Regression III: Advanced Methods

William G. Jacoby

Michigan State University

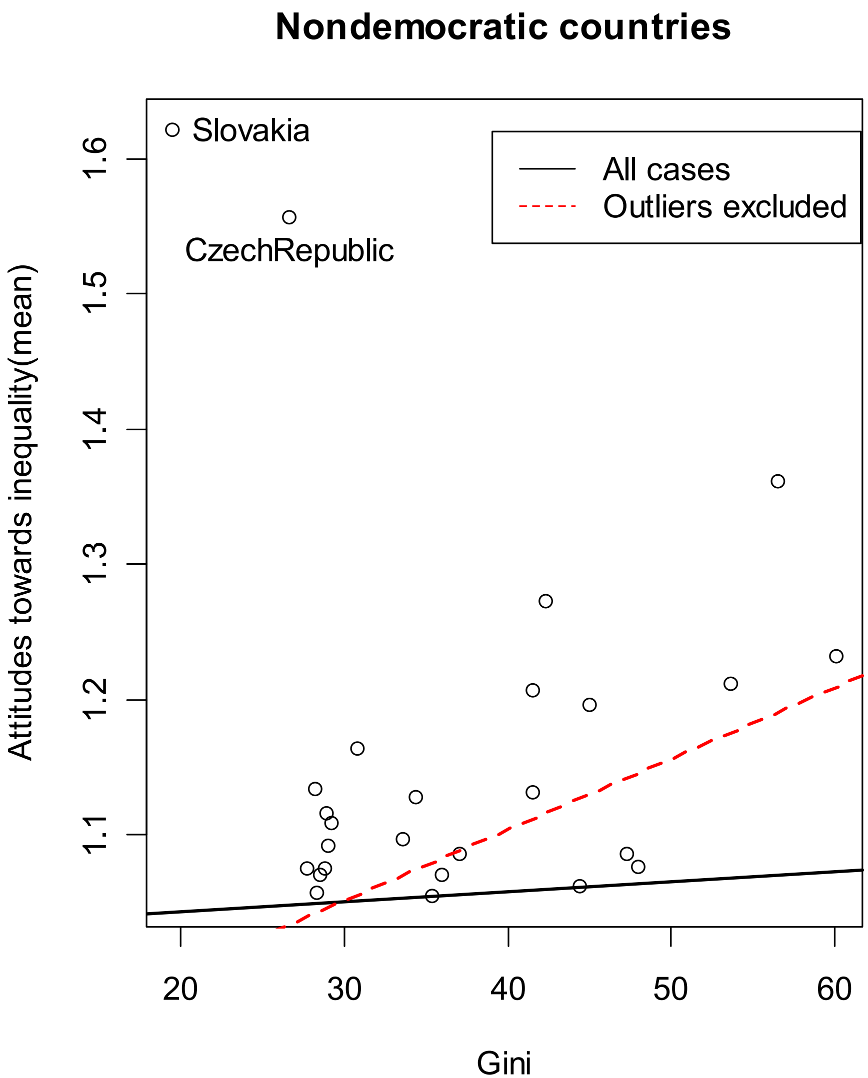
http://polisci.msu.edu/jacoby/icpsr/regress3

Outlying Observations: Why pay attention?

- Can cause us to misinterpret patterns in plots
  - Outliers can affect visual resolution of remaining data in plots (forces observations into "clusters")
  - Temporary removal of outliers, and/or transformations can "spread out" clustered observations and bring in the outliers (if not removed)
- More importantly, separated points can have a strong *influence* on statistical models—deleting 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

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

- 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



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

Model including all cases

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0283	0.1278	8.05	0.0000
gini	0.0007	0.0028	0.27	0.7908
gdp	0.0000	0.0000	2.19	0.0387

Residual standard error: 0.138

Multiple R-Squared: 0.175

Model excluding Czech Rep. & Slovakia

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8931	0.0578	15.45	0.0000
gini	0.0053	0.0013	4.07	0.0005
gdp	0.0000	0.0000	1.69	0.1050

