Data Science for Economists

Lecture 7: Introduction to Regression

Drew Van Kuiken
University of North Carolina | ECON 370

Table of contents

with thanks to blog posts from Joshua Loftus today.

What Is Regression?

Some options:

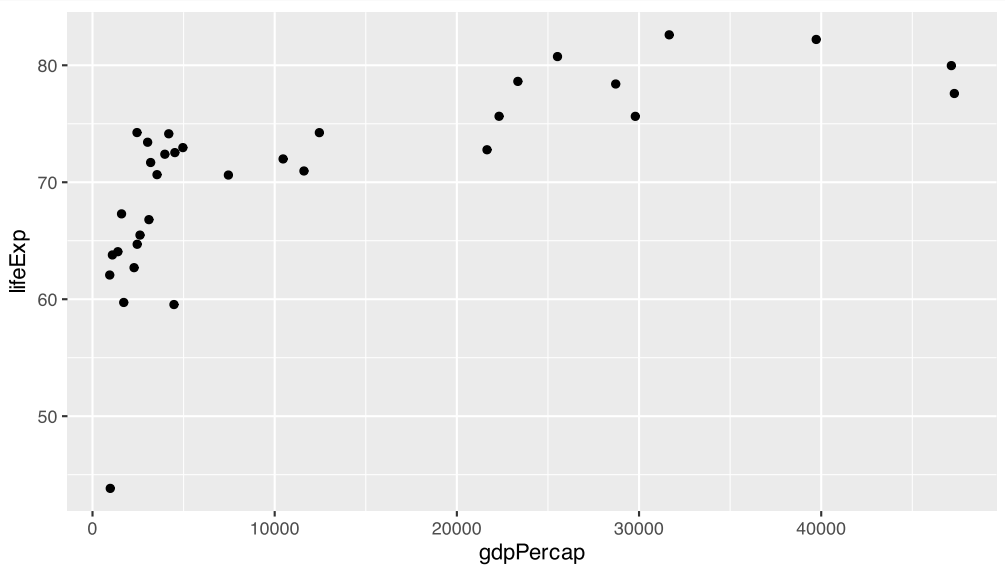
- A way of comparing **treatment** and **control** subjects who have the same observed characteristics
- A way to assess the *relationship* between independent variables and a dependent variable
- Group means (for Ordinary Least Squares, at least)

... helpful, but not very concrete

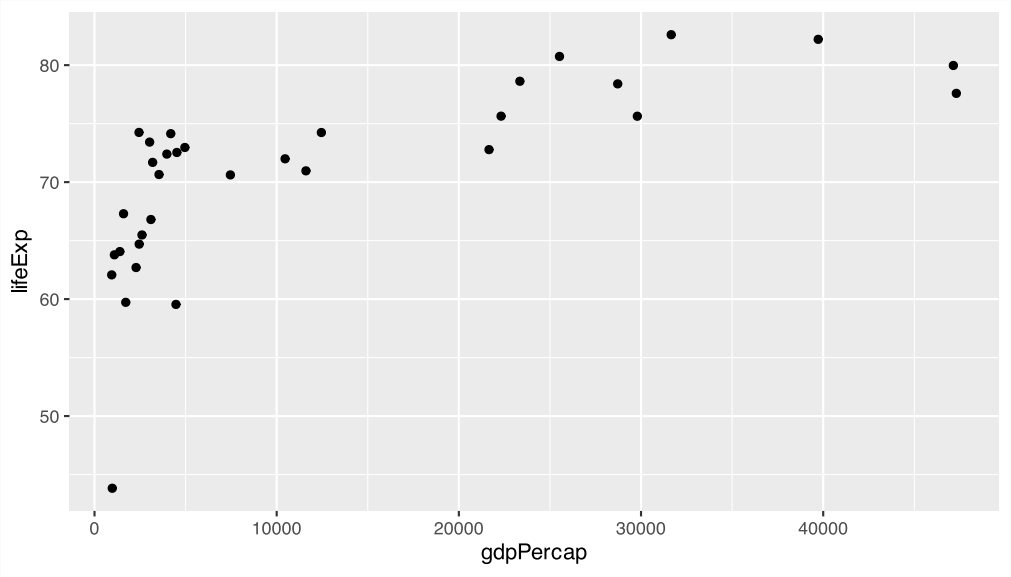
We've seen an example already...

Let's go back to our OG `gapminder` dataset.

```
library(gapminder)
gp_subset ← gapminder[gapminder$continent=="Asia"&gapminder$year=2007,]
g ← ggplot(gp_subset,aes(x=gdpPercap, y=lifeExp)) +
  geom_point()
g
```



We've seen an example already...



