Lecture 8

Generic predictor $\boldsymbol{z} \in \mathbb{R}^d$

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.*

This is our first lecture on statistical properties of regression on a generic vector of predictors. This setting is sometimes called 'multiple regression' to distinguish it from the case of a single predictor x_i with $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.

It is an annoyance that the intercept is sort of a predictor and sort of not a predictor. Weisberg handles this by having p predictors and p', generally $p+1$, model parameters and then there are $n - p' = n - p - 1$ degrees of freedom for error. Some people number their coefficients $\beta_0, \dots, \beta_p$ and others use $\beta_0, \dots, \beta_{p-1}$.

For this lecture we will suppose that observation i comes with variables $\boldsymbol{z}_i \in \mathbb{R}^d$ and we have turned them into variables $\boldsymbol{x}_i \in \mathbb{R}^p$, so $\boldsymbol{x}_i = \phi(\boldsymbol{z}_i)$ for some function $\phi : \mathbb{R}^d \to \mathbb{R}^p$. We always have $x_{i1} = 1$ in this lecture. Then our regression model is

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \tag{8.1}$$

with $\boldsymbol{\varepsilon} \sim (\boldsymbol{0}, \sigma^2 I_n)$ possibly Gaussian for our inferences.

8.1 Interpretations

This lecture is largely about how face-value interpretations of (8.1) can be misleading. We will look at some problems that can arise. Later lectures will be about how to fix (some of) those problems.

Naive face-value interpretations of (8.1) are that

$$\frac{\partial}{\partial x_{ij}} y_i = \beta_j \quad \text{or just} \quad \frac{\partial}{\partial x_{ij}} \mathbb{E}(y_i) = \beta_j$$

where as before we consider $X \in \mathbb{R}^{n \times p}$ to be non-random and we exclude the intercept from this interpretation.

Such interpretations might actually be reasonable in some settings such as the one in Forbes' data that we discussed earlier. It is quite reasonable that a linear regression of boiling point on (log) atmospheric pressure would give a linear model with a reasonably accurate face-value interpretation as above if not for y_i then for $\mathbb{E}(y_i)$.

The Forbes regression was actually in the other direction predicting atmospheric pressure from the boiling point, so it is not causal. It still has a reasonable face-value interpretation about what would happen for two measurements taken, say 1°F apart in repeats of his procedure.

8.2 Bodyfat example

As a worked example I looked at bodyfat data. We saw R^2 values and regression coefficients' significance levels. The example touches on outliers and results that are too good to be true, so we make slight digressions there. Here is a link to the data: `https://jse.amstat.org/datasets/fat.txt`. I think there may be more than one version of this dataset on the internet.

I looked at the data and found an obvious outlier. Somebody was listed with height 29.5 inches and weight 205 pounds. There were also some anomalous relationships between percent bodyfat and density. I removed those values from the regressions. This is a 'researcher degree of freedom'. We sometimes are forced to make that kind of choice and they affect our results. That's actually why we make those choices: to get better results. Different researchers will get different answers.

We should generally expect bad data might be included in our dataset and we should look for it. Time permitting, later lectures will consider robust regression methods that are relatively immune to bad data but those have other issues.

The website actually talks about those specific outliers. It speculates that the 29.5 inch height was probably really 69.5. It also reports errors like one person having bodyfat 0 because a negative value got moved up to 0 and some other cases having arithmetic errors in their computation. In my analysis, before I saw that explanation, I had removed those cases, including the one with bodyfat 0, but I had not noticed that case had 0 bodyfat or corrected the errors in the formula.

I showed the scatterplot matrix Figure 8.1 in class. The height outlier is quite visible. There are a few other unusual cases there, such as extremely large ankles or biceps or extremely small forearms. If you look at the row for 'thigh' there is somebody in the data set who is clearly largest on many of the measures. Most of the pairwise correlations are positive with points falling in a roughly ellipsoidal pattern. A few of them show 'joint outliers' where somebody is in the range on both variables but is outside the scatterplot ellipse. See for example, large wrists with small forearms and vice versa. These might be outliers but the evidence isn't compelling and so, again exercising a researcher degree of freedom, they were not removed from the data set.

