Gov 2000: 7. What is Regression?

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

October 15, 2016

Where are we? Where are we going?

- What we've been up to: estimating parameters of population distributions. Generally we've been learning about a single variable.
- This week and for the rest of the term, we'll be interested in the relationships between variables. How does one variable change we change the values of another variable? These will be the bread and butter of the class moving forward.

RELATIONSHIPS BETWEEN TWO VARIABLES

What is a relationship and why do we care?

Most of what we want to do in the social science is learn about how two variables are related. For example, we might have the following questions about relationships between different variables:

- Does turnout vary by types of mailers received?
- Is the quality of political institutions related to average incomes?
- Does conflict mediation help reduce civil conflict?

Notation and conventions

We'll use the following notational conventions:

- Y_i - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Log GDP per capita
  - Number of battle deaths
- X_i - the independent variable or covariate or explanatory variable or regressor or right-hand-side variable or treatment or predictor
  - Social pressure mailer versus Civic Duty Mailer
  - Average Expropriation Risk
  - Presence of conflict mediation

Generally our goal with regression is to understand how Y_i varies as a function of X_i:

$$Y_i = f(X_i) + \text{error}$$

Population-first approach

Before we learn too much about how to run a regression, it's first good to understand what we are estimating. In the last few weeks, we've been focusing on estimating the true difference in means between two populations (in the case of the social pressure experiment, this was the treatment effect). We have also thought about estimating the population mean of some random variable from samples of that random variable.

Regression works generally the same way. There is some **population relationship** between two variable and we are going to use sample data to estimate it. Before we can start building estimators and getting our estimates, though, it's good to understand the thing that we are estimating. That's why we'll focus on thinking about the population version of regression to start.

CONDITIONAL EXPECTATION

When we turn to predicting an outcome or estimating the effect of a covariate on an outcome, one way to describe relationships takes center stage: the conditional expectation function.

Definition 1. The **conditional expectation function** (CEF) or the **regression function** of Y given X, denoted $\mu(x) = \mathbb{E}[Y|X = x]$ is the function that gives the mean of Y at various values of x.

Note that this is a function of the *population* distributions. Regression at its most fundamental is about how the mean of Y changes as a function of X.

Discrete covariates

It's easiest to think about the CEF when X_i is discrete. Let's pin down a specific example. Suppose that you were interested in exploring the relationship between race and how long a person waits in line to vote. Let Y_i be the amount of time, in minutes, that a person waits in line to vote. And we'll code X_i as a binary measure of race/ethnicity with $X_i = 1$ for a white respondent and $X_i = 0$ for a non-white respondent. With this, there only exists two conditional expectations, which we can write with words to make things more clear:

$$\begin{aligned}\mu(\text{white}) &= E[Y_i|X_i = \text{white}] \\ \mu(\text{non-white}) &= E[Y_i|X_i = \text{non-white}]\end{aligned}$$

The first value here is the (population) average wait time for the sub-population of whites, whereas the second is the (population) average wait time for the sub-population of non-whites. We might look at the difference between these values $\mu(\text{white}) - \mu(\text{non-white})$ as measure of relationship between race and voter wait times. Notice here that since X_i can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.

Imagine now we broke up the race/ethnicity variance into a couple of categories, such as $X_i \in \{\text{white, black, hispanic, asian, other}\}$. Then we would have a different CEF:

$$\begin{aligned}\mu(\text{white}) &= E[Y_i|X_i = \text{white}] \\ \mu(\text{black}) &= E[Y_i|X_i = \text{black}] \\ \mu(\text{hispanic}) &= E[Y_i|X_i = \text{hispanic}] \\ \mu(\text{asian}) &= E[Y_i|X_i = \text{asian}] \\ \mu(\text{other}) &= E[Y_i|X_i = \text{other}]\end{aligned}$$

Moving away from categorical X_i values, we could also imagine a discrete variable such as $X_i =$ the number of polling booths at respondent i's polling place. Then, $\mu(x)$ is the average wait times for respondents in polling stations with x polling booths.

Why does the CEF measure the relationship between variables? Imagine that the CEF was constant in x, so that $\mu(\text{white}) = \mu(\text{non-white})$. This would mean that wait times are on average the same for whites and non-whites, which seems to indicate a

lack of a relationship. On the other hand, if there is a big difference, this implies that race/ethnicity can help explain or predict wait times.

CEFs with multiple covariates

In the next few weeks, we'll often want to explore the CEF conditioning on multiple variables. For instance, we might want to look at the CEF of wait times conditional on race, X_i, and gender, Z_i:

$$\begin{aligned}
\mu(\text{white, man}) &= \mathbb{E}[Y_i|X_i = \text{white}, Z_i = \text{man}] \\
\mu(\text{white, woman}) &= \mathbb{E}[Y_i|X_i = \text{white}, Z_i = \text{woman}] \\
\mu(\text{non-white, man}) &= \mathbb{E}[Y_i|X_i = \text{non-white}, Z_i = \text{man}] \\
\mu(\text{non-white, woman}) &= \mathbb{E}[Y_i|X_i = \text{non-white}, Z_i = \text{woman}]
\end{aligned}$$

Now, the CEF is a function of two variables, $\mu(x, z)$. With this CEF, we might be interested in the difference between in average wait times between white and non-white voters *of the same gender*, which is what we sometimes call the **ceteris paribus** (all else equal) difference:

$$\mu(\text{white, man}) - \mu(\text{non-white, man})$$

We also sometimes call this the conditional effect of race, though this use of "effect" is loose and not yet tied to any notion of causality.

