ECON 253: INTRODUCTION TO ECONOMETRICS

Sebastian Anti, Assistant Professor of Economics
Bryn Mawr College

Lecture 4: Multivariate Regression

Summary

- **We know:**
  - The core estimator used in econometrics
  - A measure of the "goodness of fit"
  - The assumptions necessary for it to be unbiased, efficient, and precise
  - Basic intuition of the standard error as a measure of precision (but not how it is typically used for statistical inference)
- **All of this in the bivariate context with STRONG assumptions of a simplistic true model**

Ceteris Paribus Effects

- **Remember, we want to measure the causal effect of some variable X on Y**
  - In other words, we want to hold all other determinants of Y constant and just measure the effect of X
  - This is called a "*ceteris paribus*" effect
- **In the real world, you can't hold other determinants constant (like a lab scientist could)**
  - People with higher schooling have more social connections and on average higher ability
  - Villages with health clinics are closer to roads and their benefits
  - Populations that receive cash transfers might have worse coping mechanisms
  - **Why is this a problem?**

Why Multivariate Regression?

- **As we've said many times in studying bivariate regressions: In reality many variables determine any outcome**
  - If any of these variables is correlated with our variable of interest X and is not accounted for in the model then our slope parameter, $\hat{\beta}_1$, is biased
- **So, we want to account for these other variables**
- **Multivariate regression extends the bivariate approach to account for these other variables**
  - Add other relevant variables to the right hand side

Why Multivariate Regression?

- Multiple variable regression models allow us to estimate something approaching a *ceteris paribus* measure even though we can't hold these other variables constant in a lab
- By including multiple variables on the RHS of the regression equation we account for other influences on Y
- The estimated effect of X on Y is a measure "holding those other factors constant"

Notation

- For a multivariate model we have the following assumptions about the true model and the error term,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad (1)$$

$$E(u_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = 0 \;\forall\; i \qquad (2)$$

- This implies the following,

$$E(Y_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \qquad (3)$$

  - The error term falls out because of (2)

Enabling Ceteris Paribus Estimates

$$E(Y_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \qquad (3)$$

- **$\boldsymbol{\beta_j}$ (for $\boldsymbol{j = 1 \ldots k}$) is equal to the partial derivative of $\boldsymbol{Y}$ with respect to $\boldsymbol{X_j}$,**

$$\beta_j = \frac{\partial Y}{\partial X_j} \qquad (4)$$

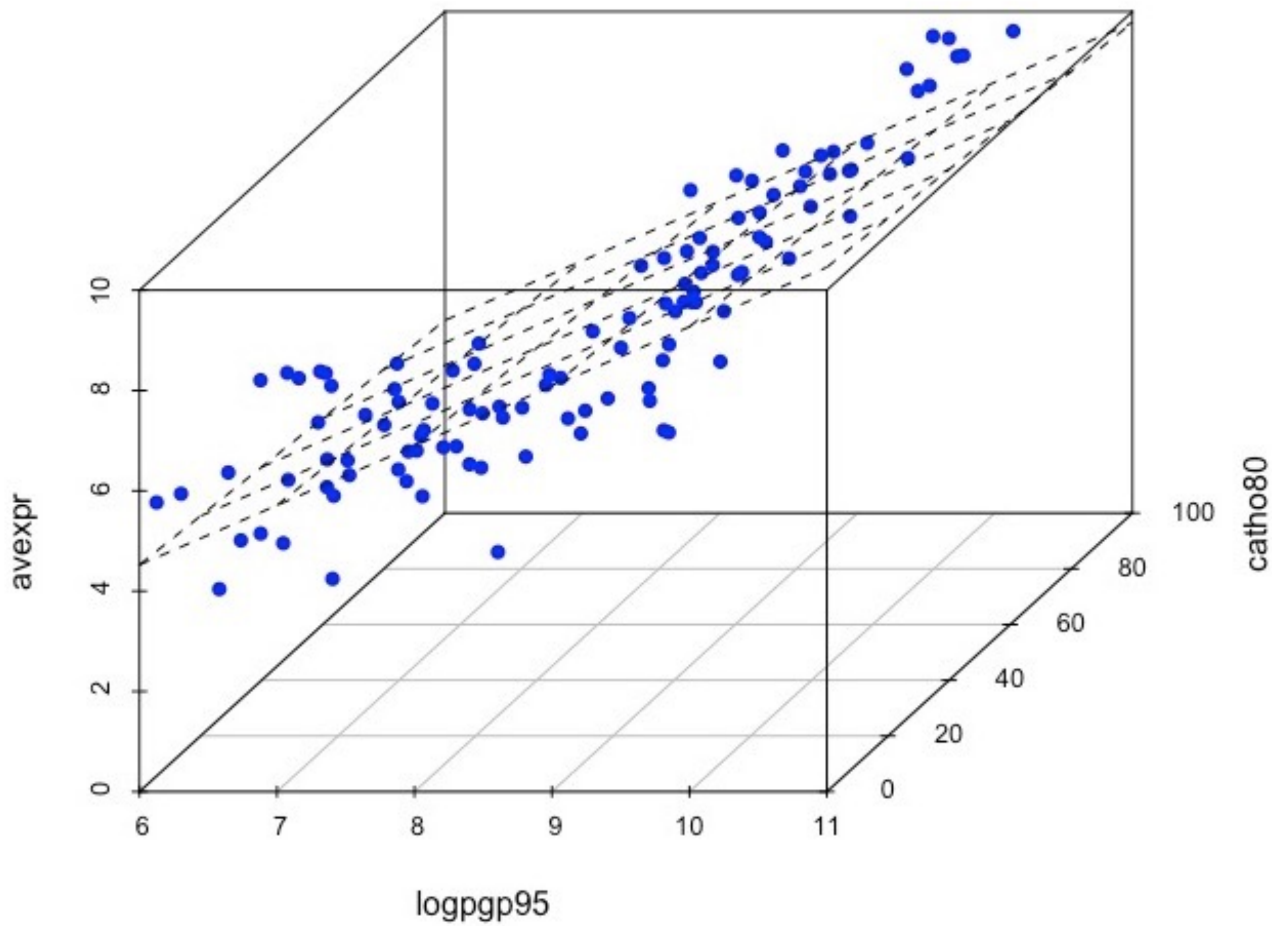
  - In other words, if you increased X_j by one unit, while holding all the other Xs constant, β_j is how Y would change

- **This is a *ceteris paribus* effect!**
- **For interested students, the formal theorem that proves that each coefficient is an effect on $\boldsymbol{Y}$ partialled out from the effects of the variables in the model is the Frisch-Waugh Theorem**

What's Going On Geometrically?

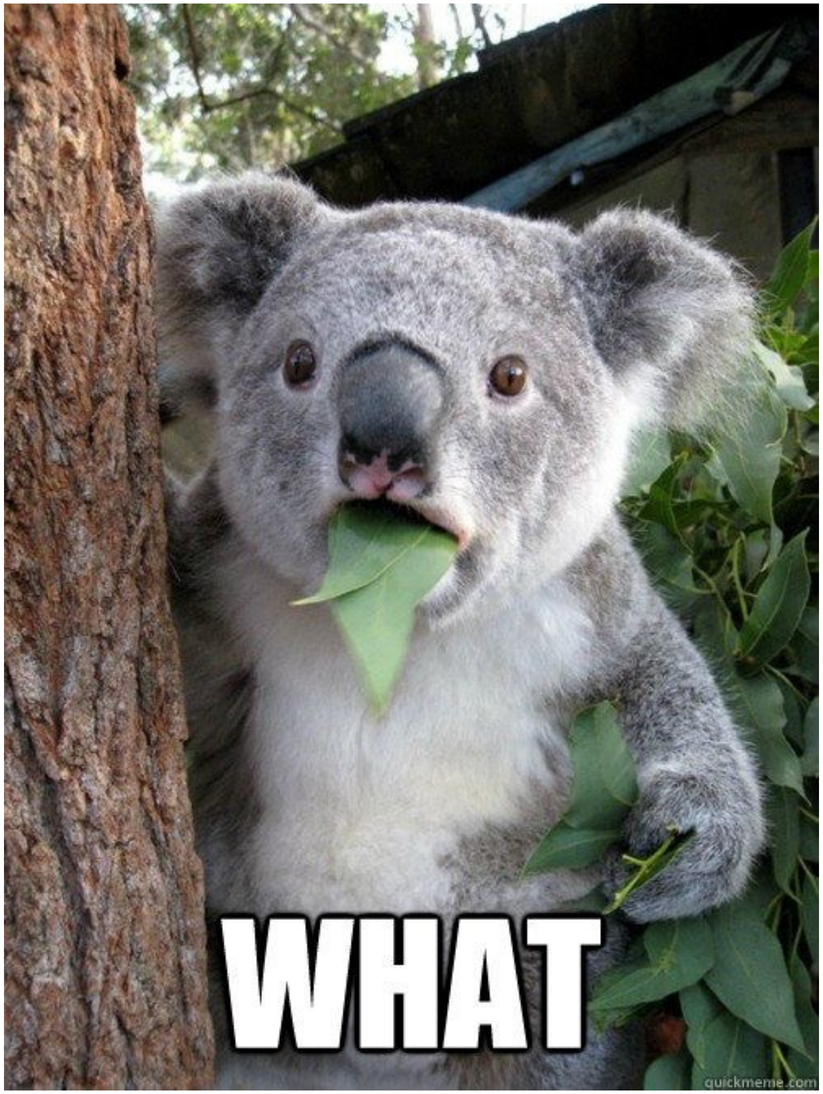
- **OLS in the bivariate linear regression model puts a line through the middle of a 2-dimensional scatter plot**
  - What does it do in 3 dimensions?
- **OLS for a 3-variable linear regression model puts a 2-dimensional plane through the middle of a 3-dimensional scatter plot of data**
- **More than 3 dimensions is impossible to visualize, but is easy to handle mathematically**
  - The estimated model provides the slopes of a multidimensional surface

Visualizing the Plane



- *X* axis is *ln*(GDP per Capita)
- *Y* axis is percent of population that is Catholic in 1980
- *Z* axis is Protection against Expropriation risk in 1990
- Data from Acemoglu et al. (2001)

How are we doing?



How Do We Estimate a Multiple Regression Model?

- **Consider the 3-variable multiple regression model,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \qquad (5)$$

- **To estimate $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_2}$, we calculate the residuals,**

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right) \qquad (6)$$

- **Then choose the parameter estimates to minimize the sum of squared residuals just as we did previously substituting the above expression**

How Do We Estimate a Multiple Regression Model?

- Here is the resulting minimization problem,

$$Min_{\{\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2\}} \sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right)^2 = \Phi\,. \tag{7}$$

- Taking derivatives provides three first-order conditions,

$$\frac{\partial \Phi}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{N} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right) = 0 \tag{8}$$

$$\frac{\partial \Phi}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{N} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right) X_{1i} = 0 \tag{9}$$

$$\frac{\partial \Phi}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^{N} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}\right)\right) X_{2i} = 0 \tag{10}$$

How Do We Estimate a Multiple Regression Model?

