

## Ordinary Least Squares

Here are some proofs of theorems in Chapters 4–5 that may shed more light on why they are true. The first proofs are the same as in the book, but with notation that should make them easier to follow.

Recall that  $X$  is an  $n \times p$  matrix of rank  $p$ . The columns of  $X$  form a basis of  $W := \text{col } X$ . Define

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

This is a  $p \times n$  matrix. We saw in Problem 3 of the Linear Algebra Homework (LinAlg#3) that  $Q$  gives us  $\hat{\beta}$ , i.e.,  $QY = \hat{\beta}$ . This means that  $Q$  takes a vector,  $Y$ , projects it orthogonally onto  $W$ , and then gives us the coordinates of the result with respect to the columns of  $X$ . In particular,  $Q(X\gamma) = \gamma$  for all  $\gamma$ . (You can also think of this as a special case of  $QY = \hat{\beta}$  with  $Y := X\gamma$ . It's also obvious from  $QX = I$ , but that matrix multiplication does not give particular insight.)

**Theorem 4.2.** *If  $E(\epsilon | X) = \mathbf{0}_n$ , then  $E(\hat{\beta} | X) = \beta$ .*

*Proof.* Since  $Y = X\beta + \epsilon$ , we have

$$\hat{\beta} = QY = Q(X\beta + \epsilon) = \beta + Q\epsilon.$$

Also, since  $Q$  is a function of  $X$ , we have  $E(Q\epsilon | X) = QE(\epsilon | X)$ . Thus,  $E(\hat{\beta} | X) = E(\beta + Q\epsilon | X) = \beta + QE(\epsilon | X) = \beta + Q\mathbf{0}_n = \beta$ . ■

**Theorem 4.3.** *If  $E(\epsilon | X) = \mathbf{0}_n$  and  $\text{Cov}(\epsilon | X) = \sigma^2 I_n$ , then  $\text{Cov}(\hat{\beta} | X) = \sigma^2 (X'X)^{-1}$ .*

*Proof.* We saw in the preceding proof that  $\hat{\beta} - \beta = Q\epsilon$ . Thus,

$$\begin{aligned} \text{Cov}(\hat{\beta} | X) &= E((\hat{\beta} - \beta)(\hat{\beta} - \beta)' | X) = E(Q\epsilon\epsilon'Q' | X) = QE(\epsilon\epsilon' | X)Q' = Q\sigma^2 I_n Q' \\ &= \sigma^2 QQ'. \end{aligned}$$

Now  $QQ' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$ , which completes the proof. ■

We also saw in the homework LinAlg#3 that if  $H := X(X'X)^{-1}X' = XQ$ , then  $H$  is the matrix of  $P_W$ , the orthogonal projection onto  $W$ . This matrix carries out the calculations of OLS:  $Y = X\hat{\beta} + e$ , with  $X\hat{\beta} \in W$  and  $e \perp W$ , and  $HY = XQY = X\hat{\beta}$ .

In the handout on the  $t$ -distribution, we proved some things for the normal distribution that actually hold more generally. Namely, if  $Z$  is a random vector with  $E(Z) = \mathbf{0}_n$  and  $\text{Cov}(Z) = I_n$ , then these same two equations hold for the vector  $RZ$  when  $R$  is any

orthogonal matrix. In addition, if  $P$  is an orthogonal projection onto a subspace  $W$  of dimension  $r$ , then  $PZ$ , in coordinates corresponding to an orthonormal basis for  $W$ , is a vector with mean  $\mathbf{0}_r$  and covariance matrix  $I_r$ . (For a normal distribution, all we added to this was the fact that the mean and covariance matrix determine the distribution once we know it is normal, and we had separate arguments that  $RZ$  and  $PZ$  were normal.) Look at the handout again to verify that we proved these things in general.