The scatterplot matrix is a good way to spot suspicious data but that doesn't work for a very large number of variables. I mentioned in class a genomic setting with tens of thousands of variables (gene measuring probes). Each subject had some demographic variables along with their genomic variables. To find questionable data it was necessary to make up a statistic quantifying 'questionableness' and look at the extreme cases. The extreme cases identified some female subjects expressing genes that are only located on the Y chromosome. That's probably an error though there are biological possibilities such as having recently carried a male fetus. There were also male subjects expressing some X chromosome genes that are used to silence one of the two copies of the X chromosome in women. The likely explanation is that some arrays were mislabeled.

Somebody asked what to do if we knew that the gender was wrong. Could we just correct it? Maybe we could if the data were just (gender, genes) but if it is (gender, age, $\cdots$, genes) then correcting the gender might make another bad data point. If we had just one male and one female subject with anomalous genetic data we could swap their genetic data, but with several of each that would be error prone producing a demographic-genomic person that never existed. [I don't remember what we did. Most likely the points were dropped or a correct labeling was found.]

I showed some regression models for the bodyfat data. The first model regressed bodyfat on all other variables in the dataset. It scored a phenome-

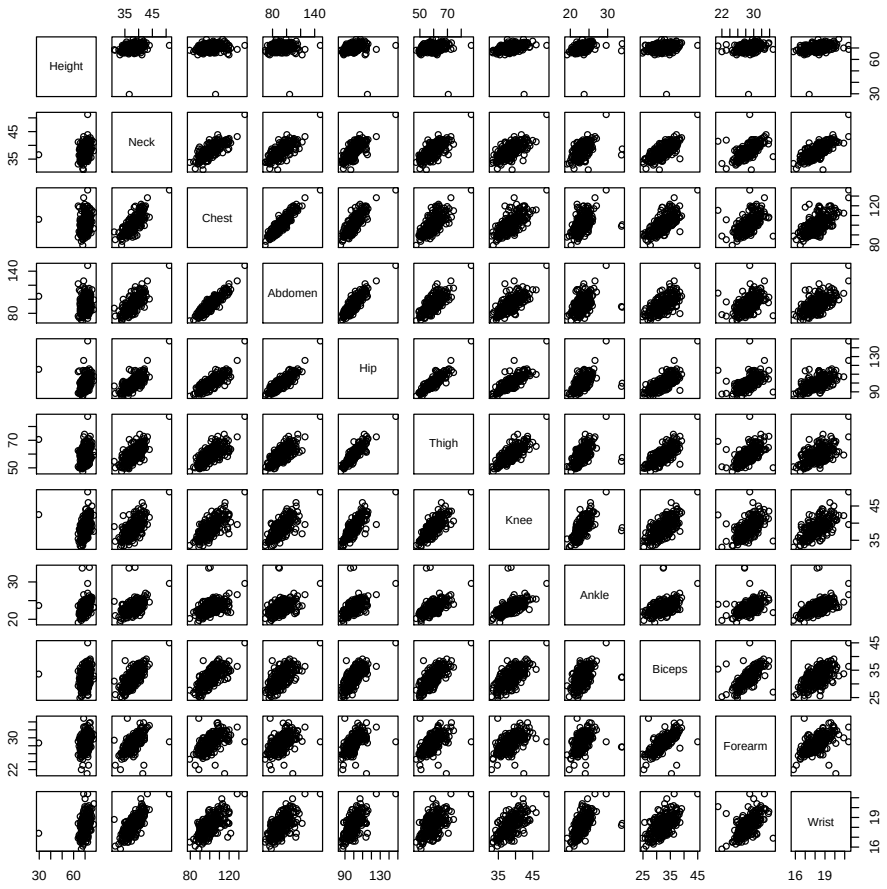


Figure 8.1: Scatterplot matrix of the bodyfat data dimensions

nal $R^2 = 0.9996$ which would let you predict somebody's bodyfat percentage to within a fraction of a percent.

