Categorical Variables

EC 320: Introduction to Econometrics

Philip Economides

Winter 2022

Prologue

Categorical Variables

Categorical Variables

Goal: Make quantitative statements about qualitative information.

- *e.g.*, race, gender, being employed, living in Oregon, *etc.*

Approach: Construct binary variables.

- *a.k.a.* dummy variables or indicator variables.
- Value equals 1 if observation is in the category or 0 if otherwise.

Regression implications

1. Binary variables change the interpretation of the intercept.
2. Coefficients on binary variables have different interpretations than those on continuous variables.

Continuous Variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + u_i$$

where

- Pay_i is a continuous variable measuring an individual's pay
- School_i is a continuous variable that measures years of education

Interpretation

- β_0: y-intercept, *i.e.*, Pay when $\text{School} = 0$
- β_1: expected increase in Pay for a one-unit increase in School

Continuous Variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + u_i$$

Derive the slope's interpretation:

$$\begin{aligned}
&\mathbb{E}[\text{Pay}|\text{School} = \ell + 1] - \mathbb{E}[\text{Pay}|\text{School} = \ell] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1(\ell + 1) + u] - \mathbb{E}[\beta_0 + \beta_1 \ell + u] \\
&\quad = [\beta_0 + \beta_1(\ell + 1)] - [\beta_0 + \beta_1 \ell] \\
&\quad = \beta_0 - \beta_0 + \beta_1 \ell - \beta_1 \ell + \beta_1 \\
&\quad = \beta_1.
\end{aligned}$$

The slope gives the expected increase in pay for an additional year of schooling.

Continuous Variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + u_i$$

Alternative derivation

Differentiate the model with respect to schooling:

$$\frac{d\text{Pay}}{d\text{School}} = \beta_1$$

The slope gives the expected increase in pay for an additional year of schooling.

Continuous Variables

If we have multiple explanatory variables, *e.g.*,

$$\mathbf{Pay}_i = \beta_0 + \beta_1 \mathbf{School}_i + \beta_2 \mathbf{Ability}_i + u_i$$

then the interpretation changes slightly.

$$\begin{aligned}
&\mathbb{E}[\mathbf{Pay}|\mathbf{School} = \ell + 1 \wedge \mathbf{Ability} = \alpha] - \mathbb{E}[\mathbf{Pay}|\mathbf{School} = \ell \wedge \mathbf{Ability} = \alpha] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha + u] - \mathbb{E}[\beta_0 + \beta_1\ell + \beta_2\alpha + u] \\
&\quad = [\beta_0 + \beta_1(\ell + 1) + \beta_2\alpha] - [\beta_0 + \beta_1\ell + \beta_2\alpha] \\
&\quad = \beta_0 - \beta_0 + \beta_1\ell - \beta_1\ell + \beta_1 + \beta_2\alpha - \beta_2\alpha \\
&\quad = \beta_1
\end{aligned}$$

The slope gives the expected increase in pay for an additional year of schooling, **holding ability constant**.

Continuous Variables

If we have multiple explanatory variables, *e.g.*,

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Ability}_i + u_i$$

then the interpretation changes slightly.

Alternative derivation

Differentiate the model with respect to schooling:

$$\frac{\partial \text{Pay}}{\partial \text{School}} = \beta_1$$

The slope gives the expected increase in pay for an additional year of schooling, **holding ability constant**.

Categorical Variables

Consider the relationship

$$\mathbf{Pay}_i = \beta_0 + \beta_1 \mathbf{Female}_i + u_i$$

where $\mathbf{Pay}_i$ is a continuous variable measuring an individual's pay and $\mathbf{Female}_i$ is a binary variable equal to 1 when i is female.

Interpretation

β_0 is the expected $\mathbf{Pay}$ for males (*i.e.*, when $\mathbf{Female} = 0$):

$$
\begin{aligned}
&\mathbb{E}[\mathbf{Pay} | \mathbf{Male}] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&\quad = \mathbb{E}[\beta_0 + 0 + u_i] \\
&\quad = \beta_0
\end{aligned}
$$

Categorical Variables

Consider the relationship

$$\mathbf{Pay}_i = \beta_0 + \beta_1 \mathbf{Female}_i + u_i$$

where $\mathbf{Pay}_i$ is a continuous variable measuring an individual's pay and $\mathbf{Female}_i$ is a binary variable equal to 1 when i is female.

