Statistical Methods for Insurance: Linear Models

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

W5.C2

Fit all possible models

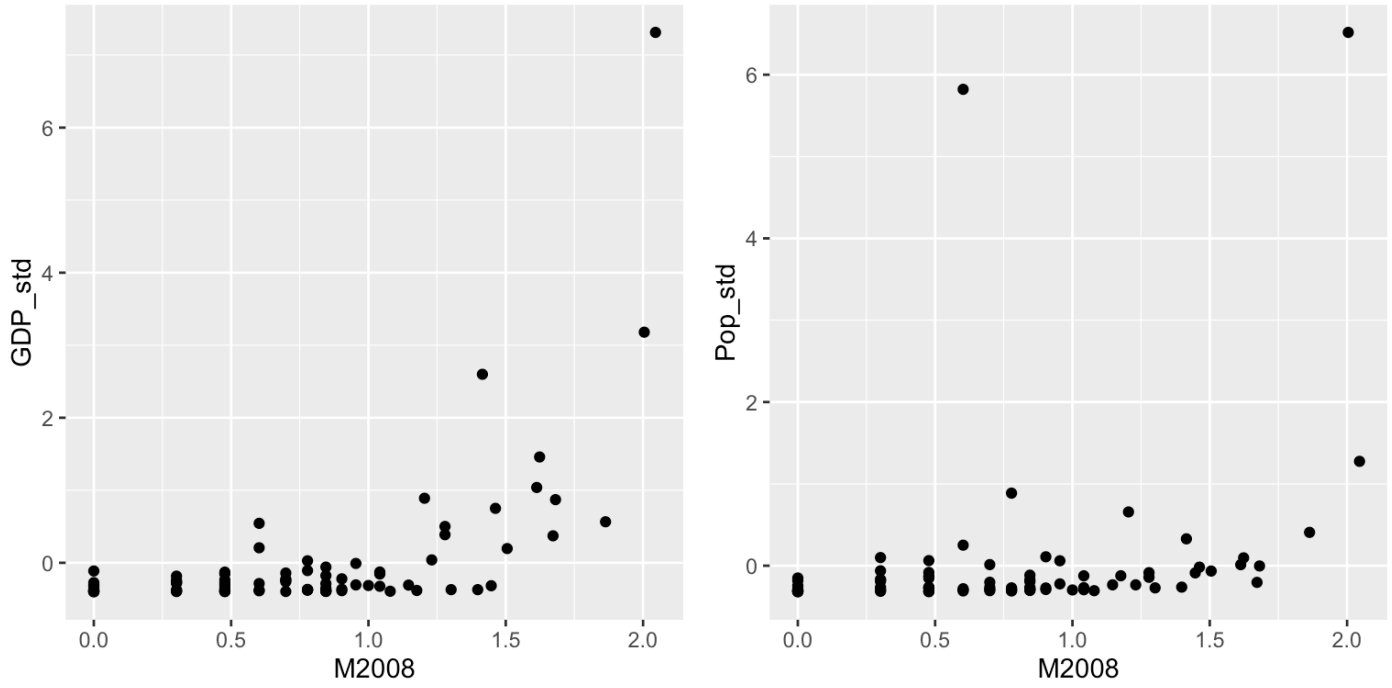
Olympic medal tallies:

- Response: M2012
- Explanatory variables: M2008, M2004, Pop, GDP
- Pop and GDP are standardised; Medal counts are on the log10 scale

