Outliers, High Leverage Observations, and Influential Observations

For this tutorial, we will go over the delivery time example from the textbook. The textbook describes the data as:

A soft drink bottler is analyzing the vending machine service routes in his distribution system. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time (y) are the number of cases of product stocked (x_1) and the distance walked by the route driver (x_2).

Download the data file, `delivery.txt`, from Collab and read the data in.

```
Data<-read.table("delivery.txt", header=TRUE)
result<-lm(Delivery~Number+Distance, data=Data)
```

1) Outliers

a) Residuals

Residuals can be used to help detect outliers. To extract residuals from `lm()`

```
##residuals, e_i
res<-result$residuals
```

Recall that residuals are defined as $e_i = y_i - \hat{y}_i$, and depend on the unit associated with the response variable. They can be scaled in a way to make them unitless. There are three primary ways to standardize the residuals.

b) Standardized residuals

The standardized residuals are defined as $d_i = \frac{e_i}{\sqrt{MS_{res}}}$. Since the residual standard error reported in the `summary` output of a model fit by `lm()` is equal to $\sqrt{MS_{res}}$, we can find the standardized residuals with

```
##standardized residuals, d_i
standard.res<-res/summary(result)$sigma
```

Even though the errors in a regression model have constant variance, the residuals do not theoretically have constant variance. In fact, the variance of the residuals is $MS_{res}(1-h_{ii})$. An implication is that observations that have high leverages tend to actually have small residuals, and so residuals cannot detect all types of outliers. Thus, a further refinement in scaling the residuals has been proposed.

c) Studentized residuals

Studentized residuals are defined as $r_i = \frac{e_i}{\sqrt{MS_{res}(1-h_{ii})}}$ to recognize that the residuals do not have constant variance. Studentized residuals can be found by

```
##studentized residuals, r_i
student.res<-rstandard(result)
```

Comparing the formulas for standardized residuals, d_i, and studentized residuals, r_i, we can see that they will be similar for observations that have low leverages.

d) Externally studentized residuals

Another way to scale residuals is to consider removing observation i from the data set. So instead of using MS_{res}, we use $S^2_{(i)}$, which is the MS_{res} of the regression model with observation i removed. Thus, we have externally studentized residuals, $t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}$

The externally studentized residuals can be found by

```
##externally studentized residuals, t_i
ext.student.res<-rstudent(result)
```

If observation i is influential, we expect $S^2_{(i)}$ to differ from MS_{res}, so differing values of studentized residuals, r_i, and externally studentized residuals, t_i, is a sign an observation is influential.

Now that we have found the different residuals, we can combine them into a data frame to make some comparisons.

```
res.frame<-data.frame(res,standard.res,
                      student.res,ext.student.res)
```

