4
Describing Relationships

What is a Relationship?

For most research questions, we are not just interested in the distribution of a single variable.[1] Instead, we are interested in the *relationship* we see in the data between two or more variables.

[1] Get lost, Chapter 3, nobody likes you.

What does it mean for two variables to have a relationship? The relationship between two variables shows you *what learning about one variable tells you about the other.*

For example, take height and age among children. Generally, the older a child is, the taller they are. So, learning that one child is 13 and another is 6 will give you a pretty good guess as to which of the two children is taller.

We can call the relationship between height and age *positive*, meaning that for higher values of one of the variables, we expect to see higher values of the other, too (older children are taller). There are also negative relationships, where higher values of one tend to go along with lower values of the other (older children cry less). There are also null relationships where the variables have nothing to do with each other (older children aren't any more or less likely to live in France). All kinds of other relationships are positive sometimes and negative other times, or *really* positive at first and then only slightly positive later. Or perhaps one of the variables is categorical and there's not really a "higher" or "lower", just "different" (older children are more likely to use a bike for transportation). Lots of options here.

The goal in this chapter is to figure out how to describe the relationship between two variables, so that we can accurately relay what we see in the data about our research question, which, once again, very likely has to do with the relationship between two variables. Once we know how to describe the relationship we see *in the data*, we can work in the rest of the book to make sure that the relationship we've

described does indeed answer our research question.

Throughout this chapter, we're going to use some example data from a study by Emily Oster,[2] who used the National Health and Nutrition Examination Survey. Her research question was: do the health benefits of recommended medications look better than they actually are because already-otherwise-healthy people are more likely to follow the recommendations?

To study this question, she looked at Vitamin E supplements, which were only recommended for a brief period of time. She then answers her research question by examining the relationship between taking Vitamin E, other indicators of caring about your health like not smoking, and outcomes like mortality, and how those relationships change before, during, and after the time Vitamin E was recommended.[3]

We can start off with an example of a very straightforward way of showing the relationship between two continuous variables, which is a scatterplot in Figure 4.1. Scatterplots simply show you every data point there is to see! They can be handy for getting a good look at the data, and trying to visualize from them what kind of relationship the two variables have. Does the data tend to slope up? Does it slope down a lot? Or slope down just a little like in Figure 4.1? Or go up and down?

[2] Emily Oster. Data and code for: Health recommendations and selection in health behaviors: Replicationfiles. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020a; and Emily Oster. Health recommendations and selection in health behaviors. *American Economic Review: Insights*, 2(2):143–60, 2020b

[3] In this chapter I'll add some analyses that weren't exactly in the original study but are in the same spirit, wherever it helps explain how to describe relationships. It's almost like she had other purposes for her study besides providing good examples for my textbook. Rude if you ask me.

Scatterplot. A graph that plots every data point for an x-axis variable and a y-axis variable.

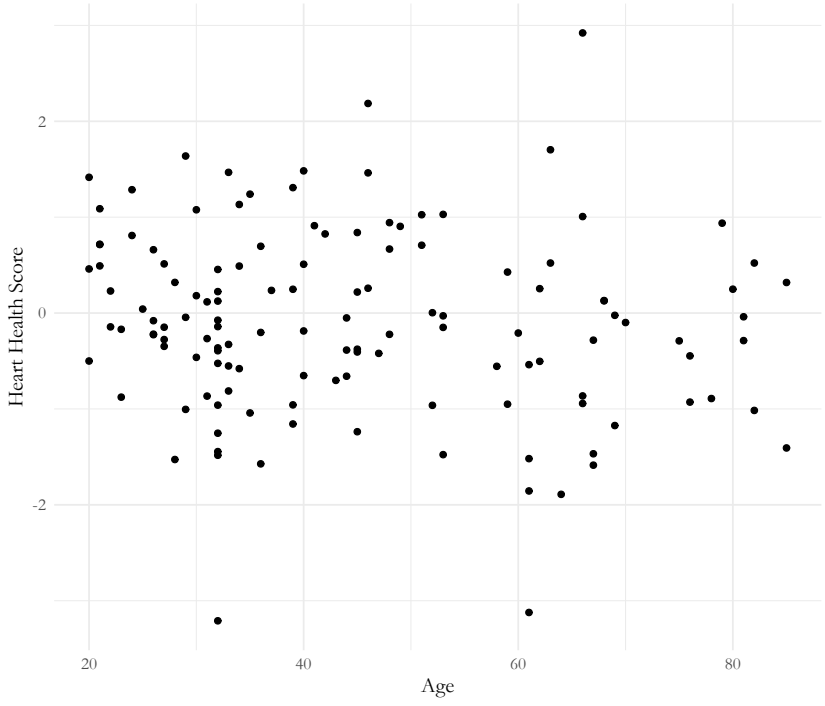


Figure 4.1: Age and Heart Health, 150 observations

A scatterplot is a basic way to show *all* the information about a

relationship between two continuous variables, like the density plots were for a single continuous variable in Chapter 3.[4] And they're usually a great place to start describing a relationship.

But scatterplots imply two things beyond what they actually show. One is bad, and one is good. The bad one is that it's very tempting to look at a relationship in a scatterplot and assume that it means that the x-axis causes the y-axis. Even if we know that's not true, it's very tempting. The good one is that it encourages us to use the scatterplot to imagine other ways of describing the relationship that might give us the information we want in a more digestible way. That's what the rest of this chapter is about.

[4] Unlike density plots, though, they tend to get very hard to read if you have a lot of data. That's why I only used 150 observations for that graph, not all of them.

Conditional Distributions

CHAPTER 3 WAS ALL ABOUT DESCRIBING the distributions of variables. However, the distributions in those chapters were what are called *unconditional* distributions.[5]

A *conditional* distribution is the distribution of a variable *given the value of another variable.*

Let's start with a more basic version—conditional probability. The probability that someone is a woman is roughly 50%. But the probability that someone *who is named Sarah* is a woman is much higher than 50%. You can also say "*among all Sarahs*, what proportion are women?" We would say that this is the "probability that someone is a woman conditional on being named Sarah."

Learning that someone is named Sarah changes the probability that we can place on them being a woman. Conditional distributions work the same way, except that this time, instead of just a single probability changing, an entire distribution changes.

Take Figure 4.2 for example. In this graph, we look at the distribution of how much Vitamin E someone takes, among people who take any. We then split out the distribution by whether someone has engaged in vigorous exercise in the last month.

We can see a small deviation in the distribution for those who exercise and those two don't.[6] In particular, those who exercise vigorouly take take larger doses of Vitamin E when they take it. The distribution is different between exercisers and non-exercisers, telling us that Vitamin E and exercise are *related* to each other in this data.

[5] These are also called "marginal" distributions but I really dislike this term, as I think it sounds like the opposite of what it means.

Conditional distribution. The distribution of a variable conditional on another variable taking a certain value.

[6] It doesn't look enormous, but this is actually how a lot of fairly prominent differences look in the social sciences. That rightward shift can be deceptively larger than it looks!

