Notes on Chap. 4: Multiple Regression

§4.1. Introduction.

We often analyze coin tossing by the mathematical theory of probability. In other words, we *model* the coin tossing of the real world by the ideal theory. We can test whether this is a good model by finding predictions of the model and then seeing whether they are borne out in the real world. Is the result of coin tossing truly random? No, it is not. It is determined by the laws of physics, and quantum mechanics (which does have true randomness in it—in theory), plays little role. The result of a coin toss is determined by how it leaves one's hand and where it falls. So why do we think of it as random? One reason is that the mathematical model turns out to be a good one; another, more primitive, reason is that we don't know how a tossed coin will land. In this sense, we are modeling our ignorance. Determining weights (Chap. 6 of FPP) and Hooke's law (Sec. 12.2 of FPP and Sec. 2.3 of SM) are good examples of modeling our ignorance: we don't know the source of measurement error, nor the exact distribution of such error, but models can be useful in making deductions from such noisy measurements (Chap. 24 of FPP and all of SM).

To a large extent, it is to model ignorance that people use statistical models, whether to analyze observational or experimental data. In other words, we don't (fully) know what determined the data, nor what future observations will be. It might be that a statistical model describes well the observations, in which case the model might be useful for prediction and possibly even for causal explanation. More generally, a statistical, mathematical, or scientific model is an approximation of reality. Especially in statistics, one does not expect the model to be a faithful representation of reality. In this course, we will be especially concerned with how well the model approximates reality. In brief, since we are usually modeling our ignorance, it is usually very difficult to tell how good the approximation is. That brief summary requires a lot of explanation; we will look at this in detail.

In Chapters 4–5, we develop the mathematical theory behind the most common statistical model, multiple linear regression. Then in Chapter 6, we look in detail at applications and at how they relate to the theory. These models make assumptions (often because we don't have experimental data). If the assumptions are not good, then the model might not

be good. Furthermore, it can be very hard to evaluate how good the assumptions really are. They are much more complicated than the assumptions behind the theory of tossing a coin. For this reason, we will look quite closely at the assumptions and at what happens when they are violated. This makes our journey ahead hard: you have probably never before been asked to consider what happens when the hypotheses of theorems you learned don't hold. A mathematician does this as part of his or her research, but undergrads don't. So even from the purely mathematical point of view, we will do things you were never asked to do before. But we will also need to think about the real world, which is quite complicated, as we saw in FPP. When we try to understand how well the complicated math reflects the complicated real world, we will have a lot to think about!

We will begin with a short review of basic statistics from FPP, but set in some new notation to prepare for the more complicated statistics to come.

Suppose we have some data, a list $(y_1, \ldots, y_n)$ of numbers. We could summarize the data by their mean, $\bar{y}$. Then we could look at the difference (deviation) of each number from the mean, $e_i := y_i - \bar{y}$. These deviations sum to 0; their typical size could be measured by their root mean square (r.m.s.), which is the standard deviation (sd) of the data, $\mathrm{sd}(y)$. Thus, we are summarizing the 1-dimensional data y_i by a single point, $\bar{y}$, which is a 0-dimensional subset of $\mathbb{R}$, and we are summarizing the spread around that summary point by the r.m.s. of the e_i. Now, we could write

$$y_i = \bar{y} + e_i \,.$$

There are n such equations, one for each data point. We could put them all together by using lists, i.e., vectors, of length n:

$$y = \bar{y}\mathbf{1}_n + e \,,$$

where $y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{1}_n := \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ denotes the vector of n 1s, and $e := \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$. There need not be any probability model for such data.

Now, however, suppose that these data came as a result of IID sampling from a box (or from some other probability mechanism). Then we could model the numbers by the distribution of the box, or, more simply, by the mean μ of the box and the SD σ of the box. This would be a better model than using the mean and sd of the data. However, if we don't know the mean and SD of the box, then we could estimate them by the data.

To write this as a model, call Y_i the ith sample, so that in the data, we found $Y_i = y_i$. (We usually will write Y_i for both.) We would say that

$$Y_i = \mu + \epsilon_i \,,$$

where ϵ_i is the ith deviation from μ, so ϵ_i are IID with mean 0 and SD σ. We don't observe ϵ_i and, unless we know μ, we can't calculate ϵ_i either. Note that $\bar{y} \neq \mu$ and $\epsilon_i \neq e_i$. Using vectors, these equations become

$$Y = \mu \mathbf{1}_n + \epsilon \,,$$

where $Y := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ and $\epsilon := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$. Here we call μ a parameter of the model; if we don't know it, then we estimate it. We would use $\bar{y}$ to estimate μ. In the notation we are going to use, we would write this as $\hat{\mu} := \bar{y}$. Properties of the vector ϵ form part of our assumptions about our sampling (maybe our sampling wasn't as good as we thought!). Namely, we assume that ϵ_i are IID with mean 0. If we want to know how far off $\bar{y}$ is likely to be from μ and if we don't know σ, then we would use the sd of the data to estimate σ with the equation

$$\hat{\sigma} := \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} e_i^2} = \frac{\|e\|}{\sqrt{n-1}} \,. \tag{N1}$$

