2.3 Adding Probabilistic Assumptions

The usual treatment of linear regression adds many more probabilistic assumptions, namely that

$$Y \mid \vec{X} \sim \mathcal{N}(\vec{X} \cdot \beta, \sigma^2) \tag{2.41}$$

and that Y values are independent conditional on their $\vec{X}$ values. So now we are *assuming* that the regression function is exactly linear; we are *assuming* that at each $\vec{X}$ the scatter of Y around the regression function is Gaussian; we are *assuming* that the variance of this scatter is constant; and we are *assuming* that there is no dependence between this scatter and anything else.

None of these assumptions was needed in deriving the optimal linear predictor. None of them is so mild that it should go without comment or without at least some attempt at testing.

Leaving that aside just for the moment, why make those assumptions? As you know from your earlier classes, they let us write down the likelihood of the observed responses $y_1, y_2, \ldots y_n$ (conditional on the covariates $\vec{x}_1, \ldots \vec{x}_n$), and then estimate β and σ^2 by maximizing this likelihood. As you also know, the maximum likelihood estimate of β is exactly the same as the β obtained by minimizing the residual sum of squares. This coincidence would not hold in other models, with non-Gaussian noise.

We saw earlier that $\widehat{\beta}$ is consistent under comparatively weak assumptions — that it converges to the optimal coefficients. But then there might, possibly, still be other estimators are also consistent, but which converge faster. If we make the extra statistical assumptions, so that $\widehat{\beta}$ is also the maximum likelihood estimate, we can lay that worry to rest. The MLE is generically (and certainly here!) **asymptotically efficient**, meaning that it converges as fast as any other consistent estimator, at least in the long run. So we are not, so to speak, wasting any of our data by using the MLE.

A further advantage of the MLE is that, as $n \to \infty$, its sampling distribution is itself a Gaussian, centered around the true parameter values. This lets us calculate standard errors and confidence intervals quite easily. Here, with the Gaussian assumptions, much more exact statements can be made about the distribution of $\widehat{\beta}$ around β. You can find the formulas in any textbook on regression, so I won't get into that.

We can also use a general property of MLEs for model testing. Suppose we have two classes of models, Ω and ω. Ω is the general case, with p parameters, and ω is a special case, where some of those parameters are constrained, but $q < p$ of them are left free to be estimated from the data. The constrained model class ω is then **nested** within Ω. Say that the MLEs with and without the constraints are, respectively, $\widehat{\Theta}$ and $\widehat{\theta}$, so the maximum log-likelihoods are $L(\widehat{\Theta})$ and $L(\widehat{\theta})$. Because it's a maximum over a larger parameter space, $L(\widehat{\Theta}) \geq L(\widehat{\theta})$. On the other hand, if the true model really is in ω, we'd expect the constrained and unconstrained estimates to be converging. It turns out that the difference in log-likelihoods has an asymptotic distribution which doesn't depend on any of the

model details, namely

$$2\left[L(\widehat{\Theta}) - L(\widehat{\theta})\right] \rightsquigarrow \chi^2_{p-q} \tag{2.42}$$

That is, a χ^2 distribution with one degree of freedom for each extra parameter in Ω (that's why they're called "degrees of freedom").[10]

This approach can be used to test particular restrictions on the model, and so it is sometimes used to assess whether certain variables influence the response. This, however, gets us into the concerns of the next section.

2.3.1 Examine the Residuals

By construction, the errors of the optimal linear predictor have expectation 0 and are uncorrelated with the regressors. Also by construction, the residuals of a *fitted* linear regression have sample mean 0, and are uncorrelated, in the sample, with the regressors.

If the usual probabilistic assumptions hold, however, the errors of the optimal linear predictor have many other properties as well.

1. The errors have a Gaussian distribution at each $\vec{x}$.
2. The errors have the *same* Gaussian distribution at each $\vec{x}$, i.e., they are *independent* of the regressors. In particular, they must have the same variance (i.e., they must be homoskedastic).
3. The errors are *independent of each other*. In particular, they must be *uncorrelated* with each other.

