Diagnostics for Outliers and Influential Points

BIOS 6611

CU Anschutz

Week 14

Outliers

Introduction to Outliers

Outliers are observations with a residual that is much larger than residuals from the rest of the data. Potential explanations for outliers:

1. Human error: the value(s) for the observation was measured, recorded, or entered incorrectly.
    - In this case, the value(s) for the observation should be corrected or the observation should be deleted from the analysis (only delete if it is **known** to be wrong!)
2. Inadequacies in the model.
    - The model may fail to fit the data well for certain values of the predictor due to non-linearity, non-homogeneity of variance, or an important variable or strong interaction may have been omitted from the model.
    - In this case, deletion of the observation from the analysis could be very very very very bad.
3. Outliers can occur because of poor sampling of observations in the tail of the distribution.

Deleting Outliers

Use extreme caution when deleting observations unless the values are not plausible or you are POSITIVE a coding error occurred.

An observation appearing unusual does not mean it should be excluded. Deleting observations can lead to an underestimation of the variability and p-values that are optimistically small.

If not due to human error, you can report model results with and without deleted outliers (i.e., a form of "sensitivity analysis").

Assessing Outliers

Jackknife residuals outside the ± 3 range are often considered potential outliers; outside the ± 4 range are of greater concern.

- Some use ± 2 as a conservative range for points of potential concern.

The expected number of observations outside a given range will depend on the sample size (recall, jackknife residuals follow a t-distribution).

The observation should be evaluated for its effect on the analysis. Depending on the location in the prediction space, an outlier can have severe effects on the regression model.

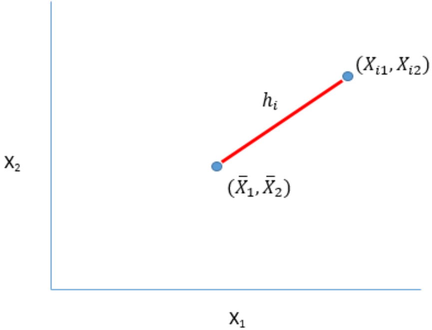
Leverage and Influential Points

Leverage

The *leverage*, h_i, of an observation is a measure of the geometric distance of the observation's predictor point $(X_{i1}, X_{i2}, \ldots, X_{ik})$ from center point (mean) of the predictor space $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$. h_i is calculated by finding the i^{th} diagonal element of the **hat matrix**:

$$\begin{aligned} \mathbf{H} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ h_i &= [\mathbf{H}]_{ii} \\ 0 \leq\ h_i &\leq 1 \end{aligned}$$

We say a point is a high leverage point if $h_i > 2(p+1)/n$.



Influential Points

An *influential observation* is defined as an observation that has a notable effect on the coefficients of the fitted regression line.

High leverage observations have the *potential* to be very influential, but not necessarily.

Low leverage observations *cannot* have dramatic influence on the regression coefficients (especially the slope coefficients).

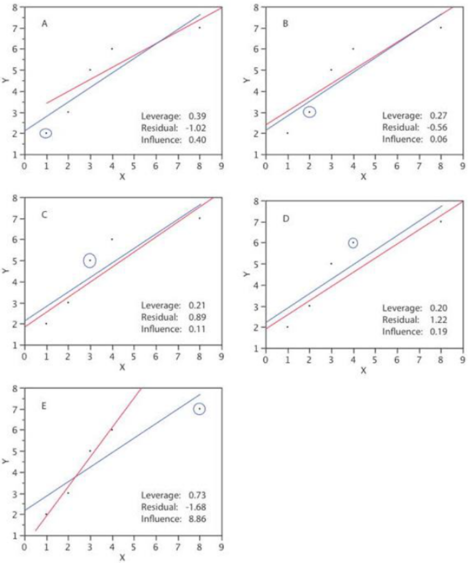
Observations with leverage greater than $2(p+1)/n$ should be further inspected.

