9: k-Variable Linear Model Miscellany
ECON 837

Brian Krauth (adapted from notes by Simon Woodcock), Spring 2010

Specification Error

Suppose the data generating process is $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ but the model we fit is $\mathbf{y} = \mathbf{X}^*\beta^* + \varepsilon$ with least squares estimator

$$\mathbf{b}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y} = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{X}\beta + (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\varepsilon.$$

Then

$$\begin{aligned} E[\mathbf{b}^*] &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{X}\beta \\ Var[\mathbf{b}^*] &= \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}. \end{aligned}$$

Omission of Relevant Variables

Let $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ and $\mathbf{X}^* = \mathbf{X}_1$. That is, the DGP is $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$ but we regress $\mathbf{y}$ on $\mathbf{X}_1$ only. Consider $\mathbf{b}^*$ as an estimator of β_1.

Proposition 1 $\mathbf{b}^*$ *is biased.*

Proof.

$$\begin{aligned} \mathbf{b}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y} \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon) \\ &= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\varepsilon \\ E[\mathbf{b}^*] &= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 \\ &\neq \beta_1. \end{aligned}$$

■

How do we interpret the bias term $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$? Notice that $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is the least squares estimate of the coefficient vector in the regression of $\mathbf{X}_2$ on $\mathbf{X}_1$. Thus the bias is zero only when this coefficient vector is zero, i.e., if $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal (so that $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$). In general, however, we cannot sign the bias since it depends on the sign of $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ and the sign of β_2. However, it is very common in applied work to make an educated guess of these two signs, and thus the direction of bias.

Note that $Var[\mathbf{b}^*] = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$, so that if $\beta_2 = \mathbf{0}$ there is an efficiency gain from imposing the (true) restriction and leaving $\mathbf{X}_2$ out of the model (recall our earlier results on constrained estimation).

Example 2 *Suppose the DGP is $y_i = \beta_0 + \beta_1 x_i + a_i + \varepsilon_i$ where y_i is the natural logarithm of earnings, x_i is a measure of schooling, and a_i is ability. Ability a_i is unobserved and omitted,*

but is positively correlated with schooling x_i. Suppose we normalize the coefficient on ability to 1 and $\sum_i a_i = 0$. Then

$$\begin{aligned} E[\mathbf{b}^*] &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i a_i \\ \sum_i a_i x_i \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \sum_i a_i x_i \end{bmatrix}. \end{aligned}$$

*The bias on the schooling coefficient is positive (it measures the return to schooling **and** ability), and the intercept is unaffected.*

Estimating σ^2

Letting $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1 (\mathbf{X}_1'\mathbf{X}_1)^{-1} \mathbf{X}_1'$, the residuals in the misspecified model are

$$\begin{aligned} \mathbf{e}^* &= \mathbf{M}_1\mathbf{y} = \mathbf{M}_1 (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon) \\ &= \mathbf{M}_1\mathbf{X}_2\beta_2 + \mathbf{M}_1\varepsilon \\ &\neq \mathbf{M}_1\varepsilon. \end{aligned}$$

Therefore

$$\mathbf{e}^{*\prime}\mathbf{e}^* = \beta_2'\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\beta_2 + \varepsilon'\mathbf{M}_1\varepsilon + 2\beta_2'\mathbf{X}_2'\mathbf{M}_1\varepsilon$$

where the last term is zero in expectation. This shows that even if the bias is zero ($\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$), our usual estimators of σ^2 will be biased since $\mathbf{M}_1\mathbf{X}_2 \neq \mathbf{0}$.