That is obviously too good to be true. Inspection of the regression output reveals that the 'density' variable was highly significant. As I noted in class the formula for bodyfat proportion was

$$\text{bodyfat} = \frac{4.95}{\text{density}} - 4.5$$

and the density was computed using subjects' weight in water, weight in air and lung volume. This explains why the model using density fit so well.

The reason that it is too good to be true is that we were using all those measurements in order to avoid weighing people in both air and water. The density measure requires us to do both of those weighings so it would not be available if we avoid doing those weighings.

It is pretty common for a problem like this to arise and a very high R^2 is a tipoff, though the issue can arise without the result being so blatant. In a setting where an R^2 of 4% is considered good, 8% could be a sign of trouble. Another way it can happen is that our data matrix might include quantities that were measured after y was observed. Those could be highly correlated with y but they will not be available to us in the future when we predict y.

A medical system could predict sepsis on the basis of diagnostic tests that were run, but those tests might have been ordered because a caregiver suspected sepsis.

Things like this happen with black box models. One example is a predictor of cancer from an image that, upon inspection, strongly relied on the presence of a ruler near the lesion. So it is right for the wrong reason because the ruler was more likely to be placed beside a lesion already thought to be cancerous. Another example is a system that was astonishingly good at predicting what country a photo was taken in, in part because it had learned about lens flaws in the cameras that had been used in the different countries. That is right for the wrong reason because it won't generalize as well to different camera choices.

Another form of leakage from the future can arise in finance. If your dataset excludes companies that went bankrupt then you might find a strategy of buying stocks that have declined greatly appears to work very well. Some of those decliners would have gone bankrupt, and some would have recovered but with the bankrupt ones removed buying-the-dip starts to look way better than it should.

8.3 'Wrong' signs

We looked at ways that the sign of a regression coefficient can be different from the one we think is right. That can easily happen due to the presence or absence of other variables in the model.

If $X^\mathsf{T}X$ is diagonal then $\hat\beta_j = \sum_{i=1}^n x_{ij}y_i / \sum_{i=1}^n x_{ij}^2$. If we remove variable k from the model $\hat\beta_j$ does not change. We would usually see a change in s^2 so the presence or absence of x_k in our model could affect the significance of $\hat\beta_j$ but not the estimate itself. For non-diagonal $X^\mathsf{T}X$ there is a change to $\hat\beta_j$ when x_k is removed from the model.

In class I showed some penguin data relating two dimensions of their bill (beak) in Figure 8.2. It is available as `penguins` in the `palmerpenguins` R package. The slope within each group is positive while lumping all the data together gives a negative slope. **Puzzler**: determine the collection of variables that we would have to add to the global regression to get the three within-species regression lines. It's not just adding 'species' as a 3 level variable.

The dataset includes several other variables, such as the island the birds were on and their sex, mass and flipper length. Any collection of adjustment variables can yield a different slope. The question we ask about β_j depends on whether x_k is in or out of the model.

When our parameter of interest is β_j and we have included x_j in the model, there are still 2^{p-1} different regressions we could do, one for each $S \subset \{1,2,\dots,p\} \setminus \{j\}$ of other variables we could have included, though ordinarily only the 2^{p-2} of them that include the intercept would get serious consideration. Each of those regressions will generally give a different value of $\hat\beta_j$. Each will generally have a different 'true' value of $\beta_{j(S)}$ as the minimizer of

$$\sum_{i=1}^n \mathbb{E}\Big(\Big(y_i - \beta_j x_{ij} - \sum_{k\in S}\beta_k x_{ik}\Big)^2\Big)$$

over β_j and β_k for $k \in S$.

Which of them is right? They all address different questions, so then the issue is about which is appropriate to a given use case. In the example above one slope is about penguins in general and the other is species-by-species.

Until now we have supposed that $\mathbb{E}(Y) = X\beta$ holds. Leaving $x_k\beta_k$ out of a model violates this if $\mathbb{E}(Y)$ includes $\beta_k \neq 0$.

