Residuals, Outliers, and Diagnostics

Dr. Michael Fix
mfix@gsu.edu

Georgia State University

3 April 2025

Some Definitions

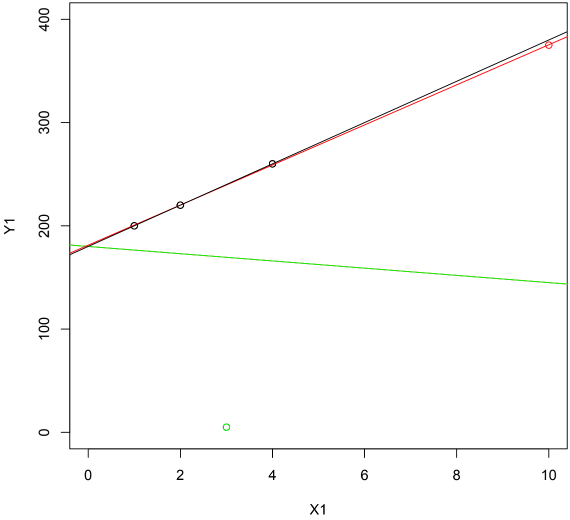
Leverage: the degree of influence an observation **CAN**—but not necessarily does—have on coefficient estimates

Discrepancy: the degree to which an observation is different from the rest of the data

Influence: Leverage * Discrepancy. What is the effect of an observations values for Y and $\mathbf{X}$ have on the coefficient estimates

Outlier: An observation with an unusual value for Y given its values for $\mathbf{X}$

An Illustration



Leverage

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$$h_i = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i'$$

Residuals

Variation:

$$\widehat{\text{Var}(\hat{u}_i)} = \hat{\sigma}^2[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i']$$

$$\begin{aligned}\widehat{\text{s.e.}(\hat{u}_i)} =& \hat{\sigma}\sqrt{[1 - \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i']} \\ =& \hat{\sigma}\sqrt{1 - h_i}\end{aligned}$$

"Standardized" Residuals:

$$\tilde{u}_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Residuals

"Studentized" Residuals:

$$\hat{\sigma}^2_{-i} = \text{Variance for the } N-1 \text{ observations} \neq i$$
$$= \frac{\hat{\sigma}^2(N-K)}{N-K-1} - \frac{\hat{u}_i^2}{(N-K-1)(1-h_i)}.$$

Then:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}_{-i}\sqrt{1-h_i}}$$

Why Does this Matter?

- The $\hat{u}'_i$ follow a t distribution with $N - K - 1$ degrees of freedom
- This means that approximately 95% fall on the interval [-2,2]
- This allows for hypothesis testing

Influence

- If influence is effective a measure of how unusual an observation is (discrepancy) combined with where it is located (leverage), then how can we measure this?
- DFBETA and DFBETAS (the "S" is for standardized) do this
  - Where positive values correspond with observations that *decrease* the value of $\hat{beta}_k$
  - And negative values correspond with observations that *increase* the value of $\hat{beta}_k$
- Plots of DFBETAs and DFBETASs generally reveal when specific observations are highly influential

Influence

"DFBETA":

$$D_{ki} = \hat{\beta}_k - \hat{\beta}_{k(-i)}$$

"DFBETAS" (the "S" is for "standardized):

$$D^*_{ki} = \frac{D_{ki}}{\widehat{\text{s.e.}(\hat{\beta}_{k(-i)})}}$$

Influence

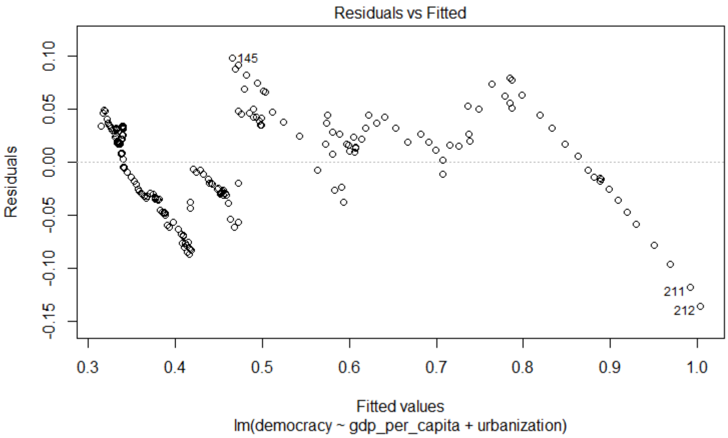
- Cook's D is a summary statistic calculated from DFBETAs to measure each observations influence on the overall regression model

$$
\begin{aligned}
D_i =& \frac{\tilde{u}_i^2}{K} \times \frac{h_i}{1 - h_i} \\
=& \frac{h_i \hat{u}_i^2}{K \hat{\sigma}^2 (1 - h_i)^2}
\end{aligned}
$$

