**5 – Simple Linear Regression (SLR)**

In the previous section we saw that when the joint distribution of two continuous numeric variables () is bivariate normal (BVN) then,

However, it is not necessary that joint distribution of () be BVN for the to be a line and the to be constant. Furthermore we are not restricted to fit a line to our data. As we saw with the Bulging Rule it is a possible to use a power transformation for (and/or ) to improve linearity. For example, the model:

may be more appropriate for a given regression situation. Both of the models above are actually **simple linear regressio**n models. **Regression** refers to the fact we are trying understand or model properties of the conditional distribution of (or ). **Simple** means that we are using single predictor () to model . **Linear** means the model is linear in the parameters, NOT necessarily in . As an example of model that is NOT linear in the parameters consider the following model for the mean function:

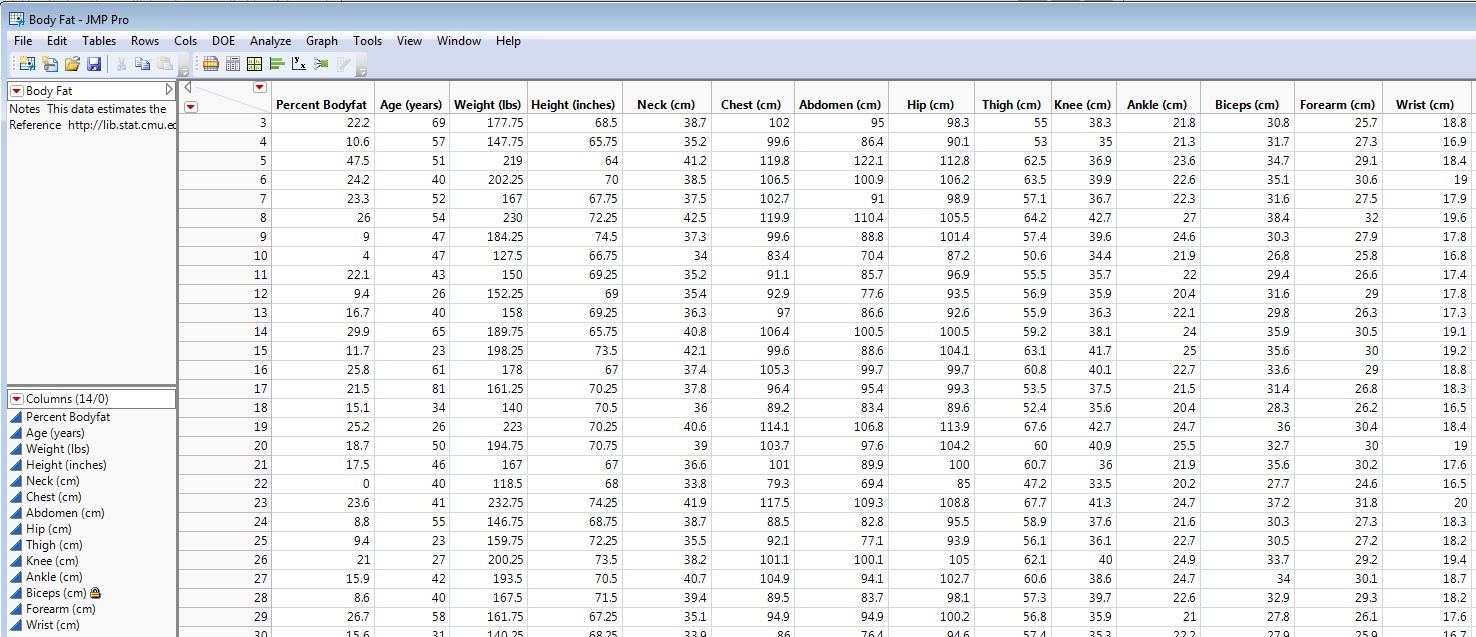
This model is NOT linear in the parameters as the model parameters and are embedded in the transcendental function . This model is a form of the logistic regression model which we will see later in the course and is an example of a **nonlinear regression** model.

In the next section we will examine all of the details of fitting the first model discussed above:

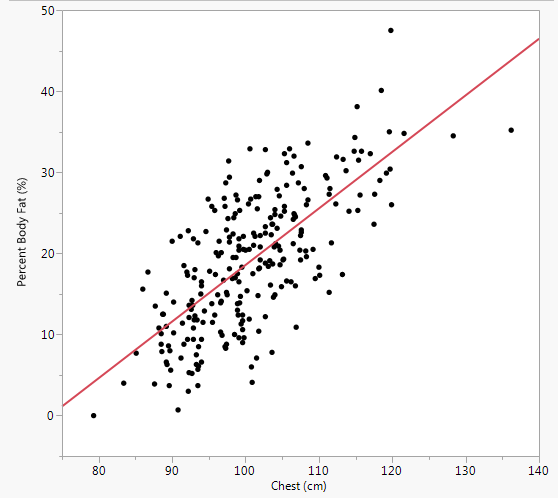
**5.2 – Simple Linear Regression Model**

**Example 5.1 –Body fat (%) and Chest Circumference (cm) (Datafile: Bodyfat.JMP)**

These data can be used to relate percent body fat found by determining the subject’s density by weighing them underwater. This is an expensive and inconvenient way to measure a subject’s percent body fat. Regression techniques can be used to develop models to predict percent body fat (Y) using easily measured body dimensions. In this study *n = 252* men were used and we will focus on the relationship between percent body fat (Y) and chest circumference (X).  
 **Datafile: Body Fat.JMP**



A scatterplot of percent body fat (Y) vs. chest circumference for our random sample is shown below with an estimated regression line added.



We saw in the previous section when we assume that the joint distribution of chest circumference and percent body fat is BVN then estimated regression line is given by:

The estimated intercept () and slope () were computed using the BVN assumption and replacing BVN parameters in the theoretical model by their estimates, i.e. using summary statistics .

This method of estimation is called *maximum likelihood estimation* (MLE) and requires that we assume joint distribution of percent body fat and chest circumference in the population is truly BVN.

Many other methods have been suggested for obtaining estimates of parameters in a regression model. The method that is typically used is called *ordinary least squares* (OLS) in which parameter estimates are found by minimizing the sum of the squared residuals or *residual sum of squares* (RSS). We will now examine the details of OLS for simple linear regression where a line is fit to represent .

**Population Model**

, , and are population parameters, i.e. numerical characteristics of the population, that will need to be estimated using our random sample .

or

This divides the *observable* random variable into two unknown parts, a mean and an error.

**Sample Model**

Given a random sample from the joint distribution of () - the model above says

and

which also implies .

**Estimated Model**

Once we obtain estimates (using some method, e.g. MLE or OLS) for the population parameters we can obtain the following:

The *fitted values* are estimates of the conditional mean of . They can also be viewed as the predicted value of Y when , thus the *fitted values* are also referred to as the *predicted values*.

*Fitted Values*

*Residuals*

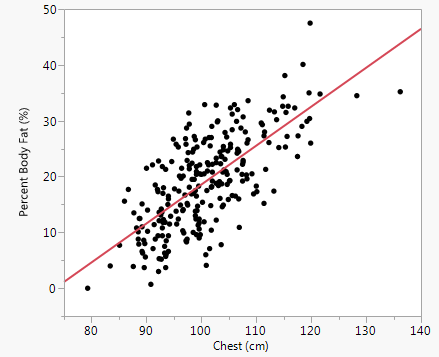
Putting these to quantities together we obtain the following relationship:

This divides the *observed values* of the response into two parts, an estimated mean and observed value of the random error.

