Statistical Methods for Insurance: Linear Models

Di Cook & Souhaib Ben Taieb, Econometrics and Business Statistics, Monash University
W5.C1

Overview of this class

- Quiz 3
- Linear model diagnostics
- Transformations
- READING: Ch 6, Diez, Barr, Cetinkaya-Rundel

Modeling Olympic medal counts

We fit the medal count for 2012, purely on the counts from 2008, to illustrate the influence diagnostics.

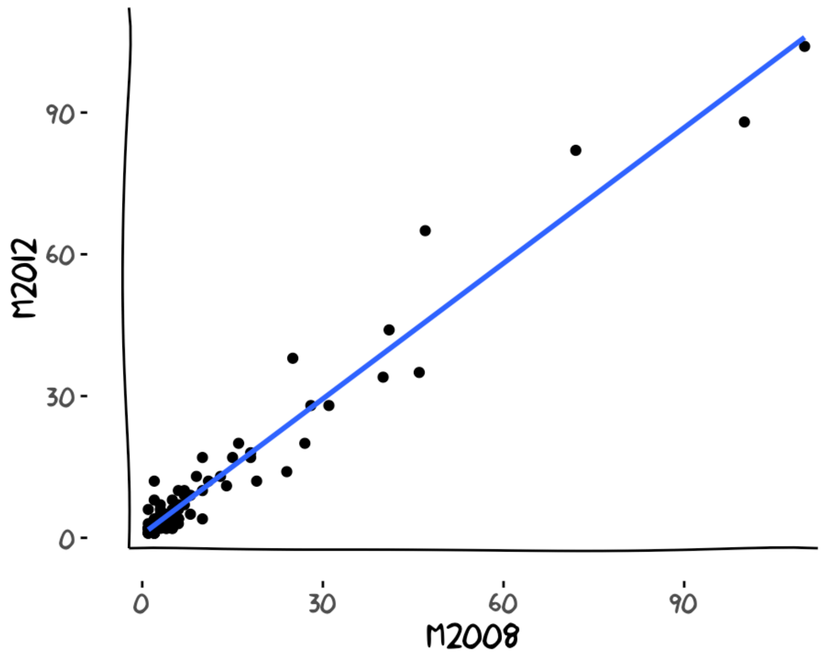
term	estimate	std.error	statistic	p.value
(Intercept)	0.74	0.64	1.1	0.25
M2008	0.96	0.03	35.5	0.00

Giving the model,

$$M_{2012} = 0.74 + 0.96\, M_{2008} + \varepsilon$$

Your turn

- Should the model be re-fit with the intercept forced to ZERO?



