Further topics in linear regression

Part 2

David Barron

Hilary Term 2018

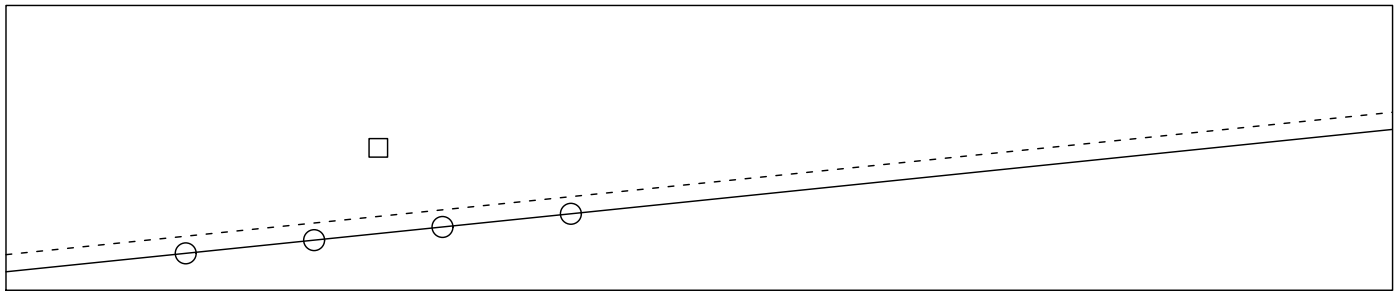
Regression diagnostics

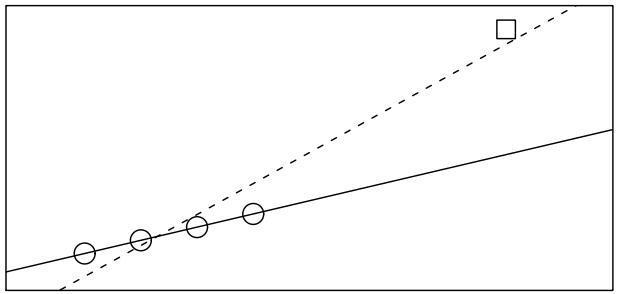
Outliers

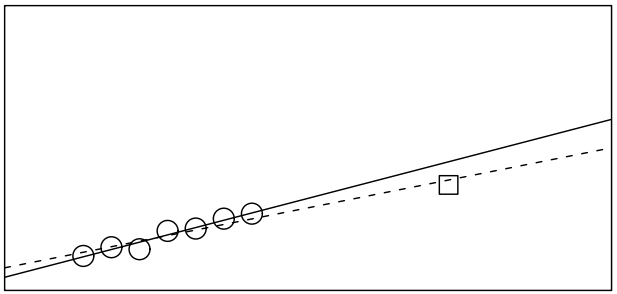
What to look for

We must identify observations with high **leverage**; that is, with an unusual x value *and* that is out of line with the other observations. In the figure, the first graph shows an outlier with low leverage because it is close to the centre of the x values. The second graph shows a high leverage outlier. The third graph doesn't really have an outlier. Although there is one unusual observation, it is in line with the other cases. Only in the second graph does deletion of the outlier have much of an impact on the regression line.

