PLSC 503: "Multivariate Analysis for Political Research"

Collinearity, etc.
March 2, 2017

(Multi)Collinearity

You might hear, at some point, that *multicollinearity* is a problem with this or that model. In fact, it is something that – except in purely experimental settings, in which randomization is possible – is *always* present, to one degree of another. Multicollinearity is nothing more than having elements of $\mathbf{X}$ that are correlated with one another. Seen in that light, it is not so much a problem with data as a characteristic of it. Remember that as we discuss the matter.

As a general matter, multicollinearity actually comprises three interrelated assumptions:

1. No perfect linear relationship among the regressors (that is, $\mathbf{X}$ is of full column rank).
2. A greater number of observations than parameters (variables) ($N > K$).
3. "Sufficient" variability in the values of the regressors.

We'll talk about each of these....

"Perfect" Multicollinearity

The first of the requirements listed above essentially requires that there are no perfect linear dependencies in $\mathbf{X}$. A compact way of writing the requirement is to say that there cannot be any set of λs such that:

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

where $\mathbf{1}$ indicates the constant vector and the λs are constants not all zero.

Think about what this implies: If (1) holds, then any variable $\mathbf{X}_j$ can be written as a deterministic linear function of the other $K-1$ variables:

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

What happens if we then attempt to estimate a regression? By substitution, note that:

$$
\begin{aligned}
\mathbf{Y} &= \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + ... + \beta_j\mathbf{X}_j + ... + \beta_K\mathbf{X}_K + \mathbf{u} \\
&= \beta_0\mathbf{1} + \beta_1\mathbf{X}_1 + ... + \beta_j\left(\frac{-\lambda_0}{\lambda_j}\mathbf{1} + \frac{-\lambda_1}{\lambda_j}\mathbf{X}_1 + ... + \frac{-\lambda_K}{\lambda_j}\mathbf{X}_K\right) + ... + \beta_K\mathbf{X}_K + \mathbf{u} \\
&= \left[\beta_0 + \beta_j\left(\frac{-\lambda_0}{\lambda_j}\right)\right]\mathbf{1} + \left[\beta_1 + \beta_j\left(\frac{-\lambda_1}{\lambda_j}\right)\right]\mathbf{X}_1 + ... + \left[\beta_K + \beta_j\left(\frac{-\lambda_K}{\lambda_j}\right)\right]\mathbf{X}_K + \mathbf{u} \\
&= \left(\beta_0 + \frac{\gamma_0}{\lambda_j}\right)\mathbf{1} + \left(\beta_1 + \frac{\gamma_1}{\lambda_j}\right)\mathbf{X}_1 + ... + \left(\beta_K + \frac{\gamma_K}{\lambda_j}\right)\mathbf{X}_K + \mathbf{u} \qquad (2)
\end{aligned}
$$

where we combine estimates $\gamma_\ell = (-\lambda_\ell\beta_j)$.

Now, we can't estimate all the parameters of (2), because we now effectively have a system of $K-1$ equations (one for each $\mathbf{X}$ left in the model) with K unknowns. We can, however, estimate the $K-1$ linear combinations $\left(\beta_0 + \frac{\gamma_0}{\lambda_j}\right) ... \left(\beta_K + \frac{\gamma_K}{\lambda_j}\right)$, which, as we'll see later, can be a useful thing to do.

All of this, however, begs a question: *What happened to* $\mathbf{X}_j$?

The answer is that $\mathbf{X}_j$ is no longer there; or, at least, its influence is no longer there. Because you can perfectly "predict" / "explain" $\mathbf{X}_j$ with the other elements of $\mathbf{X}$, it is impossible to disentangle the unique contribution of $\mathbf{X}_j$ to $\mathbf{Y}$ from that owed to the other variables in $\mathbf{X}$.

Put another way: To estimate $\hat{\boldsymbol{\beta}}$, we need:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

which of course requires that

- we invert $\mathbf{X}'\mathbf{X}$, the variance-covariance matrix of $\mathbf{X}$. That, in turn, requires that
- the determinant of $\mathbf{X}'\mathbf{X} \neq 0$, which in turn requires that
- none of the columns of $\mathbf{X}'\mathbf{X}$ are perfectly linearly dependent on any others. However,
- if such a dependency exists in $\mathbf{X}$, that dependency will also exist in $\mathbf{X}'\mathbf{X}$ (why?), and
- so $\mathbf{X}'\mathbf{X}$ will not be invertible.

In addition, because

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

we also cannot compute the variance-covariance estimates for $\hat{\boldsymbol{\beta}}$, which means no standard error estimates, etc.

A Quick Example

Consider again our 2001 Africa data ($N = 43$). We might want to recode the *GDP* variable so that it was expressed in dollars (rather than thousands) and "centered" around its mean:

```
> Africa$newgdp<-(Africa$gdppppd-mean(Africa$gdppppd))*1000
```

Now, suppose we ran a regression adding the new *GDP* variable, but also forgot to leave out the old one:

```
> 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
```

Note that:

- `newgdp` and `gdppppd` are perfectly collinear, and so the model parameters can't be estimated when both are present in the regression.
- When we run the model anyway, R automatically notices the perfect collinearity, and (arbitrarily) drops one of the variables. (Stata does the same thing, FYI...).

- The same thing would happen if, for example, we'd created a variable that was a perfect linear combination of two or more other variables in the data (for example, if we created a variable called `anywar = war + internalwar`).

