**2 – Intro to Regression and Conditional Distributions**

**2.1 – Conditional Distributions**

As discussed in Section 0, regression in general refers to the process of estimating conditional expectations (i.e. means) and variances of a given outcome of interest. This is a bit restrictive, more generally we are interested understanding how the distribution of the response variable () relates to or is associated with a set of variables called predictors (’s). How the mean of the response and the variance of the response relates to the ’s is generally what we consider, but we could also consider how the median or the 75th percentile of relates to the ’s as well.

For simplicity in this section we will restrict our attention to the case where we have single predictor . In this case, regression involves trying to understand how the distribution of the response varies across the subpopulations defined behind the predictor . We call this the *conditional distribution of Y given X*.

Notation:

~ conditional distribution of given

~ conditional distribution of given , where = specific value of

Note: If we write it is always understood we mean , but we will be explicit when were are interested a certain subpopulation where specifically.

Our focus will be on mean and variance of the condition distribution of

Mean and Variance Functions

In regression we first need specify a functional form for the mean and variance functions that we wish use.

~ this is the population mean of the response for the   
 subpopulation where . Other common notations for   
 this would be or

~ this is the population variance of the response for the   
 subpopulation where . Other common notations   
 for this would be or .

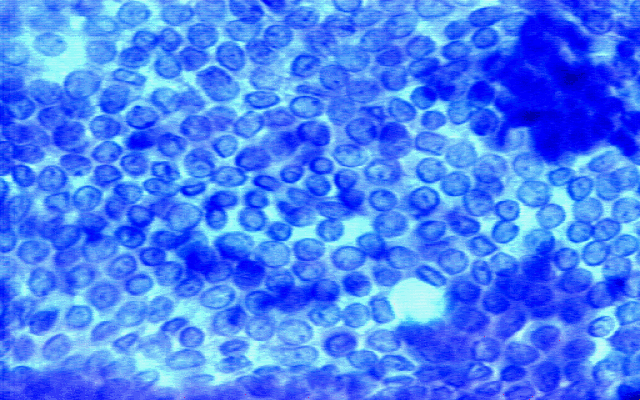
Both of these are functions of in a mathematical sense as will see in Example 2.3 (pg. 21).

**Example 2.1 – Dimensions of Malignant and Benign Breast Tumor Cells**

These data come from a study regarding the use of fine needle aspiration (FNA) to determine whether a breast tumor is malignant or benign. This work is the result of a collaboration at the University of Wisconsin-Madison between [Prof. Olvi L. Mangasarian](http://www.cs.wisc.edu/~olvi/olvi.html)of the Computer Science Department and Dr. William H. Wolberg of the departments of Surgery and Human Oncology who collected and analyzed these data in several published works.

Here is link to website explaining more about this data and the FNA method (<http://pages.cs.wisc.edu/~olvi/uwmp/cancer.html>).

In FNA tumor cells are extracted from a breast tumor using a needle and the cells are then examined under a microscope and a digitized image of the tumor cells is taken. The cells nuclei in the image are then traced using a mouse and the computer then computes twelve size, shape, and texture readings from the traced cells. The final data consists of mean value, SE, and worst case (largest) value for each of these twelve cell characteristics from a sample of 579 tumors (212 Malignant and 357 Benign).

**Data File:** **BreastDiag.JMP**

Variable Descriptions:

* Id – patient ID number
* Diagnosis – M = malignant, B = benign (212 M, 357 B)

Mean of the measured cells on each of the characteristics below:

* Radius – mean of distance from center to points on the perimeter (100 m)
* Perimeter – perimeter of cell
* Area – area of the cell
* Smoothness – local variation in radius lengths
* Compactness -
* Concavity – severity of concave portions of the contour
* ConcavePts – number of concave portions of the contour
* Symmetry – measure of cell symmetry
* FracDim – “coastline approximation” – 1.0

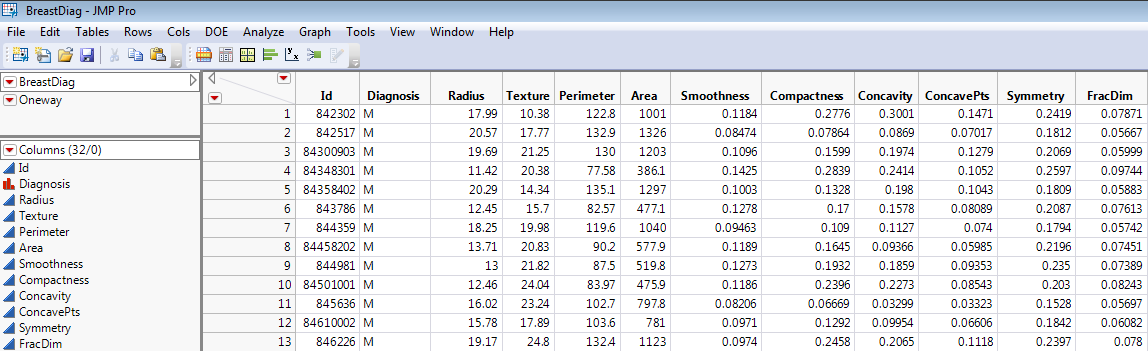
SE of the measured cells on each of the characteristics below:

