Regression III
Outliers and Influential Data

Dave Armstrong

University of Wisconsin – Milwaukee
Department of Political Science

e: armstrod@uwm.edu
w: www.quantoid.net/ICPSR.php

Regression Diagnostics

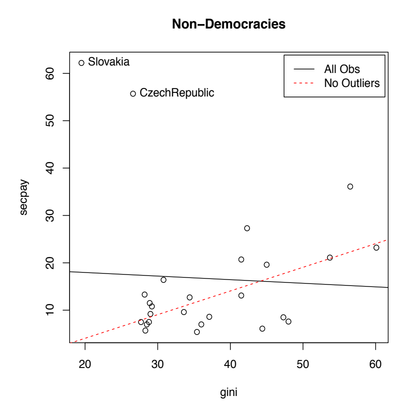
- Today's lecture deals specifically with unusual data and how they are identified and measured
- Regression Outliers
  - Studentized residuals (and the Bonferroni adjustment)
- Leverage
  - Hat values
- Influence
  - DFBETAs, Cook's D, influence plots, added-variable plots (partial regression plots)
- Robust and resistant regression methods that limit the effect of such cases on the regression estimates will be discussed later in the course

Outlying Observations: Who Cares?

- Can cause us to misinterpret patterns in plots
  - Temporarily removing them can sometimes help see patterns that we otherwise would not have
  - Transformations can also spread out clustered observations and bring in the outliers
- More importantly, separated points can have a strong influence on statistical models - removing outliers from a regression model can sometimes give completely different results
  - Unusual cases can substantially influence the fit of the OLS model - Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model
  - Outliers may also indicate that our model fails to capture important characteristics of the data

Example 1. Influence and Small Samples: Inequality Data

- Small samples are especially vulnerable to outliers - there are fewer cases to counter the outlier
- With Czech Republic and Slovakia included, there is no relationship between Attitudes towards inequality and the Gini coefficient
- If these cases are removed, we see a positive relationship



Code for Previous Figure

```
> weakliem2 <- read.csv("http://www.quantoid.net/702/Weakliem2.txt")
> rownames(weakliem2) <- weakliem2[,1]
> outs <- which(rownames(weakliem2) %in% c("CzechRepublic", "Slovakia"))
> weakliem2$gdp <- weakliem2$gdp/10000
> plot(secpay ~ gini, data=weakliem2, main="Non-Democracies")
> abline(lm(secpay ~ gini, data=weakliem2))
> abline(lm(secpay ~ gini, data=weakliem2, subset=-outs), lty=2, col="red"
> with(weakliem2, text(gini[outs], secpay[outs],
+  rownames(weakliem2)[outs], pos=4))
> legend("topright", c("All Obs", "No Outliers"),
+  lty=c(1,2), col=c("black","red"), inset=.01)
```


Ex 1. Influence and Small Samples: Inequality Data (2)

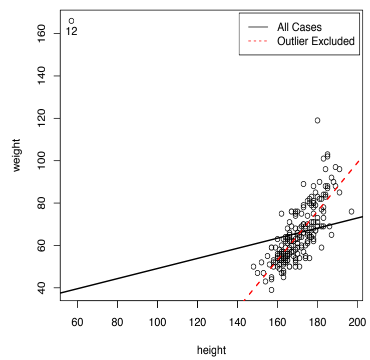
	All Obs	No Outliers
(Intercept)	19.4771	−5.9234
	(11.0655)	(5.2569)
gini	−0.0759	0.4995*
	(0.2884)	(0.1336)
N	26	24
R^2	0.0029	0.3887
adj. R^2	−0.0387	0.3609
Resid. sd	14.8520	6.2702

Standard errors in parentheses

* indicates significance at $p < 0.05$

Example 2. Influence and Small Samples: Davis Data (1)

- These data are the Davis data in the car package
- It is clear that observation 12 is influential
- The model including observation 12 does a poor job of representing the trend in the data; The model excluding observation 12 does much better
- The output on the next slide confirms this