What This Implies

Two things, really:

- On the one hand, "perfect" multicollinearity is not just a threat to regression models, it is a nuclear bomb: It utterly prevents one from obtaining $\hat{\boldsymbol{\beta}}$s.
- On the other hand, as a practical matter, "perfect" multicollinearity is not really much of a problem at all. If it happens, you'll almost certainly know about it, and will be able to deal with it.

In fact, it is almost always the case that perfect collinearity arises because the researcher has done something bone-headed that ought to be fixed anyway. In that sense, we can think of it as a blessing: it probably prevents more "bad" analyses than it causes.

The Philosophical and Statistical Importance of $N > K$

What if you have more regressors than observations? In fact, this is a special case of what we just discussed.

Recall that multivariate regression is really a solution to a system of K equations in K unknowns (the elements of $\boldsymbol{\beta}$). Moreover, the $\mathbf{X}$s are assumed to be fixed. Taken together, these things imply that

- if there are fewer observations (that is, *values* of $\mathbf{X}_i$) than variables, there are not enough "fixed" values to allow us to solve for the unknowns.
- In statistical terms, we lack adequate *degrees of freedom* to estimate the parameters.
  - ○ Think back to the "world's simplest regression," with two data points. There, we needed two observations ($N = 2$) to get a slope and an intercept ($K = 2$).
  - ○ When we wanted to talk about variability around that regression line (that is, about σ^2), that required a third data point (since we then had three parameters to estimate, we needed at least three observations).

A system in which $N \leq K$ is said to be *overdetermined*, in that there is insufficient information in the data (the $\mathbf{X}$s) to uniquely estimate all the parameters (the $\boldsymbol{\beta}$s and σ^2). As a result, you cannot solve the system of equations given the data, and so cannot get a full set of parameter estimates.

Conceptually (as well as statistically) this should make a good bit of sense: If you have as many variables doing the "explaining" as you have cases on which to base the explanation, you can't come to a unique conclusion about the relative influence of the various factors. E.g., if we observe that

- Our 23-year-old poor Ph.D.-student cousin Kenny doesn't vote, and
- Our wealthy 64-year-old Aunt June (who never even finished high school, but made a mint in the office supply business) does, then
- Is non-voting due to age? gender? education? wealth?
- Answer: without more data, we can't tell...

Another Quick Example

Consider the following regression on our Africa data:

```
> 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
```

Here, $N = K$: we have as many covariates (including the constant term) as we have observations. Note:

- With $N = K$ we can estimate a set of coefficients $\hat{\boldsymbol{\beta}}$. However,
- Because we have as many variables as cases, the covariates can perfectly predict the outcome $\mathbf{Y}$. As a result,
- The RSS is zero, which means that $\hat{\sigma}^2$ is also zero and the R^2 is 1.0.
- Because $\hat{\sigma}^2 = 0$, the variance-covariance estimates cannot be calculated.

High (Non-Perfect) Multicollinearity

Multicollinearity is also the name we give to the problem of nearly perfect linear relationships among our regressors; and, in fact, this is a far more common (and typically more serious) problem than is perfect multicollinearity. And, in fact, the last example is a special case of (and, conceptually, begins to get at) the third potential issue raised by multicollinearity: *Insufficient variability in* $\mathbf{X}$.

Near-Perfect Collinearity: The Statistics

Note at the outset that – unlike perfect collinearity – near-perfect collinearity holds absolutely no evil ramifications for our ability to estimate $\boldsymbol{\beta}$ and σ^2, and to do so in an unbiased and efficient manner. That is, **so long as X is of full column rank, $N > K$, and the other OLS assumptions hold, the OLS estimates of $\hat{\boldsymbol{\beta}}$ are BLUE**, no matter how high the correlations among the $\mathbf{X}$s.

That said, high collinearity among the variables in $\mathbf{X}$ can still cause some headaches. Recall that

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

As I've said a million times before, $(\mathbf{X}'\mathbf{X})^{-1}$ is the variance-covariance matrix of the covariates $\mathbf{X}$. If the multicollinearity among the $\mathbf{X}$s is nearly perfect, then the elements of $\mathbf{X}'\mathbf{X}$ will be very large, and so the elements of $(\mathbf{X}'\mathbf{X})^{-1}$ will be very large as well.

- Intuitively, because the $\mathbf{X}$s are so highly related, their covariation is quite high.
- Since $\widehat{\text{Var}(\hat{\boldsymbol{\beta}})}$ is inversely proportional to the amount of (independent) variability in $\mathbf{X}$, small amounts of independent variability among the $\mathbf{X}$s will yield large elements of $\widehat{\text{Var}(\hat{\boldsymbol{\beta}})}$.

Put somewhat differently, one can note that the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ can be written as

$$\frac{1}{(\mathbf{X}_k'\mathbf{X}_k)(1 - \hat{R}_k^2)} \tag{3}$$

where $\mathbf{X}_k'\mathbf{X}_k$ is the variance of the kth variable $\mathbf{X}_k$ and $\hat{R}_k^2$ is the R^2 from the regression of $\mathbf{X}_k$ on all the other variables in $\mathbf{X}$. This means that:

- If the relationship is perfect – that is, if one can perfectly predict $\mathbf{X}_k$ with the other $\mathbf{X}$s – then $\hat{R}_k^2 = 1$ and so that element of $(\mathbf{X}'\mathbf{X})^{-1}$ is inestimable.
- $\lim\limits_{\hat{R}_k^2 \to 1} \frac{1}{(\mathbf{X}_k'\mathbf{X}_k)(1-\hat{R}_k^2)} = \infty$. That means that
- If the relationship is nearly perfect, then
  - ○ $\hat{R}_k^2$ is nearly 1.0, and so
  - ○ the denominator of (3) is small, and
  - ○ the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ (that is, the estimated variance of $\hat{\beta}_k$) will be large.

