STAT 462

Applied Regression Analysis

7.1 - Log-transforming Only the Predictor for SLR

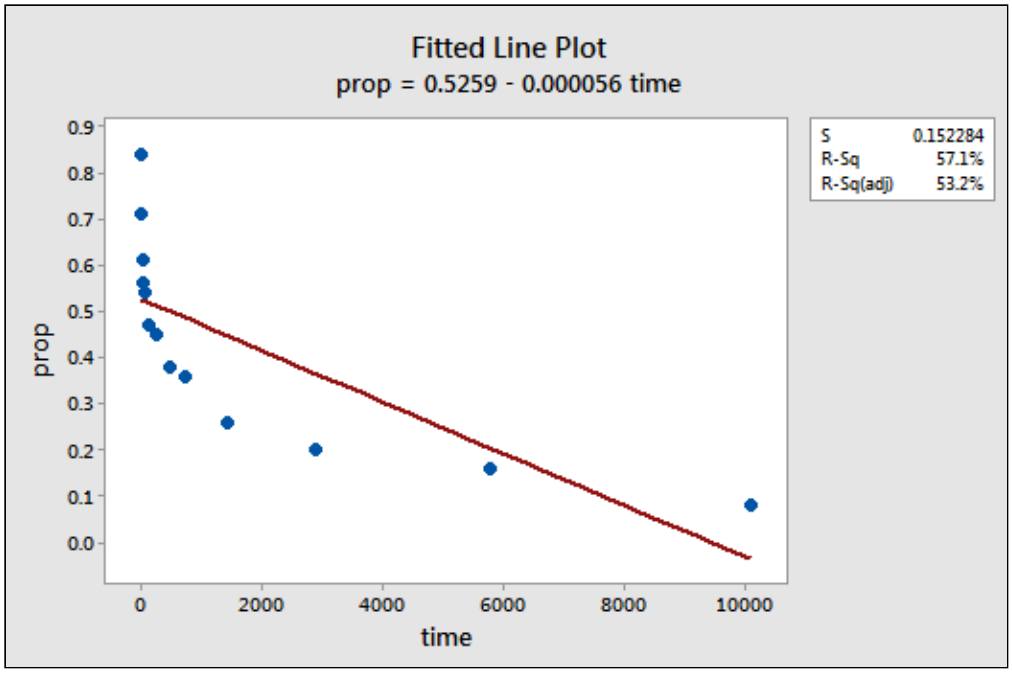
In this section, we learn **how to build** and **use** a simple linear regression model by transforming the predictor x values. This might be the first thing that you try if you find a non-linear trend in your data. That is, transforming the x values is appropriate **when non-linearity is the only problem** (i.e., the independence, normality, and equal variance conditions are met). Note, though, that it may be necessary to correct the non-linearity before you can assess the normality and equal variance assumptions. Also, while some assumptions may appear to hold prior to applying a transformation, they may no longer hold once a transformation is applied. In other words, using transformations is part of an iterative process where all the linear regression assumptions are re-checked after each iteration.

Keep in mind that although we're focussing on a simple linear regression model here, the essential ideas apply more generally to multiple linear regression models too. We can consider transforming any of the predictors by examining scatterplots of the residuals versus each predictor in turn.

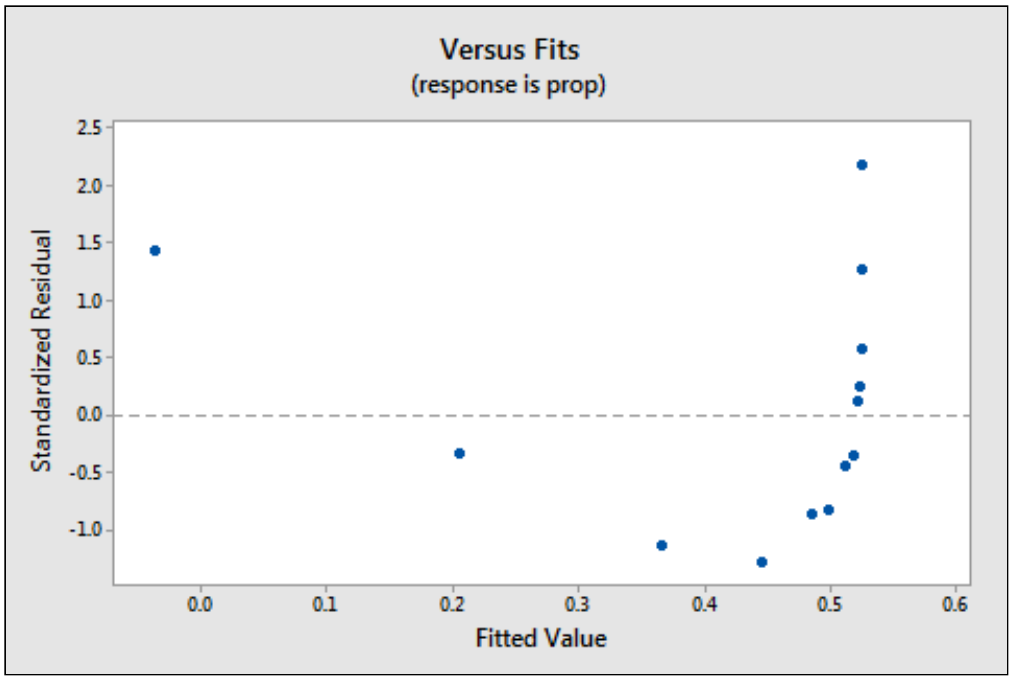
Building the model

An example. The easiest way to learn about data transformations is by example. Let's consider the data from a memory retention experiment in which 13 subjects were asked to memorize a list of disconnected items. The subjects were then asked to recall the items at various times up to a week later. The proportion of items (y = *prop*) correctly recalled at various times (x = *time*, in minutes) since the list was memorized were recorded (wordrecall.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/wordrecall.txt)) and plotted. Recognizing that there is no good reason that the error terms would not be independent, let's evaluate the remaining three conditions — linearity, normality, and equal variances — of the model.

The resulting fitted line plot suggests that the proportion of recalled items (y) is not linearly related to time (x):

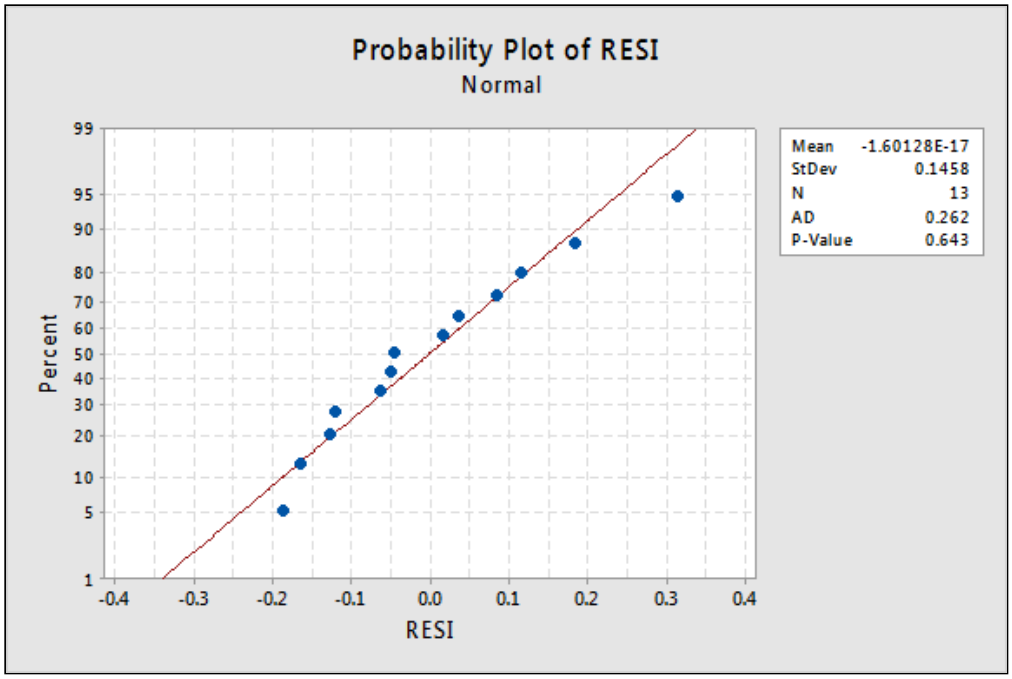


The residuals vs. fits plot also suggests that the relationship is not linear:



Because the lack of linearity dominates the plot, we cannot use the plot to evaluate whether or not the error variances are equal. We have to fix the non-linearity problem before we can assess the assumption of equal variances.

What about the normal probability plot of the residuals? What does it suggest about the error terms? Can we conclude that they are normally distributed?



The Anderson-Darling *P*-value for this example is 0.643, which suggests that we fail to reject the null hypothesis of normal error terms. There is not enough evidence to conclude that the errors terms are not normal.

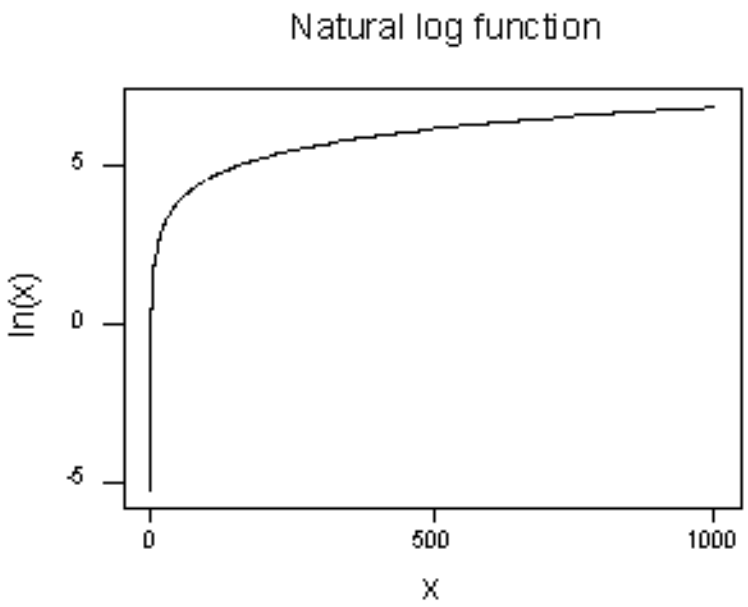
In summary, we have a data set in which non-linearity is the only major problem. This situation screams out for transforming only the predictor's values. Before we do so, let's take an aside and discuss the "**logarithmic transformation**," since it is the most common and most useful data transformation available.

The logarithmic transformation. The default logarithmic transformation merely involves taking the natural logarithm — denoted *ln* or log_e or simply *log* — of each data value. One could consider taking a different kind of logarithm, such as log base 10 or log base 2. However, the natural logarithm — which can be thought of as log base *e* where *e* is the constant 2.718282... — is the most common logarithmic scale used in scientific work.

The general characteristics of the natural logarithmic function are:

- The natural logarithm of x is the power of $e = 2.718282...$ that you have to take in order to get x. This can be stated notationally as $ln(e^x) = x$. For example, the natural logarithm of 5 is the power to which you have to raise $e = 2.718282...$ in order to get 5. Since $2.718282^{1.60944}$ is approximately 5, we say that the natural logarithm of 5 is 1.60944. Notationally, we say $ln(5) = 1.60944$.
- The natural logarithm of e is equal to one, that is, $ln(e) = 1$.
- The natural logarithm of one is equal to zero, that is, $ln(1) = 0$.

The plot of the natural logarithm function:



suggests that the effects of taking the natural logarithmic transformation are:

- Small values that are close together are spread further out.
- Large values that are spread out are brought closer together.

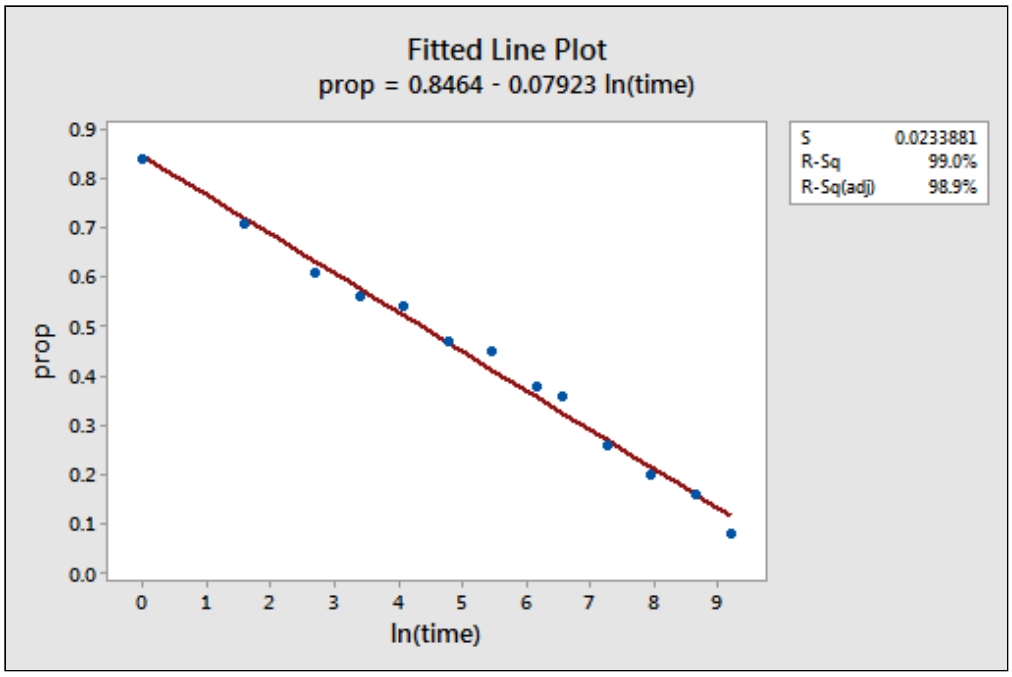
Back to the example. Let's use the natural logarithm to transform the *x* values in the memory retention experiment data. Since *x* = time is the predictor, all we need to do is take the natural logarithm of each time value appearing in the data set. In doing so, we create a newly transformed predictor called *lntime*:

time	***prop***	***lntime***
1	0.84	0.00000
5	0.71	1.60944
15	0.61	2.70805
30	0.56	3.40120
60	0.54	4.09434
120	0.47	4.78749
240	0.45	5.48064
480	0.38	6.17379
720	0.36	6.57925
1440	0.26	7.27240
2880	0.20	7.96555
5760	0.16	8.65869
10080	0.08	9.21831

We take the natural logarithm for each value of time and place the results in their own column. That is, we "transform" each predictor ***time*** value to a ***ln(time)*** value. For example, $ln(1) = 0$, $ln(5) = 1.60944$, and $ln(15) = 2.70805$, and so on.

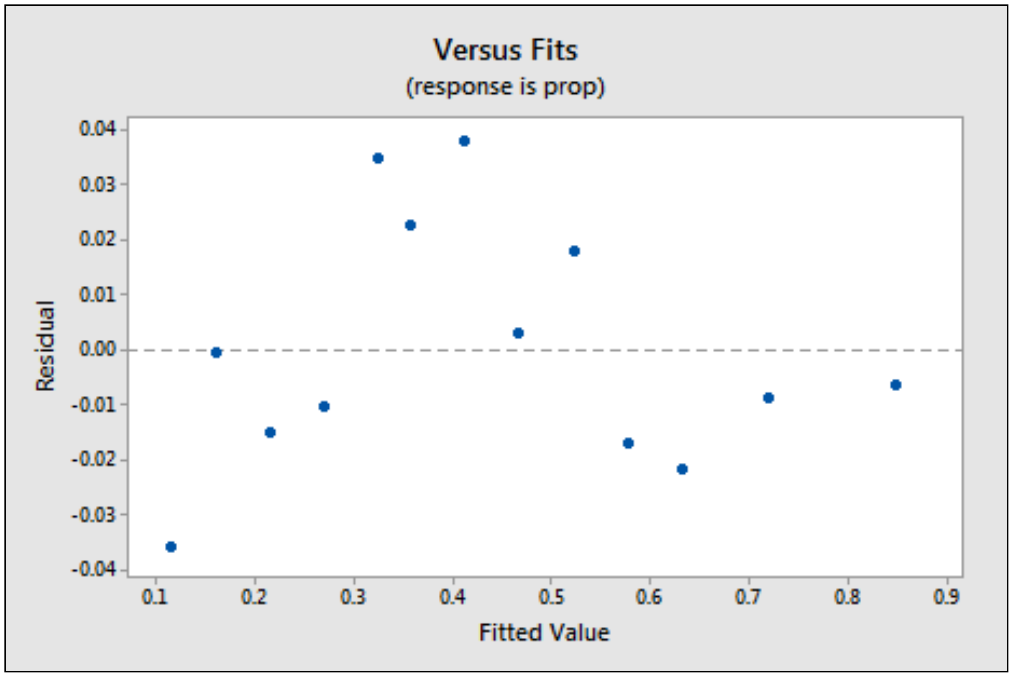
Now that we've transformed the predictor values, let's see if it helped correct the non-linear trend in the data. We re-fit the model with y = prop as the response and x = lntime as the predictor.