- **This is just three equations with three unknowns!**
- **The solutions:**

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2 \tag{11}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1)\sum_{i-1}^{N}(X_{2i} - \bar{X}_2)^2 - \sum_{i=1}^{N}(Y_i - \bar{Y})(X_{2i} - \bar{X}_2)\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{N}\ (X_{1i} - \bar{X}_1)^2(X_{2i} - \bar{X}_2)^2 - (\sum_{i=1}^{N}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2))^2} \tag{12}$$

- **The formula for $\widehat{\boldsymbol{\beta}}_2$ is the same as the equation for $\widehat{\boldsymbol{\beta}}_1$ just replacing $\boldsymbol{X_{1i}}$ for $\boldsymbol{X_{2i}}$**
- **The derivations for this and working with it to derive interesting properties are much easier using matrix notation (for this reason look to take a class in linear algebra if you are interested in PhD programs in economics):**

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X'X)^{-1}(X'Y) \tag{13}$$

- **Obviously equation (12) is really tedious to do by hand**
  - especially for large datasets and models with many variables
- **Luckily Stata and *R* can do these calculations in a matter of seconds (sometimes minutes for highly computationally intense regressions)!**

Coefficient Interpretation

- **The estimates of the model are,**

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \tag{14}$$

- **Let's say,**
  - Y is wage in USD per month
  - X_1 is years of education,
  - X_2 is years of work experience

- **What's the interpretation of $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\beta}}_1$, and $\widehat{\boldsymbol{\beta}}_2$?**
  - Don't forget to say "*ceteris paribus*" or "all else constant"

Goodness of Fit Measure in Multivariate Regressions

- **We can still use the R^2 measure we examined before,**

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = \frac{Explained\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} \quad (15)$$

- **Recall its properties and interpretation,**
  - Don't make too much of it in economics research
  - "Good" depends on the context
  - Range is between 0 and 1
  - The percentage of variation in Y explained by variation in RHS variables

Goodness of Fit Measure in Multivariate Regressions

- **Note: As one adds variables to the RHS the R^2 must stay the same or increase**
  - The TSS (the denominator) stays the same
  - The ESS (the numerator) can't go down. The worst that can happen is the coefficient is 0. In which case the ESS stays the same
    - This means you've added a variable with no relevant variation that explains Y
    - In the extreme, if we add enough variables we can fit the data perfectly
- **Implies: It's not fair to compare R^2 values for models with different numbers of variables on the RHS**

The Adjusted R^2

- **To compensate for the this "unfairness" we use the "adjusted R^2"**
  - Bounded by 0 and 1 (inclusive) and interpreted same as the R^2
  - Builds in a "penalty" for adding more RHS variables
  - This is frequently reported but not always. Often economists just report the R^2 and ignore this "unfairness" issue
    - This should give you an idea of how much economists actually care about the R^2

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(N - 1)}{(N - k - 1)} \tag{16}$$

Where to Find the Adjusted R^2


```
. reg earn_wk educ health married exper fthr_ed mthr_ed age

      Source |       SS           df       MS      Number of obs   =       741
-------------+----------------------------------   F(7, 733)       =      8.45
       Model |  3.5360e+09         7   505146714   Prob > F        =    0.0000
    Residual |  4.3810e+10       733  59767574.2   R-squared       =    0.0747
-------------+----------------------------------   Adj R-squared   =    0.0658
       Total |  4.7346e+10       740  63980620.1   Root MSE        =    7730.9

------------------------------------------------------------------------------
     earn_wk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -192.5014   1856.202    -0.10   0.917    -3836.608    3451.605
      health |  -317.6428   230.9344    -1.38   0.169    -771.0145    135.7288
     married |   2678.738   867.4809     3.09   0.002     975.6942    4381.781
       exper |  -547.4978   1870.382    -0.29   0.770    -4219.443    3124.447
     fthr_ed |   149.0461    86.1277     1.73   0.084    -20.04031    318.1324
     mthr_ed |   143.1557   102.8334     1.39   0.164     -58.7275    345.0389
         age |   667.4253   1864.065     0.36   0.720    -2992.117    4326.967
       _cons |  -7888.647   11287.14    -0.70   0.485    -30047.62    14270.33
------------------------------------------------------------------------------
```

How are we doing?

Generalizing the Conditions for OLS to be Unbiased

- **In the multivariate context, what must be true for OLS to be unbiased?**
  - **Assumption 1:** $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{Ki} + u_i$ is the true model
  - **Assumption 2:** $E(u_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = 0 \; \forall \; i$

- **These are just generalizations of the conditions we needed in the bivariate case**
  - But now we have much more reason to hope they are true!

Generalizing the Conditions for OLS to be Efficient

- **Recall we need all four assumptions for efficiency,**
  - **Assumption 1:** $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{Ki} + u_i$ is the true model
  - **Assumption 2:** $E(u_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = 0 \; \forall \; i$
  - **Assumption 3:** $Var(u_i | X_{1i} + X_{2i} + \cdots + X_{ki}) = \sigma^2$
  - **Assumption 4:** $Cov(u_i, u_j | X_{1i} + X_{2i} + \cdots + X_{ki}) = 0 \; \forall \; i \neq j$.

Generalizing the Conditions for OLS to be Precise

- **One can show that if Assumptions 1 through 4 hold, the following is true,**

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_{ji} - \bar{X}_j)^2 (1 - R_j^2)}. \qquad (17)$$

j **here denotes any one of the other RHS variables**

- **Where the R_j^2 is the R^2 from a regression of X_j on all the other RHS variables (not the R^2 from the regression you are running)**
  - It tells us how much variation in this RHS variable is explained by variation in the other RHS variables

Generalizing the Conditions for OLS to be Precise

- **Equation (17) says that the OLS estimator will be more precise (have lower variance) if:**
  - The true variance of the errors is smaller (less noise)
  - The dataset contains more variation in X_j around its mean
  - The sample size (N) is larger
  - X_j suffers from less <u>multicollinearity</u> with other RHS variables
    - In other words X_j is less related to the other variables on the RHS

Generalizing the Conditions for OLS to be Precise

- **Ideally you want:**
  - Less noise in the data
  - A large sample size
  - Low multicollinearity
  - Lots of variation in X_j

- **Note that you can't even run OLS to estimate $\widehat{\beta}_j$ if:**
  - Your sample size is smaller than the number of variables on the RHS $(n < k)$
  - X_j has no variation
  - X_j is perfectly multicollinear with the other RHS variables

Generalizing the Conditions for OLS to be Precise

- **Note that this requirement that X_j not be multicollinear says nothing about the relationship between X_j and the error term u**
- **Which of the following models violates perfect multicollinearity?**

$$InfantMort_i = \beta_0 + \beta_1 GDPperCap_i + \beta_2 DocsperCap_i + u_i \qquad (18)$$

$$\boxed{InfantMort_i = \beta_0 + \beta_1 GDPperCap_i + \beta_2 10 \times GDPperCap_i + u_i \qquad (19)}$$

$$InfantMort_i = \beta_0 + \beta_1 GDPperCap_i + \beta_2 GDPperCap_i^2 + u_i \qquad (20)$$

How to Estimate σ^2 in a Multivariate Model

- **If Assumptions 1 through 4 hold, then the following is an unbiased estimator of σ^2,**

$$\hat{\sigma}^2 = \frac{1}{N-k-1} \times \sum_{i=1}^{N} \hat{u}_i^2 \tag{21}$$

- **Again, this is just a generalization of the formula we got for the bivariate model**
  - Plug $k = 1$ and you'll just derive the formula in the bivariate context

Standard Errors in Multivariate Models

- **The standard error for the coefficient on variable X_j is therefore,**

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\frac{\frac{1}{N-k-1} \times \sum_{i=1}^{N} \hat{u}_i^2}{\sum_{i=1}^{N}\left(X_{ji}-\bar{X}_j\right)^2 (1-R_j^2)}} \qquad (22)$$

- **Tedious to calculate**
  - Stata or R does this for you!
- **Summarizes our confidence in our results**
  - We'll later use the standard error extensively in interval estimation and hypothesis testing

Summary So Far

- **So, multivariate models are an extension of bivariate analysis**
  - All our previous assumptions and formulas translate nicely to this new framework
  - Calculations become tedious but computers solve this issue for us
    - But you need to know well what the computer is doing before implementing any of this!
- **Allows us to be more confident that the assumptions needed to implement precise, efficient and unbiased OLS are satisfied**
- **But…**

This Opens Up So Many Questions!

- **How do we choose the right variables for the RHS?**
  - What happens if we include irrelevant variables?
  - What happens if we omit relevant variables?
- **What transformation should we use of the variables?**
  - What should be the functional form of the model?

How to Choose Your RHS Variables

- **Although you can find many efforts at an empirical approach to selecting the RHS variables, <u>there is no universally accepted approach to this in economics</u>**
  - Choosing the set of RHS variables that maximizes the R^2 is *not* a valid way to choose
- **Generally, the RHS variables should be every variable in your dataset that *theory* suggests is related to the outcome variable**
  - Yes. This is up to the subjective judgement of the analyst and their peer reviewers

Including Irrelevant Variables

- **This is called "overspecifying the model"**
  - Why might this happen?
    - Theory suggests the variable is relevant but in reality it is not
- **Suppose the true model is as follows,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{23}$$

But we estimate,

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \varepsilon_i \tag{24}$$

Including Irrelevant Variables

- **The parameter estimates remain unbiased,**
  - $E(\hat{\alpha}_j) = E(\hat{\beta}_j)\ for\ j = 1, 2$
  - $E(\hat{\alpha}_3) = 0$
- **However,** $\boldsymbol{Var(\hat{\alpha}_j) > Var(\hat{\beta}_j)\ for\ j = 1, 2}$
  - This diminishes precision and means that our standard errors will be larger. This means that including an irrelevant variable will *lower* our confidence in our estimates
    - **This isn't such a bad thing. Indeed it is an extremely desirable property of OLS:** If we are worried about an irrelevant variable but our estimates are statistically significant, addressing the irrelevant variable can only increase our statistical significance!

Omitting Relevant Variables

- **Why might we omit a relevant variable?**
  - We forgot to include it because we failed to think carefully about what the model should be. (i.e. we took an ”ad hoc” approach)
  - The theory we used was too simple or just plain wrong
  - We assumed a functional form that was too simple (example: we include X but not X^2)
  - We don’t have a measure of the variable in our dataset
    - This is by far the most common reason for an omitted variable
      - Either the dataset is limited and didn’t measure the variable, or
      - The variable is “unobservable” and therefore not measurable
        - Can you think of an example of this?

Omitting Relevant Variables

- **Suppose the true model is,**

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \varepsilon_i \qquad (25)$$

But we estimate,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \qquad (26)$$

- **Well… That meant the error term in the model we estimated is:** $\boldsymbol{u_i = \alpha_3 X_{3i} + \varepsilon_i}$
  - If the omitted variable is uncorrelated with the variables in the regression (in other words, if $Cov(X_{3i,}\ X_{1i}) = 0$, and $Cov(X_{3i,}\ X_{2i}) = 0$) then there is no problem
  - If the omitted variable is correlated with our variable(s) of interest on the RHS, well that means Assumptions 1 and 2 are violated. **This is a huge problem**

Omitting Relevant Variables

- **If Assumptions 1 and 2 are violated, that means our coefficient estimates are biased ($E(\widehat{\boldsymbol{\beta}}_j) \neq \boldsymbol{\beta}_j$)**
  - In other words, our measurement is wrong
  - Obviously, this is very bad because getting a good estimate is the whole point of doing this
  - It also renders all statistical inference based on $\hat{\beta}_j$ invalid
- **Dealing with Omitted Variable Bias (OVB) is a major obstacle when doing econometrics**
  - Many more advanced econometric methods address OVB. Most research using econometrics spends a lot of time discussing OVB in the context of the study
    - We'll get to these advanced methods in the last part of the class. Also we'll see in more detail why OVB is so bad. <u>For now, its enough to just know that it is a major challenge for any applied econometrician</u>

Choosing The Correct Functional Form

- **For OLS to be unbiased, the structure of the model we estimate must be an accurate representation of true model**
- **To this point we've only looked at linear models**
  - But many real world relationships are nonlinear! So we need models that allow us to estimate these relationships
- **OLS only requires the model be "linear in its parameters". This means we can still transform our variables to estimate nonlinear relationships using OLS!**

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

What Do We Want to Understand?

