Lecture 7

One scalar predictor $x \in \mathbb{R}$

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

Here we consider classical linear regression where there is one real valued predictor and one real valued response whose mean we think follows a linear model (with intercept) in the first variable. Our response remains y. We need another letter for the predictor. We used w for binary variables as is common in causal inference. We used i for a more general categorical variable in ANOVA. We could use t but that has connotations of time where often $t \geqslant 0$ is assumed. Let's use x, making $\boldsymbol{x}_i = (1, x_i)^\mathsf{T}$.

7.1 Regression

The term regression originates from Galton's analysis comparing the heights of children to heights of parents. Figure 7.1 is a scatter plot of some similar data from Karl Pearson.

Figure 7.1 includes a linear regression line that we will study later. These data are reasonably close to a bivariate Gaussian. We know that for bivariate Gaussian $(x, y)^\mathsf{T}$,

$$\mathbb{E}(y \,|\, x) = \mu_y + \frac{\rho\sigma_y}{\sigma_x}(x - \mu_x).$$

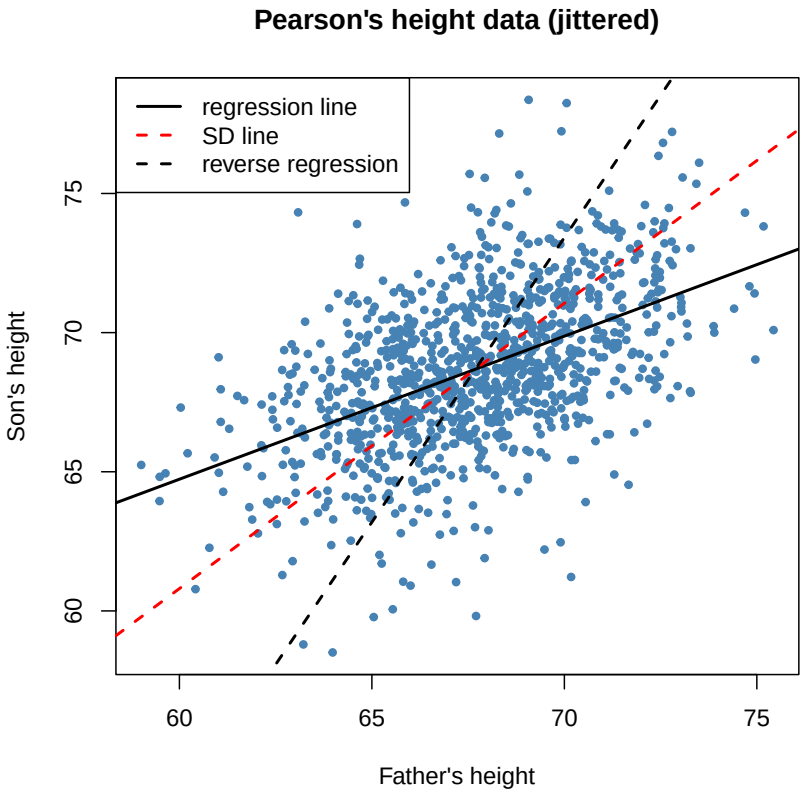


Figure 7.1: Pearson's data on heights of sons and fathers from `father.son` in package `UsingR`. Given to 10^{-5} inches by jittering.

The figure also shows the SD (standard deviation) line

$$\mu_y + \frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

It goes through the bivariate mean with a slope equal to the standard deviation of sons' height divided by the standard deviation of fathers' height. When $\sigma_x = \sigma_y = 1$ this is a '45 degree line'.

Taller than average fathers tended to have taller than average sons who were nonetheless not as many standard deviations taller than average as the fathers were. Similarly shorter than average fathers had shorter than average sons but the sons tended to be closer to average height than their fathers were. When the fathers are $k\sigma_x$ standard deviations from the mean,

the sons are on average $\rho k \sigma_y$ standard deviations from their mean. In this data $\rho \approx 0.5$.

Galton had done this earlier plotting child height versus mid-parent height (the average of both parents). He also used a multiplicative adjustment factor of 1.08 to make mother/daughter heights comparable to father/son heights, so the Pearson data is easier for us to work with.

