Notes on Chap. 5: More on Multiple Regression

§5.1. Introduction.

As in chapter 4, we have an $n \times p$ design matrix X of rank p and a response vector $Y \in \mathbb{R}^n$. OLS writes a relationship between X and Y as $Y = X\hat{\beta} + e$ with $e \perp X$. A model for this is $Y = X\beta + \epsilon$.

In this chapter, we discuss another way that OLS is best. Then we weaken some assumptions to discuss GLS, where "G" stands for "generalized". The most important parts come last, where we discuss statistical hypothesis testing. There will be t-statistics again, as well as new ones called F-statistics. In the handout on t-distributions, we found the distribution of t, provided that the samples came from a normal distribution. Likewise, these t- and F-tests depend on stronger assumptions, namely, that error terms are IID normal random variables.

§5.2. OLS is BLUE.

In chapter 4, we saw that OLS is best in giving the best possible approximation (of a certain form) to the data; this was theorem 4.1. We also saw that OLS was good in that it provided unbiased estimates of the parameters (with appropriate statistical assumptions on the model). In this section, we show that (again with assumptions) OLS is best in its class. What do we mean? First, we are talking about a property of $\hat{\beta}$, the OLS estimate of β. Second, we are going to say that $\hat{\beta}$ is best among all estimators of β that are unbiased. But actually, we are going to restrict our class of estimators a little more. Recall that $\hat{\beta} = QY$, where $Q := (X'X)^{-1}X'$. That is, supposing X is fixed (not random), $\hat{\beta}$ is a linear transformation of Y. The key aspect here is that it is linear in Y, not some more complicated function of Y. (It is indeed a complicated function of X, but we are treating X as fixed, like a constant.) Thus, $\hat{\beta}$ is an LUE of β, where "LUE" stands for "linear unbiased estimator". We are going to prove it is best in the class of all LUEs, so we will say it is BLUE, where "B" stands for "best". But what do we mean by "best"?

This is a little complicated. After all, $\hat{\beta}$ is a random vector. We want to capture the idea that $\hat{\beta}$ has the smallest variance in the entire class LUE. Of course, as a vector, $\hat{\beta}$ does not have a variance. It does have a covariance matrix, but what would it mean to

say it is smaller than another covariance matrix? This can indeed be done, but we will do it another way that is more concrete.

Namely, we use $\hat\beta$ to estimate things like β_1, β_2, $\beta_1-\beta_2$, $3\beta_4+2\beta_6$, etc. All these are linear combinations of coordinates of β. In general, a linear combination of coordinates of β is a number of the form $\sum_{i=1}^p c_i\beta_i = c'\beta$, where c is a column vector $[c_1\ \cdots\ c_p]'$. Since we estimate β by $\hat\beta$, we naturally estimate $c'\beta$ by $c'\hat\beta$. For example, we estimate β_1 by $\hat\beta_1$; this corresponds to using $c = [1\ 0\ \cdots\ 0]'$. Likewise, we estimate $\beta_1-\beta_2$ by $\hat\beta_1-\hat\beta_2$, which corresponds to $c = [1\ -1\ 0\ \cdots\ 0]'$. Not only is the estimate $c'\hat\beta$ natural, but we will show it is the best estimate of $c'\beta$ in its class.

So fix some $c\in\mathbb{R}^p$. We are going to focus on estimating $c'\beta$ instead of estimating β. Note that $E(c'\hat\beta) = c'E(\hat\beta) = c'\beta$, so $c'\hat\beta$ is an unbiased estimate of $c'\beta$. Also, $c'\hat\beta$ is a linear estimator of $c'\beta$ since

$$c'\hat\beta = c'QY\,. \tag{N1}$$

Note that $c'Q$ is a row vector, so $c'\hat\beta$ is a linear combination of the coordinates of Y. That is, $c'Q$ gives the OLS estimator of $c'\beta$ by multiplying Y on the left. In general, a *linear estimator* has the form $\sum_{i=1}^n d_iY_i = d'Y$ for some column vector $d\in\mathbb{R}^n$. (We replace $c'Q$ in (N1) by a general d'.) It would be *unbiased* if $E(d'Y) = c'\beta$. Remember we are thinking of c as fixed; we might be estimating β_1, so we would say that $d'Y$ is an unbiased estimator of β_1 if $E(d'Y)=\beta_1$.

Now $c'\hat\beta$ is a random variable, not a random vector. So $c'\hat\beta$ does have a variance. It makes sense to ask whether $\mathrm{Var}(c'\hat\beta)$ is smallest among all LUEs of $c'\beta$. It does! This is why $c'\hat\beta$ is called *best*. We state and prove this formally in the next theorem, called the *Gauss-Markov theorem*.

Theorem 5.1. *Suppose that X is fixed, $E(\epsilon)=\mathbf{0}_n$, and $\mathrm{Cov}(\epsilon)=\sigma^2 I_n$ with $\sigma^2>0$. Consider any $c\in\mathbb{R}^p$. Then for every $d\in\mathbb{R}^n$ that has the property that $E(d'Y)=c'\beta$ for all β, we have that $\mathrm{Var}(c'\hat\beta)\le\mathrm{Var}(d'Y)$ with equality iff $d'=c'Q$. Here, $Y=X\beta+\epsilon$ and d can depend on X but not on β.*

Proof. First, what is $\mathrm{Var}(d'Y)$? Since $d'X\beta$ is constant, we have

$$\mathrm{Var}(d'Y) = \mathrm{Var}\big(d'(X\beta+\epsilon)\big) = \mathrm{Var}(d'\epsilon) = \mathrm{Cov}(d'\epsilon) = d'\,\mathrm{Cov}(\epsilon)d = d'\sigma^2 I_n d = \sigma^2\|d\|^2\,. \tag{N2}$$

We used the fact that the covariance matrix of a random variable is the same as its variance, in order to use what we know about random vectors and their covariance matrices.

We next look at what it means that $d'Y$ is unbiased, i.e., that $E(d'Y)=c'\beta$ for all β. Substitute $Y=X\beta+\epsilon$ here to get $E(d'Y) = d'X\beta + d'E(\epsilon) = d'X\beta$, so that $d'X\beta = c'\beta$.

Remember that this holds for all β. It follows that

$$d'XQ = c'Q \tag{N3}$$

because we can choose for β any column of Q; the matrix multiplication on each side in (N3) works column by column on Q. Recall that $H = XQ$ is the matrix of the orthogonal projection P_W, where $W := \mathrm{col}(X)$. In particular, H is symmetric, so taking the transpose of both sides of (N3) gives $Hd = Q'c$, i.e., $P_W(d) = Q'c$. This implies that $\|d\| \geq \|Q'c\|$ with equality iff $d = Q'c$ iff $d' = c'Q$.

Therefore, (N2) gives that $\mathrm{Var}(d'Y) \geq \sigma^2\|Q'c\|^2 = \mathrm{Var}(c'QY) = \mathrm{Var}(c'\hat{\beta})$ with equality iff $d' = c'Q$. That is, the minimum variance is indeed achieved uniquely at the OLS estimator of $c'\beta$. ∎

Optional exercise: Show that the assumption in theorem 5.1 that $\sigma^2 > 0$ is not needed; the conclusion is true even when $\sigma = 0$.

Optional exercise: The Gauss-Markov theorem is sometimes stated in a more elegant, though weaker, form. To understand it, we must return to the question left unanswered above, namely, how do we compare two covariance matrices? Recall that covariance matrices are non-negative definite, and that such matrices are like non-negative real numbers. We say that one number is larger than another if their difference is positive. Likewise, we say that one matrix A is at least as large as another B if their difference is non-negative definite, written $A \geq B$ if $A - B \geq 0$. Recall that this means: for all vectors v, we have $v'(A-B)v \geq 0$. This inequality is the same as $v'Av \geq v'Bv$. Prove the following: Suppose that X is fixed, $E(\epsilon) = \mathbf{0}_n$, and $\mathrm{Cov}(\epsilon) = \sigma^2 I_n$ with $\sigma^2 > 0$. Then for every $n \times n$ matrix M with the property that $E(MY) = \beta$ for all β, we have that $\mathrm{Cov}(\hat{\beta}) \leq \mathrm{Cov}(MY)$ with equality iff $M = Q$. Here, $Y = X\beta + \epsilon$ and M can depend on X but not on β. Hint: What does $c'MY$ estimate?

---

§5.3. Generalized least squares.

Here we loosen the assumption on the covariance matrix of ϵ. This is natural when we think that not all subjects have the same variability, or when we think there is some dependence among the subject measurements. This is common in practice. A simple and beautiful mathematical example is in exercise 5C1. (That example is a special case of the *heteroscedastic regression model*, which is where G is an unknown diagonal matrix.) A consequence of our changed assumptions will be that the OLS estimator will not be BLUE.

Our assumptions in this section will be that

$$E(\epsilon \mid X) = \mathbf{0}_n \text{ and } \operatorname{Cov}(\epsilon \mid X) = G\,,$$

where G is a positive definite $n \times n$ matrix. This is called the *GLS regression model*. Before, we assumed that $G = \sigma^2 I_n$. Allowing G to be non-0 off the diagonal means that we allow correlations among the ϵ_i; allowing G to be non-constant on the diagonal means that we allow ϵ_i to have different variances. Recall that every covariance matrix is non-negative definite. We are assuming a little more: it is positive definite. If G were not positive definite, then there would be a linear relation among the ϵ_i, which would be strange.

Optional exercise: Show that if G is the covariance matrix of a random vector U and G is not positive definite, then there exists some constant vector $c \neq \mathbf{0}$ and some real number m such that $c'U = m$ with probability 1. Hint: G has a non-0 null space.

Let $\hat{\beta}_{\text{OLS}}$ denote the OLS estimator of β. Now theorem 4.2 tells us that

$$E(\hat{\beta}_{\text{OLS}} \mid X) = \beta\,,$$

so still $\hat{\beta}_{\text{OLS}}$ is unbiased. However, theorem 4.3 no longer applies since we changed our assumption. Instead, the first line of calculation in the proof of theorem 4.3 gives us that

$$\operatorname{Cov}(\hat{\beta}_{\text{OLS}} \mid X) = Q \operatorname{Cov}(\epsilon \mid X) Q' = QGQ'\,; \tag{N4}$$

recall that $Q = (X'X)^{-1}X'$.

Although $\hat{\beta}_{\text{OLS}}$ is still an LUE of β, the hypotheses of the Gauss-Markov theorem no longer hold and it is generally the case that $\hat{\beta}_{\text{OLS}}$ is not BLUE. That means that there is a better LUE of β. Of course, we can still use $\hat{\beta}_{\text{OLS}}$; but what reasons are there for using

a better LUE? First, one would like to get a smaller error in one's estimate of anything. Second, when we look at statistical hypothesis testing, as in FPP, the error is directly related to the P-value, with a larger error leading to a larger P-value. Therefore, one might not reject the null hypothesis when one should. This means that one might miss a discovery, or that one might miss the opportunity to publish. Therefore, people often use a different estimator than $\hat{\beta}_{\text{OLS}}$ for the GLS model.