Interpretation

β_1 is the expected difference in $\mathbf{Pay}$ between females and males:

$$\begin{aligned}
&\mathbb{E}[\mathbf{Pay}|\mathbf{Female}] - \mathbb{E}[\mathbf{Pay}|\mathbf{Male}] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1 \times 1 + u_i] - \mathbb{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1 + u_i] - \mathbb{E}[\beta_0 + 0 + u_i] \\
&\quad = \beta_0 + \beta_1 - \beta_0 \\
&\quad = \beta_1
\end{aligned}$$

Categorical Variables

Consider the relationship

$$\mathbf{Pay}_i = \beta_0 + \beta_1 \mathbf{Female}_i + u_i$$

where $\mathbf{Pay}_i$ is a continuous variable measuring an individual's pay and $\mathbf{Female}_i$ is a binary variable equal to 1 when i is female.

Interpretation

$\beta_0 + \beta_1$: is the expected $\mathbf{Pay}$ for females:

$$\begin{aligned}
&\mathbb{E}[\mathbf{Pay}|\mathbf{Female}] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&\quad = \mathbb{E}[\beta_0 + \beta_1 + u_i] \\
&\quad = \beta_0 + \beta_1
\end{aligned}$$

Categorical Variables

Consider the relationship

$$\text{Pay}_i = \beta_0 + \beta_1 \text{Female}_i + u_i$$

Interpretation

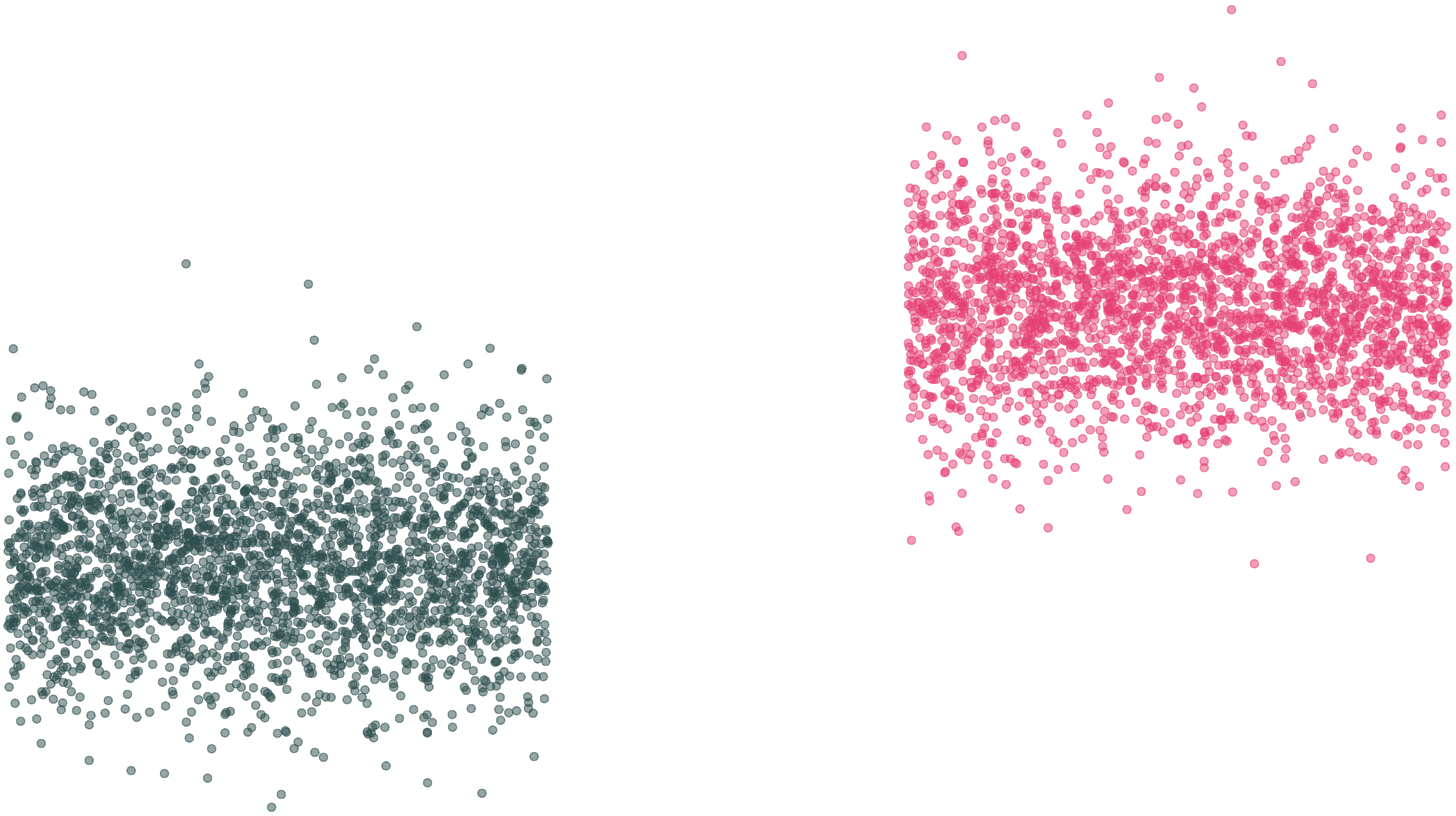
- β_0: expected Pay for males (*i.e.*, when $\text{Female} = 0$)
- β_1: expected difference in Pay between females and males
- $\beta_0 + \beta_1$: expected Pay for females
- Males are the **reference group**