Let's turn now to simple linear regression. In FPP, we used this mainly as a summary of data and, if it was football shaped, then for prediction as well. To say that the data is football shaped means that it *seems to come* from a bivariate normal distribution. This is a model assumption. It is easy to see where probability enters in Hooke's law, even though the cause of measurement error is unknown. It is a little harder, but still pretty easy, to see probability in the height data of fathers and sons. It is a lot harder to see probability in much of the other data, such as education level versus income. But we can still imagine how randomness is part of everyday life, even if "randomness" here is not the ideal mathematical notion of randomness.

So suppose now that we have data points (x_i, y_i) for $i = 1, \ldots, n$. We could draw them in a scatter plot. We could summarize the data by the means, $\bar{x}$ and $\bar{y}$, the sds, $\mathrm{sd}(x)$ and $\mathrm{sd}(y)$, and the correlation coefficient, r. Alternatively, we could summarize the 2-dimensional data by the regression line, a 1-dimensional set which is the line of best fit. "Best fit" means that its slope m and intercept b have the property that if we write

$$y_i = m x_i + b + e_i \,,$$

then the root mean square of the vertical deviations, $e_1, \ldots, e_n$, is minimized. We could summarize how well the regression line fits the data either by r or by the root-mean-square deviation from the line, $\sqrt{1 - r^2}\sigma_y$. Recall that there are two regression lines; we choose

which one to use based on which variable we want to predict from the other. In vector notation, we can write

$$y = mx + b\mathbf{1}_n + e\,,$$

where $x := \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. We could also write this by using matrices:

$$y = [\mathbf{1}_n \; x] \begin{bmatrix} b \\ m \end{bmatrix} + e\,.$$

Another form for the n equations is

$$y_i = [1 \; x_i] \begin{bmatrix} b \\ m \end{bmatrix} + e_i$$

for each i. For example, in table 12.1 of FPP (table 2.1 of SM) for Hooke's law, we have the data points (x_i, y_i) for $i = 1, \ldots, 6$ given (in order) as (0, 439.00), (2, 439.12), (4, 439.21), (6, 439.31), (8, 439.40), (10, 439.50). Here, the units are kg and cm. This gives

$$y = \begin{bmatrix} 439.00 \\ 439.12 \\ 439.21 \\ 439.31 \\ 439.40 \\ 439.50 \end{bmatrix} \qquad \text{and} \qquad x = \begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 8 \\ 10 \end{bmatrix}.$$

If we put the regression equation in matrix form, we obtain

$$\begin{bmatrix} 439.00 \\ 439.12 \\ 439.21 \\ 439.31 \\ 439.40 \\ 439.50 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} + e\,.$$

We found that $m = 0.05$ cm/kg and $b = 439.01$ cm in FPP. Soon we will use matrix methods to do the same. We did not calculate e in FPP, but we will soon.

Now suppose that we model our data. Maybe they came from IID samples of pairs of numbers; maybe they came some other way, such as an experiment like putting weights on a spring. Maybe the data is entirely observational. In mathematical terms, we might use the model

$$Y_i = aX_i + c + \epsilon_i\,, \tag{N2}$$

where a and c are (perhaps unknown) parameters and ϵ_i are IID with mean 0. In Hooke's law, we believe that a and c describe true physical properties of the spring; we want to measure them, but because of measurement error, this requires use of statistics. In the data, we found $Y_i = y_i$ and $X_i = x_i$. But $a \neq m$, $c \neq b$ and $\epsilon_i \neq e_i$; instead, we estimate a by m and c by b from the data. We did not discuss any statistical aspects of these estimates in FPP. We will do so here in Chap. 4. Again, in vector notation the model becomes

$$Y = aX + c\mathbf{1}_n + \epsilon\,,$$

where $X := \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, or equivalently

$$Y = [\mathbf{1}_n \; X] \begin{bmatrix} c \\ a \end{bmatrix} + \epsilon\,.$$

Suppose that the distribution of each (X_i, Y_i) is bivariate normal. Rather than look at actual data, we can look at the probability distribution underlying the model and form an "ideal" regression line, the one that would be the limit of taking infinitely many IID samples. Although we did not make this distinction (between the regression line of actual data and the ideal regression line) before, we saw that the ideal regression line is the line of means of vertical slices, i.e., of conditional means. In other words, $\big(X_i, E(Y_i \mid X_i)\big)$ lies on the ideal regression line:

$$E(Y_i \mid X_i) = aX_i + c\,. \tag{N3}$$

Comparing equation (N3) to (N2), we see that we don't actually need a bivariate normal assumption; it suffices that $E(\epsilon_i \mid X_i) = 0$. Usually we will want to make slightly stronger statements: even given *all* the values $X_1, \ldots, X_n$, still the conditional means lie on the line:

$$E(Y_i \mid X) = aX_i + c\,.$$

This is then equivalent to $E(\epsilon_i \mid X) = 0$. This is a stronger assumption than $E(\epsilon_i \mid X_i) = 0$ if we do not assume that all (X_i, Y_i) are independent.