Galton described the phenomenon above as 'regression to mediocrity' and that is where the term regression came from. Should we expect to eventually see a generation where all the men have the same height? Not really. That would happen if the sons' heights hit the regression line exactly, but they don't. They are spread around it. The variance of a son's height is

$$\mathbb{E}(\mathrm{var}(y \,|\, x)) + \mathrm{var}(\mathbb{E}(y \,|\, x)).$$

In a bivariate Gaussian the variance of the regression line is

$$\mathrm{var}(\mathbb{E}(y \,|\, x)) = \mathrm{var}(\rho\sigma_y(x - \mu_x)/\sigma_x) = \rho^2\sigma_y^2 < \sigma_y^2$$

but that is complemented by $\mathbb{E}(\mathrm{var}(y \,|\, x)) = \sigma_y^2(1 - \rho^2)$. Also regression to the mean applies equally to $\mathbb{E}(x \,|\, y)$. If $|\rho| < 1$ meant that all men would reach the same height, running the argument backwards in time would mean that height variation was negligible many generations ago.

We should definitely not expect data to fall along the SD line in general. For an extreme case consider $\rho = 0$ in $\mathcal{N}(\mathbf{0}, I_2)$. Then x and y are independent. We should not expect y to be particularly close to x.

7.2 Regression fallacy

Suppose we have data well described by

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \sigma^2 \right). \tag{7.1}$$

We select out some data points with x much smaller than μ say $\mu - 2.05\sigma \leqslant x \leqslant \mu - 1.95\sigma$ and then see that the corresponding y values are on average about $\mu - 2\rho\sigma$ which is then higher than $x \approx \mu - 2\sigma$. If $|\rho|$ is small the average of those y values may even be close to μ. Something similar happens when we select some data points with unusually large x. We get $\mathbb{E}(y - x \,|\, x \text{ was large}) < 0$.

This can lead to some common misinterpretations. The rookie athlete who outperformed may do far worse in their sophomore year. The mutual

fund that got the best return this year may be quite poor the next year. Those seem like a mean reversal but we could also get them just sampling from equation (7.1).

Suppose that x is an earlier measure and y is a later measure. For example we might have changed the way we teach children how to divide fractions after the x measure and before the y measure. If that intervention does no good at all, we still see the worst performers improve on average (while the best performers got worse on average). Just reporting the improvement on the worst performers would be misleading.

Praising good work could make the next performance get worse and criticizing bad work could make the next performance get better. Or it could be the regression fallacy.

Putting speed cameras in the most dangerous places we might see a safety improvement next year. That could be real or it could be the regression fallacy. If we can randomize which dangerous places get the safety improvement then we have a better chance to learn the causal effect. We might do this randomization if we are not sure that safety will really improve, or, if we can only afford to improve some of the places this year.

The placebo effect is that sometimes people feel better even after taking a placebo treatment that actually does nothing. We can think of a sugar pill. There might be some psychological effect from expecting an improvement. The placebo effect might however be partly due to the regression fallacy. You feel better, then worse, then better then worse and then one day while feeling worse you take some treatment and soon after you're feeling better.

7.3 Linear regression model

We will set up the model as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ under the strongest assumptions we make. This simple model uses only 3 numbers to describe $p(y \mid x)$. Or for our Gauss-Markov assumptions they might only have mean zero constant variance σ^2 and be uncorrelated.

Our design matrix is

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

So

$$X^{\mathsf{T}}X = n\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix} \quad \text{and} \quad X^{\mathsf{T}}Y = n\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

for

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad \text{and} \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

Our least squares estimates

$$\hat{\beta} = \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}^{-1}\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix} = \frac{1}{\overline{x^2} - \bar{x}^2}\begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} \bar{y} \\ \overline{xy} \end{pmatrix}$$

are well defined provided that $\overline{x^2} - \bar{x}^2 \neq 0$. That is a sample variance of the x_i so it cannot be negative but it could be zero if the x_i were all identical. In that case X does not have full rank: the column of x_i would be a constant multiple of the column of 1s.

