Multicollinearity

Adam J Sullivan, PhD
02/14/2018

Multicollinearity

What is Multicollinearity

- Multicollinearity exists when 2 or more covariates in a model are moderately or highly correlated.
- This may be viewed as an easy issue to deal with as many things we may want to control for are just highly correlated.
- For example, education and income are highly correlated.

Types of Multicollinearity

- Data based:
  - Could be poorly designed study
  - observational data where only variables collected are all correlated.
- Structural:
  - Duplicate variables so they are mathematically the same.
  - Variables that were created from others
    - For example, weight and height are highly correlated with BMI.

Consider the following data:

- This data has been simulated so that it is not collinear:

```
##            response predictor1 predictor2
## response      1.000        0.8      0.202
## predictor1    0.800        1.0      0.000
## predictor2    0.202        0.0      1.000
```

- Let's look at the regressions

Regression on Uncorrelated Data

- We will consider the following regressions:

$$\text{Model 1: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_1$$

$$\text{Model 2: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_2$$

$$\text{Model 3: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_1 + + \hat{\beta}_2 Predictor_2$$

$$\text{Model 4: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_2 + \hat{\beta}_2 Predictor_1$$

Regression on Uncorrelated Data

term	estimate	p.value	conf.low	conf.high
predictor1	0.8003296	0	0.7966066	0.8040526
predictor2	0.2016739	0	0.1956074	0.2077403
predictor1	0.8004000	0	0.7968944	0.8039057
predictor2	0.2019523	0	0.1984514	0.2054533
predictor2	0.2019523	0	0.1984514	0.2054533
predictor1	0.8004000	0	0.7968944	0.8039057

Regression on Uncorrelated Data

r.squared	adj.r.squared	sigma	statistic	p.value
0.6396798	0.6396762	0.6008585	177527.401	0
0.0407269	0.0407173	0.9803905	4245.513	0
0.6805193	0.6805129	0.5657864	106500.764	0
0.6805193	0.6805129	0.5657864	106500.764	0

Sum Squares of Models

term	df	sumsq	meansq
predictor1	1	64092.895	64092.895
predictor2	1	4080.641	4080.641
predictor1	1	64092.895	64092.895
predictor2	1	4091.918	4091.918
predictor2	1	4080.641	4080.641
predictor1	1	64104.172	64104.172

What Do We Notice?

- Coefficients do not change in models.
- Sums of Squares added to model remain consistent

Consider the following data:

- This data has been simulated so that it is highly collinear:

```
##            response predictor1 predictor2
## response      1.000      0.846      0.188
## predictor1    0.846      1.000      0.639
## predictor2    0.188      0.639      1.000
```

- Let's look at the regressions

Regression on Correlated Data

- We will consider the following regressions:

$$\text{Model 1: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_1$$

$$\text{Model 2: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_2$$

$$\text{Model 3: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_1 + +\hat{\beta}_2 Predictor_2$$

$$\text{Model 4: } Response = \hat{\beta}_0 + \hat{\beta}_1 Predictor_2 + +\hat{\beta}_2 Predictor_1$$

Regression on Correlated Data

term	estimate	p.value	conf.low	conf.high
predictor1	0.9289612	0.0000000	0.8117562	1.0461662
predictor2	0.2349541	0.0606592	-0.0107236	0.4806318
predictor1	1.3475785	0.0000000	1.2697073	1.4254497
predictor2	-0.7444484	0.0000000	-0.8329782	-0.6559187
predictor2	-0.7444484	0.0000000	-0.8329782	-0.6559187
predictor1	1.3475785	0.0000000	1.2697073	1.4254497

Regression on Correlated Data

r.squared	adj.r.squared	sigma	statistic	p.value
0.7162667	0.7133715	0.5868151	247.394788	0.0000000
0.0354504	0.0256081	1.0819531	3.601826	0.0606592
0.9267136	0.9252025	0.2997678	613.287170	0.0000000
0.9267136	0.9252025	0.2997678	613.287170	0.0000000

Sum Squares of Models

term	df	sumsq	meansq
predictor1	1	85.190885	85.190885
predictor2	1	4.216378	4.216378
predictor1	1	85.190885	85.190885
predictor2	1	25.030002	25.030002
predictor2	1	4.216378	4.216378
predictor1	1	106.004508	106.004508

What did we Notice?

- Coefficients change a lot
- Sum of Squares Depends on the order in which data is in the model.

Signs of Multicollinearity

- Estimates of the coefficients vary from model to model.
- t-tests of individual slopes are non-significant but overall F-test is significant.
- Correlations among covariates are large.

How Can we Detect this?

- Consider the model with just one covariate:

$$y_i = \beta_0 + \beta_k x_{ik} + \varepsilon_i$$

- We can see this variance:

$$Var(b_k)_{min} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}$$

- This is the smallest variance will be.

Then the larger model

- Consider the model with just one covariate:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots \beta_p x_{ip} \varepsilon_i$$

- We can see this variance:

$$Var(b_k) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2} \times \frac{1}{1 - R_k^2}$$

- R_k^2 is the R^2 value of the kth predictor on the remaining.

What does this tell us?

- How much is our variance inflated by?

$$\frac{Var(b_k)}{Var(b_k)_{min}} = \frac{1}{1 - R_k^2}$$

- Variance Inflation Factor

$$VIF_k = \frac{1}{1 - R_k^2}$$

Variance Inflation Factor

- Rule of thumb
  - 1 = not correlated.
  - Between 1 and 5 = moderately correlated.
  - Greater than 5 = highly correlated.
- Some suggest anything more than 2.5 should cause concern and definitely over 10.

Variance Inflation Factor

- Be careful just judging by it alone
- For example x and x^2 may have a high VIF but this would not hurt your model.
- Also Indicator variables often have a high VIF with each other but this is not an issue.

Calculating in R

```
library(car)
vif1 <- vif(mod3)
vif2 <- vif(mod4)
knitr::kable(bind_rows(vif1,vif2))
```

Calculating in R

	predictor1	predictor2
	1	1
	1	1

Calculating in R

	predictor1	predictor2
	1.691149	1.691149
	1.691149	1.691149

How Can We Deal with it?

- Remove Multicollinear variables from model.
  - What might the effects of this be?
- Create a summed score of the collinear variables.
- Create a score based on something like Principal Component Analysis.