- **What possibilities does each model allow for?**
- **What assumptions does each model impose on the ways X and Z relate to Y?**
- **In other words, what does the model say about:**
  - What happens to the size of the effect of X on Y as X increases?
  - What happens to the size of the effect of X on Y as Z increases?
- **How do we interpret the estimated coefficient for each model?**
- **How do economic relationships drive our decision over functional form?**

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Starting Point

- **We'll start with the model we know well**
  - The "Linear-Linear" Model
- **We'll demonstrate the ways this model is highly restrictive**
- **We'll then build on this to consider the complexities allowed or disallowed by other functional form decisions**

Dimensions to Consider

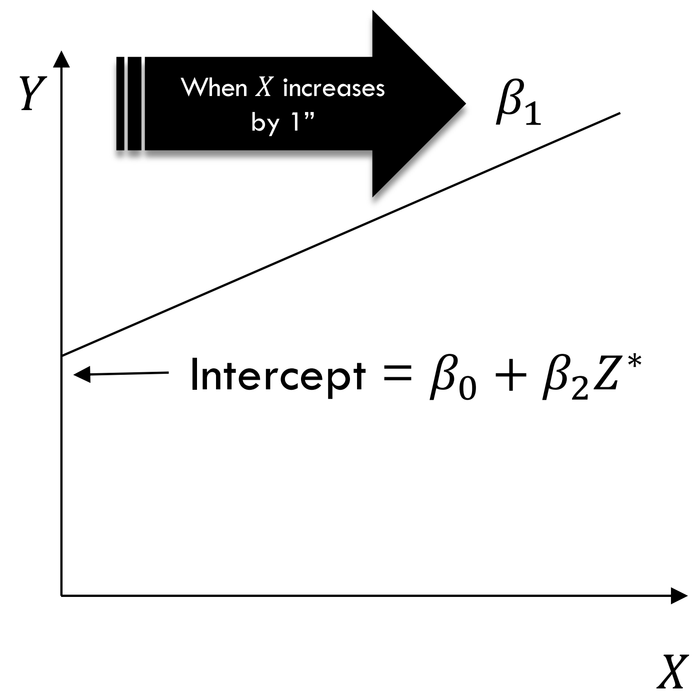
- What happens to the size of the effect of X on Y as the level of X increases?
- How do we interpret the units and sizes of the estimated coefficient?
- How do socio-economic relationships we are studying lead us to be interested in or reject specific functional form choices?

The "Linear-Linear" Model

- **Recall the "Linear-Linear" Model:**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \qquad (27)$$

  - How Y changes as X changes $\left(\frac{\partial Y}{\partial X}\right)$ is β_1
  - How Y changes as Z changes $\left(\frac{\partial Y}{\partial X}\right)$ is β_2
  - When we look at the slope for our variable of interest, we assume the second explanatory variable is held constant at Z^*



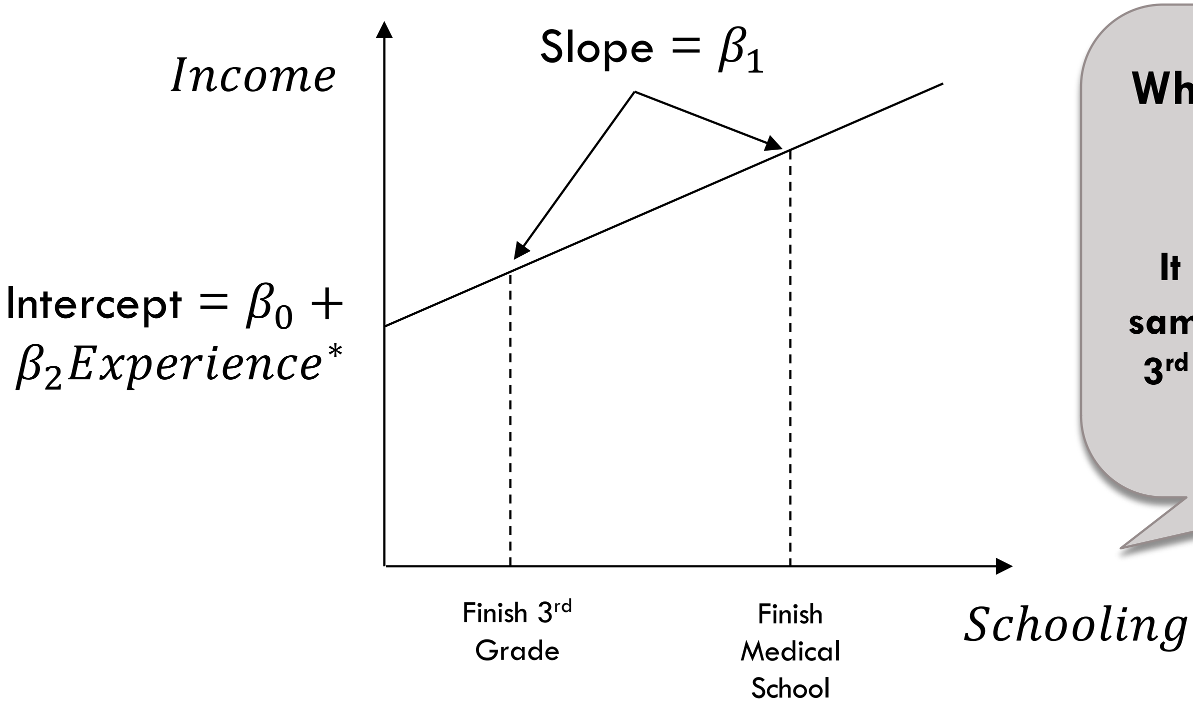
The "Linear-Linear" Model: Options Allowed

- **When we assume a linear relationship between Y, X, and Z, we aren't making any assumptions about the signs or magnitudes of β_0, β_1, or β_2**
  - All three can be negative or positive
  - All three can take on any magnitude
    - This is pretty flexible!
- **However, there are restrictive assumptions in here**

Assumption 1: Effect of X on Y is Constant for All Levels of X

- **Consider the following regression:**

$$Income_i = \beta_0 + \beta_1 Schooling_i + \beta_2 Experience_i + u_i \tag{28}$$



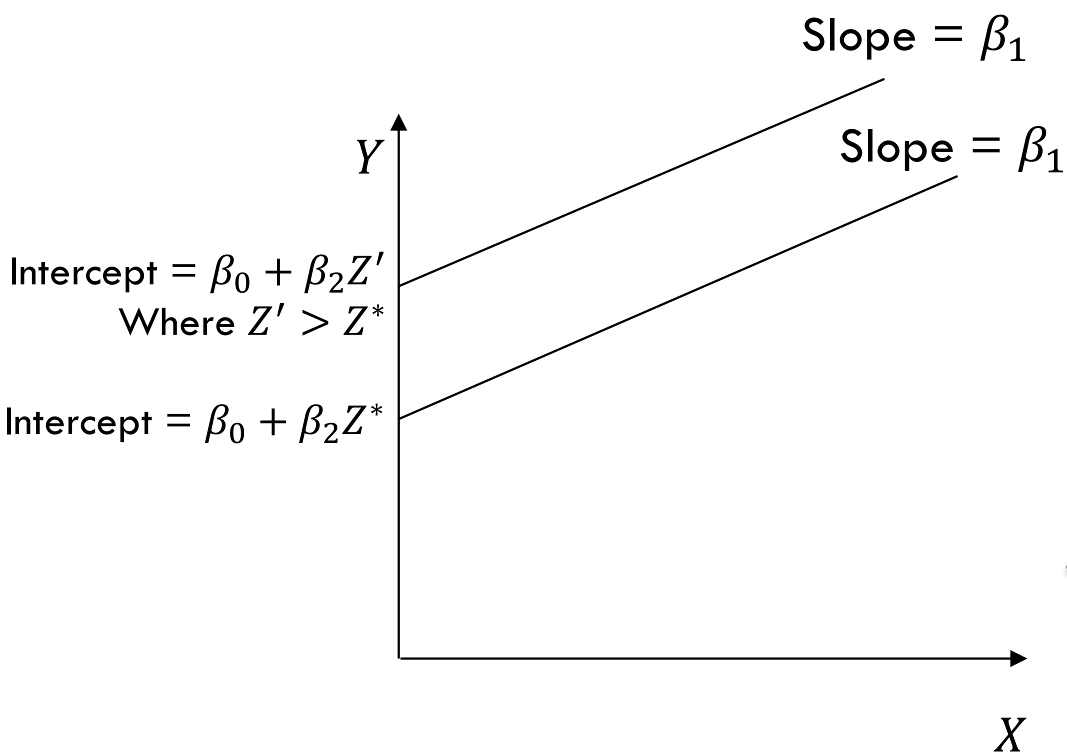
What is worrisome about this?

It says income changes the same amount after completing 3rd grade as after completing medical school

Assumption 2: Effect of X on Y is Constant for All Levels of Z

- **Again here is the model,**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \quad (29)$$



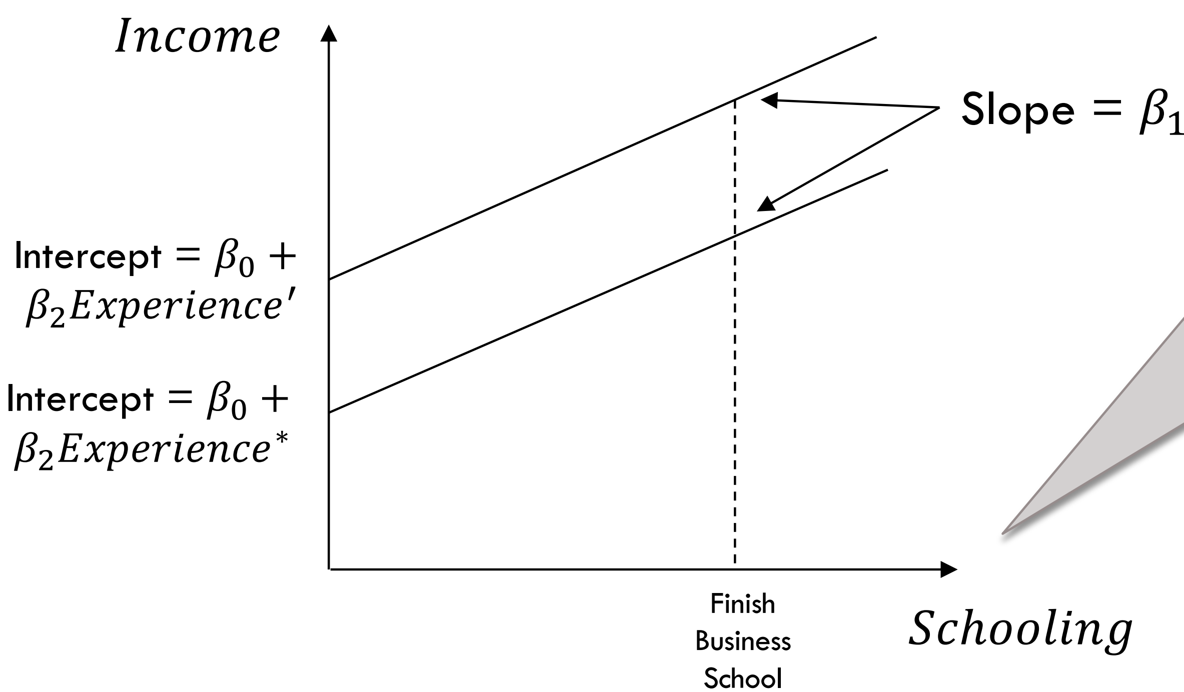
What is worrisome about this?

