Lecture 4: Simple Linear Regression Models, with Hints at Their Estimation

36-401, Fall 2017, Section B

1 The Simple Linear Regression Model

Let's recall the simple linear regression model from last time. This is a statistical model with two variables X and Y, where we try to predict Y from X. The assumptions of the model are as follows:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ.
3. $\mathbb{E}[\epsilon|X = x] = 0$ (no matter what x is), $\mathrm{Var}[\epsilon|X = x] = \sigma^2$ (no matter what x is).
4. ϵ is uncorrelated across observations.

To elaborate, with multiple data points, $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$, then the model says that, for *each* $i \in 1 : n$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

where the noise variables ϵ_i all have the same expectation (0) and the same variance (σ^2), and $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$ (unless $i = j$, of course).

1.1 "Plug-In" Estimates

In lecture 1, we saw that the optimal linear predictor of Y from X has slope $\beta_1 = \mathrm{Cov}[X, Y]/\mathrm{Var}[X]$, and intercept $\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$. A common tactic in devising estimators is to use what's sometimes called the "plug-in principle", where we find equations for the parameters which would hold if we knew the full distribution, and "plug in" the sample versions of the population quantities. We saw this in the last lecture, where we estimated β_1 by the ratio of the sample covariance to the sample variance:

$$\widehat{\beta_1} = \frac{c_{XY}}{s_X^2} \tag{2}$$

We also saw, in the notes to the last lecture, that so long as the law of large numbers holds,

$$\widehat{\beta_1} \to \beta_1 \tag{3}$$

as $n \to \infty$. It follows easily that

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X} \tag{4}$$

will also converge on β_0.

1.2 Least Squares Estimates

An alternative way of estimating the simple linear regression model starts from the objective we are trying to reach, rather than from the formula for the slope. Recall, from lecture 1, that the true optimal slope and intercept are the ones which minimize the mean squared error:

$$(\beta_0, \beta_1) = \underset{(b_0,b_1)}{\operatorname{argmin}} \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right] \tag{5}$$

This is a function of the complete distribution, so we can't get it from data, but we can approximate it with data. The **in-sample**, **empirical** or **training** MSE is

$$\widehat{MSE}(b_0, b_1) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \tag{6}$$

Notice that this is a function of b_0 and b_1; it is also, of course, a function of the data, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, but we will generally suppress that in our notation.

If our samples are all independent, for any fixed (b_0, b_1), the law of large numbers tells us that $\widehat{MSE}(b_0, b_1) \to MSE(b_0, b_1)$ as $n \to \infty$. So it doesn't seem unreasonable to try minimizing the in-sample error, which we can compute, as a proxy for minimizing the true MSE, which we can't. Where does it lead us?

Start by taking the derivatives with respect to the slope and the intercept:

$$\frac{\partial \widehat{MSE}}{\partial b_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(-2) \tag{7}$$

$$\frac{\partial \widehat{MSE}}{\partial b_0} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(-2x_i) \tag{8}$$

Set these to zero at the optimum $(\hat{\beta}_0, \hat{\beta}_1)$:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(x_i) &= 0 \end{aligned} \tag{9}$$

These are often called the **normal equations** for least-squares estimation, or the **estimating equations**: a system of two equations in two unknowns, whose solution gives the estimate. Many people would, at this point, remove the factor of $1/n$, but I think it makes it easier to understand the next steps:

$$\overline{y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{x} = 0 \tag{10}$$

$$\overline{xy} - \hat{\beta}_0 \overline{x} - \hat{\beta}_1 \overline{x^2} = 0 \tag{11}$$

The first equation, re-written, gives

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{12}$$

Substituting this into the remaining equation,

$$0 = \overline{xy} - \bar{y}\bar{x} + \hat{\beta}_1 \bar{x}\bar{x} - \hat{\beta}_1 \overline{x^2} \tag{13}$$

$$0 = c_{XY} - \hat{\beta}_1 s_X^2 \tag{14}$$

$$\hat{\beta}_1 = \frac{c_{XY}}{s_X^2} \tag{15}$$

That is, the least-squares estimate of the slope is our old friend the plug-in estimate of the slope, and thus the least-squares intercept is also the plug-in intercept.

Going forward The equivalence between the plug-in estimator and the least-squares estimator is a bit of a special case for linear models. In some non-linear models, least squares is quite feasible (though the optimum can only be found numerically, not in closed form); in others, plug-in estimates are more useful than optimization.

