2002.5 Regression - Covariates

Intro To Covariates

Why have more than one variable? In general, it adds to the flexibility of an estimator. One important case is that of *controlling* for things.

Confounders

Baby's first causal inference (it's me, I'm the baby)

Recall from 2001.14 Research Design the possibility of imbalanced treatments, or confounders. For this purpose, let's say a confounder is anything that affects both the probability of being treated and the outcome variable. This will interfere with our estimation of the treatment.

Consider the classic example of estimating the differential effect of pipe and cigarette smoking by comparing mortality in the two populations. The cigarette smokers were found to have less mortality than the pipe smokers, despite cigarettes being more addictive and nicotine-intensive.

The problem is that being old affects both mortality and the choice of pipe vs cigarettes. So the estimator above is picking up a mix of cigarette vs pipe effects and age effects.

For a more rigorous definition of confounders see Confounding. For an intuitive graphics-based, see 🚧DAGs 101.

Controlling for Confounders

Covariates and Controls

So how to deal with confounders? There are, of course, a lot of ways. In the regression context, we can think about adding a covariate. Consider the cigarette-pipe example below. We can estimate this model:

$$\text{Mortality} = \hat{\beta}_0 + \hat{\beta}_1[\text{Cigarette Smoker}] + \hat{u} \tag{1}$$

I'd like to note that this is going to indicate the same $\hat{\beta}_1$ as an uncontrolled difference-in-means estimator. Given that our sample is entirely smokers of some sort, $\hat{\beta}_0$ will pick up the general mortality effect of smoking anything at all; $\hat{\beta}_1$ will pick up the effects of cigarette smoking - including the degree to which cigarette smoking is related to age, which effects mortality. So our coefficient will be biased downward.

We can deal with this by adding a covariate: age. We can call this 'controlling for age'

$$\text{Mortality} = \hat{\beta}_0 + \hat{\beta}_1[\text{Cigarette Smoker}] + \hat{\beta}_2[\text{Age}] + \hat{u} \tag{2}$$

This will help us get rid of the Omitted Variable Bias. Having done so, we will correctly recover the increased mortality associated with cigarette smoking (unless, of course, we're Don Rubin).

The same logic applies to adding a quadratic term - except in this case, it's not age you're missing, but X^2.

Practically, you can also try and convince people that whatever potential confounders they are proposing don't matter, or, if they do matter, point in a direction favorable to your conclusion. Make sure not to give talks right before lunch.

Covariates, graphically

In (1), we're making a single line representing the slope and intercept connecting Mortality and Cigarette Smoking (pretend it's continuous). This will be a line through a series of points in the $\text{Mortality} \times \text{Cigarette Smoking}$ plane, each point being one observation with its own values of Mortality and Cigarette Smoking.

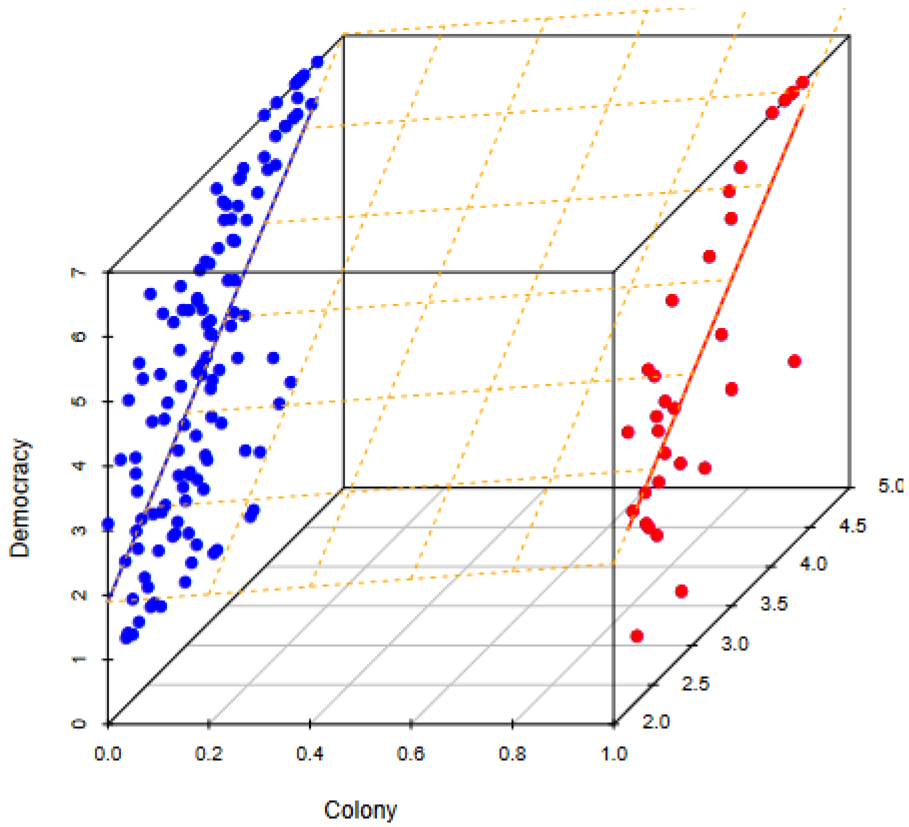
But then we realize this doesn't make sense, and we want to account for age. How can we do that?

