Reviewing Regression (Pt. 2)

ECON 490 (*Spring 2024*)

Taylor Mackay || Email: tmackay@fullerton.edu

Slides Overview

In these slides, we'll discuss:

- Including multiple variables in a regression
- Logs and percentage changes
- Including factor variables in regressions

Setting the Stage

Last class, we discussed regression as a way of drawing a fitted line

- Given two variables, we can draw a scatterplot
- Then use OLS to calculate fitted line and summarize relationship

Key point – we've seen other ways of summarizing relationships

- A natural question is, "What's so special about regression?"
- Why not just use the other tools we've seen?

That's what we'll talk about today

Unpacking our Regression Equation

Last class, we unpacked the following regression equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

What does it mean to include multiple variables in a regression?

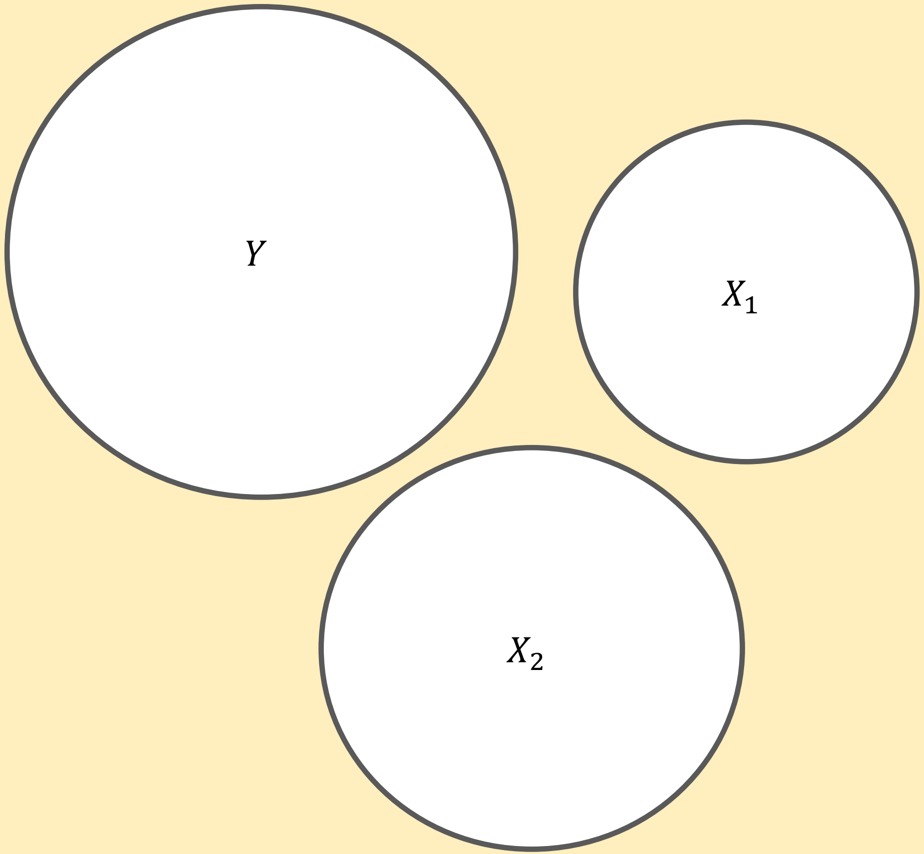
- A phrase you've probably heard before: *"Controlling for X…"*
- Let's unpack what this means using Venn diagram model