* serad, seperi, searea, sesmoo, secomp, seconc, seconpts, sesym, sefd

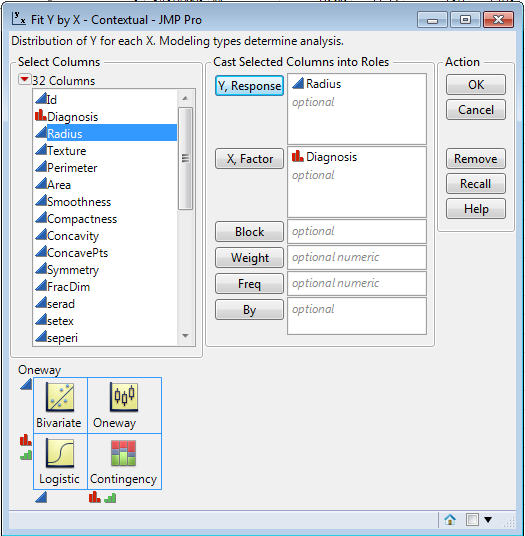
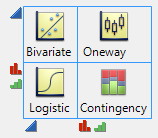
Worst case values of the measured characteristics (mean of three largest values):

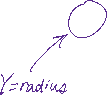
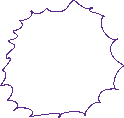
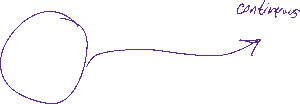
* wrad, wperi, warea, wsmoo, wcomp, wconc, wconpts, wsym, wfd

**Data in JMP – BreastDiag.JMP**

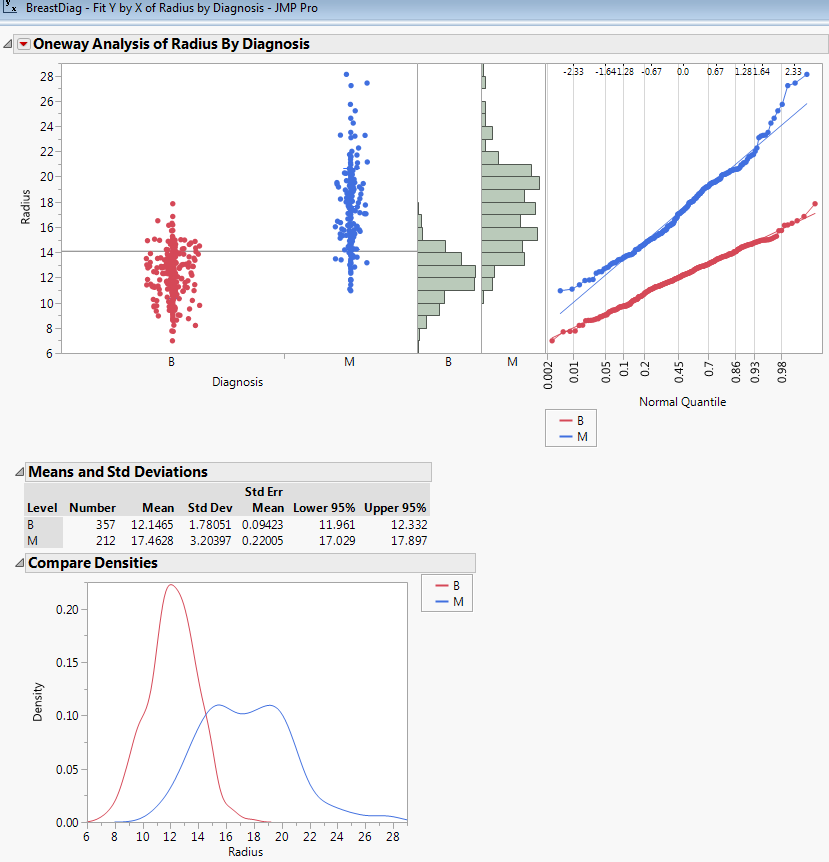


For this example we will use = cell radius as the response and = tumor status (B or M) as the predictor. For the regression of on were interested in understanding conditional distribution of cell **radius** given **diagnosis** of tumor cell (B or M), i.e. *Y|X* or *Radius|Diagnosis.* Select **Analyze > Fit Y by X** in JMP and put Radius in the , Response box and Diagnosis in the , Factor box as shown below.



