CHAPTER 14

More Than Two Explanatory Variables

14.0 What We Need to Know When We Finish This Chapter

The addition of a second explanatory variable in chapter 11 adds only four new things to what there is to know about regression. First, regression uses only the parts of each variable that are unrelated to all of the other variables. Second, omitting a variable from the sample relationship that appears in the population relationship almost surely biases our estimates. Third, including an irrelevant variable does not bias estimates but reduces their precision. Fourth, the number of interesting joint tests increases with the number of slopes. All four remain valid when we add additional explanatory variables.

1. **Equations (14.11), (14.1), and (14.2), section 14.2:** The general form of the multivariate population relationship is

$$y_i = \alpha + \sum_{l=1}^{k} \beta_l x_{li} + \varepsilon_i.$$

The corresponding sample relationship is

$$y_i = a + \sum_{l=1}^{k} b_l x_{li} + e_i.$$

The predicted value of y_i is

$$\hat{y}_i = a + \sum_{l=1}^{k} b_l x_{li}.$$

2. **Equations (14.3) and (14.4), section 14.2:** When we minimize the sum of squared errors in the multivariate regression, the errors sum to zero,

$$0 = \sum_{i=1}^{n} e_i,$$

and are uncorrelated in the sample with all explanatory variables

$$0 = \sum_{i=1}^{n} e_i x_{pi} = \text{COV}\left(e_i, x_{pi}\right) \quad \text{for all } p = 1, \ldots, k.$$

3. **Equations (14.5) and (14.8), section 14.2:** The intercept in the multivariate regression is

$$a = \bar{y} - \sum_{l=1}^{k} b_l \bar{x}_l.$$

The slopes are

$$b_p = \frac{\sum_{i=1}^{n}\left(e_{(x_p,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)i} - \bar{e}_{(x_p,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)}\right)\left(e_{(y,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)i} - \bar{e}_{(y,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)}\right)}{\sum_{i=1}^{n}\left(e_{(x_p,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)i} - \bar{e}_{(x_p,x_1\ldots x_{p-1}x_{p+1}\ldots x_k)}\right)^2}.$$

4. **Equations (14.9) and (14.17), section 14.2:** R^2 is

$$R^2 = \text{COV}\left(y_i, \hat{y}_i\right).$$

The adjusted R^2 is

$$\text{adjusted } R^2 = 1 - \frac{s^2}{\text{V}\left(y_i\right)}.$$

5. **Section 14.2:** Regression is not limited to two explanatory variables. However, the number of observations must exceed the number of estimators, and each explanatory variable must have some part that is not related to all of the other explanatory variables in order to calculate meaningful regression estimators.
6. **Equations (14.12) and (14.14), section 14.2:** If the regression is specified correctly, estimators are unbiased:

$$\text{E}\left(a\right) = \alpha \quad \text{and} \quad \text{E}\left(b_p\right) = \beta_p \quad \text{for all } p = 1, \ldots, k.$$

If the regression omits explanatory variables, estimators are biased:

$$\text{E}\left(b_p\right) = \beta_p + \sum_{m=k-q}^{k} \left[\beta_m \frac{\sum_{i=1}^{n} e_{\left(x_p, x_1 \ldots x_{p-1} x_{p+1} \ldots x_{k-q-1}\right)i} e_{\left(x_m, x_1 \ldots x_{p-1} x_{p+1} \ldots x_{k-q-1}\right)i}}{\sum_{i=1}^{n} e^2_{\left(x_p, x_1 \ldots x_{p-1} x_{p+1} \ldots x_{k-q-1}\right)i}} \right].$$

7. **Equations (14.16), (14.19), and (14.20), section 14.3:** The estimator of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}.$$

With this estimator, the standardized value of b_p is a t random variable with $n - k - 1$ degrees of freedom:

$$\frac{b_p - \beta_p}{\sqrt{\dfrac{s^2}{\sum_{i=1}^{n} e^2_{\left(x_p, x_1 \ldots x_{p-1} x_{p+1} \ldots x_k\right)i}}}} \sim t^{\left(n-k-1\right)}.$$

If n is sufficiently large, this can be approximated as a standard normal random variable. If the sample regression is correctly specified, b_p is the best linear unbiased (BLU) estimator of β_p.

8. **Equations (14.21) and (14.22), section 14.3:** The general form of the test between an unrestricted alternative hypothesis and a null hypothesis subject to j restrictions is

$$\frac{\left(\dfrac{\left(\sum_{i=1}^{n} e_i^2\right)_{\mathrm{R}} - \left(\sum_{i=1}^{n} e_i^2\right)_{\mathrm{U}}}{j}\right)}{\left(\dfrac{\left(\sum_{i=1}^{n} e_i^2\right)_{\mathrm{U}}}{n-k-1}\right)} \sim F(j, n-k-1).$$

For the null hypothesis that all coefficients are equal to zero, this reduces to

$$\frac{R_{\mathrm{U}}^2}{1-R_{\mathrm{U}}^2}\frac{n-k-1}{k} \sim F\left(k, n-k-1\right).$$

9. **Equation (14.26), section 14.3:** The Chow test for differences in regimes is

$$\frac{\left(\dfrac{\left(\sum_{i=1}^{n} e_i^2\right)_{\mathrm{R}} - \left(\left(\sum_{i=1}^{n} e_i^2\right)_{x_{1i}=0} + \left(\sum_{i=1}^{n} e_i^2\right)_{x_{1i}=1}\right)}{k-1}\right)}{\left(\dfrac{\left(\sum_{i=1}^{n} e_i^2\right)_{x_{1i}=0} + \left(\sum_{i=1}^{n} e_i^2\right)_{x_{1i}=1}}{n-2k}\right)} \sim F(k-1, n-2k).$$

10. **Section 14.6:** If omitted explanatory variables are fixed for each entity in the sample, multiple observations on each entity may allow us to use *panel data* estimation techniques. These techniques can purge the effects of unobserved heterogeneity and yield unbiased estimators for the effects of included explanatory variables.