STAC67: Regression Analysis

Lecture 21

Sohee Kang

Mar. 30, 2021

Case Study: Influential Observations in Regression

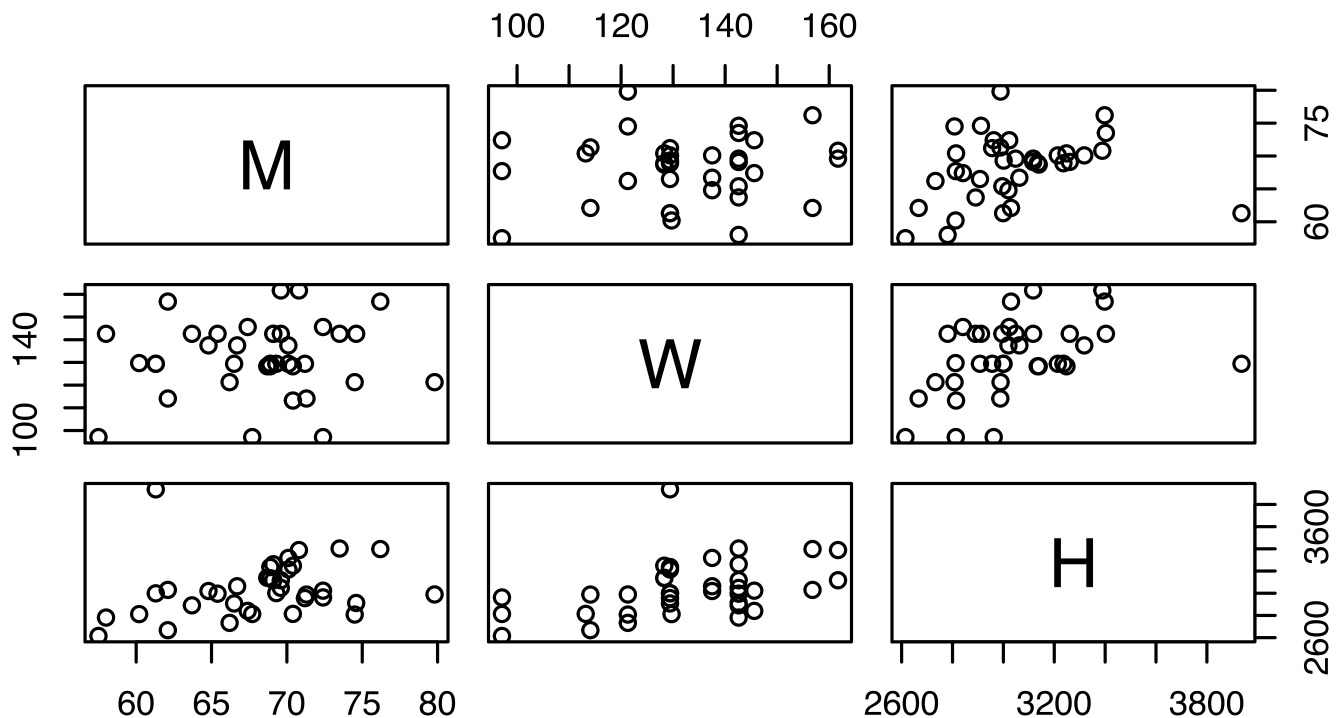
- Measurements on Heat Production as a Function of Body Mass and Work Effort. (M. Greenwood (1918). "On the Efficiency of Muscular Work," Proc. Roy. Soc. Of London, Series B, Vol. 90, #627, pp. 199-214)

- Study involved Algerians accustomed to heavy labor. Experiment consisted of several hours on stationary bicycle.
- Dependent (Response) Variable: Heat Production (Calories)
- Independent (Explanatory/Predictor) Variables:
  - Work Effort (Calories)
  - Body Mass (kg)
- Model:

$$H = \beta_0 + \beta_1 W + \beta_2 M + \epsilon$$

R codes

```
muscle <- read.table("muscle.txt",  header=F, col.names=c("M","W","H"))
attach(muscle)
par(mfrow=c(2,2))
pairs(muscle)
```