Here I have selected the following options from the Oneway Analysis pull-down menu: **Means and Std Dev**, **Normal Quantile Plots**, **Densities > Compare Densities**, and from the **Display Options** pull-out menu I have selected **Histograms** & **Points Jittered**.



Radius|M

Radius|B

What can we say about the conditional distribution of *Y|X* or *Radius|Diagnosis*?



**Estimated Conditional Mean and Variances**



*Diagnosis=B) =*



*Diagnosis=M) =*

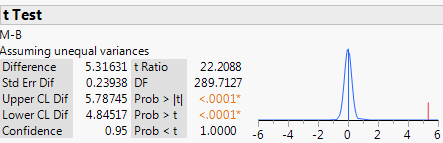
In your introductory statistics course what test would be used to compare the conditional expectations for these two populations? Sounds strange stated this way, but I am essentially asking what test would be used to conduct the following hypothesis test?

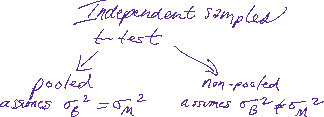
or

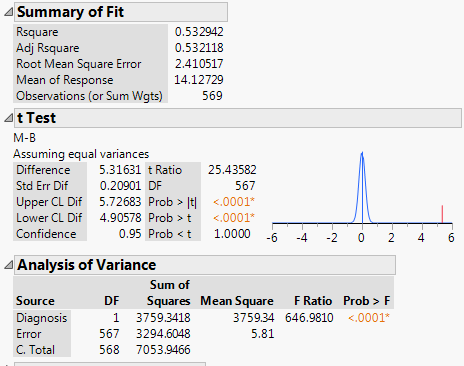
or

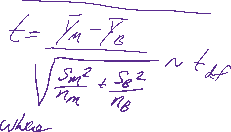
**Notes:**



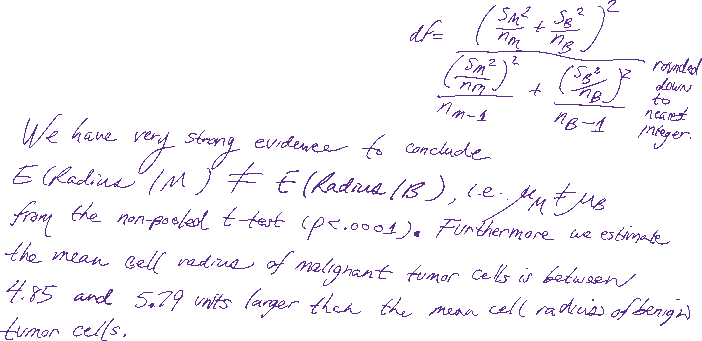








**Notes:**

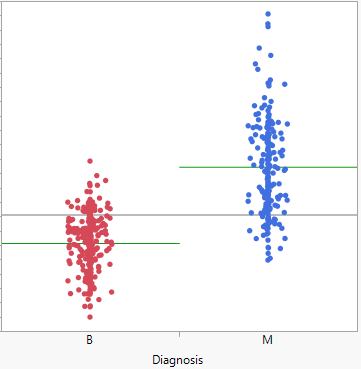
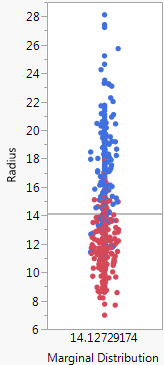


**2.2 – Coefficient of Determination (R-Square**

In the output from the pooled t-test above you will notice in the **Summary of Fit** box there is quantity labeled ***Rsquare = .532942*** or ***53.2942%***. This presents the **variation explained** by conditioning or regressing cell radius on the diagnosis/tumor type. What do we mean by variation explained? The plots below demonstrate this important concept.

**Conditional Distributions (Radius|Tumor Type)**

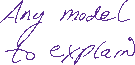
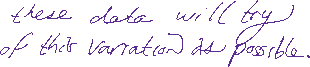
Radius|Benign Radius|Malignant





The total variation in the cell radii in this study is measured by sum of the squared deviations around the grand or overall mean of the cell radii.

This also sometimes denoted as .



By taking diagnosis into account, i.e. regressing cell radii on diagnosis we consider the variation around the mean cell radius for each group.



For benign tumors we have: For malignant tumors we have:



Thus variation left unaccounted for after taking tumor type into consideration is:

.

This is also denoted as = residual sum of squares, thus .