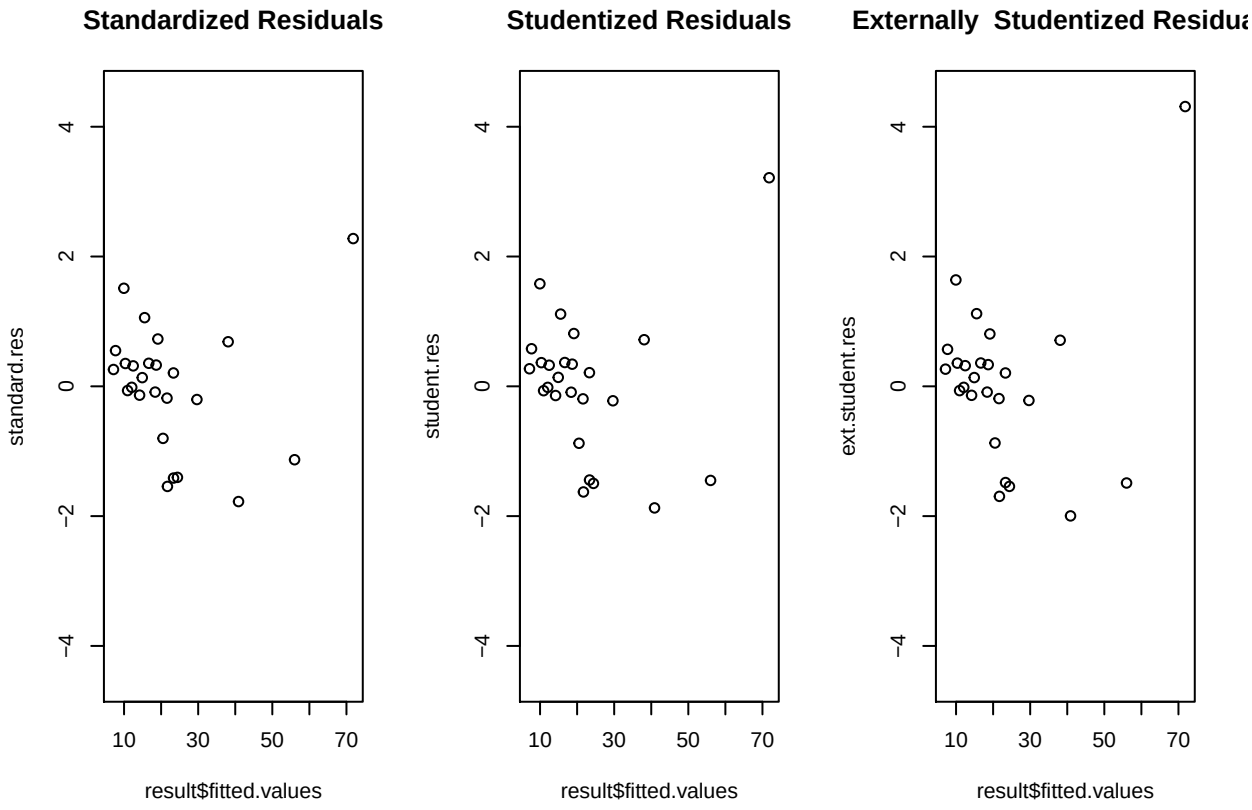
We highlight the 9th observation. Notice it's externally studentized residual is a lot larger than it's studentized residual, which in turn is a lot larger than its standardized residual (which itself is large). So this observation is likely to have high leverage and is likely to be influential.

```
res.frame[9,]
```

```
##        res standard.res student.res ext.student.res
## 9 7.419706     2.276351    3.213763         4.31078
```

We can create scatterplots of the standarized residuals, studentized residuals, and externally studentized residuals

```
par(mfrow=c(1,3))
plot(result$fitted.values,standard.res,
     main="Standardized Residuals",
     ylim=c(-4.5,4.5))
plot(result$fitted.values,student.res,
     main="Studentized Residuals",
     ylim=c(-4.5,4.5))
plot(result$fitted.values,ext.student.res,
     main="Externally  Studentized Residuals",
     ylim=c(-4.5,4.5))
```



Notice the plots look very similar, other than observation 9.

e) Outlier detection

We use the t distribution and the Bonferroni procedure to find a cut off value for outlier detection using externally studentized residuals. If $|t_i| > t_{1-\frac{\alpha}{2n},n-1-p}$, observation i is deemed an outlier.

```
##critical value using Bonferroni procedure
n<-dim(Data)[1]
p<-3
crit<-qt(1-0.05/(2*n), n-p-1)
```

To identify observations that satisfies the cutoff

```
##identify
ext.student.res[abs(ext.student.res)>crit]
```

```
##       9
## 4.31078
```

So we see that observation 9 is an outlier.

2) Leverages

Leverages, h_{ii}, are used to identify how far observation i is from the centroid of the predictor space. If $h_{ii} > \frac{2p}{n}$, then observation i is deemed to have high leverage and is outlying in the predictor space. High leverage observations are data points that are most likely to be influential.