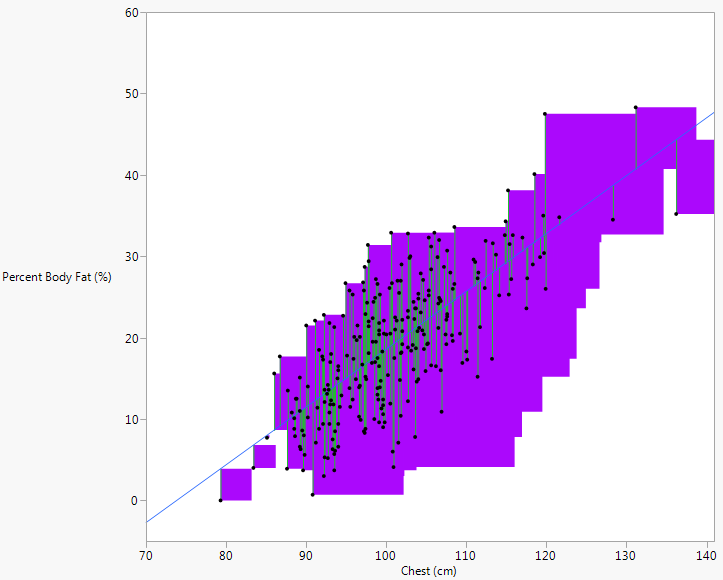
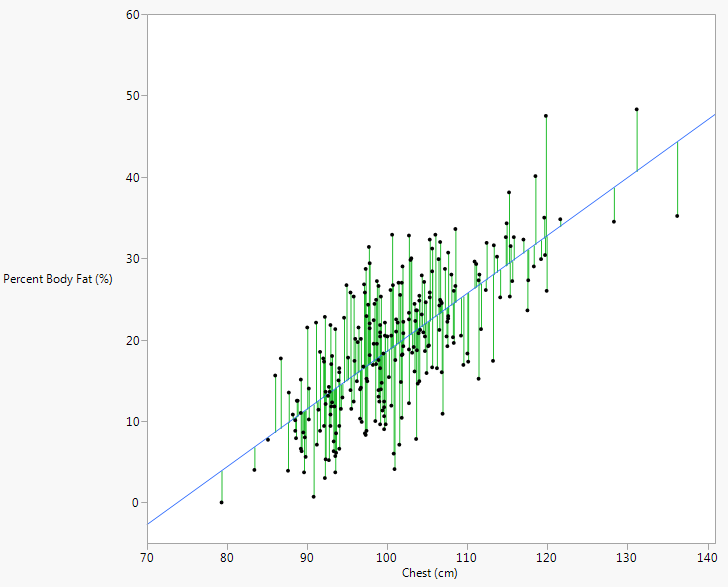
**5.2 - Ordinary Least Squares (OLS) and Parameter Estimates**

OLS regression chooses model parameters to minimize the *residual sum of squares* (RSS). Assuming the models on the previous page this means choosing and to minimize

Graphically: (see also PCTBFregressiondemo.jsl, a JMP script file I sent via e-mail.)



Plot with all residuals drawn Plot showing least squares criterion graphically.



To minimize a function of the two model parameters we need to take the partial derivatives of and set them equal to zero, then solving the system of equations simultaneously. As this requires knowledge of multivariable calculus (MATH 312) we will not go through these details here.

After wading through the calculus the solutions are:

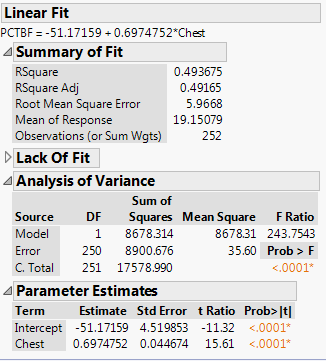
Q: Where have we seen these equations before?

The RSS criterion evaluated at these solutions is given by:

Solving this equation for the correlation squared we obtain

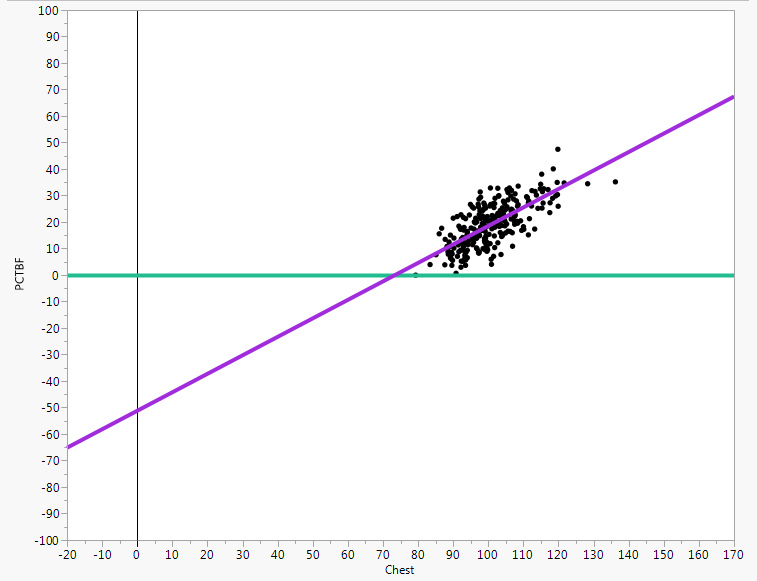
which is the proportion of variation in the response explained by the regression model.

Here are the results of fitting the model using OLS in JMP



Estimates and summary statistics:

**Interpretation of Parameter Estimates**



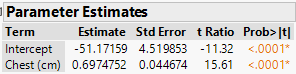
**Estimating the**

Because the variance is essentially the mean of the squared errors, we should expect that an estimate of the conditional variance would be the average of the squared residuals . An unbiased estimator of the common conditional variance is RSS divided by its degrees of freedom (*df*), which is the sample size () minus the number of parameters in the mean function which is 2 in this case, thus .

This is quantity is also called the *residual mean square* or the *mean squared error.*  The square root of this quantity is an estimate of the , and is called the *root mean squared error* (RMSE) in JMP, the *residual standard error* in R, and the *standard error of regression* by Weisberg (2014).

**5.3 - Estimated Variances & Standard Errors of Parameter Estimates**

Below is a summary of parameter estimates from the simple linear regression of percent body fat on chest circumference.



The variance and standard errors formulae for the estimated intercept and slope are given below.

and the standard error

and the standard error

For notational convenience we will drop the conditional dependence on X and denote the standard errors as and .

A is given by: .

By right-clicking on the table above you can add Lower 95% and Upper 95% confidence limits to the output.



The reported p-values typically reported in regression output is for testing the following null and alternative hypotheses.

Test statistic

In general we can test specific values of population parameters.

Test statistic

**5.4 – Confidence and Prediction Intervals**

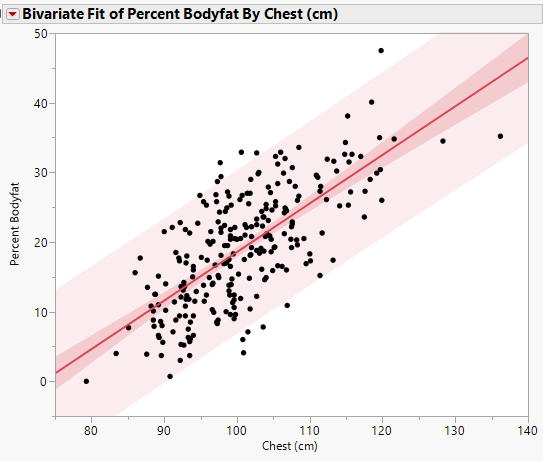
Confidence intervals for the **conditional mean** give a range of values we are confident will cover the true conditional mean of when . To obtain an estimate of the conditional mean we simply use the estimated model,

and the estimated standard error is given by

🡨 Weisberg (2014)

Prediction intervals are for estimating an **individual value** for the random variable given . The point estimate is the same as above, but the standard error needs to account for both the variance of the population and the estimate of the conditional mean .

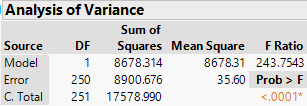
The plot on the next page shows the confidence and prediction intervals added to the regression of percent body fat on chest circumference.



The table below displays 95% CIs for and as well as 95% PIs for individuals with these chest measurements. Notice as we move away from the CI’s and PI’s get wider. It should be clear from the standard error formulae for confidence and prediction why this happens.  
  
  


**5.4 – Analysis of Variance Table**

The ANOVA table below is from the regression of percent body fat on chest circumference.



|  |  |
| --- | --- |
| **Quantity** | **Value / Description** |
| Sum of Squares  for C. Total |  |
| Sum of Squares for  Error |  |
| Sum of Squares for  Model |  |
| Degrees of Freedom (*df*) for C. Total |  |
| Degrees of Freedom for Error |  |
| Degrees of Freedom for Model |  |
| Mean Square for  Error | Note: This is our best estimate for , which is the variance in the conditional distribution of Percent Body Fat|Chest, i.e. the variability that remains in Percent Body Fat (%) that cannot be explained by Chest Circumference (cm). |
| Mean Square for  Model |  |

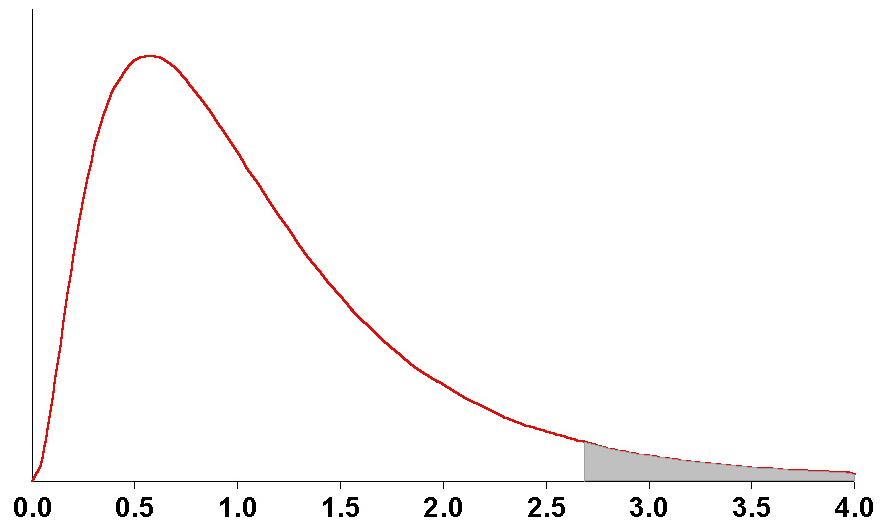
The quantities in the ANOVA table can be used to compare rival models, which will be particularly useful when we consider multiple regression models.