How do we get a better estimator? Recall from exercise 3D7 that there is a matrix $G^{-1/2}$. Take our model $Y = X\beta + \epsilon$ and multiply both sides by $G^{-1/2}$. We get

$$\underbrace{G^{-1/2}Y}_{\text{new response}} = \underbrace{G^{-1/2}X}_{\text{new design}}\,\beta + \underbrace{G^{-1/2}\epsilon}_{\text{new error}}\,. \tag{N5}$$

This looks complicated, but the calculations are easily done by MATLAB. However, we need to check some assumptions. The parameter vector is still β. The new design matrix has full rank because if $c \in \mathbb{R}^p$ has the property that $G^{-1/2}Xc = \mathbf{0}_n$, then $(Xc)'G^{-1/2}Xc = 0$; because $G^{-1/2}$ is positive definite (exercise 3D7), it follows that $Xc = \mathbf{0}_n$ by definition of "positive definite"; because X has full rank, it follows that $c = \mathbf{0}_p$; this means that the new design matrix has full rank. The new error term has

$$E(G^{-1/2}\epsilon \mid X) = G^{-1/2}E(\epsilon \mid X) = G^{-1/2}\mathbf{0}_n = \mathbf{0}_n$$

and

$$\operatorname{Cov}(G^{-1/2}\epsilon \mid X) = G^{-1/2}\operatorname{Cov}(\epsilon \mid X)G^{-1/2} = G^{-1/2}GG^{-1/2} = G^{-1/2}G^{1/2}G^{1/2}G^{-1/2} = I_n\,.$$

Thus, the assumptions of theorem 4.4 and of theorem 5.1 hold with $\sigma = 1$ for the transformed equation (N5).

That means the conclusions hold for the OLS estimator of (N5), which we will call $\hat{\beta}_{\text{GLS}}$. In order to calculate $\hat{\beta}_{\text{GLS}}$, we apply the formula from LinAlg#3 to (N5):

$$\hat{\beta}_{\text{GLS}} := \left[(G^{-1/2}X)'(G^{-1/2}X)\right]^{-1}(G^{-1/2}X)'G^{-1/2}Y\,.$$

Let's simplify this mess. We have $(G^{-1/2}X)' = X'G^{-1/2}$ and $G^{-1/2}G^{-1/2} = G^{-1}$. Thus, we get

$$\hat{\beta}_{\text{GLS}} = (X'G^{-1}X)^{-1}X'G^{-1}Y\,. \tag{N6}$$

By theorem 4.2, we have

$$E(\hat{\beta}_{\text{GLS}} \mid X) = \beta \tag{N7}$$

and by theorem 4.3, we have

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}} \mid X) = (X'G^{-1}X)^{-1} \tag{N8}$$

since $\left[(G^{-1/2}X)'(G^{-1/2}X)\right]^{-1} = (X'G^{-1}X)^{-1}$ and since $\sigma = 1$ for (N5). Furthermore, if X is fixed, then by theorem 5.1, $\hat{\beta}_{\mathrm{GLS}}$ is BLUE. (Note that $\hat{\beta}_{\mathrm{GLS}}$ is a LUE by (N6) and (N7).)

Exercise: If X has an intercept, does the transformed model (N5) also have an intercept?

Everything is looking good, even if messy. However, there's a catch: usually we don't know G, just like in OLS we usually don't know σ. That wasn't much of a problem for OLS. But it is a serious problem for GLS. After all, G is $n \times n$, so not knowing G means not knowing the n^2 entries of G. Since G is symmetric, this really means not knowing $n(n-1)/2 + n = n(n+1)/2$ numbers. At the same time, the data, X and Y, give us only $np + n = (n+1)p$ numbers. We have $p \le n$, and the case $p = n$ is not of interest. If $p > n/2$, then we will have more data than unknowns, and maybe we could get some sort of estimates. However, it is usually not the case that $p > n/2$, and even if it were the case, we would have fewer than twice as many data points as numbers to estimate (not even counting β), so we could not expect our estimates to be very accurate.

Consequently, we need to assume some constraints on G, i.e., to assume that many entries are zero, or are equal to each other, or some other constraints to cut down the number of unknowns. We will look at some examples in the next section. But once we do impose constraints and once we find some way to estimate G by some matrix $\hat{G}$, then we simply substitute $\hat{G}$ for G in (N6) to get what is called a *feasible GLS* (or *Aitken*) estimator, $\hat{\beta}_{\mathrm{FGLS}}$:

$$\hat{\beta}_{\mathrm{FGLS}} := (X'\hat{G}^{-1}X)^{-1}X'\hat{G}^{-1}Y\,. \tag{N9}$$

We would estimate the covariance of $\hat{\beta}_{\mathrm{FGLS}}$ by substituting $\hat{G}$ for G in (N8):

$$\widehat{\mathrm{Cov}}(\hat{\beta}_{\mathrm{FGLS}} \mid X) := (X'\hat{G}^{-1}X)^{-1}\,. \tag{N10}$$

The name "feasible" is used regardless of the method employed to get $\hat{G}$. These might be good estimates, or they might not be. There is no good general theory about them, nor is there a general way to get FGLS estimates. These estimates are not linear since $\hat{G}$ will depend on Y, and they are usually biased without further assumptions, though if the bias is small, it may not be too important. None of the theorems we had about OLS and GLS apply to FGLS. However, we can test how well FGLS does by simulation. (This is

illustrated for exercise 5C1 below in these notes, as well as in chapter 8.) Also, in special cases there is good theory about $\hat{G}$; see the optional exercises below for some of that.

Consider a special case of GLS, where G is a diagonal matrix. That is the same as saying that ϵ_i are uncorrelated given X, which is also equivalent to saying that Y_i are uncorrelated given X, since $X\beta$ is constant given X and therefore $X\beta$ does not affect the correlations that are conditional on X. One says that the model is *homoscedastic* if we have further that G is constant on the diagonal, i.e., we are back in the OLS model. Otherwise, one says the model is *heteroscedastic*.

Optional exercise: Rederive the results of exercise 4.5.17 from (N4).

Optional exercise: Often, as in the next section, $\hat{\beta}_{\text{FGLS}}$ is a function of the OLS residuals, $e = Y - X\hat{\beta}_{\text{OLS}}$. Namely, there is some matrix function $M(e)$ such that $\hat{\beta}_{\text{FGLS}} = M(e)Y$. More precisely, there is a function $M\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$, i.e., for every vector $v \in \mathbb{R}^n$, there is an $n \times n$ matrix $M(v)$, such that $\hat{\beta}_{\text{FGLS}} = M(e)Y$ for all $Y \in \mathbb{R}^n$, where $e := P_{\text{col}(X)^\perp}Y$. It will be convenient to write e as $e(Y)$ since e is determined by Y; we are regarding X as fixed. The property $\hat{\beta}_{\text{FGLS}} = M(e)Y$ makes it looks like $\hat{\beta}_{\text{FGLS}}$ is a linear estimator, but it is not, since e depends on Y. When we write it more completely, we have $\hat{\beta}_{\text{FGLS}} = M\big(e(Y)\big)Y$. Usually the function M satisfies $M(v)X = I_p$ for all $v \in \mathbb{R}^n$, in other words, $M(v)X\beta = \beta$ for all β and all v. These properties of $\hat{\beta}$ are not statistical properties. Now suppose that M has two more (non-statistical) properties: the range of M is bounded in $\mathbb{R}^{n\times n}$ and $M(av) = M(v)$ for all non-0 real a and all $v \in \mathbb{R}^n$. Note that $\hat{\beta}_{\text{GLS}}$ has all 4 properties: by (N6), we have that $M(v) = (X'G^{-1}X)^{-1}X'G^{-1}$ does not depend on v at all. The first example of the next section is similar. Regression with serial correlation gives examples where M is not constant, but we will not discuss that type of regression in this course. Prove that $\hat{\beta}_{\text{FGLS}}$ has a finite second moment. (Technically, we need to assume that M is measurable.) Prove that if in addition the law of ϵ equals the law of $-\epsilon$ given X (e.g., ϵ is normal), then $E(\hat{\beta}_{\text{FGLS}} \mid X) = \beta$. Hints: The bias is $E\big(M(e(Y))\epsilon \mid X\big)$. We have $e(Y) = e(\epsilon)$ when $Y = X\beta + \epsilon$. Finally, $e(-\epsilon) = -e(\epsilon)$ for all ϵ.

---

§5.4. GLS examples.

The first example of GLS we consider is like OLS in that we know G up to an unknown constant factor. That is, suppose $G = \lambda\Gamma$, where Γ is a known positive definite matrix and $\lambda > 0$ is unknown. In this case, (N6) becomes

$$\begin{aligned}\hat{\beta}_{\mathrm{GLS}} &= (X'\lambda^{-1}\Gamma^{-1}X)^{-1}X'\lambda^{-1}\Gamma^{-1}Y = \lambda(X'\Gamma^{-1}X)^{-1}X'\lambda^{-1}\Gamma^{-1}Y \\ &= (X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Y\,.\end{aligned}$$

In particular, the unknown λ cancelled, and so we can compute $\hat{\beta}_{\mathrm{GLS}}$ even though G is not completely known. Furthermore, $\hat{\beta}_{\mathrm{GLS}}$ is BLUE (for fixed X) and (N8) becomes

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}} \mid X) = \lambda(X'\Gamma^{-1}X)^{-1}\,.$$

How do we estimate λ here? We use a modification of (N5), namely,

$$\underbrace{\Gamma^{-1/2}Y}_{\text{new response}} = \underbrace{\Gamma^{-1/2}X}_{\text{new design}}\,\beta + \underbrace{\Gamma^{-1/2}\epsilon}_{\text{new error}}\,. \tag{N11}$$

That is, we regress $\Gamma^{-1/2}Y$ on $\Gamma^{-1/2}X$, where we can apply OLS. The error term satisfies $\mathrm{Cov}(\Gamma^{-1/2}\epsilon \mid X) = \Gamma^{-1/2}G\Gamma^{-1/2} = \Gamma^{-1/2}\lambda\Gamma\Gamma^{-1/2} = \lambda$. That is, λ is like σ^2 and thus has the estimate $\|e\|^2/(n-p)$, where e is the residual for the OLS estimate of (N11).