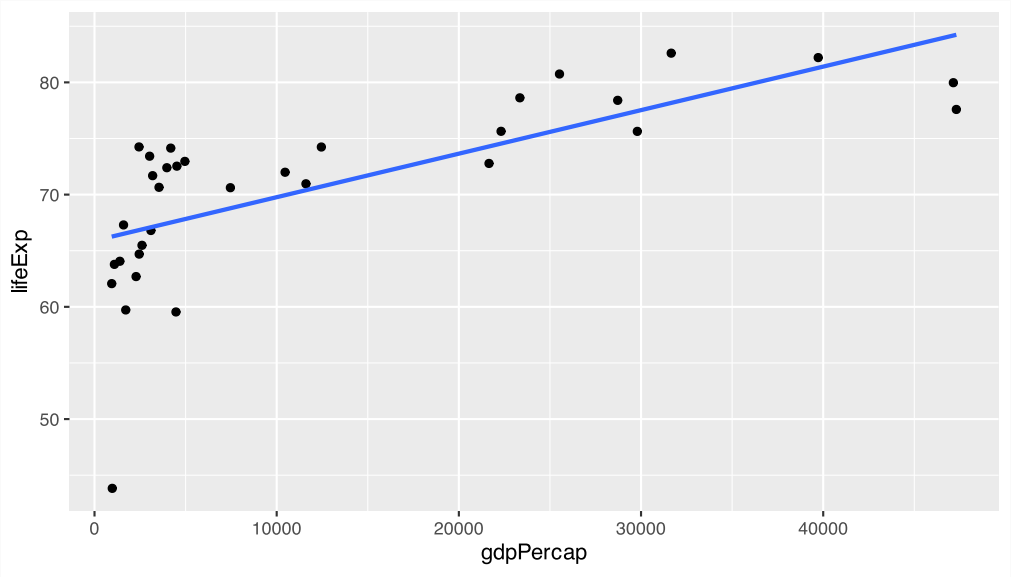
How can we assess the relationship between GDP per capita and life expectancy? If a country in Asia in 2007 had a GDP per capita of 15,000, what would we expect its life expectancy to be?

We've seen an example already...

In our first class, we added a line of best fit:

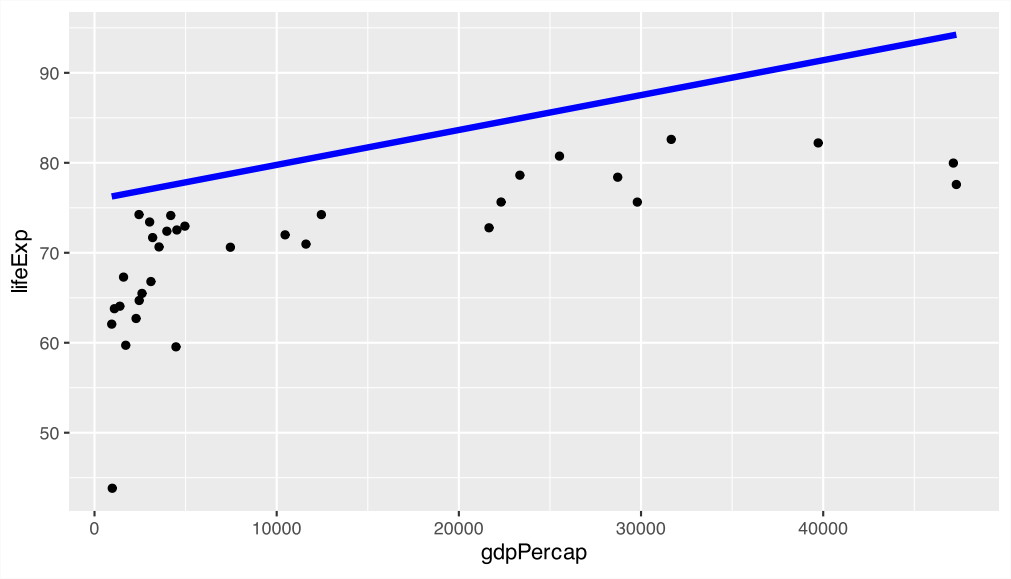
```
g + geom_smooth(method='lm',se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

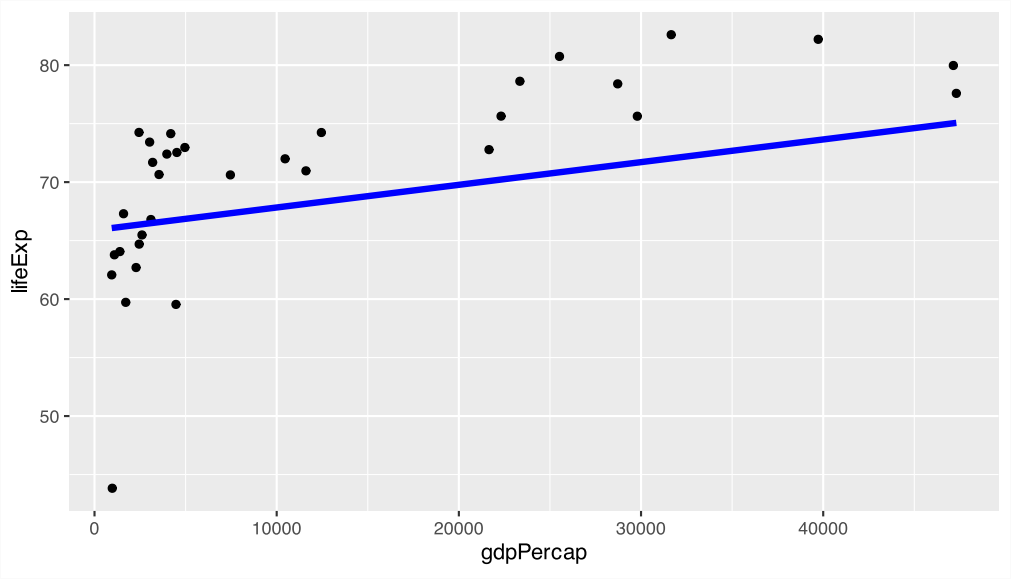


So: how would we characterize the relationship? If a country in Asia in 2007 had a GDP per capita of 15,000, what would we expect its life expectancy to be?

We can try other lines



We can try other lines

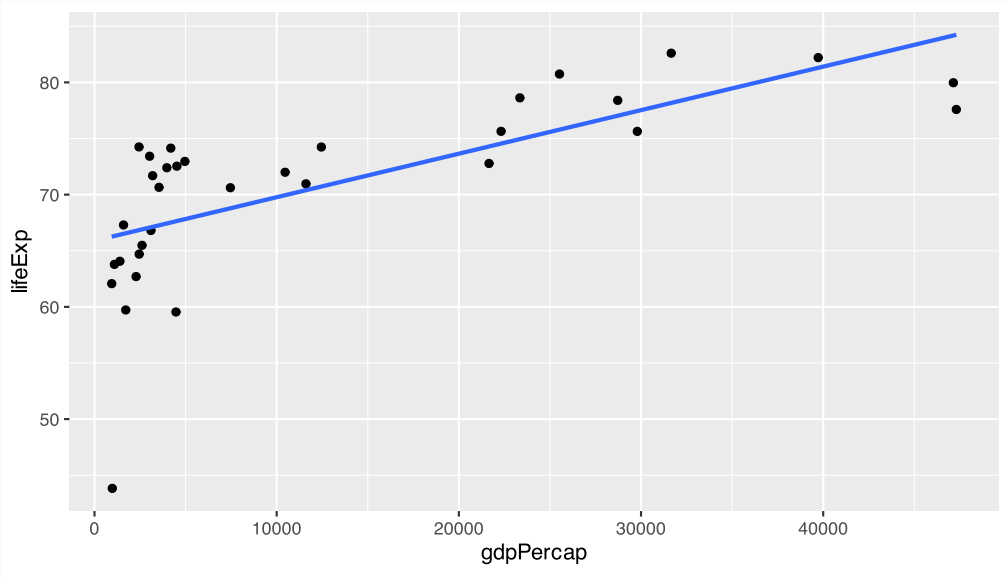


OLS

When we fit a line of best fit, we minimize the ***squared deviations*** between our predictions and our data