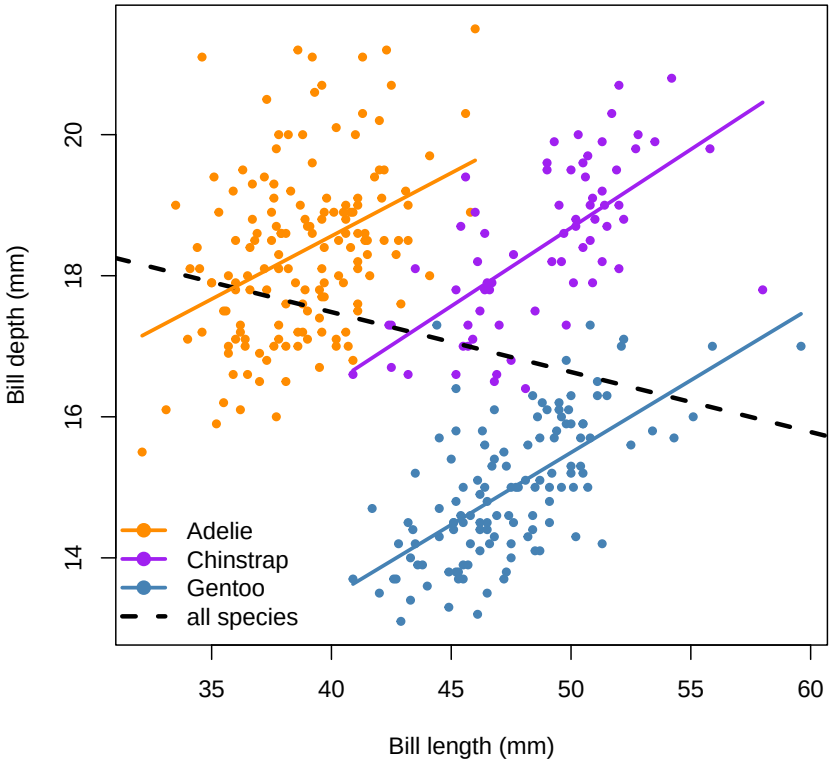


Figure 8.2: This example shows a positive regression slope within each species of penguin but a negative slope overall.

8.4 Causally questionable signs

The 'wrong' signs in the prior section were about the difference between $\mathbb{E}(y \mid x)$ averaged over z and $\mathbb{E}(y \mid x, z = z_0)$ conditionally on z taking the values z_0 in some set. There, wrong is in the eye of the beholder. The different slopes are used to answer different questions. Maybe one question is most worthy of an answer, or maybe several different formulations considering the slope with x averaging over some but not all of another set of variables are all of interest.

Another kind of 'wrong sign' issue arises when we think the regression can reveal the nature of a causal link from x to y. If x and y are associated it could be

1) because x causes y, or
2) because y causes x, or
3) because some third variable z causes both x and y.

Even if x directly causes y, we may not see it in our data if the only values we get are subject to sampling biases, such as Berkson setting in Lecture 7 where we only get to see certain special (x_i, y_i) pairs.

We might have some idea of the causal direction of x on y such as positive or negative or zero. The sign of the corresponding regression coefficient can also be positive or negative or zero, and all nine possibilities can arise. We might also have ideas about the magnitude of the effect that do not match what we see in the regression. There are some examples below.

A company might offer a discount of size x to a potential customer and get an order of size y. It is possible that larger discounts are associated with smaller order sizes. An explanation might come from a variable z that indicates whether the potential customer has bought anything in the past. Even a large discount might not be enough to get a former non-customer to make any order at all while existing customers may need no incentive to remain.

Chapter 13E of Mosteller and Tukey (1977) reports an example from World War II where there was a positive coefficient on a measure of success of bombing raids for a variable describing the amount of fighter opposition trying to shoot down the bombers. Their explanation was that non-cloudy weather made the bombers more accurate while also making them more visible to the opposition.