Inclusion of Irrelevant Variables

Now suppose that $\mathbf{X} = \mathbf{X}_1$ and $\mathbf{X}^* = [\mathbf{X}_1\ \mathbf{X}_2]$ where $\mathbf{X}_1$ is $n \times k_1$ and $\mathbf{X}_2$ is $n \times k_2$. That is, the DGP is $\mathbf{y} = \mathbf{X}_1\beta_1 + \varepsilon$ but we regress $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$. We know that

$$E[\mathbf{b}^*] = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \mathbf{X}^{*\prime}\mathbf{X}_1\beta_1.$$

The k^{th} column of $(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \mathbf{X}^{*\prime}\mathbf{X}_1$ is the vector of least-squares coefficients from the regression of the k^{th} column of $\mathbf{X}_1$ on all columns of $\mathbf{X}_1$ and $\mathbf{X}_2$. That is, the coefficient vector from the regression

$$\mathbf{x}_k = \beta_0^k + \beta_1^k\mathbf{x}_1 + \beta_2^k\mathbf{x}_2 + \cdots + \beta_k^k\mathbf{x}_k + \cdots + \beta_{k_1}^k\mathbf{x}_{k_1} + \tilde{\beta}^k\mathbf{X}_2 + \varepsilon$$

where $\mathbf{x}_k$ is the k^{th} column of $\mathbf{X}_1$. In this regression, $\beta_k^k = 1$ and all other coefficients are zero. The fit is perfect. Therefore

$$E[\mathbf{b}^*] = \begin{bmatrix} \beta_1 \\ \mathbf{0} \end{bmatrix}.$$

Let's check this.

Proposition 3 $\mathbf{b}^*$ *is unbiased.*

Proof.

$$E\begin{bmatrix}\beta_1^*\\ \beta_2^*\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1\\ \mathbf{X}_2'\mathbf{X}_1\end{bmatrix}\beta_1$$

Applying the partitioned inverse formula,

$$\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1} = \begin{bmatrix}(\mathbf{X}_1'\mathbf{X}_1)^{-1}\left(\mathbf{I}_{k_1}+\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\right) & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\\ -\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{D}\end{bmatrix}$$

where $\mathbf{D} = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}$. Working through the matrix multiplication gives

$$\begin{aligned}E\begin{bmatrix}\beta_1^*\\ \beta_2^*\end{bmatrix} &= \begin{bmatrix}\mathbf{I}_{k_1}+(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1-(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1\\ -\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1+\mathbf{D}\mathbf{X}_2'\mathbf{X}_1\end{bmatrix}\beta_1\\ &= \begin{bmatrix}\mathbf{I}_{k_1}+(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1-(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1\\ -\mathbf{D}\mathbf{X}_2'\mathbf{X}_1+\mathbf{D}\mathbf{X}_2'\mathbf{X}_1\end{bmatrix}\beta_1\\ &= \begin{bmatrix}\mathbf{I}_{k_1}\\ \mathbf{0}\end{bmatrix}\beta_1.\end{aligned}$$

■

The following Proposition follows immediately from our earlier work on constrained estimators. Think of the model that includes the irrelevant variables $\mathbf{X}_2$ as the unrestricted model, and the model that (correctly) excludes them as the restricted model.

Proposition 4 *Including the irrelevant variables increases the sampling variance of the coefficients on* $\mathbf{X}_1$. *That is,* $Var[\mathbf{b}^*] \geq Var\left[\hat{\beta}_1\right]$, *where* $\hat{\beta}_1$ *is the least squares estimator of* β_1 *in the model* $\mathbf{y} = \mathbf{X}_1\beta_1 + \varepsilon$.

Proof. We know

$$\begin{aligned}Var[\mathbf{b}^*] &= \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = \sigma^2\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\\ Var\left[\hat{\beta}_1\right] &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.\end{aligned}$$

Applying the partitioned inverse formula again, we see the upper-left $k_1 \times k_1$ block of $(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$ is $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\left(\mathbf{I}+\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\right)$. Therefore, the variance of the first k_1 coefficients (i.e., the coefficients on $\mathbf{X}_1$) in the regression including the irrelevant variables is $\sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\left(\mathbf{I}+\mathbf{X}_1'\mathbf{X}_2\mathbf{D}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\right) \geq \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ (why?). ■

