Outliers

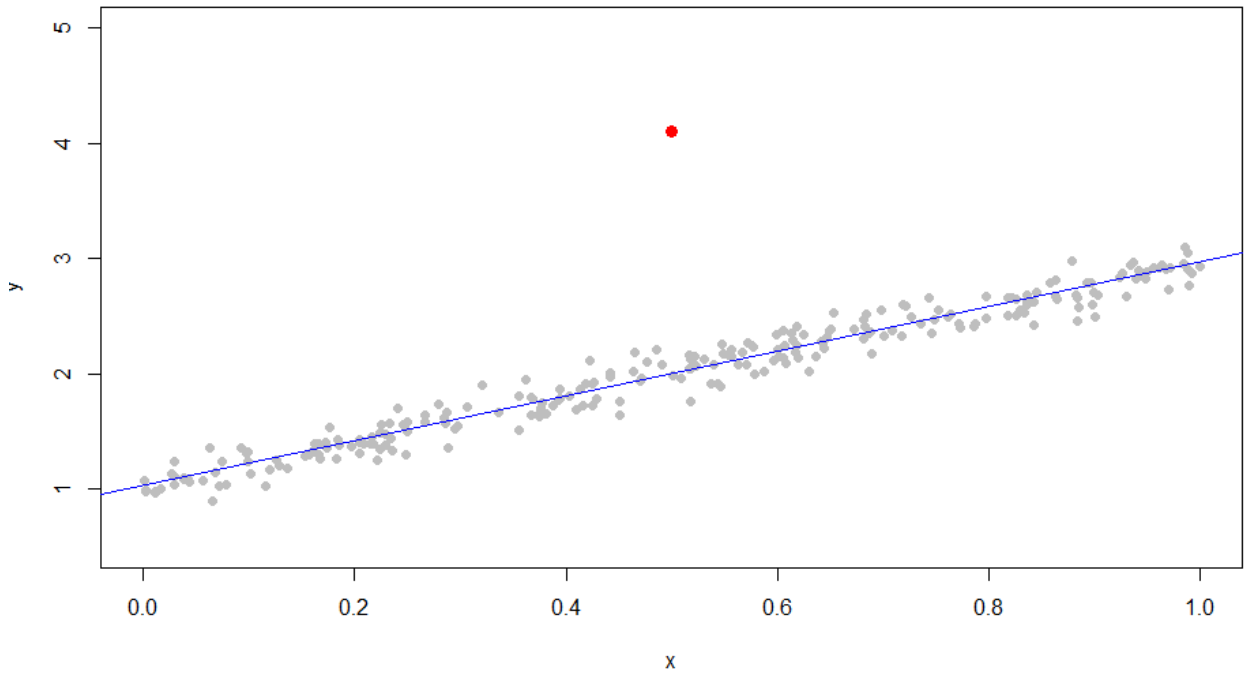
Lecture 1

Piotr Dybka

Econometrics in Practice

What are outliers?

An outlier (also called an unusual or extreme observation) is an observation that differs substantially from the rest and may distort the estimation results.



Causes

Errors in data collection or processing

1. Errors when entering variables (a missing decimal point, a flipped sign)
2. Coding conventions for variables (e.g. coding a missing observation as 9999)
3. Misunderstanding a survey question (e.g. annual instead of monthly income)

Independent factors

1. An unusual situation (e.g. the salary of the CEO of a listed company)
2. The effect of one-off shocks (e.g. elevated price volatility during a financial crisis)

Consequences – introduction

This lecture focuses on the OLS estimator, but the problems described also apply to other estimators (e.g. logistic regression).

When estimating a model we may notice that a given observation belongs to one of two groups:

- **Outliers** – observations for which we observe very large residuals
- **Influential observations** – removing an observation of this type leads to substantial differences in the estimated parameters, even in large samples

Remember: an outlier $\neq$ an influential observation

Consequences

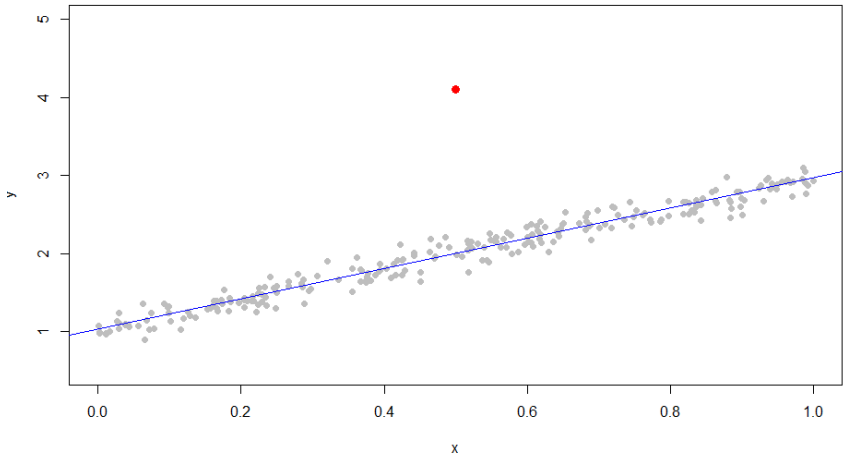
Suppose we are dealing with an outlier that is not influential:

```
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28758 -0.08277 -0.00638  0.06456  2.09959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03039    0.02188    47.1   <2e-16 ***
x            1.94004    0.03767    51.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1728 on 248 DF
Multiple R-squared:  0.9145
Adjusted R-squared:  0.9142
F-statistic: 2653 on 1 and 248 DF, p-value: <2.2e-16
```