Near-Perfect Collinearity: The Intuition

In a different regression textbook, Damodar Gujarati talks about "micronumerosity" (i.e., too few observations), rather than "multicollinearity." And, in fact, that is one way of thinking about the problem of near-perfect multicollinearity:

- If you have a lot of correlation among your independent variables, there is little *independent* variation in them.
- That is, each one "explains" the other(s) very, very well.

Intuitively, this means that:

- It is difficult to separate the impact of one variable (call it $\mathbf{X}_1$) on $\mathbf{Y}$ from that of another, highly-collinear one (say, $\mathbf{X}_2$).
- Because of this difficulty, it is more difficult to be sure that the estimates you get in $\hat{\boldsymbol{\beta}}$ are in fact close to the "true" population parameters $\boldsymbol{\beta}$.
- Hence, near-perfect multicollinearity leads to larger standard error estimates, and correspondingly wider confidence intervals, for your parameter estimates.

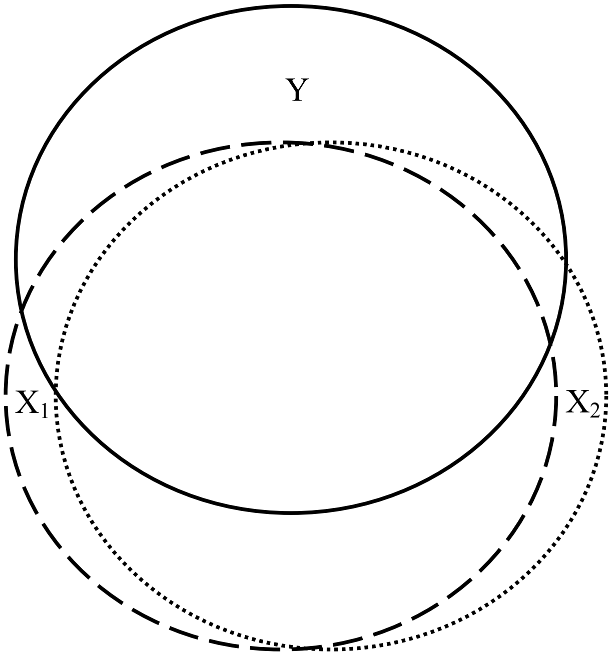
The classic way of illustrating this is through a Venn diagram, as in Figure 1. Here,

- The circles are supposed to represent the amount of variability in the variables $\mathbf{Y}$, $\mathbf{X}_1$, and $\mathbf{X}_2$.
- Here, $\mathbf{X}_1$ and $\mathbf{X}_2$ are highly collinear: Most of their variation is *shared* variation.
- Both $\mathbf{X}_1$ and $\mathbf{X}_2$ explain a relatively large amount of $\mathbf{Y}$. However,

- Because of their collinearity, neither of them explains very much of $\mathbf{Y}$ *over and above that explained by the other.*

- Thus, if we want to know the effect of (say) a change in $\mathbf{X}_1$ on $\mathbf{Y}$, it is very hard to say what that effect is with any precision (because it might, in fact, be change due not to $\mathbf{X}_1$, but to $\mathbf{X}_2$ instead).

Figure 1: The Obligatory Multicollinearity Venn Diagram



So, consider what happens if (say) we want to study presidential influence on federal judicial decision making. One approach to this would be to see if – once we've controlled for the ideology of the judge in question – the ideology of the president that appointed that judge correlates with their voting. Since political party identification is a standard indicator of political ideology, we might estimate a regression that looks like:

$$\text{Voting}_j = \beta_0 + \beta_1(\text{Judge's Party}_j) + \beta_2(\text{President's Party}_j) + u_j$$

Now, of course, Presidents tend to appoint judges from their own political party. This means that:

- The two variables are highly correlated ($r \geq 0.90$ or so, in most cases).

- This means I have high multicollinearity, and so will get large standard error estimates for $\hat{\beta}_1$ and $\hat{\beta}_2$.
- Conceptually: If I observe (say) a Democratic Carter appointee voting liberally, I can't tell whether that effect is due to her own political ideology or that of her appointing president.
- This, in turn, suggests that in order to determine if my hypotheses are correct, I need sufficient data on judges who are *not* of the same party as their appointing president.
- If I have such data, then my estimates will be more precise, and I can be more confident in them. But, of course,
- If that is the case, it also means that the two variables aren't so collinear after all...

Near-Perfect Collinearity: Implications

The preceding little discussion suggests two things about the multicollinearity problem:

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

In the first case, note that:

- Multicollinearity is typically (but not always) isolated to a particular sample.
- That is, if we have a high correlation between two variables in our data, it may or may not be the case that the correlation is equally high in the population.
- More to the point: One could, in theory, ameliorate the multicollinearity by endeavoring to pick a sample that had a lower correlation among those variables (e.g., one that oversampled opposite-party judicial appointees).