Visualizing Variance

The size of each circle reflects the ***variance*** of that variable

If all 3 variables were unrelated, then they ***won't*** overlap

Let's assume they're related

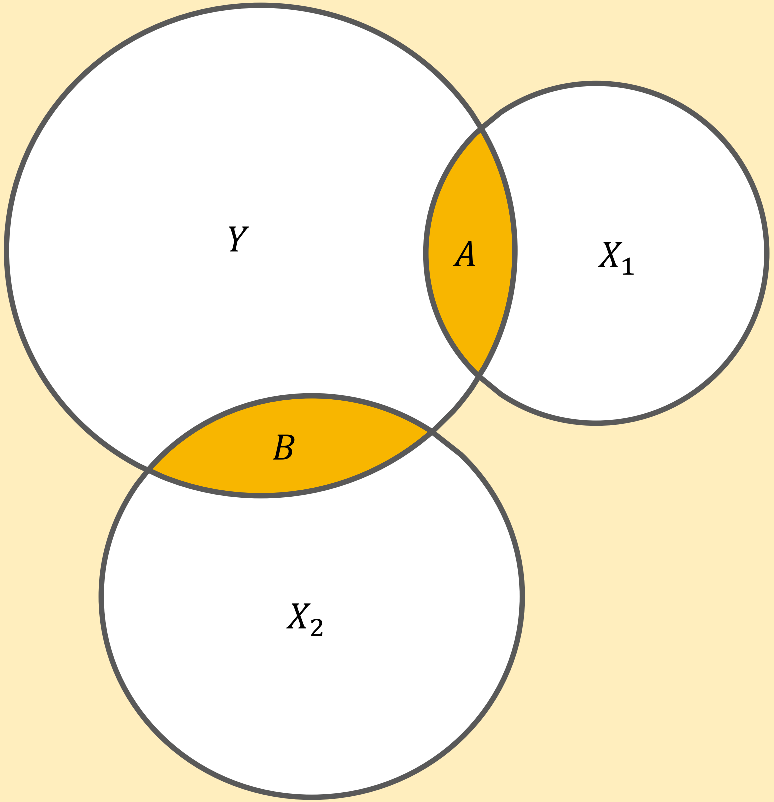


Visualizing Covariance

Let's start by imagining that both of our X's are related to Y but ***not*** to each other

We can use the shaded regions to show the **covariance** of Y and both X's:

- $Cov(Y, X_1) = A$
- $Cov(Y, X_2) = B$

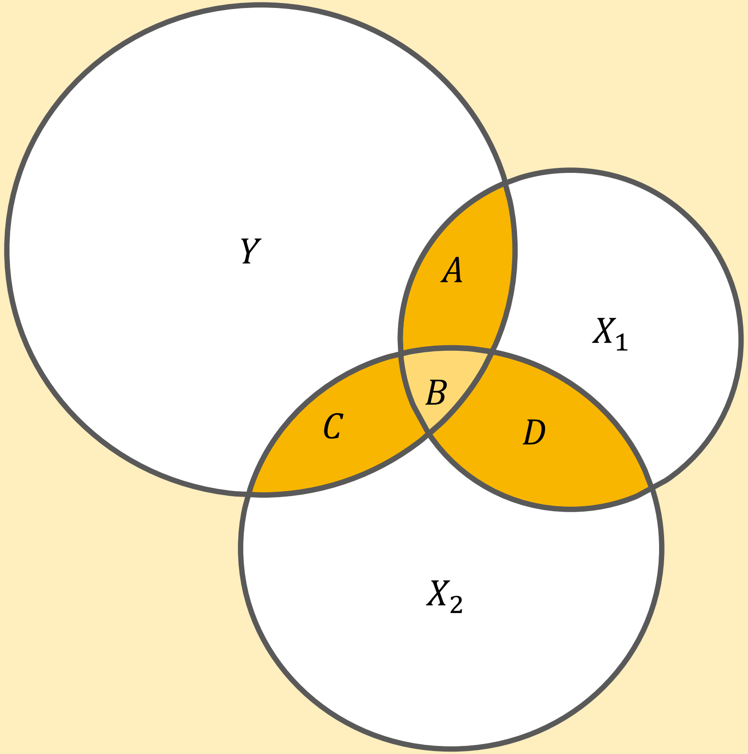


Interdependence

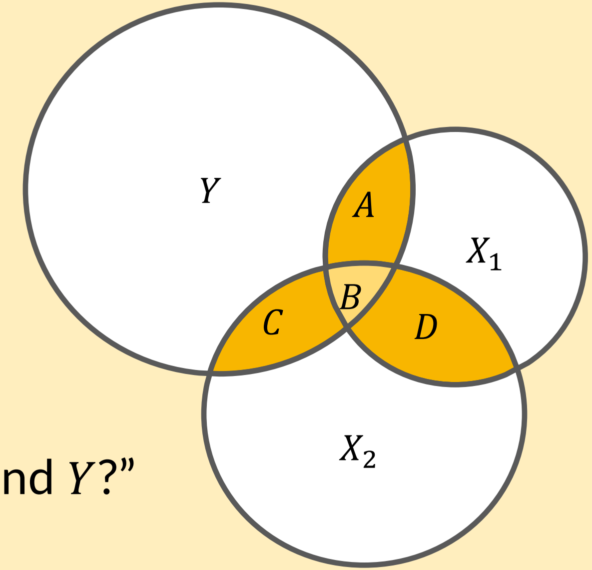
Suppose our X's are also related to ***each other***