High leverage outliers







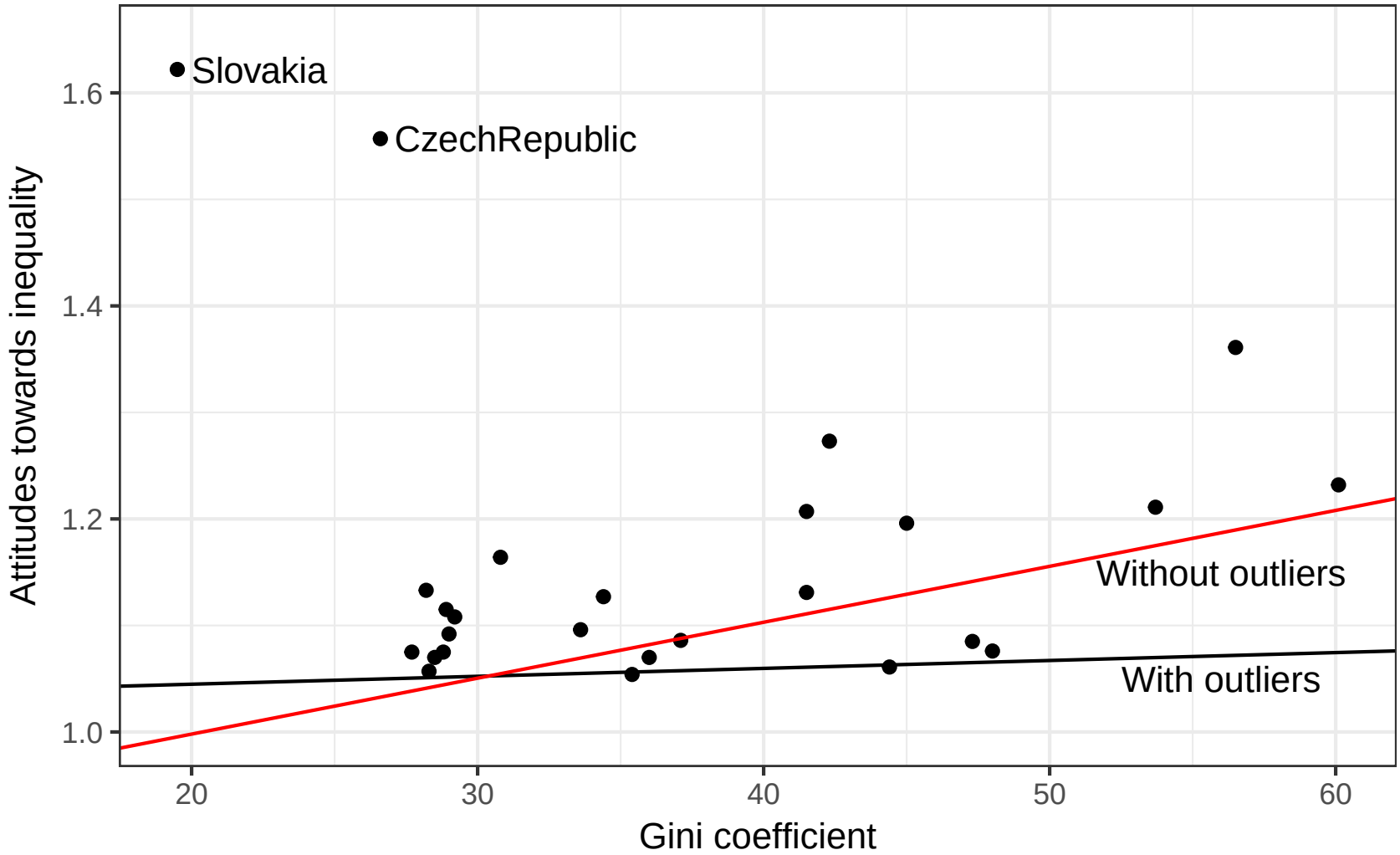
Example

Attitudes to inequality

Data from World Values Survey 1990. *secpay*: attitude to two secretaries with the same jobs getting paid different amounts if one is better at the job than the other. 1=Fair, 2=Unfair. Variable is the national average. *gini*: the gini coefficient of income inequality in the country. 0=perfect equality, 1=perfect inequality. *gdp*: GDP per capita in US dollars; *democracy*: 1=experienced democratic rule for at least 10 years. **Here we look only at non-democratic countries**.

```
Call:
lm(formula = secpay ~ gini + gdp, data = weak.nondem)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1924 -0.0789 -0.0196  0.0382  0.4093

Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 1.02826650 0.12778830    8.05 0.000000039
gini        0.00074237 0.00276544    0.27       0.791
gdp         0.00001752 0.00000799    2.19       0.039

Residual standard error: 0.138 on 23 degrees of freedom
Multiple R-squared:  0.175, Adjusted R-squared:  0.104
F-statistic: 2.45 on 2 and 23 DF,  p-value: 0.109
```

Scatterplot



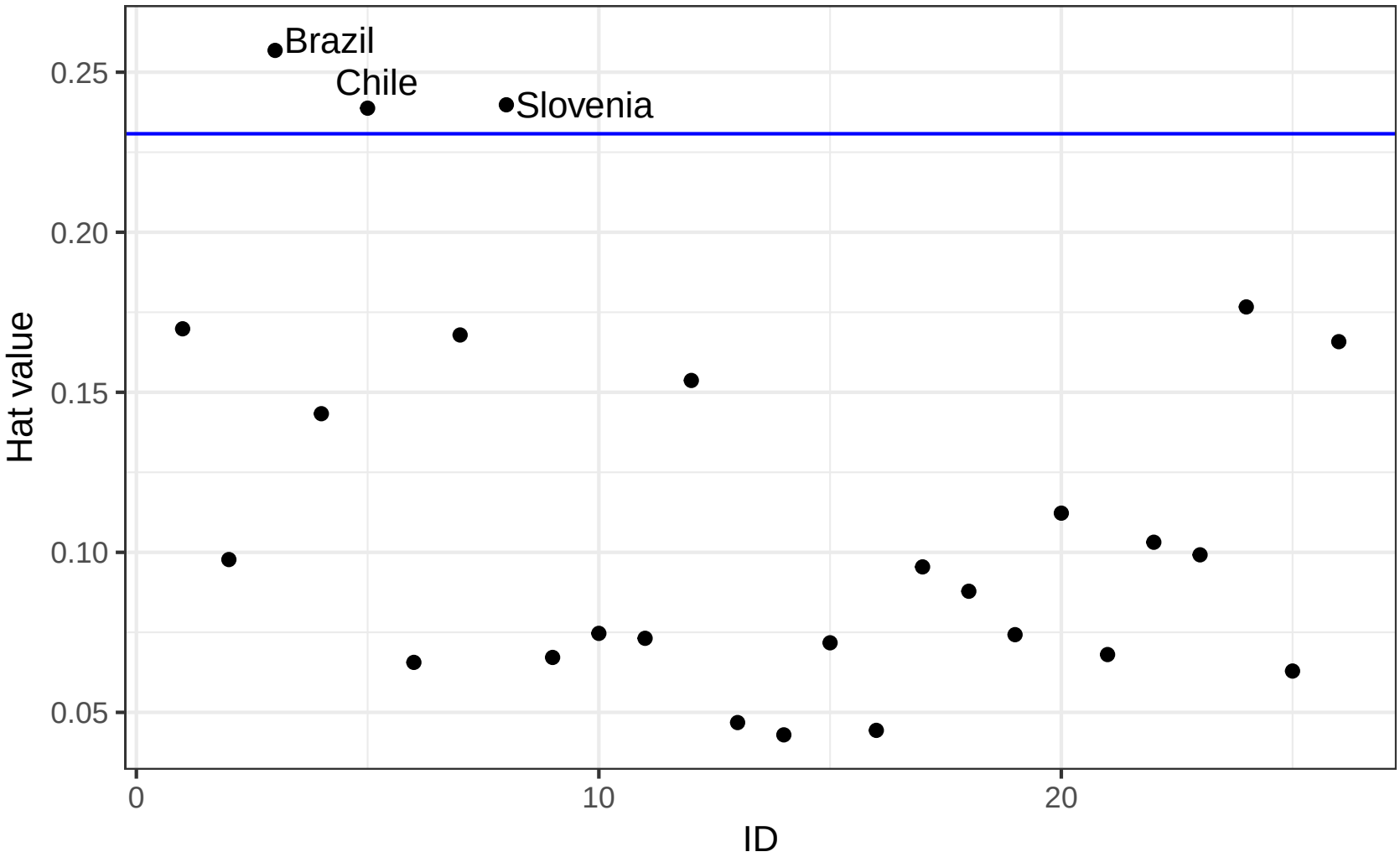
Hat values

The hat value is a common way of measuring leverage. Fitted values can be expressed in terms of observed values:

$$\hat{y}_j = h_{1j}y_1 + h_{2j}y_2 + \cdots + h_{jj}y_j + \cdots + h_{nj}y_n = \sum_{i=1}^{n} h_{ij}y_i.$$

So, the weight, h_{ij}, captures the extent to which y_i can affect $\hat{y}_j$. It may be shown that h_i summarizes the potential influence of y_i on all the fitted values. They are bounded by $1/n$ and 1. The average hat-value is $(k+1)/n$. Values twice this considered noteworthy (some people use three times).

Hat values plot



Note that there are some cases with bigger hat values that the two influential cases. Shows limitation of hat values.

Studentized residuals

If we refit the model deleting the ith observation, we obtain estimate of the standard deviation of residuals, σ_{-i} based on $n-1$ cases.