Consequences

Let us start from the formula for the variance-covariance matrix:

$$D^2\left(\hat{\beta}\right) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

which is estimated as

$$\widehat{D^2}\left(\hat{\beta}\right) = S^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

where S^2 is the **observed** variance of the model residuals, given by

$$S^2 = \frac{1}{N-K}\sum_{i=1}^{N}(e_i - \bar{e})^2 = \frac{1}{N-K}\sum_{i=1}^{N}e_i^2$$

N – number of observations, e_i – model residuals, $\bar{e}$ – mean residual (in a model with an intercept $\bar{e} = 0$), K – number of parameters (including the intercept).

The diagonal elements of the variance-covariance matrix give the variance of the estimated parameters. If we have **outliers** (large values of e_i), the variance of the estimated parameters increases (**low precision of the estimates**).

Consequences

So let us remove the outlier. We obtain the following regression results:

Regression WITH the outlier:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.28758 -0.08277 -0.00638  0.06456  2.09959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03039    0.02188    47.1   <2e-16 ***
x            1.94004    0.03767    51.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression WITHOUT the outlier:

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.279271 -0.074650  0.001844  0.071560  0.269638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02180    0.01392   73.42   <2e-16 ***
x[-125]      1.94035    0.02395   81.02   <2e-16 ***
```

[-125] means dropping the 125th observation (the outlier)

For the variable x the standard errors are **36.4%** lower!

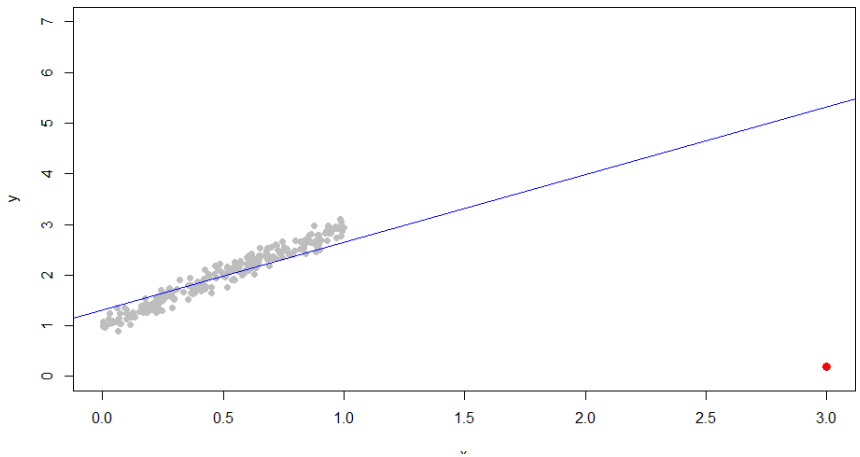
Consequences

Suppose we are dealing with an observation that is **both an outlier and influential**:

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1095 -0.1499  0.0416  0.1762  0.5004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.30612    0.04509   28.97   <2e-16 ***
x            1.33445    0.07390   18.06   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



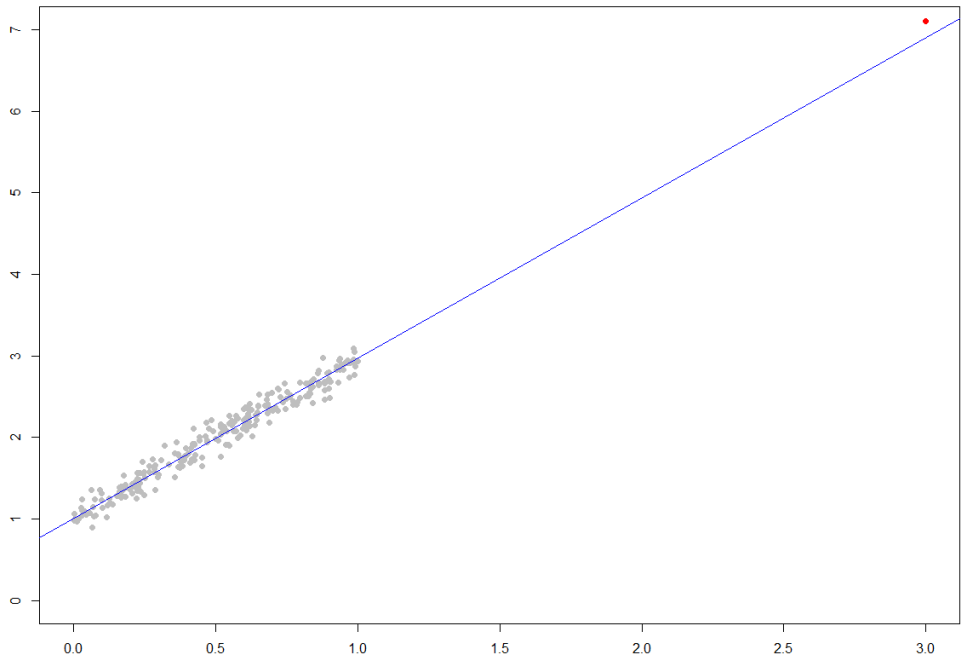
