Multivariate Regression: Part IV - Assumptions, Dummies, and Fixed Effects

EC420 MSU

Justin Kirkpatrick

Last updated February 22, 2021

This lecture

Goal:

1. Briefly review last week (linear hypotheses and F-test, restricted/unrestricted model)

2. Relax assumption of normal errors (MLR.6)

    - *Asymptotic Normality*

3. Dummy variables

4. Fixed Effects

5. Panel Data

Single hypothesis

- Uses just one β_j
- Usually $H_0 : \beta_j = 0$

Linear Hypothesis

- Testing for equality: $\beta_j = \beta_k$
  - $\beta_j - \beta_k = 0$
  - `car` packages, `linearHypothesis`

F-test for test of joint significance

Answers the question **do these coefficients jointly equal to zero**?

- $\beta_j = \beta_k = \beta_l = 0$
- Which is another way of saying "do they explain y"
- Based on
  - SSR_{UR}, the SSR from an unrestricted model
  - SSR_R, the SSR from the restricted ($\beta = 0$) model

R gives us an F-test for all coefficients jointly

- Compared to a restricted model where *only* β_0, the intercept.

Gauss-Markov Regression Assumptions:

MLR.1	The population, y is a linear function of the parameters x and u: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
MLR.2	The sample $(y_i, x_i) : i = 1, 2, \cdots, n$ follows the population model and are independent
MLR.3	No multicolinearity / "full rank": x_j is not a linear transformation of x_k for all j, k.
MLR.4	Zero conditional mean: $E[u\|x_1, x_2, \cdots, x_k] = 0$ for all x.
MLR.5	$Var[u\|x_1, \cdots, x_k] = \sigma_u^2$ for all x.
MLR.6	u is normally distributed ($u \sim N$)

If we combine MLR.6 with MLR.4 and MLR.5, we are assuming "exact normality"

Exact Normality:

- The population error u is *mean independent* of the explanatory variables $x_1, x_2, \cdots, x_k$
- And it is normally distributed with zero mean and variance σ^2: $u \sim N(0, \sigma^2)$
  - Let's call this "exact normality"
  - We need this *only* for inference (t's, F-tests)

Mean Independence:

"Mean independence" is $E[u|x_1, \cdots, x_k] = c$ and $E[u] = 0$ (therefore $c = 0$)

"Asymptotic" just means "pertaining to very large N's"

- That is, very large samples.

The "asymptotic properties" of an estimator are:

- "What it does when $N \to \infty$"
  - When "N gets larger and larger"
- Particularly, does it get *closer and closer* to some desirable value?

MLR6, the "exact normality" assumption, may not be necessary with a very large N

- Which is good, because it probably doesn't hold in many cases!

Let's look at one where exact normality doesn't hold

Example 3.5 in Wooldridge

$$NumArrests = \beta_0 + \beta_1 pcnv + \beta_2 avgsentence + \beta_3 ptime + \beta_4 qemp + u$$

Example 3.5 in Wooldridge discusses regressing *Number times arrested* on some variables of interest. Since most people are arrested zero times, $y|x_1, x_2, \cdots, x_k$ and the associated errors, $u|x_1, x_2, \cdots, x_k$ are most definitely not normally distributed!

So, the estimators are still:

- Unbiased (MLR1-4)
- Have valid variances (MLR5 or HC-robust)
- But we do not know the exact distribution to use: u is not necessarily normal, so β is not necessarily normally distributed. Therefore, our t-test is not valid.

The Central Limit Theorem to the rescue

The CLT states that any average, once standardized, is distributed standard normal when n gets very large.

- By **average**, we mean anything that is the form $\frac{1}{N}\sum_{i=1}^{N} x_i$
- By **standardized**, we mean anything that subtracts the true mean and divides by the standard deviation
- We used this fact in looking at the *se of the mean*:

$$\frac{\bar{Y} - \mu_Y}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

$\hat{\beta}$ is also an average

- $\widehat{Cov}(Y, X)$ is an average: $\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$
- $\widehat{Var}(X)$ is also an average just the same
  - β depends on a bunch of averages!

So if we *properly standardize* it, we know it is asymptotically normal *regardless* of the distribution of u

- This is true even if u is very obviously not normal.

This only applies as $n \to \infty$. It is an asymptotic result

MICHIGAN STATE UNIVERSITY

Even when MLR.6 doesn't hold

We can say that our estimator, β, has a normal **asymptotic variance**

Which means it is normally distributed **when** $n \to \infty$.

- Asymptotic standard error
- Asymptotic 95% Confidence Interval, etc.

And, since a $t_{\infty-K-1}$ is the same as a $N(0,1)$, we can use the normal tables instead of the t-tables.

When n is small and u is not normal, then we use "small sample" properties, which we won't cover in this class.

Consistency is a property of an estimat**or**, much like "unbiased"

- It is about what happens to the estimator when n gets larger and larger.
- On the other hand, *bias* is about the expected value of the estimator.