Finally, we move to multiple linear regression, such as Yule's on p. 11 of SM. If we have 3-dimensional data, (u_i, v_i, y_i) for $i = 1, \ldots, n$, and we want to predict y_i from (u_i, v_i) via a linear relationship, say,

$$y_i = a + bu_i + cv_i + e_i\,, \tag{N4}$$

where e_i is the error of the prediction (called the ith *residual*), then we might want to find (a, b, c) to minimize the r.m.s. error. The equation (to change briefly the notation)

$$z = a + bx + cy$$

is the equation of a plane in 3 dimensions. (If $a = 0$, then it consists of the points (x, y, z) such that $bx + cy - z = 0$, i.e., $(b, c, -1) \cdot (x, y, z) = 0$. These are the points that are orthogonal to $(b, c, -1)$. For $a \neq 0$, we just translate each such point in the vertical direction by a. Thus, a is the intercept with the vertical axis.) The plane that minimizes the r.m.s. error is called the *regression plane*, and we can summarize how well it fits the data by the r.m.s. error. In matrix notation, (N4) becomes

$$y_i = [1 \ u_i \ v_i] \begin{bmatrix} a \\ b \\ c \end{bmatrix} + e_i$$

or

$$y = [\mathbf{1}_n \ u \ v] \begin{bmatrix} a \\ b \\ c \end{bmatrix} + e\,,$$

where $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$, and $v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$.

Yule's equation has even more variables on the right-hand side. Rather than introduce notation for each one, we use a common notation. The left-hand side will generally be y_i, while we will use a *row* vector x_i for all the right-hand side variables. In addition, rather than different names for each parameter, we use a column vector α for all of them. Thus, the linear equation becomes

$$y_i = x_i\alpha + e_i$$

for each i, or, altogether,

$$y = x\alpha + e\,,$$

where $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ is a matrix, called the *design matrix*. It usually has a column of 1s. The equation $y_i = x_i\alpha$ is a hyperplane, a set of dimension $n - 1$. Generally we will use capital letters Y and X, both for data and for the model, but α comes from the data, whereas the model has a parameter vector β. We will estimate β by α, which will make the hyperplane fit the data best in the sense of minimizing the r.m.s. error.

More names: Y is called the *dependent variable*, the *response variable*, or the *regressand*. We are "explaining" or "predicting" or "modeling" Y by the variables that form the columns of X. The columns of X are called the *explanatory variables*, the *independent variables*, the *control variables*, the *regressors*, or the *covariates*. Note that the names "dependent" and "independent" have nothing to do with probability. They may have nothing to do with causality either. But it is true that the columns of X will be linearly independent; otherwise linear combinations of them can be written in more than one way and it does not make sense to have separate parameters for all the columns.

So the model is

$$Y_i = X_i\beta + \epsilon_i$$

for each i, or, altogether,

$$Y = X\beta + \epsilon\,.$$

Names for ϵ are *random errors* or *disturbance terms*. We usually do not know β and cannot observe or calculate ϵ_i. But we assume that y_i and x_i are *observed values* of the random variable Y_i and the random vector X_i. This is a statement of how the model relates to reality: y_i and x_i come from reality, whereas Y_i and X_i are part of the model. (It is *not true* that e_i is an observed value of ϵ_i because $\alpha \neq \beta$.) If we assume that $E(\epsilon_i \mid X) = 0$, then we get

$$E(Y_i \mid X) = X_i\beta\,,$$

i.e., the conditional means lie on a hyperplane. Alternatively, we can write this as

$$E(Y \mid X) = X\beta\,,$$

which says that the conditional expectation of Y lies in the column space of X. Probably the easiest way to think of this is that $E(Y_i \mid X_i)$ is linear in X_i, i.e., it equals $X_i\beta$ for some β; that this β is the same for all subjects i; and that this conditional expectation $E(Y_i \mid X_i)$ does not change even if we condition on the covariates of all other subjects, i.e., $E(Y_i \mid X_i) = E(Y_i \mid X)$.

There are various forms of the assumptions that come with the model:

(A1) The simplest is that ϵ_i are IID, that ϵ is independent of X, and that $E(\epsilon_i) = 0$.

(A2) A weaker set of assumptions is that ϵ_i are independent and $E(\epsilon_i \mid X) = 0$.