Note: If there are no other variables to condition on, then $\hat{\beta}_1$ equals the difference in group means, *e.g.*, $\bar{X}_{\text{Female}} - \bar{X}_{\text{Male}}$.

Note$_2$: The *holding all other variables constant* interpretation also applies for categorical variables in multiple regression settings.

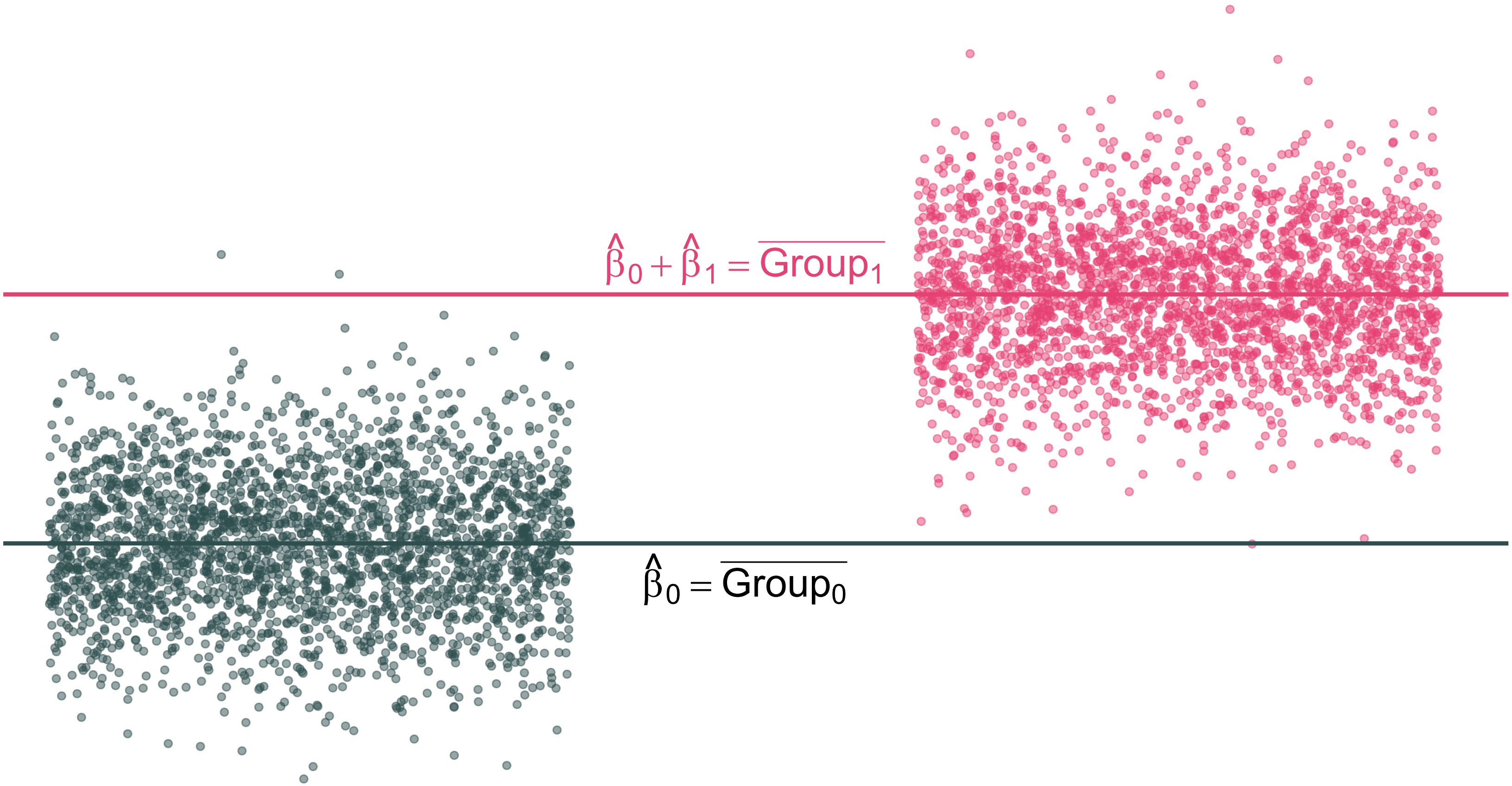
Categorical Variables

$Y_i = \beta_0 + \beta_1 X_i + u_i$ for binary variable $X_i = \{0, 1\}$



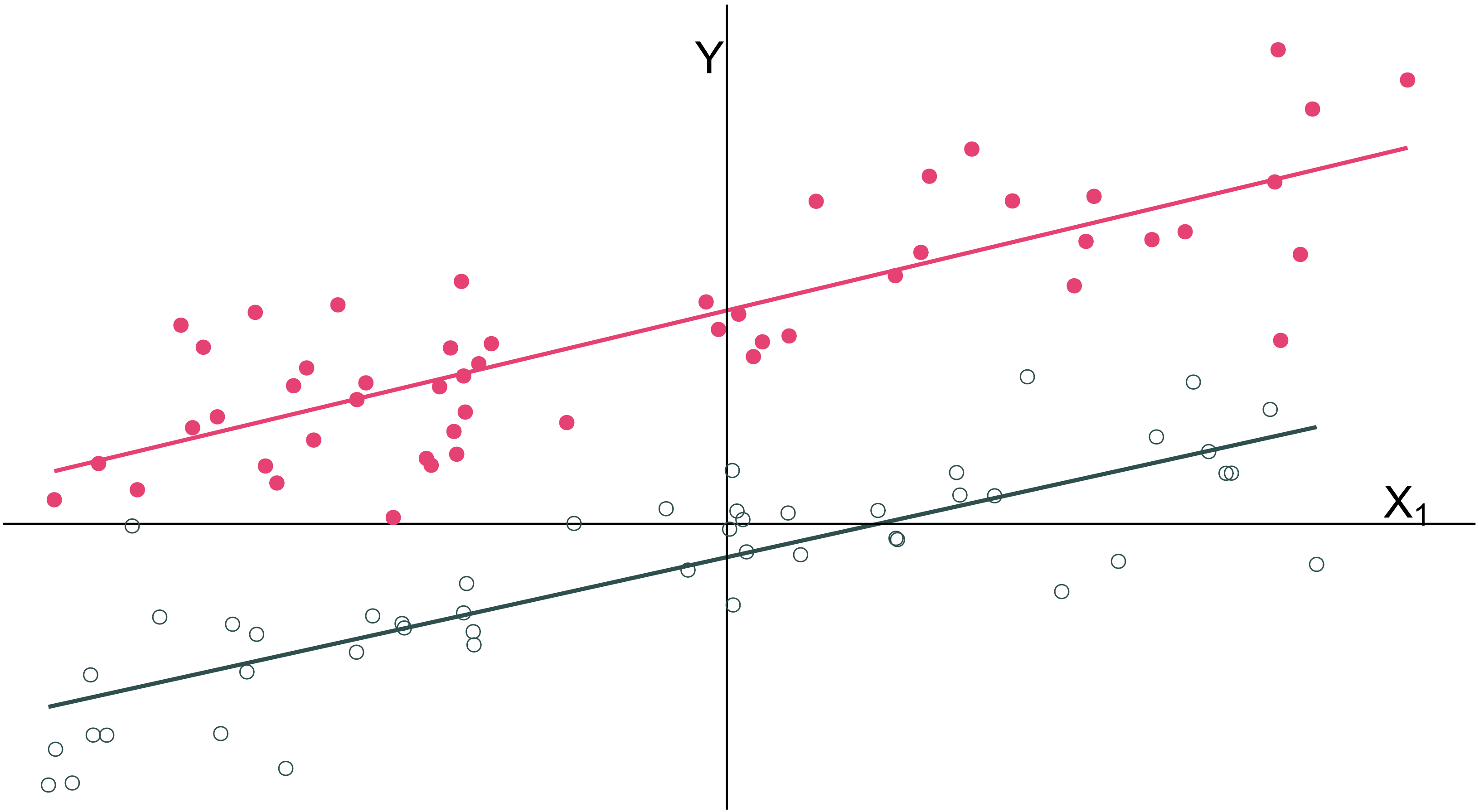
Categorical Variables

$Y_i = \beta_0 + \beta_1 X_i + u_i$ for binary variable $X_i = \{0, 1\}$



Multiple Regression

Another way to think about it:



Question: Why not estimate $\text{Pay}_i = \beta_0 + \beta_1 \text{Female}_i + \beta_2 \text{Male}_i + u_i$?

Answer: The intercept is a perfect linear combination of Male_i and Female_i.

- Violates no perfect collinearity assumption.
- OLS can't estimate all three parameters simultaneously.
- Known as **dummy variable trap**.

Practical solution: Select a reference category and drop its indicator.