With respect to the second point,

- You will always have some covariation among the regressors in your data (again, unless you're doing randomized experiments).
- Thus, the important question is not *whether* you have multicollinearity, but rather *how much*, and whether or not it matters for your ability to make inferences about your quantities of interest.

Near-Perfect Collinearity: An Example

Suppose that we're (once again) interested in explaining HIV/AIDS rates in Africa, and specifically that we want to learn about the relationship between civil wars and AIDS. We might ask two questions:

1. Do countries that experience *civil wars* have higher HIV/AIDS rates than those that do not?
2. Does the *intensity* of the civil war affect the HIV/AIDS rate?

At a conceptual level, these are distinct questions: we could find, for example, that

- There is no difference in AIDS rates between countries that have civil wars and those that don't,
- There is a difference, but the extent of that difference doesn't vary with the intensity of that war,
- There is a threshold effect: Countries that have minor civil wars are not different from those that have none, but those that have major/intense civil wars are.
- Countries experiencing civil wars have (e.g.) higher HIV rates than those that do not, but the effect of intensity is negative (that is, HIV/AIDS rates go *down* as the intensity of the war increases).

A natural way to begin to evaluate these hypotheses is to estimate a model like:

$$\text{HIV}_i = \beta_0 + \beta_1(\text{Civil War}_i) + \beta_2(\text{Intensity}_i) + u_i \tag{4}$$

Of course, by construction, these two variables are going to be highly correlated with each other:

```
> table(internalwar,intensity)
           intensity
internalwar  0  1  2  3
          0 30  0  0  0
          1  0  6  2  5
```

It is telling to examine the bivariate regressions of these variables, as well as that specified by Equation (4):
Note a few things about these three regressions:

- In all three, the estimates of $\hat{\boldsymbol{\beta}}$ are unbiased and smallest-variance; that is, they are all "good" estimates.

Table 1: Three Models

	Dependent variable:		
	adrate		
	(1)	(2)	(3)
internalwar	−4.459		−2.849
	(3.274)		(6.682)
intensity		−1.955	−0.837
		(1.481)	(3.018)
Constant	10.713^{***}	10.502^{***}	10.713^{***}
	(1.800)	(1.734)	(1.821)
Observations	43	43	43
R^2	0.043	0.041	0.045
Adjusted R^2	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<0.1; **p<0.05; ***p<0.01

- The standard error estimate for `internalwar` more than doubles when `intensity` is introduced into the equation, and that for `intensity` more than doubles in the multivariate model as well.
- The R^2 statistics tell us that:
  - both variables do about an equally good job of explaining `adrate`, but
  - the addition of each to the model including the other effectively doesn't improve the model fit at all.
  - The latter point is most clearly conveyed by the $R^2_{adj.}$, which is actually *negative* in the multivariate model.
- The same thing is conveyed by the `RMSE`, which actually increases with the addition of both variables to the model (relative to either of the bivariate ones). (Recall that this can occur because $\hat{\sigma}^2$ contains $N - K$ in its denominator; in other words, the addition of either variable to the model containing the other doesn't even reduce the overall error variability by an amount equal to $1/N$ here).

All of these characteristics, as it turns out, are classic indicators of a multicollinearity problem.

(Near-Perfect) Multicollinearity: Detection

So how do we detect it?

1. *High R^2, but nonsignificant coefficients.*
   - A model with relatively high R^2, but no covariates that are statistically distinguishable, exhibits a classic symptom of multicollinearity. But...
   - One doesn't necessarily need a high R^2 to have a multicollinearity problem...
     - It's perfectly possible to have a model that only explains 10 percent of the variance in $\mathbf{Y}$ and still have massively collinear covariates.
     - For example, consider what would happen if we regressed SAT performance on height, weight, shoe size, and hat size...
   - This means that detection is a bit harder, since we tend to take insignificant ts (along with low R^2) as signs of a crappy model, not multicollinearity.
2. *High pairwise correlations among independent variables.*
   - An easy way to look for simple multicollinearity is just to display a correlation matrix of $\mathbf{X}$. However,
   - This is also not always effective...

- ○ High pairwise correlations are a sufficient but not necessary condition for multicollinearity to be a problem.
- ○ More complex linear relationships may exist among the independent variables.
- ○ E.g. Suppose that we have four variables in $\mathbf{X}$, and the regression $\mathbf{X}_1 = -17 + 2(\mathbf{X}_2) - 0.5(\mathbf{X}_3) - 100(\mathbf{X}_4)$ yields an $R^2 = 0.97$...
- ○ Here, the linear combination of $\mathbf{X}$s is nearly perfect, but the individual correlations between $\mathbf{X}_1$ and the other variables may not be that high.

3. *High partial correlations among the* $\mathbf{X}$*s.*

- This is the natural way of getting around the problem with (2).
- It can certainly be useful, though it can also get quite complicated as the number of variables in $\mathbf{X}$ increases.
- An easier way to get at the same thing is...