Definition

> An estimator is consistent when it converges in probability to the correct population value as the sample size grows.

Converges in probability

For any tiny, tiny number we can choose, say ϵ, a consistent estimator $\hat{\beta}$ will have some n large enough that $Pr(|\hat{\beta} - \beta| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$

Remember our *standard error of the mean*

$$se(\bar{X}) = \sqrt{\frac{\sigma^2}{n}}$$

If we had a small n

- We had a pretty big std. err on $\bar{X}$

But if we had a really big n

- We got a std. error that was smaller and smaller...

With a big enough n, the std. error of the mean becomes very very small

- And a plot looks like a "spike"

That's the concept of consistent

Consistency

A good example of *biased* but *consistent* is the use of the population variation formula on a sample:

$$\hat{\sigma}^2_{biased} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Biased, yes. But *consistent* since the estimate goes to the correct value as $n \to \infty$

The proof showing why the 1/N calculation is biased is long and drawn-out. Just remember that 1/N is biased.

The end result is that we can relax MLR6 in large samples and not worry about u being normally distributed and **still:**

- Know that $\hat{\beta}$ is normally distributed
- Know that we can use a t-statistic (since we are still estimating $\hat{\sigma^2}$)
- And know that since $\hat{\sigma}^2$ is consistent, with large samples, $\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma}^2}{SST_x}}} \sim N(0,1)$

Asymptotic Normality

Any questions?

Dummy Variables

Dummy Variables

A dummy is any variable that takes **only** one of two values:

$$\{0, 1\}$$

- This is also called a **binary** variable

Sometimes called an "indicator variable" as well

- Because it "indicates" if something *qualitative* is true.
- Also, sometimes written as $\mathbb{1}(condition)$ e.g. $\mathbb{1}(age > 65)$
  - It is equal to 1 for that observation if that observation's age is greater than 65.
  - It is equal to 0 otherwise

In Wooldridge Ch. 7.1

- He uses the example of $male$ and $female$, with a variable equal to 1 if $female == TRUE$.

Since it takes on numeric values, we can use it in a regression:

$$y = \beta_0 + \beta_1 educ + \delta_0 1(female) + u$$

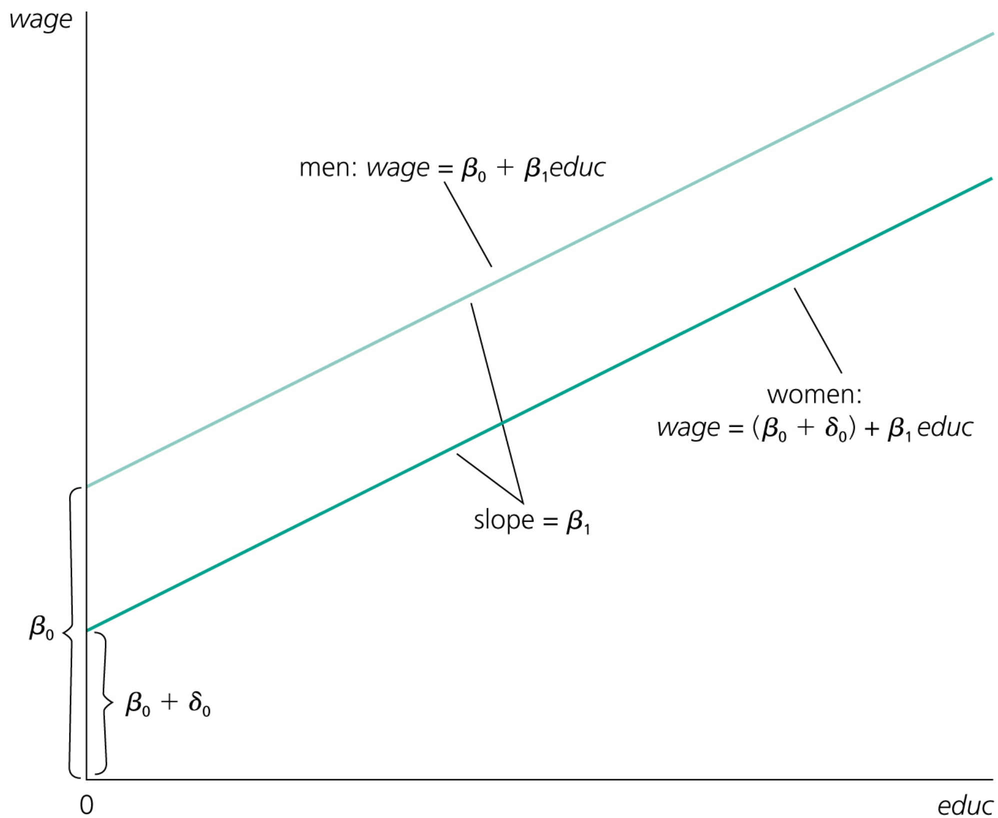
- Sometimes, it will just say that " x_2 is a binary indicator variable that takes on the value of 1 if..."
- In Wooldridge, it just says $y = \beta_0 + \delta_0 female + \cdots + u$
  - You are left to infer that $female$ is either $\{0, 1\}$.
- There are other ways that a dummy variable may be indicated as well, but almost all authors will describe when a dummy/binary/indicator is being used.
- The "dummy' allows y to vary by one discrete amount (δ_0 here) when the condition is true.

