Multiple Regression Analysis

Multiple Regression

- ***Multiple regression*** is used when the analyst is faced with more than one ***X*** (independent or predictor) variable.

Multiple Regression

- The multiple regression equation is an algebraic expression used to describe the relationship between the response (***Y***) and multiple predictor (***X'***s) variables.

Multiple Regression

The Multiple Regression Equation

predictors

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + \varepsilon$$

intercept

slopes

error

Multiple Regression

- Example of using multiple regression:

  Suppose that a production engineer believes that increased **head box consistency** and increased **calendar pli** will increase **mullen**.

  Data is gathered during production of 12 jumbos.

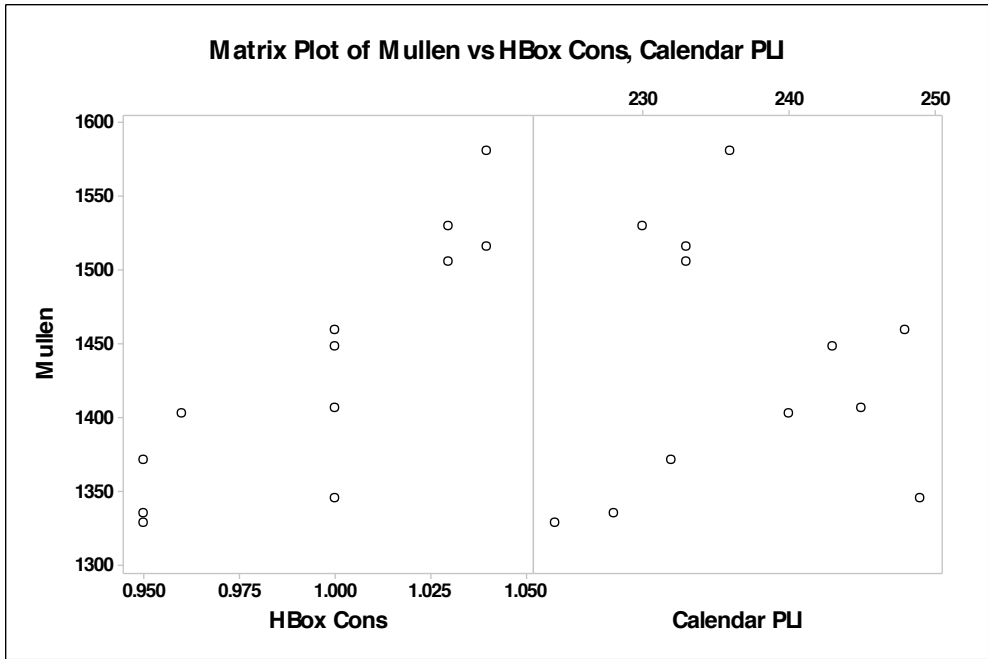
  - This was just an informal trial, not a designed experiment.

Multiple Regression

- Import ***Minitab Dataset 10.xls*** Excel file.
- Suppose we want to understand the relationship that ***Calendar PLI*** and ***HBox Cons*** (the ***X's***) have with each other and whether they are related to ***Mullen*** (the ***Y***)?
- Graphing and multiple regression analysis can help us understand this relationship.

Multiple Regression

- The first step is to examine the variables using a **Matrix Plot, Each Y versus each X**.
- Create a Matrix Plot of the variables.
- Thoughts?