The name of $\hat{\beta}_{\mathrm{GLS}}$ in this case is the *weighted least squares estimate*. The reason for the name comes from the special case where G is diagonal. As shown in exercise 5C2, the diagonal of Γ enters as weights for minimizing a sum of squares. Note too that an even more special case is where $\Gamma = I_n$; then we are back in OLS from the start.

This example was very good in that all the theory applied and we did not need to use feasible GLS. The next example forces us to use FGLS, even though, like the first example, it is a minor modification of the OLS assumptions and very little will be unknown about G.

Suppose that the subjects come in pairs, or that each subject is measured twice. Assume that given X, the response variables are uncorrelated from one pair to the next, and that the covariances within each pair are the same. An economic example is in chapter 8. That is, we assume that

$$G = \begin{bmatrix} K & \mathbf{0}_{2\times2} & \cdots & \mathbf{0}_{2\times2} \\ \mathbf{0}_{2\times2} & K & \cdots & \mathbf{0}_{2\times2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \cdots & K \end{bmatrix},$$

where K is an unknown 2×2 positive definite matrix. Here, G has the 2×2 block K along the diagonal and 0 elsewhere. Since K is symmetric, its 4 entries contain 3 unknown numbers, one of which is repeated; we need to estimate these from the data. How do we estimate K? At this juncture, there is no established theory to guide us, but we can give some heuristics.

First, let's use OLS (on the untransformed equation) to get $\hat{\beta}_{\text{OLS}}$; because we are going to make several passes and get successive estimates of β and of K, denote $\hat{\beta}_{\text{OLS}}$ by $\hat{\beta}^{(0)}$. Now by definition, $K_{11} = \text{Var}(\epsilon_1) = \text{Var}(\epsilon_3) = \cdots = \text{Var}(\epsilon_{n-1}) = E(\epsilon_1^2) = \cdots = E(\epsilon_{n-1}^2)$. If we knew ϵ, we would estimate K_{11} therefore by

$$\frac{\epsilon_1^2 + \epsilon_3^2 + \cdots + \epsilon_{n-1}^2}{n/2} = \frac{2}{n} \sum_{j=1}^{n/2} \epsilon_{2j-1}^2 \,.$$

Of course, we can't observe ϵ, but we can calculate the first set of residuals, $e^{(0)} := Y - X\hat{\beta}^{(0)}$. So we could use the estimate

$$\hat{K}_{11}^{(0)} := \frac{2}{n} \sum_{j=1}^{n/2} \left(e_{2j-1}^{(0)}\right)^2 .$$

Note that this is what we would do if we paid attention only to the data in the odd-numbered rows; according to our model, they satisfy the assumptions of OLS. In fact, we would divide not by $n/2$, but by $n/2 - p$, but put that aside for the moment. We could similarly use the estimate

$$\hat{K}_{22}^{(0)} := \frac{2}{n} \sum_{j=1}^{n/2} \left(e_{2j}^{(0)}\right)^2 .$$

What about $K_{12} = K_{21}$? This is where the dependencies between the odd and even rows enter. Again, if we knew ϵ, we could estimate $K_{12} = \text{Cov}(\epsilon_1, \epsilon_2) = \text{Cov}(\epsilon_3, \epsilon_4) = \cdots = \text{Cov}(\epsilon_{n-1}, \epsilon_n) = E(\epsilon_1\epsilon_2) = \cdots = E(\epsilon_{n-1}\epsilon_n)$ by the sample mean product

$$\frac{2}{n} \sum_{j=1}^{n/2} \epsilon_{2j-1}\epsilon_{2j} \,,$$

but since we don't know ϵ, we can use instead the residuals:

$$\hat{K}_{12}^{(0)} := \hat{K}_{21}^{(0)} := \frac{2}{n} \sum_{j=1}^{n/2} e_{2j-1}^{(0)} e_{2j}^{(0)} \,.$$

Thus, with these three estimates, we get an estimate $\hat{K}^{(0)}$ of K, and therefore of G, call it $\hat{G}^{(0)}$. If we substitute $\hat{G}^{(0)}$ into our feasible GLS estimate (N9), we get a new estimate of

β, call it $\hat{\beta}^{(1)}$. This estimate $\hat{\beta}^{(1)}$ is called *one-step* GLS; it is an FGLS estimate. It used our estimate $\hat{K}^{(0)}$ of K, so it is not GLS, but only FGLS. (One small point: if it happens that $\hat{K}$ is not invertible, then we can't find the inverse of $\hat{G}$. In that unlikely case, do anything reasonable.)

Optional remark: We could also motivate $\hat{K}$ as follows: Define the *second moment matrix* of a random vector ζ as $E(\zeta\zeta')$. When ζ has mean $\mathbf{0}$, this is equal to the covariance matrix of ζ. Define a sample second moment matrix similarly for vector-valued data. In the present case, K equals the second moment matrix of $\begin{bmatrix}\epsilon_1\\ \epsilon_2\end{bmatrix}$, whereas $\hat{K}$ equals the sample second moment matrix of the data $\left(\begin{bmatrix}e_1^{(0)}\\ e_2^{(0)}\end{bmatrix}, \ldots, \begin{bmatrix}e_{n-1}^{(0)}\\ e_n^{(0)}\end{bmatrix}\right)$.

We could now repeat this procedure: Define new residuals $e^{(1)} := Y - X\hat{\beta}^{(1)}$, use them in the same way as before to get a new estimate $\hat{K}^{(1)}$ and thus $\hat{G}^{(1)}$ and finally $\hat{\beta}^{(2)}$. This is called (surprise!) the *two-step* GLS estimate of β. One can keep going, as people usually do, until the estimate stops changing much. The final estimate is then called *iteratively weighted least squares*. However, not much is known about how well this works.

Now about the issue of the divisor, $n/2$ or $n/2-p$. First, notice that the divisor cancels when we substitute our estimate of K (and thus of G) into (N9), just as λ cancelled in the first example above. Thus, it won't affect our estimates of β at all. It will matter at the end, however, for estimating the covariance via (N10). Unfortunately, there is little theory to guide us; see pp. 174–5 for some more discussion of this issue.

An optional example: In the *equi-correlated* GLS, we assume that all diagonal elements of G are equal, and that all off-diagonal elements are also equal. Suppose that X is merely $\mathbf{1}_n$. Then we are in the situation of exercise 4.5.17, with constant correlation r instead of merely average correlation r. This might seem like an easy case. If we know r, then it becomes a special case of our first example above, so there is no problem doing GLS. However, if we do not know r, then it is impossible to estimate our error; that is, there is no FGLS in this situation. Recall that this case includes when Y_i are IID plus a systematic random bias. We get only one unobserved sample of the bias, which makes it impossible to know how large it is, never mind knowing how large it is typically.

A comment on exercise C1: I compared OLS, GLS, FGLS, and iterated FGLS by simulation for the following situation. I took 50 data points; the first 20 were IID from an exponential distribution of mean 1, whereas the last 30 were IID from an exponential distribution of mean 2, with 1 subtracted from each of the last 30 to make their mean also 1. The variance of the first distribution is 1, while the variance of the second distribution is 4. I took the straight average (OLS) and also the GLS and the FGLS estimates using the method given in the answer at the back of the book. (To estimate the two variances for FGLS, I divided by 19 and 29, rather than 20 and 30.) I also did two-step FGLS (2FGLS), as well as computed the limiting value of iterative FGLS (∞FGLS), which is the root of a cubic equation in this case. I computed all of these 10,000 times, which gave the following results:

Method	**Mean**	**rmse**	**Mean $\widehat{\text{SE}}$**	**SE**
OLS	1.0005	0.234	0.231	0.237
GLS	0.9995	0.190	—	0.191
FGLS	0.9591	0.199	0.180	?
2FGLS	0.9562	0.200	0.177	?
∞FGLS	0.9561	0.201	0.177	?

The first column reports the mean of the 10,000 simulations. How far off from the true value of 1 were the estimates on average? The second column reports the r.m.s. error (averaged over all 10,000 simulations). So indeed the GLS estimates are a little more accurate than OLS and the FGLS estimates are also more accurate than OLS and less accurate than GLS, but the FGLS estimates are definitely biased. Finally, iterated FGLS, whether two-step or the limit, is slightly more biased than FGLS and has a slightly larger error: iterating makes everything a little worse here. This shows how well these methods did.

How well did these methods predict they were doing? This is reported in the third column. The $\widehat{\text{SE}}$ for OLS is computed from (12) of chapter 4, i.e., $\hat\sigma/\sqrt{50}$; the mean of the 10,000 such $\widehat{\text{SE}}$s is given in the third column. Likewise, the $\widehat{\text{SE}}$ for FGLS and its iterated variants is computed via (N10). There is no such thing for GLS, because in GLS, we use the true values of the variances. For GLS, we can compute the actual SE instead by using (N8); that is given in column 4. Similarly, the actual SE for OLS is computed from (N4). We do not have a formula for the SE of FGLS. The fourth column is not very important here. Comparing the second and third columns, we see that OLS predicts its error correctly on average, while FGLS predicts, on average, that its error is smaller than

it really is. Although iterated FGLS does worse than FGLS, it predicts it does even better.

Optional remarks: We said above that the FGLS estimates were "definitely biased". How do we know that the mean of the 10,000 estimates (for FLGS without iteration, this was 0.9591) is not within sampling error? Since we do not have a formula for the SE of FGLS, we use the r.m.s. error instead. Dividing it by $\sqrt{10,000} = 100$ gives our estimated SE of 0.000199 on 0.9591; we find that the mean differs from the true value, 1, by -21 estimated SEs: definitely biased.

Let's consider the fourth column of the first row more carefully. (Warning: this is going to get confusing.) Are the differences between it, 0.237, and the others, 0.234 and 0.231, within sampling error? One might expect that with 10,000 IID simulations, one would have smaller differences that those we see here. In fact, however, neither column 2 nor column 3 give unbiased estimates of column 4. Recall that our formulas give unbiased estimates not of SEs, but of variances. (This is natural, since variance is an expectation, but SE, like SD, is not, and the definition of "unbiased" involves an expectation.) So we should look instead at squares. In this case, we have that the mean squared error is 0.05474, the mean squared $\widehat{\text{SE}}$ is 0.05589, and $\text{SE}^2 = 0.05600$. These are the three numbers to compare. We start by comparing the second to the third. The sd of the 10,000 $\widehat{\text{SE}}^2$s is 0.0248; divide by $\sqrt{10,000} = 100$ to get the estimated SE on 0.05589 as 0.000248. So it turns out that 0.05589 is -0.45 estimated SEs off of the true value, 0.05600: well within sampling error. Similarly, the sd of the 10,000 squared errors is 0.0787, which means that the estimated SE on 0.05474 is 0.000787. Thus, it turns out that the mean squared error is -1.6 estimated SEs off of the true SE^2: again well within sampling error.