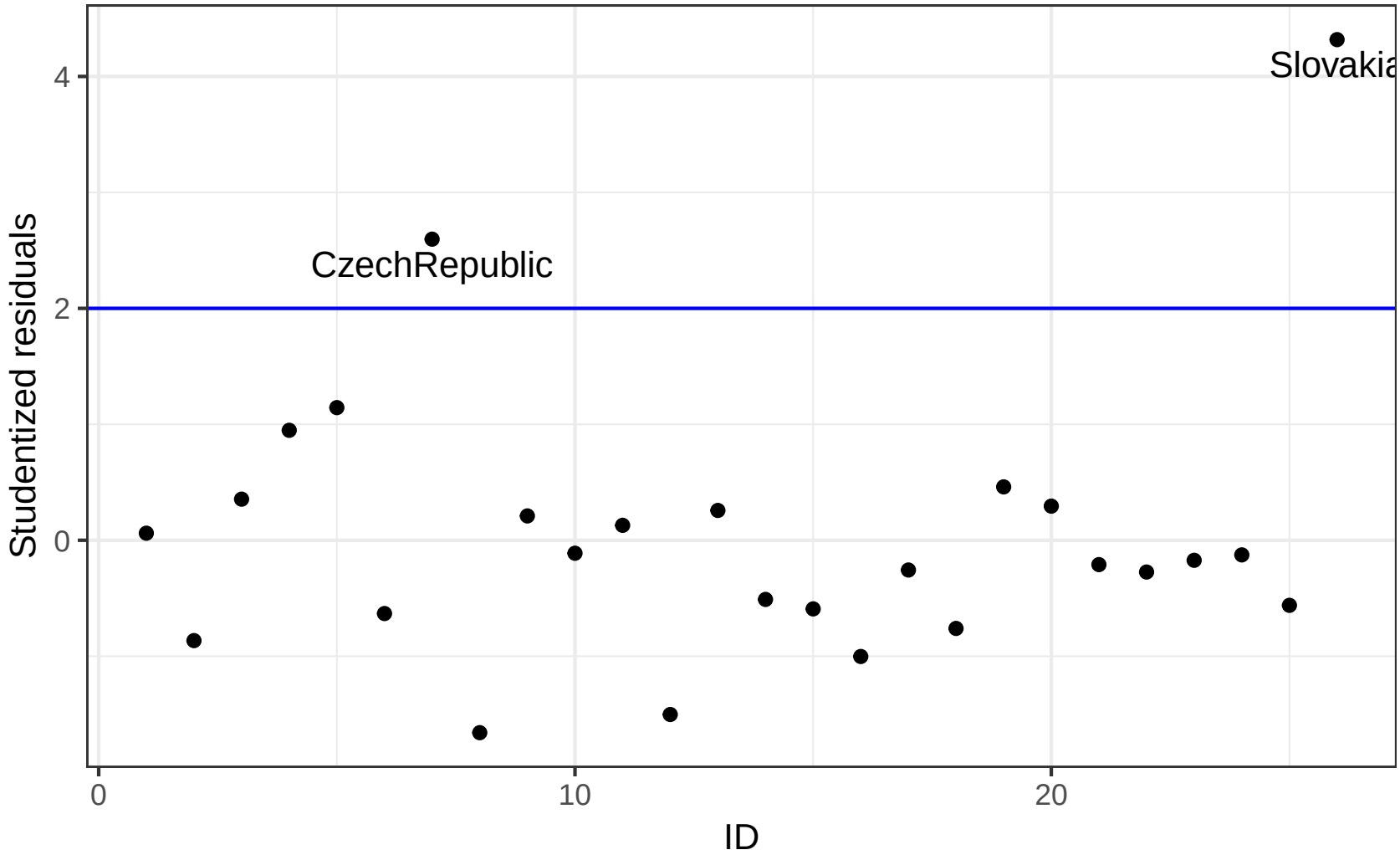
$$\epsilon_i^t = \frac{\epsilon_i}{\sigma_{-i}\sqrt{1-h_i}}$$

Studentized residuals follow a t-distribution with $n-k-2$ degrees of freedom. Observations outside ± 2 range statistically significant.

Significance tests have to be corrected for multiple comparisons. This is done for you using the `outlierTest` function in the `car` package.

```
             rstudent unadjusted p-value Bonferonni p
Slovakia         4.32           0.000278      0.00722
CzechRepublic    2.60           0.016461      0.42798
```

Studentized residuals plot



DFBETA

A direct measure of the influence of an observation on regression parameter estimates is:

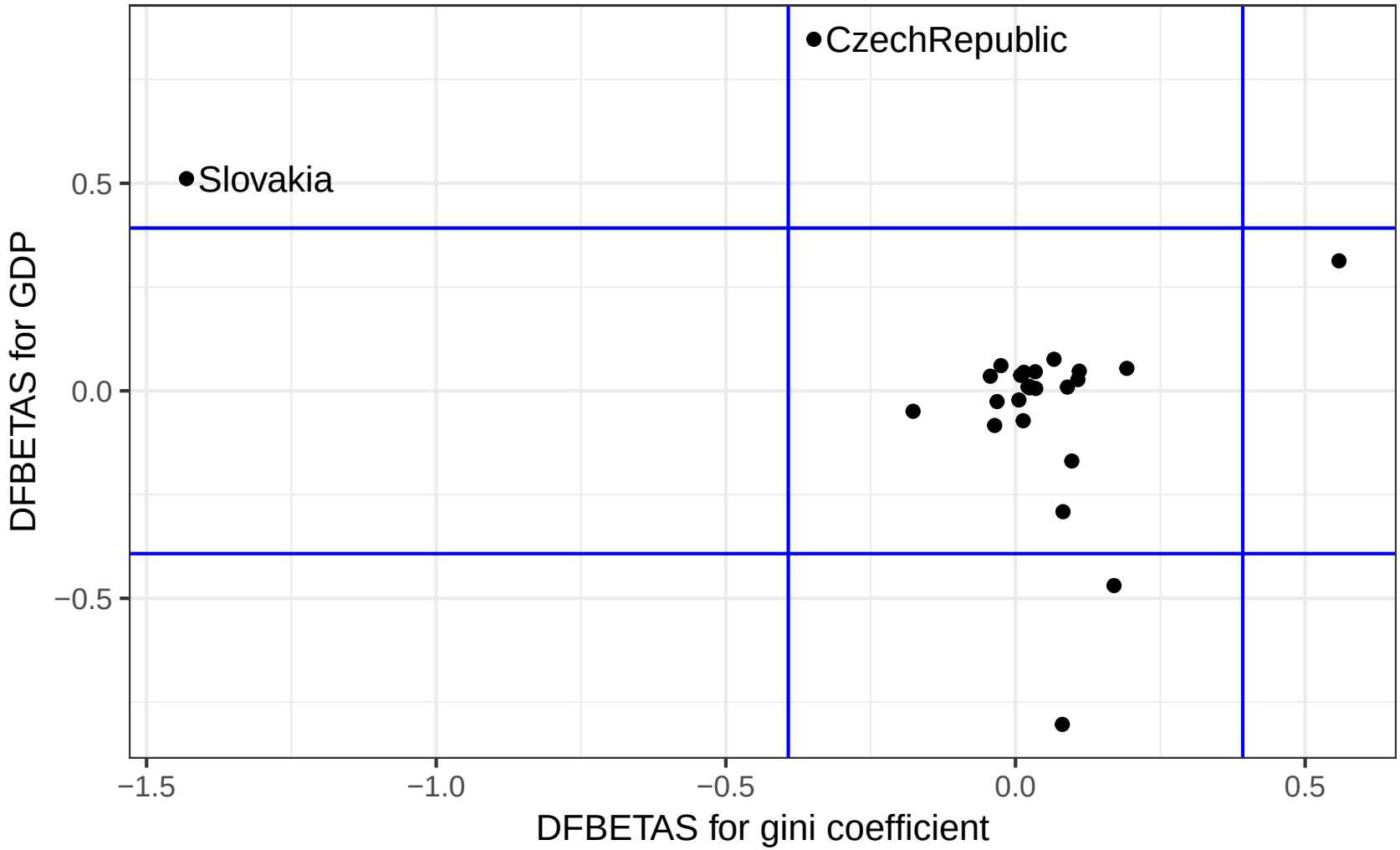
$$d_{ij} = b_j - b_{j(-i)}$$

where $b_{j(-i)}$ is the estimate of β_j with the ith observation omitted. These differences are usually scaled by (omitted) estimates of the standard error of b_j:

$$d^*_{ij} = \frac{d_{ij}}{s_{(-i)}(b_j)}.$$

The d_{ij} are often termed DFBETA and the d^*_{ij} are called DFBETAS.

