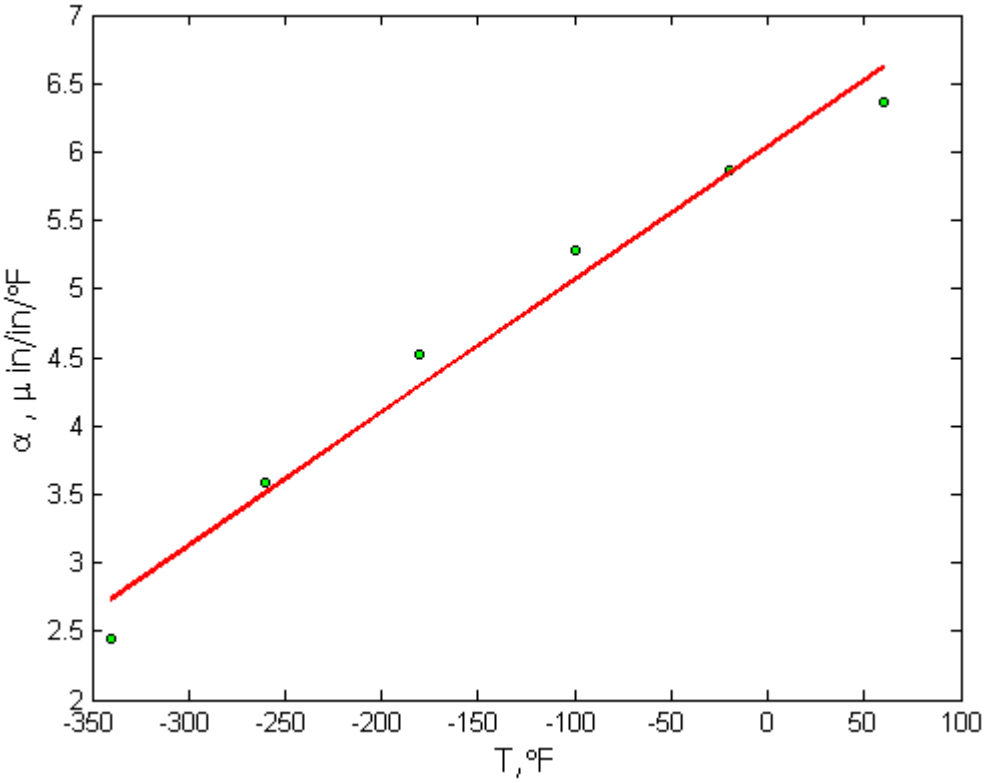

Figure 1 Plot of coefficient of thermal expansion vs. temperature data points and regression line.

2. **Calculate the standard error of estimate**.
The standard error of estimate is defined as

$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$

where

$$S_r = \sum_{i=1}^{n} (\alpha_i - a_0 - a_1 T_i)^2$$

Table 2 Residuals for data.

T_i	α_i	$a_0 + a_1 T_i$	$\alpha_i - a_0 - a_1 T_i$
-340	2.45	2.7357	-0.28571
-260	3.58	3.5114	0.068571
-180	4.52	4.2871	0.23286
-100	5.28	5.0629	0.21714
-20	5.86	5.8386	0.021429
60	6.36	6.6143	-0.25429

Table 2 shows the residuals of the data to calculate the sum of the square of residuals as

$$S_r = (-0.28571)^2 + (0.068571)^2 + (0.23286)^2 + (0.21714)^2 + (0.021429)^2 + (-0.25429)^2$$
$$= 0.25283$$

The standard error of estimate

$$s_{\alpha/T} = \sqrt{\frac{S_r}{n-2}}$$
$$= \sqrt{\frac{0.25283}{6-2}}$$
$$= 0.25141$$

The units of $s_{\alpha/T}$ are same as the units of α. How is the value of the standard error of estimate interpreted? We may say that on average the difference between the observed and predicted values is $0.25141\,\mu\text{in/in/}^\circ\text{F}$. Also, we can look at the value as follows. About 95% of the observed α values are between $\pm 2s_{\alpha/T}$ of the predicted value (see Figure 2). This would lead us to believe that the value of α in the example is expected to be accurate within $\pm 2s_{\alpha/T} = \pm 2 \times 0.25141 = \pm 0.50282\,\mu\text{in/in/}^\circ\text{F}$.