Comparing the total variation to the error variation after taking tumor type into account in the form of a ratio we have:

This represents the percentage of the total variation in the response **not explained** by taking tumor type into account, i.e. not explained by the regression of cell radius (Y) on tumor type (X). Thus the percentage of total variation **explained** by the regression of cell radius on tumor type is given by:

Thus by taking tumor type into account we can explain 53.29% of the total variation in the observed cell radii in this study.

A general formula for the coefficient of determination or R-square is:

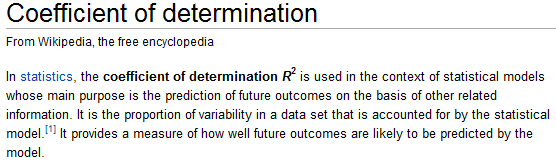


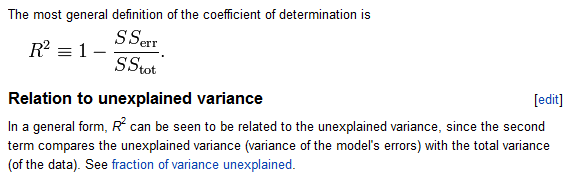
The difference between the total variation in the response and the variation left unexplained by the regression is called the Sum of Squares for Regression (, i.e.

The is the variation explained by regression model.

Wiki entry for *Coefficient of Determination*

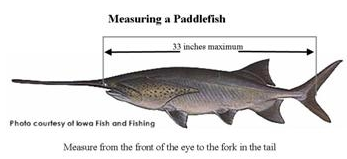
<http://en.wikipedia.org/wiki/Coefficient_of_determination>





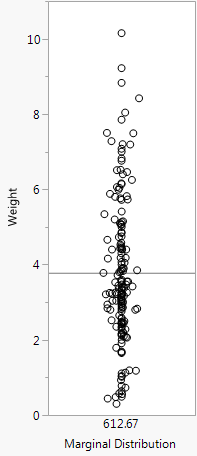
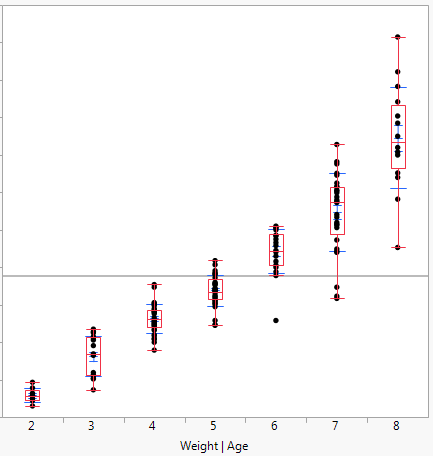
**Example 2.2 – Age (yrs.) and Weight (kg) of Paddlefish**

These data were collected by Ann Runstrom of the Iowa DNR/USFWS. Paddlefish were sampled from three different pools (sections) of the Mississippi River. Each fish sampled was weighed and their fork length measured. Also a determination of their age (yrs.) was made by again looking at growth rings on their scales.



Consider the regression of = weight (kg) on = age (yrs.), restricting our attention to paddlefish 2 – 8 years of age. (**Data File: Paddlefish (2-8).JMP**)

Marginal Distribution (Y) Conditional Distributions Y|X

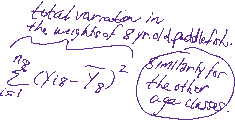
 



Total variation in the response (Y)



The residual variation (RSS) is found by sum the squared deviations from the subpopulation means determined by age (yrs.)



Here,



Thus the for the regression of weight () on age () is,

Thus 86.6% of the variation in paddlefish weight can be explained by the regression on age.

This formula above is a bit clunky and probably feels like it fell out of the sky. Let’s introduce some cleaner notation for our regression of weight on age.

The mean function for our regression model is:

for

The model for a random sample of paddlefish between 2 and 8 yrs. of age is

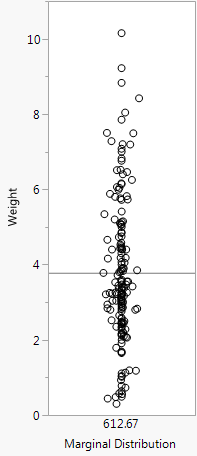
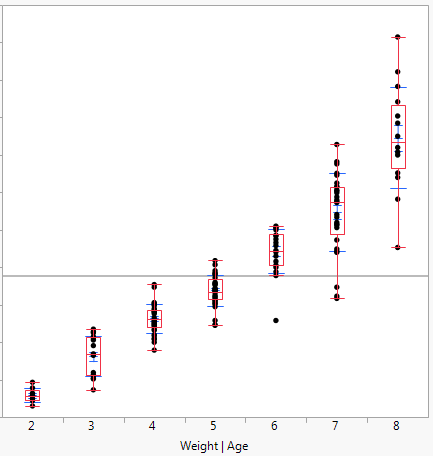
and