This is the same data as in the previous example, except for 1 observation.

Influential observations – always remove them?

On the previous slide we saw that an influential outlier can lead to a wrong result (a biased estimator). Should we therefore always remove influential observations?

If the outlier arose from an error – definitely yes. This is why procedures that trim extreme observations exist.

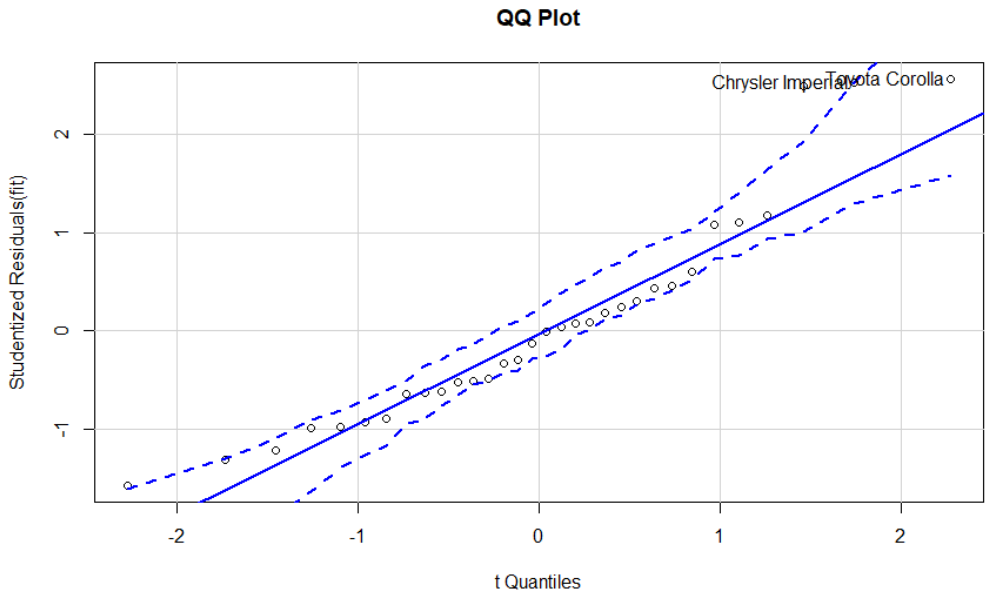
If, however, an influential observation takes unusual values for both the explanatory and the dependent variable, then it constitutes a test of our theory outside the "typical" sample. If the residual is not large (it is not an outlier), it carries valuable information and most likely did not arise from an error.



Diagnostics – the quantile plot (QQ plot)

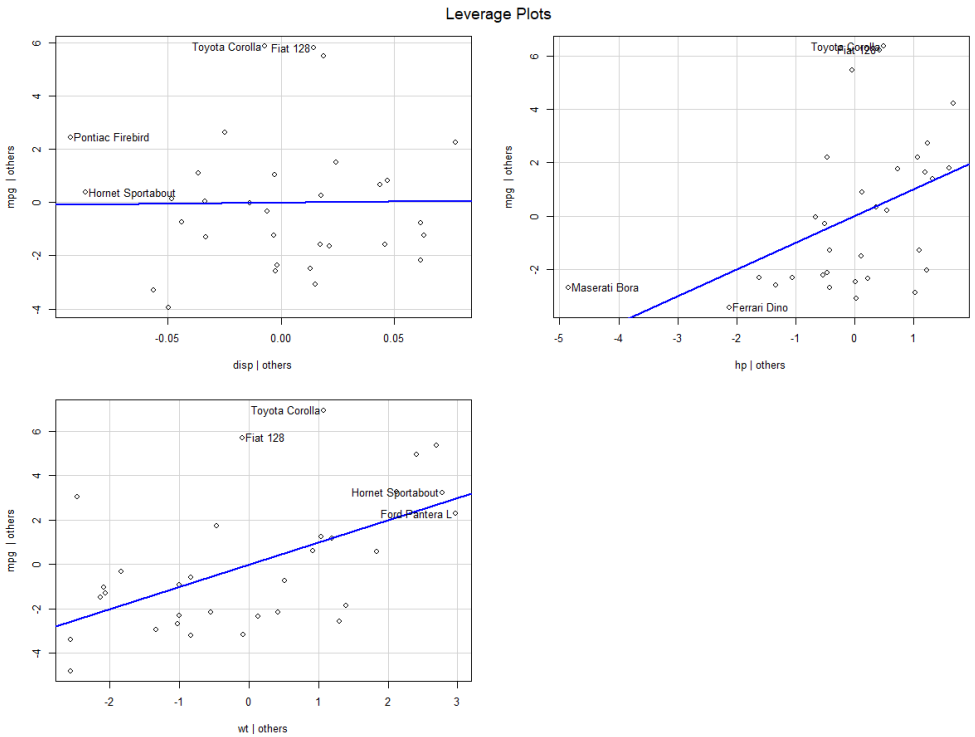
Diagnostics (detecting outliers) is based on the analysis of the model residuals. One of the tools is the so-called quantile plot (QQ plot).

1. The plot shows how the empirical quantiles of the residuals compare with the quantiles of the theoretical distribution*.
2. The points should line up along a straight line – visible departures at the ends of the plot indicate the presence of outliers.



*quantiles measure the position of a value in the sample once it has been sorted. For example quantiles of order 100, called percentiles, tell us what % of the observations in the sample take a value equal to or smaller than the given one.

Leverage plot (added-variable plot)



- The Y axis shows the residuals from a regression of the dependent variable on the **remaining** explanatory variables (excluding the variable under study)
- The X axis shows the residuals from a regression of the **variable under study** on the remaining explanatory variables
- The slope of the blue line equals the estimated coefficient on that variable in the full model – the flatter it is, the less the variable matters (flat = statistically insignificant)

1. The plot indicates which observations have the largest effect on the estimate.
2. Points at the ends of the plot matter more for the estimated parameters.

