Categorical Variables

EC 320: Introduction to Econometrics

Winter 2022

Prologue

Housekeeping

1. **Problem Set 3** grade posted
2. **Midterm** grade to be posted by Wednesday
3. **Problem Set 4** to be posted by tomorrow, stay tuned
   - Due next Monday
4. **Lab**
   - **Lab** held on Wednesday
   - **Lab material** available on Github, **Ex7** available on Canvas
   - **Ex7** due Wednesday

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.*,

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

then the interpretation changes slightly.

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

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

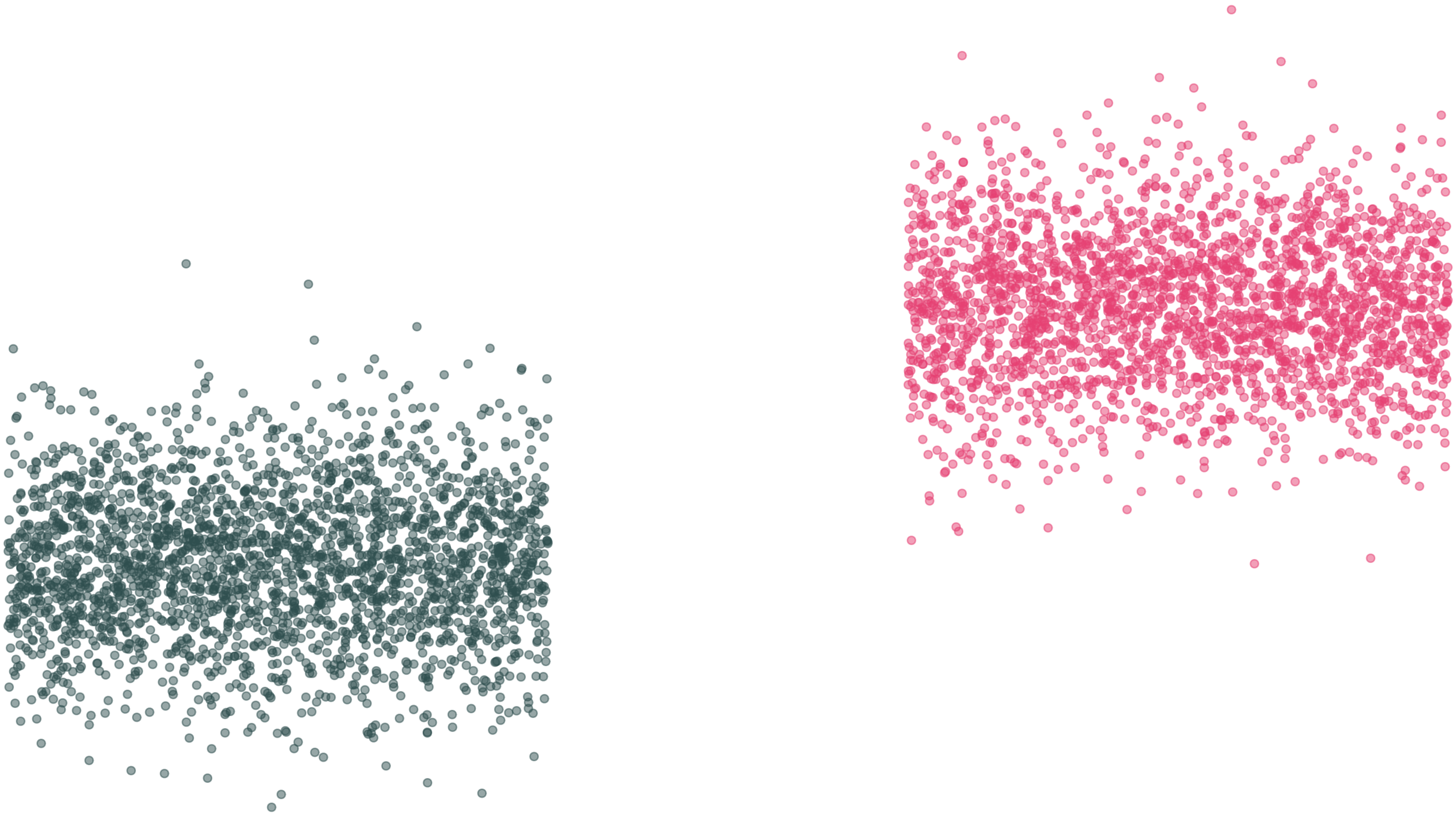
- β_0: expected Pay for males (*i.e.*, when $\text{Female} = 0$)
- β_1: expected difference in Pay between females and males
- $\beta_0 + \beta_1$: expected Pay for females
- Males are the **reference group**

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.