Clearly, the indicator must refer to something observable in the data

- They aren't magical!

The "separate intercept" interpretation

Wooldridge frames the coefficient on the binary variable as **intercept shift** between females and males.



For males, the intercept is β_0. For females, $\beta_0 + \delta_0$ (here $\delta_0 < 0$).

Since a binary variable is always either $\{0, 1\}$, it always shifts by one constant amount

- Just like the Wooldridge Fig 7-1

It doesn't alter the *slope* of the line directly, but it *can* account for variation (higher average wages for men) that then allows the slope to be better estimated

- So the slope may be different with the dummy included:

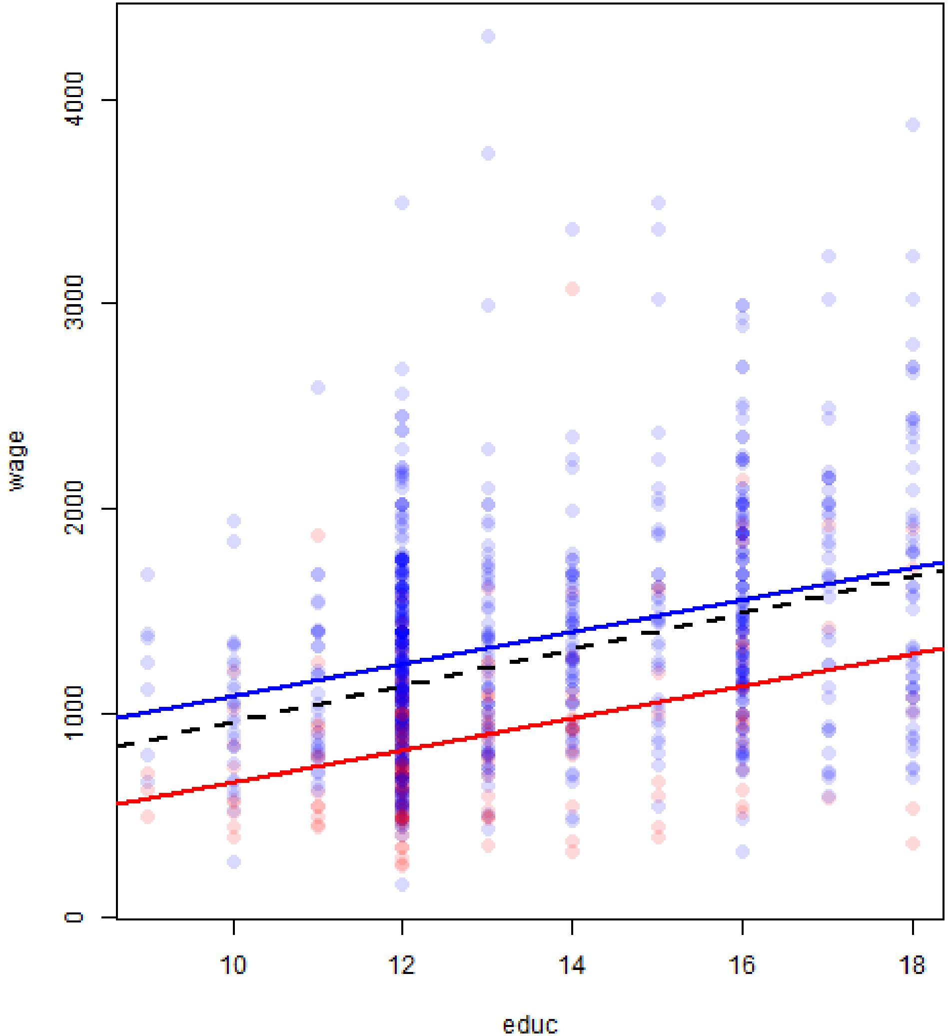
$$wage = \beta_0 + \beta_1 educ + u$$

  and

$$wage = \alpha_0 + \alpha_1 1(female) + \alpha_2 educ + u$$

will not result in $\beta_1 = \alpha_2$. They will be different estimates.

Dummy Variables



The black dashed line is the combined regression ignoring $female$

The blue is the fitted regression for $female == 0$, the red for $female == 1$

Remember, you're adding a variable, and adding a variable can only *help* explain more variation (see our discussion on R2 and F-tests)

Dummy Variables *with* continuous variables

$$OutOfPocket = \beta_0 + \beta_1 1(age > 65) + \beta_2 cigarettes + u$$

Here, $OutOfPocket$ is the annual dollars spent out of pocket on healthcare.

- We think it is affected by number of cigarettes smoked
- We think it might be affected by age

So why not just use the variable itself?