Continuous covariates

Up until now, there have been a discrete number of levels of the covariates so we could always each possible combination. But imagine now that instead of race and gender, we want to look at the CEF conditional on income (X_i). Because there are many, many possible values of income, then there are many, many possible conditional expectations. So now, $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ is the CEF of wait times for the population of people with some level of income x. Enumeration of each possible value of the CEF is no longer possible given the large (essentially infinite) number of values income can take.

Now that we can't write out each possible value of the CEF, we have think a little bit about the CEF as a function, $\mu(x)$. What does that function look like? It could be linear so that $\mu(x) = \alpha + \beta x$, which means that the conditional expectation increases as a linear function of income. Or it might be a quadratic function: $\mu(x) = \alpha + \beta x + \gamma x^2$. Or it might be some crazy non-linear function $\mu(x) = \alpha/(\beta + x)$. Again, remember that these are unknown functions in the population, which we don't get to observe. So a key challenge is that when we try to estimate the CEF, $\widehat{\mu}(x)$, we will have to estimate a function whose form we don't know.

CEF error

We can always decompose Y_i into the CEF and an error:

$$Y_i = \mathbb{E}[Y_i|X_i] + e_i,$$

where, $\mathbb{E}[e_i|X_i] = \mathbb{E}[e_i] = 0$. This statement about the CEF error is **definitional**, not an assumption. That is, if we define the error as $e_i = Y_i - \mathbb{E}[Y_i|X_i]$, then we can see that:

$$\mathbb{E}[e_i|X_i] = \mathbb{E}[Y_i - \mathbb{E}[Y_i|X_i]|X_i] = \mathbb{E}[Y_i|X_i] - \mathbb{E}[Y_i|X] = 0,$$

where we used the fact that $\mathbb{E}[\mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X]$ here (which follows because the CEF is a function of X). From this identity, we can also verify that e_i is uncorrelated with any function of X_i.

Intuitively, this says that Y_i can be decomposed into the part "explained by X_i" and a part that is uncorrelated with X_i.

Best predictor

One other reason to focus on the CEF is that it gives the best predictions for Y_i using the information in X_i. Suppose you wanted to make such a prediction. You would figure out some function of X_i, $g(X_i)$, that generates prediction. How good is that prediction? Well, we could figure that out by looking at the the mean squared error (MSE) of the prediction as:

$$\mathbb{E}[(Y_i - g(X_i))^2]$$

What function should you pick? It turns out that the CEF minimizes this prediction error:

$$\mathbb{E}[(Y_i - g(X_i))^2] \geq \mathbb{E}[(Y_i - \mu(X_i))^2]$$

In this sense, we say the CEF is the *best predictor* of Y_i among functions of X_i at least in terms of squared error.

ESTIMATING THE CEF

Before moving on, it is helpful to understand the trade-offs involved in different ways to model the CEF. When we say that we will "model the CEF," we mean that we will make some assumptions about the functional form of $\mu(x)$. That is, we might assume it is linear. But why would we do this? Don't we want to avoid assumptions? Well, sometimes the form of our covariates will force us to make some assumptions, which we will see now.

How do we estimate these discrete CEFs? Well, the easiest way is to simply use the sample mean within each group. Using $X_i = 1$ for white, and $X_i = 0$ for non-white, then we use the following:

$$\widehat{\mathbb{E}}[Y_i|X_i = 1] = \frac{1}{n_1} \sum_{i:X_i=1} Y_i$$
$$\widehat{\mathbb{E}}[Y_i|X_i = 0] = \frac{1}{n_0} \sum_{i:X_i=0} Y_i$$

Here we have $n_1 = \sum_{i=1}^{n} X_i$ is the number of men in the sample and $n_0 = n - n_1$ is the number of women. The sum here $\sum_{i:X_i=1}$ is just summing only over the observations i such that have $X_i = 1$, meaning that i is a man.

This is very straightforward: estimate the mean of Y conditional on X by just estimating the means within each group of X.