THE EXAMPLE I'VE GIVEN is for a continuous variable, but it works just as well for a categorical variable. Instead of looking at how large the doses are, let's look at whether someone takes Vitamin E at all! Oster's hypothesis is that people who take Vitamin E at all should be

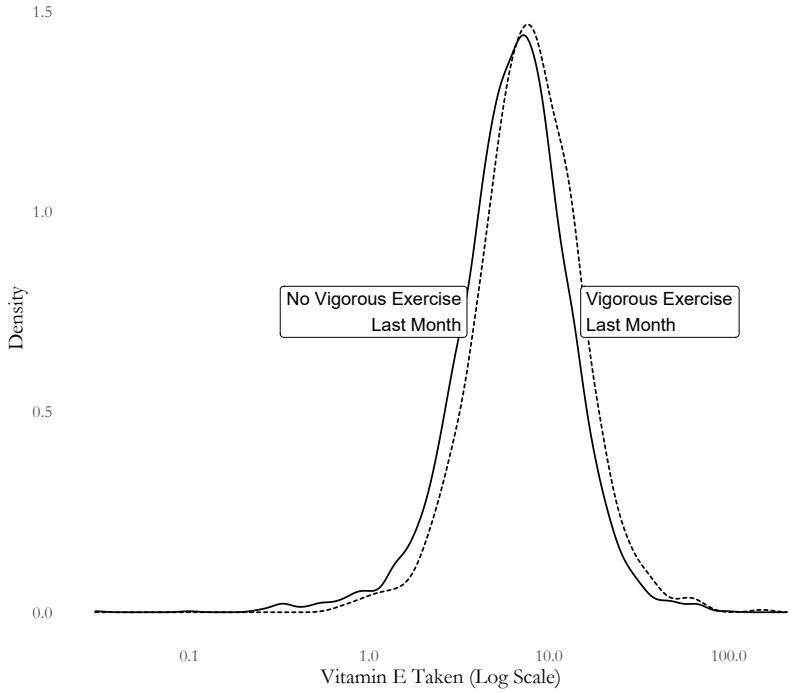


Figure 4.2: Distribution of Amount of Vitamin E Taken by Exercise Level

more likely to do other healthy things like exercise, because both are driven by how health-conscious you are.

Figure 4.3 shows an example of this. The distribution of whether you take Vitamin E or not is shown twice here, once for those who currently smoke, and one for those who don't smoke. The distributions are clearly different, with a higher proportion taking Vitamin E in the non-smoking crowd, exactly what Oster would expect.

Conditional Means

With the concept of a conditional distribution under our belt, it should be clear that we can then calculate *any* feature of that distribution conditional on the value of another variable. What's the 95th percentile of Vitamin E taking overall and for smokers? What's the median? What's the standard deviation of mortality for people who take 90th-percentile levels of Vitamin E, and for people who take 10th-percentile levels?

While all those possibilities remain floating in the air, we will focus on the conditional mean. Given a certain value of X, what do I expect the mean of Y to be?[7]

Once we have the conditional mean, we can describe the relation-

Conditional mean. The mean of one variable given that another variable takes a certain value.

[7] Why the mean? One reason is that the mean behaves a bit better in small samples, and once we start looking at things separately by specific values of X, samples get small. Another reason is that it helps us weight prediction errors and and so figure out how to minimize those errors. It's just handy.

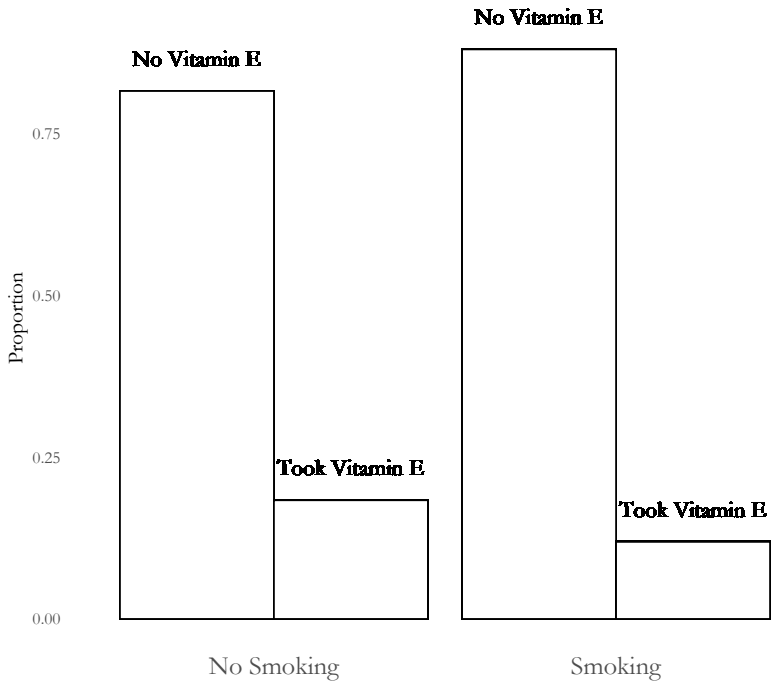


Figure 4.3: Distribution of Whether Vitamin E is Taken by Whether you Smoke

ship between the two variables fairly well. If the mean of Y is higher conditional on a higher value of X, then Y and X are positively related. Going further, we can map out all the conditional means of Y for each value of X, giving us the full picture on how the mean of one variable is related to the values of the other.

In some cases, this is easy to calculate. If the variable you are conditioning on is discrete (or categorical), you can just calculate the mean for all observations with that value. See Figure 4.4 for example, which shows the proportion taking Vitamin E conditional on whether the observations are from before Vitamin E was recommended, during recommendation, or after.[8] I just took all the observations in the data from before the recommendation and calculated the proportion who took Vitamin E. Then I did the same for the data during the recommendation, and after the recommendation.

Figure 4.4 shows the relationship between the taking of Vitamin E and the timing of the recommendation. We can see that the relationship between the taking of Vitamin E and the recommendation being in place is positive (the proportion taking Vitamin E is higher during the recommendation time). We also see that the relationship between Vitamin E and *time* is at first positive (increasing as the recommendation goes into effect) and then negative (decreasing as the

[8] "Proportion" is the mean of a binary variable.

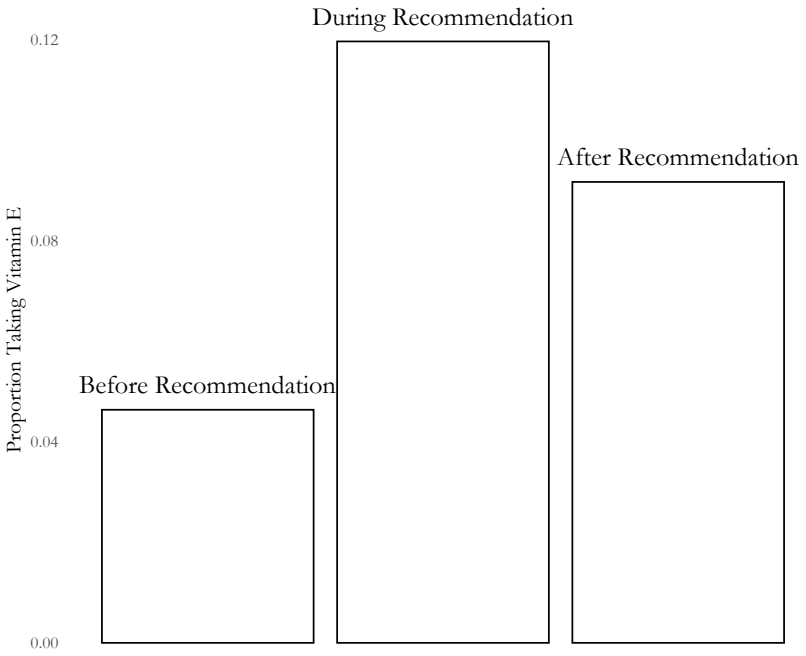


Figure 4.4: Proportion Taking Vitamin E Before it was Recommended, During, and After

recommendation is removed).

THINGS GET A LITTLE MORE COMPLEX when you are conditioning on a continuous variable. After all, I can't give you the proportion taking Vitamin E among those making $84,325 per year, because there's unlikely to even be more than one person with that exact number. And lots of numbers would have nobody at all to take the mean over!

There are two approaches we can take here. One approach is to use a *range* of values for the variable we're conditioning on rather than a single value. Another is to use some sort of shape or line to fill in those gaps with no observations.

Let's focus first on using a range of values. Table 4.1 shows the proportion of people taking Vitamin E conditional on Body Mass Index (BMI). Since BMI is continuous, I've cut it up into ten equally-sized ranges (bins), and calculated the proportion taking Vitamin E within each of those ranges. Cutting the data into bins to take a conditional mean isn't actually done that often in real research, but it gives a good intuitive sense of what we're trying to do.