- Why a dummy $1(age > 65)$ and not just age as a RHS x?

So why not just use the variable itself?

- Why a dummy $1(age > 65)$ and not just age as a RHS x?

- First, we may not want to impose that constant marginal effect - sure, we could have β_{age}, but it means we'd be assuming the same effect of age from 10 years old to 11 years old as we do from 64 years old to 65!
- Second, there may be a "threshold" we're interested in
  - For example, Medicare starts at 65 years old.
  - Then being over 65 (and being on Medicare) would have an important effect to account for.
  - And we certainly wouldn't want age alone to try to explain it!

In fact, we could include age and the dummy variable:

$$OutOfPocket = \beta_0 + \beta_1 1(age > 65) + \beta_2 age + \beta_3 cigarettes + u$$

Here's what that data would look like:

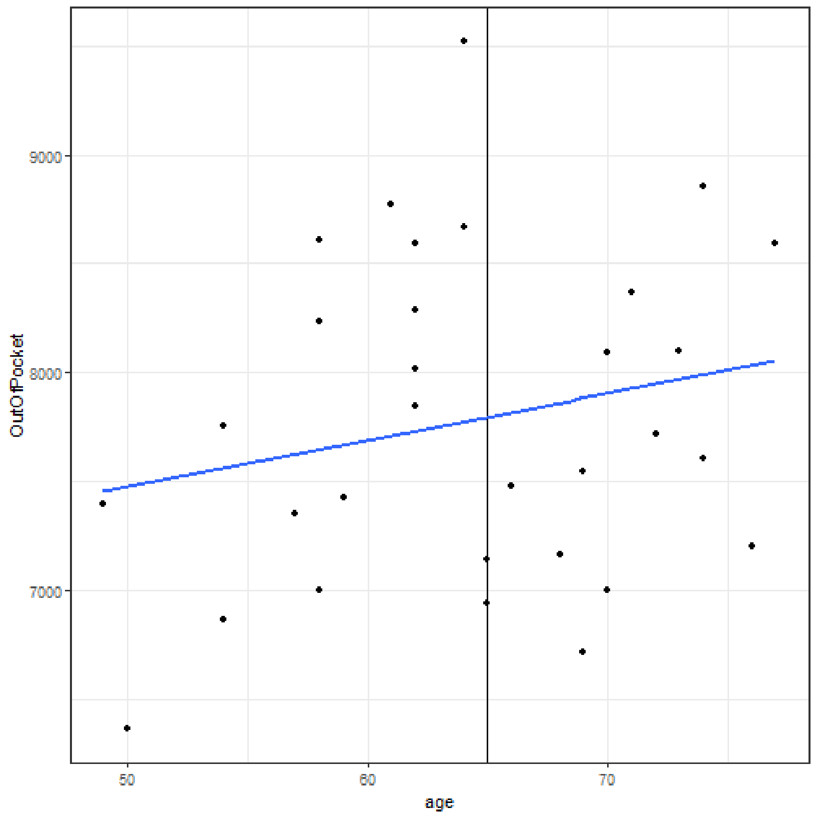
Out of Pocket	Age	1(age>65)
7782	48	0
8136	63	0
9730	86	1
7928	66	1
...	...	...

As you can see, $Over65$ is fully determined by age, but that's OK. They will not be perfectly correlated (correlation is a linear concept).

Let's see how this compares to

- Just using age
- Just using the dummy
- Both

First, ignoring the over65 dummy, just using Out-Of-Pocket health spending on age:



I'm not going to include *cigarettes* here since it adds another dimension to plot