R-script for previous slide

```
> library(car)
> data(Davis)
> plot(weight ~ height, data=Davis)
> with(Davis, text(height[12], weight[12], "12", pos=1))
> abline(lm(weight ~ height, data=Davis),
+  lty=1, col=1, lwd=2)
> abline(lm(weight ~ height, data=Davis, subset=-12),
+  lty=2, col=2, lwd=2)
> legend("topright", lty=c(1,2), col=c(1,2),
+  legend=c("All Cases", "Outlier Excluded"), inset=.01)
```

Example 2. Influence and Small Samples: Davis Data (2)

	All Obs	No Outliers
(Intercept)	25.27	-130.75^*
	(14.95)	(11.56)
height	0.24^*	1.15^*
	(0.09)	(0.07)
N	200	199
R^2	0.04	0.59
adj. R^2	0.03	0.59
Resid. sd	14.86	8.52

Standard errors in parentheses

* indicates significance at $p < 0.05$

Example 3. Large Datasets: Contrived Data

- Although regression models from small datasets are most vulnerable to unusual observations, large datasets are not completely immune
- An unusually high (or low) x or y value could easily result from miscoding during the data entry stage. This could in turn influence the findings
  - Imagine a dataset with 1001 observations, where a variable, X_1, ranges from 0.88-7.5.
  - Assume also that Y is perfectly correlated with X_1.
  - Even if there is just one miscode - e.g., A "55" is wrongly entered instead of "5" - the distribution of X_1 is drastically misrepresented. This one miscode also seriously distorts the regression line.

```
> set.seed(123)
> x<-c(rnorm(1000,mean=4,sd=1))
> x1<-c(x,55)
> y<-c(x,5)
> range(x1)
[1]  1.190225 55.000000
> range (y)
[1] 1.190225 7.241040
```

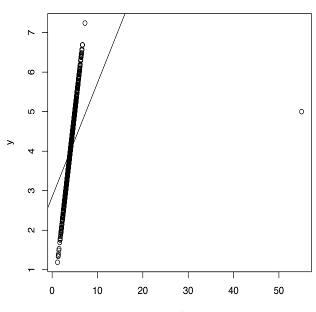

Example 3. Large Datasets: Contrived Data (2)

```
> mod1 <- lm(y ~ x1)
> apsrtable(mod1, model.names="", Sweave=T)
```

(Intercept)	2.84^*
	(0.06)
x1	0.29^*
	(0.01)
N	1001
R^2	0.30
adj. R^2	0.30
Resid. sd	0.83

Standard errors in parentheses

* indicates significance at $p < 0.05$



Example 4. Large Datasets: Marital Coital Frequency (1)

- Jasso, Guillermina (1985) 'Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences,' *American Sociological Review*, 50: 224-241.
  - Using panel data, estimates age and period effects - controlling for cohort effects - on frequency of sexual relations for married couples from 1970-75
- Major Findings:
  - Controlling for cohort and age effects, there was a negative period effect;
  - Controlling for period and cohort effects, wife's age had a positive effect
  - Both findings differ significantly from previous research in the area

Example 4. Large Datasets: Marital Coital Frequency (2)

- Kahn, J.R. and J.R. Udry (1986) ‘Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions,’ *American Sociological Review*, 51: 734-737, critiques and replicates Jasso’s research. They claim that Jasso:
  1. Failed to check the data for influential outliers
     - 4 cases were seemingly miscoded 88 (must be missing data - coded 99 - since no other value was higher than 63 and 99.5% were less than 40)
     - 4 additional cases had very large studentized residuals (each was also largely different from the first survey)
  2. Missed an interaction between length of marriage and wife’s age
- Dropping the 8 outliers (from a sample of more than 2000) and adding the interaction drastically changes the findings

