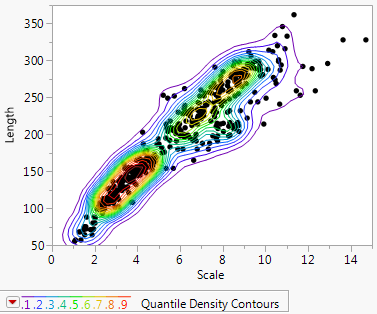
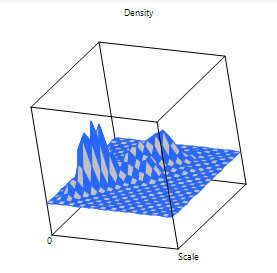
**4 – Bivariate Normality and Regression**

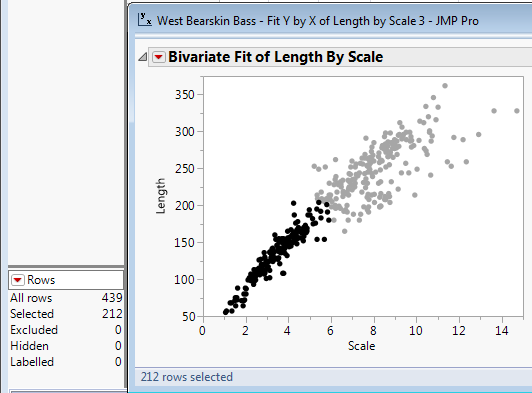
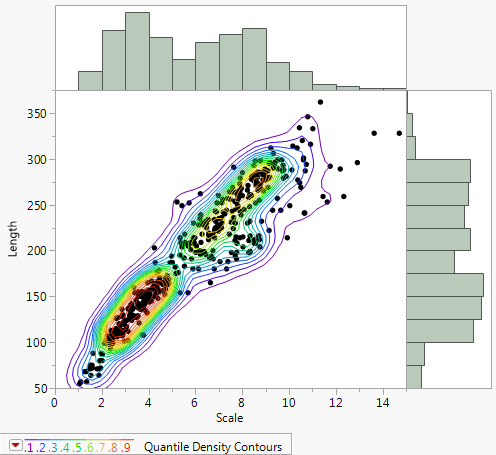
**4.1 – Bivariate or Joint Distributions**

Regression as we have said is the study of conditional distributions and properties of the conditional distribution of Y|X such as mean E(Y|X) and the variance Var(Y|X). Before we return to regression we will consider the joint distribution of the random variables X and Y. A scatterplot of Y vs. X shows how X and Y vary jointly. For example, does it appear that as X increases so does Y and vice versa? Are certain values of (X,Y) more likely to be seen than others? What is ? These are all questions that deal with the joint distribution of (X,Y). We can visualize the joint distribution of (X,Y) by constructing a 2-D histogram.

**Example 4.1 –Scale Radius (X) and Length (Y) of Smallmouth Bass**

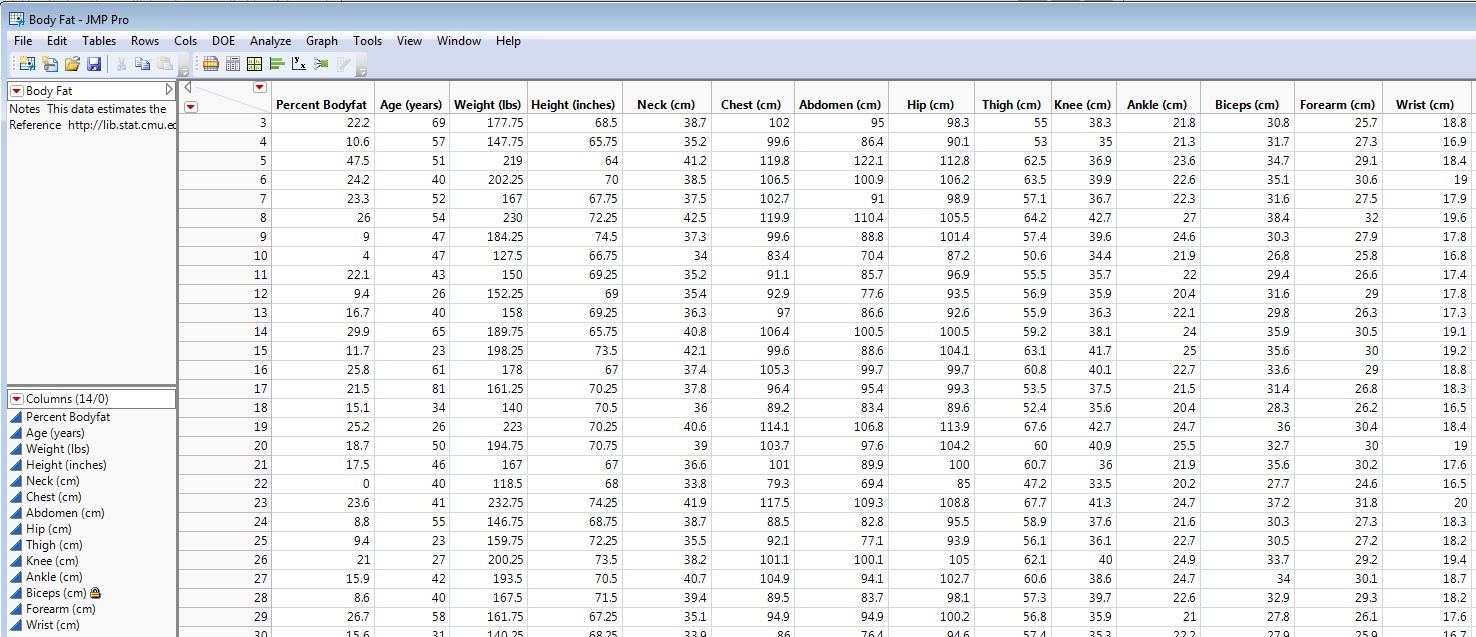
Here we see the joint distribution appears to be bimodal. We can use highlighting to estimate joint probabilities, e.g. .

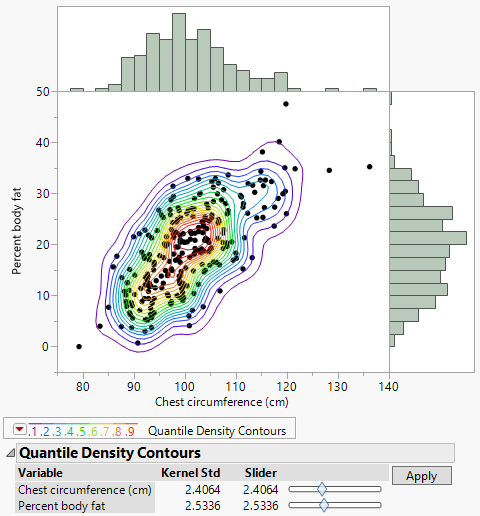
Joint density estimate with marginal distributions

**Example 4.2 –Body fat (%) and Chest Circumference (cm)**

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 and chest circumference.  
 **Datafile: Body Fat.JMP**

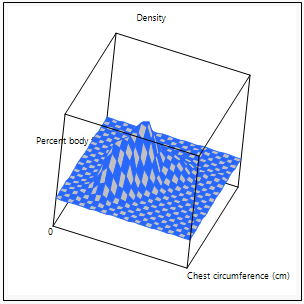


Below is a scatterplot with histogram borders of percent body fat (Y) vs. chest circumference (X) with a joint density estimate added.



How would you characterize the marginal distribution of percent body fat?

How would characterize the marginal distribution of chest circumference?