Residual standard error: 0.0602

Multiple R-Squared: 0.462

R script for Ex. 1

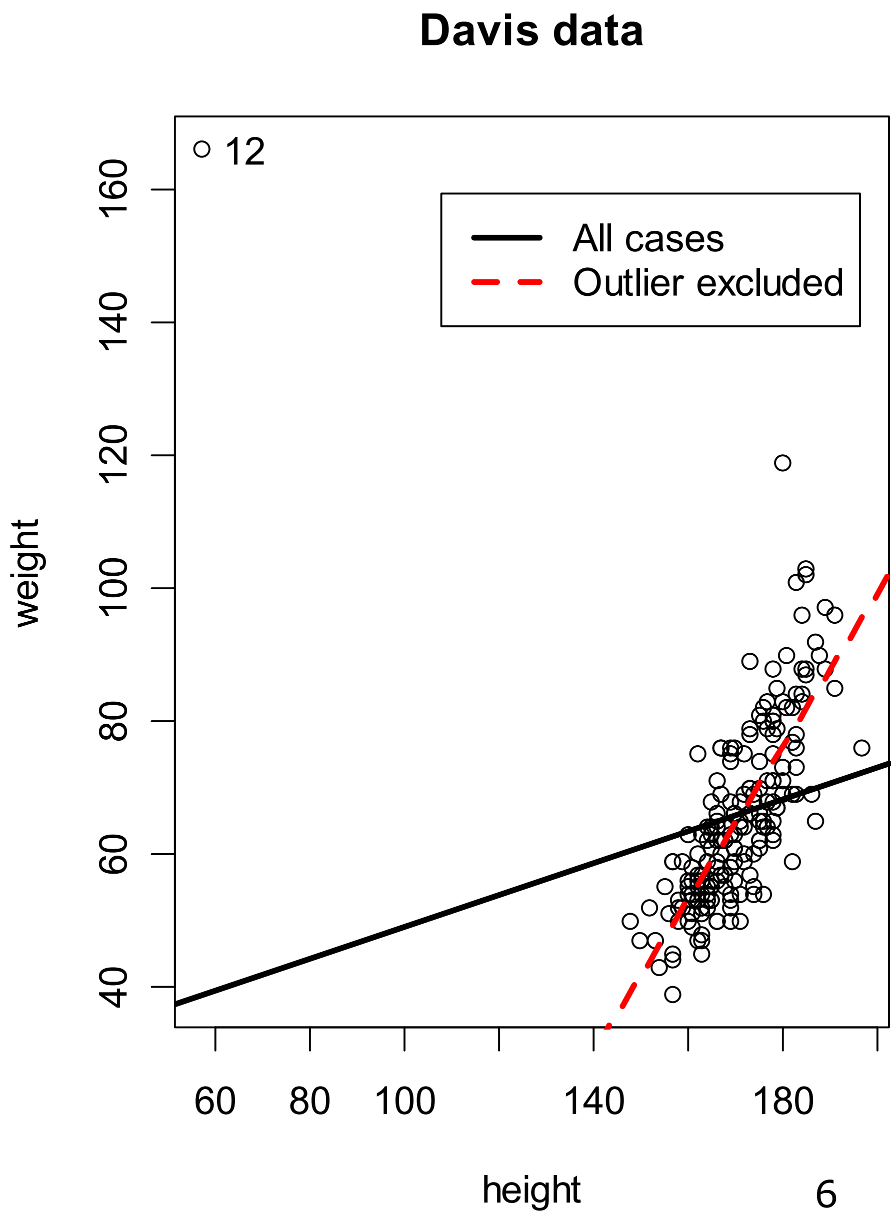
```
Weakliem2<-read.table('C:/data/Weakliem2.txt', header=T)
attach(Weakliem2)

plot(gini, secpay, main='Nondemocratic countries', xlab='Gini',
ylab='Attitudes towards inequality(mean)')
Weakliem.model1<-lm(secpay~gini+gdp)
abline(Weakliem.model1, lwd=2, lty=1, col=1)
identify(gini,secpay, row.names(Weakliem2))
#"identify" returns cases 7, 26 as outliers
Weakliem.model2<-update(Weakliem.model1, subset=-c(7,26))
abline(Weakliem.model2, lwd=2, lty=2, col=2)
legend(locator(1), lty=1:2, col=1:2,
       legend=c('All cases', 'Outliers excluded'))

library(xtable)#Prints LaTeX code for the output table
print(xtable(Weakliem.model1))
print(xtable(Weakliem.model2))
```

Ex 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



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

Model including all cases

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	25.2662	14.9504	1.69	0.0926
height	0.2384	0.0877	2.72	0.0072

Residual standard error: 14.86

Multiple R-Squared: 0.0359

Model excluding observation #12

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−130.7470	11.5627	−11.31	0.0000
height	1.1492	0.0677	16.98	0.0000

Residual standard error: 8.523

Multiple R-Squared: 0.594

R script for Ex. 2

```
>library(car)
>data(Davis)
>attach(Davis)
>davis.model.1<-lm(repwt~weight)

>plot(height, weight, main="Davis data")
>Model1<-lm(weight~height)
>identify(height, weight, row.names(Davis))
      #observation 12 returned as outlier
>abline(Model1, lty=1, col=1, lwd=3)
>Model2<-update(Model1, subset=-12)
>abline(Model2, lty=2, col=2, lwd=3)
>legend(locator(1), lty=1:2, col=1:2, lwd=3,
   legend=c('All cases', 'Outlier excluded'))
```

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 dependent variable Y, conditional on its value of the independent variable X***
  - In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
- A regression outlier will have a large residual but not necessarily affect the regression slope coefficient

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 (i.e., the potential to influence) the regression line
- The further away from from the mean of X (either in a positive or negative direction), the more leverage an observation has on the regression fit
- 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

Types of Unusual Observations (3)