Those same ranges can be graphed, as in Figure 4.5. The flat lines reflect that we are assigning the same mean to every observation in that range of BMI values. They show the mean conditional on being in that BMI bin. We see from this that BMI has a positive relation-

BMI Bin	Proportion Taking Vitamin E
(11.6,20.6]	0.133
(20.6,29.5]	0.159
(29.5,38.4]	0.171
(38.4,47.3]	0.178
(47.3,56.2]	0.203
(56.2,65.1]	0.243
(65.1,74]	0.067
(74,83]	0.143

Table 4.1: Proportion Taking Vitamin E by Range of Body Mass Index Values

ship with taking Vitamin E up until the 70+ ranges, at which point the conditional mean drops.

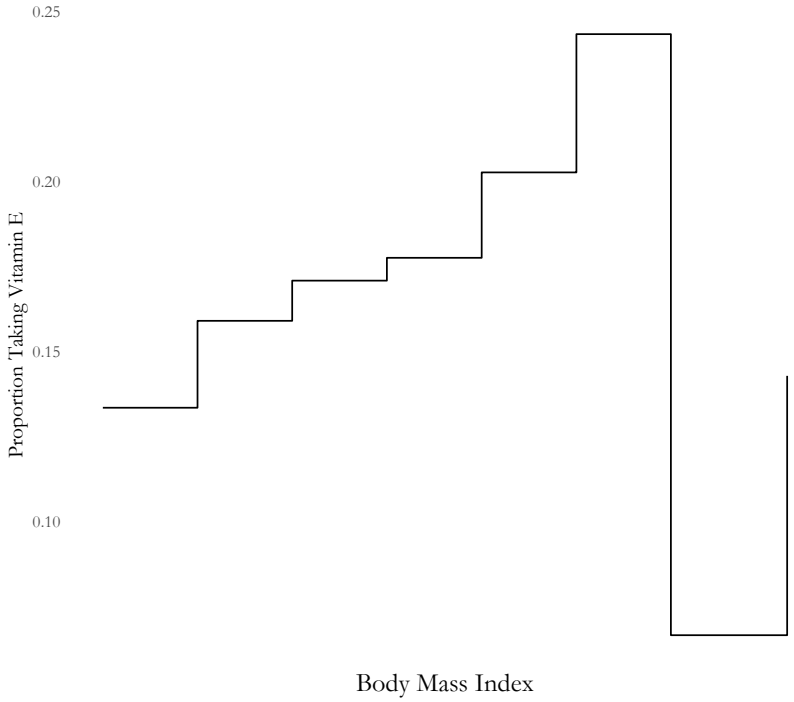


Figure 4.5: Proportion Taking Vitamin E by Range of Body Mass Index Values

Of course, while this approach is simple and illustrative, it's also fairly arbitrary. I picked the use of ten bins (as opposed to nine, or eleven, or...) out of nowhere. It's also arbitrary to use evenly-sized bins; no real reason I had to do that. Plus, it's rather choppy. Do I really think that if someone is at the very top end of their bin, they're more like someone at the bottom of their bin than like the person at the very bottom end of the next bin?

Instead, we can use a range of X values to get conditional means of

Y using *local means*. That is, to calculate the conditional mean of Y at a value of, say, $X = 2.5$, we take the mean of Y for all observations with X values *near* 2.5. There are different choices to make here—how close do you have to be? Do we count you equally if you're *very* close vs. *kind of* close?

A common way to do this kind of thing is with a LOESS curve, also known as LOWESS.[9] LOESS provides a local mean, weighting very-close observations more than kind-of close observations, and adding a little curvature so the end result is nice and smooth.

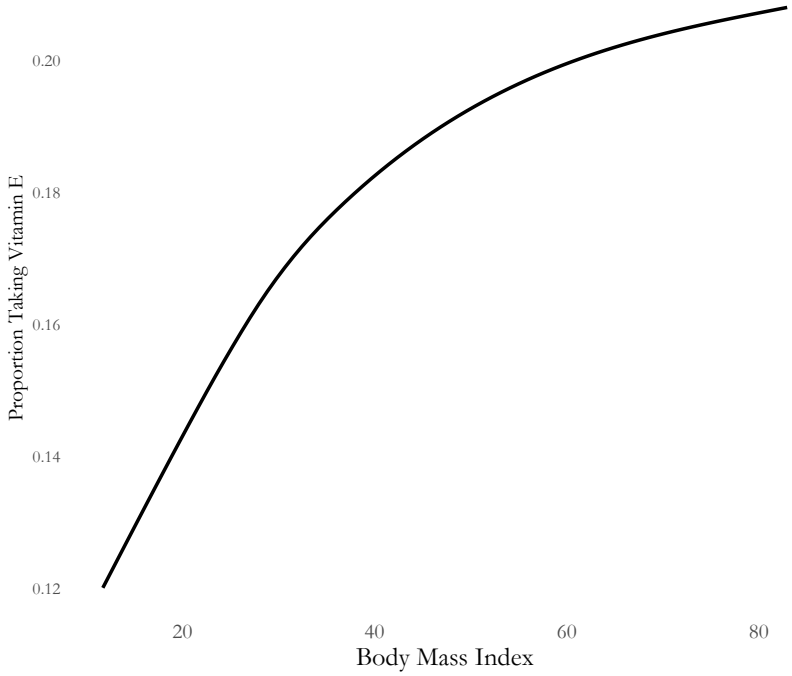
Figure 4.6 shows the LOESS curve for proportion taking Vitamin E and BMI.

Local mean. The mean of one variable calculated using only observations within a short range of another variable.

[9] "Locally Estimated Scatterplot Smoothing"

LOESS. A curve that uses local averages to smooth out the relationship between two variables.

Figure 4.6: Proportion Taking Vitamin E by BMI with a LOESS Curve



From Figure 4.6 we can see a clear relationship, with higher values of BMI being associated with more people taking Vitamin E. The relationship is very strong at first, but then flattens out a bit, although it remains positive.[10] It got there by just calculating the proportion of people taking Vitamin E among those who have BMIs in a certain range, with "a certain range" moving along to the right only a bit at a time while it constructed its conditional means.

[10] Why doesn't this dip down at the end like Figure 4.5? Basically, there are very very few observations in those really-high BMI bins. LOESS doesn't let that tiny number of observations pull it way down, and so sort of ignores them in a way that Figure 4.5 doesn't.

Line Fitting

Showing the mean of Y among local values of X is valuable, and can produce a highly detailed picture of the relationship between X and Y. But it also has limitations. There still might be gaps in your data it has trouble filling in, for one. Also, it can be hard sometimes to concisely describe the relationship you see.[11]

Enter the concept of *line-fitting,* also known as *regression.*[12]

Instead of thinking locally and producing estimates of the mean of Y conditional on values of X, we can assume that the underlying relationship between Y and X can be represented by some sort of *shape*. In basic forms of regression, that shape is a straight line. For example, the line

$$Y = 3 + 4X \tag{4.1}$$

tells us that the mean of Y conditional on, say, $X = 5$ is $3 + 4(5) = 23$. It also tells us that the mean of Y conditional on a given value of X would be 4 higher if you instead made it conditional on a value of X one unit higher!

In Figure 4.7 we repeat the Vitamin E/BMI relationship from before but now have a straight line fit to it. That particular straight line has a slope of 0.002, telling us that someone with a BMI one unit higher is .2% more likely to take a Vitamin E supplement than someone with a BMI one unit lower than them.

[11] Not to mention, it can be difficult, although certainly not impossible, to do what we do in the "Conditional Conditional Means" section with those methods.

[12] These two concepts are not the exact same thing, really. But they're close enough in most applications. Also, while I repeatedly mention conditional means in this section, there are versions of line-fitting that give conditional medians or percentiles or what-have-you as well.