The F-test summarized in an ANOVA table for a regression model is testing the following hypotheses.

As you might expect this test will usually result in rejection of NH as generally using a model will produce a statistically significant reduction in the variation in the response. As an example, for the regression of percent body fat on chest circumference the hypotheses tested are:

The test statistic is given by:

Recall the F-distribution has two *df* values that control its shape. The larger the F-statistic the more evidence we have in support of AH model.



The generic regression ANOVA table is given by

Source df SS MS F p-value

Regression (see above) (see above)

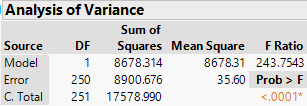
Residual or

Error

Total or

C. Total

Below is the ANOVA table from the regression of percent body fat on chest circumference.



Calculating the F-test:

We will see later when we consider multiple regression that this test can be used to compare any two models where the NH model is a subset of or is nested within the AH model. A NH model is nested within the AH model if the NH model can be formed by setting certain parameters in the AH model equal to zero, i.e. the NH model is formed by dropping terms from the AH model. This test will be referred to as the “Big F-test”.

**Big F-test for Comparing Two Nested Regression Models**

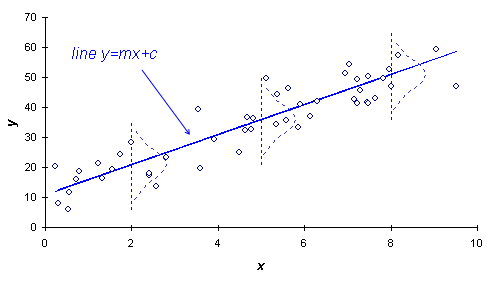
**5.5 – A First Look at Checking Model Assumptions**

In fitting a Simple Linear Regression Model using OLS we generally make the following assumptions:

1. The model for E(Y|X) is correct. If we are fitting a line to model the conditional expectation this becomes .
3. The observations are independent. This also says that the random errors () are independent. This is typically violated when X = “time”.
4. For purposes of statistical inference, i.e. CI’s and p-values derived using the t- and F-distributions, we assume normality. Specifically we assume the conditional distribution of is normal.

Putting assumptions 1, 2, and 4 together we are assuming

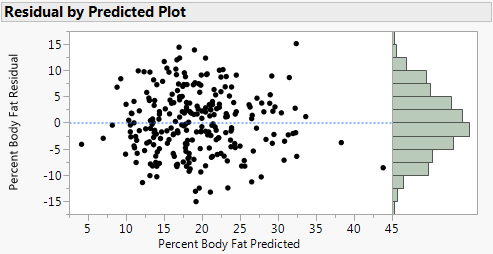
or equivalently . Again we can refer to the diagram below to visualize these assumptions.



Mean Function

To graphically check these assumptions we examine plots the residuals ().

To assess model assumptions (1) – (3) graphically we can plot the residuals vs. the fitted values . If we add a histogram of the residuals on the side as JMP does, we can actually assess assumption (4) as well.

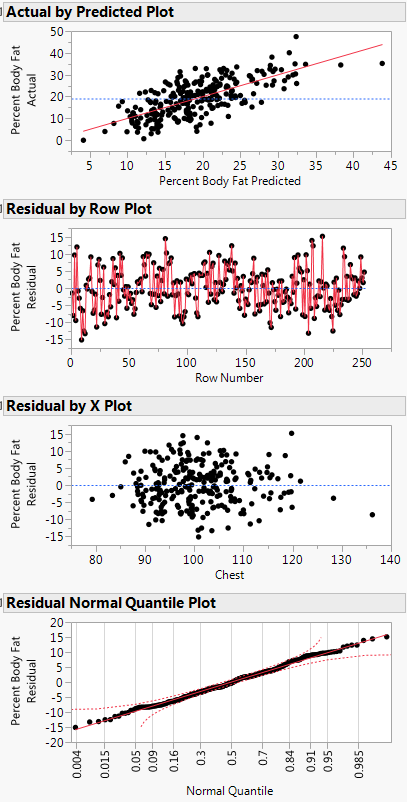


The residual vs. fitted/predicted value plot above suggests no violation of the model assumptions. The variation in the residuals appears constant, there is no discernible pattern in the residuals which would suggest inadequacies of the assumed mean function, the residuals appear to be normally distributed, and there appears to be no dependence between the residual values.

Other plots that can be useful:

* Actual vs. Fitted/Predicted Values – ideally this is will be the line y = x.
* Residuals by Row Plot - This plot can used to identify outlying cases (more on this later) and if rows are chronological (time related) lack of independence can be identified.
* Residual vs. predictor values ( – in the case of simple linear regression where the mean function is a line, this plot is “identical” to the residuals vs. the fitted values because .
* Normal quantile plot of the residuals – a better tool assess normality assumption.

These plots are shown for the regression of percent body fat on chest circumference on the following page.

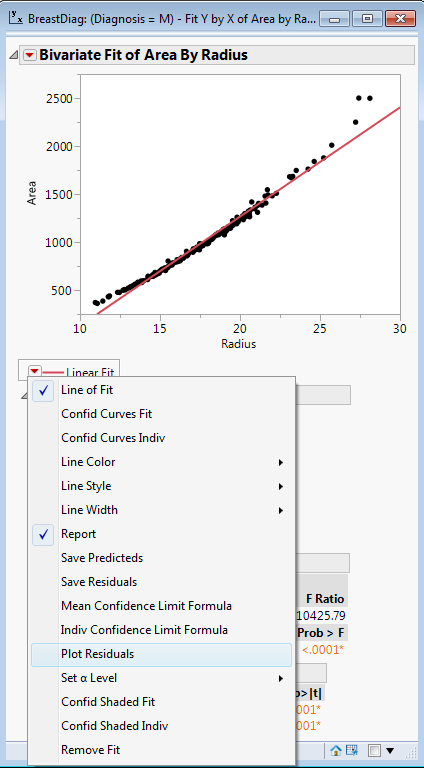


**Comments:**

To better understand the role of these plots in checking model assumptions it is instructive to consider cases where the model assumptions are violated. We will return to earlier examples for this purpose.

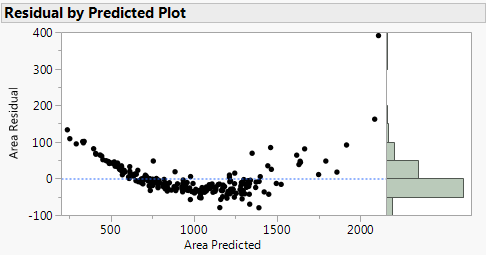
**Example 5.2 – Cell Area and Radius from Breast Tumor Study (Malignant Only)**

**(Datafile: BreastDiag (Malignant).JMP)**Below is a scatterplot of area vs. radius with the fitted model,

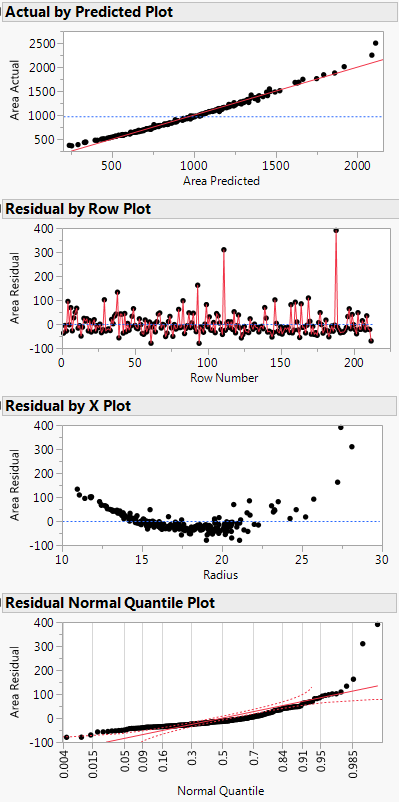


The **Plot Residuals** option creates the sequence of plots that can used to for checking model assumptions.