DFBETA plot


CzechRepublic
Slovakia
0.5
0.0
−0.5
DFBETAS for GDP
−1.5
−1.0
−0.5
0.0
0.5
DFBETAS for gini coefficient

Cook's distance

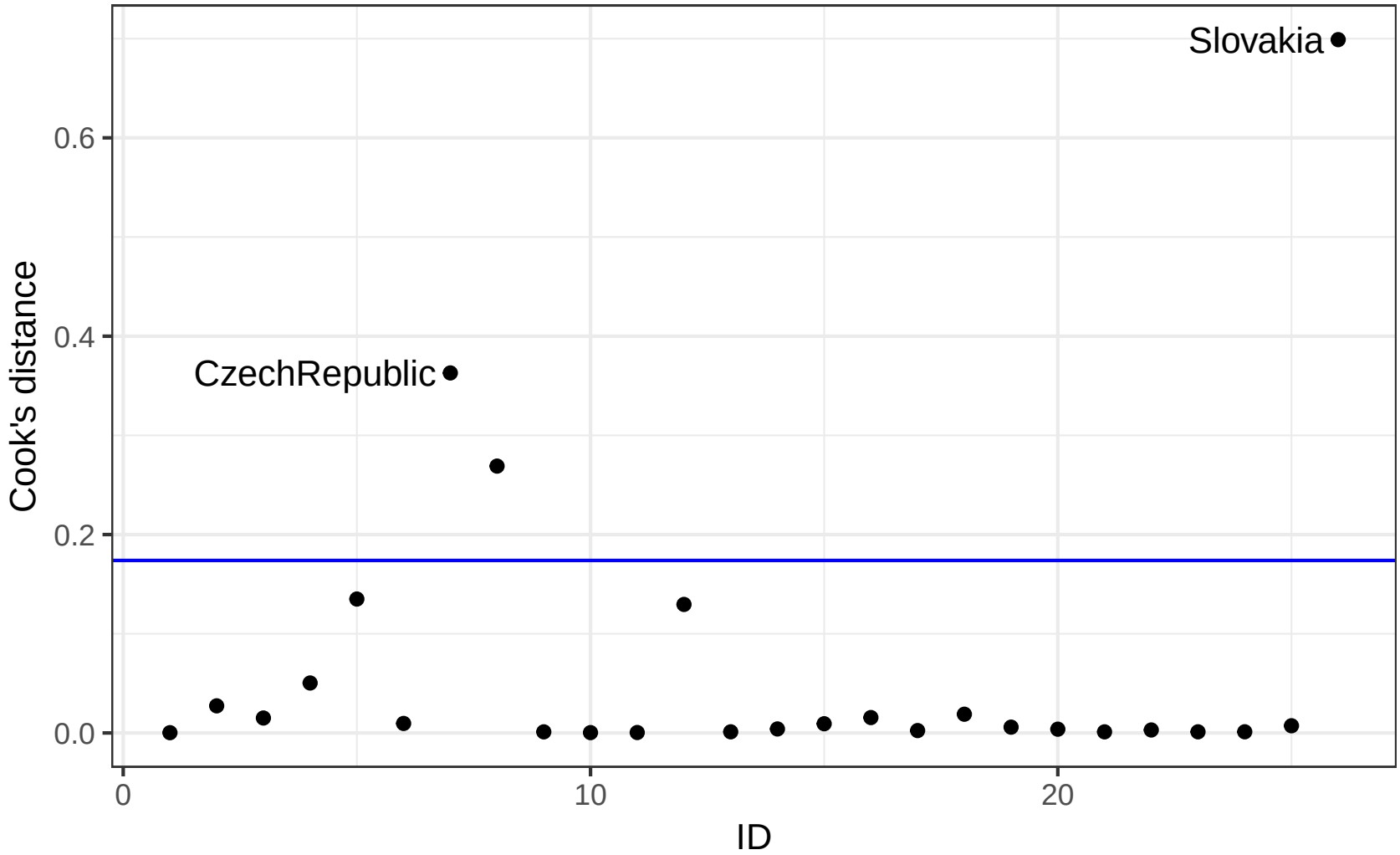
One way to use DFBETAS is to plot them for each independent variable. Another is to construct an index. Cook's distance is essentially an F statistic for the "hypothesis" that $\beta_j = b_{j(-i)}, j = 0, 1, \ldots, k$. This is calculated using:

$$D_i = \frac{\epsilon_i^{*2}}{k+1} \times \frac{h_i}{1-h_i},$$

where ϵ_i^* is the standardized residual. No formal hypothesis test, but rule of thumb is

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

Cook's distance plot



Rules of thumb

- **Hat-values** Values exceeding twice the average ($[k+1]/n$) are noteworthy.
- **Studentized residuals** About 5% of these should fall outside the range $|t_i| \leq 2$.
- **DFBETAS** $|d_{ij}^*| > 2/\sqrt{n}$
- **Cook's D** $D_i > 4/(n-k-1)$.

Heteroskedasticity

Definition

Heteroskedasticity occurs when $\mathrm{var}(\epsilon_i) \neq \sigma^2$, but varies across observations. It is especially problematic when this is related systematically to an explanatory variable.

Problems

- Increases standard errors of parameter estimates.
- Estimated standard errors are **biased**.

Solutions

- Use a different estimator for standard errors.
- Use a different estimator for regression parameters: weighted least squares.

What to do?

