**7.1 – The Concepts of Potential, Influence, and Outliers**

In Section 5 we learned how residual plots can be used to assumed mean and variance functions. A related issue is the importance of each observation or case on estimation and other important aspects of the analysis. In some data sets, the observed statistics may change in important ways if a few (or possible one) case are deleted from the data. Such cases are called *influential* and identifying cases that have undue influence on results is important*.* Some cases in our regression may not seem to correspond to the model fit to the bulk of the data. Cases that don’t follow the same model as the rest of the data are called *outliers*, and identifying these cases may be useful.

Beautifully hand drawn scatterplots go here:

In the last section we saw the matrix formulation of OLS SLR and found that the parameter estimates are given by:

$\hat{β}=\left(X^{T}X\right)^{-1}X^{T}Y$The **fitted values** $(\hat{Y})$ are given by multiplying the Hat Matrix ($H$) times the response vector ($Y$).

$$\hat{Y}=X\hat{β}=X\left(X^{T}X\right)^{-1}X^{T}Y=HY \rightarrow \hat{y}\_{i}=\sum\_{j=1}^{n}h\_{ij}y\_{j}$$

The **residuals** $(\hat{e})$are given by $\hat{e}=\left(Y-\hat{Y}\right)=\left(Y-HY\right)=\left(I-H\right)Y$.

**7.2 - Leverage or Potential –** *Case diagnostic measuring the potential for influence.*
One can show that the matrix results above imply that the following important result for case diagnostics.

$$Var\left(\hat{e}\_{i}\right)=\hat{σ}^{2}\left(1-h\_{ii}\right) where h\_{ii}=i^{th} diagonal element of H i=1,…,n$$

This may seem like a violation of the assumption that $Var\left(e\right)=σ^{2}$, but keep in mind that the residuals ($\hat{e}\_{i})$ are actually estimated quantities like the parameter estimates ($\hat{β}\_{o} \& \hat{β}\_{1})$ , etc. Thus they have estimated variances as do the parameters. The diagonal elements of the hat matrix ($h\_{ii}$) is called the ***potential*** or ***leverage*** value for the $i^{th}$ case. One can show that,

1. $\frac{1}{n}\leq h\_{ii}\leq 1$
2. The mean of the leverage values is:

 $\overbar{h}=\frac{k}{n}$

where $k=$ # of terms/parameters in the model.
Note: for the model where $E\left(X\right)=β\_{o}+β\_{1}X$ the mean leverage value is $\overbar{h}=^{2}/\_{n}$.

1. In SLR the leverage values ($h\_{ii}$) measure the distance from the sample mean for $X$ ($i.e. \overbar{x}$).

$$h\_{ii}=\frac{1}{n}+\frac{\left(x\_{i}-\overbar{x}\right)^{2}}{SXX}=h\_{i}$$

Thus the closer the leverage value for the $i^{th}$ case is to 1, the closer the variance of the residual associated with it $Var(\hat{e}\_{i})$ is to zero.

More beautiful scatterplots showing leverage or potential go here:

**Rule of Thumb:** Leverage values ($h\_{i}$) exceeding twice the average ($\overbar{h}$) should be considered noteworthy. However, when $n$ is large it is unlikely any single observation will exceed this cutoff.

**7.3 – “Standardized” Residuals and Outliers**

The value of residuals ($\hat{e}\_{i})$ are **scale dependent**, for example if the response is in millions are residual may be in the hundreds of thousands, while if the response is a decimal value the residual may be in the hundredths. Thus putting the residuals in a standardized scale would allow us to more easily determine if a residual $(\hat{e}\_{i})$ is extreme (i.e. if the $i^{th}$ case is a potential outlier).

Because $Var\left(\hat{e}\_{i}\right)=\hat{σ}^{2}\left(1-h\_{i}\right)$ and the sum (and hence the mean) of the residuals $(\hat{e}\_{i})$ is ALWAYS zero a standardized residual (think *z-score*) is given by:

$$r\_{i}=\frac{\hat{e}\_{i}}{\hat{σ}\sqrt{1-h\_{i}}}$$

The $r\_{i}'s$ are called the ***standardized residuals***. If the errors are normally distributed then we would expect approximately 95% the observations to have $r\_{i}\in (-2,2)$ and 99.7% to have $r\_{i}\in (-3,3)$. Any observations with $r\_{i}$ beyond these ranges are potentially outlying. HOWEVER, this is form of the standardized residual is not satisfactory for identifying outlier.