1. Plot of the residuals vs. the fitted/predicted values



**Comments:**

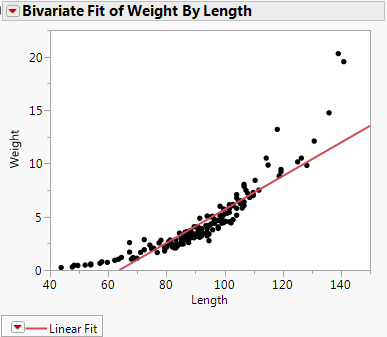


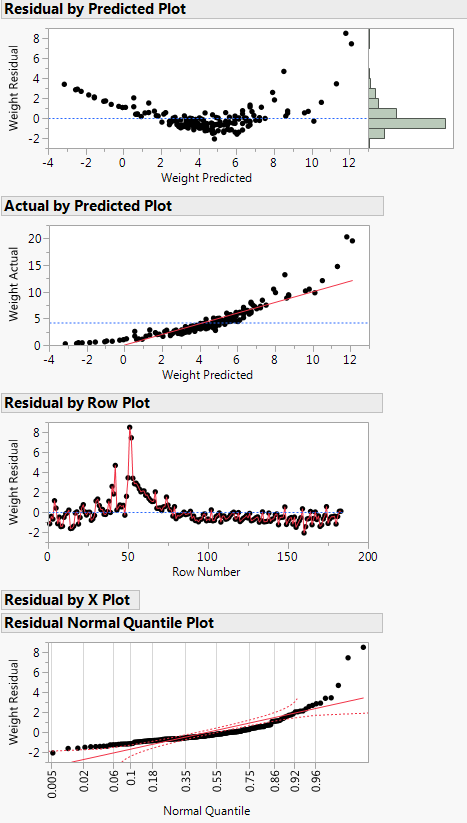
**Comments:**

These plots suggest that the assumed model is **inappropriate** for these data. In particular we conclude , is NOT constant, and the normality assumption is violated.

**Example 5.3 – Weight (kg) and Length (cm) of Mississippi River Paddlefish  
(Datafile: Paddlefish (clean).JMP)**

Below is a scatterplot of weight vs. length with the fitted model,

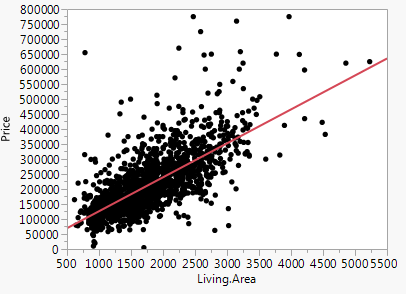


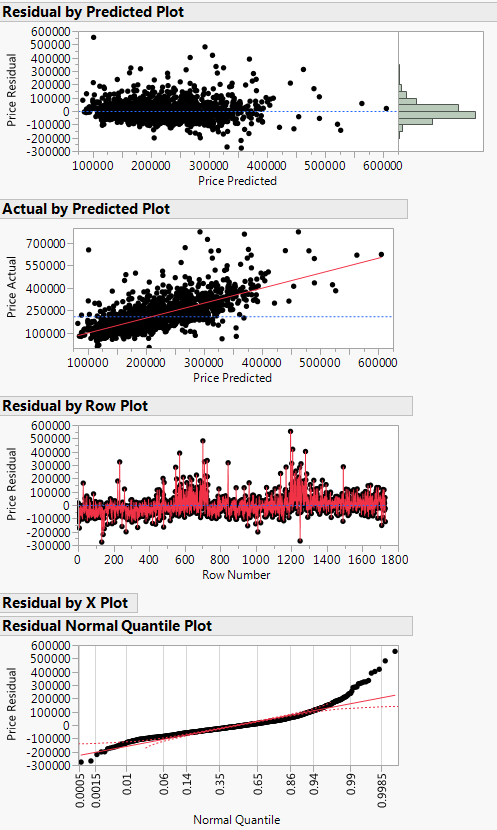


**Example 5.4 – List Price of Homes ($) and Living Area (ft2) in Saratoga, NY**

**(Datafile: Saratoga NY Homes.JMP)**

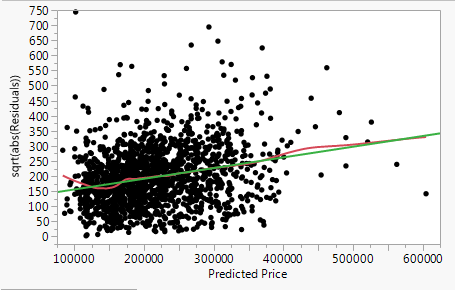
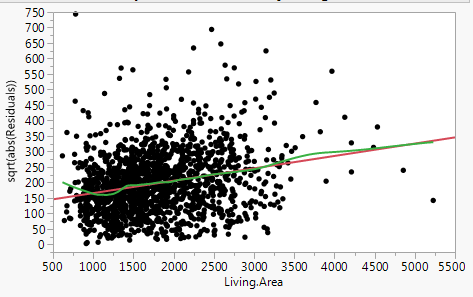
In this example we consider the simple linear regression of the list price of homes in Saratoga, NY vs. the square footage of the total living area.

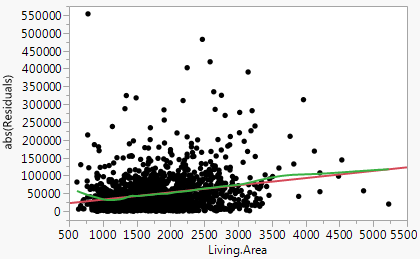




**Nonconstant Variance Plot**

A better plot to assess nonconstant variation is to plot or vs. the fitted values or vs. the predictor values . In JMP we will need to save the residuals and fitted values to the data table and then perform the necessary residual calculations in order to construct the plot. A kernel smooth added to the plot will help us see increasing (or decreasing) variation when present. Here are a sequence of such plots using the absolute residuals from the regression of price on living area.

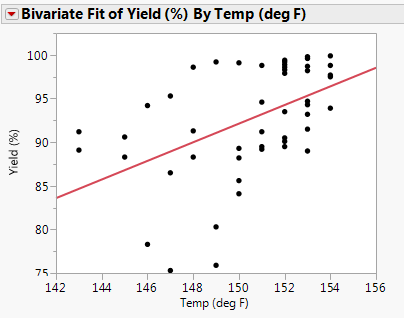
 

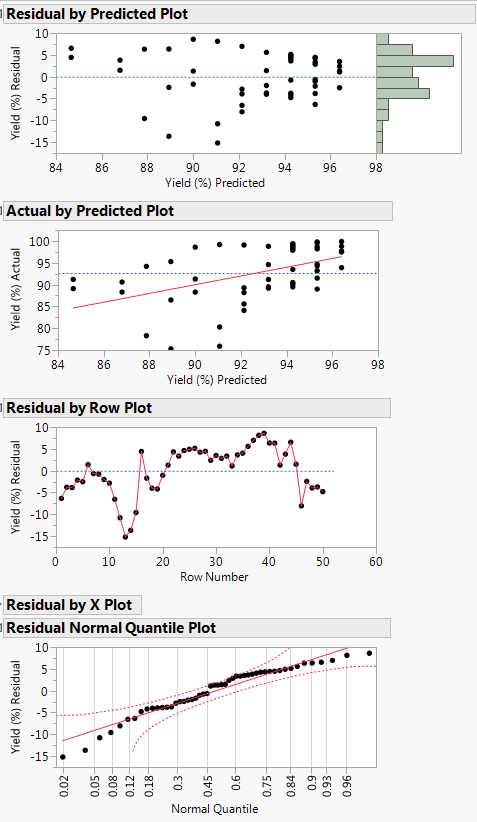
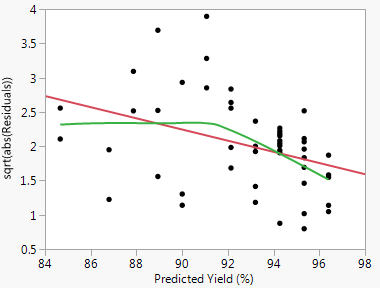
These plots clearly show that the variation in the list price increases with living area is larger or equivalently when the list price is larger.

**Example 5.5 – Chemical Yield (%) vs. Temperature (oF)   
(Datafile: Chemical Yield.JMP)**

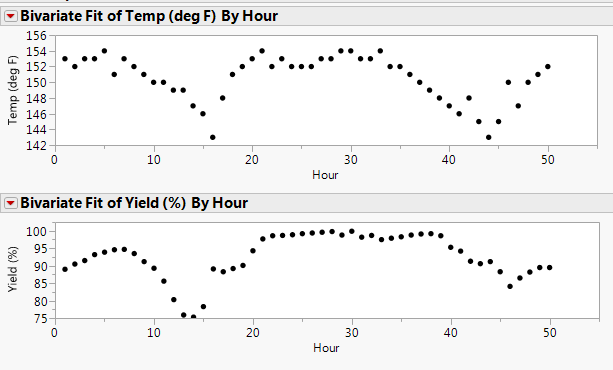
This is a study of the percent yield from a chemical reaction (Y) and temperature under which the reaction is taking place (X). Below is a scatterplot of Yield (%) vs. Temperature (oF).