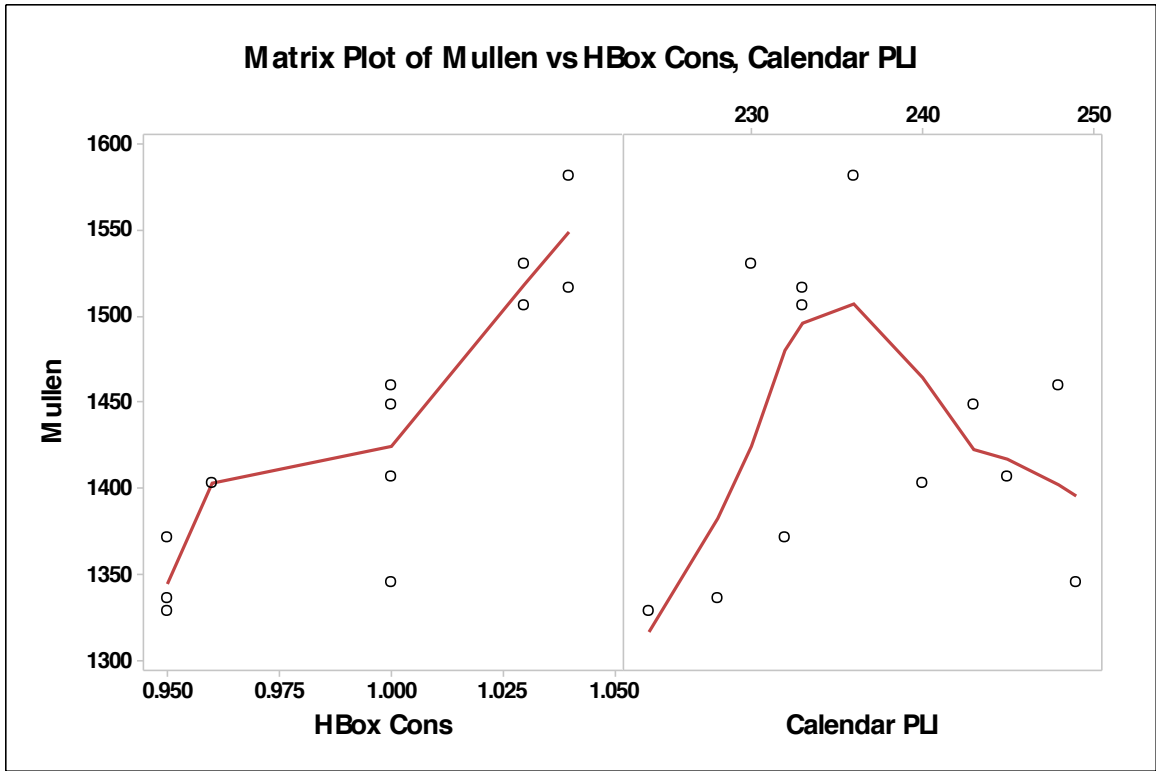


Multiple Regression

- Sometimes a trend line can help visualize the relationship
  - A Lowess smoother is preferred.
- To add a lowess smoother to the Matrix Plot **right click** on the graph → select **Add** → **Smoother…** → click **OK**.
  - The defaults are fine.

Multiple Regression

- Interpretation?

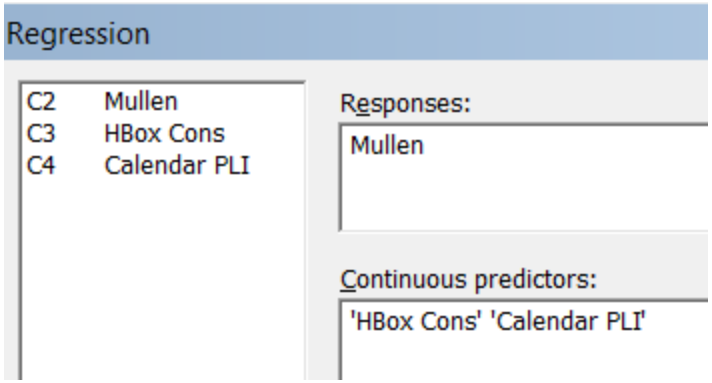


Multiple Regression

- To do a regression analysis in Minitab select **Stat** → **Regression** → **Regression** → **Fit Regression Model…**
- Minitab calls the ***Y*** variable the **Response**.
- Minitab calls the ***X*** variables the **Predictors**.

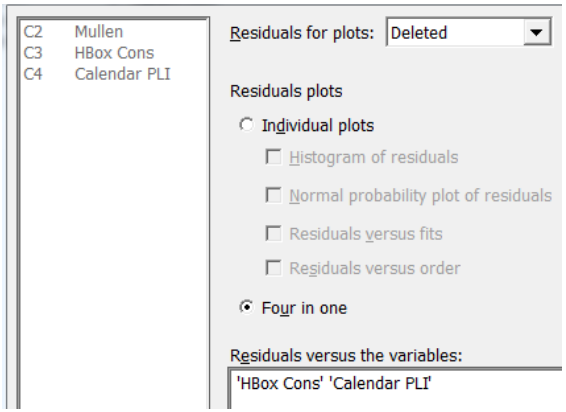
Multiple Regression

- Choose ***Mullen*** as the **Response.**
- Choose ***HBox Cons*** and ***Calendar PLI*** as the **Predictors**.



Multiple Regression

- Click the **Graphs…** button → select **Deleted** as the **Residuals for Plots**
- Select **Four in One** to plot.
- Select ***HBox Cons*** and ***Calendar PLI*** for the **Residuals versus the variables** → click **OK** once.



Multiple Regression

- Multiple regression analysis requires additional model checking/assessing and diagnostics.
- First, we will discuss some of model assessment statistics.

Multiple Regression

- One concern is the potential correlation between the ***X's*** (called multicollinearity).
  - Multicollinearity can cause bad coefficient estimates.
- The <u>variance inflation factor</u> (***VIF***) is used to check for multicollinearity effects associated with each predictor ***X***.

VIF > 10 indicates severe multicollinearity.

Multiple Regression

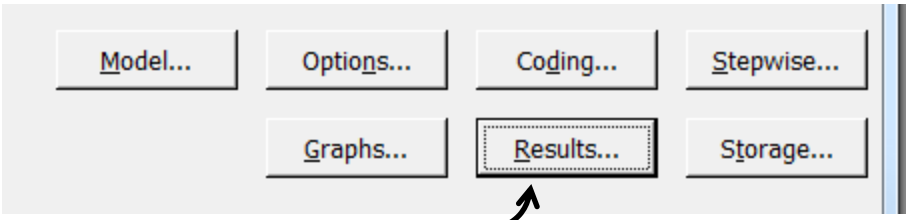
- When you are building a prediction model you need a statistic to tell you if one model is better than another model.
- ***PRESS*** assesses your models predictive ability. In general, the smaller the prediction error sum of squares (PRESS) value, the better the model's predictive ability.