Intuitively the estimated mean function is given by

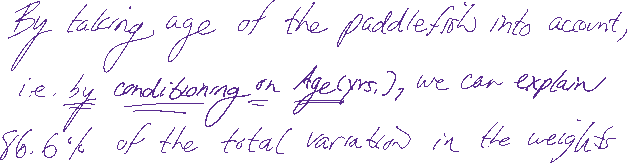
In general, the estimated mean of the response (Y) given the predictors (X) is called the ***fitted value (*** and the difference between the observed response value and the fitted value is called the ***residual ()***.

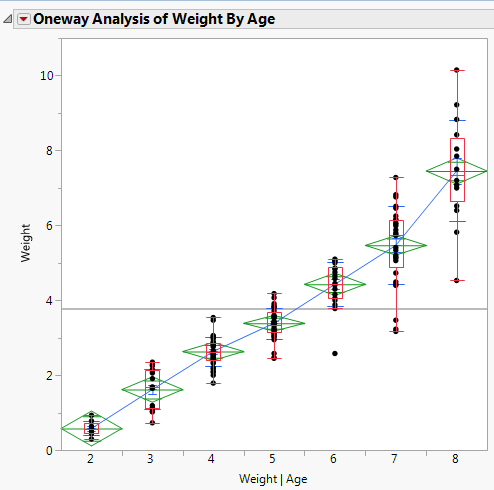
With the understanding that for this example the fitted values are the sample means for the given ages (yrs.) we can express the RSS as

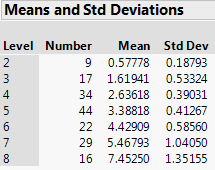
Marginal Distribution (Y) Conditional Distributions Y|X

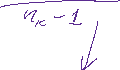
 

**Notes:**

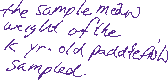
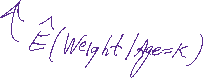






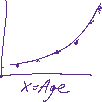


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age (X) | Sample Size |  |  | RSS Contribution |
| 2 yrs. | 9 | 0.578 kg | .0353 | .2826 |
| 3 yrs. | 17 | 1.619 kg | .2843 | 4.549 |
| 4 yrs. | 34 | 2.636 kg | .1523 | 5.027 |
| 5 yrs. | 44 | 3.388 kg | .1703 | 7.323 |
| 6 yrs. | 22 | 4.429 kg | .3429 | 7.202 |
| 7 yrs. | 29 | 5.468 kg | 1.0826 | 30.314 |
| 8 yrs. | 16 | 7.453 kg | 1.8267 | 27.401 |
|  | | | | RSS = 82.10 |



**Additional Questions:**

Does weight increase linearly with age, i.e. is the average weight gain per year constant?



Is the variation constant, i.e.



**2.3 – More on the Mean and Variation Functions:**



One can show that the following relationships between the mean and variance of the response and the conditional distribution of Y|X hold:



|X))

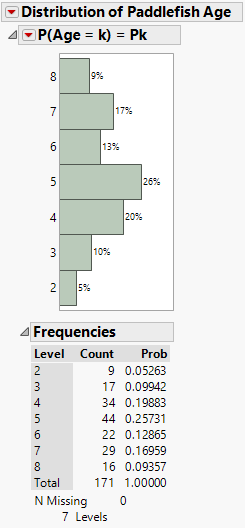
and for the variance of the response we have

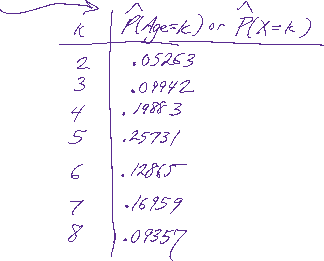
We can demonstrate these relationships using summary statistics and estimates of these quantities from the example above.

From the example above we have

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age (X) | Sample Size |  |  | RSS Contribution |
| 2 yrs. | 9 | 0.578 kg | .0353 | .2826 |
| 3 yrs. | 17 | 1.619 kg | .2843 | 4.549 |
| 4 yrs. | 34 | 2.636 kg | .1523 | 5.027 |
| 5 yrs. | 44 | 3.388 kg | .1703 | 7.323 |
| 6 yrs. | 22 | 4.429 kg | .3429 | 7.202 |
| 7 yrs. | 29 | 5.468 kg | 1.0826 | 30.314 |
| 8 yrs. | 16 | 7.453 kg | 1.8267 | 27.401 |
|  | | | | RSS = 82.10 |

and the estimated distribution of = Age (yrs.) is given by:





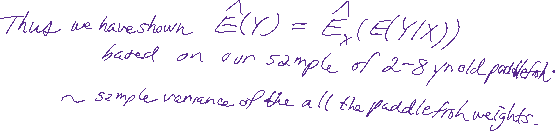
The sample-based version of the expectation relationship is given by:

Here we have,

and



.