**Comments:**



Actually these data were collected over 50 one-hour periods.



Clearly there are patterns over time in both the response chemical yield and the predictor reaction temperature. The non-independence of the observations in this study is clearly seen by examining these data in this manner.

**5.6 – Parameter Estimate Interpretation when Transformations are Used**

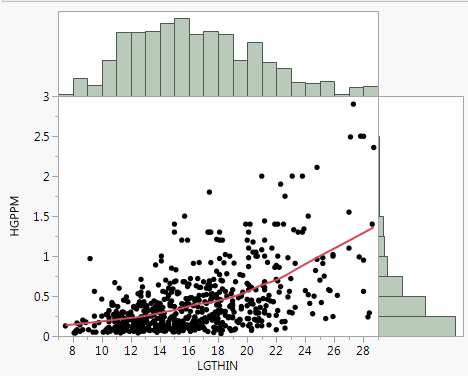
When transformations are used to improve to stabilize the variance   
(e.g. ) or address deficiencies in the model for , interpretation of the estimated parameters for the regression is an issue. In certain cases however we can still interpret the parameters in a meaningful way.

1. **Log Transformation of the Response (Log-Linear SLR Model)**

The use the log transformation for the response is typically done to stabilize the conditional variance (, to improve the adequacy of the model for the mean function , or both. As an example we will again consider the MN walleye data from assignment 1.

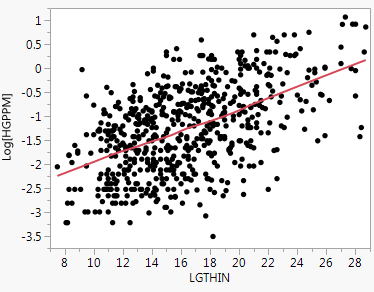
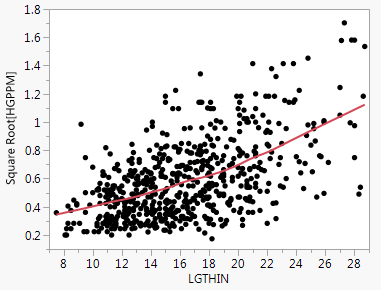
**Example 5.6: Mercury Contamination in MN Walleyes   
(Datafile: MN Walleyes.JMP)**

Below is a plot of mercury concentration (HGPPM) vs. walleye length in inches (LGTHIN) with marginal distributions and a nonparametric estimate of added to the scatterplot. Clearly the model and is adequate in both respects.



The Bulging Rule and Ladder of Powers would suggest lowering the power on the response HGPPM. As the variance is non-constant would be good choices to consider.

Below are plots with the response transformed using both of these power transformations.



Clearly the log transformation for the mercury contamination level is deals with both the curvature and non-constant variance issues found in the untransformed relationship.

The following simple linear regression model is then suggested by the considerations above.

or more simply we can write

solving for Y

🡨 Note: this is a nonlinear regression model!

The intercept is actually a constant multiplier in front of the exponential function. Also the errors are multiplicative as well, thus as the mean increases so does the error! Thus by taking the logarithm of the response we have addressed the nonconstant variation in the response!

For the slope consider the a unit increase in the predictor .

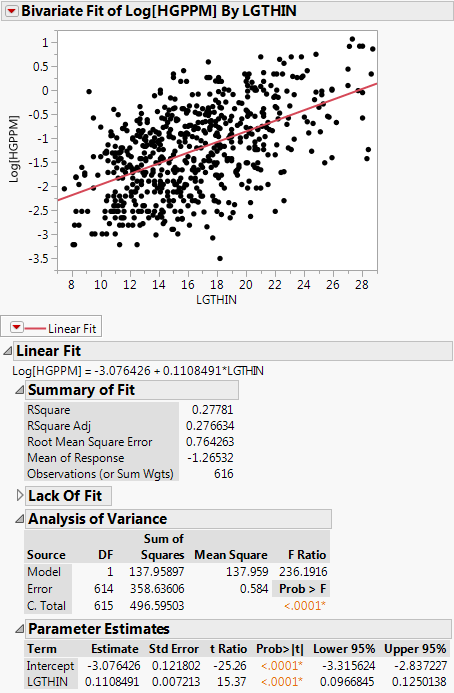
and

In order to quantify or interpret we need to consider the ratio of these conditional expectations.

Thus is the multiplicative increase in the response associated with a unit increase in the predictor . This can also be expressed a percent increase in the response.

For a c-unit increase in the predictor variable (e.g. c = 5 inches) the multiplicative increase in the response is given by and for a percent increase we have,

Below are the results of this model fit to these data.



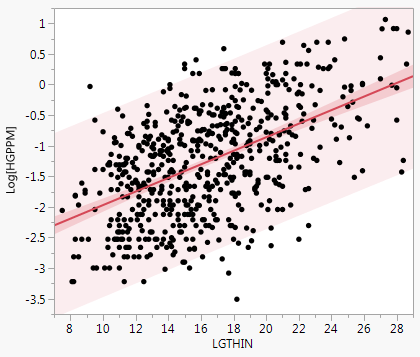
The fitted model gives

Which implies

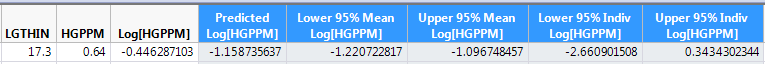
A 1 inch increase in the length of walleyes corresponds to an estimated 1.117 times higher mean mercury level or an 11.7% increase.

Using the CI for we estimate with 95% confidence a 1 inch increase in length is associated with between times multiplicative increase in the typical mercury level. In terms of percentage increase, we estimate with 95% confidence the percentage increase in Hg level associated with a one inch increase in length is between (10.2%, 13.3%).

**Scatterplot with Confidence and Prediction Intervals**



Below is a portion of a table containing the fitted value (, confidence interval for the mean response in the log scale, and prediction interval for the log mercury contamination found in a fish with a length = 17.3 inches.



We can use the fact that to back transform the fitted values, CI’s, and PI’s to the original parts per million (ppm) for mercury level.

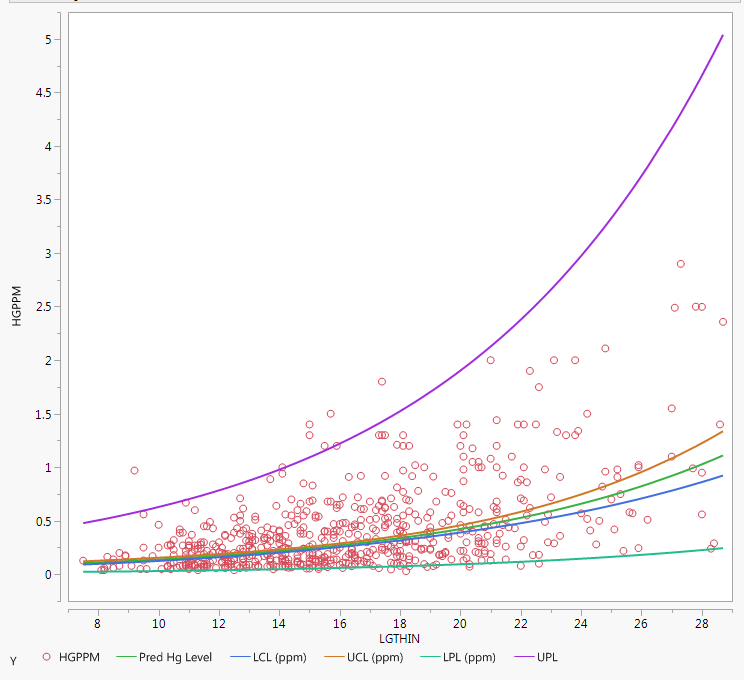
For LGTHIN = 17.3 inches we have….

Back transforming the fitted value from the log scale model we have

Back transforming the endpoints of the CI for

Back transforming the endpoints of the PI for

We can use JMP to back transform all of the highlighted columns to visualize this fitted model in the original response scale (ppm) and then plot them all simultaneously using **Graph > Overlay Plots**.



**Question:** What is the back-transformed CI for actually estimating? Is it estimating ?

In general, !! However it may be reasonable to assume that for some cases, in which case we can interpret these estimates of the conditional mean in the original scale. Personally, I think cases where the approximation above holds are rare!

**A Better Interpretation**

However, if we let *med*(Y) denote the population median of Y then one can easily show/argue that .

Now if we believe that in the log-scale mercury levels in the population of MN walleyes are normally distributed we can view the fitted values and the associated CI’s as estimates of the !

