

# PLSC 503 – Spring 2017

## Cases and Variables

March 2, 2017

# Under the Hood of **X**

OLS (and regression methods more generally) requires:

- **X** is full column rank.
- $N > K$ .
- “Sufficient” variability in **X**.

# “Perfect” Multicollinearity

Formally: There cannot be any set of  $\lambda$ s such that:

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

If there was, it would imply

$$\mathbf{X}_j = \frac{-\lambda_0}{\lambda_j} \mathbf{1} + \frac{-\lambda_1}{\lambda_j} \mathbf{X}_1 + \dots + \frac{-\lambda_K}{\lambda_j} \mathbf{X}_K$$

which means

$$\begin{aligned} Y &= \beta_0 \mathbf{1} + \beta_1 \mathbf{X}_1 + \dots + \beta_j \mathbf{X}_j + \dots + \beta_K \mathbf{X}_K + \mathbf{u} \\ &= \beta_0 \mathbf{1} + \beta_1 \mathbf{X}_1 + \dots + \beta_j \left( \frac{-\lambda_0}{\lambda_j} \mathbf{1} + \frac{-\lambda_1}{\lambda_j} \mathbf{X}_1 + \dots + \frac{-\lambda_K}{\lambda_j} \mathbf{X}_K \right) + \dots + \beta_K \mathbf{X}_K + \mathbf{u} \\ &= \left[ \beta_0 + \beta_j \left( \frac{-\lambda_0}{\lambda_j} \right) \right] \mathbf{1} + \left[ \beta_1 + \beta_j \left( \frac{-\lambda_1}{\lambda_j} \right) \right] \mathbf{X}_1 + \dots + \left[ \beta_K + \beta_j \left( \frac{-\lambda_K}{\lambda_j} \right) \right] \mathbf{X}_K + \mathbf{u} \\ &= \left( \beta_0 + \frac{\gamma_0}{\lambda_j} \right) \mathbf{1} + \left( \beta_1 + \frac{\gamma_1}{\lambda_j} \right) \mathbf{X}_1 + \dots + \left( \beta_K + \frac{\gamma_K}{\lambda_j} \right) \mathbf{X}_K + \mathbf{u} \end{aligned}$$

# In Practice

```
> Africa$newgdp<-(Africa$gdppppd-mean(Africa$gdppppd))*1000  
  
> fit<-with(Africa, lm(adrate~gdppppd+newgdp+healthexp+subsaharan+  
+ muslperc+literacy))  
> summary(fit)
```

Call:

```
lm(formula = adrate ~ gdppppd + newgdp + healthexp + subsaharan +  
    muslperc + literacy)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-15.291	-4.329	-1.412	2.723	20.682

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-7.78020	10.33872	-0.753	0.4565
gdppppd	0.36142	0.58214	0.621	0.5385
newgdp	NA	NA	NA	NA
healthexp	1.87001	0.75667	2.471	0.0182 *
subsaharanSub-Saharan	3.64354	4.54163	0.802	0.4275
muslperc	-0.07908	0.05967	-1.325	0.1932
literacy	0.12445	0.09867	1.261	0.2151

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.665 on 37 degrees of freedom

Multiple R-squared: 0.4782, Adjusted R-squared: 0.4077

F-statistic: 6.782 on 5 and 37 DF, p-value: 0.0001407

So...

- Perfect multicollinearity is terrible, but
- Perfect multicollinearity not a problem at all.

$$N > K \dots$$

Statistically,

- we lack sufficient degrees of freedom to identify  $\hat{\beta}$ .
- $\hat{\beta}$  is “overdetermined.”

Conceptually:

- Variables  $>$  Cases means
- ...no unique conclusion about explanatory / causal factors.

# $N = K$ in Practice

```
> smallAfrica<-subset(Africa,subsaharan=="Not Sub-Saharan")
> fit2<-with(smallAfrica,lm(adrate~gdppppd+healthexp+muslperc+
+                           literacy+war))
> summary(fit2)
```

Call:

```
lm(formula = adrate ~ gdppppd + healthexp + muslperc + literacy +
    war)
```

Residuals:

ALL 6 residuals are 0: no residual degrees of freedom!

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.12430	NA	NA	NA
gdppppd	-0.97906	NA	NA	NA
healthexp	-0.45166	NA	NA	NA
muslperc	0.01413	NA	NA	NA
literacy	0.09512	NA	NA	NA
war	-0.96429	NA	NA	NA

Residual standard error: NaN on 0 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: NaN

F-statistic: NaN on 5 and 0 DF, p-value: NA

# High (Non-Perfect) Multicollinearity

Recall that

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

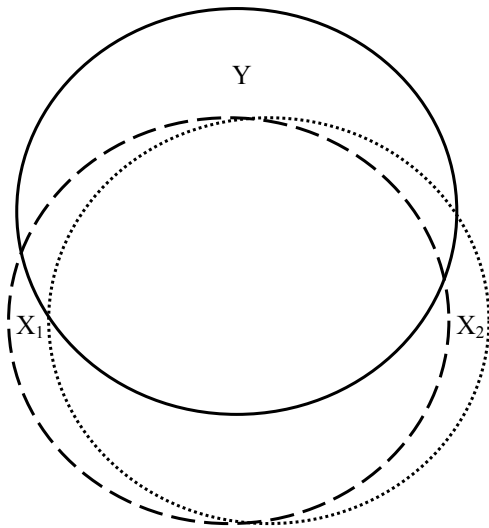
We can write the  $k$ th diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$  as:

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

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



# The Obligatory Venn Diagram



# High (Non-Perfect) Multicollinearity

Things to understand:

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

# Near-Perfect Collinearity: An Example

$$\text{HIV}_i = \beta_0 + \beta_1(\text{Civil War}_i) + \beta_2(\text{Intensity}_i) + u_i$$

```
> with(Africa, table(internalwar,intensity))
```

	intensity			
internalwar	0	1	2	3
0	30	0	0	0
1	0	6	2	5

Table: Three Models

	<i>Dependent variable:</i>		
	adrate		
	(1)	(2)	(3)
internalwar	-4.459 (3.274)		-2.849 (6.682)
intensity		-1.955 (1.481)	-0.837 (3.018)
Constant	10.713*** (1.800)	10.502*** (1.734)	10.713*** (1.821)
Observations	43	43	43
R <sup>2</sup>	0.043	0.041	0.045
Adjusted R <sup>2</sup>	0.020	0.017	-0.003
Residual Std. Error	9.860 (df = 41)	9.873 (df = 41)	9.973 (df = 40)
F Statistic	1.855 (df = 1; 41)	1.743 (df = 1; 41)	0.945 (df = 2; 40)

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

# (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.*

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:  $\text{VIF} > 10$  is a problem...

# What To Do?

Don't:

- **Drop Covariates!!!**
- Restrict  $\beta$ s...

Do:

- **Add data.**
- **Transform the covariates**
  - Data resuction
  - First differences
  - Orthogonalize
- **Shrinkage Methods** (e.g., “ridge regression”)