More fully, we are assuming that *there exists some vector* β such that the vector $\epsilon := Y - X\beta$ satisfies one of the previous sets of assumptions. Often it is also assumed that ϵ is normal.

Although this is such a common model, I cannot give you any interesting examples of it (beyond simple linear regression) where we know the assumptions hold: there don't

seem to be any such examples! (If you find any, please let me know. Multivariate normal examples do exist in scientific and engineering situations, but those are too technical to be of wide interest.) In fact, no one believes that they hold precisely in most applications of the model. Nevertheless, the hyperplane of best fit is often a useful summary and it can even be useful in prediction. If one can frequently test a model by new data, then one can see how well it works in practice, regardless of any understanding of the model. But for causal explanation, an understanding is essential. Also, keep in mind the dictum that the difference between theory and practice is greater in practice than in theory, even for online experiments that do not involve any regression techniques.

Optional exercise: "Football shaped" in FPP meant that data was similar to a bivariate normal distribution. That implied several good things about simple linear regression, as reviewed above. Higher-dimensional footballs mean multivariate normality. This exercise shows that all the assumptions we would want for multiple linear regression hold for IID samples from a multivariate normal distribution. Suppose that $(Y_1, U_1, \ldots, U_k)$ are jointly normal random variables. Prove that there exist numbers $\alpha, \beta_1, \ldots, \beta_k$ and a normal random variable ϵ_1 that has mean 0 and is independent of $(U_1, \ldots, U_k)$ such that

$$Y_1 = \alpha + \sum_{i=1}^{k} \beta_i U_i + \epsilon_1 \,.$$

Hint: Match means and covariances. Alternatively, show that with $X_1 := [1 \; U_1 \; \cdots \; U_k]$, we can use $\begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = E(X_1'X_1)^{-1}E(X_1'Y_1)$ when there is no linear relationship among $1, U_1, \ldots, U_k$.

What can we say about this model in general? Denote the number of right-hand side variables by p; this includes the "1" if there is an "intercept". In the preceding development, we began with $p = 1$, representing just 1-dimensional data, where we summarized the data by the mean. Then we moved to $p = 2$ for 2-dimensional data, summarized by a line.

In general, the design matrix X is an $n \times p$ matrix of rank p. The other data is an $n \times 1$ column vector Y. We are going to summarize the n-dimensional data by an $(n-1)$-dimensional hyperplane.

The columns of X form a basis of $W := \operatorname{col} X$. This is a p-dimensional subspace of $\mathbb{R}^n$; it is *not* the hyperplane of best fit. Define

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

This is a $p \times n$ matrix. You saw in Problem 3 of the Linear Algebra Homework (LinAlg#3) that Q gives us the coefficients $\hat{\beta}$ for the hyperplane of best fit, 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 γ.

That is, you proved the following theorem:

Theorem 4.1. *If X has rank p and we define $e := Y - X\hat{\beta}$, then $e \perp X$ (shorthand for $e \perp W$) and $\|Y - X\gamma\|$ is minimized exactly when $\gamma = \hat{\beta}$.*

So $\hat{\beta}$ is "best" in a certain precise sense.

To see the calculations written out with data for Hooke's law, see example 4.1 in SM. We get that

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix}' \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \\ 1 & 10 \end{bmatrix}' \begin{bmatrix} 439.00 \\ 439.12 \\ 439.21 \\ 439.31 \\ 439.40 \\ 439.50 \end{bmatrix} = \begin{bmatrix} 439.01 \text{ cm} \\ .05 \text{ cm/kg} \end{bmatrix} .$$

The method of finding $\hat{\beta}$ that we used is called "ordinary least squares", or "OLS" for short. It may be best to think of this as simply the orthogonal decomposition of Y with respect to $\mathbb{R}^n = W \oplus W^\perp$, i.e., as *defining* $\hat{\beta}$ and e by the conditions

$$Y = X\hat{\beta} + e \text{ with } e \perp W .$$

This does indeed define them because there is a unique way to write $P_W(Y)$ as a linear combination of the columns of X and then it must be that what is left, e, equals $P_{W^\perp}(Y)$.

Now we discuss another way that $\hat{\beta}$ is "good". This involves statistical assumptions. It says that $\hat{\beta}$ is unbiased, even conditioning on X. This is a nice feature to have, though it is not crucial.

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

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

$$\hat{\beta} = QY = Q(X\beta + \epsilon) = \beta + Q\epsilon. \tag{N5}$$

Also, since Q is a function of X, we have $E(Q\epsilon \mid X) = QE(\epsilon \mid X)$. Thus, taking conditional expectation in (N5) yields

$$E(\hat{\beta} \mid X) = E(\beta + Q\epsilon \mid X) = \beta + QE(\epsilon \mid X) = \beta + Q\mathbf{0}_n = \beta\,. \qquad \blacksquare$$

A comment on exercises 4A3 and 4A4(b): The reason why we might want ways to validate our assumptions is that it is very hard to know how well they hold. When we form a model, we can test its assumptions by seeing what deductions, or predictions, follow from the model, and then seeing how well these predictions in fact hold in the real world. If they fit well, that's good, though not proof that our model is correct; if these predictions deviate a lot from the real world, then we are alerted that our model may be misleading.

