Model Building Process Part 1: Checking Model Assumptions
V 1.1

Authored by: Sarah Burke, PhD

24 October 2017

Revised 30 October 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Rev 1, 30 Oct 2018: format and minor editing

Executive Summary

A linear regression model is a valuable method to characterize and analyze test data; however, the conclusions reached are only valid if the underlying assumptions about the data and model hold true. After you fit a model, you must verify that the underlying model assumptions have been met by graphing the data and then checking the assumptions of data independence, constant variance, and normality. Finally, the data must be examined for outliers and leverage points and adjusted if appropriate. The purpose of this best practice is to highlight several techniques, mostly graphical, used to evaluate these assumptions.

Keywords: statistical model, linear regression, independence, constant variance, normality, outliers, leverage points

Introduction

This document is the first part in a series on the steps of the (statistical) model building process. This first part focuses on checking the assumptions of a model, with an emphasis on assessing the validity of the assumptions for linear regression models. Subsequent parts of this series will discuss model diagnostics, model comparisons, and model selection.

Statistical models are valid only if the assumptions about the data or population of interest hold true. Checking the adequacy of a statistical model involves not only ensuring that the model is a good fit to the data, but also verifying that all model assumptions are met. This first task may involve selecting between several potential models to choose the final model, a topic to be discussed in a subsequent part of this series. First, we introduce the model-building process for the case of linear regression and discuss how to assess potential violations of model assumptions for linear regression. The purpose of this best practice is to highlight several techniques, mostly graphical, used to evaluate these assumptions. We also discuss potential remedial measures if model assumptions are violated.

Background

The Linear Regression Model

A linear regression model is used to model, characterize, optimize, and/or predict a continuous response as a function of a set of independent variables. The model has the following (basic) form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

where y_i is the i^{th} observed response in the dataset, $\beta_0, \beta_1, \cdots, \beta_k$ are the model parameters we wish to estimate, $x_{i1}, x_{i2}, \dots x_{ik}$ are the k independent variables (also called factors or predictors) for the i^{th}observation, and ε_i represents the unknown error. The error represents the noise present in the system and thus the noise in the measured response values. The model may also contain interactions between predictors (e.g., $x_1 x_2$), quadratic terms (e.g., x_1^2), or even higher-order terms (e.g., $x_1^2 x_2$).

Consider the following dataset of 10 observations with one predictor and one response, shown in Table 1. Figure 1 shows this dataset with the linear regression line. The equation of this line is $\hat{y} = 1.641 + 3.319x$. This regression equation can be used to make predictions of the response for a given value of the predictor, x. For a one unit increase in the predictor x, there is a 3.319 unit increase in the response.

Table 1: Example dataset

Obs #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
x	6	3	5	10	4	15	8	9	13	4
y	29	8	23	25	9	53	29	36	44	16

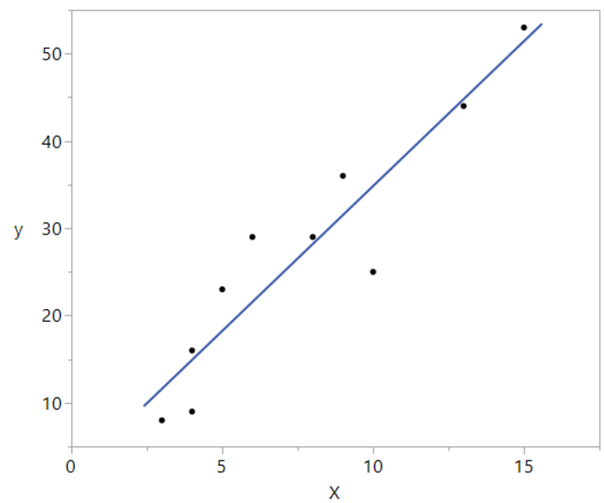

Figure 1: Example of a simple linear regression model

There are three key assumptions in a linear regression model: 1) independent observations; 2) constant variance of error terms; and 3) normally distributed error terms. An additional, implicit assumption is that a linear regression model is appropriate for the data (we will address this assumption in part II of this series). Not all assumptions are treated equal; some have a larger effect on the interpretation and validity of the model. We discuss methods to evaluate the validity of each of these three key assumptions, discuss potential consequences of violating the model assumptions, and provide suggestions on remedial measures if the assumptions are violated.