Estimating σ^2

Verify for yourself that $\mathbf{e}^* = \mathbf{M}^*\mathbf{y} = \mathbf{M}^*\varepsilon$. Then under normality

$$\frac{\mathbf{e}^{*\prime}\mathbf{e}^*}{\sigma^2} = \frac{\varepsilon'\mathbf{M}^*\varepsilon}{\sigma^2} \sim \chi^2_{n-k_1-k_2}.$$

So if we unnecessarily include regressors $\mathbf{M}_2$ we have the following unbiased estimator of σ^2:

$$s^{*2} = \frac{\mathbf{e}^{*\prime}\mathbf{e}^*}{n-k_1-k_2}.$$

Heteroskedasticity

Suppose that our assumption of spherical errors is violated, so that

$$Var\left[\mathbf{y}\right] = Var\left[\varepsilon\right] = \mathbf{\Sigma} \neq \sigma^2 \mathbf{I}_n$$

where the matrix $\mathbf{\Sigma}$ is diagonal. Is the least squares estimator still unbiased? Is it BLUE?

Proposition 5 *Under heteroskedasticity,*

$$Var\left[\hat{\beta}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{1}$$

Proof.

$$\begin{aligned} Var\left[\hat{\beta}\right] &= E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

■

Notice that when

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}$$

the middle term $\mathbf{X}'\mathbf{\Sigma}\mathbf{X} = \sum_i \sigma_i^2 \mathbf{x}_i\mathbf{x}_i' = E\left[\sum_i \varepsilon_i^2 \mathbf{x}_i\mathbf{x}_i'\right]$ where $\mathbf{x}_i'$ is the i^{th} row of $\mathbf{X}$. This suggests using $\sum_i e_i^2\mathbf{x}_i\mathbf{x}_i'$ as an estimator of this term. This is the basis of the White (1980) heteroskedasticity-consistent estimator

$$\widehat{Var}\left[\hat{\beta}\right]_{White} = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_i e_i^2\mathbf{x}_i\mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

The formal justification of this estimator is asymptotic (we'll get there soon), but the intuition is pretty obvious even without knowing any asymptotic theory.

Testing for Heteroskedasticity

It is good practice to plot the estimated residuals against regressors x_i. If the magnitude of the residuals appears to vary with any regressor, then a formal test is in order.

1. **Goldfeld-Quandt Test**. Suppose we suspect that σ_i^2 varies with x_i, where x_i is some regressor in our model (in a time series setting, it can be time). If we rank the observations on the basis of x_i, we can separate the observations into groups having high and low variances. The intuition behind the test is to test whether the variances are (very) different in the two groups. For example, suppose n is even. If we reorder

the data on the basis of x_i, and if we could observed ε, then under the null of constant variance σ^2 (and normality)

$$\frac{\frac{1}{\sigma^2}\left(\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_{n/2}^2\right)/(n/2)}{\frac{1}{\sigma^2}\left(\varepsilon_{(n/2)+1}^2+\varepsilon_{(n/2)+2}^2+\cdots+\varepsilon_n^2\right)/(n/2)}=\frac{\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_{n/2}^2}{\varepsilon_{(n/2)+1}^2+\varepsilon_{(n/2)+2}^2+\cdots+\varepsilon_n^2}\sim F_{n/2,n/2}$$

could be the basis of our test. If the variances in the two groups were very different, the distribution of the test statistic would differ greatly from the $F_{n/2,n/2}$. However we don't observe ε. The temptation of course is to use $\mathbf{e}$, but we can't because the first $n/2$ residuals are not independent of the last $n/2$ residuals. The Goldfeld-Quandt "trick" is to separate the observations into two groups as above, but run separate regressions on $\mathbf{X}$ in each group. This gives independent estimates of the residual variances. Let $\mathbf{e}_1$ denote the vector of residuals obtained from the regression on the first group, and $\mathbf{e}_2$ denote the residuals in the second group. The test statistic is

$$\frac{\mathbf{e}_1'\mathbf{e}_1/(n_1-k)}{\mathbf{e}_2'\mathbf{e}_2/(n_2-k)}\sim F_{n_1-k,n_2-k}$$

where n_1 and n_2 are the number of observations in the two groups. The test turns out to be more powerful if some observations in the "middle" of the sample are excluded. Of course, the more observations you drop the smaller are the degrees of freedom of the test, which in turn reduces its power. A good rule of thumb is to omit the middle 1/3 observations, although the optimal number will vary in application.
The problem with the Goldfeld-Quandt test is, of course, that we need to know the variable x_i that influences heteroskedasticity. What if we think that multiple variables matter?

