Multivariate Regression

EC420 MSU Online

Justin Kirkpatrick
Last updated June 05, 2021

This Deck

Lectures:

(1) Motivation

(2) Estimating Multivariate Regression and Partialling Out

(3) Goodness of fit and Specification Errors

(4) Variance of the estimator

(5) Perfect multicolinearity

(6) Distribution of the estimator and Gauss-Markov for MLR

(7) Heteroskedasticity

(8) Testing Multiple β's

Next page...

This Deck

Lectures:

(9) F-tests

(10) Testing for Heteroskedasticity

(11) Asymptotic normality and consistency

(12) Dummy variables

(13) Panel Data

(14) Fixed effects with multiple groups

(15) Interactions with dummies

Motivation

top

Multivariate regression

Multivariate regression is the estimation of the PRF:

$$E[Y|X_1, X_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Where we previously had PRF:

$$E[Y|X] = \beta_0 + \beta_1 x$$

The SRF for two variables is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

We still have one error term, u

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + u_i$$

And we estimate $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$ the same way.

Multivariate regression

Examples

- We want to explain country-level *life expectancy* as a function of *gdp per capita* and *population growth*.

$$LifeExp_i = \beta_0 + \beta_1 gdppc_i + \beta_2 popgrowth_i + u_i$$

- We want to explain *mortality rate* with *number of cigarettes smoked* and *average daily caloric intake*

$$Mortality_i = \beta_0 + \beta_1 cigarettes_i + \beta_2 calories_i + u_i$$

- We want to explain *wage* with *education* and *ability*:

$$Wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

Ceteris Paribus - *all else held equal*

$$Wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

We interpret β_1 as "the effect of $educ$ on the expectation of $wage$, *all else held equal*"

What other random variables are we holding equal?:

- $ability$
- u too!

This means:

$$\beta_1 = \frac{\Delta Wage}{\Delta educ} \text{ when } \underbrace{\Delta ability = \Delta u = 0}_{\text{all else held equal}}$$

And a similar interpretation for β_2

We interpret β_2 as "the effect of $\boldsymbol{ability}$ on the expectation of $\boldsymbol{wage}$, *all else held equal*"

This means:

$$\beta_2 = \frac{\Delta Wage}{\Delta ability} \text{ when } \underbrace{\Delta educ = \Delta u = 0}_{\text{all else held equal}}$$

Does this require that $\frac{\Delta educ}{\Delta ability}$ be zero?

Nope. But we are measuring the effect of one *while holding the other equal to zero.*

We can interpret β_0 as:

$\beta_0 = \boldsymbol{E}[\boldsymbol{wage}]$ when $\boldsymbol{educ}$ and $\boldsymbol{ability}$ **and** $u = 0$

This is because:

$$\boldsymbol{E}[\boldsymbol{wage}|\boldsymbol{educ}, \boldsymbol{ability}] = \beta_0 + \beta_1 \boldsymbol{educ} + \beta_2 \boldsymbol{ability}$$

or, in general notation

$$\boldsymbol{E}[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

A slightly different interpretation, and unique to the β_0.

What if we *should* use two variables, but we only use one?

We could run the regression $wage_i = \beta_0 + \beta_1 educ_i + u_i$, but we probably think $ability_i$ also affects wages.

What if we don't observe $ability_i$?

- Just because we don't observe it doesn't mean it isn't affecting $wage_i$
- It *is present in the error term.*
- Let's make a new variable called $\tilde{u} = \delta_1 ability_i + u_i$

$$wage_i = \beta_0 + \beta_1 educ_i + \underbrace{\delta_1 ability_i + u_i}_{\tilde{u}_i}$$

- Note: usually the $\sim$ over a coefficient or variable (like $\tilde{u}$) will indicate it is related to, but different from, the non- $\sim$ version.

- δ_1 is just the effect of $ability$ on $wage$

We can naively write this as a single variable regression:

$$wage_i = \beta_0 + \beta_1 educ_i + \tilde{u}_i$$

But wait!

- We think $E[ability|educ] > 0$
  - Then this violates the assumption that $E[\tilde{u}|X] = 0$

Because

1. $\frac{\Delta \tilde{u}}{\Delta ability} = \delta_1 \neq 0$
2. $\frac{\Delta ability}{\Delta educ} \neq 0$

$\Rightarrow \frac{\Delta \tilde{u}}{\Delta educ} \neq 0$

Bias

Recall that we could show $E[\hat{\beta}_1] = \beta_1$ if and only if $E[u|X] = 0$.

Looking at $\tilde{u}_i = \delta_1 ability_i + u_i$, we can see why $E[\tilde{u}|educ] \neq 0$

- Thus, β_1 in the single-variable regression was **biased**.

Multivariate regression

Adding in $ability$ as a second variable fixes this:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

because $\boldsymbol{u_i}$ does not change with $\boldsymbol{educ}$ or $\boldsymbol{ability}$.

Now, $\boldsymbol{E[u|X_1, X_2] = E[u|educ, ability] = 0}$

We brought the problem, $\boldsymbol{ability}$, out of the error term.

Multivariate regression allows us to account for the effect of both $ability$ and $educ$

$$wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

And we can calculate the change in $wage$ from any change in $ability$ and $educ$ using the SRF:

$$\Delta\widehat{wage} = \widehat{\beta_1}\Delta educ + \widehat{\beta_2}\Delta ability$$

This also means our assumption is now:

$$E[u|educ, wage] = 0$$

Which we write in general as:

$$E[u|x_1, x_2] = 0$$

Which is to say that we think we've got everything that could potentially be correlated with x_1 and x_2 out of the error term.

Since we want to work with some sample data (wage2.dta from Wooldridge), let's replace $ability$ with $experience$, which is in the dataset....

Multivariate regression

When we estimate $wage_i = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + u_i$, we are fitting a *plane*:

Multivariate regression

When we estimate $wage_i = \beta_0 + \beta_1 educ_i + \beta_2 experience_i + u_i$, we are fitting a *plane*:

Multivariate regression

The best fit is no longer a line, but a *plane* with a slope in the $\boldsymbol{educ}$ and the $\boldsymbol{exper}$ axis of $\beta_{\mathbf{1}}$ and $\beta_{\mathbf{2}}$

Fitted values from a regression on the data in the previous slide where $\beta_{educ} = \mathbf{76.22}$ and $\beta_{exper} = \mathbf{17.64}$.

Estimating Multivariate Regression

and Partialling Out

top

How do we estimate $\hat{\beta}$?

Remember our two assumptions that let us derive $\hat{\beta}_1$ and $\hat{\beta}_0$:

$$E[u] = 0, \quad E[u|x] = 0$$

Now, we want to estimate $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$

- And we have two x's: x_1 and x_2.

$$E[u] = 0, \quad E[u|x_1] = 0, \quad E[u|x_2] = 0$$

We have **three** moment conditions, and three unknowns to estimate. We can do that!

These three moment conditions give us the following to start with:

$$E[y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2] = E[u] = 0$$

$$E[x_1(y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2)] = E[x_1 u] = 0$$

$$E[x_2(y_i - \beta_0 - \beta_1 x_1 - \beta_1 x_2)] = E[x_2 u] = 0$$

Multivariate regression

But don't worry, we won't derive them directly from this, but that's how we would do it.

Multivariate regression

Let's talk notation for a second:

- I will use β as the coefficients we are estimating ($\hat{\beta}$)
- When talking about the right hand side (the covariates), I'll either call them x_1, $x_2, \cdots$
  - Or sometimes just using the variable names: $wage = \beta_0 + \beta_1 educ$
  - Or sometimes with subscripts: $y = \beta_0 + \beta_{educ} x_{educ} + u$
- And sometimes, if we want to emphasize that two regressions are wholly different, I will use δ or γ instead of β
- u and v will represent errors in two different regressions

Partialing out

Imagine if we had two x's, x_{temp} and x_{rain} that both had an effect on y, but were closely related.

We could look at β_{temp} in

$$y = \beta_0 + \beta_{temp} x_{temp} + u$$

And β_{rain} in

$$y = \beta_0 + \beta_{rain} x_{rain} + u$$

We learned from the previous section that β_{temp} is biased when x_{rain} is in the error term u, and vice versa.

That is, β_{temp} is going to "pick up some of the effect" of x_{rain}.

Partialing out

So, would $\tilde{\beta}_{temp}$ in the following (correct) specification equal β_{temp} from the previous slide?

$$y = \tilde{\beta}_0 + \tilde{\beta}_{temp} x_{temp} + \tilde{\beta}_{rain} x_{rain} + \tilde{u}$$

No!

Once we include both variables, we will get a different estimate for $\tilde{\beta}$ than before since each effect is isolated (*ceteris paribus*)

So, to calculate the correct $\tilde{\beta}_{temp}$, we need estimates that take the effect of $\tilde{\beta}_{rain}$ into consideration.

This is called **partialing out**.

One way we can estimate unbiased β_{temp} is the following way:

First, estimate the regression of x_{temp} on x_{rain}

$$x_{temp} = \delta_0 + \delta_{rain} x_{rain} + v$$

Couple of things:

- x_{temp} is on the left hand side.
- **We are "explaining temperature with rainfall"**
- Using δ to show that these are different coefficients

That error term, v has an interpretation

- v is the $temp$ that is **not explained by** $rain$
- $\delta_{rain} x_{rain}$ is the $temp$ that **is** explained by $rain$

v is $temp$ that has had $rain$ "partialed out"

Of course, we have a sample analog for v, the SRF residuals:

$$\hat{v} = x_{temp} - (\hat{\delta}_0 + \hat{\delta}_{rain} x_{rain})$$

Remember, $\boldsymbol{v}$ still varies along with $\boldsymbol{x}_{temp}$, but it is not correlated at *all* with $\boldsymbol{x}_{rain}$.

Multivariate regression

Now, if we want to get the correct value for β_{temp} in the full regression:

$$y = \gamma_0 + \gamma_1 \hat{\boldsymbol{v}} + \boldsymbol{u}$$

- We do not use $\boldsymbol{x}_{rain}$ directly.
- We use $\hat{\boldsymbol{v}}$ and leave $\boldsymbol{x}_{rain}$ out.
- $\hat{\boldsymbol{v}}$ is correlated with $\boldsymbol{x}_{temp}$, but not with $\boldsymbol{x}_{rain}$
- Put another way, $\hat{\boldsymbol{v}}$ contains only the part of $\boldsymbol{x}_{temp}$ that is not correlated with $\boldsymbol{x}_{rain}$.

One can show that $\gamma_1 = \tilde{\beta}_{temp}$

One can show that $\gamma_1 = \tilde{\beta}_{temp}$, the unbiased estimate.

That is, we get the (unbiased) coefficient one would get from regressing

$$\boldsymbol{y} = \tilde{\beta}_0 + \tilde{\beta}_{temp}\boldsymbol{x}_{temp} + \tilde{\beta}_{rain}\boldsymbol{x}_{rain} + \tilde{\boldsymbol{u}}$$

by first "partialing" $\boldsymbol{x}_{rain}$ out of $\boldsymbol{x}_{temp}$ then regressing what is left on $\boldsymbol{y}$.

Multivariate regression

Similarly, one can do the same for $\boldsymbol{x}_{rain}$:

$$\boldsymbol{x}_{rain} = \boldsymbol{\kappa}_0 + \boldsymbol{\kappa}_1 \boldsymbol{x}_{temp} + \boldsymbol{w}$$

Then use the residuals, $\hat{\boldsymbol{w}}$, in a regression:

$$\boldsymbol{y} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \hat{\boldsymbol{w}} + \boldsymbol{\epsilon}$$

And $\tilde{\beta}_{rain} = \boldsymbol{\alpha}_1$

Since $\hat{\tilde{\beta}}_{rain} = \hat{\boldsymbol{\alpha}}_1 = \frac{Cov(y, \hat{w})}{Var(\hat{w})}$, we can say that β_{rain} is the effect of $\boldsymbol{x}_{rain}$ *once we've taken out the effect of* $\boldsymbol{x}_{temp}$ and vice versa.

Multivariate regression

Since we get the same $\hat{\beta}_1$ if we

- Partial out the effect of x_2 and run a single variable regression, or
- Run a two-variable regression

Then we can think of the $\tilde{\beta}_1$ in:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{u}$$

As the effect of x_1 on y **after partialing** x_2 **out of** x_1, and vice versa.

Multivariate regression automatically partials out each of the $x's$.

Let's compare a simple and multiple regression estimates

- $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{u}$
- $y = \beta_0 + \beta_1 x_1 + u$

How will $\tilde{\beta}_1$ differ from β_1?

It depends on the relationship between x_2 and x_1

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If we take $x_2 = \delta_0 + \delta_1 x_1 + v$ and substitute it into the first equation above:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2(\delta_0 + \delta_1 x_1 + v) + \tilde{u}$$

$$y = \tilde{\beta}_0 + \tilde{\beta}_2 \delta_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 \delta_1 x_1 + \tilde{\beta}_2 v + \tilde{u}$$

$$y = \tilde{\tilde{\beta}}_0 + (\tilde{\beta}_1 + \tilde{\beta}_2 \delta_1) x_1 + \tilde{v}$$

Multivariate regression

Therefore it is true that:

- $\hat{\beta}_1 = \hat{\tilde{\beta}}_1 + \hat{\tilde{\beta}}_2 \hat{\delta}_1$
  - In words: to whatever extent $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are correlated ($\boldsymbol{\delta_1}$), our naive $\hat{\beta}_1$ will include that correlation.

Knowing this, when would the simple regression estimate $\hat{\beta}_1$ **equal** the multiple regression (multivariate) estimate $\hat{\tilde{\beta}}_1$?

When $\boldsymbol{\delta_1 = 0}$

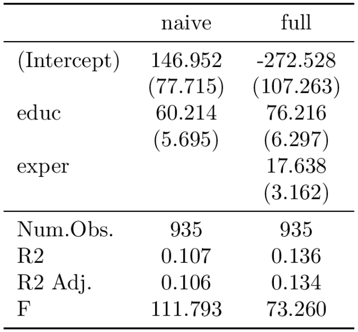
When $\boldsymbol{\tilde{\beta}_2 = 0}$

Will this hold empirically?

Will $\hat{\beta}_1$ change when you add in $\hat{\beta}_2$?

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

$$wage = \tilde{\beta}_0 + \tilde{\beta}_1 educ + \tilde{\beta}_2 exper + \tilde{u}$$



	naive	full
(Intercept)	146.952	-272.528
	(77.715)	(107.263)
educ	60.214	76.216
	(5.695)	(6.297)
exper		17.638
		(3.162)
Num.Obs.	935	935
R2	0.107	0.136
R2 Adj.	0.106	0.134
F	111.793	73.260

Multivariate regression

$$exper = \delta_0 + \delta_1 educ + u$$

Will this hold empirically?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.78	0.79	30.03	0
educ	-0.91	0.06	-15.63	0

Using the coefficients from the previous slide, we see it works:

$$60.21 = 76.22 + (\underbrace{17.64}_{\hat{\tilde{\beta}}_2} \times \underbrace{-.91}_{\hat{\delta}_1})$$

Multivariate regression

To summarize:

- Leaving a variable out can bias our estimate of β_1
- We need to account for the variation in x_1 explained by x_2
- We can do this by partialling out x_2 from x_1
- The result will change our estimate of β_1
- We can even sign and calculate the bias

Goodness of fit and Specification Errors

We can still use the same formula for R^2

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \hat{\bar{y}})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

This is because R^2 only uses the fit of the whole model, determined by how well $\hat{y}$ fits.