Chapter 2 of Sen and Srivastava (1990) have an example where y_i is the selling price of a home and x_{ij} is the number of bedrooms. They got $\hat{\beta}_j < 0$. An explanation is that adding a bedroom would almost always involve adding some square feet, so a causal change should include changing

both of those variables. While I like the logic in that explanation I'm less fond of this example than before because I see that $n = 26$ and there were up to 13 predictors though not all were used in each example regression. Of course changing both the number of bedrooms and the number of square feet at once will still have no assurance of giving us the causal answer.

One practice that you might encounter is where somebody decides whether or not to include x_k based on whether the resulting sign of $\hat{\beta}_j$ is the desired one. In some cases there is also a desire to get the magnitude $|\hat{\beta}_j|$ into a specific range. That is unlikely to get a good model and if there ever is a reason to do a constrained optimization of this sort it should be prominently announced so that nobody thinks it was the result of ordinary least squares without those constraints. A risk for that practice is that every time there is a new data set it is likely that the variable set will change.

Regression gives us predictions tuned to minimize squared error. We are finding a hyperplane in $\mathbb{R}^{p+1}$ where the last component y is clustered as close as we can get it to $\boldsymbol{x}^{\mathsf{T}}\beta$ a linear combination of the first p variables. There is no general reason why β should then have a causal interpretation. It can be very predictive.

8.5 Partial correlation

In class I gave an example with three variables measured on children and teens in Boston: a measure y_i of lung volume, smoking level x_i and age z_i. The example is from Kahn (2005) and it includes links to the data. A regression of y_i on $\boldsymbol{x}_i = (1, x_i)^{\mathsf{T}}$ yields a positive coefficient for smoking. Smoking more is associated with greater lung volume. A regression of y_i on $\boldsymbol{x}_i = (1, x_i, z_i)^{\mathsf{T}}$ yields a negative coefficient for smoking. If we adjust for age, smoking is associated with lesser lung volume. Older kids tended to have both greater lung capacity and a higher rate of smoking than younger kids. The positive correlation between x_i and y_i in the data is very likely to hold for the population as a whole but it (obviously) does not imply that smoking is beneficial for lung health. It's not answering the question we might have expected to be answered.

What we would rather know is what the slope is at any given age. There are not enough data points to do a separate regression of y on x at every age z, especially if we slice z very finely. What we can do instead is contemplate whether an age-adjusted y correlates positively or negatively with an age-adjusted x.

Let $y_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i + \hat{\varepsilon}_i$ and $x_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i + \hat{\eta}_i$ be two least squares models

fit to z_i. We can then interpret $\tilde{y}_i = \hat{\varepsilon}_i$ as 'age adjusted lung volume', the lung volume of subject i above or below what is typical for a subject of age z_i. Similarly $\tilde{x}_i = \hat{\eta}_i$ is an 'age adjusted smoking'. Whether these are good adjustments or bad adjustments depends on the quality of the regression models they are based on. We can then find the correlation of $\tilde{x}_i$ and $\tilde{y}_i$ and use it as a measure of whether lung volume and smoking are correlated after adjusting for age.

This is sometimes described as 'controlling for age', but that makes it sound like a perfect adjustment which we do not reasonably expect. Andrew Gelman has made that point frequently in his blog and his publications. The idea generalizes to adjustments for more than one variable like age here by regressing them all out of both y and x and correlating the residuals. We could use nonlinear prediction models to adjust both x and y before correlating the residuals.

It is useful to think of what the partial correlation is estimating. For this we suppose that we have IID observations $(x_i, y_i, z_i)^{\mathsf{T}}$ with finite variance of full rank hence a well defined correlation matrix of

$$\begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & 1 & \rho_{yz} \\ \rho_{xz} & \rho_{yz} & 1 \end{pmatrix}.$$

In this case the population version of the partial correlation between x and y adjusting for z is

$$\rho_{xy\bullet z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}. \tag{8.2}$$

It is a well-known formula. I didn't see it in the books I checked but it must be out there. Wikipedia lists it with 'citation needed' as of October 2026.