Binary covariate example

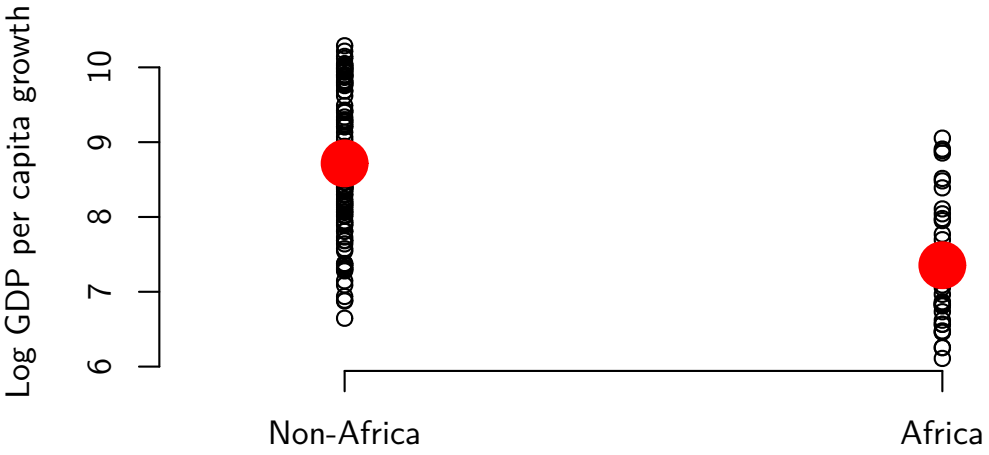
```
ajr <- foreign::read.dta("../data/ajr.dta")
## mean of log GDP among non-African countries
mean(ajr$logpgp95[ajr$africa == 0], na.rm = TRUE)
```

```
## [1] 8.716383
```

```
## mean of log GDP among African countries
mean(ajr$logpgp95[ajr$africa == 1], na.rm = TRUE)
```

```
## [1] 7.355197
```

```
plot(ajr$africa, ajr$logpgp95, xlab = "", ylab = "Log GDP per capita growth", xaxt = "n", xlim = c(-0.25, 1.25),
axis(side = 1, at = c(0,1), labels = c("Non-Africa", "Africa"))
points(x = 0, y = mean(ajr$logpgp95[ajr$africa == 0], na.rm = TRUE), pch = 19, col = "red", cex = 3)
points(x = 1, y = mean(ajr$logpgp95[ajr$africa == 1], na.rm = TRUE), pch = 19, col = "red", cex = 3)
```



Discrete covariate: sample conditional expectations

What if X_i has more than two categories? We can use the same logic to create a similar estimator. That is, we can still estimate $\mathbb{E}[Y_i|X_i = x]$ with the sample mean among those who have $X_i = x$:

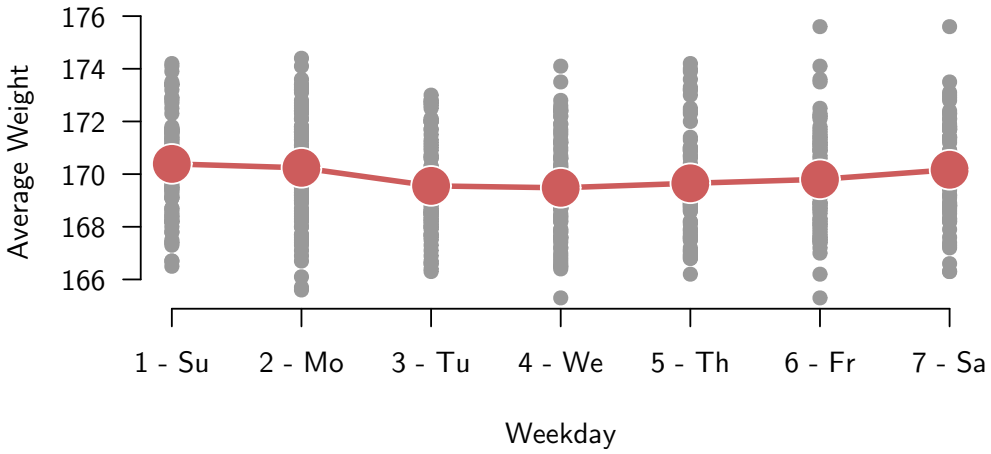
$$\widehat{\mathbb{E}}[Y_i|X_i = x] = \frac{1}{n_x} \sum_{i:X_i=x} Y_i$$

For example, I've been collecting data on my own weight for a little under a year. What if I wanted to know how my weight (Y_i) varied by the day of the week (X_i)? This is a discrete, ordered variable. - Well, we could just calculate the mean weight for each day of the week:

```
weight <- read.csv("../data/weight.csv", stringsAsFactors = FALSE)
weight$weekday <- as.numeric(format(as.Date(weight$date, format = "%m/%d/%y%n%H:%M"), "%w"))+1
weight$date <- as.Date(weight$date, format = "%m/%d/%y%n%H:%M")
day.means <- rep(NA, times = 7)
names(day.means) <- c("1 - Su", "2 - Mo", "3 - Tu", "4 - We", "5 - Th", "6 - Fr", "7 - Sa")
for (i in 1:7) {
    day.means[i] <- mean(weight$weight[weight$weekday == i])
}
day.means
```

```
##    1 - Su    2 - Mo    3 - Tu    4 - We    5 - Th    6 - Fr    7 - Sa
## 170.3883 170.2398 169.5532 169.4810 169.6513 169.8014 170.1657
```

```
plot(x = weight$weekday, y = weight$weight, bty = "n", xaxt = "n", xlab = "Weekday", ylab = "Average Weight", las
lines(x = 1:7, y = day.means, pch = 19, col = "indianred", lwd = 3)
points(x = 1:7, y = day.means, pch = 21, col = "white", cex = 3, bg = "indianred")
axis(side = 1, at = 1:7, labels = names(day.means))
```