When we add one more variable to our model

- The denominator doesn't change (SST doesn't change)
- The numerator, though, gets weakly larger (SSE increases and SSR decreases)
- Intuitively, SSR decreases because more variables can only provide more explanatory power. $\hat{u}_i$ can only get smaller with more variables. If it didn't, the Least Squares estimate would be to set the new variable's $\beta = 0$.
    - $\Rightarrow$ R^2 never gets smaller when variables are added.

Multiple regression is easily extended to many variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

And multiple variables, we can extend the partialing out in the following manner:

- $x_1 = \delta_0 + \beta_1 x_2 + \beta_2 x_3 + \cdots + \beta_k x_{k+1} + v$
  - v is the part of x_1 that has had $x_2, x_3, \cdots$ partialed out
- $y = \alpha_0 + \alpha_1 \hat{v}$
- β_1 = α_1

You can "partial out" multiple variables, leaving only variation that is uncorrelated with the other variables.

OLS is easily extended from 2 to >2 variables

We might be worried about two *specification errors*

- Including an irrelevant variable.
  - Suppose the "true model" is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  - And we estimate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

  - Then OLS is still an unbiased estimator, since *unbaisedness* holds regardless of the true value of the parameters, even if $\beta_j = 0$ for some j.
  - Including an irrelevant variable will, however, impact the variance of the OLS estimator.

We might be worried about two *specification errors*

- Omitting a relevant variable.

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$$

  - We showed that $\tilde{\beta}_1 = \beta_1$ only when $\beta_2 = 0$ or $\delta_1 = 0$.
  - Size and direction depend on the sign and size of $\beta_2 \delta_1$, which depends on the relationship of the omitted variable and the included variable, x_1, and the outcome variable, y.
  - With multiple regressors, the sign and size may not be clear.
  - We can usually *"sign the bias"* if we
    1. have an idea of what is omitted,
    2. have an idea of how it's correlated with y, and
    3. have an idea of how it's correlated with one or more of $x_1, x_2, \cdots, x_k$

Variance of the estimator

top

Variance of the estimator

Last video we saw:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

is equivalent to "partialing out" each one:

$$y = \beta_0 + \beta_1 \hat{v} + \beta_2 \hat{w} + u$$

Where:

$$x_1 = \delta_0 + \delta_1 x_2 + v$$

and

$$\hat{v} = x_1 - \hat{\delta}_0 - \hat{\delta}_1 x_2$$

$\hat{v}$ is correlated with x_1, but is not correlated with x_2

$$x_2 = \gamma_0 + \gamma_1 x_1 + w$$

and

$$\hat{w} = x_2 - \hat{\gamma}_0 - \hat{\gamma}_1 x_1$$

$\hat{w}$ is correlated with x_2, but is not correlated with x_1

Variance of the estimator

What if we want to do inference?

We need to know the variance of the estimator, $\hat{\beta}_1$ and $\hat{\beta}_2$.

Variance of the estimator

Recall Gauss-Markov Assumptions for **single** variable x

SLR.1: In the population, y is a linear function of the parameters, x, and u:
$y = \beta_0 + \beta_1 x + u$

SLR.2: the sample $(y_i, x_i) : i = 1, 2, \cdots, n$ follows the population model and are independent.

SLR.3: "Sample Variation in the Explanatory (X) Variable". That is, x_i is not the same for all i's.

SLR.4: "Zero conditional mean". $E[u|x] = 0$

SLR.1-SLR.4 $\Rightarrow \hat{\beta}$ is unbiased estimate of β.

SLR.5: $Var[u|x] = \sigma_u^2$ for all x. (*homoskedasticity*)

$\Rightarrow Var(\hat{\beta}) = \frac{\sigma_u^2}{SST_x}$

$Var(\hat{\beta})$ for single-variable regression:

We formulated β up to this point earlier:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum((x_i - \bar{x})u_i)}{SST_x}$$

And then took the variance of this, noting that $Var(\beta_1) = 0$ because it is a (constant) population parameter.

$$Var(\hat{\beta}_1) = \frac{1}{SST_x^2} \times Var\left[\sum(x_i - \bar{x})u_i\right] = \frac{SST_x}{SST_x^2}\sigma_u^2 = \frac{1}{SST_x}\sigma_u^2$$

Last time (or earlier)

And we could estimate σ_u^2:

$$\hat{\sigma}_u^2 = \frac{1}{(N-2)} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{SSR}{N-2}$$

The $N-2$ is because we lost two *degrees of freedom* due to the two restrictions:

- $\sum \hat{u} = 0$
- $\sum \hat{u}_i x_i = 0$

Variance of the estimator

Let's start by looking at σ_u^2

In multiple regression, we have more restrictions:

- $\sum \hat{u} = 0$
- $\sum \hat{u}_i x_{i,1} = 0$
- $\sum \hat{u}_i x_{i,2} = 0$
- $\sum \hat{u}_i x_{i,\dots} = 0$

One for each β. So, when we have $\{\beta_0, \beta_1, \beta_2\}$, we lose 3 degrees of freedom:

$$\hat{\sigma}_u^2 = \frac{1}{(N-3)} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{SSR}{N-3}$$

Generalizing to K variables

- We always count 1 for β_0
- We call the number of X's K
- We would use $\frac{1}{N-K-1}$ if there are K x's.
- N is the number of observations in our regression.

If $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$

- Then we have $N - 2 - 1 = N - 3$ degrees of freedom

If $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$

- Then we have $N - 3 - 1 = N - 4$ degrees of freedom

Variance of the estimator

Since

$$\widehat{Var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{SST_x}$$

and

$$\hat{\sigma}_u^2 = \frac{SSR}{N - K - 1}$$

More X's (more regressors) means the denominator on $\hat{\sigma}_u^2$ gets smaller

$\Rightarrow$ $\hat{\sigma}_u^2$ gets **larger**

$\Rightarrow$ $\widehat{Var}(\hat{\beta})$ gets **larger**

And thus our confidence intervals get larger, rejection region gets smaller, and we lose *precision*

Variance of the estimator

So we know how to calculate our multivariate $\hat{\sigma}_u^2$.

What about the rest of $\frac{\hat{\sigma}_u^2}{SST_x}$? What is SST_x?

- Before, we had one x, so SST_x was straightforward.
- Now, we have 2 or more x's.

Each x has its own SST_x:

$$SST_{x_k} = \sum_{i=1}^{n} (x_{i,k} - \bar{x}_k)^2$$

Note that we are summing the x_k over all i.

Variance of the estimator

SST_x example

The data:

x1	x2
1	-2
5	0
6	2

$$\bar{x}_1 = 4$$
$$\bar{x}_2 = 0$$

Which results in SST_x's of:

SST_{x_1} = 14
SST_{x_2} = 8

Variance of the estimator

We need to make one more adjustment

We need to account for how unique the variance in each of the x_k's is.

- Imagine if we had two x's: x_j, x_k, but they were *almost* always the same number.
  - Think: temperature and rainfall.
- The estimates of the corresponding β's: β_j and β_k, should have a lot of variance to them because we aren't sure which is *actually* explaining the variation in y.

So, we are going to weight each SST_{x_j} by $(1 - R_j^2)$

Where R_j^2 is the R^2 of the regression $x_j = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v$.

That is, we will weight it by "how well is this variable, x_j, explained by all the other variables"

Variance of the estimator

So with that weighted SST_{x_j}, $Var(\beta_j)$ is:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_{x_j}(1 - R_j^2)}$$

And we can estimate this easily:

$$\widehat{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}_u^2}{SST_{x_j}(1 - \hat{R}_j^2)}$$

Variance of the estimator

When R_j^2 is high:

- Then x_j is explained almost completely by $x_1, \cdots, x_k$ (the other x's)
- R_j^2 is very high
- $(1 - R_j^2)$ is very small, close to 0
- $SST_{x_j}(1 - R_j^2)$ is very small, close to 0
- And thus, $Var(\hat{\beta}_j) = \frac{\hat{\sigma}_u^2}{SST_{x_j}(1-R_j^2)}$ is **very high** when R_j^2 is very high.
  - It is division by a small number near 0

Perfect multicolinearity

top

Perfect multicolinearity

What if $x_j = x_k$?

- What is the R_j^2?
  - What is the R^2 of the regression: $x_j = \beta_0 + \beta_1 x_k$?

If $R_j^2 = 1$, what is $Var(\hat{\beta}_j)$?

A problem, that's what it is.

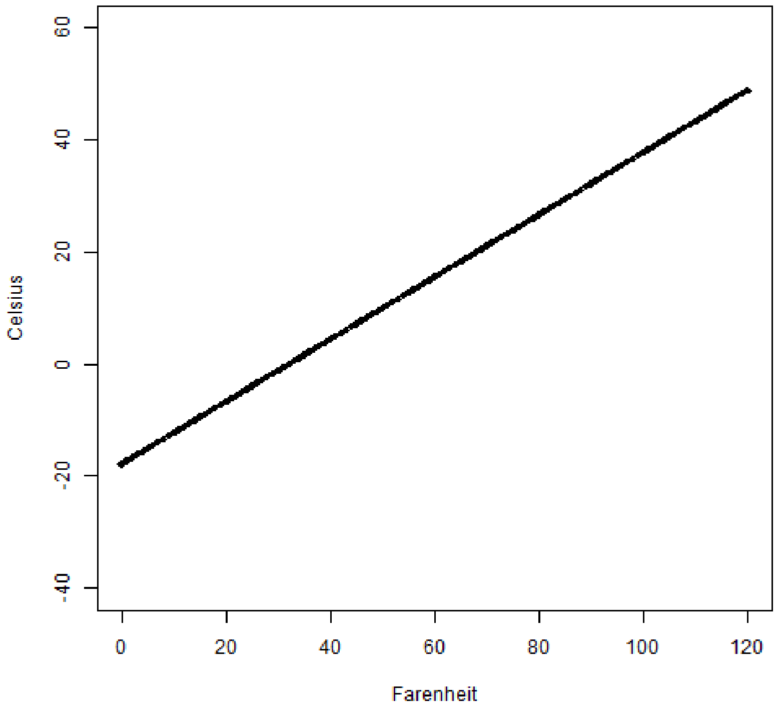
When two x's are perfectly correlated, you have *multicolinearity*

- **Perfect** correlation occurs when $x_j = c + bx_k$, an *affine* transformation
- Degrees farenheit and degrees celsius are a perfect example

Perfect multicolinearity

X degrees Farenheit to Y degrees Celsisus conversion:

$$(X^{\circ}F - 32) \times \frac{5}{9} = Y^{\circ}C$$



```
##
## Call:
## lm(formula = C ~ F, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.210e-14 -7.180e-16  7.020e-16  1.401e-15  6.060e-15
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.778e+01  8.305e-16 -2.141e+16   <2e-16 ***
## F            5.556e-01  1.196e-17  4.644e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.596e-15 on 119 degrees of freedom
## Multiple R-squared:      1,	Adjusted R-squared:      1
## F-statistic: 2.157e+33 on 1 and 119 DF,  p-value: < 2.2e-16
```

Regression of C on F (perfect fit, **note the R^2**)

Perfect multicolinearity

y	degC	degF
2	34	93.2
3	27	80.6
1	30	86.0
3	36	96.8
3	27	80.6

This matrix is not *full rank*

Perfect multicolinearity

So the regression doesn't go so well

```
summary(lm(y ~ degC + degF, degdf))
```

```
## 
## Call:
## lm(formula = y ~ degC + degF, data = degdf)
## 
## Residuals:
##       1       2       3       4       5 
## -0.3234  0.5090 -1.4192  0.7246  0.5090 
## 
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13772    3.89581   0.805    0.480
## degC        -0.02395    0.12561  -0.191    0.861
## degF              NA         NA      NA       NA
## 
## Residual standard error: 1.027 on 3 degrees of freedom
## Multiple R-squared:  0.01198,	Adjusted R-squared:  -0.3174 
## F-statistic: 0.03636 on 1 and 3 DF,  p-value: 0.8609
```

Perfect multicolinearity

We have a bit of a problem when two of our x's are perfectly correlated.

What do we do about this?

In practice, we omit one of the x's

Won't this bias the result?

- Yes and no. If they are perfectly correlated, then *one* of the $\boldsymbol{x}$'s explains exactly what *both* $\boldsymbol{x}$'s could explain.

- But we will never, ever be able to tell *which* one is causal.
    - Think of the temperature example.
    - Can we tell if degrees F has an effect while degrees C doesn't?
    - Of course not!

Distribution of the estimator

and Gauss-Markov for MLR

top

Distribution of the estimator

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: $E[u \vert x_1, x_2, \cdots, x_k] = 0$ for all x.
MLR.5	$Var[u \vert x_1, \cdots, x_k] = \sigma_u^2$ for all x.

A neat thing happens when assumptions 1-5 hold

OLS is B.L.U.E.

- **B**est
  - Has the lowest variance
- **L**inear
  - β is a linear function of the data
- **U**nbiased
  - Is unbiased (showed for single; holds for multiple)
- **E**stimator

Of all linear, unbiased estimators, OLS is the most efficient

Distribution of the estimator

Remember what we needed for inference

- $E[\hat{\beta}] = \beta$
- $Var(\hat{\beta})$
- That $\hat{\beta} \sim N(\beta, Var(\hat{\beta}))$

How do we know it's Normal?

- We will need more assumptions
    - Chapter 5 has weaker assumptions with a similar result

Distribution of the estimator

Assumption MLR.6: Normality of u

We can assume a normal distribution for the OLS estimator, $\hat{\beta}$, by assuming that the errors, u, are normally distributed in the population.

Assume:

$$u|x_1, x_2, \cdots, x_k \sim N(0, \sigma_u^2)$$

Then:

$$y|x_1, \cdots, x_k \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma_u^2)$$

Note that this is the distribution of y conditional on the x's. All of the random variation comes from u. x's conditionally shift the mean deterministically.

I'm adding the subscript u to σ_u^2 for emphasis, but since u is the only source of random variation once we condition on x's, it is implied to be the only σ^2.

How do we get from normal u's to normal β?

Define $\hat{v}_j$ to be the residual of a regression of x_j on all other x's. In a two variable (x_j, x_k) example for observation i:

$$x_{i,j} = \hat{\delta}_0 + \hat{\delta}_1 x_{i,k} + \hat{v}_{i,j}$$

Then $\hat{\beta}_j$ is:

$$\hat{\beta}_j = \frac{\widehat{Cov}(\hat{v}, y)}{\widehat{Var}(\hat{v})} = \frac{\sum_{i=1}^n \hat{v}_{ij} y_i}{\sum_{s=1}^n \hat{v}_{sj}^2} = \sum_{i=1}^n w_{ij} u_i$$

Where:

$$w_{ij} = \frac{\hat{v}_{ij}}{\sum_{s=1}^n \hat{v}_{sj}^2}$$

Thus, $\hat{\beta}_j$ is a linear combination of normals!

Distribution of the estimator

And a linear combination of normals is....normal!

(See Stats Review notes. Told you that property would come in handy).