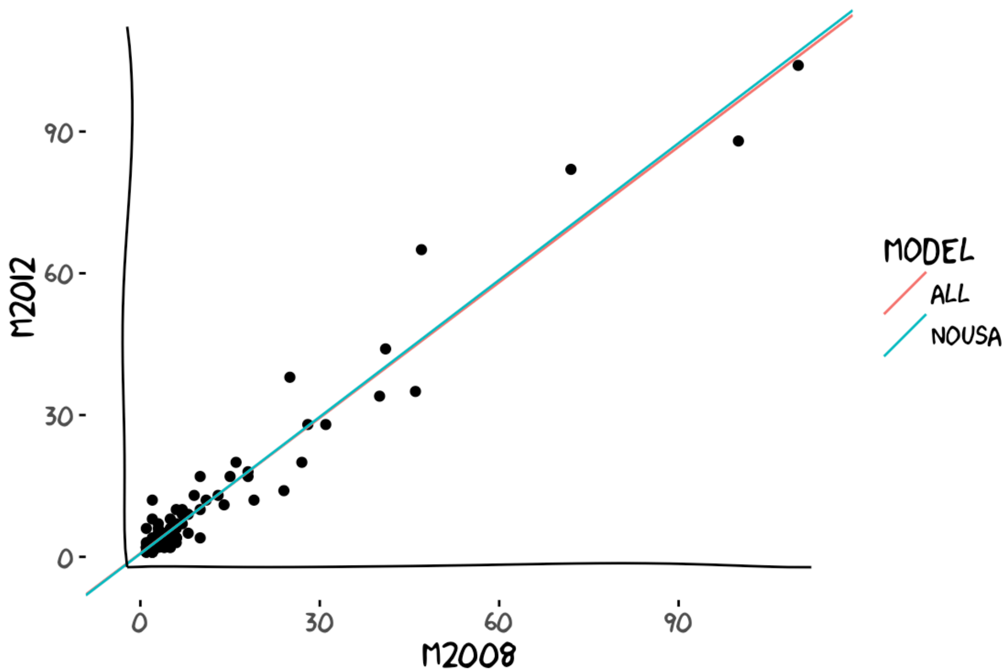
M2012
90
60
30
0
0
30
60
90
M2008

Model diagnostics

- Based on `leave-one-out` statistics
- For n observations, fit n models where each model has one observation removed.
- Let's take a look at fitting the medal tallies, without the USA.

	all	noUSA	estimate
1	0.74	0.66	intercept
2	0.96	0.97	slope

- Parameter estimates change a little

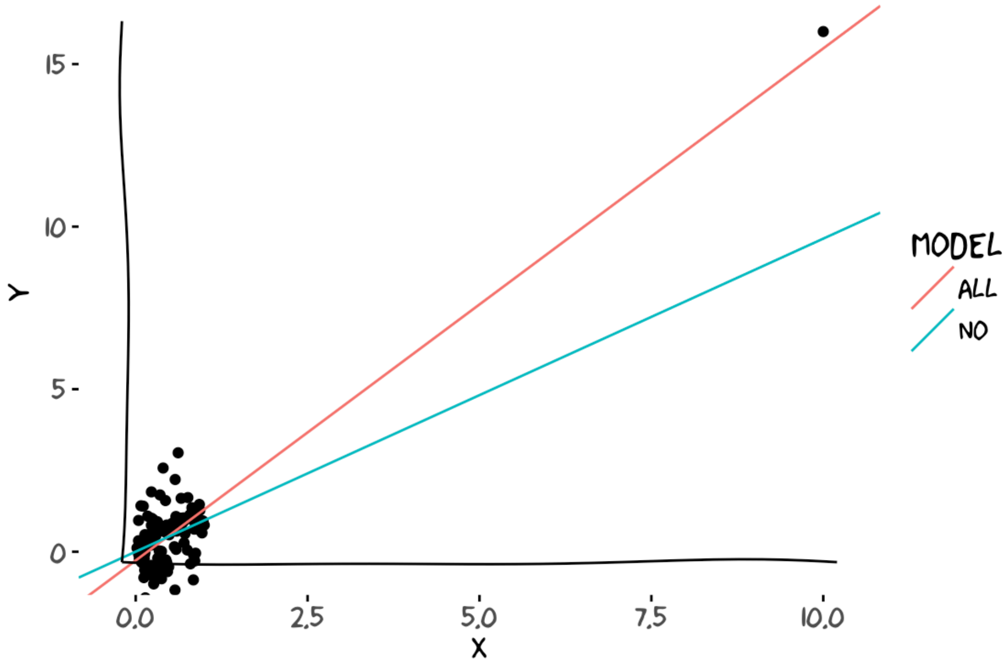

M2012
M2008
90
60
30
0
0
30
60
90
MODEL
ALL
NOUSA

Other model fit parameters

- deviance
- predicted values, residuals

	null.dev	deviance	fitted	resid
All	28811	1533	106	-1.9
No USA	20412	1528	107	-2.9

What it could look like



Leverage

Leverage h_{ii} is defined for each observation, $1, \ldots, n$, and is the i^{th} diagonal element of the hat matrix:

$$H = X(X^T X)^{-1} X^T$$

where X is the design matrix, e.g. for $\beta_0 + \beta_1 x$,

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Intuitively, observations which are far from the mean of the explanatory variables will have higher leverage.