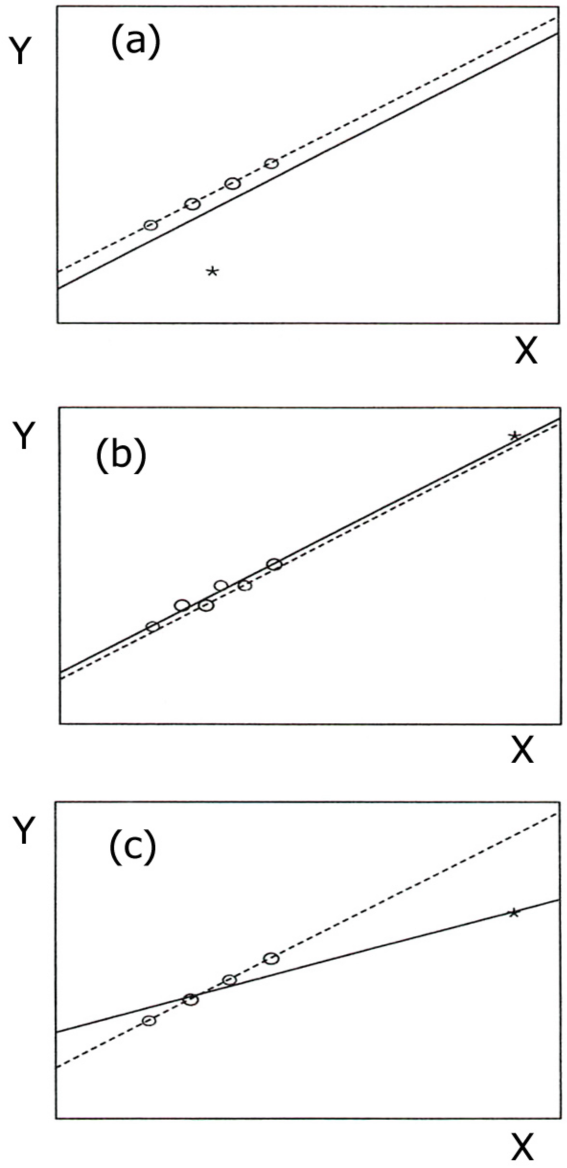
3. Influential Observations

- Only when an observation has ***high leverage*** and is an ***outlier in terms of Y-value*** will it 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***

Influence=Leverage X Discrepancy

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 (Y-hat):

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

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

$$\begin{aligned}\hat{\mathrm{y}} &= \boldsymbol{X}\mathbf{b} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{y} \\ &= \mathrm{H}\mathrm{y}\end{aligned}$$

$$\underset{(n\times n)}{\mathrm{H}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

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 $\mathbf{H}$ is symmetric and idempotent, the diagonal entries represent both the i_{th} row and the i_{th} column:

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

- This implies, then, 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: $\bar{h} = (k+1)/n$
- Hat values are bounded between $1/n$ and 1
- In simple regression hat values measure distance from the mean of X:

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

- In multiple regression, h_i measures the distance from the centroid point of X*'s* (point of means)
- ***<u>Rule of Thumb</u>:***
  - ***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 independent 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 ***dependent variable values are irrelevant in calculating*** h_i

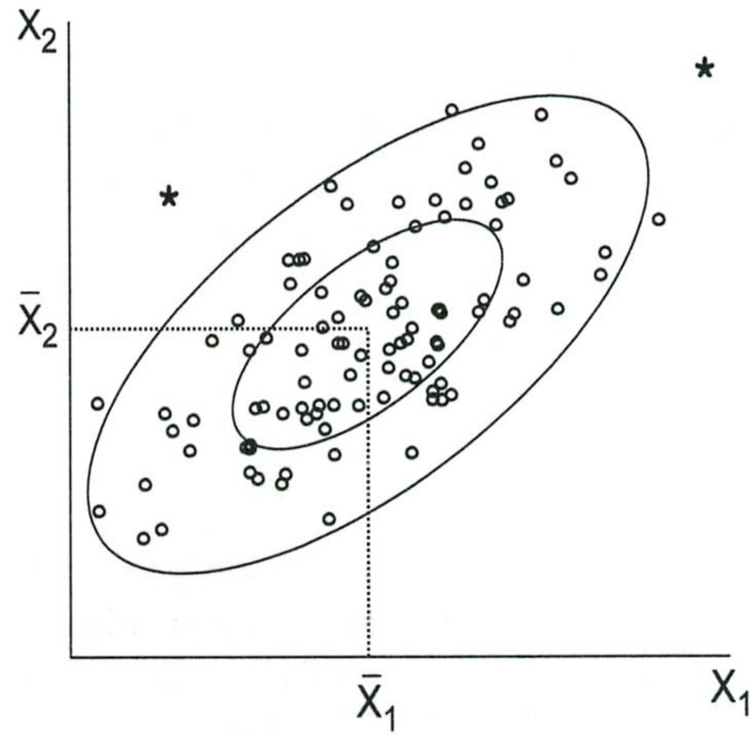


Figure 11.3 from Fox (1997)

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:

Model including all cases

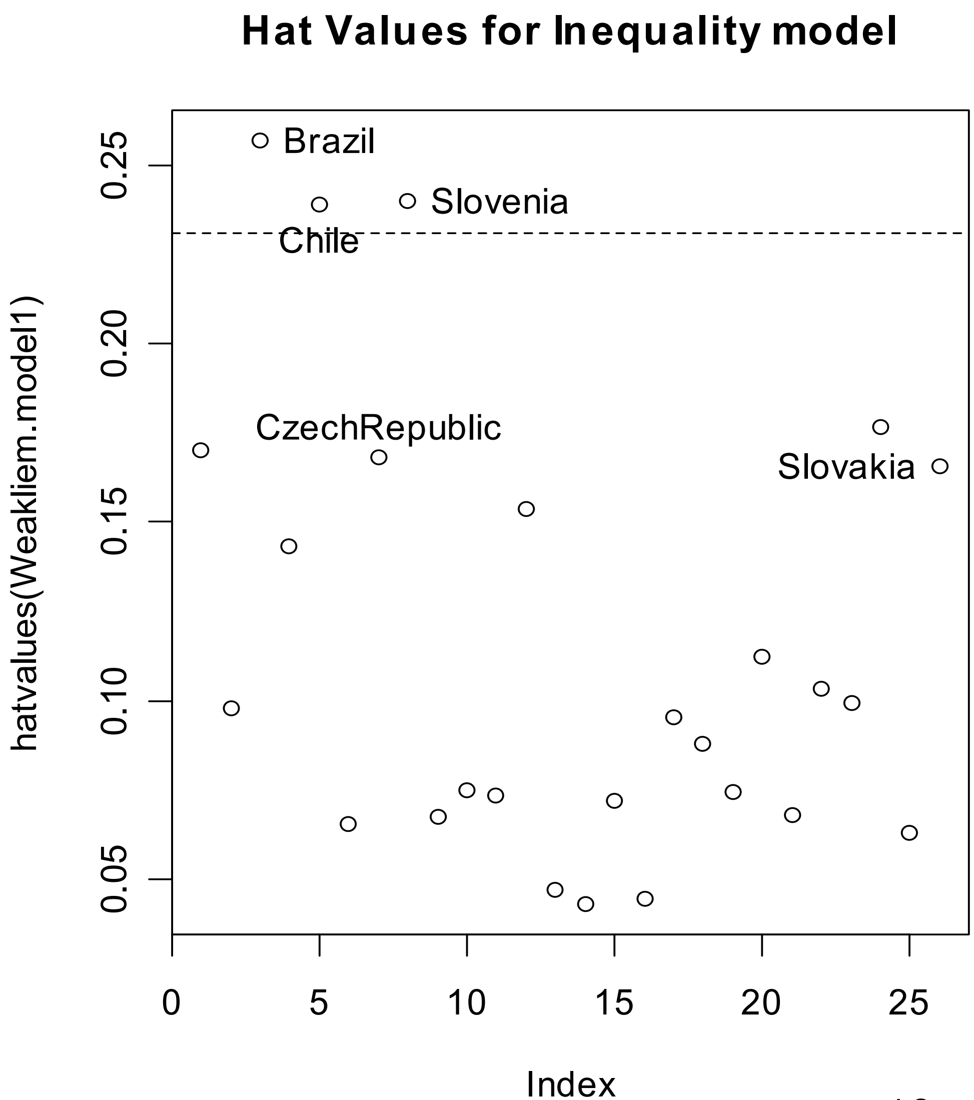
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0283	0.1278	8.05	0.0000
gini	0.0007	0.0028	0.27	0.7908
gdp	0.0000	0.0000	2.19	0.0387

R script for plot of Hat Values

```
>library(car)
>plot(hatvalues(Weakliem.model1),
    main="Hat Values for Inequality model")
>abline(h=c(2,3)*3/length(secpay), lty=2)
#"h" signifies horizontal line
#the average hat value=(k+1)/n.
#A rule of thumb is that 2*average hat value
#for large samples, and 3*average hat value
#for small samples should be examined
>identify(1:length(secpay),
   hatvalues(Weakliem.model1),
   row.names(Weakliem2))
```

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***



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:

$$S_E = \sqrt{\sum E_i^2/(n - k - 1)}$$

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% *a* level

Studentized Residuals (2)

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

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \gamma D + \varepsilon$$

  Here D is a dummy regressor coded 1 for observation *i* and 0 for all other observations.

- We test the null hypothesis that the outlier i does not differ from the rest of the observations, H_0:γ=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)
The 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
- A special t-table is needed if you do this calculation by hand, but the `outlier.test` function in the `car` package for **R** will give it to you automatically

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*:

```
> outlier.test(Weakliem.model1)
 max|rstudent| df unadjusted p Bonferroni p
      4.317504 22 0.0002778084  0.007223019

Observation: 26
> row.names(Weakliem2)[26]
[1] "Slovakia"
```