4. *Auxilliary regressions of the* $\mathbf{X}$*s.*

- Remember that, in the multivariate context:

$$\widehat{\text{Var}(\hat{\beta}_k)} = \frac{\hat{\sigma}^2}{(\mathbf{X}_k'\mathbf{X}_k)(1 - \hat{R}_k^2)} \tag{5}$$

  where $\mathbf{X}_k'\mathbf{X}_k$ is the variance of the kth variable $\mathbf{X}_k$ and $\hat{R}_k^2$ is the R^2 from the regression of $\mathbf{X}_k$ on the other $K - 1$ variables in $\mathbf{X}$.
- That "auxiliary" R^2 – $\hat{R}_k^2$ – is quite useful, in that it can tell you a lot about the collinearity of that variable.
- This suggests that one option would be to regress each of the $\mathbf{X}_j$s on all the other elements of $\mathbf{X}$ to see if there are any strong linear dependencies. One could then:
  - ○ Conduct an F-test on that regression:

$$F_k = \frac{\frac{\hat{R}_k^2}{K-2}}{\frac{1-\hat{R}_k^2}{N-K+1}}$$

    is distributed according to an F distribution, with $K - 2$ and $N - K - 2$ degrees of freedom; this amounts to a test of whether the other variables in $\mathbf{X}$ have any significant (joint) linear relationship to $\mathbf{X}_j$.
  - ○ One can also adopt "Klein's Rule": *If any of the* $\hat{R}_k^2$ *are larger than the* $\hat{R}^2$ *of the original model, you have a problem...*
- An even better, simpler approach, however, is to consider:

5. *VIF and Tolerance.*

- Note that if $\hat{R}^2_k = 0$, then (5) reduces to

$$\widehat{\text{Var}(\hat{\beta}_k)} = \frac{\hat{\sigma}^2}{\mathbf{X}'_k \mathbf{X}_k}; \tag{6}$$

  think of this as a "lower bound" on the degree to which $\mathbf{X}_k$ can be collinear with the other variables in $\mathbf{X}$.

- Dividing (6) by (5) yields a measure of *relative* collinearity known as the "Variance Inflation Factor" (VIF):

$$\text{VIF}_k = \frac{1}{1 - \hat{R}^2_k} \tag{7}$$

- VIF_k tells how much the variance of $\hat{\beta}_k$ is "inflated" by the multicollinearity of $\mathbf{X}_j$ with the other elements of $\mathbf{X}$.
- While there are no hard and fast rules, a VIF of larger than, say, 10 (that is, an $\hat{R}^2_k > 0.90$) is usually an indication that multicollinearity may be a problem with that variable.
- *Tolerance* is just $\frac{1}{\text{VIF}_k}$; it rescales VIF to range between zero and one.
  - Tolerance = 0 ⟶ perfect multicollinearity.
  - Tolerance = 1 ⟶ no multicollinearity.

Finally (and happily), we usually don't have to calculate these things for ourselves; most good software packages (yes, including R and Stata both) will calculate the VIF_ks and tolerances following a regression model automatically.

(Near-Perfect) Multicollinearity: What Do We Do About It?

One thing that you'll often hear said in statistics classes is:

> **"If you have multicollinearity, and your coefficients are significant anyway, stop right there."**

The logic behind this idea is that, because all multicollinearity does is inflate one's standard error estimates, if one's estimates are still "small enough" to yield reasonably precise coefficient estimates then there's nothing to be worried about. As a general matter, this is not bad practical advice, but its only mediocre statistical advice.

- Practically speaking, since
  - the estimates are BLUE,
  - your coefficients are still "statistically significant," and

    - reviewers/readers know that this is what multicollinearity does

  then you're probably OK.

- Statistically, however,

    - we still want the most precise estimates we can get (not just for big t-statistics, but in general terms, for precision).
    - This means that even if your estimates are significant, it still makes sense at least to conduct some diagnostics, and (if multicollinearity is present) to consider some of the possible "remedies" to improve that precision.

What Not To Do

1. **Dropping one or more of the collinear variables from the model.**

    - **Do Not Do Not Do Not Do This!!!**
    - It can be very tempting, largely because, when you do, your results will likely be "significant."
    - However, it is also – very directly – specification error.
    - In other words, *it is tantamount to taking good (BLUE) estimates and exchanging them for biased, inefficient, crappy ones.*
    - As in all things, let *theory* be your guide...
        - If your theory suggests that one of the variables can be dropped, that's one thing.
        - If not, **do not do it.**

2. **A priori restrictions on coefficients.**

    - That is, restricting some coefficients to be equal to zero (or some other value) on the basis of theory.
    - This is often substantively similar (and statistically identical) to dropping variables.
    - This sort of thing requires a lot of theory; it's sometimes possible for natural scientists, and even economists, but not likely for us...

What To Do

1. **Add data**.

    - This is always a good idea...

    - ○ At the very least, will decrease $\hat{\sigma}^2$ and give you more precise estimates.
    - ○ Moreover, if the data you add yield more "odd" observations, it will also reduce the degree of multicollinearity and improve them even further.
- Note, however, that this does *not* imply that one should go out with the idea of collecting *only* data which "go against" the pattern(s) of multicollinearity in the independent variables (so, for example, collecting only data on jurists whose political party is not that of the president that appointed them).

2. **Transform the covariates**.