The resulting fitted line plot suggests that taking the natural logarithm of the predictor values is helpful.



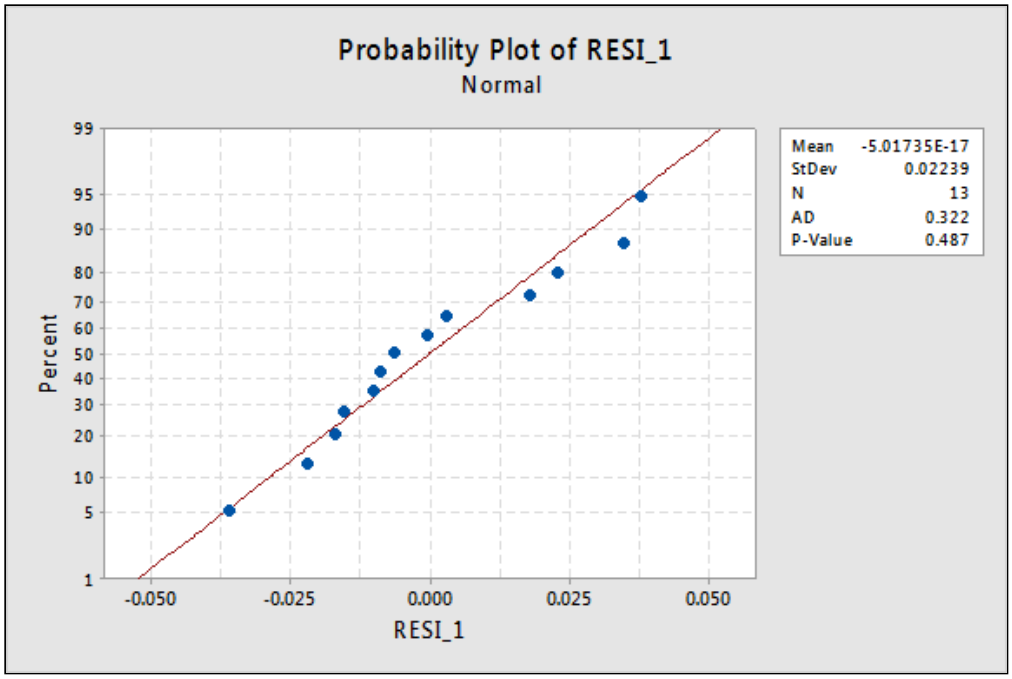
Indeed, the R^2 value has increased from 57.1% to 99.0%. It tells us that 99% of the variation in the proportion of recalled words (*prop*) is reduced by taking into account the natural log of time (*lntime*)!

The new residual vs. fits plot shows a significant improvement over the one based on the untransformed data.



You might become concerned about some kind of a up-down cyclical trend in the plot. I caution you again not to over-interpret these plots, especially when the data set is small like this. You really shouldn't expect perfection when you resort to taking data transformations. Sometimes you have to just be content with significant improvements. By the way, the plot also suggests that it is okay to assume that the error variances are equal.

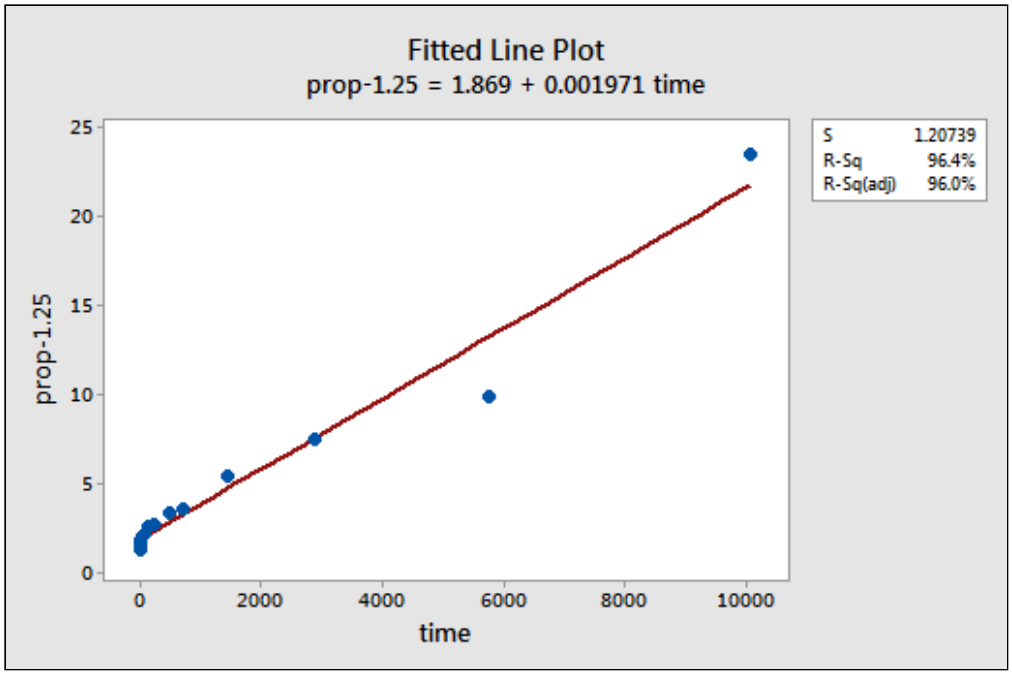
The normal probability plot of the residuals shows that transforming the x values had no effect on the normality of the error terms:



Again the Anderson-Darling P-value is large, so we fail to reject the null hypothesis of normal error terms. There is not enough evidence to conclude that the errors terms are not normal.

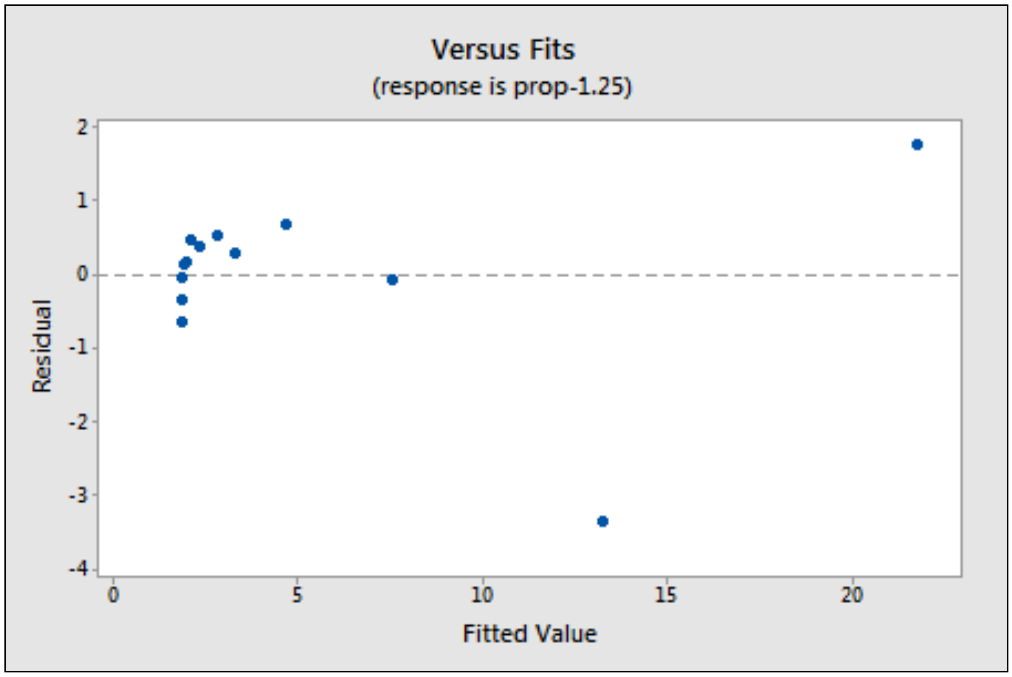
What if we had transformed the y values instead? Earlier I said that while some assumptions may appear to hold prior to applying a transformation, they may no longer hold once a transformation is applied. For example, if the error terms are well-behaved, transforming the y values could change them into badly-behaved error terms. The error terms for the memory retention data prior to transforming the x values appear to be well-behaved (in the sense that they appear approximately normal). Therefore, we might expect that transforming the y values instead of the x values could cause the error terms to become badly-behaved. Let's take a quick look at the memory retention data to see an example of what can happen when we transform the y values when non-linearity is the only problem.

By trial and error, we discover that the power transformation of y that does the best job at correcting the non-linearity is $y^{-1.25}$. The fitted line plot illustrates that the transformation does indeed straighten out the relationship — although admittedly not as well as the log transformation of the x values.

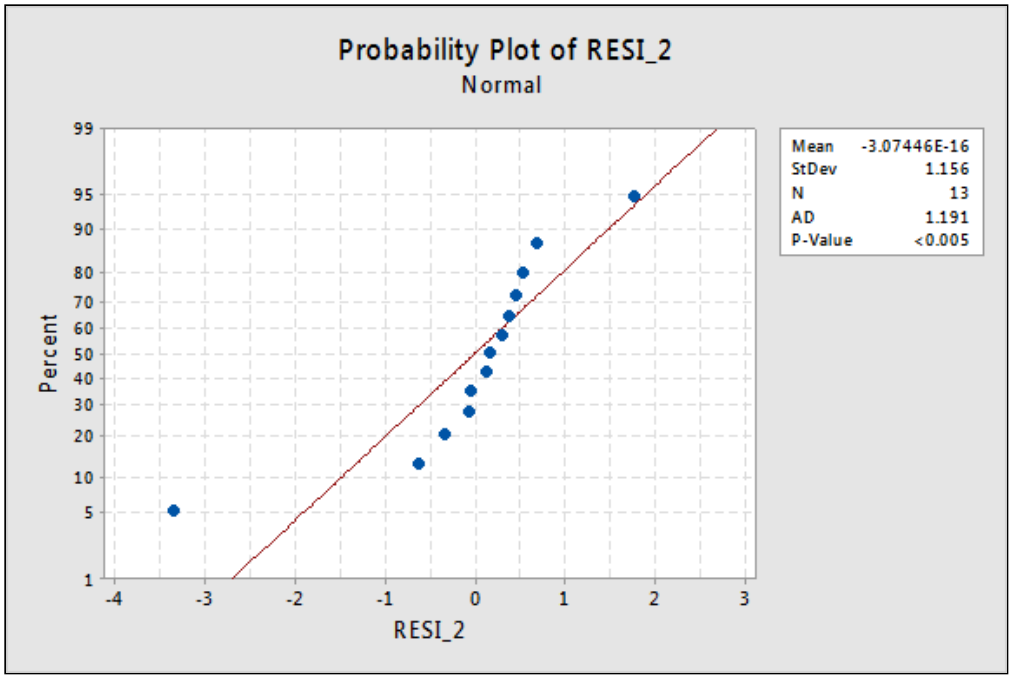


Note that the R^2 value has increased from 57.1% to 96.4%.

The residuals show an improvement with respect to non-linearity, although the improvement is not great...



...but now we have non-normal error terms! The Anderson-Darling *P*-value is less than 0.005, so we reject the null hypothesis of normal error terms. There is sufficient evidence to conclude that the error terms are not normal:



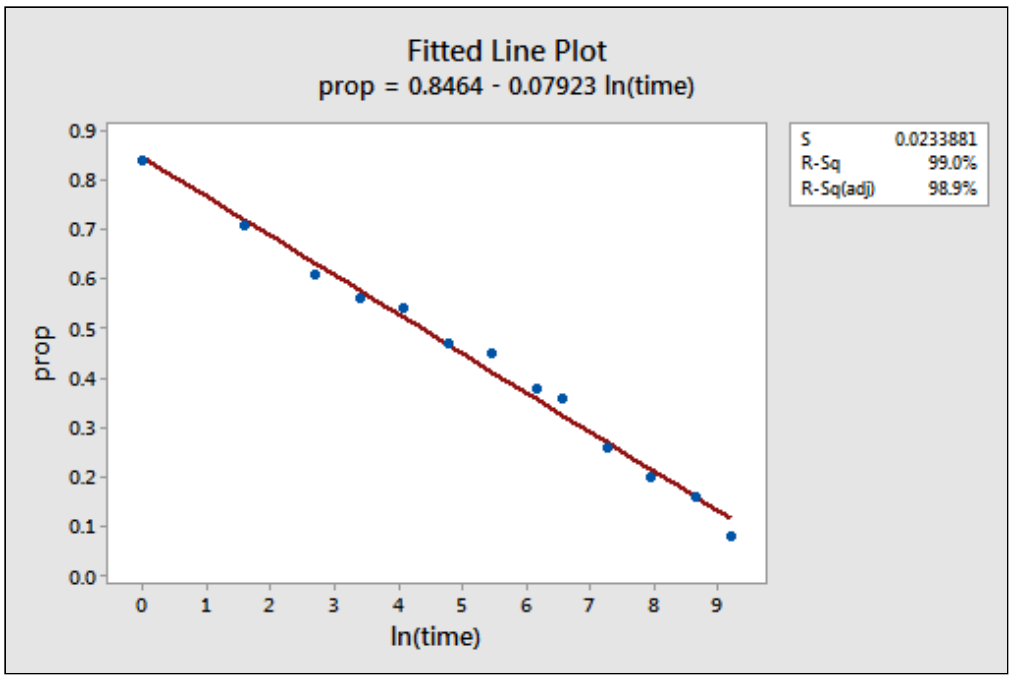
Again, if the error terms are well-behaved prior to transformation, transforming the y values can change them into badly-behaved error terms.

Using the model

Once we've found the best model for our regression data, we can then use the model to answer our research questions of interest. If our model involves transformed predictor (x) values, we may or may not have to make slight modifications to the standard procedures we've already learned.

Let's use our linear regression model for the memory retention data—with $y = prop$ as the response and $x = lntime$ as the predictor—to answer four different research questions.

Research Question #1: What is the nature of the association between time since memorized and the effectiveness of recall?



To answer this research question, we just describe the nature of the relationship. That is, the proportion of correctly recalled words is negatively linearly related to the natural log of the time since the words were memorized. Not surprisingly, as the natural log of time increases, the proportion of recalled words decreases.

Research Question #2: Is there an association between time since memorized and effectiveness of recall?

In answering this research question, no modification to the standard procedure is necessary. We merely test the null hypothesis H_0: $\beta_1 = 0$ using either the F-test or the equivalent t-test:

```
The regression equation is
prop = 0.846 - 0.0792 lntime

Predictor        Coef     SE Coef          T        P
Constant      0.84642     0.01419      59.63    0.000
lntime      -0.079227    0.002416     -32.80    0.000

S = 0.02339     R-Sq = 99.0%     R-Sq(adj) = 98.9%

Analysis of Variance
Source           DF      SS        MS        F       P
Regression        1   0.58841   0.58841  1075.70   0.000
Residual Error   11   0.00602   0.00055
Total            12   0.59443
```

As the software output illustrates, the P-value is < 0.001. There is significant evidence at the 0.05 level to conclude that there is a linear association between the proportion of words recalled and the natural log of the time since memorized.

Research Question #3: What proportion of words can we expect a randomly selected person to recall after 1000 minutes?

We just need to calculate a prediction interval — with one slight modification — to answer this research question. Our predictor variable is the natural log of time. Therefore, when we use statistical software to calculate the prediction interval, we have to make sure that we specify the value of the predictor values *in the transformed units*, not the original units. The natural log of 1000 minutes is 6.91 log-minutes. Using software to calculate a 95% prediction interval when *lntime* = 6.91, we obtain:

```
Predicted Values for New Observations

New Obs    Fit     SE Fit       95.0% CI          95.0% PI
1       0.29896   0.00766  (0.282, 0.316)  (0.245, 0.353)

Values of Predictors for New Observations

New Obs    lntime
1            6.91
```

The output tells us that we can be 95% confident that, after 1000 minutes, a randomly selected person will recall between 24.5% and 35.3% of the words.