Can you think of an example that where this wouldn't apply?

Assumption 2: Effect of X on Y is Constant for All Levels of Z

- Consider (again) the following regression:

$$Income_i = \beta_0 + \beta_1 Schooling_i + \beta_2 Experience_i + u_i \tag{30}$$



What is worrisome about this?

It says that the effect of business school on earnings is the same no matter how much work experience start with

"Linear-Linear" Model: Interpretation of Coefficients

- **As we've said before,**
  - β_1 describes how much Y changes as X changes by one unit $\left(\frac{\partial Y}{\partial X}\right)$

- **β_1 answers the question: "By how many units does Y increase when we increase X by one unit?"**

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Some Quick Math

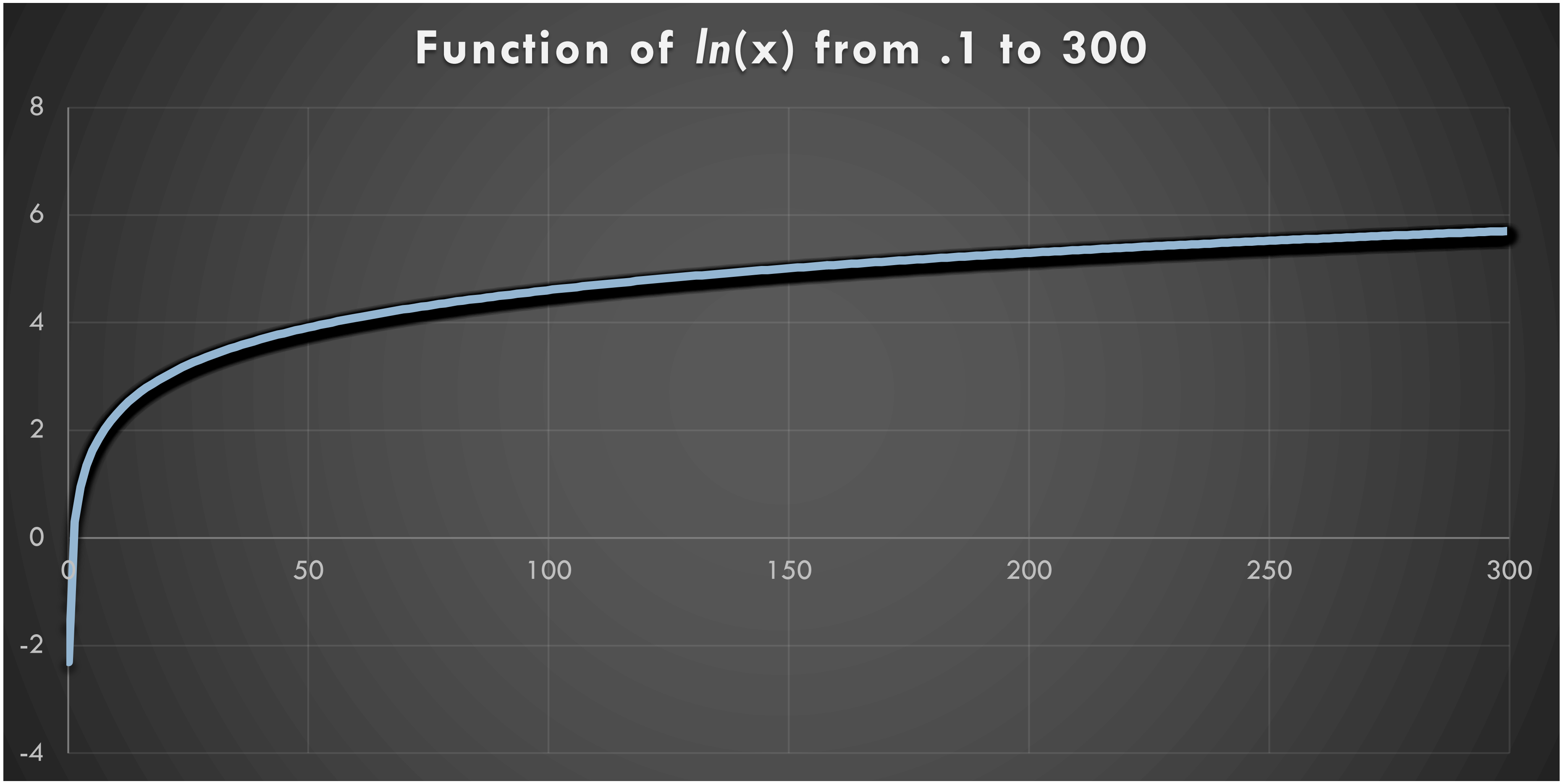
- **A logarithmic transformation of base b of variable X results in the power b would need to be raised to in order to produce X, so**

$$log_b(X) = y, where\ b^y = X \tag{31}$$

  - Usually we use a base of e (also known as "Euler's Number" ≈ 2.718). This is also known as the "natural log" and is abbreviated $ln()$
  - This is a "monotonic" transformation in that it preserves the order of X

The Logarithmic Function

- $f(x) = ln(x)$ **has the following shape:**



The Logarithmic Function

- **Note that $ln(0)$ is undefined**
- **This means that logging a variable that contains meaningful 0s will mean a loss of that observation in your data**
- **To address this, researchers have two approaches**
  1. Replace 0s with some very small positive number like 0.00001
  2. Use the "Inverse Hyperbolic Sine" (IHS) Transformation,

$$IHS(x) = ln\left(x + \sqrt{x^2 + 1}\right) \tag{32}$$

Logs and the Inverse Hyperbolic Sine

- **In the past, the first approach has been the most common, but the IHS transformation is now the more accepted way to deal with this**
  - You can treat a variable transformed by the IHS transformation ***exactly*** like a log-transformed variable!
  - See the following for more on the IHS transformation,

Bellemare, Marc F. and Casey J. Wichman (2020), "Elasticities and the Inverse Hyperbolic Sine Transformation," *Oxford Bulletin of Economics and Statistics* 82 (1): 50-61.

How are we doing?

Log-Linear Model

- These are models of the form,

$$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i \tag{33}$$

- You can rewrite this to have Y on the LHS as follows,

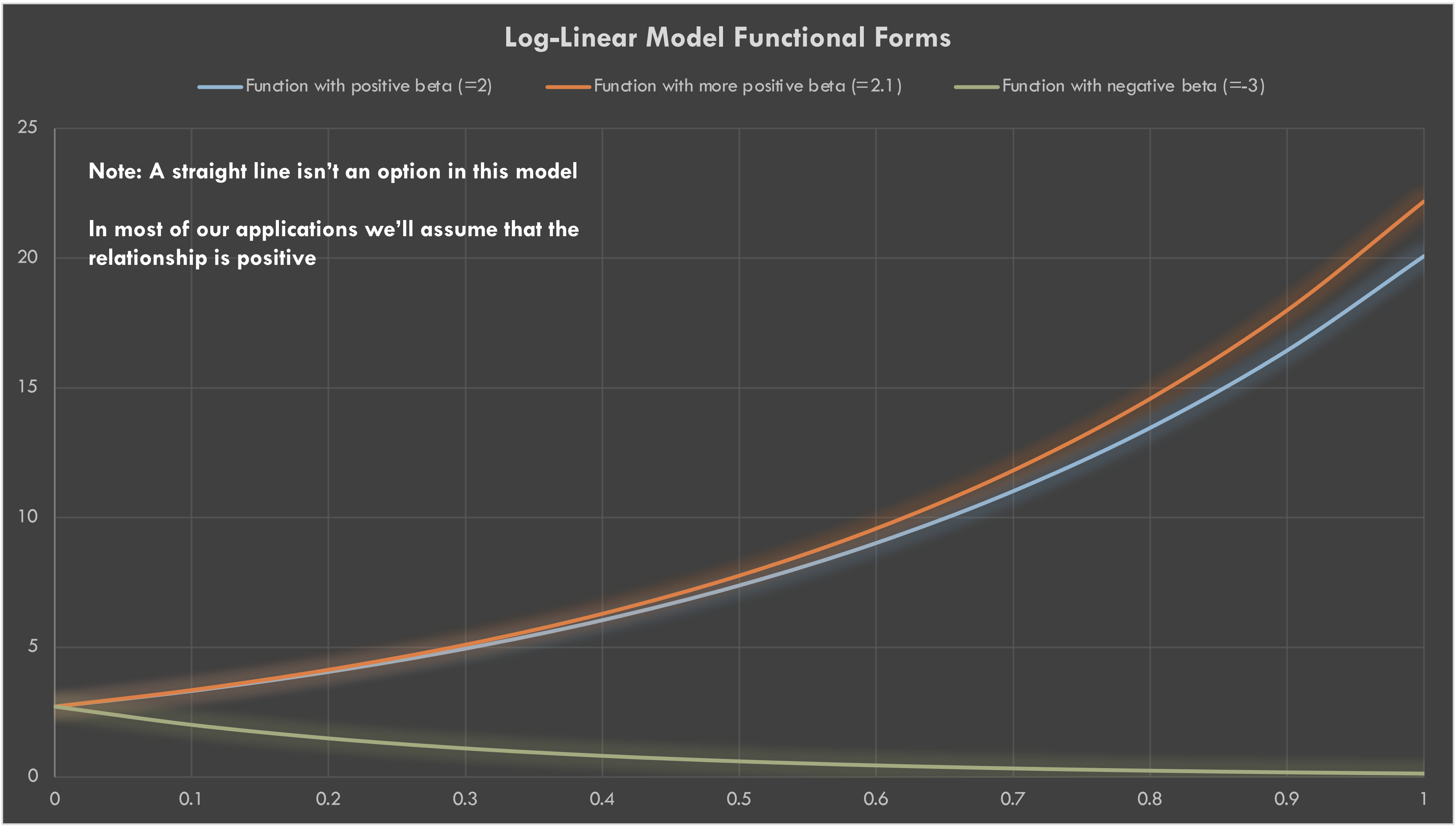
$$Y_i = e^{\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i} = e^{\beta_0} e^{\beta_1} e^{\beta_2} e^{u_i} \tag{34}$$

- The partial derivatives are:

$$\frac{\partial Y}{\partial X} = \beta_1 e^{\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i} = \beta_1 Y_i \tag{35}$$

$$\frac{\partial Y}{\partial Z} = \beta_2 e^{\beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i} = \beta_2 Y_i \tag{36}$$

