STAT 462

Applied Regression Analysis

Lesson 4: SLR Assumptions, Estimation & Prediction

Overview of this Lesson

A typical regression analysis involves the following steps:

1. Model formulation
2. Model estimation
3. Model evaluation
4. Model use

So far, we have learned how to formulate and estimate a simple linear regression model. We have also learned about some methods for evaluating the model. The first part of this lesson continues the topic of *evaluating the model*.

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 dataset 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 graphical residual analysis. When we revisit this topic in the context of multiple linear regression in Lesson 6 we'll also study some statistical tests for assessing the assumptions. We'll consider various remedies for when linear regression model assumptions fail throughout the rest of the course, but particularly in Lesson 7.

In the second part of this lesson, we focus our efforts on *using the model* to answer two specific research questions, namely:

- What is the average response for a given value of the predictor x?
- What is the value of the response likely to be for a given value of the predictor x?

In particular, we will learn how to calculate and interpret:

- A confidence interval for estimating the mean response for a given value of the predictor x.
- A prediction interval for predicting a new response for a given value of the predictor x.

Key Learning Goals for this Lesson:
▪ 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.
- Distinguish between estimating a mean response (confidence interval) and predicting a new observation (prediction interval).
- Understand the various factors that affect the width of a confidence interval for a mean response.
- Understand why a prediction interval for a new response is wider than the corresponding confidence interval for a mean response.
- Know the formula for a prediction interval depends strongly on the condition that the error terms are normally distributed, while the formula for the confidence interval is not so dependent on this condition for large samples.
- Know the types of research questions that can be answered using the materials and methods of this lesson.

4.1 - Residuals › (/stat462/node/116)

STAT 462

Applied Regression Analysis

4.1 - Residuals

In the first part of this lesson, we 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 , $\mathrm{E}(Y_i)$, at each value of the predictor, x_i, is a **Linear function** of the x_i.
- The errors, ϵ_i, are **Independent**.
- The errors, ϵ_i, at each value of the predictor, x_i, are **Normally distributed**.
- The errors, ϵ_i, at each value of the predictor, x_i, have **Equal variances** (denoted σ^2).

An equivalent way to think of the first (linearity) condition is that the mean of the error, $\mathrm{E}(\epsilon_i)$, at each value of the predictor, x_i, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2.

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

- The population regression function is **not linear**. That is, the response Y_i is not a function of linear trend ($\beta_0 + \beta_1 x_i$) plus some error ϵ_i .
- 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 β_0 and β_1 are fairly "**robust**" (that is, forgiving) against departures from normality.
- Prediction intervals are quite sensitive to departures from normality.

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 *x* and *y*, *i.e.* test that the slope β_1 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 y_{new} , 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! And when you are worried, there are remedies available, which we'll learn more about later in the course. For example, one thing to try is transforming either the response variable, predictor variable, or both - there is an example of this in Section 4.8 and we'll see more examples in Lesson 7.

This is definitely a topic in which you are exposed to the idea that data analysis is an art (subjective decisions!) based on science (objective tools!). We might therefore call data analysis "an artful science!" Let's get to it!

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 data point y_i and its estimated value $\hat{y}_i$ is its observed "**residual**":

$$e_i = y_i - \hat{y}_i$$

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

$$\epsilon_i = Y_i - E(Y_i)$$

Let's look at an illustration of the distinction between a residual e_i and an unknown true error term ϵ_i. 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 $n = 15$ students. The **residual error** is the distance from the data point to to the dashed line. **Click on the "Zoom in!" icon** to see the two types of errors — the true error and 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.

‹ Lesson 4: SLR Assumptions, Estimation & Prediction (/stat462/node/81) up (/stat462/node/81) 4.2 - Residuals vs. Fits Plot › (/stat462/node/117)

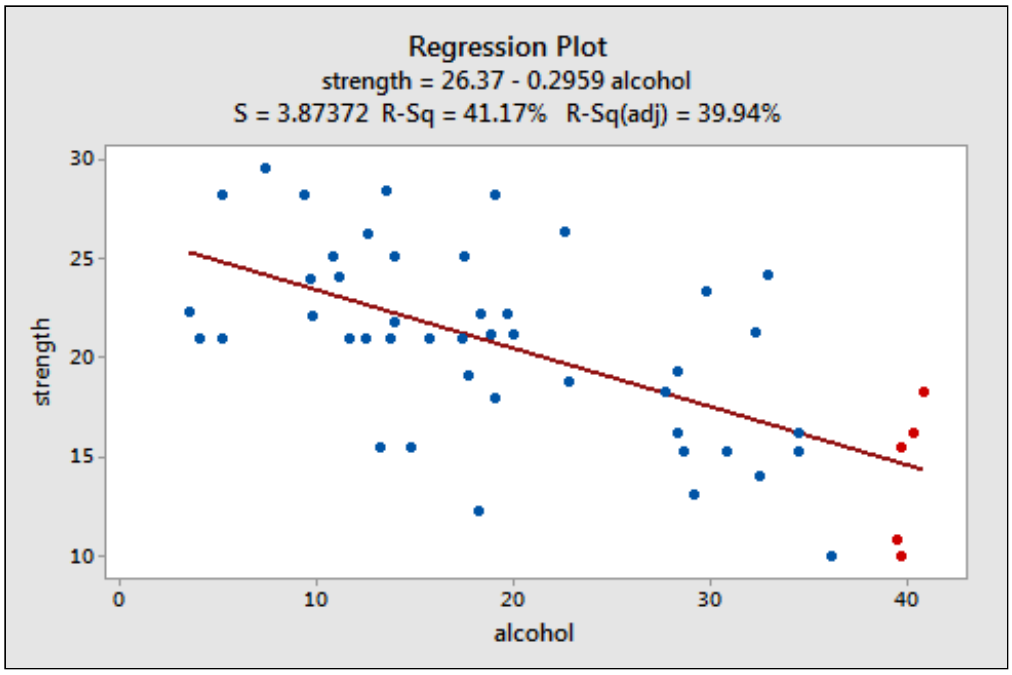
STAT 462

Applied Regression Analysis

4.2 - Residuals vs. Fits Plot

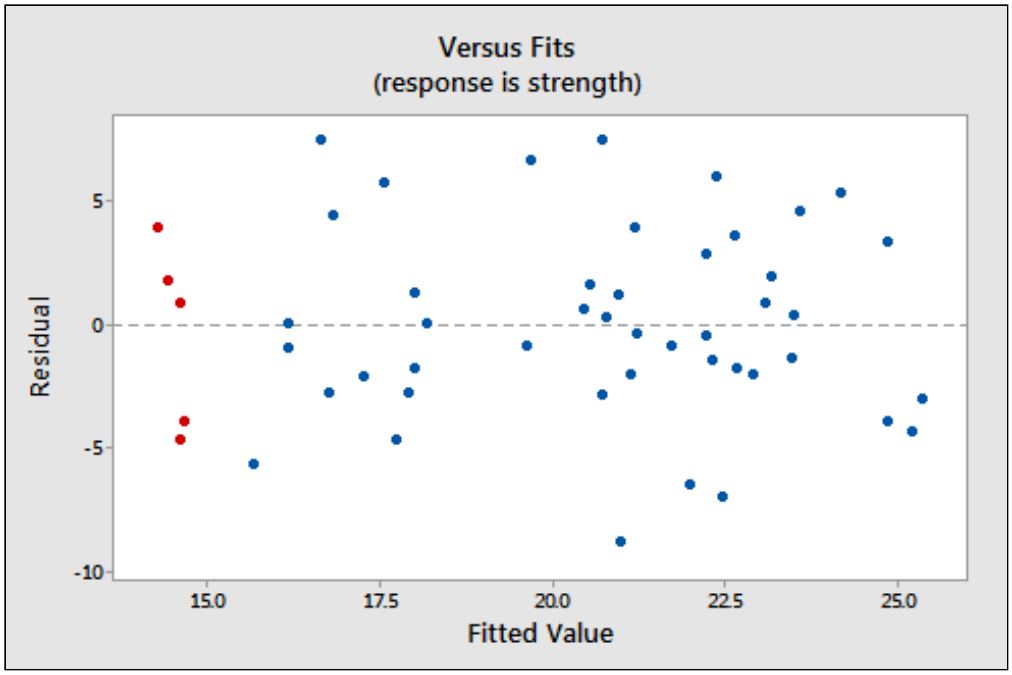
When conducting a residual analysis, a "**residuals versus fits plot**" is the most frequently created plot. It is a scatter plot of residuals on the *y* axis and fitted values (estimated responses) on the *x* axis. The plot is used to detect non-linearity, unequal error variances, and outliers.

Let's look at an example to see what a "well-behaved" residual plot looks like. Some researchers (Urbano-Marquez, *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 (x) on a random sample of $n = 50$ alcoholic men. They also measured the strength (y) of the deltoid muscle in each person's nondominant arm. A fitted line plot of the resulting data, (alcoholarm.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/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 error 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 *y* axis and the fitted values appear on the *x* 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 versus 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 classical example of a well-behaved residuals vs. fits plot. Here are the characteristics of a well-behaved residual 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 students 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 interpretation of a residuals vs. fits plot worthwhile. Don't worry! You will learn — with practice — how to "read" these plots.

‹ 4.1 - Residuals (/stat462/node/116) up (/stat462/node/81) 4.3 - Residuals vs. Predictor Plot › (/stat462/node/118)

STAT 462

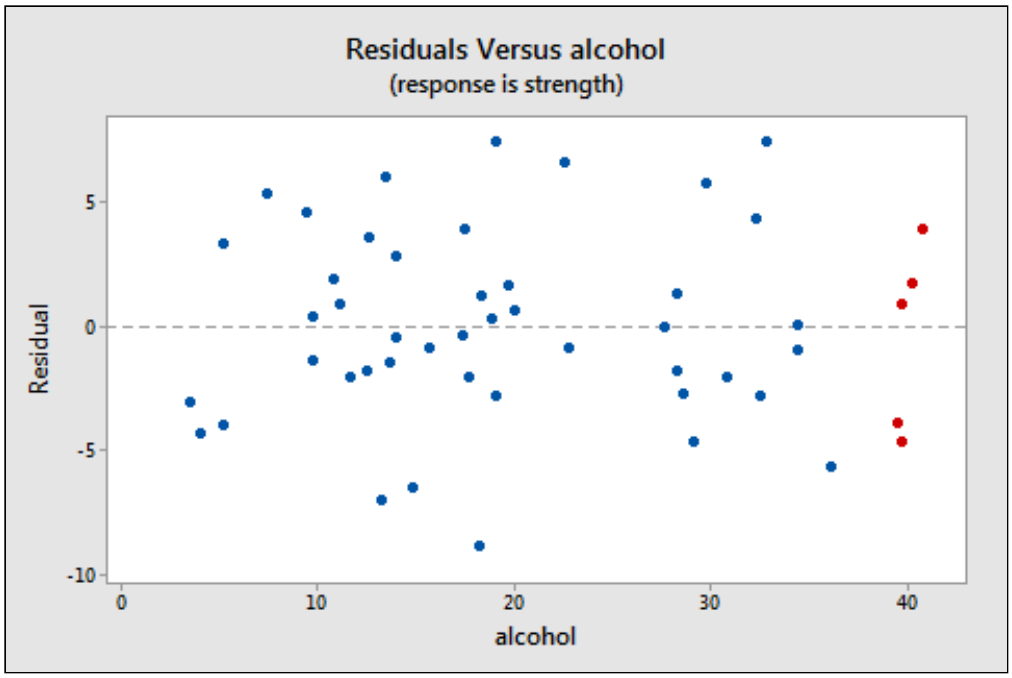
Applied Regression Analysis

4.3 - Residuals vs. Predictor Plot

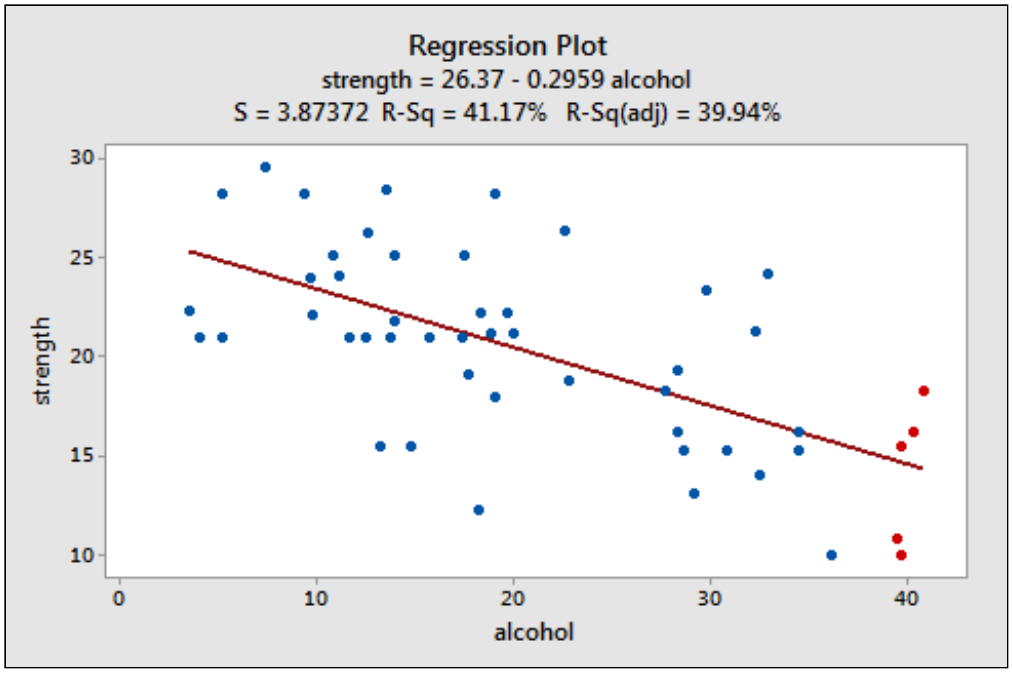
An alternative to the residuals vs. fits plot is a "**residuals vs. predictor plot**." It is a scatter plot of residuals on the y axis and the predictor (x) values on the x axis. For a simple linear regression model, if the predictor on the x axis is the same predictor that is used in the regression 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 x axis is a new and different predictor, the residuals vs. predictor plot can help to determine whether the predictor should be added to the model (and hence a multiple regression model used instead).

The interpretation of a "residuals 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.

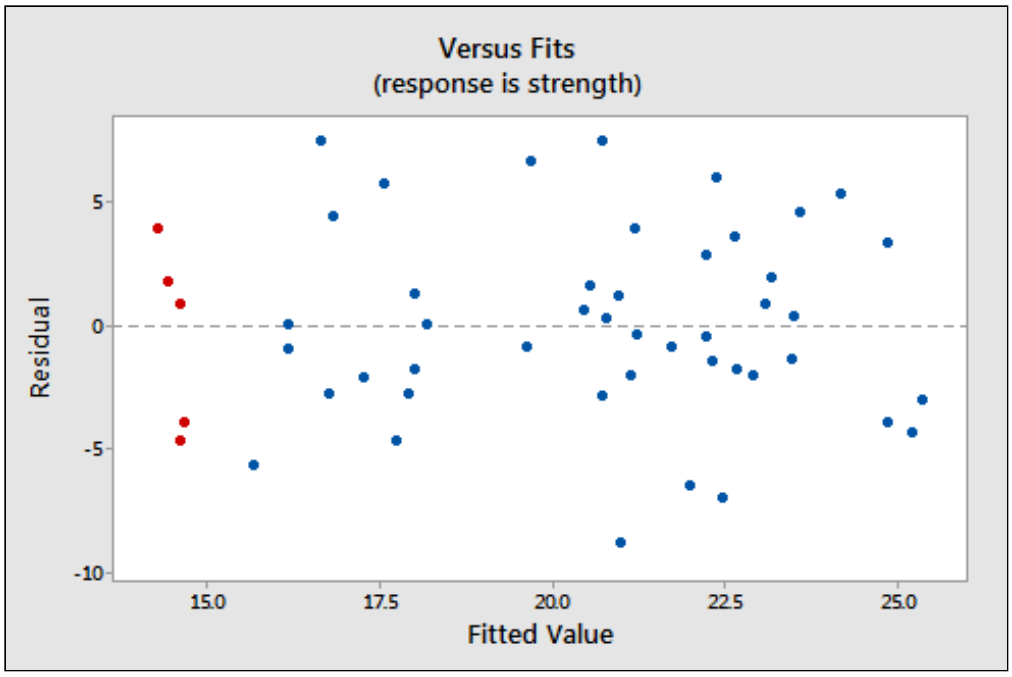
Here's the residuals vs. predictor plot for the simple linear regression model with arm strength as the response and level of alcohol consumption as the predictor:



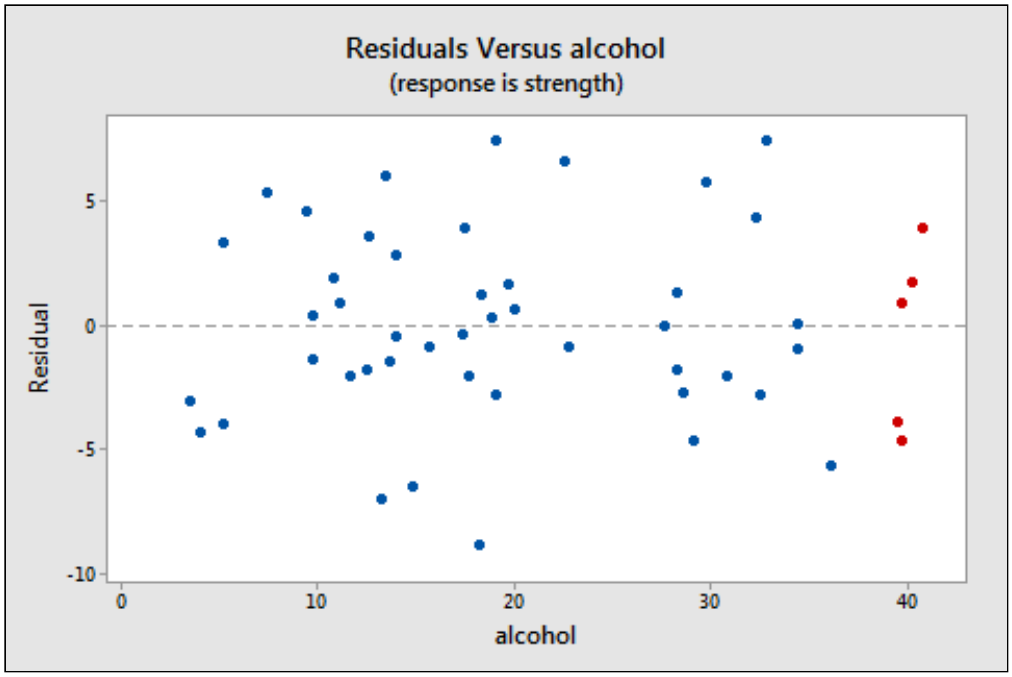
Note that, as defined, the residuals appear on the y axis and the predictor values — the lifetime alcohol consumptions for the men — appear on the x axis. Now, you should be able to look back at the scatter plot of the data:



and the residuals vs. fits plot:



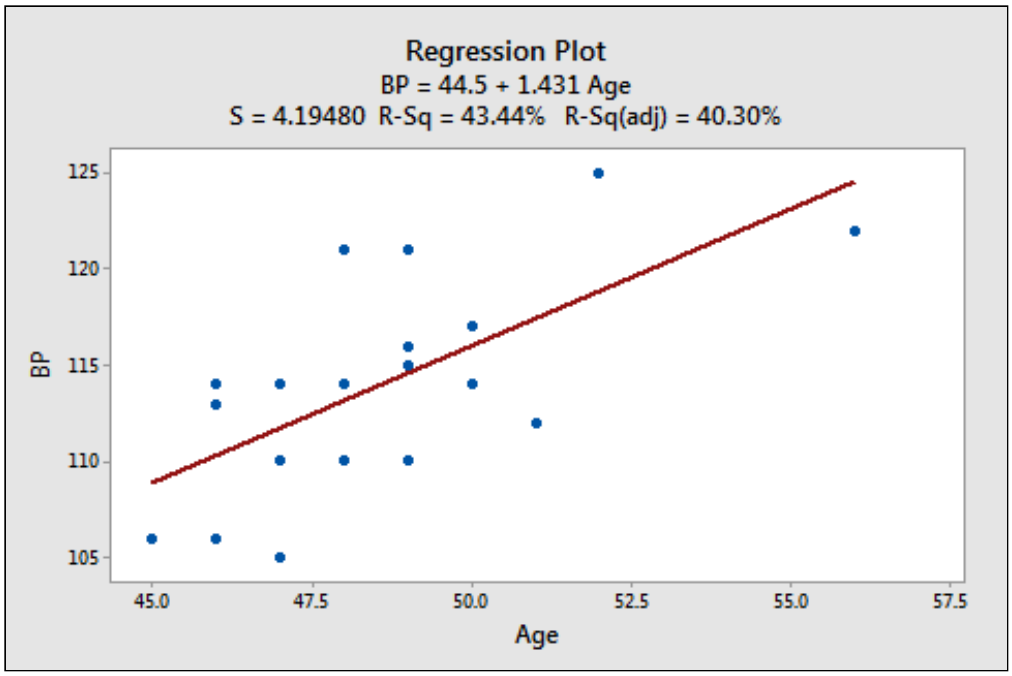
to see how the data points there correspond to the data points in the residuals versus predictor plot:



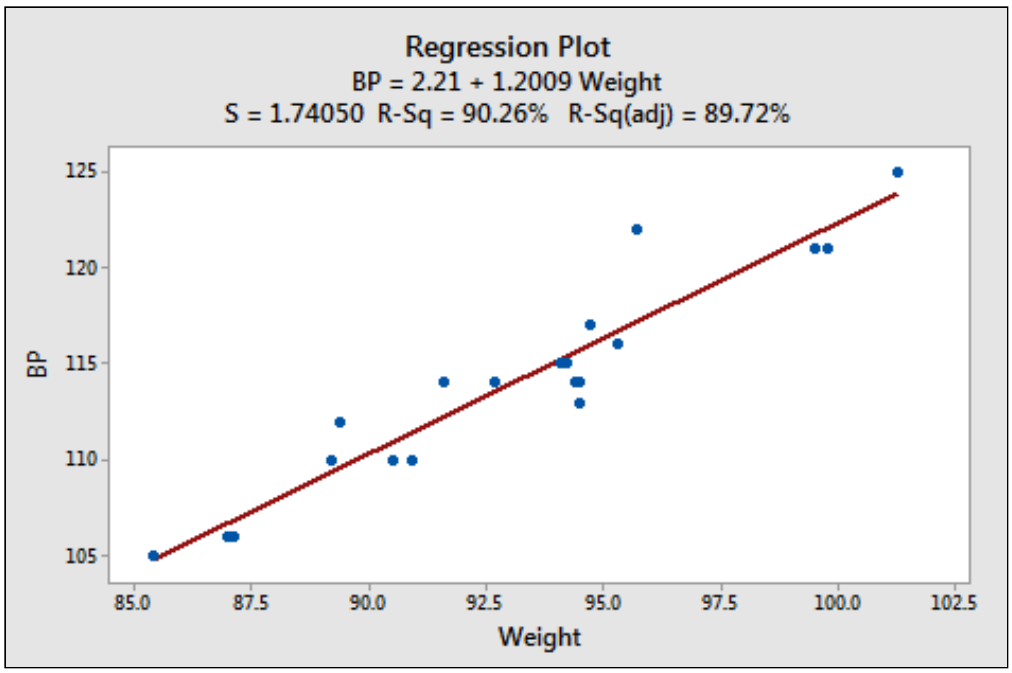
The five red data points should help you out again. The alcohol consumption of the five men is about 40, and hence why the points now appear on the "right side" of the plot. In essence, for this example, the residuals vs. predictor plot is just a mirror image of the residuals vs. fits plot. The residuals 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 of 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 diastolic blood pressure (in mm Hg) on a sample of $n = 20$ hypertensive individuals (bloodpress.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/bloodpress.txt)).

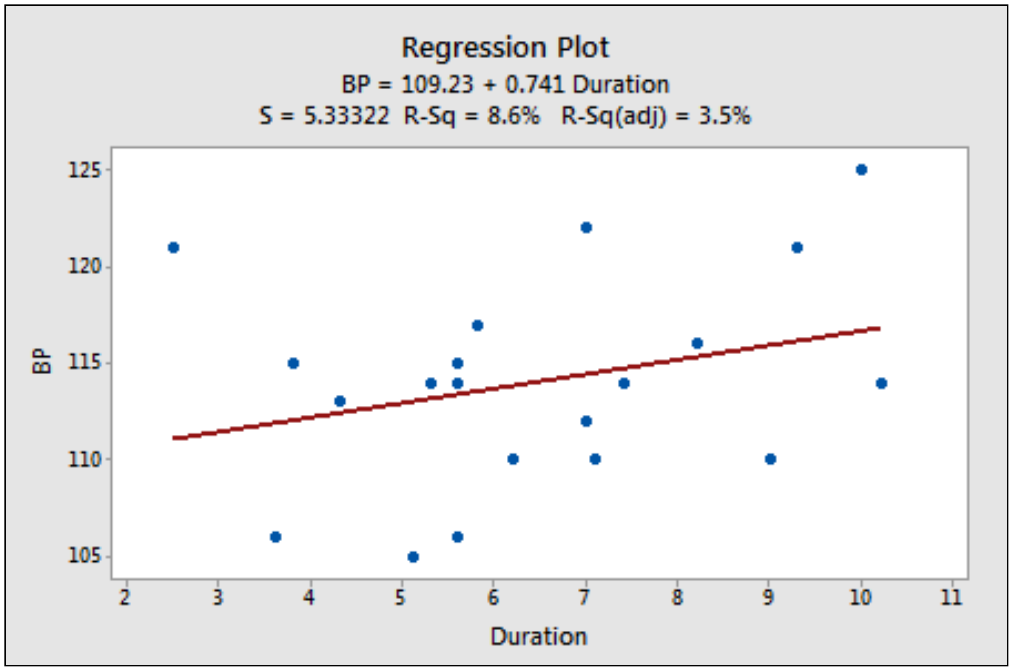
The regression of the response diastolic blood pressure (BP) on the predictor age:



suggests that there is a moderately strong linear relationship (r^2 = 43.4%) between diastolic blood pressure and age. The regression of the response diastolic blood pressure (BP) on the predictor weight:

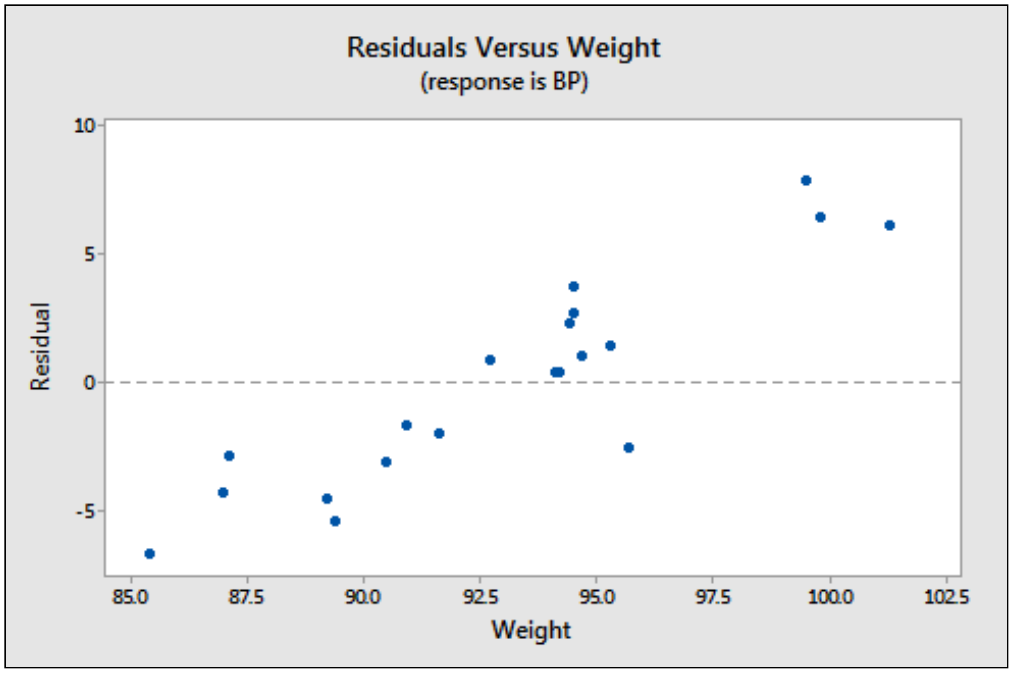


suggests that there is a strong linear relationship (r^2 = 90.3%) between diastolic blood pressure and weight. And, the regression of the response diastolic blood pressure (BP) on the predictor duration:



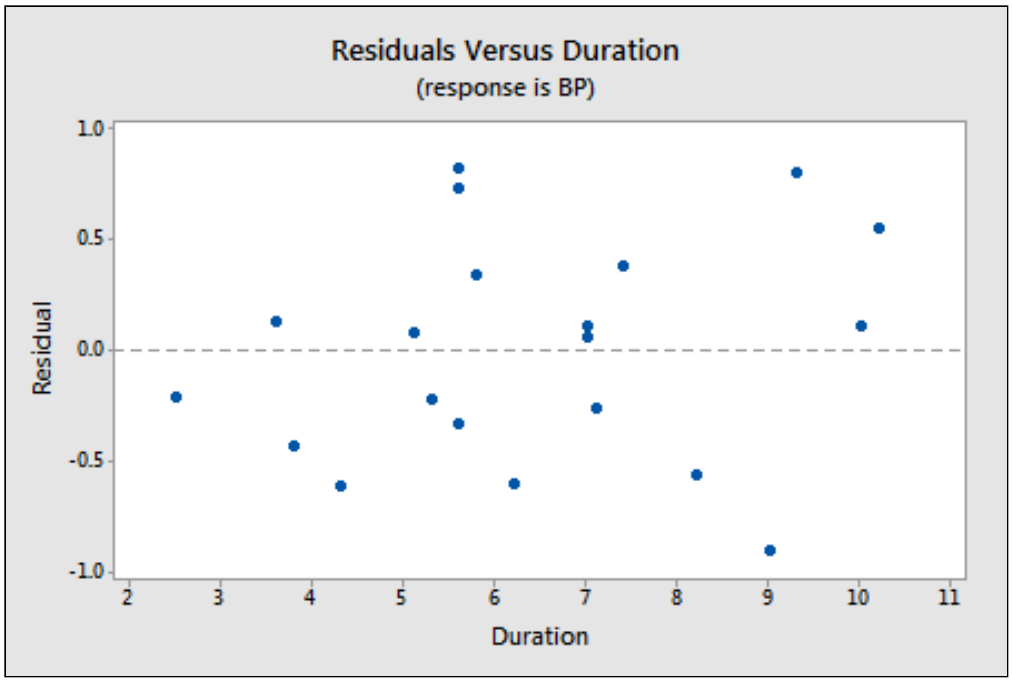
suggests that there is little linear association (r^2 = 8.6%) between diastolic blood pressure and duration of hypertension. In summary, it appears as if weight has the strongest association with diastolic blood pressure, age has the second strongest association, and duration the weakest.

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 versus weight" plot:



This "residuals versus 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 may be worthwhile adding the predictor to the model. In essence, you can think of the residuals on the *y* axis as a "new response," namely the individual's diastolic blood pressure adjusted for their age. If a plot of the "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 simple linear regression. Suppose we fit 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 versus duration" plot:



The points on the plot show no 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 individuals' blood pressures by taking into account the individuals' ages and weights, none of the remaining variability can be explained by the individuals' durations.

‹ 4.2 - Residuals vs. Fits Plot (/stat462/node/117) up (/stat462/node/81) 4.4 - Identifying Specific Problems Using Residual Plots › (/stat462/node/120)

STAT 462

Applied Regression Analysis

4.4 - Identifying Specific Problems Using Residual Plots

In this section, we learn how to use residuals versus 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 (providing the predictor is the one in the model).

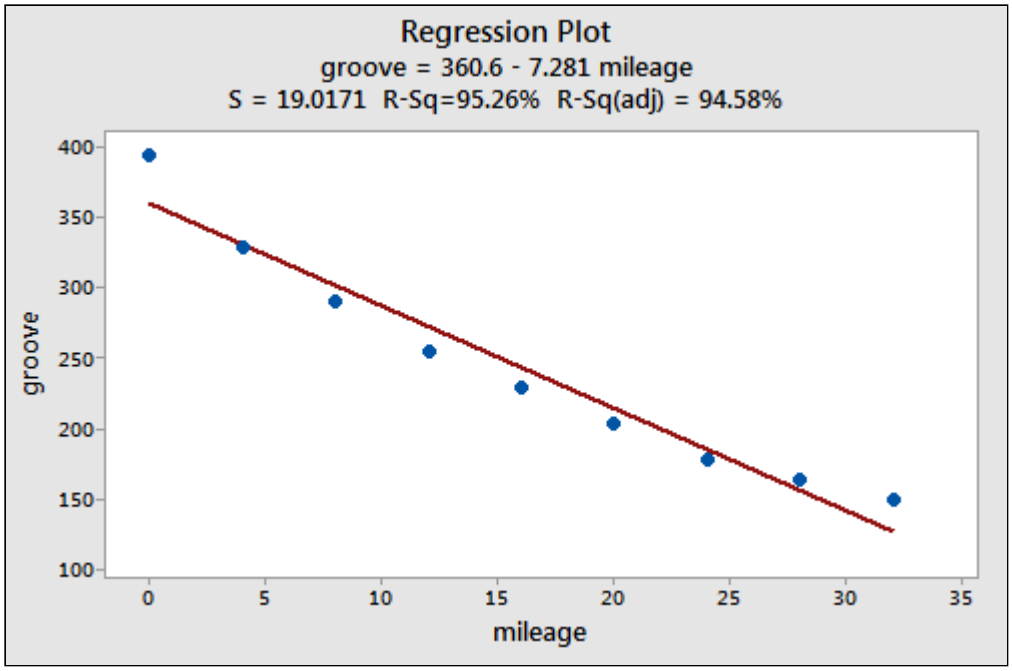
How does a non-linear regression function show up on a residual vs. fits plot?