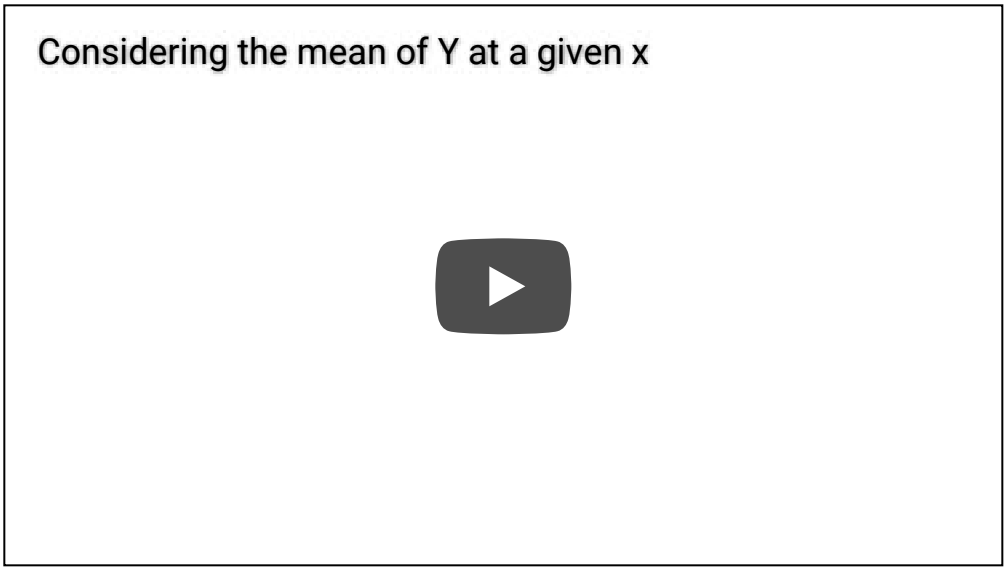
Research Question #4: How much does the expected recall change if time increases ten-fold?

If you think about it, answering this research question merely involves estimating and interpreting the slope parameter β_1. Well, not quite—there is a slight adjustment. In general, a k-fold increase in the predictor x is associated with a:

$$\beta_1 \times ln(k)$$

change in the mean of the response y.

This derivation that follows might help you understand and therefore remember this formula.



That is, a ten-fold increase in x is associated with a $\beta_1 \times ln(10)$ change in the mean of y. And, a two-fold increase in x is associated with a $\beta_1 \times ln(2)$ change in the mean of y.

In general, you should only use multiples of k that make sense for the scope of the model. For example, if the x values in your data set range from 2 to 8, it only makes sense to consider k multiples that are 4 or smaller. If the value of x were 2, a ten-fold increase (*i.e.*, $k = 10$) would take you from 2 up to $2 \times 10 = 20$, a value outside the scope of the model. In the memory retention data set, the predictor values range from 1 to 10080, so there is no problem with considering a ten-fold increase.

If we are only interested in obtaining a point estimate, we merely take the estimate of the slope parameter ($b_1 =$ -0.079227) from the software output:

Predictor	Coef	SE Coef	T	P
Constant	0.84642	0.01419	59.63	0.000
lntime	-0.079227	0.002416	-32.80	0.000

and multiply it by $ln(10)$:

$b_1 \times ln(10) = -0.079227 \times ln(10) =$ **-0.182**

We expect the percentage of recalled words to decrease (since the sign is negative) 18.2% for each ten-fold increase in the time since memorization took place.

Of course, this point estimate is of limited usefulness. How confident can we be that the estimate is close to the true unknown value? Naturally, we should calculate a 95% confident interval. To do so, we just calculate a 95% confidence interval for β_1 as we always have:

$-0.079227 \pm 2.201(0.002416) =$ (**-0.085, -0.074**)

and then multiply each endpoint of the interval by $ln(10)$:

-0.074 × $ln(10)$ = **- 0.170** and -0.085 × $ln(10)$ = **- 0.l95**

We can be 95% confident that the percentage of recalled words will decrease between 17.0% and 19.5% for each ten-fold increase in the time since memorization took place.

‹ Lesson 7: Transformations & Interactions (/stat462/node/85) up (/stat462/node/85) 7.2 - Log-transforming Only the Response for SLR › (/stat462/node/153)

STAT 462

Applied Regression Analysis

7.2 - Log-transforming Only the Response for SLR

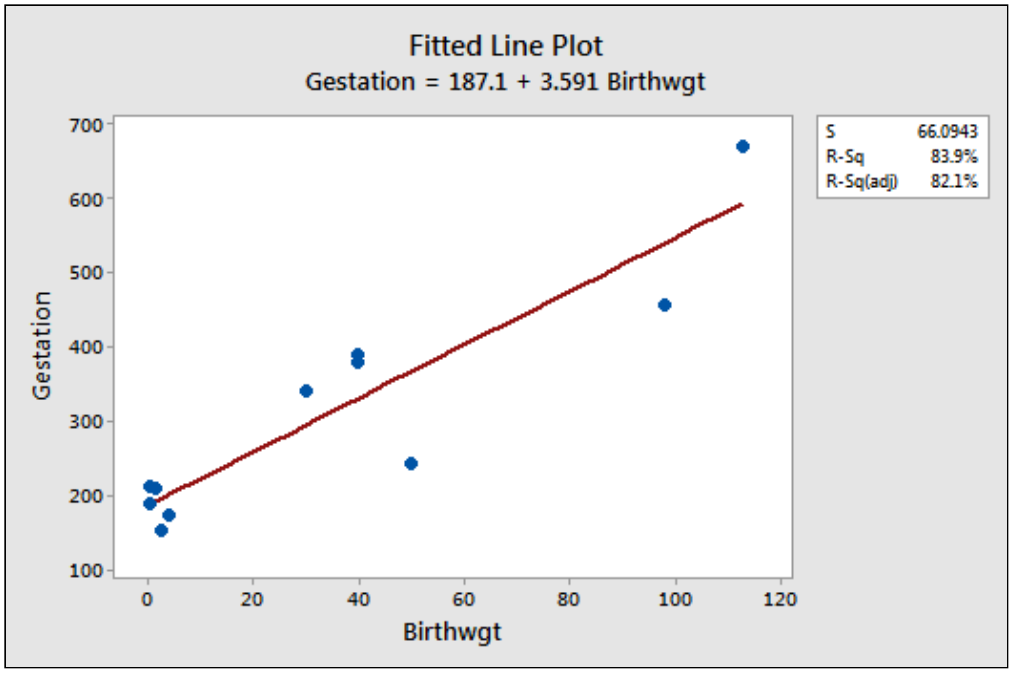
In this section, we learn **how to build** and **use** a model by transforming the response y values. Transforming the y values should be considered when non-normality and/or unequal variances are the problems with the model. As an added bonus, the transformation on y may also help to "straighten out" a curved relationship.

Again, keep in mind that although we're focussing on a simple linear regression model here, the essential ideas apply more generally to multiple linear regression models too.

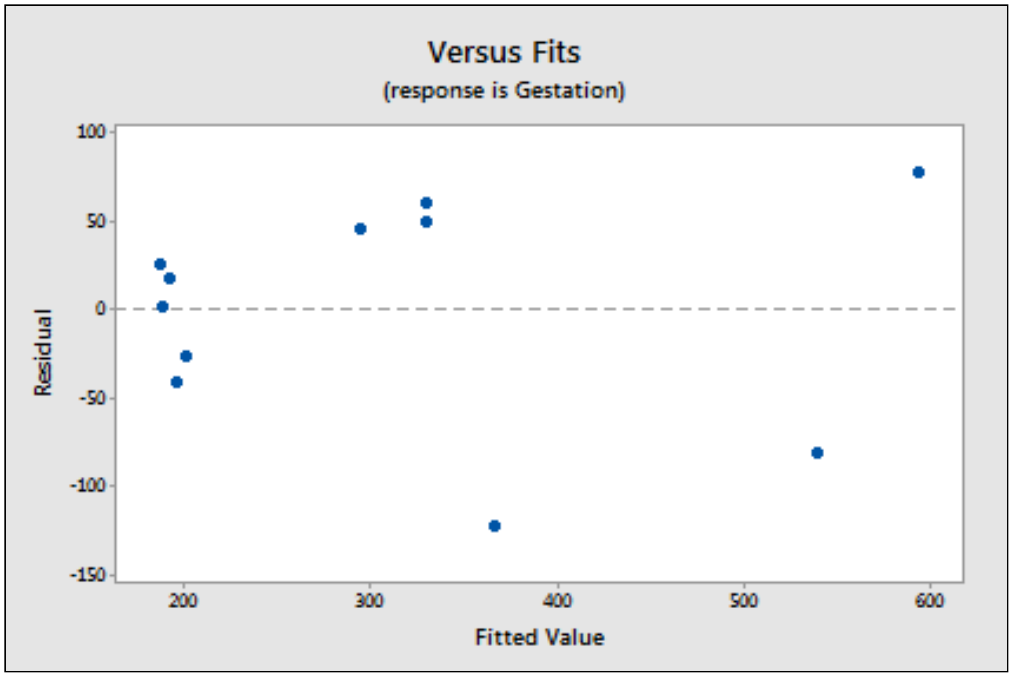
Building the model

An example. Let's consider data (mammgest.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/mammgest.txt)) on the typical birthweight and length of gestation for various mammals. We treat the birthweight (x, in kg) as the predictor and the length of gestation (y, in number of days until birth) as the response.

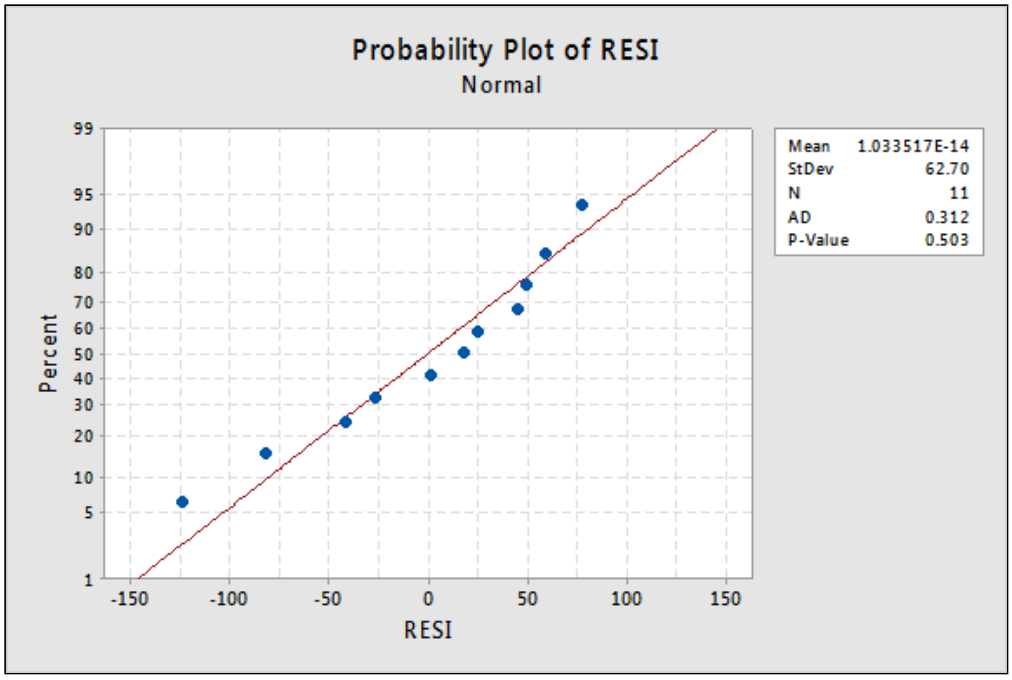
The fitted line plot suggests that the relationship between gestation length (y) and birthweight (x) is linear, but that the variance of the error terms might not be equal:



The residuals vs. fits plot exhibits some fanning and therefore provides yet more evidence that the variance of the



The normal probability plot supports the assumption of normally distributed error terms:



The line is approximately straight and the Anderson-Darling *P*-value is 0.503. We fail to reject the null hypothesis of normal error terms. There is not enough evidence to conclude that the errors terms are not normal.

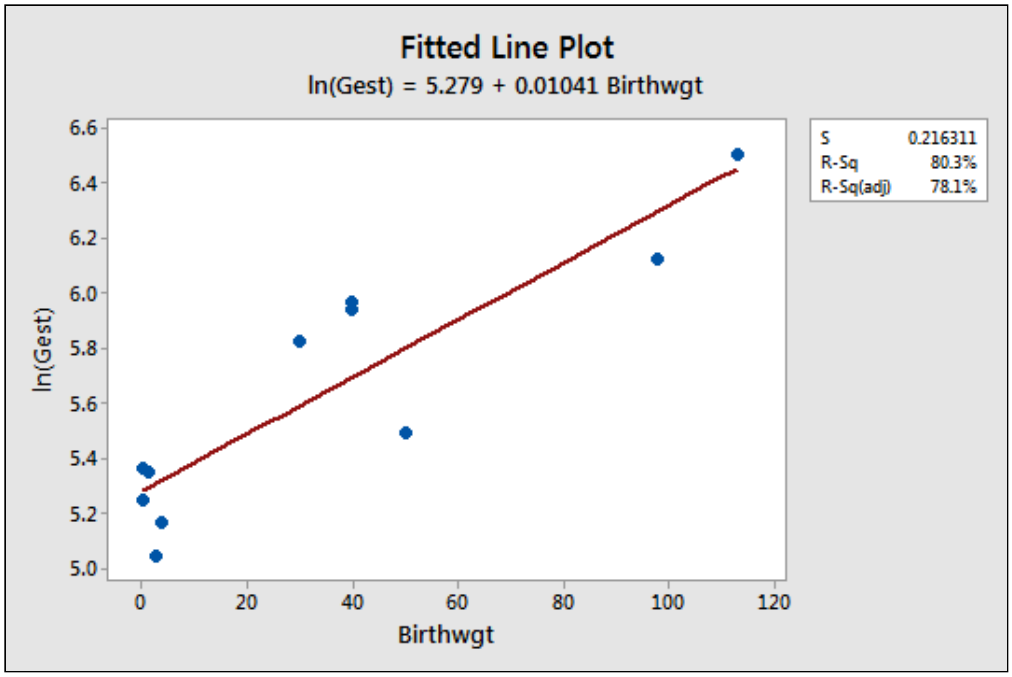
Let's transform the *y* values by taking the natural logarithm of the lengths of gestation. Doing so, we obtain the new response *y* = *lnGest*:

Mammal	***Birthwgt***	***Gestation***	***lnGest***
Goat	2.75	155	5.04343
Sheep	4.00	175	5.16479
		190	5.24702

Porcupine	1.50	210	5.34711
Bear	0.37	213	5.36129
Hippo	50.00	243	5.49306
Horse	30.00	340	5.82895
Camel	40.00	380	5.94017
Zebra	40.00	390	5.96615
Giraffe	98.00	457	6.12468
Elephant	113.00	670	6.50728

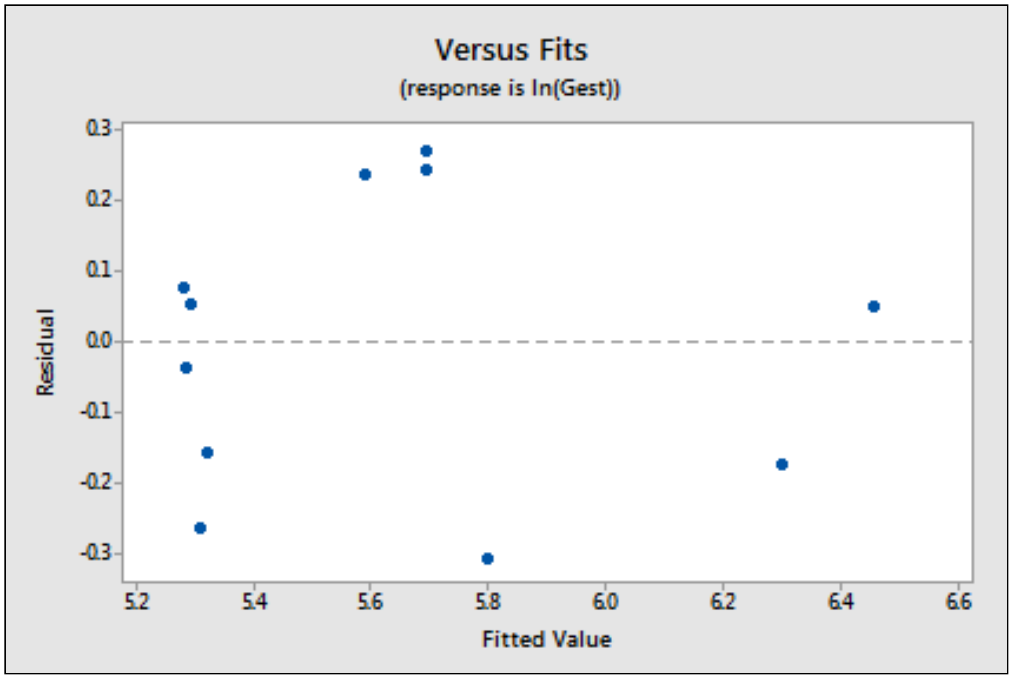
For example, $ln(155) = 5.04343$ and $ln(457) = 6.12468$. Now that we've transformed the response y values, let's see if it helped rectify the problem with the unequal error variances.