Measuring the influence of an observation (leverage)

To measure the influence of an observation we can use a measure of how far x_i departs from $\bar{x}$ (for a simple regression):

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

In general, h_i is the i-th diagonal element of the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, that is $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$.

For a model with K parameters:

$$\sum_{i=1}^{N} h_i = K, \qquad \text{where for every } i: \ \frac{1}{N} \leq h_i \leq 1$$

This means that the average leverage (h_i) equals K/N, and a value above $2K/N$ (for small samples $3K/N$) can be treated as potentially influential.

Studentized residuals

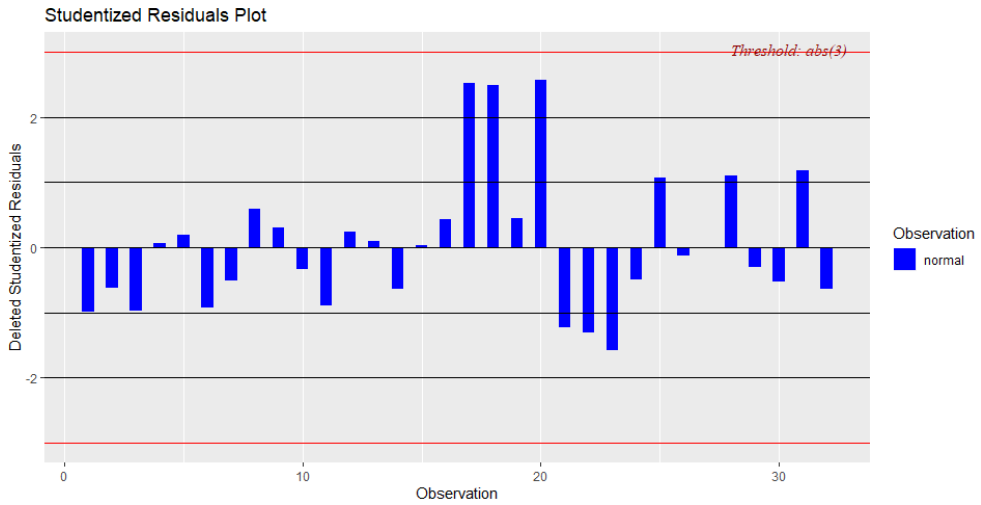
Standardised (internally studentized) residuals are residuals divided by the estimated standard deviation of the residual

$$r_i = \frac{e_i}{SE_{e_i}}, \qquad SE_{e_i} = S\sqrt{1 - h_i}$$

where

$$h_i = \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Studentized residuals should fall within the interval $[-2; 2]$.



If we replace S with $S_{(i)}$ (estimated without the i-th observation), we obtain the studentized deleted residual (RStudent) – and that is what is plotted above.

Bonferroni outlier test

The $[-2; 2]$ rule is a heuristic. Formally, we can **test** the hypothesis:

$$H_0: \text{ observation } i \text{ is not an outlier} \qquad H_1: \text{ observation } i \text{ is an outlier}$$

The test statistic is the studentized deleted residual, which under H_0 follows

$$t_i = \frac{e_i}{S_{(i)}\sqrt{1-h_i}} \sim t(N-K-1)$$

- **The problem**: we are not testing a single, pre-selected observation – we are looking for the *largest* among N residuals. At $\alpha = 0.05$ and $N = 32$ we would expect about 1.6 "outliers" purely by chance (this is exactly *swamping*).
- **The Bonferroni correction**: we multiply the p-value of the largest $|t_i|$ by the number of tests performed:

$$p_{Bonf} = \min\left(N \cdot 2P\left(t_{N-K-1} > |t_i|\right),\ 1\right)$$

- In R: `car::outlierTest(fit)`

Bonferroni test – an example

For our model `lm(mpg~disp+hp+wt, data=mtcars)`:

```
> outlierTest(fit)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
               rstudent unadjusted p-value Bonferroni p
Toyota Corolla 2.564129           0.016225      0.51919
```

Interpretation:

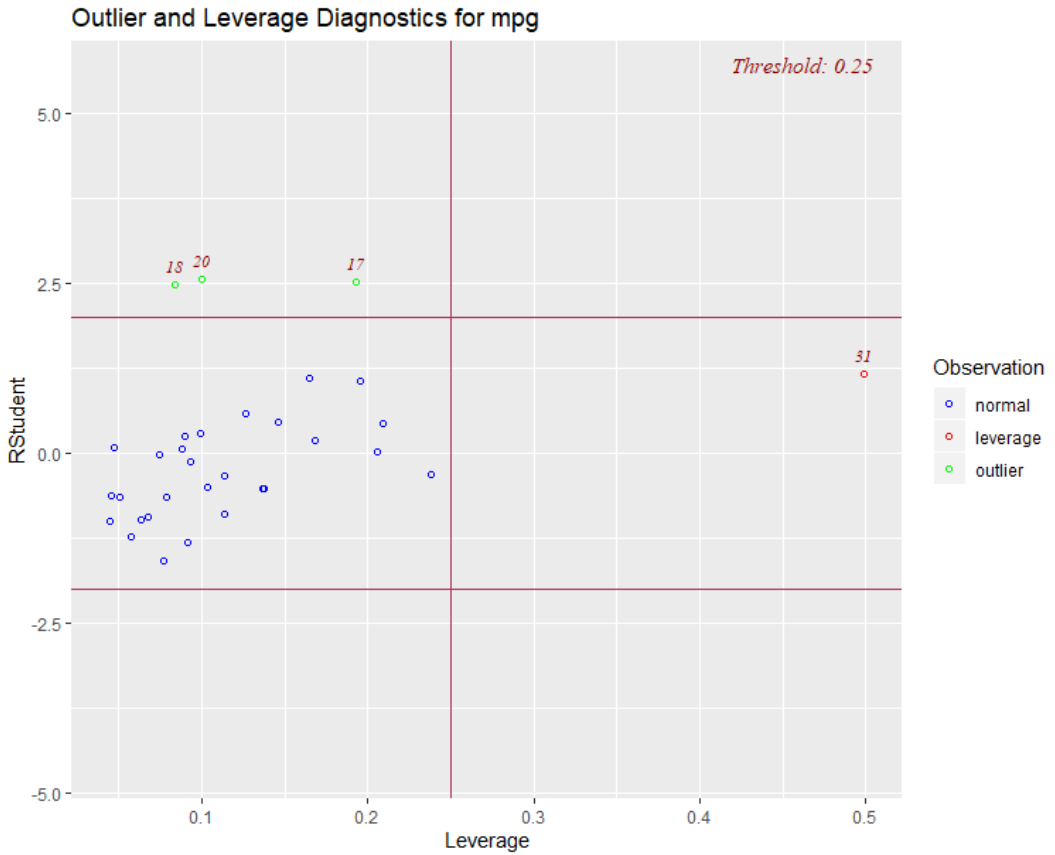
- The unadjusted $p = 0.016$ would suggest that the Toyota Corolla is an outlier.
- Once we account for having screened all $N = 32$ observations, $p_{Bonf} = 0.52$ – **no grounds to reject** H_0.
- The critical value of $|t|$ here is 3.52, not 2 – the plots on the previous slides flagged observations 17, 18 and 20 as suspicious, but the formal test does not confirm this.

Note: the test points to only *one*, most extreme observation – with several outliers it may fail (*masking*). In that case we apply it iteratively.

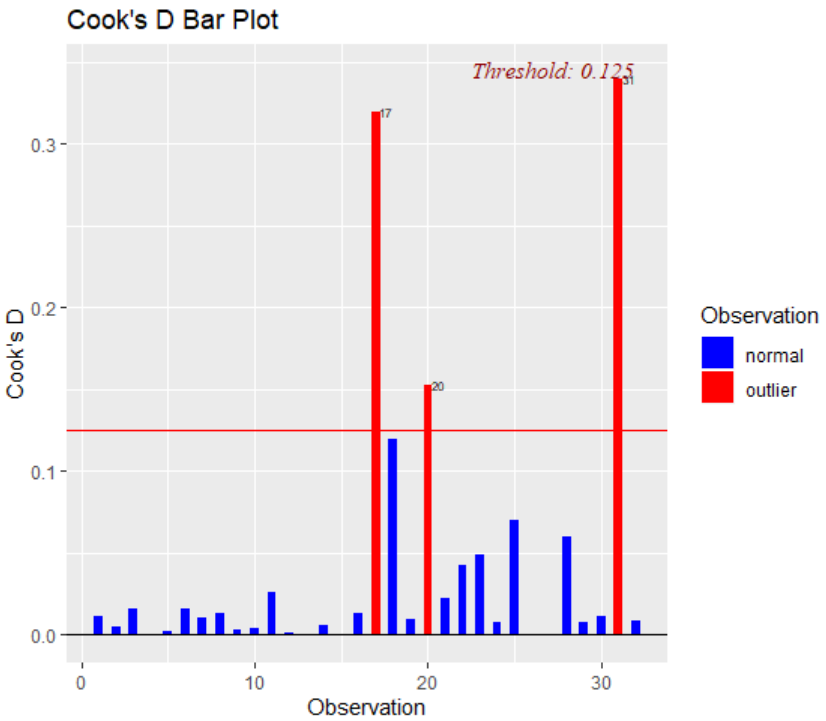
Studentized residuals vs leverage

We can also plot the studentized residuals against leverage – this lets us divide the observations into:

- ordinary ones,
- outliers that are not influential,
- outliers that are influential,
- influential observations that are not outliers.



Cook's distance (Cook's D)



Cook's distance measures the change in the regression coefficients when the i-th observation is dropped:

$$D_i = \frac{e_i^2}{K \cdot MSE} \left(\frac{h_i}{(1-h_i)^2} \right)$$

where MSE = mean square error.

- $\frac{e_i^2}{K \cdot MSE}$ measures how far the observation lies from the fit
- $\frac{h_i}{(1-h_i)^2}$ measures leverage (influence)

The threshold can be set as:

1. $4/N$ (plot on the left)
2. $3\bar{D}$ (three times the average Cook's distance)
3. 1.0
4. $F(0.5; K, N-K)$ (values above the 50th percentile of the F distribution)

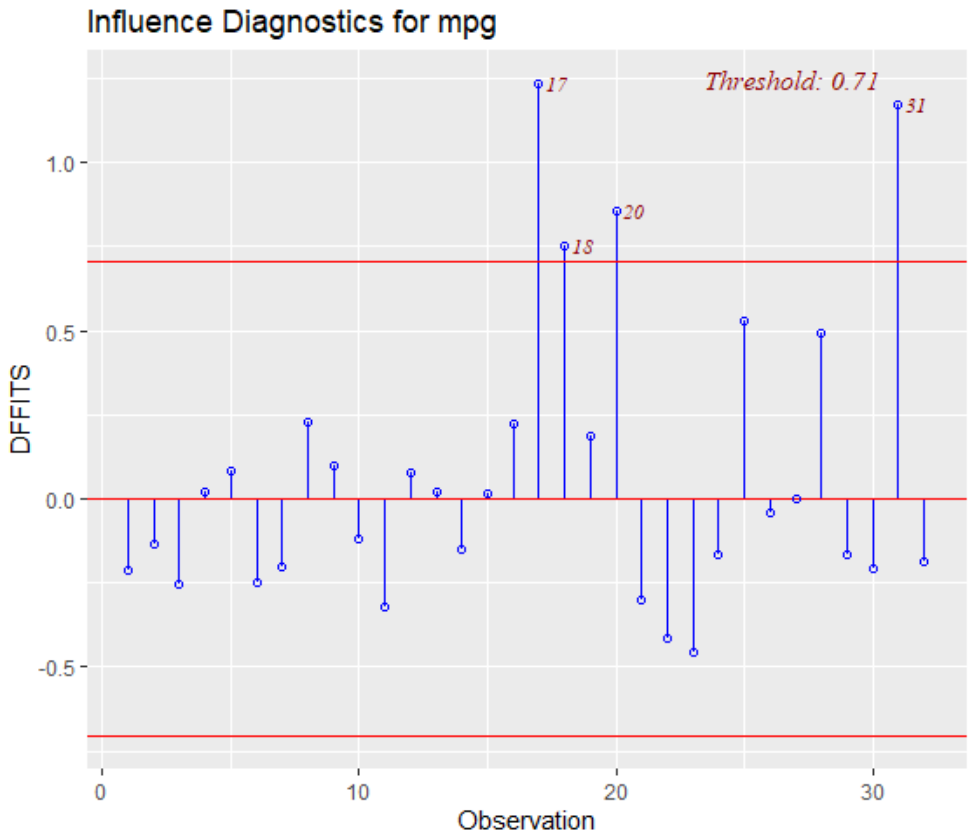
DFFITS (Difference in Fits)

DFFITS checks how the i-th observation affects the fitted value obtained from the estimated regression equation:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{S_{(i)}\sqrt{h_i}} = \frac{e_i}{S_{(i)}} \cdot \frac{\sqrt{h_i}}{1 - h_i}$$