Multiple Regression

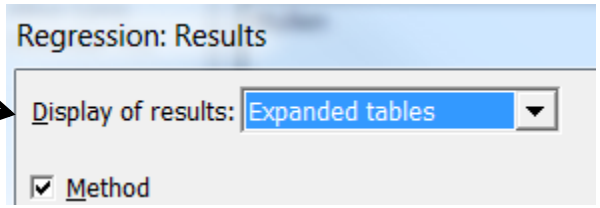
- ***Predicted R^2*** indicates how well the model predicts responses for new observations. Higher the better (the range is 0 to 1).
- ***VIF*** and ***Predicted R^2*** are displayed in the regression output.
- ***Press*** has to be chosen in the **Results…** dialog.

Multiple Regression

- Click the **Results…** button.



- Select **Expanded tables** from the **Display of results** pull-down menu and click **OK** once.

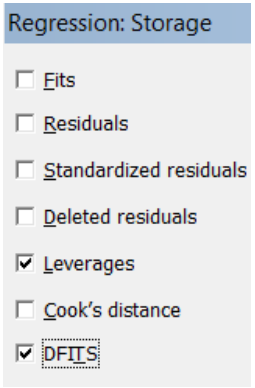


Multiple Regression

- The model diagnostics that assess how the observations fit the model are ***leverage*** and ***DFITS***.
- Model diagnostics for the observations are available in the **Storage…** dialog.

Multiple Regression

- Click the **Storage…** button → select **Leverages** and **DFITS** → click **OK** once.



- This will store these model diagnostics in the current worksheet in variables named HI*x* or DFIT*x* (where *x* is a number).

Multiple Regression

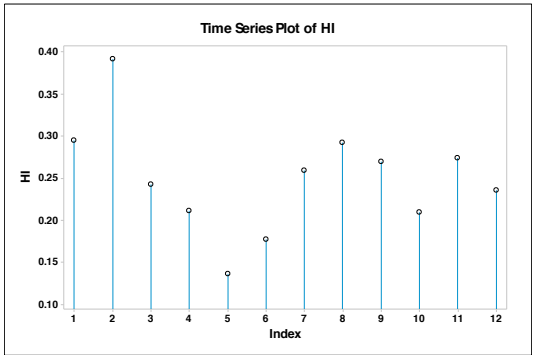
- ***Leverage*** identifies observations with unusual ***X***-values. Observations with large leverage may exert influence on the fitted value and the model.
  - Leverage values fall between 0 and 1.
- **Large leverage values suggests you examine the corresponding observation.** Large is defined as point greater than a ***3p/n*** limit, where ***p*** is the number of *predictors* or factors plus the constant and ***n*** is the number of *observations*.

Multiple Regression

- ***DFITS*** is a measure of the influence of each observation on the fitted value. It represents the number of standard deviations that the fitted value changes when each case is removed from the data set.
- Observations with the ***absolute value*** of the ***DFITS*** greater than ***2*sqrt(p/n)*** are considered large and should be examined, where ***p*** is the number of predictors and the constant, and ***n*** is the number of observations.

Multiple Regression

- The best way to examine both ***leverage*** and ***DFIT*** is with an ***index plot*** with the respective limits superimposed on them.
- An ***index plot*** is a time series plot, but instead of connecting the points, the lines are projected onto the x-axis.



Multiple Regression

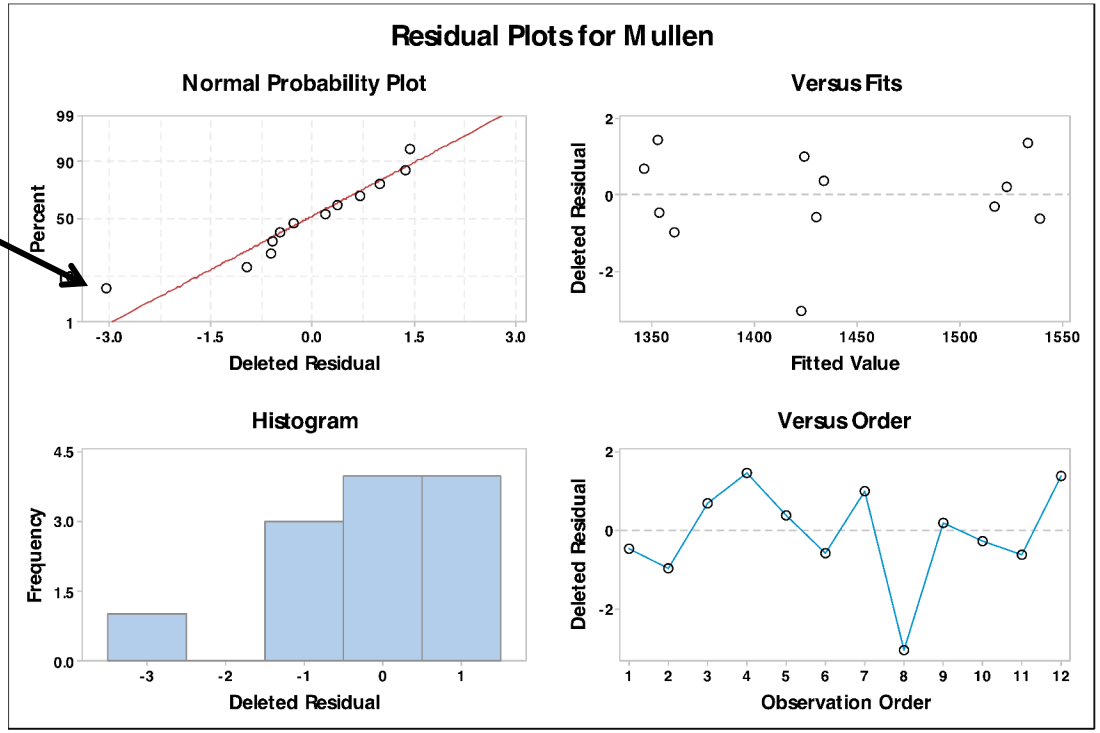
- Look for **leverage** (**H_i**) and **DFIT** values that are **far away from the rest of the cluster of points**.
- We will create these plots after we run the regression analysis.