Log-Linear Model: Options Allowed

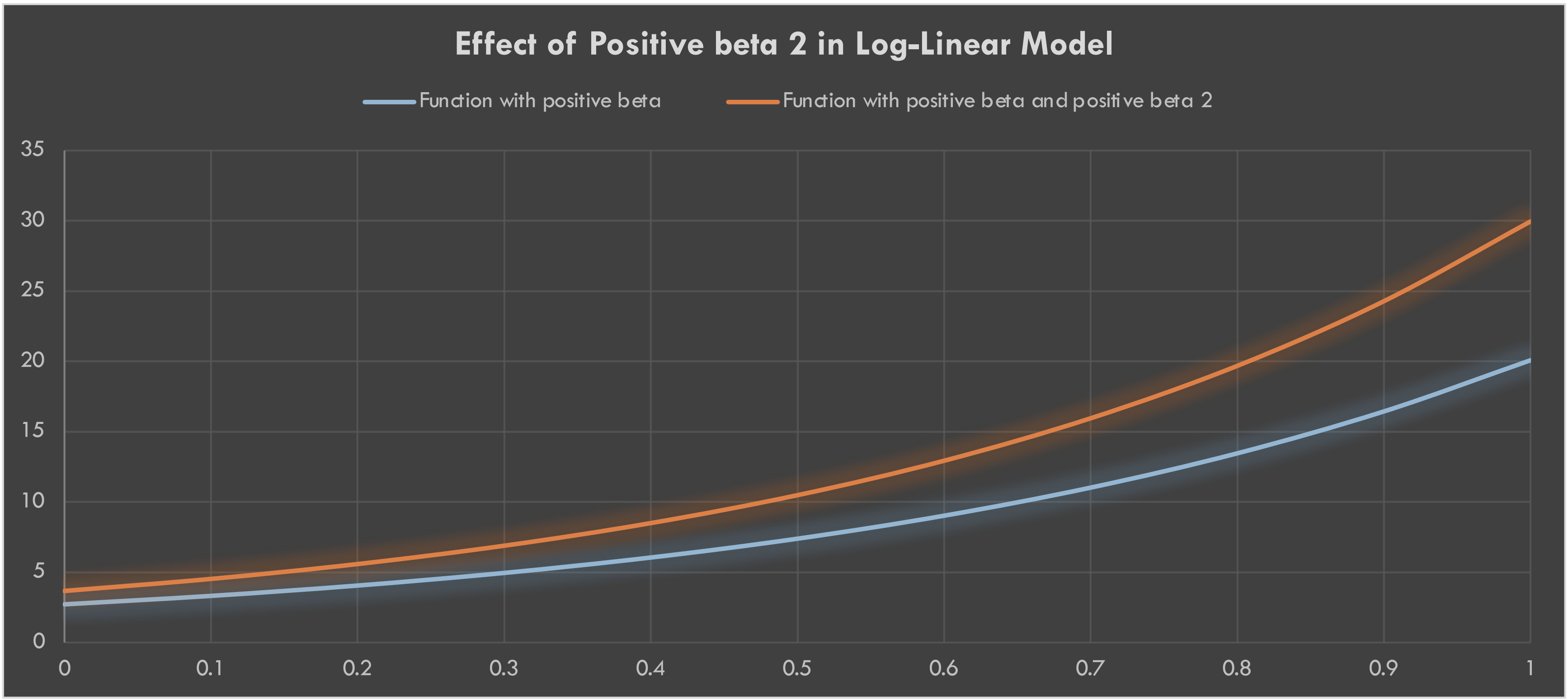


Log-Linear Model

- **So, as we move to the right on the graph with a positive coefficient, the slope gets steeper**
  - Why might this be useful for thinking about the relationship between schooling and wages?
  - Why might this be a bad functional form for a production function?

Log-Linear Model: Effect of X on Y Rises with Z if $\beta_2 > 0$

- As Z increases, the intercept shifts up and the graph swivels up. At any level of X, the effect of an additional unit of X on Y is now larger.



- Why might this be a useful property in modelling the effect of schooling and experience on earnings?

Example of a Log-Linear Model

- So let's consider the classic regression,

$$ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exp_i + u_i \qquad (37)$$

- Let's re-express this with Y on the LHS,

$$Wage_i = e^{\beta_0 + \beta_1 Educ_i + \beta_2 Exp_i + u_i} \qquad (38)$$

- Now, I'll estimate using data again from Blattman and Annan (2010)

Log-Linear Model: Interpretation of Coefficients

- **The slope coefficient on X is interpreted as a percentage in increase in Y for a one unit change in X**
  - It answers the question "By what proportion does Y increase when X increases by one unit?"
- **So if $\widehat{\beta}_1 = 0.3$ then a one unit increase in X corresponds with a 30 percent increase in wages**
  - You'll often hear this referred to as a "semi-elasticity" in papers

Log-Linear Model: Interpretation of Coefficients and Miscellaneous Points

- **No matter what unit Wage is in, the coefficient is always interpreted as a *percentage increase***
- **Percentages are free of units, this is nice because it allows us to infer whether the effect size is large or small easily**
  - We still need to consider the units of X though

- **Miscellaneous points on the Log-Linear Model:**
  - All Y values must be positive
    - Inverse hyperbolic sine or adding 0.1 or something smaller is a good alternative if data contains meaningful zeros (You'll get experience with this in problem sets)

A Brief Summary of the Log-Linear Model

- **So, based on this what types of relationships can be best captured in a Log-Linear Model?**
  - When the effects of the RHS variables get bigger as the levels of the RHS variables get larger
  - When the RHS variables are complementary

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Log-Log Models

- These are models of the form,

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i \qquad (39)$$

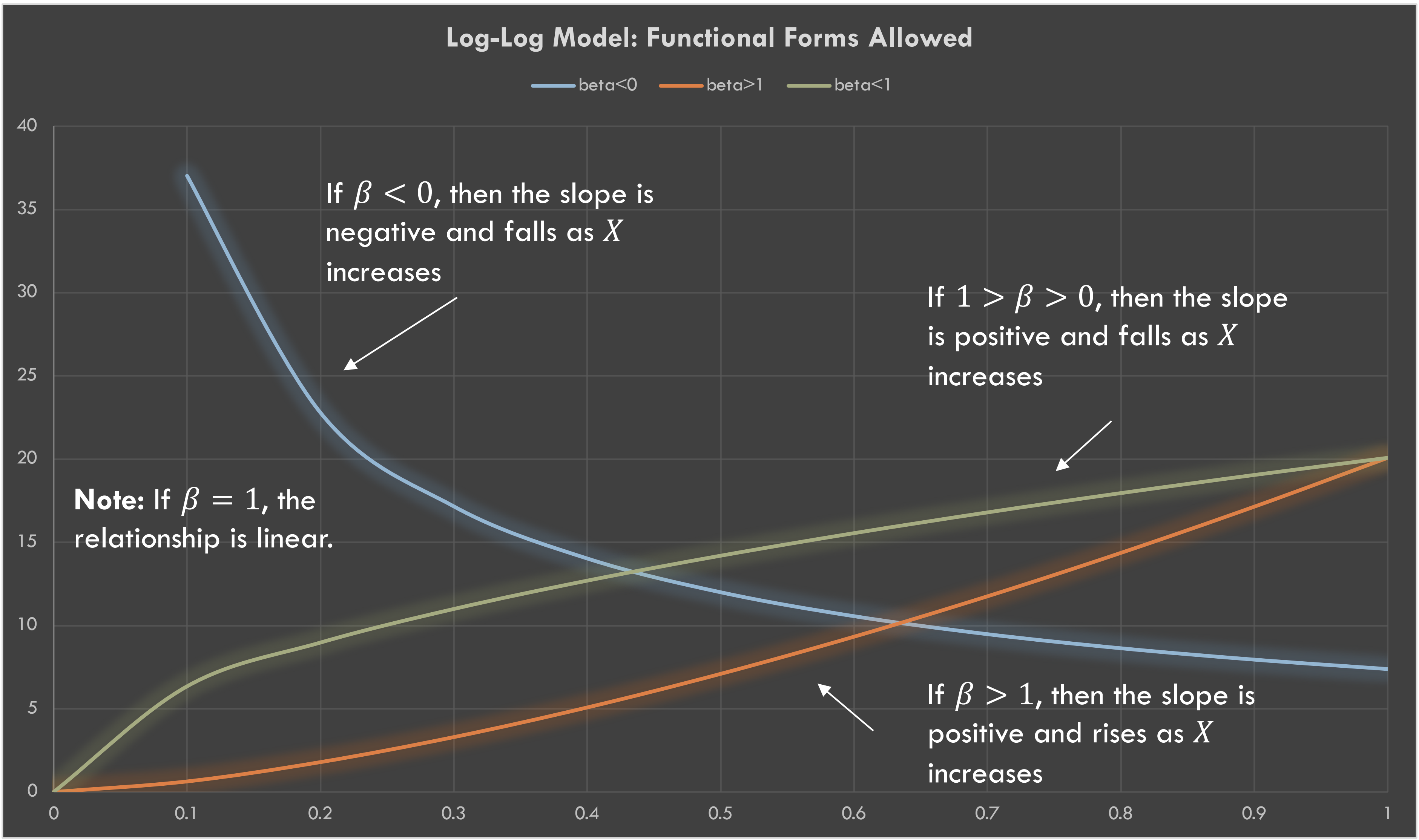
- You can rewrite Y to be on the LHS as follows,

$$Y_i = e^{\beta_0} X_i^{\beta_1} Z_i^{\beta_2} e^{u_i}. \qquad (40)$$

- The partial derivatives are of the form:

$$\frac{\partial Y}{\partial X} = \beta_1 e^{\beta_0} X_i^{(\beta_1 - 1)} Z_i^{\beta_2} e^{u_i} = \frac{\beta_1 e^{\beta_0} X_i^{\beta_1} Z_i^{\beta_2} e^{u_i}}{X_i} = \beta_1 \frac{Y_i}{X_i} \qquad (41)$$

Log-Log Models: Options Allowed



Log-Log Model Production Function

- Log-Log Models are common. One place you'll see them a lot is estimates of production functions
- To see why, let's start with a production function with a Cobb-Douglas functional form,

$$Y_i = e^{\beta_0} L_i^{\beta_1} K_i^{\beta_2} e^{u_i} \tag{42}$$

- Using log rules, we can re-express this as a linear equation as follows,

$$ln(Y_i) = \beta_0 + \beta_1\, ln(L_i) + \beta_2\, ln(K_i) + u_i \tag{43}$$

Log-Log Model: Interpretation of Coefficients

- **The slope coefficient answers the question "When X increases by 1 percent, by what percentage does Y increase?"**
- **Thus, the slope coefficient in the Log-Log Model has the interpretation of an elasticity. Remember those from Econ 105?**
  - They're actually great because they don't have units