Method

Preliminary Data Analysis: Graph your data!

The first step when modeling data is to graph it. Histograms, dotplots, or boxplots of your response and predictors are recommended. This preliminary visual check will help you find any data entry errors as well as identify potential outliers (discussed in more detail in the final section). Plots of the response will also provide information on the shape and symmetry of the response which will help determine whether a linear regression model is appropriate. A pairwise scatterplot matrix of your factors and response(s), as shown in Figure 2, is also useful to understand the relationships between your predictors and the response. The bottom row of this plot shows scatterplots of the response by each predictor. In the bottom right corner, for example, the data appear to approximate a curved line. This pattern is indicative of a quadratic relationship between the response and variable X4. Once you have resolved any data entry errors, preliminary plots of the response versus each predictor can indicate potential terms that should be included in the model. Additional information on various types of plots can be found at the following link: http://www.statisticshowto.com/types-graphs/

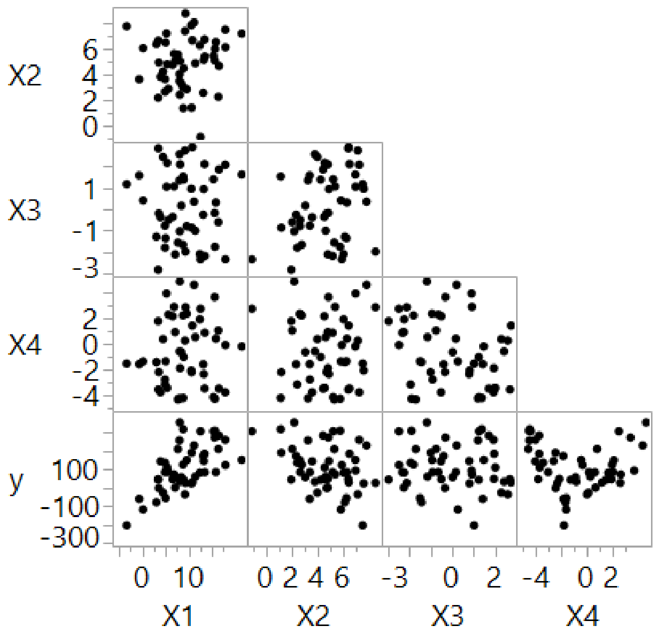

Figure 2: Example of a Scatterplot Matrix

Assessing Linear Regression Model Assumptions

The assumptions of a linear regression model are that the (unknown) true error terms, ε_i, are independent and normally distributed with a mean of 0 and a constant variance σ^2. Diagnostic checks of the model assumptions (and model fit) are often done using the model residuals. The residual is defined for a given observation as the difference between the observed response value, y, and the estimated value of the response, $\hat{y}$, which is determined by the regression model. The residual is, in a sense, the observed error. Consider the previous example as illustrated in Figure 1. This dataset produces the linear regression equation $\hat{y} = 1.641 + 3.319x$. Observation 4 has an x value of 10 and response value 25. The predicted response is $\hat{y} = 1.641 + 3.319(10) = 34.83$. Therefore, the residual for this value is $r_4 = y_4 - \hat{y}_4 = 25 - 34.83 = -9.84$, illustrated in Figure 3 with the label "r_4."

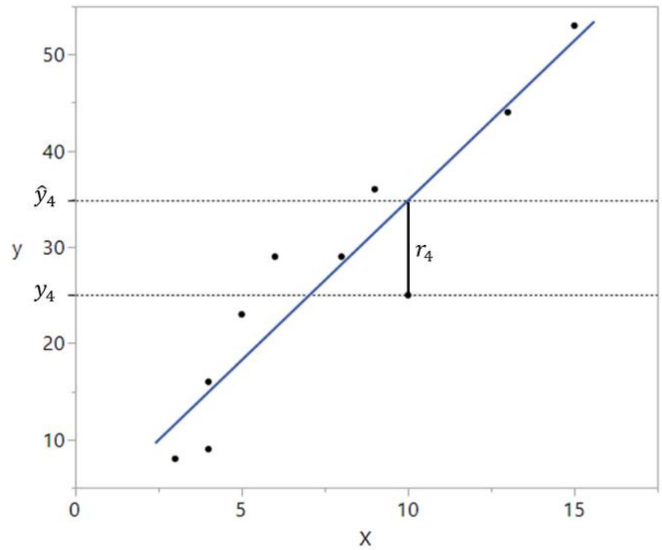

Figure 3: Example of a residual for a simple linear regression model

In the following sub-sections, we discuss various techniques to assess the validity of the model assumptions using residuals.