So now we know :

- $E[\hat{\beta}]$
- $Var(\hat{\beta})$
- That $\hat{\beta}$ really is normally distributed

That's what we need to start testing things!

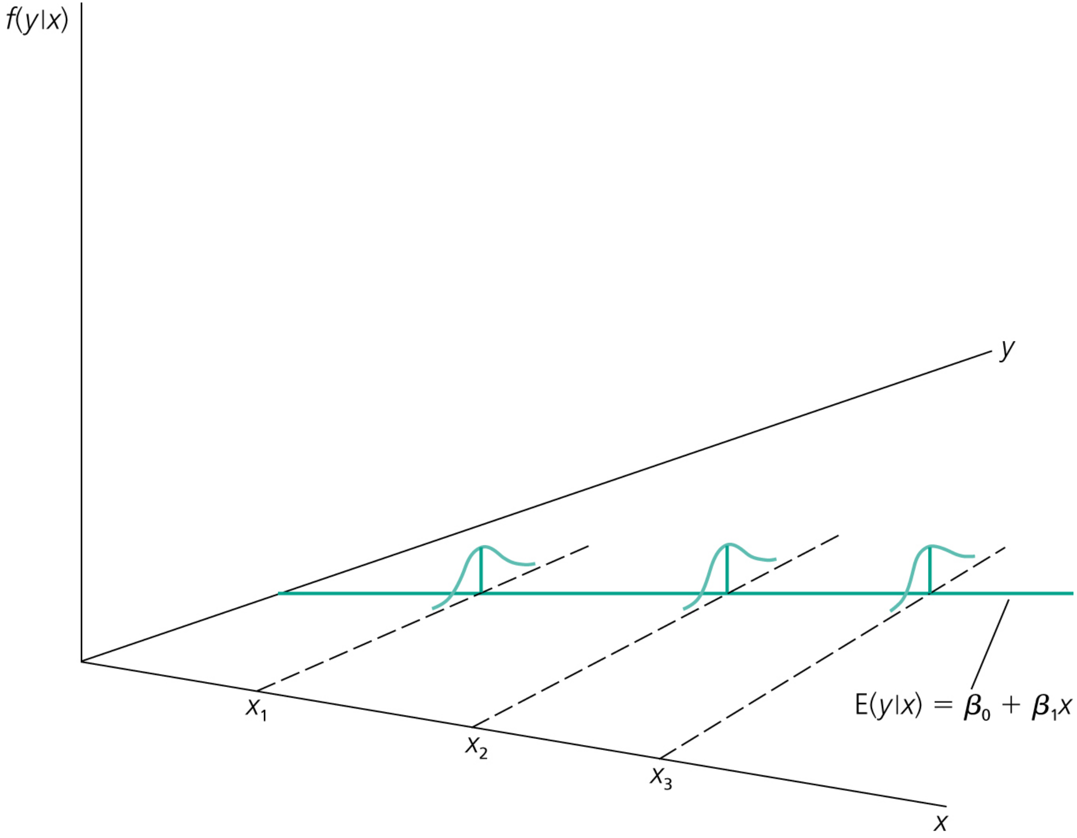
$$\frac{\hat{\beta}-\beta}{se(\hat{\beta})} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\beta}-\beta}{\hat{se}(\hat{\beta})} \sim t_{N-K-1}$$

Heteroskedasticity

top

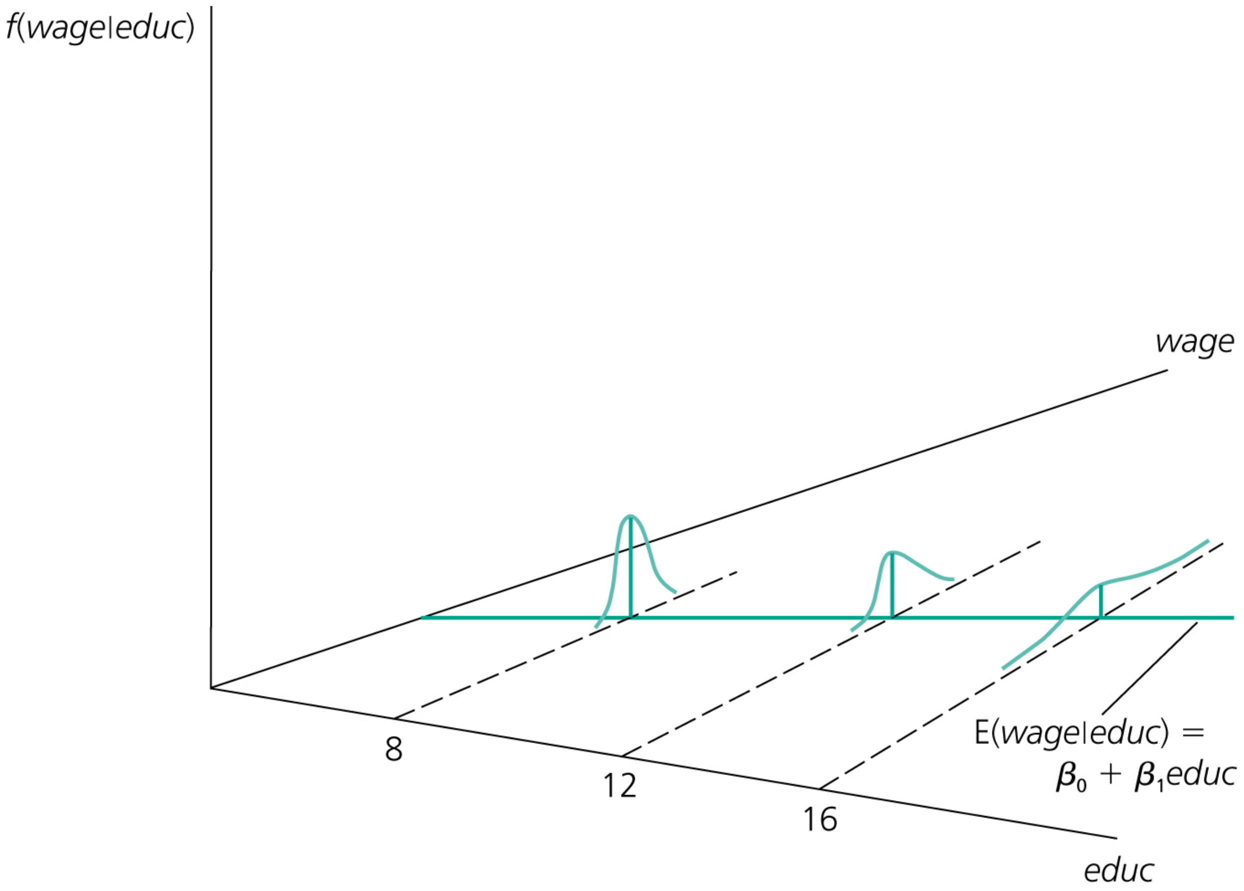
Heteroskedasticity

Now is a good time to revisit MLR.5, homoskedasticity:



Homoskedasticity (from Wooldridge)

And what to do about heteroskedasticity:



Heteroskedasticity (from Wooldridge)

In practice, we have a very useful method of "correcting" for heteroskedasticity called "robust standard errors"

- Eicker-Huber-White Heteroskedasticity-Consistent (HC) errors (1980)

In R, we can compute these errors fairly easily

We'll see in a few slides.

It comes at a cost, though: it inflates errors (make larger)

- Less likely to be "significant" (reject H_0) even if there is evidence to reject H_0.
- That's what it's supposed to do **if** there is heteroskedasticity
- But if there **isn't** heteroskedasticity, you are wasting some power.

Heteroskedasticity-robust standard errors: how do they work?

The problem is that x_j may be correlated with u and thus $\sigma_m^2 \neq \sigma_n^2$ - there is no common, single σ^2.

In the **single variable regression case**, we would account for this:

$$\widehat{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \hat{u}_i^2}{(SST_x)^2}$$

Note that we have squared the sum-of-squares total in the denominator. The numerator looks a little like covariance, but it's more like the covariance of squared terms.

Heteroskedasticity

In multiple regression, things get complicated:

$$\hat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{N} \hat{v}_{ij}^2 \hat{u}_i^2}{(SSR_j)^2}$$

Which looks like the multivariate variance error, but with the extra SSR_j in the denominator, and $\hat{v}$, the residual from x_j on x_i

If we adjust with heteroskedasticity-consistent errors (HC)

then we can relax MLR.5 and still have a *valid* estimate of the variance of $\hat{\beta}$.

Note that **heteroskedasticity-consistent errors** do not **ever** affect the point estimate of $\hat{\beta}$.

- The point estimates remain the same, but the error (and thus the significance) changes.
  - "Point estimate" refers to the value of $\hat{\beta}$, regardless of the variance.

Heteroskedasticity-consistent errors in R

- `install.packages(c('sandwich','lmtest'))`
- `require(sandwich)`
- `require(lmtest)`
- `myOLS = lm(Y ~ X1 + X2, df)`
- `coeftest(myOLS, vcov = vcovHC(myOLS, 'HC1'))`
    - `myOLS` is your linear regression object
    - `vcov` stands for "variance-covariance"
    - The `HC1` gives a specific type of HC errors
    - It is identical to the `, robust` errors in Stata.

If you do not adjust your standard errors, you must justify exactly why you are assuming homoskedasticity.

Heteroskedasticity

```
wage2 = wooldridge::wage2
myOLS = lm(wage ~ educ + exper, wage2)
summary(myOLS)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -924.38 -252.74  -40.88  198.16 2165.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -272.528    107.263  -2.541   0.0112 *
## educ          76.216      6.297  12.104  < 2e-16 ***
## exper         17.638      3.162   5.578 3.18e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 376.3 on 932 degrees of freedom
## Multiple R-squared:  0.1359,    Adjusted R-squared:  0.134
## F-statistic: 73.26 on 2 and 932 DF,  p-value: < 2.2e-16
```

Using HC errors

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)    
## (Intercept) -272.5279   109.8965 -2.4799   0.01332 *  
## educ          76.2164     6.7468 11.2966 < 2.2e-16 ***
## exper         17.6378     3.1126  5.6666 1.941e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 'HC1' yields Stata-type robust errors
  - If you are planning on taking EC422 with Prof. Imberman, use 'HC1'.

There are two meanings of the word "robust" in econometrics

- Robust standard errors, which is what we are discussing here
- A "robust" regression is one that is not affected by a particular specification issue
    - When we saw that we could include unrelated x's and not worry about getting bias, our regression was "robust"

Bonus

Using the `fixest` package

```r
library(fixest)
myFEOLS = feols(wage ~ educ + exper, wage2)
# feols is fixed-effect OLS. We will get to fixed effects soon
summary(myFEOLS, se = 'hetero') # se = 'standard'
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 935
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -272.530   109.9000 -2.4799  0.013319 *
## educ          76.216     6.7468 11.2970 < 2.2e-16 ***
## exper         17.638     3.1126  5.6666  1.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 375.7   Adj. R2: 0.133999
```

`fixest` package's `feols` lets you list the std. error correction in the `summary(...)` call, which is handy. It does a lot more as well.

Testing Multiple β's

top

Testing multiple betas

In economics, we sometimes want to test

- $H_0 : \beta_1 = \beta_2$
- $H_A : \beta_1 > \beta_2$
    - A one-tailed test

Let's look at how we can set this up using the hypothesis testing tools we have:

- $H_0 : \beta_1 = \beta_2$ is the same as $H_0 : \beta_1 - \beta_2 = 0$
- $H_A : \beta_1 - \beta_2 > 0$ follows.

Super - if we have $\hat{\beta}_1$ and $\hat{\beta}_2$, all we need is the $se(\hat{\beta}_1 - \hat{\beta}_2)$.

- This is **not, I repeat NOT** $se(\hat{\beta}_1) - se(\hat{\beta}_2)$.

Testing multiple betas

The formula for the variance of the sum of two random variables

$$Var(\hat{\beta}_1 - \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$$

And the se is the square root of the variance.

We have not been given the $Cov(\hat{\beta}_1, \hat{\beta}_2)$

- It is calculated by R in the variance-covariance matrix
- We won't worry about how. Just know that it is possible

R can be asked to generate that t-statistic and test it

```
vcov(myOLS)
```

Testing multiple betas

The example in Wooldridge, Section 4-4:

$$log(wage) = \beta_0 + \beta_{jc} jc + \beta_{univ} univ + \beta_{exper} exper$$

We'd like to test to see if *years in junior college, jc* have the same effect on log-wages as *years in university*, controlling for experience (ceteris paribus!)

We'd like to test $\beta_{jc} = \beta_{univ}$, which is:

- $H_0 : \beta_{jc} - \beta_{univ} = 0$
- $H_A : \beta_{jc} - \beta_{univ} \neq 0$

Testing multiple betas

Sure, we **could** rewrite the equation so that we get a coefficient that is equivalent to $\beta_{jc} - \beta_{univ}$:

$$log(wage) = \beta_0 + \beta_{jc} jc + \beta_{totcollege}(univ + jc) + \beta_{exper} exper$$

Since $\beta_{totcollege}$ captures both JC and University, β_{jc} captures the difference between the two - exactly what we want to test! The t-stat for that coefficient is our test.

But we could do it with a **linear hypothesis test**

Testing multiple betas

Of course, R does it for us as well without re-writing the equation:

```
library(car)
myOLS = lm(wage ~ jc + univ + exper, df)
linearHypothesis(model = myOLS, hypothesis.matrix = "jc - univ = 0" )
```

This tests if the coefficient β_{jc} is the same as β_{univ}.

- That is, it tests if the effect of $\boldsymbol{jc}$ is the same as the effect of $\boldsymbol{univ}$
- It does so by calculating the $\hat{\boldsymbol{se}}(\hat{\beta}_{jc} - \hat{\beta}_{univ})$

Testing multiple betas

```r
NN = 400
df1 = data.frame(jc = rpois(2, NN),
                 univ = rpois(4, NN),
                 u = rnorm(NN, mean=0, sd = 5)) %>% dplyr::mutate(wage = 10 + 2.5*jc + 2

myOLS<-lm(wage ~ jc + univ, df1)
linearHypothesis(model = myOLS, hypothesis.matrix = 'jc - univ = 0')
```

```
## Linear hypothesis test
##
## Hypothesis:
## jc - univ = 0
##
## Model 1: restricted model
## Model 2: wage ~ jc + univ
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    398 9316.3
## 2    397 9306.4  1    9.9516 0.4245 0.5151
```

Here, I create fake data where I know there are equal effects, and I test $H_0 : \beta_{jc} = \beta_{univ}$. We fail to reject ($p > .05$)

Testing multiple restrictions

and F-tests

top

Testing multiple restrictions

But what if we want to know if more than one coefficient is zero?

- Let's say we have run `lm(khomeprice ~ bedrooms + bathrooms + sqft, df)`
- And we want to know if $\beta_{bedrooms} = \beta_{bathrooms} = 0$.
  - Are all of these coefficients *jointly* zero?
  - Once we account for $sqft$, which is not being tested
  - Is the effect of $bedrooms$ **and** $bathrooms$ zero **all together**.

Testing multiple restrictions

This differs from asking about each of the separately

- It might be that each one has no statistically significant effect, but taken together, they might jointly have some effect.
- It is also asking if these coefficients, together, explain much of y.