Assuming X has full rank we get a sample slope of

$$\hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{SXY}}{\mathrm{SXX}}$$

where we define

$$\mathrm{SXX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i(x_i - \bar{x}) \quad \text{and} \tag{7.2}$$

$$\mathrm{SXY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i. \tag{7.3}$$

We will use these sums of squares and cross products later. When you need a quantity like SXY you have three choices for which one to use. The third one above is convenient because the y values appear in only one place. The reason we can do that is that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ from which $\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = 0$ too. We will also use

$$\mathrm{SYY} = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The sample intercept is

$$\begin{aligned} \hat{\beta}_0 &= \frac{\overline{x^2}\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2} \\ &= \frac{(\overline{x^2} - \bar{x}^2)\bar{y} + \bar{x}^2\bar{y} - \bar{x}\overline{xy}}{\overline{x^2} - \bar{x}^2} \\ &= \bar{y} - \bar{x}\hat{\beta}_1. \end{aligned}$$

It follows that $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, so that the regression line goes through $(\bar{x}, \bar{y})$.

7.4 Centering the x_i

If we center the x_i then the derivation is easier. We write

$$y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i.$$

Our new β is our old β_1 so we don't really need a new symbol for it. We then write

$$y_i = \alpha + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

where $\alpha = \beta_0 + \beta_1 \bar{x}$. Now our design matrix is

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$$

so

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{pmatrix} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y = \begin{pmatrix} n & 0 \\ 0 & \mathrm{SXX} \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ \mathrm{SXY} \end{pmatrix}$$

from which we can read off $\hat{\beta}_1 = \mathrm{SXY}/\mathrm{SXX}$ and $\hat{\alpha} = \bar{y}$.

The least squares estimate of β_0 must obviously be $\hat{\alpha} - \hat{\beta}_1 \bar{x}$ because $\beta_0 = \alpha - \beta_1 \bar{x}$. To justify this obvious result write $\beta_0 = \alpha - \beta_1 \bar{x} = \boldsymbol{c}(\alpha, \beta_1)^\mathsf{T}$ for $\boldsymbol{c} = (1, -\bar{x})$. The least squares estimate of $\boldsymbol{c}(\alpha, \beta_1)^\mathsf{T}$ is $\boldsymbol{c}(\hat{\alpha}, \hat{\beta}_1)^\mathsf{T} = \hat{\alpha} - \hat{\beta}_1 \bar{x}$ as expected.

7.5 Distribution of $\hat{\beta}_1$

We are usually more interested in β_1 than β_0. Of course $\mathbb{E}(\hat{\beta}_1) = \beta_1$. Because only y_i are random

$$\mathrm{var}(\hat{\beta}_1) = \mathrm{var}\Big(\frac{\mathrm{SXY}}{\mathrm{SXX}}\Big) = \frac{\mathrm{var}(\mathrm{SXY})}{\mathrm{SXX}^2}$$

and

$$\mathrm{var}\left(\sum_{i=1}^n (x_i - \bar{x}) y_i\right) = \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = \mathrm{SXX}\sigma^2$$

so

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\mathrm{SXX}}.$$

Here is where we used SXY $= \sum_{i=1}^n (x_i - \bar{x})y_i$. If we had used SXY $= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ then we would have gotten the answer after a longer argument using $\mathrm{cov}((x_i - \bar{x})(y_i - \bar{y}), (x_{i'} - \bar{x})(y_{i'} - \bar{y}))$ for $i \neq i'$.

For an interpretation we can write this as

$$\mathrm{var}(\hat{\beta}_1) = \frac{1}{n}\frac{\sigma^2}{\hat{\sigma}_x^2} \quad \text{for} \quad \hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

We want a small $\mathrm{var}(\hat{\beta}_1)$. The first and obvious way to get that is to take a larger value of n. Here, as quite commonly in statistics, the estimate has a variance proportional to $1/n$. Taking larger n may be expensive. It is also impossible if we already have all n data points. For example all $n = 50$ states.

Another way to reduce $\mathrm{var}(\hat{\beta}_1)$ is to have a smaller value of $\sigma^2 = \mathrm{var}(y_i \,|\, x_i)$. We don't often think of reducing the variance of y_i, but in some contexts we can do that. In industrial settings, more expensive equipment or longer testing could reduce $\mathrm{var}(y_i)$. If each y_i is itself an average of other quantities such as m repeated measurements on subject i then larger m reduces variance also at some increased cost.