For the third row, FGLS, we don't have a formula for the SE, but we could use the mean-squared error (i.e., the square of the r.m.s. error) to get an unbiased estimate of the theoretical mean squared error. What is that? You often see the statement that

$$\text{estimate} \;=\; \text{parameter} \;+\; \text{bias} \;+\; \text{sampling error}.$$

The error of the estimate is thus partly due to bias and partly to sampling error. This is made precise by the following equation:

$$\text{MSE}(T) = \;\text{bias}(T)^2 + \text{SE}(T)^2\,. \tag{N12}$$

Here, T is an estimator of a parameter, θ. The *bias* of T is $E(T) - \theta$ and

$$\text{MSE}(T) := E\big[(T-\theta)^2\big]\,.$$

One also defines

$$\text{RMSE}(T) := \sqrt{\text{MSE}(T)}$$

to get this to be the right size and the right units. As usual, $\mathrm{SE}(T)^2 = \mathrm{Var}(T) = E\big[\big(T - E(T)\big)^2\big]$. In order to derive (N12), we use the abbreviation $m := E(T)$ to compute

$$\begin{aligned}\mathrm{MSE}(T) &= E\big[(T-\theta)^2\big] = E\big[\big((T-m)+(m-\theta)\big)^2\big] \\ &= E\big[(T-m)^2\big] + 2E[(T-m)(m-\theta)] + E\big[(m-\theta)^2\big] \\ &= \mathrm{SE}(T)^2 + 0 + (m-\theta)^2\,.\end{aligned}$$

Now, for FGLS, we estimate the unknown MSE by the mse, 0.199^2. Is the mean $\widehat{\mathrm{SE}}^2$ (which turned out to be 0.0341, with an sd of 0.0146) within sampling error of this? A similar calculation as the preceding ones shows that the difference is -38 estimated SEs: not at all possible from sampling error. The $\widehat{\mathrm{SE}}^2$ is definitely biased. (Note that $\widehat{\mathrm{SE}}$ is an estimate of RMSE, as it does not involve an estimate of the mean, but, rather, is obtained from a formula for an unbiased estimator.)

A comment on exercise C2: It is easier to deduce parts (b) and (c) from example 1, where we noted that the estimates agree.

For exercise C3, you might prefer to build the design matrix row by row instead. Concerning part (ii) of the answer, note first that the divisor (2 or 3) cancels in computing $\hat\beta_{\mathrm{FGLS}}$ as long as we always use the same divisor, just as λ cancelled in example 1. For SEs, however, it will not cancel. It is reasonable to use 2 instead of 3 because there are 3 observations for each a_i; although there is a second unknown, b, there are 2400 observations for it, so it is known quite precisely and we can regard p as 1 in $n - p$. (This is just a heuristic.)

Optional lab: Replicate the above table by your own simulations. You may choose to change "20" and "30", as well as the distributions (instead of exponential).

Optional exercise: Suppose that $\hat\beta_{\mathrm{FGLS}}$ is given by the form $\hat\beta_{\mathrm{FGLS}} = M(e)Y$, where e is the OLS residual of Y, as in the optional exercise at the end of section 5.3 above. Assume also that if M has the property that $M(v)X = I_p$ for all v. Suppose we regard this as one-step FGLS and denote it by $\hat\beta^{(1)} := M\big(e(Y)\big)Y$. Now iterate in the sense that we define new (non-OLS) residuals $e^{(1)} := e^{(1)}(Y) := Y - X\hat\beta^{(1)}$, use them in the same way as before to get a new estimate $\hat\beta^{(2)} := M\big(e^{(1)}(Y)\big)Y$. Define the matrix function $M^{(2)}$ by $M^{(2)}(v) := M\big(v - XM(v)v\big)$.

(a) Show that $\hat\beta^{(2)} = M^{(2)}\big(e(Y)\big)Y$ for all Y, so that the two-step FGLS estimator $\hat\beta^{(2)}$ also has the form of the previous optional exercise. Hint: Show that $e^{(1)}(Y) = e(Y) - XM\big(e(Y)\big)e(Y)$ by showing $\hat\beta^{(1)} = \hat\beta_{\rm OLS} + M\big(e(Y)\big)e(Y)$.

(b) Show that $M^{(2)}$ also has the property that $M^{(2)}(v)X = I_p$ for all v.

(c) Show the same two properties hold with each continued iteration, where we define successively $e^{(k)}(Y) := Y - X\hat\beta^{(k)}$, $\hat\beta^{(k+1)} := M\big(e^{(k)}(Y)\big)Y$, and $M^{(k+1)}(v) := M\big(v - XM^{(k)}(v)v\big)$.

(d) Show that if $M(av) = M(v)$ for all scalars a and vectors v, then the same holds for each function $M^{(k)}$.

(e) Show that if the range of M is bounded, then so is the range of each $M^{(k)}$.

(f) Prove that $\hat\beta^{(k)}$ has a finite second moment. Prove that if in addition the law of ϵ equals the law of $-\epsilon$ given X (e.g., ϵ is normal), then $E(\hat\beta^{(k)} \mid X) = \beta$.

Optional exercise: Consider a heteroscedastic regression model where the diagonal of G is $(\sigma_1^2, \dots, \sigma_1^2, \sigma_2^2, \dots, \sigma_2^2, \dots, \sigma_r^2, \dots, \sigma_r^2)$. Here, σ_k^2 occurs n_k times, with $n_1 + \cdots n_r = n$. Exercise 5C1 is an example with $r = 2$ and $X = \mathbf{1}_n$. Exercise 5C3 is another example, with each $n_k = 3$ and $r = 800$. As in the solutions to these two exercises, suppose that we estimate σ_k^2 by $\hat\sigma_k^2 := \|e_k\|^2/n_k$ for each k, where e_k contains the corresponding coordinates of the OLS residual vector e, so that $e = \begin{bmatrix} e_1 \\ \vdots \\ e_r \end{bmatrix}$. We then use all $\hat\sigma_k^2$ to obtain $\hat\beta_{\rm FGLS}$. Show that $\hat\beta_{\rm FGLS} = M(e)Y$ for some matrix function M that satisfies all 4 properties in the optional exercise at the end of the preceding section. Hint: If $\hat G^{-1}$ is large, divide by its largest diagonal element in order to show the boundedness property.

§5.6. Normal theory.

In the next two sections, we make not only the usual OLS assumptions, but also that the error terms are normal. More precisely, ϵ has a $\mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$ distribution. The reason is that we will derive statistics like the t-statistic in FPP, and their distribution requires such assumptions. We will treat the $n \times p$ design matrix X as fixed, so the only other assumption is that X has rank $p < n$.

The most common statistical hypothesis test is whether a coefficient in the regression, say β_k, is 0. The interest in such a test depends on the reason that the kth variable is in the model to start with: One may be particularly interested in the "effect" of that variable

on Y, or merely whether there is any effect at all after "controlling" for other variables, i.e., whether $\beta_k = 0$. Or the kth variable may be there only to "control" for a confounding factor; in this case, one may want to know whether it is really necessary to include that variable in the model. If $\beta_k = 0$, then it can be omitted.

Thus, the null hypothesis of interest in this section is $\beta_k = 0$; the alternative is $\beta_k \neq 0$. (Of course, another possibility is that the model is wrong.) We form the t-statistic

$$t := \frac{\hat{\beta}_k}{\widehat{\mathrm{SE}}(\hat{\beta}_k)} .$$

Here, $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ is our estimate of $\mathrm{SE}(\hat{\beta}_k \mid X)$. Now $\mathrm{Var}(\hat{\beta}_k \mid X)$ is the (k,k)-element of $\mathrm{Cov}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$ (according to theorem 4.3); we take its square root to get the SE. However, since we don't know σ, we need to estimate it. Thus, $\widehat{\mathrm{SE}}$ is $\hat{\sigma}$ times the square root of the (k,k)-element of $(X'X)^{-1}$. Recall from (N6) of the notes on chapter 4 the estimate $\hat{\sigma} := \|e\|/\sqrt{n-p}$, where $e = Y - X\hat{\beta}$ is the OLS residual vector.

Sometimes, as in FPP, we would test instead whether β_k takes some other value, say, b. Then we would use $t := (\hat{\beta}_k - b)/\widehat{\mathrm{SE}}(\hat{\beta}_k)$, just as in FPP where $X = \mathbf{1}_n$, and the null hypothesis would be that $\beta_k = b$. The theory will be the same as when $b = 0$.

What is the distribution of t under the null hypothesis? We first need the following theorem, which extends one that we proved in the handout on the t-distribution. Recall that χ^2_d is the distribution of the sum of d IID standard normal random variables. That is, if $Z \sim \mathrm{N}(\mathbf{0}_d, I_d)$, then $\|Z\|^2 \sim \chi^2_d$.

Theorem 5.2. *Condition on X. If ϵ has distribution $\mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$, then $\hat{\beta}$ and e are independent with distributions $\hat{\beta} \sim \mathrm{N}\big(\beta, \sigma^2 (X'X)^{-1}\big)$ and $\|e\|^2 \sim \sigma^2 \chi^2_{n-p}$.*

Proof. We've already shown that the mean of $\hat{\beta}$ is β (theorem 4.2) and the covariance of $\hat{\beta}$ is $\sigma^2 (X'X)^{-1}$ (theorem 4.3). Clearly $Y \sim \mathrm{N}(X\beta, \sigma^2 I_n)$ has independent normal coordinates, whence $QY = \hat{\beta}$ has a normal distribution by definition—see the handout on correlation and normal distributions. This shows that the distribution of $\hat{\beta}$ is as claimed.

Let $W := \mathrm{col}(X)$, as usual. Since $\epsilon \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$, the orthogonal projections

$$P_W(\epsilon) \text{ and } P_{W^\perp}(\epsilon) \text{ are independent normals;} \tag{N13}$$

in orthonormal coordinates of W and $W^\perp$, they have means $\mathbf{0}_p$ and $\mathbf{0}_{n-p}$ and covariances $\sigma^2 I_p$ and $\sigma^2 I_{n-p}$. We proved this in the handout on the t-distribution. Now

$$P_W(\epsilon) = P_W(Y - X\beta) = X\hat{\beta} - X\beta \qquad \text{and} \qquad P_{W^\perp}(\epsilon) = e\,, \tag{N14}$$

as we saw in (N8) of the notes on chapter 4. As in the proof of theorem 4.4, write U for the vector e in orthonormal coordinates of $W^\perp$. We saw that $U \sim \mathrm{N}(\mathbf{0}_{n-p}, \sigma^2 I_{n-p})$, so by definition of χ^2_{n-p}, we have $\|e\|^2 = \|U\|^2 \sim \sigma^2 \chi^2_{n-p}$, as claimed. Finally, (N13) and (N14) together imply that e is independent of $X\hat\beta - X\beta$, whence of $X\hat\beta$ (since $X\beta$ is constant) and therefore of $QX\hat\beta = \hat\beta$. ∎

Now recall that Student's t-distribution with d degrees of freedom is the distribution of Z/S, where Z and S are independent with $Z \sim \mathrm{N}(0,1)$ and $S \sim \chi_d/\sqrt{d}$.