Let's fix the graph first. This time we have three coordinates per observation - $\mathrm{Mortality}$, $\mathrm{Cigarette\ Smoking}$ and Age. So now we need three dimensions to represent each observation fully. Think of a series of $\mathrm{Mortality} \times \mathrm{Cigarette\ Smoking}$ planes, each having a different value of Age. We can glue these planes to each other (a technical term) to make a 3-dimensional space of $\mathrm{Mortality} \times \mathrm{Cigarette\ Smoking} \times \mathrm{Age}$.

So, how do I start drawing a regression? For simplicity, let's assume that age doesn't change the marginal impact of smoking - but aging makes people more mortal. So, thinking of the many $\mathrm{Mortality} \times \mathrm{Cigarette\ Smoking}$ planes that make up our 3D plane, each of them has a different intercept for its value of Age. So we have a large number of different regression lines, all parallel, each higher than the last.

Now, think of the $\mathrm{Age} \times \mathrm{Mortality}$ planes we are implicitly making. These have the same property! Each line connecting Age and $\mathrm{Mortality}$ is parallel to all the others. But, given a certain age, smoking more cigarettes increases your $\mathrm{Mortality}$. So, instead of a single regression line, we have a plane that's increasing linearly in each covariate.

Here's a simple example from class, with one covariate being binary (a dummy).



We can think of 'iso-Mortality lines' in the $\text{Age} \times \text{Smoking}$ plane, defining exchange rates. So we can say that 10 cigarettes a day creates as much mortality as being 3 years older than the baseline. You can also attain the same level of mortality with 5 cigarettes a day, and being 1.5 years older.

Visually, I like to think of two triangles, set upright, at right angles, with their points touching. The hypotenuses have slopes $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively - the same as each cross-section paralle to one of the axes (representing a *ceteris paribus* estimate). If you were to run a number of strings between those triangles - the iso-mortality lines- and combine them into a sheet, you'd get this regression surface. These triangles are the marginal conditional expectations, while the surface is the joint.

🚧The bivariate estimator

We get the estimator the same way as before. We want to minimize $Y - \beta_0 - X_1\beta_1 - X_2\beta_2$. We take partial derivatives, get FOCs, and solve.