When these properties — Gaussianity, homoskedasticity, lack of correlation — hold, we say that the errors are **white noise**. They imply strongly related properties for the residuals: the residuals should be Gaussian, with variances and covariances given by the hat matrix, or more specifically by $\mathbf{I} - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$ (§1.5.3.2). This means that the residuals will not be *exactly* white noise, but they should be *close* to white noise. You should check this! If you find residuals which are a long way from being white noise, you should be extremely suspicious of your model. These tests are much more important than checking whether the coefficients are significantly different from zero.

Every time someone uses linear regression with the standard assumptions for inference and does *not* test whether the residuals are white noise, an angel loses its wings.

2.3.2 On Significant Coefficients

If all the usual distributional assumptions hold, then t-tests can be used to decide whether particular coefficients are statistically-significantly different from zero.

[10] If you assume the noise is Gaussian, the left-hand side of Eq. 2.42 can be written in terms of various residual sums of squares. However, the equation itself remains valid under other noise distributions, which just change the form of the likelihood function.

Pretty much any piece of statistical software, R very much included, reports the results of these tests automatically. It is far too common to seriously over-interpret those results, for a variety of reasons.

Begin with exactly what hypothesis is being tested when R (or whatever) runs those t-tests. Say, without loss of generality, that there are p predictor variables, $\vec{X} = (X_1, \ldots X_p)$, and that we are testing the coefficient on X_p. Then the null hypothesis is not just "$\beta_p = 0$", but "$\beta_p = 0$ in a linear, Gaussian-noise model which also includes $X_1, \ldots X_{p-1}$, and nothing else". The alternative hypothesis is not just "$\beta_p \neq 0$", but "$\beta_p \neq 0$ in a linear, Gaussian-noise model which also includes $X_1, \ldots X_{p-1}$, but nothing else". The optimal linear coefficient on X_p will depend not just on the relationship between X_p and the response Y, but also on which other variables are included in the model. The test checks whether adding X_p really improves predictions more than would be expected, under all these assumptions, if one is already using all the other variables, and only those other variables. It does not, cannot, test whether X_p is important in any absolute sense.

Even if you are willing to say "Yes, all I really want to know about this variable is whether adding it to the model really helps me predict in a linear approximation", remember that the question which a t-test answers is whether adding that variable will help *at all*. Of course, as you know from your regression class, and as we'll see in more detail in Chapter 3, expanding the model never hurts its performance on the *training* data. The point of the t-test is to gauge whether the improvement in prediction is small enough to be due to chance, or so large, *compared to what noise could produce*, that one could confidently say the variable adds *some* predictive ability. This has several implications which are insufficiently appreciated among users.

In the first place, tests on individual coefficients can seem to contradict tests on groups of coefficients. Adding multiple variables to the model could significantly improve the fit (as checked by, say, a partial F test), even if *none* of the coefficients is significant on its own. In fact, every single coefficient in the model could be insignificant, while the model as a whole is highly significant (i.e., better than a flat line).

In the second place, it's worth thinking about which variables will show up as statistically significant. Remember that the t-statistic is $\widehat{\beta}_i/\mathrm{se}(\widehat{\beta}_i)$, the ratio of the estimated coefficient to its standard error. We saw above that $\mathbb{V}\left[\widehat{\beta} \mid \mathbf{X} = \mathbf{x}\right] = \frac{\sigma^2}{n}(n^{-1}\mathbf{x}^T\mathbf{x})^{-1} \to n^{-1}\sigma^2\mathbf{v}^{-1}$. This means that the standard errors will shrink as the sample size grows, so more and more variables will become significant as we get more data — but how much data we collect is irrelevant to how the process we're studying actually works. Moreover, at a fixed sample size, the coefficients with smaller standard errors will tend to be the ones whose variables have more variance, and whose variables are less correlated with the other predictors. High input variance and low correlation help us *estimate* the coefficient precisely, but, again, they have nothing to do with whether the input variable actually *influences* the response a lot.