Corollary. *Condition on X. If ϵ has distribution $\mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$ and $\beta_k = 0$, then the distribution of $t := \hat\beta_k / \widehat{\mathrm{SE}}(\hat\beta_k)$ is Student's t-distribution with $n-p$ degrees of freedom.*

Proof. Let's denote the square root of the (k,k)-element of $(X'X)^{-1}$ by a; it is not random. Thus,

$$t = \frac{\hat\beta_k}{\widehat{\mathrm{SE}}(\hat\beta_k)} = \frac{\hat\beta_k}{\hat\sigma a} = \frac{\hat\beta_k}{\|e\| a/\sqrt{n-p}}\,.$$

By theorem 5.2, it follows from this that the numerator of t and the denominator of t are independent, and $\hat\beta \sim \mathrm{N}\big(\beta, \sigma^2 (X'X)^{-1}\big)$. In particular, $\hat\beta_k \sim \mathrm{N}(\beta_k, \sigma^2 a^2)$. Since the null hypothesis says that $\beta_k = 0$, we get that

$$\hat\beta_k \sim \mathrm{N}(0, \sigma^2 a^2) = \sigma a \mathrm{N}(0,1)\,. \tag{N15}$$

Theorem 5.2 again tells us that $\|e\| \sim \sigma \chi_{n-p}$, so

$$\widehat{\mathrm{SE}}(\hat\beta_k) \sim \sigma a \chi_{n-p}/\sqrt{n-p}\,. \tag{N16}$$

Putting together (N15) and (N16), we obtain

$$t = \frac{\hat\beta_k}{\widehat{\mathrm{SE}}(\hat\beta_k)} \sim \frac{\sigma a \mathrm{N}(0,1)}{\sigma a \chi_{n-p}/\sqrt{n-p}} = \frac{\mathrm{N}(0,1)}{\chi_{n-p}/\sqrt{n-p}}\,,$$

where the numerator and denominator are independent. This is precisely the t-distribution with $n-p$ degrees of freedom. ∎

In fact, even without the null hypothesis that $\beta_k = 0$, the statistic $t := (\hat\beta_k - \beta_k)/\widehat{\mathrm{SE}}(\hat\beta_k)$ has the same distribution, i.e., Student's t-distribution with $n-p$ degrees of freedom. More generally, suppose we want to estimate some linear combination of the coordinates of β, i.e., $c'\beta$ for some fixed $c \neq \mathbf{0}_p$. A similar proof shows that $t := (c'\hat\beta - c'\beta)/\widehat{\mathrm{SE}}(c'\hat\beta)$ has Student's t-distribution with $n-p$ degrees of freedom. Here, $\widehat{\mathrm{SE}} = \hat\sigma\sqrt{c'(X'X)^{-1}c}$ since $\mathrm{Cov}(c'\hat\beta \mid X) = \sigma^2 c'(X'X)^{-1}c$. For example, we might be

interested in testing whether $\beta_1 = \beta_2$. We would use $c = (1, -1, 0, 0, \ldots, 0)$ and we would compute $t := (\hat{\beta}_1 - \hat{\beta}_2)/\widehat{\mathrm{SE}}(\hat{\beta}_1 - \hat{\beta}_2)$ since our null hypothesis would be that $c'\beta = 0$.

Note that in the course of proving the corollary, we showed that our estimate $\hat{\beta}_k$ is independent of our estimate $\widehat{\mathrm{SE}}(\hat{\beta}_k)$ of its error. This is nice, because we could be quite deceived if we thought our error was large when it was in fact small, and thought the error was small when it was in fact large. These deceptive possibilities had not been ruled out just by the fact that we have an unbiased estimate of σ^2. (Of course, it would be even nicer if our error were positively correlated with our estimate of our error.)

When is t so large that the null hypothesis should be rejected? This depends on $n - p$. When $n - p$ is large, then the distribution of t is close to the standard normal distribution (we saw this in FPP, section 26.6), in which case the 5% significance level of $|t|$ is approximately 2. When $n - p$ is large, the t-test is also referred to as the z-test, just as in FPP. For small $n - p$, see p. A-105 in FPP for the significance levels.

If one tests two variables and rejects the null hypothesis (that their coefficient is 0) for each, then people say that those two variables have "independent effects" on Y. Note, however, that this has nothing to do with either stochastic independence or linear independence. The meaning will be explored more thoroughly in chapter 6.

All the above theory depended on the assumption that ϵ is normal. What if ϵ is not normal? With small n, we saw already with $p = 1$ in the handout on the t-distribution that the distribution of t changes in a way that is important for making hypothesis tests. But what if n is large? When $p = 1$, we know that t is approximately normal; we used this frequently in FPP. What is the story for $p > 1$? We assume that ϵ_i are IID with mean 0 and finite variance (given X). The distribution of $\hat{\beta}$ depends on X, but in most cases, the sizes of the entries in each column of X are fairly similar, with none much larger than all the rest, and then $\hat{\beta}$ will be approximately normal by a multivariate central limit theorem when n is large and p is fixed. This implies that also $\hat{\beta}_k$ is approximately normal. Thus, the non-normality of ϵ makes little difference when n is large. To see a little why $\hat{\beta}$ is approximately normal, recall that $\hat{\beta} = (X'X)^{-1}X'\epsilon$. Now the jth coordinate of $X'\epsilon$ equals $\sum_{i=1}^{n} X_{i,j}\epsilon_i$. This is therefore a sum of many independent random variables, and thus approximately normal by a version of the CLT. A multivariate CLT shows that the entire vector $X'\epsilon$ is approximately normal. When we take a linear transformation of it by multiplying by $(X'X)^{-1}$, it remains approximately normal. The upshot is that $(\hat{\beta}_k - \beta_k)/\mathrm{SE}(\hat{\beta}_k) \approx \mathrm{N}(0, 1)$. Since $\widehat{\mathrm{SE}}(\hat{\beta}_k)/\mathrm{SE}(\hat{\beta}_k) \approx 1$, it follows that $t \approx \mathrm{N}(0, 1)$.

When the errors are not assumed to be normal and one wants to know whether n is large enough, or X is good enough, then one can simulate: take X from the data, assign the ϵ_i to be IID samples from other distributions of interest, and compute the corresponding

t-statistics. Another approach is to simulate via bootstrapping, as in chapter 8.

A comment on exercise D2: 95% confidence that b lies in the interval $3.79 \pm 2 \times 1.88$ is far different than 95% confidence that $b \neq 0$: those are two quite different statements about where b is. More importantly, the second doesn't make any sense. To see why, recall what confidence means. The statement that we are 95% confident that b lies in $3.79 \pm 2 \times 1.88$ means that 95% of the time, given that the model is true, the procedure we used will have the property that b lies in $\hat{b} \pm 2\widehat{\mathrm{SE}}$. Note: we do not assume the null hypothesis that $b = 0$; instead, we are concerned here with estimating b. Confidence intervals like this are for estimation, not for hypothesis testing. The reason that the event $b \in [\hat{b} - 2\widehat{\mathrm{SE}}, \hat{b} + 2\widehat{\mathrm{SE}}]$ has probability about 95% is stated after the corollary to theorem 5.2: the statistic $t := (\hat{\beta}_k - \beta_k)/\widehat{\mathrm{SE}}(\hat{\beta}_k)$ has the Student t-distribution with $n - p$ degrees of freedom. The numbers that occur in the statement "we are 95% confident that b lies in $3.79 \pm 2 \times 1.88$" are (according to our model) observed values of random variables, namely, 3.79 is an observed value of $\hat{b}$ and 1.88 is an observed value of $\widehat{\mathrm{SE}}(\hat{b})$. The property "$b$ lies in $\hat{b} \pm 2\widehat{\mathrm{SE}}$" is an event involving random variables and therefore it has a probability (namely, approximately 95%). When we pass from random variables to observed values, we pass from probability to confidence. On the other hand, the statement "$b \neq 0$" involves no observed values of random variables at all. Therefore, it does not correspond to a probability statement and we cannot attach a confidence to it. You may wish to review FPP Sections 21.2,3. We *can* say that we reject the null $b = 0$ at the 5% significance level. Note that we would also be able to say we reject the null at the 10% significance level; we do not attach a precise P-value to such statements, but, of course, you can always say what the P-value is and indeed that is often better than less precise statements.

Note on Lab 5: The book now tells you about a MATLAB shortcut, the backslash operator. You can now use this (but not other statistics shortcuts). The book suggests using

$$\texttt{betahatSim} = \texttt{X}\backslash(\texttt{X} * \texttt{beta} * \texttt{ones}(1, 1000) * \texttt{sigma} * \texttt{randn}(32, 1000));$$

The math of $X\backslash Y$ is $(X'X)^{-1}X'Y$, with which you are familiar. Why the backslash notation? Note that if X were a scalar, then $(X'X)^{-1}X'$ would equal $1/X$, so we would be dividing by X. More generally, $X\backslash Y = \hat{\beta}$, where $\hat{Y} = X\hat{\beta}$, so in a sense we are dividing $\hat{Y}$ by X, where it makes sense since $\hat{Y}$ is indeed X times something. While Y is not X times something, $X\backslash Y$ gives the nearest multiple of X to Y, so it is the best inverse that exists. A name for $(X'X)^{-1}X'$ is a *pseudoinverse* of X. The second aspect of the above MATLAB suggestion is to use 1000 columns. In thinking about why this works, remember

that since the backslash operator is just multiplication by a matrix, this multiplication works column by column.

The reason for the strange-looking suggestion $\epsilon_i \sim \sigma \times (\chi^2_5 - 5)/\sqrt{10}$ is that this distribution is somewhat similar to a normal distribution with mean 0 and SD σ: Recall that χ^2_5 is the distribution of the sum of 5 independent standard normal random variables. Therefore, its mean is 5. To calculate the variance, write $\sum_{i=1}^5 (U_i^2 - 1)$ for $\chi^2_5 - 5$, where U_i are independent standard normal random variables. We then have that the variance equals $\sum_{i=1}^5 E\big((U_i^2-1)^2\big)$. When you do the algebra and use the fact that the 4th moment of U_i equals 3, you get that the variance is $5 \times 2 = 10$. [Factoid: the kth moment of a standard normal random variable is 0 when k is odd (by symmetry) and is $(k-1)!! := 1 \cdot 3 \cdot 5 \cdots (k-1)$ when k is even. You can prove this via integration by parts.] You could, of course, simulate with whatever distribution you like.