- Statistical tests
  - Goldfeld-Quandt test
  - Breusch-Pagan test
- Remedial action
  - Heteroskedasticity-consistent standard errors
  - Weighted least squares

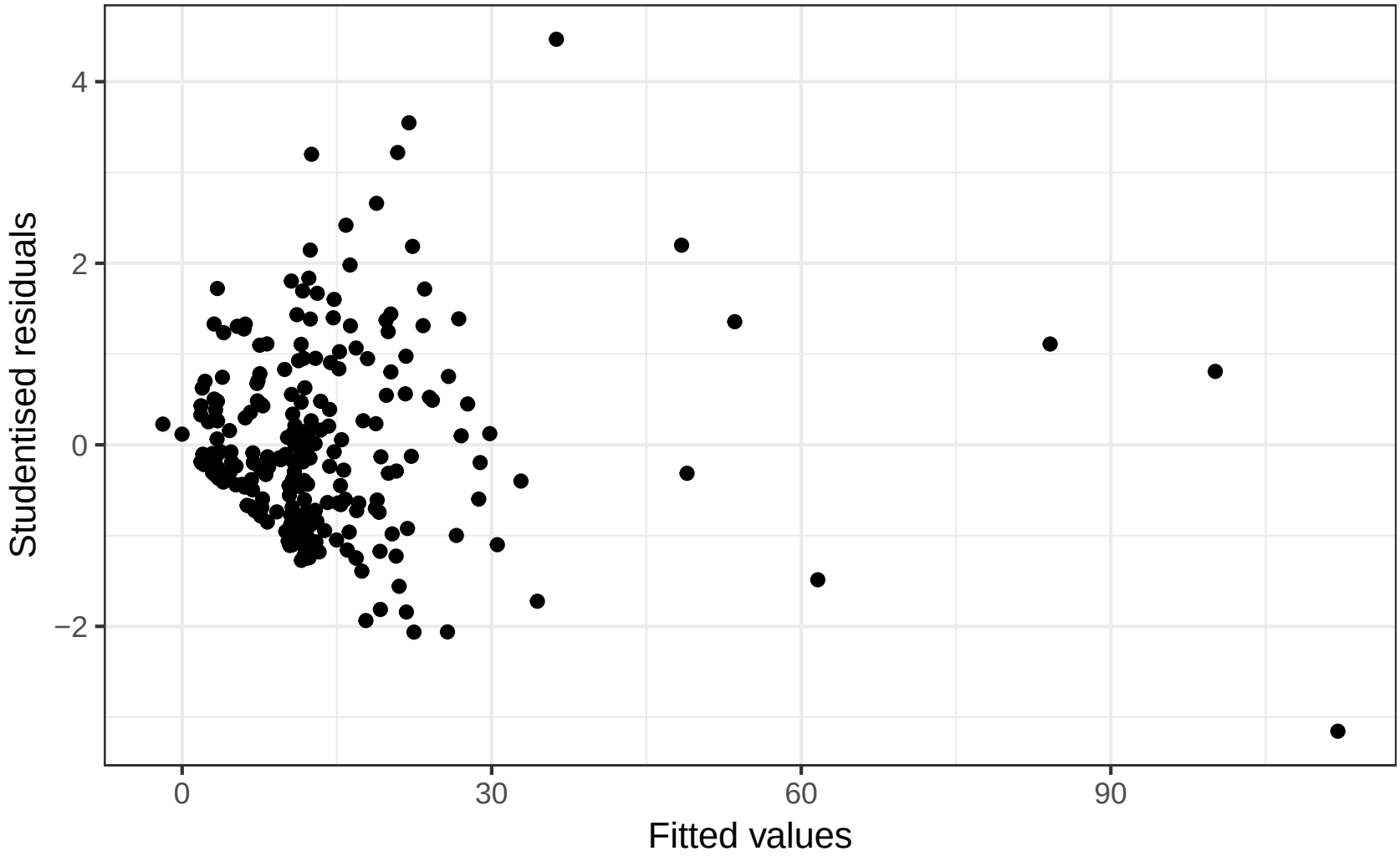
Director interlocks example

Data on the 248 largest Canadian firms in the mid-1970s. *interlocks*: the number of board members shared with other major firms; *assets*: Assets in millions of dollars; *sector*: a factor with levels BNK=banking, CON=construction, FIN=other financial, HLD=holding company, MAN=manufacturing, MER=merchandising, MIN=mining, TRN=transport, WOD=wood and paper; *nation*: nation of control, a factor with levels CAN=Candian, OTH=Other, UK, US.

```
Call:
lm(formula = interlocks ~ I(assets/1000) + sector + nation, data = Ornstein)

Residuals:
   Min     1Q Median     3Q    Max
-25.00  -6.60  -1.63   4.78  40.73

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.2669     1.5615    6.58  3.1e-10
I(assets/1000)   0.8096     0.0612   13.23  < 2e-16
sectorBNK      -17.8139     5.9065   -3.02   0.0028
sectorCON       -4.7087     4.7282   -1.00   0.3203
sectorFIN        5.1527     2.6457    1.95   0.0527
sectorHLD        0.8777     4.0041    0.22   0.8267
sectorMAN        1.1487     2.0645    0.56   0.5785
sectorMER        1.4915     2.6359    0.57   0.5721
sectorMIN        4.8803     2.0670    2.36   0.0190
sectorTRN        6.1713     2.7599    2.24   0.0263
sectorWOD        8.2283     2.6786    3.07   0.0024
nationOTH       -1.2413     2.6953   -0.46   0.6456
nationUK        -5.7752     2.6745   -2.16   0.0318
nationUS        -8.6181     1.4963   -5.76  2.6e-08

Residual standard error: 9.83 on 234 degrees of freedom
Multiple R-squared:  0.646, Adjusted R-squared:  0.627
F-statistic: 32.9 on 13 and 234 DF,  p-value: <2e-16
```

Heteroskedastic errors

A common diagnostic is to plot studentised residuals against fitted values. The cone shape is characteristic of heteroskedastic errors.



Goldfeld-Quandt test

Based on the idea that if the sample observations have been generated under the conditions of homoscedasticity, then the variance of the disturbances of one sub-sample is the same as the variances of any other sub-sample. Order cases by the variable you think variance is associated with (often fitted values from regression).

$$R = \frac{\text{SSE}_2}{\text{SSE}_1}.$$

SSE from the 1st regression:	4245.1
SSE from the 2nd regression:	17187.4
The F-statistic for this test:	4.04
The p-value for this test:	$\ll 0.05$

```
gqtest(inter1, order.by = int1.fit)
```

```
    Goldfeld-Quandt test

data:  inter1
GQ = 4, df1 = 100, df2 = 100, p-value = 9e-13
alternative hypothesis: variance increases from segment 1 to 2
```

Breusch-Pagan test

Model variances using variables thought to be related to the heteroskedasticity. First obtain residuals by OLS, then divide these by an estimate of the variance $\hat{s}^2$. Use as the dependent variable, with either the fitted values or some other variable as "explanatory" variable; the B-P statistic is the explained variance of this regression divided by 2. This has a χ^2 distribution with degrees of freedom equal to number of regressors in the second regression.

```
ncvTest(inter1)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 47    Df = 1     p = 7.15e-12
```

Transform standard errors

A common way of dealing with heteroskedasticity is to transform standard errors—recall it is standard errors *not* parameter estimates that are affected by this problem.

$$V(b) = (X'X)^{-1}X'\text{diag}(e^2)X(X'X)^{-1}.$$

The variance-covariance matrix of the parameter estimates is transformed by the square of the residuals. The square root of the diagonal of this matrix is the standard errors of the parameter estimates. This is very commonly used in practice now. They are called the heteroskedasticity-consistent standard errors or robust standard errors.