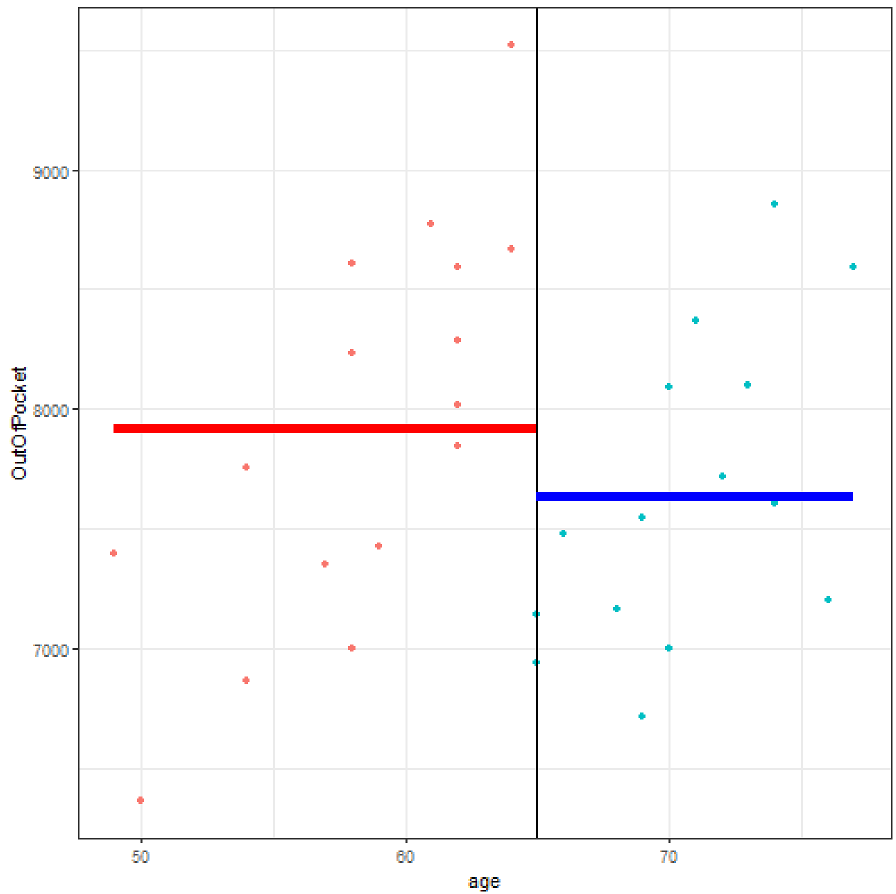
Dummy Variables

$$OutOfPocket = \beta_0 + \beta_1 age + u$$

```
coeftest(lm1, vcov = vcovHC(lm1, 'HC1'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 6408.317   1070.088  5.9886 1.643e-06 ***
## age           21.368     16.419  1.3014    0.2034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here's what just including $1(age > 65)$ looks like

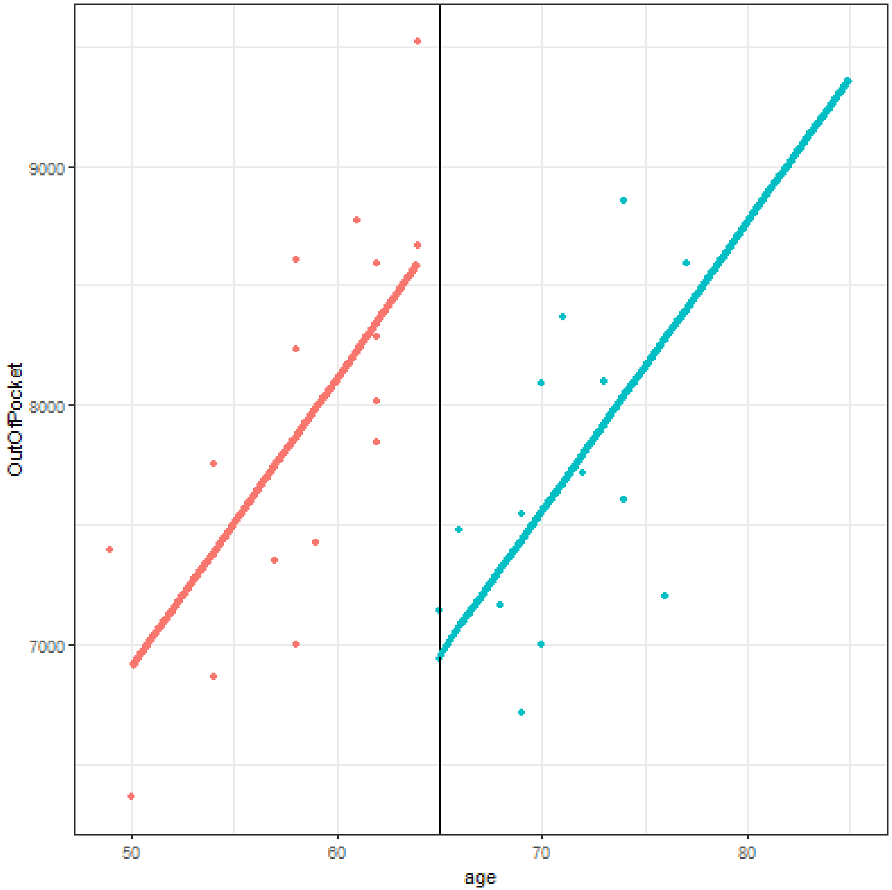


$$OutOfPocket = \beta_0 + \beta_1 1(age > 65) + u$$

```
coeftest(lm1b, vcov = vcovHC(lm1b, 'HC1'))
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7919.62     207.21 38.2207   <2e-16 ***
## over65TRUE   -284.28     265.53 -1.0706   0.2932
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here's what that looks like including both age and $1(age > 65)$:



Dummy Variables

```r
coeftest(lm2, vcov = vcovHC(lm2, 'HC1'))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    869.162   1456.278  0.5968    0.5554
## age            120.779     24.721  4.8857 3.791e-05 ***
## over65TRUE   -1760.798    316.821 -5.5577 6.055e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One interpretation of β_0 is "the expected value of y when $x = 0$"

- I'm going add $cigarettes$ back in here:

$$OutOfPocket = \beta_0 + \beta_1 1(age > 65) + \beta_2 cigarettes + u$$

- When does $x = 0$ here?
- So, what is the $E[Y|age < 65, cigarettes == 0]$?
- What is the $E[Y|age > 65, cigarettes == 0]$?

That seems like a comparison of means because it is.