The most interesting way to reduce $\mathrm{var}(\hat{\beta}_1)$ is to increase $\hat{\sigma}_x^2$. If we must have $A \leqslant x_i \leqslant B$ then for even n the optimal choice is to take $n/2$ observations with $x_i = A$ and $n/2$ with $x_i = B$. Moreover we can do even better by taking $B - A$ extremely large.

We should be cautious about actually using something optimal. A mathematical optimum is usually found with a combination of assumptions about what is true and assumptions about our goals. Taking half the data at A and half the data at B leaves us no way to test for nonlinearity in $\mathbb{E}(y \,|\, x)$. There could be some curvature that we would want to detect and that does not affect $\mathrm{var}(\hat{\beta}_1)$ so the optimization does not take account of it. It makes sense to include some intermediate values of x unless we are very confident in the linearity from past data or theory. We might use three or four levels. An extreme choice would be to use n different equispaced levels. That would increase $\mathrm{var}(\hat{\beta}_1)$ (by a factor that would be a simple exercise) and it might be logistically difficult in a planned experiment. Similarly, taking A and B far apart may take us into x values where linearity is more questionable.

We can run the logic in reverse. If $\hat{\beta}_1$ is not significantly different from 0 that could be because $\hat{\sigma}_x^2$ is small. A confidence interval for β_1 will give information about whether the non-significance comes with confidence that $|\beta_1|$ is small or whether the interval includes some very large values of $|\beta_1|$.

One way for $\hat{\sigma}_x^2$ to be small is when we have selected only (x_i, y_i) pairs with x_i in a tiny range. A classic example has y_i equal to a student or employee performance measure with x_i equal to a predictor such as GPA or an SAT score. A university that only admits students with a narrow range of SAT scores is less likely to see a significant relationship between SAT score and first year GPA.

In class I mentioned that student selection at the university will be based in part on other variables. The example I gave was strong extracurriculars. If that happens then it can bias the regression on SAT alone. We will look at missing variables later.

7.6 Distribution of $\hat{\beta}_0$ and $\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$

We can go through a similar analysis and find that

$$\mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2}{n}\frac{\overline{x^2}}{\overline{x^2} - \bar{x}^2}.$$

In class I sketched two point clouds with x_i tightly clustered. In one they were clustered far from 0 as measured by $\hat{\sigma}_x$. In the other they were clustered near zero. Drawing possible regression lines through the point clouds shows that the first case will have an enormous $\mathrm{var}(\hat{\beta}_0)$ and the second case will have it small.

Now suppose that a fresh new x value is being contemplated. We call it x_{n+1} though it could be a repeat of a previous x_i. We don't yet have the response y_{n+1} so we would estimate it by

$$\begin{aligned}
\widehat{\mathbb{E}}(y \,|\, x_{n+1}) &= \bar{y} + \hat{\beta}_1(x_{n+1} - \bar{x}) \\
&= \sum_{i=1}^{n}\Big(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})(x_i - \bar{x})}{\mathrm{SXX}}\Big) y_i \\
&= \sum_{i=1}^{n} \lambda_i y_i
\end{aligned}$$

for $\lambda_i = 1/n + (x_{n+1} - \bar{x})(x_i - \bar{x})/\mathrm{SXX}$. In the first expression above we

have used centered x_i to make it a bit easier. Then

$$\begin{aligned}\operatorname{var}(\widehat{\mathbb{E}}(y\,|\,x_{n+1})) &= \sum_{i=1}^{n}\lambda_i^2\sigma^2\\ &= \sum_{i=1}^{n}\Big(\frac{1}{n^2}+\frac{2}{n}\frac{(x_{n+1}-\bar{x})(x_i-\bar{x})}{\mathrm{SXX}}+\frac{(x_{n+1}-\bar{x})^2(x_i-\bar{x})^2}{\mathrm{SXX}^2}\Big)\sigma^2\\ &= \sigma^2\Big(\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\mathrm{SXX}}\Big).\end{aligned}$$

What we see here is that predictions at x_{n+1} near $\bar{x}$ have less variance while extrapolation for x_{n+1} far from $\bar{x}$ increases variance. We also see that large n and large SXX help at every x_{n+1}. The variance at $x_{n+1}=\bar{x}$ is σ^2/n the same as we would have gotten putting all n observations there.