- 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

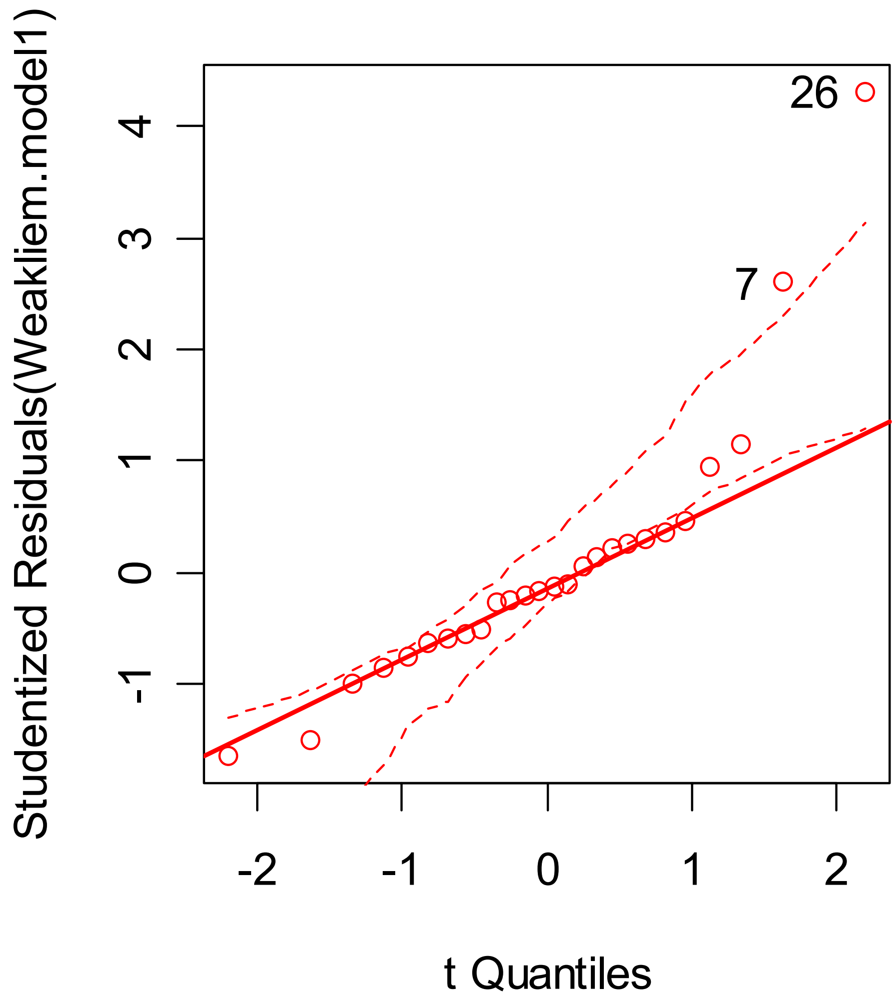
Quantile Comparison Plots (1)

- Recall that we used 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

```
>library(car)
>qq.plot(Weakliem.model1, simulate=T,
     lables=row.names(Weakliem2))
       #simulate=T specifies a bootstrap
       # 95% confidence envelope
```

Quantile-Comparison Plots (2): Example: Inequality data

- Here we can again see that two cases appear to be outliers: 7 and 26, which represent the Czech Republic and Slovakia



Influential Observations: DFBeta and DFBetas (1)

- Recall that ***an influential observation is one that combines discrepancy with leverage***
- Therefore, examine how regression coefficients change if outliers are omitted from the model
- We can use D_{ij} (often termed ***DFBeta$_{ij}$***) to do so:

$$D_{ij} = B_j - B_{j(-i)}$$

for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, k$

where the B_j are for all the data and the $B_{j(-i)}$ are with the ith observation removed

- D^*_{ij} (Dfbetas$_{ij}$) standardizes the measure, by dividing by $S_{Bj(-i)}$
- A standard cut-off for an influential observation is:

$$D^*_{ij} = 2\, n^{-.5}$$

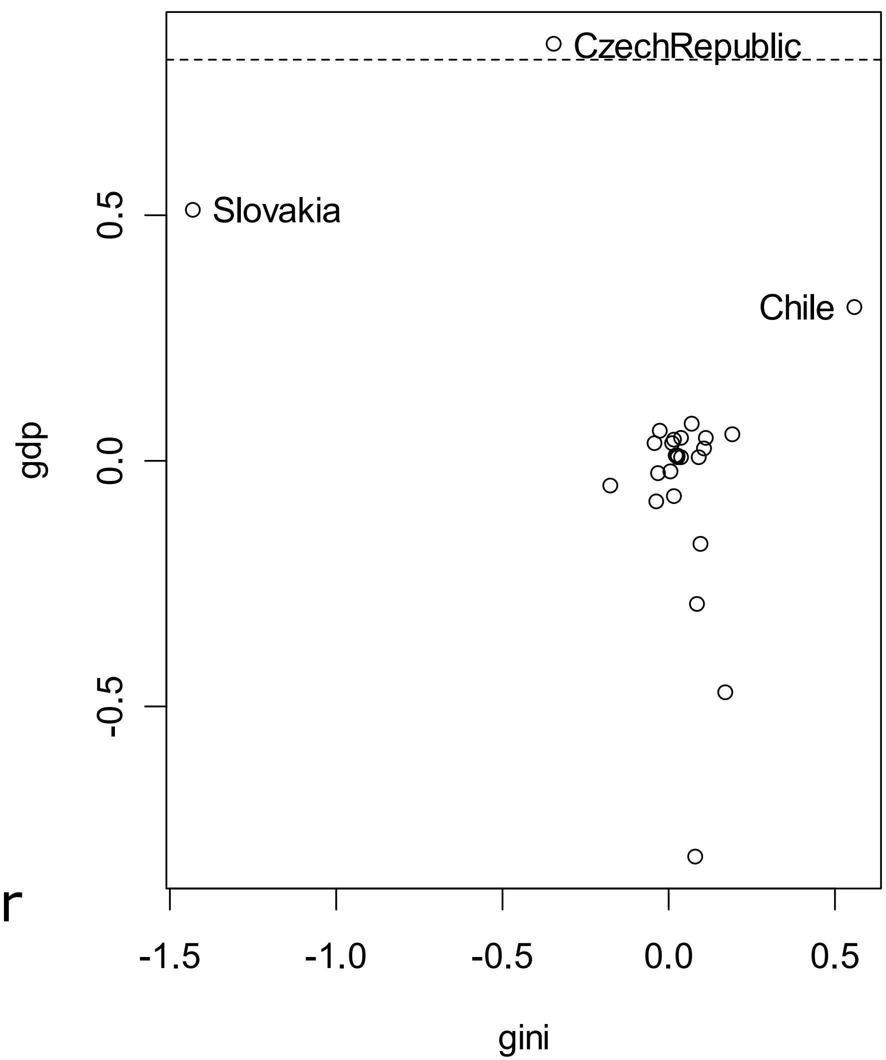
R script for DFBetas plot

```
>library(car)
>Weakliem.dfbetas<-dfbetas(Weakliem.model1)
>plot(Weakliem.dfbetas[,c(2,3)],
    main="DFBetas for the Gini and GDP coefficients")
     #c(2,3) specifies the coefficients of interest
>abline(h=2/sqrt(length(Weakliem2)), lty=2)
     #adds the rule of thumb cut-off line
>identify(Weakliem.dfbetas[,2], Weakliem.dfbetas[,3],
    row.names(Weakliem2))
```