```
t.test(OutOfPocket ~ over65, data=df)
```

```
##
##  Welch Two Sample t-test
##
## data:  OutOfPocket by over65
## t = 1.0708, df = 28.076, p-value = 0.2934
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -259.4510  828.0067
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7919.622            7635.345
```

Compare that to the first regression with only a dummy for $1(age < 65)$

Interpretation of Dummy Variables

The dummy variable has a "base" level that is *included in* β_0

- And the coefficient on the dummy **is the difference between the base level and the "dummy is true" level**
  - This is because $\beta_0 = E[Y|X=0]$ for all X
- If there are two dummies, x_1 and x_2:
  - β_0 is the $E[Y|x_1=0, x_2=0]$
  - That is, it is the value when both are "false"
  - And β_1 is the relative value if **only** x_1 were true, **ceteris paribus**
  - Same for β_2, **ceteris paribus**
- It does *not* tell us anything about x_1 and x_2 being true together, except that we can add the effects of x_1 being true and x_2 being true.

Dummy Variables fall under the category of "specification"

- All of the rules about x's still hold
  - MLR3 - No Multicolinearity
- Dummies don't change the way we estimate equations or coefficients
- Dummies don't change our assumptions or use of the residuals $\hat{u}$
- Dummies don't change *how* we calculate $\hat{\beta}$, $se(\hat{\beta})$, or SSR etc.

Dummies *do* (hopefully) improve our model

- By accounting for and explaining variation that continuous variables don't
- And by being "interpretable"
  - Lots of ways we can account/explain variation, but not all are "interpretable"

The dummy variable trap

What if we add a variable for under 65 as well?

Out of Pocket	Age	Over65	Under65
7782	48	0	1
8136	63	0	1
9730	86	1	0
7928	66	1	0
...	...	...	...

MICHIGAN STATE UNIVERSITY

So we can't have $1(age > 65)$ and $1(age < 65)$

- Because MLR.3, no multicolinearity