There is another danger in extrapolation. As we move x_{n+1} far from the data there is the possibility that the linearity of $\mathbb{E}(y\,|\,x)$ breaks down. That would make $\widehat{\mathbb{E}}(y\,|\,x_{n+1})$ differ from our linear model adding a bias that our modeling does not account for.

7.7 Berkson's paradox

Suppose that x and y have some joint distribution but our data only include those with $\lambda x+(1-\lambda)y > t$ for some weight $\lambda\in(0,1)$ and some threshold t.

Classic, though hypothetical examples, refer to Hollywood actors scored on talent and looks. If x and y are uncorrelated prior to selection at some high level t then they are negatively correlated after selection. In that case a regression of y on x would follow a different typically nonlinear model under selection than it does without the selection. Similarly restaurants might be selected based on a combination of location quality and food quality.

A similar selection effect can happen if we need $\max(x,y)>t$ to select the data. For example a site might be selected if either one of two pollutants exceeds a high threshold. We can then get a negative correlation even if prior to selection the correlation was zero.

This becomes an issue of external validity. Regression on the selected data can still be predictive ***in the selected population*** but the findings then fail to generalize to a setting without the selection.

The selection can change the correlation. It is not always about turning uncorrelated values into negatively correlated values.

7.8 R^2

The R^2 is the proportion of variation 'explained' by the regression. It is

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where we have written it as 1 minus the proportion unexplained. Algebra like we have seen already gives

$$R^2 = \frac{\mathrm{SXY}^2}{\mathrm{SXXSYY}} = \hat{\rho}_{x,y}^2$$

where $\hat{\rho}_{x,y}$ is the sample correlation between x_i and y_i.

We generally prefer models with high explanatory power. Whether R^2 is large enough for a model to be useful depends on context. Predicting stock returns with $R^2 = 0.04$ might be quite useful. In other settings an R^2 of 0.99 might not be accurate enough. That can happen if our response is the logarithm of the quantity we really want to predict. If we exponentiate upper and lower limits from a model on $y = \log(\tilde{y})$ the result might not be very precise.

7.9 Regression through the origin

We didn't cover this case. Regression through the origin is useful pedagogically. We will see the effects of some model errors most simply without an intercept.

My goto examples for this case include having x_i equal to the population of county i and y_i equal to a number of cancer deaths. It is pretty clear that $x = 0 \Rightarrow y = 0$ and so we might entertain a model like

$$y_i = x_i\beta + \varepsilon_i$$

with no intercept. This setting would typically have $\mathrm{var}(y_i)$ growing with x_i. We will consider how to handle that later.

Another example has x_i equal to an amount of fuel and y_i equal to distance traveled. Once again we may expect $x = 0 \Rightarrow y = 0$. This could be a bad choice. Suppose that none of the data actually had x_i anywhere near 0. Then forcing the line to go through $(0, 0)$ could bring significant bias, much more than $\beta_0 + \beta_1 x_i$ over the relevant range of x_i.

It is hard to get a meaningful R^2 for regression through the origin or for other models that do not include an intercept. When there is an intercept we are assured that $\sum_{i=1}^n (y_i - \hat{y}_i)^2 \leqslant \sum_{i=1}^n (y_i - \bar{y})^2$. Without an intercept, this can fail and leave us with $R^2 < 0$. We also lose $R^2 = \rho_{x,y}^2$.

7.10 Errors in variables

The notion of errors in variables describes the setting where both x_i and y_i are measured with error. So x_i is the variable that now has the new kind of error. An example from the work of Dan Kluger runs as follows. We want to learn the presumably negative effect of a pollutant on crop yield. The best data on pollutant levels come from areas in cities where human health is being monitored. Near the farms we must make some educated guesses about what the pollution level is.

We set this up as

$$\begin{aligned} x_i &= u_i + \delta_i \\ y_i &= v_i + \varepsilon_i \\ v_i &= \beta_0 + \beta_1 u_i. \end{aligned}$$

We want to learn the effects of the real pollutant level u_i on the real agricultural yield y_i or on its expected value v_i. But now we only get noisy versions of both variables. Up to this point we have had exact x_i and y_i scattered around a line. We assume that $\delta_i \sim (0, \sigma_0^2)$ independently of $\varepsilon_i \sim (0, \sigma_1^2)$ and different observations have independent errors $(\delta_i, \varepsilon_i)$. We also assume that δ is independent of u.