The Independence Assumption

The first assumption in a linear regression model is that your observations are independent; that is, the value of the response for one observation does not depend on the response of another observation. If your data are collected in a sequence, such as in the run order of a designed experiment or over a given (tracked) time period in an observational study, the independence assumption can be verified. A run chart (also called a time sequence plot) of the residuals indicates if there is a relationship between the residuals over time. Ideally, the points will be randomly scattered around 0 over time. For some observational studies, there may not be information on the time that each data point was collected, making these graphical methods ineffective. In these cases, utilize knowledge of the system to assess whether the independence assumption is satisfied.

Figure 4 depicts several common patterns of run charts. In Figure 4.a, the residuals are randomly scattered around zero, indicating the observations are independent. In Figure 4.b, the residuals tend to increase as the previous value increases and decrease as the previous value decreases. This relationship is called (positive) autocorrelation. Autocorrelation occurs when an observation is correlated with the observation that preceded it. Figure 4.c shows a cyclic pattern of the residuals, potentially indicating systematic changes in the environment, such as a shift change. The Durbin-Watson test, a statistical test for autocorrelation, is used in time series modeling and can also accompany this graphical analysis. The null hypothesis of the Durbin-Watson test is that the data are random (i.e., not autocorrelated). The alternative hypothesis is that the data are autocorrelated.

Figure 4 depicts several common patterns of run charts. In Figure 4.a, the residuals are randomly scattered around zero, indicating the observations are independent. In Figure 4.b, the residuals tend to increase as the previous value increases and decrease as the previous value decreases. This relationship is called (positive) autocorrelation. Autocorrelation occurs when an observation is correlated with the observation that preceded it. Figure 4.c shows a cyclic pattern of the residuals, potentially indicating systematic changes in the environment, such as a shift change. The Durbin-Watson test, a statistical test for autocorrelation, is used in time series modeling and can also accompany this graphical analysis. The null hypothesis of the Durbin-Watson test is that the data are random (i.e., not autocorrelated). The alternative hypothesis is that the data are autocorrelated.

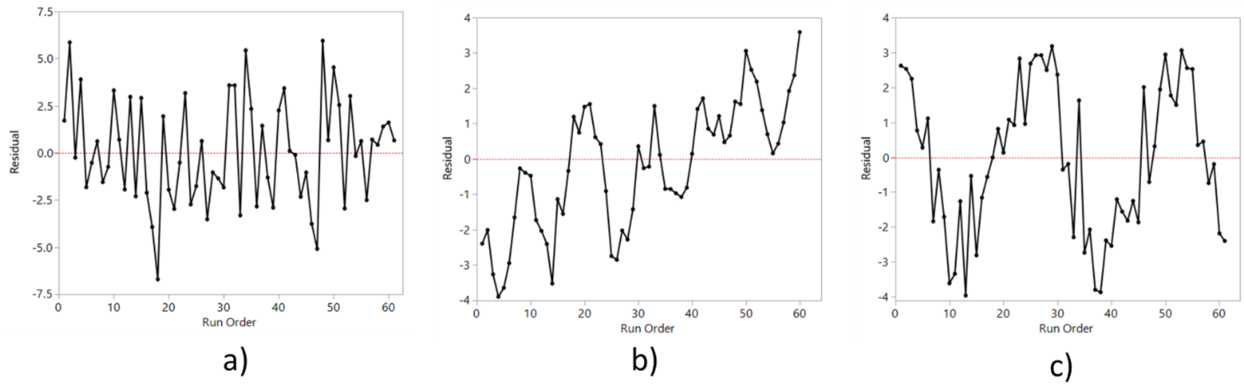

Figure 4: Run Charts of (a) indepenent residuals, (b) autocorrelated residuals, and (c) cyclic residuals

When there is a relationship among the residuals over time as in Figures 4.b and 4.c, the error terms are not independent and there is not an easy correction for this. For example, if there is positive autocorrelation as in Figure 4.b, the regression model may severely underestimate the true variance σ^2. This makes it more likely to conclude a model term is significant when in fact it is not.