§4.2. Standard errors.

Now that we have discussed how to estimate the parameters β, we can ask how far off $\hat{\beta}$ is likely to be from β, i.e., what are the SEs. We want the SE of each $\hat{\beta}_i$, or, equivalently, $\mathrm{Var}(\hat{\beta}_i)$. The variances are the diagonal elements of the matrix $\mathrm{Cov}(\hat{\beta})$. This is a statistical question, so it depends on statistical assumptions. In fact, it is because of our model that $\hat{\beta}$ is random: $\hat{\beta}$ comes from the data, but our model says that the data come from a random mechanism and that there are true values for the parameters β.

We can assume a little less than that ϵ_i are IID and independent of X. Instead, we assume only that conditioned on X, the random errors ϵ_i have mean 0, a common SD σ, and are uncorrelated. This is weaker than (A1), but stronger than (A2). [Note: "sd", "SD", and "SE" all refer to standard deviations. We are using "sd" for data, "SD" for random variables, and "SE" for estimators, in order to help distinguish them, even though an estimator is a random variable.] Since we are conditioning on X, the answer will be a conditioned covariance matrix.

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

Proof. By (N5), we have $\hat{\beta} - \beta = Q\epsilon$. Since $\beta = E(\hat{\beta} \mid X)$, it follows that

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

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

In the case of FPP, the covariance matrix is just a variance, $X = \mathbf{1}_n$, and we get the variance σ^2/n, as usual.

We are accustomed to estimates having less error with more data; typically the SE decreases at the rate $1/\sqrt{n}$. Is this what theorem 4.3 tells us in general? Yes, it does, depending on what we assume about X as $n \to \infty$. To see this, write

$$\sigma^2 (X'X)^{-1} = \frac{\sigma^2}{n} \left(\frac{X'X}{n} \right)^{-1} .$$

Note that $X'X$ is a $p \times p$ matrix and p is not changing as $n \to \infty$. Furthermore, the (i, j)-entry of $X'X$ is the dot product of the ith column of X with the jth column. So it is reasonable to expect that $X'X/n$ has a limit, say, L, and that L is also invertible; then $\mathrm{Cov}(\hat{\beta} \mid X) \approx (\sigma^2/n) L^{-1}$. In the case of FPP, $X = \mathbf{1}_n$, so $L = 1$.

Optional exercise: If the rows X_i are IID (with finite first moment), then the limit L exists and is invertible. Hint: Use the law of large numbers and that X has full rank; show $L = E(X_i X_i')$.

The off-diagonal elements of the covariance matrix are also useful, though not nearly as often as the diagonal elements. For example, if we want to find the SE for $\hat{\beta}_3 - \hat{\beta}_4$, we will need the (3, 4)-entry: see exercise 4B14. The reason we might want the SE for such a difference is to test whether $\beta_3 = \beta_4$; we will look at tests in Chap. 5 and an example in Chap. 6. The off-diagonal entries are also useful for the multi-dimensional CLT, but that will not concern us in this course.

Now there is a problem with using Theorem 4.3: we usually don't know σ, just like in FPP, where we had $p = 1$ and $X = \mathbf{1}_n$. We solved that problem in FPP by using (N1). We divided by $n - 1$ instead of n because it was a little more accurate. It gave us an unbiased estimate of the variance: $E(\hat{\sigma}^2) = \sigma^2$.

In the general problem, we will again use the residuals e to estimate σ^2. It turns out that the correct replacement for (N1) consists in merely replacing "1" by p:

$$\hat{\sigma} := \sqrt{\frac{1}{n-p} \sum_{i=1}^{n} e_i^2} = \frac{\|e\|}{\sqrt{n-p}} . \tag{N6}$$

If we use $\hat{\sigma}$ in place of σ in the formula of Theorem 4.3 in order to *estimate* the covariance matrix of $\hat{\beta}$, then we obtain

$$\widehat{\mathrm{Cov}}(\hat{\beta} \mid X) := \hat{\sigma}^2 (X'X)^{-1} . \tag{N7}$$

For example, with the data for Hooke's law, we obtain the residuals $e := Y - X\hat{\beta}$ to be (approximately)

$$e = \begin{bmatrix} -.011 \\ .011 \\ .002 \\ .004 \\ -.004 \\ -.002 \end{bmatrix},$$

so that $\hat{\sigma}^2 = \|e\|^2/(6-2) = .00007$ by (N6) and

$$\widehat{\mathrm{Cov}}(\hat{\beta} \mid X) = \begin{bmatrix} 4 \times 10^{-5} & -5 \times 10^{-6} \\ -5 \times 10^{-6} & 1 \times 10^{-6} \end{bmatrix}$$

by (N7). Therefore $\widehat{\mathrm{SE}}(\hat{\beta}_1) = \sqrt{4 \times 10^{-5}} = .006$ and $\widehat{\mathrm{SE}}(\hat{\beta}_2) = \sqrt{1 \times 10^{-6}} = .001$. The model fits the data very well! (This was clear from figure 12.5 of FPP.)