The subscript (i) denotes dropping the i-th **observation**.

The threshold is usually taken to be $2\sqrt{\frac{K}{N}}$.



DFBETA(S) (Difference in Betas)

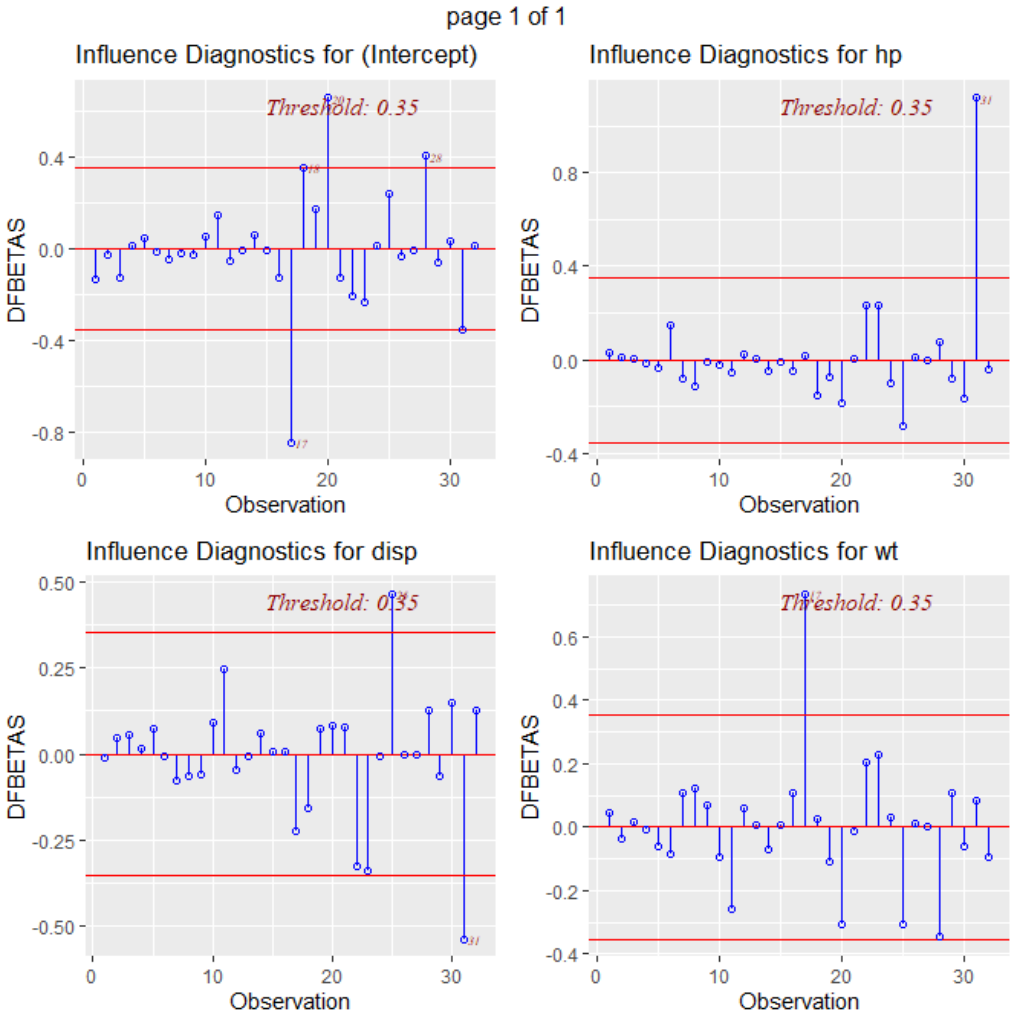
DFBETA analyses changes in the regression coefficients (β). For each $\hat{\beta}_j$ (j – variables) we can compute the change in $\hat{\beta}_j$ when the i-th observation is dropped:

$$DFBETA_{j(i)} = \hat{\beta}_j - \hat{\beta}_{j(i)}$$

After standardisation we obtain DFBETAS:

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}_{jj}}}$$

Threshold: $\frac{2}{\sqrt{N}}$



An outlier, or a misspecified model?

Before you remove outliers, check whether **they have something in common**. In our model all three suspicious observations have **positive** residuals – the model under-predicts them:

observation	*wt*	e_i	RStudent
Toyota Corolla	1.835	+5.86	2.56
Fiat 128	2.200	+5.79	2.50
Chrysler Imperial	5.345	+5.49	2.53

- The range of *wt* is [1.51; 5.42] – two observations sit at the *bottom* and one at the *top* of the weight distribution.
- Under-prediction at **both ends** of an explanatory variable is a typical symptom of a **convex** relationship that a linear term cannot capture.
- A check: replacing *wt* with ln(*wt*) raises R^2 from 0.827 to 0.867, and the RStudent for the Chrysler falls from 2.53 to 2.06.

Conclusion: removing these observations would improve the fit, but it would *hide* the specification error instead of fixing it. A pattern among the outliers is a signal about the form of the model, not a data-cleaning task.

What can we do about them?

1. Removing erroneous information
2. Winsorizing (modifying observations)
3. Least trimmed squares (LTS)
4. Least absolute deviations (LAD)
5. Quantile regression

1. Removing erroneous information