- In practice, there are lots of ways to do this. Here, I'll focus on three.
- <u>Data reduction</u>.
    - ○ Often, we might have multiple variables that are all indicators of the same underlying concept.
    - ○ So, for example, political unrest might be measured in terms of numbers of general strikes, riots, assassinations, demonstrations, and so forth.
    - ○ When that is the case, it is usually possible (and desirable) to combine related variables into an "index" of some sort.
    - ○ The simplest way to do this is to create an *additive index* of some kind, either by
        - · Summing up comparable covariates, or
        - · Summing up binary indicators of the presence/absence of such things.
    - ○ Additive indices get used a lot (e.g., the POLITY scores, and the Segal-Cover Supreme Court ideology scores, are examples).
    - ○ However, they also typically imply some pretty stringent restrictions, such that there are often better ways of aggregating such variables.
    - ○ A more sophisticated way to create an index is through *factor* (also called *principal components*) *analysis.*
        - · This is too complicated to go into here, but
        - · Suffice it to say that it's a method for combining data across a number of variables that takes into account how highly correlated each is with some underlying (latent) "factor(s)."
        - · For more, ask me outside of class (or, preferably, attend Bill Jacoby's class at ICPSR).
    - ○ Irrespective of which approach you use, it *must* be guided by theory: *do not* index/factor variables that are theoretically distinct, even if / just because they're highly correlated.
- <u>Take "first differences</u>."
    - ○ This is often useful in time-series data, to eliminate time-related multicollinearity.

- ○ E.g. trending variables are often collinear with each other.
- ○ Differencing allows you to estimate the same parameters, but without the multicollinearity. However,
- ○ It can cause some other problems (as we'll see when we discuss autocorrelation, next week).

- Orthogonalize the covariates.
  - ○ If we think back to our Venn diagram, we could imagine "netting out" the variation in a particular covariate $\mathbf{X}_k$ that is "independent" of some other variable of set of variables.
  - ○ Statistically, this is done by *orthogonalizing* the variables; this involves
    - · Ordering the variables, say: $\{\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_K\}$.
    - · Creating a new variable $\mathbf{O}_1$ that is essentially a "normalized" version of $\mathbf{X}_1$ (that is, $O_{1i} = \frac{X_{1i}-\bar{X}_1}{s_{\mathbf{X}_1}}$).
    - · Creating $\mathbf{O}_2$ that is the remaining variation in (normalized) $\mathbf{X}_2$ after the variation in $\mathbf{X}_2$ that is shared with $\mathbf{X}_1$ is removed. (Think of regressing $\mathbf{X}_2$ on $\mathbf{X}_1$ and then normalizing the residuals...).
    - · Repeating this for $\mathbf{O}_3$, $\mathbf{O}_4$, etc. up through $\mathbf{O}_K$.
  - ○ By construction, the variables in $\mathbf{O}$ will be uncorrelated with each other.
  - ○ These "orthogonalized" variables can then be included in the right-hand side of the model in place of the $\mathbf{X}$s.
  - ○ The resulting regression results will thus *not* suffer from multicolllinearity.
  - ○ However, a key factor in orthogonalization is that *the order in which it occurs is critical to the results one will get.*
  - ○ There's an example of how to do this in Stata at the end of the notes, but we won't go into it much more here...

3. **Shrinkage Methods.**

- Think for a moment about the idea behind OLS estimation – that is, choosing $\hat{\boldsymbol{\beta}}$ to minimize the squared errors. We can think of this as:

$$\begin{aligned} \text{MSE} &= \text{E}\{[\mathbf{Y} - \text{E}(\mathbf{Y})]^2\} \\ &= \text{E}[(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^2] \end{aligned}$$

which can be further written as:

$$\text{MSE} = [Y_i - \text{E}(\mathbf{X}_i\hat{\boldsymbol{\beta}})]^2 + \{\text{E}[(\mathbf{X}_i\hat{\boldsymbol{\beta}}) - \text{E}(\mathbf{X}_i\hat{\boldsymbol{\beta}})]\}^2 \qquad (8)$$

This is often restated – intuitively – as:

$$\text{MSE} = (\text{Bias})^2 + \text{Variance}$$

which reflects the fact that the total mean squared error in a model is a function of both the amount of bias (that is, expected error) in the model and the variability of the estimates, with the former receiving comparably more weight.

- When multicollinearity is high, the mean squared error is being driven up by the second term (because the estimator $\hat{\boldsymbol{\beta}}$ has a high degree of variance).
- This leads us to consider whether we might be willing to "trade off" some bias in exchange for smaller variance in our estimates.
- The most common way of doing this is "ridge regression" (see e.g. Hoerl and Kennard 1970 *Technometrics*):

$$\hat{\boldsymbol{\beta}}^R = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y} \tag{9}$$

- The idea here is to introduce a bias factor equal to a constant λ into the diagonal elements of $\mathbf{X}'\mathbf{X}$ prior to inverting it. This has two effects:
  (a) It *biases* the point estimates, so that $\text{E}(\hat{\boldsymbol{\beta}}^R \neq \boldsymbol{\beta})$.
  (b) It increases the (perceived) independent variability in $\mathbf{X}$, which in turn decreases the standard errors of the estimated coefficients.
  (c) Specifically, the variances of the parameter estimates from the ridge regression can be written as:

$$\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_\ell^R)} = \frac{\hat{\sigma}^2}{(\mathbf{X}_\ell'\mathbf{X}_\ell + \lambda)(1 - R_\ell^2)} \tag{10}$$

- Statistically, ridge regression has the effect of reducing the effective number of parameters which the model can account for; in that sense, it is similar (in some respects) to dropping covariates.
- The larger the value of λ, the greater the bias, but the greater the level of variance reduction. In theory, one can achieve *any* level of $\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_\ell^R)}$ by choosing a large enough value of λ.
- As a practical matter, one often standardizes the variables in $\mathbf{X}$ before doing a ridge regression; otherwise, a particular value of λ can have a much greater impact on some covariates than others.
- Finally, a source of much controversy has been the choice of λ.
  - There's a large literature on this, and on more general (2- and 3-parameter) shrinkage models, in statistics.
  - The "obvious" alternative – choose a value of λ that yields the smallest MSE – is problematic, since it conflates model specification and model results (which is generally an inferential no-no).