Now, our covariances are:

- $Cov(Y, X_1) = A + B$
- $Cov(Y, X_2) = B + C$
- $Cov(X_1, X_2) = B + D$



Thinking About Relationships

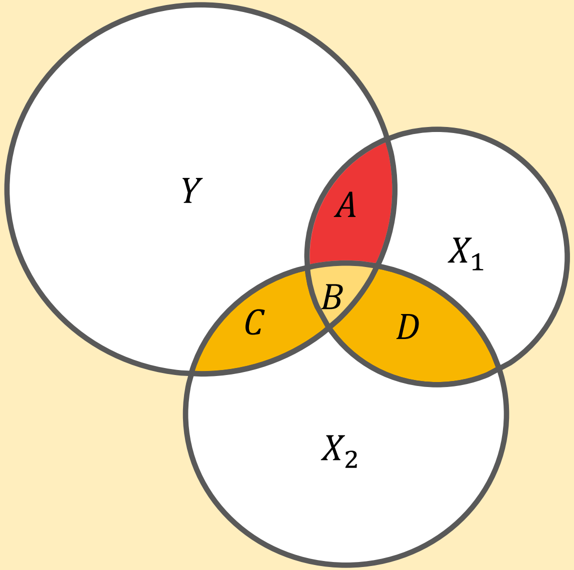


Given our diagram, what's the "effect" of X_1 on Y?

Let's start by thinking about $A + B$ as an answer
- Answers the question, "What's the covariance of X_1 and Y?"
- Do we really want to count B?

What if we applied the same definition for the effect of X_2?
- The effect of X_2 would be $B + C$...
- We'd have counted B twice!

Isolating Variation



What's the solution?

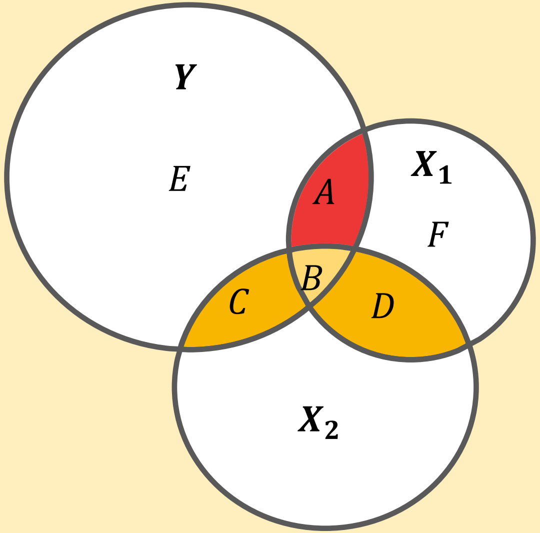
- Remove the covariance between X_1 and X_2 in B
- Isolate the ***unique*** covariance between X_1 and Y in A

This is what we mean when we say, "The effect of X_1 on Y controlling for X_2"

Revisiting Our Regression Equation

Let's return to the equation we started with: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u$

- Regression ***isolates*** variation in X_1 to estimate its effect on Y
- How do we calculate α_1?



Now the effect of X_1 on Y is α_1:

$$\alpha_1 = \frac{Cov(Y, \widetilde{X_1})}{Var(\widetilde{X_1})} = \frac{A}{A+F}$$

Where $\widetilde{X_1}$ represents the unique variation in X_1 after controlling for X_2

Introducing the Error Term

There's one last component of our regression that we haven't explored:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

The error term u_i reflects everything that isn't included in our model

Suppose we just estimated the relationship between Y and X_1

- Then our error term would include X_2!

$$Y_i = \beta_0 + \beta_1 X_{1i} + \boxed{\tilde{u}_i} = \beta_0 + \beta_1 X_{1i} + \boxed{(\beta_2 X_{2i} + u_i)}$$

Error Terms

In general, economic outcomes depend on lots of factors

- It's difficult to include every possible explanatory variable in a regression
- Relevant variables might be hard to collect … or impossible to observe

The key question – is there a variable missing in our regression that's correlated with our explanatory variables?