Dummy Variable *Trap?*

Don't worry, R will bail you out if you include perfectly collinear indicators.

Example

```
lm(wage ~ black + nonblack, data = wage_data) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)     617.      5.27     117.   0
#> 2 black          -168.     10.9      -15.4  7.78e-52
#> 3 nonblack          NA     NA         NA   NA
```

Thanks, R.

Multiple Categories

So far we have only discussed **binary** categorical variables represented by dummies.

In many cases, there is a wide variety of categories by which we can characterize a set of observations.

For example

- Transport Modes: Rail, Highway, Air, Water
- Income Range: 1st quartile, 2nd quartile, 3rd quartile, 4th quartile
- Geographic Regions: Alabama, Idaho, Oregon etc.

When addressing product diversification and trade, we can end up with an incredible number of categories to consider. Trade Statistics by Product (HS 6-digit)

Categorical Variable Types

Type of Variable	Represents	Examples
Binary Variables	*Yes/no outcomes*	Heads/tails in a coin flip Win/lose in a football game
Nominal Variables	*Groups with no rank or order between them*	Specific names Colors Brands
Ordinal Variables	*Groups that are ranked in a specific order*	Rankings in a competition Rating scale responses in survey

Beyond Binary

How do we deal with heaps of categories? **It depends**.

Are these categories your **outcome variable**?

Binary: *Logistic Regression Model*, where we are determining the probability of an event, given individual characterisitics of i.

Ordinal: *Cumulative/Ordered Logit Model* for categorical variables with an implied order and J choices.

Nominal: *Generalized Logit Model* which holds characteristics fixed across choices and *Multinomial/Conditional Logit Model* which allows characteristics to differ for different choices.

These items **will not be covered** in this class, nor will their descriptions be tested upon. This is guidance for those interested in reading further and understanding what future econometrics classes deliver.

Beyond Binary

Are these categories part of an **explanatory variable**?