- Click **OK** now to run the regression analysis.

Multiple Regression

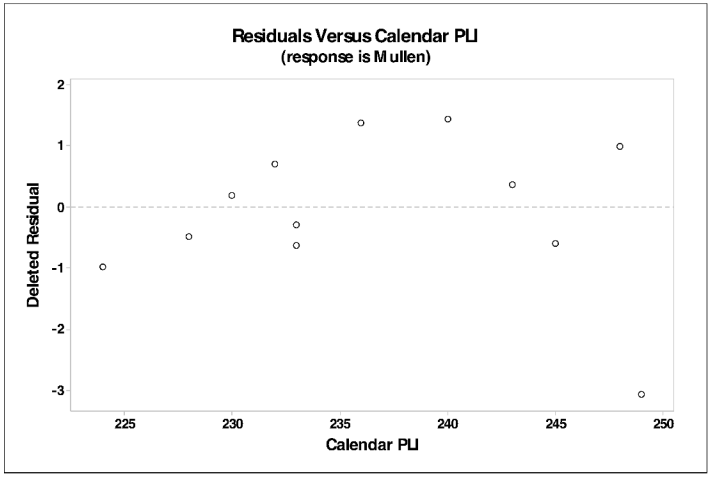
There is a large deleted residual value > +/-3

The **Histogram** looks unusual, but the sample size is small.

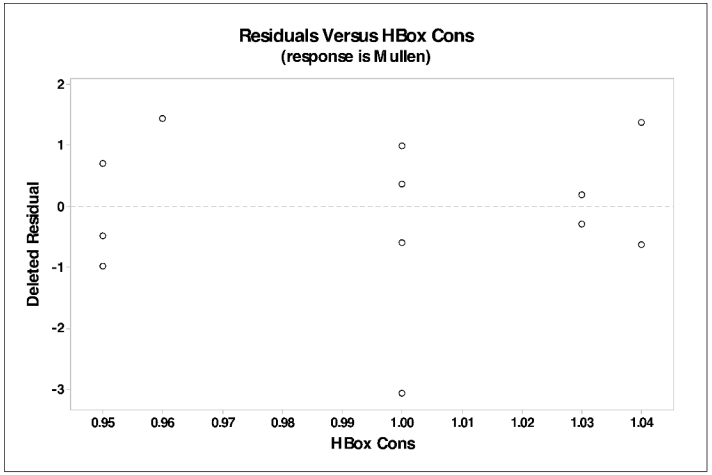


Multiple Regression

The residual versus variable plot showed that ***Calendar PLI*** may have a curved relationship to ***Mullen***, but is it due to the outlier?



Hbox Cons looks ok.

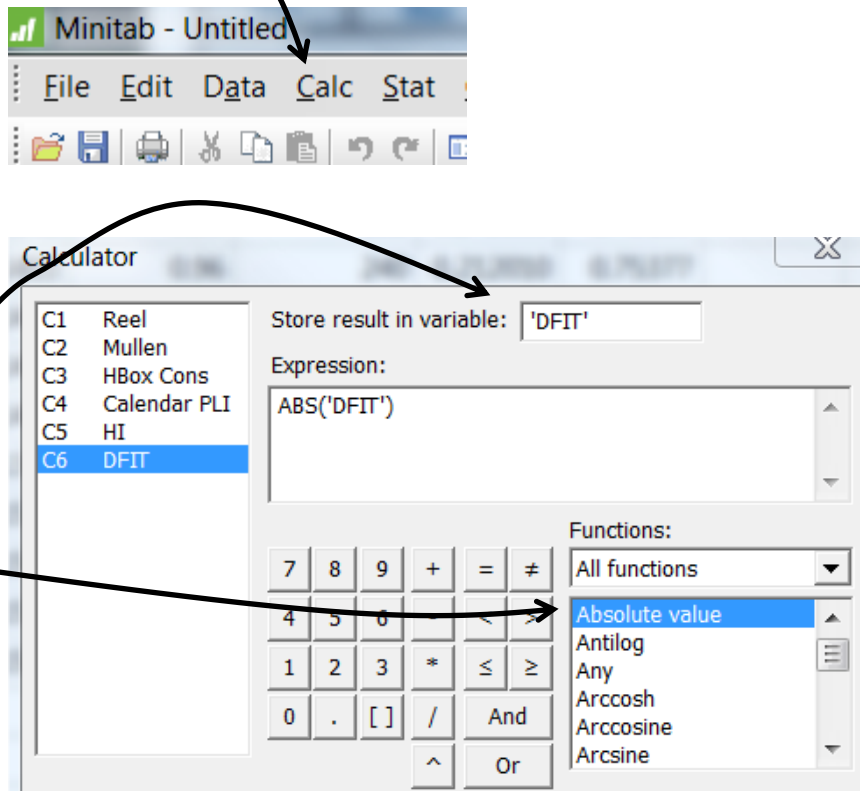


Multiple Regression

- The outlier is in row 8.
  - This is just happenstance data and not a designed experiment. Maybe delete ***sample #8***?
- Now is a good time to look at **Leverage** and **DFIT** using an **index plot**.
- First we must take the absolute value of **DFIT**.

