Chapter 4 SLR Model Assumptions

# Introduction

How do we evaluate a model? How do we know if the model we are using is good? One way to consider these questions is to assess whether the assumptions underlying the simple linear regression model seem reasonable when applied to the data set in question. Since the assumptions relate to the (population) prediction errors, we do this through the study of the (sample) estimated errors, the residuals.

We focus in this lesson on the graphical residual analysis. When we revisit this topic in the context of multiple linear regression in Chapter 7 we will also study some statistical tests for assessing the assumptions. We will consider various remedies for when linear regression model assumptions fail throughout the rest of the course, but particularly in Chapter 9.

# Learning Objectives and Outcomes

Upon completion of this chapter you should:

* Understand why we need to check the assumptions of our model.
* Know the things that can go wrong with the linear regression model.
* Know how we can detect various problems with the model using a “residuals vs fits” plot.
* Know how we can detect various problems with the model using a “residuals vs. predictor” plot.
* Know how we can detect a certain kind of dependent error terms using a “residuals vs. order” plot.
* Know how we can detect non-normal error terms using a normal probability plot.

# Background: Residuals

In this section, we will learn how to check the appropriateness of a simple linear regression model. Recall that the four conditions (“LINE”) - that comprise the simple linear regression model – are:

* The mean of the response, , at each value of the predictor, , is a **Linear function** of the .
* The errors, , are **Independent**.
* The errors, , at each value of the predictor, , are **Normally distributed**.
* The errors, , at each value of the predictor, , have **Equal variances** (denoted ).

An equivalent way to think of the first (linearity) condition is that the mean of the error, , at each value of the predictor, , is zero. An alternative way to describe all four assumptions is that the errors, , are independent normal random variables with mean zero and constant variance, .

The four conditions of the model pretty much tell us what can go wrong with the model, namely:

* The population regression function is not linear. That is, the response is not a function of linear trend plus some error .
* The error terms are not independent
* The error terms are not normally distributed.
* The error terms do not have equal variance.

In this lesson, we learn ways to detect the above four situations, as well as learn how to identify the following two problems:

* The model fits all but one or a few unusual observations. That is, are there any “outliers?”
* An important predictor variable has been left out of the model. That is, could we do better by adding a second or third predictor into the model, and instead use a multiple regression model to answer our research questions?

Before jumping in, let’s make sure it’s clear why we have to evaluate any regression model that we formulate and subsequently estimate. In short, it’s because:

* All of the estimates, intervals, and hypothesis tests arising in a regression analysis have been developed assuming that the model is correct. That is, all the formulas depend on the model being correct!
* If the model is incorrect, then the formulas and methods we use are at risk of being incorrect.

The good news is that some of the model conditions are more forgiving than others. So, we really need to learn when we should worry the most and when it’s okay to be more carefree about model violations. Here’s a pretty good summary of the situation:

* All tests and intervals are very sensitive to even minor departures from independence.
* All tests and intervals are sensitive to moderate departures from equal variance.
* The hypothesis tests and confidence intervals for and are fairly “robust” (that is, forgiving) against departures from normality.
* Prediction intervals are quite sensitive to departures from normality (Do they mean linearity?)

The important thing to remember is that the severity of the consequences is always related to the severity of the violation. And, how much you should worry about a model violation depends on how you plan to use your regression model. For example, if all you want to do with your model is test for a relationship between and , i.e. test that the slope of is 0, you should be okay even if it appears that the normality condition is violated. On the other hand, if you want to use your model to predict a future response , then you are likely to get inaccurate results if the error terms are not normally distributed.

In short, you’ll need to learn how to worry just the right amount. Worry when you should, and don’t overworry when you shouldn’t! When you are worried, there are remedies available, which we will learn about in an in-depth manner later. For now, one thing to try is transforming either the response variable, predictor variable, or both – there is an example of this later in this chapter, and we will see more examples in chapter 9.

## The Basic Idea of Residual Analysis

Recall that not all of the data points in a sample will fall right on the least squares regression line. The vertical distance between any one point and its estimated value is its observed “**residual**”:

Each observed residual can be thought of as an estimate of the actual unknown “**true error**” term:

Let’s look at an illustration of the distinction between a residual and an unknown true error term . The solid line on the plot describes the true (unknown) linear relationship in the population. Most often, we can’t know this line. However, if we could, the true error would be the distance from the data point to the solid line.

|  |  |
| --- | --- |
|  |  |

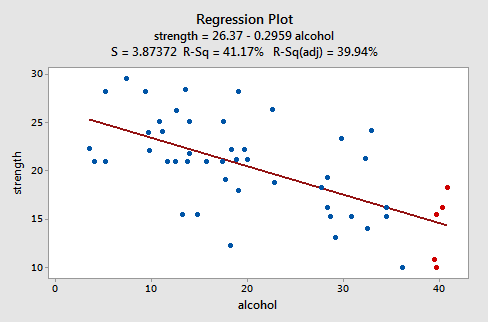
On the other hand, the dashed line on the plot represents the estimated linear relationship for the sample of students. The residual error is the distance from the data point to the dashed line. Compare the two types of errors – the true error and the residual error – depicted for the blue data point.

The observed residuals should reflect the properties assumed for the unknown true error terms. The basic idea of residual analysis, therefore, is to investigate the observed residuals to see if they behave “properly.” That is, we analyze the residuals to see if they support the assumptions of linearity, independence, normality, and equal variances.

# Residuals vs. Fits Plot

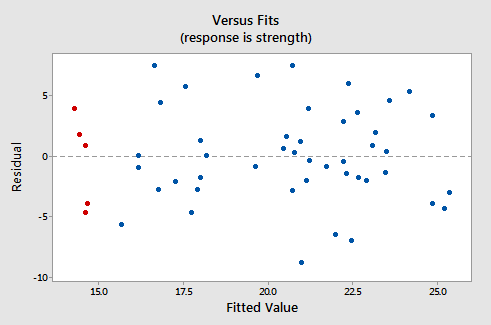
When conducting a residual analysis, a “residuals versus fists plot” is the most frequently created plot. It is a scatter plots of residuals on the axis and fitted values (estimated responses) on the axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

Let’s look at an example to see what a “well-behaved” residual plot looks like. Some researchers (Urbano-Marques, et al., 1989) were interested in determining whether or not alcohol consumption was linearly related to muscle strength. The researchers measured the total lifetime consumption of alcohol, , on a random sample of alcoholic men. They also measured the strength, , of the deltoid muscle in each person’s nondominant arm. A fitted line plot of the resulting data, ([alcoholarm.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/alcoholarm.txt)), looks like:



The plot suggests that there is a decreasing linear relationship between alcohol and arm strength. It also suggests that there are no unusual data points in the data set. And, it illustrates that the variation around the estimated regression line is constant suggesting that the assumption of equal variances is reasonable.