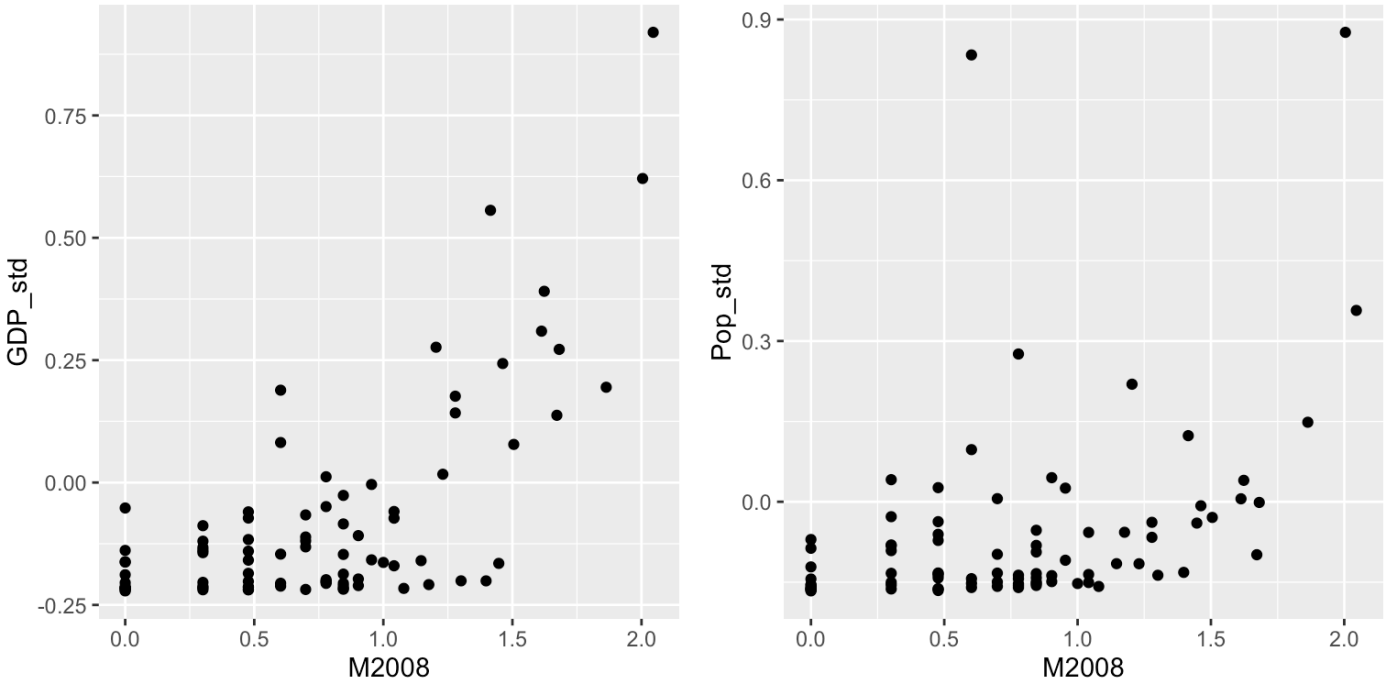
Your turn

What are the possible models? How many are there?

Quick check



Need to log transform GDP and Pop, too.


GDP_std
0.75
0.50
0.25
0.00
-0.25
0.0
0.5
1.0
1.5
2.0
M2008
Pop_std
0.9
0.6
0.3
0.0
0.0
0.5
1.0
1.5
2.0
M2008