1.3 Bias, Variance and Standard Error of Parameter Estimates

Whether we think of it as deriving from pluging-in or from least squares, we work out some of the properties of this estimator of the coefficients, using the model assumptions. We'll start with the slope, $\hat{\beta}_1$.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{c_{XY}}{s_X^2} && (16)\\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{s_X^2} && (17)\\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i + \epsilon_1) - \bar{x}(\beta_0 + \beta_1\bar{x} + \bar{\epsilon})}{s_X^2} && (18)\\
&= \frac{\beta_0\bar{x} + \beta_1\overline{x^2} + \frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i - \bar{x}\beta_0 - \beta_1\bar{x}^2 - \bar{x}\bar{\epsilon}}{s_x^2} && (19)\\
&= \frac{\beta_1 s_X^2 + \frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_X^2} && (20)\\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i\epsilon_i - \bar{x}\bar{\epsilon}}{s_X^2} && (21)
\end{aligned}$$

Since $\bar{x}\bar{\epsilon} = n^{-1}\sum_i \bar{x}\epsilon_i$,

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{s_X^2} \tag{22}$$

This representation of the slope estimate shows that it is equal to the true slope (β_1) plus something which depends on the noise terms (the ϵ_i, and their sample average $\bar{\epsilon}$). We'll use this to find the expected value and the variance of the estimator $\hat{\beta}_1$.

In the next couple of paragraphs, I am going to treat the x_i as non-random variables. This is appropriate in "designed" or "controlled" experiments, where we get to chose their value. In randomized experiments or in observational studies, obviously the x_i aren't necessarily fixed; however, these expressions will be correct for the conditional expectation $\mathbb{E}\left[\hat{\beta}_1 | x_1, \ldots x_n\right]$ and conditional variance $\mathrm{Var}\left[\hat{\beta}_1 | x_1, \ldots x_n\right]$, and I will come back to how we get the unconditional expectation and variance.

Expected value and bias Recall that $\mathbb{E}\left[\epsilon_i | X_i\right] = 0$, so

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}\left[\epsilon_i\right] = 0 \tag{23}$$

Thus,

$$\mathbb{E}\left[\hat{\beta}_1\right] = \beta_1 \tag{24}$$

Since the **bias** of an estimator is the difference between its expected value and the truth, $\hat{\beta}_1$ is an **unbiased** estimator of the optimal slope.

(To repeat what I'm sure you remember from mathematical statistics: "bias" here is a technical term, meaning no more and no less than $\mathbb{E}\left[\hat{\beta}_1\right] - \beta_1$. An unbiased estimator could still make systematic mistakes — for instance, it could underestimate 99% of the time, provided that the 1% of the time it over-estimates, it does so by much more than it under-estimates. Moreover, unbiased estimators are not necessarily superior to biased ones: the total error depends on both the bias of the estimator and its variance, and there are many situations where you can remove lots of bias at the cost of adding a little variance. Least squares for simple linear regression happens not to be one of them, but you shouldn't expect that as a general rule.)

Turning to the intercept,

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \mathbb{E}\left[\overline{Y} - \hat{\beta}_1 \overline{X}\right] && (25)\\
&= \beta_0 + \beta_1 \overline{X} - \mathbb{E}\left[\hat{\beta}_1\right] \overline{X} && (26)\\
&= \beta_0 + \beta_1 \overline{X} - \beta_1 \overline{X} && (27)\\
&= \beta_0 && (28)
\end{aligned}$$

so it, too, is unbiased.

Variance and Standard Error Using the formula for the variance of a sum from lecture 1, and the model assumption that all the ϵ_i are uncorrelated with each other,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\beta}_1\right] &= \mathrm{Var}\left[\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{s_X^2}\right] && (29)\\
&= \mathrm{Var}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{s_X^2}\right] && (30)\\
&= \frac{\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \mathrm{Var}\left[\epsilon_i\right]}{(s_X^2)^2} && (31)\\
&= \frac{\frac{\sigma^2}{n} s_X^2}{(s_X^2)^2} && (32)\\
&= \frac{\sigma^2}{n s_X^2} && (33)
\end{aligned}$$

In words, this says that the variance of the slope estimate goes up as the noise around the regression line (σ^2) gets bigger, and goes down as we have more observations (n), which are further spread out along the horizontal axis (s_X^2); it should not be surprising that it's easier to work out the slope of a line from many, well-separated points on the line than from a few points smushed together.