We establish this formula in a few steps. We define the population adjusted quantities to be

$$\tilde{y} = y - \alpha_0 - \alpha_1 z \quad \text{and} \quad \tilde{x} = x - \gamma_0 - \gamma_1 z.$$

where α_0 and α_1 minimize $\mathbb{E}((y - \alpha_0 - \alpha_1 z)^2)$ and similarly for γ_0 and γ_1. We know that $\alpha_1 = \rho_{yz}\sigma_y/\sigma_z$ and $\gamma_1 = \rho_{xz}\sigma_x/\sigma_z$, where σ_x^2, σ_y^2 and σ_z^2 are variances of x and y and z. To test yourself, verify α_1 and γ_1 by going back if necessary to the lecture notes on least squares or on a scalar predictor and thinking of 'population' versions of least squares.

Now

$$\begin{aligned}\rho_{xy\bullet z} &= \mathrm{cor}(y-\alpha_0-\alpha_1 z, x-\gamma_0-\gamma_1 z)\\ &= \mathrm{cor}(y-\alpha_1 z, x-\gamma_1 z)\\ &= \frac{\mathrm{cov}(y-\alpha_1 z, x-\gamma_1 z)}{\sqrt{\mathrm{var}(y-\alpha_1 z)\mathrm{var}(x-\gamma_1 z)}}\end{aligned}$$

and

$$\mathrm{var}(y-\alpha_1 z) = \sigma_y^2 - 2\alpha_1\rho_{yz}\sigma_y\sigma_z + \alpha_1^2\sigma_z^2 = \sigma_y^2(1-\rho_{yz}^2).$$

Then by the same argument $\mathrm{var}(x-\gamma_1 z) = \sigma_x^2(1-\rho_{xz}^2)$. Next

$$\begin{aligned}\mathrm{cov}(y-\alpha_1 z, x-\gamma_1 z) &= \rho_{xy}\sigma_x\sigma_y - \alpha_1\rho_{xz}\sigma_x\sigma_z - \gamma_1\rho_{yz}\sigma_y\sigma_z + \alpha_1\gamma_1\sigma_z^2\\ &= \rho_{xy}\sigma_x\sigma_y - \rho_{xz}\rho_{yz}\sigma_x\sigma_y - \rho_{xz}\rho_{yz}\sigma_x\sigma_y + \rho_{xz}\rho_{yz}\sigma_x\sigma_y\\ &= (\rho_{xy}-\rho_{xz}\rho_{yz})\sigma_x\sigma_y.\end{aligned}$$

Finally

$$\rho_{xy\bullet z} = \frac{(\rho_{xy}-\rho_{xz}\rho_{yz})\sigma_x\sigma_y}{\sqrt{\sigma_x^2(1-\rho_{xz}^2)\sigma_y^2(1-\rho_{yz}^2)}} = \frac{\rho_{xy}-\rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)(1-\rho_{yz}^2)}}$$

as required.

A possibly easier route is to use the conditional distribution of (x, y) given z if they were jointly Gaussian. But the equation (8.2) does not require Gaussianity.

For the lung volume data we naturally prefer $\rho_{xy\bullet z}$ over ρ_{xy} as a description of the association of x and y. We really wanted to adjust for the effect of age to see the effect of smoking on people of the same age.

We don't always prefer a partial correlation over a correlation. Suppose that x is the amount of pesticide applied in April, y is the yield of our apple harvest in September and z is the intensity of moths and other insects on our apple trees in August. If the pesticide is very effective we expect a positive ρ_{xy}. The way the pesticide is effective is probably that it is reducing z. Then if we adjust for z we may be adjusting out the main way that x has been beneficial to y. We could reasonably adjust for variables prior to x.

Suppose that we see $\rho_{xy} > 0$. Somebody might then speculate that this is misleading because there is some missing variable z that would give $\rho_{xy\bullet z} = 0$ if only we had measured it. It is easy to adjust for variables we have, but we cannot adjust for variables that we did not measure. We can however say that if $\rho_{xy} > 0$ and $\rho_{xy\bullet z} = 0$ then equation (8.2) implies that both $|\rho_{xz}| > \rho_{xy}$ and $|\rho_{yz}| > \rho_{xy}$ must be true. That missing variable must

be more strongly associated with each of x and y than they are to each other.