One potential remedial measure for autocorrelated residuals is to add a missing predictor into the model because including predictors that have a time effect on the response can correct autocorrelated residuals. This additional term may be a variable not originally considered, particularly if the data came from an observational study. Modeling the differences between consecutive observations sometimes sufficiently removes the autocorrelation of the error terms. Rather than use the response values, you can model the difference in consecutive responses as a function of the predictors in a linear regression model. If your error terms are not independent, the best approach may be to use a time series model that incorporates the autocorrelation present in the data. In these cases, consult your local STAT expert for assistance because time series models can be tricky to deal with. If your data resulted from a designed experiment, this issue of independence highlights the importance of completely randomizing the runs in your design!

The Constant Variance Assumption

The second assumption in a linear regression model is that the variance of the error terms across the range of predictors and/or response is the same value. Nonconstant variance can occur frequently in practice, often when the normality assumption is also violated. In these cases, the variance may be a function of the mean (Montgomery, 2013 p. 243). A plot of the residuals versus each independent variable in the model can indicate issues related to nonconstant variance. A plot of the residuals versus the predicted responses can also identify nonconstant variance issues. Ideally, the residuals are randomly dispersed around 0 for different values of the predictors or the predicted response (Figure 5.a). A non-random pattern in this plot indicates that the magnitude of the residual changes with values in the predictor(s) or the response. A common pattern of nonconstant variance is the “megaphone” or “funnel” pattern as shown in Figure 5.b. This plot indicates that the variance increases as the predictor X increases. Another commonly seen pattern of nonconstant variance is a diamond shape seen in Figure 5.c. In this case, the variability is larger for central values of the predictor and smaller for the extreme values. A statistical hypothesis test for nonconstant variance, such as the Brown-Forsythe test or Bartlett’s test, can also accompany these diagnostic plots. These tests are commonly available in statistical software.

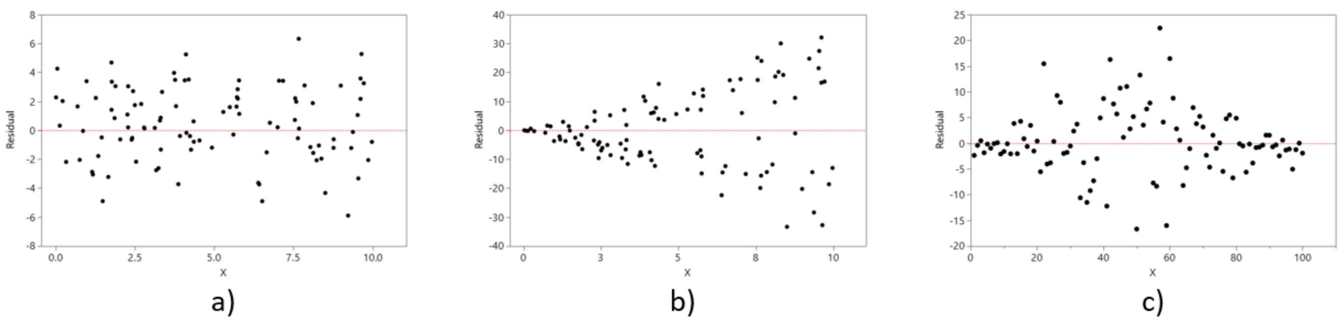

Figure 5: Residual by predictor plot (a) random scatter indicates constant variance; (b) funnel shape indicates nonconstant variance; (c) diamond shape indicates nonconstant variance

The effect of nonconstant variance on the analysis of a linear regression model is less serious than the independence, but still causes issues. The estimated value of the regression coefficient is not affected by a violation in this assumption. However, if there is not constant variance across the responses and/or predictors, the variance of the parameter estimate will not be accurate. This can sometimes lead to the wrong conclusions in a hypothesis test on the significance of a predictor.