Why?

The punchline is this: if you back-transform results from the regression where the was used as the response we interpret parameter estimates multiplicatively and I would advocate that we interpret back-transformed estimates & CI’s as being for conditional medians, not means!

1. **Other Log Transformation SLR Models**

Other forms of SLR models where logarithmic transformations for and/or are summarized in the table below:

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

**Linear-Log Model**

Here is not transformed (linear) and only is log-transformed (log). For interpretation purposes base 2 and base 10 transformations are recommended.

Here represents the change in the mean of when is increased by 1 unit. Increasing by 1 corresponds to doubling , thus represents the increase in associated with doubling . If base 10 is used, represents the increase in associated with a 10-fold increase in . If base is used the multiplicative increase in is .. , which is rather clunky.

**Log-Log Model**

Here and are both log-transformed so the model is given by

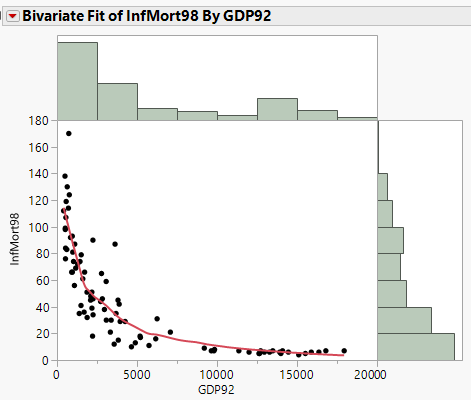
.

In instances where both the dependent variable and independent variable(s) are log-transformed variables, the interpretation is a combination of the linear-log and log-linear cases above. In other words, the interpretation is given as an expected percentage change in when increases by some percentage. Such relationships, where both and are log-transformed, are commonly referred

to as elastic in econometrics, and the coefficient of is referred to as an *elasticity*.   
  
If increases by 1% then increases %. If increases by 10% then increases by %.

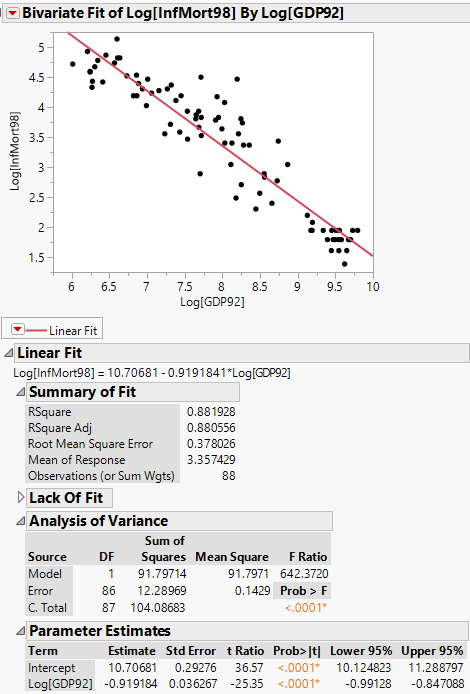
**Example 5.7: Infant Mortality Rate and GDP per Capita**

**(Datafile: Infmort and GDP.JMP)**



**Comments:**

The Bulging Rule suggests lowering the power on and/or . As both have very skewed right distributions we will transform both to the log scale and fit the model.

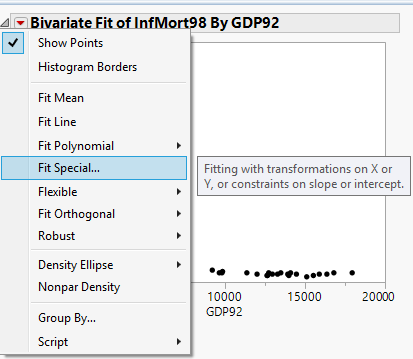
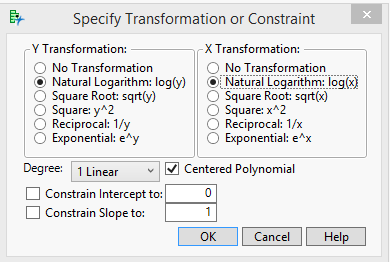


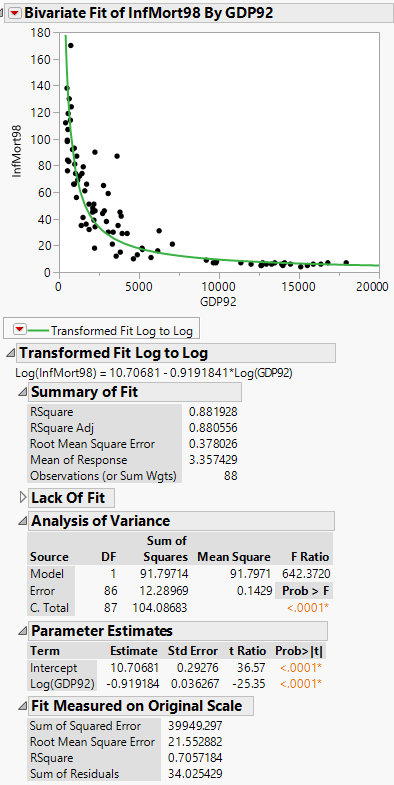
Interpretation of for a unit increase in

Interpretation of for a 10% increase in

Interpretation of 50% increase in GDP

Another way to obtain the regression summary for the Log-Log Model fit to these data is to use the **Bivariate Fit > Fit Special…** option.



Here we can see the log-log fit on a plot of these data in the original scale.

Notice the parameter estimates and other regression summary statistics are all the same as those from fitting the model in the log-scale for the both variables, as they should be.

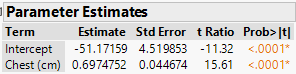
**5.7 – Understanding Sampling Variability in Regression**

From introductory statistics we know that if we are sampling from a population with mean , the sample mean is the *best* estimator of the unknown population parameter (). The theoretical sampling distribution of the sample mean () has standard deviation

We also know that when is sufficiently large that . The standard error (*SE*) is a measure of precision of our estimate () and is used in test statistics and CI’s for making inferences about the population mean (). We saw above that there are also standard error formulae for the parameter estimates and . We can (and will) see where these formulae come from when we consider simple linear regression using matrices in the next section. In general, any statistic used to estimate a parameter has an associated standard error which is the standard deviation of the sampling distribution of that statistic.

The variance and standard errors formulae for the estimated intercept and slope are given below:

Computed for the regression of percent body fat on chest circumference we obtain the following from JMP.



How do we interpret the standard error for the estimated slope

?

To approximate the process of drawing repeated samples from the population, which in practice is never done as we almost always draw a single sample, is to use the ***bootstrap***. **Bootstrapping** in general refers to the process of resampling your original random sample with replacement. A **bootstrap sample** is a random sample of size *n* drawn with replacement from your original sample of size *n*.

For a simple linear regression we can construct a bootstrap sample in of one of two ways which are outlined below.

**1) Case Resampling Bootstrap for Estimating**

Step 0 - Original Random Sample:

Step 1 – Draw Case Bootstrap Sample: each pairs is one of the data points in the original sample. Because we are sampling with replacement some of the original data points will appear in the bootstrap sample more than once and some may not appear in the bootstrap sample at all.

Step 2 - We then fit our model to the bootstrap sample and obtain parameter estimates and .

Step 3 – We then repeat steps 1 and 2 B times, where B is large (e.g. B = 1000). This will give use B estimates of the intercept ( and B estimates of the slope (.

Step 4 – The bootstrap estimate of the standard error for either of these parameters will be the standard deviation of B bootstrap estimates, i.e.

Step 5 – Visualize the distribution of ’s and in the case of simple linear regression with , we can plot the bootstrap sample regression lines to a scatterplot of .

1. **Residual Resampling Bootstrap for Estimating**

Step 0 – Fit a model to the original random sample and save the residuals ().

Step 1 – Draw a Bootstrap Sample of the residuals to obtain . Add these residuals to the original to obtain . Our final bootstrap sample which we will fit a model to is then: .

Step 2 - We then fit our model to the bootstrap sample and obtain parameter estimates and .

Step 3 – We then repeat steps 1 and 2 B times, where B is large (e.g. B = 1000). This will give use B estimates of the intercept ( and B estimates of the slope (.

Step 4 – The bootstrap estimate of the standard error for either of these parameters will be the standard deviation of B bootstrap estimates, i.e.

Step 5 – Visualize the distribution of ’s and in the case of simple linear regression with , we can plot the bootstrap sample regression lines to a scatterplot of .

Note: In general residual resampling will work better than case resampling, however the residual bootstrap requires that the assumed model and is correct! Violations to these assumptions requires we use case resampling vs. residual resampling.

We will now consider some examples.