  - There's a lot more to read on this if you really want to; however...

- I cannot think of one single use of ridge regression in political science. So, its practical value is somewhat limited as well.

An Example: House Voting on Impeachment

For our example, we'll look at some data on the vote on the impeachment of President Clinton in the U.S. House of Representatives. The dependent variable is the number of articles of impeachment that each member of the House voted in favor of; that means it ranges from zero (voted against all four articles) to four (voted in favor of all four). We'll consider five covariates:

- `pctbl96` – the percentage of that member's House district that was African-American, as of 1996.

- `unionpct` – the percentage of that member's House district that were members of organized labor,

- `clint96` – the percentage of the two-party presidential vote that President Clinton received in that district in the 1996 election,

- `GOPmember` – a dummy variable coded 1 if the representative was a member of the Republican party, and 0 if they were a Democrat, and

- `ADA98` – the member's 1998 Americans for Democratic Action score; a 0-100 measure of how liberal that member's voting behavior is.

The summary statistics for these data look like this:

```
> summary(impeachment)
     name              state              district     votesum
 Length:433         Length:433         Min.   : 1   Min.   :0.00
 Class :character   Class :character   1st Qu.: 3   1st Qu.:0.00
 Mode  :character   Mode  :character   Median : 6   Median :2.00
                                       Mean   :10   Mean   :1.85
                                       3rd Qu.:13   3rd Qu.:4.00
                                       Max.   :52   Max.   :4.00
    pctbl96        unionpct         clint96      GOPmember          ADA98
 Min.   : 0.0   Min.   :0.0257   Min.   :26.0   Min.   :0.000   Min.   :  0.0
 1st Qu.: 2.0   1st Qu.:0.0930   1st Qu.:42.0   1st Qu.:0.000   1st Qu.:  5.0
 Median : 5.4   Median :0.1690   Median :48.0   Median :1.000   Median : 30.0
 Mean   :11.9   Mean   :0.1636   Mean   :50.3   Mean   :0.527   Mean   : 46.3
 3rd Qu.:14.0   3rd Qu.:0.2150   3rd Qu.:57.0   3rd Qu.:1.000   3rd Qu.: 90.0
 Max.   :74.0   Max.   :0.3733   Max.   :94.0   Max.   :1.000   Max.   :100.0
```

We begin by estimating the five-variable regression model:

```
> fit<-lm(votesum~ADA98+GOPmember+clint96+pctbl96+unionpct)
> summary(fit)

Call:
lm(formula = votesum ~ ADA98 + GOPmember + clint96 + pctbl96 +
    unionpct)

Residuals:
   Min     1Q Median     3Q    Max
-3.271 -0.259  0.133  0.337  2.731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.51785    0.23246   10.83   <2e-16 ***
ADA98       -0.02144    0.00238   -9.00   <2e-16 ***
GOPmember    1.59981    0.18043    8.87   <2e-16 ***
clint96     -0.00935    0.00433   -2.16    0.031 *
pctbl96      0.00347    0.00270    1.29    0.199
unionpct    -0.52544    0.48065   -1.09    0.275
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.629 on 427 degrees of freedom
Multiple R-Squared: 0.883,Adjusted R-squared: 0.882
F-statistic:  647 on 5 and 427 DF,  p-value: <2e-16
```

At the outset, we probably don't have any strong reasons – on the basis of these results – to suspect that multicollinearity is a big problem. Note that:

- The R^2 is relatively high; however,
- A number of the covariates are also statistically differentiable from zero, and in the "direction" we'd expect them to be.

At the same time, a quick look at the correlation matrix of the variables in the model suggests that there might be some issues:

```
> idata=impeachment[c(-1,-2)]
> cor(idata)
          district  votesum  pctbl96 unionpct clint96 GOPmember    ADA98
district   1.00000 -0.03496 -0.06759  0.09155  0.1044  -0.02881  0.04988
votesum   -0.03496  1.00000 -0.28765 -0.26199 -0.6408   0.91977 -0.92795
```

```
pctbl96   -0.06759 -0.28765  1.00000 -0.09394  0.6165  -0.30911  0.30288
unionpct   0.09155 -0.26199 -0.09394  1.00000  0.3331  -0.19406  0.27563
clint96    0.10437 -0.64084  0.61651  0.33305  1.0000  -0.61196  0.67033
GOPmember -0.02881  0.91977 -0.30911 -0.19406 -0.6120   1.00000 -0.93918
ADA98      0.04988 -0.92795  0.30288  0.27563  0.6703  -0.93918  1.00000
```

The very strong, negative correlation between `GOPmember` and `ADA98` (i.e., that Republicans tend to vote conservatively and Democrats liberally) is a likely source of multicollinearity. Happily, R (and Stata) will calculate the VIFs and tolerances for each of the variables in the regression for us, using a variant of the `estat` command:

```
> vif(fit)
    ADA98 GOPmember   clint96   pctbl96  unionpct
   10.292     8.878     3.313     1.998     1.371
```

These VIFs suggest that the `ADA98` and (to a lesser extent) the `GOPmember` variables are highly collinear with the others, including each other. The `ADA98` variable, in particular, is very well-explained by the other four covariates.