We will now prove that (N7) does give an unbiased estimate, i.e., that $E(\hat{\sigma}^2 \mid X) = \sigma^2$. Another name: $n - p$ is called the number of "degrees of freedom". Don't worry why.

You 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 $HY = X\hat{\beta}$.

In the handout on the t-distribution, we proved that if Z is a random vector with $E(Z) = \mathbf{0}_n$ and $\mathrm{Cov}(Z) = I_n$, 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 V of dimension r, then PZ, in coordinates corresponding to an orthonormal basis for V, 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.)

Recall that our estimate of σ comes from $\hat{\sigma}^2 := \|e\|^2/(n-p)$.

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

Proof. As we just reviewed, we have that $e = P_{W^\perp}(Y)$. Since $X\beta \in W$, it follows that

$$e = P_{W^\perp}(Y) - \mathbf{0}_n = P_{W^\perp}(Y) - P_{W^\perp}(X\beta) = P_{W^\perp}(Y - X\beta) = P_{W^\perp}(\epsilon)\,. \tag{N8}$$

Since $e = P_{W^\perp}(\epsilon)$, we can write e in orthonormal coordinates of $W^\perp$ as an $(n-p)$-dimensional vector, $U = (U_1, U_2, \ldots, U_{n-p})$. What we discussed before this theorem tells us that $E(U \mid X) = \mathbf{0}_{n-p}$ and $\mathrm{Cov}(U \mid X) = \sigma^2 I_{n-p}$. In particular, $E(U_i^2 \mid X) = \sigma^2$ for each i. Since $\|e\| = \|U\|$, it follows that

$$E\big(\|e\|^2 \mid X\big) = E\big(\|U\|^2 \mid X\big) = E\Big(\sum_{i=1}^{n-p} U_i^2 \mid X\Big) = \sum_{i=1}^{n-p} E(U_i^2 \mid X) = (n-p)\sigma^2\,.$$

This gives the result: $E(\hat{\sigma}^2 \mid X) = E\big(\|e\|^2/(n-p) \;\big|\; X\big) = (n-p)\sigma^2/(n-p) = \sigma^2$. ∎

There is another use of the hat notation: We write $\hat{Y} := X\hat{\beta}$. This is not an estimate of Y; after all, we know Y (once we have our data). Instead, $\hat{Y}$ is called the "predicted" or "fitted" value of Y. It is, of course, the same as $P_W(Y)$, the point in W closest to Y. The word "fitted" comes from this. While Y is not really being predicted, $\hat{Y}$ is what the model would predict from X and $\hat{\beta}$ if we were to make such a prediction (but recall that $\hat{\beta}$ comes from X and Y).

§4.3. Explained variance.

We have discussed the hyperplane of best fit for data. How well does it fit? While we could use the r.m.s. error $\sqrt{\|e\|^2/n}$ (or we could divide by $n-p$), is there a scale-free way to measure how well it fits, like the correlation coefficient r for simple linear regression? Yes, there is, and it is called R^2. It is always non-negative, unlike r. For simple linear regression, we will have that $R^2 = r^2$. Note that we are now discussing merely data and the way that we are approximating it; there are no statistical assumptions.

Suppose X has a column of 1s, which we denote as usual by $\mathbf{1}_n$. Since $Y = X\hat{\beta} + e$ is the orthogonal decomposition of Y with respect to $\mathrm{col}(X)$, we have

$$\|Y\|^2 = \|X\hat{\beta}\|^2 + \|e\|^2 \tag{N9}$$

by the Pythagorean theorem. Now the *sample mean* of Y is $\mathbf{1}_n'Y/n$ and the *sample variance* of Y is

$$\mathrm{var}(Y) = \frac{\|Y\|^2}{n} - \left(\frac{\mathbf{1}_n'Y}{n}\right)^2 . \tag{N10}$$