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Linear-Log Models

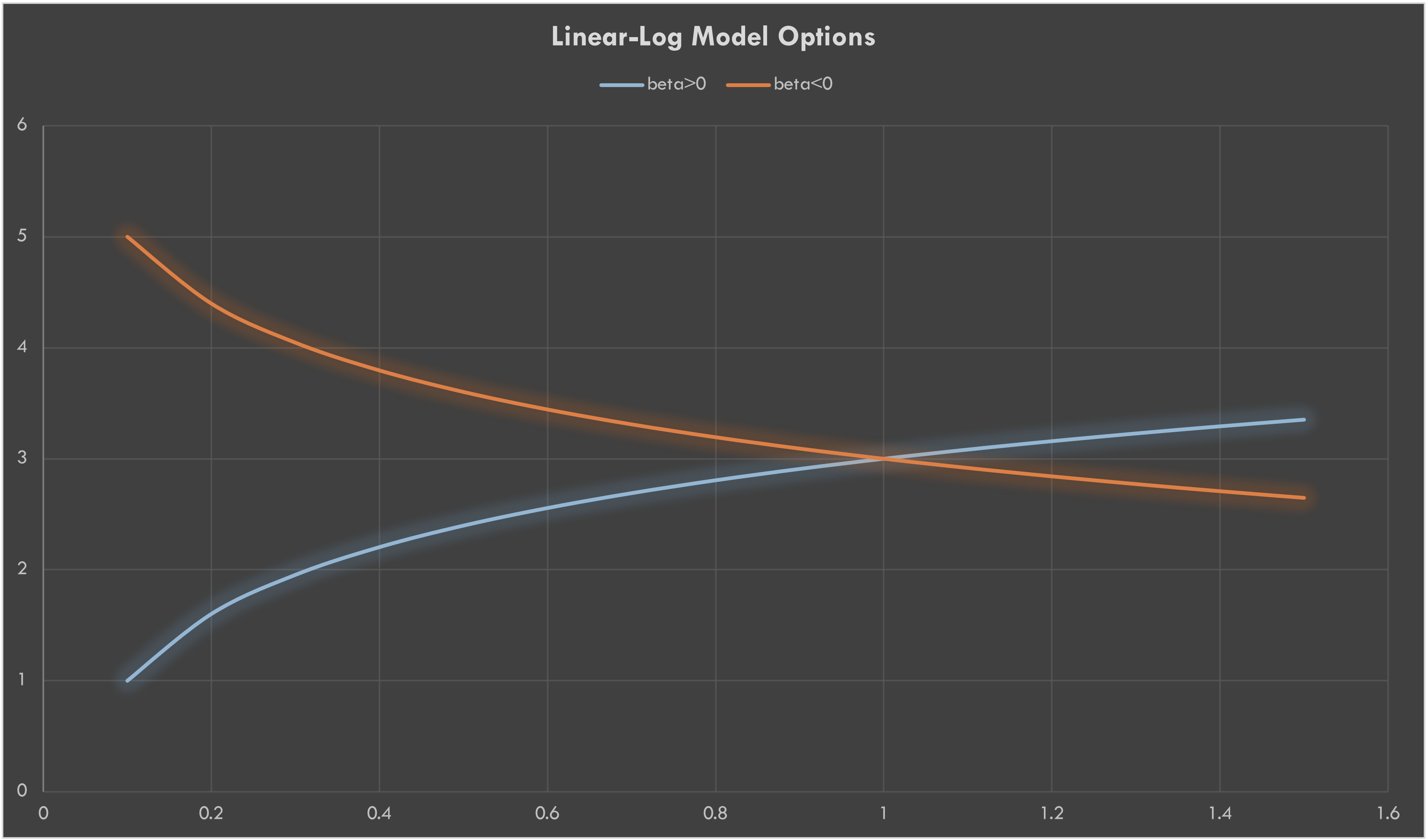
- These are models of the form,

$$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i \tag{44}$$

- The partial derivatives are of the form,

$$\frac{\partial Y}{\partial X} = \beta_1 \frac{1}{X_i} \tag{45}$$

Linear-Log Models: Options Allowed



The Effect of Z on the Relationship between X and Y and Coefficient Interpretation

- **Unlike the other functional forms, the effect of X on Y does not depend of Z**
  - Easy!

- **Coefficient Interpretation**
  - The slope coefficient here answers the question "When X changes by 100 percent, by how many units does Y change?" or
  - Whenever X doubles, by how many units does Y change?

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Quadratic Models

- These are models of the form,

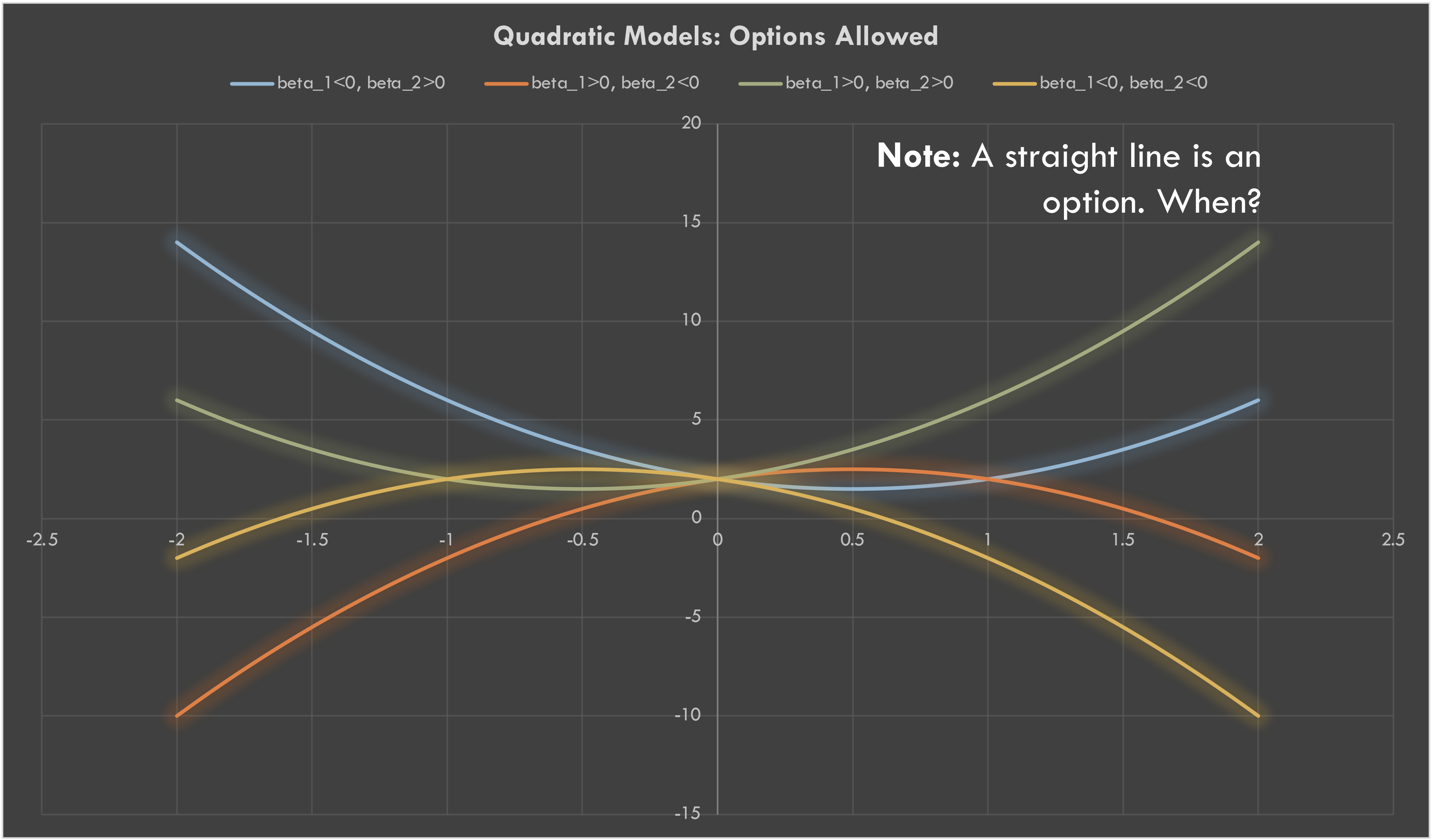
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i \qquad (46)$$

- The partial derivatives are,

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X_i \qquad (47)$$

$$\frac{\partial Y}{\partial Z} = \beta_3 \qquad (48)$$

Quadratic Models: Options Allowed



Quadratic Models: Options Allowed

- **There are some things to note here:**
  - If Z changes, the intercept changes but that's all
    - The effect of X on Y does not depend on Z
  - A straight line is an option. What's the value of β_2 if this is the case?
    - 0!

Quadratic Models: Interpretation of Coefficients

- **Notice that in quadratic models two coefficients are involved in describing the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$: $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$**
  - If β_1 is negative, then when $X = 0$, the slope is negative
    - If β_2 is negative then the effect of X on Y gets more negative as X increases, if β_2 is positive then it gets less negative
  - If β_1 is positive then when $X = 0$, the slope is positive
    - If β_2 is negative then the effect of X on Y gets less positive as X increases, if β_2 is positive then it gets more positive

Quadratic Models: Interpretation of Coefficients

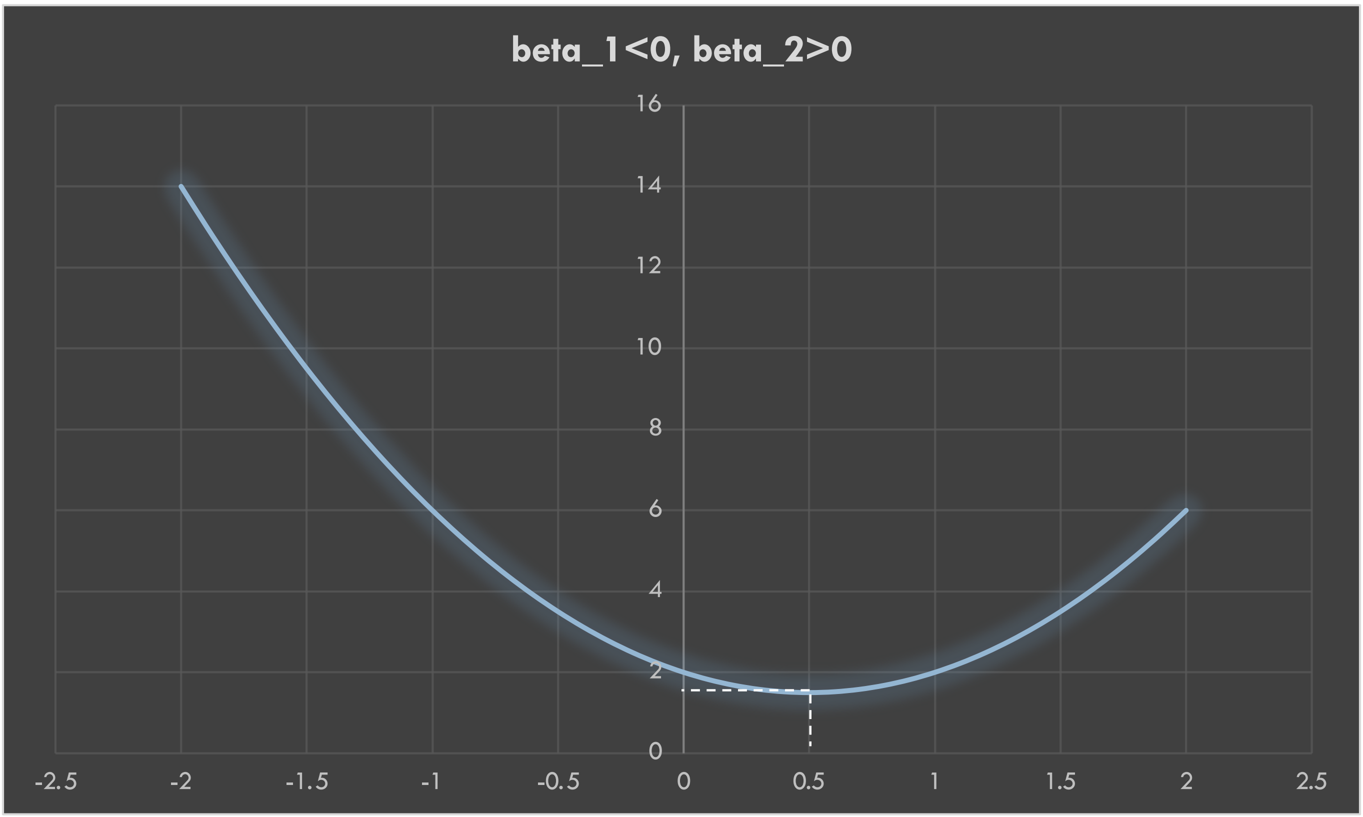
- **Notice that in quadratic models two coefficients are involved in describing the effect of X on Y: $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$**
  - So, the average effect of X on Y is the following

$$\hat{\beta}_1 + 2\hat{\beta}_2\bar{X} \tag{49}$$

    - You can replace $\bar{X}$ for some interesting level of X in the context of your study

Quadratic Models: Interpretation of Coefficients

- **An obviously interesting point in a quadratic model is the point where the slope of the function turns from positive to negative (or negative to positive)**



Quadratic Models: Interpretation of Coefficients

- This point is very easy to find once you have estimates. Just set equation (49) equal to zero and solve for X,

$$\hat{\beta}_1 + 2\hat{\beta}_2 X_{peak} = 0 \tag{50}$$

$$\Longrightarrow X_{peak} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \tag{51}$$

Quadratic Models: Example

- **Let's take a look at a paper from (my friend) Jeff Bloem of the USDA's Economic Research Service:**
  - "Aspirations and Investment in Rural Myanmar", published in the *Journal of Economic Inequality*
- **Jeff's research question is "What is the relationship between the "aspirations gap" and investment in one's home?"**
  - "Aspirations gap" means the difference between one's standard of living and the standard of living that they aspire towards