To sum up, it is *never* the case that statistical significance is the same as

scientific, real-world significance. The most important variables are *not* those with the largest-magnitude t statistics or smallest p-values. Statistical significance is always about what "signals" can be picked out clearly from background noise[11]. In the case of linear regression coefficients, statistical significance runs together the size of the coefficients, how bad the linear regression model is, the sample size, the variance in the input variable, and the correlation of that variable with all the others.

Of course, even the limited "does it help linear predictions enough to bother with?" utility of the usual t-test (and F-test) calculations goes away if the standard distributional assumptions do not hold, so that the calculated p-values are just wrong. One can sometimes get away with using bootstrapping (Chapter 6) to get accurate p-values for standard tests under non-standard conditions.

2.4 Linear Regression Is Not the Philosopher's Stone

The philosopher's stone, remember, was supposed to be able to transmute base metals (e.g., lead) into the perfect metal, gold (Eliade, 1971). Many people treat linear regression as though it had a similar ability to transmute a correlation matrix into a scientific theory. In particular, people often argue that:

1. because a variable has a significant regression coefficient, it must influence the response;
2. because a variable has an insignificant regression coefficient, it must not influence the response;
3. if the input variables change, we can predict how much the response will change by plugging in to the regression.

All of this is wrong, or at best right only under very particular circumstances.

We have already seen examples where influential variables have regression coefficients of zero. We have also seen examples of situations where a variable with no influence has a non-zero coefficient (e.g., because it is correlated with an omitted variable which does have influence). *If* there are no nonlinearities and *if* there are no omitted influential variables and *if* the noise terms are always independent of the predictor variables, are we good?

No. Remember from Equation 2.6 that the optimal regression coefficients depend on both the marginal distribution of the predictors and the joint distribution (covariances) of the response and the predictors. There is no reason whatsoever to suppose that if we *change* the system, this will leave the conditional distribution of the response alone.

A simple example may drive the point home. Suppose we surveyed all the cars in Pittsburgh, recording the maximum speed they reach over a week, and how often they are waxed and polished. I don't think anyone doubts that there will be a positive correlation here, and in fact that there will be a positive regression

[11] In retrospect, it might have been clearer to say "statistically *detectable*" rather than "statistically *significant*".

coefficient, even if we add in many other variables as predictors. Let us even postulate that the relationship is linear (perhaps after a suitable transformation). Would anyone believe that polishing cars will make them go faster? Manifestly not. But this is exactly how people interpret regressions in all kinds of applied fields — instead of saying polishing makes cars go faster, it might be saying that receiving targeted ads makes customers buy more, or that consuming dairy foods makes diabetes progress faster, or .... Those claims might be *true*, but the regressions could easily come out the same way were the claims false. Hence, the regression results provide little or no *evidence* for the claims.

Similar remarks apply to the idea of using regression to "control for" extra variables. If we are interested in the relationship between one predictor, or a few predictors, and the response, it is common to add a bunch of other variables to the regression, to check both whether the apparent relationship might be due to correlations with something else, and to "control for" those other variables. The regression coefficient is interpreted as how much the response would change, on average, if the predictor variable were increased by one unit, "holding everything else constant". There is a very particular sense in which this is true: it's a prediction about the difference in expected responses (conditional on the given values for the other predictors), assuming that the form of the regression model is right, *and* that observations are randomly drawn from the same population we used to fit the regression.

In a word, what regression does is *probabilistic* prediction. It says what will happen if we keep drawing from the same population, but *select* a sub-set of the observations, namely those with given values of the regressors. A **causal** or **counter-factual** prediction would say what would happen if we (or Someone) *made* those variables take those values. Sometimes there's no difference between selection and intervention, in which case regression works as a tool for causal inference[12]; but in general there is. Probabilistic prediction is a worthwhile endeavor, but it's important to be clear that this is what regression does. There are techniques for doing causal prediction, which we will explore in Part III.

Every time someone thoughtlessly uses regression for causal inference, an angel not only loses its wings, but is cast out of Heaven and falls in extremest agony into the everlasting fire.

[12] In particular, if our model was estimated from data where Someone *assigned* values of the predictor variables in a way which breaks possible dependencies with omitted variables and noise — either by randomization or by experimental control — then regression can, in fact, work for causal inference.