Multiple Regression

- Choose **Calc → Calculator…**
- Type ***DFIT*** as the **Store result in variable** entry.
- Select **Absolute Value** from the **Functions** menu → select ***DFIT*** for the **number** → click **OK**.

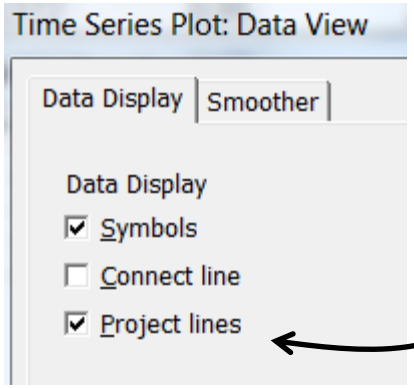


Multiple Regression

- Now choose **Graph** → **Time Series Plot…** → select **Simple** → click **OK**.
- Select ***DFIT*** as the **Series**.
- We now need to change the connect lines to project lines to enhance visualization.

Multiple Regression

- Click the **Data View** button → uncheck **Connect line** and check **Project lines**
- Click **OK** twice.

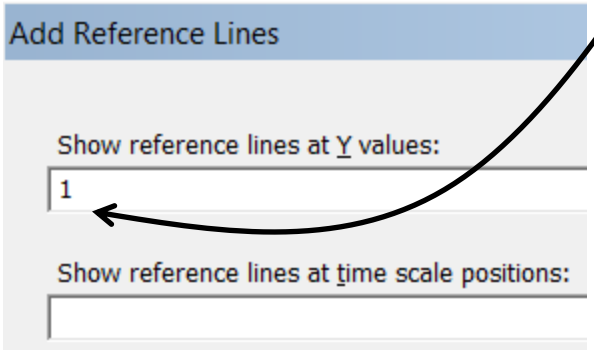


Multiple Regression

- Remember the limit for large DFIT is 2*sqrt(*p*/*n*).
  - For this example, for [p=3, n=12] the **limit** is 2*sqrt(3/12) = **1.0**
- We will learn how to add this as a **Y=1** as a **reference line** on the plot.

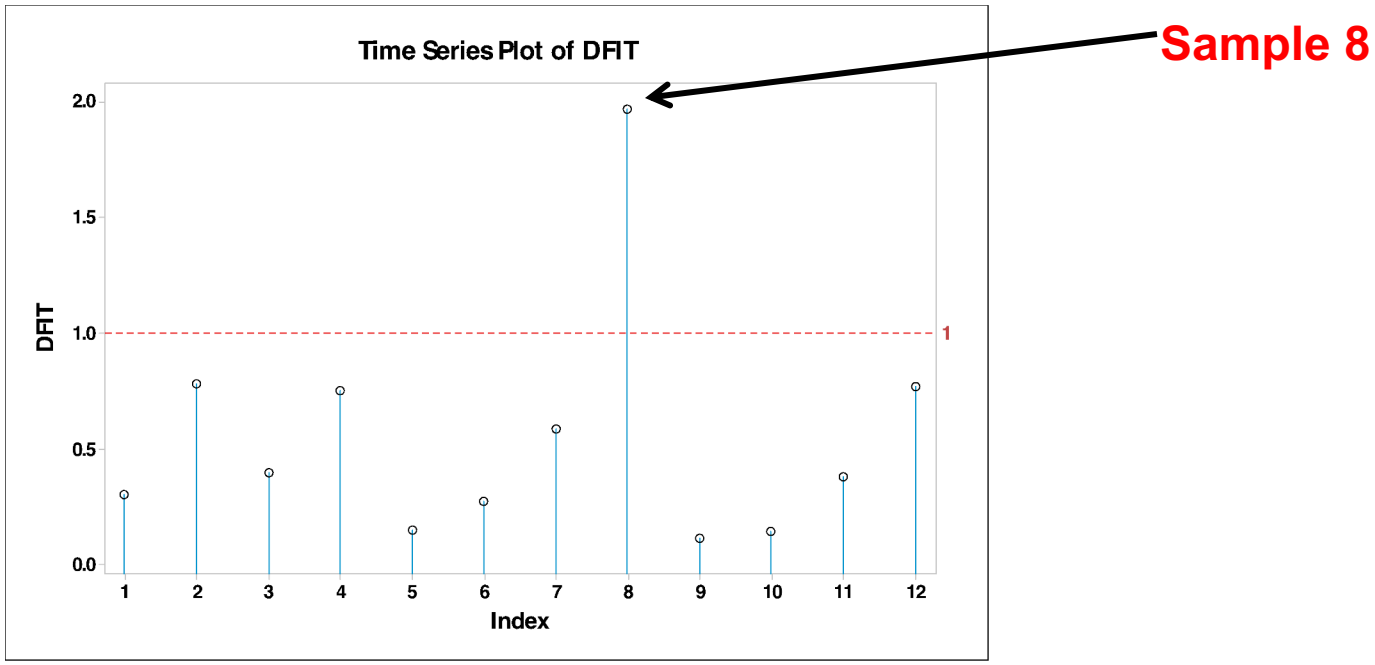
Multiple Regression

- Right click in graph → select **Add** → **Reference Lines…**
- Enter ***1*** in the **Show reference lines at Y values** box → click **OK**



Multiple Regression

- Look for points a large distance away from the 2*sqrt(p/n) limit.