Example: HCCM results

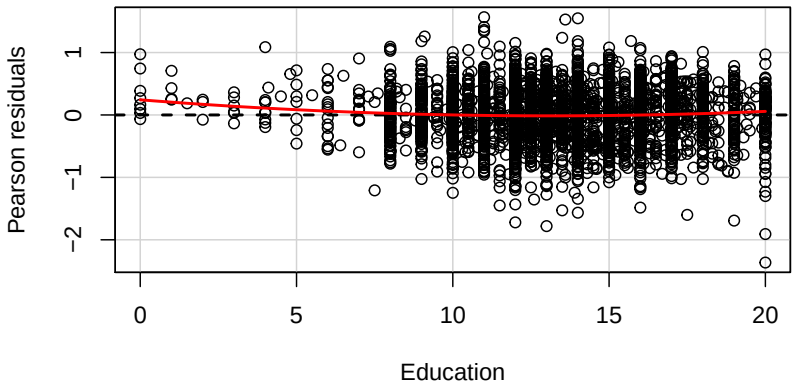
The idea is that the HC ("heteroskedasticity-consistent") standard errors are used instead of the usual ones to calculate t-statistics and hence p-values. You sometimes see these referred to as "robust" standard errors, or "White-corrected" standard error (after their inventor). The method of calculating them is sometimes referred to as a "sandwich estimator."

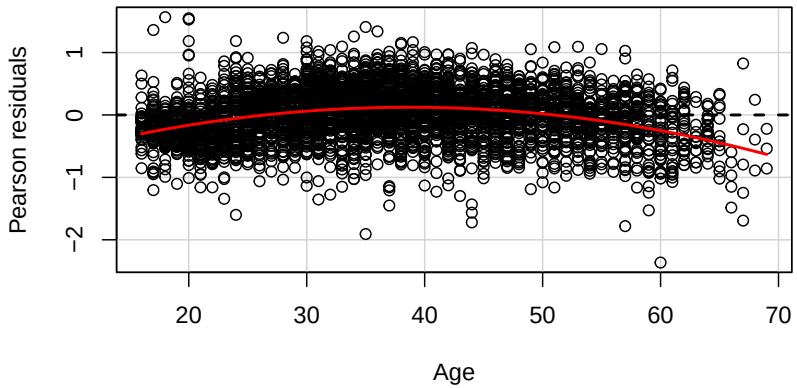
	Estimate	H-C Std. Error	t-value	Pr(>|t|)
(Intercept)	10.27	1.50	6.83	0.00
I(assets/1000)	0.81	0.09	9.32	0.00
sectorBNK	-17.81	5.10	-3.50	0.00
sectorCON	-4.71	2.68	-1.76	0.08
sectorFIN	5.15	2.70	1.91	0.06
sectorHLD	0.88	4.47	0.20	0.84
sectorMAN	1.15	1.74	0.66	0.51
sectorMER	1.49	2.20	0.68	0.50
sectorMIN	4.88	1.81	2.70	0.01
sectorTRN	6.17	3.07	2.01	0.04
sectorWOD	8.23	3.27	2.51	0.01
nationOTH	-1.24	2.81	-0.44	0.66
nationUK	-5.78	2.06	-2.81	0.00
nationUS	-8.62	1.38	-6.25	0.00

Linearity

Residual plots

The most straightforward thing to do is plot residuals against each of the explanatory variables to look for evidence of non-linearity.





	Education	Age
Test	3.66	-21.4
Pvalue	0.00	0.0

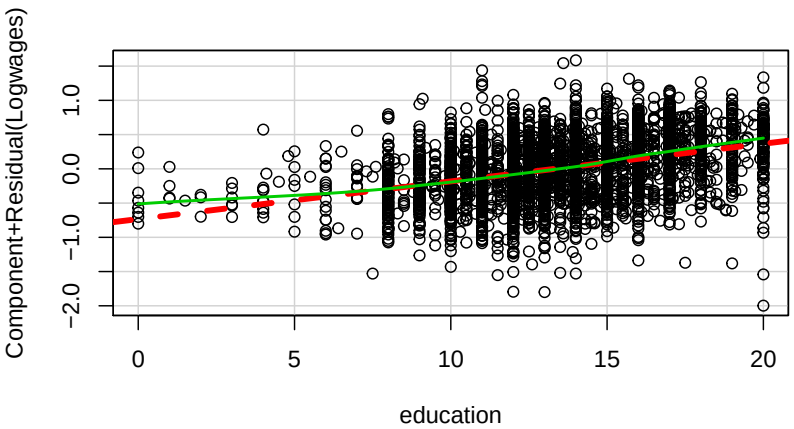
This function also provides a test of whether adding a quadratic term would be statistically significant.

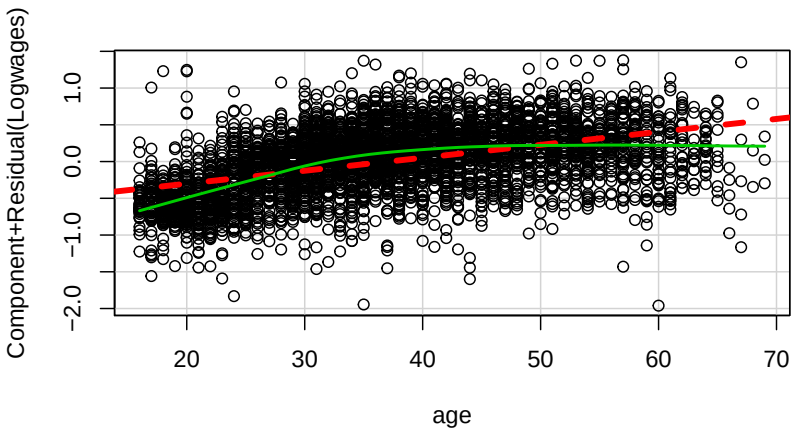
Component plus residual plots

The y axis is:

$$e + \hat{\beta}_i X_i$$

where e are residuals, $\hat{\beta}_i$ is the estimated regression parameter for the ith explanatory variable, X_i, which is plotted on the x-axis. The augmented plots shown also have the linear fit (red dotted line) and a non-parametric 'smoother' (green solid line). This can also show a departure from linearity.