If the constant variance assumption is violated, a transformation on the response that stabilizes the variance on the response is recommended. This transformation may be a log transformation of the response variable so that you model the logarithm of the response as a function of the predictors. Alternatively, a square root transformation of the response variable could be used so that you model the square root of the response as a function of the predictors. Other transformations called power transformations (modeling the response raised to some power) may also be appropriate. The square root transformation is an example of a power transformation (since y is raised to the ½ power). The Box-Cox method to select an appropriate power transformation of your response is also available in many statistical software packages. This method searches through all possible powers and mathematically chooses an optimal power level. We caution against always using the exact value resulting from the Box-Cox method in your transformation as they do not always have practical interpretations. For example, the Box-Cox method may indicate the transformation $y^* = y^{0.6}$ is the best choice. In other words, model your response raised to the 0.6 power as a function of the predictors. However, this transformation is much harder to interpret compared to a square root transformation. Once you select a transformation, you should perform the analysis on the transformed response, and verify that the assumptions have been satisfied with the transformation. Remember that the conclusions drawn from this analysis apply to the transformed data, not the original data.

The Normality Assumption

The final assumption of linear regression models is that the error is normally distributed with a mean of zero. If this assumption holds (and the model is a good fit to the data), the residuals should also be normally distributed with a mean of 0. A histogram or dotplot of the residuals can provide a visual check on the shape and symmetry of the residuals around 0. A reasonably large dataset is most effective to detect deviations from the normality assumption; small datasets often exhibit fluctuations that may appear as violations of normality even when there is not a serious violation. Figure 6 shows 3 histograms of residuals, illustrating three commonly seen distributions of residuals. In Figure 6.a, the data is approximately normal. Figure 6.b indicates that the data is skewed to the right while Figure 6.c shows that the tails of the distribution appear to be longer than they should if the residuals truly followed a normal distribution.

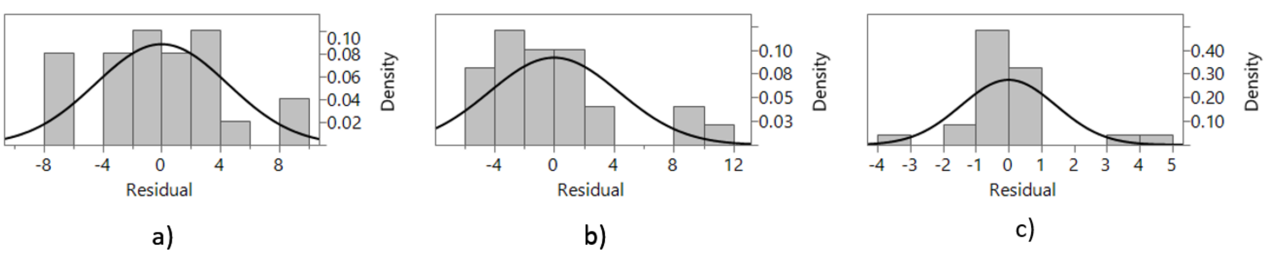

Figure 6: Histograms of model residuals with a normal curve overlaid for (a) normally distributed residuals, (b) residuals skewed right, (c) residuals where the tails are larger than expected in a normal distribution

Analyzing a histogram alone can sometimes be difficult to assess the normality assumption. A normal probability plot can be used to better assess deviations from the normality assumption. In this plot, each residual is plotted against its expected value assuming the normal distribution holds. If the points fall along the plotted line, then the residuals agree with the assumption of normality; if the points deviate from this line, the residuals do not agree with the assumption. Because this is a visual assessment, there is some subjectivity to assessing whether the residuals agree with the normality assumption or not. The "fat pen" test is often used to make the assessment. If you can cover up the points with a pen along the line, the normality assumption holds; otherwise, it does not. Using a normal probability plot is subjective. These plots are often accompanied with confidence bounds. In general, if a straight line reasonably fits through these bounds along the data, the normality assumption is valid. Patterns, however, indicate violations in the assumption, whether all the points fall inside the confidence bounds or not.

Figure 7 shows three normal probability plots for the same model residuals shown in Figure 6. In Figure 7.a, the data agree with the normality assumption since all the points fall reasonably close to the line. The points in Figure7.b have a concave shape, indicating the data is skewed to the right; in other words, the largest residuals are larger than expected and the smallest residuals are not as large as expected. In Figure 7.c, the points form an "S" shape, indicating the tails of the observed values are larger in magnitude than would be expected in a normal distribution; that is, the tails of the distribution are heavier than those for a normal distribution.