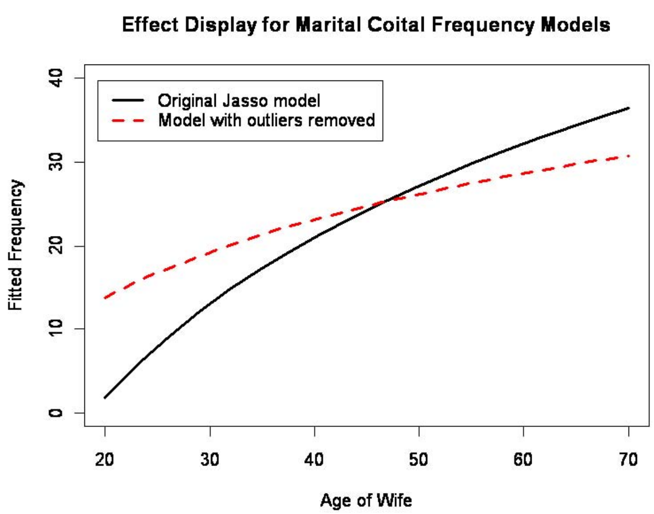
Example 4. Large Datasets: Marital Coital Frequency (3)

Table 1. Fixed-Effects Estimates of the Determinants of Marital Coital Frequency

	(1) Jasso’s Results	(2) Our Replication	(3) Drop 4 Miscodes	(4) Drop 4 Miscodes & 4 Outliers	(5) Marital Duration ≤ 2	(6) Marital Duration > 2
Period	−0.72***	−0.72***	−0.75***	−.67***	−3.06**	−0.08
Log Wife’s Age	27.61**	27.50**	21.99*	13.56	29.49	−1.62
Log Husband’s Age	−6.43	−6.38	1.87	7.87	57.89	−5.23
Log Marital Duration	−1.50***	−1.51***	−1.61***	−1.56***	−1.51*	1.29
Wife Pregnant	−3.71***	−3.70***	−3.71***	−3.74***	−2.88***	−3.95*
Child under 6	−0.56**	−0.55**	−0.73***	−0.68***	−2.91***	−0.55**
Wife Employed	0.37	0.38	0.17	0.23	0.86	0.02
Husband Employed	−1.28**	−1.27**	−1.29**	−1.10**	−4.11***	−0.38
R^2	.0475	.0474	.0568	.0612	.2172	.0411
N	2062	2063	2059	2055	243	1812

* $p < .10$.
** $p < .05$.
*** $p < .01$.

Example 4. Large Datasets: Marital Coital Frequency (4)



Example 4. Large Datasets: Marital Coital Frequency (5)

- Jasso, Guilermina (1986) ‘Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality,’ *American Sociological Review*, 51:738-42.
  - Claims that Kahn and Udry’s analysis generates a new problem of sample truncation bias
    - The outcome variable has been confined to a specified segment of its range
    - She argues that we should not remove data just because they don’t conform to our beliefs
  - She doesn’t believe that the 88’s are miscodes, claiming that 2 of the complete n=5981 were coded 98, so 88 is possible
  - She claims that having sex 88 times a month - which is only 22 times a week (or about 3 times a day) is not unrealistic :
    - There are large differences in coital frequencies, especially due to cultural/regional difference

Types of Unusual Observations (1)

1. Regression Outliers

- An observation that is unconditionally unusual in either its Y or X value is called a univariate outlier, but it is not necessarily a regression outlier
- A regression outlier is an observation that has an unusual value of the outcome variable Y, conditional on its value of the explanatory variable X
  - In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
- Regression outliers often have large residuals but do not necessarily affect the regression slope coefficient
- Also sometimes referred to as vertical outliers

Types of Unusual Observations (2)

2. Cases with Leverage

- An observation that has an unusual X value - i.e., it is far from the mean of X - has leverage on the regression line
  - The further the outlier sits from the mean of X (either in a positive or negative direction), the more leverage it has
- High leverage does not necessarily mean that it influences the regression coefficients
  - It is possible to have a high leverage and yet follow straight in line with the pattern of the rest of the data. Such cases are sometimes called "good" leverage points because they help the precision of the estimates. Remember, $V(B) = \sigma_\varepsilon^2(\mathbf{X'X})^{-1}$, so outliers could increase the variance of X.