Regression. The practice of fitting a shape, usually a line, to describe the relationship between two variables.

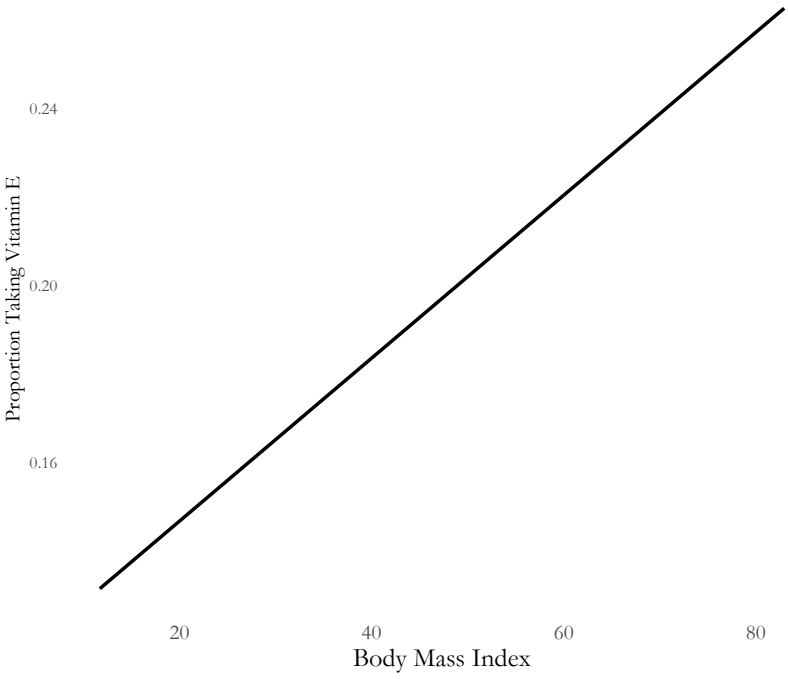


Figure 4.7: Proportion Taking Vitamin E by BMI with a Fitted Straight Line

This approach has some real benefits. For one, it gives us the conditional mean of Y for *any* value of X we can think of, even if we don't have data for that specific value.[13] Also, it lets us very cleanly describe the relationship between Y and X. If the slope coefficient on X (4 in that previous equation) is positive, then X and Y are positively related. If it's negative, they're negatively related.

[13] Although if we don't have data anywhere near that value, we probably shouldn't be trying to get the conditional mean there.

Those are pragmatic upsides for using a fitted line. There are more upsides in statistical terms in using a line-fitting procedure to estimate the relationship. Since the line is estimated using *all* the data, rather than just local data, the results are more precise. Also, the line can be easily extended to include more than one variable (more on that in the next section).

There is a downside as well, of course. The biggest downside is that fitting a line requires us to *fit a line.* We need to pick what kind of shape the relationship is—a straight line? A curved line? A line that wobbles up and down and up and down?—and then the line-fitting procedure picks the best version of that shape. But if the shape is all wrong, our estimate of the conditional mean will be all wrong. Imagine trying to describe the relationship in Figure 4.6 using a straight line!

The weakness here isn't necessarily that straight lines aren't always correct—line-fitting procedures will let us use curvy lines. But we have to be aware ahead of time that a curvy line is the right thing to use, and then pick which kind of curvy line it is ahead of time. But this does remain a weakness of line-fitting.

That weakness is, of course, set against the positives, which are strong enough that line-fitting is an extremely common practice across all applied statistical fields. So, then, how do we do it?

Ordinary Least Squares (OLS) is the most well-known application of line-fitting. OLS picks the line that gives the lowest *sum of squared residuals.* A residual is the difference between an observation's actual value and the conditional mean assigned by the line.[14]

Ordinary Least Squares. A regression method that uses a straight line and minimizes the sum of squared residuals.

[14] Or if you prefer, the difference between the actual value and the prediction.

Take that $Y = 3 + 4X$. We determined that the conditional mean of Y when $X = 5$ was $3 + 4(5) = 23$. But what if we see someone in the data with $X = 5$ and $Y = 25$? Well then their *residual* is $25 - 23 = 2$. OLS takes that number, squares it into a 4, then adds up all the predictions across all your data. Then it picks the values of β_0 and β_1 in the line $Y = \beta_0 + \beta_1 X$ that make that sum of squared residuals as small as possible, as in Figure 4.8.

How does it do this?[15] It takes advantage of information about how the two variables move together or apart, encoded in the *covariance*.

[15] Calculus, for one. But besides that.

If you recall the variance from Chapter 3, you'll remember that

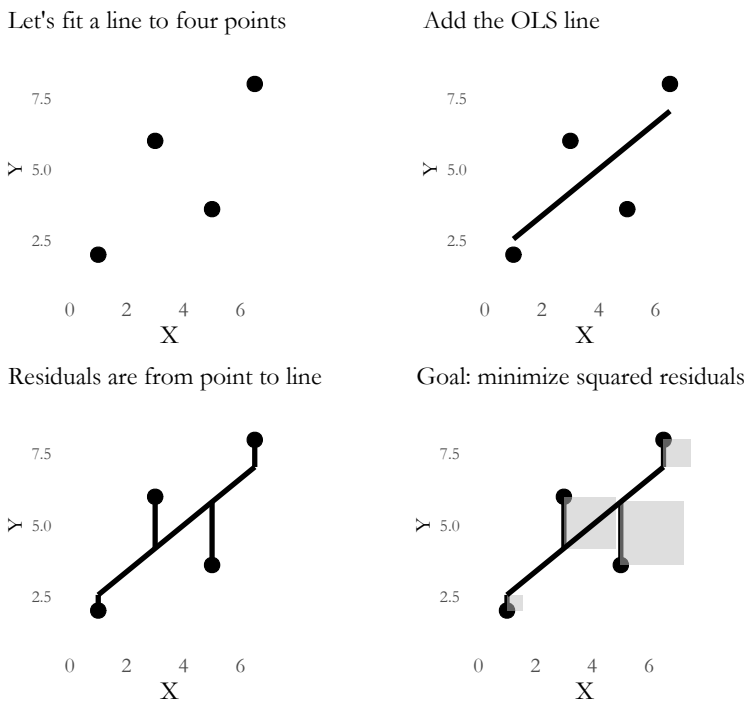


Figure 4.8: Fitting an OLS Line to Four Points

to calculate the variance of X, we: (a) subtracted the mean of X from X, (b) squared the result, (c) added up the result across all the observations, and (d) divided by the sample size minus one. The resulting variance shows how much a variable actually varies.

The covariance is the exact same thing, except that in step (a) you subtract the mean from *two* separate variables, and in step (b) you multiply the result from one variable by the result from the other. The resulting covariance shows how much two variables move together or apart. If they tend to be above average at the same time, then multiplying one by the other will produce a positive result for most observations, increasing the covariance. If they have nothing to do with each other, then multiplying one by the other will give a positive result about half the time and a negative result the other half, which will cancel out in step (c) and give you a covariance of 0.

How does OLS use covariance to get the relationship between Y and X? It just takes the covariance and divides it by the variance of X, i.e. $cov(X,Y)/var(X)$. That's it![16] This is roughly saying "of all the variation in X, how much of it varies along with Y?"[17] Then, once it has its slope, it picks an intercept for the line that makes the mean of the residuals (not the squared residuals) 0, i.e. the conditional mean is at least right on average.

The result from OLS is then a line with an intercept and a slope,

Covariance. A measurement of how much two variables vary with each other, as opposed to how much a single variable varies as in the variance. Technically the average of the summed products of de-meaned variables.

[16] For the two-variable version. We'll get to more complex ones in a bit.