**Theorem 4.4.** *If  $E(\epsilon | X) = \mathbf{0}_n$  and  $\text{Cov}(\epsilon | X) = \sigma^2 I_n$ , then  $E(\hat{\sigma}^2 | X) = \sigma^2$ .*

*Proof.* Recall that our estimate of  $\sigma$  comes from  $\hat{\sigma}^2 = \|e\|^2/(n-p)$ . Now,  $Y - X\beta = \epsilon$  and the covariance matrix of  $\epsilon$  gives  $\sigma$ . The problem in using this is that we don't know  $\beta$ . But if we project onto  $W^\perp$ , we can eliminate this difficulty:

$$P_{W^\perp}(\epsilon) = P_{W^\perp}(Y - X\beta) = P_{W^\perp}(Y),$$

which does not depend on  $\beta$ . Also, we know what this projection is by LinAlg#3: it is  $e$ . Since  $e = P_{W^\perp}(\epsilon)$ , what we discussed before this theorem tells us that, in orthonormal coordinates of  $W^\perp$ ,  $E(e | X) = \mathbf{0}_{n-p}$  and  $\text{Cov}(e | X) = \sigma^2 I_{n-p}$ . In particular, the expected squares of these coordinates of  $e$  are each  $\sigma^2$ , so their sum, which is  $E(\|e\|^2 | X)$ , is  $(n-p)\sigma^2$ . This gives the result:  $E(\hat{\sigma}^2 | X) = E(\|e\|^2/(n-p) | X) = (n-p)\sigma^2/(n-p) = \sigma^2$ . ■

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

*Proof.* We've already shown that the mean of  $\hat{\beta}$  is  $\beta$  (Theorem 4.2) and the covariance of  $\hat{\beta}$  is  $\sigma^2(X'X)^{-1}$  (Theorem 4.3). Clearly  $Y$  is normal, whence so is  $QY = \hat{\beta}$ . This shows that the distribution of  $\hat{\beta}$  is as claimed.

Furthermore, since  $\epsilon$  is  $N(\mathbf{0}_n, \sigma^2 I_n)$ , the orthogonal projections  $H\epsilon$  and  $(I-H)\epsilon$  are independent normals; 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  $\sigma^2 I_{n-p}$ . Now

$$H\epsilon = H(Y - X\beta) = X\hat{\beta} - X\beta \quad \text{and} \quad (I-H)\epsilon = (I-H)Y = e.$$

The second of these tells us that  $e \sim N(\mathbf{0}_{n-p}, \sigma^2 I_{n-p})$ , so  $\|e\|^2 \sim \sigma^2 \chi_{n-p}^2$ , as claimed. Finally, the two equations together also imply that  $e$  is independent of  $X\hat{\beta} - X\beta$ , whence of  $X\hat{\beta}$  and therefore of  $QX\hat{\beta} = \hat{\beta}$ . ■

Now fix  $k$  between 1 and  $p$  and consider the  $k$ th coordinate of  $\hat{\beta}$ , which we call  $\hat{\beta}_k$ . This is our estimate of  $\beta_k$ . When we test the null hypothesis that  $\beta_k = 0$ , we consider the statistic  $t := \hat{\beta}_k / \widehat{\text{SE}}$ . What's  $\widehat{\text{SE}}$ ? It's our estimate of the SE of  $\hat{\beta}_k$ . Now  $\text{Var}(\hat{\beta}_k | X)$  is

the  $(k, k)$ -element of  $\text{Cov}(\hat{\beta} \mid X) = \sigma^2(X'X)^{-1}$  (according to Theorem 4.3); we'll take its square root to get the SE. However, since we don't know  $\sigma$ , we need to estimate it. Thus,  $\widehat{\text{SE}}$  is  $\hat{\sigma}$  times the square root of the  $(k, k)$  element of  $(X'X)^{-1}$ .

**Corollary.** *Condition on  $X$ . Under the null hypothesis, the distribution of  $t$  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  $a$ . Thus,

$$t = \frac{\hat{\beta}_k}{\widehat{\text{SE}}} = \frac{\hat{\beta}_k}{\hat{\sigma}a} = \frac{\hat{\beta}_k}{\|e\|a/\sqrt{n-p}}.$$

By Theorem 5.2, it follows from this that the numerator of  $t$  and the denominator of  $t$  are independent, and  $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$ . Since the null hypothesis says that  $\beta_k = 0$ , we get that  $\hat{\beta}_k \sim N(0, \sigma^2a^2) = \sigma aN(0, 1)$ . Again, Theorem 5.2 tells us that  $\|e\| \sim \sigma\chi_{n-p}$ , so  $\widehat{\text{SE}} \sim a\sigma\chi_{n-p}/\sqrt{n-p}$ . This means that