Here’s what the corresponding residuals versus fits plot looks like for the data set’s simple linear regression model with arm strength as the response and level of alcohol consumption as the predictor:



Note that, as defined, the residuals appear on the axis and the fitted values appear on the axis. You should be able to look back at the scatter plot of the data and see how the data points there correspond to the data points in the “residual vs. fits” plot here. In case you’re having trouble with doing that, look at the five data points in the original scatter plot that appear in red. Note that the predicted response (fitted value) of these men (whose alcohol consumption is around 40) is about 14. Also, note the pattern in which the five data points deviate from the estimated regression line.

Now look at how and where these five data points appear in the “residuals versus fits” plot. Their fitted value is about 14 and their deviation from the residual = 0 line shares the same pattern as their deviation from the estimated regression line. Do you see the connection? Any data point that falls directly on the estimated regression line has a residual of 0. Therefore, the residual = 0 line corresponds to the estimated regression line.

This plot is a classic example of a well-behaved “residuals vs. fits” plot. Here are the characteristics of a well-behaved residuals vs. fits plot and what they suggest about the appropriateness of the simple linear regression model:

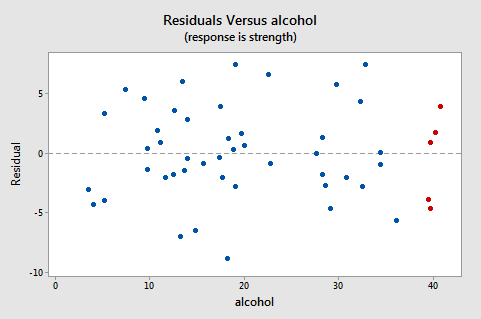
* The residuals “bounce randomly” around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
* The residuals roughly form a “horizontal band” around the 0 line. This suggests that the variances of the error terms are equal.
* No one residual “stands out” from the basic random pattern of residuals. This suggests that there are no outliers.

In general, you want your “residual vs. fits” plots to look something like the above plot. Don’t forget though that interpreting these plots is subjective. My experience has been that student’s learning residual analysis for the first time tend to over-interpret these plots, looking at every twist and turn as something potentially troublesome. You’ll especially want to be careful about putting too much weight on “residual vs. fits” plots based on small data sets. Sometimes the data sets are just too small to make interpretations of a “residuals vs. fits” plot worthwhile. Don’t worry! You will learn – with practice – how to “read” these plots.

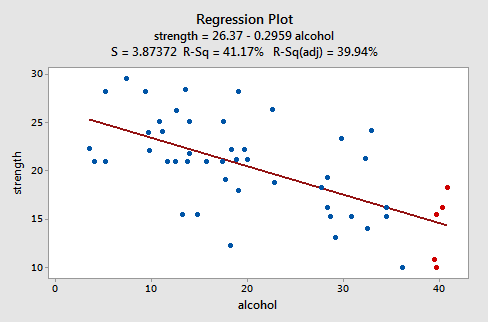
# Residual vs. Predictor Plot

An alternative to the “residuals vs. fits” plot is a “**residuals vs. predictor**” plot. It is a scatter plot of residuals on the axis and the predictor () on the axis. For a simple linear regression model, if the predictor on the axis is the same predictor that is used in the regression model, the “residuals vs. predictor” plot offers no new information to that which is already learned by the “residuals vs. fits” plot. On the other hand, if the predictor on the axis is a new and different predictor, the “residuals vs. predictor” plot can help determine whether the predictor should be added to the model (and hence a multiple regression model used instead).

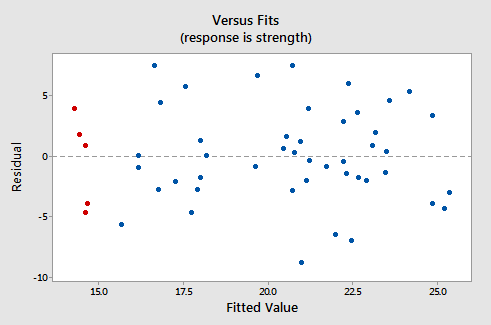
The interpretation of a “residual vs. predictor” plot is identical to that for a “residuals vs. fits” plot. That is, a well-behaved plot will bounce randomly and form a roughly horizontal band around the residual = 0 line. And, no data points will stand out from the basic random pattern of the other residuals.



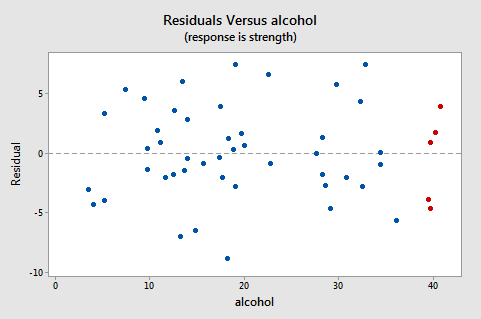
Note that, as defined, the residuals appear on the axis and the predictor values – the lifetime alcohol consumptions for the men – appear on the axis. Now, you should be able to look back at the scatter plot of the data:



And the residuals vs. fits plot:



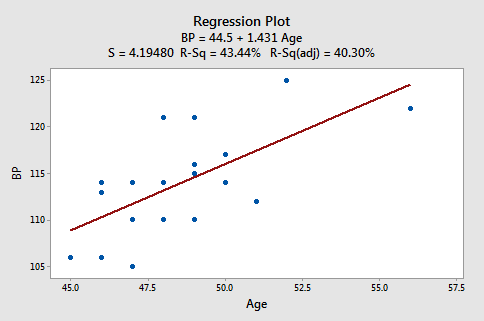
To see how the data points there correspond to the data points in the residuals versus predictor plot:



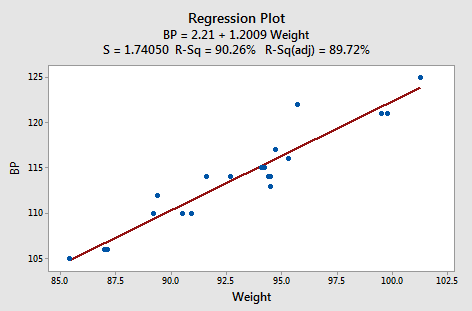
The five red data points should help you out again. The alcohol consumption of the five men is about, and hence why the points now appear on the “right side” of the plot. In essence, for this example, the “residuals vs. predictor” plot is just a mirror image of the “residuals vs. fits” plot. The “residuals vs. predictor” plot offers no new information

Let’s take a look at an example in which the “residuals vs. predictor” plot is used to determine whether or not another predictor should be added to the model. A researcher is interested in determining which if the following – age, weight, and duration of hypertension – are good predictors of the diastolic blood pressure of an individual with high blood pressure. The researcher measured the age (in years), weight (in pounds), duration of hypertension (in years), and a diastolic blood pressure (in mm Hg) on a sample of hypertensive individuals ([bloodpress.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/bloodpress.txt)).