**4.2 - Correlation**

One characteristic of the joint distribution of two random variables of particular interest in regression is the population correlation (). The correlation measures to what degree two random variables are linearly related. The population correlation is defined as:

The covariance is defined as:

To understand what the covariance is measure consider the diagrams below:

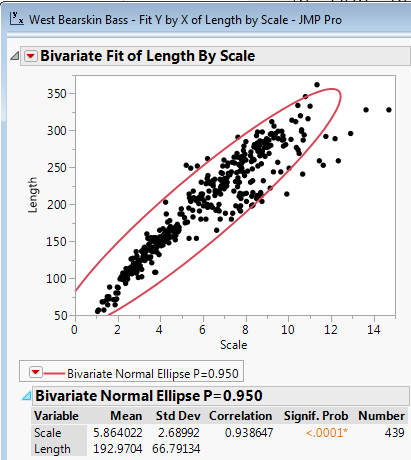
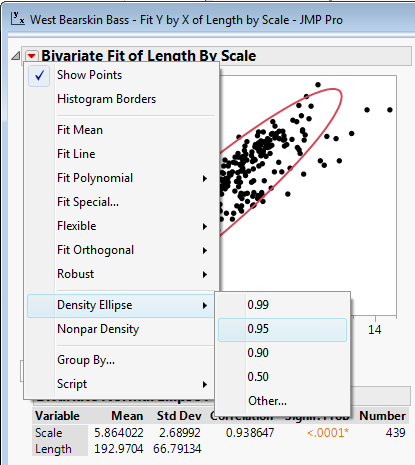
The sample correlation based on random sample of size *n* from the bivariate population is defined to be:

Random sample:

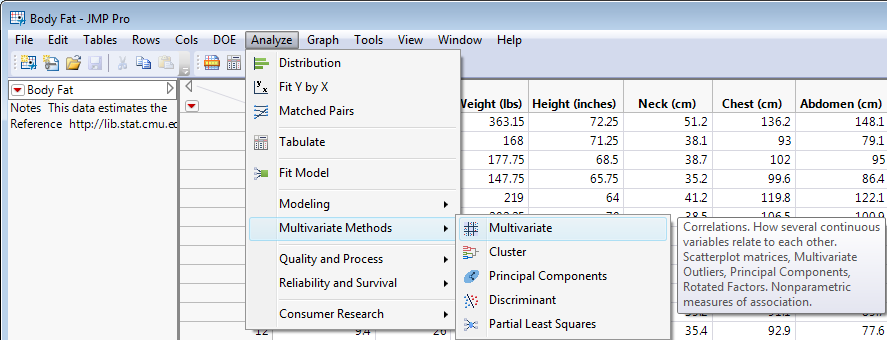
This is called *Pearson’s Product Moment Correlation* and is the default correlation, though other correlation measures exist.

To compute correlations in JMP there are two approaches. From **Analyze > Fit Y by X** select **Density Ellipse** from the Bivariate Fit pull-down menu as shown below as is the resulting density ellipse. The density ellipse is drawn assuming the joint distribution is *bivariate normal* (section 4.3).

**Example 4.1 (cont’d):** Here we see that the sample correlation between scale radius and length is (p < .0001).

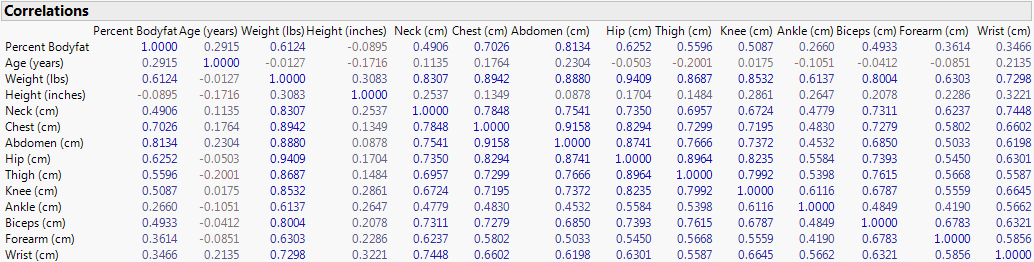


The other way to obtain correlations between numeric variables in JMP is to use **Analyze > Multivariate Methods > Multivariate**. This approach will compute pairwise correlations between a given set of numeric random variables. It also creates a scatterplot matrix show all possible scatterplots for the set of variables.



Examining the scatterplot matrix of the response (Y) and set of numeric predictors (**X**) is an important first step in beginning a multiple regression analysis. We will discuss this in more detail when begin discussing multiple regression. On the next page are the correlations between all variables in the body fat dataset (Example 4.2) along with the associated scatterplot matrix.

**Example 4.2 (cont’d):**  Sample Correlation Matrix for the Body Fat Study

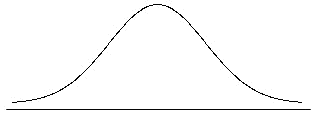


Scatterplot Matrix for the Body Fat Study (with marginal histograms added)

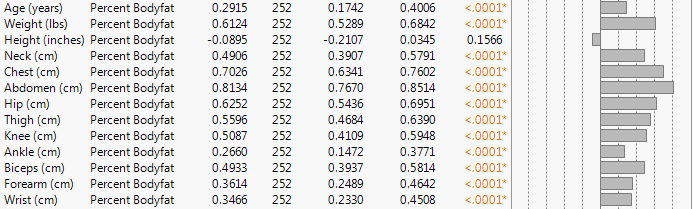


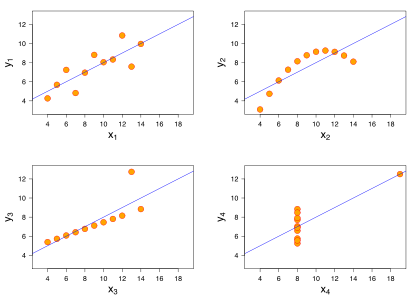
**Notes:**

To test whether the population correlation is statistically significantly different from 0, we can use the following test procedure.



**Example 4.2 (cont’d):** The results of all possible correlation tests for the body fat data are obtained by selecting **Multivariate > Pairwise Correlations** option. The correlations with the response are tested in the output below. A 95% CI for each is also provided.

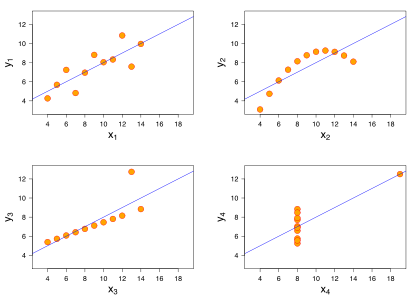
 It is interesting to note that all of the correlations are statistically significantly different from 0 with the exception of height (in.). Why do you suppose that is?

This example presents an opportunity to present some cautionary notes about the correlation as a measure of the strength of the association between two numeric variables. NEVER COMPUTE CORRELATIONS WITHOUT PLOTTING THE RELATIONSHIPS THEY REPRESENT**!  
  
Example 4.3: Anscombe’s Quartet**

All of these plots represent relationships where the correlation .

Furthermore all plots have the same means and standard deviations for the two variables displayed. The only meaningful correlation is the one for the first plot in this sequence. The other three plots all represent things to watch out for when computing correlations (see discussion on the following page).

**Example 4.3: Anscombe’s Quartet (cont’d)**



**Plot 4**

**Plot 3**

**Plot 2**

**Plot 1**

Plot 1 –

Plot 2 –

Plot 3 –

Plot 4 -

Note: We will explore some of these ideas in further detail in Sections 7 and 8.

**4.3 – Bivariate Normal Distribution -**

The joint distribution of percent body fat and chest circumference from Example 4.2 is approximately **bivariate normal**. A **bivariate normal distribution** (BVN) has the following properties:

1. The marginal distributions of both variables are normal distributions.
2. E(Y|X) and E(X|Y) are both **linear** functions of the conditioning variable.
3. Var(Y|X) and Var(X|Y) are **constant**, i.e. they do not depend on the conditioning variable.

More specifically consider the following, let:

1. Then the **marginals** are and .
2. The **mean functions** are:

where,

and

Note: Though not usually of interest the mean of X given Y is given by

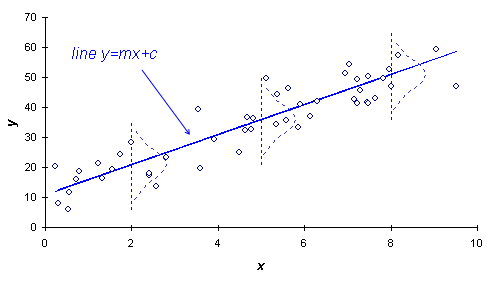
and the intercept and slope for the mean function change accordingly.

