with $\boldsymbol{\beta}_{20} \neq 0$. Then it is easy to see that the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ will, in general, be biased. Under this DGP,

$$\begin{aligned} E(\tilde{\boldsymbol{\beta}}_1) &= E\Big(\big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{y}\Big) \\ &= E\Big(\big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1^{\top}\big(\boldsymbol{X}_1\boldsymbol{\beta}_{10} + \boldsymbol{X}_2\boldsymbol{\beta}_{20} + \boldsymbol{u}\big)\Big) \\ &= \boldsymbol{\beta}_{10} + \big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\beta}_{20}. \end{aligned} \tag{3.57}$$

Unless $\boldsymbol{X}_1^{\top}\boldsymbol{X}_2$ is a zero matrix or $\boldsymbol{\beta}_{20}$ is a zero vector, $\tilde{\boldsymbol{\beta}}_1$ will be a biased estimator. The magnitude of the bias will depend on the matrices $\boldsymbol{X}_1^{\top}\boldsymbol{X}_1$ and $\boldsymbol{X}_1^{\top}\boldsymbol{X}_2$ and the vector $\boldsymbol{\beta}_{20}$.

Results very similar to (3.57) are available for all types of restrictions, not just for linear restrictions, and for all sorts of models in addition to linear regression models. We will not attempt to deal with nonlinear models here because that requires a good deal of technical apparatus, which will be developed in Chapter 12. Results analogous to (3.57) for nonlinear regression models and other types of nonlinear models may be found in Kiefer and Skoog (1984). The important point is that imposition of false restrictions on some of the parameters of a model generally causes all of the parameter estimates to be biased. This bias does not go away as the sample size gets larger.

Even though $\tilde{\boldsymbol{\beta}}_1$ is biased when the DGP is (3.56), it is still of interest to ask how well it performs. The analog of the covariance matrix for a biased estimator is the **mean squared error matrix**, which in this case is

$$\begin{aligned} & E\big(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\big)\big(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\big)^{\top} \\ &= E\big((\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}(\boldsymbol{X}_2\boldsymbol{\beta}_{20} + \boldsymbol{u})\big)\big((\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}(\boldsymbol{X}_2\boldsymbol{\beta}_{20} + \boldsymbol{u})\big)^{\top} \\ &= \sigma_0^2\big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1} + \big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\boldsymbol{\beta}_{20}\boldsymbol{\beta}_{20}^{\top}\boldsymbol{X}_2^{\top}\boldsymbol{X}_1\big(\boldsymbol{X}_1^{\top}\boldsymbol{X}_1\big)^{-1}. \end{aligned} \tag{3.58}$$

The third line here is the sum of two matrices: the covariance matrix of $\tilde{\boldsymbol{\beta}}_1$ when the DGP satisfies the restrictions, and the outer product of the second term in the last line of (3.57) with itself. It is possible to compare (3.58) with $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$, the covariance matrix of the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$, only if σ_0 and $\boldsymbol{\beta}_{20}$ are known. Since the first term of (3.58) is smaller in the matrix sense than $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$, it is clear that if $\boldsymbol{\beta}_{20}$ is small enough (3.58) will be smaller than $\boldsymbol{V}(\hat{\boldsymbol{\beta}}_1)$. Thus it may be desirable to use the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ when the restrictions are false, provided they are not too false.

Applied workers frequently find themselves in a situation like the one we have been discussing. They want to estimate $\boldsymbol{\beta}_1$ and do not know whether or not $\boldsymbol{\beta}_2 = 0$. It then seems natural to define a new estimator,

$$\check{\boldsymbol{\beta}}_1 = \begin{cases} \tilde{\boldsymbol{\beta}}_1 & \text{if } F_{\boldsymbol{\beta}_2=\mathbf{0}} < c_\alpha; \\ \hat{\boldsymbol{\beta}}_1 & \text{if } F_{\boldsymbol{\beta}_2=\mathbf{0}} \geq c_\alpha. \end{cases}$$

Here $F_{\boldsymbol{\beta}_2=\mathbf{0}}$ is the usual F test statistic for the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$, and c_α is the critical value for a test of size α given by the $F(r, n-k)$ distribution. Thus $\check{\boldsymbol{\beta}}_1$ will be the restricted estimator $\tilde{\boldsymbol{\beta}}_1$ when the F test does not reject the hypothesis that the restrictions are satisfied and will be the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$ when the F test does reject that hypothesis. It is an example of what is called a **preliminary test estimator** or **pretest estimator**.

Pretest estimators are used all the time. Whenever we test some aspect of a model's specification and then decide, on the basis of the test results, what version of the model to estimate or what estimation method to use, we are employing a pretest estimator. Unfortunately, the properties of pretest estimators are, in practice, very difficult to know. The problems can be seen from the example we have been studying. Suppose the restrictions hold. Then the estimator we would like to use is the restricted estimator, $\tilde{\boldsymbol{\beta}}_1$. But, $\alpha\%$ of the time, the F test will incorrectly reject the null hypothesis and $\check{\boldsymbol{\beta}}_1$ will be equal to the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$ instead. Thus $\check{\boldsymbol{\beta}}_1$ must be less efficient than $\tilde{\boldsymbol{\beta}}_1$ when the restrictions do in fact hold. Moreover, since the estimated covariance matrix reported by the regression package will not take the pretesting into account, inferences about $\check{\boldsymbol{\beta}}_1$ may be misleading.

On the other hand, when the restrictions do not hold, we may or may not want to use the unrestricted estimator $\hat{\boldsymbol{\beta}}_1$. Depending on how much power the F test has, $\check{\boldsymbol{\beta}}_1$ will sometimes be equal to $\tilde{\boldsymbol{\beta}}_1$ and sometimes be equal to $\hat{\boldsymbol{\beta}}_1$. It will certainly not be unbiased, because $\tilde{\boldsymbol{\beta}}_1$ is not unbiased, and it may be more or less efficient (in the sense of mean squared error) than the unrestricted estimator. Inferences about $\check{\boldsymbol{\beta}}_1$ based on the usual estimated OLS covariance matrix for whichever of $\tilde{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_1$ it turns out to be equal to may be misleading, because they fail to take into account the pretesting that occurred previously.

In practice, there is often not very much that we can do about the problems caused by pretesting, except to recognize that pretesting adds an additional element of uncertainty to most problems of statistical inference. Since α, the level of the preliminary test, will affect the properties of $\check{\boldsymbol{\beta}}_1$, it may be worthwhile to try using different values of α. Conventional significance levels such as .05 are certainly not optimal in general, and there is a literature on how to choose better ones in specific cases; see, for example, Toyoda and Wallace (1976). However, real pretesting problems are much more complicated than the one we have discussed as an example or the ones that have been studied in the literature. Every time one subjects a model to any sort of test, the result of that test may affect the form of the final model, and the implied pretest estimator therefore becomes even more complicated. It is hard to see how this can be analyzed formally.

Our discussion of pretesting has been very brief. More detailed treatments may be found in Fomby, Hill, and Johnson (1984, Chapter 7), Judge, Hill, Griffiths, Lütkepohl, and Lee (1985, Chapter 21), and Judge and Bock (1978). In the remainder of this book, we entirely ignore the problems caused