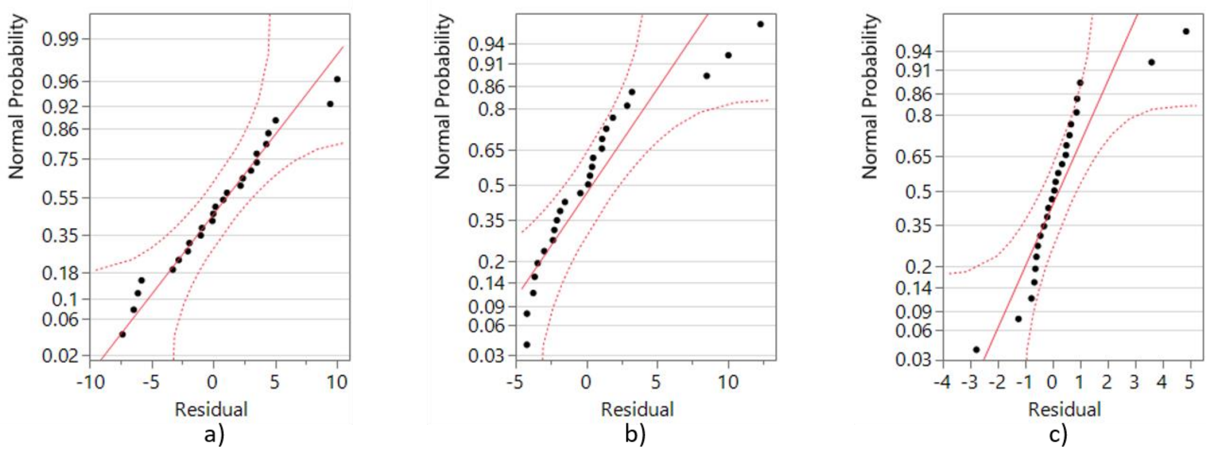

Figure 7: Normal probability plots of error distribution (a) normality assumption holds, (b) error distribution is skewed right, (c) tails of error distribution are larger than expected in normal distribution

A goodness-of-fit test such as the Shapiro-Wilk, Kolmogorov-Smirnov, or Anderson-Darling test can be used to accompany the normal probability plot to assess the normality assumption. The null hypothesis associated with these tests is that the residuals are normally distributed. The alternative hypothesis is that the residuals are not normally distributed. A small p-value, therefore, provides evidence in favor of the alternative hypothesis, that the model residuals are not normally distributed. Use the results of these tests with some caution, however. Normal probability tests have assumptions of their own, which includes that the observations in the test are drawn from a random sample. Model residuals are not drawn from a random sample; they are derived from a model. However, large samples can overcome the limitations of this assumption for these statistical tests. These tests are commonly available in statistical software and can be used to assess normality of any variable, not just model residuals. Figure

8 shows the results of the Shapiro-Wilk test for the same model residuals shown in Figures 6 and 7. These tests agree with the graphical analysis of the normal probability plots in Figure 7.

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.970979	0.6700

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

a)

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.847546	0.0016*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

b)

Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.828452	0.0007*

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

c)

Figure 8: Shapiro-Wilk test for normality (a) normal distribution assumption holds, (b) and (c) normal distribution does not hold

Small to moderate deviations from normality do not heavily influence the analysis of a regression model. (It is said that the *F*-test of the test for significance of regression is "robust" to the normality assumption). If your sample size is very large, even larger deviations of your data from the normal distribution will likely not affect your analysis drastically. A general rule of thumb is a sample size greater than 40. However, large deviations do matter when the sample size is small and should be dealt with. A transformation of the response can help account for discrepancies in the normal distribution. A linear model would then be fit to the transformed response. Another alternative is to adjust the analysis method, for example by using a non-parametric test or an alternative model. Note that non-parametric models are less powerful than linear regression when the normality assumption holds true. In addition, non-parametric does not mean "assumption-free." It only means that there are no assumptions on the distribution of the error terms in the model. If another distribution, such as the exponential or Weibull distribution, has a better fit for the data, it may be more meaningful to use an alternate modeling technique that accounts for this distribution (a generalized linear model, for example). In these more advanced cases, we recommend contacting your local STAT expert or the STAT COE (COE@afit.edu) for assistance.