1. The **variance functions** are given by:

Note that is the proportion of variation in Y not explained by the   
regression on X. Thus is the proportion of variation in Y explained by the regression on X, i.e. it is the R-square or coefficient of determination. This is the same for the regression of on .

1. Additionally the **conditional distributions** and are normal distributions.

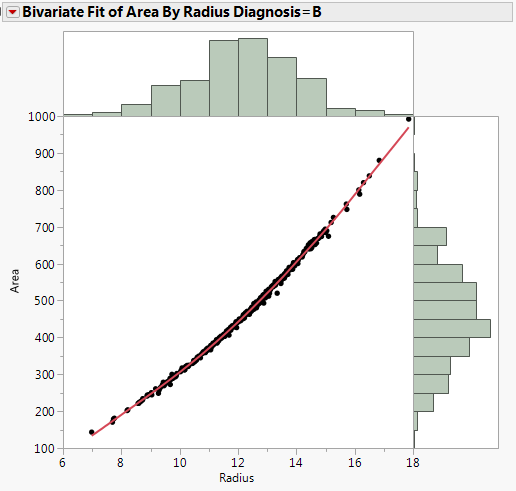
All of these properties (1) – (4) are conveyed in the diagram on the next page.



Mean Function

Thus if it is reasonable to assume that the joint distribution of (X,Y) is bivariate normal then mean function is the equation of line in X, i.e. and the variation function is constant, i.e. .

How do we assess bivariate normality (BVN), i.e. how do I know if the joint distribution of (X,Y) is approximately BVN? We know that both X and Y must have reasonably normal marginal distributions, but that does not guarantee that (X,Y) are BVN. As an example of this consider the joint distribution of (Radius,Area) for benign tumors only in the breast tumor study (see Example 3.4).

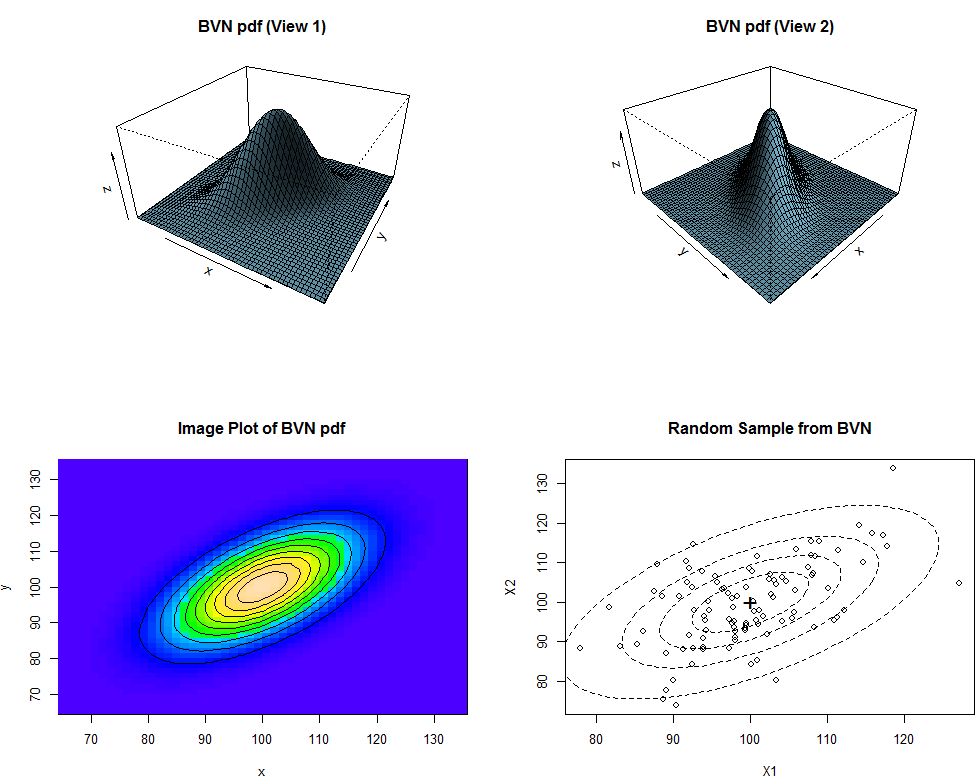


How would we characterize the marginal distributions of cell radius and cell area?

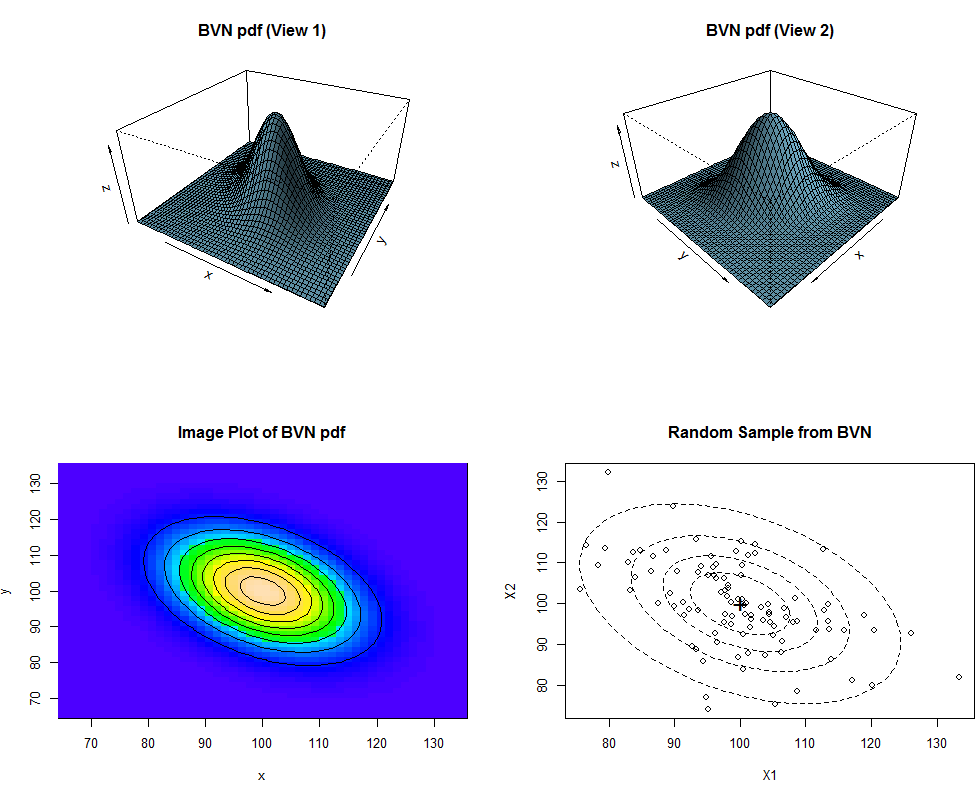
However (Radius,Area) cannot be BVN!   
Why not?

Note: We are actually considering Area|Radius,Tumor Type in this example, i.e. we are conditioning on both cell radius and the fact the cell tumor type = B.