The fitted line plot with $y = lnGest$ as the response and $x = Birthwgt$ as the predictor suggests that the log transformation of the response has helped:

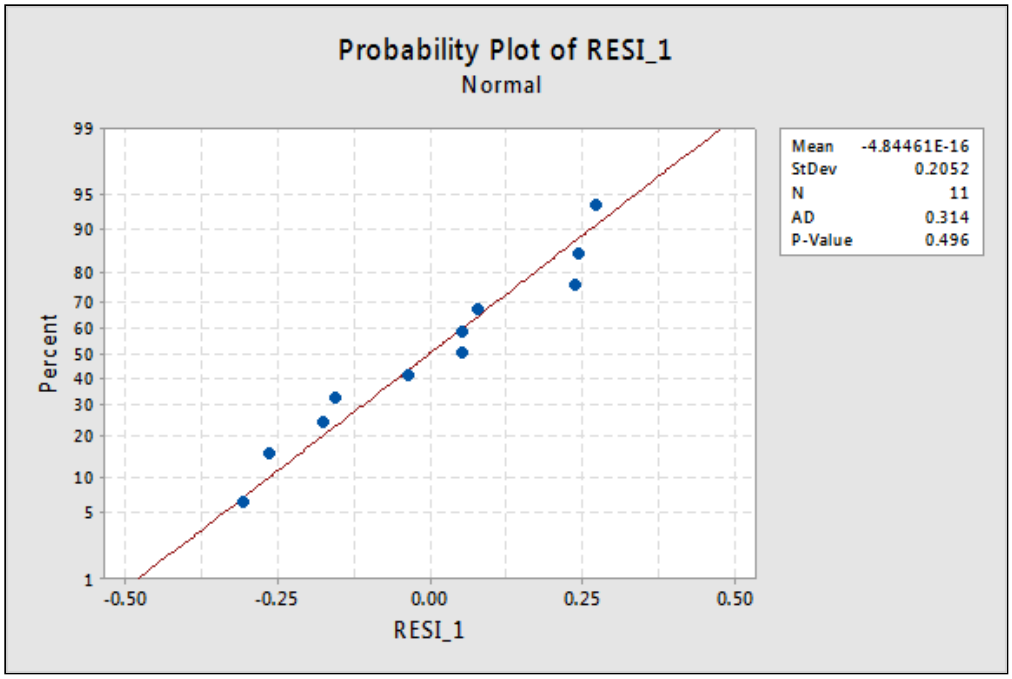


Note that, as expected, the log transformation has tended to "spread out" the smaller gestations and tended to "bring in" the larger ones.

The new residual vs. fits plot shows a marked improvement in the spread of the residuals:



The log transformation of the response did not adversely affect the normality of the error terms:



The line is approximately straight and the Anderson-Darling *P*-value is 0.496. We fail to reject the null hypothesis of normal error terms. There is not enough evidence to conclude that the errors terms are not normal.

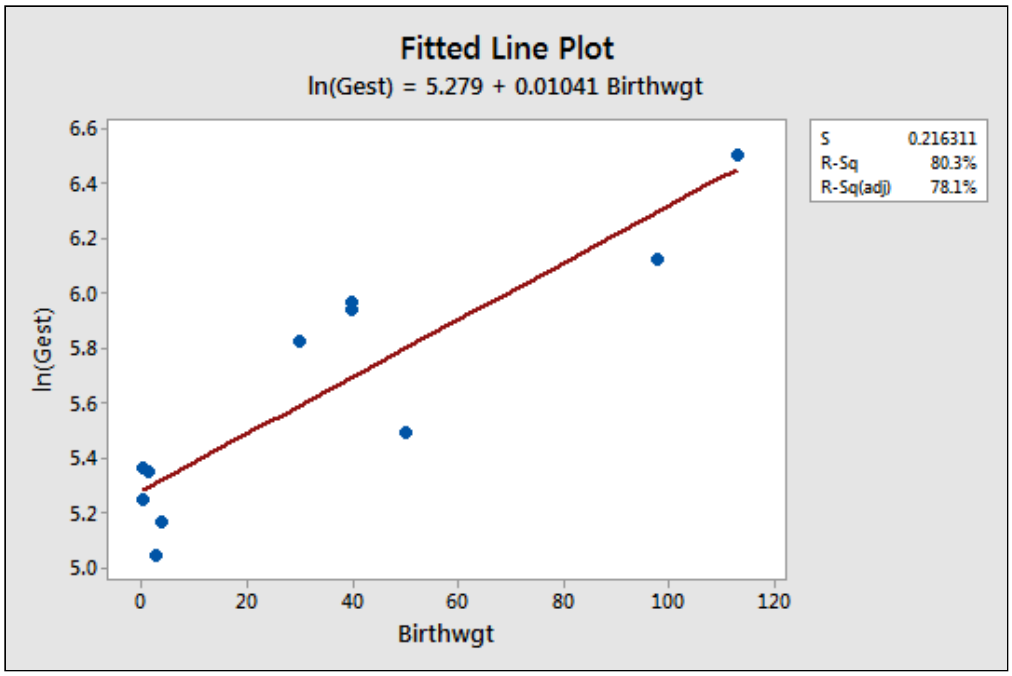
Note that the R^2 value is lower for the transformed model than for the untransformed model (80.3% versus 83.9%). This does *not* mean that the untransformed model is preferable. Remember the untransformed model failed to satisfy the equal variance condition, so we should not use this model anyway.

Again, transforming the *y* values should be considered when non-normality and/or unequal variances are the main problems with the model.

We've identified what we think is the best model for the mammal birthweight and gestation data. The model meets the four "LINE" conditions. Therefore, we can use the model to answer our research questions of interest. We may or may not have to make slight modifications to the standard procedures we've already learned.

Let's use our linear regression model for the mammal birthweight and gestation data—with $y = lnGest$ as the response and $x = birthwgt$ as the predictor—to answer four different research questions.

Research Question #1: What is the nature of the association between mammalian birth weight and length of gestation?



Again, to answer this research question, we just describe the nature of the relationship. That is, the natural logarithm of the length of gestation is positively linearly related to birthweight. That is, as the average birthweight of the mammal increases, the expected natural logarithm of the gestation length also increases.

Research Question #2: Is there an association between mammalian birth weight and length of gestation?

Again, in answering this research question, no modification to the standard procedure is necessary. We merely test the null hypothesis H_0: $\beta_1 = 0$ using either the F-test or the equivalent t-test:

```
The regression equation is
lnGest = 5.28 + 0.0104 Birthwgt

Predictor        Coef     SE Coef          T        P
Constant      5.27882     0.08818      59.87    0.000
Birthwgt     0.010410    0.001717       6.06    0.000

S = 0.2163      R-Sq = 80.3%     R-Sq(adj) = 78.1%

Analysis of Variance
Source           DF      SS        MS       F       P
Regression        1   1.7193    1.7193   36.75   0.000
Residual Error    9   0.4211    0.0468
Total            10   2.1405
```

As the software output illustrates, the *P*-value is < 0.001. There is significant evidence at the 0.05 level to conclude that there is a linear association between the mammalian birthweight and the natural logarithm of the length of gestation.

Research Question #3: What is the expected gestation length of a new 50 *kg* mammal?

In answering this research question, if we are only interested in obtaining a point estimate, we merely enter $x = 50$ into the estimated regression equation:

$$\widehat{ln(Gest)} = 5.28 + 0.0104 \times Birthwgt$$

to obtain:

$$\widehat{ln(Gest)} = 5.28 + 0.0104 \times 50 = 5.8$$

That is, we predict the length of gestation of a 50 *kg* mammal to be 5.8 log-days! Well, that's not very informative! We need to transform the answer back into the original units. This just requires recalling one of the fundamental properties of the natural logarithm, namely that e^x and $\ln(x)$ "cancel each other out." That is:

$$\widehat{Gest} = e^{\widehat{ln(Gest)}}$$

Furthermore, if we exponentiate the left side of the equation:

$$\widehat{ln(Gest)} = 5.28 + 0.0104 \times 50 = 5.8$$

we also have to exponentiate the right side of the equation. Doing so, we obtain:

$$\widehat{Gest} = e^{\widehat{ln(Gest)}} = e^{5.8} = 330.3$$

We predict the gestation length of a 50 *kg* mammal to be 330 days. That sounds better!

Again, a point estimate is of limited usefulness. It doesn't tells us how confident we can be that the prediction is close to the true unknown value. We should calculate a 95% prediction interval:

```
Predicted Values for New Observations

New   Fit     SE Fit      95.0% CI           95.0% PI
1   5.7993   0.0704   (5.6401, 5.9586)  (5.2847, 6.3139)

Values of Predictors for New Observations

New Obs  Birthwgt
1           50.0
```

So, we can be 95% confident that the gestation length of a 50 *kg* mammal is predicted to be between 5.2847 and 6.3139 log-days! Again, we need to transform these predicted limits back into the original units. Doing so, we obtain:

$$e^{5.2847} = 197.3 \text{ and } e^{6.3139} = 552.2$$

We can be 95% confident that the gestation length for a 50 *kg* mammal will be between 197.3 and 552.2 days.

is the expected change in length of gestation for each one pound increase in birth weight?

Figuring out how to answer this research question takes a little bit of work — and some creativity, too! If you only care about the end result, this is it:

- The median of the response changes by a factor of e^{β_1} for each one unit increase in the predictor x. Although you won't be required to duplicate the derivation, it might help you understand—and therefore remember—the result.

- And, therefore, the median of the response changes by a factor of $e^{k\beta_1}$ for each k-unit increase in the predictor x. Again, although you won't be required to duplicate the derivation, it might help you understand—and therefore remember—the result.

- As always, we won't know the slope of the population line, β_1, so we'll have to use b_1 to estimate it.

For the mammalian birthweight and gestation data, the software output tells us that $b_1 = 0.01041$:

Predictor	Coef	SE Coef	T	P
Constant	5.27882	0.08818	59.87	0.000
Birthwgt	0.010410	0.001717	6.06	0.000

and therefore:

Loading [MathJax]/extensions/MathMenu.js

$$e^{b_1} = e^{0.01041} = 1.0105$$

The result tells us that the predicted median gestation changes by a factor of 1.0105 for each one unit increase in birthweight. For example, the predicted median gestation for a mammal weighing 3 *kgs* is 1.0105 times the median gestation for a mammal weighing 2 *kgs*. And, since there is a 10-unit increase going from a 20 *kg* to a 30 *kg* mammal, the median gestation for a mammal weighing 30 *kgs* is $1.0105^{10} = 1.110$ times the median gestation for a mammal weighing 20 *kgs*.

So far, we've only calculated a point estimate for the expected change. Of course, a 95% confidence interval for β_1 is:

$$0.01041 \pm 2.2622(0.001717) = (0.0065, 0.0143)$$

Because:

$$e^{0.0065} = 1.0065 \text{ and } e^{0.0143} = 1.0144$$

we can be 95% confident that the median gestation will increase by a factor between 1.0065 and 1.0144 for each one kilogram increase in birth weight. And, since:

$$1.0065^{10} = 1.067 \text{ and } 1.0144^{10} = 1.154$$

we can be 95% confident that the median gestation will increase by a factor between 1.067 and 1.154 for each 10-kilogram increase in birth weight.

‹ 7.1 - Log-transforming Only the Predictor for SLR (/stat462/node/152) | up (/stat462/node/85) | 7.3 - Log-transforming Both the Predictor and Response for SLR › (/stat462/node/154)

STAT 462

Applied Regression Analysis

7.3 - Log-transforming Both the Predictor and Response for SLR

In this section, we learn **how to build** and **use** a model by transforming both the response y and the predictor x. You might have to do this when everything seems wrong — when the regression function is **not linear** and the error terms are **not normal** and have **unequal variances**. In general (although not always!):

- Transforming the y values corrects problems with the error terms (and may help the non-linearity).
- Transforming the x values primarily corrects the non-linearity.

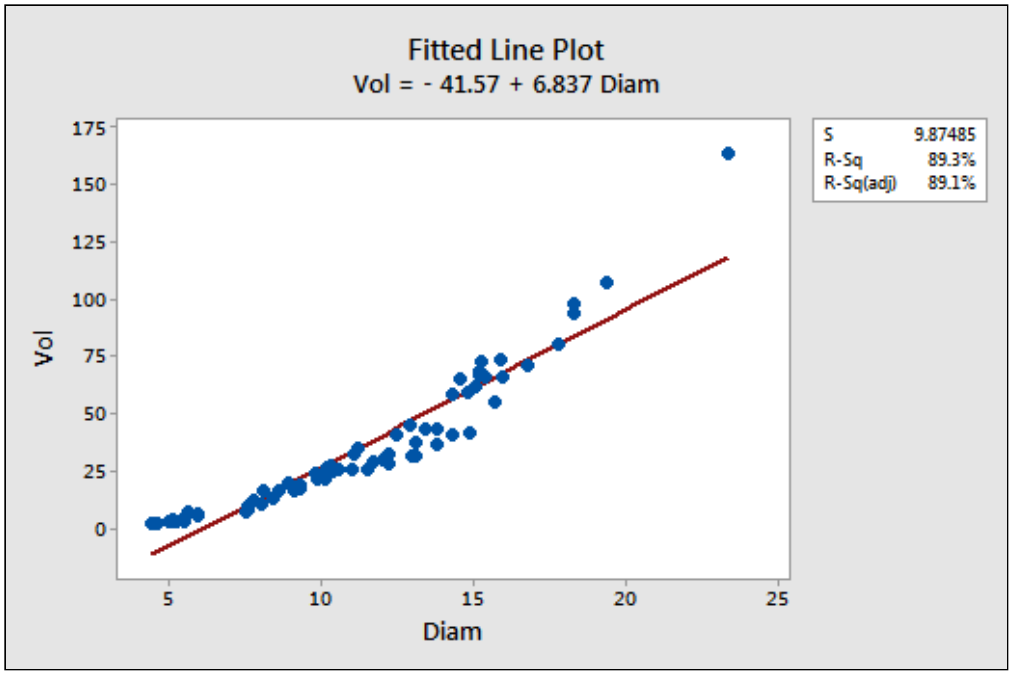
Again, keep in mind that although we're focussing on a simple linear regression model here, the essential ideas apply more generally to multiple linear regression models too.

As before, let's learn about transforming both the x and y values by way of example.