The estimator comes out as follows (I'm replacing $\text{Var}(a)$ with $\text{Cov}(a, a)$ because I think it's clarifying

$$
\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 \\
\hat{\beta}_1 &= \frac{\mathrm{Cov}(x_1, y)\mathrm{Var}(x_2) - \mathrm{Cov}(x_2, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \mathrm{Cov}(x_1, x_2)^2} \\
\hat{\beta}_2 &= \frac{\mathrm{Cov}(x_2, y)\mathrm{Var}(x_1) - \mathrm{Cov}(x_1, y)\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)\mathrm{Var}(x_2) - \mathrm{Cov}(x_1, x_2)^2}
\end{aligned}
$$

🚧What the hell is this? I don't know yet

First, remember the univariate formula:

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(x_1, y)}{\mathrm{Var}(x_1)}$$

$\mathrm{Cov}(x_1, y)$ represents the average size of the joint variations from their means. So that's what we want! But we want to normalize it to a slope coefficient, to represent changes in y per unit x_1. So we divide by $\mathrm{Var}(x_1)$, which is dividing by x_1's variation from its mean - twice.

🚧Finding the Bivariate Estimator with Matrix Algebra

Partialling Out

It can be shown (maybe I will sometime) that running a regression of y on x_1, x_2 gives the same coefficient on x_2 as if you ran a regression of x_2 on x_1, then regressed y on the residuals.

In other words, once you take the x_1 'part' out of x_2, you can look at the relationship between what's left and y to get the x_1 'part' of y the same way you would in a regular regression.

🚧I think this is about linear algebra so I will come back to it later.

Multicollinearity

Once we add more covariates, one can begin to worry about multicollinearity - the degree to which one covariate is a linear combination of others. If that's absolutely true, then the coefficients cannot be found (consider what would happen to the denominator if $\mathrm{Var}(x_1) = \mathrm{Cov}(x_1, x_2)$).

But if it's not strictly true? Then we can generate estimates. But the regression will struggle to differentiate between the factors. Might kill your standard errors for variables involved.

Substantively, I'd be interested to hear about cases where x_1 and x_2 are very closely related but I want to disambiguate them. I suspect many times when this is the case, I need to apply theory or research design, rather than econometric tools.

Flavors of Covariate

Dummies

A discrete-valued/indicator covariate. If dummy covariate d_i can take on k values, then we will need to create $k-1$ covariates. The k^{th} is in the base case - consider the smoking example, in which the effect of pipe smoking was hidden in the intercept, and the differential effect of smoking cigarettes was a coefficient.

I can use these to control for specific things - usually some sort of fixed effect. If I'm regressing foreign aid on GDP flows, and I think that former colonial countries get more aid, I can have a colonial dummy. If I am regressing income on kindergarten class size, and I think years of primary/secondary education has a separable effect, I can create a 12-valued dummy variable for years of education. Note that, in this case, I have them all in one dummy because I expect a linear effect. If I think that the gap between 9th and 10th grade is much more important than that between 3rd and 4th, then I would want these in different dummies.

Interaction Terms

The approach in the graphical section above makes it a lot easier for me to think about interaction terms. Suppose our model to be:

$$\begin{gathered}\text{Mortality} = \hat{\beta}_0 + \hat{\beta}_1[\text{Cigarette Smoker}] + \hat{\beta}_2[\text{Age}] + \\ \hat{\beta}_3[\text{Age} \times \text{Cigarette Smoker}] + \hat{u} \qquad \text{(2a)}\end{gathered}$$

Interpret this as saying that, if you're older, each cigarette hurts you more. Our isomortality lines are now curves. Say as before, that 10 cigarettes a day creates as much mortality as 3 years of age above baseline. But now, if I smoke 5 cigarettes and age 1.5 years, I'll be worse off than that previous level, because the age and the cigarettes make each other worse! So our isomortality curves are convex.

Visually, imagine the same two triangles. Set another one between the first two at an angle, with a slope steeper than the other two. Then draw lines between the Age triangle and the middle triangle, and the middle and the Smoking triangle. This is a linear approximation to the humped plane above.

The triangles still have hypotenuses with slopes $\hat{\beta}_1$ and $\hat{\beta}_2$. But now each cross-section parallel to a triangle has a different slope. The differences in the slopes come from the interaction term.

Polynomial Terms

Suppose I think Y increases with X. Not only that, but the marginal effect of X on Y also increases with X. I can capture this with a model like this:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

This really is just another sort of interaction term. The marginal effect of X increases as X increases. Think of it as fitting a polynomial with least-squares rather than a line.