This is a *multiple linear restriction* test (W. 4.5)

We can do this in R with `linearHypothesis` as well.

Testing multiple restrictions

```
myOLS3 = lm(khomeprice ~ bedrooms + bathrooms + sqft, df)
coeftest(myOLS3, vcov=vcovHC(myOLS3, 'HC1'))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.272328 237.133362  0.3301   0.7440
## bedrooms      2.265803  50.977959  0.0444   0.9649
## bathrooms     0.926042   1.699717  0.5448   0.5905
## sqft          0.084879   0.105242  0.8065   0.4273
```

Testing multiple restrictions

```
linearHypothesis(myOLS3, c('bedrooms=0','bathrooms=0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## bedrooms = 0
## bathrooms = 0
##
## Model 1: restricted model
## Model 2: khomeprice ~ bedrooms + bathrooms + sqft
##
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     28 2373617
## 2     26 2348144  2     25473 0.141 0.8691
```

We *fail to reject* the null hypothesis that both $\beta_{bedrooms}$ and $\beta_{bathrooms} = 0$ jointly.

Testing multiple restrictions

$H_0 : \beta_{bedrooms} = \beta_{bathrooms} = 0$

$H_A : H_0$ is not true

Rejecting the null hypothesis doesn't tell us which "part" of the hypothesis rejects.

- It doesn't say that bedrooms isn't zero,
- Or that bathrooms isn't zero.

We can think of these joint tests as *restrictions* - we are asking "do these coefficients, jointly, explain any of y?"

Testing multiple restrictions

Testing if a group of coefficients all *jointly* equal zero is a special situation

Saying that β_1, β_2 are jointly zero is the same as saying that β_1, β_2 do not explain any variation in y.

- If they don't explain any variation in y (jointly), then *they can be left out of the model*.
- Testing if they are jointly zero is the same as testing if you can leave them out of the regression entirely.

So how would we test this?

Testing multiple restrictions

Testing these "restrictions"

First, we run the *unrestricted* model:

```
lm(khomeprice ~ bedrooms + bathrooms + sqft, df)
```

Then, we take the SSR, the Sum of Squared Residuals

- $\sum \hat{u}^2$. Call it SSR_{UR}
- The unrestricted (UR) model also has degrees of freedom $N-K-1=N-4-1$

Testing multiple restrictions

Then, we run the *restricted* model:

```
lm(khomeprice ~ sqft, df)
```

- This is "restricted" because we are making a restrictive statement about $\beta_{bedrooms}, \beta_{bathrooms}$
- Specifically, we are saying **that they are equal to zero** in this model!
    - Doesn't get much more restrictive than that, does it?
- The restricted (R) model also has degrees of freedom $N - K - 1 = N - 1 - 1$.

We can compare the SSR of each model.

"Do the increased number of parameters (2) explain enough variance (reduce the variance of u) sufficiently to include them?"

What should our test do?

- If $SSR_R - SSR_{UR}$ is very big, then the unrestricted model (more β's) is more "explanatory", and that **set** of β's are not, jointly, zero.
  - Our test should reject the null that $\beta_{bedrooms} = \beta_{bathrooms} = 0$
- It should account for the difference in degrees of freedom

Testing multiple restrictions

Here's our test statistic, F:

$$F = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{N - K_{UR} - 1}}$$

Where q is the **number of restrictions we are testing**. Here, it is 2.

- Because we are testing $\beta_{bedrooms} = \beta_{bathrooms} = 0$
- q = $df_R - df_{UR}$

Testing multiple restrictions

Here's our test statistic, F:

$$F = \frac{\frac{(SSR_R - SSR_{UR})}{q}}{\frac{SSR_{UR}}{N - K_{UR} - 1}}$$

F has a known distribution:

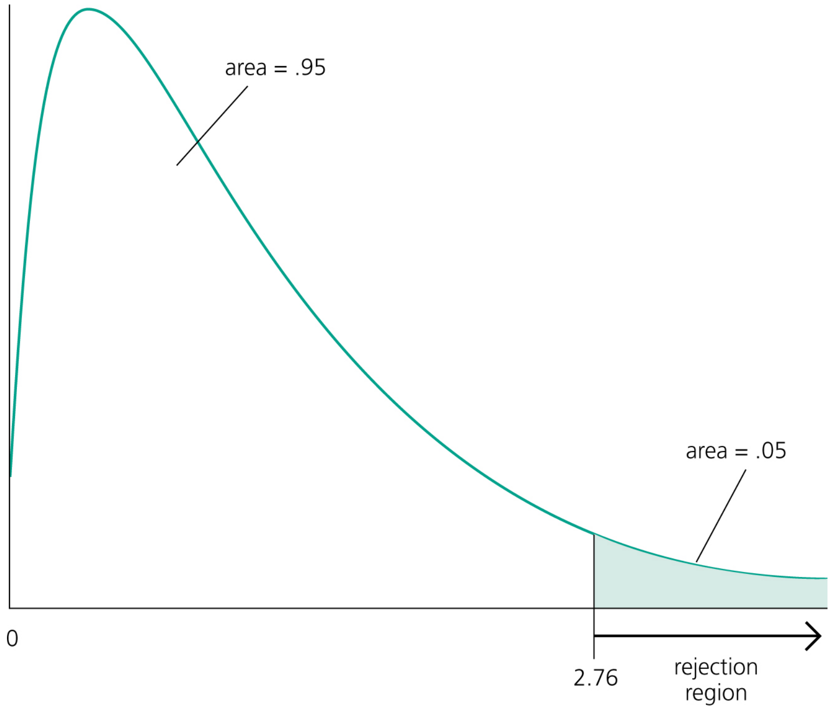
If you recall (you probably don't), we introduced an F-distribution back in stats review.

F is like t - it is defined only by its degrees of freedom.

F, unlike t, takes *two* degrees of freedom: the numerator $(q)$and the denominator $(N - K_{UR} - 1)$.

- $F \sim F_{q, N-K_{UR}-1}$

Testing multiple restrictions



This is the $F_{3,60}$ distribution from Wooldridge Fig 4-7.

- The rejection region is always only on the right ($SSR_R - SSR_{UR} > 0$ always)
- When SSR_R is big relative to SSR_{UR} (the β's being tested explain a lot, making $\hat{u}^2$ smaller), the F-stat is larger
  - Which means it is further out to the right, closer or in the rejection region.

Testing multiple restrictions

If F is big, it is more likely to be in the rejection region

When F is in the rejection region, we reject the $H_0 : \beta_{x_1} = \beta_{x_2} = \beta_{x_3} = 0$

When we reject H_0, all of these are true:

- Jointly, the coefficients are not all zero
- **The unrestricted model (the one with all coefficients in it) is a better model**
- It has sufficiently better explanatory power to justify the extra coefficients.

R automatically gives us an F-stat

It is the F test where the restriction is that all β's except the constant are 0

- Which is saying "does this model, with all it's coefficients and RHS variable x's, explain y any better than just using β_0.
- A model with only β_0 is equivalent to just guessing $\bar{y}$, and not using *any* of the x's.
- So the F test that R outputs is a test for whether or not all the coefficients (except the intercept, β_0) are zero.

Testing multiple restrictions

```
##
## Call:
## lm(formula = khomeprice ~ bedrooms + bathrooms + sqft, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.68 -157.45  -35.56  205.13  615.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 78.27233  252.34541   0.310    0.759
## bedrooms     2.26580   57.98750   0.039    0.969
## bathrooms    0.92604    1.81426   0.510    0.614
## sqft         0.08488    0.13514   0.628    0.535
##
## Residual standard error: 300.5 on 26 degrees of freedom
## Multiple R-squared:  0.1535,    Adjusted R-squared:  0.05582
## F-statistic: 1.572 on 3 and 26 DF,  p-value: 0.2201
```

Testing multiple restrictions

So we know

- How to test a hypothesis about a single coefficient
- How to test a joint hypothesis about multiple coefficients
- How to test if many coefficients are jointly zero
- What the F test R gives us is testing:
    - Whether or not all the coefficients, jointly, are zero
    - Which is the same as saying whether or not all the coefficients, jointly, explain y any better than just using β_0

Testing for heteroskedasticity

top

Testing for Heteroskedasticity

How do we know for sure

that we should be concerned about heteroskedasticity?

Like the F test for whether or not some coefficients are jointly zero, we have a test for heteroskedasticity.

We can use it to see if we need to apply our Heteroskedasticity-consistent errors (HC)

The test follows from the notion that σ_u^2 might increase **with one of our X**'s.

Breusch-Pagan Test for Heteroskedasticity

1. Estimate our model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

2. Calculate $\hat{u}$ and $\hat{u}^2$.

3. Regress $\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + v$

4. Check the p-value of the F statistic from this test.

   - If it is small, we reject the H_0
   - The H_0 is homoskedasticity
   - So rejecting $\rightarrow$ we have heteroskedasticity and should use HC (robust) errors

This works because we are doing a joint test for whether or not $x_1, \cdots, x_2$ explain (jointly) the magnitude (variance) of $\hat{u}$.

Asymptotic normality and consistency

top

Asymptotic Normality

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: $E[u \vert x_1, x_2, \cdots, x_k] = 0$ for all x.
MLR.5	$Var[u \vert x_1, \cdots, x_k] = \sigma_u^2$ 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*:

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

Asymptotic Normality

$\hat{\beta}$ 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 $n \rightarrow \infty$. It is an asymptotic result

Asymptotic normality

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** $n \rightarrow \infty$.

- 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

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 $Pr(|\hat{\beta} - \beta| > \epsilon) \to 0$ as $n \to \infty$

Consistency

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, $\frac{\hat{\beta}-\beta}{\sqrt{\frac{\hat{\sigma}^2}{SST_x}}} \sim N(0,1)$

Dummy variables

top

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 $\mathbf{1}(condition)$ e.g. $\mathbf{1}(age > 65)$
  - 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 $female == TRUE$.

Dummy Variables

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

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

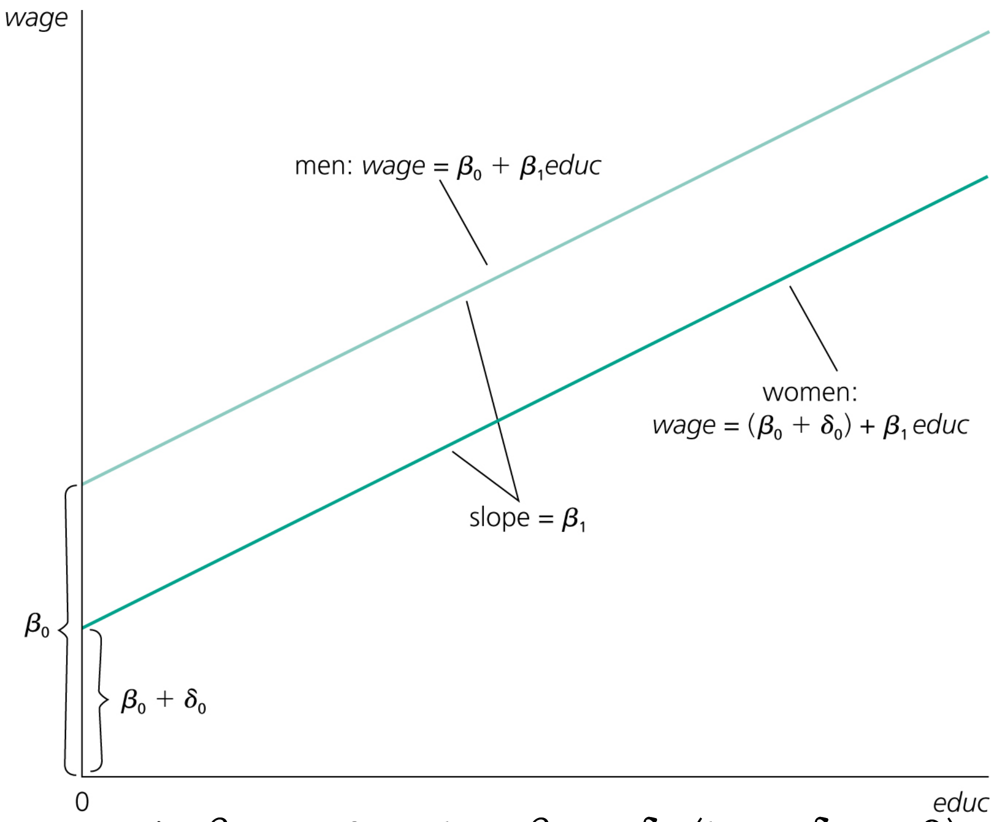
- Sometimes, it will just say that " x_2 is a binary indicator variable that takes on the value of 1 if..."
- In Wooldridge, it just says $y = \beta_0 + \delta_0 female + \cdots + u$
    - You are left to infer that $female$ is either $\{0, 1\}$.
- There are other ways that a dummy variable may be indicated 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.


wage
men: $wage = \beta_0 + \beta_1 educ$
women:
$wage = (\beta_0 + \delta_0) + \beta_1 educ$
slope = β_1
β_0
$\beta_0 + \delta_0$
0
educ

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:

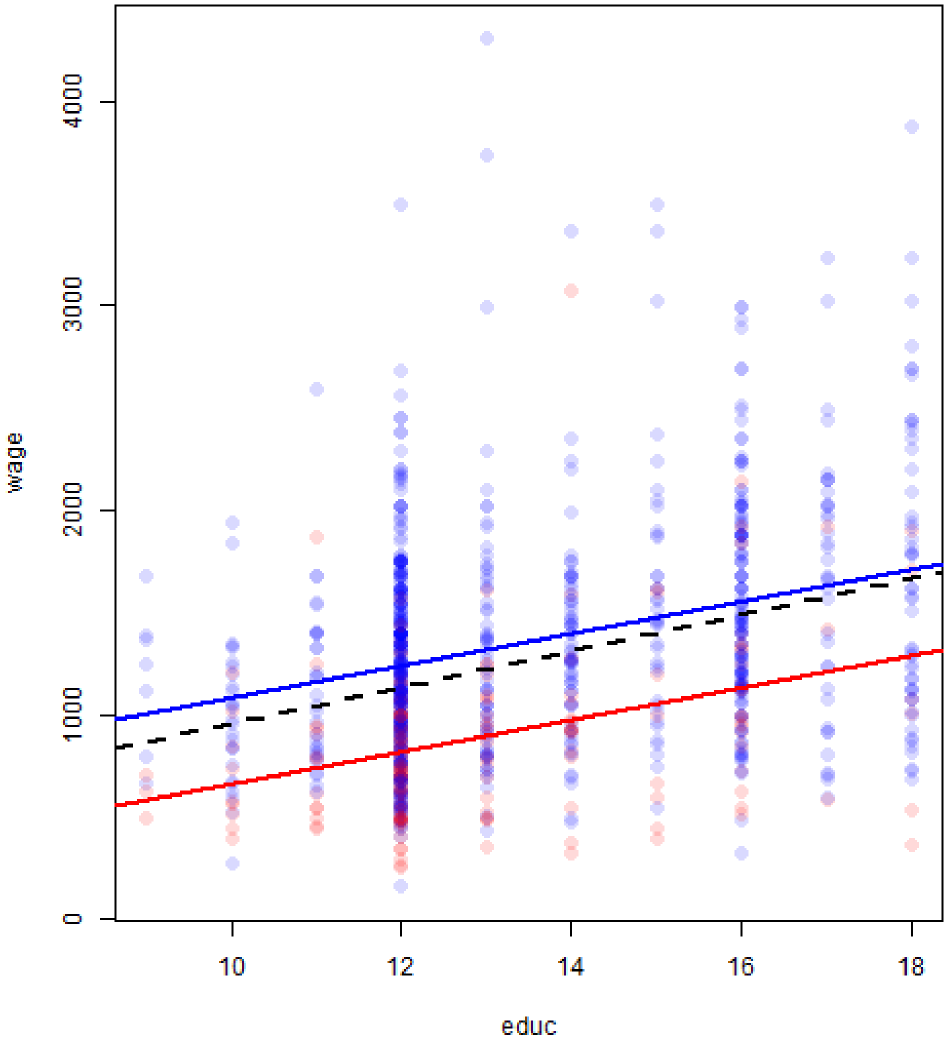
$$wage = \beta_0 + \beta_1 educ + u$$

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!