Regression Output

```r
muscle.reg = lm(H~M + W, data=muscle)
summary(muscle.reg)
```

```
##
## Call:
## lm(formula = H ~ M + W, data = muscle)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -275.4 -133.3  -34.9  116.0  981.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1536.510    584.499   2.629   0.0128 *
## M             10.141      7.683   1.320   0.1957
## W              6.156      2.366   2.602   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 228.7 on 34 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.1681
## F-statistic: 4.637 on 2 and 34 DF,  p-value: 0.01657
```

R codes

```
muscle.rstandard <- rstandard(muscle.reg)
muscle.rstudent <- rstudent(muscle.reg)
muscle.inf <- influence.measures(muscle.reg)
cbind (muscle.rstandard, muscle.rstudent)
```

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

$$d_i = e_i\sqrt{\frac{(n-p'-1)}{SSE(1-h_{ii})-e_i^2}}$$

```
##    muscle.rstandard muscle.rstudent
## 1        0.58563043     0.579886056
## 2        0.11842487     0.116694398
## 3       -0.32219190    -0.317904101
## 4       -1.05819989    -1.060125719
## 5       -1.18074112    -1.187856269
## 6        0.48796380     0.482426540
## 7        0.65047123     0.644859117
## 8       -0.47957241    -0.474073372
## 9        0.51463172     0.508993419
## 10      -0.36591523    -0.361205888
## 11       0.92466805     0.922643410
## 12       0.00829637     0.008173463
## 13      -0.76543328    -0.760675366
## 14      -0.92799699    -0.926051220
## 15       1.10098858     1.104545180
## 16       0.20157243     0.198704776
## 17       0.99680755     0.996711285
## 18      -0.50671098    -0.501099385
## 19       4.47250058     6.867425403
## 20      -0.09004197    -0.088718519
## 21      -0.61445053    -0.608736315
## 22       0.45455892     0.449191331
## 23       0.90583576     0.903382685
## 24       0.75573743     0.750874078
## 25       0.64868128     0.643062346
## 26      -0.44198249    -0.436690557
```

R codes

```
muscle.inf
```

$DFBETAS_{(i)}$

h_{ii} $H = X(X'X)^{-1}X'$

```
## Influence measures of
##   lm(formula = H ~ M + W, data = muscle) :
##
##        dfb.1_     dfb.M     dfb.W    dffit  cov.r   cook.d    hat inf
## 1  -0.207960  0.154473  0.142251  0.24404 1.2488 2.02e-02 0.1505
## 2   0.000771  0.015080 -0.024251  0.03392 1.1846 3.95e-04 0.0779
## 3   0.025193 -0.012379 -0.032720 -0.06449 1.1283 1.42e-03 0.0395
## 4  -0.290650  0.407014 -0.160887 -0.46551 1.1799 7.20e-02 0.1616
## 5   0.271727 -0.255362 -0.104618 -0.35162 1.0491 4.07e-02 0.0806
## 6   0.004159  0.014255 -0.021927  0.08433 1.1036 2.43e-03 0.0296
## 7  -0.041760  0.014051  0.067416  0.12909 1.0956 5.65e-03 0.0385
## 8  -0.035315  0.116785 -0.139433 -0.19305 1.2493 1.27e-02 0.1422
## 9   0.007315  0.011566 -0.022791  0.08844 1.1006 2.67e-03 0.0293
## 10 -0.015311  0.038119 -0.042461 -0.08170 1.1361 2.28e-03 0.0487
## 11 -0.031947  0.074671 -0.046680  0.17603 1.0501 1.04e-02 0.0351
## 12 -0.000529  0.000178  0.000854  0.00164 1.1375 9.19e-07 0.0385
## 13 -0.070370  0.125814 -0.094523 -0.19839 1.1087 1.33e-02 0.0637
## 14 -0.260191  0.180236  0.165008 -0.30301 1.1211 3.07e-02 0.0967
## 15 -0.215148  0.193430  0.100684  0.29527 1.0509 2.89e-02 0.0667
## 16  0.045307 -0.047084 -0.001748  0.05846 1.1841 1.17e-03 0.0797
## 17 -0.069121  0.060946  0.047105  0.18530 1.0352 1.14e-02 0.0334
## 18  0.150276 -0.225700  0.085812 -0.25209 1.3397 2.17e-02 0.2020   *
## 19  1.565858 -1.627281 -0.060405  2.02030 0.0829 5.77e-01 0.0797   *
## 20 -0.007439  0.010699 -0.005794 -0.01897 1.1429 1.24e-04 0.0437
## 21 -0.159273  0.169565  0.001104 -0.20042 1.1723 1.36e-02 0.0978
## 22  0.026977  0.089002 -0.189276  0.21814 1.3271 1.62e-02 0.1908   *
## 23  0.003108  0.025661 -0.030610  0.15545 1.0465 8.10e-03 0.0288
## 24 -0.023371  0.052148 -0.028501  0.13793 1.0746 6.42e-03 0.0326
## 25 -0.137400  0.040593  0.202080  0.23924 1.1994 1.94e-02 0.1216
## 26 -0.031669  0.023346  0.011316 -0.07778 1.1090 2.07e-03 0.0307
```