2. **Breusch-Pagan/Godfrey Test.** This test allows the error variance to vary with a set of p variables $\mathbf{Z}$. The errors are assumed to be independently normally distributed with variance $\sigma_i^2=h(\alpha_0+\mathbf{z}_i'\alpha)$ where $\mathbf{z}_i'$ is the i^{th} row of $\mathbf{Z}$, and h is an unknown function. When $\alpha=\mathbf{0}$, the errors are homoskedastic. Some of the variables in $\mathbf{Z}$ can be the same as in $\mathbf{X}$. The procedure is to estimate the model of interest, collect the residuals $\mathbf{e}$, then regress e_i^2/σ_{ML}^2 on $\mathbf{Z}$ and a constant, and compute the explained sum of squares (ESS). Then under the null of homoskedasticity (and normality),

$$\frac{1}{2}ESS\overset{a}{\sim}\chi_{p-1}^2$$

where $\overset{a}{\sim}$ means "is asymptotically distributed" (we'll see lots of this starting next lecture). The factor $\frac{1}{2}$ appears because under normality $Var\left[\varepsilon_i^2/\sigma^2\right]=2$ (that is, $E\left[\varepsilon_i^4\right]=2\sigma^4$). An asymptotically equivalent but computationally simpler alternative is to regress e_i^2 on $\mathbf{Z}$, and use the alternative test statistic $nR^2\overset{a}{\sim}\chi_{p-1}^2$ under the null.
This test is quite sensitive to the assumption of normality, and so Koenker (1981) and others suggest estimating the variance of e_i^2 directly by

$$V=\frac{1}{n}\sum_i\left(e_i^2-\sigma_{ML}^2\right)^2$$

and weighting the test statistic by $\frac{1}{V}$ instead of $\frac{1}{2}$.

Multicollinearity

Multicollinearity is just the problem that arises when (some of) the regressors $\mathbf{X}$ are highly correlated. This makes it difficult to identify their separate effects on $\mathbf{y}$. Many economic data move together, and hence multicollinearity problems can be endemic in econometric applications.

When the columns of $\mathbf{X}$ are (near) linear combinations of one another, $\mathbf{X}'\mathbf{X}$ is (nearly) singular. We say that $\mathbf{X}'\mathbf{X}$ is poorly conditioned. In extreme circumstances, this prevents calculation of $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $Cov\left[\hat{\beta}\right] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In less extreme circumstances, the usual symptoms of multicollinearity are: [1] small changes in data lead to large changes in coefficient estimates; [2] estimated coefficients have large standard errors, but are jointly significant; [3] coefficients take implausible values. Symptom [2] is quite easy to see algebraically. Suppose $\mathbf{X}$ contains a constant, all that all other regressors are in deviations from means. Let $\mathbf{x}_k$ denote the k^{th} column of $\mathbf{X}$, let $\mathbf{X}_{(k)}$ denote the other columns of $\mathbf{X}$, and $\mathbf{M}_{(k)} = \mathbf{I}_n - \mathbf{X}_{(k)}\left(\mathbf{X}'_{(k)}\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}'_{(k)}$. The variance of $\hat{\beta}_k$ is the k^{th} diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Using the partitioned inverse formula, you can show that the k^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ is