Leverage and Influence

High Leverage → Potential Influence

Blue line: includes circled point
Red line: excludes circled point

A) High leverage & low influence
B) Lower leverage → lower influence
C) Low leverage → low influence
D) Low leverage → low influence
E) High leverage & high influence



Cook's Distance (Cook's D)

We will examine three ways to measure influence: Cook's Distance, DFFITS, DFBETAS

Cook's Distance (d_i) measures how much the regression coefficients are changed by deleting an observation.

d_i measures the influence of the i^{th} observation on all n fitted values:

$$d_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T(\mathbf{X}^T\mathbf{X})(\hat{\beta} - \hat{\beta}_{(-i)})}{(p+1)MSE}$$

It can also be expressed in terms of its residual, e_i, and its leverage h_i:

$$d_i = \frac{e_i^2 h_i}{(p+1)MSE(1-h_i)^2}$$

Observations with Cook's Distance values > 1.0 should be examined, although recent work has called into question the usefulness of this measure.

DFFITS (Difference in Fits)

Where Cook's Distance measures the influence of the ith observation on all n fitted values, $(\text{DFFITS})_i$ is a measure of the influence of the ith observation on the fitted value $\hat{Y}_i$. The measure is given by

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{(-i)}}{\sqrt{MSE_{(-i)}h_i}}$$

where $\hat{Y}_{(-i)}$ is the fitted value of Y_i from the regression model fit with the ith observation deleted.

The denominator is the estimated standard deviation of $\hat{Y}_i$ and is based on the MSE calculated from the regression model fit with the ith observation deleted.

The resulting standardization represents the number of estimated standard deviations of Y_i that the fitted value increases or decreases with the inclusion of the ith observation in the model.

DFFITS (cont.)

Using the jackknife residual, we can calculate $(DFFITS)_i$ without refitting the model n times:

$$(DFFITS)_i = r_{(-i)}\sqrt{\frac{h_i}{1-h_i}}$$

where $r_{(-i)}$ is the studentized residual from the model fit without the ith observation.

If h_i or $r_{(-i)}$ is near 0, then there is little effect from the observation, and DFFITS is close to 0.

Any observation with $(DFFITS)_i$ outside the range of $\pm 2\sqrt{(p+1)/n}$ warrants further investigation.

DFBETAS (Difference in Betas)

Both DFFITS and Cook's Distance measure the influence of an observation on the fitted values, whereas **DFBETAS** measures the influence on the individual coefficient estimates.

DFBETAS measures the difference between the coefficient estimated with and without the *i*th observation and standardizes difference by dividing by an estimate of the standard error:

$$(DFBETAS)_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k(-i)}}{\sqrt{C_{kk} MSE_{(-i)}}}$$

where C_{kk} is the kth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. A large value indicates the ith observation has a sizable impact on the kth regression coefficient. The sign (positive or negative) is meaningful.

Any observation with $(DFBETAS)_{k,i}$ outside the range of $\pm 2/\sqrt{n}$ warrants further investigation. In smaller data sets, larger may be considered meaningful.

Example

Example

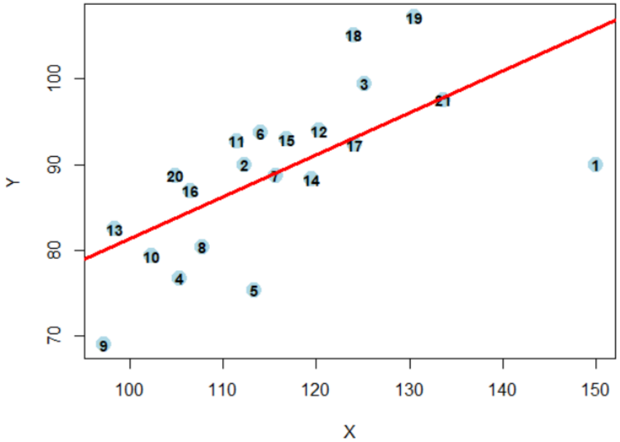
A data set with 21 observations and one predictor:

```
oildat <- read.delim(file="oildata.txt",sep=" ", header=FALSE)[,1:2]
colnames(oildat) <- c("X","Y")
lm0 <- lm(Y~X, data=oildat)
```

Y	X	Y	X
90.00	150.00	93.98	120.34
90.01	112.26	82.51	98.4
99.47	125.20	88.31	119.52
76.76	105.31	92.95	116.84
75.36	113.35	86.93	106.52
93.70	114.08	92.31	124.16
88.72	115.68	105.15	124.04
80.41	107.80	107.19	130.47
68.96	97.27	88.75	104.86
79.33	102.35	97.54	133.61
92.79	111.44	-	-

