Collinearity

Dr. Michael Fix
mfix@gsu.edu

Georgia State University

25 March 2025

Under the Hood of $\mathbf{X}$

OLS (and regression methods more generally) requires:

- $\mathbf{X}$ is full column rank.
- $N > K$.
- "Sufficient" variability in $\mathbf{X}$.

"Perfect" Multicollinearity

First a formal definition:
There cannot be any set of λs such that:

$$\lambda_0 \mathbf{1} + \lambda_1 \mathbf{X}_1 + \ldots + \lambda_K \mathbf{X}_K = \mathbf{0}$$

A Toy Model

Let's see if there is a relationship between gas milage and car performance.

```
> data("mtcars")
> model1 <- lm(qsec ~ mpg, mtcars)
> summary(model1)

Call:
lm(formula = qsec ~ mpg, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8161 -1.0287  0.0954  0.8623  4.7149 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15.35477    1.02978  14.911 2.05e-15 ***
mpg          0.12414    0.04916   2.525   0.0171 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.65 on 30 degrees of freedom
Multiple R-squared:  0.1753,Adjusted R-squared:  0.1478 
F-statistic: 6.377 on 1 and 30 DF,  p-value: 0.01708
```

A Toy Model

Now let's redo that using Kilograms/Liter instead of Miles/Gallon, but accidentally include both measures as predictor variables. What happens?

```
> mtcars$kgL <- mtcars$mpg * .425
> model2 <- lm(qsec ~ mpg + kgL, mtcars)
> summary(model2)

Call:
lm(formula = qsec ~ mpg + kgL, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8161 -1.0287  0.0954  0.8623  4.7149

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.35477    1.02978  14.911 2.05e-15 ***
mpg          0.12414    0.04916   2.525   0.0171 *
kgL               NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.65 on 30 degrees of freedom
Multiple R-squared:  0.1753,Adjusted R-squared:  0.1478
F-statistic: 6.377 on 1 and 30 DF,  p-value: 0.01708
```

What Does This Tell Us?

1. Perfect Multicollinearity is a very big problem (Theoretically)
2. Prefect Multicollinearity is NOT a problem at all (In Practice)

$N > K$

- Statistically, if $N < K$, then:
  - We lack sufficient degrees of freedom to identify $\hat{\beta}$.*
  - $\hat{\beta}$ is "overdetermined."
- Conceptually, $N < K$ means that:
  - Our number of variables > Cases
  - Which means there can be no unique conclusion about explanatory / causal factors.

*Note: "identification" is used in statistics and econometrics to mean several different things, I am using it here in the most basic sense to mean that the parameters (here the $\hat{\beta}$s) cannot be determined from the variables

Another Toy Model

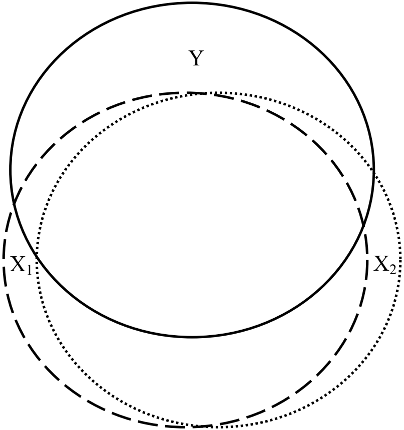
Let's subset the `mtcars` data to only look at lightweight cars and add some more predictor variables:

```
> rm(list=ls())
> data("mtcars")
> lightweight <- subset(mtcars, wt<2)
> model3 <- with(lightweight, lm(qsec ~ mpg + disp + hp))
> summary(model3)

Call:
lm(formula = qsec ~ mpg + disp + hp)

Residuals:
ALL 4 residuals are 0: no residual degrees of freedom!

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.54944        NaN     NaN      NaN
mpg         -0.14716        NaN     NaN      NaN
disp        -0.25649        NaN     NaN      NaN
hp           0.05502        NaN     NaN      NaN

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA
```

What Does This Tell Us?

As with "perfect" multicollinearity, having $N > K$ will result in a model specification that is impossible to estimate. Thus, you cannot violate this assumption in practice

Intuition



High (Non-Perfect) Multicollinearity

Recall that

$$\widehat{\text{Var}(\hat{\beta})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

We can write the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ as:

$$\frac{1}{(\mathbf{X}_k'\mathbf{X}_k)(1 - \hat{R}_k^2)}$$

where $\hat{R}_k^2$ is the R^2 from the regression of $\mathbf{X}_k$ on all the other variables in $\mathbf{X}$.

High (Non-Perfect) Multicollinearity

Things to understand:

1. Multicollinearity is a *sample problem*.
2. Multicollinearity is a matter of *degree*.

(Near-Perfect) Multicollinearity: Detection

1. *High R^2, but nonsignificant coefficients.*
2. *High pairwise correlations among independent variables.*
3. *High partial correlations among the* $\mathbf{X}$*s.*
4. *VIF and Tolerance.*

VIF / Tolerance

If $\hat{R}_k^2 = 0$, then

$$\widehat{\text{Var}(\hat{\beta}_k)} = \frac{\hat{\sigma}^2}{\mathbf{X}_k'\mathbf{X}_k};$$

So:

$$\text{VIF}_k = \frac{1}{1 - \hat{R}_k^2}$$

$$\text{Tolerance} = \frac{1}{\text{VIF}_k}$$

Rule of Thumb: VIF > 10 is a problem.