5/31

Fit them - first one

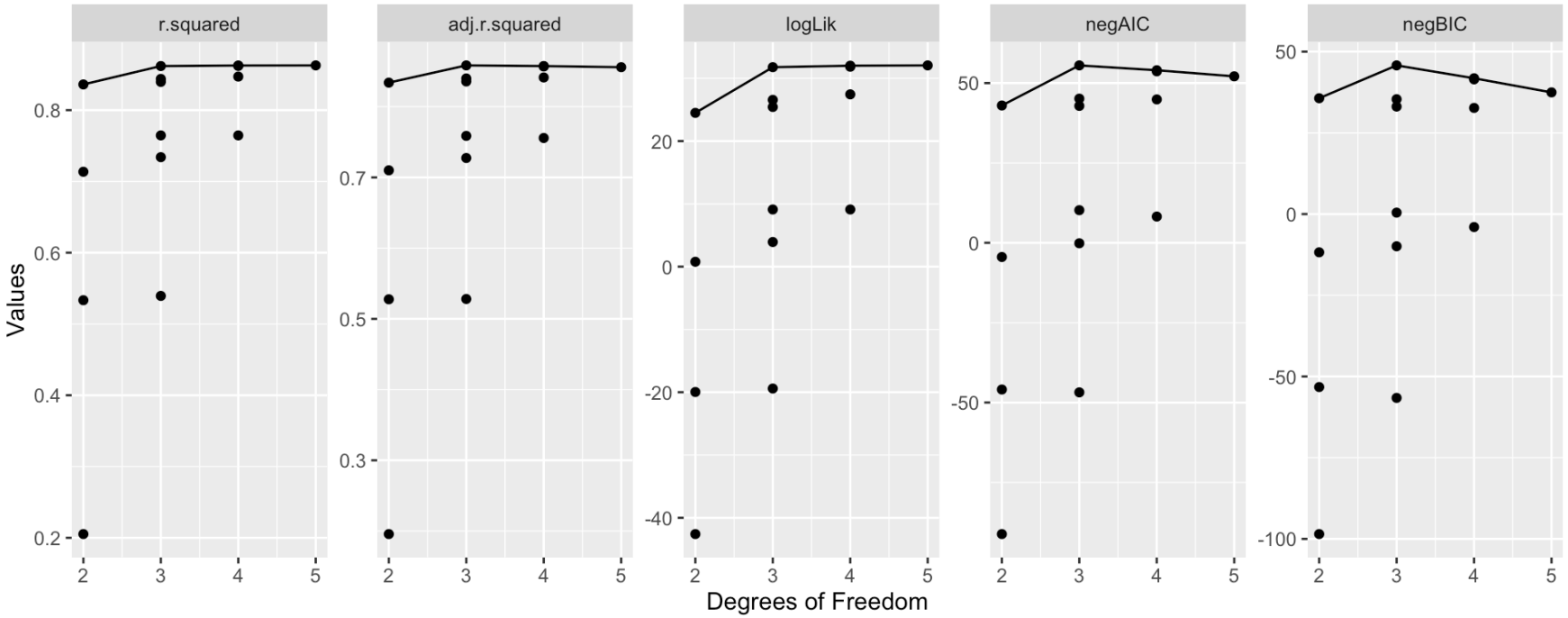
```
#>
#> Call:
#> lm(formula = y ~ M2008, data = data, model = FALSE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.3546 -0.1460  0.0301  0.0967  0.5249
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.2043     0.0356    5.74  1.5e-07 ***
#> M2008         0.8062     0.0392   20.57  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.18 on 83 degrees of freedom
#> Multiple R-squared:  0.836,  Adjusted R-squared:  0.834
#> F-statistic:  423 on 1 and 83 DF,  p-value: <2e-16
```

Second one

```
#>
#> Call:
#> lm(formula = y ~ M2004, data = data, model = FALSE)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.7013 -0.1153 -0.0075  0.1574  0.6647
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.3353     0.0423    7.92  9.3e-12 ***
#> M2004         0.6853     0.0477   14.38  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.24 on 83 degrees of freedom
#> Multiple R-squared:  0.714,  Adjusted R-squared:  0.71
#> F-statistic:  207 on 1 and 83 DF,  p-value: <2e-16
```

Model fit summaries

- Extract the model fit statistics, adjusted R^2, R^2, AIC, BIC, for each model, and display these against the degrees of freedom
- Flip AIC and BIC, so that the direction matches other statistics
- Helps choose best model
- How different is the best model from the next best model


r.squared
adj.r.squared
logLik
negAIC
negBIC
Values
Degrees of Freedom
0.8
0.6
0.4
0.2
0.7
0.5
0.3
20
0
-20
-40
50
0
-50
-100
2
3
4
5

9/31

What do we learn?

- Not a lot of improvement by adding more variables to the model with one variable
- Maybe gain a small amount with two variables
- There is a big difference between the best and worst model