```
g + geom_smooth(method='lm',se=FALSE)
```

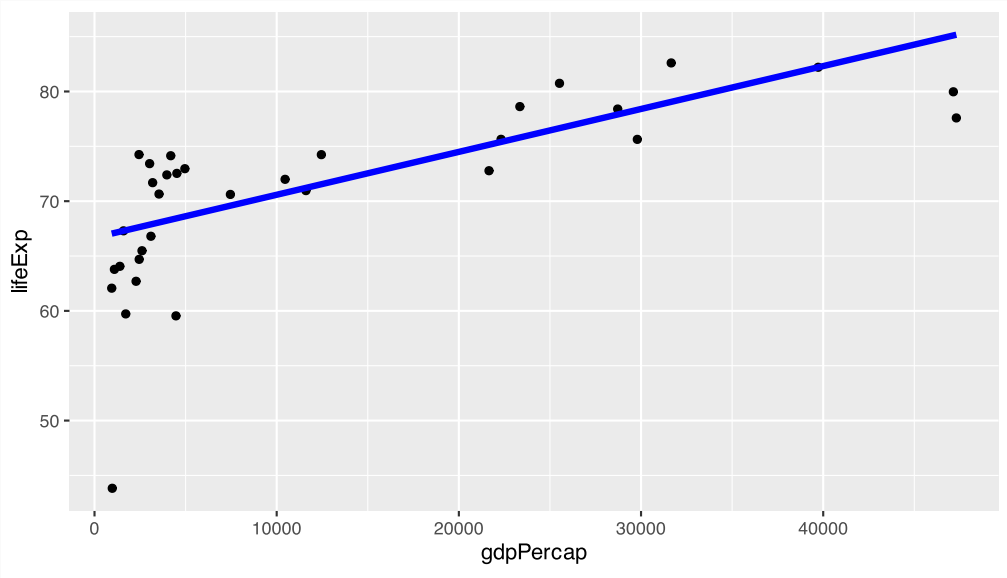
```
## `geom_smooth()` using formula = 'y ~ x'
```



OLS = Ordinary Least *Squares*. We'll get into the math of this later.

Other Options Can Work!

We could just as easily minimize the ***absolute deviations***.

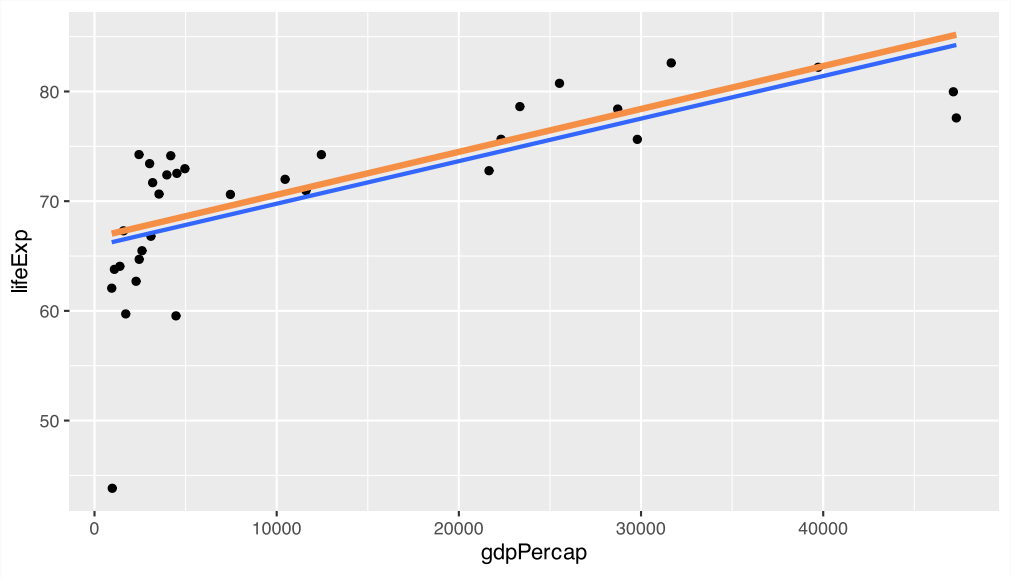


This line looks a lot better than our ad hoc lines! We could certainly use it to assess relationships or predict life expectancies.

We typically don't though. OLS has some *very* nice properties, and so it's become the first thing in every data scientist's toolkit.

In Case You Were Curious:

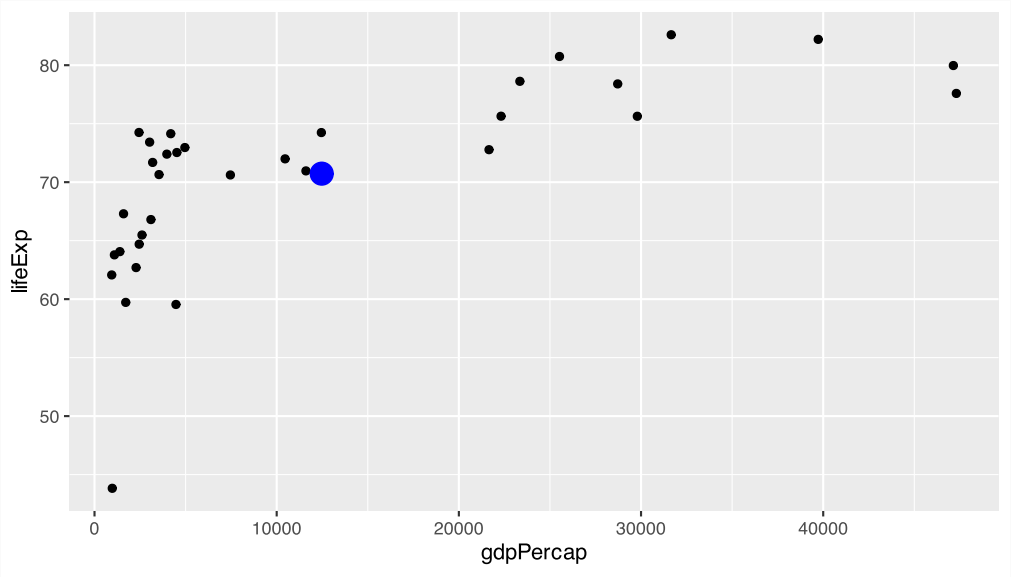
```
## `geom_smooth()` using formula = 'y ~ x'
```



Here's the difference between minimizing ***absolute*** deviations and minimizing ***squared*** deviations.