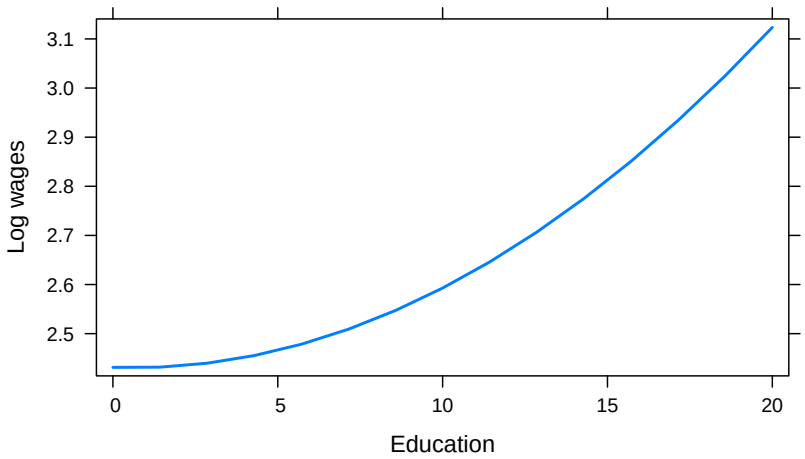


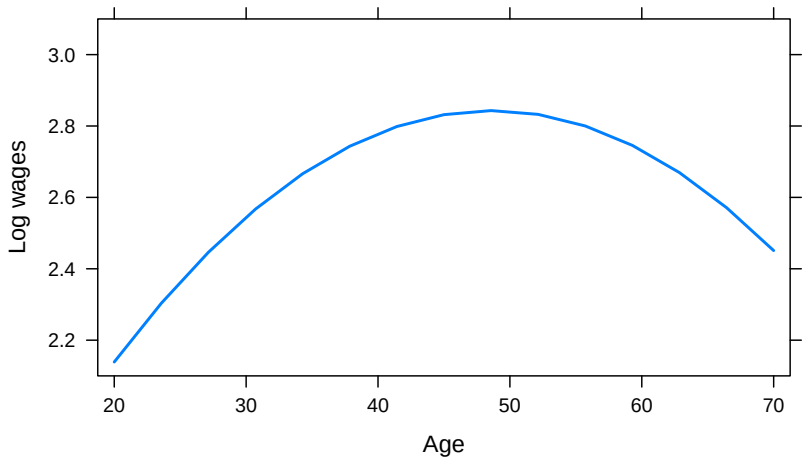
Add quadratic terms

```
Call:
lm(formula = Logwages ~ sex + language + poly(education, 2, raw = TRUE) +
    poly(age, 2, raw = TRUE), data = SLID)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0855 -0.2404  0.0223  0.2515  1.7813

Coefficients:
                                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)                      0.4055703  0.0920226    4.41 0.0000107
sexMale                          0.2215223  0.0125469   17.66   < 2e-16
languageFrench                  -0.0133857  0.0255896   -0.52      0.60
languageOther                   -0.0089290  0.0197519   -0.45      0.65
poly(education, 2, raw = TRUE)1 -0.0023108  0.0110123   -0.21      0.83
poly(education, 2, raw = TRUE)2  0.0018456  0.0004094    4.51 0.0000067
poly(age, 2, raw = TRUE)1        0.0835514  0.0031156   26.82   < 2e-16
poly(age, 2, raw = TRUE)2       -0.0008590  0.0000398  -21.57   < 2e-16

Residual standard error: 0.395 on 3979 degrees of freedom
Multiple R-squared:  0.384, Adjusted R-squared:  0.383
F-statistic:  354 on 7 and 3979 DF,  p-value: <2e-16
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
3981	697	NA	NA	NA	NA
3979	622	2	75.1	240	0

Effect plots



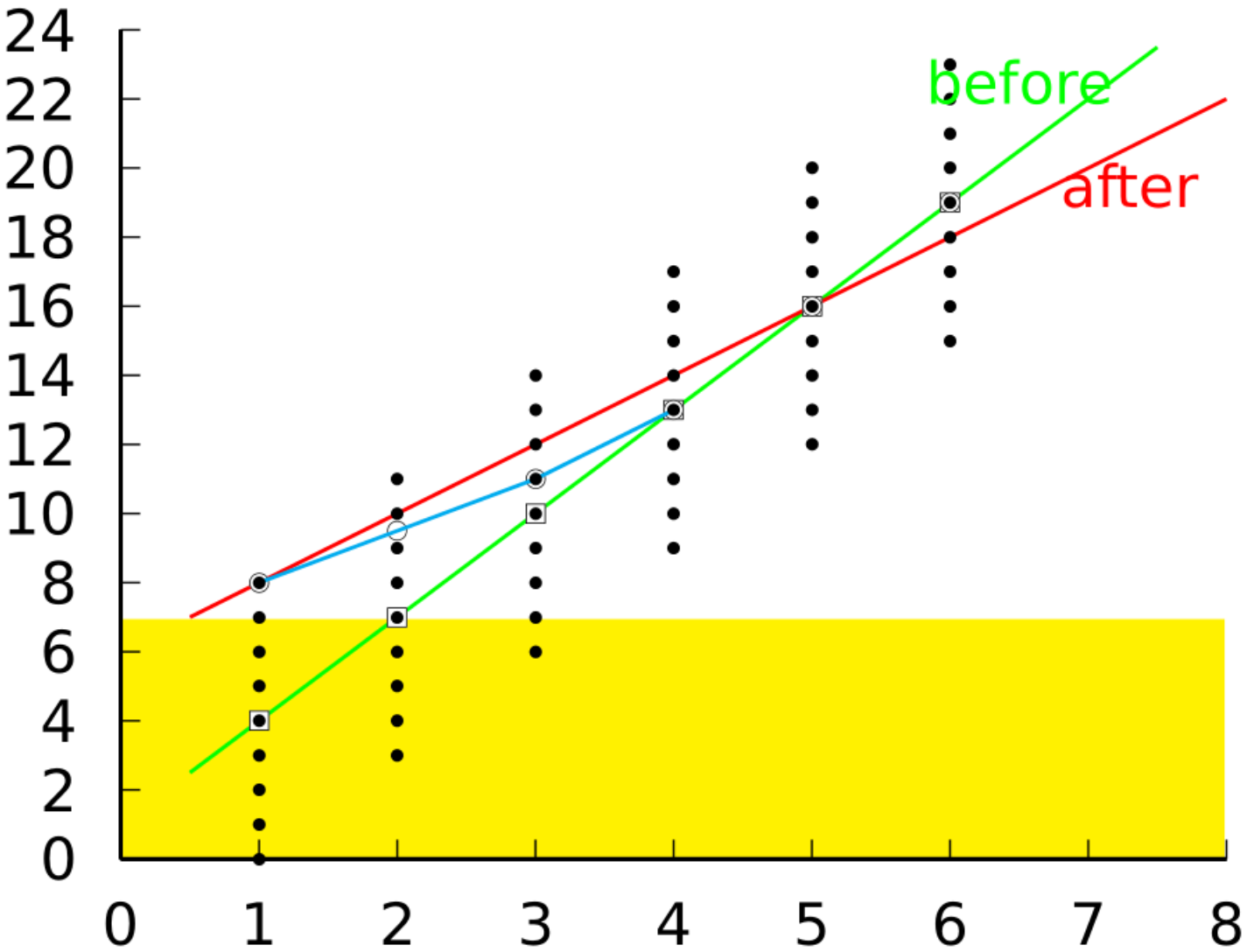


Selection models

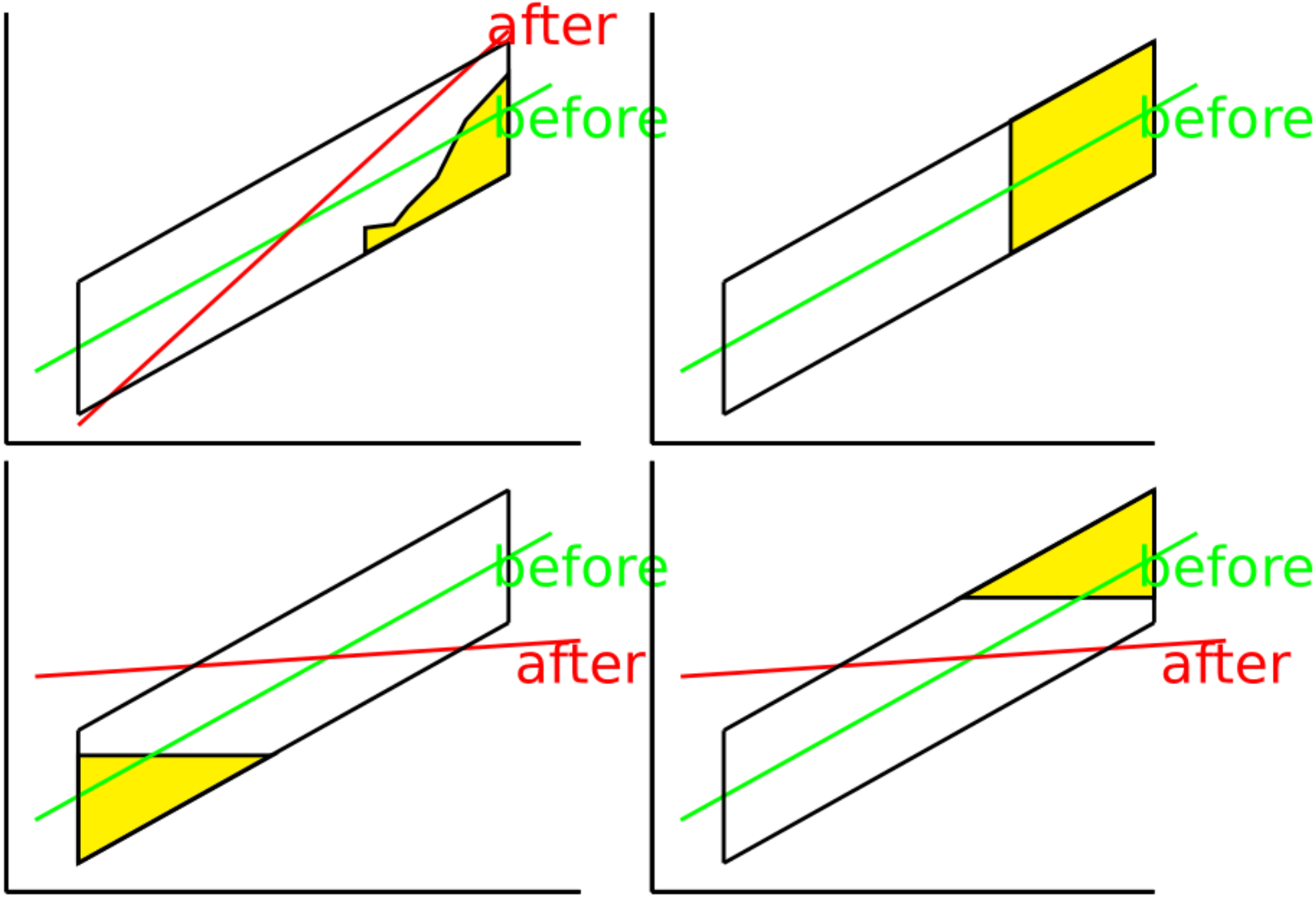
Sample selection bias

A general issue, not only concerning linear regression. It is important because it undermines external and internal validity. That is, the problem is not solved by claiming to be interested only in a sub-set of the population. In effect sample selection excludes a regressor that is correlated with an included regressor.

Illustration



Types of selection



Intuition

- Non-random selection—inference may not extend to the unobserved group
- Example: Suppose we observe that college grades are uncorrelated with success in graduate school
- Can we infer that college grades are irrelevant?
- No: applicants admitted with low grades may not be representative of the population with low grades
- Unmeasured variables (e.g. motivation) used in the admissions process might explain why those who enter graduate school with low grades do as well as those who enter graduate school with high grades

Selection equation

- z_i^* =latent variable, DV of selection equation; the propensity to be including in the sample;
- w_i' = vector of covariates for unit i for selection equation;
- α = vector of coefficients for selection equation;
- ϵ_i = random disturbance for unit i for selection equation;

$$z_i^* = w_i'\alpha + \epsilon_i.$$

Outcome equation

- y_i = DV of outcome equation;
- x_i' = vector of covariates for unit i for outcome equation;
- β = vector of coefficients for outcome equation;
- u_i = random disturbance for unit i for outcome equation;

$$y_i = x_i'\beta + u_i$$

Heckman model

Assume that y_i is observed if and only if a second, unobserved latent variable, z_i^* exceeds a particular threshold:

$$z_i = \begin{cases} 1 \text{ if } z_i^* > 0; \\ 0 \text{ otherwise} \end{cases}$$

So, we first estimate the probability that $z_i = 1$, and use a transformation of this predicted probability (known as the Mills ratio) as an independent variable in the outcome equation.

Sample selection bias: Conclusions

- If potential observations from some population of interest are excluded on a nonrandom basis, one risks sample selection bias.
- It is difficult to anticipate whether the biased regression estimates overstate or understate the true causal effects.
- Problems caused by nonrandom exclusion of observations are manifested in the expected values of the endogenous variable.

Example

A common example of sample selection is when studying wages. In order to earn a wage, you have to have a job. You are more likely to have a job if you are able to earn a good wage. So, there is likely to be sample selection. This is the ordinary regression.