The above is a continuous data version of an argument that Cornfield made when Fisher postulated that the association of smoking and lung cancer could be due to some third variable like genetic predisposition. In that case both of those associations would have to extremely strong and then it is hard to believe such a z could exist. [I keep thinking that somebody could or should have asked Fisher why all those lung cancers started appearing after smoking caught on. Presumbably the genetic explanation should have been operating before that.]

8.6 Variables competing and collaborating

Our regression output usually includes estimates $\hat{\beta}_j$ for each β_j along with a t statistic for $H_{0j} : \beta_j = 0$ and the corresponding p-value. Recall from Lecture 3 that the square of this t statistic is a random variable with the $F_{1,n-p}$ distribution under H_{0j}. That F statistic has an extra sum of squares interpretation.

We could fit least squares by constraining $\beta_j = 0$ and minimizing the sum of squared errors over $(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) \in \mathbb{R}^{p-1}$. That is like dropping X_j (column j of X) from our model and refitting the regression: we get the same least squares solution, the same fits and residuals and sum of squared errors either way.

Let the full model be a regression on X and a submodel be a regression on $X_{-j} \in \mathbb{R}^{n\times(p-1)}$ is the matrix X after we have removed column j. Then let $\mathrm{SS}_{\mathrm{FULL}}$ be $(n-p)s^2$, the sum of squared errors in the full model with all p predictors and SS_{-j} be the sum of squared errors in our sub-model where we have removed variable j from the regression.

Then H_{0j} is rejected for

$$\frac{\mathrm{SS}_{-j} - \mathrm{SS}_{\mathrm{FULL}}}{\frac{1}{n-p}\mathrm{SS}_{\mathrm{FULL}}} > F^{1-\alpha}_{1,n-p}.$$

The value for numerator degrees of freedom is one above. The point of this equation is that variable X_j is scored by how much it reduces the sum of squares after **all** $p-1$ other variables have been included.

Now suppose that x_{ij} and x_{ik} are highly correlated, so $x_{ik} \approx a + bx_{ij}$. Adding X_k to a model that has X_j in it might not reduce the sum of squares much making $\hat{\beta}_k$ insignificant. At the same time $\hat{\beta}_j$ might also be insignificant.

Now suppose that x_{ij} is highly predictive of y_i. A regression on X_j may be highly significant if X_k is excluded but insignificant if X_k is included and vice versa. The variables 'compete' in that if one is in the model the other does not need to be. This can happen for positively or negatively correlated variables.

Let's see how that works out algebraically. We take $\boldsymbol{x}_i = (1, x_{i2}, x_{i3})^{\mathsf{T}}$ with $\sum_{i=1}^n x_{ij} = 0$ for $j = 2, 3$, so we have centered those variables. Then introducing some more moments of $\boldsymbol{x}_i$ we get

$$X^{\mathsf{T}}X = n\begin{pmatrix} 1 & 0 & 0 \\ 0 & \hat{\sigma}_2^2 & \hat{\rho}\hat{\sigma}_2\hat{\sigma}_3 \\ 0 & \hat{\rho}\hat{\sigma}_2\hat{\sigma}_3 & \hat{\sigma}_3^2 \end{pmatrix}$$

and so

$$\mathrm{var}(\hat{\beta}) = (X^{\mathsf{T}}X)^{-1}\sigma^2 = \frac{1}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & \hat{\sigma}_3^2/D & -\hat{\rho}\hat{\sigma}_2\hat{\sigma}_3/D \\ 0 & -\hat{\rho}\hat{\sigma}_2\hat{\sigma}_3/D & \hat{\sigma}_2^2/D \end{pmatrix}\sigma^2$$

for a determinant $D = \hat{\sigma}_2^2\hat{\sigma}_3^2(1-\hat{\rho}^2)$, so

$$\mathrm{var}\begin{pmatrix}\hat{\beta}_2 \\ \hat{\beta}_3\end{pmatrix} = \frac{1}{n}\begin{pmatrix} \hat{\sigma}_3^2 & -\hat{\rho}\hat{\sigma}_2\hat{\sigma}_3 \\ -\hat{\rho}\hat{\sigma}_2\hat{\sigma}_3 & \hat{\sigma}_2^2 \end{pmatrix}\frac{\sigma^2}{D}.$$

When $\hat{\rho} > 0$, we see a negative correlation between $\hat{\beta}_2$ and $\hat{\beta}_3$. That means that the confidence ellipse for $(\beta_2, \beta_3)^{\mathsf{T}}$ will have a negative slope to it.