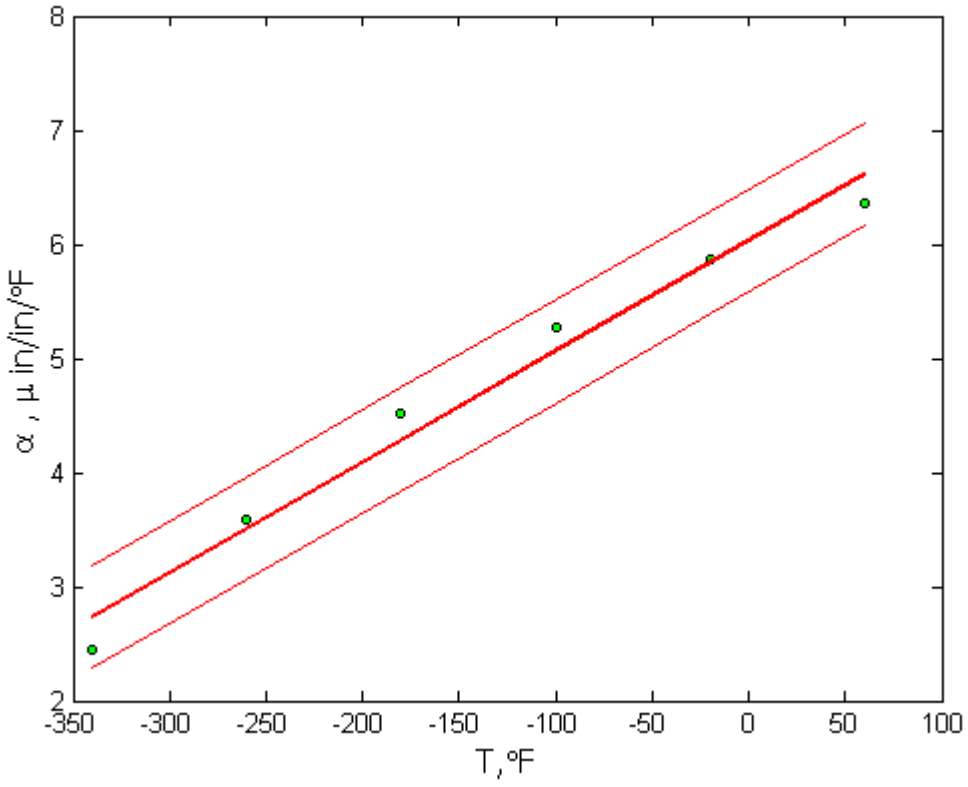

Figure 2 Plotting the linear regression line and showing the regression standard error.

One can also look at this criterion as finding if 95% of the scaled residuals for the model are in the domain [-2,2], that is

$$\text{Scaled residual} = \frac{\alpha_i - a_0 - a_1 T_i}{s_{\alpha/T}}$$

For the example,

$$s_{\alpha/T} = 0.25141$$

Table 4 Residuals and scaled residuals for data.

T_i	α_i	$\alpha_i - a_0 - a_1 T_i$	Scaled Residuals
-340	2.45	-0.28571	-1.1364
-260	3.58	0.068571	0.27275
-180	4.52	0.23286	0.92622
-100	5.28	0.21714	0.86369
-20	5.86	0.021429	0.085235
60	6.36	-0.25429	-1.0115

and the scaled residuals are calculated in Table 4. All the scaled residuals are in the [-2,2] domain.

3. **Calculate the coefficient of determination**.

Denoted by r^2, the coefficient of determination is another criterion to use for checking the adequacy of the model.