R codes for graphical diagnostics

```
library(ggpubr)
library(olsrr)
ols_plot_cooksd_chart(muscle.reg) # Cook's Distance #
```

```
ols_plot_resid_stud(muscle.reg) # t*: deleted studentized residual #
```

```
ols_plot_dfbetas(muscle.reg) # DFBETAS #
```

```
ols_plot_dffits(muscle.reg)  # DIFFITS #
```

```
ols_plot_resid_lev(muscle.reg) # Studentized Residual vs Leverages #
```

```
ols_plot_resid_stud_fit(muscle.reg) # Deleted Studentized Residual vs Predicted valu
```

```
p1 <- ols_plot_cooksd_chart(muscle.reg)
```

```
p2 <- ols_plot_resid_stud(muscle.reg)
```

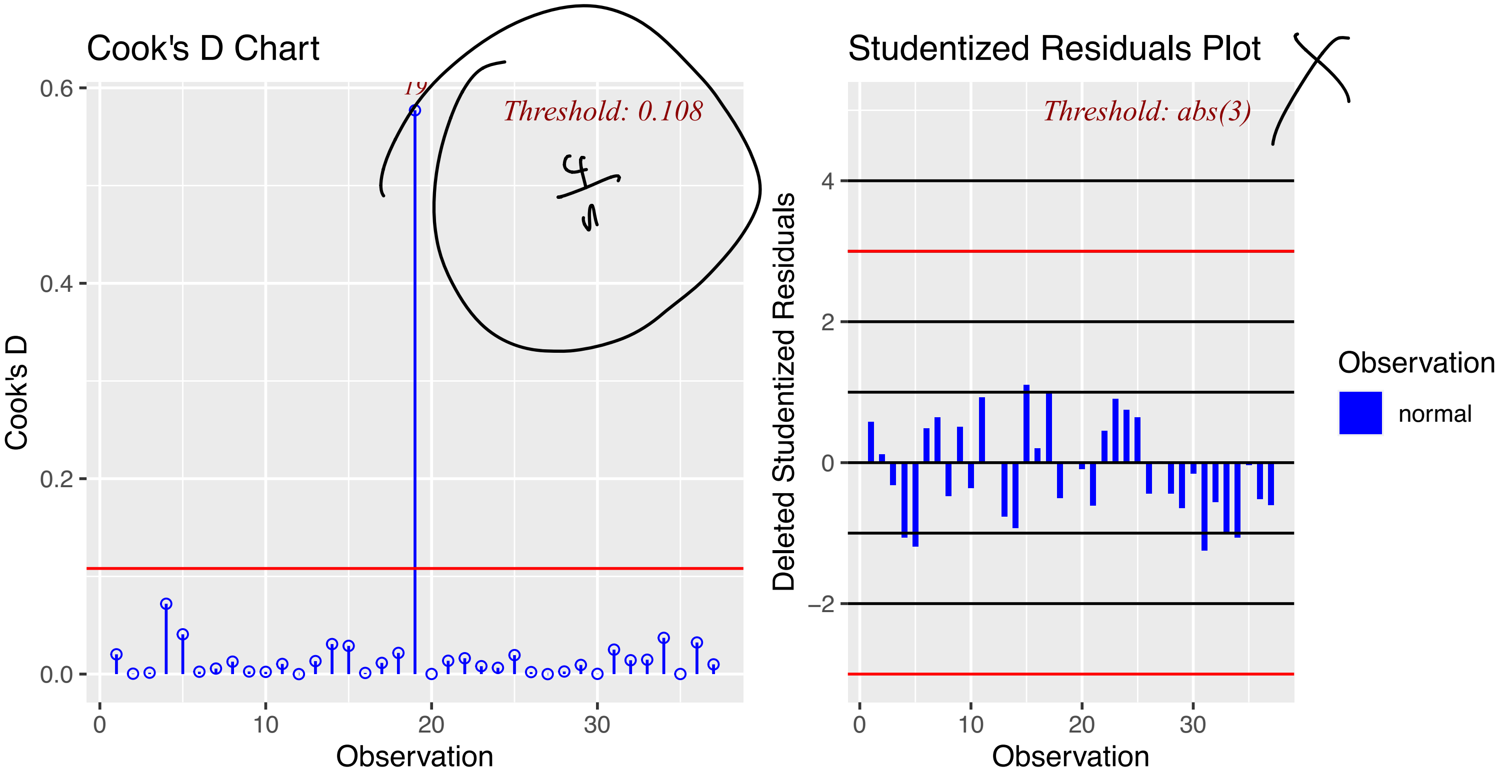
```
p4 <- ols_plot_dffits(muscle.reg)
```

```
p5 <- ols_plot_resid_lev(muscle.reg)
```

```
p6 <- ols_plot_resid_stud_fit(muscle.reg)
```