Quadratic Models: Example

- **Why would there be a relationship between these two variables?**
  - Well, investment in one's home is a way one can attempt to realize one's aspirations
  - If you're far away from what you aspire towards, likely you will behave differently in how you invest than if you are much closer
- **Jeff thinks the relationship might be a quadratic one. Why does this seem plausible?**
  - If you are farther away from your aspiration you might say "I'm too far away from my goal, so to heck with it, I'm not investing much"
  - If you are closer to your goal then you don't need to investment much to achieve what you aspire towards
  - In between the two you'll invest more

Quadratic Models: Example

- **Jeff estimates the following log-quadratic model,**

$$IHS(Inv_{ie}) = \beta_0 + \beta_1 AspGap_{ie} + \beta_2 AspGap_{ie}^2 + Z_i'\Gamma + \theta_e + \varepsilon_{ie} \qquad (52)$$

- **Here, there is some stuff you aren't yet familiar with:**
  - The "e" subscript indicates that these observations are coming from a variety of enumeration areas
  - Z_i' is compact notation saying he is including multiple Z variables. Γ contains all their coefficients
  - θ_e is again compact notation for "fixed effects" for the enumeration areas
    - Jeff includes this because he is worried that there are omitted variables at the level of enumeration area that he doesn't have measures for (like differences in local governance, resources, etc.) and this is a way to control for these variables
    - We'll get to what this is later in the semester. For now we'll just ignore it

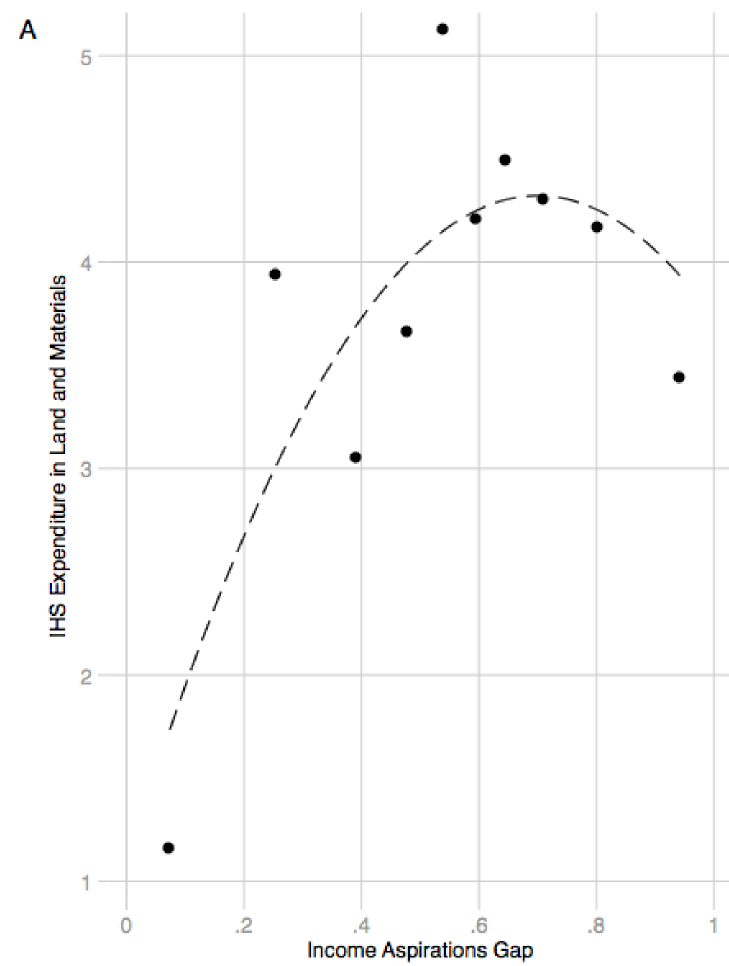
Quadratic Model: Example

	(1) IHS Investment
Income aspirations gap	13.657***
	(2.511)
Squared income aspirations gap	-11.087***
	(2.403)
Alt. income aspirations gap	
Squared alt. income aspirations gap	
Observations	445
R-squared	0.371
EA fixed effects?	Yes
Additional controls?	Yes
U-test results:	
Turning point	0.616
Fieller 95% C.I.	[0.497; 0.814]
Sasabuchi p-value	0.003
Slope at Min	13.657
Slope at Max	-8.516

- **Jeff's Results:**
  - $\hat{\beta}_1$=13.657
  - $\hat{\beta}_2$=-11.087
- **Turning Point:**
  - $X_{peak} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} = -\frac{13.657}{2\times(-11.087)} = 0.616$

Quadratic Model: Example

- Here is a neat illustration of the data displaying this relationship Jeff includes in the paper
- If you'd like to read the whole paper, you can read a "working paper" version here:
  - https://jeffbloem.files.wordpress.com/2020/11/aspirations-myanmar.pdf
- More on Jeff's work here:
  - https://jeffbloem.com/



How are we doing?

Different Functional Forms

Function	Form
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Linear-Linear
$ln(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i$	Log-Linear
$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Log-Log
$Y_i = \beta_0 + \beta_1 ln(X_i) + \beta_2 ln(Z_i) + u_i$	Linear-Log
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + u_i$	Quadratic
$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$	Interaction

Model with Interaction Effects

- These are models of the form,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + u_i \tag{53}$$

- The partial derivatives are,

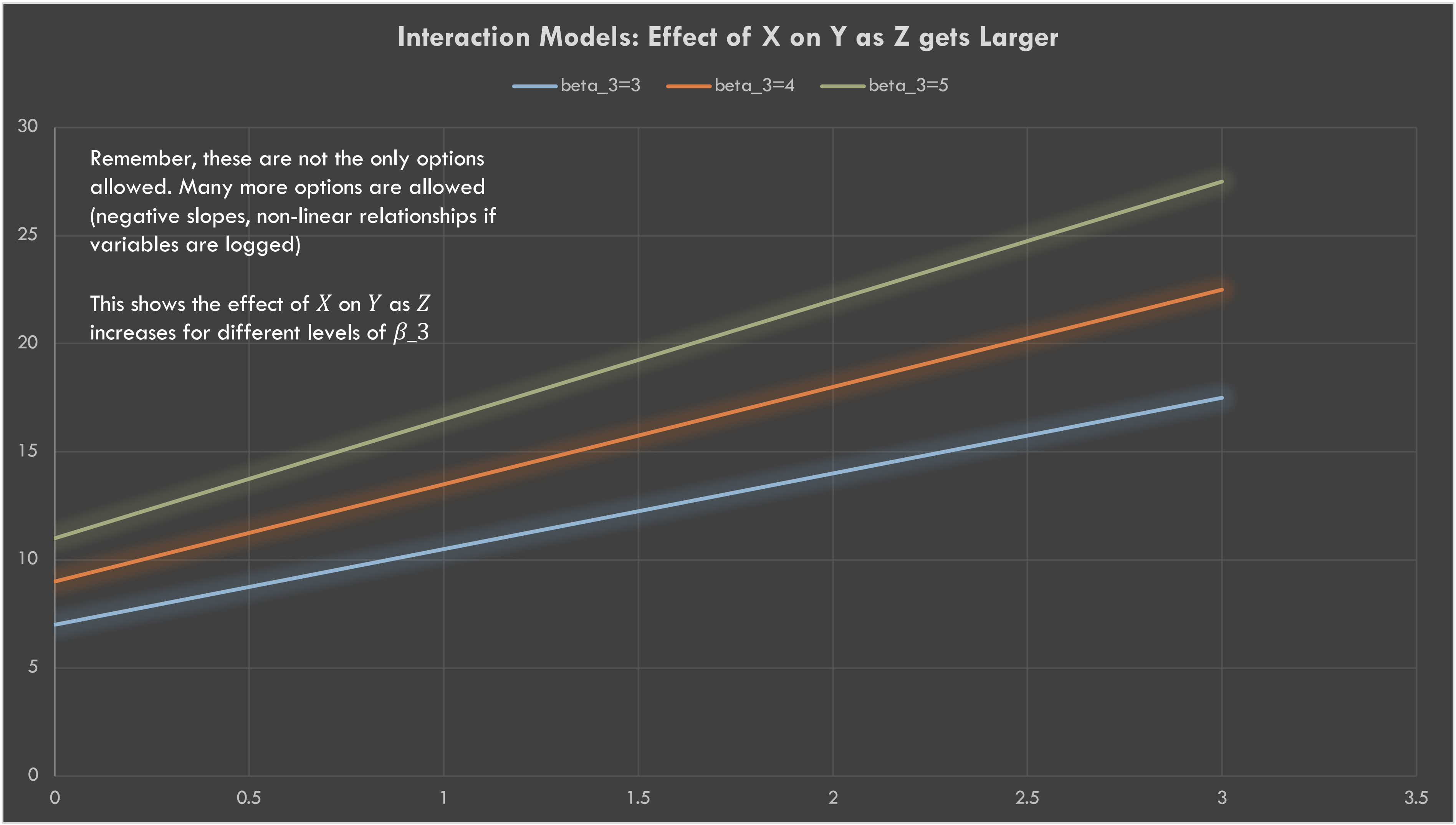
$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z_i \tag{54}$$

$$\frac{\partial Y}{\partial Z} = \beta_2 + \beta_3 X_i \tag{55}$$

Model with Interaction Effects

- **Note that this is a linear formulation so only linear relationships are allowed**
  - You can log variables and have interaction terms if you think non-linear relationships are likely
  - Just make sure in your code that you calculate $ln(X_i) \times ln(Z_i)$ and not $ln(X_i \times Z_i)$, because by log rules, $ln(X_i \times Z_i) = ln(X_i) + ln(Z_i)$
- **In these models the effect of X on Y depends on Z**

How the Effect of X on Y changes with Z



Interaction Effects Models: Example

- **Interaction effects are used widely**
  - They will show up again prominently when we discuss "Differences-in-Differences" later in the semester
  - They also are employed when looking at "heterogeneous effects" (how effects differ by group)
- **Generally, you want to use them when you think there is reason to believe that the two variables have an effect together on the change in Y (beyond the effects they offer independently). Example:**
  - The effect of education (X) varies by level of experience (Z)

How are we doing?