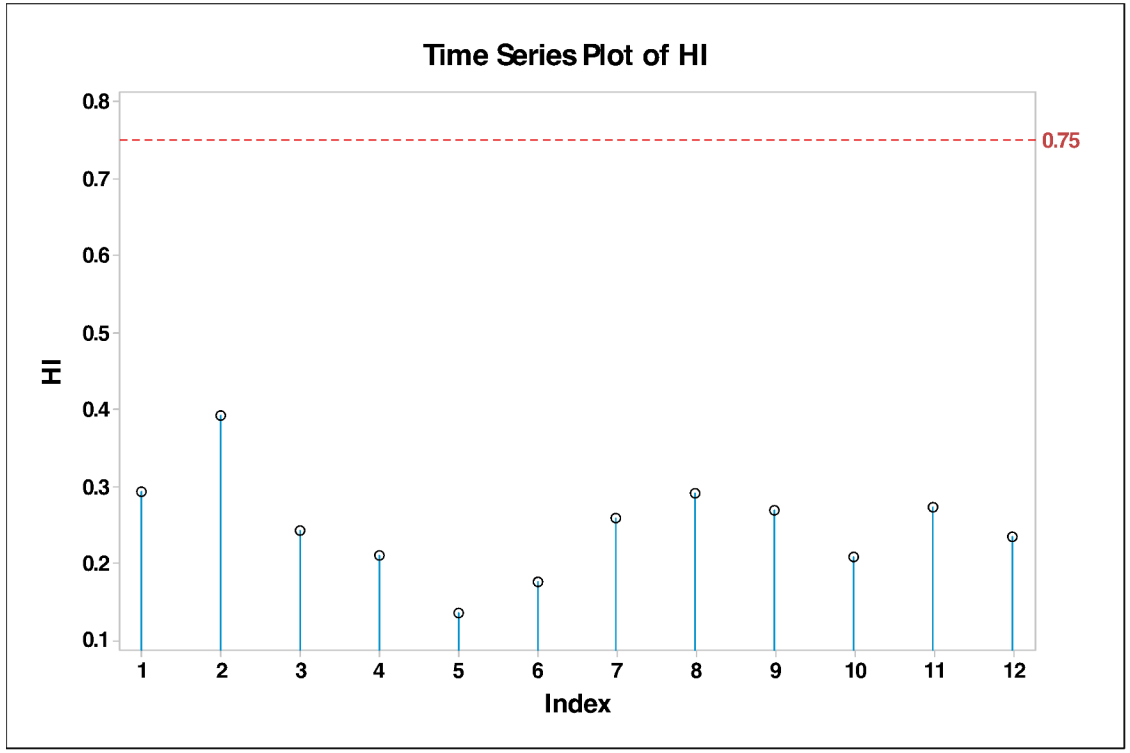
Multiple Regression

- Now make an ***Index Plot*** of the **Leverage** (**HI**). There is no need to take the absolute value for **Leverage**, its always positive.
- The limit for a large **Leverage** is **3*(*p*/*n*).**

  3*(3/12) = ***0.75***, for our example
- Add this limit as a **Y reference line** on the graph.

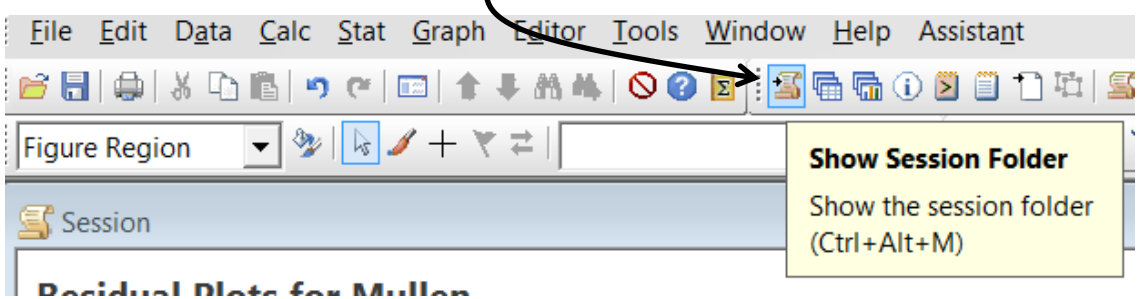
Multiple Regression

- The **leverage** plot looks ok.