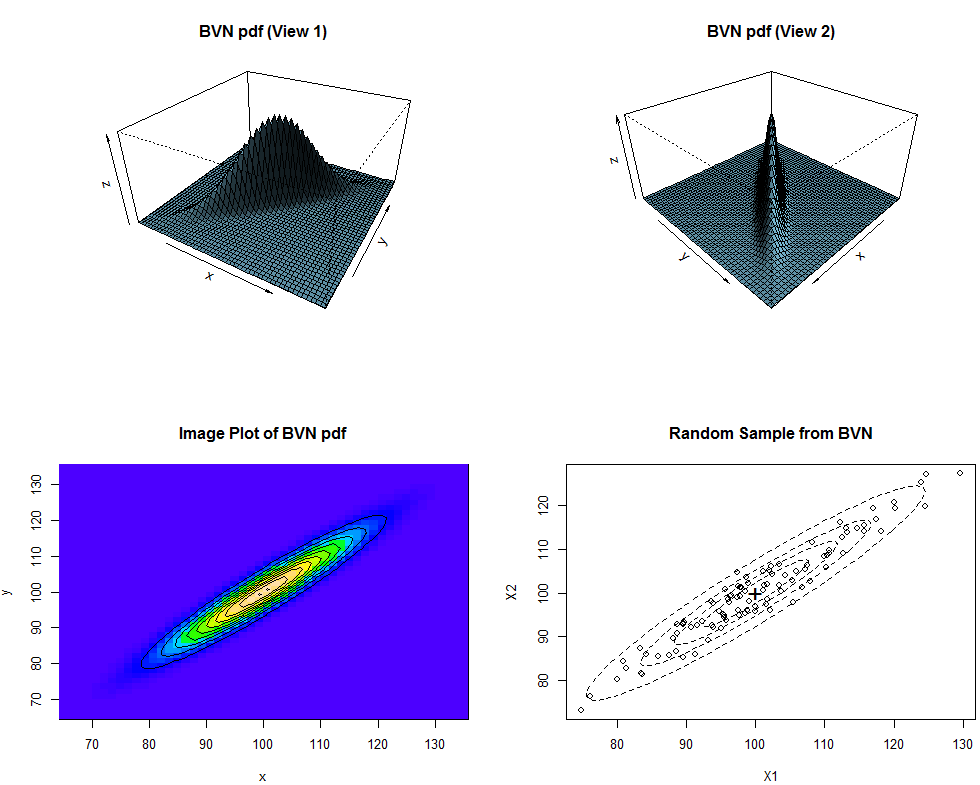
Here are some plots of random variables that show what a joint distribution that is BVN looks like. This function allows you enter means and standard deviations for both X and Y as well as the correlation between them. The function then displays what a theoretical BVN population looks like with those parameter values. BVNplot3d(mux,muy,sigx,sigy,cor) 🡨 need to set 5 parameter values

> BVNplot3d(mux=100,muy=100,sigx=10,sigy=10,cor=0.6)  


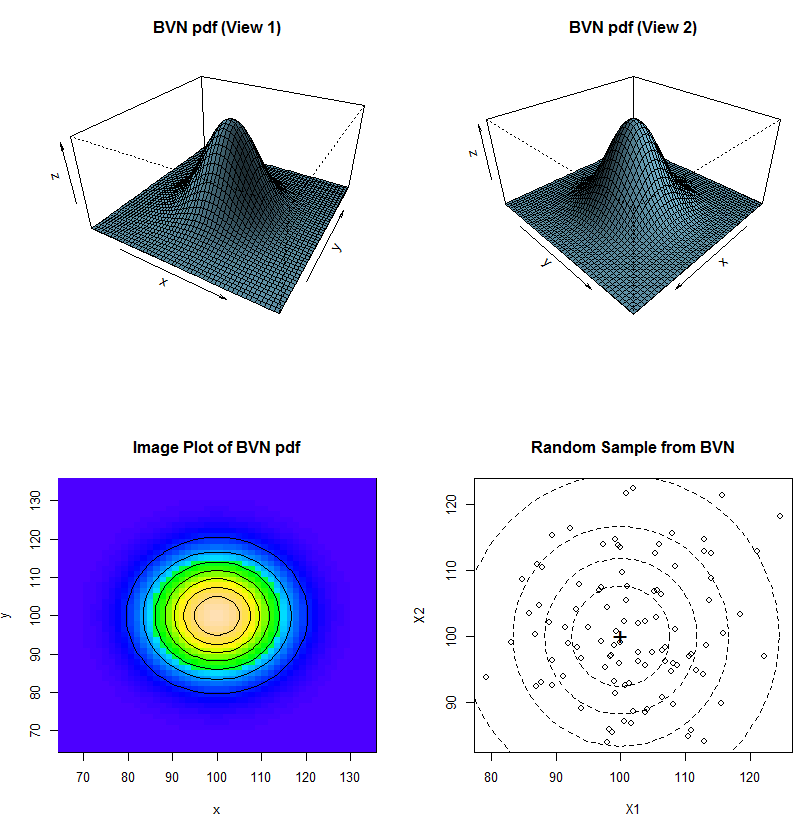
> BVNplot3d(mux=100,muy=100,sigx=10,sigy=10,cor=-.4)



> BVNplot3d(100,100,10,10,.95)



> BVNplot3d(100,100,10,10,0)

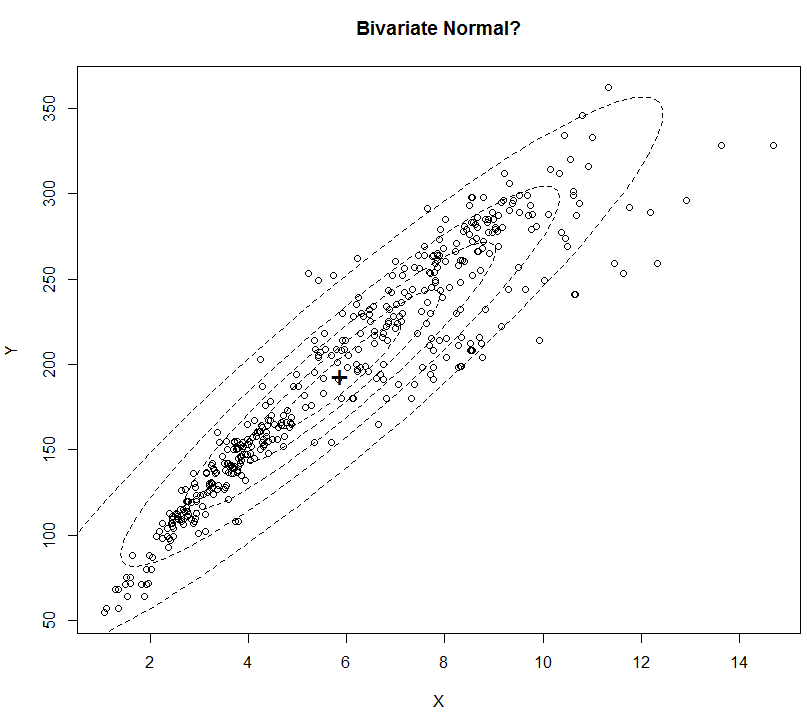


The function BVNcheck takes two numeric variables as input, computes the sample means, standard deviations, & correlation between them and then superimposes density contours of a BVN with those parameter values. This is essentially what the **Density Ellipse** option from the **Fit Y by X** platform in JMP does as well.

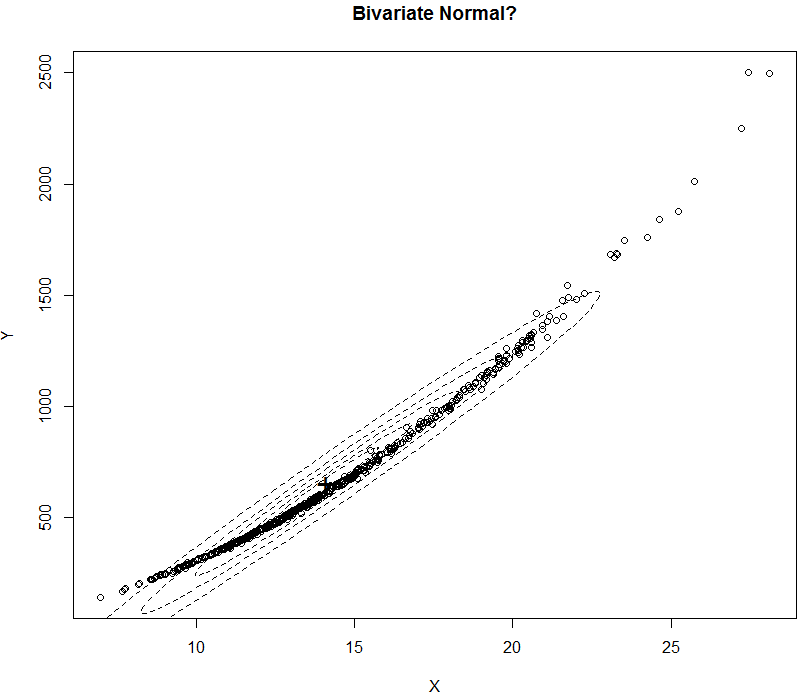
**Example 4.1 – Scale Radius and Length of Smallmouth Bass**  
> names(wblake)