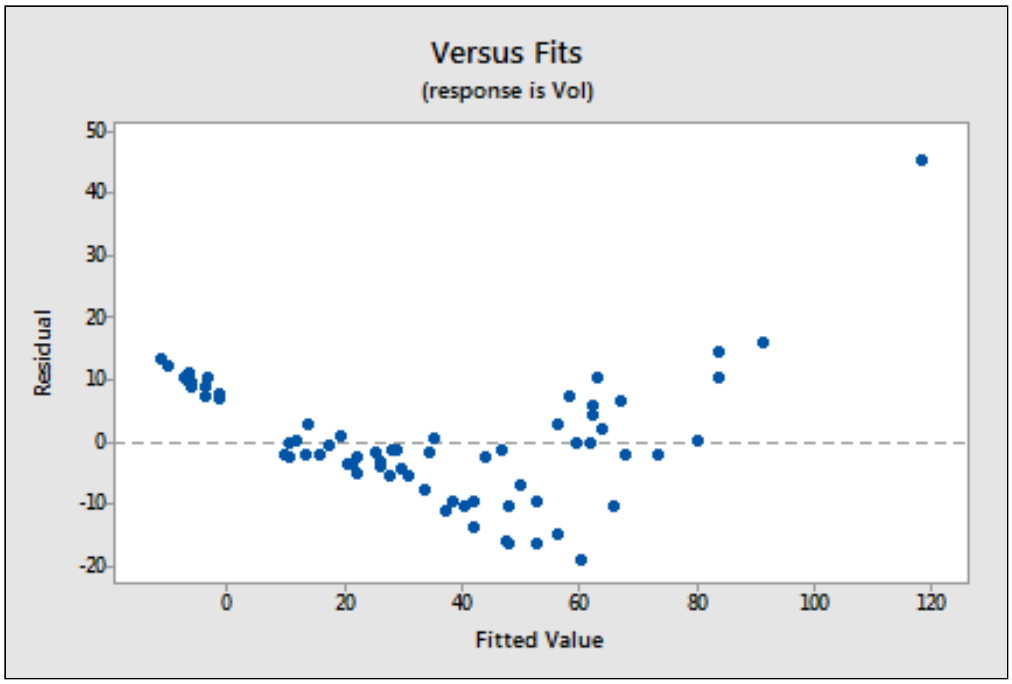
Building the model

An example. Many different interest groups — such as the lumber industry, ecologists, and foresters — benefit from being able to predict the volume of a tree just by knowing its diameter. One classic data set (shortleaf.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/shortleaf.txt)) — reported by C. Bruce and F. X. Schumacher in 1935 — concerned the diameter (x, in inches) and volume (y, in cubic feet) of $n = 70$ shortleaf pines. Let's use the data set to learn not only about the relationship between the diameter and volume of shortleaf pines, but also about the benefits of simultaneously transforming both the response y and the predictor x.

Although the r^2 value is quite high (89.3%), the fitted line plot suggests that the relationship between tree volume and tree diameter is not linear:

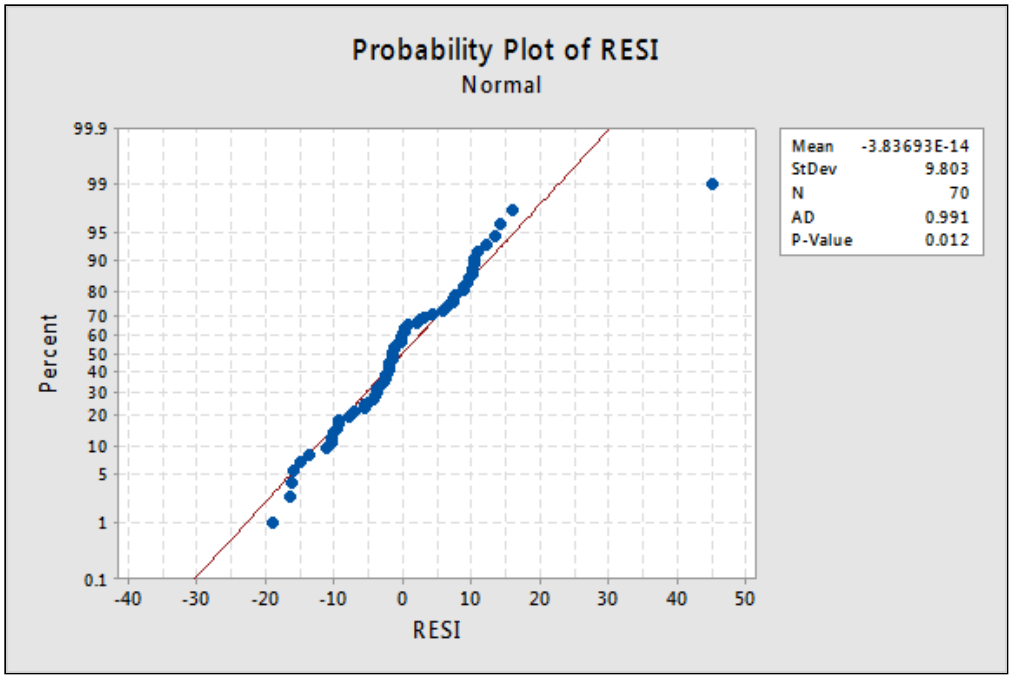


The residuals vs. fits plot also suggests that the relationship is not linear:



Because the lack of linearity dominates the plot, we can not use the plot to evaluate whether the error variances are equal. We have to fix the non-linearity problem before we can assess the assumption of equal variances.

The normal probability plot suggests that the error terms are not normal. The plot is not quite linear and the Anderson-Darling *P*-value is 0.012. There is sufficient evidence to conclude that the error terms are not normally distributed:



The plot actually has the classical appearance of residuals that are predominantly normal but have one outlier. This illustrates how a data point can be deemed an "outlier" just because of poor model fit.

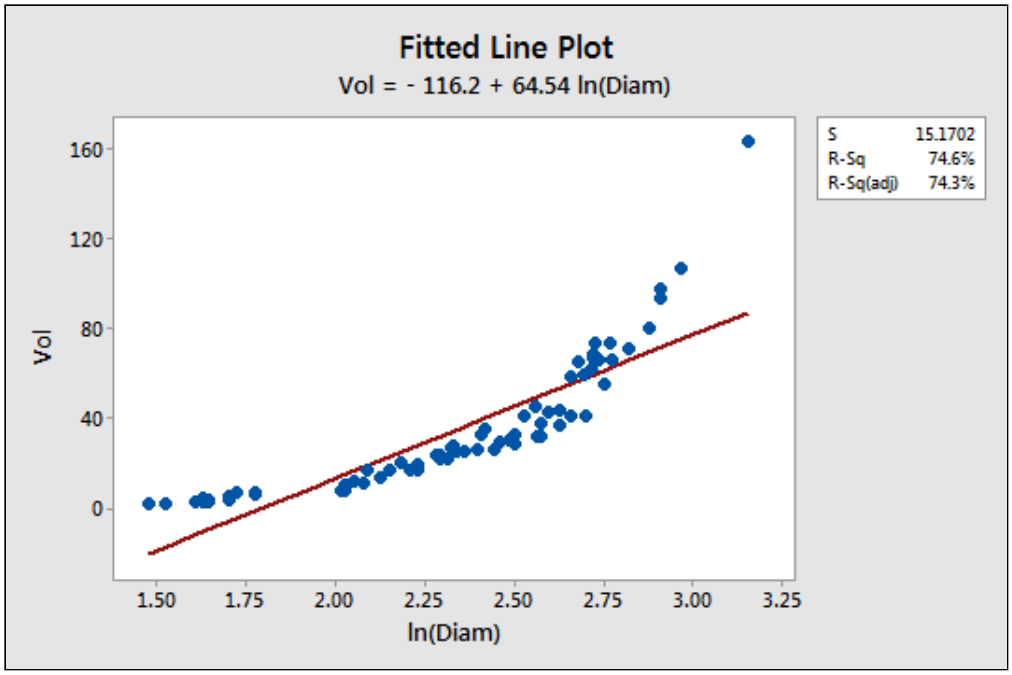
In summary, it appears as if the relationship between tree diameter and volume is not linear. Furthermore, it appears as if the error terms are not normally distributed.

Let's see if we get anywhere by transforming only the x values. In particular, let's take the natural logarithm of the tree diameters to obtain the new predictor $x = lnDiam$:

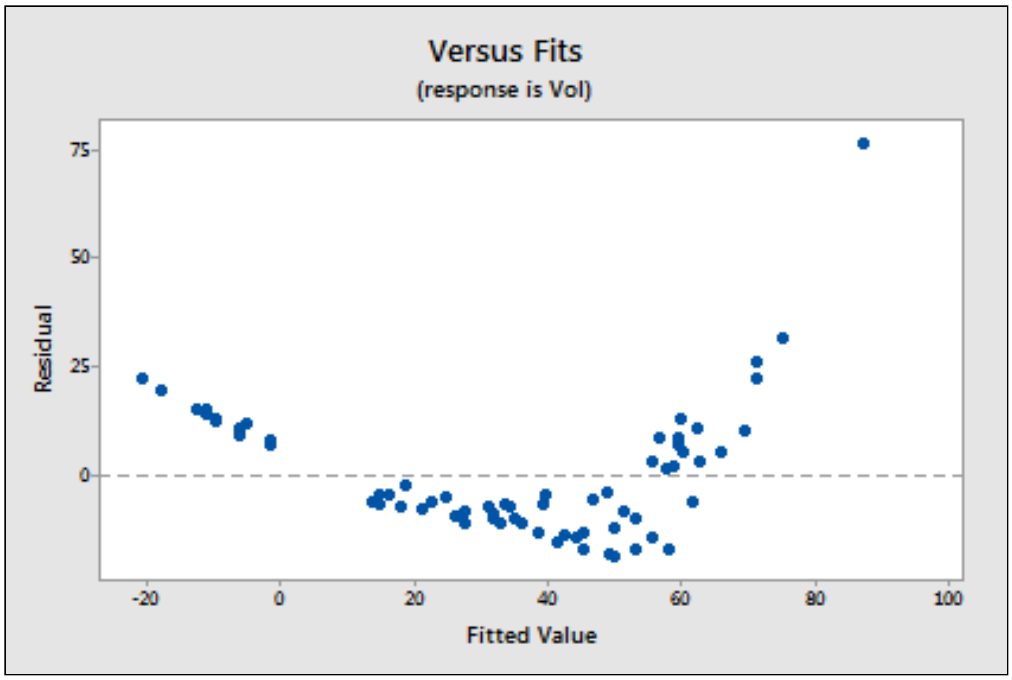
Diameter	*Volume*	*lnDiam*
4.4	2.0	1.48160
4.6	2.2	1.52606
5.0	3.0	1.60944
5.1	4.3	1.62924
5.1	3.0	1.62924
5.2	2.9	1.64866
5.2	3.5	1.64866
5.5	3.4	1.70475
5.5	5.0	1.70475
5.6	7.2	1.72277
5.9	6.4	1.77495
5.9	5.6	1.77495
7.5	7.7	2.01490
7.6	10.3	2.02815

For example, $ln(5.0) = 1.60944$ and $ln(7.6) = 2.02815$. How well does transforming only the x values work? Not very!

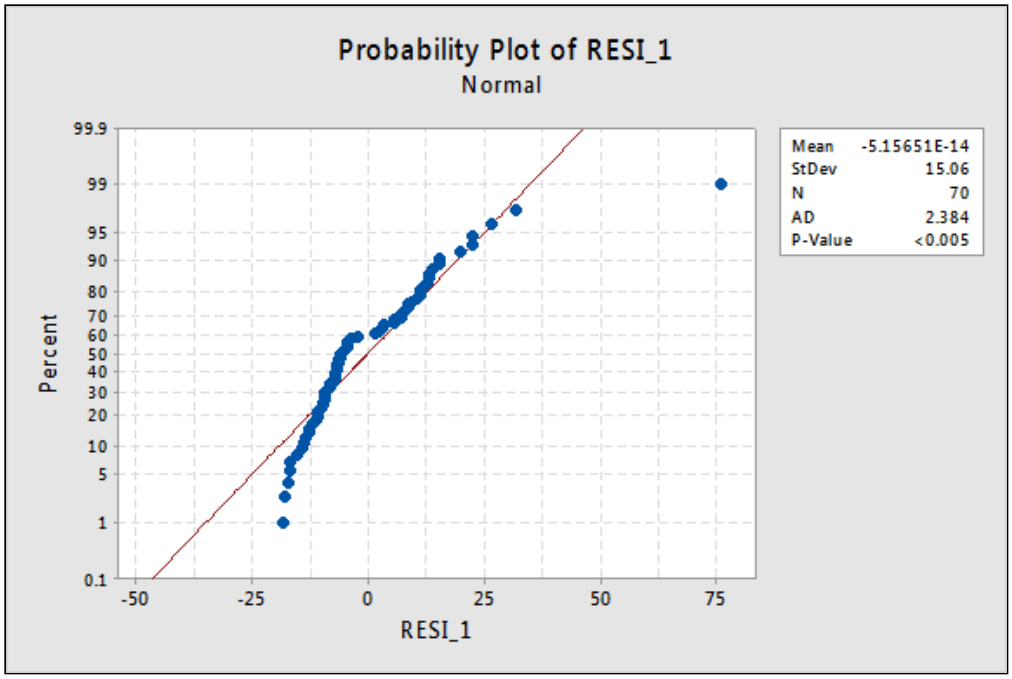
The fitted line plot with $y = volume$ as the response and $x = lnDiam$ as the predictor suggests that the relationship is still not linear:



Transforming only the x values didn't change the non-linearity at all. The residuals vs. fits plot also still suggests a non-linear relationship ...



... and there is little improvement in the normality of the error terms:

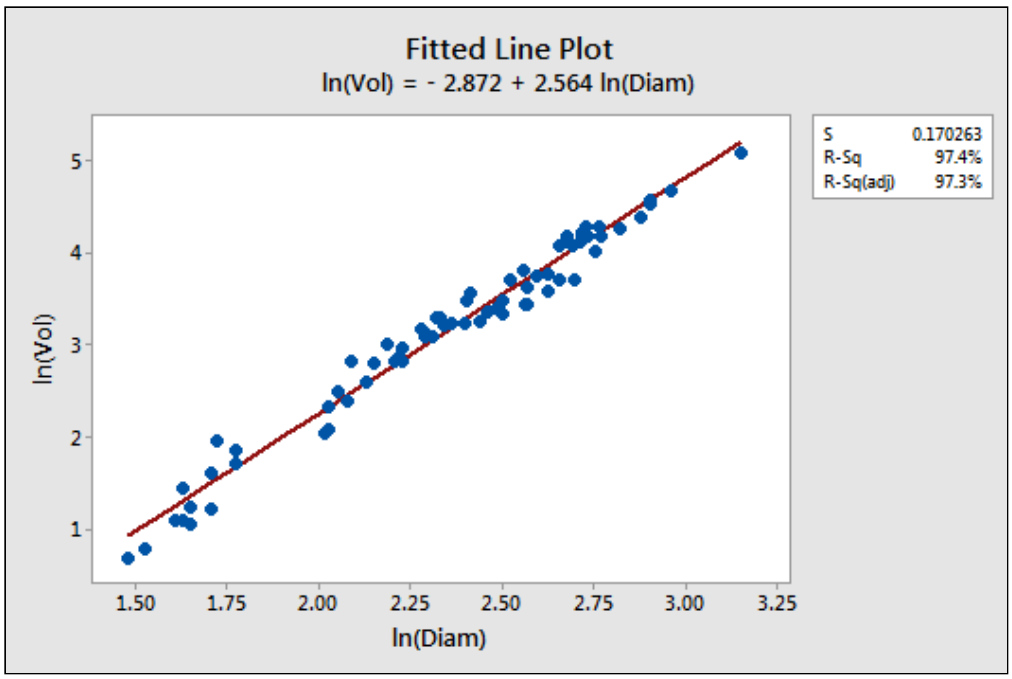


The pattern is not linear and the Anderson-Darling *P*-value is less than 0.005. There is sufficient evidence to conclude that the error terms are not normally distributed.