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

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  $t := (\hat{\beta}_k - \beta_k)/\widehat{\text{SE}}$  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  $\beta$ , i.e.,  $c'\beta$  for some fixed  $c$ . A similar proof shows that  $t := (c'\hat{\beta} - c'\beta)/\widehat{\text{SE}}$  has Student's  $t$ -distribution with  $n - p$  degrees of freedom. Here,  $\widehat{\text{SE}} = \hat{\sigma}\sqrt{c'(X'X)^{-1}c}$  since  $\text{Cov}(c'\hat{\beta} \mid X) = \sigma^2c'(X'X)^{-1}c$ . For example, we might be interested in testing whether  $\beta_1 = \beta_2$ . We would use  $c = (1, -1, 0, 0, \dots, 0)$  and we would have  $t = (\hat{\beta}_1 - \hat{\beta}_2)/\widehat{\text{SE}}$  since our null hypothesis would be that  $c'\beta = 0$ .

For the  $F$ -test, we consider a smaller design matrix  $X^{(s)}$ . This is formed from the first  $p - p_0$  columns of  $X$ . Its column space is a smaller space  $W^{(s)}$  inside of  $W$ . We form the statistic

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

**Theorem 5.3.** *Condition on  $X$ . Assume the null hypothesis  $Y = X^{(s)}\beta + \epsilon$ , where  $\epsilon$  has distribution  $N(\mathbf{0}_n, \sigma^2I_n)$ . Then the numerator and denominator of  $F$  are independent with  $\|\hat{Y}\|^2 - \|\hat{Y}^{(s)}\|^2 \sim \sigma^2\chi_{p_0}^2$  and  $\|e\|^2 \sim \sigma^2\chi_{n-p}^2$ .*

*Proof.* Recall that  $Y = \hat{Y} + e$  with  $e = P_{W^\perp}(Y)$ . Likewise,  $Y = \hat{Y}^{(s)} + e^{(s)}$  with  $e^{(s)} = P_{(W^{(s)})^\perp}(Y)$ . Thus,

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

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

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

Since  $W^{(s)} \subseteq W$ , we have  $(W^{(s)})^\perp \supseteq W^\perp$ . Let's write  $(W^{(s)})^\perp = W^\perp \oplus U$  with  $U \perp W^\perp$ . Here,  $\dim U = \dim (W^{(s)})^\perp - \dim W^\perp = (n - (p - p_0)) - (n - p) = p_0$ . Now

$$e = P_{W^\perp}(Y) = P_{W^\perp} P_{(W^{(s)})^\perp}(Y) = P_{W^\perp}(e^{(s)}).$$

This means that

$$e^{(s)} = e + (e^{(s)} - e)$$

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

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

As we saw in the proof of Theorem 5.2, the null hypothesis implies that  $e^{(s)} \sim N(\mathbf{0}_{n-(p-p_0)}, \sigma^2 I_{n-(p-p_0)})$  if we write  $e^{(s)}$  in orthonormal coordinates of  $(W^{(s)})^\perp$ . This implies (as in the proof of Theorem 5.2 again) that  $e$  and  $e^{(s)} - e$  are independent normal random vectors. If we write  $e^{(s)} - e$  in orthonormal coordinates of  $U$ , we have  $e^{(s)} - e \sim N(\mathbf{0}_{p_0}, \sigma^2 I_{p_0})$ . This means that  $\|e^{(s)} - e\|^2 \sim \sigma^2 \chi_{p_0}^2$ . Likewise,  $\|e\|^2 \sim \sigma^2 \chi_{n-p}^2$ , so the proof is complete. ■

Here are some questions to prepare for discussion.

1. Suppose that  $X$  and  $Y$  are *jointly normal random vectors* in  $\mathbb{R}^n$ ; in other words, the set of all the coordinates of  $X$  and  $Y$  together forms a jointly normal collection of (real-valued) random variables. If  $X \perp Y$  holds with probability 1, must it be that  $X$  and  $Y$  are independent? Or must it be that  $X$  and  $Y$  are not independent?
2. Suppose that  $X \sim N(0, I_n)$  and  $W_1, W_2$  are two fixed subspaces of  $\mathbb{R}^n$  with  $W_1 \perp W_2$ . Let  $Y_i := P_{W_i}(X)$  for  $i = 1, 2$ . Is  $Y_1 \perp Y_2$ ? Are  $Y_1$  and  $Y_2$  two independent normal random vectors? Are they jointly normal? What if  $X \sim N(0, G)$  and  $G \neq I_n$ ?