[1] "Age" "Length" "Scale"

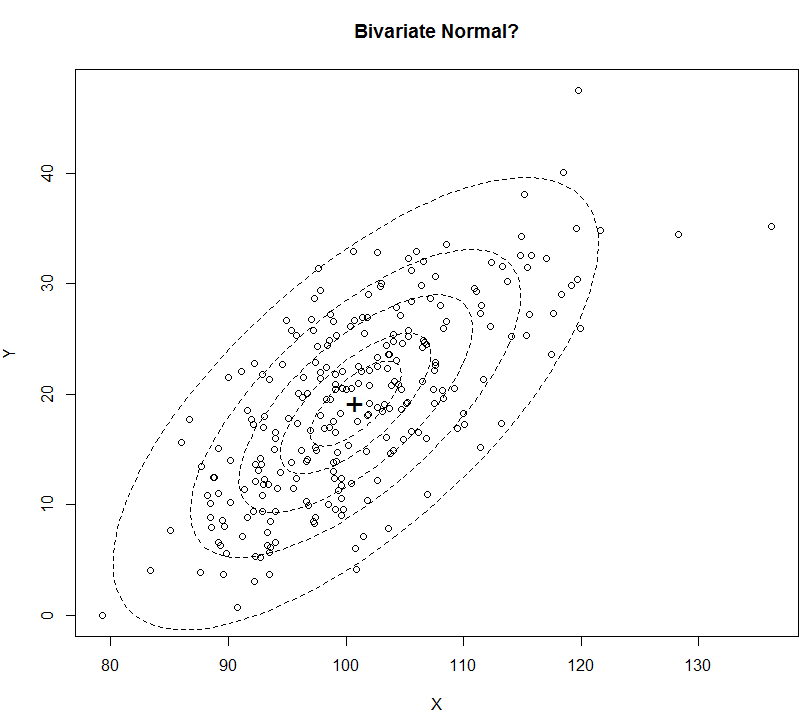
> BVNcheck(wblake$Scale,wblake$Length)



**Example 3.4 (cont’d) – Cell Radius and Area for Breast Tumor Data**> BVNcheck2(BreastDiag$Radius,BreastDiag$Area)



**Example 4.2 (cont’d) – Chest Circumference and Percent Body Fat**> BVNcheck(Bodyfat$chest,Bodyfat$bodyfat)

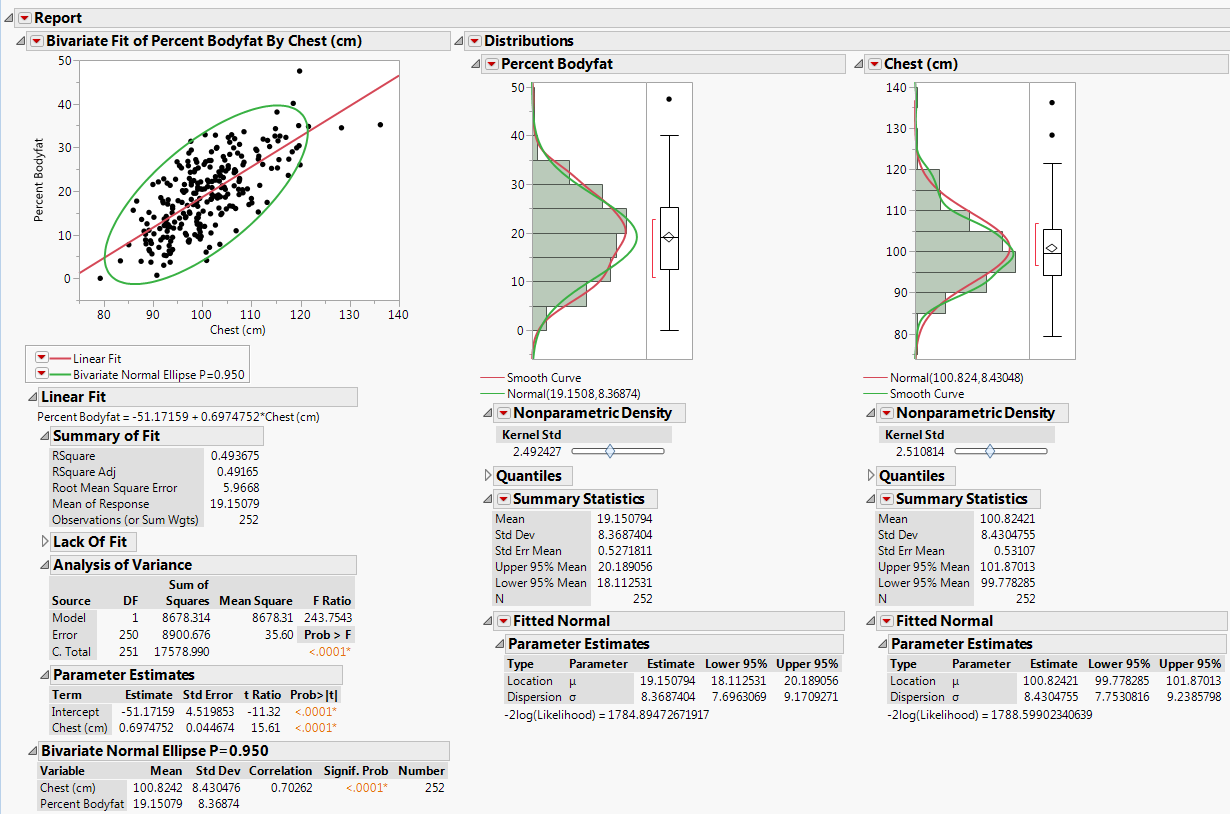


**4.4 – Bivariate Normal Distribution and Simple Linear Regression**

As we have seen above, when we can assume that the joint distribution of (X,Y) is bivariate normal and consider the regression of Y on X we are guaranteed to have a mean function that is the equation of a line in X and that the variance function is constant, i.e.

, i.e. does not depend on the value of .

**Example 4.2 (cont’d): Percent Body Fat (Y) and Chest Circumference (X)**



In discussion in section 4.3 we looked at population versions of the E(Y|X) and the Var(Y|X). Let’s consider sample-based versions of all of these formulae and compute them for the regression of percent body fat on chest circumference (cm).

**Model –** assuming

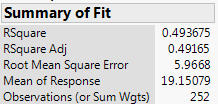
**Data Model**

we are assuming

We will discuss fitting this model in more detail in Section 5, but for now we use the fact we are assuming the joint distribution (Bodyfat,Chest) is BVN to estimate and and all of the parameters involved in estimating these functions.

Key results from Section 4.3:

and

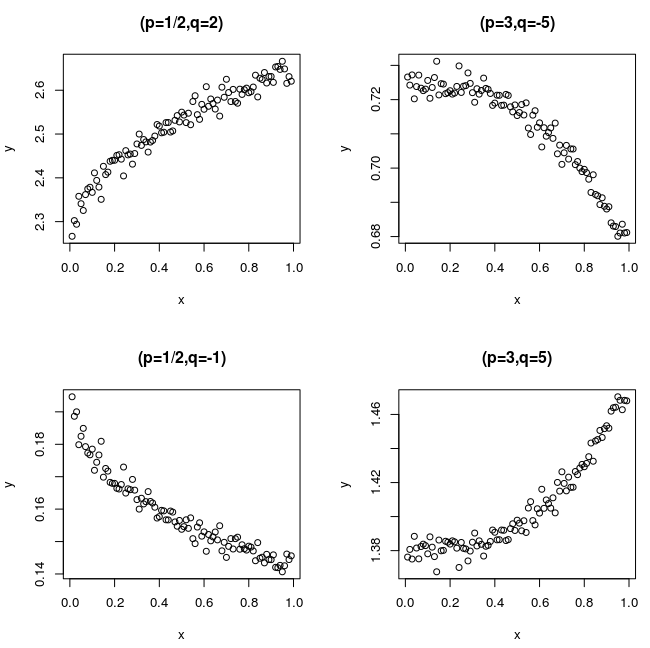
 

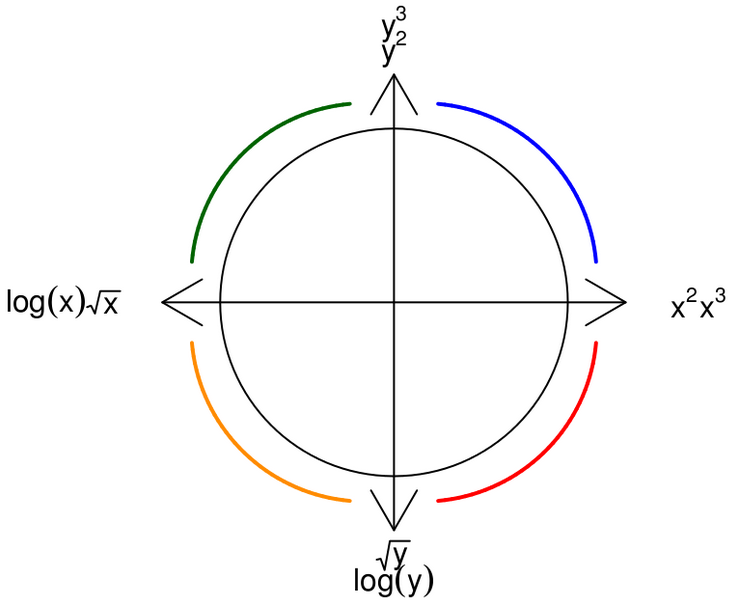
This example shows that when we consider the regression of Y on X and the joint distribution of (X,Y) is BVN everything basically works great! If the joint distribution of (X,Y) is not BVN, can we still consider the regression of Y on X? The answer is yes, but we may need to make changes to our assumed regression model. Also if your bivariate sample does not represent a random sample from a BVN distribution, it is possible a transformation of X and/or Y will result in a joint distribution that closer approximates a BVN.

**4.5 – Bulging Rule (*Tukey & Mosteller, 1977*)**

The Bulging Rule provides guidance in making power transformations of the response (Y) and the predictor (X) to improve linearity and potentially bivariate normality & variance stabilization. The diagram below shows how the Bulging Rule is used.

***Bulging Rule Diagram Sample Scatterplots***





When applying the power transformations you may just transform X, just transform Y or both. Sometimes you apply one transformation and you move diagonally in the diagram from one corner to the opposite. This can happen if you apply to strong of a transformation to one of the variables, i.e. go too far up or down the ladder of powers when transforming.

Here are some general rules of thumb that I employ:

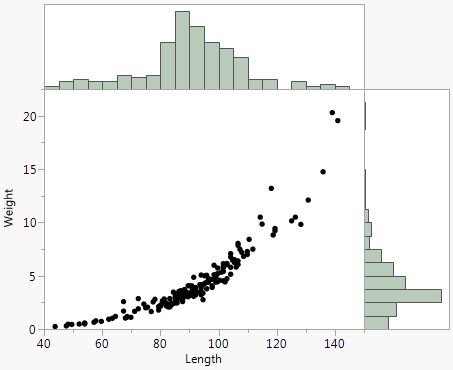
* If you can transform X only, that is generally preferable for interpretation reasons.
* If you see visual evidence that is not constant, then transforming Y instead of X maybe preferable. Typically this will be using either or as the response.
* If you do transform Y to stabilize the variance, you may still need to transform X.
* I generally avoid negative powers.

**A Word of Caution**: Transformations can make interpretation of models where they are used difficult! However, if prediction accuracy is the goal (and not interpretation) then you can generally torture your data at will.

**Example 4.3: Fork Length and Weight of Paddlefish**

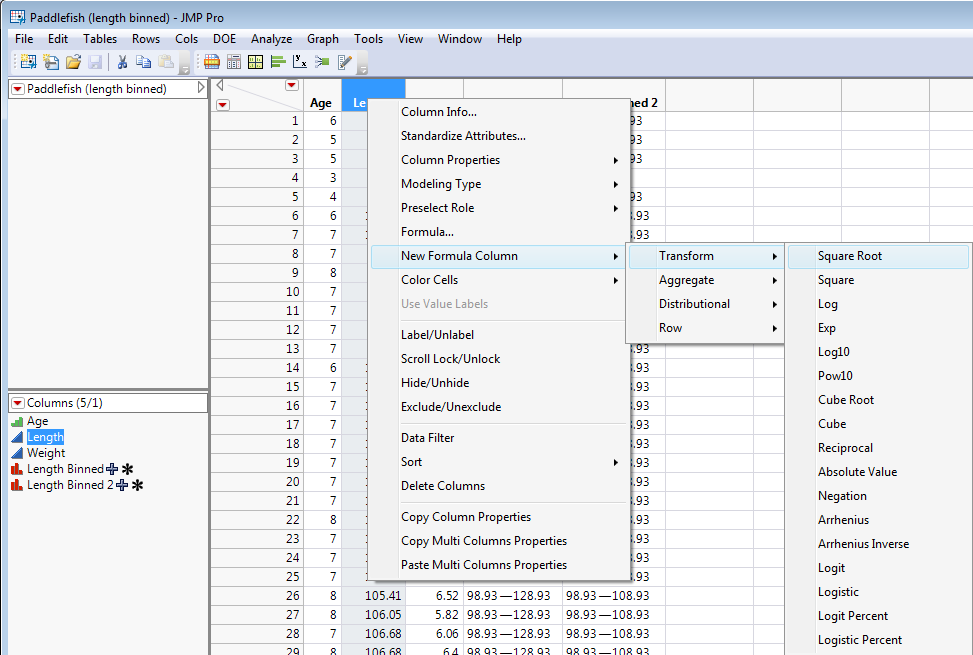
Consider again the regression of weight (kg) on the fork length (cm) of paddlefish in the Mississippi River. Below is a scatterplot with histogram borders. This relationship certainly appears to be a candidate for applying the Bulging Rule.

What guidance does the Bulging Rule give for transforming weight (Y) and length (X)?

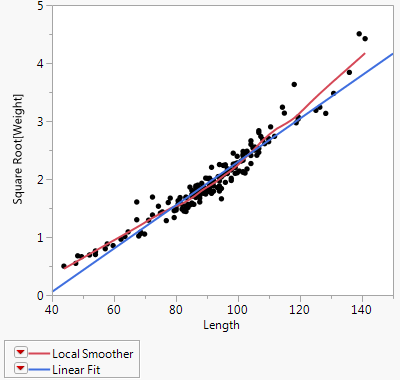
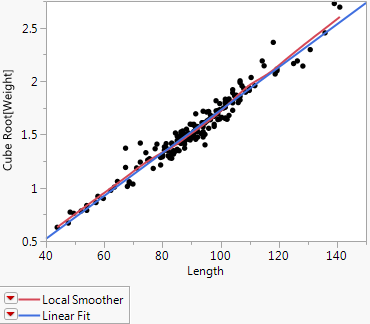


As there is evidence is not constant, I will try lowering the power on the response (Y) to strengthen the linear trend.

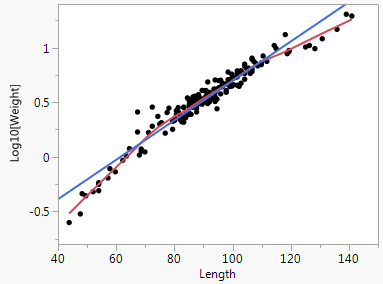
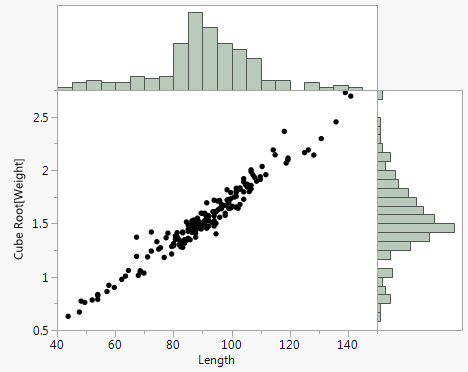
Power transformations of a continuous variables in JMP are easy to form by right-clicking at the top of the column for the variable to be transformed and selecting **New Formula Column > Transform > …** as shown below.