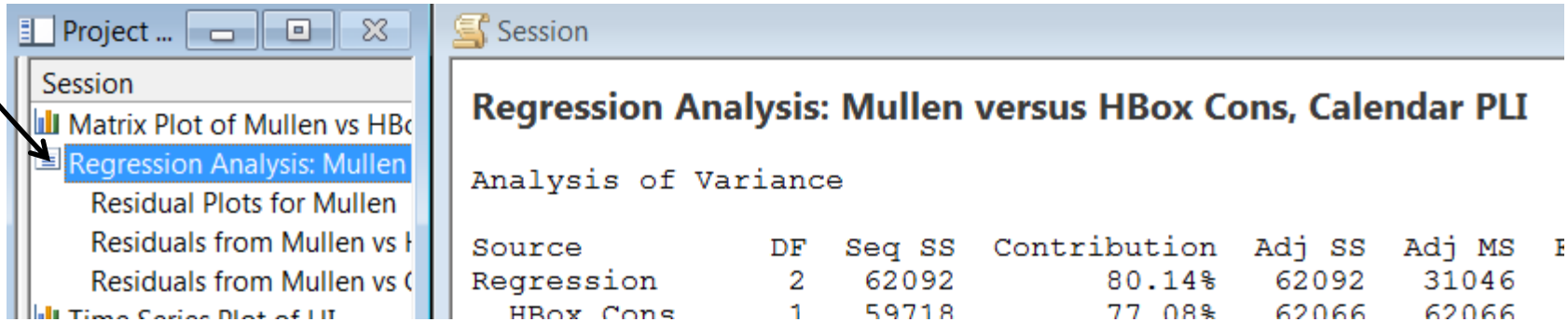
Multiple Regression

- ***Graphing and model diagnostics are important steps in doing any data analysis.***
- Lets look at the important statistical results for the first cut at model building.
- Click the **Show Session Folder** shortcut.



Multiple Regression

- Double-click on the **Regression Analysis: Mullen** entry in the **Project Manager**. This brings you to the regression output.



Multiple Regression

- Let's look at the statistical outputs.

- We are looking for **P** values less than ***0.05***

  P value = .000 for ***HBox Cons***, **P** < 0.05, so the regression coefficient is significant, ***Calendar PLI*** is not.
- We are looking for **VIF** less than 10.

  VIF = 1, so **VIF** is okay.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-282	457	(-1315, 751)	-0.62	0.552	
HBox Cons	2165	359	(1352, 2978)	6.02	0.000	1.05
Calendar PLI	-1.85	1.57	(-5.39, 1.70)	-1.18	0.269	1.05

Multiple Regression

- More statistical output.
- Looking for **adjusted R^2** bigger than 50%.

  R-Sq(adj) = 75.7%, which is good.

```
Model Summary

      S     R-sq  R-sq(adj)    PRESS  R-sq(pred)
41.3494  80.14%     75.73%  28914.7      62.68%
```

- **PRESS** will be used to compare to another model.

- Looking for **predicted R^2** bigger than 50% and fairly close to **R-Sq(adj)**.

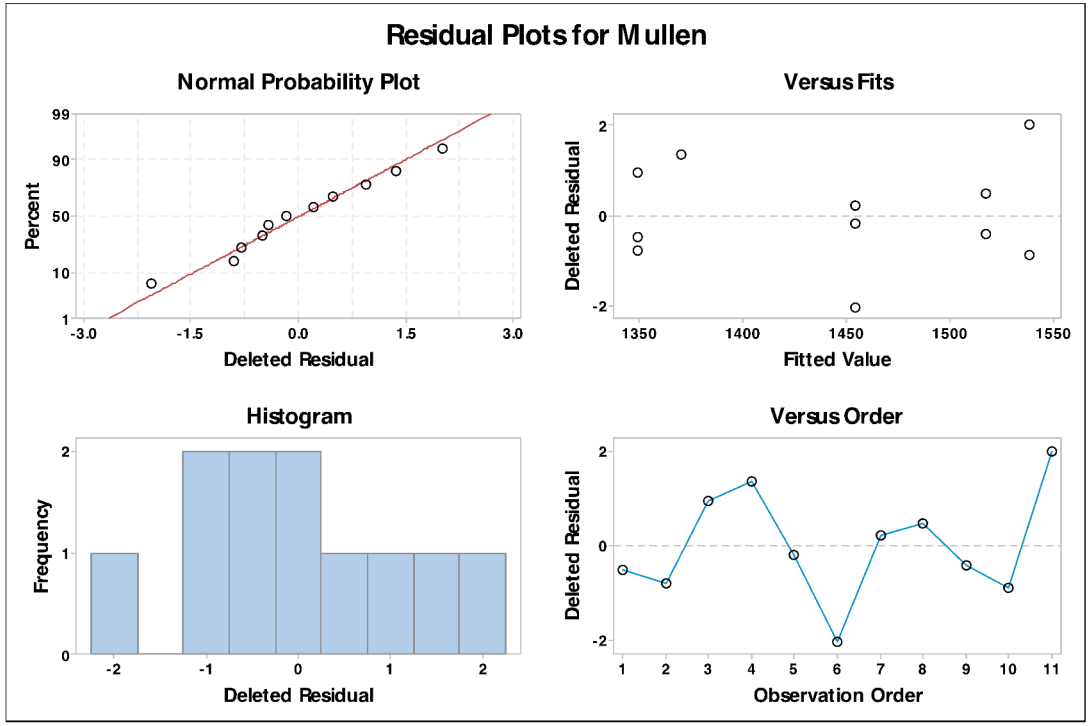
  R-Sq(pred) is okay.

Multiple Regression

- After examining the regression analysis, model assessment and diagnostic statistics, I would do the following:
  - Remove ***sample #8*** from the data.
  - Remove ***Calendar PLI*** from the model.
  - Run the regression model again and examine the model checking and diagnostic statistics.