Since $e \perp X$ and $\mathbf{1}_n$ is one of the columns of X, we know that $e \perp \mathbf{1}_n$. Thus, we can express the sample sum as

$$\mathbf{1}_n'Y = \mathbf{1}_n'(X\hat{\beta}) + \mathbf{1}_n'e = \mathbf{1}_n'X\hat{\beta}\,, \tag{N11}$$

so that the sample mean of Y is equal to the sample mean of $X\hat{\beta}$. (We are just reproving and using that the sample mean of e is 0.) Therefore, by (N9), (N10), and (N11), we have

$$\begin{aligned}\mathrm{var}(Y) &= \frac{\|Y\|^2}{n} - \left(\frac{\mathbf{1}_n'Y}{n}\right)^2 = \frac{\|Y\|^2}{n} - \left(\frac{\mathbf{1}_n'X\hat{\beta}}{n}\right)^2 = \frac{\|X\hat{\beta}\|^2 + \|e\|^2}{n} - \left(\frac{\mathbf{1}_n'X\hat{\beta}}{n}\right)^2 \\ &= \frac{\|X\hat{\beta}\|^2}{n} - \left(\frac{\mathbf{1}_n'X\hat{\beta}}{n}\right)^2 + \frac{\|e\|^2}{n} = \mathrm{var}(X\hat{\beta}) + \mathrm{var}(e)\,,\end{aligned}$$

where we have used again the fact that the sample mean of e is 0. This is equation (4.22) in the book. We can also write it by dividing both sides by $\mathrm{var}(Y)$ as

$$1 = \frac{\mathrm{var}(X\hat{\beta})}{\mathrm{var}(Y)} + \frac{\mathrm{var}(e)}{\mathrm{var}(Y)}\,. \tag{N12}$$

This leads to the definition

$$R^2 := \mathrm{var}(X\hat{\beta})/\mathrm{var}(Y)\,.$$

Equation (N12) gives implies that

$$1 - R^2 = \mathrm{var}(e)/\mathrm{var}(Y)\,. \tag{N13}$$

(Note that if Y is a multiple of $\mathbf{1}_n$, then $\mathrm{var}(Y) = 0$ and the above definitions do not make sense. But this is not going to occur in practice.)

Now recall that with simple linear regression, the r.m.s. error is $\sqrt{1-r^2}\cdot \mathrm{sd}(Y)$. The definition of the r.m.s. error, in our new notation, is $\sqrt{\|e\|^2/n} = \mathrm{sd}(e)$. In other words, we proved before for simple linear regression that $\mathrm{sd}(e) = \sqrt{1-r^2}\cdot \mathrm{sd}(Y)$, i.e., $1 - r^2 = \mathrm{var}(e)/\mathrm{var}(Y)$. Thus, comparing to (N13), we conclude that $R^2 = r^2$.

Exercise: Show that $R^2 = 1$ iff all the data lies on a hyperplane, i.e., $Y = X\gamma$ for some γ, which we can also express by saying that $Y \in \mathrm{col}(X)$. Show that $R^2 = 0$ iff all coordinates of $\hat{\beta}$ are 0 except the intercept.

Optional exercise: Show that R, the non-negative square root of R^2, is the sample correlation between Y and $\hat{Y}$, where, as usual, $\hat{Y} = X\hat{\beta}$. (Hint: Write $Y = \hat{Y} + e$.)

The name for R^2 is *explained variance*. The reason for this name is that the sample variance is decomposed in (N12) as part from $X\hat{\beta}$ and part from e. If we are trying to explain Y via X, leaving an unexplained part e, then the portion of variance thus explained is indeed R^2. However, this is just terminology. We might not really be explaining anything. Even in the case of simple linear regression, r doesn't necessarily tell us to what extent the regression explains the relationship. Remember that association is not causation. For example, there is a high correlation between shoe size and reading ability in children (Sec. 9.5 of FPP). Does this mean that shoe size is greatly explained by reading ability? Or that reading ability is greatly explained by shoe size? For another example, consider a regression of rectangle area on perimeter, or a multiple regression of rectangle area on

perimeter and diagonal (Sec. 12.3 of FPP). How much is being explained? How much is being mis-directed? A final example is in SM: the purchasing power of the US dollar and lung cancer rates in the US, both over the period 1950–1999, have a very high correlation ($r = -0.95$). But neither explains the other.

Another strange thing about the terminology "explained variance" is that the units are wrong: variance always has the wrong units. Rather, sd has the right units. However, it is *not true* that $\mathrm{sd}(Y) = \mathrm{sd}(X\hat{\beta}) + \mathrm{sd}(e)$, so it does not make sense to parcel out the sd of Y between the two parts. Does it really make sense to parcel out the variance? This is what the discussion in SM is about, where he discusses the Pythagorean theorem and the "area between San Francisco and Sacramento".

Nevertheless, R^2 does measure how well the regression fits the data.

All the above assumed that the regression had an intercept. If not, one defines $R^2 := \|\hat{Y}\|^2/\|Y\|^2$. However, almost all regressions have an intercept.

§4.5. Discussion questions.

In 3, we are assuming that $E(\epsilon_i \mid X) = 0$ for all i. In (b), we should really ask about "the estimated squares of standard errors". You might want to think about the whole problem in the simplest case, where $X = \mathbf{1}_n$.