[17] The sheer intuitive nature of this calculation might give a clue as to why we focus on minimizing the sum of squared residuals rather than, say, the residuals to the fourth power, or the product, or the sum of the absolute values. OLS gets some flak in some statistical circles for being restrictive, or for some of its assumptions. But the way that it seems to pop up everywhere and be linked to everything—it's the π of multivariate statistical methods if you ask me. I could write a whole extra chapter just on cool stuff going on under the hood of OLS. Look at me, starstruck over a ratio.

like $Y = 3 + 4X$. You can plug in a value of X to get the conditional mean of Y. And, crucially, you can describe the relationship between the variables using the slope. Since the line has $4X$ in it, we can say that a one-unit increase in X is associated with a 4-unit increase in Y.

Sometimes we may find it useful to rescale the OLS result. This brings us to the concept of *correlation*. Correlation, specifically Pearson's correlation coefficient, takes this exact concept and just rescales it, multiplying the OLS slope by the standard deviation of X and dividing it by the standard deviation of Y. This results in the covariance between X and Y divided by the standard deviation of X and the standard deviation of Y.

Correlation. A measurement of how two variables vary linearly together or apart, scaled to be between -1 and 1.

The correlation coefficient also relies on this concept of fitting a straight line. It just reports the result a little differently. We lose the ability to interpret the slope in terms of the units of X and Y.[18] However, we gain the ability to more easily tell how strong the relationship is. The correlation coefficient can only range from -1 to 1, and the interpretation is the same no matter what units the original variables were in. The closer to -1 it is, the more strongly the variables move in opposite directions (downward slope). The closer to 1 it is, the more strongly the variables move in the same direction (upward slope).

[18] Why? Well, the slope of a straight line tells you the change in units-of-Y-*per*-units-of-X. You can read that "per" as "divided by." When we multiply by the standard deviation of X, that's in units of X, so the units cancel out with the per-units-of-X, leaving us with just units-of-Y. Then when we divide by the standard deviation of Y, that's in units of Y, canceling out with units-of-Y and leaving us without any units!

How about for Vitamin E and BMI? OLS estimates the line

$$VitaminE = \beta_0 + \beta_1 BMI \tag{4.2}$$

and selects the best-fit values of β_1 and β_2 to give us

$$VitaminE = 0.110 + 0.002 BMI \tag{4.3}$$

So for a one-unit increase in BMI we'd expect a 0.002 increase in the conditional mean of Vitamin E. Since Vitamin E is a binary variable, we can think of a 0.002 increase in conditional mean as being a 0.2 percentage point increase in the proportion of people taking Vitamin E.

Then, since the standard deviation of taking Vitamin E is 0.369 and the standard deviation of BMI is 6.543, the Pearson correlation between the two is $0.002 \times 6.543/0.369 = 0.355$.

Regression slope coefficient. The linear relationship between two variables, estimated by regression. A one-unit change in one variable is associated with a (coefficient)-unit change in the other.

SOMETIMES BEING STRAIGHT IS INSUFFICIENT. OLS fits a straight line, but many sets of variables do not have a straight-line relationship! In fact, as shown in Figure 4.6, our Vitamin E/BMI relationship is one of them! What to do?

Two heroes come to our rescue.[19]

[19] These are two heroes that will not really receive the attention necessary in this book, which in general covers regression just enough to get to the research design. See a little more in 15, or check out a more dedicated book on regression like Bailey's Real Econometrics.

The first of them is apparently also the villain, OLS. Turns out OLS doesn't actually have to fit a *straight* line. Haha, gotcha. It just needs to fit a line that is "linear in the coefficients," meaning that the

slope coefficients don't have to do anything wilder than just being multiplied by a variable.

Asking it to estimate the β values in $Y = \beta_0 + \beta_1 X$ is fine, as before. But so is $Y = \beta_0 + \beta_1 X + \beta_2 X^2$—not a straight line! Or $Y = \beta_0 + \beta_1 ln(X)$—also not a straight line! And so on. What would be something that's *not* linear in coefficients? That would be something like $Y = \beta_0 + X_1^{\beta}$ or $Y = \frac{\beta_0}{1+\beta_1 X}$.

So that scary looking curved line in Figure 4.6? Not a problem, actually. As long as we take a look at our data beforehand to see what kind of shape makes sense (do we need a squared term for a parabola? Do we need a log term to rise quickly and then level out?), we can mimic that shape. For Figure 4.6 we could probably do with $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ to get the nice flexibility of the LOESS with the OLS bonuses of having fit a shape.

The second hero is "nonlinear regression" which can take many, many forms. Often it is of the form $Y = F(\beta_0 + \beta_1 X)$ where $F()$ is... some function, depending on what you're doing.

Nonlinear regression is commonly used when Y can only take a limited number of values. For example, we've been using all kinds of line-fitting approaches for the relationship between Vitamin E and BMI, but Vitamin E is *binary*—you take it or you don't! So a straight line like OLS will give us, or even a line that obeys the curve like $VitaminE = \beta_0 + \beta_1 BMI + \beta_2 BMI^2$ will be a bit misleading. Follow that line out far enough and eventually you'll predict that people with really high BMIs are more than 100% likely to use Vitamin E, and people with really low BMIs are less than 0% likely. Uh-oh.

You can solve this by using an $F()$ that doesn't go above 100% or below 0%, like a "probit" or "logit" function.

There are many other functions you could use, of course, for all kinds of different Y variables and the values they can take. I won't be spending much time on them in this book, but do be aware that they're out there, and they represent another important way of fitting a (non-straight) line.

Conditional Conditional Means, a.k.a. "Controlling for a Variable"

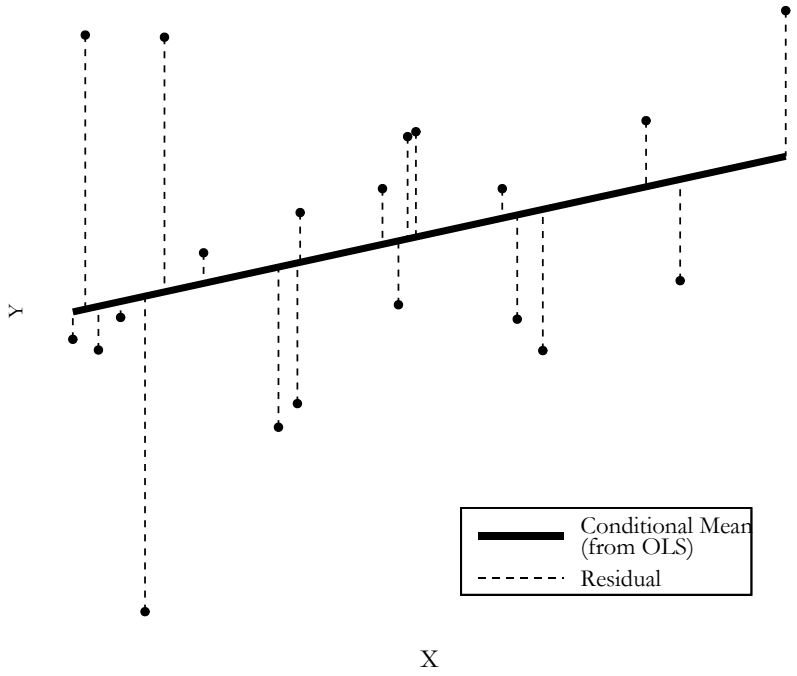
Let us enter the land of the unexplained. By which I mean the residuals.

When you get the mean of Y conditional on X, no matter how you actually do it, you're splitting each observation into two parts—the part *explained by* X (the conditional mean), and the part *not explained by* X (the residual). If the mean of Y conditional on $X = 5$

is 10, and we get an observation with $X = 5$ and $Y = 13$, then the prediction is 10 and the residual is $13 - 10$. Figure 4.9 shows how we can distinguish the conditional mean from the residual.

Residual. The difference between the actual and predicted values of an observation.

Figure 4.9: An OLS Line and its Residuals



It might seem like those residuals are just little nuisances or failures, the parts we couldn't predict. But it turns out there's a little magic in there. Because we can also think of the residual as *the part of Y that has nothing to do with X.* After all, if the conditional mean is 10 and the actual value is 13, then X can only be responsible for the 10. The extra 3 must be because of some other part of the data generating process.