$$\begin{aligned}
\left(\mathbf{x}'_k\mathbf{M}_{(k)}\mathbf{x}_k\right)^{-1} &= \left[\mathbf{x}'_k\mathbf{x}_k - \mathbf{x}'_k\mathbf{X}_{(k)}\left(\mathbf{X}'_{(k)}\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}'_{(k)}\mathbf{x}_k\right]^{-1} \\
&= \left[\mathbf{x}'_k\mathbf{x}_k\left(1 - \frac{\mathbf{x}'_k\mathbf{X}_{(k)}\left(\mathbf{X}'_{(k)}\mathbf{X}_{(k)}\right)^{-1}\mathbf{X}'_{(k)}\mathbf{x}_k}{\mathbf{x}'_k\mathbf{x}_k}\right)\right]^{-1} \\
&= \left[\mathbf{x}'_k\mathbf{x}_k\left(1 - \frac{\mathbf{x}'_k\left(\mathbf{I} - \mathbf{M}_{(k)}\right)\mathbf{x}_k}{\mathbf{x}'_k\mathbf{x}_k}\right)\right]^{-1} \\
&= \left[\mathbf{x}'_k\mathbf{x}_k\left(1 - \frac{\hat{\mathbf{x}}'_k\hat{\mathbf{x}}_k}{\mathbf{x}'_k\mathbf{x}_k}\right)\right]^{-1} \\
&= \frac{1}{\mathbf{x}'_k\mathbf{x}_k\left(1 - R^2_k\right)}
\end{aligned}$$

where $\hat{\mathbf{x}}_k$ are predicted values from the regression of $\mathbf{x}_k$ on $\mathbf{X}_{(k)}$, and R^2_k is the squared correlation coefficient in this regression. When $\mathbf{x}_k$ is highly correlated with the other regressors $\mathbf{X}_{(k)}$, then R^2_k approaches one, and the variance of $\hat{\beta}_k$ is inflated. In the extreme case where $R^2_k = 1$, its variance is infinite.

Multicollinearity is a controversial topic among textbook authors. One of my professors, the late Art Goldberger, was fond of pointing out that the consequences of multicollinearity are identical to those of having a small number of observations (a situation Art called "micronumerosity"). Outside of the extreme cases of perfect multicollinearity (the matrix $\mathbf{X}'\mathbf{X}$ is singular) or perfect micronumerosity (the sample size is $n = 0$), neither one makes the OLS estimator biased. They just make the standard errors bigger.

There are some commonly discussed (though infrequently applied) "remedies" for multicollinearity. The most palatable is to get more data. This is never a bad idea. A less

palatable suggestion is to exclude some of the highly colinear variables from your model. However, doing so produces omitted variables bias. Note that our previous results imply that omitted variables bias tends to be large precisely when the omitted variable is highly collinear with the other explanatory variables.

There are some purely "technical" solutions, including ridge regression, partial least squares, or principal components regression. None of these should be applied blindly. They all "work" by imposing arbitrary and sometimes hidden information.

Dummy Variables

Dummy variables are regressors that take value 0 or 1. Usually these indicate the absence or presence of a characteristic. Dummy variables allow the intercept and/or slope parameters to differ across groups of observations that share a characteristic. Common examples include dummy variables for male/female, fulltime/parttime worker, a policy is in effect/not, etc.

The **dummy variable trap** is well known. It arises when a set of dummy variables is a linear combination of the intercept.

Example 6 *Let* $\mathbf{x}_1 = \mathbf{1}$ *be the column of ones in* $\mathbf{X}$ *(our intercept). Suppose*

$$x_{2i} = \begin{cases} 1 & \text{if characteristic } A \text{ is present} \\ 0 & \text{if characteristic } A \text{ is not present} \end{cases}$$

$$x_{3i} = \begin{cases} 0 & \text{if characteristic } A \text{ is present} \\ 1 & \text{if characteristic } A \text{ is not present} \end{cases}$$