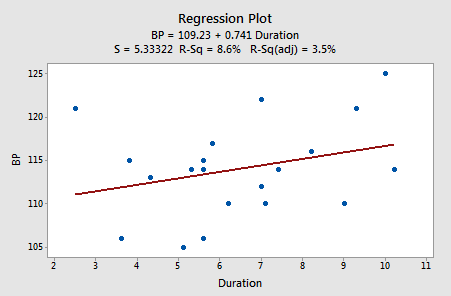
The regression of the response diastolic blood pressure (BP) on the predictor age:



Suggests that there is a moderately strong linear relationship (43.4%) between diastolic blood pressure and age. The regression of the response diastolic blood pressure (BP) on the predictor weight:

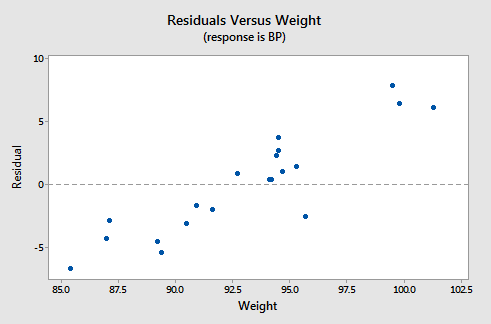


Suggest that there is a strong linear relationship (90.3%) between diastolic blood pressure and weight. And, the regression of the response diastolic blood pressure (BP) on the predictor duration:



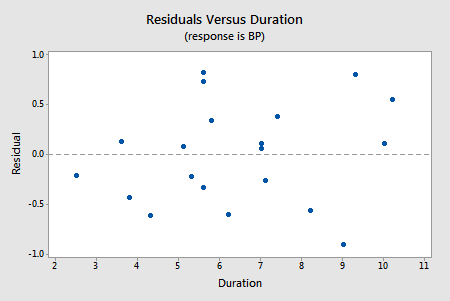
Suggest that there is little linear association (8.6%) between diastolic blood pressure and duration of hypertension. In summary, it appears as if weight has the strongest association with diastolic blood pressure, age has the second strongest association, with duration the weakest.

Let’s investigate various “residuals vs. predictors” plots to learn whether adding predictors to any of the above three simple linear regression models is advised. Upon regressing blood pressure on age, obtaining the residuals, and plotting the residuals against the predictor weight, we obtain the following “residuals vs. weight” plot:



This “residuals vs. weight” plot can be used to determine whether we should add the predictor weight to the model that already contains the predictor age. In general, if there is some non-random pattern to the plot, it indicates that it would be worthwhile adding the predictor to the model. In essence, you can think of the residuals on the axis as a “new response” against a predictor shows a non-random pattern, it indicates that the predictor explains some of the remaining variability in the new (adjusted) response. Here, there is a pattern in the plot. It appears that adding the predictor weight to the model already containing age would help to explain some of the remaining variability in the response.

We haven’t yet learned about multiple linear regression models – regression models with more than one predictor. But, you’ll soon learn that it’s a straightforward extension of a simple linear regression. Suppose we fir the model with blood pressure as the response and age and weight as the two predictors. Should we also add the predictor duration to the model? Let’s investigate! Upon regressing blood pressure on weight and age, obtaining the residuals, and plotting the residuals against the predictor duration, we obtain the following “residuals vs. duration” plot:



The points on the plot show no clear pattern or trend, suggesting that there is no relationship between the residuals and duration. That is, the “residuals vs. duration” plot tells us that there is no sense in adding duration to the model that already contains age and weight. Once we’ve explained the variation in the individual’s blood pressures by taking into account the individual’s ages and weights, none of the remaining variability can be explained by the individual’s durations.

# Identifying Specific Problems Using Residual Plots

In this section, we will learn how to use “residuals vs. fits (or predictor)” plots to detect problems with our formulated regression model. Specifically, we investigate:

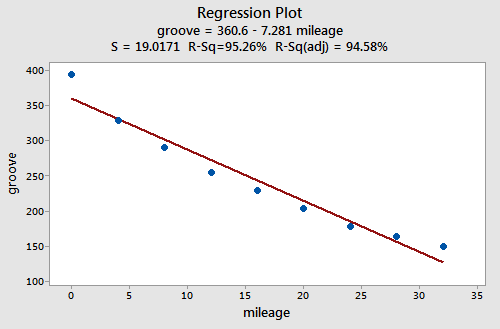
* How a non-linear regression function shows up on a “residuals vs. fits” plot
* How unequal error variances show up on a “residuals vs. fits” plot
* How an outlier show up on a “residuals vs. fits” plot

Note that although we will use “residuals vs. fits” plots throughout our discussion here, we just as easily could use “residuals vs. predictor” plots (provided the predictor is the one in the model).

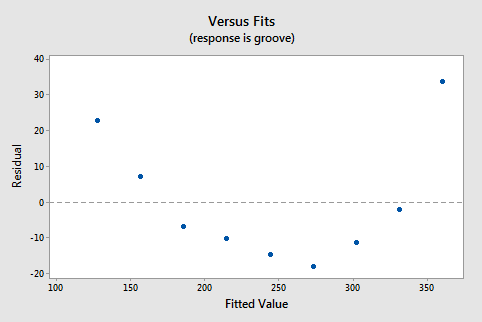
## How does a non-linear regression function show up on a “residual vs. fits” plot?

**The Answer**: The residuals depart from 0 in some systematic manner, such as being positive for small values, negative medium values, and positive again for large values. Any systematic (non-random) pattern is sufficient to suggest that the regression function is not linear.

**An Example**: Is tire tread wear linearly related to mileage? A laboratory (Smith Scientific Services, Akron, OH) conducted an experiment in order to answer this research question. As a result of the experiment, the researchers obtained a data set ([treadwear.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/treadwear.txt)) containing the mileage (, 1000 miles) driven and the depth of the remaining groove (, in mils). The fitted line plot of the resulting data:



Suggests that there is a relationship between groove depth and mileage. The relationship is just not linear. As is generally the case, the corresponding “residuals vs. fits” plot accentuates this claim:



Note that the residuals depart from 0 in a systematic manner. They are positive for small values, negative for medium values, and positive again for large values. Clearly, a non-linear model would better describe the relationship between the two variables.

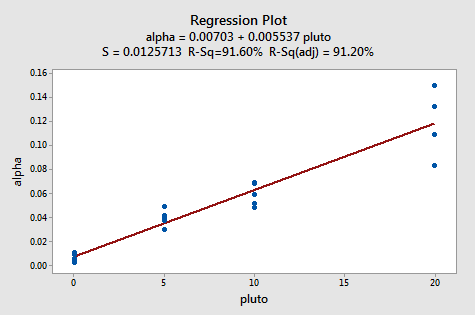
Incidentally, did you notice that the value is very high (95.26%)? This is an excellent example of the caution “a large value should not be interpreted as meaning that the estimated regression line fits the data well.” The large value tells you that if you wanted to predict groove depth, you’d better off taking into account mileage than not. The “residuals vs. fits” plot tells you, though, that your prediction would be better if you formulated a non-linear model rather than a linear one.