Example

```
plot(Y~X, col="lightblue",pch=19,cex=2,data=oildat)
text(Y~X, labels=rownames(oildat),cex=0.9,font=2,data=oildat)
abline(lm0,col="red",lwd=3)
```



Example

Recommendations for further investigation:

- Leverage: $2(p+1)/n = 2(2)/21 = 0.19$
- Jackknife Residual: ± 3; ± 4
- DDFITS: $\pm 2\sqrt{(p+1)/n} = 2\sqrt{2/21} = \pm 0.62$
- DFBETAS: $\pm 2/\sqrt{n} = 2/\sqrt{21} = 0.44$
- Cook's Distance: $d_i > 1.0$

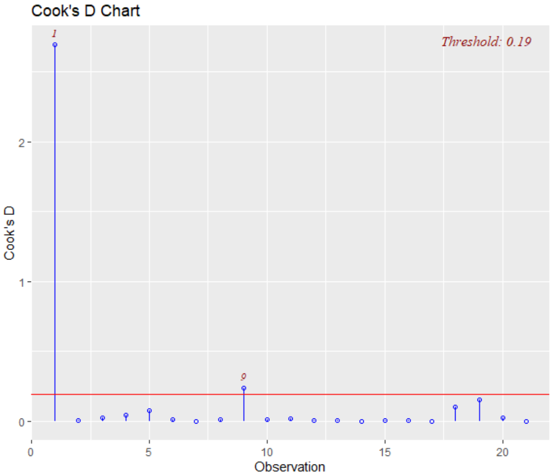
Influential Measures

```
# Cases which are influential with respect to any of these measures are marked with an asterisk.
# "hat" is leverage, "cov.r" is covariance ratios
influence.measures(lm0)
```

```
## Influence measures of
##   lm(formula = Y ~ X, data = oildat) :
##
##      dfb.1_    dfb.X    dffit cov.r   cook.d    hat inf
## 1   2.63735 -2.759026 -2.93463 0.663 2.69e+00 0.4102   *
## 2   0.03236 -0.023850  0.08487 1.158 3.77e-03 0.0517
## 3  -0.11760  0.137773  0.22891 1.122 2.67e-02 0.0747
## 4  -0.21972  0.196451 -0.30233 1.088 4.57e-02 0.0824
## 5  -0.12719  0.084077 -0.41929 0.832 7.82e-02 0.0496
## 6   0.04209 -0.024429  0.16950 1.101 1.47e-02 0.0486
## 7  -0.00104  0.000138 -0.00852 1.170 3.83e-05 0.0476
## 8  -0.11120  0.096134 -0.17573 1.141 1.59e-02 0.0680
## 9  -0.64515  0.605862 -0.72740 0.981 2.41e-01 0.1555
## 10 -0.12048  0.110397 -0.14955 1.218 1.17e-02 0.1046
## 11  0.08380 -0.064972  0.19222 1.097 1.88e-02 0.0538
## 12 -0.02066  0.029512  0.08700 1.160 3.97e-03 0.0538
## 13  0.10046 -0.093941  0.11507 1.288 6.96e-03 0.1428
## 14  0.01458 -0.022997 -0.08146 1.159 3.48e-03 0.0517
## 15  0.00297  0.007934  0.10260 1.144 5.49e-03 0.0479
## 16  0.06460 -0.056940  0.09434 1.190 4.67e-03 0.0749
## 17  0.01469 -0.017574 -0.03159 1.195 5.26e-04 0.0690
## 18 -0.22103  0.265163  0.48147 0.865 1.04e-01 0.0684
## 19 -0.40444  0.447186  0.58621 0.953 1.58e-01 0.1139
## 20  0.15935 -0.143120  0.21510 1.154 2.38e-02 0.0854
## 21  0.01139 -0.012378 -0.01509 1.304 1.20e-04 0.1455
```