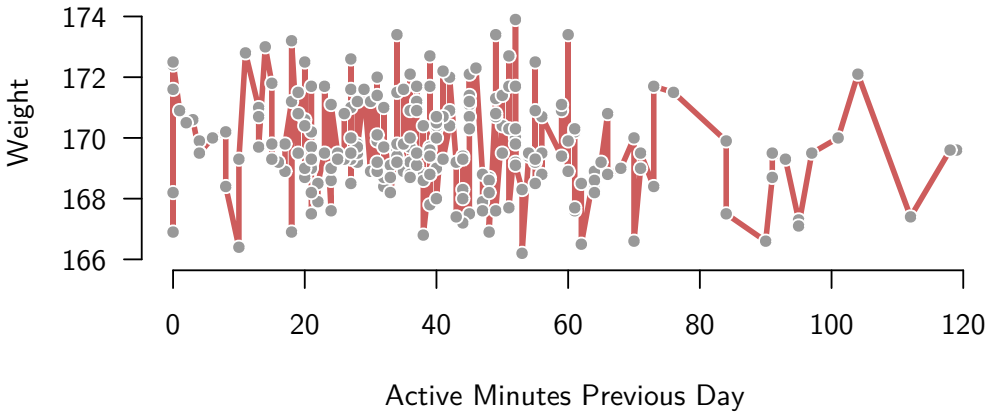


Continuous covariate (I): each unique value gets a mean

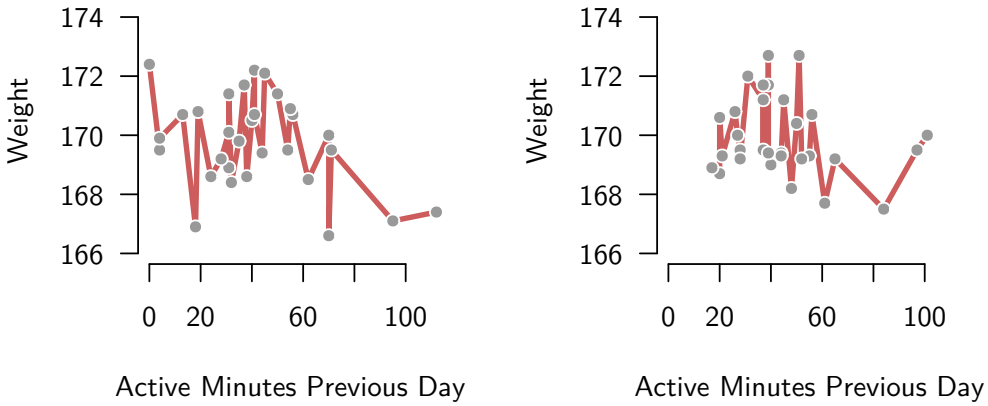
What if X_i is continuous? Can we calculate a mean for every value of X? Not really, because remember the probability that two values will be the same in a continuous variable is 0. Thus, we'll end up with a very "jumpy" function, $\widehat{\mathbb{E}}[Y_i|X_i = x]$, since n_x will be at most 1 for any value of x. Let's look at the relationship between my weight and my active minutes in the previous day using this approach:

```
fitbit <- read.csv("../data/fitbit.csv", stringsAsFactors = FALSE)
fitbit$date <- as.Date(fitbit$date, format = "%m/%d/%y")
## lag fitbit by one day
fitbit$date <- fitbit$date + 1
## merge fitbit and weight data
weight <- merge(weight, fitbit, by = "date")

plot(weight$active.mins[order(weight$active.mins)], weight$weight[order(weight$active.mins)], type = "l", lwd = 
points(weight$active.mins, weight$weight, pch = 21, col = "white", bg = "grey60")
```



You can imagine that this will jump around a lot from sample to sample. The estimates, $\widehat{\mathbb{E}}[Y_i|X_i = x]$, will have high sampling variance. Here are two different samples of this data fit with this interpolation. While both functions are "bumpy," the bumps are coming at different points and it's very difficult to tell what is going on:

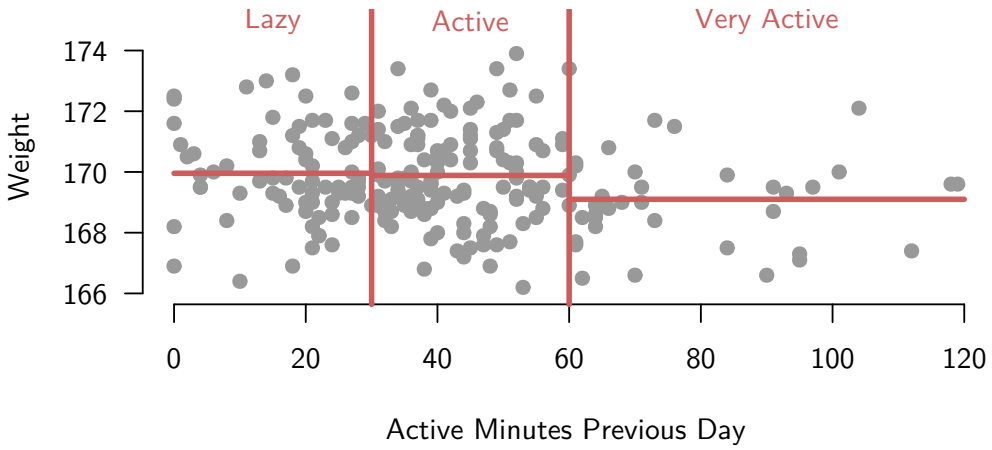


Continuous covariate (II): stratify and take means

So, that seems like each value of X won't work, but maybe we can take the continuous variable and turn it into a discrete variable. We call this **stratification**. Once it's discrete, we can just calculate the means within each **strata**. For instance, we could break up the "Active Minutes" variable into 3 categories: lazy (< 30mins), active (30-60mins), and very active (>60min).

```
lowactivity.mean <- mean(weight$weight[weight$active.mins < 30])
medactivity.mean <- mean(weight$weight[weight$active.mins >= 30 & weight$active.mins < 60])
hiactivity.mean <- mean(weight$weight[weight$active.mins >= 60])
```

```
plot(weight$active.mins, weight$weight, pch = 19, las = 1, bty = "n",
     xlab = "Active Minutes Previous Day", ylab = "Weight", ylim = c(166, 175),
     col = "grey60")
abline(v = c(30,60), col = "indianred", lwd = 3)
text(x = c(15, 45, 90), y = c(175, 175, 175), c("Lazy", "Active", "Very Active"),
     col ="indianred")
segments(x0=c(0,30,60), x1 = c(30,60,120),
         y0 = c(lowactivity.mean, medactivity.mean, hiactivity.mean),
         col ="indianred", lwd = 3)
```



Now we're starting to see that there seems to be a negative relationship. But can we make this even more simple?

Continuous covariate (III): model relationship as a line

The stratification approach was fairly crude: it assumed that means were constant within strata, but that seems wrong. Can we get a more global model for the regression function? Well, maybe we could assume that it is linear:

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \beta_0 + \beta_1 x$$

Why might we do this? Parsimony, first and foremost: 2 numbers to predict any value. The parameters of this linear model also have a nice interpretation:

1. Intercept: the average outcome (weight) among units with $X = 0$ is β_0:

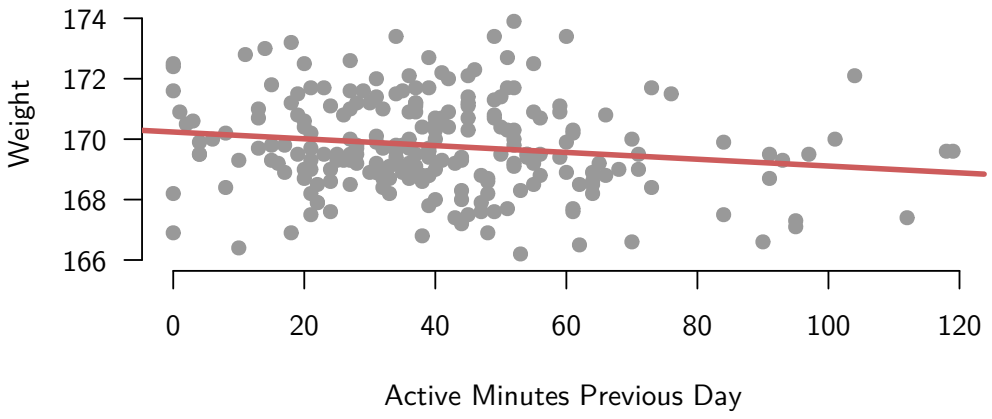
$$\mathbb{E}[Y|X = 0] = \beta_0 + \beta_1 0 = \beta_0$$

2. Slope: a one-unit change in X (active minutes) is associated with a β_1 change in Y (weight):

$$\begin{aligned}\mathbb{E}[Y|X = x+1] - \mathbb{E}[Y|X = x] &= (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1\end{aligned}$$

Here is the linear regression function for the weight-active minutes relationships:

```
plot(weight$active.mins, weight$weight, pch = 19, las = 1, bty = "n",
     xlab = "Active Minutes Previous Day", ylab = "Weight", ylim = c(166, 175),
     col = "grey60")
abline(lm(weight ~ active.mins, data = weight), col = "indianred", lwd = 3)
```



We'll see soon how we estimate this line. It's a bit more complicated that the stratify and calculate means.

Parametric vs. nonparametric models

The estimation of the CEF for discrete independent variables we presented was **nonparametric** because they make no assumptions about the functional form of $\mu(x) = \mathbb{E}[Y_i|X_i = x]$. We just estimate the mean among each value of x. With continuous independent variables, this approach breaks down because of the number of values at which we want to evaluate $\mu(x)$, so we make **parametric** assumptions about the functional form of $\mathbb{E}[Y_i|X_i = x]$ in order to make progress.

LINEAR CEF

Given the discussion about estimation, we are going to have to make some assumptions about the functional form of the CEF to make progress on estimation. We say that a CEF (conditional on one independent variable) is linear when:

$$\mu(x) = \beta_0 + \beta_1 x_1.$$

In this case, we can interpret β_1 as the average change in the mean of Y_i given a one-unit change in X_i and β_0 (called the **intercept**) is the conditional mean of Y_i when $X_i = 0$. For illustration, imagine that Y_i is annual income, measured in dollars, and X_i is years of education. Then, β_1 the expected difference in Y_i between two adults that differ by 1 year of education. And β_0 would be the expected income for someone with 0 years of education.

To see why linearity is an assumption, think about the income-education example. If we assume that the CEF is linear, $\mu(x) = \beta_0 + \beta_1 x$, then we are assuming the difference in average income between $x = 11$ and $x = 12$ is the same as the difference in average incomes between $x = 15$ and $x = 16$. In words, this says that the difference between no high school degree and having a high school degree is the same as the difference between some college and having a 4-year college degree. Of course, this might not be true in the population! But our linearity assumption imposes that.