**Example 5.8 – Percent Body Fat and Chest Circumference**

> Bodyfat = read.table(file.choose(),header=T,sep=",")

> names(Bodyfat)

[1] "Density" "PCTBF" "Age" "Weight" "Height" "Neck" "Chest"

[8] "Abdomen" "Hip" "Thigh" "Knee" "Ankle" "Biceps" "Forearm"

[15] "Wrist"

> lm1 = lm(PCTBF~Chest,data=Bodyfat)

> with(Bodyfat,plot(Chest,PCTBF))

> abline(lm1)  
> summary(lm1)

Call:

lm(formula = PCTBF ~ Chest, data = Bodyfat)

Residuals:

Min 1Q Median 3Q Max

-15.104 -4.123 -0.312 3.767 15.114

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -51.1716 4.5199 -11.3 <2e-16 \*\*\*

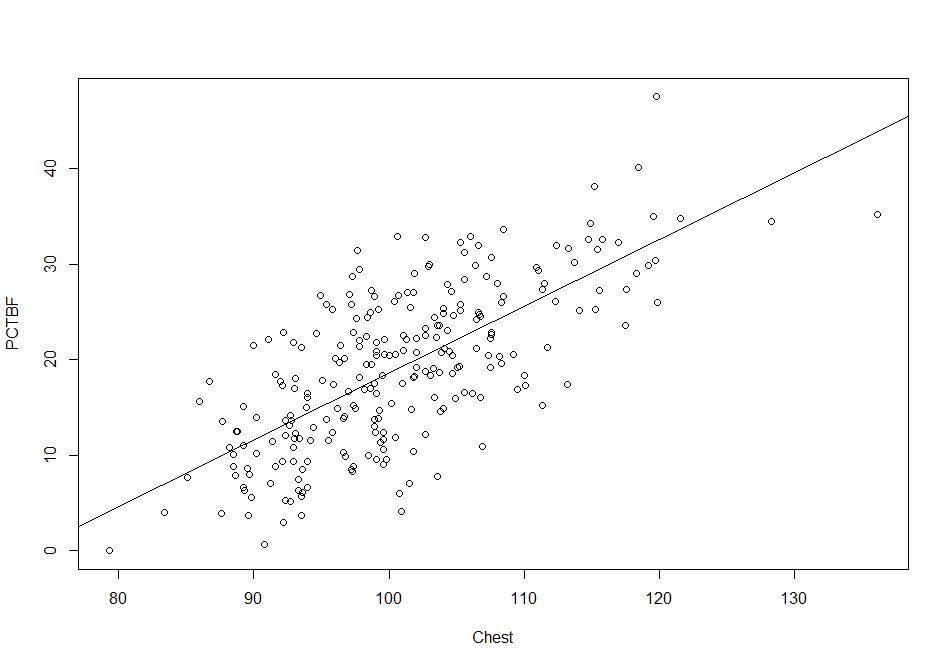
Chest 0.6975 0.0447 15.6 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

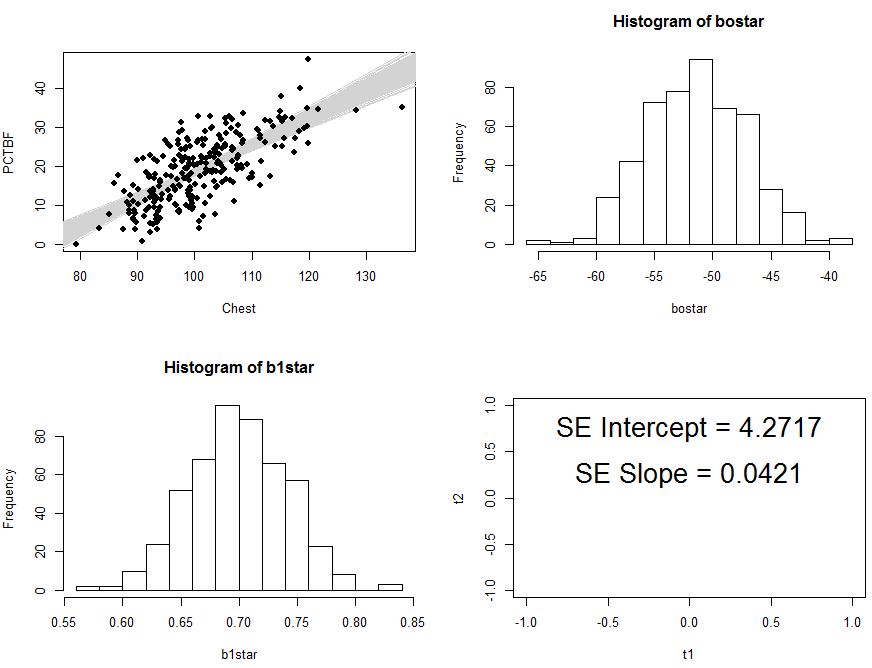
Residual standard error: 5.97 on 250 degrees of freedom

Multiple R-squared: 0.494, Adjusted R-squared: 0.492

F-statistic: 244 on 1 and 250 DF, p-value: <2e-16  


As the model appears correct we will use the residual bootstrap to approximate the standard errors of the estimate regression coefficients.

> BootResid(lm1)



Example 5.9 – Length and Scale Radius of 4-year Old Bass  
> lm2 = lm(Length~Scale,data=Bass4)

> summary(lm2)

Call:

lm(formula = Length ~ Scale, data = Bass4)

Residuals:

Min 1Q Median 3Q Max

-35.08 -8.12 3.24 14.10 22.92

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 102.74 22.22 4.62 0.00048 \*\*\*

Scale 14.66 3.48 4.21 0.00102 \*\*

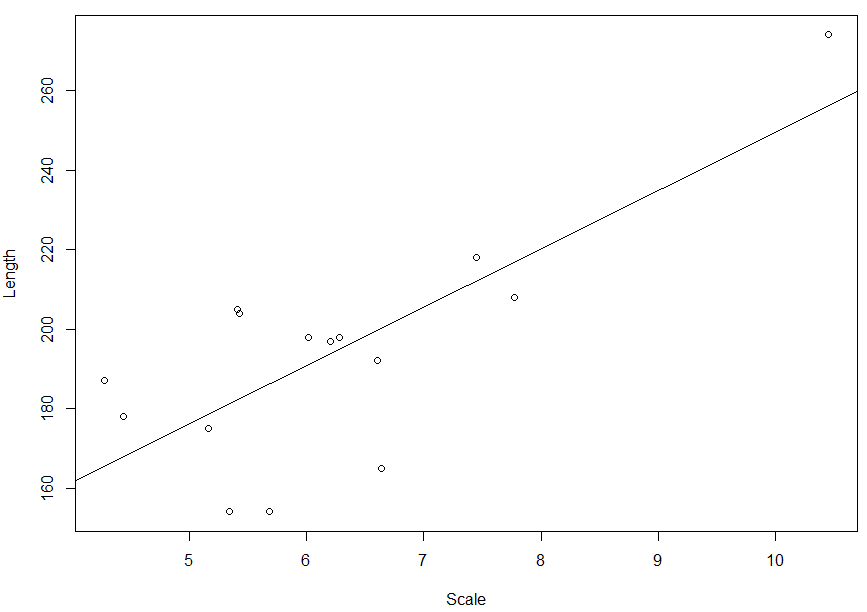
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

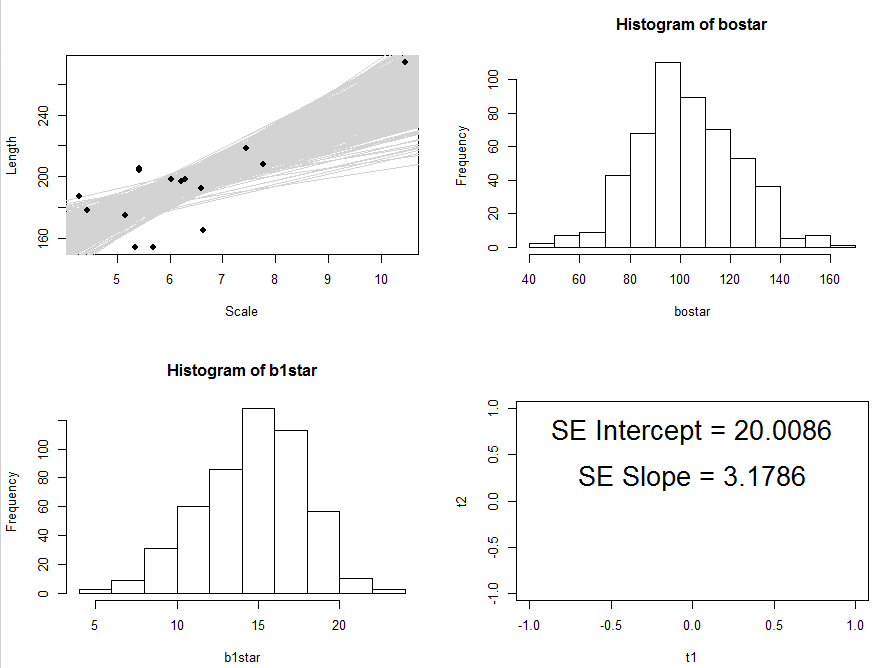
Residual standard error: 19.9 on 13 degrees of freedom