The **standard error** of an estimator is just its standard deviation, or the square root of its variance:

$$\mathrm{se}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{n s_X^2}} \quad (34)$$

I will leave working out the variance of $\hat{\beta}_0$ as an exercise.

Unconditional-on-X Properties The last few paragraphs, as I said, have looked at the expectation and variance of $\hat{\beta}_1$ conditional on $x_1, \ldots x_n$, either because the x's really are non-random (e.g., controlled by us), or because we're just interested in conditional inference. If we do care about unconditional properties, then we still need to find $\mathbb{E}\left[\hat{\beta}_1\right]$ and $\mathrm{Var}\left[\hat{\beta}_1\right]$, not just $\mathbb{E}\left[\hat{\beta}_1 | x_1, \ldots x_n\right]$ and $\mathrm{Var}\left[\hat{\beta}_1 | x_1, \ldots x_n\right]$. Fortunately, this is easy, *so long as the simple linear regression model holds.*

To get the unconditional expectation, we use the "law of total expectation":

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}_1\right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_1 | X_1, \ldots X_n\right]\right] & (35) \\ &= \mathbb{E}\left[\beta_1\right] = \beta_1 & (36) \end{aligned}$$

That is, the estimator is *unconditionally* unbiased.

To get the unconditional variance, we use the "law of total variance":

$$\begin{aligned} \mathrm{Var}\left[\hat{\beta}_1\right] &= \mathbb{E}\left[\mathrm{Var}\left[\hat{\beta}_1 | X_1, \ldots X_n\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[\hat{\beta}_1 | X_1, \ldots X_n\right]\right] & (37) \\ &= \mathbb{E}\left[\frac{\sigma^2}{n s_X^2}\right] + \mathrm{Var}\left[\beta_1\right] & (38) \\ &= \frac{\sigma^2}{n}\mathbb{E}\left[\frac{1}{s_X^2}\right] & (39) \end{aligned}$$

1.4 Parameter Interpretation; Causality

Two of the parameters are easy to interpret.

σ^2 is the variance of the noise around the regression line; σ is a typical distance of a point from the line. ("Typical" here in a special sense, it's the root-mean-squared distance, rather than, say, the average absolute distance.)

β_0 is the simply the expected value of Y when X is 0, $\mathbb{E}\left[Y | X = 0\right]$. The point $X = 0$ usually has no special significance, but this setting does ensure that the line goes through the point $(\mathbb{E}\left[X\right], \mathbb{E}\left[Y\right])$.

The interpretation of the slope is both very straightforward and very tricky. Mathematically, it's easy to convince yourself that, for any x

$$\beta_1 = \mathbb{E}\left[Y | X = x\right] - \mathbb{E}\left[Y | X = x - 1\right] \quad (40)$$

or, for any x_1, x_2,

$$\beta_1 = \frac{\mathbb{E}\left[Y | X = x_2\right] - \mathbb{E}\left[Y | X = x_1\right]}{x_2 - x_1} \quad (41)$$

This is just saying that the slope of a line is "rise/run".

The tricky part is that we have a *very* strong, natural tendency to interpret this as telling us something about causation — "If we change X by 1, then on average Y will change by β_1". This interpretation is usually completely unsupported by the analysis. If I use an old-fashioned mercury thermometer, the height of mercury in the tube usually has a nice linear relationship with the temperature of the room the thermometer is in. This linear relationship goes both ways, so we could regress temperature (Y) on mercury height (X). But if I manipulate the height of the mercury (say, by changing the ambient pressure, or shining a laser into the tube, etc.), changing the height X will not, in fact, change the temperature outside.

The right way to interpret β_1 is not as the result of a *change*, but as an expected *difference*. The correct catch-phrase would be something like "If we select two sets of cases from the un-manipulated distribution where X differs by 1, we expect Y to differ by β_1." This covers the thermometer example, and every other I can think of. It is, I admit, much more inelegant than "If X changes by 1, Y changes by β_1 on average", but it has the advantage of being true, which the other does not.

There *are* circumstances where regression can be a useful part of causal inference, but we will need a lot more tools to grasp them; that will come towards the end of 402.