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.

$$S_r = \sum_{i=1}^{n} (\alpha_i - \hat{\alpha}_i)^2 = \sum_{i=1}^{n} (\alpha_i - a_0 - a_1 T_i)^2 \quad (1)$$

$$S_t = \sum_{i=1}^{n} (\alpha_i - \overline{\alpha})^2 \quad (2)$$

where

$$\overline{\alpha} = \frac{\sum_{i=1}^{n} \alpha_i}{n}$$

For the example data

$$\overline{\alpha} = \frac{\sum_{i=1}^{6} \alpha_i}{6} = \frac{2.45 + 3.58 + 4.52 + 5.28 + 5.86 + 6.36}{6} = 4.6750 \, \mu\text{in/in/}^\circ\text{F}$$

$$S_t = \sum_{i=1}^{n} (\alpha_i - \overline{\alpha})^2$$
$$= (-2.2250)^2 + (-1.0950)^2 + (-0.15500)^2 + (0.60500)^2 + (1.1850)^2 + (1.6850)^2$$
$$= 10.783$$

Table 5 Difference between observed and average value.

T_i	α_i	$\alpha_i - \overline{\alpha}$
-340	2.45	-2.2250
-260	3.58	-1.0950
-180	4.52	-0.15500
-100	5.28	0.60500
-20	5.86	1.1850
60	6.36	1.6850

where S_r is the sum of the square of the residuals (residual is the difference between the observed value and the predicted value), and S_t is the sum of the square of the difference between the observed value and the average value.

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 regression 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

$$r^2 = \frac{S_t - S_r}{S_t} \qquad (5)$$
$$= \frac{10.783 - 0.25283}{10.783}$$
$$= 0.97655$$

Based on the value obtained above, we can claim that 97.7% of the original uncertainty in the value of α can be explained by the straight-line regression model of $\alpha(T) = 6.0325 + 0.0096964T$.

Going back to the definition of the coefficient of determination, one can see that S_t is the variation without any relationship of y vs. x, while S_r is the variation with the straight-line relationship.

The limits of the values of r^2 are between 0 and 1. What do these limiting values of r^2 mean? If $r^2 = 0$, then $S_t = S_r$, which means that regressing the data to a straight line does nothing to explain the data any further. If $r^2 = 1$, then $S_r = 0$, which means that the straight line is passing through all the data points and is a perfect fit.

Caution in the use of r^2

a) The coefficient of determination, 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.
b) 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.
c) Large regression slope will also yield artificially high r^2.
d) The coefficient of determination, r^2 does not measure the appropriateness of the linear model. r^2 may be large for nonlinearly related x and y values.
e) Large coefficient of determination r^2 value does not necessarily imply the regression will predict accurately.
f) The coefficient of determination , r^2 does not measure the magnitude of the regression slope.
g) These statements above imply that one should not choose a regression model solely based on consideration of r^2.

4. Find if the model meets the assumptions of random errors.

These assumptions include that the residuals are negative as well as positive to give a mean of zero, the variation of the residuals as a function of the independent variable is random, the residuals follow a normal distribution, and that there is no auto correlation between the data points.

To illustrate this better, we have an extended data set for the example that we took. Instead of 6 data points, this set has 22 data points (Table 6). Drawing conclusions from small or large data sets for checking assumption of random error is not recommended.

Table 6 Instantaneous thermal expansion coefficient as a function of temperature.

Temperature	Instantaneous Thermal Expansion
°F	μin/in/°F
80	6.47
60	6.36
40	6.24
20	6.12
0	6.00
-20	5.86
-40	5.72
-60	5.58
-80	5.43
-100	5.28
-120	5.09
-140	4.91
-160	4.72
-180	4.52
-200	4.30
-220	4.08
-240	3.83
-260	3.58
-280	3.33
-300	3.07
-320	2.76
-340	2.45