Building Intuition: Mechanics

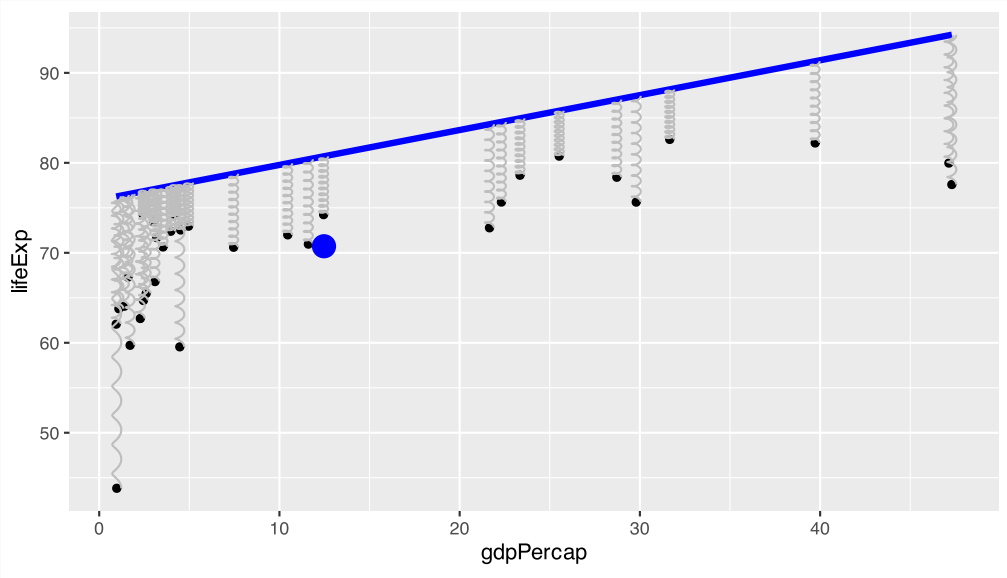
Imagine each of our data points is a physical object. The large blue dot below is the average of our dataset:



Now, imagine we've attached springs between each point and a line running through our data. Each spring has equal strength.

What If Our Line Is Wrong?

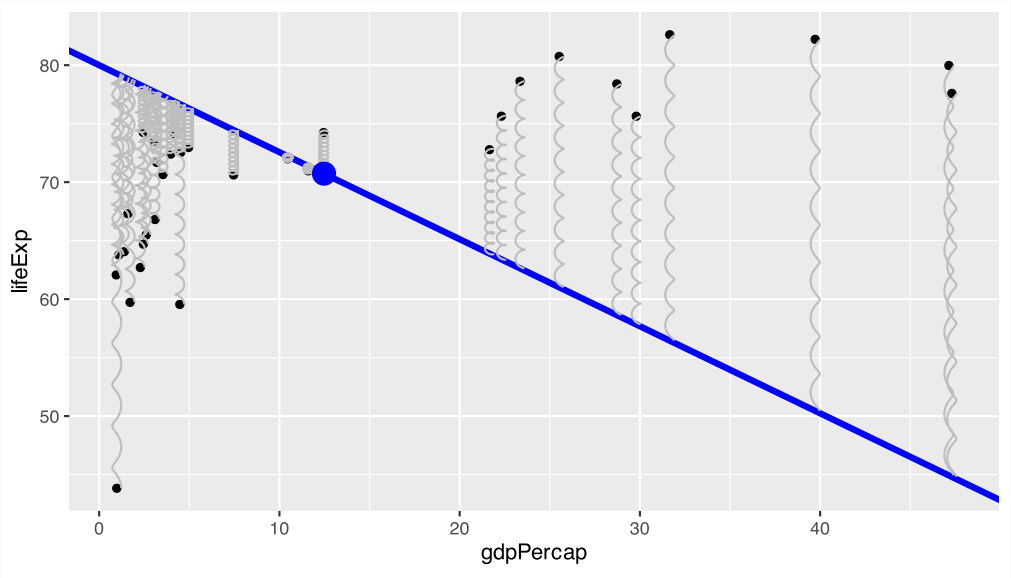
What happens if we use the line that was too high up?



All the springs pull downwards on the line, until it runs through the big blue dot

What If Our Line Is Wrong?

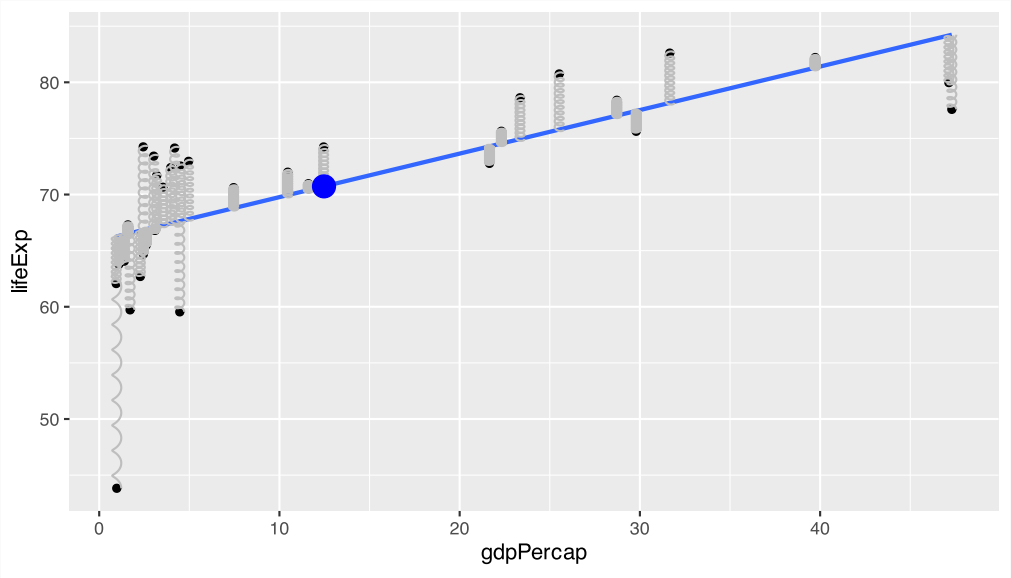
The springs will exert pressure even if the line runs through the mean of the data:



The springs will exert torque on our line until the torque balances out!

What If Our Line Is Wrong?