Multiple R-squared: 0.577, Adjusted R-squared: 0.545

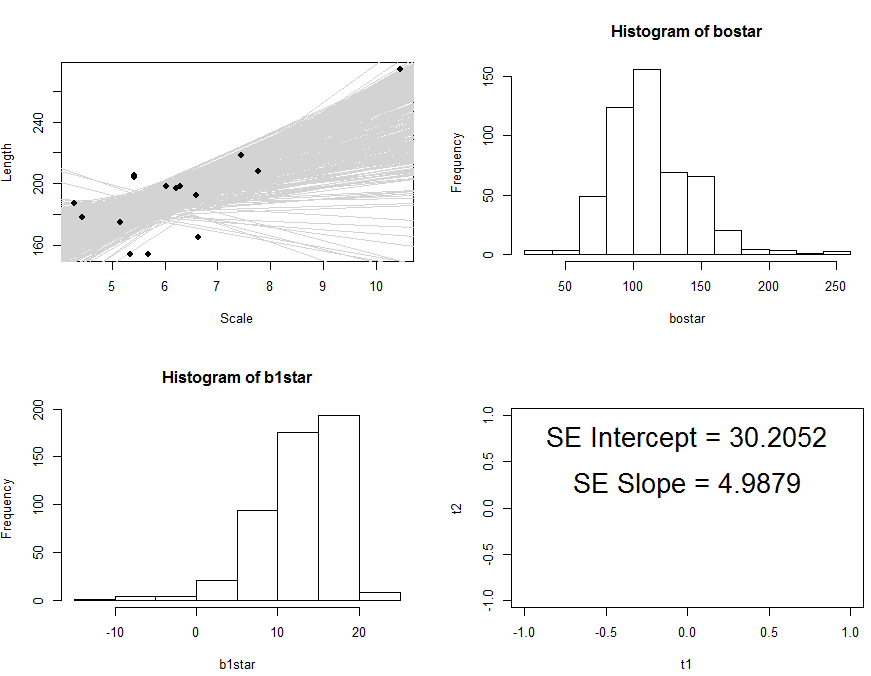
F-statistic: 17.7 on 1 and 13 DF, p-value: 0.00102



> BootResid(lm2)



> BootCase(lm2)

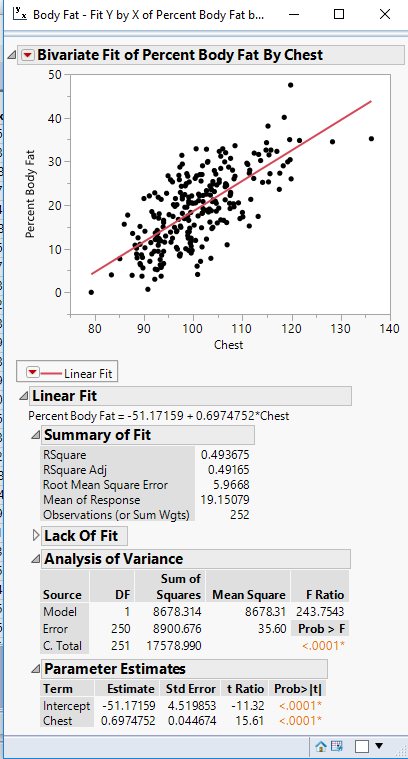


Notice the contrast in the results from the two bootstrap approaches for this example. The main difference is point circled in the plot above. If our case bootstrap sample does not include this point, the regression has the potential become quite different. This illustrates the influence that this case has on the regression. We will examine case diagnostics in Section 7.

**Bootstrapping in JMP**

In JMP you can bootstrap a lot of the statistics that reported in output from an analysis such as regression.

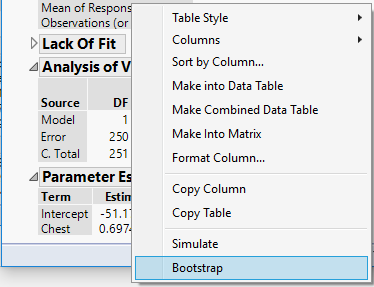
To illustrate consider again the simple linear regression of percent body fat () on chest circumference (). The output from conducting a simple linear regression using the **Analyze > Fit Y by X > Fit Line** approach.

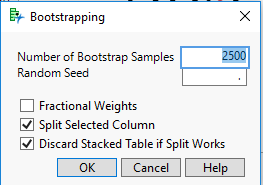


Suppose we would like to draw bootstrap samples from our data and fit the model

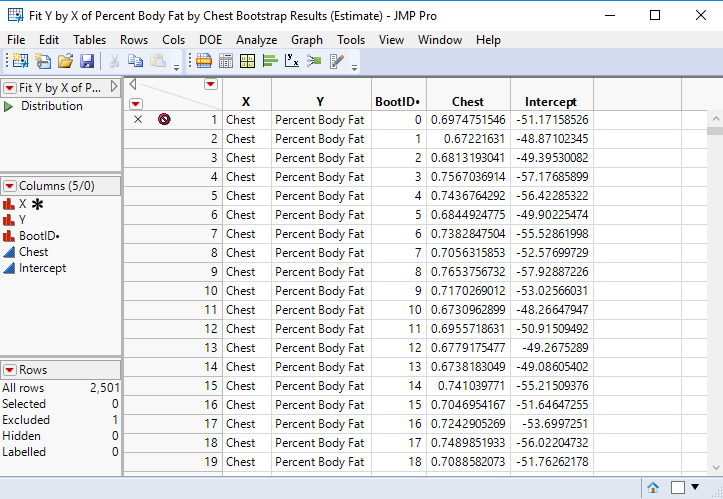
to each bootstrap sample and store the estimated slope and intercept fit to each sample.

This can easily be accomplished by right-clicking on the Estimates in the Parameter Estimates section of output and selecting Bootstrap from the pop-out menu as shown below.

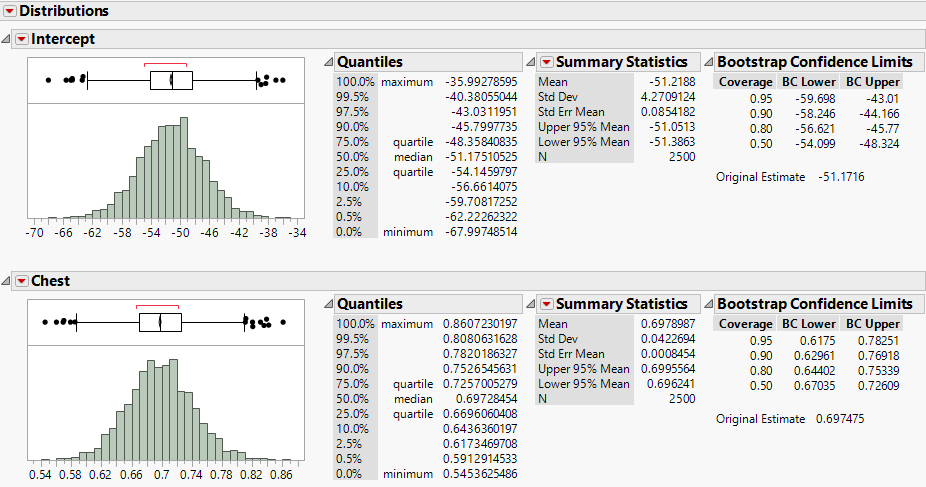


The bootstrap dialog box will allow you specify the number of bootstrap samples which is set to by default.  
  


A portion of the resulting data table containing the bootstrap estimates of the intercept and slope . We can then use **Analyze > Distribution** to examine the bootstrap sampling distribution for these parameter estimates.

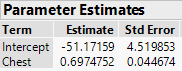


The distribution of the bootstrap parameters estimates for the intercept and slope are shown below.



We can compare the standard deviation of the bootstrap estimates for the intercept and slope to their standard errors ).

From Original Fit From Bootstrap Samples

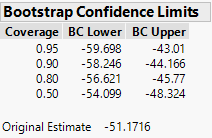
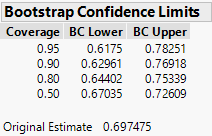
We can also compare the Bootstrap Confidence Limits to those obtained using the standard confidence interval formulae:

**Confidence intervals for and from original fit**



**Bootstrap Intervals for and**

Intercept Slope

The bootstrap is really not needed here, but it is a powerful tool for understanding variability of more complex estimators.

**4-year old bass from West Bearskin Lake**

**Original Fit**



**Bootstrap Results**