What to do? We could reasonably...

- ...do nothing. Since the variables most affected by collinearity remain important (and statistically significant) predictors of `votesum`, we could defensibly restrict our response to noting (maybe in a footnote) that there was some significant correlation among the **X**s, but that it didn't appreciably affect our results.
- ... drop `GOPmember` or `ADA98`. Bear in mind that we should only do this if (a) we have a theoretical reason to do so (e.g., that we believe that both variables are just proxies for the same thing – that being left-right ideology), and (b) we have no other good alternatives. In this case, this looks to be a particularly bad idea, since both `GOPmember` and `ADA98` still have "statistically significant" effects on **Y** even in the presence of the other.

    If we do so, we get:

```
> fit2<-lm(votesum~ADA98+clint96+pctbl96+unionpct)
> summary(fit2)

Call:
lm(formula = votesum ~ ADA98 + clint96 + pctbl96 + unionpct)

Residuals:
   Min     1Q Median     3Q    Max
-3.300 -0.300  0.179  0.383  2.913
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02775    0.17198   23.42   <2e-16 ***
ADA98       -0.04052    0.00111  -36.60   <2e-16 ***
clint96     -0.00658    0.00469   -1.40     0.16
pctbl96      0.00165    0.00293    0.56     0.57
unionpct     0.08300    0.51706    0.16     0.87
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.684 on 428 degrees of freedom
Multiple R-Squared: 0.862,Adjusted R-squared: 0.861
F-statistic:  667 on 4 and 428 DF,  p-value: <2e-16
```

- Note that
  - `ADA98` is now massively significant – in large part because its standard error has been more than cut in half, while its coefficient estimate is almost twice as large.
  - The model fit is roughly the same as it was before – the R^2 changed by only 0.02.
  - When we do this, the VIFs drop to more reasonable levels:

```
> vif(fit2)
   ADA98  clint96  pctbl96 unionpct
   1.883    3.296    1.986    1.343
```

- Finally, we can consider estimating an ordinary ridge regression; this can get complicated in a hurry... There's no simple way I've found to use R to do orthogonalized regression, but that's not a really big deal anyway. R is just fine, however, at doing ridge regression. The relevant command is `lm.ridge()`:

```
> ridge.vote<-lm.ridge(votesum~ADA98+GOPmember+clint96+pctbl96+unionpct,
  lambda=seq(0,5000,10))
> select(ridge.vote)
modified HKB estimator is 0.8365
modified L-W estimator is 0.4018
smallest value of GCV  at 10
```

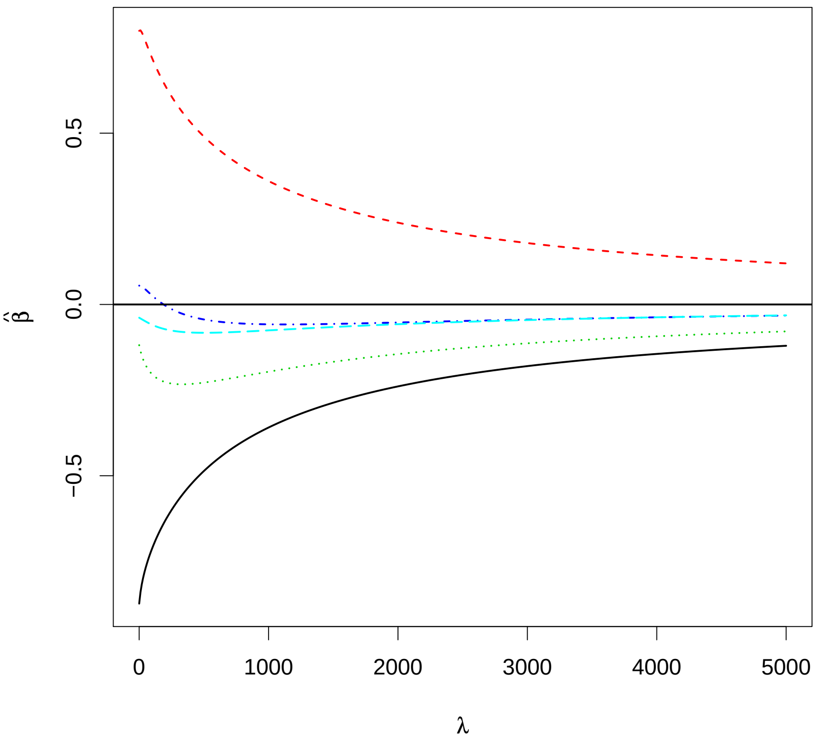
The latter are "automatically" selected values of λ; oftentimes they will be (relatively) small, and the analyst will prefer a greater degree of shrinkage.

A `lm.ridge` object contains the $\hat{\beta}^R_\ell$s for each different value of λ; The standard practice is to use these in a "ridge plot":

```
> matplot(ridge.vote$lambda,t(ridge.vote$coef),type="l",xlab=expression(lambda),
  ylab=expression(hat(beta)),lwd=2)
> abline(h=0)
```

which yields:

Values of $\hat{\beta}_k^R$, by λ



This plots the values of each of the $\hat{\beta}$s against λ, the degree of shrinkage. So, for example, the black line is $\hat{\beta}_{\texttt{ADA98}}$, the red line is $\hat{\beta}_{\texttt{GOPmember}}$, and so forth. Note that they all start off at their OLS values, and eventually wind up at zero, but in between they can take on different values.