R Diagnostic Graphs

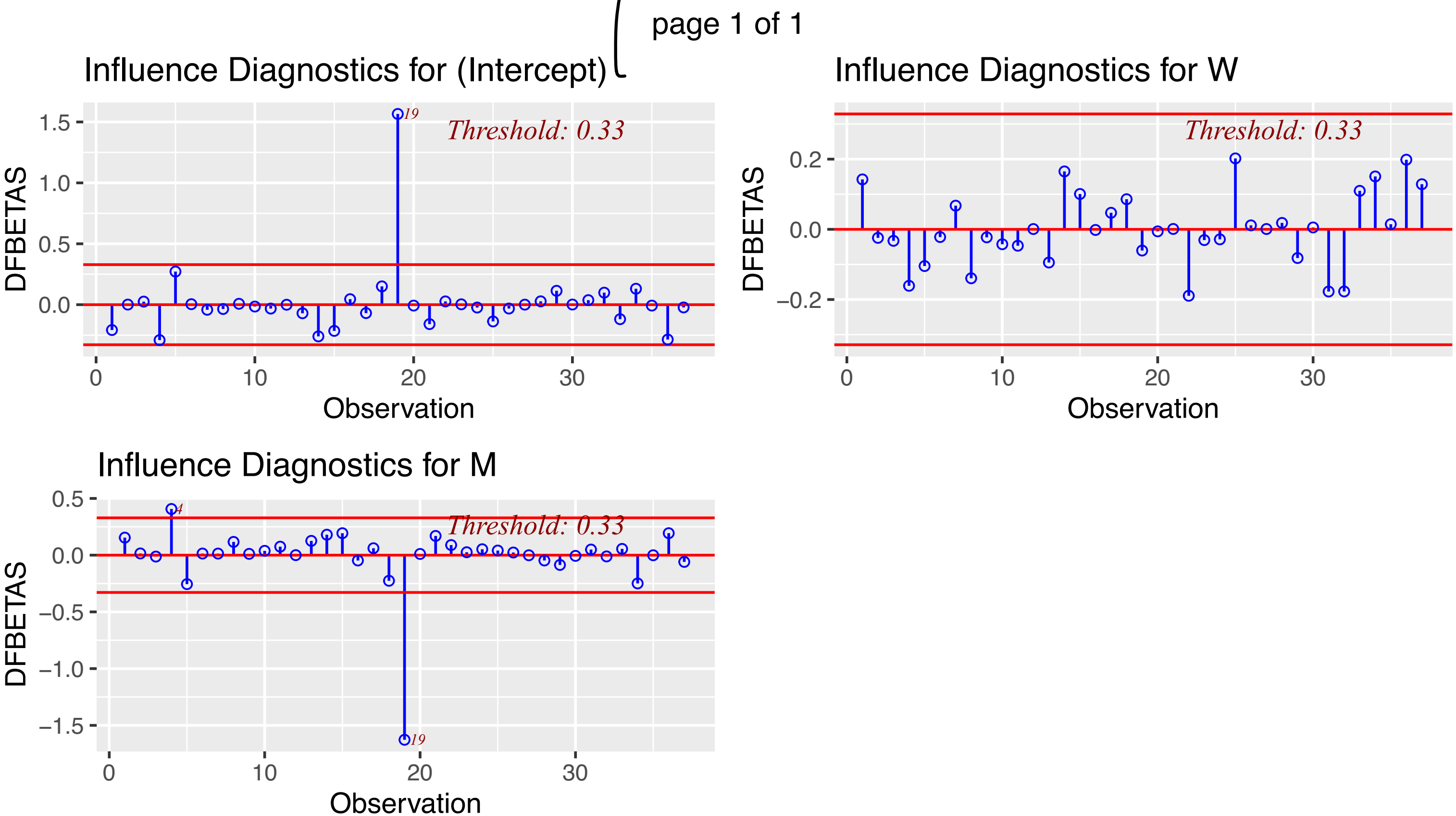
```
ggarrange(p1, p2, ncol=2, nrow=1)
```