- When we are certain that an observation has been miscoded, arose from a misunderstood question or arose under exceptional circumstances that we do not want to include in the model, we should remove that **observation** (regardless of its influence; even if it is not influential, it will reduce the precision of the estimates).
- When, however, we have no grounds to claim that the observation is erroneous, removing it may mean losing valuable information (in particular when it is only influential and not an outlier).
- Two phenomena can arise when identifying outliers:
  - *Masking* – with a larger number of outliers the methods may fail to detect them (they are "hidden"); in that case the methods presented here may only start detecting outliers after several observations of this type have been removed.
  - *Swamping* – incorrectly classifying an observation as an outlier.

2. Winsorizing (modifying observations)

- When we do not want to remove an observation, we can use so-called winsorizing.
- Instead of removing outliers, we can replace their value with the value of the 1st or 99th percentile of the distribution of the given variable (depending on which side the variable lies out).
- The idea is that if we have an observation for which the level of one variable appears to be badly measured, we look for a reasonable value for that variable so as not to lose the observation.

3. Least trimmed squares

- The removal of outliers can also be built into the estimation procedure itself – LTS (Least Trimmed Squares).
- The idea behind LTS is that we minimise the sum of *only* the h smallest squared residuals (the remaining, largest ones are discarded):

$$S_{LTS}(\tilde{\beta}) = \sum_{i=1}^{h} \left(e^2\right)_{i:N}, \qquad \left(e^2\right)_{1:N} \leq \cdots \leq \left(e^2\right)_{N:N}$$

- In practice this means: we estimate the model, sort the observations by the absolute value of the residuals and discard a given % of the observations with the largest residuals; the procedure is repeated until the subsample giving the smallest sum of squared residuals is found.