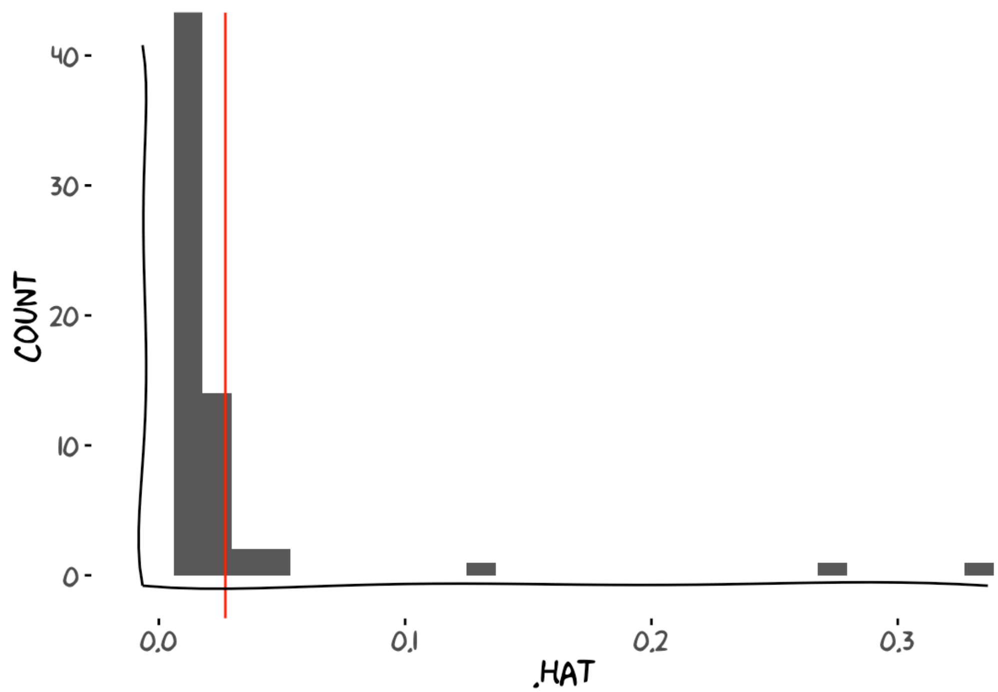
YOU CAN CALCULATE THIS WITHOUT FITTING ALL n MODELS!

Highest leverage for medal tally model

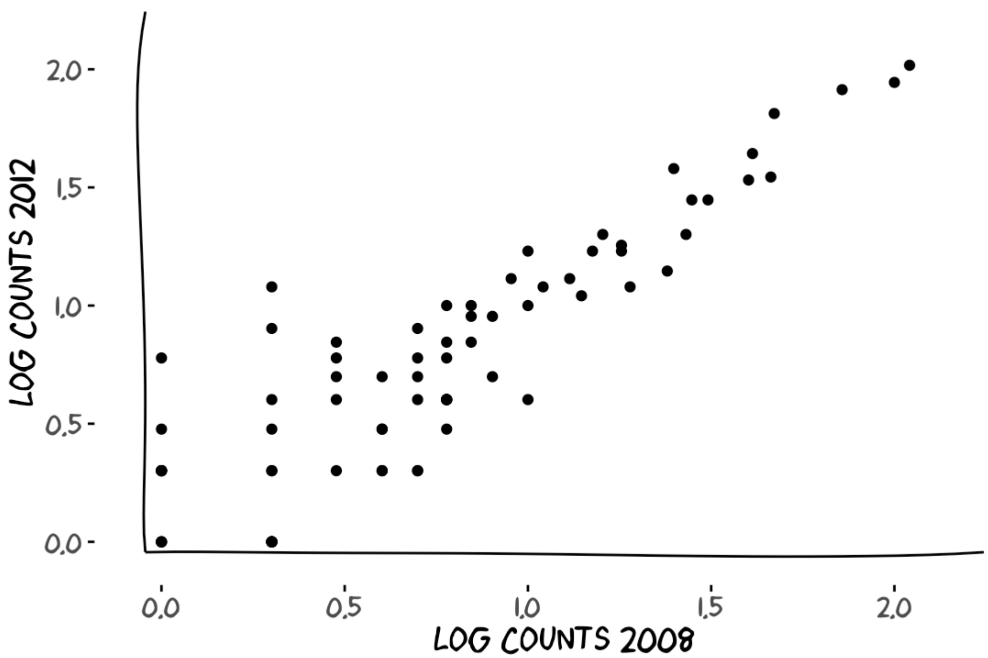
```
#>           Country  .hat
#> 1   UnitedStates 0.330
#> 2          China 0.268
#> 3     RussianFed 0.131
#> 4   GreatBritain 0.053
#> 5      Australia 0.051
#> 6        Germany 0.040
#> 7         France 0.038
#> 8          Korea 0.025
#> 9          Italy 0.021
#> 10       Ukraine 0.020
#> 11         Japan 0.019
#> 12       Bahrain 0.018
#> 13         Egypt 0.018
#> 14      Malaysia 0.018
#> 15 Rep.ofMoldova 0.018
```