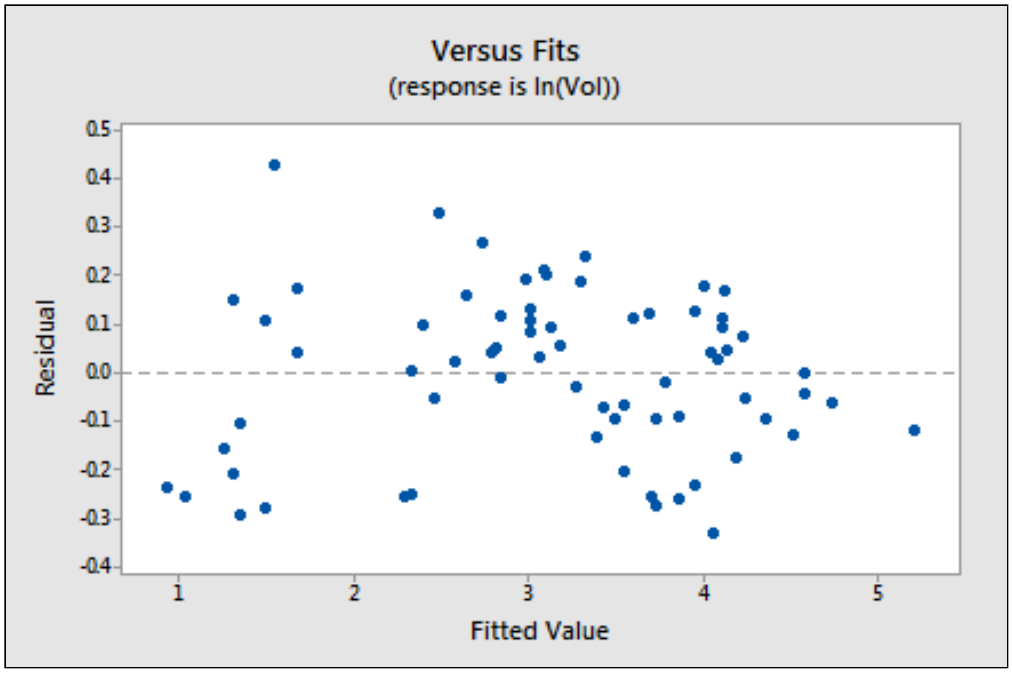
So, transforming *x* alone didn't help much. Let's also try transforming the response (*y*) values. In particular, let's take the natural logarithm of the tree volumes to obtain the new response *y* = *lnVol*:

Diameter	*Volume*	*lnDiam*	*lnVol*
4.4	2.0	1.48160	0.69315
4.6	2.2	1.52606	0.78846
5.0	3.0	1.60944	1.09861
5.1	4.3	1.62924	1.45862
5.1	3.0	1.62924	1.09861
5.2	2.9	1.64866	1.06471
5.2	3.5	1.64866	1.25276
5.5	3.4	1.70475	1.22378
5.5	5.0	1.70475	1.60944
5.6	7.2	1.72277	1.97408
5.9	6.4	1.77495	1.85630
5.9	5.6	1.77495	1.72277
7.5	7.7	2.01490	2.04122
7.6	10.3	2.02815	2.33214
... and so on ...			

Let's see if transforming both the x and y values does it for us. Wow! The fitted line plot should give us hope! The relationship between the natural log of the diameter and the natural log of the volume looks linear and strong (r^2 = 97.4%):

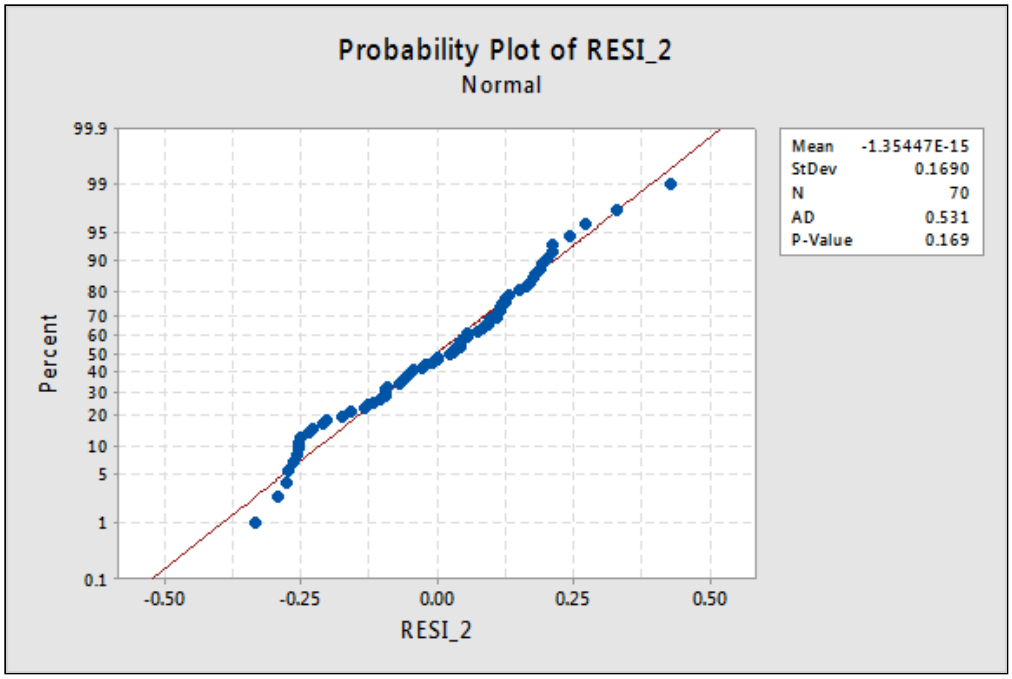


The residuals vs. fits plot provides yet more evidence of a linear relationship between *lnVol* and *lnDiam*:



Generally, speaking the residuals bounce randomly around the residual = 0 line. You might be a little concerned that some "funneling" exists. If it does, it doesn't appear to be too severe, as the negative residuals do follow the desired horizontal band.

The normal probability plot has improved substantially:



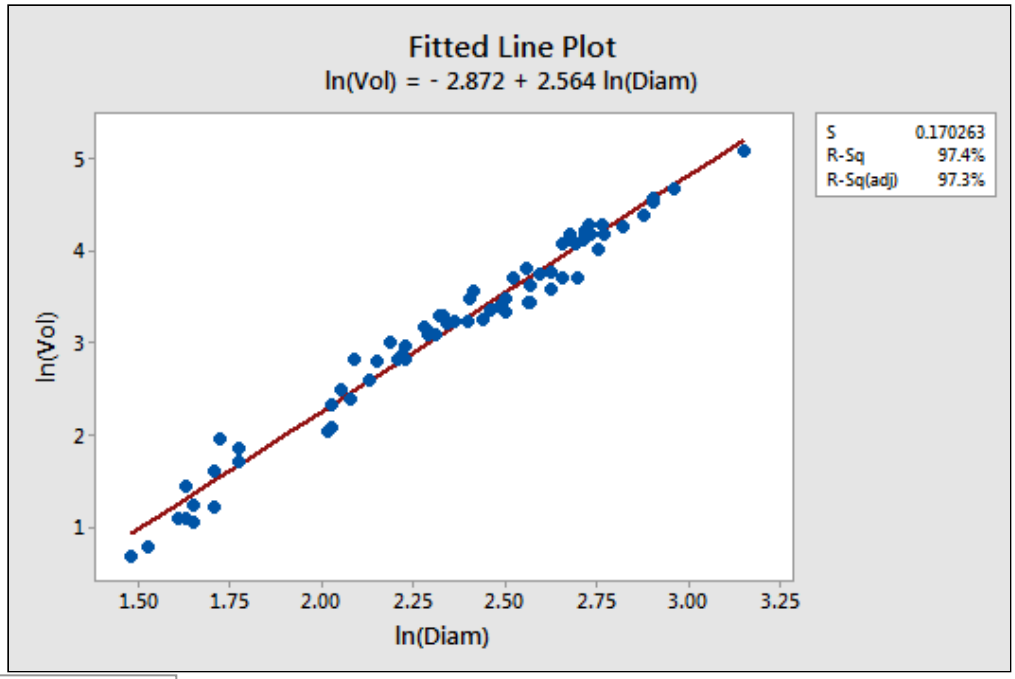
The trend is generally linear and the Anderson-Darling *P*-value is 0.169. There is insufficient evidence to conclude that the error terms are not normal.

In summary, it appears as if the model with the natural log of tree volume as the response and the natural log of tree diameter as the predictor works well. The relationship appears to be linear and the error terms appear independent and normally distributed with equal variances.

Using the model

Let's now use our linear regression model for the shortleaf pine data—with $y = lnVol$ as the response and $x = lnDiam$ as the predictor—to answer four different research questions.

Research Question #1: What is the nature of the association between diameter and volume of shortleaf pines?



Again, to answer this research question, we just describe the nature of the relationship. That is, the natural logarithm of tree volume is positively linearly related to the natural logarithm of tree diameter. That is, as the natural log of tree diameters increases, the average natural logarithm of the tree volume also increases.

Research Question #2: Is there an association between diameter and volume of shortleaf pines?

Again, in answering this research question, no modification to the standard procedure is necessary. We merely test the null hypothesis H_0: $\beta_1 = 0$ using either the F-test or the equivalent t-test:

```
The regression equation is
lnVol = - 2.87 + 2.56 lnDiam

Predictor        Coef     SE Coef          T        P
Constant      -2.8718      0.1215     -23.63    0.000
lnDiam        2.56442     0.05120      50.09    0.000

S = 0.1703      R-Sq = 97.4%     R-Sq(adj) = 97.3%

Analysis of Variance
Source           DF      SS      MS         F        P
Regression        1  72.734  72.734   2509.00    0.000
Residual Error   68   1.971   0.029
Total            69  74.706
```

As the software output illustrates, the P-value is < 0.001. There is significant evidence at the 0.01 level to conclude that there is a linear association between the natural logarithm of tree volume and the natural logarithm of tree diameter.

Research Question #3: What is the "average" volume of all shortleaf pine trees that are 10" in diameter?

In answering this research question, if we are only interested in a point estimate, we put $x = ln(10) = 2.303$ into the estimated regression function:

$$ln(\text{Vol}) = -2.8718 + 2.56442\ ln(\text{Diam})$$

to obtain:

$$ln(\text{Vol}) = -2.8718 + 2.56442 \times ln(10) = 3.034$$

That is, we estimate the average of the natural log of the volumes of all 10"-diameter shortleaf pines to be 3.034 log-cubic feet. Of course, this is not a very helpful conclusion. We have to take advantage of the fact, as we showed before, that the average of the natural log of the volumes approximately equals the natural log of the median of the volumes. Exponentiating both sides of the previous equation:

$$\text{Vol} = e^{ln(\text{Vol})} = e^{3.034} = 20.8 \text{ cubic feet}$$

we estimate the median volume of all shortleaf pines with a 10" diameter to be 20.8 cubic feet. Helpful, but not sufficient! A 95% confidence interval for the average of the natural log of the volumes of all 10"-diameter shortleaf pines is:

```
Predicted Values for New Observations

New  Fit     SE Fit       95.0% CI          95.0% PI
1    3.0330  0.0204   (2.9922, 3.0738)  (2.6908, 3.3752)

Values of Predictors for New Observations

New Obs    lnDiam
1            2.30
```

Exponentiating both endpoints of the interval, we get:

$$e^{2.9922} = 19.9 \text{ and } e^{3.0738} = 21.6.$$

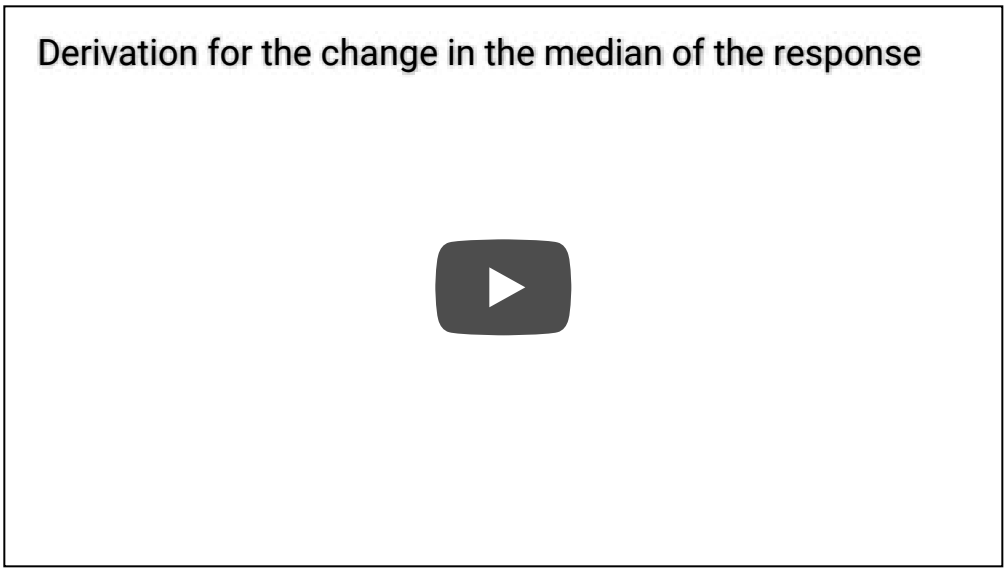
We can be 95% confident that the median volume of all shortleaf pines, 10" in diameter, is between 19.9 and 21.6 cubic feet.

Research Question #4: What is expected change in volume for a two-fold increase in diameter?

Figuring out how to answer this research question also takes a little bit of work. The end result is:

- In general, the median changes by a factor of k^{β_1} for each k-fold increase in the predictor x.
- Therefore, the median changes by a factor of 2^{β_1} for each two-fold increase in the predictor x.
- As always, we won't know the slope of the population line, β_1. We have to use b_1 to estimate it.

Again, you won't be required to duplicate the derivation, shown below, of this result, but it may help you to understand it and therefore remember it.



For the shortleaf pine data, the software output tells us that $b_1 = 2.56442$:

```
Predictor      Coef     SE Coef          T        P
Constant    -2.8718      0.1215    -23.63    0.000
lnDiam       2.56442     0.05120    50.09     0.000
```

and therefore:

Loading [MathJax]/extensions/MathZoom.js

$$2^{b_1} = 2^{2.56442} = 5.92$$

The result tells us that the estimated median volume changes by a factor of 5.92 for each two-fold increase in diameter. For example, the median volume of a 20"-diameter tree is estimated to be 5.92 times the median volume of a 10" diameter tree. And, the median volume of a 10"-diameter tree is estimated to be 5.92 times the median volume of a 5"-diameter tree.

So far, we've only calculated a point estimate for the expected change. Of course, a 95% confidence interval for β_1 is:

$$2.56442 \pm 1.9955(0.05120) = (2.462, 2.667)$$

Because:

$$2^{2.462} = 5.51 \text{ and } 2^{2.667} = 6.35$$

we can be 95% confident that the median volume will increase by a factor between 5.51 and 6.35 for each two-fold increase in diameter.

‹ 7.2 - Log-transforming Only the Response for SLR (/stat462/node/153) | up (/stat462/node/85) | 7.4 - Other Data Transformations › (/stat462/node/155)

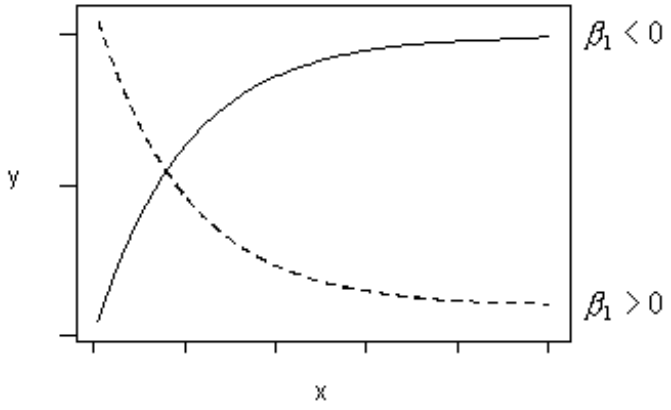
STAT 462

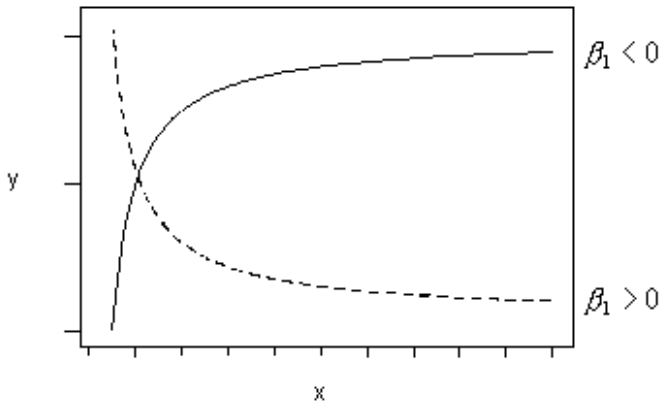
Applied Regression Analysis

7.4 - Other Data Transformations

Is the natural log transformation the only transformation available to you? The answer is no—it just happens to be the only transformation we have investigated so far. We'll try to take care of any misconceptions about this issue in this section, in which we briefly enumerate other transformations you could try in an attempt to correct problems with your model. One thing to keep in mind though is that transforming your data almost always involves lots of trial and error. That is, there are no cut-and-dried recipes. Therefore, the best we can do is offer advice and hope that you find it helpful!