***Attempt 1*** - vs. ***Attempt 2*** -

***Attempt 3*** - & histogram borders

**Bulging Rule and Tukey Power Functions in R**

These functions require that install and load the R packages: manipulate and lattice. You then need to copy the code for these functions and paste them at the R prompt (>) in R-Studio.

tukeyLadder = function(x, q = NULL) {

if (is.null(q)) {

return(x)

}

if (q == 0) {

x.new = log(x)

} else {

if (q < 0) {

x.new = -x^q

} else {

x.new = x^q

}

}

return(x.new)

}

tukeyPlot = function(x, y, q.x = 1, q.y, ...) {

ytran = tukeyLadder(y, q.y)

xtran= tukeyLadder(x, q.x)

y.center = mean(ytran, na.rm = TRUE)

x.center = mean(xtran, na.rm = TRUE)

x.bottom = 0.1 \* (max(ytran) - min(ytran)) + min(ytran)

y.left = 0.1 \* (max(xtran) - min(xtran)) + min(xtran)

xyplot(ytran ~ xtran, panel = function(x, y, ...) {

panel.xyplot(x, y, pch = 19, alpha = 0.2, cex = 2)

panel.loess(x,y,span=0.2)

panel.text(y.left, y.center, paste("y.lam =", q.y), col = "red", cex = 2)

panel.text(x.center, x.bottom, paste("x.lam =", q.x), col = "red", cex = 2)

})

}

bulgePlot = function(x,y,step=0.25){

manipulate(tukeyPlot(x,y,q.x,q.y),q.x=slider(-2,2,step=step,initial=1),  
q.y=slider(-2,2,step=step,initial=1))

}

To use them you simply have to specify a variable x and y that you wish to apply the Bulging Rule to.

> library(manipulate)

> library(lattice)

Read the paddlefish data from the file Paddlefish (clean).csv into a data frame called Paddlefish.

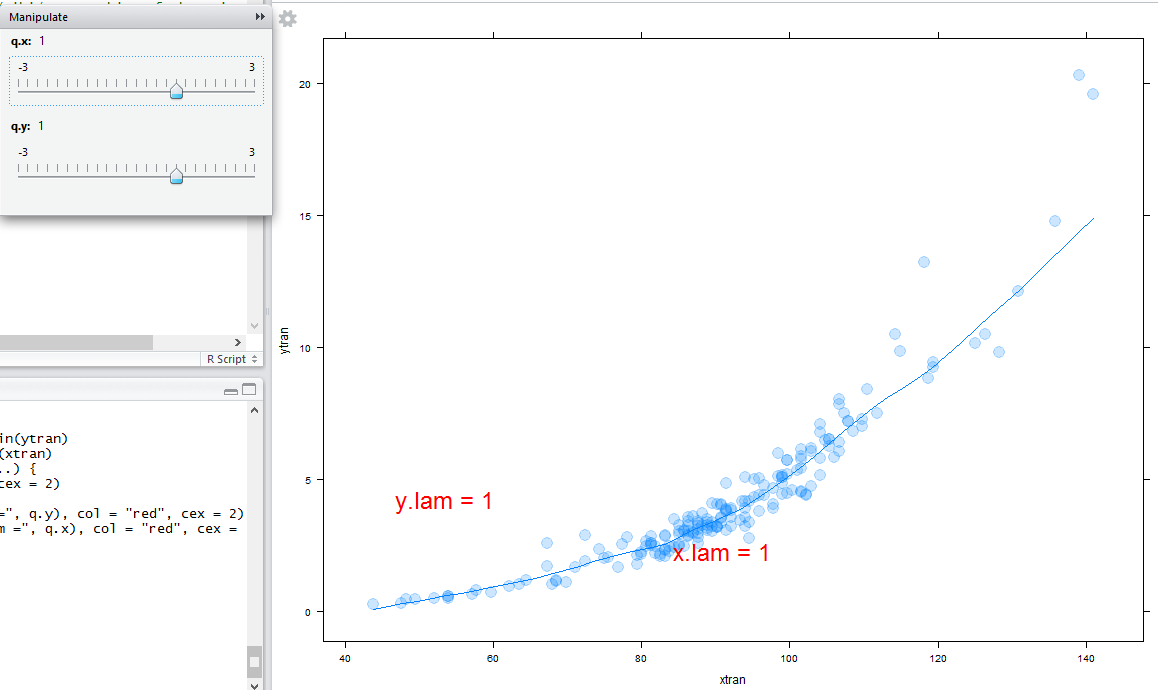
> Paddlefish = read.table(file.choose(),header=T,sep=”,”)

> names(Paddlefish)

[1] "Age" "Length" "Weight"

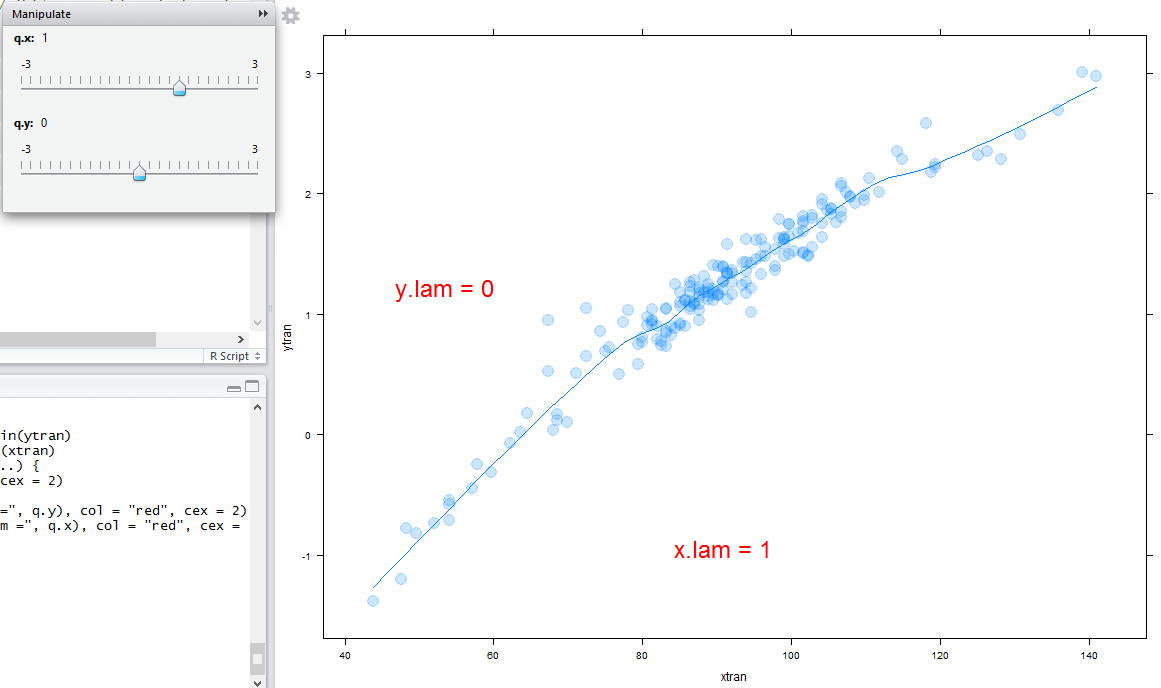
> with(Paddlefish,bulgePlot(Length,Weight))