If so, then our β's won't be right – we'll explore what this means later

Staying Organized

Always remember (1) what's in your data set and (2) what you can calculate

Things that are in your data set:

- Outcome variable and explanatory variable(s) == your Y and X variables

Things you can calculate by estimating a regression:

- ***Coefficients*** describing the relationship between Y and X variables
- ***Predicted values*** of Y given your estimated coefficients and values of X
- ***Residuals*** == the difference between your actual Y's and the predicted Y's

What you ***can't*** see in your data or calculate is the regression error term u

Non-Linearity

We can use OLS to estimate regression equations like the following:

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

This let's Y be a function of X **and** X squared

OLS requires equations that are "linear in coefficients"

- In other words, equal to the sum of intercept + variables multiplied by β's
- OLS is ***flexible*** w.r.t. the functional form of individual explanatory variables

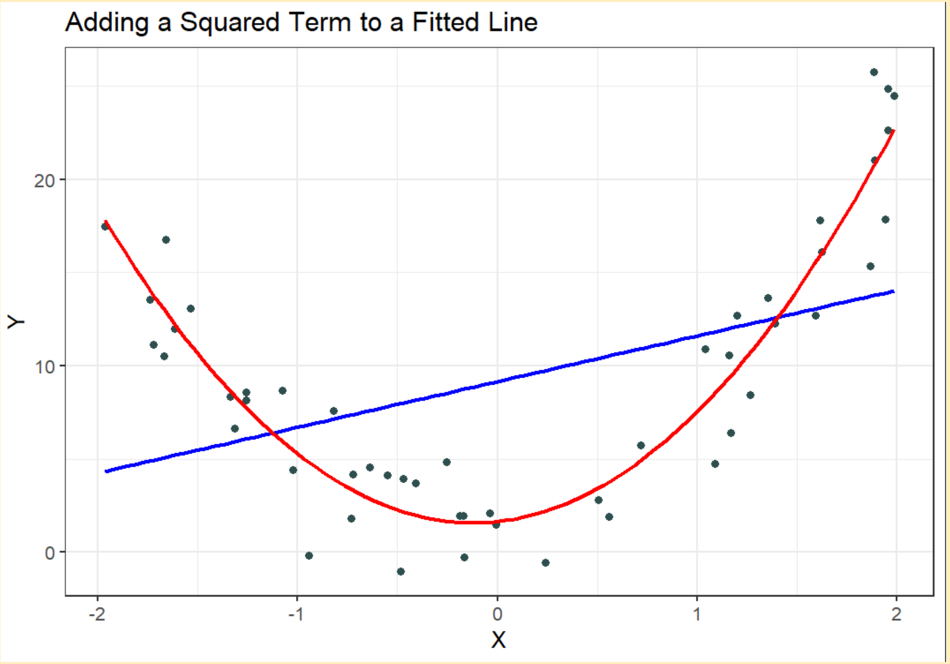
Squared Term Example

Equation for the blue line:

$$Y = \beta_0 + \beta_1 X + u$$

Equation for the red line:

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



(Natural) Logs

Another common data transformation is using the natural log function

- Widely used in metrics + stats, generally just referred to as log
- In R, calculated using the `log()` function

When using logs, we'll refer to the ***level*** of a variable X, and the log value $\log(X)$

Two useful things to remember:

1. Logs can make skewed distributions "better behaved" = more normal-looking
2. We can interpret small log differences as percentage changes