What To Do?

Don't:

- **Blindly drop covariates!!!**
- Restrict βs...

Do:

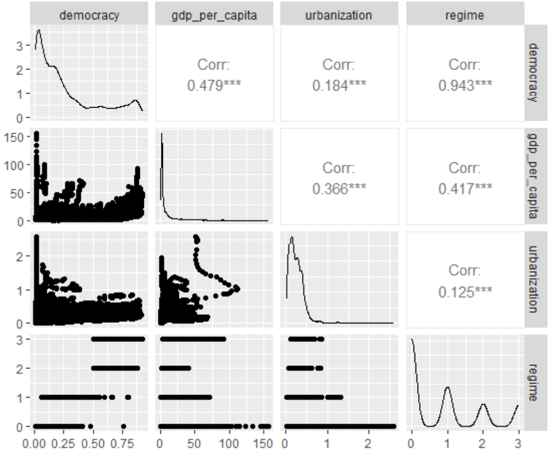
- **Add data**.
- **Transform the covariates**
  - Data reduction
  - First differences
  - Orthogonalize
- **Shrinkage / Regularization Methods**

Toy Model

	Dependent variable:	
	democracy	
	US sample	Full sample
	(1)	(2)
gdp_per_capita	0.008	0.002
	(0.001)	(0.0001)
	t = 15.551	t = 18.264
	p = 0.000***	p = 0.000***
urbanization	0.399	-0.016
	(0.158)	(0.004)
	t = 2.521	t = -3.716
	p = 0.014**	p = 0.0003***
regime	0.090	0.228
	(0.009)	(0.001)
	t = 9.675	t = 234.418
	p = 0.000***	p = 0.000***
Constant	0.161	0.099
	(0.040)	(0.002)
	t = 4.027	t = 58.892
	p = 0.0002***	p = 0.000***
Observations	101	10,810
R2	0.972	0.877
Adjusted R2	0.971	0.877
Residual Std. Error	0.027 (df = 97)	0.095 (df = 10806)
F Statistic	1,128.081*** (df = 3; 97)	25,701.890*** (df = 3; 10806)
Note:		*p<0.1; **p<0.05; ***p<0.01

Correlation Matrix

```r
# Correlation matrix ----
my_data |>
  select(democracy, gdp_per_capita, urbanization, regime) |>
  ggpairs()
```



Correlation

```
# Correlation basics ----

# cor() computes the correlation coefficient
# cor.test() test for association/correlation between paired samples.
# It returns both the correlation coefficient and the significance level
# (or p-value) of the correlation.

# cor.test(x, y, method=c("pearson", "kendall", "spearman"))

# Pearson - normal distribution, continuous
# Spearman - non-parametric, ordinal variables
# Kendall - non-parametric, continuous

# The nice thing about the Spearman correlation is that relies on nearly all
# the same assumptions as the pearson correlation, but it doesn't rely on
# normality, and your data can be ordinal as well.

# The Kendall correlation is similar to the spearman correlation in that it is
# non-parametric. It can be used with ordinal or continuous data.
cor.test(my_data$democracy, my_data$regime,
    use = "complete.obs",
    method = c("pearson"))
```

Correlation

```
> cor.test(my_data$democracy, my_data$regime,
+     use = "complete.obs",
+     method = c("pearson"))

        Pearson's product-moment correlation

data:  my_data$democracy and my_data$regime
t = 391.26, df = 19041, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9414762 0.9446191
sample estimates:
      cor
0.9430687
```

Variance Inflation Factor (VIF)

```
> # Variance Inflation Factor (VIF) ----
> # VIF value starts from 1
> # A value of 1 indicates there is no correlation
> # A value between 1 and 5 indicates moderate correlation
> # A value greater than 5 indicates potentially severe correlation
> vif(us_model)
gdp_per_capita   urbanization         regime
      5.023951       1.633371       6.213308
> vif(my_model)
gdp_per_capita   urbanization         regime
      1.446900       1.131696       1.297502
```

First differences I

```r
# Taking the first difference ----
us_data$diff_regime <- us_data$regime - lag(us_data$regime, n = 1)

# OR in tidy language
us_data <- us_data |>
  mutate(diff_regime = regime - lag(regime, n = 1))
```

First differences II

	Dependent variable: democracy	
	US Sample	US Sample - First difference
	(1)	(2)
gdp_per_capita	0.008 (0.001) p = 0.000 t = 15.551***	0.012 (0.0003) p = 0.000 t = 37.626***
urbanization	0.399 (0.158) p = 0.014 t = 2.521**	1.351 (0.185) p = 0.000 t = 7.313***
regime	0.090 (0.009) p = 0.000 t = 9.675***	
diff_regime		0.007 (0.027) p = 0.810 t = 0.242
Constant	0.161 (0.040) p = 0.0002 t = 4.027***	-0.017 (0.053) p = 0.749 t = -0.322
Observations	101	100
R2	0.972	0.945
Adjusted R2	0.971	0.943
Residual Std. Error	0.027 (df = 97)	0.038 (df = 96)
F Statistic	1,128.081*** (df = 3; 97)	545.046*** (df = 3; 96)
Note:		*p<0.1; **p<0.05; ***p<0.01

First differences II

```
> vif(us_model)
gdp_per_capita   urbanization         regime
      5.023951       1.633371       6.213308
> vif(us_model2)
gdp_per_capita   urbanization    diff_regime
      1.038942       1.038071       1.001096
>
```