Cutoff for high leverage is $2p/n = 2 * 1/73 = 0.027$.

Plot of leverage

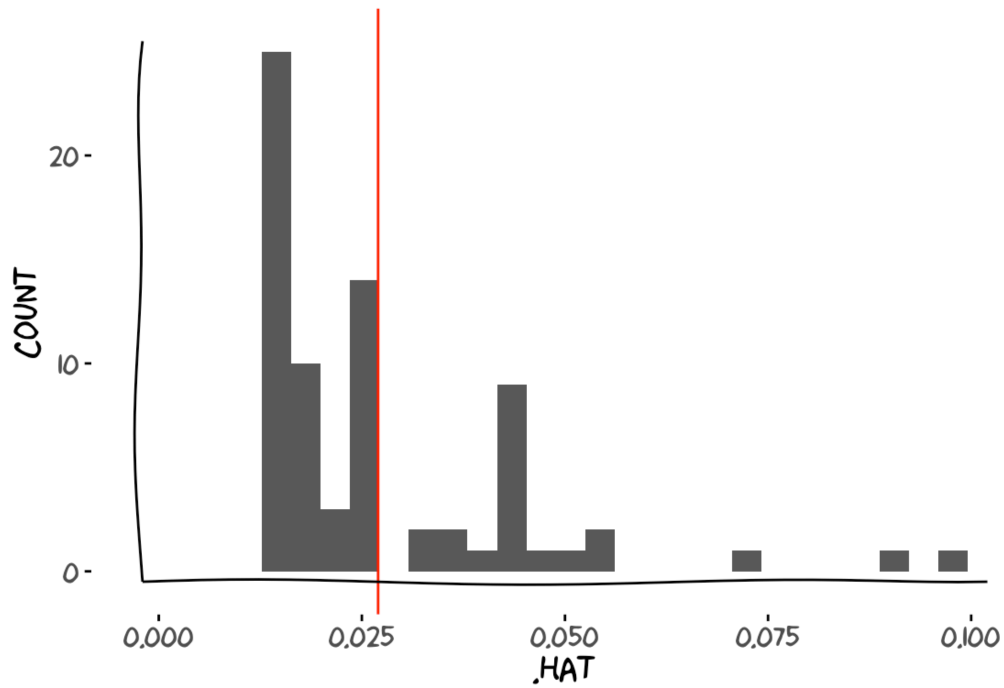


Log-tranform the counts



13/33

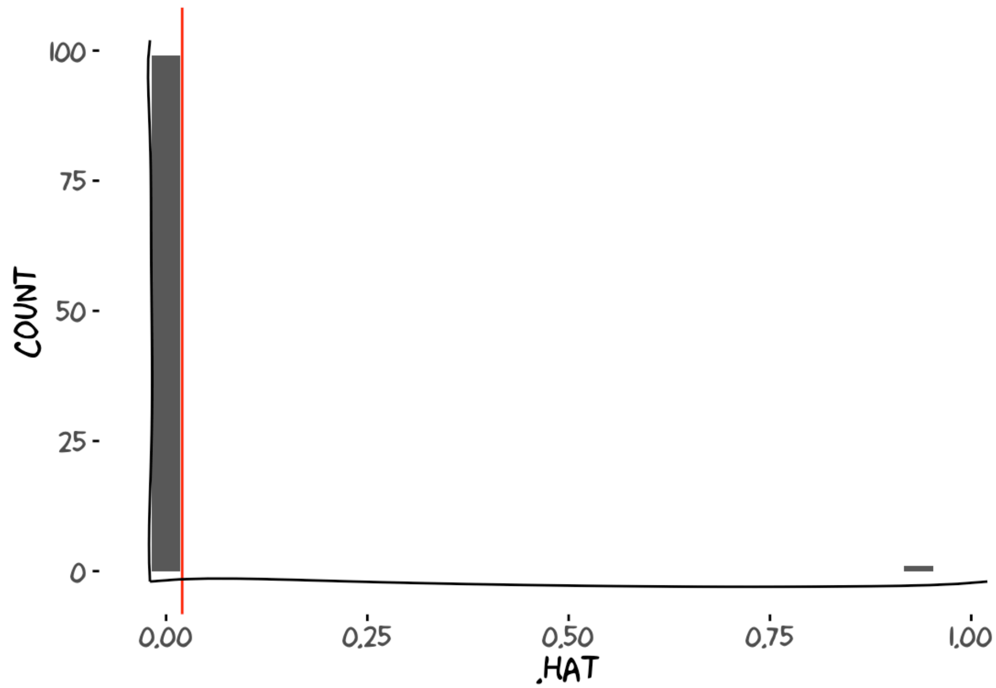
```
#>          Country  .hat
#> 1   UnitedStates 0.096
#> 2          China 0.091
#> 3     RussianFed 0.074
#> 4   GreatBritain 0.055
#> 5      Australia 0.054
#> 6        Germany 0.050
#> 7         France 0.049
#> 8    Afghanistan 0.044
#> 9        Bahrain 0.044
#> 10         Egypt 0.044
#> 11      Malaysia 0.044
#> 12 Rep.ofMoldova 0.044
#> 13     Singapore 0.044
#> 14   SouthAfrica 0.044
#> 15       Tunisia 0.044
```