First piece of advice. If the primary problem with your model is non-linearity, look at a scatter plot of the data to suggest transformations that might help. (This only works for simple linear regression models with a single predictor. For multiple linear regression models, look at residual plots instead.) Remember, it is possible to use transformations other than logarithms:

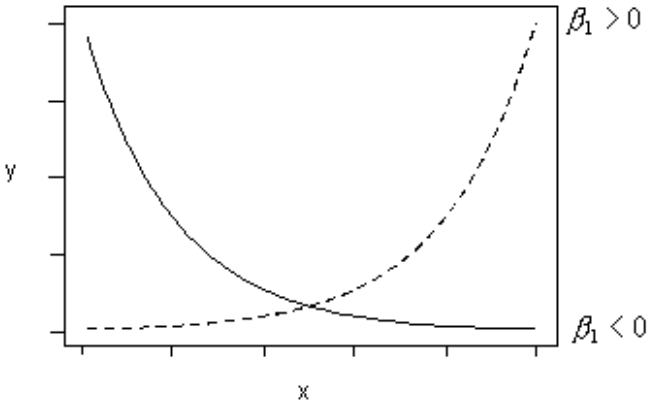
If the trend in your data follows either of these patterns, you could try fitting this regression function: $\mu_Y = \beta_0 + \beta_1 e^{-x}$ to your data.	
Or, if the trend in your data follows either of these patterns, you could try fitting this regression function: $\mu_Y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$ to your data. (This is sometimes transformation.)	



Or, if the trend in your data follows either of these patterns, try fitting this regression function:

$\mu_{lnY} = \beta_0 + \beta_1 x$

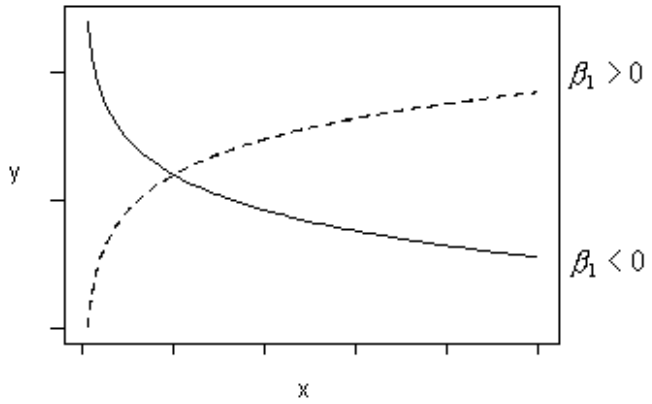
to your data. That is, fit the model with *ln*(*y*) as the response and *x* as the predictor.



Or, try fitting this regression function:

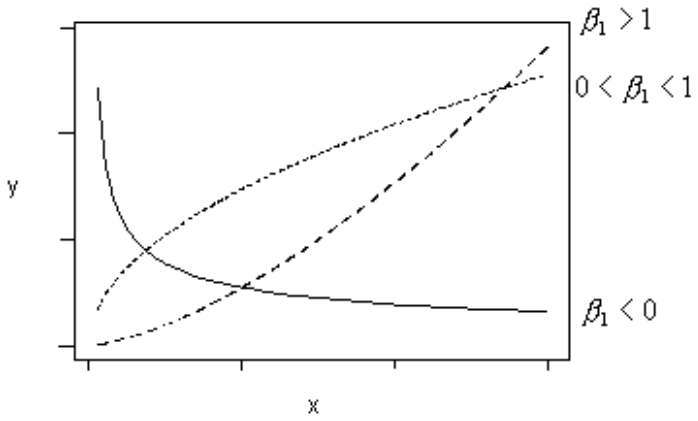
$\mu_Y = \beta_0 + \beta_1 ln(x)$

if the trend in your data follows either of these patterns. That is, fit the model with *y* as the response and *ln*(*x*) as the predictor.



And, finally, try fitting this regression function:

$\mu_{lnY} = \beta_0 + \beta_1 ln(x)$

if the trend in your data follows any of these patterns. That is, fit the model with *ln*(*y*) as the response and *ln*(*x*) as the predictor.	

Second piece of advice. If the variances are unequal and/or error terms are not normal, try a "**power transformation**" on *y*. A **power transformation on *y*** involves transforming the response by taking it to some power λ. That is $y^* = y^\lambda$. Most commonly, for interpretation reasons, λ is a "meaningful" number between -1 and 2, such as -1, -0.5, 0, 0.5, (1), 1.5, and 2 (i.e., it's rare to see $\lambda = 1.362$, for example. **When λ = 0**, the transformation is taken to be the natural log transformation. That is $y^* = ln(y)$. One procedure for estimating an appropriate value for λ is the so-called **Box-Cox Transformation**, which we'll explore further in the next section.

Third piece of advice. If the error variances are unequal, try "**stabilizing the variance**" by transforming *y*:

- If the response *y* is a Poisson count, the variances of the error terms are not constant but rather depend on the value of the predictor. A common (now archaic?) recommendation is to transform the response using the "**square root transformation**," $y^* = \sqrt{y}$, and stay within the linear regression framework. Perhaps, now, the advice should be to use "**Poisson regression**" (which we'll cover in Lesson 15).
- If the response *y* is a binomial proportion, the variances of the error terms are not constant but rather depend on the value of the predictor. Another common (now archaic?) recommendation is to transform the response using the "**arcsine transformation**," $\hat{p}^* = sin^{-1}\left(\sqrt{\hat{p}}\right)$, and stay within the linear regression framework. Perhaps, now, the advice should be to use a form of "**logistic regression**" (which we'll cover in Lesson 12).
- If the response *y* isn't anything special, but the error variances are unequal, a common recommendation is to try the natural log transformation $y^* = ln(y)$ or the "**reciprocal transformation**" $y^* = \frac{1}{y}$.

And two final pieces of advice.

- It's not really okay to remove some data points just to make the transformation work better, but if you have a good reason to do so, make sure you report the scope of the model.
- It's better to give up some model fit than to lose clear interpretations. Just make sure you report that this is what you did.

‹ 7.3 - Log-transforming Both the Predictor and Response for SLR (/stat462/node/154) | up (/stat462/node/85) | 7.5 - Further Transformation Advice and Box-Cox › (/stat462/node/156)

STAT 462

Applied Regression Analysis

7.7 - Polynomial Regression

In our earlier discussions on multiple linear regression, we have outlined ways to check assumptions of linearity by looking for curvature in various plots.

- For instance, we look at the scatterplot of the residuals versus the fitted values.
- We also look at a scatterplot of the residuals versus each predictor.

Sometimes, a plot of the residuals versus a predictor may suggest there is a nonlinear relationship. One way to try to account for such a relationship is through a **polynomial regression** model. Such a model for a single predictor, X, is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_h X^h + \epsilon,$$

where h is called the **degree** of the polynomial. For lower degrees, the relationship has a specific name (i.e., $h = 2$ is called **quadratic**, $h = 3$ is called **cubic**, $h = 4$ is called **quartic**, and so on). Although this model allows for a nonlinear relationship between Y and X, polynomial regression is still considered linear regression since it is linear in the regression coefficients, $\beta_1, \beta_2, \ldots, \beta_h$!

In order to estimate the equation above, we would only need the response variable (Y) and the predictor variable (X). However, polynomial regression models may have other predictor variables in them as well, which could lead to interaction terms. So as you can see, the basic equation for a polynomial regression model above is a relatively simple model, but you can imagine how the model can grow depending on your situation!

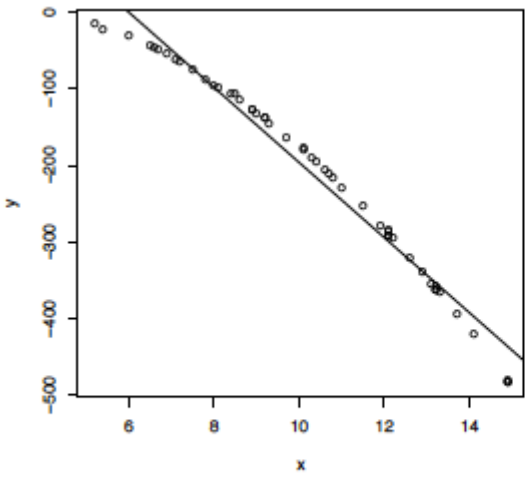
For the most part, we implement the same analysis procedures as done in multiple linear regression. To see how this fits into the multiple linear regression framework, let us consider a very simple data set of size $n = 50$ that was simulated:

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	6.6	-45.4	21	8.4	-106.5	41	8	-95.8
2	10.1	-176.6	22	7.2	-63	42	8.9	-126.2
3	8.9	-127.1	23	13.2	-362.2	43	10.1	-179.5
4	6	-31.1	24	7.1	-61	44	11.5	-252.6
5	13.3	-366.6	25	10.4	-194	45	12.9	-338.5
6	6.9	-53.3	26	10.8	-216.4	46	8.1	-97.3
7	9	-131.1	27	11.9	-278.1	47	14.9	-480.5
8	12.6	-320.9	28	9.7	-162.7	48	13.7	-393.6
9	10.6	-204.8	29	5.4	-21.3	49	7.8	-87.6
10	10.3	-189.2	30	12.1	-284.8	50	8.5	-105.4
11	14.1	-421.2	31	12.1	-287.5			
12	8.6	-113.1	32	12.1	-290.8			
13	14.9	-482.3	33	9.2	-137.4			
14	6.5	-42.9	34	6.7	-47.7			
15	9.3	-144.8	35	12.1	-292.3			
16	5.2	-14.2	36	13.2	-356.4			
17	10.7	-211.3	37	11	-228.5			
18	7.5	-75.4	38	13.1	-354.4			
19	14.9	-482.7	39	9.2	-137.2			
20	12.2	-295.6	40	13.2	-361.6			

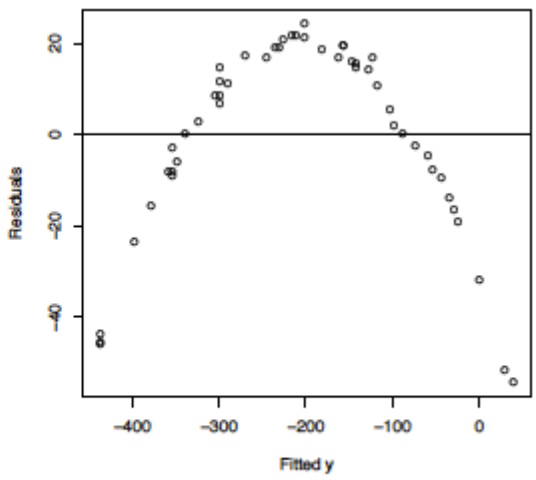
The data was generated from the quadratic model

$$y_i = 5 + 12x_i - 3x_i^2 + \epsilon_i,$$

where the $\epsilon_i s$ are assumed to be normally distributed with mean 0 and variance 2. A scatterplot of the data along with the fitted simple linear regression line is given below (a). As you can see, a linear regression line is not a reasonable fit to the data.

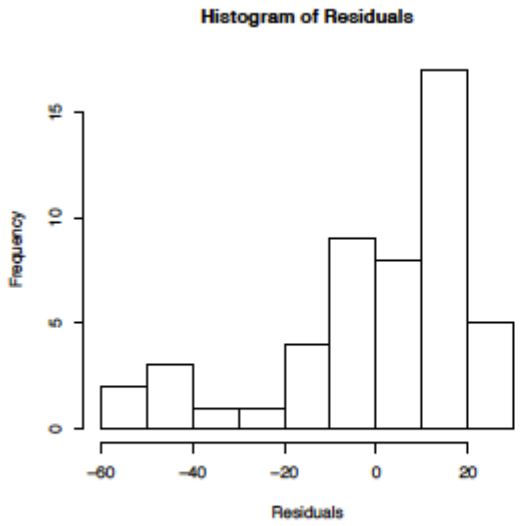


(a)

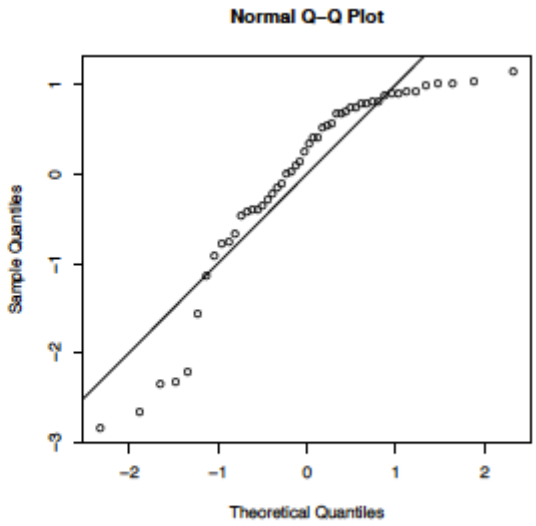


(b)

Loading [MathJax]/extensions/MathZoom.js



(c)



(d)

Residual plots of this linear regression analysis are also provided in the plot above. Notice in the residuals versus fits plot (b) how there is obvious curvature and it does not show uniform randomness as we have seen before. The histogram (c) appears heavily left-skewed and does not show the ideal bell-shape for normality. Furthermore, the normal probability plot (d) seems to deviate from a straight line and curves down at the extreme percentiles. These plots alone suggest that there is something wrong with the model being used and indicate that a higher-order model may be needed.

The matrices for the second-degree polynomial model are:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{50} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{50} & x_{50}^2 \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{50} \end{pmatrix}$$

where the entries in **Y** and **X** would consist of the raw data. So as you can see, we are in a setting where the analysis techniques used in multiple linear regression are applicable.

Some general guidelines to keep in mind when estimating a polynomial regression model are:

- The fitted model is more reliable when it is built on a larger sample size *n*.
- Do not extrapolate beyond the limits of your observed values, particularly when the polynomial function has a pronounced curve such that an extraploation produces meaningless results beyond the scope of the model.
- Consider how large the size of the predictor(s) will be when incorporating higher degree terms as this may cause numerical overflow for the statistical software being used.
- Do not go strictly by low *p*-values to incorporate a higher degree term, but rather just use these to support your model only if the resulting residual plots looks reasonable. This is an example of a situation where you need to determine "practical significance" versus "statistical significance".
- In general, as is standard practice throughout regression modeling, your models should adhere to the **hierarchy principle**, which says that if your model includes X^h and X^h is shown to be a statistically significant predictor of *Y*, then your model should also include each X^j for all $j < h$, whether or not the coefficients for these lower-order terms are significant. In other words, when fitting polynomial regression functions, fit a higher-order model and then explore whether a lower-order (simpler) model is adequate. For example, suppose [illegible] cubic polynomial regression function:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

Then, to see if the simpler first order model (a "straight line") is adequate in describing the trend in the data, we could test the null hypothesis:

$$H_0 : \beta_2 = \beta_3 = 0$$

But then … if a polynomial term of a given order is retained, then *all related lower-order terms are also retained*. That is, if a quadratic term (x^2) is deemed significant, then it is standard practice to use this regression function:

$$\mu_Y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

and not this one:

$$\mu_Y = \beta_0 + \beta_2 x_i^2$$

whether or not the linear term (x) is significant. That is, we always fit the terms of a polynomial model in a hierarchical manner.