Multiple Regression

- The residual plots look good.
  - With only 11 data points, the **histogram** is bound to look a little strange.

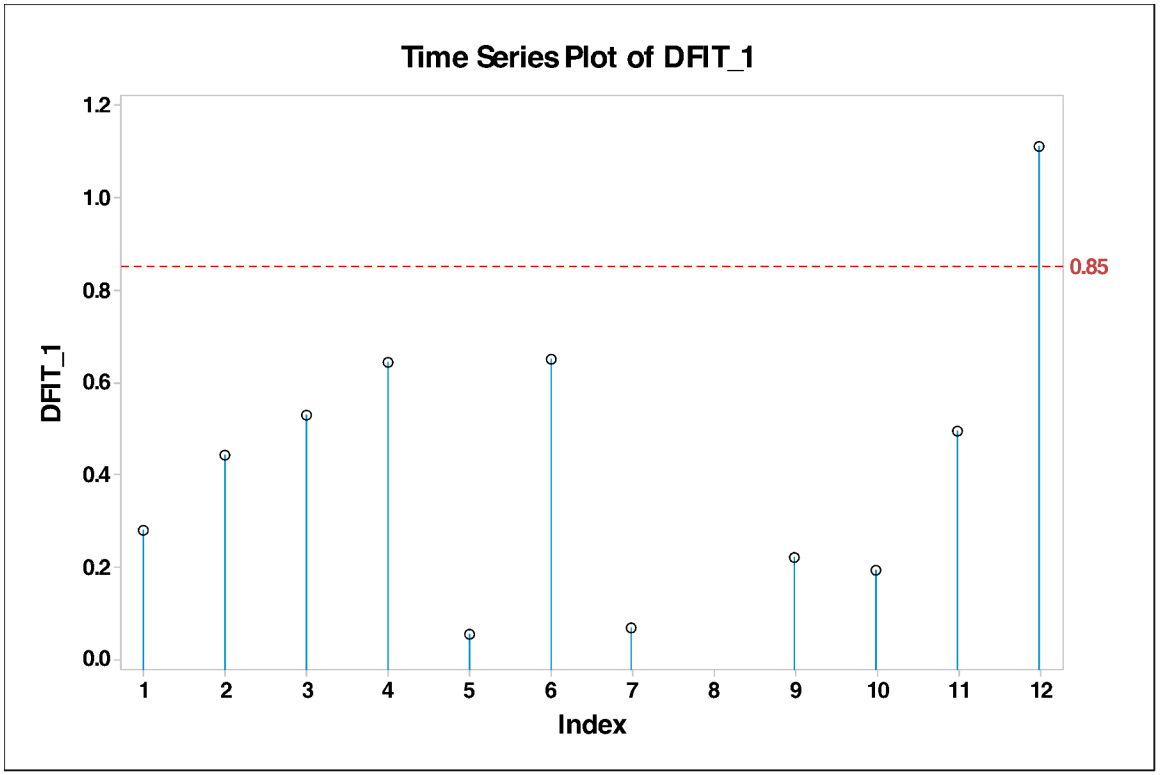


Multiple Regression

- The **leverage** (HI_1) values shouldn't change much.
- Create the index plot for ***DFIT_1***.
- Don't forget to create the absolute value of the ***DFIT_1*** and to calculate the new ***DFIT_1*** limit (hint: n=11, p=2).

Multiple Regression

- The **DFIT** plot looks okay. No big spikes.



Multiple Regression

- The ***P*** value indicates **Hbox Cons** is significant.
- ***PRESS*** decreased and ***R-Sq(pred)*** increased compared to the previous model, which is good.

```
Model Summary

      S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
28.0467  89.68%     88.53%  10680.3      84.43%

Coefficients

Term        Coef  SE Coef      95% CI      T-Value  P-Value   VIF
Constant    -652      237  (-1188, -115)    -2.75    0.023
HBox Cons   2106      238  ( 1567, 2644)     8.84    0.000  1.00

Regression Equation

Mullen = -652 + 2106 HBox Cons
```

Multiple Regression

- I didn't want to let you go without a brief discussion of the Analysis of Variance table.
  - But you'll have to take notes. ☺

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	61515	89.68%	61515	61515.2	78.20	0.000
HBox Cons	1	61515	89.68%	61515	61515.2	78.20	0.000
Error	9	7080	10.32%	7080	786.6		
Lack-of-Fit	**3**	**2050**	**2.99%**	**2050**	**683.2**	**0.82**	**0.531**
Pure Error	6	5030	7.33%	5030	838.3		
Total	10	68595	100.00%				

Multiple Regression

- The second regression analysis with only HBox Cons used to explain Mullen is a **better** prediction model.
- If this were the final model the prediction equation would be:

Mullen = - 652 + 2106*HBox Cons

Multiple Regression

- Given these results, a next step would be to devise a designed experiment that can be run in the mill to extract the maximum amount of data with minimum amount of runs.
  - Good idea to call a statistician to help with the design of this experiment.

EXERCISE

- Suppose a mill technical person believes the moisture and weight of a bale of pulp effect the bulkiness of the fiber.
- Use multiple regression to analyze the dataset ***Exercise 5.xls*** to see if there is a relationship.
- What are your conclusions?