Dummy Variables

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:

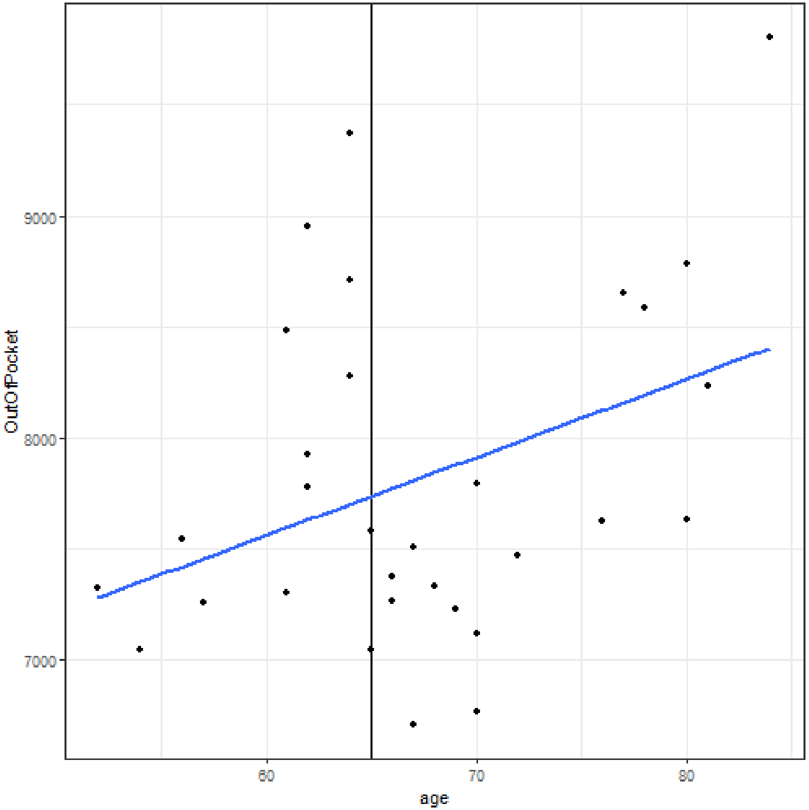
Out of Pocket	Age	1(age>65)
7782	48	0
8136	63	0
9730	86	1
7928	66	1
...	...	...

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

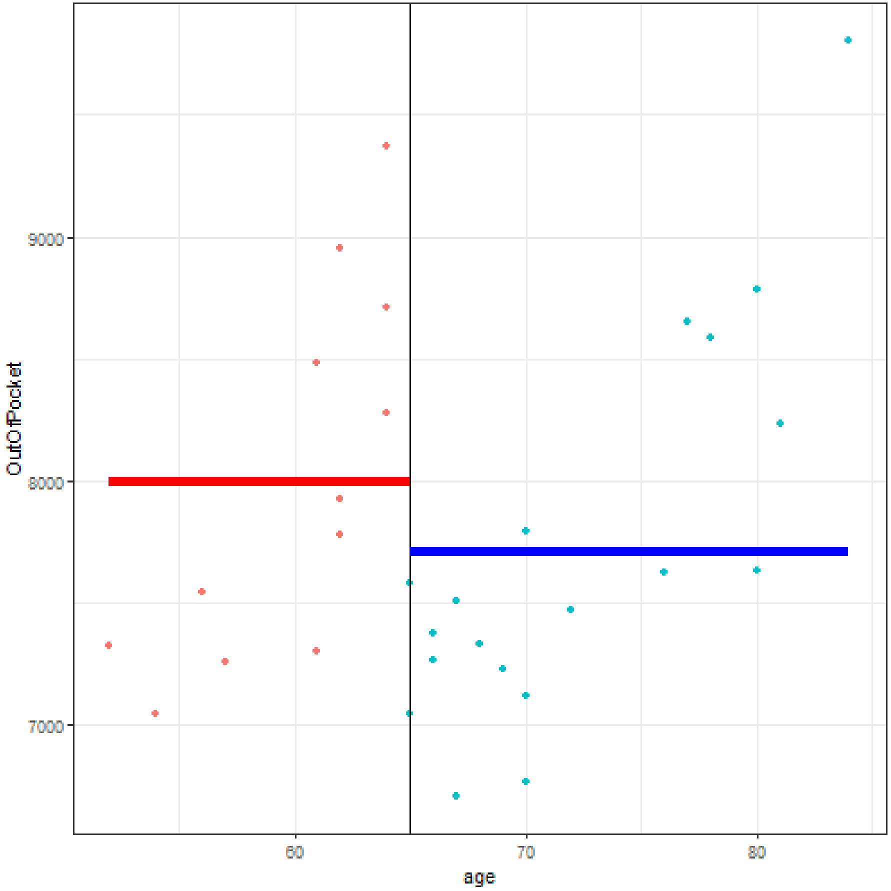
Dummy Variables

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

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

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 5449.643   1037.603  5.2521 1.259e-05 ***
## age           35.182     15.605  2.2545   0.03189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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



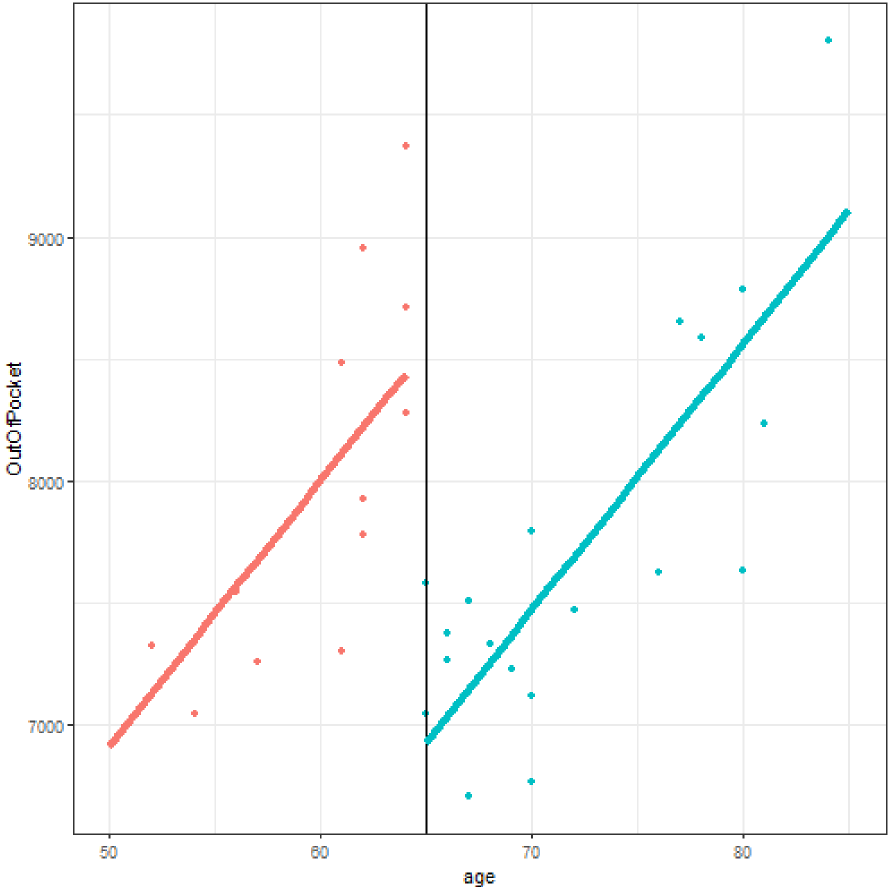
Dummy Variables

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

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

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7997.00     215.87 37.0447   <2e-16 ***
## over65TRUE   -286.17     281.29 -1.0173   0.3174
## ---
## 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

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

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1461.715   1100.388  1.3284    0.1948
## age           109.073     18.689  5.8363 2.845e-06 ***
## over65TRUE  -1621.360    285.891 -5.6713 4.449e-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 $\boldsymbol{cigarettes}$ back in here:

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

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

Dummy Variables

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 = 1.0138, df = 24.144, p-value = 0.3207
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -296.2129  868.5601
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            7997.000            7710.827
```

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

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

Dummy Variables fall under the category of "specification"

- All of the rules about $\boldsymbol{x}$'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 $\hat{\boldsymbol{u}}$
- Dummies don't change *how* we calculate $\hat{\beta}$, $\boldsymbol{se}(\hat{\beta})$, or $\boldsymbol{SSR}$ 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"

Dummy Variables

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
...	...	...	...

Remember MLR3? No perfect colinearity? Uh-oh.

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*.

Let's sum this all up.

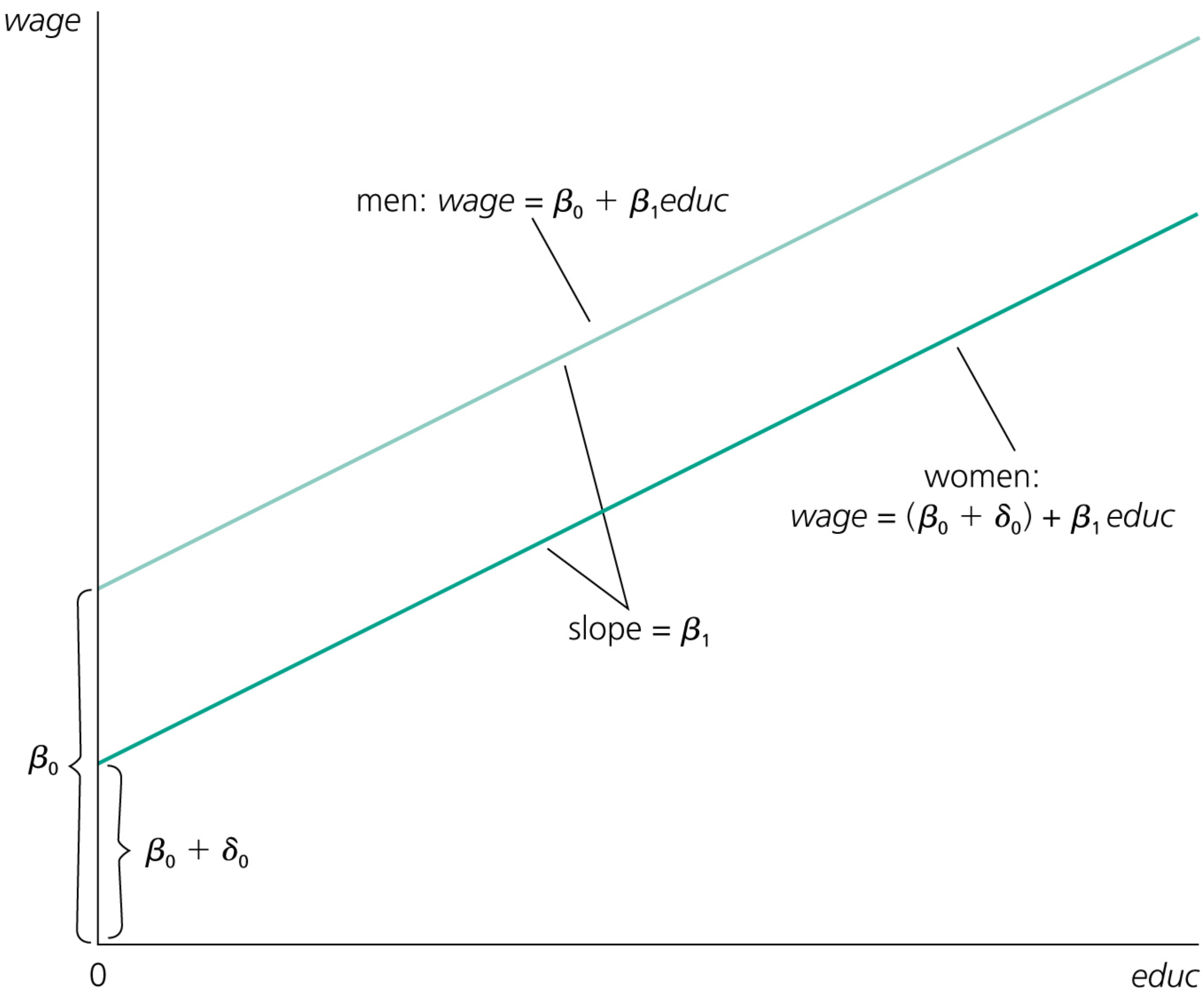
- A dummy is one variable taking the value of $\{0, 1\}$. For example
  - $1(female)$
  - $1(age > 65)$

Interpretation

- A shift in the intercept **relative to the base group**
  - Base group: $male$
  - Base group: $age < 65$
- The "base group" (or "base level") is represented in the intercept: β_0
- The other group(s) ($female$, $age > 65$) have shifted intercepts:
  - β_0 for the base (male, under 65)
  - $\beta_0 + \beta_1$ for a female under 65
  - $\beta_0 + \beta_1 + \beta_2$ for a female over 65
  - $\beta_0 + \beta_2$ for a male over 65

$$y = \beta_0 + \beta_1 1(female) + \beta_2 1(age > 65) + u$$

Dummy Variables


wage
men: $wage = \beta_0 + \beta_1 educ$
women:
$wage = (\beta_0 + \delta_0) + \beta_1 educ$
slope = β_1
β_0
$\beta_0 + \delta_0$
0
educ

Dummy Variables

Ceteris Paribus still applies

Interpretation of the dummy variable coefficient is:

> "The change in the expectation of Y from being in the group relative to being in the base group, *ceteris paribus*"

We are using "in the group" here to mean "the observations for which the dummy is true"

The "base level" is very important

- Since the in-group intercept is $E[Y|in - group] = \beta_0 + \beta_1$, but the coefficient is β_1, we have to be careful.
- The coefficient is the *difference* between the base level and the in-group.
- The "base" group is sometimes called **the omitted level**

Interpretation of $\{0, 1\}$ dummy variables

In the Wooldridge example

$$wage = \beta_0 + \beta_1 1(female) + \beta_2 educ + u$$

> "Conditional on education, females make on average β_1 more/less than males, ceteris paribus"

More/less depending on whether or not the coefficient is negative

In the age example:

$$Out - of - pocket = \beta_0 + \beta_1 age + \beta_2 1(age > 65) + u$$

> "Individuals over 65 years of age pay β_2 more/less in out-of-pocket expenses relative to those under 65, controlling for the linear effect of age, ceteris paribus"

Here, we have to be a little more specific since the dummy variable and the continuous variable, age, both refer to age. It would be strange to say "conditional on age, being 65 means paying β_1 more/less".

Dummy Variables

We can have more than one dummy variable:

$$wage = \beta_0 + \beta_1 1(female) + \beta_2 1(age > 65) + u$$

- $E[wage|male, under65]$ = β_0
- $E[wage|female, under65]$ = ??
- $E[wage|female, over65]$ = ??

Panel Data

top

Panel Data is what we call 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 $N = 100 \times 12 = 1200$

Or

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

Contrast Panel Data with other types of data:

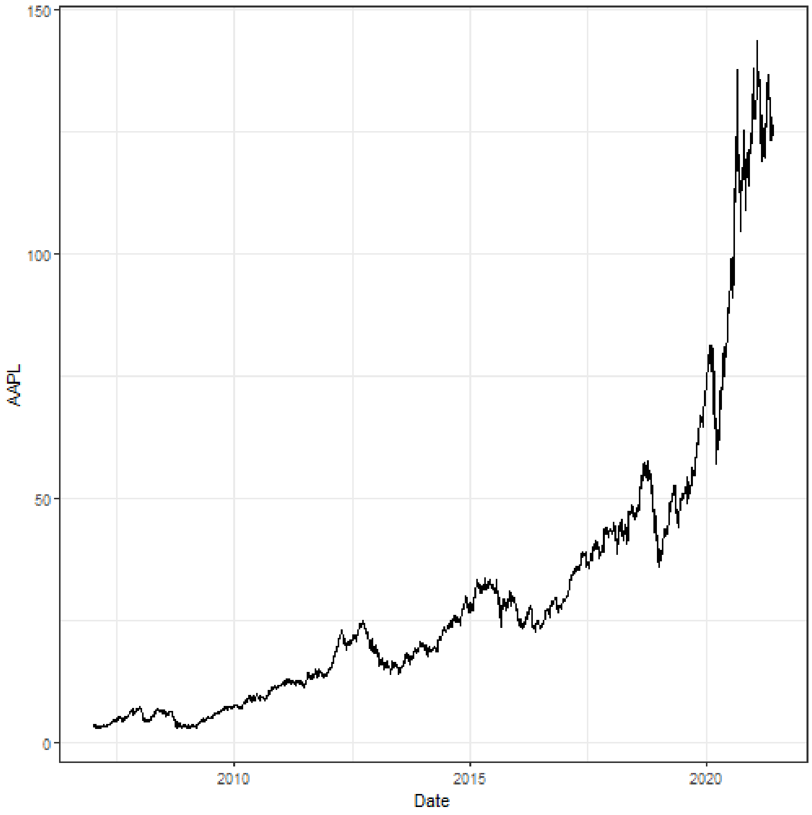
Time series data

- 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

AAPL



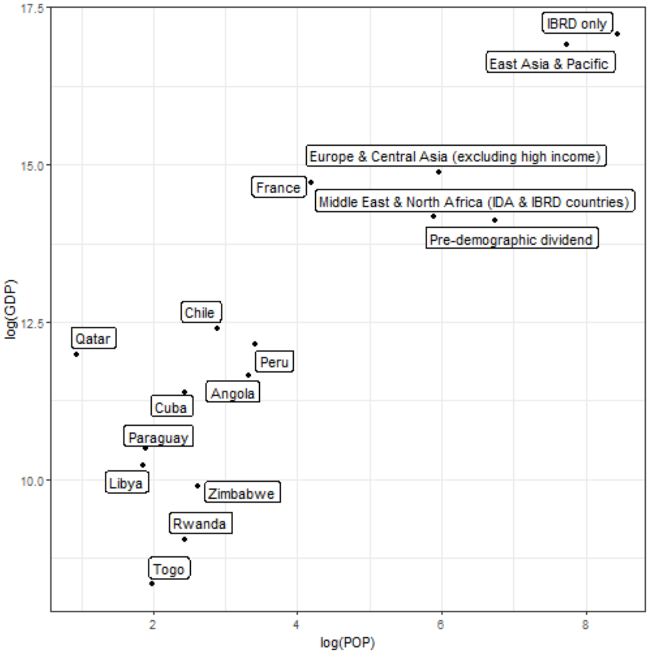
Time series, not panel data.

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



Panel Data

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

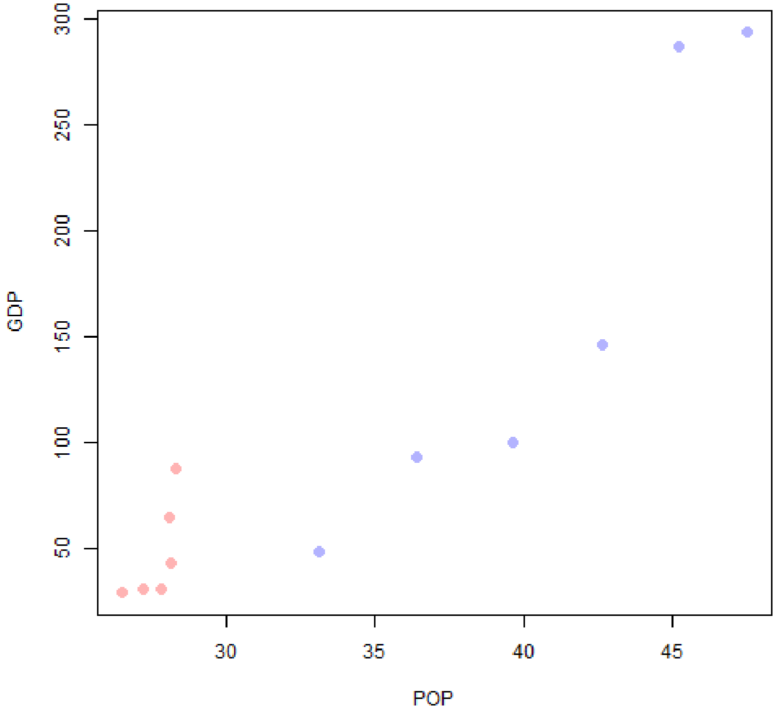
- We will get to time series later on
- Let's focus on Panel Data today

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

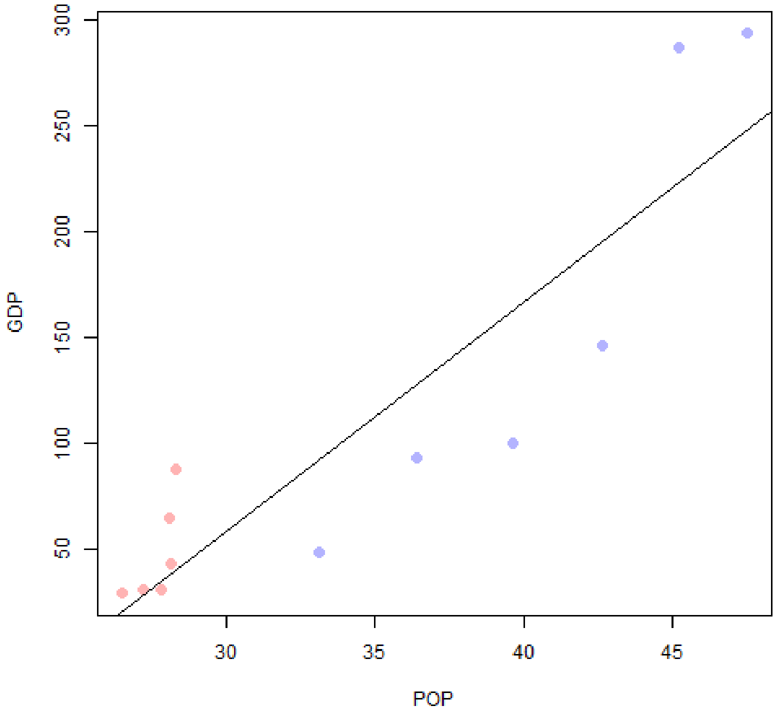
Panel Data

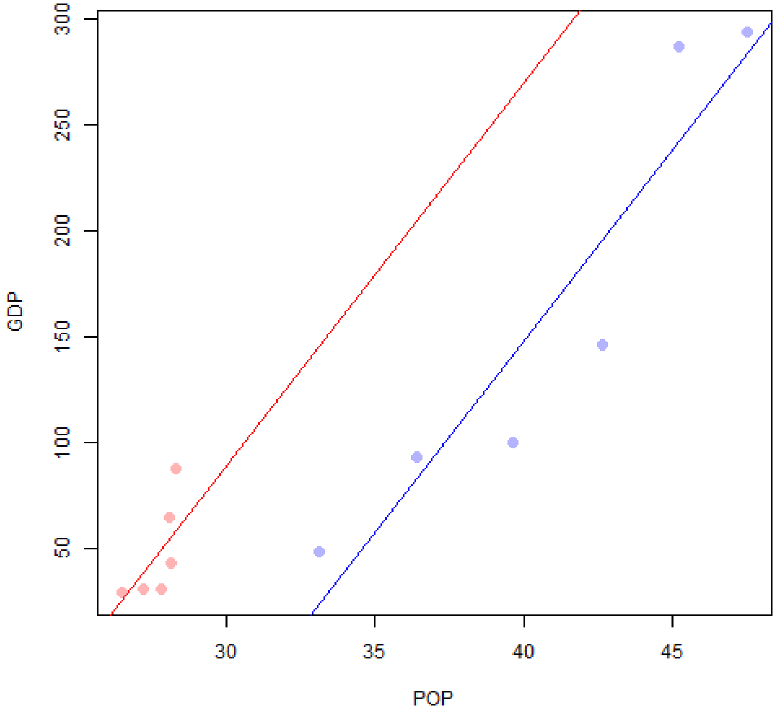
Country	GDP	POP	Year	GDPPC
Colombia	47.8	33.1	1990	1445.3
Colombia	92.5	36.4	1995	2539.9
Colombia	99.9	39.6	2000	2520.5
Colombia	145.6	42.6	2005	3414.5
Colombia	286.6	45.2	2010	6336.7
Colombia	293.5	47.5	2015	6175.9
Cuba	28.6	26.5	1990	2703.2
Cuba	30.4	27.2	1995	2794.7
Cuba	30.6	27.8	2000	2747.1
Cuba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





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)

Panel Data

```
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".

Panel Data

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...

Fixed effects with multiple groups

top

Fixed Effects

What if we have three groups?

Take *education* as an example - we can "bin" education into:

High School or less		2- or 4-year college degree		Graduate degree
"HS"		"College"		"Graduate"

When this is represented with one variable, it's called a **categorical** variable

Fixed Effects

Our three groups would work as follows:

wage	experience	educ	education
9000	0	12	HS
20000	5	16	College
60000	12	14	College
27000	2	18	Graduate
32000	10	9	HS

In the US, primary (required) education is 12 years, undergraduate is 4 additional years, and graduate school is 2-5+ additional years.

Base level with categorical variable

- There is still a "base level" (or "omitted level")
- It is *your* choice as to which one is the "base level"
  - Coefficient estimates will still add up the same.
  - Interpretability is easier if you choose wisely
  - We should choose "HS" as the "base level" here, so that estimates are relative to HS
  - This is incorporating *ordinal* information since we think $wage_{grad} > wage_{college} > wage_{HS}$

Numeric representation

- To represent a categorical variable with 3 categories, we need to create **two** more columns
  - If there are K categories, then we need $K - 1$ new columns
  - Whichever one we don't create a column for is the "base"
  - It's effect will be found in the β_0 (the intercept)

Using "HS" as the base level:

wage	experience	education	education==College	education==Graduate
9000	0	HS	0	0
20000	5	College	1	0
60000	12	College	1	0
27000	2	Graduate	0	1
32000	10	HS	0	0

If we run this in R (leaving out the "education" column), we would get a coefficient for $education == College$ and $education == Graduate$

- These would be the increase in the expected wage resulting from moving between the HS group to the College (or Graduate, respectively) group, *ceteris paribus*.

Fixed Effects

In R, categorical variables are a special type of variable called "**factor**"

```
df$education = as.factor(df$education)
```

- R stores the labels separately, but will let you refer to them
- If we use `str(df)`, we can see the factor structure
- I'm going to switch to a dataset that has a categorical in it

```
census = wooldridge::census2000
str(census)
```

```
## 'data.frame':    29501 obs. of  6 variables:
##  $ state   : Factor w/ 51 levels "Alabama","Alaska",..: 41 39 11 29 3 5 38 27 14 19 ...
##  $ puma    : int  100 2502 1800 100 206 1601 1309 100 3301 1600 ...
##  $ educ    : int  13 13 12 13 16 12 13 13 16 16 ...
##  $ lweekinc: num  6.47 6.09 7.03 6.69 7.34 ...
##  $ exper   : int  37 14 21 12 18 15 29 14 22 26 ...
##  $ expersq : int  1369 196 441 144 324 225 841 196 484 676 ...
```

To go from a factor to a character string

```
census$state = as.character(census$state)
head(census)
```

```
##             state puma educ lweekinc exper expersq
## 1 South Carolina  100   13 6.471038    37    1369
## 2   Pennsylvania 2502   13 6.087648    14     196
## 3        Georgia 1800   12 7.034049    21     441
## 4         Nevada  100   13 6.694181    12     144
## 5        Arizona  206   16 7.338538    18     324
## 6     California 1601   12 6.422247    15     225
```

```
str(census)
```

```
## 'data.frame':    29501 obs. of  6 variables:
##  $ state   : chr  "South Carolina" "Pennsylvania" "Georgia" "Nevada" ...
##  $ puma    : int  100 2502 1800 100 206 1601 1309 100 3301 1600 ...
##  $ educ    : int  13 13 12 13 16 12 13 13 16 16 ...
##  $ lweekinc: num  6.47 6.09 7.03 6.69 7.34 ...
##  $ exper   : int  37 14 21 12 18 15 29 14 22 26 ...
##  $ expersq : int  1369 196 441 144 324 225 841 196 484 676 ...
```

More important, how to go from character string to factor

```
census$state = as.factor(census$state)
head(census)
```

```
##            state puma educ lweekinc exper expersq
## 1 South Carolina  100   13 6.471038    37    1369
## 2   Pennsylvania 2502   13 6.087648    14     196
## 3        Georgia 1800   12 7.034049    21     441
## 4         Nevada  100   13 6.694181    12     144
## 5        Arizona  206   16 7.338538    18     324
## 6     California 1601   12 6.422247    15     225
```

```
str(census)
```

```
## 'data.frame':    29501 obs. of  6 variables:
##  $ state   : Factor w/ 51 levels "Alabama","Alaska",..: 41 39 11 29 3 5 38 27 14 19 ...
##  $ puma    : int  100 2502 1800 100 206 1601 1309 100 3301 1600 ...
##  $ educ    : int  13 13 12 13 16 12 13 13 16 16 ...
##  $ lweekinc: num  6.47 6.09 7.03 6.69 7.34 ...
##  $ exper   : int  37 14 21 12 18 15 29 14 22 26 ...
##  $ expersq : int  1369 196 441 144 324 225 841 196 484 676 ...
```

Fixed Effects

If you use a factor variable in a regression, R will construct the additional columns

```
census.small = census[census$state=='South Carolina'|census$state=='Arizona'|
                      census$state=='Nevada',c('lweekinc','state','educ', 'exper')]
census.small$statefactor = as.factor(census.small$state)
head(census.small, 10)
```

```
##      lweekinc          state educ exper    statefactor
## 1   6.471038 South Carolina   13    37 South Carolina
## 4   6.694181         Nevada   13    12         Nevada
## 5   7.338538        Arizona   16    18        Arizona
## 17  7.129298        Arizona   12    44        Arizona
## 31  6.738426 South Carolina   12     8 South Carolina
## 40  6.827713 South Carolina   12    42 South Carolina
## 58  7.495542         Nevada   13    15         Nevada
## 81  6.357709        Arizona   12    25        Arizona
## 119 6.645391         Nevada   12    32         Nevada
## 152 5.860730        Arizona   13    27        Arizona
```

How does R convert factors to data columns?

```
head(model.matrix(lweekinc ~ educ + exper + statefactor, data = census.small)) # we won
```

```
##    (Intercept) educ exper statefactorNevada statefactorSouth Carolina
## 1            1   13    37                 0                         1
## 4            1   13    12                 1                         0
## 5            1   16    18                 0                         0
## 17           1   12    44                 0                         0
## 31           1   12     8                 0                         1
## 40           1   12    42                 0                         1
```

Question: What is the base level?