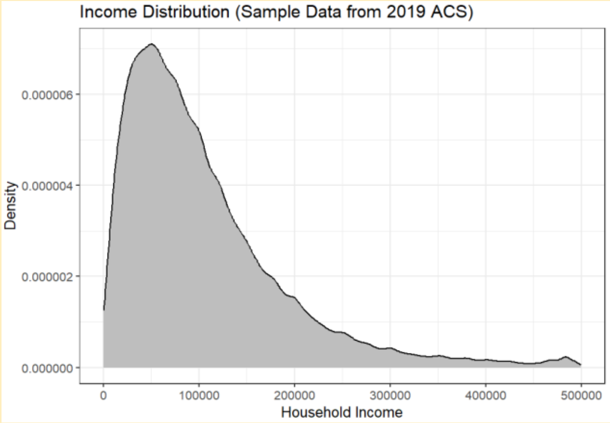
Logs and Distributions

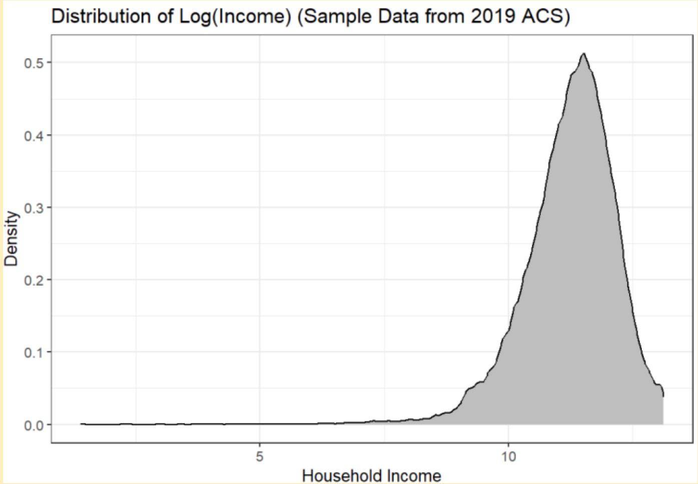
Distribution of household income in the top graph is skewed to the right

- In the bottom graph, distribution of log(income) is closer to normal
- There's a bit of a tail on the left (but mass is smaller)

One caveat – you can only calculate log of values > 0

- Means people with income <= 0 aren't in sample





Logs and Percentage Changes

We can interpret (small) increases in log values as percentage changes

- Log increase of 0.01 is ***approximately*** a $0.01 \times 100\% = 1\% \uparrow$ in original variable

If your income is \$22,026, your log income is 10

- Suppose your income rises by 1% to \$22,247 → log income is now 10.01
- Log difference 10.01 – 10 = 0.01

Year	Avg. Hourly Wage in CA	% Change YoY	Log(GDP)	Difference in Log Values
2022	\$37.44	3.80%	3.623	3.623 – 3.585 = 0.037
2021	\$36.07	4.85%	3.585	3.585 – 3.538 = 0.047
2020	\$34.40	.	3.538	

Percentage vs. Percentage Points

Related to logs, an important distinction you should know

- ***Percentage points*** = difference between two percentages

Suppose an NBA player makes 40% percent of their 3's in year 1 and 50% in year 2

- Did their 3-point shooting percentage increase by 10%? ***No!***

How to correctly characterize this change?

- Their shooting percentage increased by 25% → 100%x(50%-40%)/40% = 25%
- Their shooting percentage increased by 10-***percentage points*** (p.p.) → 50%-40% = 10 p.p.

Practice Problems

If the log value of X rises from 4.60 to 4.63, then the level (non-log) value of X has:

- Increased by approximately 3% → 4.63 – 4.60 = 0.03 log difference ≈ 3% increase
- Note this is an approximation (its close enough for up to log differences of ~0.2)

Suppose the inflation rate was 4% in Year 1 and 6% in year 2. We can say that:

1. The inflation rate increased by 50% → 100% x (6%-4%)/4% = 50%
2. The inflation rate increased by 2 ***percentage points*** → 6%-4% = 2 p.p.

Discrete Variables in OLS Regression

Until now, I've intentionally used examples with ***continuous*** variables

- What if we want to include a discrete variable?
- From our ACS data, suppose we want to control for employment status:

$$Household\ Income_i = \beta_0 + \beta_1 Employed_i + u_i$$

In words, this says "We're interested in conditional mean of household income, given employment status."

- Remember that if person i is working, $Employed_i = 1$ (and 0 otherwise)

Binary Variables in OLS Regression

Let's think about the equation from last slide separately for two groups of people:

$$Income_i = \beta_0 + \beta_1 Employed_i + u_i$$

For every person i who is ***not*** working, $Employed_i = 0$, so we have:

$$Income_i = \beta_0 + u_i$$

Remember that our intercept β_0 is just a single number

For everyone who's working, we'll have, $Employed_i = 1$:

$$Income_i = \beta_0 + \beta_1 + u_i$$

Key point = everyone gets the same β_0, so the term that differentiates people working and not working is β_1

Binary Variables in OLS Regression

In R, run `lm(hhincome ~ employed, acs.data)` → gives us the following output:

- β_0= `(Intercept)` = 96,012
- β_1= `Employed` = 33,970

We can calculate the conditional mean of earnings for…

- People who are ***not*** working = $\beta_0 = \$96{,}012$
- People who ***are*** working = $\beta_0 + \beta_1 = \$96{,}012 + \$33{,}970 = \$129{,}982$

Let's take a step back – what's β_1? It's just the average difference in earnings between workers and non-workers.