Transforming skewed variables reduces the influence of any one, or few points. The distribution is more even, and the highest leverage value is much lower now.

15/33

Hat values for simulated data



Cooks D

Leverage takes no notice of the response variable. So the USA did not have a huge influence because its medal count in 2012 was similar to that in 2008, so it was close to the trend. If for some reason the medal count in 2012 was 0, the line with the USA would be much more drawn away from the other countries.

Cooks D, and DFFITS, also use the response variable, to assess influence.

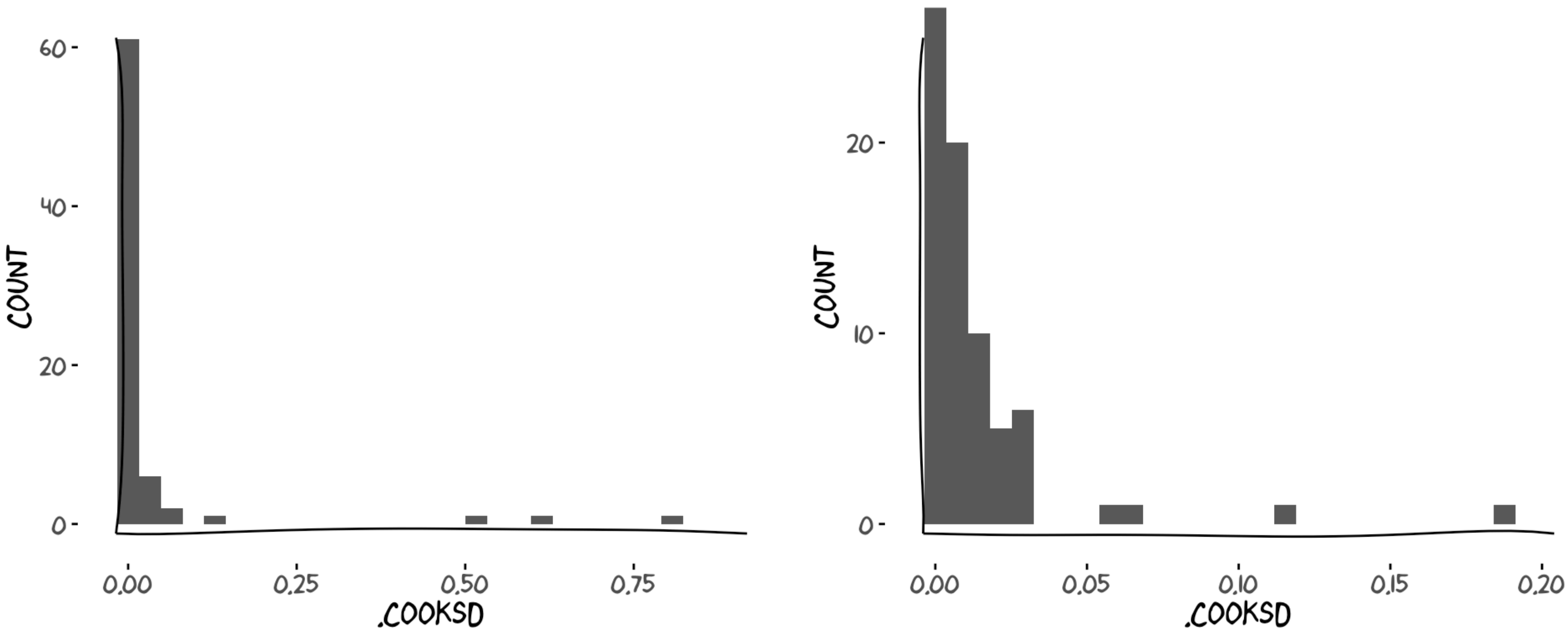
$$D_i = \frac{e_i^2}{MSE^2 p} \frac{h_{ii}}{(1-h_{ii})^2}$$