Fixed Effects

```
summary(lm(lweekinc ~ educ + exper + statefactor, data = census.small))
```

```
## 
## Call:
## lm(formula = lweekinc ~ educ + exper + statefactor, data = census.small)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.6167 -0.3284  0.0254  0.3749  3.2501 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                5.529636   0.177034  31.235  < 2e-16 ***
## educ                       0.070970   0.011758   6.036 2.14e-09 ***
## exper                      0.004664   0.001954   2.387   0.0172 *  
## statefactorNevada          0.043311   0.054585   0.793   0.4277    
## statefactorSouth Carolina -0.059640   0.044963  -1.326   0.1850    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6706 on 1119 degrees of freedom
## Multiple R-squared:  0.03657,    Adjusted R-squared:  0.03312 
## F-statistic: 10.62 on 4 and 1119 DF,  p-value: 1.903e-08
```

A "within-group" interpretation

- Group fixed effects explain the *mean* of the y variable within that group
  - E.g. our Cuba/Colombia example on Monday
  - The intercept is just the difference in means (conditional on the other x's)
- The group fixed effect accounts for the averge difference *between* groups
  - And leaves the rest of the x's to explain the variation in y *within* the group
- If we think of "partialling out" the fixed effect, this makes even more sense.

Fixed Effects

Let's go to our wage/education/experience example. We might think there is a "gender experience gap" where men tend to be more experienced (e.g. due to not giving birth):

$$wage = \beta_0 + \beta_1 1(female) + \beta_2 experience + u$$

Partial out the fixed effect:

$$experience = \delta_0 + \delta_1 1(female) + v$$

$\hat{v}$ is $experience$ that isn't associated with being female. It has had the "gender experience gap" *removed*.

That is, the variation in $\hat{v}$ does not reflect the "male experience gap", so we are *identifying* β_2 off of variation within the group, eliminating variation between $male$ and $female$.

So, a regression using $\hat{v}$ in place of $experience$:

$$wage = \beta_0 + \beta_2 \hat{v}$$

Gives us the correct β_2 (remember our "partialling out" of x_1, x_2) using the "within group" variation in $experience$.

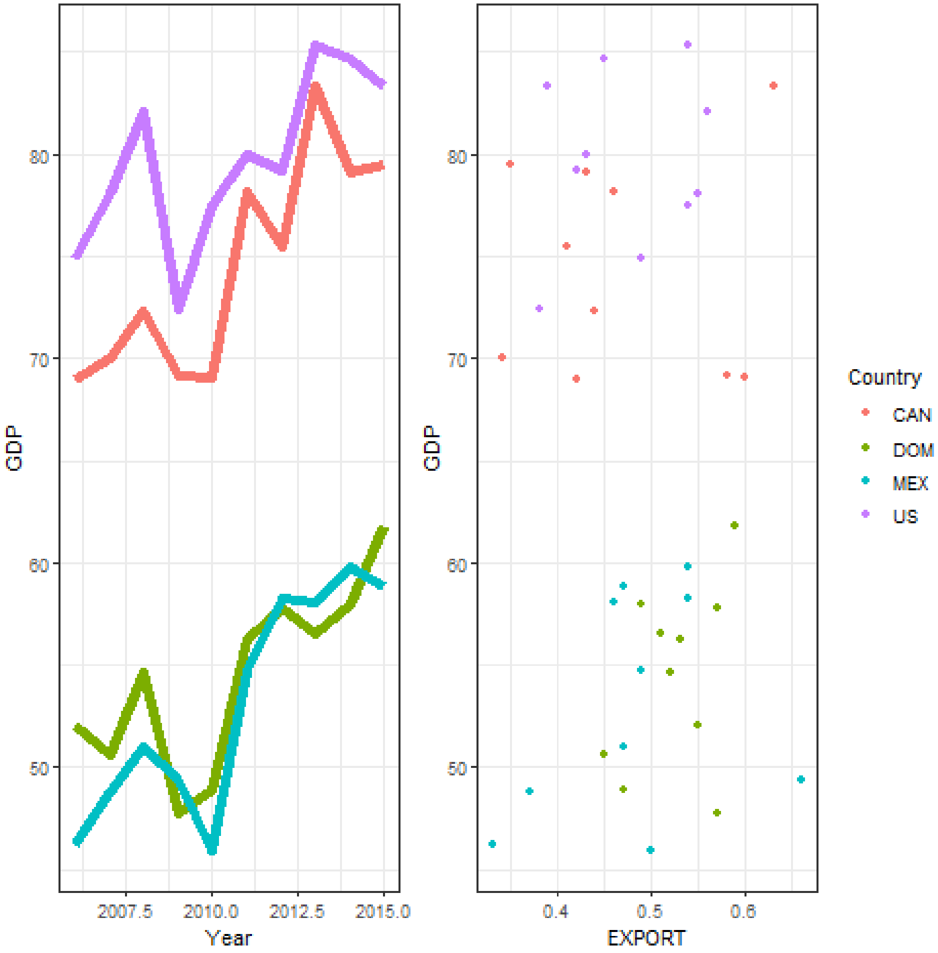
Time fixed effects

What if we have N observations and T time periods (a common type of panel data), but instead of worrying about group-level differences giving us biased estimates, we worried that some time trend or time-specific shock is making one time period different from the others?

Here, let's look at (entirely fake) data on North American GDP and EXPORT (share of GDP from exports).

We want to know if higher EXPORTS are associated with higher GDP.

Fixed Effects



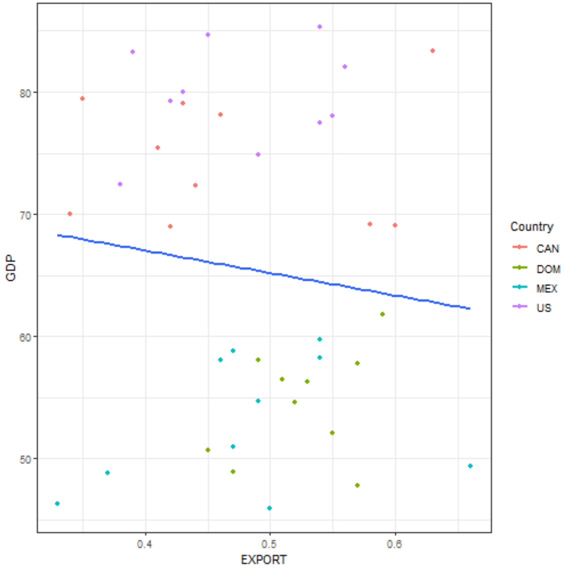
Since this is constructed (fake) data, I know the right coefficient on $EXPORT$, $\beta_{export} = 20$

Fixed Effects

MICHIGAN STATE UNIVERSITY

```
coeftest(lm(GDP ~ EXPORT, df), vcov = vcovHC, type = 'HC1')
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   74.415     13.843  5.3757 4.085e-06 ***
## EXPORT       -18.417     27.902 -0.6601    0.5132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Note I used a simplier call to `coeftest`. Before we had `vcov = vcovHC(OLSobject, 'HC1'))`, but that required two steps: one to create the OLS object, and one to call `coeftest`. This does it all at once.

Fixed Effects

```
coeftest(lm(GDP ~ EXPORT + as.factor(Year), df), vcov = vcovHC, type='HC1')
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          69.69517   16.29177  4.2779 0.0001872 ***
## EXPORT              -20.42546   30.00917 -0.6806 0.5014976
## as.factor(Year)2007   0.91621   10.83901  0.0845 0.9332170
## as.factor(Year)2008   5.43941   10.38025  0.5240 0.6042500
## as.factor(Year)2009   1.16664    9.38704  0.1243 0.9019494
## as.factor(Year)2010   1.41325   11.04306  0.1280 0.8990514
## as.factor(Year)2011   7.32123    9.72125  0.7531 0.4574514
## as.factor(Year)2012   7.90148    8.67860  0.9105 0.3700877
## as.factor(Year)2013  12.03960   11.23519  1.0716 0.2927386
## as.factor(Year)2014  10.44856    9.64676  1.0831 0.2876807
## as.factor(Year)2015  10.34100    9.05938  1.1415 0.2630128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# note - you can use "as.factor" in the ~ formula
```

Fixed Effects

```
coeftest(lm(GDP ~ EXPORT + as.factor(Year) + as.factor(Country), df), vcov = vcovHC, typ
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error  t value  Pr(>|t|)    
## (Intercept)            61.51985    1.32211  46.5317 < 2.2e-16 ***
## EXPORT                 18.99747    3.33745   5.6922 5.466e-06 ***
## as.factor(Year)2007     1.70466    0.59633   2.8586 0.0082725 ** 
## as.factor(Year)2008     3.46827    0.92018   3.7691 0.0008512 ***
## as.factor(Year)2009    -2.77565    0.93890  -2.9563 0.0065428 ** 
## as.factor(Year)2010    -1.74059    1.30822  -1.3305 0.1949056    
## as.factor(Year)2011     6.13854    0.85611   7.1702 1.293e-07 ***
## as.factor(Year)2012     6.42312    0.97667   6.5766 5.659e-07 ***
## as.factor(Year)2013     8.59009    1.07620   7.9819 1.846e-08 ***
## as.factor(Year)2014     9.26587    0.65546  14.1365 1.023e-13 ***
## as.factor(Year)2015    10.24245    0.70164  14.5978 4.860e-14 ***
## as.factor(Country)DOM -21.18461    0.68779 -30.8008 < 2.2e-16 ***
## as.factor(Country)MEX -21.73305    0.58530 -37.1312 < 2.2e-16 ***
## as.factor(Country)US    5.05193    0.66545   7.5917 4.656e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, you can specify more than one set of categorical variables

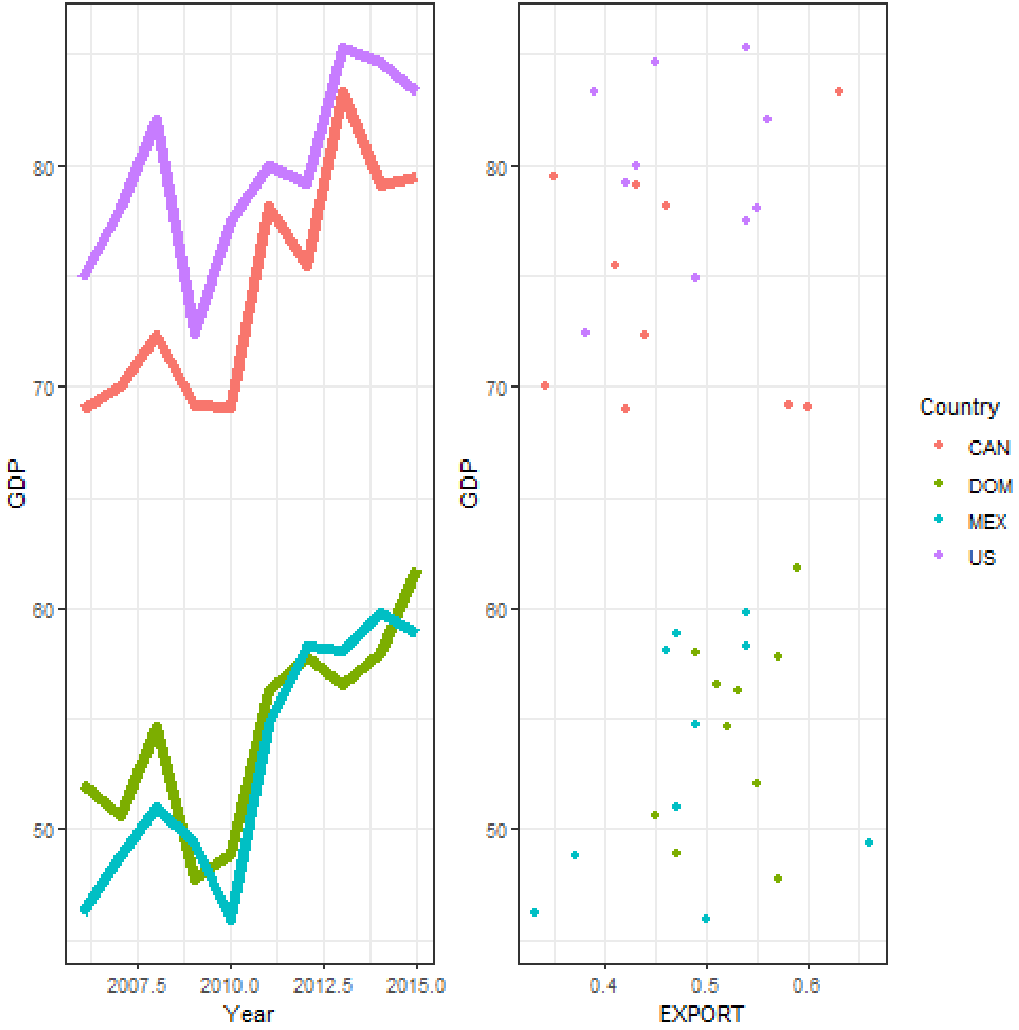
- Just as you can have more than one dummy variable
- The interpretation of each one is still the same: the effect of being in the group/time period relative to the base group/time period, *ceteris paribus*.
- These are called **two-way fixed effects** (TWFE)
  - When used on panel data
  - And when there is one fixed effect for each of the panel data's dimensions
  - N countries and T years here.