To identify high leverage observations

```
##leverages
lev<-lm.influence(result)$hat
##identify high leverage points
lev[lev>2*p/n]
```

```
##         9        22
## 0.4982922 0.3915752
```

Observations 9 and 22 have high leverage. Observation 9 is not surprising given our earlier analysis.

3) Influential observations

a) Cook's distance

Cook's distance, D_i, can be interpreted as the squared Euclidean distance that the vector of fitted values moves when observation i is removed from the regression model. A cutoff rule for an influential observation is $D_i > F_{0.5,p.n-p}$.

To find Cook's distance and influential observations

```
##cooks distance
COOKS<-cooks.distance(result)
COOKS[COOKS>qf(0.5,p,n-p)]
```

```
##        9
## 3.419318
```

So observation 9 is deemed to be influential.

b) DFFITs

$DFFITS_i$ measures how much the fitted value of observation i changes when it is removed from the regression model. Observation i is influential if $|DFFITS_i| > 2\sqrt{p/n}$.

To find $DFFITS_i$ and influential observations

```
##dffits
DFFITS<-dffits(result)
DFFITS[abs(DFFITS)>2*sqrt(p/n)]
```

```
##         9        22
##  4.296081 -1.195036
```

Observations 9 and 22 are influential based on $DFFITS_i$.

c) DFBETAs

Another measure of influence is $DFBETAS_{j,i}$, which measures how much the estimated coefficient $\hat{\beta}_j$ changes when observation i is removed from the regression. The cutoff used is $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$.

To find $DFBETAS_{j,i}$ and influential observations

```
##dfbetas
DFBETAS<-dfbetas(result)
abs(DFBETAS)>2/sqrt(n)
```

```
##    (Intercept) Number Distance
## 1        FALSE   TRUE     TRUE
## 2        FALSE  FALSE    FALSE
## 3        FALSE  FALSE    FALSE
## 4         TRUE  FALSE    FALSE
## 5        FALSE  FALSE    FALSE
## 6        FALSE  FALSE    FALSE
## 7        FALSE  FALSE    FALSE
## 8        FALSE  FALSE    FALSE
## 9         TRUE   TRUE     TRUE
## 10       FALSE  FALSE    FALSE
## 11       FALSE  FALSE    FALSE
## 12       FALSE  FALSE    FALSE
## 13       FALSE  FALSE    FALSE
## 14       FALSE  FALSE    FALSE
## 15       FALSE  FALSE    FALSE
```

```
## 16       FALSE  FALSE    FALSE
## 17       FALSE  FALSE    FALSE
## 18       FALSE  FALSE    FALSE
## 19       FALSE  FALSE    FALSE
## 20       FALSE  FALSE    FALSE
## 21       FALSE  FALSE    FALSE
## 22       FALSE   TRUE     TRUE
## 23       FALSE  FALSE    FALSE
## 24       FALSE   TRUE     TRUE
## 25       FALSE  FALSE    FALSE
```

We see that observation 1 is influential in estimating β_1 and observation 4 is influential in estimating β_0. Observation 9 is influential in estimating all the coefficients. Observations 22 and 24 are influential in estimating β_1, β_2.

If you have a lot more observations, an alternative way to find infuential observations based on $DFBETAS_{j,i}$ would be to evaluate each coefficient

```
##for beta0
DFBETAS[abs(DFBETAS[,1])>2/sqrt(n),1]
```

```
##          4          9
##  0.4519647 -2.5757398
```

```
##for beta1
DFBETAS[abs(DFBETAS[,2])>2/sqrt(n),2]
```

```
##          1          9         22         24
##  0.4113119  0.9287433 -1.0254141  0.4046226
```

```
##for beta2
DFBETAS[abs(DFBETAS[,3])>2/sqrt(n),3]
```

```
##          1          9         22         24
## -0.4348621  1.5075506  0.5731402 -0.4654469
```