Why would we want that? It turns out there are a number of uses for the residual. Just off the bat, perhaps we don't just want to know the variation in Vitamin E alone. Maybe what we want is to know how much variation there is in Vitamin E-taking that *isn't explained by BMI*. Looking at the residuals from Figure 4.7 would answer exactly that question!

But things get real interesting when we look at the residuals of two variables at once.

So what if we take the explained part out of two different variables? Let's expand our analysis to include a third variable. Let's keep it simple with Y, X, and Z. So, what do we do?[20]

[20] This particular set of calculations, when applied to linear regression, is known as the Frisch-Waugh-Lovell theorem and doesn't apply precisely to regression approaches that are nonlinear in parameters, like logit or probit as previously described. However, for those regressions the concept is still the same.

1. Get the mean of Y conditional on Z.
2. Subtract out that conditional mean to get the residual of Y. Call this Y^R.
3. Get the mean of X conditional on Z.
4. Subtract out that conditional mean to get the residual of X. Call this X^R.
5. Describe the relationship between Y^R and X^R.

Now, since Y^R and X^R have had the parts of Y and X that can be explained with Z *removed*, the relationship we see between Y^R and X^R is *the part of the relationship between Y and X that is not explained by Z.*

In other words, were getting the *Mean of Y conditional on X* all conditional on Z. We're *washing out the part of the X/Y relationship that is explained by Z.*

In doing this, we are taking out all the variation related to Z, in effect not allowing Z to vary. This is why we call this process "controlling for" Z (although "adjusting for" Z might be a little more accurate).

Controlling for a variable. Removing all the variation associated with that variable from all the other variables.

Let's take our ice cream and shorts example. We see that days where more people eat ice cream also tend to be days where more people wear shorts. But we also know that the temperature outside affects both of these things.

If we really want to know if ice-cream-eating affects shorts-wearing, we would want to know *how much of a relationship is there between ice cream and shorts that isn't explained by temperature?* So we would get the mean of ice cream conditional on temperature, and then take the residual, getting only the variation in ice cream that has nothing to do with temperature. Then we would take the mean of shorts-wearing conditional on temperature, and take the residual, getting only the variation in shorts wearing that has nothing to do with temperature. Finally, we get the mean of the shorts wearing residual conditional on the ice cream residual. If the shorts mean doesn't change much conditional on different values of ice cream, then the entire relationship was just explained by heat! If there's still a strong relationship there, maybe we do have something.

THE EASIEST WAY TO DO THIS IS WITH REGRESSION. Regression allows us to control for a variable by simply adding it to the equation. Now we have "multivariate" regression. So instead of

$$Y = \beta_0 + \beta_1 X \tag{4.4}$$

we instead just give it

$$Y = \beta_0 + \beta_1 X + \beta_2 Z \tag{4.5}$$

and voila, the OLS estimate for β_1 will automatically go through the steps of removing the conditional means and analyzing the relationship between Y^R and X^R.

Even better, we can do things conditional on *more than one variable*. So we could add W and do...

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W \tag{4.6}$$

and now the β_1 that OLS picks will give us the relationship between Y and X conditional on *both* Z and W.

Let's take a quick look at how this might affect our Vitamin E/BMI relationship. Some variables that might be related to both taking Vitamin E and to BMI are gender and age. So let's add those two variables to our regression and see what we get.

Before, with only BMI, we estimated

$$VitaminE = 0.110 + 0.002BMI \tag{4.7}$$

Now, with BMI, gender, and age, we get

$$VitaminE = -0.006 + 0.001BMI + 0.002Age + 0.016Female \tag{4.8}$$

The effect of BMI has changed a bit, from 0.002 to 0.001, telling us that some of the relationship we saw between BMI and Vitamin E was explained by Age and/or gender. We also see that older people are more likely to take Vitamin E—for each additional year of age we expect the proportion taking Vitamin E to go up by .2 percentage points. Women are also more likely than men to take the supplement. A one-unit increase in "Female" (i.e. going from 0—a man—to 1—a woman) is associated with an increased proportion taking Vitamin E of 1.6 percentage points.[21]

[21] Of course, OLS by itself doesn't know *which* variable is the treatment. So the Age effect "controls for" BMI and Female, and the Female effect controls for BMI and Age. However, we don't want to get too wrapped up in interpreting the coefficients on controls, as we generally haven't put much work into identifying their effects (see Chapter 5).

So how does regression do this? Put your mental-visualization glasses on.

One way is mathematically. If you happen to know a little linear algebra (and if you don't, you can skip straight to the next paragraph), the formula for multivariate OLS is $(A'A)^{-1}A'Y$, where A is a matrix of all the variables other than Y, including the X we're interested in. In other words, it washes out the influence of all the non-X variables on the X/Y relationship by dividing out a bunch of covariances.

Another way is graphically. If you can think of a two-variable OLS line $Y = \beta_0 + \beta_1 X$ as being a line, you can think of a three-variable

OLS line as a *plane* in 3-D space (or with four variables, in 4-D space, and so on). We can visualize this by looking at each of the three sides of that 3-D image one at a time.

Figure 4.10 shows the X-Y axis on the top-left. Then to the right you can see the Z-Y axis, and below the Z-X axis. The coordinates are flipped on the Z-X axis—even though we're getting the mean of X conditional on Z here, I've put X on the x-axis to be consistent with the X-Y graph. The upward slope on the Z-Y and Z-X axes shows that Z is explaining part of both X and Y, and that we could take that explanation out to focus on the residuals.

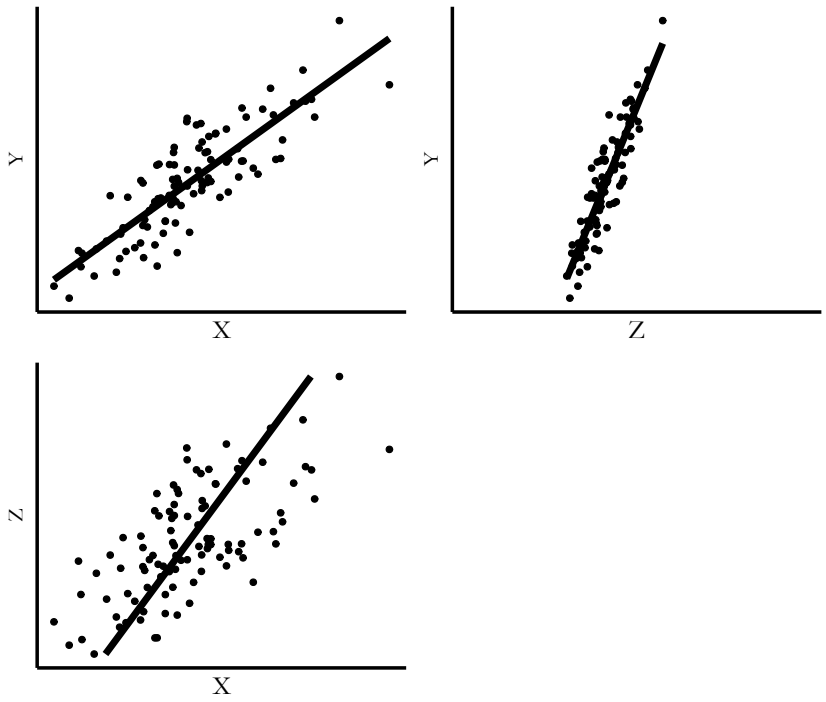


Figure 4.10: A Three-Variable Regression from All Three Dimensions

Then, in Figure 4.11, we flatten out those explanations. The upward slopes get flattened out, moving the X and Y points with them. You can see how subtracting out the parts explained by Z literally leave the X/Y relationship no part of Z to hold on to! Z has been flatlined in both directions, providing no additional "lift" to the points in the X/Y graph. What's still there on X/Y is there without Z.