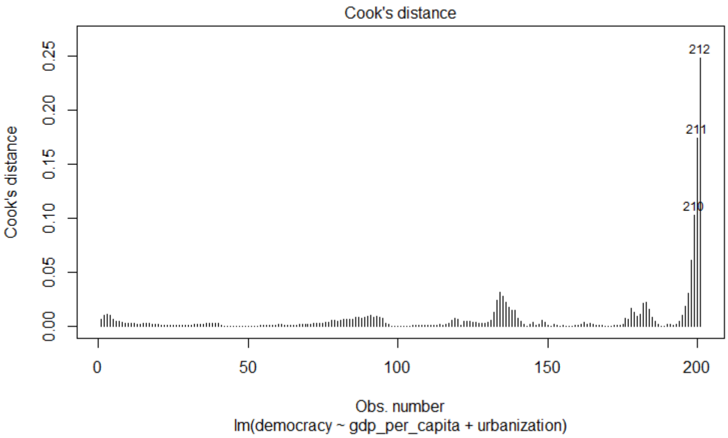
Toy Model

```
Predictors of democracy in the US
================================================
                        Dependent variable:
                    ---------------------------
                             democracy
------------------------------------------------
gdp_per_capita                 0.013
                             (0.0003)
                            t = 38.604
                           p = 0.000***

urbanization                   0.253
                              (0.056)
                             t = 4.562
                          p = 0.00001***

Constant                       0.264
                              (0.010)
                            t = 25.127
                           p = 0.000***

------------------------------------------------
Observations                    201
R2                             0.942
Adjusted R2                    0.942
Residual Std. Error      0.044 (df = 198)
F Statistic         1,615.404*** (df = 2; 198)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
>
```

plot(my_model)



plot(my_model)

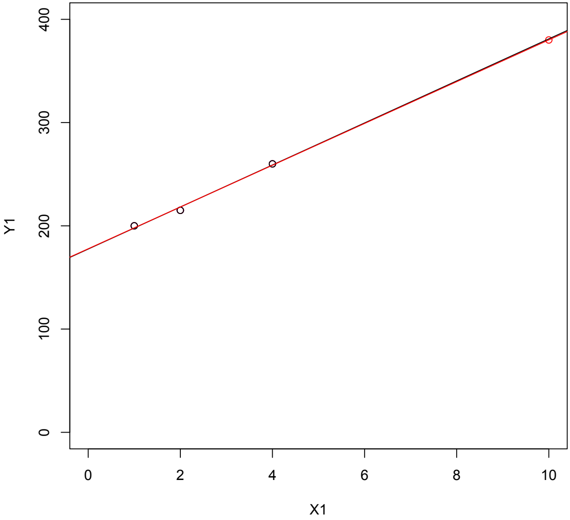


Or, do everything manually?

```r
# Or, you can manually find observations ----
# Get cook's D
cooksD <- cooks.distance(my_model)
# any values greater than 3x the mean
cooksD[(cooksD > (3 * mean(cooksD, na.rm = TRUE)))]

# Get DFBETA for gdp_per_capita
# I decided that 2/sqrt(n) is my threshold for suspicious dfbetas
my_data[which(abs(dfbetas(my_model)[,'gdp_per_capita']) > 0.15),]
```

```
> cooksD[(cooksD > (3 * mean(cooksD, na.rm = TRUE)))]
       144        145        146        147        193        194        208        209        210        211        212 
0.02368734 0.03163166 0.02753306 0.02200472 0.02124732 0.02206320 0.03023971 0.06099106 0.10293906 0.17365460 0.24792651 
>
> # Get DFBETA for gdp_per_capita
> # I decided that 2/sqrt(n) is my threshold for suspicious dfbetas
> my_data[which(abs(dfbetas(my_model)[,'gdp_per_capita']) > 0.15),]
    year democracy gdp_per_capita urbanization
133 1921     0.520         10.155        0.281
134 1922     0.522         10.459        0.282
135 1923     0.531         11.076        0.284
136 1924     0.532         11.382        0.285
178 1966     0.708         25.415        0.298
182 1970     0.709         27.314        0.296
183 1971     0.731         27.914        0.296
184 1972     0.741         28.732        0.295
195 1983     0.862         34.138        0.293
196 1984     0.863         35.674        0.292
197 1985     0.865         36.750        0.292
198 1986     0.865         37.846        0.292
199 1987     0.868         38.917        0.292
200 1988     0.868         39.853        0.292
201 1989     0.870         40.848        0.291
```