For the variance we have,

and

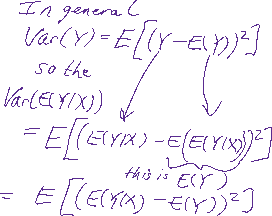


We will handle the two terms on the right-hand side separately.

*First term*



*Second Term*

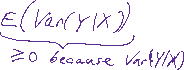


Combining the terms on the right-hand side we have,

The variance result, though more theoretically difficult, is more important. The consequence of this result is the following:



Why?

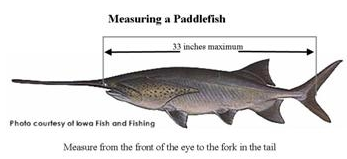


This says by conditioning on we reduce the variation in the response, i.e. by conditioning on we can explain some of the variation in the response . Equality will only hold when is completely independent of , i.e. conditional distribution of given is the same as the marginal distribution of .



**Example 2.3 – Length and Weight of Paddlefish**

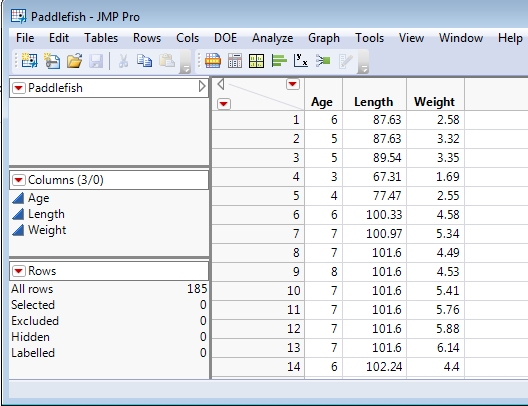
These data were collected by Ann Runstrom of the Iowa DNR/USFWS. Paddlefish were sampled from three different pools (sections) of the Mississippi River. Each fish sampled was weighed and their fork length measured.

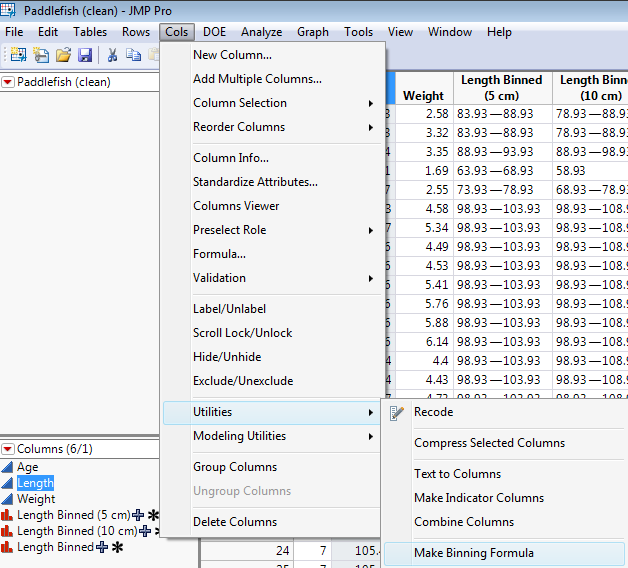


Suppose we interested in the regression of weight in kg (Y) on the fork length in cm (X) of paddlefish in these sections of the Mississippi River, i.e. what can we say about the conditional distribution of *Y|X* or *Weight|Length*. Fork length (cm) is a continuous variable in contrast to age (yrs.) which was discrete. Thus we cannot simply estimate and by simply finding the mean and variance of the weights for all sampled fish with as there may only one fish in our sample with that length.

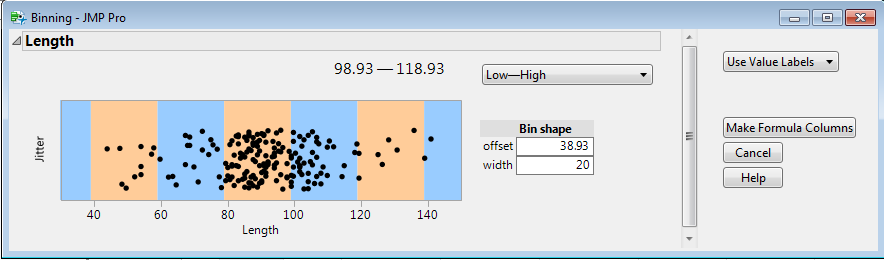
When we have a single continuous predictor a simple scatterplot of vs. will give us a nice visualization of the conditional distribution of and properties of this distribution such as and .