- We can only *identify* the *difference* between over/under 65.
    - The intercept, β_0 is the intercept for the *base* level
    - The coefficient is the *intercept shift*.

Any questions on dummies?

Panel Data

Panel Data is what we all a dataset where we have multiple observations for each unit of observation

- We have a sample of 100 people
- For each person, we have 12 years of earnings
  - We have $N = 100 \times 12 = 120$

Or

- We have a sample of 15 countries
- For each country we have 30 years of infant mortality rates
  - We have $N = 15 \times 30 = 450$

MICHIGAN STATE UNIVERSITY

Contrast Panel Data with other types of data:

Time series data

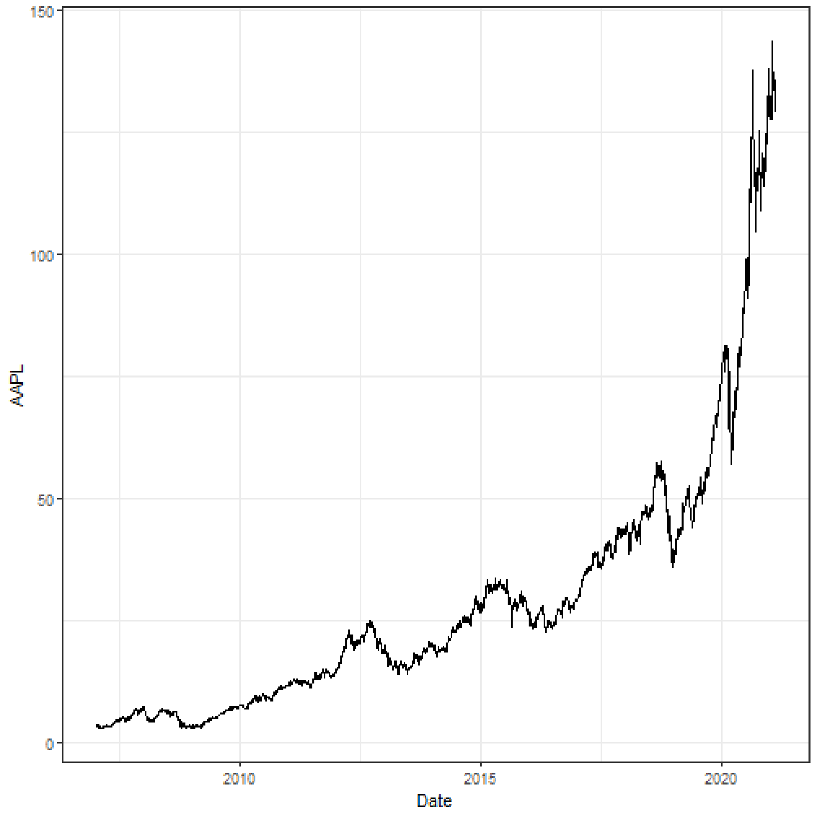
- We have one observation per time period
- But of only one thing.
  - There are no concurrent time periods.

Stock values would be a time series if talking about one stock:

- **AAPL** has one time series of data

MICHIGAN STATE UNIVERSITY

AAPL



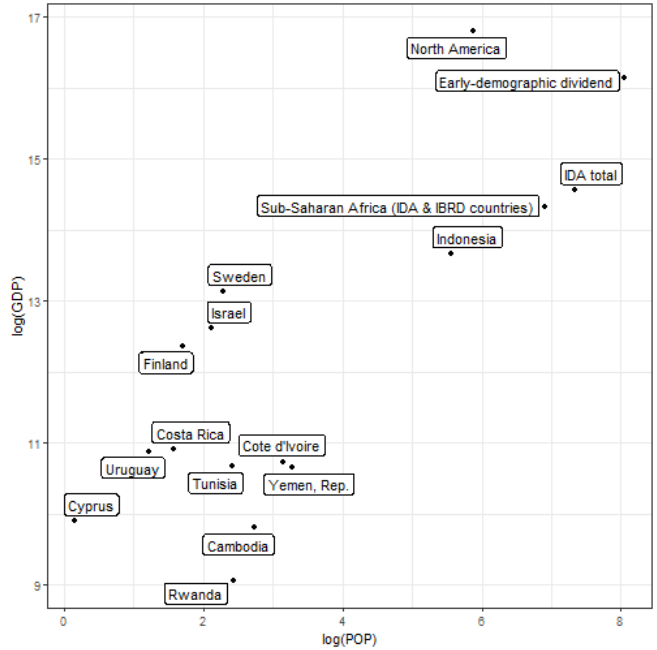
Time series, not panel data.

Contrast Panel Data with other types of data:

Cross-sectional data

- We have multiple observation units, but only one observation of each

Country-level data (for a single year, or average) would be cross-sectional



We have been working with cross-sectional data so far.

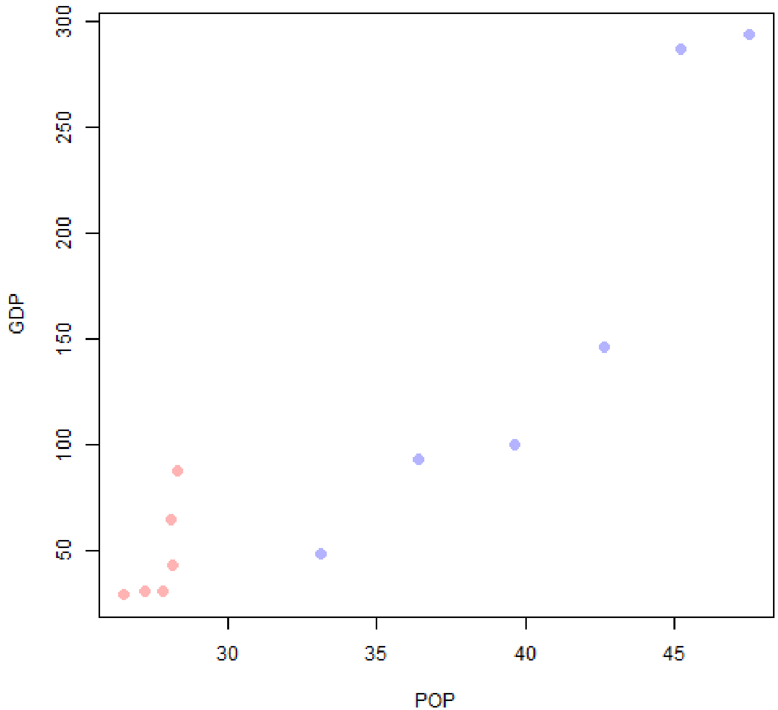
- We will get to time series later on
- Let's focus on Panel Data today

Let's say we had two countries that we observe

- Say, "Cuba" and "Colombia"
- And we observe each one once a year for five years

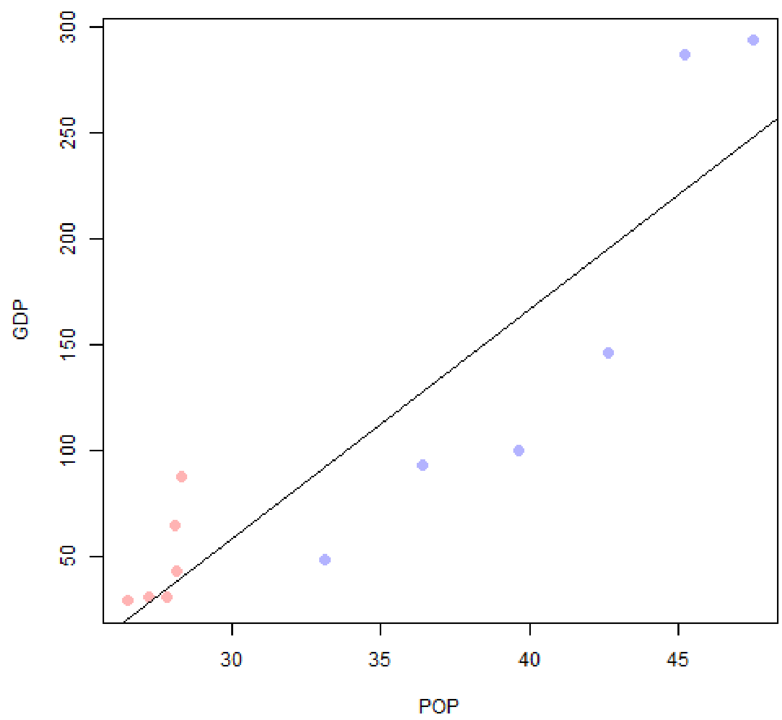
Panel Data

Country	GDP	POP	Year	GDPPC
Colombia	47.8	33.1	1990	1445.3
Colombia	92.5	36.4	1995	2539.9
Colombia	99.9	39.6	2000	2520.5
Colombia	145.6	42.6	2005	3414.5
Colombia	286.6	45.2	2010	6336.7
Colombia	293.5	47.5	2015	6175.9
Cuba	28.6	26.5	1990	2703.2
Cuba	30.4	27.2	1995	2794.7
Cuba	30.6	27.8	2000	2747.1
Cuba	42.6	28.2	2005	3786.7
Cuba	64.3	28.1	2010	5730.4
Cuba	87.1	28.3	2015	7694.0