Outliers and Leverage Points

An outlier is an extreme or unusual observation. One or more outliers can be problematic in fitting a regression model and may distort the analysis. Leverage points, also called influential points, are a special type of outlier that drive the slope of the regression line. Figure 9 presents an extreme example of an outlier and leverage point. In Figure 9.a, there are no outliers or leverage points present in the dataset. In Figure 9.b, observation (x = 13, y = 12.75) is an outlier because it does not follow the same pattern compared to the other points in the data set. However, the linear regression model is not heavily influenced by its inclusion (or exclusion) in the model. In Figure 9.c, the observation (19, 12.5) is a leverage point: removing it from the dataset would cause a large change in the regression line. By graphing the data, we see that a linear model does not adequately characterize this data.

Outliers are often identified graphically or with heuristics. For example, outliers often stand out in a normal probability plot of the residuals, in a plot of the residuals by predicted values, or in a run chart of residuals. One heuristic is to examine standardized residuals, the residual scaled by the estimated root mean square error ($d_i = e_i/\sqrt{MSE}$). If the error terms are normally distributed with mean 0 and variance σ^2, then the standardized residuals should be approximately normal with mean 0 and variance 1. Therefore, observations with a standardized residual greater than three is a potential outlier. Alternative heuristics are Cook's distance and the "hat" value. Refer to Kutner et al (2004) for more information on these heuristics.

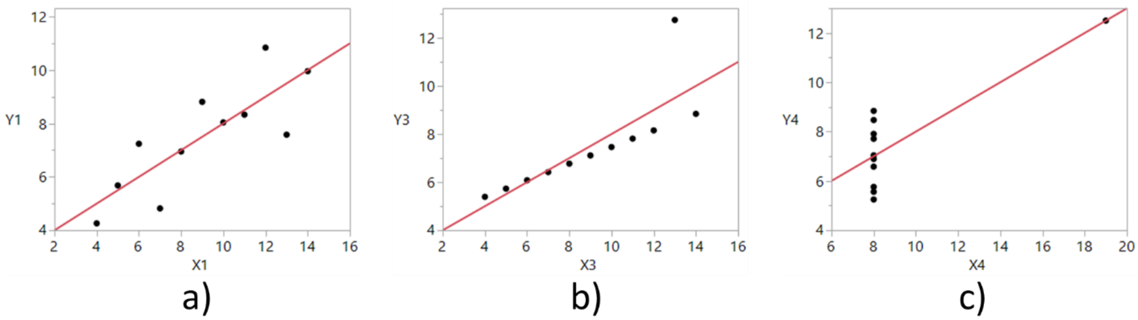

Figure 9: Example by Anscombe (1973) of (a) no outlier present, (b) an outlier, and (c) a leverage point

What should you do about outliers in your data? Investigate for potential causes of the outlier. This includes reviewing the test, checking for input errors, investigating any abnormalities that occurred during the experiment, etc. If you are certain (and there is direct evidence) that the outlying value is the result of an error (e.g., was incorrectly coded) or resulted from a deviation in a planned experiment, it is likely safe to discard the value. However, if there is uncertainty as to why the outlying value occurred, the value may represent important information about the system. Leverage points can have large impacts on the validity of the model. Performing the analysis with and without the outlier or leverage point is often done to see how the results and conclusions differ based on the unusual value. Note that the influence of an individual observation will decrease as the sample size increases.

Conclusions

Regression modeling is a powerful tool that allows us to characterize a response as a function of several independent variables. The results of this analysis, however, are dependent on the model assumptions holding. We have presented several diagnostic methods, mostly graphical, to assess the assumptions of the error terms of the model. Whenever you fit a regression model, be sure to assess the model assumptions and report them. Future documents in this series will discuss model diagnostics, model comparisons, and model selection.

References

Anscombe, F. J. "Graphs in Statistical Analysis." The American Statistician, vol. 27, no. 1, 1973, p. 17.
Kutner, Michael H., et al. Applied Linear Statistical Models. McGraw-Hill, 2004.
Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017.