**Data in JMP – Paddlefish (clean).JMP**

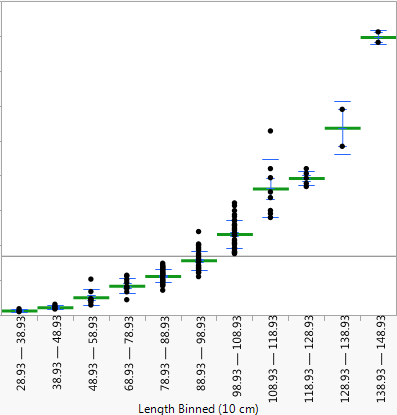
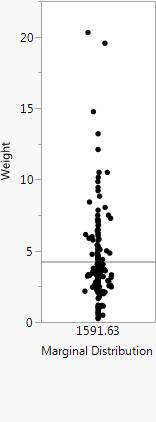


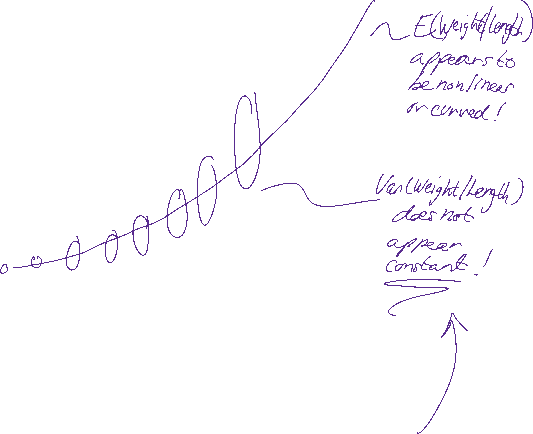
One potential work around would be to condition on ranges of similar lengths and consider the mean and variance of the weights of fish within these length ranges. This can be done easily in JMP using **Cols > Utilities > Make Binning Formula** option after first highlighting the column for the fork lengths of the paddlefish (see below).   


The below is the graphic that appears showing how the ranges of similar lengths will be formed. The width of the bins is set to 20 cm for these data, but it can be changed.

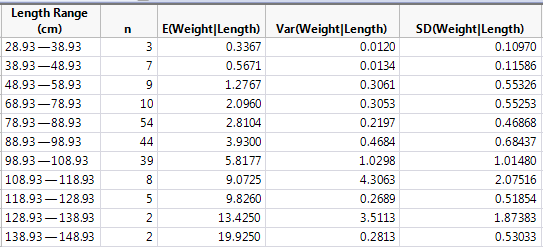


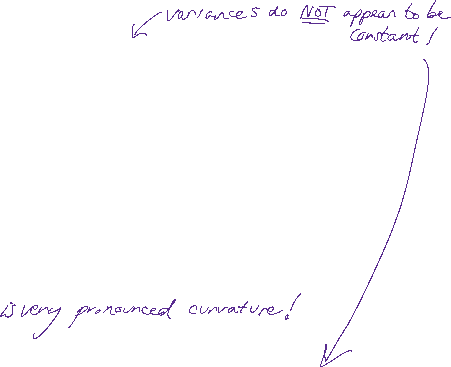
Change the field that says “*Use Value Labels*” to “*Use Range Values*” and then click **Make Formula Columns**. Using a bin width = 10 cm and using **Fit Y by X** to display the results we obtain the following.





We clearly see that by conditioning on fork length we have a sizeable reduction in the residual variation in the response. The by conditioning on 10 cm wide bins of the fork length. The table below gives the estimated mean and variance functions.

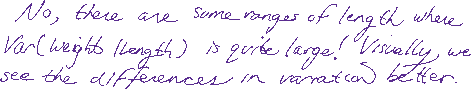




1) Does the mean weight seem to increase linearly with fork length?

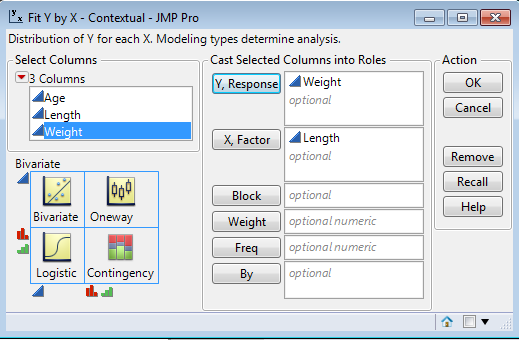
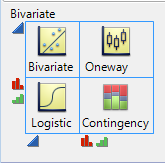


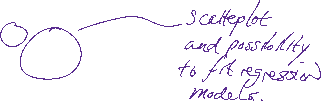
2) Is the variation in weight (kg) similar across the length ranges, i.e. is?