## How does non-constant error variance show up on a “residual vs. fits” plot?

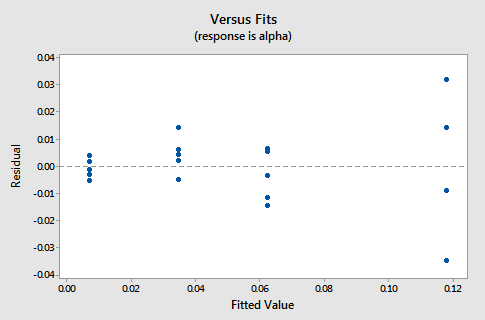
**The Answer**: Non-constant error variance shows up on a “residual vs. fits (or predictor)” plot in any of the following ways:

* The plot has a “**fanning**”. Effect. That is, the residuals are close to 0 for small values and are more spread out for large values.
* The plot has a “**funneling**” effect. That is, the residuals are spread out for small values and close to 0 for large values.
* Or, the spread of the residuals in the “residuals vs. fits” plot varies in some complex fashion.

**An example**: How is plutonium activity related to alpha particle counts? Plutonium emits subatomic particles – called alpha particles. Devices used to detect plutonium record the intensity of alpha particle strikes in counts per second. To investigate the relationship between plutonium activity (, in pCi/g) and alpha count rate (, in number per second), a study was conducted on 23 samples of plutonium. The following fitted line plot was obtained on the resulting data ([alphapluto.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/alphapluto.txt)):



The plot suggests that there is a linear relationship between alpha count rate and plutonium activity. It also suggests that the error terms vary around the regression line in a non-constant manner – as the plutonium level increases, not only does the mean alpha count rate increase, but also the variance increases. That is, the fitted line plot suggests that the assumption of equal variance is violated. As is generally the case, the corresponding “residuals vs. fits” plot accentuates this claim:

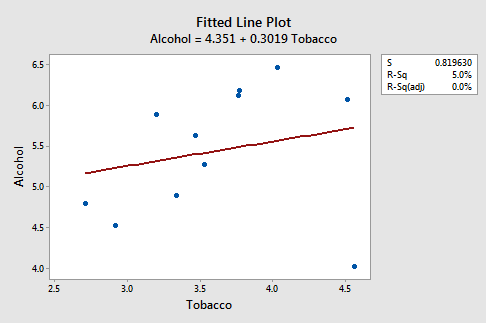


Note that the residuals “fan out” from the left to right rather than exhibiting a constant spread around the residual = 0 line. The “residual vs. fits” plot suggests that the error variances are not equal.

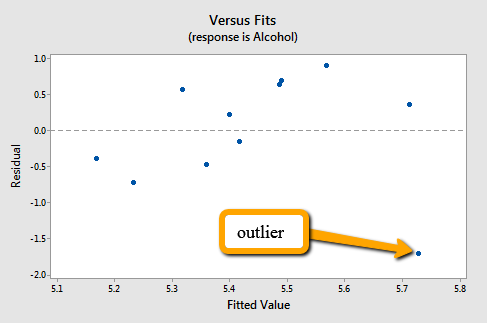
How does an outlier show up on a “residuals vs. fits” plot?

**The Answer**: The observation’s residual plot can even disappear if one outlier really deviates from the basic pattern of the rest of the residuals. The random pattern of the residual plot can even disappear if one outlier really deviates from the pattern of the rest of the data.

**An Example**: Is there a relationship between tobacco use and alcohol use? The British government regularly conducts surveys on household spending. One such survey (Family Expenditure Survey, Department of Employment, 1981) determined the average weekly spending on tobacco (, in British pounds) and the average weekly expenditure on alcohol (, in Bristish pounds) for households in different regions in the United Kingdom. The fitted line plot of the resulting data ([alcholtobacco.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/alcoholtobacco.txt)):

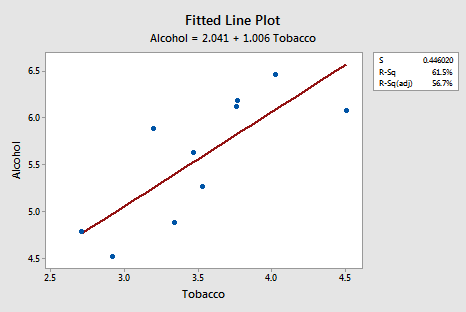


Suggests that there is an outlier – in the lower right corner of the plot – which corresponds to the Northern Ireland region. In fact, the outlier is so far removed from the pattern of the rest of the data that it appears to be “pulling the line” in its direction. As is generally the case, the corresponding “residuals vs. fits” plot accentuates this claim:



Note that Northern Ireland’s residual stands apart from the basic random pattern of the rest of the residuals. That is, the “residual vs. fits” plot suggests that an outlier exists.

Incidentally, this is an excellent example of the caution that the “coefficient of determination can be greatly affected by just one data point.” Note above that the value on the data set with all regions included is 5%. Removing Northern Ireland’s data point from the data set, and refitting the regression line, we obtain:

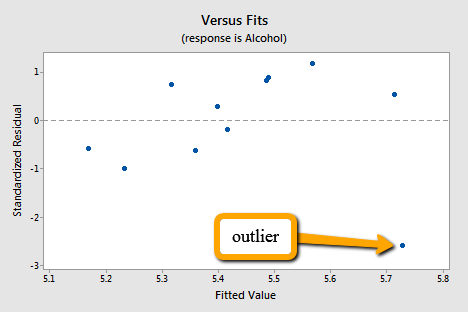


The value has jumped from 5% (“no relationship”) to 61.5% (“moderate relationship”)! Can one data point greatly affect the value of ? Clearly, it can!

Now, you might be wondering how large a residual has to be before a data point should be flagged as being an outlier. The answer is not straightforward, since the magnitude of the residuals depends on the units of the response variables. That is, if your measurements are made in pounds, then the units of the residuals are in pounds. And, if your measurements are made in inches, then the units of the residuals are in inches. Therefore, there is no “rule of thumb” that we can define to flag a residual as being exceptionally unusual.

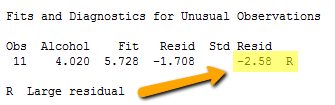
There’s a solution to this problem. We can make the residuals “unitless” by dividing them by their standard deviation. In this way, we create what are called “**standardized residuals**.” They tell us how many standard deviations above – if positive – or below – if negative – a data point is from the estimated regression line. (Note that there are a number of alternative ways to standardize residuals, which we will consider in Chapter 11.) Recall that the empirical rule tells us that, for the data that are normally distributed, 95% of the measurements fall within two standard deviations of the mean. Therefore, any observations with a standardized residual greater than 2 or smaller than -2 might be **flagged for further investigation**. It is important to note that by using this “greater than 2, smaller than -2 rule,” approximately 5% of the measurements in a data set will be flagged even though they are perfectly fine. It is in your best interest not to treat this rule of thumb as a cut-and-dry, believe-it-to-the-bone, hard-and-fast rule! So, in most cases it may be more practical to investigate further any observations with a standardized residual greater than 3 or smaller than -3 (using the empirical rule we would expect only 0.2% of observations to fall into this category).