```
## `geom_smooth()` using formula = 'y ~ x'
```



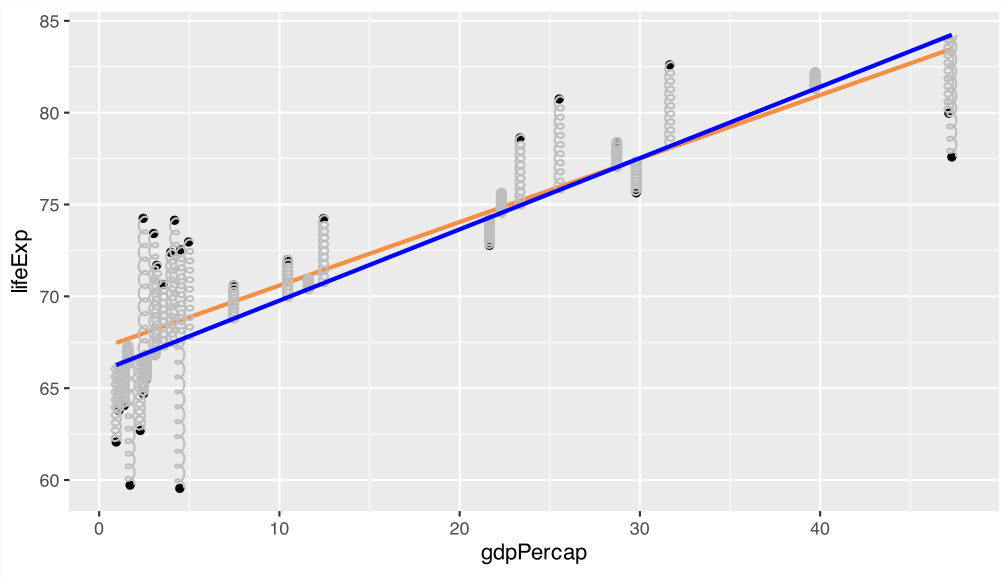
Now an equal amount of force is being exerted on the line such that it will not shift up or down or rotate further.

Outliers are important!

Check out how much pressure the point in the bottom left corner is exerting on our line.

Ordinary Least Squares minimizes the squared deviation from our line of best fit, so outliers are going to weigh disproportionately on our results. Imagine we dropped that observation:

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Looking Closer

So we've run our first regression. We've teased out some details about our relationship and we're left with some questions.

Most importantly: what's going on with that outlier? We can do some investigation:

```
head(gp_subset[gp_subset$lifeExp <= 50,1:6])
```

```
## # A tibble: 1 × 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       2007    43.8 31889923      975.
```

Why might our regression be misleading?

Regression: An Intuitive Approach

College Attendance and Earnings

How much money would a 40-year-old Massachusetts-born graduate of Harvard have made if he or she had come to UMass Amherst instead?

Harvard's average mid-career salary: $98k

UMass's average mid-career salary: $83k

Drew said regressions are grouped means:

- If you're in the Harvard Group, your salary is $98k
- And if you're in the UMass Group, your salary is $83k

Is this a good answer? What might be wrong here?

Problem 1: Selection Bias

When we run a regression, we are implicitly trying to understand a ***counterfactual***. If *Alice* had gone to UMass instead of Harvard, what would her earnings in particular look like?

When we look at group averages, we obscure what makes Harvard and UMass students different:

- Harvard students might have had higher SAT scores or high school grades
- Or Harvard students have parents who run hedge funds and get them jobs at their hedge fund

In economics, we call this ***selection bias***. It says that individuals are *selected* into our sample differently based on their characteristics. In this case, students who attend Harvard are fundamentally different from students who attend UMass.

The best way to solve this: randomization.

Imagine Harvard and UMass randomize which students they accept. Then, on average, Harvard and UMass students will have the same test scores on average and the same parental incomes. Thus any difference in average post-graduation earnings will be due to the college's *treatment effect*.

Problem 1: Selection Bias

In economics, we usually can't randomize. We can alleviate some concerns by **matching on (controlling for) observables** though. In this case, we can use SAT scores and parental occupations to match students into groups, some of whom when to Harvard and some of whom went to UMass, and compare post-graduation earnings within those groups.

In Mastering Metrics, the authors describe a different, clever matching strategy:

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

Controls

We can break a regression into three parts:

1. A dependent variable (our outcome)
2. A treatment variable (what we're trying to study)
3. Controls (variables that determine selection outside of our treatment)
    - Intuitively: the groups we want to match people into
    - Things we want to "hold fixed"

In the case of school applications, we would want to control for SAT scores, high school grades, and parental income (and probably more). In our gapminder example, we would want to control for "defender in long brutal war."

What would we want to control for in the Mastering Metrics example?

Should we control for eye color?

No! Eye color has nothing to do with how people select into treatment or what someone's post-graduation earnings look like. Be careful that you don't run "kitchen-sink" regressions.

More broadly, be careful with how you use regressions. Data is more and more available these days, but that doesn't absolve you of responsibility. A brief history of bad regressions:

- 1835: Adolphe Quetelet starts doing social physics (first social scientist...), correlates social data together, tries to predict crime, poverty, alcohol consumption, etc
- 1883: Francis Galton founds the field of eugenics
- 1991: "A glass of red wine per day is healthy"

Bad Internet Infographics vs. Economics

How can selection help us understand what's wrong with this figure?

Roy Model (Brief!)

This actually goes back to a 1951 article in econ about people choosing between being a rabbit hunter or a fisher. But its more commonly known based on a 1987 article about determinants of immigration patterns.

Think about the distribution of skill in an economy and the distribution of wages. Three different scenarios can lead someone to immigrate to the US:

1. Someone is very talented in their home country and would be very talented in the US. A person can make more money in the US, so high-talent people immigrate.
2. Someone is less talented in their home country and would be less talented in the US. The US has a better social safety net than their home country, so they'll be wealthier living in the US, and they decide to migrate here.
3. Someone is less talented in their home country and would be very talented in the US. This might happen if skills are valued differently in each country. Think about minority groups facing prejudice in their home country. Or, someone abroad who can eat 80 hotdogs in 5 minutes.

Back to the Chart

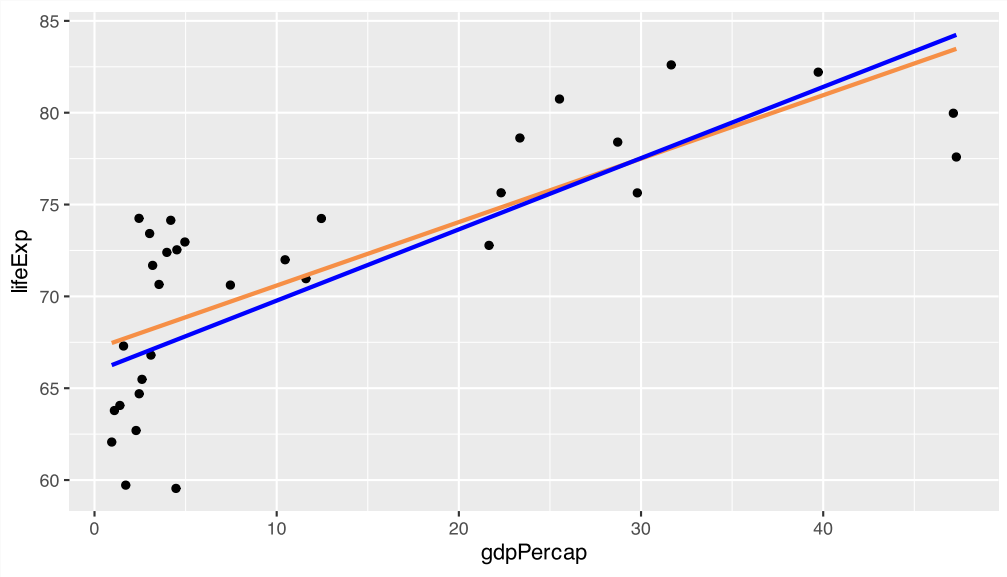
So: why doesn't just matching on observables help us here?

Without modeling the distribution of skills, the distribution of wages, and the way that societies value skills, the immigrant earning chart is total nonsense. (We *could* estimate returns to immigration using the Roy model though)

Problem 2 (or 1b?): Omitted Variables

Earlier, we said that, in our gapminder regression, we would want to control for "defender in a long and brutal war." We didn't want to control for "at war." Any guesses as to why this distinction matters?

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



Omitted Variables Bias

Anyone can be at war, no matter what your GDP is. So, while being at war certainly lowers a country's life expectancy, it merely shifts the entire line of best fit downwards (reduces the intercept). It **doesn't change the slope of the coefficient we care about**.

Fighting a guerrilla war at home for 10+ years affects a country's GDP per capita and its life expectancy. When we don't control for the long and brutal war, we omit an important part of how a country's GDP per capita and life expectancies are determined.

More formally: Omitted variable bias arises when we forget to control for a quantity that affects our treatment variable and our dependent variable.

Calculating OVB

Generally, omitted variables are hard to fix. Two methods:

1. Include the variable
2. Model the variable (a la the Roy Model)

Calculating OVB: Including the variable

What is OVB numerically?

It's the size of our treatment coefficient with the omitted variable included minus the size of our treatment coefficient with the omitted variable excluded. Using `gapminder`:

```
# Calculating OVB:
res_with_ovb ← lm(lifeExp ~ gdpPercap, data = gp_subset)
ovb_coef = res_with_ovb[[1]][2]

gp_subset$long_and_brutal_war = gp_subset[,"country"] == "Afghanistan"
res_without_ovb ← lm(lifeExp ~ gdpPercap + long_and_brutal_war, data = gp_subset)
no_ovb_coef = res_without_ovb[[1]][2]

print(cbind(ovb_coef,no_ovb_coef, ovb_coef-no_ovb_coef))
```

```
##                ovb_coef  no_ovb_coef
## gdpPercap 0.000387841 0.0003454364 4.24046e-05
```

Without controlling for the Afghanistan war, our original estimate of GDP per capita on life expectancy was too high!

Modeling the Variable

How are earnings determined? Famous question, important to get right.

We can start with Jacob Mincer's earnings function:

$$ln(\text{wage}) = \beta_1 + \beta_2 \text{education} + \beta_3 \text{experience} + \beta_4 \text{experience}^2 + u$$

What's an important missing variable here?

Ability. People with higher ability tend to obtain more years of education. To account for this, we can use some kind of standardized test scores (GRE, SAT, IQ) to model ability, which we can then put into our model of wages:

$$\text{ability} = \gamma_1 + \gamma_2 \text{SAT} + v$$

$$ln(\text{wage}) = \beta_1 + \beta_2 \text{education} + \beta_3 \text{experience} + \beta_4 \text{experience}^2 + \beta_5 \text{ability} + u$$

$$ln(\text{wage}) = \beta_1 + \beta_2 \text{education} + \beta_3 \text{experience} + \beta_4 \text{experience}^2 + \beta_5 (\gamma_1 + \gamma_2 \text{SAT} + v)$$

... of course, standardized tests bring about measurement error, which we won't cover in this class.

Two Kinds of Biases

We've covered 2 kinds of biases to watch out for so far:

1. Selection Bias
    - It says that individuals are *selected* into our sample differently based on their characteristics.
    - What is the return on earnings for a competitive hot dog eater who moves to the US? It depends on how much their own country values hot dog eating
2. Omitted Variables Bias
    - Arises when we forget to control for a quantity that affects our treatment variable and our dependent variable
    - Those with higher ability go to more school, earn more money

Are these really separate ideas? Only sort of. Imo, thinking about selection as a process makes things a little cleaner. The Roy model isn't a missing variable, it's a missing process that determines who shows up in our datasets. But things are admittedly a little muddy.

Next lecture(s): Regression (Math and Coding, Optimization)