Given (x_i, y_i) data we can estimate $\mathbb{E}(y \,|\, x)$. But what we want is $\mathbb{E}(v \,|\, u)$ which is the same as $\mathbb{E}(y \,|\, u)$ because $\mathbb{E}(\varepsilon \,|\, u) = 0$. We find that

$$\begin{aligned} \mathbb{E}(y \,|\, x) &= \mathbb{E}(\beta_0 + \beta_1 u + \varepsilon \,|\, x) \\ &= \beta_0 + \beta_1 \mathbb{E}(u \,|\, x) \\ &= \beta_0 + \beta_1 (x - \mathbb{E}(\delta \,|\, x)). \end{aligned}$$

Now $\mathbb{E}(\delta \mid x) \neq 0$ in general. A large x comes from a large $u + \delta$ so x is informative about δ. As a result

$$\mathbb{E}(y \,|\, x) \neq \beta_0 + \beta_1 x$$

and the slope we get from regressing y on x won't be the same as regression v on u.

To see how it works we suppose that

$$\begin{pmatrix} u \\ \delta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_u \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix} \right).$$

Then

$$\begin{pmatrix} u \\ x \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u \\ \delta \end{pmatrix}$$

has a bivariate Gaussian distribution from which we can find that

$$\mathbb{E}(u\,|\,x) = \frac{\sigma_u^2 x + \sigma_0^2 \mu_u}{\sigma_u^2 + \sigma_0^2}$$

after filling in some details. We end up with

$$\mathbb{E}(y\,|\,x) = \beta_0 + \beta_1 \frac{\sigma_0^2 \mu_u}{\sigma_u^2 + \sigma_0^2} + \beta_1 \frac{\sigma_u^2}{\sigma_u^2 + \sigma_0^2} x.$$

What we see above is that the regression slope for y versus x is the slope β_1 that we want multiplied by $\sigma_u^2/(\sigma_u^2 + \sigma_0^2) < 1$. The slope is attenuated towards zero by the errors in x_i.

This presentation is based on Chapter 5 of Miller (1986). Beyond the scalar x case, if component j of $\boldsymbol{x}_i$ is measured with error then the least squares estimate $\hat{\beta}_k$ for $k \neq j$ can also be biased. For more on measurement error in models see Carroll et al. (2006).

7.11 Confidence bands

For any $x \in \mathbb{R}$ and assuming Gaussian errors

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{SXX}}}} \sim t_{(n-2)}$$

by our usual recipe for a t test. At any fixed x that gives us a confidence interval of

$$\beta_0 + \beta_1 x \in \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{(n-2)}^{1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{SXX}}}.$$

The upper and lower limits trace out curves versus x, actually branches of an hyperbola. They get wider as x departs from $\bar{x}$.

For a general linear regression model with Gaussian errors we get

$$\boldsymbol{x}^{\mathsf{T}}\beta \in \boldsymbol{x}^{\mathsf{T}}\hat{\beta} \pm t_{(n-p)}^{1-\alpha/2} s \sqrt{\boldsymbol{x}^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\boldsymbol{x}}.$$

We may also want prediction bands. We are about to get a value x and we want an upper and lower limit for the actual y value not just $\mathbb{E}(y\,|\,x)$. Of course the interval will have to be wider. Now

$$y = \big(\beta_0 + \beta_1 x - \hat{\beta}_0 + \hat{\beta}_1 x\big) + \varepsilon.$$

The two terms above are independent Gaussians, so we can add their variances. The result is that

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x) - \varepsilon}{s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{SXX}}}} \sim t_{(n-2)}.$$

We have added a 1 inside the square root to account for $\mathrm{var}(\varepsilon) = \sigma^2$. Now

$$\Pr\left(y \in \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{(n-2)}^{1-\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\mathrm{SXX}}}\right) = 1 - \alpha,$$

gives us a prediction interval for the next value of y given x not just for $\mathbb{E}(y \,|\, x)$.

Bibliography

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

R. G. Miller. *Beyond ANOVA, Basics of Applied Statistics*. Wiley, 1986.