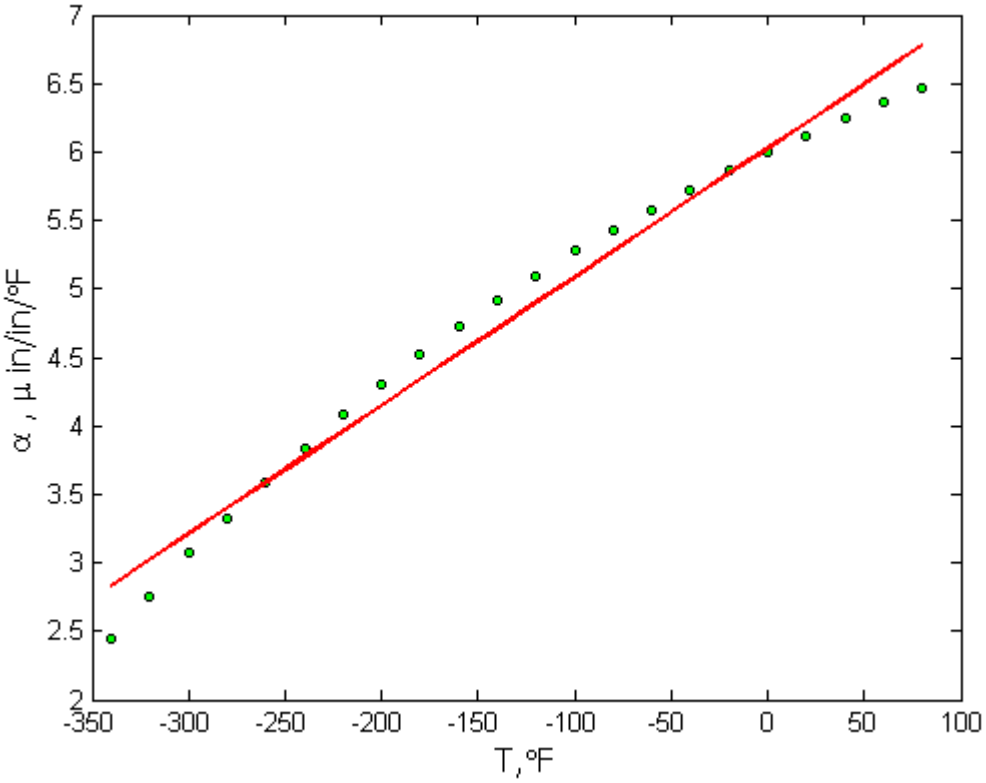

Figure 3 Plot of thermal expansion coefficient vs. temperature data points and regression line for more data points.

Regressing the data from Table 2 to the straight line regression line

$$\alpha(T) = a_0 + a_1 T$$

and following the procedure for conducting linear regression as given in Chapter 06.03, we get (Figure 3)

$$\alpha = 6.0248 + 0.0093868T$$

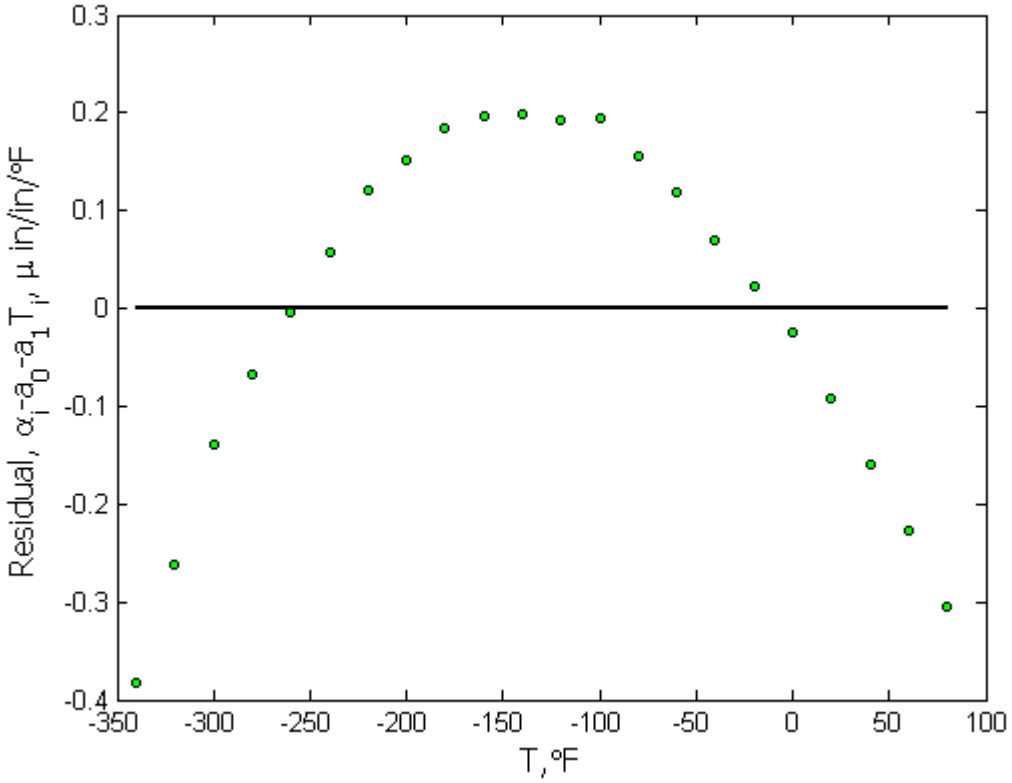

Figure 4 Plot of residuals.