2 The Gaussian-Noise Simple Linear Regression Model

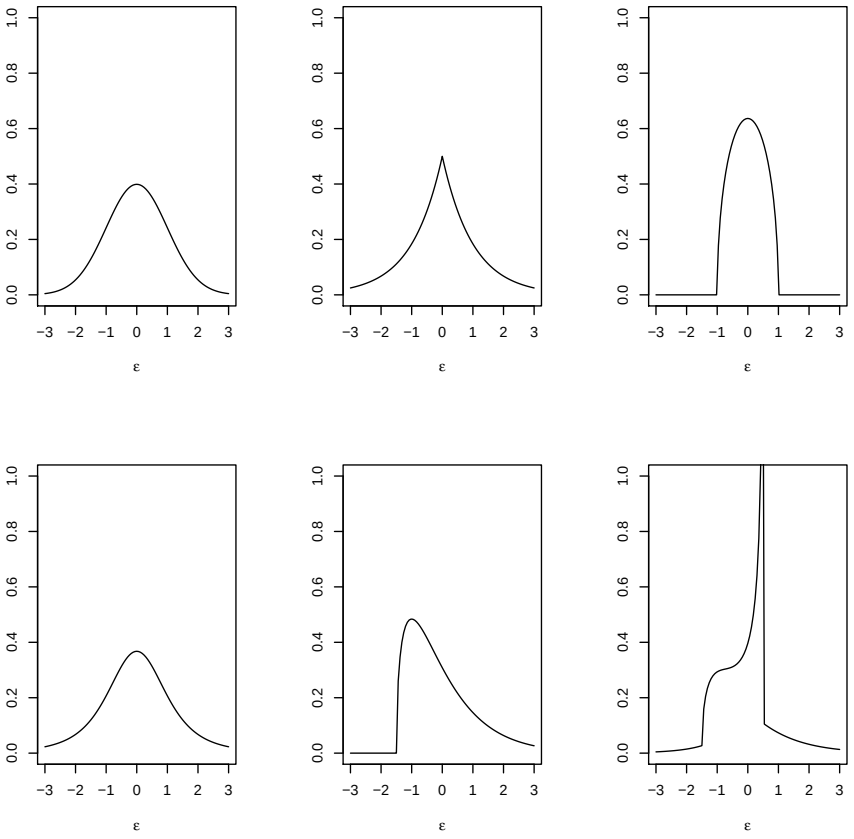
We have, so far, assumed comparatively little about the noise term ϵ. The advantage of this is that our conclusions apply to lots of different situations; the drawback is that there's really not all that much more to say about our estimator $\widehat{\beta}$ or our predictions than we've already gone over. If we made more detailed assumptions about ϵ, we could make more precise inferences.

There are lots of forms of distributions for ϵ which we might contemplate, and which are compatible with the assumptions of the simple linear regression model (Figure 1). The one which has become the most common over the last two centuries is to assume ϵ follows a Gaussian distribution.

The result is the Gaussian-noise simple linear regression model[1]:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ.
3. $\epsilon \sim N(0, \sigma^2)$, independent of X.

[1] Our textbook, rather old-fashionedly, calls this the "normal error" model rather than "Gaussian noise". I dislike this: "normal" is an over-loaded word in math, while "Gaussian" is (comparatively) specific; "error" made sense in Gauss's original context of modeling, specifically, errors of observation, but is misleading generally; and calling Gaussian distributions "normal" suggests they are much more common than they really are.




```
par(mfrow=c(2,3))
curve(dnorm(x), from=-3, to=3, xlab=expression(epsilon), ylab="", ylim=c(0,1))
curve(exp(-abs(x))/2, from=-3, to=3, xlab=expression(epsilon), ylab="",
      ylim=c(0,1))
curve(sqrt(pmax(0,1-x^2))/(pi/2), from=-3, to=3, xlab=expression(epsilon),
      ylab="", ylim=c(0,1))
curve(dt(x,3), from=-3, to=3, xlab=expression(epsilon), ylab="", ylim=c(0,1))
curve(dgamma(x+1.5, shape=1.5, scale=1), from=-3, to=3,
      xlab=expression(epsilon), ylab="", ylim=c(0,1))
curve(0.5*dgamma(x+1.5, shape=1.5, scale=1) +
      0.5*dgamma(0.5-x, shape=0.5, scale=1), from=-3,
      to=3, xlab=expression(epsilon), ylab="", ylim=c(0,1))
par(mfrow=c(1,1))
```

FIGURE 1: *Some possible noise distributions for the simple linear regression model, since all have* $\mathbb{E}[\epsilon] = 0$*, and could get any variance by scaling. (The model is even compatible with each observation taking* ϵ *from a different distribution.) From top left to bottom right: Gaussian; double-exponential ("Laplacian"); "circular" distribution; t with 3 degrees of freedom; a gamma distribution (shape 1.5, scale 1) shifted to have mean 0; mixture of two gammas with shape 1.5 and shape 0.5, each off-set to have expectation 0. The first three were all used as error models in the 18th and 19th centuries. (See the source file for how to get the code below the figure.)*

4. ϵ is independent across observations.