where e_i is the i^{th} residual, p =number of explanatory variables, and MSE is the mean squared error of the linear model.

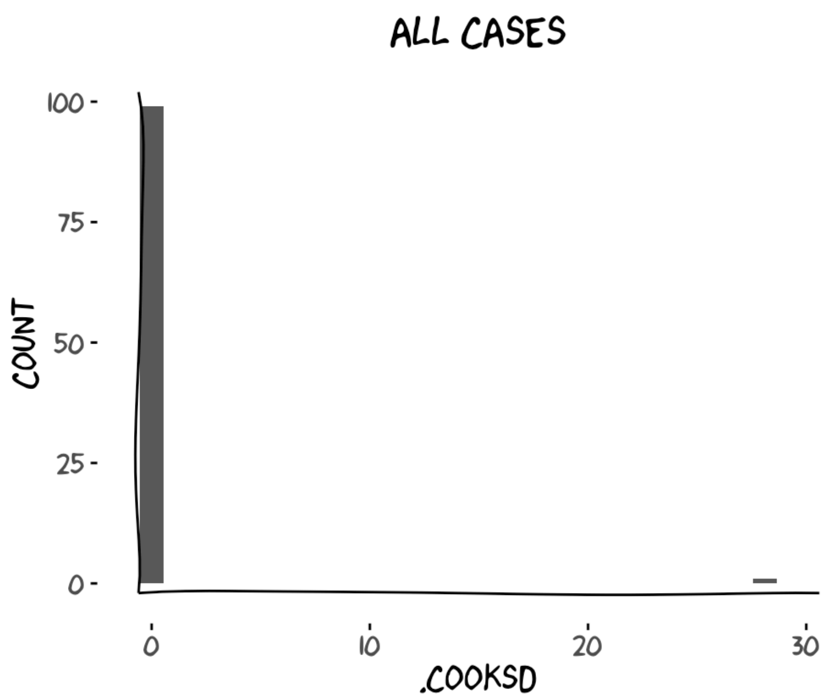
Values greater than $4/n$ are large, by a rule of thumb. Or alternatively, greater than 1 is another rule of thumb.