Types of Unusual Observations (3)

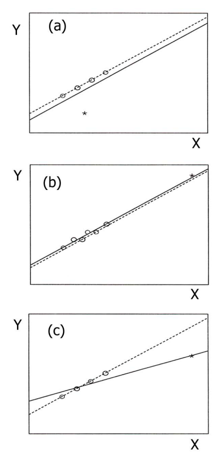
3. Influential Observations

- An observation with high leverage that is also a regression outlier will strongly influence the regression line
  - In other words, it must have an unusual X-value with an unusual Y-value given its X-value
- In such cases both the intercept and slope are affected, as the line chases the observation

Discrepancy $\times$ Leverage = Influence

Types of Unusual Observations (4)

- Figure (a): Outlier without influence. Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range
- Figure (b) High leverage because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has no influence
- Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value) results in strong influence. When this case is deleted both the slope and intercept change dramatically.



Adapted from Figure 11.1 (Fox, 1997)

Assessing Leverage: Hat Values (1)

- Most common measure of leverage is the $hat-value$, h_i
- The name $hat-values$ results from their calculation based on the fitted values ($\hat{Y}$):

$$\begin{aligned}\hat{Y}_j &= h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{nj}Y_n \\ &= \sum_{i=1}^{n} h_{ij}Y_i\end{aligned}$$

- Recall that the *Hat Matrix*, $\boldsymbol{H}$, projects the Y's onto their predicted values:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{Xb} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{Hy} \\ \underset{(n\times n)}{\boldsymbol{H}} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\end{aligned}$$

Assessing Leverage: Hat Values (2)

- If h_{ij} is large, the i^{th} observation has a substantial impact on the j^{th} fitted value
- Since $\boldsymbol{H}$ is symmetric and idempotent, the diagonal entries represent both the i^{th} row and the i^{th} column:

$$\begin{aligned}h_i &= \boldsymbol{h}_i'\boldsymbol{h}_i \\ &= \sum_{i=1}^{n} h_{ij}^2\end{aligned}$$

- This means that $h_i = h_{ii}$
- As a result, the hat value h_i measures the *potential leverage of Y_i on all the fitted values*

Properties of Hat Values

- The average hat value is: $\bar{h} = \frac{k+1}{n}$
- The hat values are bound between $\frac{1}{n}$ and 1
- In simple regression hat values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- In multiple regression, h_i measures the distance from the centroid point of all of the X's (point of means)
- Commonly used Cut-off:
  - Hat values exceeding about twice the average hat-value should be considered noteworthy
  - With large sample sizes, however, this cut-off is unlikely to identify any observations regardless of whether they deserve attention

Hat Values in Multiple Regression

- The diagram to the right shows elliptical contours of hat values for two explanatory variables
- As the contours suggest, hat values in multiple regression take into consideration the correlational and variational structure of the X's
- As a result, outliers in multi-dimensional X-space are high leverage observations - i.e., the outcome variable values are irrelevant in calculating h_i

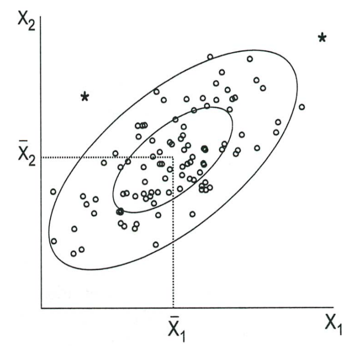


Figure 11.3 from Fox (1997)

Leverage and Hat Values: Inequality Data Revisited (1)

- We start by fitting the model to the complete dataset
- Recall that, looking at the scatterplot of Gini and attitudes, we identified two possible outliers (Czech Republic and Slovakia)
- With these included in the model there was no apparent effect of Gini on attitudes:

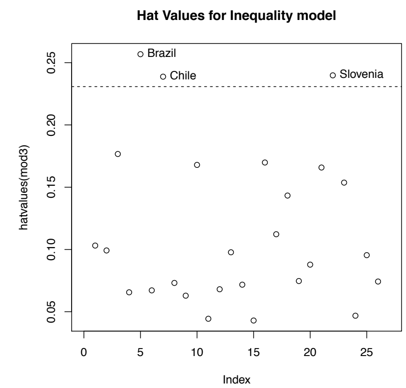
(Intercept)	2.83
	(12.78)
gini	0.07
	(0.28)
gdp	17.52*
	(7.99)
N	26
R^2	0.18
adj. R^2	0.10
Resid. sd	13.80

Standard errors in parentheses

* indicates significance at $p < 0.05$

Leverage and Hat Values: Inequality data Revisited (2)

- Several countries have large hat values, suggesting that they have unusual X values
- Notice that there are several that have much higher hat values than the Czech Republic and Slovakia
- These cases have high leverage, but not necessarily high influence



R-Script for Hat Values Plot

```
> plot(hatvalues(mod3), xlim=c(0,27),
+     main="Hat Values for Inequality model")
> cutoff <-2*3/nrow(weakliem2)
> bighat <- hatvalues(mod3) > cutoff
> abline(h=cutoff, lty=2)
> tx <- which(bighat)
> text((1:length(bighat))[tx], hatvalues(mod3)[tx],
+     rownames(weakliem2)[tx], pos=4)
```


Formal Tests for Outliers: Standardized Residuals

- Unusual observations typically have large residuals but not necessarily so - high leverage observations can have small residuals because they pull the line towards them:

$$V(E_i) = \sigma_\varepsilon^2(1 - h_i)$$

- Standardized residuals provide one possible, though unsatisfactory, way of detecting outliers:

$$E_i' = \frac{E_i}{S_E\sqrt{1 - h_i}}$$

- The numerator and denominator are not independent and thus E_i' does not follow a t-distribution: If $| E_i |$ is large, the standard error is also large:

$$S_E = \sqrt{\frac{\sum E_i^2}{n - k - 1}}$$

Studentized Residuals (1)

- If we refit the model deleting the i^{th} observation we obtain an estimate of the standard deviation of the residuals $S_{E(-1)}$ (standard error of the regression) that is based on the $n-1$ observations
- We then calculate the studentized residuals E_i^*'s, which have an independent numerator and denominator:

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$$

  Studentized residuals follow a t-distribution with $n-k-2$ degrees of freedom
- We might employ this method when we have several cases that might be outliers
- Observations that have a studentized residual outside the ± 2 range are considered statistically significant at the 95% level

Studentized Residuals (2)

- An alternative, but equivalent, method of calculating studentized residuals is the so-called *'mean-shift' outlier model*:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \gamma D_i + \varepsilon_i$$

  where D is a dummy regressor coded 1 for observation i and 0 otherwise
- We test the null hypothesis that the outlier i does not differ from the rest of the observations, $H_0 : \gamma = 0$, by calculating the t-test:

$$t_0 = \frac{\tilde{\gamma}}{\widehat{SE}(\tilde{\gamma})}$$

- The test statistic is the studentized residual E_i^* and is distributed as t_{n-k-2}
- This method is most suitable when, after looking at the data, we have determined that a particular case might be an outlier

Studentized Residuals (3): Bonferroni adjustment

- Since we are selecting the furthest outlier, it is not legitimate to use a simple t-test
  - We would expect that 5% of the studentized residuals would be beyond $t_{.025} \pm 2$ by chance alone
- To remedy this we can make a Bonferroni adjustment to the p-value.
  - The Bonferroni p-value for the largest outlier is: $p = 2np'$ where p' is the unadjusted p-value from a t-test with $n-k-2$ degrees of freedom
- The `outlierTest` function in the `car` package for **R** gives Bonferroni p-value for the largest absolute studentized residual

Studentized Residuals (4): An Example of the Outlier Test

- The Bonferroni-adjusted outlier test in `car` tests the largest absolute studentized residual.
- Recalling our inequality model:

```
> mod3 <- lm(secpay~gini + gdp, data=weakliem2)
> outlierTest(mod3)

         rstudent unadjusted p-value Bonferonni p
Slovakia 4.317504         0.00027781     0.007223
```