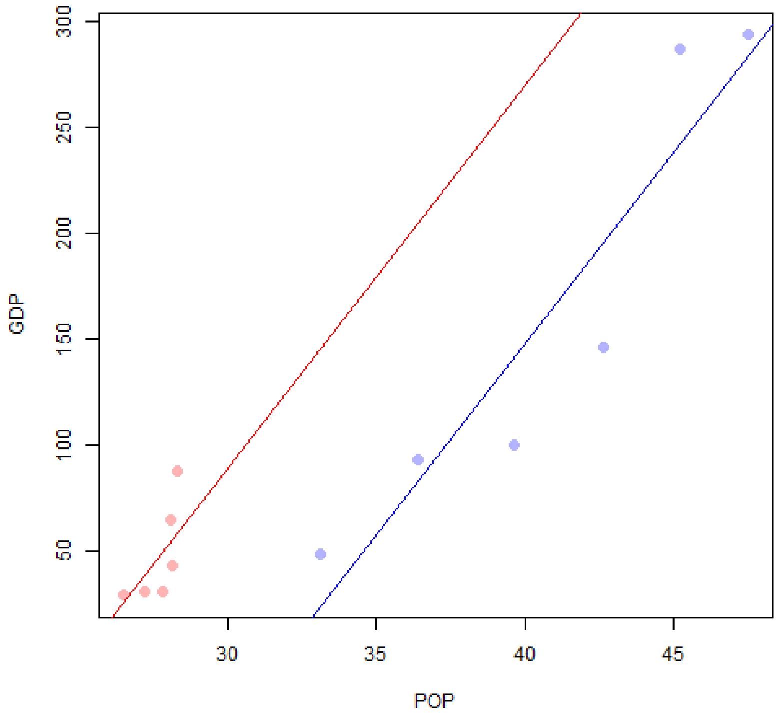

GDP
300
250
200
150
100
50
30
35
40
45
POP

A naive approach

If we are interested in the effect of population on GDP, we might try fitting a line ignoring $Country$



Panel Data



Here, we have included a dummy for $Cuba$.

- The slope is the same across countries (by our specification)
- The intercept is different (though the intercept is very far off the chart here)

```
lm1 = lm(GDP ~ POP + as.factor(Country), data = pop_two)
coeftest(lm1, vcov = vcovHC(lm1, 'HC1'))
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)             -575.5840    93.4567 -6.1588 0.0001669 ***
## POP                       18.0719     2.3715  7.6204 3.257e-05 ***
## as.factor(Country)Cuba   122.7046    31.9972  3.8349 0.0039979 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the previous slide regression:

- This is similar to the male grouping from before
- It has a slightly different interpretation
    - We think there is something unobserved about $Cuba$ that gives it a different average GDP, even conditional on POP.
    - The *dummy* is the *country-level effect* for all of the things about $Cuba$ that change it's GDP overall, independent of POP.

Now, consider that we could have three countries in the data

- We would have one β_0 (the base level)
- And we would have **two** intercept shifts - one for each of the non-base levels

When we allow there to be any number of binary indicators, we call them "fixed effects".

The most common form of Panel Data is Unit x Time

- That's what we have here: We observed Cuba over different time periods
- And Colombia over the same time periods

So the fixed effect captures things about Cuba (relative to Colombia) that do not differ over time

- Things that are always there

Of course, we can also have time fixed effects!

- If there is something different about, say, 2009 that is the same across multiple countries
- Like, say, a global recession...