The book also gives a discussion of some terminology related to hypothesis testing (pp. 299–300). Consider testing a parameter, like a mean in FPP, at the 5% significance level. If the null hypothesis is true, then by definition, the test statistic will translate to a P-value that will be 5% or smaller 5% of the time. See p. 455 of FPP for a plot of z-statistics; they could have been translated to P-values. Here, $|z| = 1.96$ translates to $P = 0.05$. In this context, 5% is called the *level* or *size* of the test and 1.96 is called the *critical value*. If the null hypothesis is true, but you reject it, then you make what is called a *type I error*. If you are testing at the 5% level, then you will make a type I error 5% of the time when the null is true. You don't want to make such an error too often, which is why we want the size of the test to be small.

What if the null is false? In general, we can't make any quantitative statements, since the alternative might not be quantitative (e.g., it might be that the model is entirely false). However, if the alternative is a specific statistical model (such as simply changing the parameter value we are estimating), then we may be able to make similar statements about probabilities. Often, the alternative is a range of parameter values; clearly, then, the probability of an event depends on the specific value and is not the same for all parameters in the alternative. But whatever the alternative is, if it does correspond to a statistical model, then there will be a probability of not rejecting the null when the null is false and the alternative is true. This kind of error is called a *type II error*, and the probability of *not making* a type II error is called the *power* of the test. Thus, with the preceding example, the probability that $|z| > 1.96$, under the alternative, is the power. This is a good thing, so we want the power to be large. The word "power" suggests the ability, or power, to distinguish between the null and the alternative.

The general framework is this: We have a parameter, θ (which might be a vector), that we want to estimate. For each value of θ, there is a probability distribution in our model, denoted $\mathbf{P}_\theta$. E.g., this gives the distribution of Y and $\hat{\beta}$ if $Y = X\beta + \epsilon$ and $\theta = \beta$ in linear regression. The null is usually a specific value of θ, but it might be a set of θ; e.g., your null might be that a mean is at most 0, rather than exactly 0. This makes things a little more complicated. The alternative is usually a set of θ, but sometimes it is a specific θ.

Let T be our test statistic (*not* our estimate of θ). If we test at level α (like 0.05), then we choose a critical value k of T so that $\mathbf{P}_\theta(T \geq k) = \alpha$ for θ in the null; actually, there might not be such a k, especially when there is more than one θ in the null. So instead, the critical value means the smallest value of k such that $\mathbf{P}_\theta(T \geq k) \leq \alpha$ for all θ in the null. The probability $\mathbf{P}_\theta(T \geq k)$ for θ in the alternative is the power; it depends on k and on θ.

When you collect data, you will compute an observed value of T, denoted T_{obs}. We translate it to an *observed significance level* $P_{\text{obs}} := \mathbf{P}_\theta(T \geq T_{\text{obs}})$. Just as an estimator of θ is random before data collection and computation, so is P_{obs}. This was illustrated (again, before translation from z to P) on p. 455 of FPP. If the null hypothesis is a specific value of θ, then usually $\mathbf{P}_\theta(P_{\text{obs}} \leq \alpha) = \alpha$; this is true whenever T has a continuous distribution. In other words, the chance is 5% that we obtain a P-value of at most 5% when the null is true, the chance is 1% that we obtain a P-value of at most 1% when the null is true, etc. Another way of saying that is that the distribution of P_{obs} is uniform on $[0, 1]$ when the null is true.

§5.7. The F-test.

In the preceding section, we discussed how to test whether one coefficient β_k was 0. Sometimes people want to test at once whether several coefficients are 0. You might imagine this is simple: just test one at a time via a t-test. But this is actually not the same, because each of those t-tests does not assume the others are 0. Thus, you might not reject on the basis of any of the individual t-tests, but maybe the data truly are unlikely if all of the tested coefficients are 0 at once. (For details on how the new test relates to the various t-tests, see the optional handout, "How F Relates to t".) So we need a different sort of test.

Before we say how to do that, let's discuss why people might want such a test. Mainly, it is to find out whether those coefficients belong in the model, or could be "safely" left out. Sometimes, *all* the coefficients are tested at once, except for the intercept: the question

then is whether one can distinguish statistically between the proposed model and just a constant as a summary of Y. In all cases, remember that if the null hypothesis is rejected, this means that taking the coefficients as 0 is hard to defend; it does not mean that the model is correct *and* (some of) the coefficients are not 0. It only means that *if* the model is correct, then (some of) the coefficients are not 0. Likewise, if the null hypothesis is *not* rejected, it does not mean the model is correct, whatever the coefficients. Thus, it is not truly "safe" to leave out coefficients which cannot be statistically distinguished from 0, unless one has more information than simply a P-value.

Suppose that the number of coefficients we want to test is $p_0 \le p < n$. For simplicity, we will take them to be the last ones, $\beta_{p-p_0+1}, \dots, \beta_p$. Suppose we put them together into a vector $\beta^{(\text{test})} \in \mathbb{R}^{p_0}$. We want to test the null hypothesis that $\beta^{(\text{test})} = \mathbf{0}_{p_0}$. As usual, fix X. We will again need to assume that $\epsilon \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$. When we tested whether $\beta_k = 0$, we compared its estimate, $\hat\beta_k$, to $\mathrm{SE}(\hat\beta_k)$, or, more likely, to $\widehat{\mathrm{SE}}(\hat\beta_k)$. In other words, $\hat\beta_k/\widehat{\mathrm{SE}}(\hat\beta_k)$ is a standardized quantity whose distribution can be calculated, whatever X is. Now, however, we are testing a vector, $\hat\beta^{(\text{test})}$. How can we judge whether it is "too large"?

This is a little subtle, so we will give two ways to think about it. The first way avoids this question directly, but does not explain as much. For the null hypothesis, we have a smaller design matrix, denoted $X^{(\text{s})}$. This is formed from the first $p-p_0$ columns of X. We could fit it, getting a residual $e^{(\text{s})}$. Suppose we compare $e^{(\text{s})}$ to the original residual e by comparing their lengths; $e^{(\text{s})}$ should be larger because there is less of X used to fit Y. More specifically, $e^{(\text{s})}$ is gotten by orthogonally projecting onto a larger subspace, $\big(\operatorname{col} X^{(\text{s})}\big)^\perp$ instead of $(\operatorname{col} X)^\perp$. Thus, form the statistic

$$F := \frac{\big(\|e^{(\text{s})}\|^2 - \|e\|^2\big)/p_0}{\|e\|^2/(n-p)}\,.$$

We will be able to calculate its distribution; it depends only on p_0 and $n-p$. We recognize the denominator of F as $\hat\sigma^2$; we will show that the numerator also estimates σ^2, but under the null hypothesis.

Unfortunately, that probably looked rather ad hoc. In order to understand better why we make that definition of F, we go back to the question: how do we test whether a random vector is abnormally "large"? Now by theorem 5.2, we know that $\hat\beta \sim \mathrm{N}\big(\beta, \sigma^2(X'X)^{-1}\big)$. Since all the coefficients of $\hat\beta$ are jointly normal, so are the coefficients of $\hat\beta^{(\text{test})}$. Write $\operatorname{Cov}(\hat\beta^{(\text{test})}) = \sigma^2 G^{(\text{test})}$. Here, $G^{(\text{test})}$ is the $p_0 \times p_0$ bottom right corner of $(X'X)^{-1}$, but we won't need that. When we standardize a random variable, we subtract its mean and divide by its standard deviation. We can certainly subtract the mean $\beta^{(\text{test})}$ from $\hat\beta^{(\text{test})}$,

but how do we divide by its standard deviation? The idea is that the SD is the square root of the variance; the analogue of variance for a random vector is covariance matrix; and the analogue of division for a matrix is multiplying by its inverse. Thus, the standardized $\hat{\beta}^{(\text{test})}$ is

$$\left(\sigma^2 G^{(\text{test})}\right)^{-1/2}\left(\hat{\beta}^{(\text{test})} - \beta^{(\text{test})}\right). \tag{N17}$$

Clearly, the mean after standardization is $\mathbf{0}_{p_0}$, and because we applied a linear transformation, we still have a normal random vector. If the analogies hold, the covariance matrix after standardization should be the analogue of 1, i.e., the identity matrix. And it is:

$$\begin{aligned}
&\operatorname{Cov}\left(\left(\sigma^2 G^{(\text{test})}\right)^{-1/2}\left(\hat{\beta}^{(\text{test})} - \beta^{(\text{test})}\right)\right) \\
&\qquad = \left(\sigma^2 G^{(\text{test})}\right)^{-1/2} \operatorname{Cov}\left(\hat{\beta}^{(\text{test})} - \beta^{(\text{test})}\right)\left(\sigma^2 G^{(\text{test})}\right)^{-1/2} \\
&\qquad = \left(\sigma^2 G^{(\text{test})}\right)^{-1/2} \sigma^2 G^{(\text{test})}\left(\sigma^2 G^{(\text{test})}\right)^{-1/2} \\
&\qquad = I_{p_0}\,.
\end{aligned}$$

Therefore,

$$\left(\sigma^2 G^{(\text{test})}\right)^{-1/2}\left(\hat{\beta}^{(\text{test})} - \beta^{(\text{test})}\right) \sim \mathrm{N}(\mathbf{0}_{p_0}, I_{p_0})\,.$$

So a reasonable way to gauge the size of the standardized estimator (N17) would be to compare its squared length to its expected squared length, which is p_0, the trace of I_{p_0}. (The expected squared length is just the sum of the variances of the coordinates of (N17).) Thus, if we divide by one more thing, namely, p_0, we get something that can be compared to 1, just like $(\hat{\beta}_k - \beta_k)/\mathrm{SE}(\hat{\beta}_k)$. There's only one catch: we need to replace σ by its estimate, $\hat{\sigma}$. We will show that this final quantity,

$$\left\|\left(\hat{\sigma}^2 G^{(\text{test})}\right)^{-1/2}\left(\hat{\beta}^{(\text{test})} - \beta^{(\text{test})}\right)\right\|^2 / p_0 \tag{N18}$$

is the same as F when $\beta^{(\text{test})} = \mathbf{0}_{p_0}$.

What is the distribution of F? When we compute the squared length of (N17), we get the squared length of a normal random vector of dimension p_0, so that has distribution $\chi^2_{p_0}$. This becomes F only after replacing σ by $\hat{\sigma}$, so we need to divide by $\hat{\sigma}^2/\sigma^2$, which, as we saw, has distribution $\chi^2_{n-p}/(n-p)$. We will show that these two random variables are independent, and this will give Fisher's F-distribution with p_0 degrees of freedom in the numerator and $n-p$ degrees of freedom in the denominator.