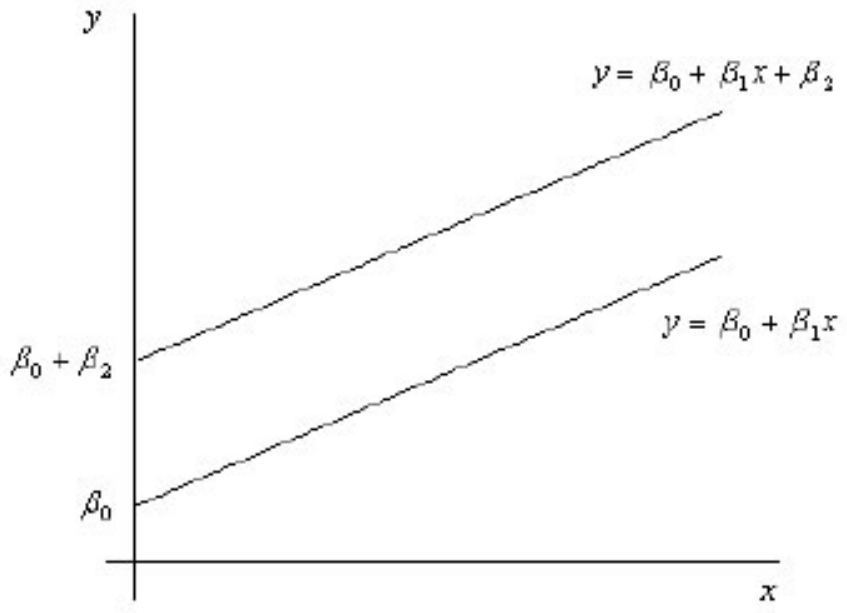

Figure 1: Regression model with a dummy variable: $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \varepsilon_i$

Then $\mathbf{x}_2 + \mathbf{x}_3 = \mathbf{1} = \mathbf{x}_1$. *Then* $\mathbf{X} = [\,\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \mathbf{x}_3]$ *has rank 2 and* $\mathbf{X}'\mathbf{X}$ *is not invertible.*

We can **interact** dummy variables for different characteristics. This allows even greater flexibility in the specification of the intercept.

Example 7 *Suppose* $\mathbf{x}_1 = \mathbf{1}$, $\mathbf{x}_2$ *is a dummy variable for characteristic* A, *and* $\mathbf{x}_3$ *is a dummy variable for characteristic* B. *Define* $\mathbf{x}_4 = \mathbf{x}_2 * \mathbf{x}_3$, *where* $*$ *denotes elementwise multiplication. Then*

$$x_{4i} = \begin{cases} 1 & \text{if characteristics } A \text{ and } B \text{ are present} \\ 0 & \text{if characteristic } A \text{ or } B \text{ is not present.} \end{cases}$$

The effect of having characteristic A *and not* B *is* β_2, *the effect of having characteristic* B *and not* A *is* β_3, *and the effect of having both* A *and* B *is* $\beta_2 + \beta_3 + \beta_4$. *We could of course set this up differently, for example*

$$\begin{aligned} x_{2i} &= 1 \text{ if characteristic } A \text{ is present but characteristic } B \text{ is not} \\ x_{3i} &= 1 \text{ if characteristic } B \text{ is present but characteristic } A \text{ is not} \\ x_{4i} &= 1 \text{ if characteristics } A \text{ and } B \text{ are present.} \end{aligned}$$

Although the two definitions of $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ *will yield different parameter estimates, the correct interpretation of these estimates will give the same estimated (marginal) effects.*

Of course we can also interact dummy variables with continuous regressors. This allows slopes to vary across groups, with or without separate intercepts.

We can also use dummy variables for non-binary characteristics.

Example 8 *Suppose we have a sample of retirees and age is reported in grouped form:* < 60 *years;* $60 - 70$ *years;* $70+$ *years. How should we code dummy variables for age? One temptation is to code*

$$d = \begin{cases} 0 & \text{if } < 60 \text{ years of age} \\ 1 & \text{if } 60 - 70 \text{ years of age} \\ 2 & \text{if } 70 + \text{ years of age} \end{cases}$$