The corresponding standardized residuals vs. fits plot for our expenditure survey example looks like:



The standardized residual of the suspicious data point is smaller than -2. That is, the data point lies more than 2 standard deviations below its mean. Since this is such a small data set the data point should be flagged for further investigation!

Incidentally, Minitab (and most other statistical software) identifies observations with large standardized residuals. Here is what a portion of Minitab’s output for our expenditure survey example looks like:

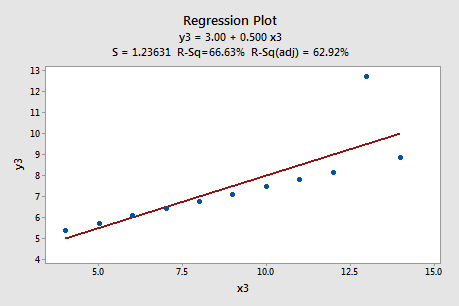


Minitab labels observations with large standardized residuals with an “R.” For our example, Minitab reports that observation #11 – for which tobacco = 4.56 and alcohol = 4.02 – has a large standardized residual (-2.58). The data point has been flagged for further investigation.

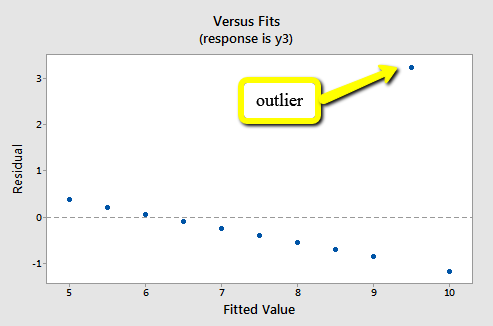
Note that I have intentionally used the phrase “flagged for investigation.” I have not said that the data point should be “removed.” Here’s, my recommendation strategy, once you’ve identified a data point as being unusual:

1. Determine whether a simple – and therefor correctable – mistake was made in recording or entering the data point. Examples include transcription errors (recording 62.1 instead of 26.1) or data entry errors (entering 99.1 instead of 9.1). Correct the mistakes you found.
2. Determine if the measurement was made in such a way that keeping the experimental unit in the study can no longer be justified. Was some procedure not conducted according to the study guidelines? For example, was a person’s blood pressure measured standing up rather than sitting down? Was the measurement made on someone not in the population of interest, i.e. was the survey completed by a man instead of a woman? If it is convincingly justifiable, remove the data point from the data set.
3. If the first two steps don’t resolve the problem, consider analyzing the data twice – once with the data point included and once with the data point excluded. Report the results of both analyses.

**Another example**: The Anscombe data set #3 ([anscombe.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/anscombe.txt)) presents us with another example of an outlier. The fitted line plot suggests that one data point does not follow the trend in the rest of the data.



Here’s what the residual vs. fits plot looks like:



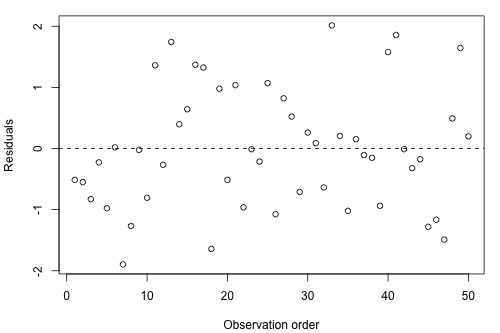
The ideal random pattern of the residual plot has disappeared, since the one outlier really deviates from the pattern of the rest of the data.

# Residuals vs. Order Plot

Recall that the second condition – the “I” condition – of the linear regression model is that the error terms are independent. In this section, we learn how to use a “**residuals vs. order plot**” as a way of detecting a particular form of non-independence of the error terms, namely **serial correlation**. If the data are obtained **in a time (or space) sequence**, a “residual vs. order plot” helps to see if there is any correlation between the error terms that are near each other in the sequence.

**The plot is only appropriate if you know the order in which the data were collected!** Highlight this, underline this, circle this, …, err, on second thought, don’t do that if you are reading it on a computer screen. Do whatever it takes to remember it though – it is a very *common* mistake made by people new to regression analysis.

So, what is this “residual vs order plot” all about? As its name suggests, it is a scatter plot with residuals on the axis and the order in which the data were collected on the axis. Here’s an example of a well-behaved “residuals vs. order plot”:

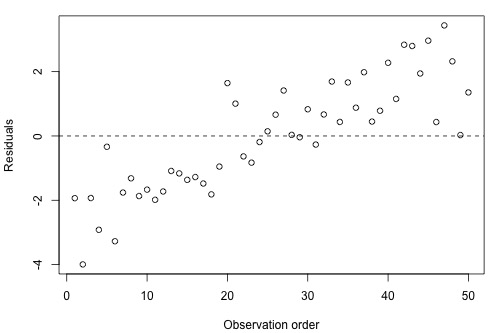


The residuals bounce randomly around the residual = 0 line as we would hope so. In general, residuals exhibiting normal random noise around the residual = 0 line suggests that there is no serial correlation.

Let’s take a look at examples of the different kinds of “residuals vs. order plots” we can obtain and learn what each tells us.

## A time trend

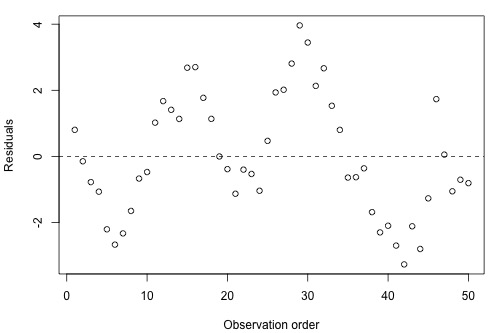
A “residuals vs. order plot” that exhibits (positive) trend as the following plot does:



Suggests that some of the variation in the response is due to times. Therefore, it may be a good idea to add the predictor line to the model. That is, you interpret this plot just as you would interpret any other residuals vs. predictor plot. It’s just that here your predictor is “time.”

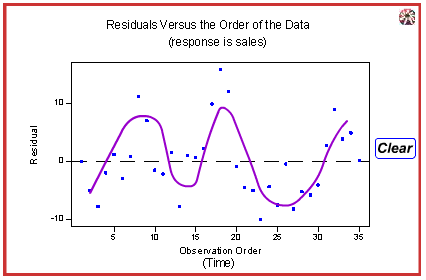
## Positive Serial Correlation

A residual vs. order plot that looks like the following plot:



Suggests that there is “positive serial correlation” among the error terms. That is, positive serial correlation exists when residuals tend to be followed, in time, by residuals of the same sign and about the same magnitude. The plot suggests that the assumption of independent error terms is violated.