Fixed effects and Partialling Out

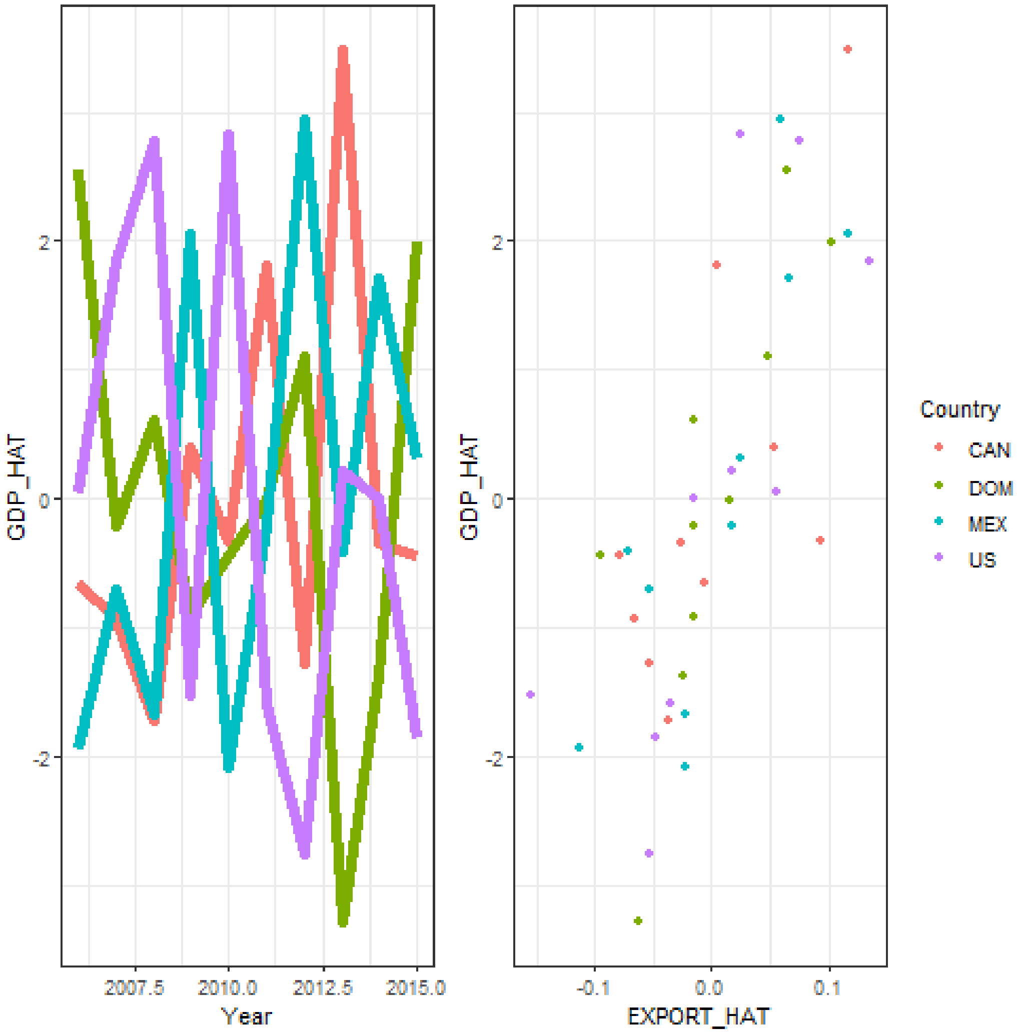
On the next slide, GDP_HAT is the residual from regressing GDP on categorical Year and Country

- Same for EXPORT_HAT

Our original data before partialling out:



After paritalling out left: YEAR and right: YEAR and COUNTRY



The code I used to parital out and plot the prior slide:

```
df = df %>%
  dplyr::mutate(EXPORT_HAT = resid(lm(EXPORT ~ as.factor(Year) + as.factor(Country), data=df)
                GDP_HAT = resid(lm(GDP ~ as.factor(Year) + as.factor(Country), data=df)))


c = ggplot(df, aes(x = Year, y=GDP_HAT, col=Country)) + geom_line(lwd=2) +
theme_bw() + theme(legend.position='none')

d = ggplot(df, aes(x = EXPORT_HAT, y = GDP_HAT, col=Country)) + geom_point() + theme_bw()

library(patchwork)
c+d
```

Interactions with Dummies

top

Dummy variables shift *the intercepts*

- Very useful when a group (or time) has a different mean
- Covers "*unobserved, time-invariant differences*"

But what if we think that the *slopes* differ

- For instance, maybe each country in our GDP/EXPORT example has *it's own unique relationship* between GDP and $EXPORT$?
- This can be *in addition* to thinking that each country has its own unique intercept
  - In fact, it would be odd to think that they'd have their own unique slope but *not* a unique intercept.

How do we let the slopes vary?

- In a way very similar to letting the intercepts vary
- Let's look at it in an example with only two categories (a single dummy)

$$y = \beta_0 + \beta_1 1(condition) + \beta_2 x_1 + \underbrace{\beta_3 \times x_1 \times 1(condition)}_{\text{The interaction term}} + u$$

A couple things to note:

- x_1 is our variable of interest here
- $condition$ is our group dummy (like $male$ or $age > 65$)
- x_1 appears twice, once with β_2, and *again* in the interaction of $x_1 \times 1(condition)$

Interactions

$$y = \beta_0 + \beta_1 1(condition) + \beta_2 x_1 + \underbrace{\beta_3 x_1 1(condition)}_{\text{The interaction term}} + u$$

Refreshing our interpretation of the intercept:

- The intercept for the base group is β_0
- The intercept for the in-group defined by $condition$ is $\beta_0 + \beta_1$

Applying the same thought process to the interaction:

- **For the base group**, the marginal change in y from a unit increase in x_1 is β_2
- **For the in-group**, the marginal change in y from a unit increase in x_1 is $\beta_2 + \beta_3$

$$\text{For the base group: } \frac{\Delta y}{\Delta x_1} = \beta_2$$

$$\text{For the in-group: } \frac{\Delta y}{\Delta x_1} = \beta_2 + \beta_3$$

Interactions

Of course, we can have >2 groups (categorical)

$$\begin{aligned} y = \beta_0 + {} & \beta_1 1(group == 2) + \beta_3 1(group == 3) + \beta_4 x_1 \\ & + \beta_5 x_1 1(group == 2) + \beta_6 x_1 1(group == 3) + u \end{aligned}$$

What does that look like?

wage	experience	educ	educ = College	educ = Graduate	experience x educ == College	experience x educ == Graduate
9000	0	HS	0	0	0	0
20000	5	College	1	0	5	0
60000	12	College	1	0	12	0
27000	2	Graduate	0	1	0	2
32000	10	HS	0	0	0	0

Interactions

And in R:

```
lm(wage ~ as.factor(educ) + exper + as.factor(educ)*exper,
data=df)
```

Here, you'll get **intercept** **shift** coefficients on:

- educ = College
- educ = Grad

And you'll get **slope** **shift** coefficients on:

- experience for educ = College
- experience for educ = Graduate

The wage/education/experience regression would be:

$$\begin{aligned} wage = \beta_0 + &\beta_1 1(educ == College) + \beta_2 1(educ == Grad) + \beta_3 exper \\ &+ \beta_4 x_1 1(educ == Coll) + \beta_5 x_1 1(educ == Grad) + u \end{aligned}$$

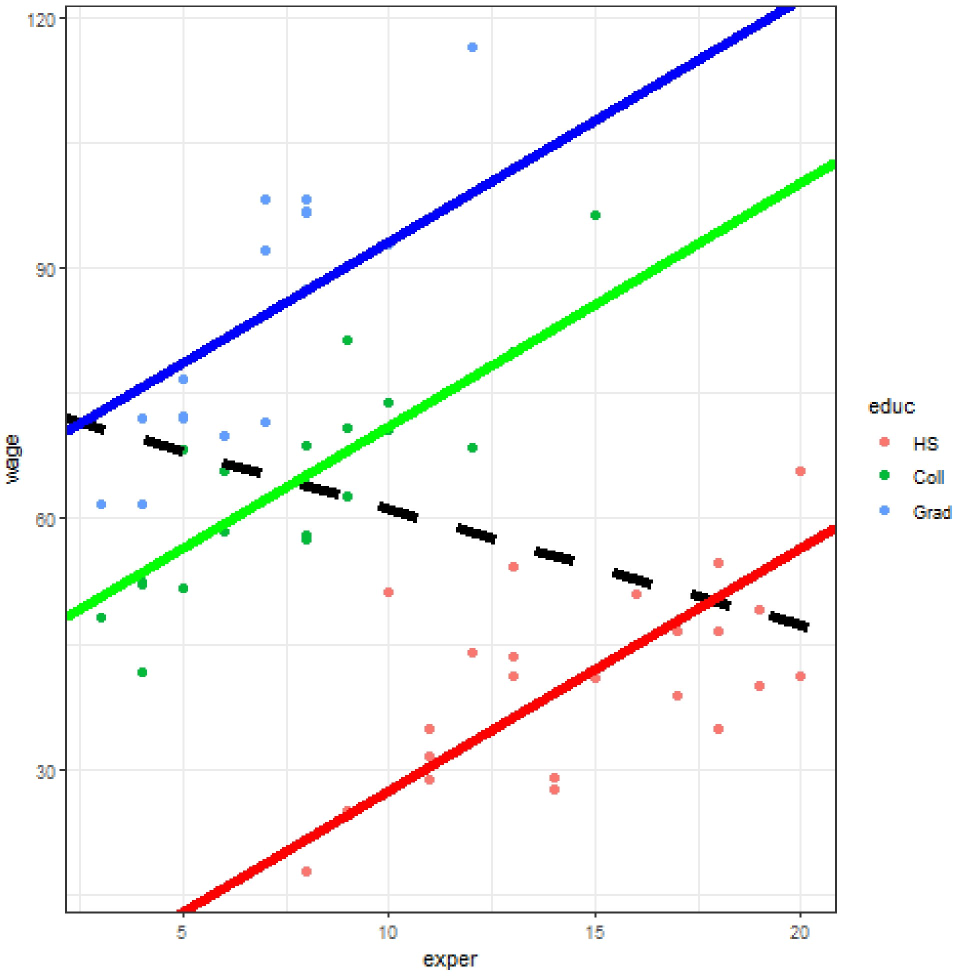
Expected Values conditional on X:

- $E[wage|exper, educ = HS] = \beta_0 + \beta_3 \times exper$
- $E[wage|exper, educ = Coll] = (\beta_0 + \beta_1) + (\beta_3 + \beta_4) \times exper$
- $E[wage|exper, educ = Grad] = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) \times exper$

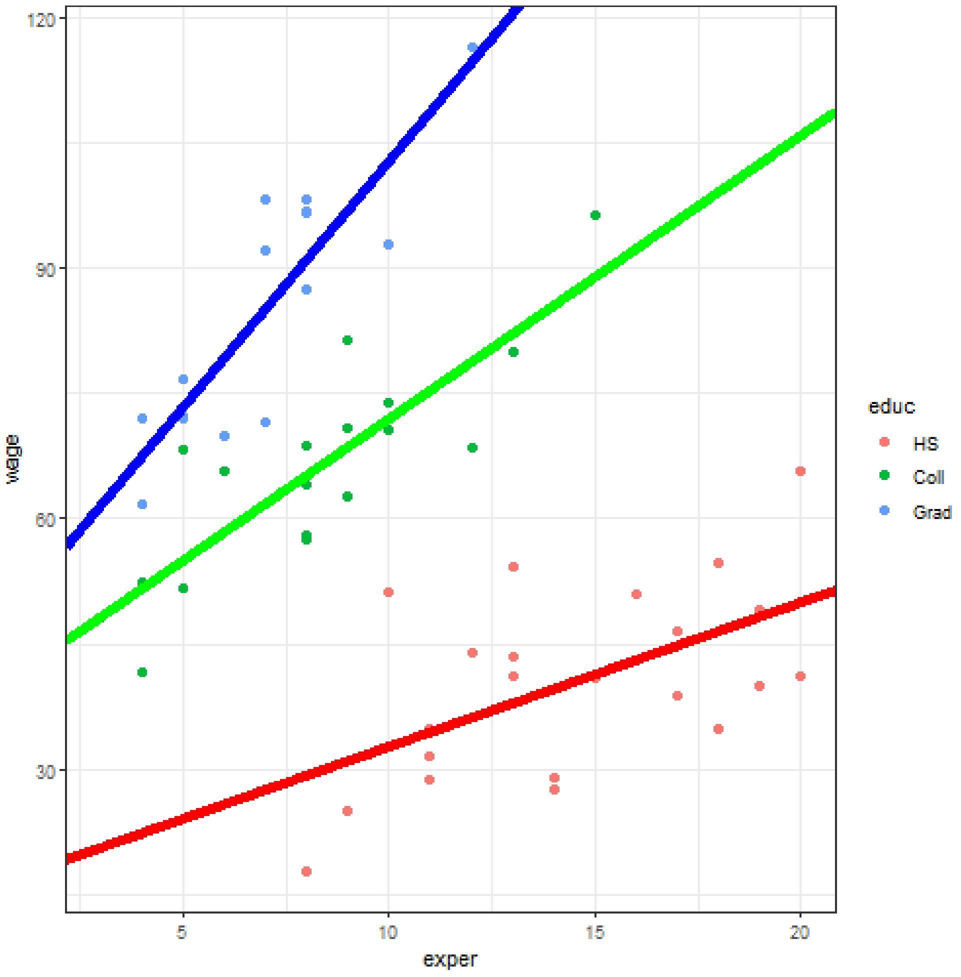
Just as we do with the intercepts, we add to the base level

- Note that when we have three categories $\{HS, Coll, Grad\}$ and we want the $E[wage|exper, educ == Grad]$, we do **not** add in the intercept-shift or slope-shift for $educ == Coll$.

The naive pooled (black) and the intercept-shift only:



And letting *intercept* and *slope* vary:



Interactions

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    15.52086    9.33517  1.6626  0.102183
## educColl       22.58345   10.14406  2.2263  0.030189 *
## educGrad       28.32628    9.97941  2.8385  0.006375 **
## exper           1.72848    0.61230  2.8229  0.006649 **
## educColl:exper  1.66493    0.77769  2.1409  0.036816 *
## educGrad:exper  4.17432    0.81558  5.1182 4.211e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interactions

The true slopes (since this is fictional data) are:

educ	Slope
HS	1
Coll	4
Grad	5

How would we say this?

> β_4 is the college-specific increase in the relationship between per-year-of-experience and wages relative to HS graduates

- We can also just think of it in terms of slope: a positive β_4 means the slope is steeper (more up) than HS

Significance

- The statistical test that is output in these regressions refers to whether or not that coefficient is zero
- For a intercept-shift (β_1 or β_2), the test tells us whether or not the *intercept* (or *mean*) outcome of the in-group is different from the base level.
- For a slope-shift (interaction, e.g. β_3 or β_4), the test tells us whether or not the *slope* is different of the in-group is different from the base level.
    - That is, it asks: "does this group have a *different relationship between exper and wage* than the base group?"

Interractions

Two-dummy interactions:

$$Out - of - pocket = \beta_0 + \beta_1 1(single) + \beta_2 1(age > 65) \\ + \beta_3 1(single) 1(age > 65) + u$$

We have the same interpretation for β_0 thru β_2

- **But** β_3 tells us the $E[Out - of - pocket | \text{both things true}]$

This means a single person over 65 adds *four* beta's together:

- $E[O - o - p | \text{married, 64 years old}] = \beta_0$
- $E[O - o - p | \text{single, 64 years old}] = \beta_0 + \beta_1$
- $E[O - o - p | \text{married, 66 years old}] = \beta_0 + \beta_2$
- $E[O - o - p | \text{single, 66 years old}] = \beta_0 + \beta_1 + \beta_2 + \beta_3$

This is because a single person over 65 is all four things at once. β_3 is interpreted as the additional effect of being *both* >65 and single.