Theorem 5.3. *Condition on X. Assume the null hypothesis $Y = X^{(\mathrm{s})}\beta^{(\mathrm{s})} + \epsilon$, where ϵ has distribution $\mathrm{N}(\mathbf{0}_n, \sigma^2 I_n)$. Then the numerator and denominator*

$$F := \frac{\left(\|e^{(\mathrm{s})}\|^2 - \|e\|^2\right)/p_0}{\|e\|^2/(n-p)}$$

are independent with $\|e^{(\mathrm{s})}\|^2 - \|e\|^2 \sim \sigma^2\chi^2_{p_0}$ *and* $\|e\|^2 \sim \sigma^2\chi^2_{n-p}$. *Thus,*

$$F \sim \frac{\chi^2_{p_0}/p_0}{\chi^2_{n-p}/(n-p)}\,,$$

where the numerator and denominator are independent. Furthermore, regardless of assumptions,

$$F = \frac{\big(\|\hat{Y}\|^2 - \|\hat{Y}^{(\mathrm{s})}\|^2\big)/p_0}{\|e\|^2/(n-p)} = \frac{\big(R^2 - (R^2)^{(\mathrm{s})}\big)/p_0}{(1-R^2)/(n-p)}\,,$$

where the superscript $^{(\mathrm{s})}$ *always refers to quantities calculated with the small design matrix.*

Optional proof. The smaller design matrix $X^{(\mathrm{s})}$ spans the smaller column space $W^{(\mathrm{s})}$ inside of W. Now $W^{(\mathrm{s})} \subseteq W \subseteq \mathbb{R}^n$, with $\dim W^{(\mathrm{s})} = p - p_0$ and $\dim W = p$. Let's take an orthonormal basis $(w_1, \ldots, w_{p-p_0})$ of $W^{(\mathrm{s})}$. Then extend it to an orthonormal basis $(w_1, \ldots, w_p)$ of W. Finally, extend it once more to an orthonormal basis $(w_1, \ldots, w_n)$ of $\mathbb{R}^n$. It will be useful to call V the linear span of $(w_{p-p_0+1}, \ldots, w_p)$, i.e., the span of the basis vectors of W that aren't in $W^{(\mathrm{s})}$. We have $\dim V = p_0$. Note that $W^\perp$ is the span of $(w_{p+1}, \ldots, w_n)$ and $(W^{(\mathrm{s})})^\perp$ is the span of $(w_{p-p_0+1}, \ldots, w_n)$. Thus, $(W^{(\mathrm{s})})^\perp = W^\perp \oplus V$. Now $e = P_{W^\perp}(Y) \in W^\perp$ and $e^{(\mathrm{s})} = P_{(W^{(\mathrm{s})})^\perp}(Y) \in (W^{(\mathrm{s})})^\perp$. Let's write $v := P_V(Y) \in V$. Then

$$e^{(\mathrm{s})} = e + v$$

is the orthogonal decomposition of $e^{(\mathrm{s})}$ with the first term in $W^\perp$ and the second term in V. In particular, the Pythagorean theorem gives

$$\|e^{(\mathrm{s})}\|^2 = \|e\|^2 + \|v\|^2\,,$$

so

$$\|e^{(\mathrm{s})}\|^2 - \|e\|^2 = \|v\|^2$$

and therefore

$$F = \frac{\|v\|^2/p_0}{\|e\|^2/(n-p)}\,.$$

As we saw in the proof of theorem 5.2, the null hypothesis implies that $e^{(s)} \sim \mathrm{N}\big(\mathbf{0}_{n-(p-p_0)}, \sigma^2 I_{n-(p-p_0)}\big)$ if we write $e^{(\mathrm{s})}$ in orthonormal coordinates of $\big(W^{(\mathrm{s})}\big)^\perp$. This implies (as in the proof of theorem 5.2 again) that e and v are independent normal random vectors. If we write v in orthonormal coordinates of V, we have $v \sim \mathrm{N}(\mathbf{0}_{p_0}, \sigma^2 I_{p_0})$.

This means that $\|v\|^2 \sim \sigma^2 \chi^2_{p_0}$. Likewise, $\|e\|^2 \sim \sigma^2 \chi^2_{n-p}$, so this proves the first part of theorem 5.3.

To prove the remainder, recall that $Y = \hat{Y} + e$ with $e = P_{W^\perp}(Y)$. Likewise, $Y = \hat{Y}^{(\mathrm{s})} + e^{(\mathrm{s})}$ with $e^{(\mathrm{s})} = P_{(W^{(\mathrm{s})})^\perp}(Y)$. Thus, the Pythagorean theorem again gives

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|e\|^2 \qquad \text{and} \qquad \|Y\|^2 = \|\hat{Y}^{(\mathrm{s})}\|^2 + \|e^{(\mathrm{s})}\|^2 .$$

Setting the right-hand sides of these equations equal and rearranging a little, we obtain

$$\|\hat{Y}\|^2 - \|\hat{Y}^{(\mathrm{s})}\|^2 = \|e^{(\mathrm{s})}\|^2 - \|e\|^2 .$$

This shows the next alternative formula for F.

Finally, recall that $1 - R^2 = \|e\|^2/\|Y\|^2$ if there is no intercept, while $1 - R^2 = \mathrm{var}(e)/\mathrm{var}(Y)$ if there is an intercept. In the latter case, since e has mean 0, we have that $\mathrm{var}(e) = \|e\|^2$. Thus, in both cases we can write $\|e\|^2 = \alpha_Y \big(1 - R^2\big)$ for the appropriate α_Y, either $\|Y\|^2$ or $\mathrm{var}(Y)$. Likewise, $\|e^{(\mathrm{s})}\|^2 = \alpha_Y \big(1 - (R^2)^{(\mathrm{s})}\big)$ for the same α_Y. Substituting these in the definition of F, we find that all occurrences of α_Y cancel, and we obtain the final formula for F. ∎

Optional remark: We did not justify that (N18) is the same as F. To do so, let's look more closely at the random vector $v = P_V(Y) = e^{(\mathrm{s})} - e$ that entered in the proof. We can also write v as $v = \hat{Y} - \hat{Y}^{(\mathrm{s})} = X\hat{\beta} - X^{(\mathrm{s})}\hat{\beta}^{(\mathrm{s})}$. What is this? Both terms are linear combinations of the columns of X. To see it more clearly, write the last p_0 columns of X as $X^{(\mathrm{test})}$. Also, write the first $p - p_0$ coordinates of $\hat{\beta}$ as the vector $\hat{\beta}^{(p-p_0)}$. The last p_0 coordinates of $\hat{\beta}$ were called $\hat{\beta}^{(\mathrm{test})}$. (In particular, $\hat{\beta}^{(p-p_0)}$ is not the same as $\hat{\beta}^{(\mathrm{s})}$; the former takes coordinates of $\hat{\beta}$, whereas the latter computes an entirely new vector from $X^{(\mathrm{s})}$.) Then $X\hat{\beta} = X^{(\mathrm{s})}\hat{\beta}^{(p-p_0)} + X^{(\mathrm{test})}\hat{\beta}^{(\mathrm{test})}$. Therefore,

$$v = X^{(\mathrm{s})}(\hat{\beta}^{(p-p_0)} - \hat{\beta}^{(\mathrm{s})}) + X^{(\mathrm{test})}\hat{\beta}^{(\mathrm{test})} . \tag{N19}$$

On the other hand, since $v \in V$, we have that $v = P_V(v)$, so when we take the orthogonal projection of the right-hand side of (N19), we obtain $v = P_V\big(X^{(\mathrm{test})}\hat{\beta}^{(\mathrm{test})}\big)$ since $X^{(\mathrm{s})} \perp V$. In the proof of the theorem, we saw that $v \sim \mathrm{N}(\mathbf{0}_{p_0}, I_{p_0})$. Since v is a linear transformation of $\hat{\beta}^{(\mathrm{test})}$, as is (N17), and both have the same standard normal distribution, they are equal in norm. (In fact, one must be an orthogonal transformation of the other.) This proves the desired relation, since F was obtained from v by dividing by $\hat{\sigma}^2/\sigma^2$, just as (N18) was obtained from (N17).

When the F-test is used without specifying which coefficients are being tested, the default is that all the coefficients are 0 except the intercept. Thus, it has $p-1$ degrees of freedom in the numerator and $n-p$ in the denominator. Sometimes research papers do not make clear the values of n and p, but they can be deduced when the papers say how many degrees of freedom are used in reported statistical tests.

As in the preceding section, this theory depends on ϵ being standard normal. If not and if n is not large, then the theory definitely breaks down. However, if p is fixed and n is large, and X is fairly reasonable, then since $\hat{\beta}$ is approximately normal, as discussed in the preceding section, we will have that F has approximately the same distribution as when ϵ is normal.

A comment on exercise 5E1: As the back of the book says, in this case $F = t^2$. Here, it is easy to see. This equation, $F = t^2$, holds whenever $p_0 = 1$, but it is harder to see in general. It is proved in the handout, "How F Relates to t", and can also be seen from (N18). One more way to see it is to use exercise 3B17 and block-matrix inversion.

A comment on exercise 5E6: Because of this, "the" F-test and R^2 are two ways of measuring the same thing. If you have only one of them reported in a paper, you can calculate the other.

Exercise: Let Z_1 be a standard normal random variable. Let U be a random variable independent of Z_1 that takes the values ± 1 with probability 1/2 each. Define $Z_2 := UZ_1$. Is Z_2 a standard normal random variable? What is the correlation between Z_1 and Z_2? Are Z_1 and Z_2 independent? Are Z_1 and Z_2 jointly normal? Is $Z_1 + Z_2$ normal?

Exercise: Consider a model with $n = 200$ and $p = 5$. A researcher reports that adding a new regressor (thus changing p from 5 to 6) increases R^2 from 0.5 to 0.55. In this new model with $p = 6$, what is t for this new regressor?

§5.8. Data snooping.

Data snooping was discussed in sec. 29.2 of FPP. It concerns testing many null hypotheses: even if all are true, eventually some of them will give statistically significant results, leading to rejection of a (true) null hypothesis. This in turn is one of the causes of publication bias: only rejections of null hypotheses get published by most journals.

Multiple linear regression leads to a new way data snooping can occur: start with many possible regressors, test all of them individually, and keep only the ones that are "significant". Then call that your model, refit, and publish. How can we understand the effect of this kind of data snooping?