Approach I: Apply a unique dummy variable for each category

For example consider $\textbf{earn}_i = \alpha + \beta_1 \textbf{HS}_i + \beta_2 \textbf{UG}_i + \beta_3 \textbf{MS}_i + \beta_3 \textbf{PhD}_i + u_i$

In this case I may have a single categorical variable, DEG_i, that lists degree types of individual i across my sample.

```
educ_df %<>% mutate(HS = if_else(DEG=="Highschool", 1, 0),
                    UG = if_else(DEG=="Undergraduate", 1, 0),
                    MS = if_else(DEG=="Masters of Science", 1, 0),
                    PhD = if_else(DEG=="Doctorate", 1 ,0)
                    )
educ_reg <- lm(data=educ_df, earn ~ HS + UG + MS + PhD)
```

Assuming i w/o any degree, would form my reference group in which for every included individual, $HS + UG + MS + PhD = 0$.

Beyond Binary

What if there are **too many** categories but I want to create individual dummies?

Jacob Kaplan (Princeton) created the `fastDummies` package, which provides a useful function `dummy_cols()` LINK

Consider the following example

numbers	gender	animals	dates
1	male	dog	2012-01-01
2	male	dog	2011-12-31
3	female	cat	2012-01-01

Beyond Binary

```
results <- fastDummies::dummy_cols(fastDummies_example)
knitr::kable(results) %>%
  kable_styling(font_size=10)
```

numbers	gender	animals	dates	gender_female	gender_male	animals_cat	animals_dog
1	male	dog	2012-01-01	0	1	0	1
2	male	dog	2011-12-31	0	1	0	1
3	female	cat	2012-01-01	1	0	1	0

```
results <- fastDummies::dummy_cols(fastDummies_example,
                                   select_columns= c("animals","gender"))
knitr::kable(results) %>%
  kable_styling(font_size=10)
```

numbers	gender	animals	dates	animals_cat	animals_dog	gender_female	gender_male
1	male	dog	2012-01-01	0	1	0	1
2	male	dog	2011-12-31	0	1	0	1
3	female	cat	2012-01-01	1	0	1	0

Beyond Binary

Are these categories part of an **explanatory variable**?

Approach II: Apply a fixed effect to your model

Consider the following model

$$\mathbf{earn}_{ij} = \alpha + \beta_1 \mathbf{Age}_i + \beta_2 \mathbf{AgeSq}_i + \beta_3 \mathbf{Educ}_i + \beta_3 \mathbf{Female}_i + u_{ij}$$

There may be **unobservable** aspects related to groups defined by j, that are **fixed** across individuals in each group $j \in \{1, 2, \ldots, J\}$.

For example, if we were regressing the earnings of service staff across J countries, the USA may see unobserved $\mathbf{tips}_{ij}$ contributing more significantly towards income due to underlying cultural/professional norms.

In this case a **country fixed effect** in our regression would do wonders.

Beyond Binary

Where $u_{ij} = \phi_j + \nu_{ij}$, our new regression would look like

$$\textbf{earn}_{ij} = \alpha + \beta_1 \textbf{Age}_i + \beta_2 \textbf{AgeSq}_i + \beta_3 \textbf{Educ}_i + \beta_3 \textbf{Female}_i + \phi_j + \nu_{ij},$$

Any **unobserved** contribution towards earnings that **varies across** J **but** is **constant within** each j for those individuals is controlled for.

How do we run regressions with fixed-effects?

> fixed.dum = lm(data=dataset, Y ~ X + factor(category_variable))

Turning your character variables into factors will automatically have the code treat each jth category as if it had its own dummy variable Example

`plm` maintained by Yves Croissant Example Code

`fixest` maintained by Laurent Berge and Grant McDermott Example Code

Beyond Binary

I estimate whether productivity rankings across different categories of firm-types, represented by **dummy variables**, are consistent with **Melitz(2003)**.

Unobs **fixed-effects** within industries, years and countries controlled for.