```
Call:
lm(formula = Logwage ~ education + age, data = wom)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3116 -0.1389  0.0223  0.1769  0.6962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.362865   0.041026   57.59   <2e-16
education   0.038897   0.002297   16.93   <2e-16
age         0.006358   0.000863    7.37    3e-13

Residual standard error: 0.251 on 1340 degrees of freedom
  (657 observations deleted due to missingness)
Multiple R-squared:  0.231, Adjusted R-squared:  0.23
F-statistic:  201 on 2 and 1340 DF,  p-value: <2e-16
```

Sample selection results

```
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
2000 observations (657 censored and 1343 observed)
11 free parameters (df = 1990)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.46737    0.19256  -12.81  < 2e-16
married      0.43086    0.07421    5.81  7.4e-09
children     0.44732    0.02874   15.56  < 2e-16
education    0.05836    0.01097    5.32  1.2e-07
age          0.03472    0.00423    8.21  3.9e-16
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.10581    0.05781    36.4   <2e-16
education    0.04303    0.00249    17.2   <2e-16
age          0.00950    0.00102     9.3   <2e-16
Multiple R-Squared:0.26,    Adjusted R-Squared:0.258
   Error terms:
              Estimate Std. Error t value Pr(>|t|)
invMillsRatio    0.192      0.028    6.85  9.6e-12
sigma            0.276         NA      NA       NA
rho              0.696         NA      NA       NA
--------------------------------------------
```