In 5, we make the usual assumptions on the model, before we exclude or add a variable. In (b), change "going" to "likely". For more discussion, see the separate handout.

In 8, see the handout on multicollinearity for more details on the comment in the back. Since collinearity leads to large SEs, it can lead to large P-values. This can lead people to accept a null hypothesis (e.g., that a coefficient is 0), but that is not justified. It is a very common error in practice (even without collinearity).

In 10–14, we look at 5(c),(d) in more detail. In 10–11, we have a model with two variables and we omit one. In 12–14, we have a model with one variable and we add one more. Note that in 11 and 14, we are considering estimates of a, even when the model is changed. The model could be correct when changed, even when the original model is also correct. Indeed, as was indicated in an earlier optional exercise, if all the variables are jointly normal, then both models are correct—but they use different parameter values. [Remember that the model says that there exists some β such that $Y - X\beta$ has certain properties. When X is changed, β will change too.] When the changed model is correct, then Tom and Dick will still find unbiased estimates of the new parameters. Thus, the issues concern not the statistical assumptions alone, but some reason that we are especially interested in a. This could be because a has a meaning from other considerations, such as

causality, as explained in chapter 6 (response schedules). This can be related to whether we ought to control for another variable, or whether we ought not to control. For example, perhaps in 11 we should control for W, whereas in 14, we should not.

In 11, 3B15 is our LinAlg#1. This problem can be extended by the same analysis to the following. Consider simple linear regression, as in 4B15, where we have $Y_i = a+bX_i+\epsilon_i$, but now assume that (Y_i, X_i, ϵ_i) are IID and that ϵ has mean 0. If $\mathrm{Cov}(X_i, \epsilon_i) \neq 0$, then we get a biased estimate of a and of b. (See the handout on 4.5.5 for a discussion of what we mean by a and b when the assumptions don't hold.) We will pay attention to b, which is the parameter of most interest in applications. The law of large numbers shows that the bias will be approximately $\mathrm{Cov}(X_i, \epsilon_i)/\mathrm{Var}(X_i)$ for large n. By 4B15, this equals $\sqrt{n}\,\mathrm{Corr}(X_i, \epsilon_i)$. Therefore, when the correlation between X_i and ϵ_i is of the order of $1/\sqrt{n}$, there will be serious trouble in how far off from the truth we think we are, and therefore if a t-test is done on b, as in chapter 5, we may be seriously led astray.

In 13, we should also assume that $P(\delta_i \neq 0) = 1$ so that $P(W = cX) = 0$. For the answer when $d \neq 0$, note that the true coefficient of W is 0 and $W \perp\!\!\!\perp \epsilon$; in other words, the model assumptions hold with the model $Y = aX + 0W + \epsilon$, which is what Dick is basing his regression estimates on.

For the answer to 14, note that $\hat{\beta} \to L^{-1}M$ as $n \to \infty$. The answer for (a) is thus approximately $1/(1+\sigma^2)$ and for (b) is approximately $(1+2\sigma^2)/(1+\sigma^2)$. For more on (c), see the handout "To Be or Not to Be (in the Model)?".

The point of 15 is that sampling does not justify the assumptions of OLS. We are interested in a and b, the true population values. In this context, OLS estimates a and b for the model $Y_j = a+bX_j+\epsilon_j$. We do not have $E(\epsilon_j \mid X) = 0$, nor any other good properties. In fact, if all the data points $x_1, \ldots, x_N$ are distinct, then $E(\epsilon_j \mid X) = E(\epsilon_j \mid X_j) = u_i$ for the i such that $X_j = x_i$. More generally, $E(\epsilon_j \mid X)$ equals the average of the values of u_i for those i where $X_j = x_i$. However, the bias is on the order of $1/\sqrt{n}$ for large n; this can be proved by a similar analysis to that for exercise 14, together with the central limit theorem.

In 17, there is no general way to estimate r. To see this, we give an example. It corresponds to a random systematic bias. Namely, let $Y_1, \ldots, Y_n$ be IID with mean α and let A be independent of all Y_i and have mean 0. For our example, we take $X_i := Y_i + A$. Then $\mathrm{Cov}(X_i, X_j) = E(A^2)$ for $i \neq j$, so $r = E(A^2)/\sigma^2$. In all our data $X_1, \ldots, X_n$, we get only one sample of A (and even that is not observed). Thus, we cannot estimate $E(A^2)$. Of course, if we make an assumption on the nature of the correlations among the X_i, then we may indeed be able to estimate r. However, such assumptions virtually always assume independence somewhere else, so the question is pushed to a less visible place.

Note also that the formula in (b) shows that when r is of the order $1/n$, then the true variance changes substantially. When we do a z-test in such a situation, our P-value will be seriously wrong. Thus, the more data we have, the smaller r can be to create havoc. Keep this in mind for chapter 5, where we will discuss t-tests.