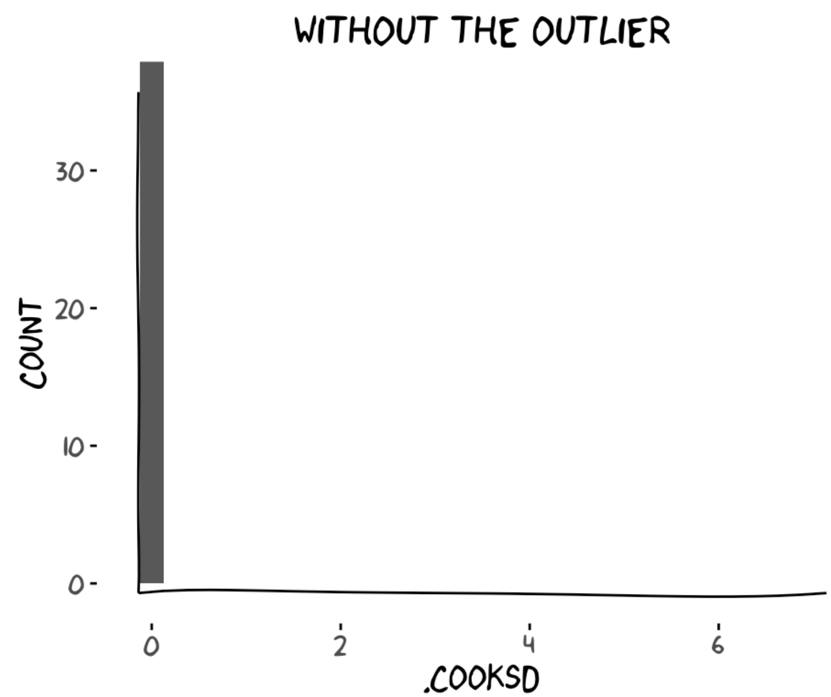
Cooks D for Olympic medal tally

Raw		*Transformed*	
Country	.cooksd	Country	.cooksd
China	0.81	SouthAfrica	0.18
RussianFed	0.62	Iran	0.11
GreatBritain	0.51	Colombia	0.06
Australia	0.12	Tunisia	0.06
Japan	0.08	Algeria	0.03
UnitedStates	0.06	Bahamas	0.03
Cuba	0.04	Morocco	0.03
Iran	0.04	Portugal	0.03


60
40
20
0
COUNT
0.00
0.25
0.50
0.75
.COOKSD
20
10
0
COUNT
0.00
0.05
0.10
0.15
0.20
.COOKSD

Cooks D for simulated data





Values are more spread, when the one extreme value is removed. No other points are influential.

Solutions

- Remove influential observations, and re-fit model
- Transform explanatory variables to reduce influence
- Use weighted regression to downweight influence of extreme observations

Your turn

What happens when there are two extreme points with virtually the same values?