In class I drew a strongly negative ellipse, centered at $(\hat{\beta}_2, \hat{\beta}_3)^{\mathsf{T}}$ in the first quadrant that intersected both the horizontal and vertical axes. Because some points in the ellipse include $\beta_2 = 0$ we do not rule out $\beta_2 = 0$ and similarly we do not rule out $\beta_3 = 0$. However the ellipse did not include the origin $(0, 0)^{\mathsf{T}}$. In that setting we can be confident that at least one of β_2 and β_3 is nonzero. Exercise: what does that ellipse argument tell you about $(\mathrm{sign}(\beta_2), \mathrm{sign}(\beta_3))^{\mathsf{T}}$?

The ellipse I drew showed that we have great uncertainty about β_2 and β_3 but, in the coordinates I drew it, we had much less uncertainty about $\beta_2 + \beta_3$. Hint: draw contours of $\beta_2 + \beta_3$ over such an ellipse. Exercise: what about $\beta_2 - \beta_3$?

What should one do about such a pair of correlated variables? There are several choices and which is best depends on context. We could drop one of the two predictors. We could form an index by taking a simple or

weighted average of them. That is popular if the index is interpretable. You probably don't want to use $\hat{\beta}_j$ to form an index because those estimates are highly variable.

You can clearly have 3 or more variables highly correlated and competing for significance.

While the phenomenon of two jointly significant but individually not significant variables is common for positively correlated variables you could also have that happen for uncorrelated variables. Just replace the confidence ellipse by a circle that intersects both axes while excluding the origin. You can just barely make it happen for negatively correlated variables!

It is less frequently noted but it is also possible that of two variables neither is significant on its own but both are significant if they are both included in the model. The collaboration effect can arise if you need both predictors to get a small enough value of s^2.

8.7 Causal issues

It would be nice if causal facts could be read off of a regression but that is not a safe assumption. The gold standard approach to studying causality is to intervene and randomly set $\boldsymbol{x}_i$ (or just some components of it) and model how y_i then depends on $\boldsymbol{x}_i$. The following quote is from Box (1966):

> To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it).

For followups, you can study 'Experimental design' to see how this works including how best to choose $\boldsymbol{x}_i$ when you have to (or get to) do that. As we remarked in Lecture 5 even randomized controlled trials must face questions of external validity.

Sometimes the hard or expensive randomization is actually done. There was in the 1970s/1980s a randomized trial that gave free healthcare to people. The free healthcare plan that a subject got was randomized. Some people got better plans than others but every subject was given something they didn't already have and didn't have to accept it.

In many settings it is expensive or unethical or impossible to randomize x_{ij} for one or more indices j, yet we must still do our best to learn about causal effects. Recent decades have seen greatly increased interest in drawing causal conclusions from purely observational data. I highly recommend studying that. In order to draw causal conclusions from observational data

we must use some assumptions about how causality works for our variables. We will look at some example methods which can be described in the language of regression.

Bibliography

G. E. P. Box. Use and abuse of regression. *Technometrics*, 8(4):625–629, 1966.

M. Kahn. An exhalent problem for teaching statistics. *Journal of Statistics Education*, 13(2), 2005.

F. Mosteller and J. W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley Series in Behavioral Science. Addison-Wesley, Reading, MA, 1977.

A. Sen and M. Srivastava. *Regression Analysis: Theory, Methods, and Applications*. Springer Texts in Statistics. Springer-Verlag, New York, 1990.