Our Earlier Illustration



Regression: Black Line

```
> summary(lm(Y1~X1))

Call:
lm(formula = Y1 ~ X1)

Residuals:
     1      2      3
 2.143 -3.214  1.071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  177.500      4.910   36.15   0.0176 *
X1            20.357      1.856   10.97   0.0579 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.009 on 1 degrees of freedom
Multiple R-squared:  0.9918,Adjusted R-squared:  0.9835
F-statistic: 120.3 on 1 and 1 DF,  p-value: 0.05787
```

Regression: Red Line

```
> summary(lm(Y3~X3))

Call:
lm(formula = Y3 ~ X3)

Residuals:
       1        2        3        4
 2.00000 -3.23077  1.30769 -0.07692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  177.769      2.239   79.41 0.000159 ***
X3            20.231      0.407   49.70 0.000405 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.842 on 2 degrees of freedom
Multiple R-squared:  0.9992,Adjusted R-squared:  0.9988
F-statistic:  2470 on 1 and 2 DF,  p-value: 0.0004046
```

A Variance-Based Statistic

- "COVRATIO" is a statistic that provides an estimate of whether a particular observation has a large effect on the variance-covariance estimates of our parameters

$$\text{COVRATIO}_i = \left[(1-h_i)\left(\frac{N-K-1+\hat{u}_i'^2}{N-K}\right)^K\right]^{-1}$$

- Observations with $\text{COVRATIO}_i > 1$ *increase* the precision of our estimates (decrease S.E. estimates)
- Observations with $\text{COVRATIO}_i < 1$ *decrease* the precision of our estimates (increase S.E. estimates)

COVRATIO statistics in R

```
> influence.measures(my_model)
Influence measures of
	 lm(formula = democracy ~ gdp_per_capita + urbanization, data = my_data) :

     dfb.1_ dfb.gd__ dfb.urbn   dffit cov.r   cook.d     hat inf
12 1.36e-01  0.04288 -0.11563 0.13742 1.036 6.31e-03 0.02939
13 1.77e-01  0.05262 -0.14814 0.17881 1.026 1.07e-02 0.02728
14 1.81e-01  0.05045 -0.14962 0.18390 1.021 1.13e-02 0.02528
15 1.69e-01  0.04426 -0.13840 0.17310 1.021 9.98e-03 0.02369
16 1.38e-01  0.03340 -0.11101 0.14156 1.024 6.68e-03 0.02218
17 1.18e-01  0.02619 -0.09346 0.12178 1.026 4.95e-03 0.02075
18 1.08e-01  0.02171 -0.08428 0.11250 1.025 4.23e-03 0.01940
19 1.02e-01  0.01840 -0.07829 0.10696 1.025 3.82e-03 0.01837
20 9.15e-02  0.01401 -0.06886 0.09720 1.025 3.16e-03 0.01715
21 8.37e-02  0.01113 -0.06196 0.08979 1.024 2.69e-03 0.01625
22 7.89e-02  0.00878 -0.05735 0.08558 1.024 2.45e-03 0.01539
23 7.91e-02  0.00845 -0.05721 0.08593 1.023 2.47e-03 0.01519
24 7.07e-02  0.00667 -0.05063 0.07736 1.024 2.00e-03 0.01479
25 6.11e-02  0.00543 -0.04355 0.06709 1.026 1.51e-03 0.01459
26 7.94e-02  0.00660 -0.05627 0.08740 1.022 2.55e-03 0.01440
27 7.90e-02  0.00543 -0.05530 0.08762 1.021 2.56e-03 0.01401
28 7.63e-02  0.00446 -0.05301 0.08510 1.021 2.42e-03 0.01382
29 7.35e-02  0.00355 -0.05062 0.08233 1.022 2.27e-03 0.01363
30 7.45e-02  0.00300 -0.05101 0.08391 1.021 2.35e-03 0.01345
31 7.30e-02  0.00239 -0.04959 0.08249 1.021 2.27e-03 0.01328
32 5.71e-02  0.00148 -0.03859 0.06487 1.024 1.41e-03 0.01311
```

So What Do We Do with Outliers?

- Two relevant questions here:
  1. Is the outlier due to a coding error or mistake of some sort?
  2. Is the outlier correctly coded but a true outlier (i.e. weird)?