**Data in the Original Scale (Weight (kg) vs. Length (cm))**



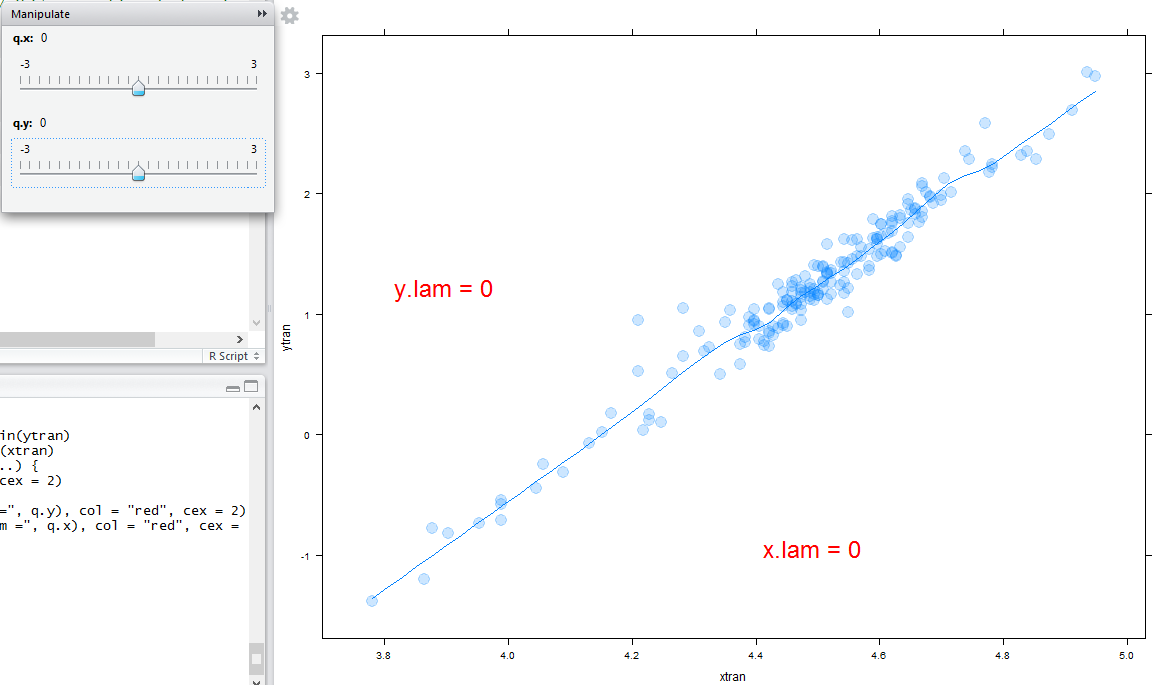
The Bulging Rule suggests lowering the power on Weight or raising the power on Length. As the variation also appears to be nonconstant transforming Weight is probably a good starting point.

**log(Weight) vs. Length (cm)**



Now we see that the bend is the other way, suggesting lowering the power on Weight (i.e. the log was too strong) or lowering the power on Length. As variance appears to be constant in the log scale, we might try lowering the power on Length as a next step. Also transforming Length (cm) to the log scale appears to result in a very linear relationship with constant variation.

**Log(Weight) vs. Log(Length)**



> logWt = log(Weight)

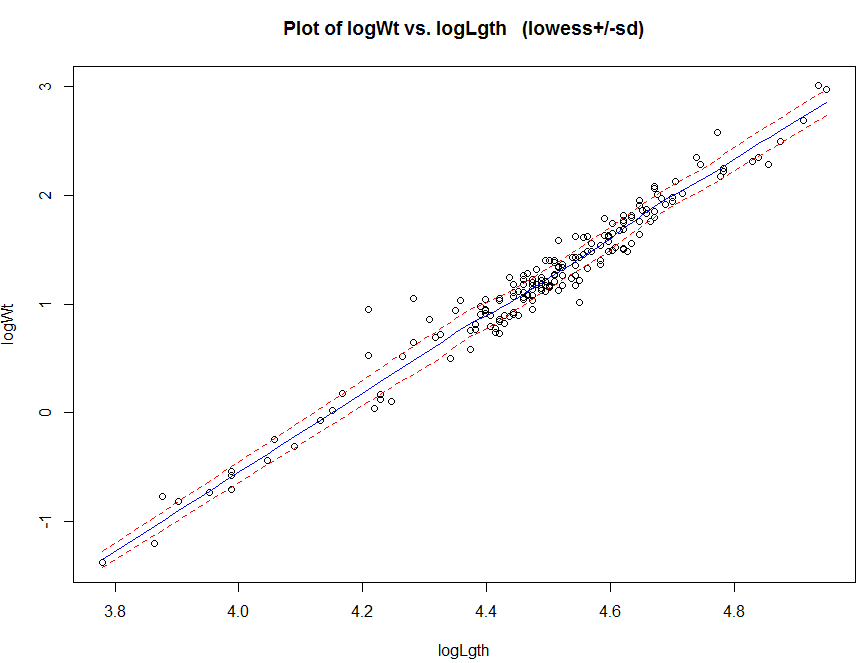
> logLgth = log(Length)

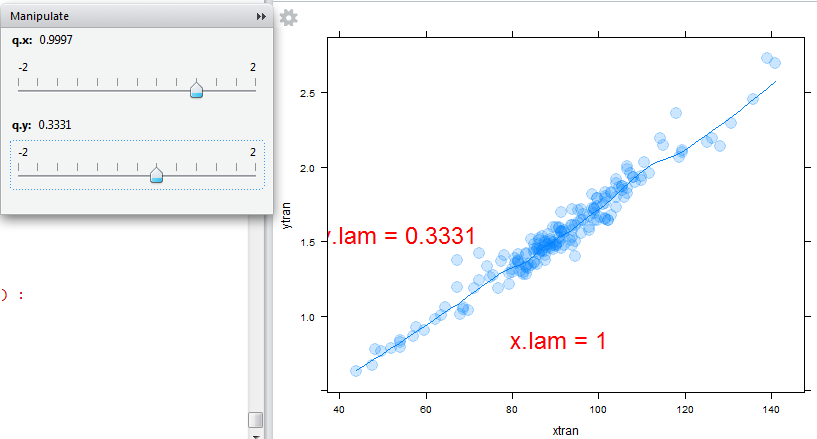
Would you characterize the joint distribution of as bivariate normal? Why or why not?

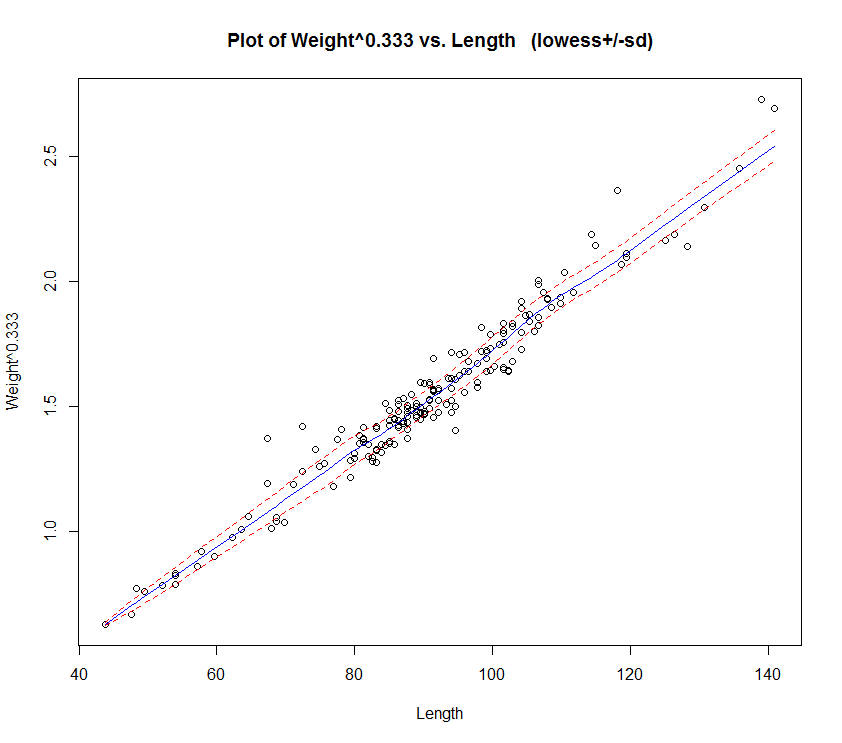
> library(s20x)

Attaching package: ‘s20x’

> trendscatter(logWt~logLgth)



Using JMP we found also looked pretty good.  
  
> with(Paddlefish,bulgePlot(Length,Weight,step=.3333))  
  


> with(Paddlefish,trendscatter(Weight^.333~Length))  
  


**Note:** The R command with the avoids need to attach(Paddlefish). The basic function call is

> with(dataset, function to run) 🡨 the function to be run can refer to variables in   
 the dataset by name.

In the next section we will be putting the concepts and tools introduced in Sections 0 – 4 to use as we begin our discussion of simple linear regression.