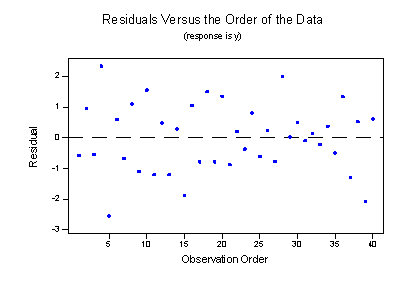
Here is another less obvious example of a dataset exhibiting positive serial correlation:



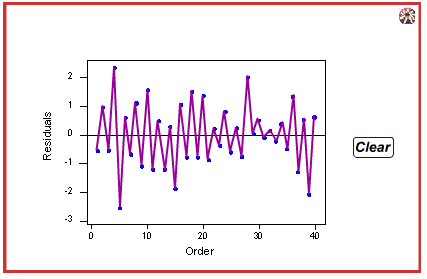
Can you see a cyclical trend – up and down up and down, and up again? Certainly, the positive serial correlation in the error terms is not as obvious here as in the previous example. These two examples taken together are a nice illustration of the “severity of the consequences is related to the severity of the violation.” The violation in the previous example is much more severe than in this example. Therefore, we should expect that the consequences of using a regression model in the previous example would be much greater than using one in this example. In either case, you would be advised to, move out of the realm of regression analysis and into that of “**time series modeling**.”

Negative Serial Correlation

A residual vs. order plot that looks like the following plot:



Suggests that there is “negative serial correlation” among the error terms. Negative serial correlation exists when residuals of one sign tend to be followed, in time, by residuals of the opposite sign. What? Can’t you see it? If you connect the dots from left to right, you should be able to see the pattern. Observe the following plot for a more detailed description:



Negative, positive, negative, positive, and so on. The plot suggests that the assumption of independent error terms is violated. If you obtain a “residual vs. order plot” that looks like this, you would again be advised to move out of the realm of regression analysis and into that of “**time series modeling**.”

# Normal Probability Plot of Residuals

Recall that the third condition – the “N” condition – of the linear regression model is that the error terms are normally distributed. In this section, we learn how to use a “**normal probability plot of the residuals**” as a way of learning whether it is reasonable to assume that the error terms are normally distributed.

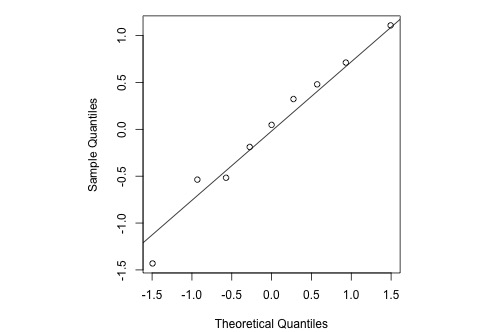
Here’s the basic idea behind any normal probability plot: if the data follow a normal distribution with mean and variance , then a plot of the **theoretical percentiles of the normal distribution** versus the observed **sample percentiles** should be approximately linear. Since we are concerned about the normality of the error terms, we create a normal probability plot of the residuals. If the resulting plot is approximately linear, we proceed assuming that the error terms are normally distributed.

The **theoretical -th percentile** of any normal distribution is the value such that % of the measurements fall below the value.

The problem is that to determine the percentile value of a normal distribution, you need to know the mean and the variance . And, of course, the parameters and are typically unknown. Statistical theory says it’s okay just to assume that and . Once you do that, determining the percentiles of the standard normal curve is straightforward. The -th percentile values reduces to just a “-score” (or “normal score”).

The sample -th percentile of any data set is , roughly speaking, the value such that % of the measurements fall below the value. For example, the median, which is just a special name for the 50th percentile, is the value so that 50%, or half, of your measurements fall below the value. Now, if you are asked to determine the 27th percentile, you take your ordered data set, and you determine the value so that 27% of the data points in your data set fall below the value. And so on.

Consider a simple linear regression model fit to a simulated data set with 9 observations, so that we’re considering the 10th, 20th, …, 90th percentiles. A normal probability plot of the residuals is a scatter plot with the theoretical percentiles of the normal distribution in the axis and the samples percentiles of the residuals on the axis, for example:



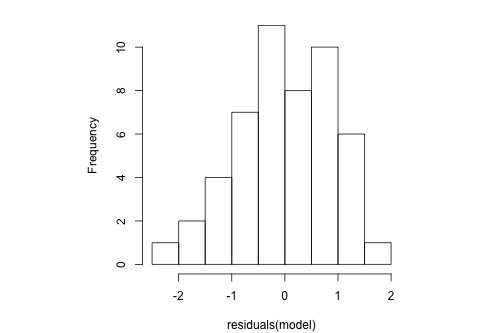
Note that the relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, the normal probability plot of the residuals suggest that the error terms are indeed normally distributed.

Statistical software sometimes provides normality tests to complement the visual assessment available in a normal probability plot (we’ll revisit normality tests in Chapter 7).

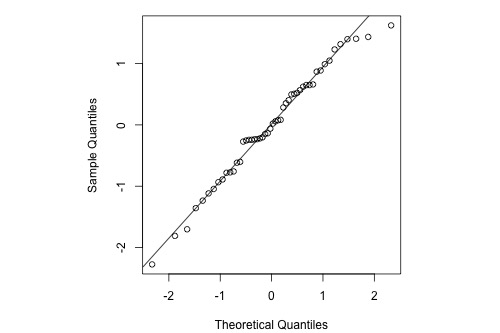
Let’s take a look at examples of the different kinds of normal plots we can obtain and learn what each tells us.

## Normally Distributed Residuals

The following histogram of residuals suggests that the residuals (and hence the error terms) are normally distributed:

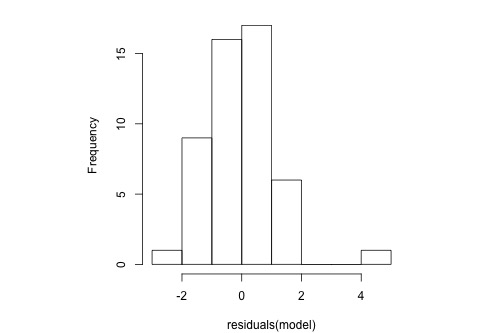


The normal probability plot of the residuals is approximately linear supporting the condition that the error terms are normally distributed.