Beautiful plots showing why the standardized residuals ($r\_{i}$) are problematic for identifying outliers:

Because outliers will inflate our estimate of the error standard deviation (RMSE) $\hat{σ} $they can potentially mask themselves. Thus a better statistic/residual for identifying outliers is the ***studentized residual*** $(t\_{i})$.

$$t\_{i}=\frac{\hat{e}\_{i}}{\hat{σ}\_{(-i) }\sqrt{1-h\_{i}}} where \hat{σ}\_{\left(-i\right)}=RMSE when i^{th}case is excluded.$$

Another expression for the studentized residual ($t\_{i}$) is:

$$t\_{i}=r\_{i}\left(\frac{n-k-1}{n-k-r\_{i}^{2}}\right)^{\frac{1}{2}}$$

Because the $i^{th}$ case is excluded when calculating the estimated error standard deviation ($\hat{σ})$ the outlier cannot inflate the estimate and subsequently mask itself. One can show that the studentized residuals $(t\_{i})$ have a t-distribution with $df=n-k-1 $, where $k=$ # of terms/parameters in the model. The NH and AH being tested by this t-statistic are as follows:

 $NH:E\left(X=x\_{i}\right)=β\_{o}+β\_{1}x\_{i}$

 $AH:E\left(X=x\_{i}\right)=β\_{o}+β\_{1}x\_{i}+δ$ 🡨 This is called the *mean shift model*.

The mean shift model says that for the $i^{th}$ case we need a special shift $\left(δ\right)$ in the mean function.

As we generally perform this test for all *n* cases we need to make a correction for the *experiment-wise error rate* (EER) when considering p-values associated with the test statistic $(t\_{i}$). Why? What is this EER thing? Suppose you have $n=100$ observations in your regression. The probability of making a type I error on any one test is $α=.05.$ Thus we have 5% chance when testing any case to flag it as an outlier $\left(i.e. conclude δ\ne 0\right)$, when in fact it is not. However, if we test all cases the probability of making at least one type I error is equal to $1-\left(.95\right)^{100}= .994$. Thus we have a 99.4% chance of finding at least one outlier when none exist! This is clearly not acceptable.

***Bonferroni Correction*** - To fix this problem we use what is called the *Bonferroni Correction* which says when conducting a series of *m* significance tests you compare your p-values to $^{α}/\_{m}$ rather than $α$. One can show that by making this correction the $P(any type I error)\leq α$.

Employing the Bonferroni Correction for outlier detection we would compare the studentized residual ($t\_{i}$) p-value to $^{α}/\_{n}$ for each of the observations. As a general rule of thumb any observation with $|t\_{i}|>3.00 $ could be flagged as an outlier.

In JMP, we only have the standardized residuals available to us, and they label them as studentized residuals. The standardized residuals ($r\_{i}$) are also called the ***internally studentized residuals*** because they use an estimate of $σ$ based on all the data, i.e. including the case in question. The studentized residuals ($t\_{i}$) are also called the ***externally studentized residuals*** because they use an estimate of $σ$ not including the case in question. Thus in practice if we are using JMP to examine case diagnostics, then if you want to formally test if a case is an outlier, we need to use the formula above relating $t\_{i}$ to $r\_{i}$, i.e.

$$t\_{i}=r\_{i}\left(\frac{n-k-1}{n-k-r\_{i}^{2}}\right)^{\frac{1}{2}}$$

We can then find the p-value for testing if the $i^{th}$ case is an outlier and compare it to the Bonferroni Corrected p-value ($^{α}/\_{n}$).

**7.4 – Measuring Influence (Cook’s Distance)**

A case is considered influential if the results of a regression change appreciably when it’s exclusion from the analysis produces markedly different results. Cook’s Distance $(D\_{i})$ measures influence by measuring changes in the fitted values $\left(\hat{y}\right)$ when the $i^{th}$case is deleted. As the fitted values $(\hat{y})$ are determined by the parameter estimates $(\hat{β}) $we can view these as changes in the parameter estimates when the $i^{th}$ case is deleted. The following two formulations of Cook’s Distance ($D\_{i})$ are equivalent:

$$D\_{i}=\frac{\left(\hat{y}\_{\left(i\right)}-\hat{y}\right)^{T}\left(\hat{y}\_{\left(i\right)}-\hat{y}\right)}{k\hat{σ}^{2}}=\frac{1}{k\hat{σ}^{2}}\sum\_{j=1}^{n}\left(\hat{y}\_{\left(i\right),j}-\hat{y}\_{j}\right)^{2}$$

$$ =\frac{\left(\hat{β}\_{\left(i\right)}-\hat{β}\right)^{T}X^{T}X(\hat{β}\_{\left(i\right)}-\hat{β})}{k\hat{σ}^{2}}$$

There is no significance test based upon Cook’s Distance ($D\_{i}$) but there are some general guidelines for identifying cases with a high degree of influence. It is generally useful to investigate cases where $D\_{i}>0.5$ and you should always investigate cases where $D\_{i}>1.0$. For large datasets it is unlikely any case with $D\_{i}$ exceeding these guidelines. For that reason some suggest using a size-adjusted cut-off of

$$D\_{i}>\frac{4}{n-k} .$$

A more “intuitive” formulation of the Cook’s Distance for the $i^{th}$ case is given in terms of the standardized residual $(r\_{i})$ and the leverage/potential value ($h\_{i})$.

$$D\_{i}=\frac{r\_{i}^{2}}{k}×\frac{h\_{i}}{1-h\_{i}}$$

This formulation allows to better understand what has to happen in order for a case to have unduly high influence on the results of a regression. On the next page we will examine some plots illustrating influence and the concept of Cook’s Distance.