‹ 7.6 - Interactions Between Quantitative Predictors (/stat462/node/157) | up (/stat462/node/85) | 7.8 - Polynomial Regression Examples › (/stat462/node/159)

STAT 462

Applied Regression Analysis

7.8 - Polynomial Regression Examples

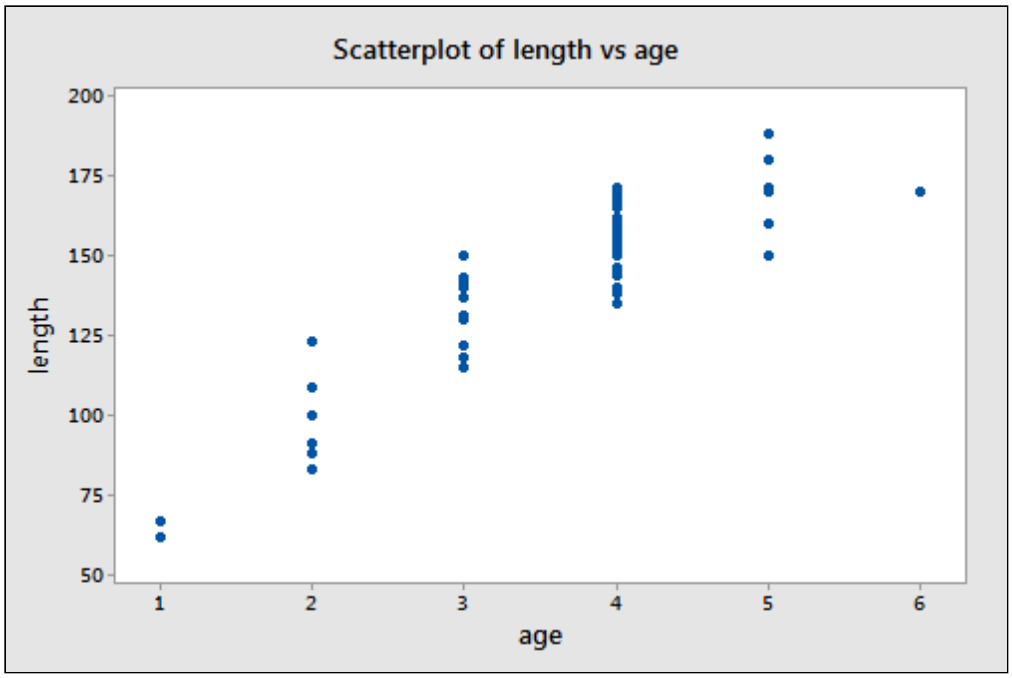
Example 1: How is the length of a bluegill fish related to its age?

In 1981, $n = 78$ bluegills were randomly sampled from Lake Mary in Minnesota. The researchers (Cook and Weisberg, 1999) measured and recorded the following data (bluegills.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/bluegills.txt)):

- Response (y): length (in mm) of the fish
- Potential predictor (x_1): age (in years) of the fish

The researchers were primarily interested in learning how the length of a bluegill fish is related to it age.

A scatter plot of the data:



suggests that there is positive trend in the data. That is, not surprisingly, as the age of bluegill fish increases, the length of the fish tends to increase. The trend, however, doesn't appear to be quite linear. It appears as if the relationship is slightly curved.

Loading [MathJax]/extensions/MathMenu.js

ature in these data is to formulate a "**second-order polynomial model**" with **one quantitative predictor**:

$$y_i = (\beta_0 + \beta_1 x_i + \beta_{11} x_i^2) + \epsilon_i$$

where:

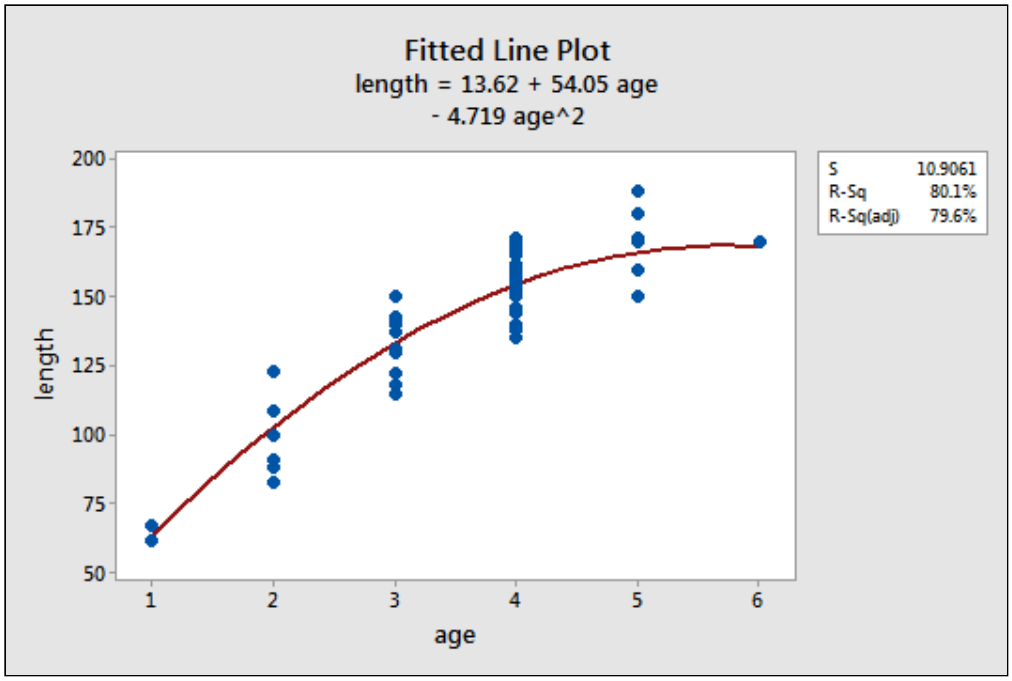
- y_i is length of bluegill (fish) i (in mm)
- x_i is age of bluegill (fish) i (in years)

and the **independent** error terms ε_i follow a **normal** distribution with mean 0 and **equal variance** σ^2.

You may recall from your previous studies that "**quadratic function**" is another name for our formulated regression function. Nonetheless, you'll often hear statisticians referring to this quadratic model as a second-order model, because the highest power on the x_i term is 2.

Incidentally, observe the notation used. Because there is only one predictor variable to keep track of, the 1 in the subscript of x_{i1} has been dropped. That is, we use our original notation of just x_i. Also note the double subscript used on the slope term, β_{11}, of the quadratic term, as a way of denoting that it is associated with the squared term of the one and only predictor.

The estimated quadratic regression function looks like it does a pretty good job of fitting the data:



To answer the following potential research questions, do the procedures identified in parentheses seem reasonable?

- How is the length of a bluegill fish related to its age? (Describe the nature—"quadratic"—of the regression function.)
- What is the length of a randomly selected five-year-old bluegill fish? (Calculate and interpret a prediction interval for the response.)

Statistical software output follows:

```
Analysis of Variance

Source         DF   Adj SS   Adj MS  F-Value  P-Value
Regression      2  35938.0  17969.0   151.07    0.000
  age           1   8252.5   8252.5    69.38    0.000
  age^2         1   2972.1   2972.1    24.99    0.000
Error          75   8920.7    118.9
  Lack-of-Fit   3    108.0     36.0     0.29    0.829
  Pure Error   72   8812.7    122.4
Total          77  44858.7


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.9061  80.11%     79.58%      78.72%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value    VIF
Constant    13.6     11.0     1.24    0.220
age        54.05     6.49     8.33    0.000  23.44
age^2     -4.719    0.944    -5.00    0.000  23.44


Regression Equation

length = 13.6 + 54.05 age - 4.719 age^2
```

Prediction for length

```
Regression Equation

length = 13.6 + 54.05 age - 4.719 age^2


Variable  Setting
age             5
age^2          25


    Fit   SE Fit        95% CI              95% PI
165.902  2.76901  (160.386, 171.418)  (143.487, 188.318)
```

The output tells us that:

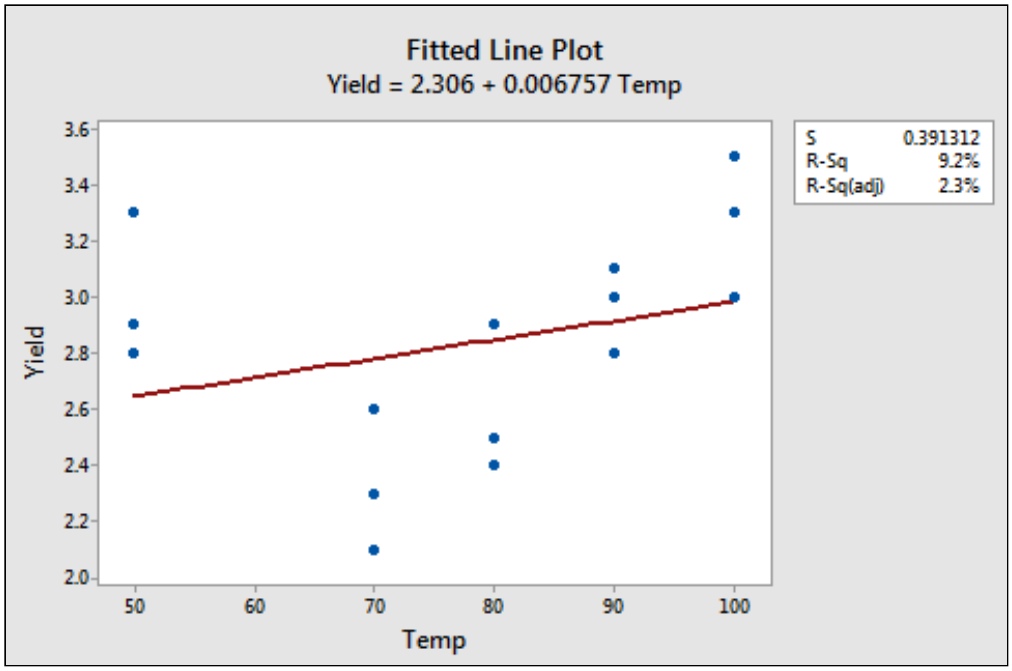
- 80.1% of the variation in the length of bluegill fish is reduced by taking into account a quadratic function of the age of the fish.
- We can be 95% confident that the length of a randomly selected five-year-old bluegill fish is between 143.5 and 188.3 mm.

Example 2: Yield Data Set

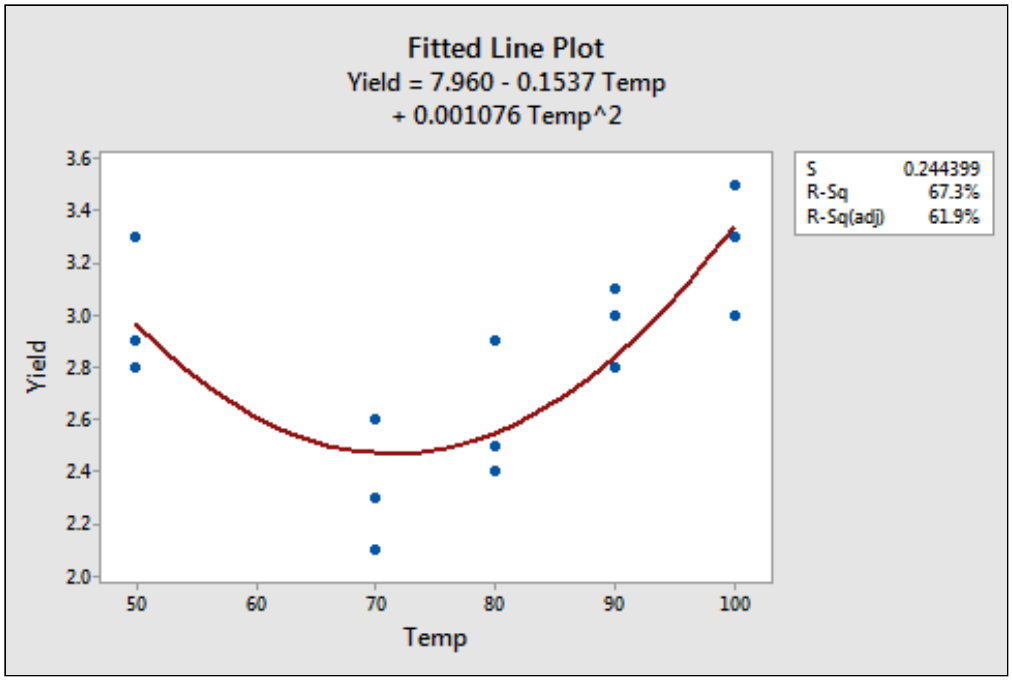
This data set of size $n = 15$ (yield.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/yield.txt)) contains measurements of yield from an experiment done at five different temperature levels. The variables are y = yield and x = temperature in degrees Fahrenheit. The table below gives the data used

i	Temperature	Yield
1	50	3.3
2	50	2.8
3	50	2.9
4	70	2.3
5	70	2.6
6	70	2.1
7	80	2.5
8	80	2.9
9	80	2.4
10	90	3.0
11	90	3.1
12	90	2.8
13	100	3.3
14	100	3.5
15	100	3.0

The figures below give a scatterplot of the raw data and then another scatterplot with lines pertaining to a linear fit and a quadratic fit overlayed. Obviously the trend of this data is better suited to a quadratic fit.



Loading [MathJax]/extensions/MathMenu.js



Here we have the linear fit results:

```
##########
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.306306   0.469075   4.917 0.000282 ***
temp        0.006757   0.005873   1.151 0.270641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3913 on 13 degrees of freedom
Multiple R-Squared: 0.09242,    Adjusted R-squared: 0.0226
F-statistic: 1.324 on 1 and 13 DF,  p-value: 0.2706
##########
```

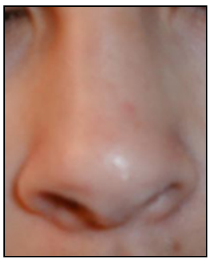
Here we have the quadratic fit results:

```
##########
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.9604811  1.2589183   6.323 3.81e-05 ***
temp        -0.1537113  0.0349408  -4.399 0.000867 ***
temp2        0.0010756  0.0002329   4.618 0.000592 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2444 on 12 degrees of freedom

Multiple R-Squared: 0.6732,     Adjusted R-squared: 0.6187
F-statistic: 12.36 on 2 and 12 DF,  p-value: 0.001218
```

We see that both temperature and temperature squared are significant predictors for the quadratic model (with *p*-values of 0.0009 and 0.0006, respectively) and that the fit is much better than for the linear fit. From this output, we see the estimated regression equation is $y_i = 7.96050 - 0.15371x_i + 0.00108x_i^2$. Furthermore, the ANOVA table below shows that the model we fit is statistically significant at the 0.05 significance level with a *p*-value of 0.0012. Thus, our model should include a quadratic term.

```
##########
Analysis of Variance Table

Response: yield
           Df  Sum Sq Mean Sq F value   Pr(>F)
Regression  2 1.47656 0.73828   12.36 0.001218 **
Residuals  12 0.71677 0.05973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##########
```

Example 3: Odor Data Set



An experiment is designed to relate three variables (temperature, ratio, and height) to a measure of odor in a chemical process. Each variable has three levels, but the design was not constructed as a full factorial design (i.e., it is not a 3^3 design). Nonetheless, we can still analyze the data using a response surface regression routine, which is essentially polynomial regression with multiple predictors. The data obtained (odor.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/odor.txt)) was already coded and can be found in the table below.

Odor	Temperature	Ratio	Height
66	-1	-1	0
58	-1	0	-1
65	0	-1	-1
-31	0	0	0
39	1	-1	0
17	1	0	-1
7	0	1	-1
-35	0	0	0
43	-1	1	0
-5	-1	0	1
43	0	-1	1
-26	0	0	0
49	1	1	0
-40	1	0	1
		1	1

First we will fit a response surface regression model consisting of all of the first-order and second-order terms. The summary of this fit is given below:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
18.7747	86.83%	76.95%	47.64%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-30.7	10.8	-2.83	0.022	
Temp	-12.13	6.64	-1.83	0.105	1.00
Ratio	-17.00	6.64	-2.56	0.034	1.00
Height	-21.37	6.64	-3.22	0.012	1.00
Temp2	32.08	9.77	3.28	0.011	1.01
Ratio2	47.83	9.77	4.90	0.001	1.01
Height2	6.08	9.77	0.62	0.551	1.01

As you can see, the square of height is the least statistically significant, so we will drop that term and rerun the analysis. The summary of this new fit is given below:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
18.1247	86.19%	78.52%	56.19%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-26.92	8.71	-3.09	0.013	
Temp	-12.13	6.41	-1.89	0.091	1.00
Ratio	-17.00	6.41	-2.65	0.026	1.00
Height	-21.37	6.41	-3.34	0.009	1.00
Temp2	31.62	9.40	3.36	0.008	1.01
Ratio2	47.37	9.40	5.04	0.001	1.01

The temperature main effect (i.e., the first-order temperature term) is not significant at the usual 0.05 significance level. However, the square of temperature is statistically significant. To adhere to the hierarchy principle, we retain the temperature main effect in the model.

‹ 7.7 - Polynomial Regression (/stat462/node/158) up (/stat462/node/85)