The simplest way is to simulate. Suppose that $Y = X\beta + \epsilon$ with $\beta = \mathbf{0}_{50}$, $\epsilon \sim \mathrm{N}(\mathbf{0}_{100}, I_{100})$, and X a 100×50 matrix, all of whose entries are IID $\mathrm{N}(0,1)$ random variables, independent of ϵ. Thus, $Y = \epsilon$. All assumptions of OLS hold, nothing is related to anything else. We will run a t-test on each of the 50 columns of X and keep those that pass at the 10% significance level. So each one, by definition, has a 10% chance of being kept. Therefore, the expected number kept is 10% of 50, i.e., 5. Now refit with only the columns that are kept and compute the new t-statistic of each kept column. Freedman did this simulation 1000 times. Sometimes by chance, no columns were kept: this happened 19 times. Among all the others, a total of 5213 columns were kept and their t-statistics are shown in the histogram on p. 79 of SM. Most of them are larger than 2 in absolute value: they are supposedly "statistically significant".

Freedman mentions that R^2 in the original regression of Y on all of X will be typically about 0.5; why is that? The definition is that $R^2 := 1 - \|e\|^2/\|Y\|^2$. Here, $\|Y\|^2 \sim \chi^2_{100}$ has mean 100. Now e is the orthogonal projection of ϵ onto $\mathrm{col}(X)^\perp$, which is a 50-dimensional subspace. Therefore $\|e\|^2 \sim \chi^2_{50}$ has mean 50. Therefore R^2 is about $1 - 50/100 = 0.5$.

Exercise: In this simulation example, suppose X were a fixed 100×50 data matrix of full rank. Keep $\epsilon \sim \mathrm{N}(\mathbf{0}_{100}, I_{100})$. Would the expected number of columns kept by t-tests at the 10% level still be 5? Would R^2 still be about 0.5?

A comment on 5F1: We know that $\mathrm{Cov}(\hat\beta \mid X) = \sigma^2(X'X)^{-1}$. Therefore, unless all the columns of X are pairwise orthogonal, the coordinates of $\hat\beta$ are dependent. But even if the columns are orthogonal, although the coordinates of $\hat\beta$ are independent (because they are jointly normal and uncorrelated), the corresponding t-statistics are dependent because each one involves dividing by the same (random) $\hat\sigma^2$. If we divided by σ^2 instead, which we could in this simulation, then orthogonal columns would give that the distribution of the number of columns kept is binomial.

§5.9. Discussion questions.

In 1, the random variables X_i are observable.

For 6, you can also see that Julia would get a slope of about 0 by symmetry: the dependence of Y_i on X_i is unchanged when X_i changes sign. To be a heroine, she should (and would) include an intercept in the regression. A scatter plot would show a non-linear pattern; however, if the SD of δ_i were larger, say, at least 2, then it would be hard to spot if the values of $|X_i|$ were not too large, say, at most 2: see the link on Oncourse. It might be even harder with more variables.

§5.10. End notes.

Since the assumptions behind OLS, t-tests and F-tests often seem unrealistic in practice, another type of test related to an informal null hypothesis is of interest. For large n (and with other good properties of the data), these other tests give results that are essentially the same. These are permutation tests. They go back to Wald and Wolfowitz, "Statistical tests based on permutations of the observations", *Ann. Math. Stat.* **15** (1944), 358–372 and are extended by Freedman and Lane, "A nonstochastic interpretation of reported significance levels", *J. Bus. Econ. Stat.* **1**, 4 (1983), 292–298. For example, suppose we regress Y on X and Z, where Z has p_0 columns. We are thinking that Z does not belong, so X would be the small design matrix and $[X\ Z]$ would be the full design matrix. The informal null hypothesis is that given X, there is no further relationship between Y and Z. The test goes as follows: Regress Y on X, getting $Y = X\hat{\beta} + e$. Now suppose that π is a permutation of $1, 2, \ldots, n$ and that e^π is the result of applying this permutation to the coordinates of e. Define $Y^\pi := X\hat{\beta} + e^\pi$. Next, compute F for Y^π regressed on X and Z, corresponding to testing whether Z belongs; call the result F^π. If we choose π at random many times, we can compute such an F^π each time and get an empirical distribution of the resulting values of F^π. The more times we simulate, the closer this empirical distribution will be to the theoretical distribution of F^π, called the permutation distribution of F^π. You can simulate permutations in MATLAB with the command `randintrlv`. The theorem (with some conditions on the data omitted here) is that as $n \to \infty$, this permutation distribution is close to the F distribution with p_0 degrees of freedom in the numerator and $n - p$ in the denominator, where p is the number of columns of $[X\ Z]$, i.e., the permutation distribution is close to $\chi^2_{p_0}/p_0$. But even without such a theorem, we can simply see how F compares without any permutation to an empirical distribution of F^π for many permutations π; this gives an idea of whether indeed the relationship of Y to Z given X is "accidental" or not. No assumptions are made on any statistical model about where the data arose.

§5.∞. ANOVA.

In FPP, we saw how to compare two sample means, e.g., to test whether the underlying population means are the same. Sometimes one wants to test whether several different means are all the same: this is like extending the t-test to an F-test. It is the first example of ANOVA, which stands for "analysis of variance". In general, ANOVA considers special cases of multiple linear regression where all columns are dummy variables, i.e., have only

1s and 0s in them. The questions of interest are usually ones like the F-test we looked at in section 7.

One-factor ANOVA, or the one-way layout, is concerned with the following model. Suppose that n subjects come in I groups, with n_i subjects in the ith group for $1 \le i \le I$. Thus, $n = \sum_{i=1}^{I} n_i$. Often all n_i are equal. The model is

$$Y_{i,j} = a_i + \epsilon_{i,j} \tag{N20}$$

for $1 \le i \le I$ and $1 \le j \le n_i$, with $\epsilon_{i,j}$ being independent IID normal random variables with mean 0, independent of assignment of subject to group. Thus, the response variables of subjects in the ith group have mean a_i. How do we test the null hypothesis that all a_i are equal? For example, the ith group may represent the result of the ith treatment; do all treatments have the same mean effect?

We put this into the OLS framework by writing $Y = X\beta + \epsilon$, where Y is a vector of all responses $Y_{i,j}$, likewise for ϵ, $\beta = [a_1 \ a_2 \ \cdots \ a_I]'$, and X is an $n \times I$ matrix of 1s and 0s. The row of X corresponding to (i,j) has a 1 only in the ith column. Clearly $X'X$ is a diagonal matrix with n_i in the ith place. The estimate of a_i is simply the sample mean of the ith group. But it is not immediately clear how to do an appropriate F-test. One way to do so would be to rewrite the equations in terms of an overall mean, a, and deviations from that mean, $a_i - a$, thereby introducing a column of 1s:

$$Y_{i,j} = a + (a_i - a) + \epsilon_{i,j}\,.$$

However, we could not then keep all I columns of X we have already, since the sum of the existing columns of X is $\mathbf{1}_n$. (In the above equation, this is reflected in the parameter constraint that $\sum_{i=1}^{I}(a_i - a) = 0$.) We would have to eliminate one of the columns. This works, but it is nicer to do it another way.

Note that the F-test did not depend very much on X and $X^{(\mathrm{s})}$: it depended exclusively on their column spaces, W and $W^{(\mathrm{s})}$. Furthermore, in the present instance, we want to compare the full model (N20) to another model,

$$Y = a\mathbf{1}_n + \epsilon\,. \tag{N21}$$

Thus, we simply ignore the fact that X does not have $\mathbf{1}_n$ as a column, because $\mathbf{1}_n \in W$. Defining $W^{(\mathrm{s})}$ as the span of $\mathbf{1}_n$, we can simply compute F using these two models and make our test. It does not matter that the columns of $X^{(\mathrm{s})}$ are not a subset of the columns of X; what matters is that $W^{(\mathrm{s})}$ is a subspace of W. We say that the model (N21) is nested in (N20) because the former column space lies inside the latter column space.

One normally sees the numerator and denominator of F written out with various terms named "SSx", which stands for "sum of squares" of x, with various possibilities for x. This is because $\|e\|^2$ is indeed a sum of squares, and $\|e^{(\mathrm{s})}\|^2 - \|e\|^2$ can also be written as a sum of squares, as in the proof of theorem 5.3. These take special forms in the case of ANOVA.

The next more complicated form of ANOVA is two factor, or two-way layout. Here, there are two different types of treatments. (Of course, there can be more than two, but the math is not much different.) One question is whether all treatments of the first type, say, have the same mean effect, so that the only differences in responses are due to the second treatment (and randomness). Another question is whether the mean effects of the two types of treatments are simply additive, or whether there is an interaction between the treatments. The notation for this model uses two subscripts, one for each of the two types of factors, and a third subscript for the individual. Thus, the full model for both preceding questions is

$$Y_{i,j,k} = a_{i,j} + \epsilon_{i,j,k} \tag{N22}$$

for $1 \le i \le I$, $1 \le j \le J$, and $1 \le k \le n_{i,j}$. The total number of subjects is $n := \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j}$. The assumptions are that $\epsilon_{i,j,k}$ are IID normal random variables with mean 0, independent of assignment of subject to group. If we want to test that $a_{i,j}$ depends only on j, then we are considering the smaller model

$$Y_{i,j,k} = a_j + \epsilon_{i,j,k} \tag{N23}$$

with the same assumptions as above. To see that this is indeed a smaller (nested) model, note that the column space of any linear regression model $Y = X\beta + \epsilon$ consists of the columns Y that have the form $X\beta$, i.e., those Y that satisfy the model equation with $\epsilon = \mathbf{0}_n$ and arbitrary β. Now it is clear that if Y can be written with $Y_{i,j,k} = a_j$, then we can also write Y as $Y_{i,,j,k} = a_{i,j}$ for some numbers $a_{i,j}$: we simply take $a_{i,j} := a_j$ for all i. Thus, the second model is indeed smaller and we can make a corresponding F-test. The other question, whether the mean effects of the two types of treatments are simply additive, is slightly more complicated. Now, the model is

$$Y_{i,j,k} = a_i + b_j + \epsilon_{i,j,k} \tag{N24}$$

with the same assumptions as above. However, this does not specify a_i and b_j uniquely: we could add a constant to all a_i as long as we subtract the same constant from all b_j. The problem is that not all columns are linearly independent the way we have written it. We need to leave out one column; any one will do. But for the purpose of seeing that the

column space of (N24) is contained within that of the full model (N22), we may keep all the columns. (We could require, say, that $\sum_j b_j = 0$ if we wished.) Similar reasoning as before shows that (N24) is indeed a smaller model than the full model (N22), so we can again make an F-test.

Finally, another model is called analysis of covariance, or ANCOVA. This is the same as ANOVA except that in addition to the dummy variables, there are one or more covariates on the right-hand side that are not dummy variables.

If we want to relax the condition that all errors have the same variance, but instead allow them to depend on the group, then we use (feasible) GLS.