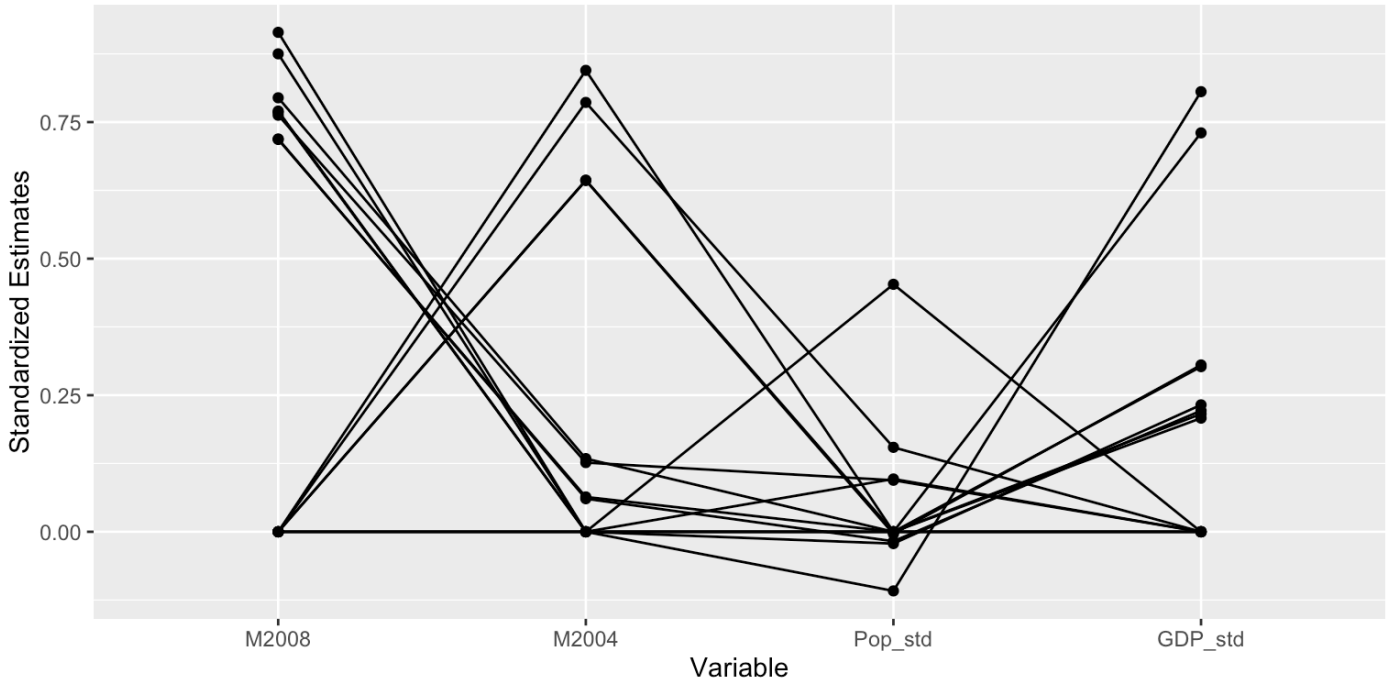
Interactive

What do we learn

- The best single variable model uses M2008. The next best single variable model uses M2004.
- The best two variable model uses M2008 and GDP, but there is very little difference between the next two models (M2008, Pop; M2008, M2004).

Coefficients for each model

- Extract the parameter estimates for each model
- Plot these against the explanatory variable name
- Connect values corresponding to the same model with lines


Standardized Estimates
0.75
0.50
0.25
0.00
M2008
M2004
Pop_std
GDP_std
Variable

16/31

What do we learn

- If M2008 is in the model, its coefficient is big and coefficients for other variables are really small.
- M2004 has a high coefficient, if M2008 is not in the model - it substitutes for M2008.
- GDP and Pop only have high coefficients when M2008 and M2004 are not in the model.

Balance complexity vs accuracy

- Choose the simplest model that explains almost as much as a more complex model.
- Choosing here between a single variable model using M2008, and adding GDP.

Model A

```
#>
#> Call:
#> lm(formula = M2012 ~ M2008, data = oly_gdp2012_tf)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.3546 -0.1460  0.0301  0.0967  0.5249
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.2043     0.0356    5.74  1.5e-07 ***
#> M2008         0.8062     0.0392   20.57  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.18 on 83 degrees of freedom
#> Multiple R-squared:  0.836,  Adjusted R-squared:  0.834
#> F-statistic:  423 on 1 and 83 DF,  p-value: <2e-16
```

Model B

```
#>
#> Call:
#> lm(formula = M2012 ~ M2008 + GDP_std, data = oly_gdp2012_tf)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -0.3925 -0.1240  0.0259  0.0894  0.4878
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   0.3357     0.0470    7.14  3.4e-10 ***
#> M2008         0.6791     0.0486   13.96  < 2e-16 ***
#> GDP_std       0.4632     0.1184    3.91  0.00019 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.17 on 82 degrees of freedom
#> Multiple R-squared:  0.862,  Adjusted R-squared:  0.858
#> F-statistic:  256 on 2 and 82 DF,  p-value: <2e-16
```