- The `robustbase` package in R; the `ltsReg()` command, where the parameter $\alpha = h/N$ sets what fraction of the observations is to remain in the sample.

4. Least absolute deviations

- Another approach is to use the LAD (Least Absolute Deviations) estimator, where we minimise the following objective function:

$$S_{LAD}(\tilde{\beta}) = \sum_{i=1}^{N} \left| y_i - \boldsymbol{x}_i' \tilde{\beta} \right|$$

- So instead of squared residuals (as in OLS) we have absolute values.
- The LAD estimator fits the parameters $\tilde{\beta}$ to the median rather than to the mean of the distribution – the median is a measure that is more robust to the presence of outliers.

5. Quantile regression

- Quantile regression is a generalisation of LAD, where we minimise the following objective function:

$$S(\tilde{\boldsymbol{\beta}}_Q) = \sum_{i:\, y_i \geq \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q} Q \left| y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q \right| + \sum_{i:\, y_i < \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q} (1 - Q) \left| y_i - \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}}_Q \right|$$

- The parameter Q tells us for which quantile of the distribution of the dependent variable we are looking for the values of $\boldsymbol{\beta}$.
- Positive residuals (above the fitted value) are weighted by Q, and negative ones by $(1 - Q)$.
- For example, for $Q = 0.7$ estimation errors above the fitted value will matter a lot and those below will matter less; the regression will therefore be fitted so as to reflect the relationship at the 70th percentile of the distribution of the variable under study.
- For $Q = 0.5$ we fit to the median and obtain the LAD estimator.

5. Quantile regression

- In this way we can see whether the relationship under study changes at the ends of the distribution of the variable.
- In particular, large changes for the extreme quantiles may indicate the presence of outliers.