Linear CEF with binary covariates

Sometimes, linearity isn't an assumption at all, but follows from the data itself. Let's go back to the wait times example with race as a binary covariate, where $X_i = 1$ means that respondent i is white and $X_i = 0$ means that they are non-white. This means that the mean wait times for whites is $\mu(1)$ and the mean wait times for non-whites is $\mu(0)$. With this, we can write the CEF as:

$$\mathbb{E}[Y_i|X_i = x] = \mu(x) = \mu(0) + (\mu(1) - \mu(0))x$$

Doing some simple plugging in reveals that when $x = 0$, then we have $\mathbb{E}[Y_i|X_i = x] = \mu(0)$ and when $x = 1$, we have $\mathbb{E}[Y_i|X_i = x] = \mu(1)$. Given this structure, we can just rewrite this with different parameters:

$$\mu(x) = \beta_0 + \beta_1 x,$$

where β_0 is the mean for non-whites and β_1 is difference between the condition mean for whites and the conditional mean for non-whites. Clearly this is a linear CEF! And we didn't have to make any assumptions since there are only two levels of X_i.

Best linear approximation

We saw that $\mu(x)$ is the best approximation to Y_i in terms of MSE. But this isn't very useful if we don't know the functional form of the CEF. Instead, we may ask the following: what **linear** function of x is the best approximation to Y_i? This question can be written mathematically as:

$$(\beta_0, \beta_1) = \underset{(b_0,b_1)}{\arg\min}\, \mathbb{E}[(Y_i - (b_0 + b_1 X_i))^2]$$

The resulting linear function $(\beta_0 + \beta_1 X_i)$ is called the **linear projection** of Y_i onto X_i. All we are doing is finding the line that best approximates Y_i and we are doing that by minimizing the (average of the) squared distance between values of Y_i.

Can we solve for the slope and intercept of the this linear projection? Yes, we can! Using some simple multivariate calculus (see the technical appendix for details), we can show that the population intercept and slope of the linear projection can be written as:

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$
$$\beta_1 = \frac{\text{Cov}[Y_i, X_i]}{\mathbb{V}[X_i]}$$

Okay, what now? Let's recap. We wanted to find a linear function of X_i that, in the population, minimized the MSE. It turns out if this function is $\beta_0 + \beta_1 X_i$, where β_0 and β_1 are defined above. Thus, this is the population line of best fit.

It's important to see that this best linear predictor exists independent of the CEF. This linear function won't, in general be equal to the CEF since the CEF might not be linear! But we can still define this linear projection. There are two ways in which the linear project relates to the CEF, though.

Theorem 1. *If the CEF is a linear function, $\mathbb{E}[Y_i|X_i] = b_0 + b_1 X_i$, then it will be equal to the linear projection: $\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$.*

Theorem 2. *The linear projection is the best linear approximation to the CEF, so that:*

$$(\beta_0, \beta_1) = \underset{(b_0,b_1)}{\arg\min}\, \mathbb{E}[(\mu(X_i) - (b_0 + b_1 X_i))^2]$$

So now we know that the linear projection has three awesome properties: (1) its the best linear predictor of Y_i given X_i, (2) it is equal to the CEF if the CEF is linear, and (3) it is the best linear approximation to the CEF even if the CEF is nonlinear.

Remember that all of these ideas are in the population. We are not trying to estimate anything yet. We are just trying to understand what the parameters in the population are that we are going to estimate with linear regression.

Linear projection error

Define the error of the linear projection as:

$$u_i = Y_i - \beta_0 - \beta_1 X_i.$$

It can be shown that this error has two properties: $\mathbb{E}[X_i u_i] = 0$ and $\mathbb{E}[u_i] = 0$. This implies that the covariance (and thus the correlation) between u_i and X_i is 0, since $\text{Cov}[X_i, u_i] = \mathbb{E}[X_i u_i] - \mathbb{E}[X_i]\mathbb{E}[u_i] = 0$.

Problem 1. Prove that $\mathbb{E}[X_i u_i] = 0$

So the linear projection error is uncorrelated with X_i, but you should note that it isn't mean independent like the CEF error, $\mathbb{E}[u_i|X_i] \neq 0$. Why is this the case? Remember that correlation/covariance are measures of **linear dependence** and if the CEF is nonlinear, then the linear projection errors might be a nonlinear function of X_i. So even if they are uncorrelated, the projection errors might be a nonlinear function of X_i.

Summary of the linear projection model

Why did we talk about linear projections? The reason is that we will soon be estimating ordinary least squares (OLS) and I wanted to make it clear that OLS estimates a sensible set of population parameters even if the CEF is nonlinear. Moving forward, we will often assume that the CEF is linear, so that the CEF and the linear projection are the same.

LEAST SQUARES

To review our approach: we have defined a linear projection model gives us the line of best fit between Y_i and X_i in the population, $\beta_0 + \beta_1 X_i$. We also know that if we assume that the CEF is linear, then it is equal to the linear projection: $\mathbb{E}[Y_i|X_i = x] = \beta_0 + \beta_1 x$. Either way, β_0 and β_1 are valid population parameters just like μ or σ^2, and we will often want to estimate them in real data.

To do this, let's first make an assumption about our data being i.i.d.:

Assumption 1. $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ *are i.i.d. draws from a population joint distribution,* $f_{(Y,X)}(y, x)$.