What We're Not Covering

In the previous chapter, on describing variables, we did a pretty good job covering a lot of what you'd want to know when describing a variable. This chapter, however, leaves out a whole lot more.

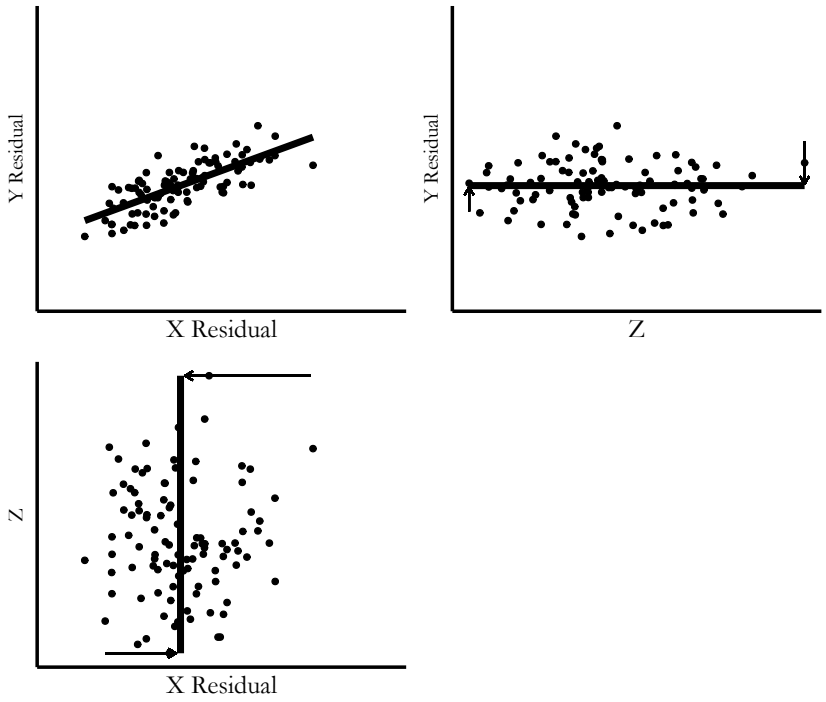


Figure 4.11: A Three-Variable Regression from All Three Dimensions After Removing the Variation Explained by Z

This is largely for reasons of focus. This book is about research design. Once you've got research design pinned down, there are certainly a lot of statistical issues you need to deal with at that point. But things like specific probability distributions (normal vs. log-normal vs. t vs. Poisson vs. a million others we didn't cover), functional form (OLS vs. probit/logit vs. many others), or standard errors and hypothesis testing can be a distraction when thinking about the broad strokes of how you're going to answer your research question.

In one case, omission is less for focus and more to cover it more appropriately later. Notice how I introduced the Oster paper as being all about how the relationship between Vitamin E and health indicators changed over time... but then I never showed how the BMI relationship changed over time? There are a number of research designs that have to do with *how a relationship changes* in different settings.[22] However, a proper treatment of this will have to wait until Part II of the book.

To be clear, you want to know all this stuff. And I will cover it more in this book in Chapters 13, 15, and many of the other Part II chapters. You can also check out a more traditional econometrics book like Bailey's Real Econometrics or Wooldridge's Introductory Econometrics. But for observational data, most of the time these are things to consider *after* you have your design and plan to take that

[22] Controlling for time would not achieve this. Controlling for time would remove the part of the relationship explained by time, but would not show how the relationship changes over time.

design to actual data.

For now, I want you think about *what you want to do* with your data—what kinds of descriptions of variables your research design requires, what kinds of relationships, what kinds of conditional means and conditional conditional means. Figure out how you want your data to *move*. Figure out the journey you're going to take first; you can pack your bags when it's actually time to leave.

Relationships in Software

In this section I'll show you how to calculate or graph the relationship between variables in three different languages: R, Stata, and Python.

The Oster data, while free to download, would require special permissions to redistribute. Instead, I will be using data from Mroz[23], which is a data set of women's labor force participation and earnings from 1975.

[23] Thomas A Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, pages 765–799, 1987

In each of these languages, I'm going to:

1. Load in the data
2. Draw a scatterplot between log women's earnings and log other earnings in the household, among women who work
3. Get the conditional mean of women's earnings by whether they attended college
4. Get the conditional mean of women's earnings by different bins of other household earnings
5. Draw the LOESS and linear regression curves of the mean of log women's earnings conditional on the log amount of other earnings in the household
6. Run a linear regression of log women's earnings on log other earnings in the household, by itself and including controls for college attendance and the number of children under five in the household

XXX MOVE DATA ONLINE

```r
# R CODE
library(tidyverse); library(jtools)

df <- read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/carData/Mroz.csv') %>%
    # Keep just working women
    filter(lfp == 'yes') %>%
    # Get unlogged earnings %>%
    mutate(earn = exp(lwg))

# 1. Draw a scatterplot
ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
```

```r
    # Use a log scale for both axes
    scale_x_log10() + scale_y_log10()

# 2. Get the conditional mean by college attendance
df %>%
    # wc is the college variable
    group_by(wc) %>%
    # Functions besides mean could be used here to get other conditionals
    summarize(earn = mean(earn))

# 3. Get the conditional mean by bins
df %>%
    # use cut() to cut the variable into 10 bins
    mutate(inc_cut = cut(inc, 10)) %>%
    group_by(inc_cut) %>%
    summarize(earn = mean(earn))

# 4. Draw the LOESS and linear regression curves
ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
    # geom_smooth by default draws a LOESS; we don't want standard errors
    geom_smooth(se = FALSE) +
    scale_x_log10() + scale_y_log10()
# Linear regression needs a 'lm' method
ggplot(df, aes(x = inc, y = earn)) +
    geom_point() +
    geom_smooth(method = 'lm', se = FALSE) +
    scale_x_log10() + scale_y_log10()

# 5. Run a linear regression, by itself and including controls
model1 <- lm(lwg ~ log(inc), data = df)
# k5 is number of kids under 5 in the house
model2 <- lm(lwg ~ log(inc) + wc + k5, data = df)
# And make a nice table
export_summs(model1, model2)
```

```
* STATA CODE
import delimited using "https://vincentarelbundock.github.io/Rdatasets/csv/carData/Mroz.csv"
* Keep just working women
keep if lfp == "yes"
* Get unlogged earnings
g earn = exp(lwg)
* Drop negative other earnings
drop if inc < 0

* 1. Draw a scatterplot
twoway scatter inc earn, yscale(log) xscale(log)

* 2. Get the conditional mean college attendance
table wc, c(mean earn)

* 3. Get the conditional mean by bins
* Create the cut variable with ten groupings
egen inc_cut = cut(inc), group(10) label
table inc_cut, c(mean earn)

* 4. Draw the LOESS and linear regression curves
* Create the logs manually for the fitted lines
g loginc = log(inc)
twoway scatter loginc lwg || lowess loginc lwg
twoway scatter loginc lwg || lfit loginc lwg
```

```
* 5. Run a linear regression, by itself and including controls
reg lwg loginc
* college needs to be turned into a numeric variable
g college = wc == "yes"
reg lwg loginc college k5
```

```
# PYTHON CODE
import pandas as pd
import numpy as np
import statsmodels.formula.api as sm
import matplotlib.pyplot as plt
import seaborn as sns

# Read in data
dt = pd.read_csv('https://vincentarelbundock.github.io/Rdatasets/csv/carData/Mroz.csv')
# Keep just working women
dt = dt[dt['lfp'] == 'yes']
# Create unlogged earnings
dt.loc[:,'earn'] = dt['lwg'].apply('exp')

# 1. Draw a scatterplot
sns.scatterplot(x = 'inc',
y = 'earn',
data = dt).set(xscale="log", yscale="log")
# The .set() gives us log scale axes

# 2. Get the conditional mean by college attendance
# wc is the college variable
dt.groupby('wc')[['earn']].mean()


# 3. Get the conditional mean by bins
# Use cut to get 10 bins
dt.loc[:, 'inc_bin'] = pd.cut(dt['inc'],10)
dt.groupby('inc_bin')[['earn']].mean()

# 4. Draw the LOESS and linear regression curves
# Do log beforehand for these axes
dt.loc[:,'linc'] = dt['inc'].apply('log')
sns.regplot(x = 'linc',
            y = 'lwg',
            data = dt,
            lowess = True)
sns.regplot(x = 'linc',
            y = 'lwg',
            data = dt,
            ci = None)


# 5. Run a linear regression, by itself and including controls
m1 = sm.ols(formula = 'lwg ~ linc', data = dt).fit()
print(m1.summary())
# k5 is number of kids under 5 in the house
m2 = sm.ols(formula = 'lwg ~ linc + wc + k5', data = dt).fit()
print(m2.summary())
```