2.5 Further Reading

If you would like to read a lot more — about 400 pages more — about linear regression from this perspective, see *The Truth About Linear Regression*, at `http://www.stat.cmu.edu/~cshalizi/TALR/`. That manuscript began as class notes for the class *before* this one, and has some overlap.

There are many excellent textbooks on linear regression. Among them, I would mention Weisberg (1985) for general statistical good sense, along with Faraway (2004) for R practicalities, and Hastie *et al.* (2009) for emphasizing connections to more advanced methods. Berk (2004) omits the details those books cover, but is superb on the big picture, and especially on what must be assumed in order to do certain things with linear regression and what cannot be done under any assumption.

For some of the story of how the usual probabilistic assumptions came to have that status, see, e.g., Lehmann (2008). On the severe issues which arise for the usual inferential formulas when the model is incorrect, see Buja *et al.* (2014).

Linear regression is a special case of both additive models (Chapter 8), and of locally linear models (§10.5). In most practical situations, additive models are a better idea than linear ones.

Historical notes

Because linear regression is such a big part of statistical practice, its history has been extensively treated in general histories of statistics, such as Stigler (1986) and Porter (1986). Farebrother (1999) is especially clear on transition from the first appearance of the method of least squares, where it was used to find parameters when there were more equations than unknowns[13], to more general linear modeling. I would particularly recommend Klein (1997) for a careful account of how regression, on its face a method for doing comparisons at one time across a population, came to be used to study causality and dynamics. The paper by Lehmann (2008) mentioned earlier is also informative.

The derivation of the optimal linear predictor in §2.1, assuming nothing beyond wanting to use a linear prediction function and $\mathbf{v}$ being invertible, is standard in

[13] The classic cases where astronomy and "geodesy", the measurement of the exact shape of the Earth (important for physics and for navigation). Take astronomy: if you have a model of the orbit of a planet, and plug in values for the parameters, you get a prediction for the position of the planet in the sky every night. Going the other direction, every observation gives you an equation with the unknown parameters on one side, and known, measured values of the planetary position on the other side. Even with a very complicated model with dozens of adjustable parameters, a few years worth of nightly observations gives you more equations than unknowns. With more equations than unknowns, there's usually no solution that fits all the data exactly. The literally-ancient approach to this embarrassing problem, going back to the ancient Greeks and to the Babylonians before them, was to try to select the *best*, most reliable observations, discarding the bad ones, until you had just as many observations as unknowns, and then solving for the parameters. The crucial innovation in the 1700s was to realize that least squares gave us a way of trying to use *all* the observations, giving parameter values that *generally* fit well but not perfectly, because even the best observations are imperfect. In this context, the emphasis on *linear* equations made sense, because of the form of the models the astronomers and geodesists were using.

the theory of time series (Ch, 23) and stochastic processes, going back there at least to Kolmogorov (1941) and Wiener (1949). Special cases were known in the 1930s in factor analysis (Ch. 16), though I believe all of them also, unnecessarily, assumed Gaussian distributions for all variables. It's possible someone else got there first, but if so, I haven't been able to find it. In spatial statistics, the same ideas were re-discovered by D. G. Krige in the 1950s (Krige, 1981), and popularized by Georges Matheron under the name "kriging" (Matheron, 2019), which has stuck in geostatistics.

Exercises

2.1
1. Write the expected squared error of a linear predictor with slopes $\vec{b}$ and intercept b_0 as a function of those coefficients.
2. Find the derivatives of the expected squared error with respect to all the coefficients.
3. Show that when we set all the derivatives to zero, the solutions are Eq. 2.6 and 2.5.

2.2 Show that the expected error of the optimal linear predictor, $\mathbb{E}\left[Y - \vec{X} \cdot \beta\right]$, is zero.

2.3 Convince yourself that if the real regression function is linear, β does not depend on the marginal distribution of X. You may want to start with the case of one predictor variable.

2.4 Run the code from Figure 2.5. Then replicate the plots in Figure 2.6.

2.5 Which kind of transformation is superior for the model where $Y \mid X \sim \mathcal{N}(\sqrt{X}, 1)$?