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

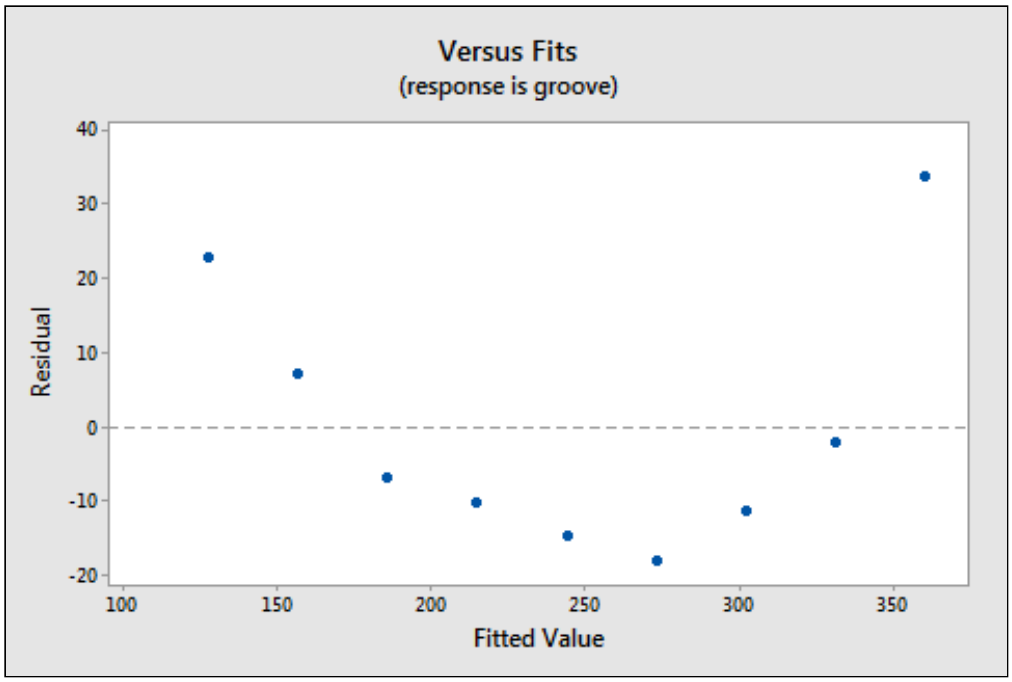
An Example: Is tire tread wear linearly related to mileage? A laboratory (*Smith Scientific Services*, Akron, OH) conducted an experiment in order to answer this research question. As a result of the experiment, the researchers obtained a data set (treadwear.txt

(https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/treadwear.txt)) containing the mileage (x, in 1000 miles) driven and the depth of the remaining groove (y, in mils). The fitted line plot of the resulting data:

Loading [MathJax]/extensions/MathZoom.js



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 x values, negative for medium x values, and positive again for large x values. Clearly, a non-linear model would better describe the relationship between the two variables.

Incidentally, did you notice that the r^2 value is very high (95.26%)? This is an excellent example of the caution "a large r^2 value should not be interpreted as meaning that the estimated regression line fits the data well." The large r^2 value tells you that if you wanted to predict groove depth, you'd be 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.

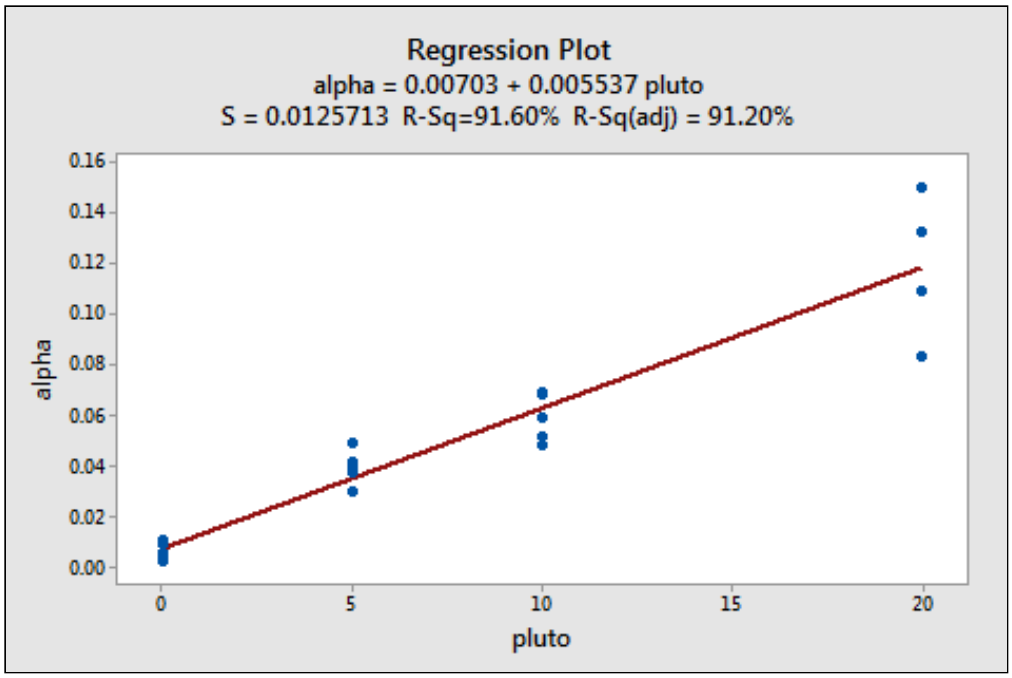
Loading [MathJax]/extensions/MathZoom.js

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

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

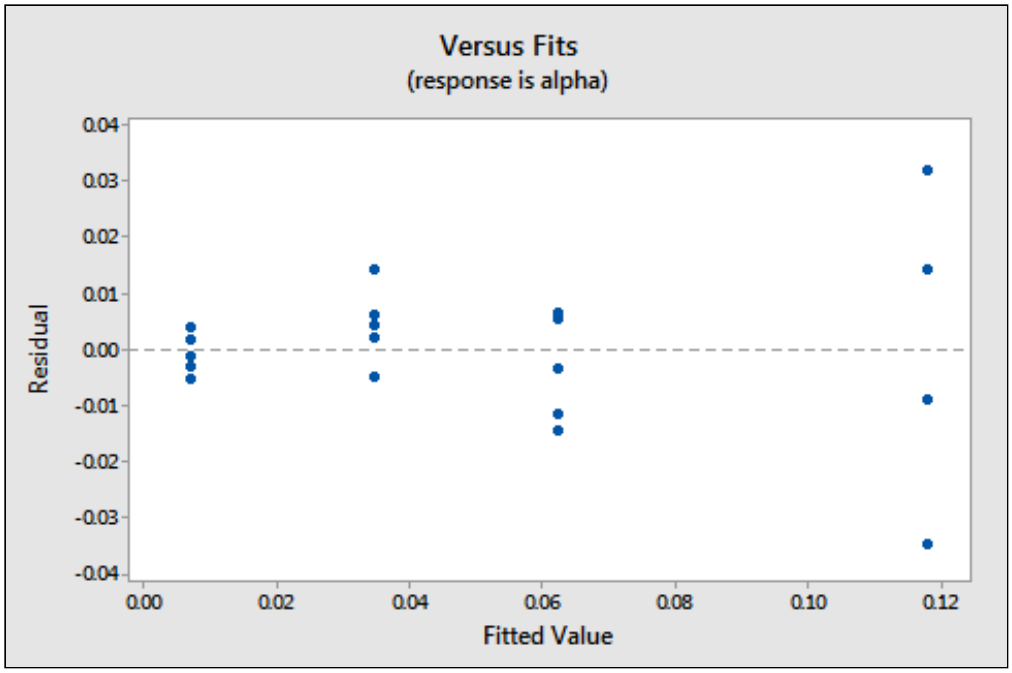
- The plot has a "**fanning**" effect. That is, the residuals are close to 0 for small x values and are more spread out for large x values.
- The plot has a "**funneling**" effect. That is, the residuals are spread out for small x values and close to 0 for large x values.
- Or, the spread of the residuals 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 (x, in pCi/g) and alpha count rate (y, 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/stat462/sites/onlinecourses.science.psu.edu.stat462/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 variances is violated. As is generally the case, the corresponding residuals vs. fits plot accentuates this claim:

Loading [MathJax]/extensions/MathZoom.js



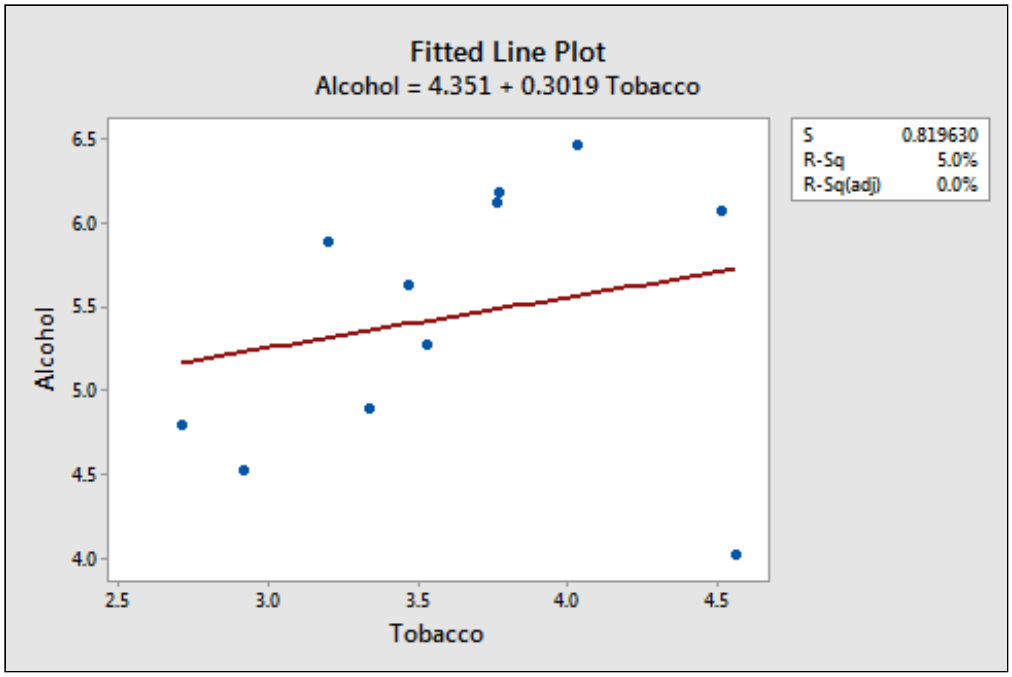
Note that the residuals "fan out" from left to right rather than exhibiting a consistent spread around the residual = 0 line. The residual vs. fits plot suggests that the error variances are not equal.

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

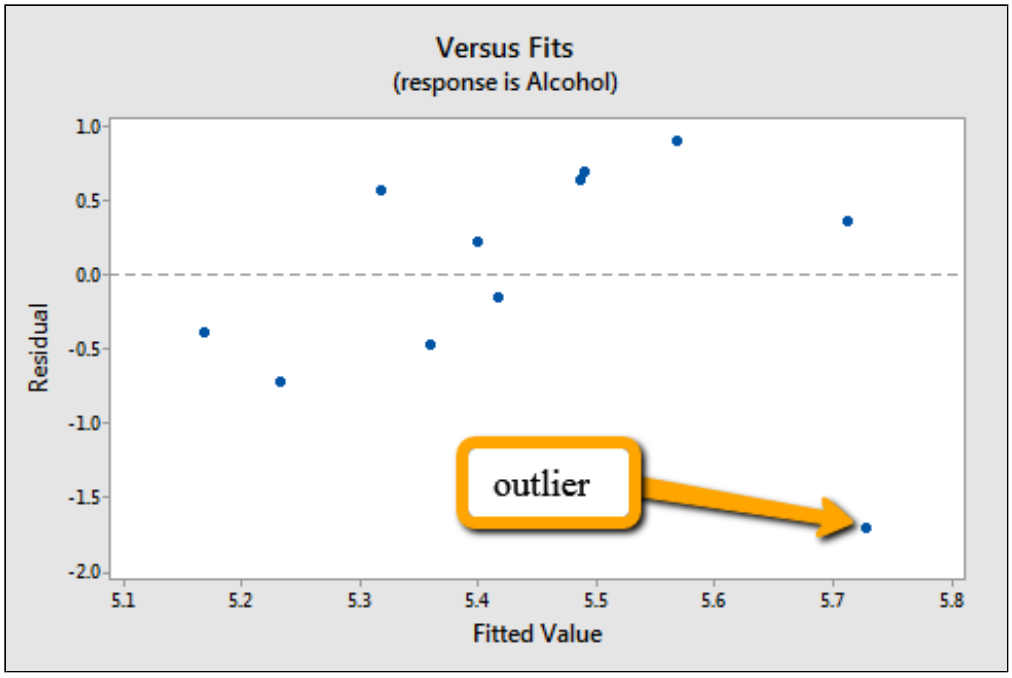
The Answer: The observation's residual stands apart from the basic random pattern of the rest of the residuals. The random pattern of the residual plot can even disappear if one outlier really deviates from the pattern of the rest of the data.

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 expenditure on tobacco (x, in British pounds) and the average weekly expenditure on alcohol (y, in British pounds) for households in $n = 11$ different regions in the United Kingdom. The fitted line plot of the resulting data (alcoholtobacco.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/alcoholtobacco.txt)):

Loading [MathJax]/extensions/MathZoom.js



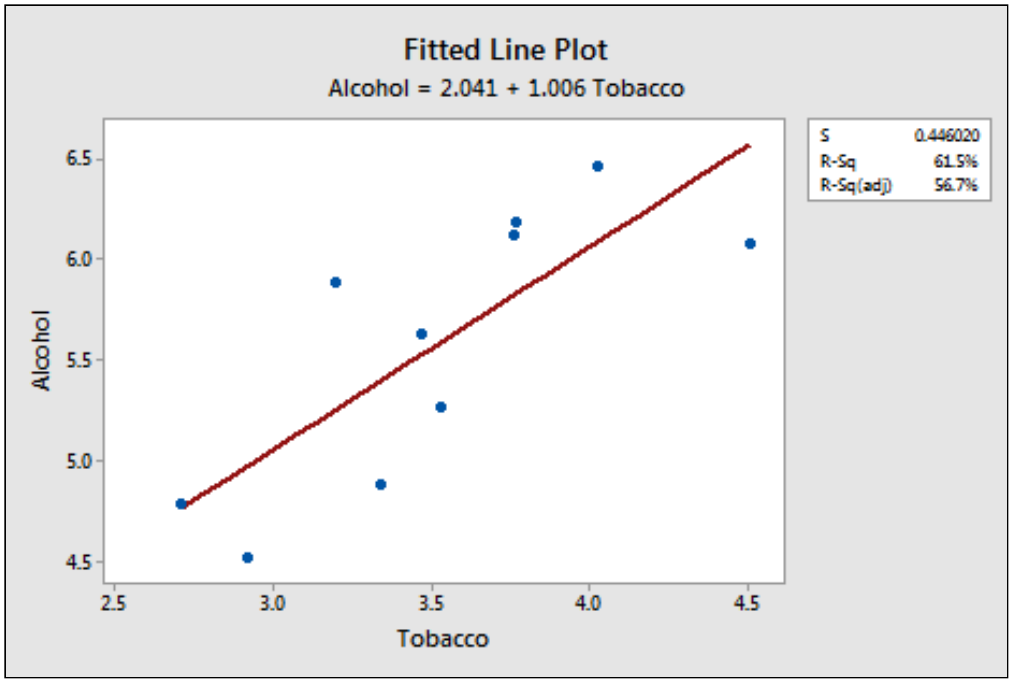
suggests that there is an outlier — in the lower right corner of the plot — which corresponds to the Northern Ireland region. In fact, the outlier is so far removed from the pattern of the rest of the data that it appears to be "pulling the line" in its direction. As is generally the case, the corresponding 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 r^2 can be greatly affected by just one data point." Note above that the r^2 value on the data set with all $n = 11$ regions included is 5%. Removing Northern Ireland's data point from the data set, and refitting the regression line, we obtain:

Loading [MathJax]/extensions/MathZoom.js



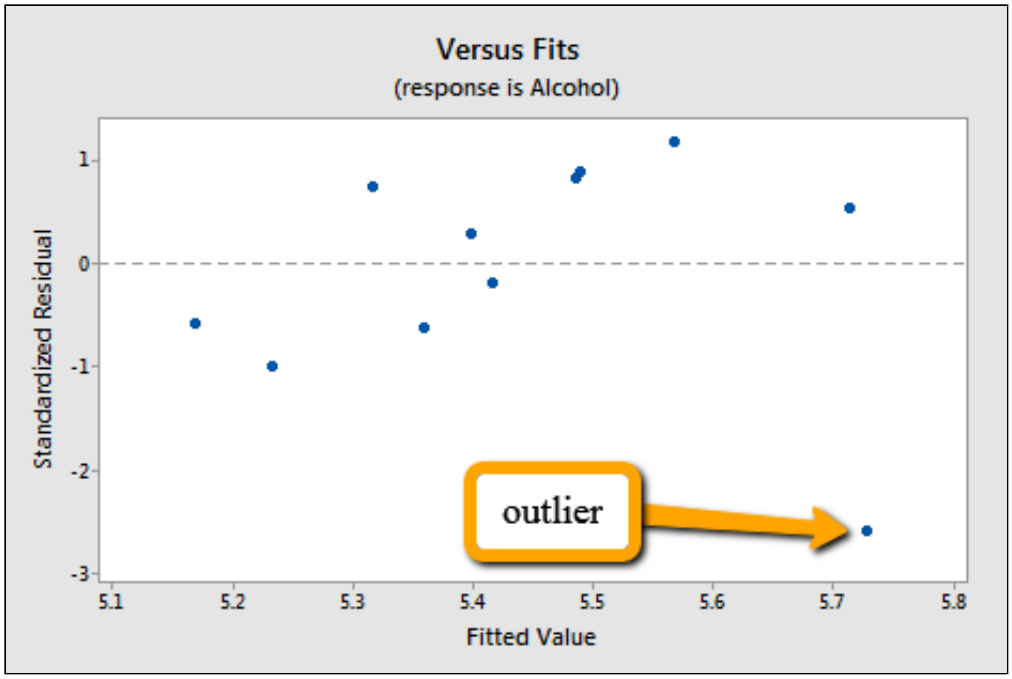
The r^2 value has jumped from 5% ("no-relationship") to 61.5% (" moderate relationship")! Can one data point greatly affect the value of r^2? Clearly, it can!

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 variable. 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 one "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 Lesson 9.) Recall that the empirical rule tells us that, for data that are normally distributed, 95% of the measurements fall within 2 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-dried, 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:

Loading [MathJax]/extensions/MathZoom.js



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 dataset the data point should be flagged for further investigation!