- It is now quite clear that Slovakia (observation 26) is an outlier, but as of yet we have not assessed whether it influences the regression line - the test statistically significant

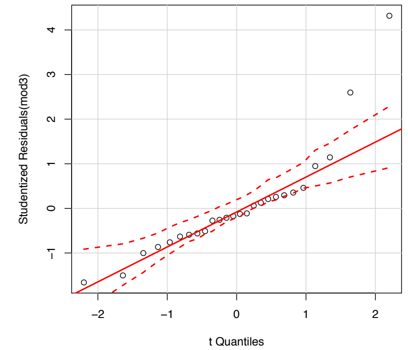
Quantile Comparison Plots (1)

- We can use a quantile comparison plots to compare the distribution of a single variable to the t-distribution, assessing whether the distribution of the variable showed a departure from normality
- Using the same technique, we can compare the distribution of the studentized residuals from our regression model to the t-distribution
- Observations that stray outside of the 95% confidence envelope are statistically significant outliers

```
> qqPlot(mod3, simulate=T, labels=F)
```


Quantile Comparison Plot (2): Inequality Data

- Here we can again see that two cases appear to be outliers: these are the Czech Republic and Slovakia



Influential Observations: DFBeta (1)

- Recall that an influential observation is one that combines discrepancy with leverage
- The most direct approach to assessing influence is to assess how the regression coefficients change if outliers are omitted from the model
- We can use D_{ij} (often termed DFBeta$_{ij}$) to do so:

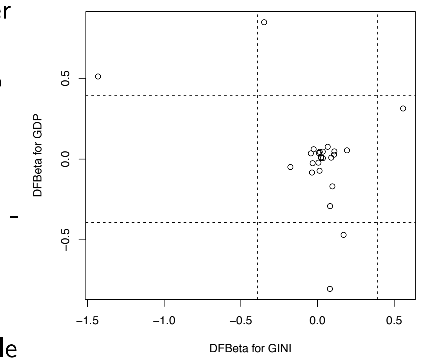
$$D_{ij} = B_j - B_{j(-i)} \quad \forall \quad i = 1, \ldots, n; \quad j = 1, \ldots, k$$

The B_j are the coefficients for all the data and the $B_{j(-i)}$ are the coefficients for the same model with the i^{th} observation removed.

- A standard cut-off for an influential observation is: $D_{ij} \geq \frac{2}{\sqrt{n}}$.

Influential Observations: DFBeta (2)

- We see here Slovakia makes the gdp coefficient larger and the coefficient for gini smaller
- The Czech Republic also makes the coefficient for gdp larger
- A problem with DFBetas is that each observation has several measures of influence - one for each coefficient $n(k + 1)$ different measures
- Cook's D overcomes the problem by presenting a single summary measure for each observation



Identifying DFBetas

```
> cutoff <- 2/sqrt(26)
> big <- with(dfb, which(abs(gini) > cutoff |
+  abs(gdp) > cutoff))
> dfb[big, ]
              (Intercept)        gini        gdp
Chile         -0.51676696  0.55836187  0.3132308
CzechRepublic  0.06163614 -0.34805553  0.8471765
Slovakia       1.14014221 -1.43107966  0.5112908
Slovenia       0.17438196  0.08084083 -0.8037418
Taiwan        -0.01827400  0.17003877 -0.4692173
```


Influential Observations: Cook's D (1)