Influential Observations: DFBetas (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

DFBetas for the Gini and GDP coefficients



Cook's Distance (Cook's D)

- Cook's D measures the 'distance' between B_j and $B_{j(-i)}$ by calculating an F-test for the hypothesis that $\beta_j = B_{j(-i)}$, for $j=0,1,\ldots,k$. An F statistic 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

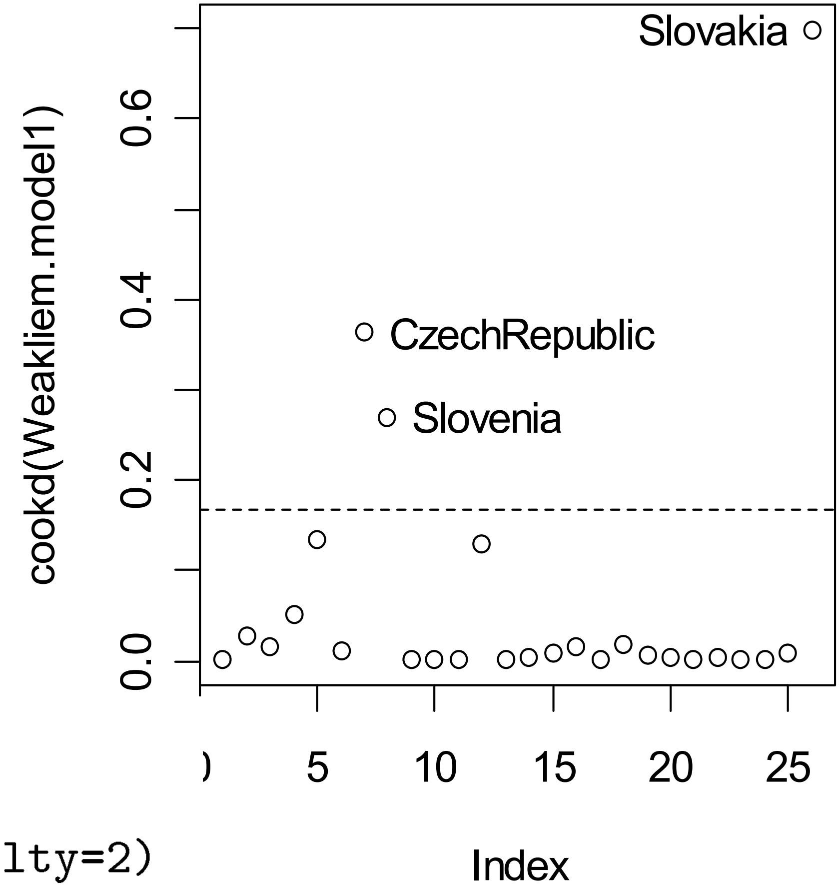
- The first fraction ***measures discrepancy***; the second fraction ***measures leverage***
- There is **no significance test** for D_i (i.e., the F value here measures only distance) but a cut-off rule of thumb is:

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

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

Cook’s D: An Example

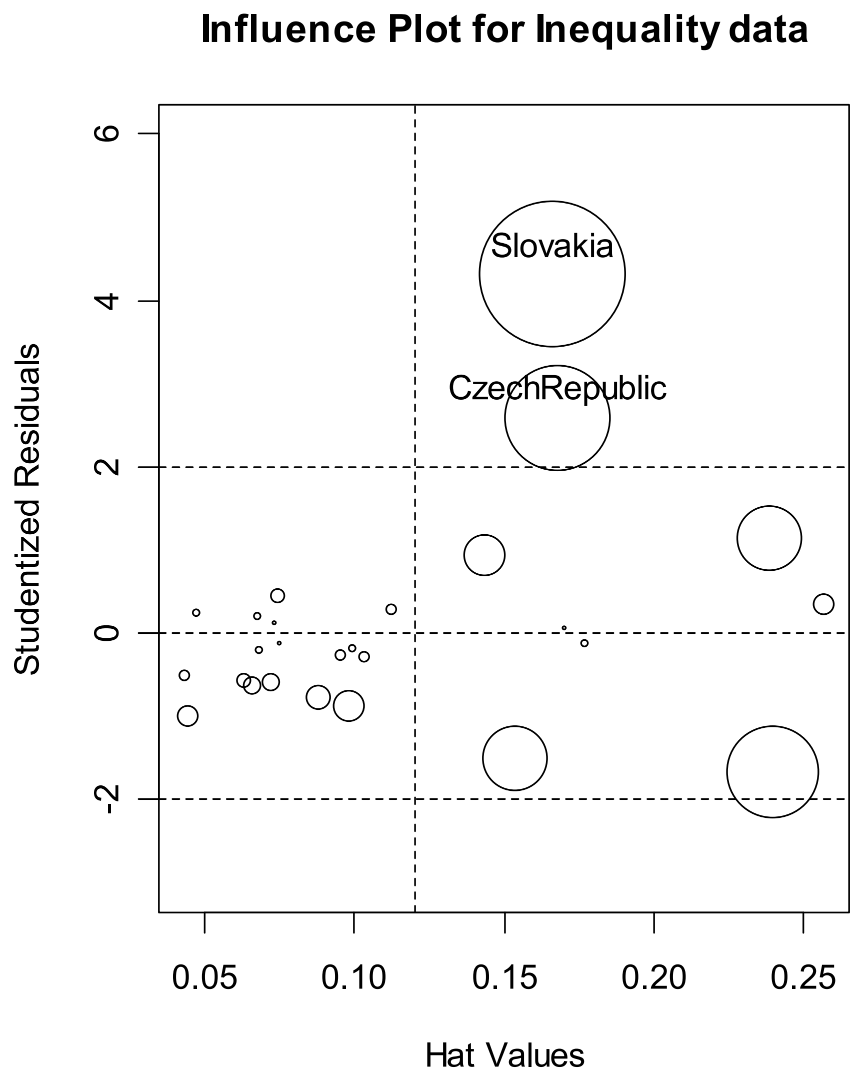
- We can see from this plot of Cook’s D against the case numbers, that Slovakia has an unusually high level of influence on the regression surface
- The Czech Republic and Slovenia also stand out




```
> library(car)
> plot(cookd(Weakliem.model1))
> abline(h=4/length(Weakliem2), lty=2)
> identify(1:26, cookd(Weakliem.model1),
      row.names(Weakliem2))
[1]  7  8 26
```

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***



R-script for the Influence Plot

```
plot(hatvalues(Weakliem.model1),
    rstudent(Weakliem.model1), ylim=c(-3,6),type='n',
    main="Influence Plot for Inequality data",
    xlab="Hat Values",
    ylab="Studentized Residuals")
cook<-sqrt(cookd(Weakliem.model1))
points(hatvalues(Weakliem.model1),
     rstudent(Weakliem.model1), cex=10*cook/max(cook))
abline(v=3/25, lty=2)#line for hatvalues
abline(h=c(-2,0,2), lty=2)
#lines for studentized residuals
identify(hatvalues(Weakliem.model1),
   rstudent(Weakliem.model1), row.names(Weakliem2))
```

Joint Influence (1)

- Subsets of cases can jointly influence a regression line, or can offset each other's influence
- Cook's D can help us determine joint influence if there are relatively few influential cases.
  - That is, we can delete cases sequentially, up-dating the model each time and exploring the Cook's Ds 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

Joint influence (2)

- The heavy solid represent 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

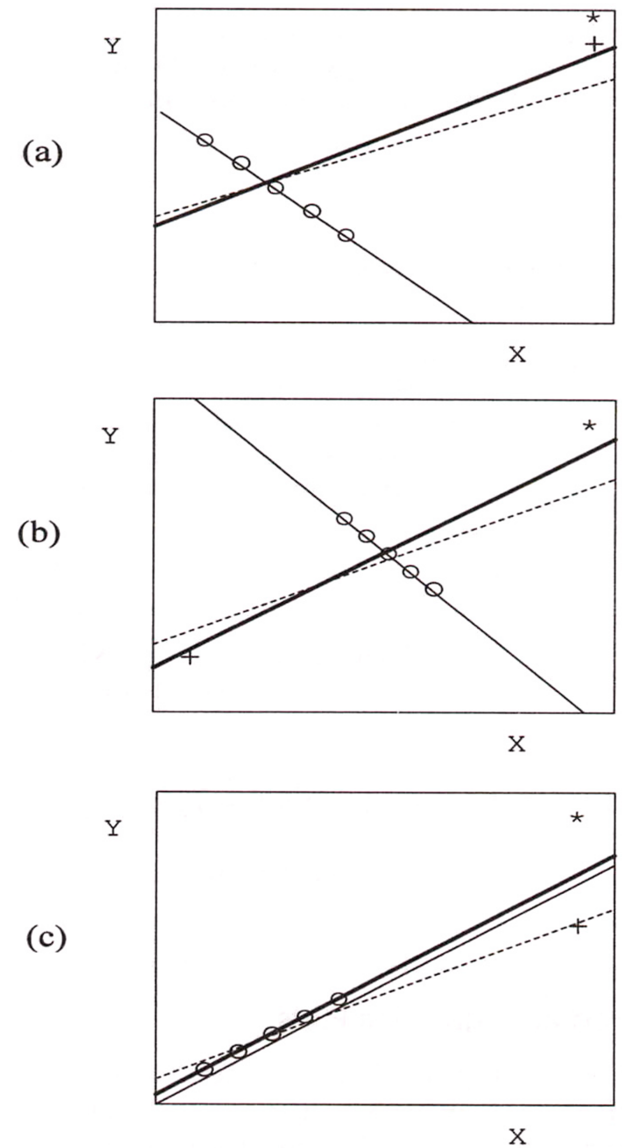


Figure 11.4 from Fox (1997)

Added-Variable Plots (1) (or partial regression plots)

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

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

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

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

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

Added-Variable Plots (2) (or partial regression plots)

- 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:

$$Y_i^{(1)} = B_1 X_1^{(1)} + E_i$$

  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.