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

```
Fits and Diagnostics for Unusual Observations

Obs  Alcohol    Fit   Resid  Std Resid
 11    4.020  5.728  -1.708      -2.58  R

R  Large residual
```

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.

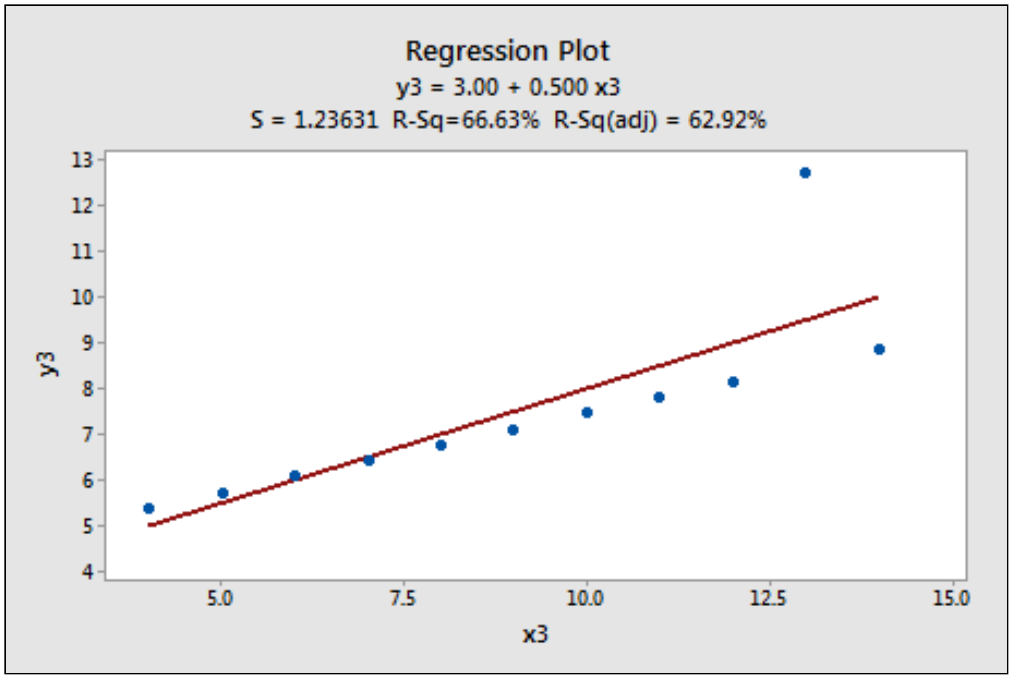
Note that I have intentionally used the phrase "flagged for futher investigation." I have not said that the data point should be "removed." Here's my recommended strategy, once you've identified a data point as being unusual:

1. Determine whether a simple — and therefore 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 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? For example, 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.

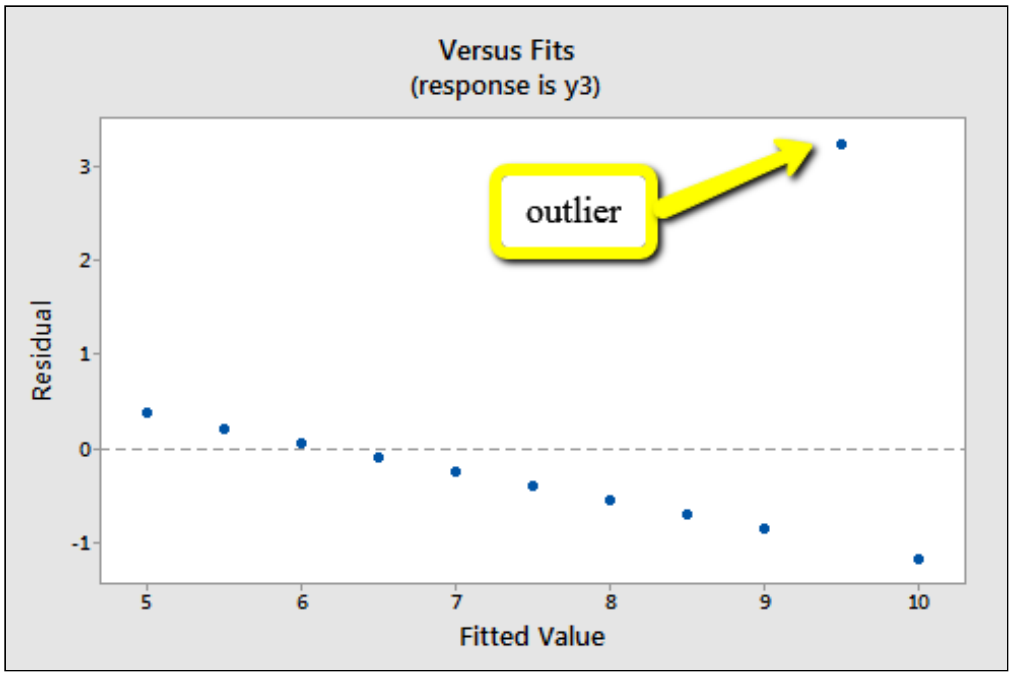
Loading [MathJax]/extensions/MathZoom.js

ombe data set #3 (anscombe.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/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.

‹ 4.3 - Residuals vs. Predictor Plot (/stat462/node/118) up (/stat462/node/81) 4.5 - Residuals vs. Order Plot › (/stat462/node/121)

Loading [MathJax]/extensions/MathZoom.js

STAT 462

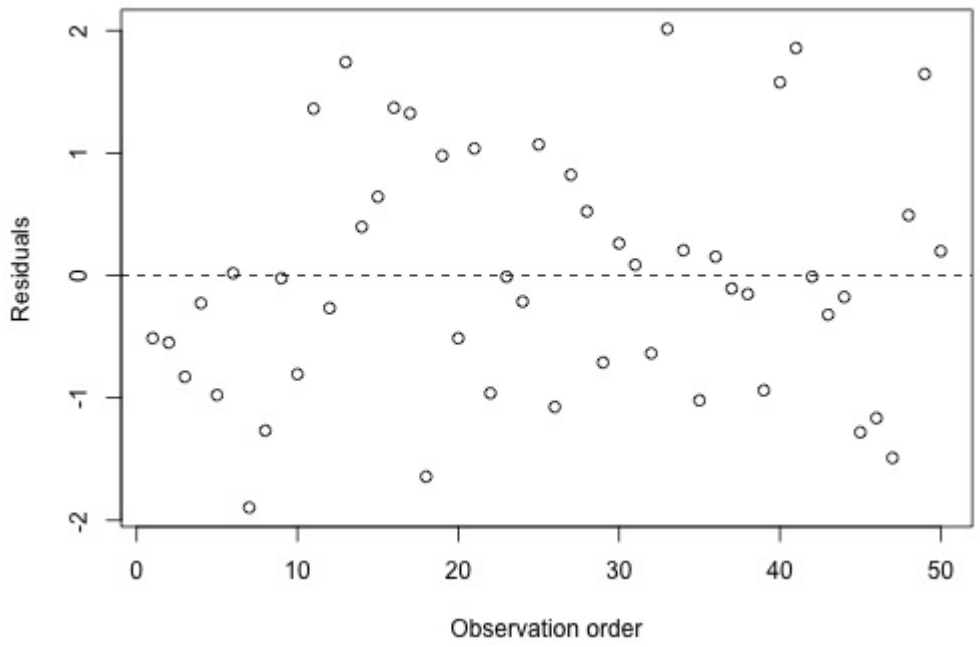
Applied Regression Analysis

4.5 - 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 residuals 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, ..., er, 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 residuals vs. order plot all about? As its name suggests, it is a scatter plot with residuals on the *y* axis and the order in which the data were collected on the *x* axis. Here's an example of a well-behaved residuals vs. order plot:



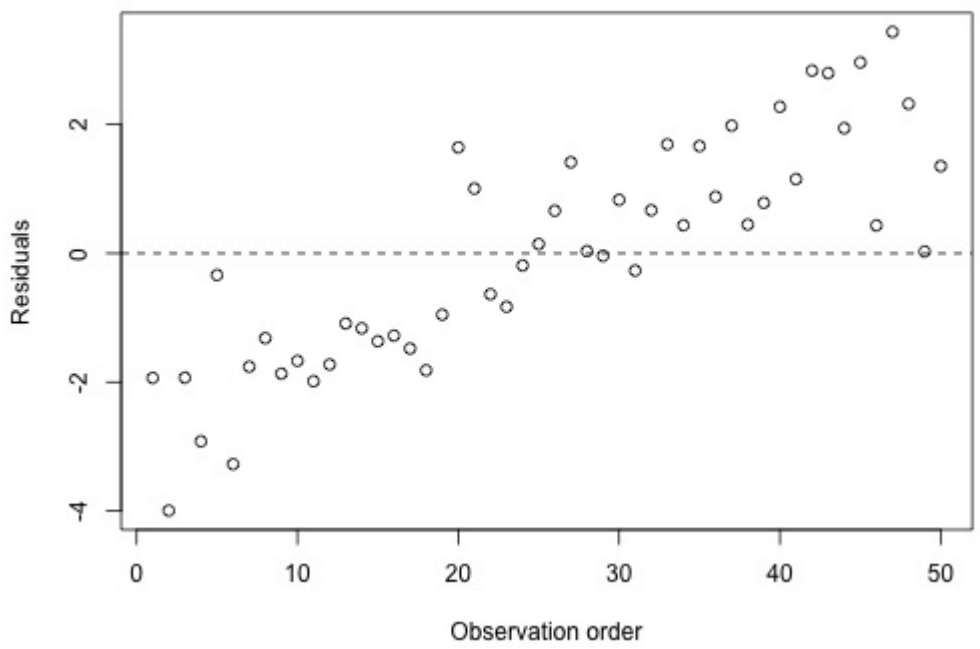
The residuals bounce randomly around the residual = 0 line as we would hope so. In general, residuals exhibiting normal random noise around the residual = 0 line suggest that there is no serial correlation.

Loading [MathJax]/extensions/MathZoom.js

f the different kinds of residuals vs. order plots we can obtain and learn what each tells us.

A time trend

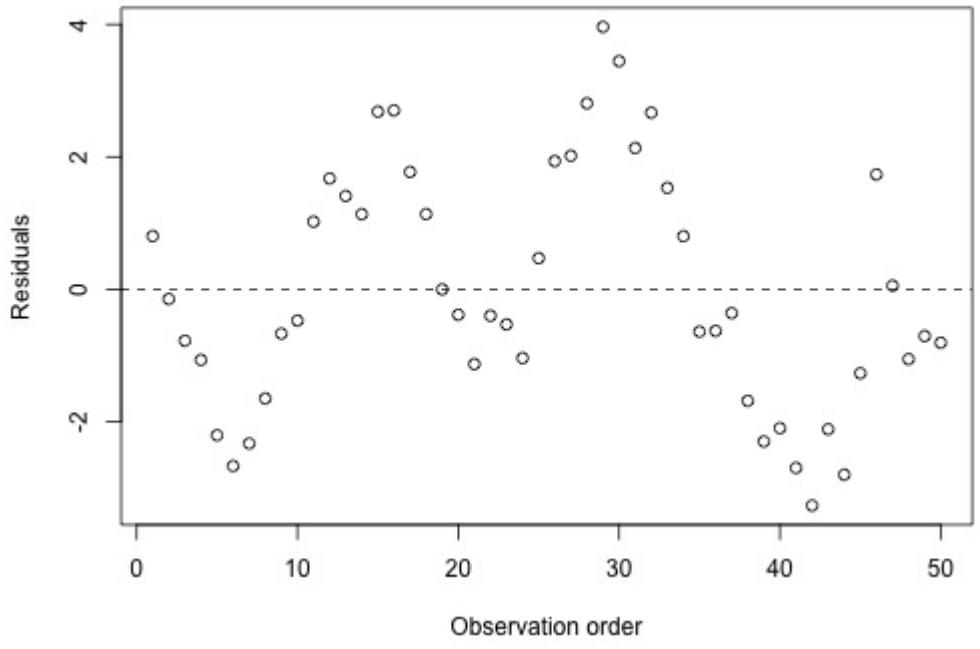
A residuals vs. order plot that exhibits (positive) trend as the following plot does:



suggests that some of the variation in the response is due to time. Therefore, it might be a good idea to add the predictor "time" to the model. That is, you interpret this plot just as you would interpret any other residual vs. predictor plot. It's just that here your predictor is "time."

Positive serial correlation

A residuals 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 wed, 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.

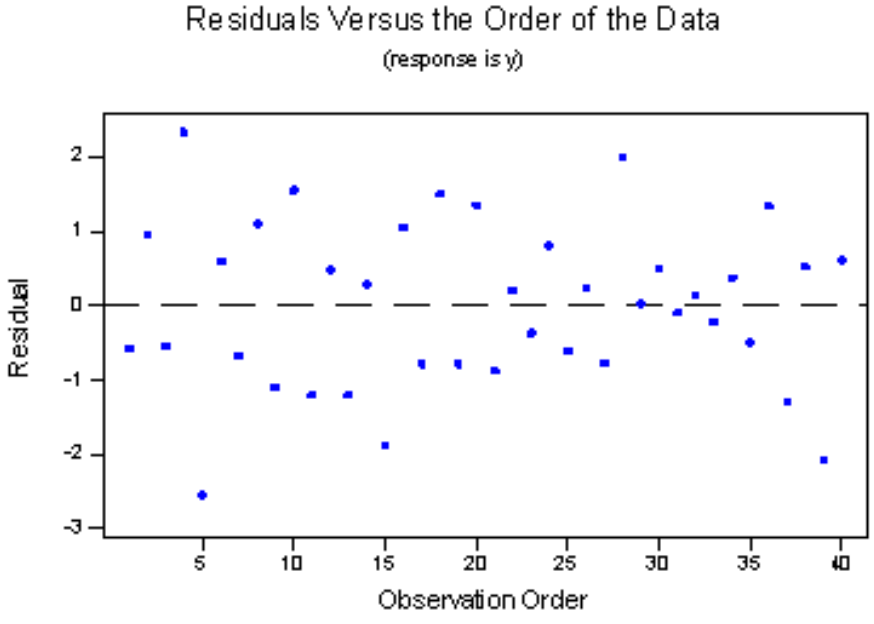
Loading [MathJax]/extensions/MathZoom.js

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

Can you see a cyclical trend -- up and then down, up and down, and up again? If not, **click on the "Draw trend!" icon**. 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 residuals 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

Loading [MathJax]/extensions/MathZoom.js

connect the dots in order from left to right, you should be able to see the pattern. If you can't see it, **click on the "Draw trend!" icon**:

Negative, positive, negative, positive, negative, positve, and so on. The plot suggests that the assumption of independent error terms is violated. If you obtain a residuals 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**."

‹ 4.4 - Identifying Specific Problems Using Residual Plots (/stat462/node/120) | up (/stat462/node/81) | 4.6 - Normal Probability Plot of Residuals › (/stat462/node/122)

Loading [MathJax]/extensions/MathZoom.js

STAT 462

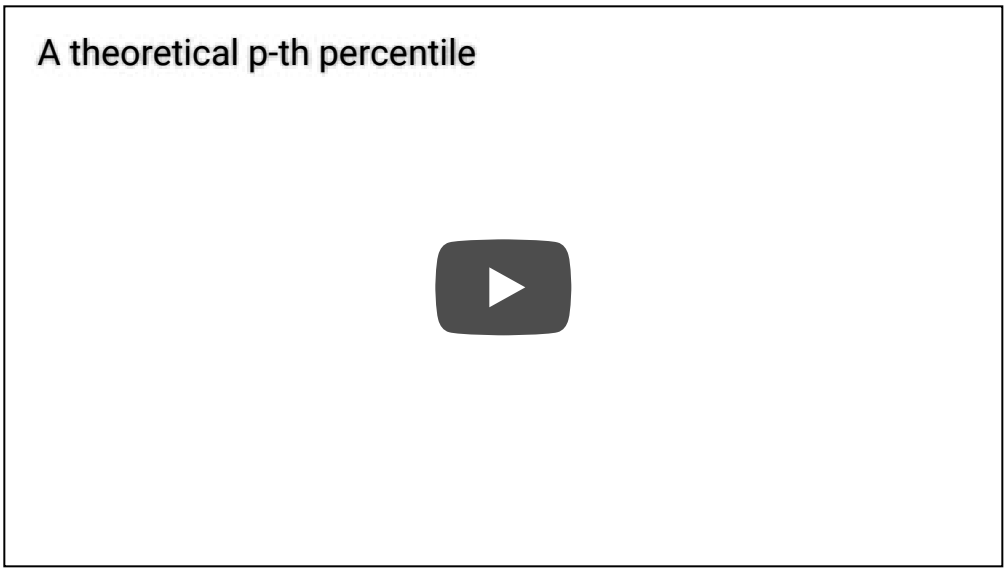
Applied Regression Analysis

4.6 - Normal Probability Plot of Residuals

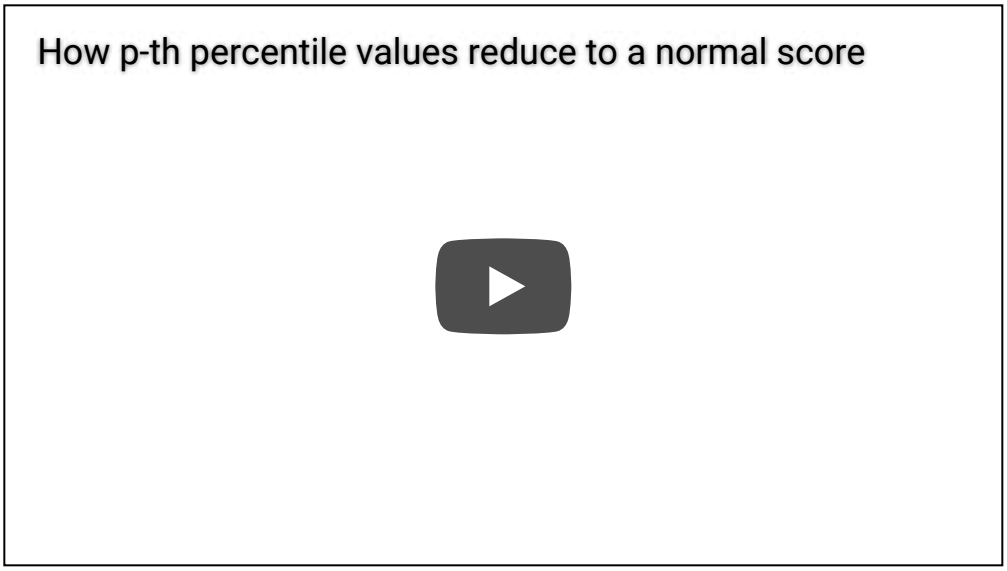
Recall that the third condition — the "N" condition — of the linear regression model is that the error terms are normally distributed. In this section, we learn how to use a "**normal probability plot of the residuals**" as a way of learning whether it is reasonable to assume that the error terms are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

Here's the basic idea behind any normal probability plot: if the error terms follow a normal distribution with mean μ and variance σ^2, then a plot of the **theoretical percentiles of the normal distribution** versus the observed **sample percentiles** of the residuals should be approximately linear. If a normal probability plot of the residuals is approximately linear, we proceed assuming that the error terms are normally distributed.