	(1)	(2)	(3)	(4)	(5)	(6)
HomeEXP	0.058*** (0.010)	0.052*** (0.011)	0.062*** (0.008)	0.057*** (0.011)	0.052*** (0.011)	0.062*** (0.009)
MNE	0.104*** (0.014)	0.105*** (0.013)	0.093*** (0.013)			
MNEDOM				0.107*** (0.015)	0.112*** (0.014)	0.080*** (0.015)
MNEEXP				0.101*** (0.017)	0.099*** (0.016)	0.103*** (0.014)
lnage	0.015** (0.006)	0.020*** (0.005)	0.011** (0.004)	0.015** (0.006)	0.020*** (0.005)	0.011** (0.004)
qcert	0.090*** (0.012)	0.079*** (0.009)	0.067*** (0.007)	0.090*** (0.012)	0.079*** (0.009)	0.066*** (0.007)
license	0.034*** (0.009)	0.030*** (0.009)	0.024*** (0.007)	0.034*** (0.009)	0.030*** (0.009)	0.024*** (0.007)
import	−0.014* (0.008)	−0.013* (0.006)	0.007 (0.008)	−0.014* (0.008)	−0.012* (0.006)	0.006 (0.008)
multi	−0.011** (0.005)	−0.009** (0.004)	−0.009** (0.004)	−0.011** (0.005)	−0.009** (0.004)	−0.009** (0.004)
Fixed Effects						
Industry	✓	✓	✓	✓	✓	✓
Year		✓	✓		✓	✓
Country			✓			✓
N	40,012	40,012	40,012	40,012	40,012	40,012
R^2	0.472	0.480	0.534	0.472	0.480	0.534

Notes: *** at 1 percent level, ** at 5 percent, * at 10 percent.

Omitted Variable Bias

Omitted variable bias (OVB) arises when we omit a variable that

1. Affects the outcome variable Y
2. Correlates with an explanatory variable X_j

Biases OLS estimator of β_j.

Omitted Variable Bias

Example

Let's imagine a simple population model for the amount individual i gets paid

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Male}_i + u_i$$

where School_i gives i's years of schooling and Male_i denotes an indicator variable for whether individual i is male.

Interpretation

- β_1: returns to an additional year of schooling (*ceteris paribus*)
- β_2: premium for being male (*ceteris paribus*)
  If $\beta_2 > 0$, then there is discrimination against women.

Omitted Variable Bias

Example, continued

From the population model

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \beta_2 \text{Male}_i + u_i$$

An analyst focuses on the relationship between pay and schooling, *i.e.*,

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + (\beta_2 \text{Male}_i + u_i)$$

$$\text{Pay}_i = \beta_0 + \beta_1 \text{School}_i + \varepsilon_i$$

where $\varepsilon_i = \beta_2 \text{Male}_i + u_i$.

We assumed exogeneity to show that OLS is unbiasedness. But even if $\mathbb{E}[u|X] = 0$, it is not necessarily true that $\mathbb{E}[\varepsilon|X] = 0$ (false if $\beta_2 \neq 0$).

Specifically, $\mathbb{E}[\varepsilon|\text{Male} = 1] = \beta_2 + \mathbb{E}[u|\text{Male} = 1] \neq 0$. **Now OLS is biased.**

Omitted Variable Bias

Let's try to see this result graphically.

The true population model:

$$\text{Pay}_i = 20 + 0.5 \times \text{School}_i + 10 \times \text{Male}_i + u_i$$