Connecting Regression and Summary Statistics

`employed` in regression output below tells us average difference in earnings between workers and non-workers

```
Call:
lm(formula = hhincome ~ employed, data = emp.data)

Residuals:
    Min      1Q  Median      3Q     Max
-136782  -70012  -31012   27418 1752918

Coefficients:
            Estimate Std. Error t value
(Intercept)    96012        760   126.2
employed       33970        847    40.1
```

We could also calculate this using the `mean()` function and calculating the difference:

```
> # Calculate average income for non-workers
>
> avg.inc.not.working <-
+   mean(acs.data[acs.data$employed == 0, ]$hhincome)
>
> avg.inc.not.working
[1] 96012
>
> # Calculate average income for workers
>
> avg.inc.working <-
+   mean(acs.data[acs.data$employed == 1, ]$hhincome)
>
> avg.inc.working
[1] 129982
>
> # Calculate difference in avg. income b/w workers & non-workers
>
> avg.inc.working - avg.inc.not.working
[1] 33970
```

Factor Variables in OLS Regression

In our ACS data, `employed` was binary (just two levels)

- `education` has 3 levels (1 = Non-HS grad, 2 = HS grad, 3 = College Grad)
- What happens if we include `education` in an OLS regression?

If we include any factor variable with more than 2 levels in a `lm()` regression:

- Intuitively, R starts by creating indicator variables for each level of variable
- Equal to 1 if an observation has that particular level of the variable (0 o/w)
- R will then include ***all but one*** of those binary variables in regression

Education Factor Variable Example

Suppose we run `lm(hhincome ~ as.factor(education), acs.data)` in R

We get the following output:

```
Coefficients:
                      Estimate Std. Error
(Intercept)              74114        957
as.factor(education)2    25948       1053
as.factor(education)3    99293       1096
```

What should we notice?

1. We had to tell R that `education` is a factor variable using `as.factor()`
2. We get ***two*** coefficients on `education` – corresponds coefficients on binary variables for the 2nd and 3rd levels of our factor variable (HS-grads & college grads)

Education Factor Variable Example

Remember that one group is always omitted

- Here, that's non-HS graduates (everyone with `education` = 1)
- What's their average earnings? It's just β_0 = `(Intercept)` = \$74,114

What about folks with higher education levels?

- For HS graduates (with `education` = 2), calculate their average earnings as:

$$\boxed{\texttt{(Intercept)} + \texttt{as.factor(education)2} = \$74{,}114 + \$25{,}948 = \$100{,}062}$$

- Finally, for college graduates (with `education` = 3), we have:

$$\boxed{\texttt{(Intercept)} + \texttt{as.factor(education)3} = \$74{,}114 + \$99{,}293 = \$173{,}407}$$

Omitted Levels

We do we need to do omit one of the levels of our education variable?

Whenever we run a regression with an intercept term, R automatically includes a column of 1's in our regression data (this happens automatically)

What happens if we included binary variables for every level of a factor variable?

- If we added all the 1's from each column, it would sum to 1 for each row
- The sum of these columns is then perfectly colinear with our intercept term

Regression Equations with Factor Variables

There's lots of different ways to write equations with factor variables. Here's one way:

$$Income_i = \beta_0 + \sum_{j \neq 1} \beta_{j-1} I(Education_i = j) + u_i$$

The $I(.)$ says, "create a binary variable equal to 1 when person i has an education level of j" – use $j \neq 1$ to be clear that the first level of $education$ won't be included

Equation above could be expanded as either of the following:

$$Income_i = \beta_0 + \beta_1 I(Education_i = 2) + \beta_2 I(Education_i = 3) + u_i$$

$$Income_i = \beta_0 + \beta_1 HS\ Grad_i + \beta_2 College\ Grad_i + u_i$$

General Tips on Handling Discrete Variables

For explanatory (X) variables:

- Any factor variable should be either (1) coded as a factor variable in your data set or (2) set as a factor variable using `as.factor()` in your `lm()` formula
- For a factor variable with N levels, your regression output should include $N-1$ coefficients corresponding to dummy variables for $N-1$ levels of your variable

For outcome (Y) variables:

- Numeric and binary (0 or 1) variables are okay to use with OLS / `lm()` models
- You ***cannot*** include factor variables with more than two levels – as an alternative, try creating a binary version of your variable (e.g., group multiple levels together)