How can we develop an estimator for the linear projection? Remember that the population linear approximation minimized the expected value of the squared projection errors:

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\arg\min}\, \mathbb{E}[(Y_i - b_0 - b_1 X_i)^2]$$

To find a good in-sample version of these quantities, we can try to produce the best linear approximation to the observed Y_i given the observed X_i. This involves replacing the population expectation with a sample expectation:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

The estimator defined by this procedure is usually called **least squares** (LS) or **ordinary least squares** (OLS). This estimator works by the plug-in or analogy principle we mentioned a few weeks ago.

We can solve for the least squares estimator in a similar way to how we solved for the linear approximation coefficients. When we do that (see the appendix), we find that:

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \\
&= \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
\end{aligned}$$

These should look very familiar! They are exactly the population coefficients where we replace population expectations with their sample versions.

Intuition of the OLS estimator

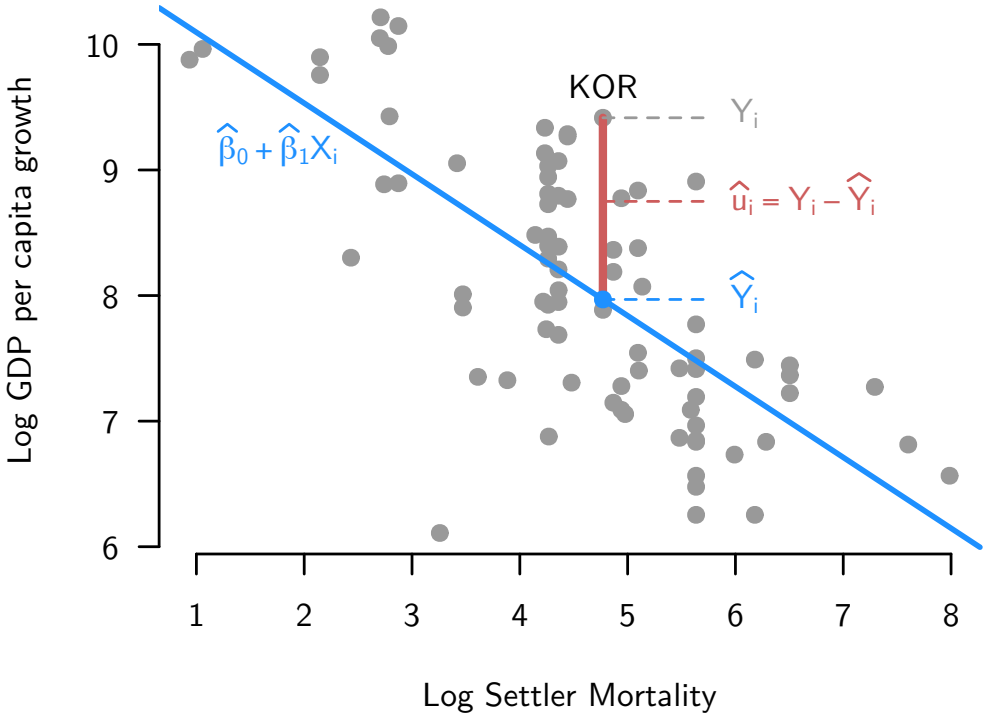
It is useful to have the following definitions:

- **Definition** A **fitted value** or **predicted value** of Y_i for a particular observation with independent variable X_i:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- **Definition** The **residual** is the difference between the actual value of Y_i and the predicted value, $\widehat{Y}_i$:

$$\widehat{u}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$



Minimize the residuals

The residuals, $\widehat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data. Larger magnitude residuals means that points are very far from the line. Residuals close to 0 mean points very close to the line. The smaller the magnitude of the residuals, the better we are doing at predicting Y_i Thus, it makes sense to choose the line that minimizes the residuals. One way to think about OLS is that it chooses the line that **minimizes the sum of the squared residuals.**

Mechanical properties of least squares

- The residuals will be 0 on average:

$$\sum_{i=1}^{n} \widehat{u}_i = 0$$

- The residuals will be uncorrelated with the predictor:

$$\sum_{i=1}^{n} X_i \widehat{u}_i = 0$$

- The residuals will be uncorrelated with the fitted values:

$$\sum_{i=1}^{n} \widehat{Y}_i \widehat{u}_i = 0$$

Sample covariance

- The sample version of population covariance, $\sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.
- **Defintion** The **sample covariance** between Y_i and X_i is

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$$

```
## cov() gets confused when you give it missing data
cov(ajr$logem4, ajr$logpgp95)
```

```
## [1] NA
```

```
## tell cov() to use only the pairwise complete observations:
cov(ajr$logem4, ajr$logpgp95, use = "pair")
```

```
## [1] -0.9881104
```

Sample correlation

- The sample version of population correlation, $\rho = \sigma_{XY} / \sigma_X \sigma_Y$.
- **Defintion** The **sample correlation** between Y_i and X_i is

$$\hat{\rho} = r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sum_{i=1}^{n} (Y_i - \overline{Y}_n)^2}}$$

```
## cor() is very similar to cov()
cor(ajr$logem4, ajr$logpgp95)
```

```
## [1] NA
```

```
## and has the same solution to NAs:
cor(ajr$logem4, ajr$logpgp95, use = "pair")
```

```
## [1] -0.7047632
```