You will notice that these assumptions are strictly stronger than those of the simple linear regression model. More exactly, the first two assumptions are the same, while the third and fourth assumptions of the Gaussian-noise model imply the corresponding assumptions of the other model. This means that everything we have done so far directly applies to the Gaussian-noise model. On the other hand, the stronger assumptions let us say more. They tell us, exactly, the probability distribution for Y given X, and so will let us get exact distributions for predictions and for other inferential statistics.

Why the Gaussian noise model? Why should we think that the noise around the regression line would follow a Gaussian distribution, independent of X? There are two big reasons.

1. *Central limit theorem* The noise might be due to adding up the effects of lots of little random causes, all nearly independent of each other and of X, where each of the effects are of roughly similar magnitude. Then the central limit theorem will take over, and the distribution of the sum of effects will indeed be pretty Gaussian. For Gauss's original context, X was (simplifying) "Where is such-and-such-a-planet in space?", Y was "Where does an astronomer record the planet as appearing in the sky?", and noise came from defects in the telescope, eye-twitches, atmospheric distortions, etc., etc., so this was pretty reasonable. It is clearly *not* a universal truth of nature, however, or even something we should expect to hold true as a general rule, as the name "normal" suggests.

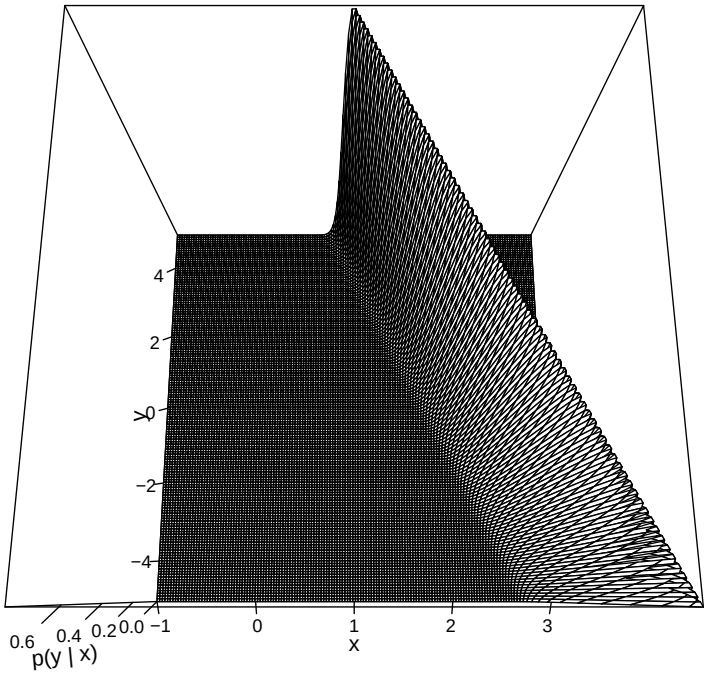
2. *Mathematical convenience* Assuming Gaussian noise lets us work out a very complete theory of inference and prediction for the model, with lots of closed-form answers to questions like "What is the optimal estimate of the variance?" or "What is the probability that we'd see a fit this good from a line with a non-zero intercept if the true line goes through the origin?", etc., etc. Answering such questions without the Gaussian-noise assumption needs somewhat more advanced techniques, and much more advanced computing; we'll get to it towards the end of the class.

2.1 Visualizing the Gaussian Noise Model

The Gaussian noise model gives us not just an expected value for Y at each x, but a whole conditional distribution for Y at each x. To visualize it, then, it's not enough to just sketch a curve; we need a three-dimensional surface, showing, for each combination of x and y, the probability density of Y around that y given that x. Figure 2 illustrates.