Chapter Problems

1. What is a conditional distribution?

2. Figure 4.12 shows the relationship between Income level and Depression from a sample of participants.

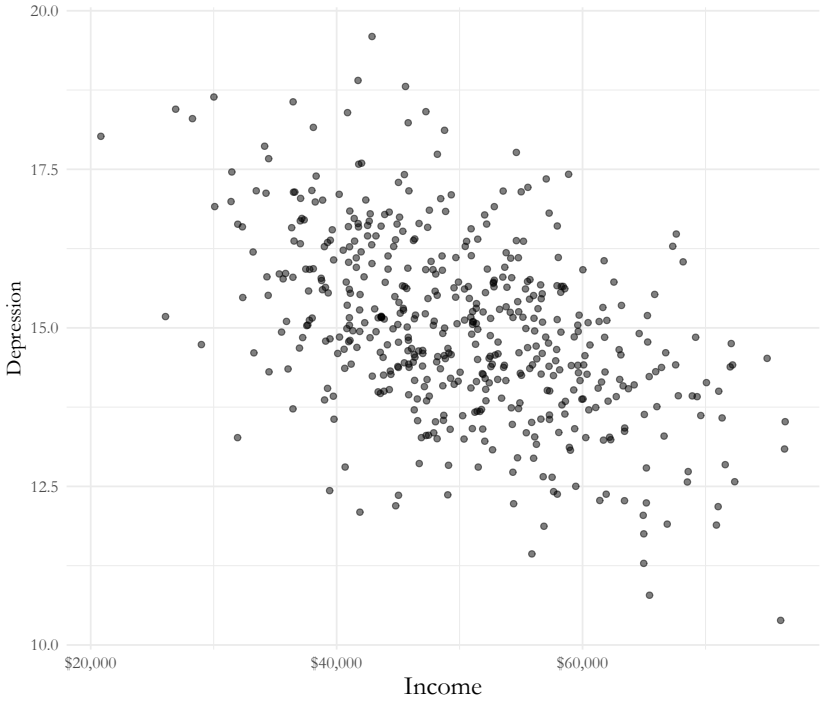


Figure 4.12: A Simulated Relationship between Income and Depression

   (a) How does the conditional mean of Depression change as the value of Income changes?

   (b) Does the graph indicate that lower income causes depression? Why or why not?

3. Table 4.2 depicts data collected from 3000 university students on their classification (Freshman, Sophomore, Junior, Senior) and whether or not they receive financial aid. The table shows a cross tabulation of classification and receipt of financial aid.

Financial Aid	Freshman	Sophomore	Junior	Senior
Y	508	349	425	288
N	371	337	384	338

Table 4.2: Financial Aid by Student Standing

   (a) From this data, calculate the probability of receiving financial aid.

(b) Calculate the probability of being a Freshman.

(c) Calculate the probability of receiving financial aid given that a student is a Senior.

(d) Calculate the probability of receiving financial aid given that a student is a Freshman.

4. Describe the advantages and disadvantages of using line-fitting methods as opposed to calculating local means.

5. Consider the line of best fit: $Y = 4 - 3.5X$.

(a) What is the conditional mean of Y when $X = 6$?

(b) What is the conditional mean of Y when $X = -2$?

6. Which of the following terms describes a measurement of how much two variables vary with each other?

(a) Variance

(b) Conditional mean

(c) Covariance

(d) Local mean

7. What is the difference between covariance and correlation?

8. Figure 4.13 Plot A and Plot B below depict the relationship between scores on a math exam and an intelligence measure from data collected from a fictional sample of 100 students.

(a) What type of line/curve is shown in Plot A and Plot B? Which one is more accurate? Why?

(b) For Plot A, describe the residuals for different ranges of math exam scores. Are the observed data below or above the line/curve? Do the residuals have negative or positive sign? Are the residuals evenly scattered around the line/curve?

9. Table 4.3 is data collected on 5 employees from a company on how well they get along with their coworkers (GetAlong) and their level of job satisfaction (Satisfaction). The Prediction variable is the predicted satisfaction level, or the conditional mean of satisfaction, for each value of GetAlong derived after fitting a line of best fit using ordinary least squares estimation.

(a) Calculate the residuals.

(b) Describe how ordinary least squares estimation uses the residuals when fitting a line.

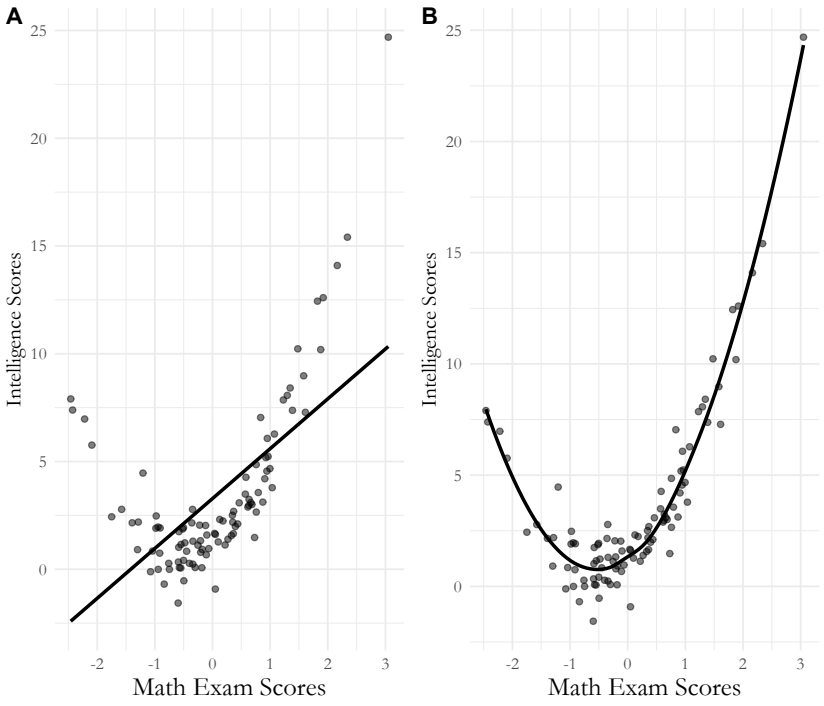


Figure 4.13: Scores and Intelligence with Two Different Lines

GetAlong	Satisfaction	Prediction
4.64	5.19	5.1
5.94	5.36	5.22
4.05	4.9	5.05
5.11	5.65	5.15
5.51	4.61	5.19

Table 4.3: Satisfaction and Getting Along with Coworkers

10. Describe the process of controlling for a variable. Consider the example: What is the relationship between first generation status and graduation rate in a population of college students?

    (a) What is a residual?

    (b) What does it mean to control for a variable? What role do residuals play in the process of controlling for a variable?

    (c) In the above example of the relationship between first generation status and graduation rate, how would our interpretation of the correlation between first generation status and graduation rate change if we controlled for race, gender and socio-economic status of the students?