Beautiful hand-drawn plots showing influence go here:

**Example 7.1 – West Bearskin Bass (Age = 4)**

Here we consider the regression of length (mm) on scale radius. One point stands out from the rest. This fish has a very large scale radius (and length) relative to the other 4-year old fish in the sample. Using the Analyze > Fit Model option to fit the simple linear regression of length on scale radius with $E\left(Scale\right)=β\_{o}+β\_{1}X$ and $Var\left(Scale\right)=σ^{2}$.







Below is a scatterplot with the suspect point deleted.







Clearly when the case in question is deleted the regression results differ substantially. This point clearly has a great deal of influence on the regression results. We will now examine the leverage ($h\_{i}$), studentized residual ($t\_{i}$), and Cook’s Distance $(D\_{i}$) for all of the cases in this regression analysis.

 

Case Diagnostics for Observation 11

For this regression there are n = 15 observations so $\overbar{h}=\frac{2}{15}=.1\overbar{333}$. We can see that for the observation in question the leverage is $h\_{i}=.619>2\overbar{h}.$

The Cook’s Distance for this case is $D\_{i}=1.744$ which is VERY large. This is not surprising given the differences in the parameters estimates between the models fit with and without this case.

The studentized residual $r\_{i}=1.467$ which is not particularly large, thus this point would not be deemed an outlier.

**What should we do?**

**Example 7.2 – Percent Body Fat (%) and Weight (lbs.)**

Here we consider the regression of percent body fat on weight. Again we see an observation (#39) that stands apart from the rest.







Without case 39 we have the following.






The case diagnostics for case 39 are shown below.


$h\_{i}=.1605>\frac{2}{252}=.0079$ \* high leverage

$D\_{i}=.6670> .50$ \*moderately high influence

$r\_{i}=-2.647 $ \*clearly poorly fit

As JMP does not actually calculate the studentized residual (externally) we have to perform the calculation by hand:

$$t\_{39}=r\_{39}\left(\frac{n-k-1}{n-k-r\_{39}^{2}}\right)^{\frac{1}{2}}=-2.647\left(\frac{252-2-1}{252-2-\left(-2.647\right)^{2}}\right)^{\frac{1}{2}}=-2.6795$$

Using the t-distribution probability calculator in R:

> pt(-2.6795,250)
[1] 0.00393143

The Bonferroni correction gives ${.05}/{252}=.0002$. Thus even though this case has an extreme “standardized” residual it is not technically an outlier using the mean shift test.

**Example 7.3 – Adaptive Score Data**

These data consist of observations made on sample *n = 21* children. Their age at their first spoken word was recorded as well as their adaptive score, which is a measure of the development of the child. The goal of the study is study the conditional distribution of adaptive score given the age at which they spoke their first word. Below is a scatterplot of these data with three interesting cases labeled. This labelling can be done in JMP by first highlighting these observations by holding down the shift to select multiple points and then selecting **Rows > Label/Unlabel**.

 Adaptive Scores vs. Age First Spoke Regression summary (all cases included) 

How would you characterize case 19?

How would you characterize cases 2 and 18?

Below are the case diagnostics for each case in this study. 

What do these case diagnostics suggest?

 $\overbar{h}$=.095

Below are the regression summaries with cases 2, 18, and 2 & 18 removed.

Case 2 removed
 

Case 18 removed
 

Case 2 & 18 removed
 

**Comments:**

Returning to the model fit to the entire sample, can we conclude that case 19 is an outlier? Again we can use the “studentized” residuals from JMP ($r\_{i}$) to find the mean shift outlier t-statistic $(t\_{i})$.

$$t\_{19}=r\_{19}\left(\frac{n-k-1}{n-k-r\_{19}^{2}}\right)^{\frac{1}{2}}=2.823\left(\frac{21-2-1}{21-2-\left(2.823\right)^{2}}\right)^{\frac{1}{2}}=3.606$$

Again using R as t-distribution probability calculator:

> 1 - pt(3.606,19)

[1] 0.000941285

The Bonferroni correction gives ${.05}/{21}=.0024$ , thus we conclude that case 19 is indeed an outlier on the basis of the mean shift model ($p< .0024)$.

The function SLRdiag function in .RData directory in the Shared folder in the Class Storage directory calculates all case diagnostics discussed above and highlights cases of particular interest.

> AdaptScore = read.table(file.choose(),header=T,sep=”,”)
> lm1 = lm(Score~Age,data=AdaptScore)

> SLRdiag(lm1)



**Section Summary:**

Case diagnostics should routinely be examined to some degree whenever we have chosen a “final” model. In simple linear regression it is generally very easy to see potentially problematic points as the examples above illustrate. However, in multiple regression it is harder to identify cases that have exceptionally high influence on the regression results in particular. Potential outliers are still generally easy diagnose when we examine plots of the residuals.