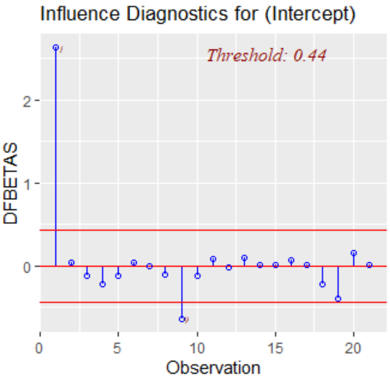
Cooks D Visualization

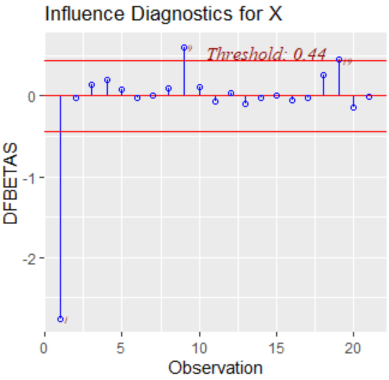
```
library(olsrr)
ols_plot_cooksd_chart(lm0)
```



DFBETAS Visualization

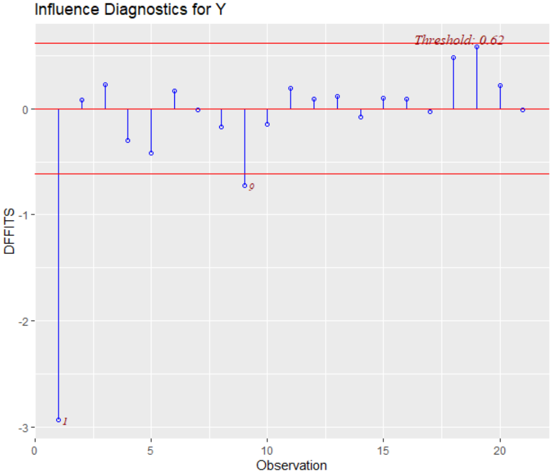
```
ols_plot_dfbetas(lm0)
```





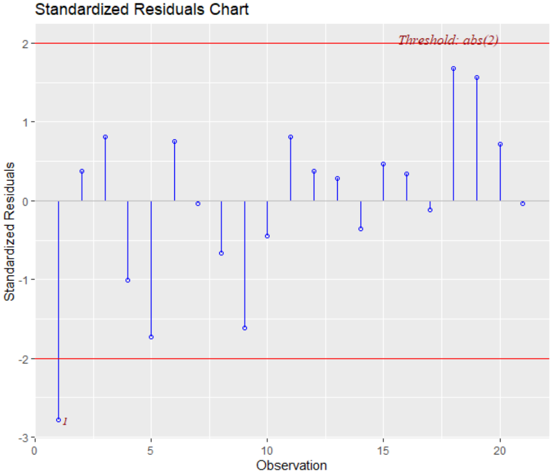
DFFITS Visualization

```
ols_plot_dffits(lm0)
```



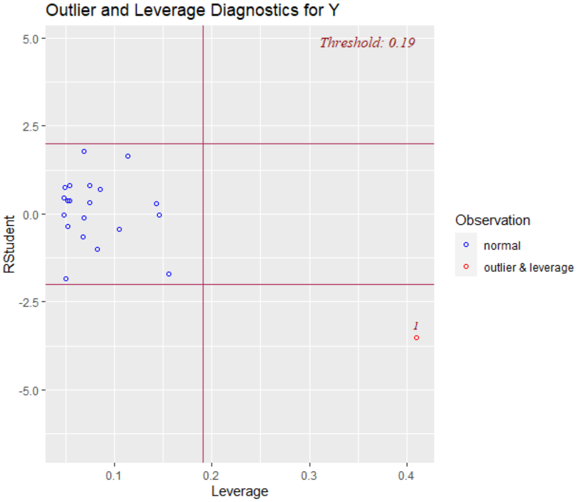
Residuals Visualization

```
ols_plot_resid_stand(lm0)
```



Residuals and Leverage Visualization

```
ols_plot_resid_lev(lm0)
```



Practice exercise

Practice exercise:

- Replace first row of oil data set with (X,Y)=(150, 115). Confirm the first observation is not an outlier, has little influence, and has high leverage
- Replace first row with (114, 115). Confirm the first observation is an outlier, has little influence, and has low leverage
- Delete first row. Confirm there are no outliers, influential points or leverage points.