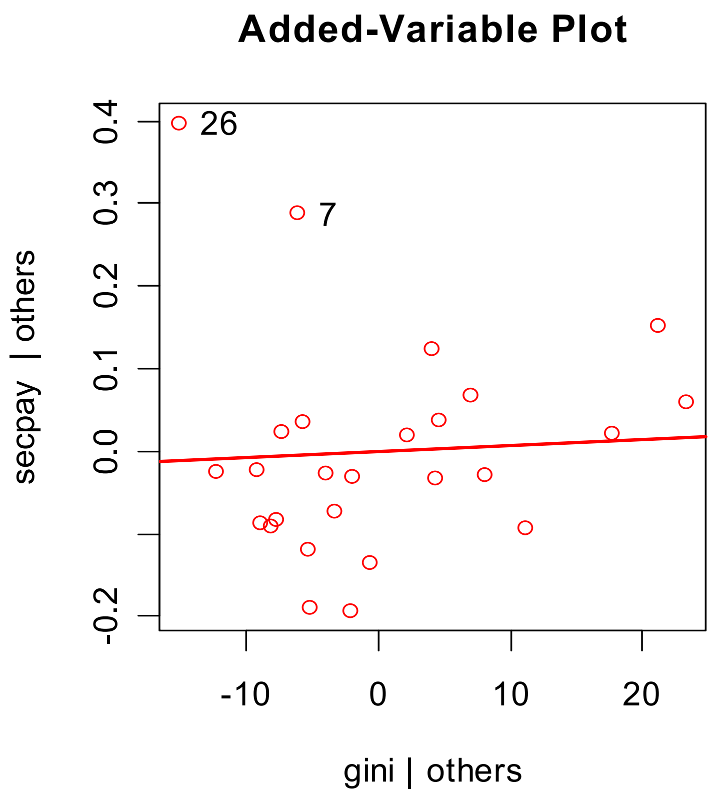
$$\widehat{SE(B_1)} = \frac{S_E}{\sqrt{\sum X_i^{(1)^2}}}$$

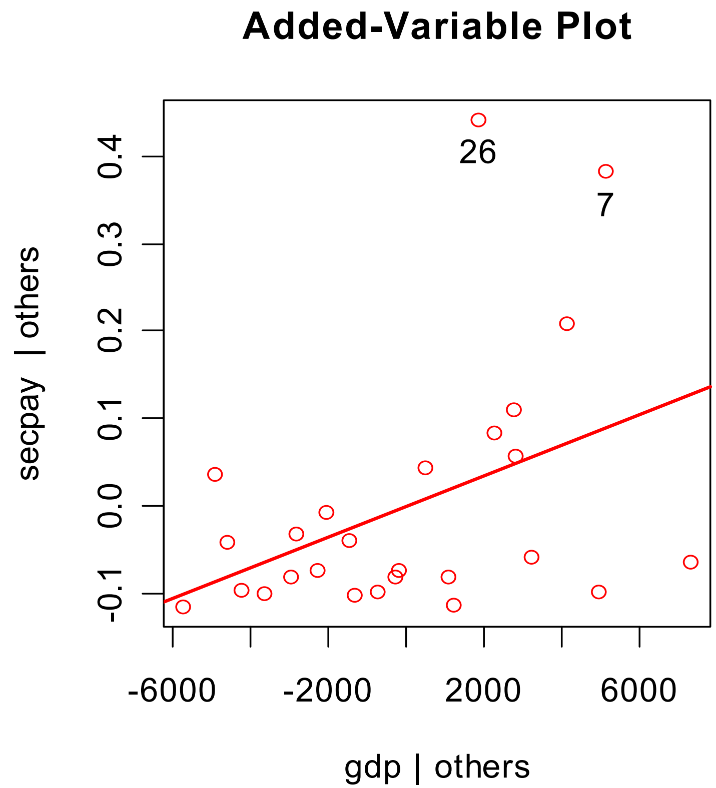
Added-Variable Plots (3)
An Example

- Once again recalling the outlier model from the Inequality data (Weakliem.model1)
- 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 ($B_1 \ldots B_k$)

```
#Added variable plots (partial regression plot)
>library(car)
>av.plots(Weakliem.model1)
#This allows you to choose the
#variables interactively
>leverage.plot(Weakliem.model1, "gini")
#This method you choose the
#variable of interest
```

Added-Variable Plots (4) Example cont'd





- We see here that cases 7 (Czech Republic) and 26 (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:

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

- 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 X 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 Ds***
- ***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)