AJR Example in R

- Let's use those simple formulas we just learned:

```
cov.xy <- cov(ajr$logem4, ajr$logpgp95, use = "pair")
var.x <- var(ajr$logem4, na.rm = TRUE)
cov.xy/var.x
```

```
## [1] -0.5816937
```

```
mean(ajr$logpgp95, na.rm = TRUE) - cov.xy/var.x * mean(ajr$logem4, na.rm = TRUE)
```

```
## [1] 10.97596
```

- Compare it to what `lm()`, the OLS function in R produces:

```
coef(lm(logpgp95 ~ logem4, data = ajr))
```

```
## (Intercept)      logem4
##  10.6602465  -0.5641215
```

- Why aren't these equal? Hint: think about missing data.

Mechanical properties of least squares in R

```
mod <- lm(logpgp95 ~ logem4, data = ajr)
mean(residuals(mod))
```

```
## [1] -2.623502e-18
```

```
## mod$model is the data used in the lm() call
cor(mod$model$logem4, residuals(mod))
```

```
## [1] -3.184875e-17
```

```
cor(fitted(mod), residuals(mod))
```

```
## [1] -1.160489e-16
```

TECHNICAL PROOFS

Proof of the linear projection

First note that we can expand the square and use the linearity of the expectation to get the following:

$$\mathbb{E}[(Y_i-(\beta_0+\beta_1 x))^2] = \mathbb{E}[Y_i^2]-2\beta_0\mathbb{E}[Y_i]-2\beta_1\mathbb{E}[Y_iX_i]+2\beta_0\beta_1\mathbb{E}[X_i]+\beta_0^2+\beta_1^2\mathbb{E}[X_i^2]$$

To find the values that minimize this function, we need to take partial derivatives with respect to each parameter, set these equal to zero, and solve for the parameters. We can write the first-order condition for β_1 as follows:

$$0 = -2\mathbb{E}[Y_i] + 2\beta_1\mathbb{E}[X_i] + 2\beta_0$$

Reorganizing gives us,

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1\mathbb{E}[X_i]$$

The first order condition for the slope β_1 is:

$$\begin{aligned}
0 &= -2\mathbb{E}[Y_iX_i] + 2\beta_0\mathbb{E}[X_i] + 2\beta_1\mathbb{E}[X_i^2] \\
0 &= -\mathbb{E}[Y_iX_i] + \beta_0\mathbb{E}[X_i] + \beta_1\mathbb{E}[X_i^2] \\
0 &= -\mathbb{E}[Y_iX_i] + (\mathbb{E}[Y_i] - \beta_1\mathbb{E}[X_i])\mathbb{E}[X_i] + \beta_1\mathbb{E}[X_i^2] \\
0 &= -(\mathbb{E}[Y_iX_i] - \mathbb{E}[Y_i]\mathbb{E}[X_i]) + \beta_1(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2)
\end{aligned}$$

Rearranging and noting that $\mathbb{V}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$ and $\text{Cov}[Y_i, X_i] = \mathbb{E}[Y_iX_i] - \mathbb{E}[Y_i]\mathbb{E}[X_i]$, we have the following:

$$\beta_1 = \frac{\text{Cov}[Y_i, X_i]}{\mathbb{V}[X_i]}$$

Implicit in these derivations are three assumptions: (1) $\mathbb{E}[Y_i^2] < \infty$, (2) $\mathbb{E}[X_i^2] < \infty$, and (3) $\mathbb{V}[X_i] > 0$.

Deriving the OLS estimator

- Define the least squares objective function:

$$S(b_0, b_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - X_ib_1)^2.$$

- How do we derive the LS estimators for β_0 and β_1?

1. Take partial derivatives of S with respect to b_0 and b_1.
2. Set each of the partial derivatives to 0
3. Solve for $\{b_0, b_1\}$ and replace them with the solutions

- The partial derivatives are:

$$\begin{aligned} S(b_0, b_1) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - X_i b_1)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2) \end{aligned}$$

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = \frac{1}{n}\sum_{i=1}^{n}(-2Y_i + 2b_0 + 2b_1 X_i)$$

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = \frac{1}{n}\sum_{i=1}^{n}(-2Y_i X_i + 2b_0 X_i + 2b_1 X_i^2)$$

- The first order conditions are:

$$0 = \frac{1}{n}\sum_{i=1}^{n}(-2Y_i + 2b_0 + 2b_1 X_i)$$

$$0 = \frac{1}{n}\sum_{i=1}^{n}(-2Y_i X_i + 2b_0 X_i + 2b_1 X_i^2)$$

now solving for b_0 and b_1 yields the **normal equations**:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

$$\widehat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} X_1^2\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - \widehat{\beta}_0 \overline{X}$$

- Plug the first into the second normal equation:

$$\widehat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} X_1^2\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - (\overline{Y} - \widehat{\beta}_1 \overline{X})\overline{X}$$

$$\widehat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} X_1^2 - \overline{X}^2\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) - \overline{YX}$$

- Use the following fact twice, once for each side of the above equals sign (challenge question: show this is true):

$$\left(\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right) - \overline{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$$

- This leaves us with:

$$\widehat{\beta}_1 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})$$

- And rearrange them to get the **OLS estimators**:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$