R Diagnostic Graphs

$\frac{2}{\sqrt{n}}$

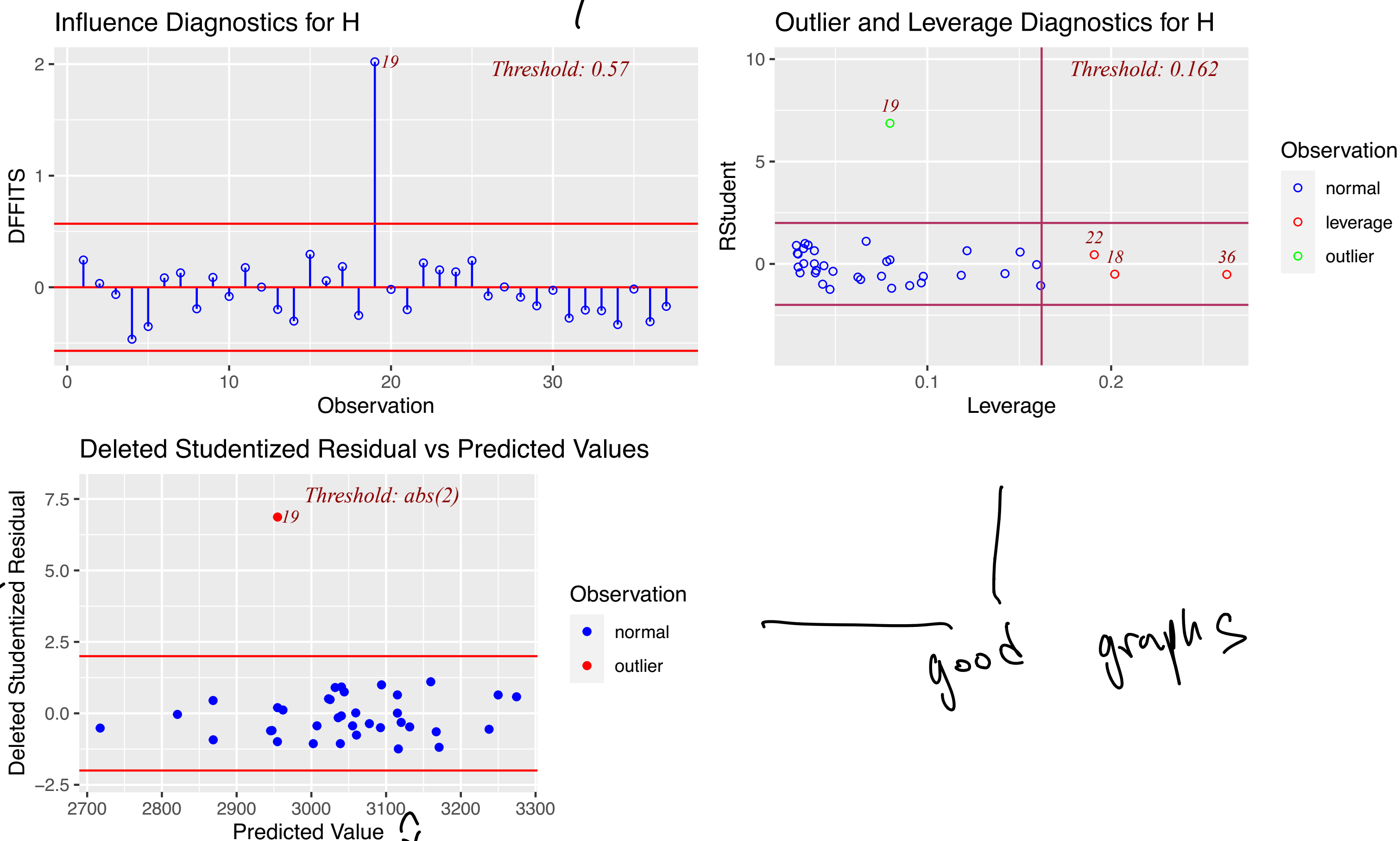
```
ols_plot_dfbetas(muscle.reg)
```



R Diagnostic Graphs

```
ggarrange(p4, p5, p6, ncol=2, nrow=2)
```

$2\sqrt{\frac{p'}{n}}$



t^*

$\hat{Y}$

good graphs

Influential Measures

$qt\left(0.05/_{2\cdot 37}, 33, \text{lower.tail}=F\right)$

- Note: n= 37, p'= 3 parameters
- Studentized residuals:

  outlier if $|t^*| > t\left(1-\frac{0.05}{2\cdot 37}; 37-3-1\right)$

- Leverage values: $h_{ii} > \frac{2p'}{n}$
- DFFITS $t_i\sqrt{\frac{h_{ii}}{(1-h_{ii})}}$ > 1 for small, $> 2\sqrt{\frac{p'}{n}}$ for large
- DFBETAS: > 1 for small, $> 2/\sqrt{n}$ for large
- Cook's Distance: $D_i = \frac{e_i^2}{p'MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2} > qf(0.5, 3, 34)$
- Covariance Ratio: highly influential on stadard Errors of Regression Coefficients $\det\left(MSE_{(i)}(X'_{(i)}X_{(i)})^{-1}\right) / \det\left(MSE(X'X)^{-1}\right)$