The **theoretical *p*-th percentile** of any normal distribution is the value such that *p*% of the measurements fall below the value. Here's a screencast illustrating a theoretical *p*-th percentile.

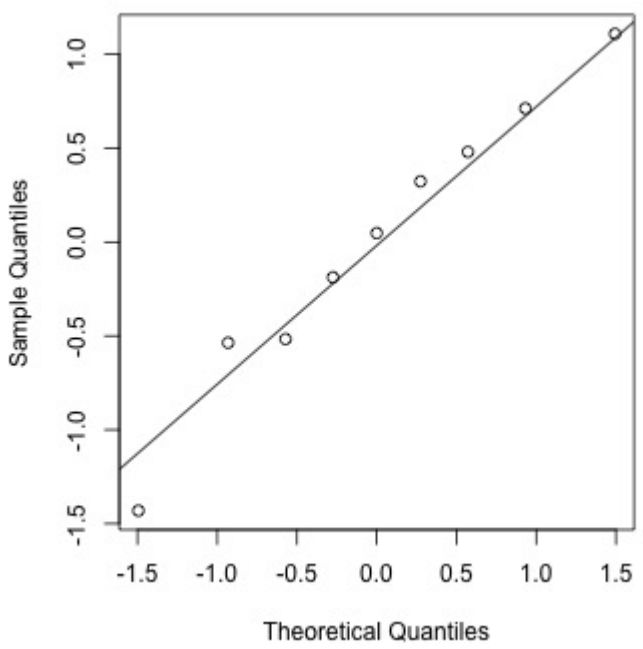


The problem is that to determine the percentile value of a normal distribution, you need to know the mean μ and the variance σ^2. And, of course, the parameters μ and σ^2 are typically unknown. Statistical theory says its okay just to assume that $\mu = 0$ and $\sigma^2 = 1$. Once you do that, determining the percentiles of the standard normal curve is straightforward. The *p*-th percentile value reduces to just a "*Z*-score" (or "normal score"). Here's a screencast illustrating how the *p*-th percentile value reduces to just a normal score.



The **sample *p*-th percentile** of any data set is, roughly speaking, the value such that p% 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 dataset fall below the value. And so on.

Consider a simple linear regression model fit to a simulated dataset 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 on the y axis and the sample percentiles of the residuals on the x axis, for example:



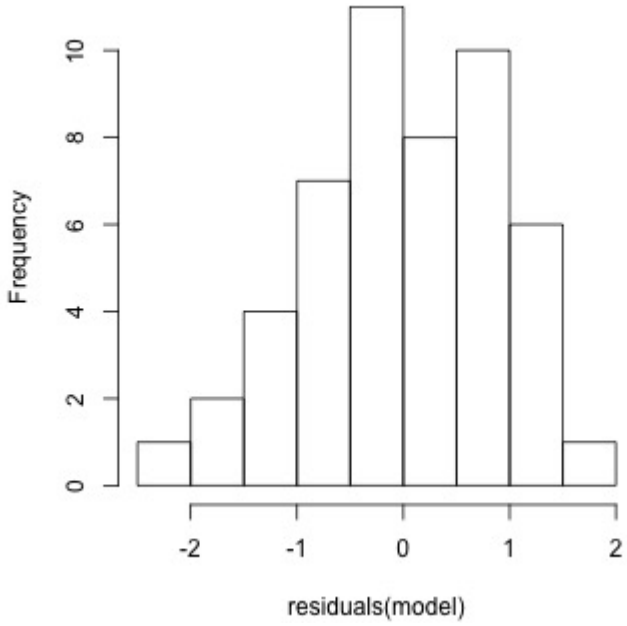
Note that the relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, the normal probability plot of the residuals suggests that the error terms are indeed normally distributed for this example.

Statistical software sometimes provides normality tests to complement the visual assessment available in a normal probability plot (we'll revisit normality tests in Lesson 6).

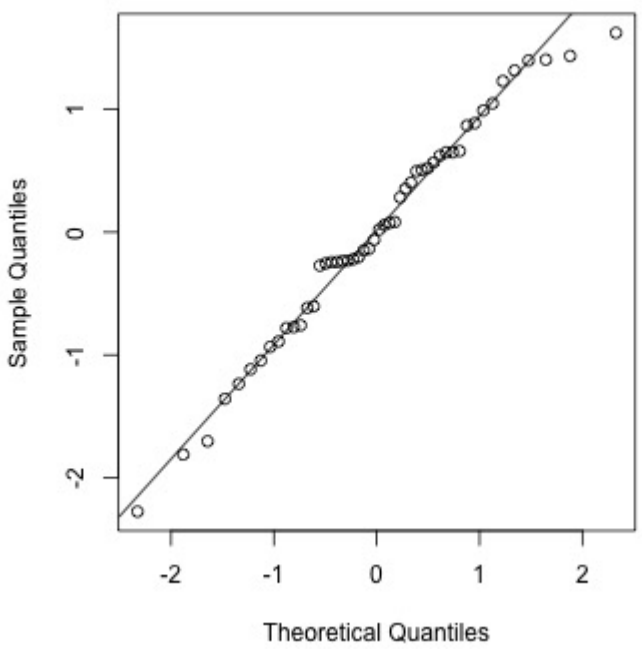
Let's take a look at examples of the different kinds of normal probability plots we can obtain and learn what each tells us.

Normally distributed residuals

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

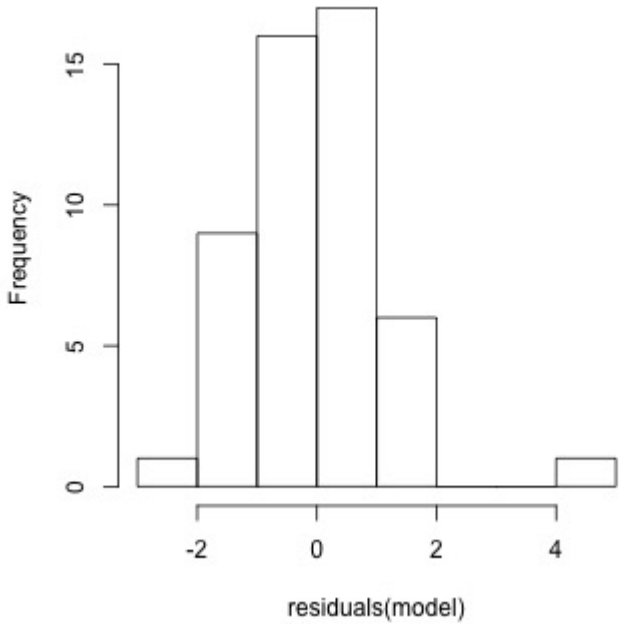


The normal probability plot of the residuals is approximately linear supporting the condition that the error terms are normally distributed.

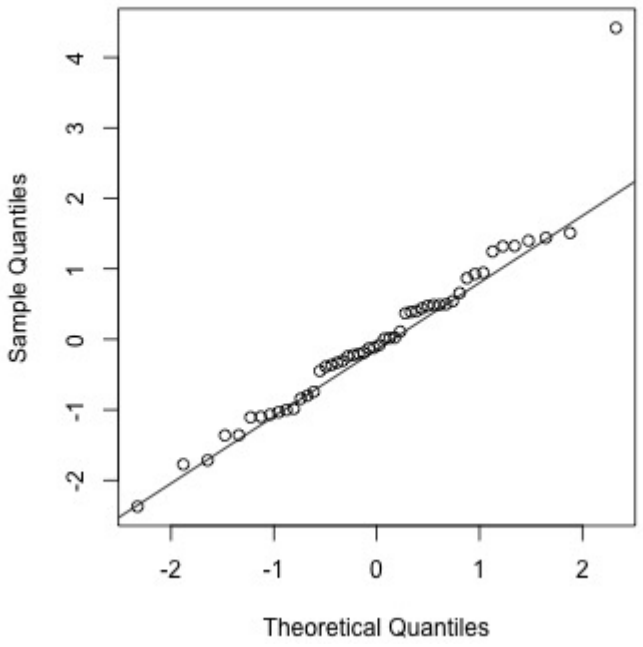


Normal residuals but with one outlier

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



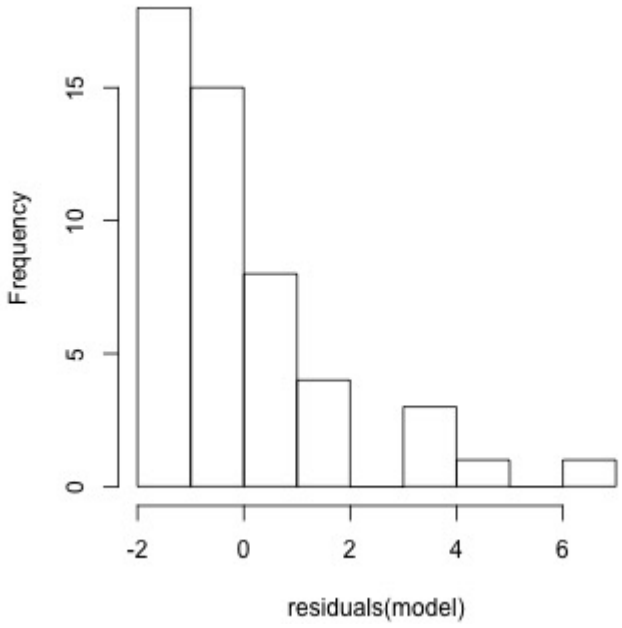
Here's the corresponding normal probability plot of the residuals:



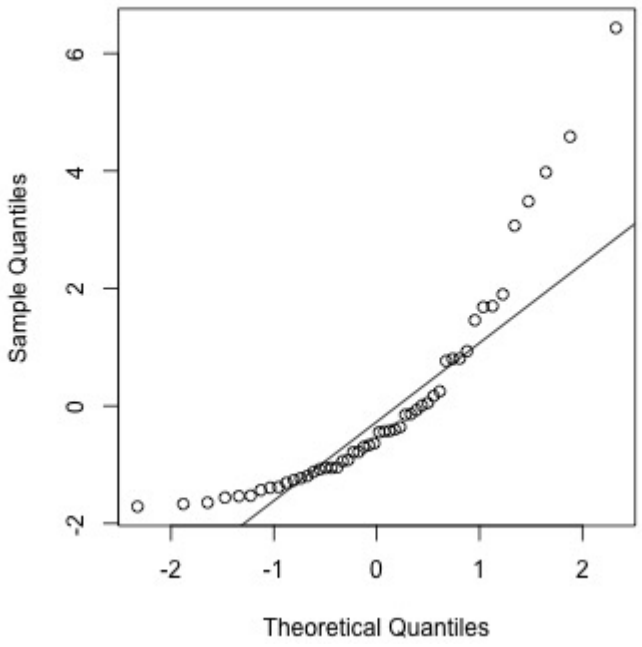
This is a classic example of what a normal probability plot looks like when the residuals are normally distributed, but there is just one outlier. The relationship is approximately linear with the exception of the one data point. We could proceed with the assumption that the error terms are normally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



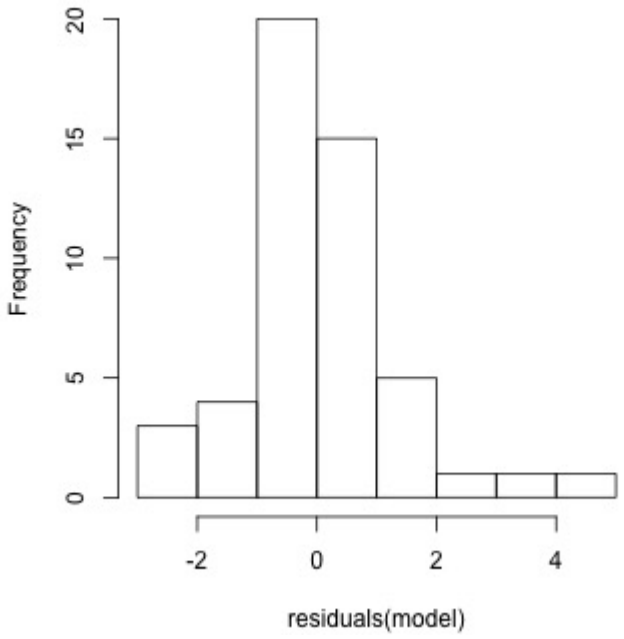
Here's the corresponding normal probability plot of the residuals:



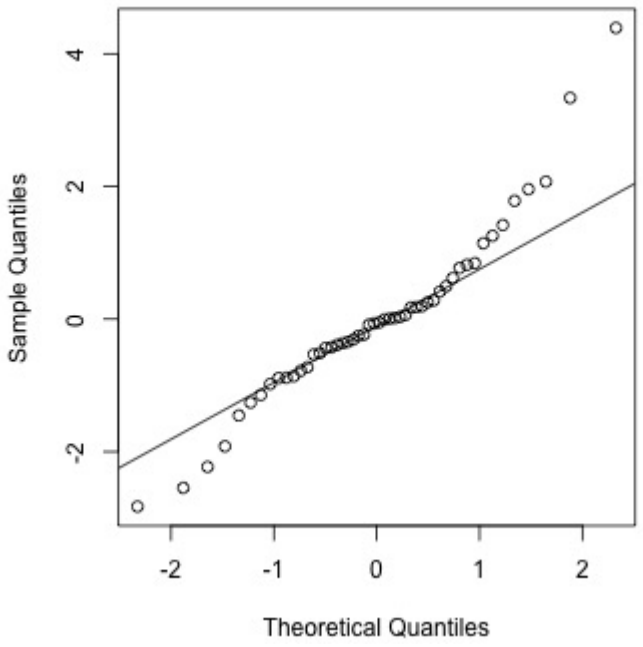
This is a classic example of what a normal probability plot looks like when the residuals are skewed. Clearly, the condition that the error terms are normally distributed is not met.

Heavy-tailed residuals

The following histogram of residuals suggests 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.

‹ 4.5 - Residuals vs. Order Plot (/stat462/node/121) up (/stat462/node/81) 4.7 - Assessing Linearity by Visual Inspection › (/stat462/node/123)

STAT 462

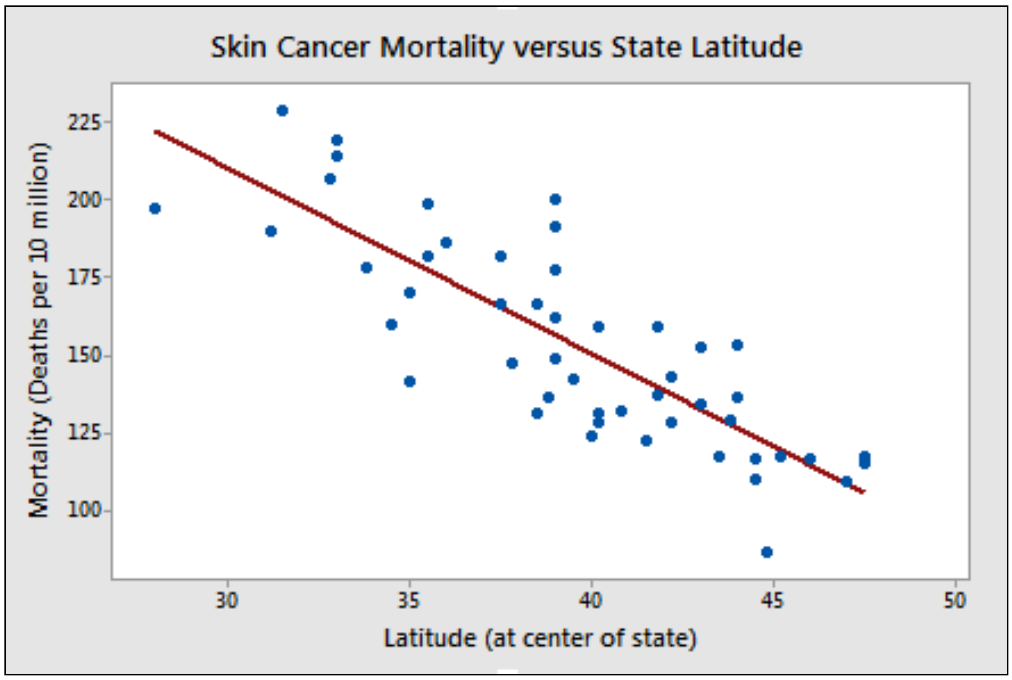
Applied Regression Analysis

4.7 - 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 y and just one predictor x — is that we can get a good feel for this condition just by looking at a simple scatter plot (so in this 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

Do 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/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt))?