2.2 Maximum Likelihood vs. Least Squares

As you remember from your mathematical statistics class, the **likelihood** of a parameter value on a data set is the probability density at the data under




```
x.seq <- seq(from=-1, to=3, length.out=150)
y.seq <- seq(from=-5, to=5, length.out=150)
cond.pdf <- function(x,y) { dnorm(y, mean=10-5*x, sd=0.5) }
z <- outer(x.seq, y.seq, cond.pdf)
persp(x.seq,y.seq,z, ticktype="detailed", phi=75, xlab="x",
      ylab="y", zlab=expression(p(y|x)), cex.axis=0.8)
```

FIGURE 2: *Illustrating how the conditional pdf of Y varies as a function of X, for a hypothetical Gaussian noise simple linear regression where $\beta_0 = 10$, $\beta_1 = -5$, and $\sigma^2 = (0.5)^2$. The perspective is adjusted so that we are looking nearly straight down from above on the surface. (Can you find a better viewing angle?) See* `help(persp)` *for the 3D plotting (especially the examples), and* `help(outer)` *for the* `outer` *function, which takes all combinations of elements from two vectors and pushes them through a function. How would you modify this so that the regression line went through the origin with a slope of $4/3$ and a standard deviation of 5?*

those parameters. We could not work with the likelihood with the simple linear regression model, because it didn't specify enough about the distribution to let us calculate a density. With the Gaussian-noise model, however, we can write down a likelihood[2] By the model's assumptions, if think the parameters are the parameters are b_0, b_1, s^2 (reserving the Greek letters for their true values), then $Y|X = x \sim N(b_0 + b_1 x, s^2)$, and Y_i and Y_j are independent given X_i and X_j, so the over-all likelihood is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i - (b_0 + b_1 x_i))^2}{2s^2}} \tag{42}$$

As usual, we work with the log-likelihood, which gives us the same information[3] but replaces products with sums:

$$L(b_0, b_1, s^2) = -\frac{n}{2}\log 2\pi - \frac{n}{\log} s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \tag{43}$$

We recall from mathematical statistics that when we've got a likelihood function, we generally want to maximize it. That is, we want to find the parameter values which make the data we observed as likely, as probable, as the model will allow. (This may not be very likely; that's another issue.) We recall from calculus that one way to maximize is to take derivatives and set them to zero.

$$\frac{\partial L}{\partial b_0} = -\frac{1}{2s^2}\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i))(-1) \tag{44}$$

$$\frac{\partial L}{\partial b_1} = -\frac{1}{2s^2}\sum_{i=1}^{n} 2(y_i - (b_0 + b_1 x_i))(-x_i) \tag{45}$$

Notice that when we set these derivatives to zero, all the multiplicative constants — in particular, the prefactor of $\frac{1}{2s^2}$ — go away. We are left with

$$\sum_{i=1}^{n} y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i) = 0 \tag{46}$$

$$\sum_{i=1}^{n} (y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))x_i = 0 \tag{47}$$

These are, up to a factor of $1/n$, *exactly* the equations we got from the method of least squares (Eq. 9). That means that the least squares solution *is* the maximum likelihood estimate under the Gaussian noise model; this is no coincidence[4].

[2]Strictly speaking, this is a "conditional" (on X) likelihood; but only pedants use the adjective in this context.

[3]Why is this?

[4]It's no coincidence because, to put it somewhat anachronistically, what Gauss did was ask himself "for what distribution of the noise would least squares maximize the likelihood?".

Now let's take the derivative with respect to s:

$$\frac{\partial L}{\partial s} = -\frac{n}{s} + 2\frac{1}{2s^3}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \tag{48}$$

Setting this to 0 at the optimum, including multiplying through by $\widehat{\sigma}^3$, we get

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i))^2 \tag{49}$$

Notice that the right-hand side is just the in-sample mean squared error.

Other models Maximum likelihood estimates of the regression curve coincide with least-squares estimates when the noise around the curve is additive, Gaussian, of constant variance, and both independent of X and of other noise terms. These were all assumptions we used in setting up a log-likelihood which was, up to constants, proportional to the (negative) mean-squared error. If any of those assumptions fail, maximum likelihood and least squares estimates can diverge, though sometimes the MLE solves a "generalized" least squares problem (as we'll see later in this course).

Exercises

To think through, not to hand it.

1. Show that if $\mathbb{E}[\epsilon|X = x] = 0$ for all x, then $\mathrm{Cov}[X, \epsilon] = 0$. Would this still be true if $\mathbb{E}[\epsilon|X = x] = a$ for some other constant a?
2. Find the variance of $\hat{\beta}_0$. *Hint:* Do you need to worry about covariance between $\overline{Y}$ and $\hat{\beta}_1$?