- Cook's D measures the 'distance' between B_j and $B_{j(-i)}$ by calculating an F-test for the hypothesis that $b_j = B_{j(-i)}$, for $j = 0, 1, \ldots, k$. An F-test is calculated for each observation as follows:

$$D_i = \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

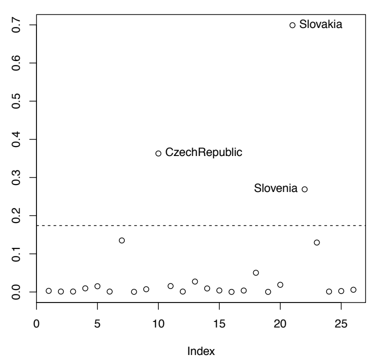
where h_i is the hat value for each observation and E_i' is the standardized residual
- The first fraction measures discrepancy; the second fraction measures leverage
- There is no significance test for D_i (i.e., the F-test here measures only distance) but a commonly used cut-off is:

$$D_i > \frac{4}{n-k-1}$$

- The cut-off is useful, but there is no substitution for examining relative discrepancies in plots of Cook's D versus cases, or of E_i' against h_i

Cook's D: An Example

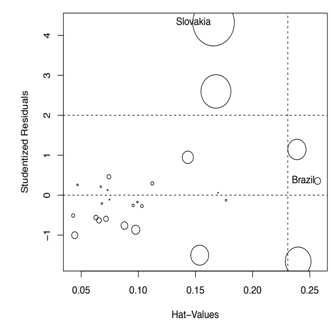
- We can see from this plot of Cook's D against the case numbers, that Slovakia has an unusually high level of influence on the regression surface
- The Czech Republic and Slovenia also stand out




```
> mod3.cook <- cooks.distance(mod3)
> plot(cooks.distance(mod3))
> cutoff <- with(mod3, 4/df.residual)
> abline(h=cutoff, lty=2)
> text(which(mod3.cook > cutoff), mod3.cook[which(mod3.cook > cutoff)],
+     names(mod3.cook[which(mod3.cook > cutoff)]), pos=c(4,4,2))
```


Influence Plot or "Bubble Plot"

- Displays studentized residuals, hat-values and Cook's D on a single plot
- The horizontal axis represents the hat-values; the vertical axis represents the studentized residuals; circles for each observation represent the relative size of the Cook's D
  - The radius is proportional to the square root of Cook's D, and thus the areas are proportional to the Cook's D



Joint Influence (1)

- Subsets of cases can jointly influence a regression line, or can offset each other's influence
- The heavy solid is the regression with all cases included; The broken line is the regression with the asterisk deleted;The light solid line is for the regression with both the plus and asterisk deleted
- Depending on where the jointly influential cases lie, they can have different effects on the regression line.
- (a) and (b) are jointly influential because they change the regression line when included together.
- The observations in (c) offset each other and thus have little effect on the regression line

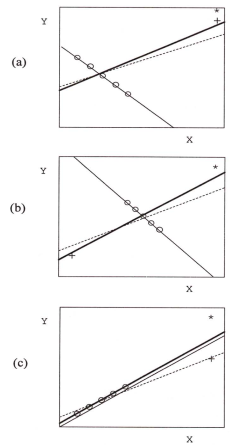


Figure 11.4 from Fox (1997)

Joint Influence (2)

- Cook's D can help us determine joint influence if there are relatively few influential cases.
  - That is, we can delete cases sequentially, updating the model each time and exploring the Cook's D's again
  - This approach is impractical if there are potentially a large number of subsets to explore, however
- Added-variable plots (also called partial-regression plots) provide a more useful method of assessing joint influence
  - These plots essentially show the partial relationships between Y and each X
  - We make one plot for each X

Added Variable Plots [or Partial Regression Plots] (1)

1. Let $Y_i^{(1)}$ represent the residuals from the least-squares regression of Y on all of the X's except for X_1:

$$Y_i = A^{(1)} + B_2^{(1)} X_{i2} + \cdots + B_k^{(1)} X_{ik} + Y_i^{(1)}$$

2. Similarly, $X_i^{(1)}$ are the residuals from the regression of X_1 on all the other X's

$$X_{i1} = C^{(1)} + D_2^{(1)} X_{i2} + \cdots + D_k^{(1)} X_{ik} + X_i^{(1)}$$

3. These two equations determine the residuals $X^{(1)}$ and $Y^{(1)}$ as parts of X_1 and Y that remain when the effects of $X_2, \ldots, X_k$ are removed.