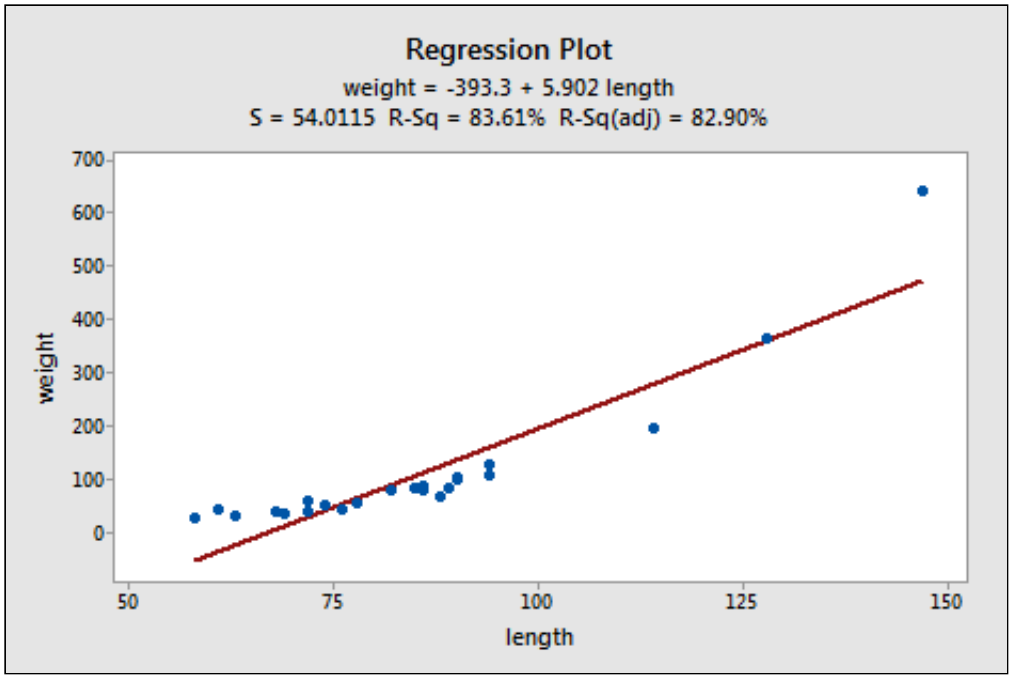
The answer is yes! It appears as if the relationship between latitude and skin cancer mortality is indeed linear, and therefore it would be best if we summarized the trend in the data using a linear function.

Alligators

Loading [MathJax]/extensions/MathZoom.js

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 then be used to predict the weight of the alligator. In taking this approach, some wildlife biologists captured a random sample of $n = 25$ alligators. They measured the length (x, in inches) and weight (y, in pounds) of each alligator. (alligator.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/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?



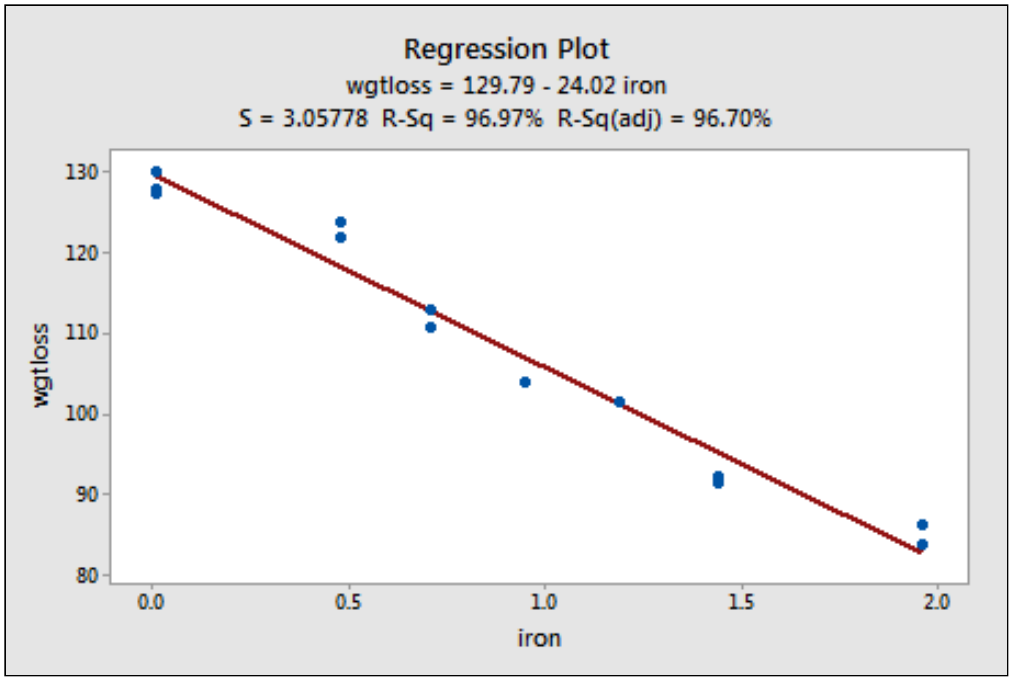
The answer is no! Don't you think a curved function would more adequately describe the trend? The scatter plot gives us a pretty good indication that a linear model is inadequate in this case.

Alloy Corrosion

Thirteen ($n = 13$) alloy specimens comprised of 90% copper and 10% nickel — 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 (x) and weight loss due to corrosion (y). (corrosion.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/corrosion.txt))

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

Loading [MathJax]/extensions/MathZoom.js



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

‹ 4.6 - Normal Probability Plot of Residuals (/stat462/node/122) up (/stat462/node/81) 4.8 - Further Residual Plot Examples › (/stat462/node/124)

Loading [MathJax]/extensions/MathZoom.js

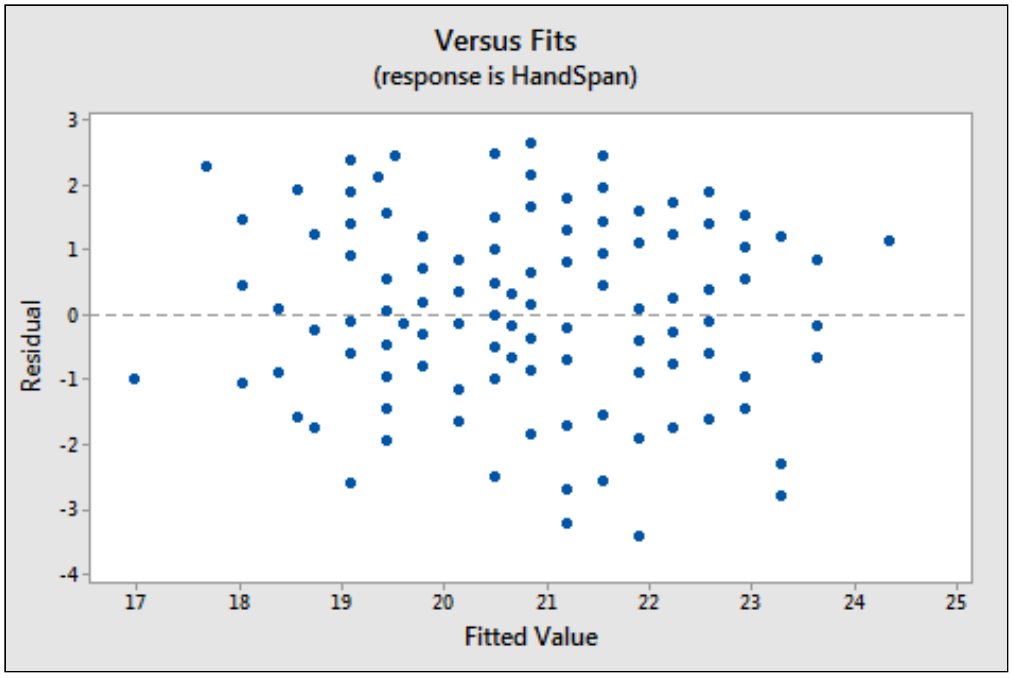
STAT 462

Applied Regression Analysis

4.8 - Further Residual Plot Examples

Example 1: A Good Residual Plot

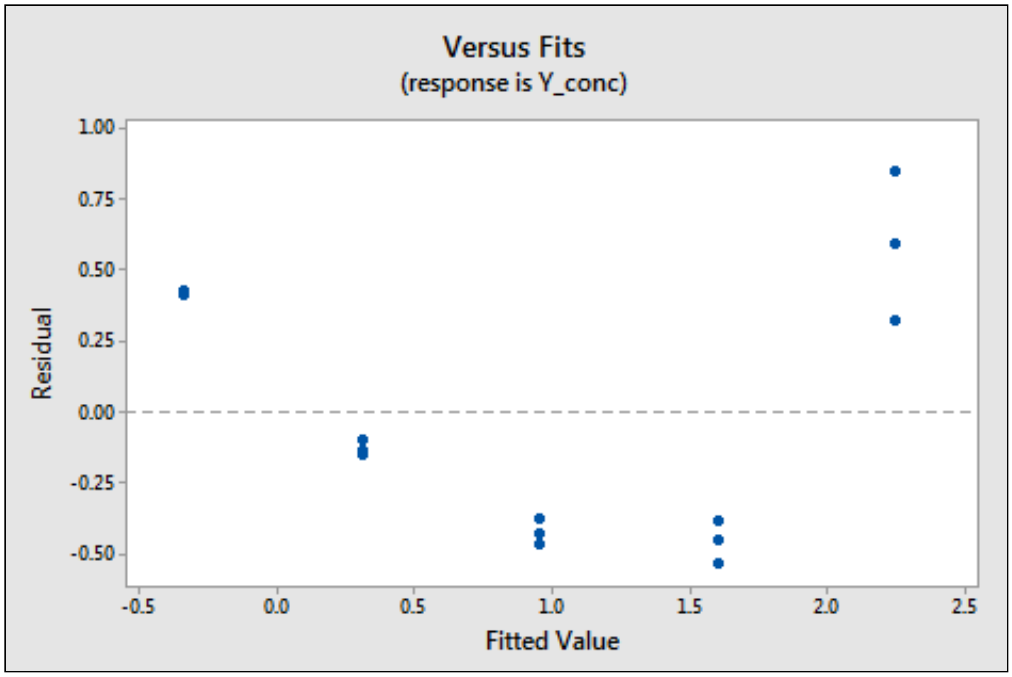
Below is a plot of residuals versus fits after a straight-line model was used on data for y = handspan (cm) and x = height (inches), for n = 167 students (handheight.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/handheight.txt)).



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

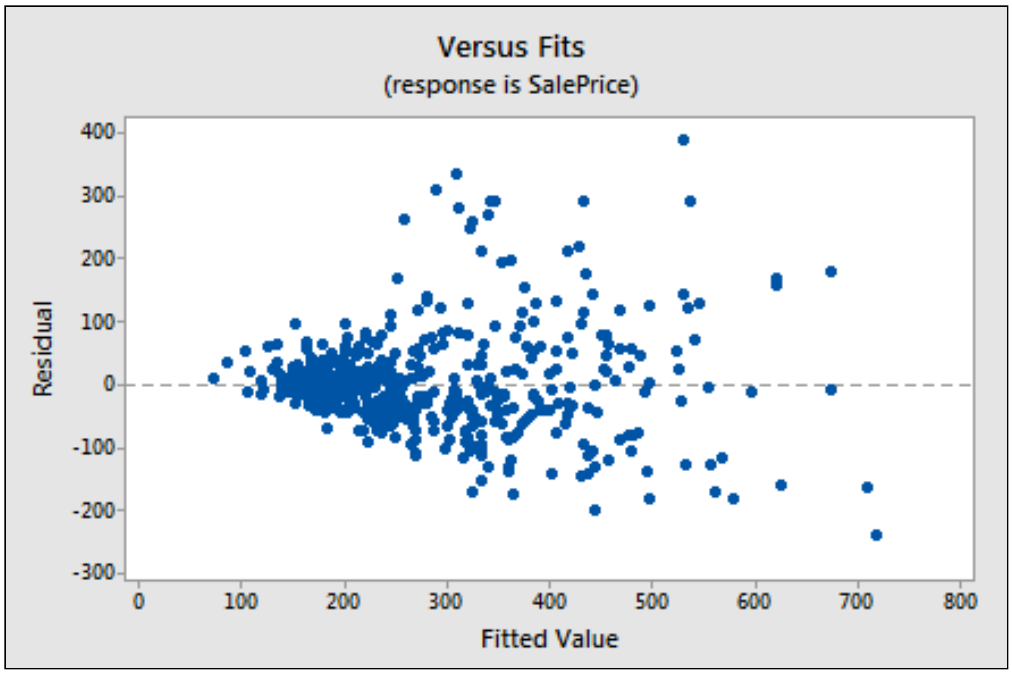
Below is a plot of residuals versus fits after a straight-line model was used on data for y = concentration of a chemical solution and x = time after solution was made (solutions_conc.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/solutions_conc.txt)).



Interpretation: This plot of residuals versus plots shows two difficulties. First, the pattern is curved which indicates that the wrong type of model 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

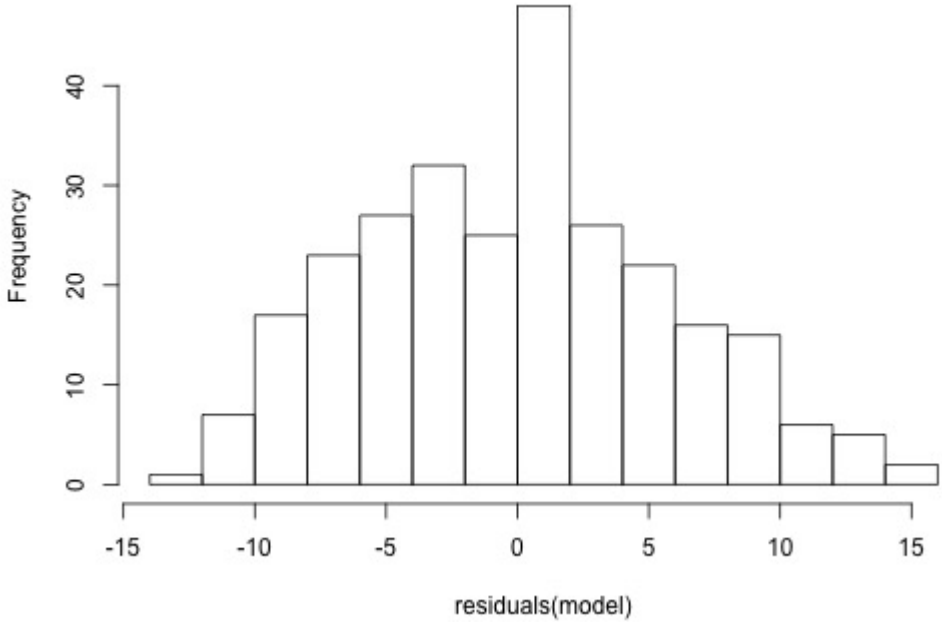
Below is a plot of residuals versus fits after a straight-line model was used on data for y = sale price of a home and x = square foot area of home (realestate.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/realestate.txt)).

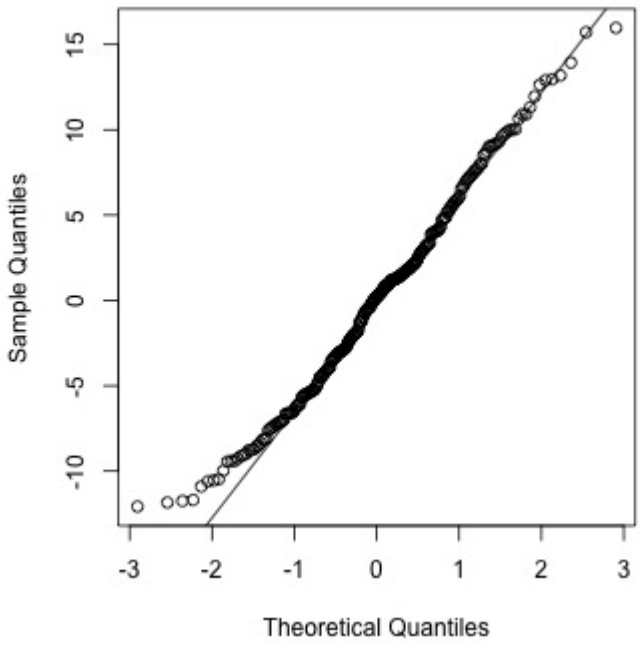


Interpretation: This plot of residuals versus fits shows that the residual variance (vertical spread) increases as the fitted values (predicted values of sale price) increase. This violates the assumption of constant error variance.

Example 5: Indications that Assumption of Normal Distribution for Errors is Valid

The graphs below are a histogram and a normal probability plot of the residuals after a straight-line model was used for fitting y = time to next eruption and x = duration of last eruption for eruptions of the Old Faithful geyser (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/oldfaithful.txt) .