This is restrictive and probably unwise. It imposes that the effect of being $70+$ *years of age is twice the effect of being* $60 - 70$ *years of age. A better way would be to use two dummy variables*

$$\begin{aligned} d_1 &= \begin{cases} 1 & \text{if } < 60 \text{ years of age} \\ 0 & \text{otherwise} \end{cases} \\ d_2 &= \begin{cases} 1 & \text{if } 60 - 70 \text{ years of age} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

which is much more flexible. We can't include a third dummy (say d_3*) for the category* $70+$ *years of age without falling into the dummy variable trap. Instead, the effect of age is measured relative to being* $70+$ *years of age. We call this the excluded category or reference category. All we are doing is imposing a linear restriction on the dummies (or equivalently their coefficients) to break any exact linear relationship between them and the intercept. That is, we implicitly impose* $\beta_3 = 0$. *The estimated effects* β_1 *and* β_2 *are measured relative to the reference category. Other linear restrictions will do the trick equally well – we just need one for each set of colinear variables. For example, we could impose instead that* $\beta_1 + \beta_2 + \beta_3 = 0$.

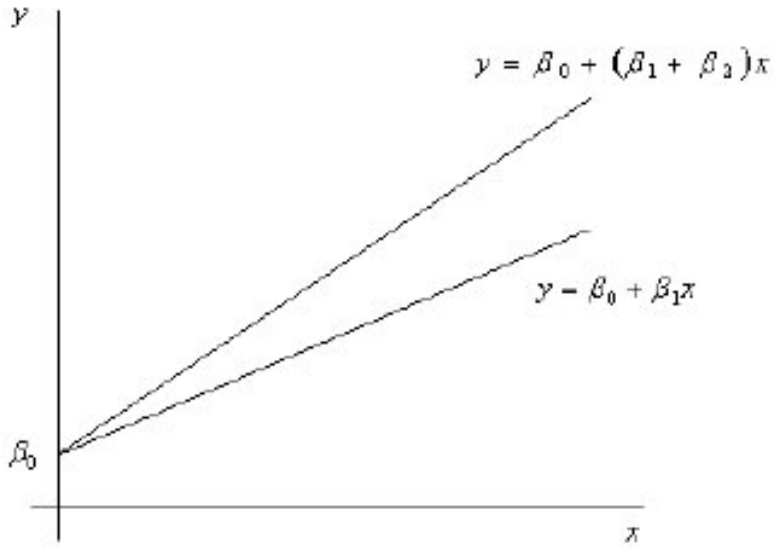

Figure 2: Regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i x_i + \varepsilon_i$

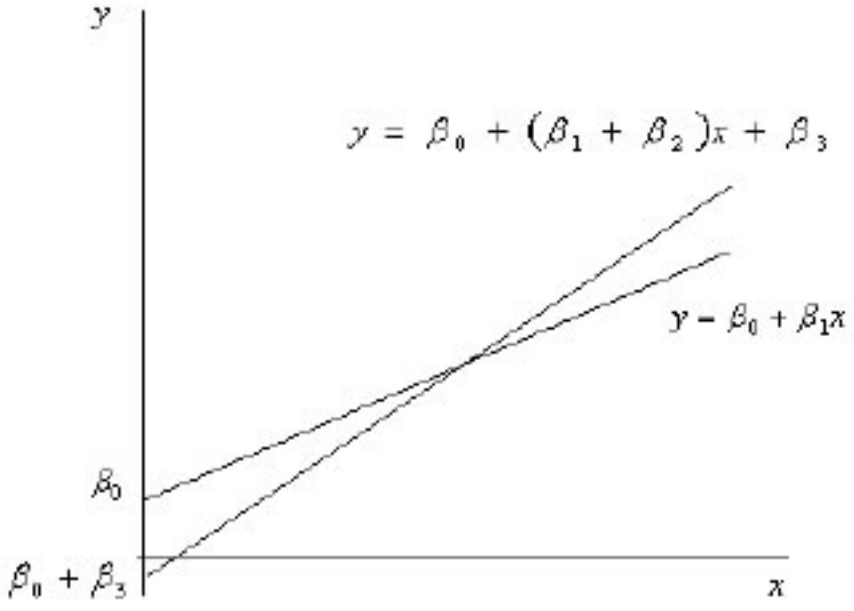

Figure 3: Regression model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i x_i + \beta_3 d_i + \varepsilon_i$