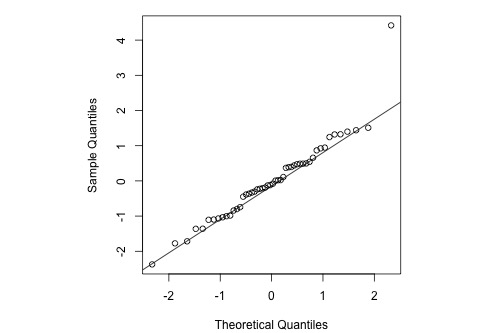


## Normal Residuals but with One Outlier

The following histogram of residuals suggests that the residuals (and hence the error terms) are normally distributed. But, there is one extreme outlier (with value larger than 4):



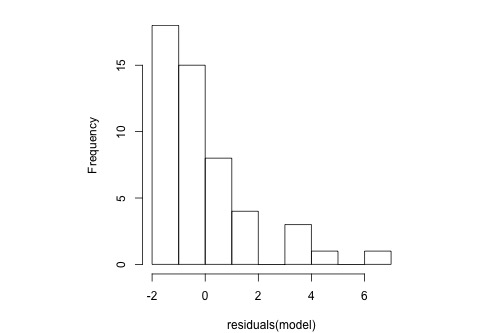
Here’s the corresponding normal probability plot of the residuals:



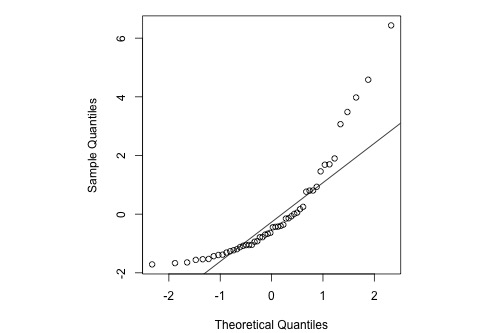
This is a classic example of what a normal probability plot looks like when the residuals are normally distributed, but there is just one outlier. The relationship is approximately linear with the exception of the one data point. We could proceed with the assumption that the error terms are normally distributed upon removing the outlier from the data set.

## Skewed Residuals

The following histogram of residuals suggests that the residuals (and hence the error terms) are not normally distributed. On the contrary, the distribution of the residuals is quite skewed.



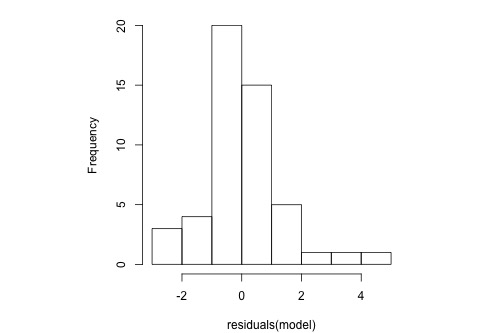
Here’s the corresponding normal probability plot of the residuals:



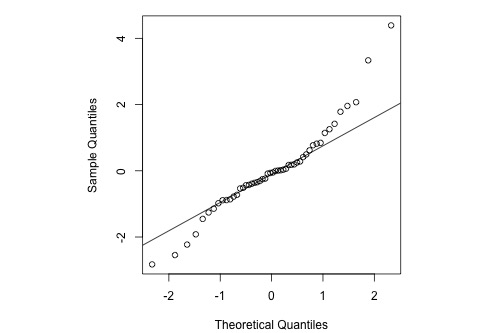
This is a classic example of what a normal probability plot looks like when the residuals are skewed. Clearly, the condition that the error terms are normally distributed is not met.

## Heavy-tailed Residuals

The following histogram of residuals suggest that the residuals (and hence the error terms) are not normally distributed. There are too many extreme positive and negative residuals. We say the distribution is “**heavy tailed**.”



Here’s the corresponding normal probability plot of the residuals:



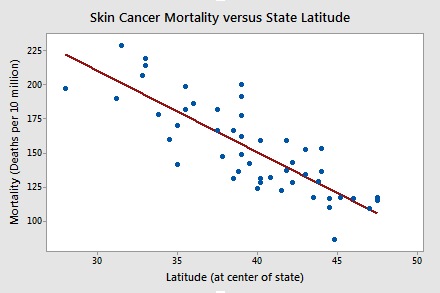
The relationship between the sample percentiles and theoretical percentiles is not linear. Again, the condition that the error terms are normally distributed is not met.

# Assessing Linearity by Visual Inspection

The first simple linear regression model condition concerns linearity” the mean of the response at each predictor value should be a linear function of the predictor. The neat thing about simple linear regression – in which there is a response and just one predictor – is that we can get a good feel for this condition just by looking at a simple scatter plot (so in case we don’t even need to look at a residual plot). Let’s start by looking at three different examples.

## Skin cancer and mortality

Does the data suggest that a linear function is adequate in describing the relationship between skin cancer mortality and latitude ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt))?

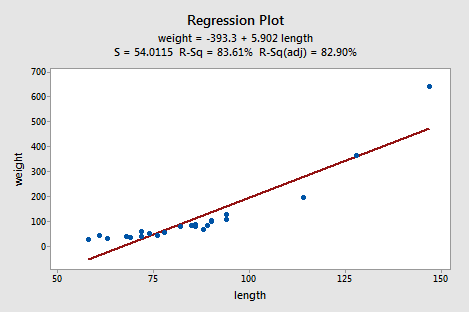


The answer is yes! It appears as if the relationship between latitude and skin cancer mortality is indeed linear, and therefore it would be best if we summarized the trend in the data using a linear function.

## Alligators

The length of an alligator can be estimated fairly accurately from aerial photographs or from a boat. Estimating the weight of the alligator, however, is a much greater challenge. One approach is to use a regression model that summarizes the trend between the length and weight of alligators. The length of an alligator obtained from an aerial photograph or boat can be used to predict the weight of the alligator. In taking this approach, some wildlife biologists captured a random sample of alligators. They measured the length (, in inches) and weight (, in pounds) of each alligator ([alligator.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/alligator.txt)).

Do the resulting data suggest that a linear function is adequate in describing the relationship between the length and weight of an alligator?

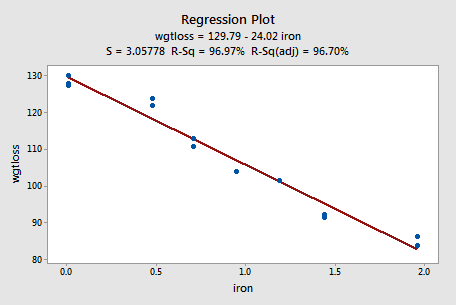


The answer is no! Don’t you think a curved function would more adequately describe the trend? The scatter plot gives us a pretty good indication that a linear model is inadequate in this case.

Alloy Corrosion

Thirteen alloy specimens comprised of 90% copper and 10% nickle – each with a specific iron content – were tested for corrosion. Each specimen was rotated in salty seawater at 30 feet per second for 60 days. The corrosion was measured in weight loss in milligrams/square decimeter/day. The researchers were interested in studying the relationship between iron content () and weight loss due to corrosion () ([corrosion.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/corrosion.txt)).

Does the resulting data, that appears in the following plot suggest that a linear function is adequate in describing the relationship between iron content and weight loss due to corrosion?