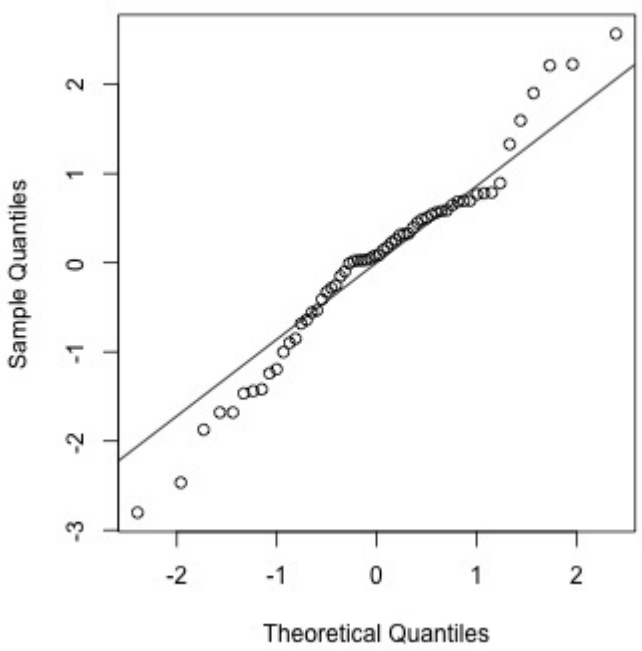


Interpretation: The histogram is roughly bell-shaped so it is an indication that it is reasonable to assume that the errors have a normal distribution. The pattern of the normal probability plot is straight, so this plot also provides evidence that it is reasonable to assume that the errors have a normal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

Example 5: Indications that Assumption of Normal Distribution for Errors is Not Valid

Below is a normal probability plot for the residuals from a straight-line regression with y = infection risk in a hospital and x = average length of stay in the hospital. The observational units are hospitals and the data are taken

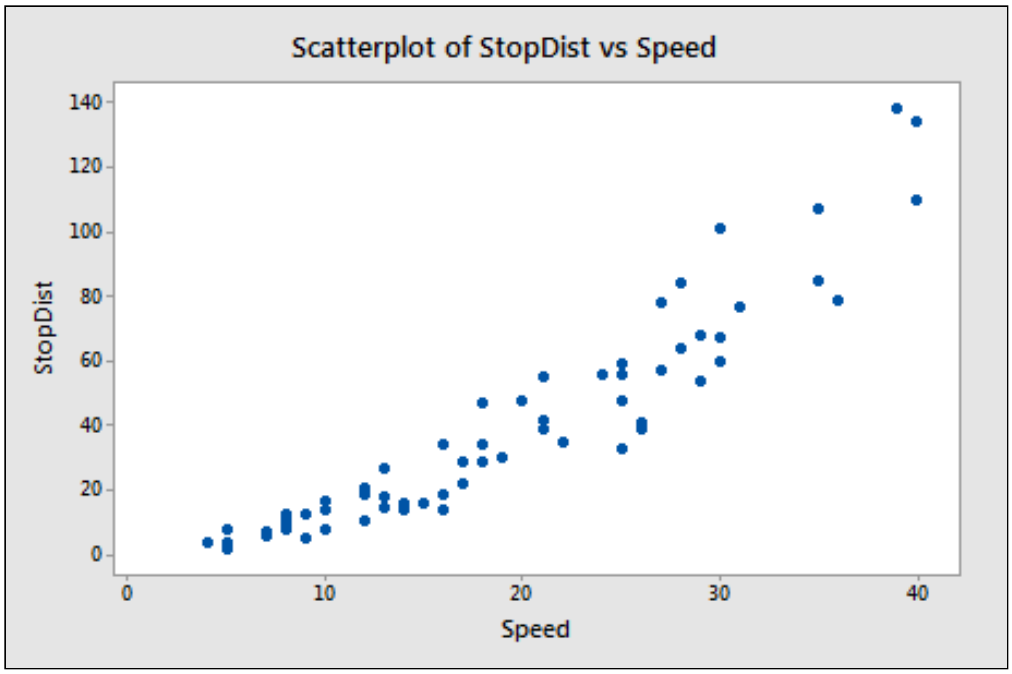
from regions 1 and 2 in the infection risk (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/infectionrisk.txt) dataset.



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 y can sometimes help us with nonconstant variance problems. We will look at the stopping distance data with y = stopping distance of a car and x = speed of the car when the brakes were applied (carstopping.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/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 nonconstant variance. One possible remedy is to transform y. With some trial and error, we find that there is an approximate

linear relationship between $\sqrt{y}$ and x with no suggestion of nonconstant variance.

The software 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.

```
The regression equation is sqrtdist = 0.918 + 0.253 Speed
Speed    Fit        95% PI
  10    3.44     1.98,  4.90
  20    5.97     4.52,  7.42
  30    8.50     7.03,  9.97
  40   11.03     9.53, 13.53
```

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).

Speed	Predicted	95% PI
10	11.83	3.92, 24.01
20	35.64	20.43, 55.06
30	72.25	49.42, 99.40
40	121.66	90.82, 156.75

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

We cover transformations like this in more detail in Lesson 7.

‹ 4.7 - Assessing Linearity by Visual Inspection (/stat462/node/123) | up (/stat462/node/81) | 4.9 - Estimation and Prediction Research Questions › (/stat462/node/125)

STAT 462

Applied Regression Analysis

4.9 - Estimation and Prediction Research Questions

In the remainder of this lesson, we are concerned with answering two different types of research questions. Our goal here — and throughout the practice of statistics — is to translate the research questions into reasonable statistical procedures.

Let's take a look at examples of these two types of research questions:

1. What is the mean weight, μ, of **all** American women, aged 18-24?

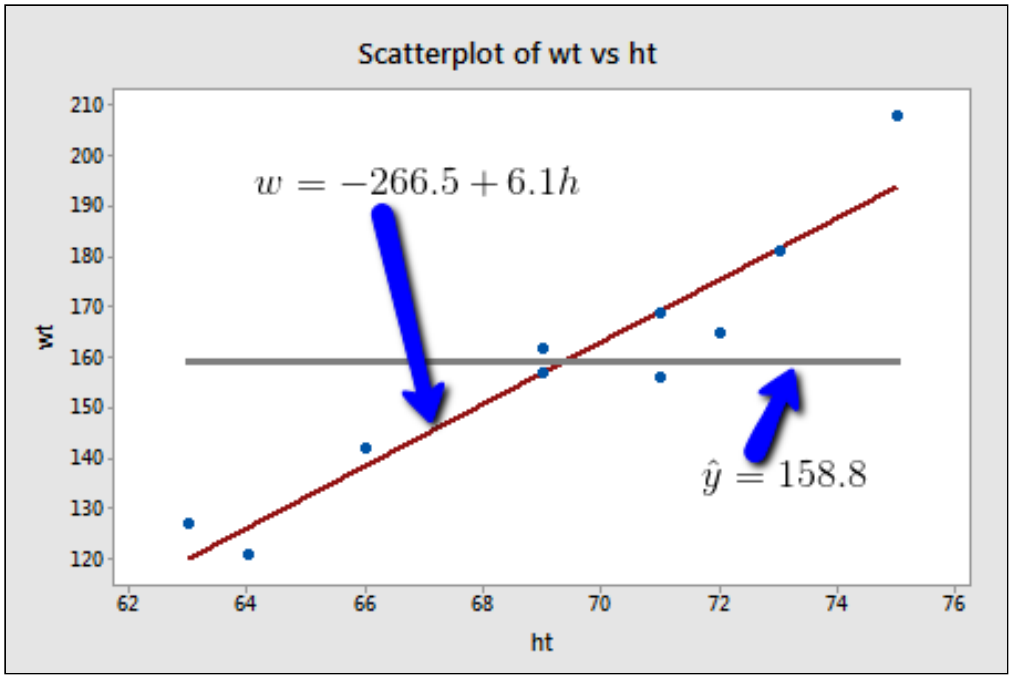
If we wanted to **estimate** μ, what would be a good estimate? It seems reasonable to calculate a confidence interval for μ using $\bar{y}$, the average weight of a random sample of American women, aged 18-24.

2. What is the weight, y, of **an individual** American woman, aged 18-24?

If we want to **predict** y, what would be a good prediction? It seems reasonable to calculate a "prediction interval" for y using, again, $\bar{y}$, the average weight of a random sample of American women, aged 18-24.

A person's weight is, of course, highly associated with the person's height. In answering each of the above questions, we likely could do better by taking into account a person's height. That's where an estimated regression equation becomes useful.

Here are some weight and height data from a sample of $n = 10$ people, (student_height_weight.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/student_height_weight.txt)):

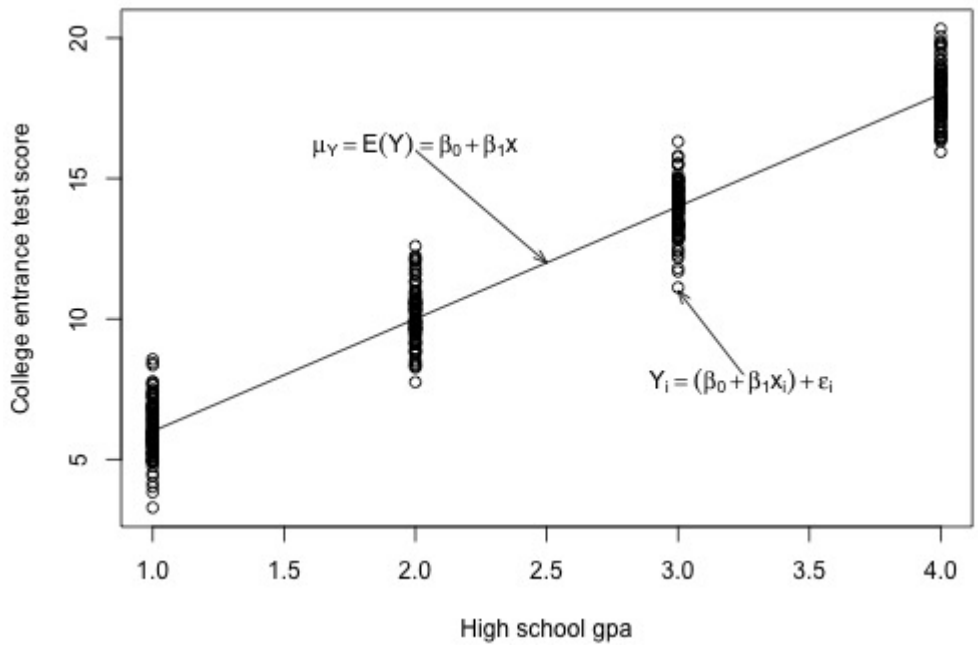


If we used the average weight of the 10 people in the sample to estimate μ, we would claim that the average weight of all American women aged 18-24 is 158.8 pounds *regardless* of the height of the women. Similarly, if we used the average weight of the 10 people in the sample to predict y, we would claim that the weight of an individual American women aged 18-24 is 158.8 pounds *regardless* of the woman's height.

On the other hand, if we used the estimated regression equation to estimate μ, we would claim that the average weight of all American women aged 18-24 *who are only 64 inches tall* is -266.5 + 6.1(64) = 123.9 pounds. [This calculation is based on rounded estimated regression coefficients; if you use unrounded estimates you'll get an answer closer to 126.4.] Similarly, we would predict that the weight y of an individual American women aged 18-24 *who is only 64 inches tall* is 123.9 pounds. This example makes it clear that we get significantly different (and better!) answers to our research questions when we take into account a person's height.

Let's make it clear that it is one thing to **estimate** μ_Y and yet another thing to **predict** y. (Note that we subscript μ with Y to make it clear that we are talking about the mean of the response Y not the mean of the predictor x.)

Let's return to our example in which we consider the potential relationship between the predictor "high school gpa" and the response "college entrance test score."



For this example, we could ask two different research questions concerning the response:

- What is the mean college entrance test score for the subpopulation of students whose high school gpa is 3? (Answering this question entails estimating the mean response μ_Y when $x = 3$.)
- What college entrance test score can we predict for a student whose high school gpa is 3? (Answering this question entails predicting the response y_{new} when $x = 3$.)

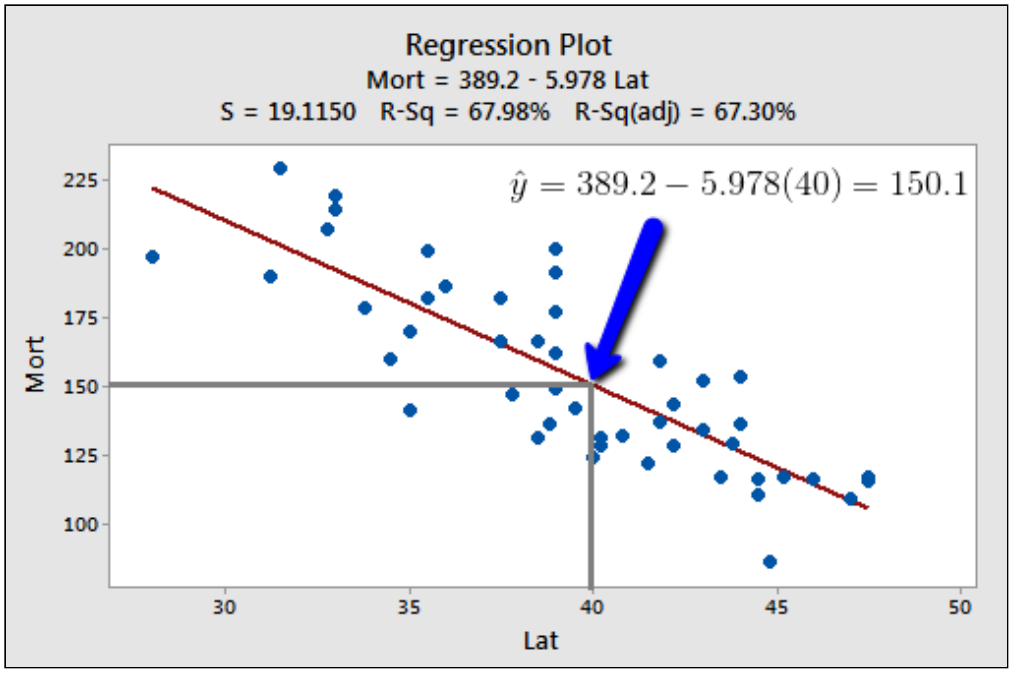
The two research questions can be asked more generally as:

- What is the mean response μ_Y when the predictor value is x_h ?
- What value will a new response y_{new} be when the predictor value is x_h?

Let's take a look at one more example, namely, the one concerning the relationship between the response "skin cancer mortality" and the predictor "latitude" (skincancer.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt)). Again, we could ask two different research questions concerning the response:

- What is the expected (mean) mortality rate due to skin cancer **for all locations** at 40 degrees north latitude?
- What is the predicted mortality rate **for an individual location** at 40 degrees north, say at Chambersburg, Pennsylvania?

At some level, answering these two research questions is straightforward. Both just involve using the estimated regression equation:



That is, $\hat{y}_h = b_0 + b_1 x_h$ is the best answer to each research question. It is the best guess of the mean response at x_h, and it is the best guess of a new response at x_h:

- Our best estimate of the mean mortality rate due to skin cancer **for all locations** at 40 degrees north latitude is 389.19 - 5.97764(40) = 150 deaths per 10 million people.
- Our best prediction of the mortality rate due to skin cancer in Chambersburg, Pennsylvania is 389.19 - 5.97764(40) = 150 deaths per 10 million people.

The problem with the answers to our two research questions is that we'd have obtained a completely different answer if we had selected a different random sample of data. As always, to be confident in the answer to our research questions, we should put an interval around our best guesses. We learn how to do this in the next two sections. That is, we first learn a "**confidence interval for μ_Y**" and then a "**prediction interval for y_{new}.**"

‹ 4.8 - Further Residual Plot Examples (/stat462/node/124) | up (/stat462/node/81) | 4.10 - Confidence Interval for the Mean Response › (/stat462/node/126)

STAT 462

Applied Regression Analysis

4.10 - Confidence Interval for the Mean Response

In this section, we are concerned with the confidence interval, called a "***t*-interval**," for the mean response μ_Y when the predictor value is x_h. Let's jump right in and learn the formula for the confidence interval. The general formula in words is as always:

Sample estimate ± (*t*-multiplier × standard error)

and the formula in notation is:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

where:

- $\hat{y}_h$ is the "**fitted value**" or "**predicted value**" of the response when the predictor is x_h
- $t_{(\alpha/2,n-2)}$ is the "***t*-multiplier**." Note that the *t*-multiplier has *n*-2 (not *n*-1) degrees of freedom, because the confidence interval uses the mean square error (*MSE*) whose denominator is *n*-2.
- $\sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$ is the "**standard error of the fit**," which depends on the mean square error (*MSE*), the sample size (*n*), how far in squared units the predictor value x_h is from the average of the predictor values $\bar{x}$, or $(x_h-\bar{x})^2$, and the sum of the squared distances of the predictor values x_i from the average of the predictor values $\bar{x}$, or $\sum(x_i-\bar{x})^2$.

Fortunately, we won't have to use the formula to calculate the confidence interval, since statistical software will do the dirty work for us. Here is some output for our example with "skin cancer mortality" as the response and "latitude" as the predictor (skincancer.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt)):

Prediction for Mort

Regression Equation

Mort = 389.2 - 5.978 Lat

Variable	Setting
Lat	40

Fit	SE Fit	95% CI	95% PI
150.084	2.74500	(144.562, 155.606)	(111.235, 188.933)

Here's what the output tells us:

- **Variable setting**: the value x_h (40 degrees north) for which we requested the confidence interval for μ_Y.
- The predicted value $\hat{y}_h$, ("**Fit**" = 150.084) and the standard error of the fit ("**SE Fit**" = 2.74500).
- **95% CI**: the 95% confidence interval. We can be 95% confident that the average skin cancer mortality rate of all locations at 40 degrees north is between 144.562 and 155.606 deaths per 10 million people.
- **95% PI**: the 95% prediction interval for a new response (which we discuss in the next section).

Factors affecting the width of the *t*-interval for the mean response μ_Y

Why do we bother learning the formula for the confidence interval for μ_Y when we let statistical software calculate it for us anyway? As always, the formula is useful for investigating what factors affect the width of the confidence interval for μ_Y. Again, the formula is:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

and therefore the width of the confidence interval for μ_Y is:

$$2 \times \left[t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}\right]$$

So how can we affect the width of our resulting interval for μ_Y?