22/33

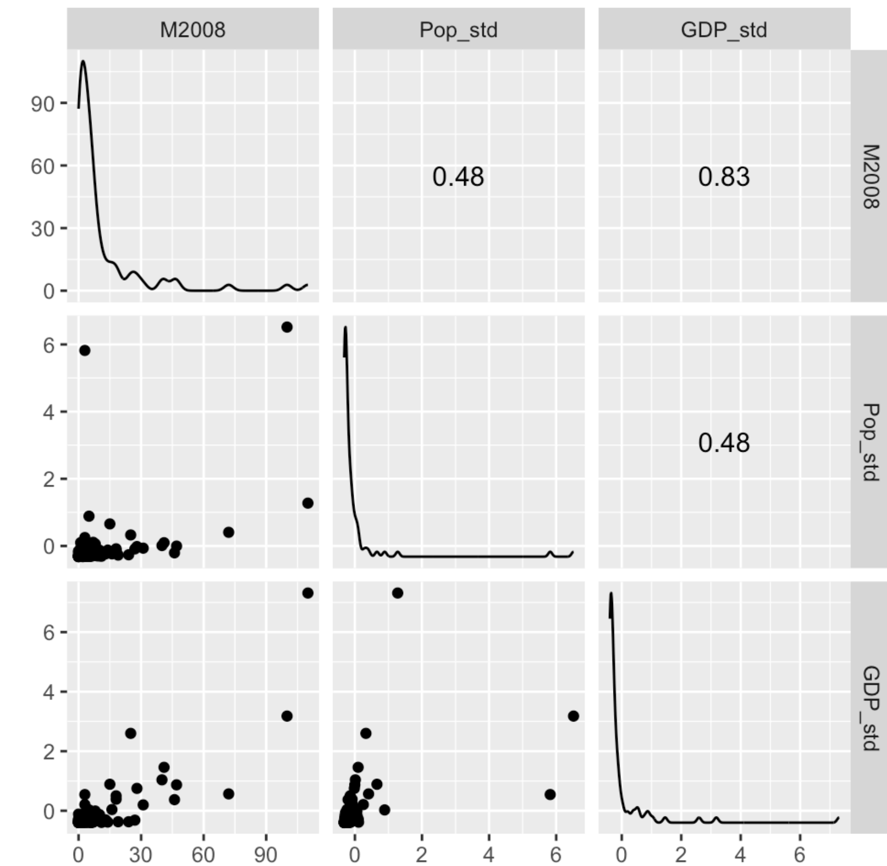
Collinearity

Population and GDP are standardised.

term	estimate	std.error	statistic	p.value
(Intercept)	1.28	0.67	1.91	0.06
M2008	0.91	0.04	20.80	0.00
Pop_std	-0.51	0.54	-0.94	0.35
GDP_std	1.27	0.85	1.50	0.14

Giving the model $M2012 =$ 1.28 + 0.91 $M2008+$ -0.51 $Pop_{std}+$ 1.27 GDP_{std} $+ \varepsilon$

Plot the explanatory variables



Explore countries

Variance inflation factor (VIF)

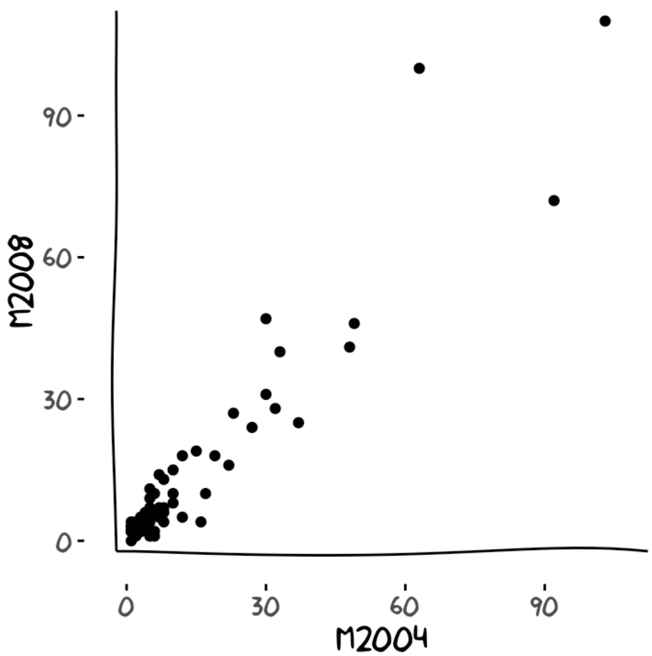
$$\frac{1}{1 - R_j^2}$$

where R_j^2 is computed by regressing variable j on all other variables. VIF is a measure the collinearity of the explanatory variables. Values greater than 10 are considered to be high.

These are the VIFs for the olympic medal tally data:

```
#>   M2008 Pop_std GDP_std
#>     3.3     1.3     3.3
```

Suppose we add 2004 counts as an explanatory variable



Model

term	estimate	std.error	statistic	p.value
(Intercept)	1.12	0.90	1.25	0.22
M2008	0.73	0.11	6.75	0.00
M2004	0.21	0.10	1.98	0.05
Pop_std	0.05	0.67	0.07	0.95
GDP_std	1.05	0.98	1.07	0.29

Giving the model $M2012 =$ 1.12 + 0.73 $M2008+$ 0.21 $M2004+$ 0.05 $Pop_{std}+$ 1.05 GDP_{std} $+ \varepsilon$

VIFs

```
#>   M2008   M2004 Pop_std GDP_std
#>    14.0    11.5     1.6     3.2
```

Notice that the VIFs for both 2004 and 2008 are high.

Your turn

- Why is it called `Variance Inflation Factor`? Look at the standard deviation of the estimates for the model with 2004 and without 2004.
- Why would multicollinearity inflate variance of estimates?

Solutions

- Drop some variables
- Use principal component regression (more advanced courses)
- Partial regression: Fit best variable. Regress next explanatory variable first variable and use the residuals from this fit as the second variable in the model. Continue with other variables.

Resources

- Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

Share and share alike