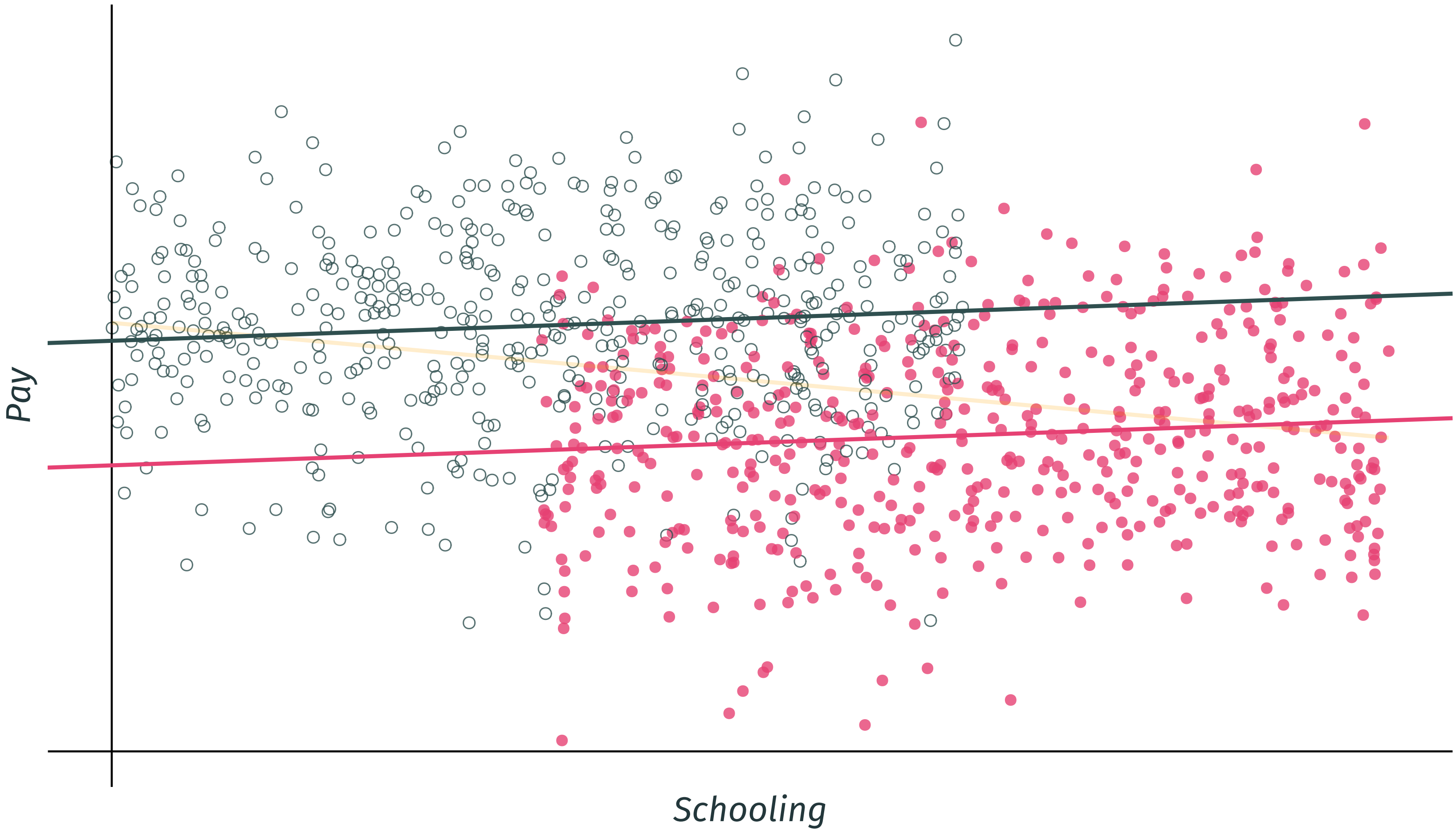
The regression model that suffers from omitted-variable bias:

$$\text{Pay}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{School}_i + e_i$$

Finally, imagine that women, on average, receive more schooling than men.

Omitted Variable Bias

Unbiased regression: $\widehat{\text{Pay}}_i = 20.9 + 0.4 \times \text{School}_i + 9.1 \times \text{Male}_i$



Categorical Variables

Example: Weekly Wages

```
lm(wage ~ south, data = wage_data) %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)     632.      6.00     105.  0
#> 2 south          -137.      9.45     -14.5 6.21e-46
```

Q_1: What is the reference category?

Q_2: Interpret the coefficients.

Q_3: Suppose you ran `lm(wage ~ nonsouth, data = wage_data)` instead. What is the coefficient estimate on `nonsouth`? What is the intercept estimate?

Categorical Variables

Example: Weekly Wages

```
lm(wage ~ south + black, data = wage_data) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    647.       6.02     107.  0
#> 2 south          -98.6      9.84     -10.0 2.89e-23
#> 3 black         -129.      11.4      -11.3 3.43e-29
```

Q_1: What is the reference category?

Q_2: Interpret the coefficients.

Q_3: Suppose you ran `lm(wage ~ south + nonblack, data = wage_data)` instead. What is the coefficient estimate on `nonblack`? What is the coefficient estimate on `south`? What is the intercept estimate?

Categorical Variables

Example: Weekly Wages

Answer to Q_3:

```
lm(wage ~ south + nonblack, data = wage_data) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    518.      11.7       44.3 0
#> 2 south          -98.6      9.84     -10.0 2.89e-23
#> 3 nonblack       129.      11.4       11.3 3.43e-29
```