- **As the mean square error (*MSE*) decreases, the width of the interval decreases.** Since *MSE* is an estimate of how much the data vary naturally around the unknown population regression line, we have little control over *MSE* other than making sure that we make our measurements as carefully as possible. (We will return to this issue later in the course when we address "model selection.")
- **As we decrease the confidence level, the *t*-multiplier decreases, and hence the width of the interval decreases.** In practice, we wouldn't want to set the confidence level below 90%.
- **As we increase the sample size *n*, the width of the interval decreases.** We have complete control over the size of our sample — the only limitation beting our time and financial constraints.
- The more spread out the predictor values, the larger the quantity $\sum(x_i-\bar{x})^2$ and hence the narrower the interval. In general, you should make sure your predictor values are not too clumped together but rather sufficiently spread out.
- The closer x_h is to the average of the sample's predictor values $\bar{x}$, the smaller the quantity $(x_h-\bar{x})^2$, and hence the narrower the interval. If you know that you want to use your estimated regression equation to

estimate μ_Y when the predictor's value is x_h, then you should be aware that the confidence interval will be narrower the closer x_h is to $\bar{x}$.

Let's see this last claim in action for our example with "skin cancer mortality" as the response and "latitude" as the predictor:

```
Predicted Values for New Observations

New  Fit   SE Fit   95.0% CI      95.0% PI
1   150.08  2.75  (144.6,155.6)  (111.2,188.93)
2   221.82  7.42  (206.9,236.8)  (180.6,263.07)X
X  denotes a row with X values away from the
center

Values of Predictors for New Observations
New Obs  Latitude
1           40.0      Mean of Lat = 39.533
2           28.0
```

The software output reports a 95% confidence interval for μ_Y for a latitude of 40 degrees north (first row) and 28 degrees north (second row). The average latitude of the 49 states in the data set is 39.533 degrees north. The output tells us:

- We can be 95% confident that the mean skin cancer mortality rate of all locations at 40 degrees north is between 144.6 and 155.6 deaths per 10 million people.
- And, we can be 95% confident that the mean skin cancer mortality rate of all locations at 28 degrees north is between 206.9 and 236.8 deaths per 10 million people.

The width of the 40 degree north interval (155.6 - 144.6 = 11 deaths) is shorter than the width of the 28 degree north interval (236.8 - 206.9 = 29.9 deaths), because 40 is much closer than 28 is to the sample mean 39.533. Note that some software is kind enough to warn us that 28 degrees north is far from the mean of the sample's predictor values.

When is it okay to use the formula for the confidence interval for μ_Y?

One thing we haven't discussed yet is when it is okay to use the formula for the confidence interval for μ_Y. It is okay:

- When x_h is a value within the range of the x values in the data set — that is, when x_h is a value within the "**scope of the model**." But, note that x_h does not have to be one of the actual x values in the data set.
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. The formula works okay even if the error terms are only approximately normal. And, if you have a large sample, the error terms can even deviate substantially from normality.

‹ 4.9 - Estimation and Prediction Research Questions (/stat462/node/125) up (/stat462/node/81) 4.11 - Prediction Interval for a New Response › (/stat462/node/127)

STAT 462

Applied Regression Analysis

4.11 - Prediction Interval for a New Response

In this section, we are concerned with the prediction interval for a new response y_{new} when the predictor's value is x_h. Again, let's just jump right in and learn the formula for the prediction interval. The general formula in words is as always:

Sample estimate ± (*t*-multiplier × standard error)

and the formula in notation is:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

where:

- $\hat{y}_h$ is the "**fitted value**" or "**predicted value**" of the response when the predictor is x_h
- $t_{(\alpha/2,n-2)}$ is the "***t*-multiplier**." Note again that the *t*-multiplier has *n*-2 (not *n*-1) degrees of freedom, because the prediction interval uses the mean square error (*MSE*) whose denominator is *n*-2.
- $\sqrt{MSE \times \left(1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$ is the "**standard error of the prediction**," which is very similar to the "standard error of the fit" when estimating μ_Y. The standard error of the prediction just has an extra MSE term added that the standard error of the fit does not. (More on this a bit later.)

Again, we won't use the formula to calculate our prediction intervals. We'll let statistical software do the calculation for us. Let's look at the prediction interval for our example with "skin cancer mortality" as the response and "latitude" as the predictor (skincancer.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt)):

Loading [MathJax]/extensions/MathZoom.js

Prediction for Mort

Regression Equation

Mort = 389.2 - 5.978 Lat

Variable	Setting
Lat	40

Fit	SE Fit	95% CI	95% PI
150.084	2.74500	(144.562, 155.606)	(111.235, 188.933)

The output reports the 95% prediction interval for an individual location at 40 degrees north. We can be 95% confident that the skin cancer mortality rate at an individual location at 40 degrees north will be between 111.235 and 188.933 deaths per 10 million people.

When is it okay to use the prediction interval for y_{new} formula?

The requirements are similar to, but a little more restrictive than, those for the confidence interval. It is okay:

- When x_h is a value within the scope of the model. Again, x_h does not have to be one of the actual x values in the data set.
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. Unlike the case for the formula for the confidence interval, the formula for the prediction interval depends **strongly** on the condition that the error terms are normally distributed.

Understanding the difference in the two formulas

In our discussion of the confidence interval for μ_Y, we used the formula to investigate what factors affect the width of the confidence interval. There's no need to do it again. Because the formulas are so similar, it turns out that the factors affecting the width of the prediction interval are identical to the factors affecting the width of the confidence interval.

Let's instead investigate the formula for the prediction interval for y_{new}:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(1+\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

to see how it compares to the formula for the confidence interval for μ_Y:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \left(\frac{1}{n}+\frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$$

Observe that the only difference in the formulas is that the standard error of the prediction for y_{new} has an extra *MSE* term in it that the standard error of the fit for μ_Y does not.

Let's try to understand the prediction interval to see what causes the extra *MSE* term. In doing so, let's start with an easier problem first. Think about how we could predict a new response y_{new} at a particular x_h if the mean of the

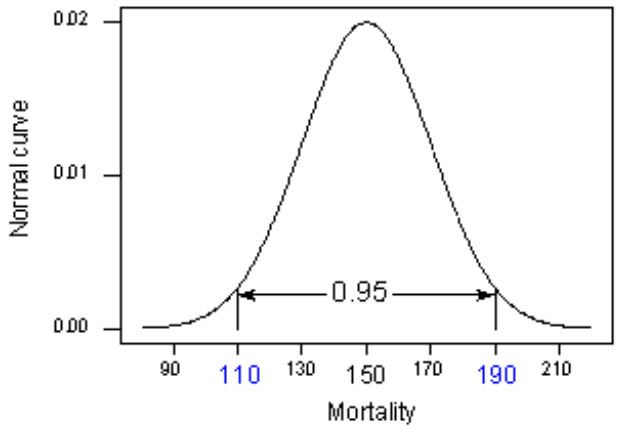
Loading [MathJax]/extensions/MathZoom.js

responses μ_Y at x_h were known. That is, suppose it were known that the mean skin cancer mortality at $x_h = 40^o$ N is 150 deaths per million (with variance 400)? What is the predicted skin cancer mortality in Columbus, Ohio?

Because $\mu_Y = 150$ and $\sigma^2 = 400$ are known, we can take advantage of the "**empirical rule**," which states among other things that 95% of the measurements of normally distributed data are within 2 standard deviations of the mean. That is, it says that 95% of the measurements are in the interval sandwiched by:

$$\mu_Y - 2\sigma \text{ and } \mu_Y + 2\sigma.$$

Applying the 95% rule to our example with $\mu_Y = 150$ and $\sigma = 20$:



95% of the skin cancer mortality rates of locations at 40 degrees north latitude are in the interval sandwiched by:

150 - 2(20) = 110 and 150 + 2(20) = 190.

That is, if someone wanted to know the skin cancer mortality rate for a location at 40 degrees north, our best guess would be somewhere between 110 and 190 deaths per 10 million. The problem is that our calculation used μ_Y and σ, population values that we would typically not know. Reality sets in:

- **The mean μ_Y is typically not known.** The logical thing to do is estimate it with the predicted response $\hat{y}$. The cost of using $\hat{y}$ to estimate μ_Y is the variance of $\hat{y}$. That is, different samples would yield different predictions $\hat{y}$, and so we have to take into account this variance of $\hat{y}$.
- **The variance σ^2 is typically not known.** The logical thing to do is to estimate it with *MSE*.

Because we have to estimate these unknown quantities, the variation in the prediction of a new response depends on two components:

1. the variation due to estimating the mean μ_Y with $\hat{y}_h$, which we denote "$\sigma^2(\hat{Y}_h)$." (Note that the estimate of this quantity is just the standard error of the fit that appears in the confidence interval formula.)
2. the variation in the responses *y*, which we denote as "σ^2." (Note that quantity is estimated, as usual, with the mean square error *MSE*.)

Adding the two variance components, we get:

Loading [MathJax]/extensions/MathZoom.js

$$\sigma^2 + \sigma^2(\hat{Y}_h)$$

which is estimated by:

$$MSE + MSE\left[\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = MSE\left[1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Do you recognize this quantity? It's just the variance of the prediction that appears in the formula for the prediction interval y_{new}!

Let's compare the two intervals again:

Confidence interval for μ_Y:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

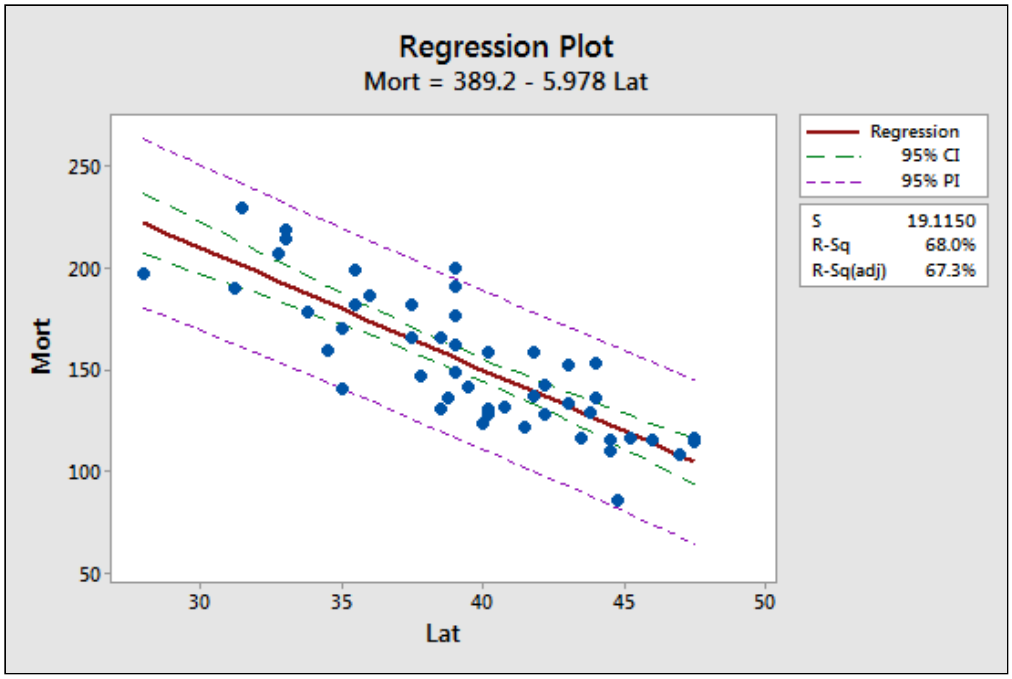
Prediction interval for y_{new}:

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

What's the practical implications of the difference in the two formulas?

- Because the prediction interval has the extra *MSE* term, a (1-α)100% confidence interval for μ_Y at x_h will always be narrower than the corresponding (1-α)100% prediction interval for y_{new} at x_h.
- By calculating the interval at the sample's mean of the predictor values ($\boldsymbol{x_h} = \bar{x}$) and increasing the sample size *n*, the confidence interval's standard error can approach 0. Because the prediction interval has the extra *MSE* term, the prediction interval's standard error cannot get close to 0.

The first implication is seen most easily by studying the following plot for our skin cancer mortality example:



Loading [MathJax]/extensions/MathZoom.js

erval (in **purple**) is always wider than the confidence interval (in **green**). Furthermore, both intervals are narrowest at the mean of the predictor values (about 39.5).

STAT 462

Applied Regression Analysis

4.12 - Further Example of Confidence and Prediction Intervals

Hospital Infection Data

The hospital infection risk (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/infectionrisk.txt) dataset consists of a sample of 113 hospitals in four regions of the U.S. The response variable is y = infection risk (percent of patients who get an infection) and the predictor variable is x = average length of stay (in days). Here we analyze $n = 58$ hospitals in the east and north central U.S (regions 1 and 2). [Two hospitals with extreme values for Stay have also been removed.] Statistical software output for a simple linear regression model fit to these data follows:

Regression Analysis: InfctRsk versus Stay

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	38.3059	38.3059	36.50	0.000
Stay	1	38.3059	38.3059	36.50	0.000
Error	56	58.7763	1.0496		
Lack-of-Fit	54	58.5513	1.0843	9.64	0.098
Pure Error	2	0.2250	0.1125		
Total	57	97.0822			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.02449	39.46%	38.38%	35.07%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.160	0.956	-1.21	0.230	
Stay	0.5689	0.0942	6.04	0.000	1.00

Regression Equation

ctRsk = -1.160 + 0.5689 Stay

Loading [MathJax]/extensions/MathZoom.js

Software output with information for $x = 10$.

Prediction for InfctRsk

Regression Equation

InfctRsk = -1.160 + 0.5689 Stay

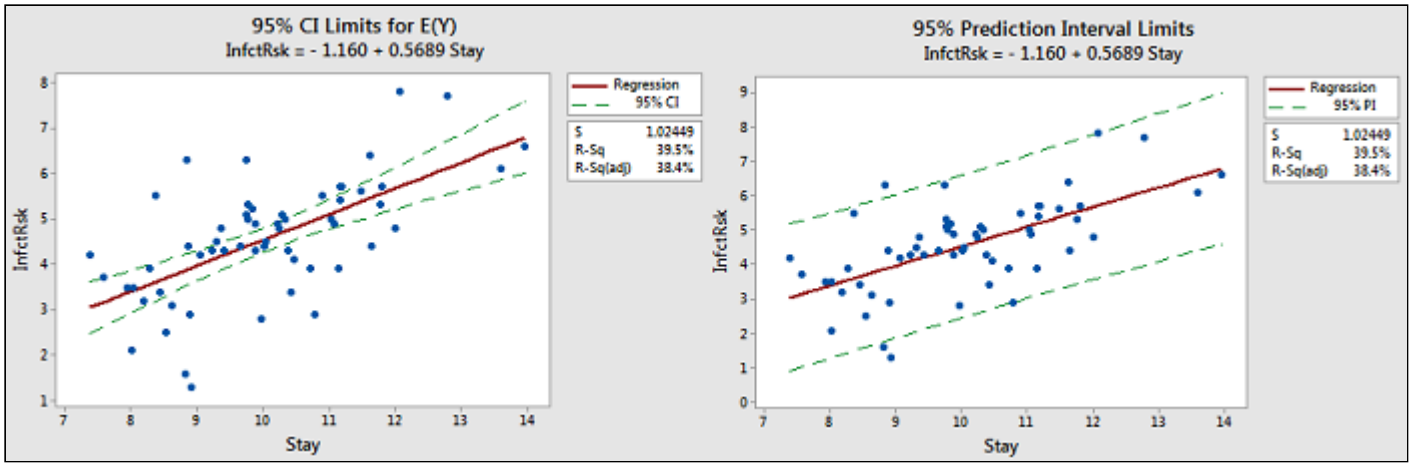
Variable	Setting
Stay	10

Fit	SE Fit	95% CI	95% PI
4.52885	0.134602	(4.25921, 4.79849)	(2.45891, 6.59878)

We can make the following observations:

1. For the interval given under 95% CI, we say with 95% confidence we can estimate that in hospitals in which the average length of stay is 10 days, the mean infection risk is between 4.25921 and 4.79849.
2. For the interval given under 95% PI, we say with 95% confidence that for any future hospital where the average length of stay is 10 days, the infection risk is between 2.45891 and 6.59878.
3. The value under Fit is calculated as $\hat{y} = -1.160 + 0.5689(10) = 4.529$.
4. The value under SE Fit is the standard error of $\hat{y}$ and it measures the accuracy of $\hat{y}$ as an estimate of E(Y).
5. Since $df = n - 2 = 58 - 2 = 56$, the multiplier for 95% confidence is 2.00324. The 95% CI for E(Y) is calculated as $4.52885 \pm (2.00324 \times 0.134602) = 4.52885 \pm 0.26964 = (4.259, 4.798)$.
6. Since S = $\sqrt{MSE}$ = 1.02449, the 95% PI is calculated as $4.52885 \pm (2.00324 \times \sqrt{1.02449^2 + 0.134602^2}) = 4.52885 \pm 0.20699 = (2.459, 6.599)$.

The following figure provides plots showing the difference between a 95% CI for E(Y) and 95% PI for y.



There are also some things to note:

1. Notice that the limits for E(Y) are close to the line. The purpose for those limits is to estimate the "true" location of the line.
2. Notice that the prediction limits (on the right) bracket most of the data. Those limits describe the location of individual y-values.
3. Notice that the prediction intervals are wider than the confidence intervals. This is something that can be noted by the formulas.

Loading [MathJax]/extensions/MathZoom.js

‹ 4.11 - Prediction Interval for a New Response (/stat462/node/127) up (/stat462/node/81)

Loading [MathJax]/extensions/MathZoom.js