Added Variable Plots (2)

- The Residuals $Y^{(1)}$ and $X^{(1)}$ have the following properties:
  1. Slope of the regression of $Y^{(1)}$ on $X^{(1)}$ is the least-squares slope B_1 from the full multiple regression
  2. Residuals from the regression of $Y^{(1)}$ on $X^{(1)}$ are the same as the residuals from the full regression:

$$Y_i^{(1)} = B_1 X_{i1}^{(1)} + E_i$$

  3. Variation of $X^{(1)}$ is the conditional variance of X_1 holding the other X's constant. Consequently, except for the df the standard error from the partial simple regression is the same as the multiple regression SE of B_1.

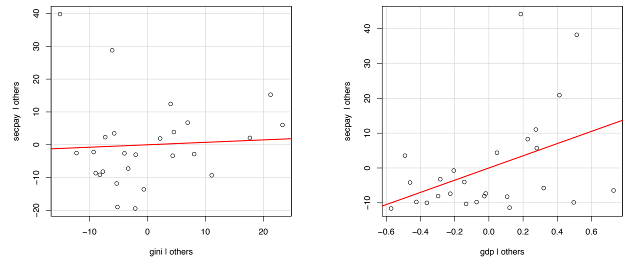
$$\widehat{SE}(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)^2}}}$$

Added Variable Plots (3): An Example

- Once again recalling the outlier model from the Inequality data
- A plot of $Y^{(1)}$ against $X^{(1)}$ allows us to examine the leverage and influence of cases on B_1
  - we make one plot for each X
- These plots also gives us an idea of the precision of our slopes $(B_1, \ldots, B_k)$

```
> avPlots(mod3, "gini")
> avPlots(mod3, "gdp")
```


Added Variable Plots (4): Example Continued



- We see here that the Czech Republic and Slovakia have unusually high Y values given their X's
- Because they are on the extreme of the X-range as well, they are most likely influencing both slopes

Unusual Observations and their impact on Standard Errors

- Depending on their location, unusual observations can either increase or decrease standard errors
- Recall that the standard error for a slope is as follows:

$$\widehat{SE}(B) = \frac{S_E}{\sqrt{\sum(X_i - \overline{X})^2}}$$

- An observation with high leverage (i.e., an X-value far from the mean of X) increases the size of the denominator, and thus decreases the standard error
- A regression outlier (i.e., a point with a large residual) that does not have leverage (i.e., it does not have an unusual X-value) does not change the slope coefficients but will increase the standard error

Unusual Cases: Solutions

- Unusual observations may reflect miscoding, in which case the observations can be rectified or deleted entirely
- Outliers are sometimes of substantive interest:
  - If only a few cases, we may decide to deal separately with them
  - Several outliers may reflect model misspecification - i.e., an important explanatory variable that accounts for the subset of the data that are outliers has been neglected
- Unless there are strong reasons to remove outliers we may decide to keep them in the analysis and use alternative models to OLS, for example robust regression, which down weight outlying data.
  - Often these models give similar results to an OLS model that omits the influential cases, because they assign very low weight to highly influential cases

Summary (1)

- Small samples are especially vulnerable to outliers - there are fewer cases to counter the outlier
- Large samples can also be affected, however, as shown by the "marital coital frequency" example
- Even if you have many cases, and your variables have limited ranges, miscodes that could influence the regression model are still possible
- Unusual cases are only influential when they are both unusual in terms of their Y value given their X (outlier), and when they have an unusual X-value (leverage):

Influence = Leverage × Discrepency

Summary (2)

- We can test for outliers using studentized residuals and quantile - comparison plots
- Leverage is assessed by exploring the hat-values
- Influence is assessed using DFBetas and, preferably Cook's D's
- Influence Plots (or bubble plots) are useful because they display the studentized residuals, hat-values and Cook's distances all on the same plot
- Joint influence is best assessed using Added Variable Plots (or partial regression plots)