Binary Variables as Regressors

- So that does it with functional form choices!
- Now, we are going to spend more time looking at binary variables
- Specifically, we'll think hard about the following:
  - Binary variables as independent variables (regressors)
  - Interaction terms with binary variables
  - Groups of binary variables
  - Interpreting units of measurement

Binary Variable Definition

- **Binary variables are variables that take on just two values: 0 and 1**
  - Often called "indicator variables," "indicators," "dummy variables," or simply "dummies"
- **They are used to measure qualitative characteristics or to indicate membership to certain groups or time periods**
- **Examples:**
  - Married, union member, living in an urban area, female or male, abuse survivor, recipient of some intervention (like an NGO program)

Binary Variable Definition

- **In data with a time dimension (where data are collected over time) you can use them to measure membership to a certain time period**
  - This allows us to control for data that come from years with shocks like the 2008 financial crisis or the COVID-19 pandemic
- **We'll get to this in a bit, but when there are multiple binaries indicating membership to multiple mutually exclusive groups, the collection of binaries is called: "Fixed Effects"**
  - This is what Jeff was using

A Linear Model with a Binary Regressor

- **Let's look at a basic linear model with a continuous regressor X_i and a binary regressor D_i,**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \tag{62}$$

- **To interpret, write down what this says about Y first when $D_i = 1$ and then when $D_i = 0$,**

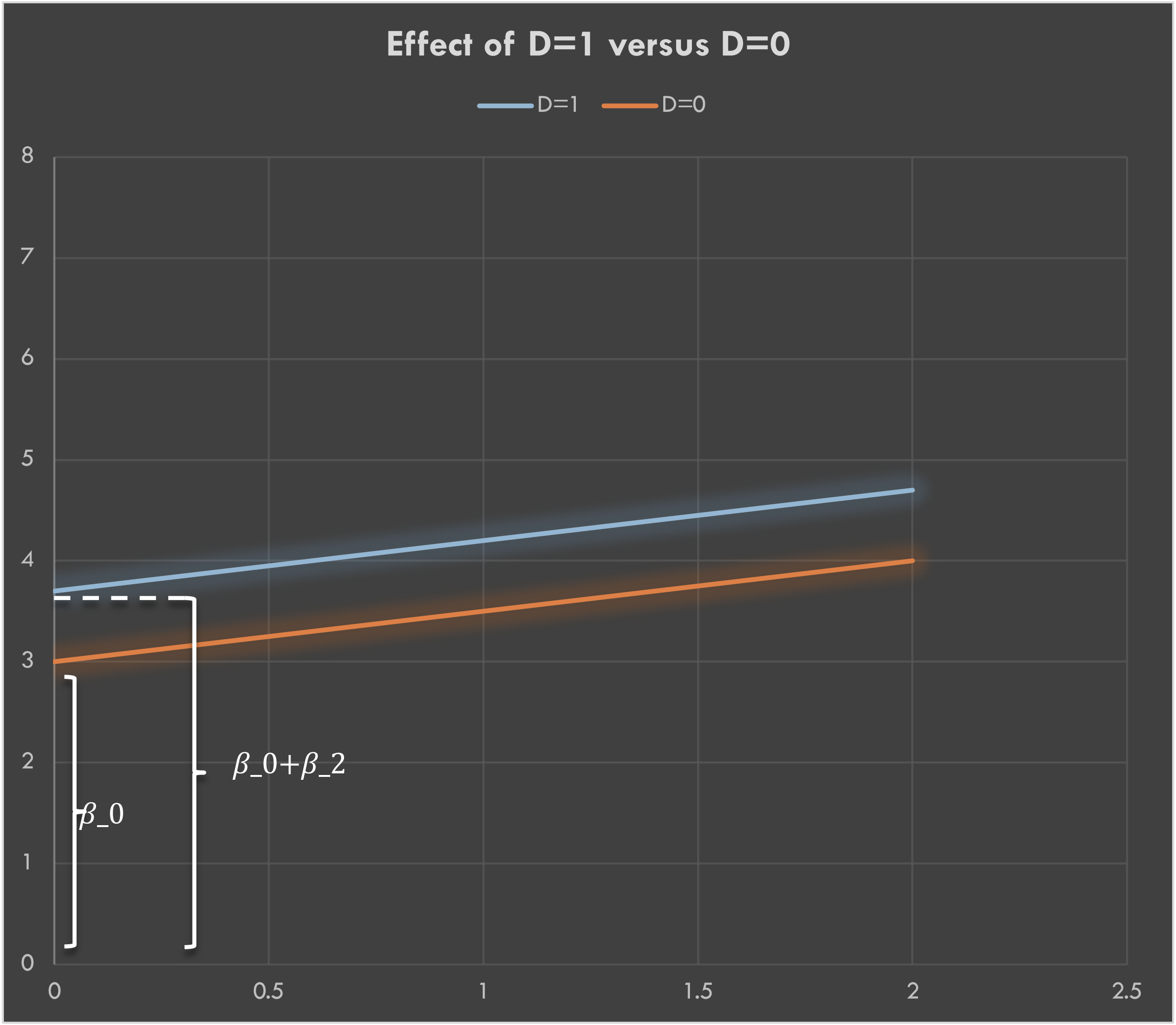
$$D_i = 1: Y_i = \beta_0 + \beta_1 X_i + \beta_2 + u_i \tag{63}$$

$$D_i = 0: Y_i = \beta_0 + \beta_1 X_i + u_i \tag{64}$$

RHS Binary Variable: Graphical Representation

- When $D_i = 1$ the graph is just a parallel shift from the people with $D_i = 0$

- So the effect of the membership to a certain group is an intercept shift relative to the other group measured



Linear Model with Binary Variable: Interpretation of β_2

- **Here is a useful derivation to remind yourself of the meaning:**

$$E(Y_i|X_i, D_i = 1) = \beta_0 + \beta_1 X_i + \beta_2 \quad (65)$$

$$E(Y_i|X_i, D_i = 0) = \beta_0 + \beta_1 X_i \quad (66)$$

$$\Longrightarrow \beta_2 = E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0) \quad (67)$$

- **In words: $\boldsymbol{\beta_2}$ is the difference in average $\boldsymbol{Y}$ between individuals with $\boldsymbol{D = 1}$ and $\boldsymbol{D = 0}$ after controlling for $\boldsymbol{X}$**

Linear Model with Binary Variable: Interpretation of β_2

- **We always interpret coefficients on binary variables as differences in average Y between the two groups after controlling for other covariates**
- **To interpret completely and precisely you need to specify which groups are being compared, in which order (who is $D = 0$ and who is $D = 1$) and which units**
  - The group against which the comparison is being made is $D = 0$. They are the "base group/category," "omitted group/category," or "comparison group/category"
    - Note that this is the same as a normal OLS model. Still don't know if it is a *causal* estimate without a lot more work and caveats

The "Binary Variable Trap"

- **You may be asking, "why do we write the model like this?...**

$$Wage_i = \beta_0 + \beta_1 Female_i + \beta_2 X_i + U_i \qquad (68)$$

and not like this?"

$$Wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Male_i + \beta_3 X_i + U_i \qquad (69)$$

The "Binary Variable Trap"

$$Wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Male_i + \beta_3 X_i + U_i \qquad (69)$$

- **Look at this formulation. Note that because individuals are either male or female in the data* if you know an individual's value for $Female_i$ you know their value for $Male_i$**
  - This means they are "collinear"
- **So if you have $Female_i$ in the regression, adding $Male_i$ offers OLS no more information**
- **Stata will simply drop second collinear variable**

* Of course sex is not a binary state as perhaps up to 1.7 percent of humans are born intersex. However, most data collection efforts ask people to choose male or female. One can dispute the ethics of this but it still remains that in the data the observations are reported as either male or female with no ambiguity. If you worry about mismeasurement as a result, OLS actually has very nice properties when certain types of measurement error are introduced. We'll get to this towards the end of the semester.

The "Binary Variable Trap"

- **Including the binary variable ($Female_i$ in this case) and its collinear converse ($Male_i$ in this case) is called the "binary variable trap" because it befuddles many novice applied econometricians**

- **Just avoid it by remembering to omit one of the categories when you code your regression**
  - This is the origin of the term "omitted group/category"

Binary Variable in the Interaction Term

- **You can estimate a model of the following form,**

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + u_i \tag{78}$$

- **To interpret, just do the same trick as before and look at Y for when $D = 1$ and $D = 0$,**

$$D = 1: Y_i^1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + u_i \tag{79}$$

$$D = 0: Y_i^0 = \beta_0 + \beta_1 X_i + u_i \tag{80}$$

Binary Variable in the Interaction Term

$$D = 1: Y_i^1 = \underbrace{(\beta_0 + \beta_2)}_{\text{Intercept Shift}} + \underbrace{(\beta_1 + \beta_3)}_{\text{Slope rotation}} X_i + u_i \qquad (79)$$

$$D = 0: Y_i^0 = \beta_0 + \beta_1 X_i + u_i \qquad (80)$$

- **So the effect of $D = 1$ is some base difference (β_2) plus a difference in how X affects Y (β_3)**

Binary Variable and Interaction Terms: Example

- **Binary variables in interaction terms are very common, especially in frontier work. So it is important to understand them well. Let's look at an example**
- **We may be interested in how returns to schooling are different for men and women**
  - Of course there is labor market discrimination against women, so it is possible that schooling simply doesn't earn as much of a return for women
  - Interaction terms with a binary variable allows us to estimate this difference

Binary Variable and Interaction Terms: Example

- **To do this, we'd estimate a model of the form,**

$$Wage_i = \beta_0 + \beta_1 Educ_i + \beta_2 Female_i + \beta_3 Educ_i \times Female_i + u_i \qquad (81)$$

- **How do we interpret the parameters here?**
  - What is β_0?
  - What is β_1?
  - What is β_2?
  - What is β_3?
- **This is the basis of "heterogeneity analysis"**
  - When one wants to examine the effect of something for various subgroups of people

Binary Variables in the Context of Multiple Related Groups

- **You may have a variable that measures membership among more than two groups. Examples:**
  - Ethnic or racial identity
  - Geographic location (states, provinces, towns, enumeration areas)
  - Occupation
  - Sector of the economy (manufacturing, healthcare, education, etc.)
- **Binary variables can be used in exactly the same way as when you have just two categories**

Binary Variables in the Context of Multiple Related Groups

- Let's say you have data of people from 5 different states (PA, NJ, NY, DE, and MD) and we wanted to estimate the previous wage regression
- We could add in binary variables for each state leaving one omitted as follows,

$$\begin{aligned} Wage_{is} &= \beta_0 + \beta_1 Educ_{is} + \beta_2 Female_{is} + \beta_4 PA_s + \beta_5 NJ_s + \beta_6 NY_s \\ &+ \beta_7 DE_s + u_i \end{aligned} \tag{82}$$

Binary Variables in the Context of Multiple Related Groups

$$Wage_{is} = \beta_0 + \beta_1 Educ_{is} + \beta_2 Female_{is} + \beta_4 PA_s + \beta_5 NJ_s + \beta_6 NY_s + \beta_7 DE_s + u_i \quad (82)$$

- **Often $\boldsymbol{\beta_4 PA_s + \beta_5 NJ_s + \beta_6 NY_s + \beta_7 DE_s}$ is just abbreviated to some Greek letter with a subscript like "$\boldsymbol{\theta_s}$"**
- **Now, the coefficients on these group variables are the difference from living in that state <u>relative to the omitted category</u> (in this case MD) controlling for one's level of education and one's sex**
- **They are often referred to as "fixed effects" in papers**
  - They are what we skipped over when we looked at Jeff's paper

Binary Variables in the Context of Multiple Related Groups

- **The coefficients on these "fixed effects" are sometimes interesting by themselves, but more often they aren't**
- **More frequently they are included in models because they control for omitted relevant variables that otherwise would be ignored and be in the error term**
  - Recall that not controlling for relevant variables violates Assumption 2 and biases our estimate

Binary Variables in the Context of Multiple Related Groups

- **To see how they do this, let's think about equation (82) again**
  - Why would wages differ by state?
    - Local economic opportunities (NY has a larger more prosperous economy than DE)
    - Local cost of living
    - Local governance
    - Local education systems
    - Various historical legacies
    - Probably many more you can think of…

Binary Variables in the Context of Multiple Related Groups

$$Wage_{is} = \beta_0 + \beta_1 Educ_{is} + \beta_2 Female_{is} + \beta_4 PA_s + \beta_5 NJ_s + \beta_6 NY_s + \beta_7 DE_s + u_i \quad (82)$$

- **You can think of the coefficients on the fixed effects as the "effect" of living in PA, NJ, NY, and DE relative to MD on wages (controlling for education and sex)**
- **What is the "effect" of living in a state? Well, it's the combined effect of everything I just mentioned (local economic prosperity, local governance, local educational quality)**
- **Including the FEs will allow OLS to partial out β_4, β_5, β_6, β_7**
- **<u>An easy shorthand for remembering what fixed effects control for is they control for "time-invariant omitted covariates at the level of the fixed effect"</u>**