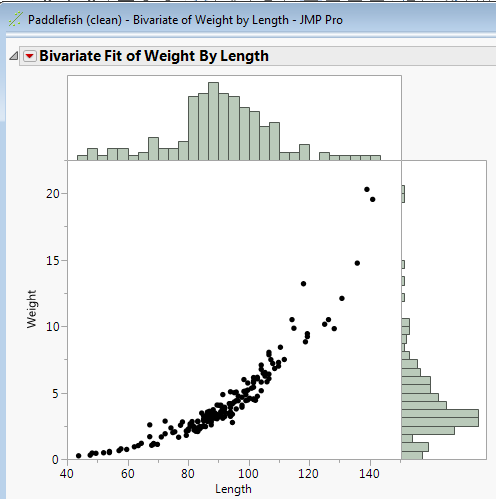
In the next section we will use the idea of “binning” to form *nonparametric* estimates of the mean and variance functions.

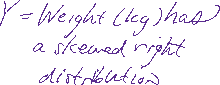
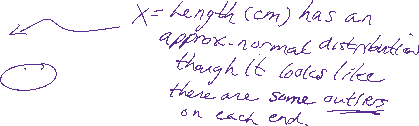
In general, we don’t discretize or use binning when working with predictors that are continuous like fork length in the paddlefish study.

To obtain a scatterplot in JMP select **Analyze > Fit Y by X** and place the response variable in the **Y, Response** box and the predictor variable in the **X, Factor** box.  
 



Below is a scatterplot of the Weight () vs. Fork Length () for these data with marginal distributions of both and added.





What can we say about the marginal distribution of the response?



What can we say about the conditional distribution *Y|X* or *Weight|Length*?

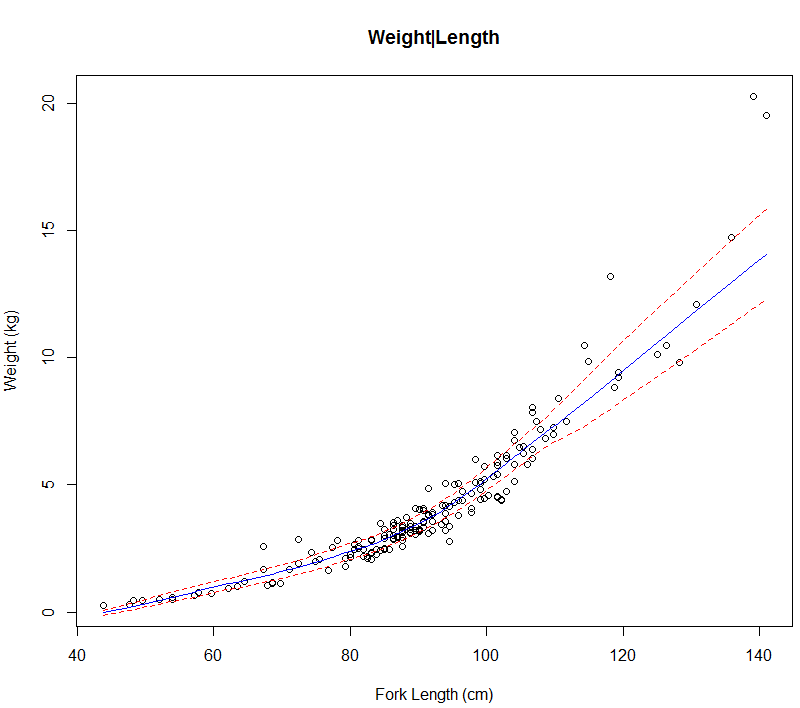


If we focus on the mean and variance functions, what can we say?



Now focusing on the mean function and realizing that this is a function ; what function of might we use for the mean function,

Y = Weight (kg)



X = Fork Length (cm)

Model 1

*E(Y|X) =*



Model 2

*E(Y|X) =*



Model 3



*E(Y|X) =*



As we will see later, we in linear regression or ordinary least squares (OLS) regression we assume the variance function is constant, i.e. *Var(Y|X) = constant.*

Here we have visual evidence that this is not the case.

**Conclusion:**

When developing a parametric regression model we first need to specify the mean and variance functions we wish to use. This is a VERY involved process and the definitely the focus of this course. It also is an iterative process! We usually try several models and choose the “best” model amongst those considered. An important famous quote from George E. P. Box (1987):

  
George E.P. Box

**“*Essentially, all models are wrong, but some are useful.”***