  criterion : outside of $(1 \pm 3\cdot p'/n)$

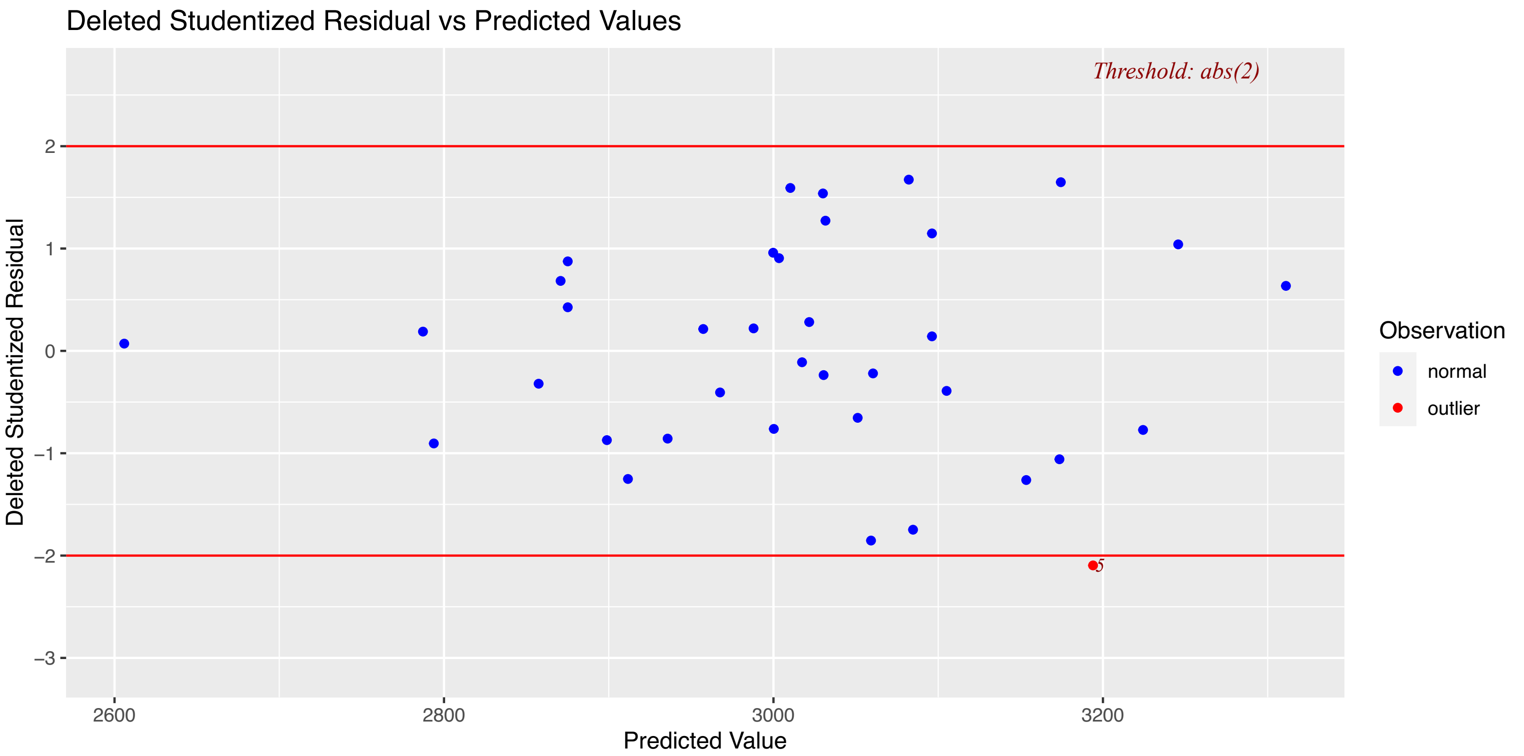
Diagnosing Influential Observations

- Clearly, Observation #19 exerts a huge influence (although it has a small hat or leverage value, so it must be near center of Mass/Work observations
- Upon further review to author's original calculations provided in paper, the mean and S.D. are much to high for H (but exactly the same for M and W). Could observation been a "typo"? Try replacing H_{19}=3936 with H_{19}=2936
- Note: Do not do this arbitrarily, check your data sources in practice

```
##
## Call:
## lm(formula = H ~ M + W, data = muscle2)
##
## Residuals:
##     Min      1Q Median      3Q     Max
## -282.0 -109.2    9.1  123.9  235.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  977.425    376.053   2.599 0.013723 *
## M             17.778      4.943   3.597 0.001011 **
## W              6.244      1.522   4.102 0.000242 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.1 on 34 degrees of freedom
## Multiple R-squared:  0.4922, Adjusted R-squared:  0.4624
## F-statistic: 16.48 on 2 and 34 DF,  p-value: 9.914e-06
```

Plot of Residuals versus Predicted Values

```
ols_plot_resid_stud_fit(fit2)
```



Chapter 11

Building the Regression Model III: Remedial Measures

Unequal (Independent) Error Variances - Weighted Least Squares (WLS)

- Case 1 - Error Variances known exactly (VERY rare)
- Case 2 - Error Variances known up to a constant
  - Occasionally information known regarding experimental units regarding the relative magnitude (unusual)
  - If "observations" are means of different numbers of units (each with equal variance) at the various X levels, Variance of observation i is σ^2/n_i where n_i is known
- Case 3 - Variance (or Standard Deviation) is related to one or more predictors, and relation can be modeled (see Breusch-Pagan Test)
- Case 4 - Ordinary Least Squares with correct variances

Checking Equal Variance

- Two tests for equal variance are the Brown-Forsyth test and the Breusch-Pagan (aka Cook-Weisberg) test.
- Breusch-Pagan Test(aka Cook-Weisberg Test) - - Fits a regression of the squared residuals on X and tests whether the variance is related to X.

$$H_0 : \qquad\qquad vs \qquad\qquad H_a :$$

- When the regression of the squared residuals is fit, we obtain SSR_{e2} , the regression sum of squares. The test is conducted as follows, where SSE is the Error Sum of Squares for the original regression of Y on X.