Figure 4 shows the residuals for the example as a function of temperature. Although the residuals seem to average to zero, but within a range, they do not exhibit this zero mean. For an initial value of T, the averages are below zero. For the middle values of T, the averages are below zero, and again for the final values of T, the averages are below zero. This may be considered a violation of the model assumption.

Figure 4 also shows the residuals for the example are following a nonlinear variance. This is a clear violation of the model assumption of constant variance.

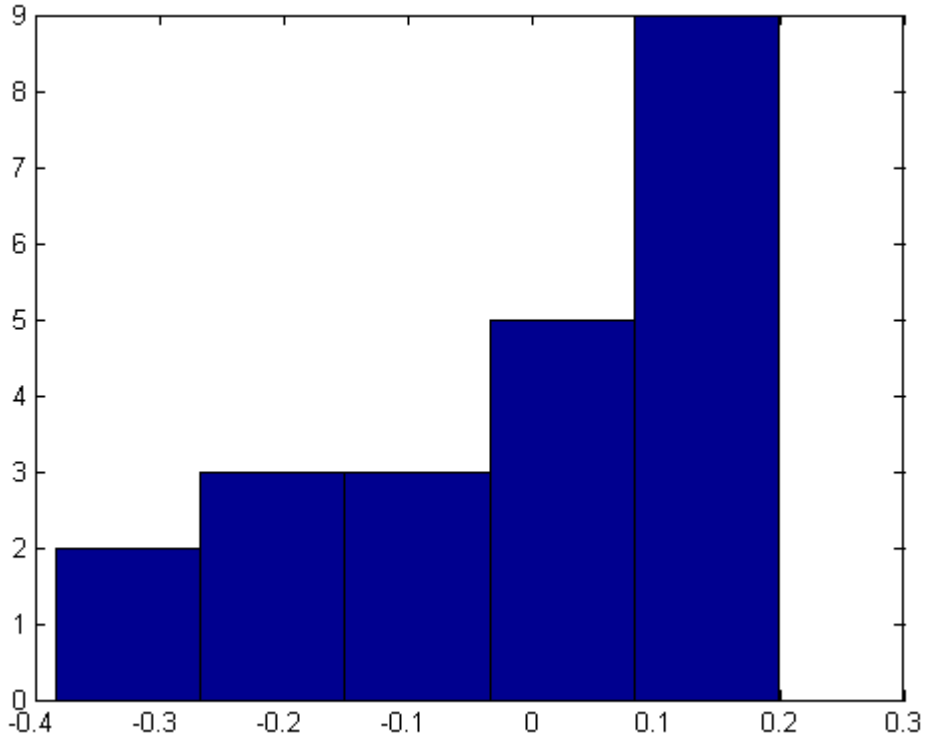

Figure 5 Histogram of residuals.

Figure 5 shows the histogram of the residuals. Clearly, the histogram is not showing a normal distribution, and hence violates the model assumption of normality.

To check that there is no autocorrelation between observed values, the following rule of thumb can be used. If n is the number of data points, and q is the number of times the sign of the residual changes, then if

$$\frac{(n-1)}{2}-\sqrt{n-1}\le q\le \frac{n-1}{2}+\sqrt{n-1},$$

you most likely do not have an autocorrelation. For the example, $n=22$, then

$$\frac{(22-1)}{2}-\sqrt{22-1}\le q\le \frac{22-1}{2}+\sqrt{22-1}$$

$$5.9174\le q\le 15.083$$

is not satisfied as $q=2$. So this model assumption is violated.

References

ADEQUACY OF REGRESSION MODELS	
Topic	Adequacy of Regression Models
Summary	Textbook notes of Adequacy of Regression Models
Major	General Engineering

Authors	Autar Kaw, Egwu Kalu
Date	May 31, 2013
Web Site	http://numericalmethods.eng.usf.edu