Dealing with Outliers: Coding Errors

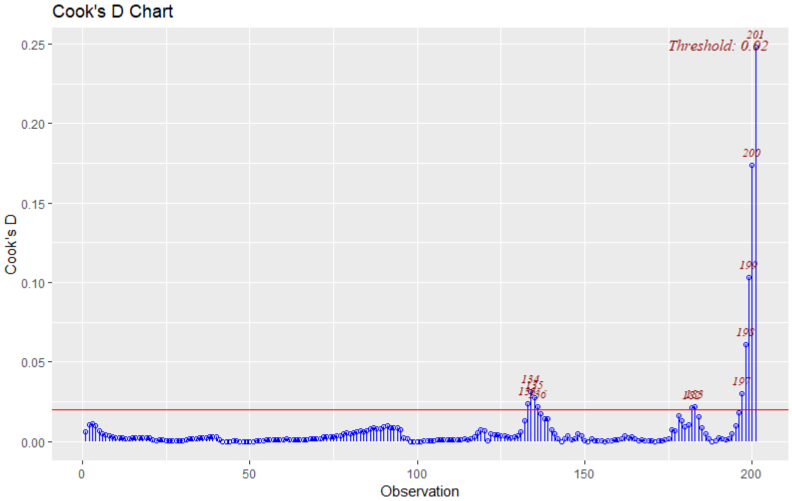
- Here the answer is simple, just fix the error and your problem is solved
- If you cannot fix the issue, drop the observation
  - Can assume the observation is "missing at random"
  - Could also use imputation approaches for solving missing data issues

Dealing with Outliers: True Outliers

- If there some reason why that observation is very different:
  - AND that reason is theoretically important, then this is now a *theoretical* issue and you may need to revisit your theory or model selection to account for it
  - AND that reason is NOT theoretically important, then you can probably safely drop it
- If there is no reason why that observation is very different:
  - There isn't an easy answer: look at that data point more deeply and make a judgement call
  - Whether you keep or drop, you probably need to run the alternative specification as a robustness check and footnote it

Cook's Distance (using *olsrr* package)

```
# Bar Plot of Cook's distance to detect observations that strongly influence
# fitted values of the model.
ols_plot_cooksd_bar(my_model)
ols_plot_cooksd_chart(my_model)
```



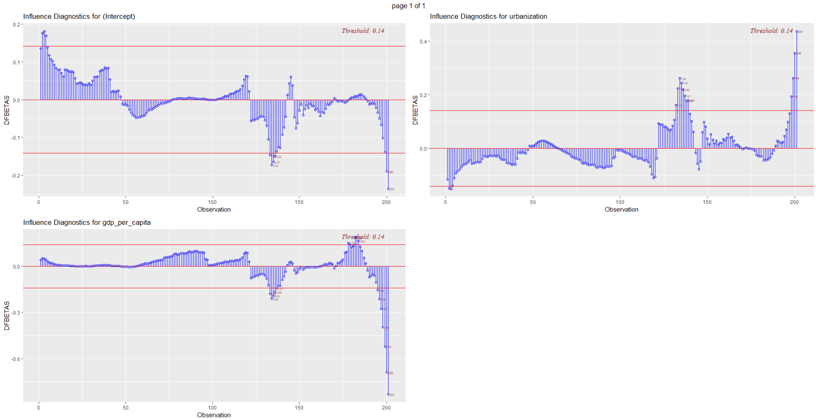
Let's check these observations

```
> my_data[197:201, ]
    year democracy gdp_per_capita urbanization
197 1985     0.865         36.750        0.292
198 1986     0.865         37.846        0.292
199 1987     0.868         38.917        0.292
200 1988     0.868         39.853        0.292
201 1989     0.870         40.848        0.291
> # maybe after-effect of oil crisis? Or, Reagan defeated Carter? More conservative economy?
```

DFBETAs (using *olsrr* package)

```
## DFBETAs ----
# DFBETA measures the difference in each parameter estimate with and without the influential point.
ols_plot_dfbetas(my_model)

# Let's check these observations
```

DFBETAs (using *olsrr* package)



Let's check these observations

```
> my_data[134:136, ]
    year democracy gdp_per_capita urbanization
134 1922     0.522         10.459        0.282
135 1923     0.531         11.076        0.284
136 1924     0.532         11.382        0.285
> # Supreme Court rejected women's right to vote? US occupies Haiti?
```

What happens if we drop influential points and outliers?

	Dependent variable: democracy	
	full sample (1)	influential/outliers removed (2)
gdp_per_capita	0.013	0.014
	(0.0003)	(0.0004)
	t = 38.604	t = 35.791
	p = 0.000***	p = 0.000***
urbanization	0.253	0.240
	(0.056)	(0.057)
	t = 4.562	t = 4.193
	p = 0.00001***	p = 0.00005***
Constant	0.264	0.265
	(0.010)	(0.011)
	t = 25.127	t = 24.884
	p = 0.000***	p = 0.000***
Observations	201	193
R2	0.942	0.937
Adjusted R2	0.942	0.936
Residual Std. Error	0.044 (df = 198)	0.044 (df = 190)
F Statistic	1,615.404*** (df = 2; 198)	1,402.193*** (df = 2; 190)
Note:		*p<0.1; **p<0.05; ***p<0.01