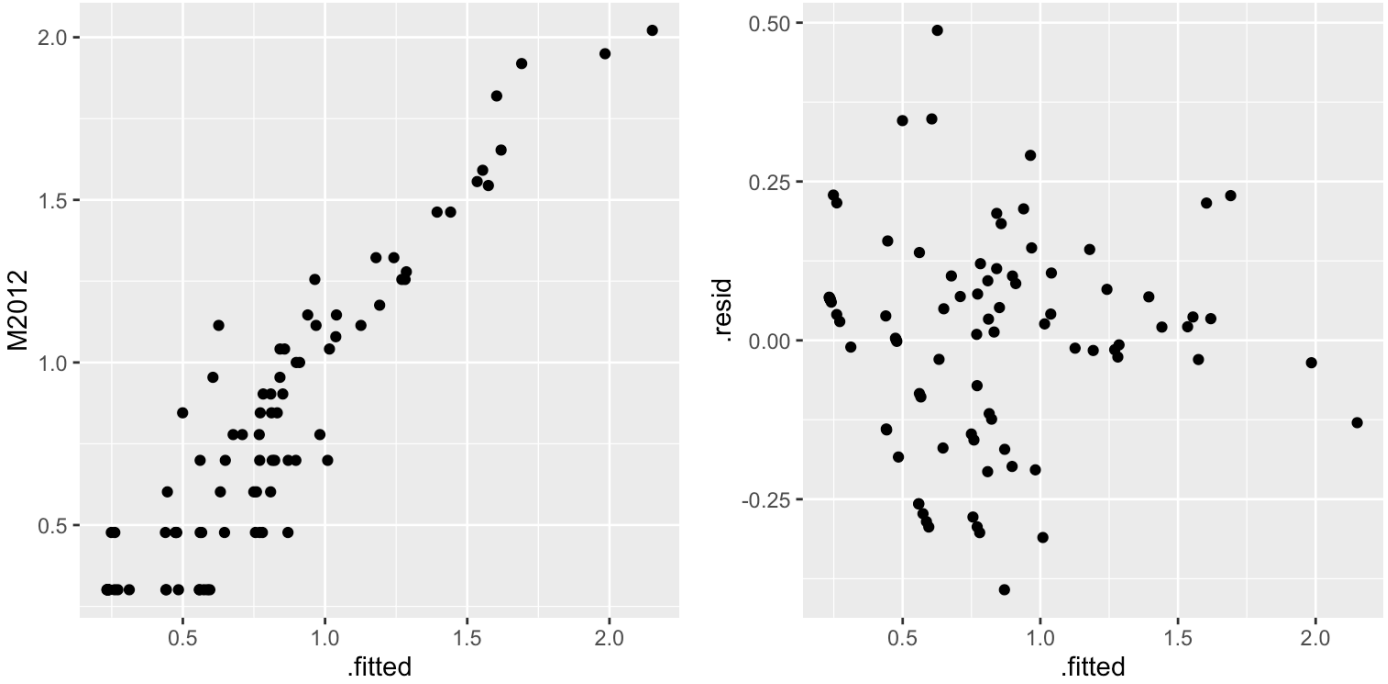
Model comparison

```
#> Analysis of Variance Table
#>
#> Model 1: M2012 ~ M2008
#> Model 2: M2012 ~ M2008 + GDP_std
#>   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
#> 1     83 2.80
#> 2     82 2.36  1      0.44 15.3 0.00019 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

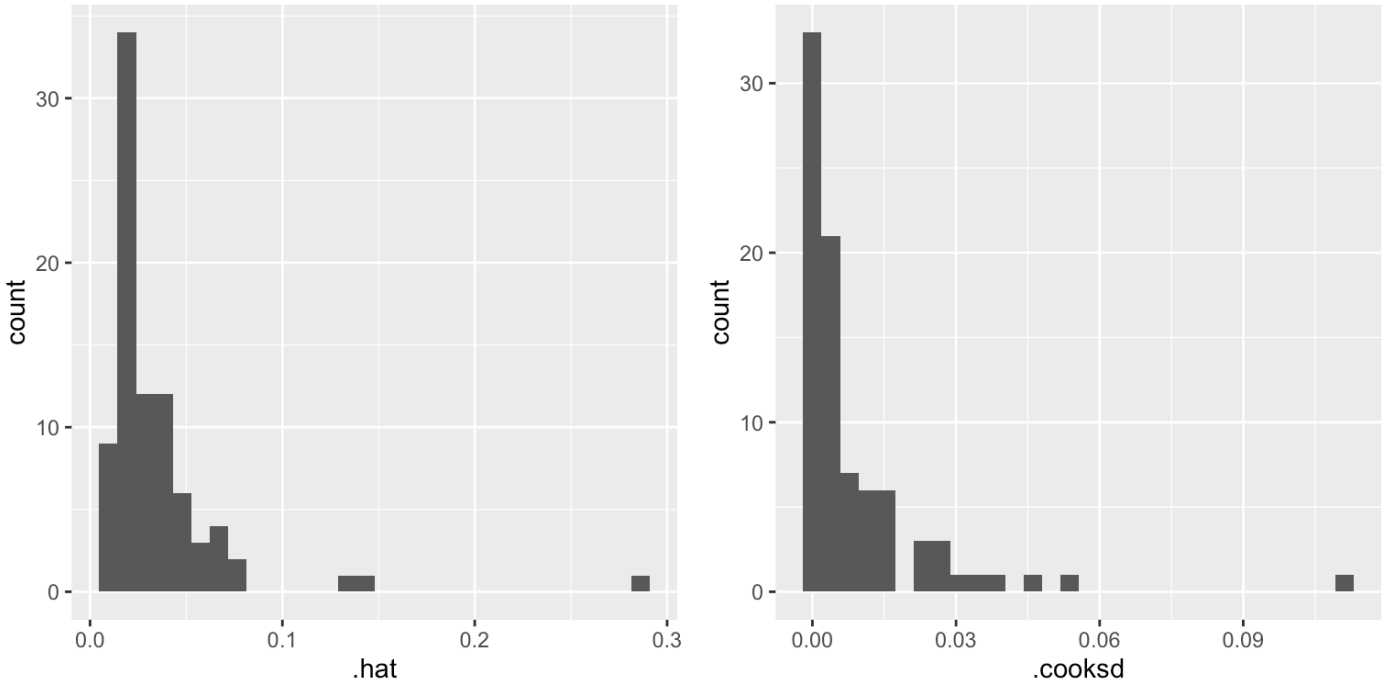
Your turn

How much more explanatory power do we get by including GDP?

Diagnostics



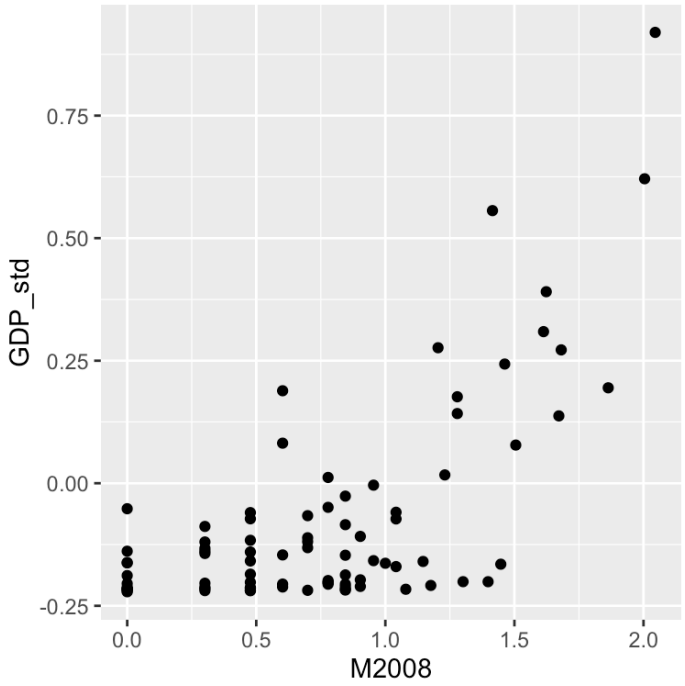
Diagnostics



Hmm, do we have a problem with one observation?

VIFs

```
#>   M2008 GDP_std
#>     1.8     1.8
```



27/31

Final model

- $log10(M2012 + 1) = 0.34 + 0.68\ log10(M2008 + 1) + 0.46\ log10(GDP_{std} + 1) + \varepsilon$
- Predict the medal count for Australia $M2008 = 46$ and $GDP = 1507000000000$ ($GDP_{std} = 0.3720791$).
- $0.34 + 0.68\ log10(46 + 1) + 0.46\ log10(0.3720791 + 1) = 1.53$
- The observed count for 2012 was 35 medals. What will the estimated medal count be?

??

Resources

- Statistics online textbook, Diez, Barr, Cetinkaya-Rundel

Share and share alike

This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/ 3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.