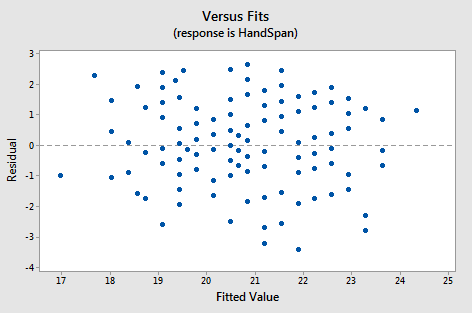


The answer is yes! As in the first example, our visual inspection if the data suggests that a linear model would be adequate in describing the trend between iron content and weight loss due to corrosion.

# Further Examples

## Example 1: A Good Residual Plot

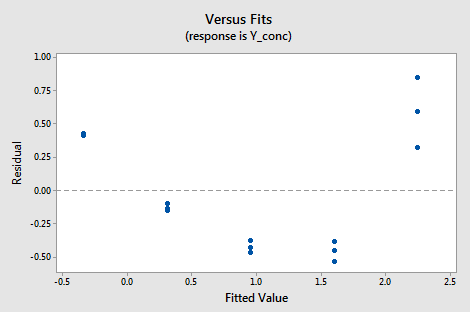
Below is a plot of residuals vs. fits after a straight-line model was used on data for = handspan (cm) and = height (inches), for students (handheight.txt).



*Interpretation*: This plot looks good in that the variance is roughly the same all the way across and there are no worrisome patterns. There seems to be no difficulties with the model or data.

## Example 2: Residual Plot Resulting from Using the Wrong Model

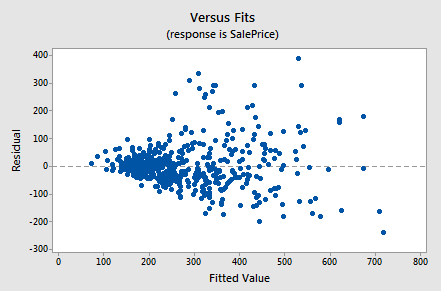
Below is a plot of residuals versus fits after a straight-line model was used on dara for = concentration of a chemical solution and = time after the solution was made ([solutions\_conc.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/solutions_conc.txt)).



*Interpretation*: This plot of residuals versus plots shows two difficulties. First, the pattern is curved which indicates that the wrong type of equation was used. Second, the variance (vertical spread) increases as the fitted values (predicted values) increase.

## Example 3: Indications that Assumption of Constant Variance is Not Valid

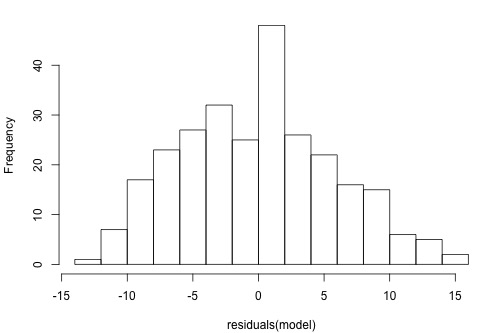
Below is plot of residuals versus fits after a straight-line model was used on data for = sale price of a home and = square foot area of home (realestate.txt).

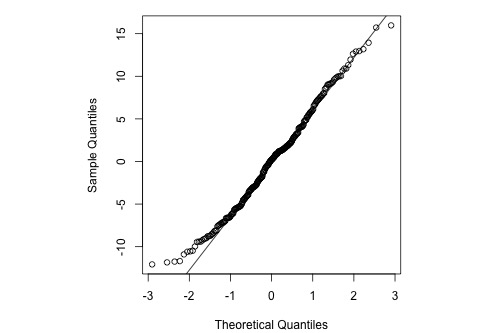


Interpretation: This plot of residuals versus fits shows that the residual variance (vertical spread) increases as the fitted values (predicted values of sale price) increases. This violates the assumption of constant error variance.

## Example 4: Indications that Assumption of Normal Distribution for Errors is Valid

The graphs below are a histogram and a normal probability plot of the resifuals after a straight-line model was used for fitting = time to next eruption and = duration of last eru[tion for eruptions of the Old Faithful geyser ([oldfaithful.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/oldfaithful.txt)).

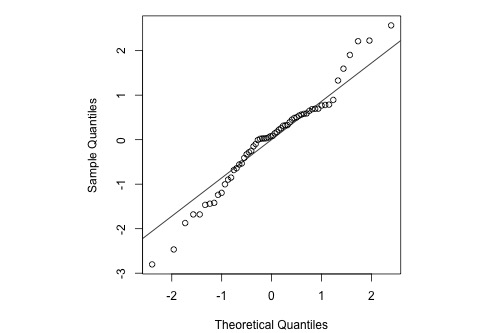




Interpretation: The histogram is roughly bell-shaped so it is an indication that it is reasonable to assume that the errors have a normal distribution. The pattern of the normal probability plot is straight, so this plot also provides evidence that it is reasonable to assume that the errors have a normal distribution.

## Example 5: Indications that Assumptions of Normal Distributions for Errors is not Valid

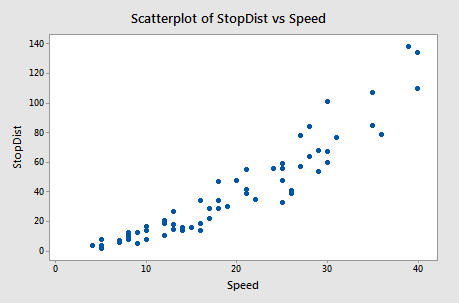
Below is a normal probability for the residuals from a straight-line regression with = infection risk in a hospital and = average length of stay in the hospital. The observational units are hospitals and the data are taken from regions 1 and 2 in the infection risk data set ([infectionrisk.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/infectionrisk.txt)).



*Interpretation*: The plot shows some deviation from the straight-line pattern indicating a distribution with heavier tails than a normal distribution.

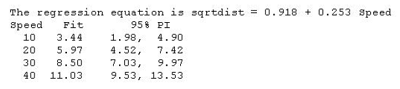
## Example 6: Stopping distance Data

We investigate how transforming can sometimes help us with non-constant variance problems. We will look at the stopping distance data with = stopping distance of a car and = speed of the car when the brakes were applied (carstopping.txt). A graph of the data is given below:

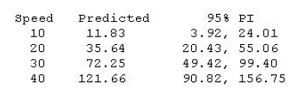


Fitting a simple linear regression model to these data leads to problems with both curvature and non-constant variance. One possible remedy is to transform . With some trial and error, we find that there is an approximate linear relationship between and with no suggestion of non-constant variance.

The Minitab output below gives the regression equation for square root distance on speed along with predicted values and prediction intervals for speeds of 10, 20, 30, and 40 mph. The predictions are for the square root of stopping distance.



Then, the output below shows predicted values and prediction intervals when we square the results (i.e., transform back to the scale of the original data).



Notice that the predicted values coincide more or less with the average pattern in the scatterplot of speed and stopping distance above. Also, note that the prediction intervals for stopping distance are becoming increasingly wide as speed increases. This reflects the non-constant variance in the original data.

We cover transformations like this in more detail in Chapter 9.

Quick reference: , , , , ,