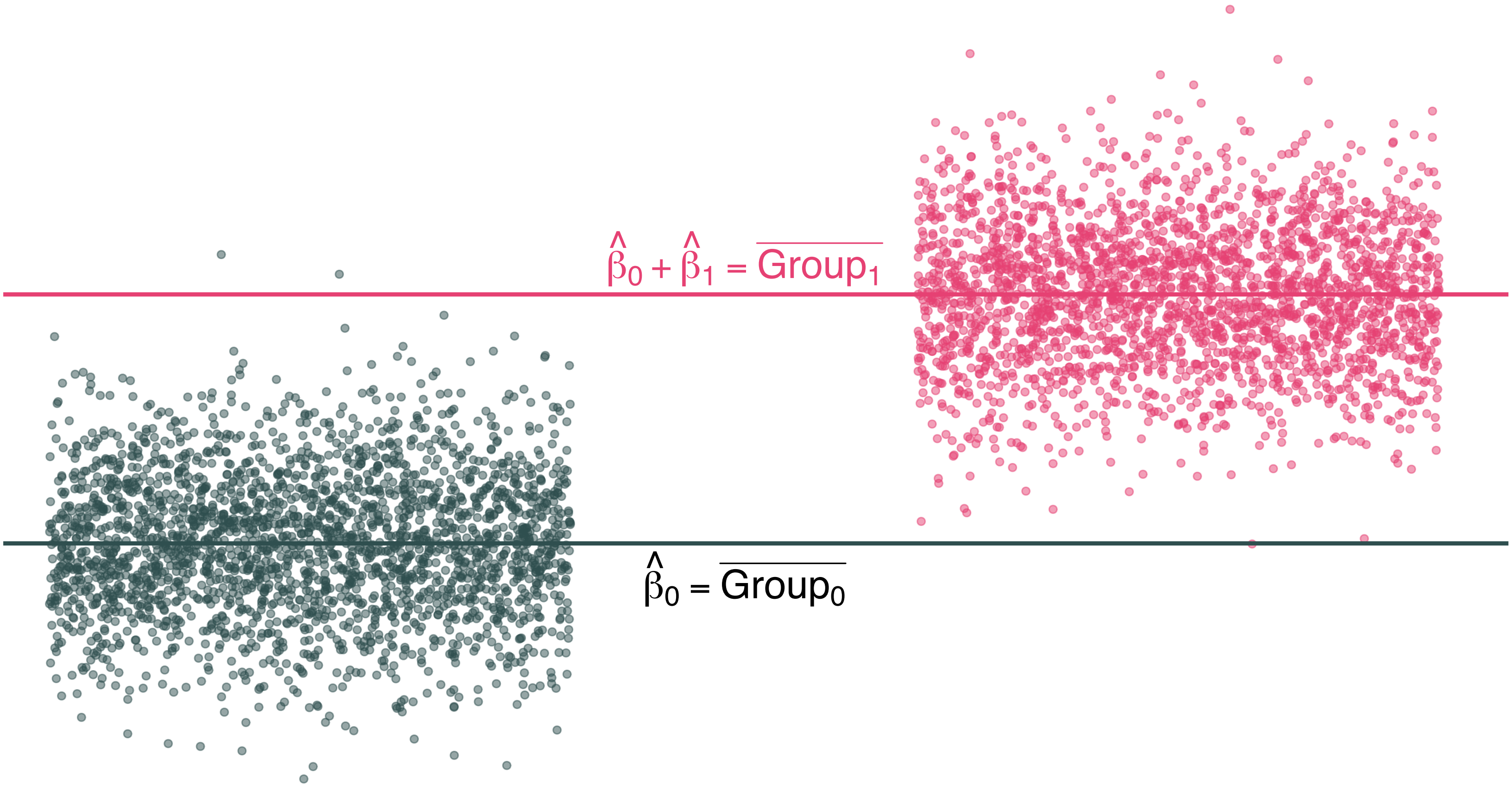
Categorical Variables

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



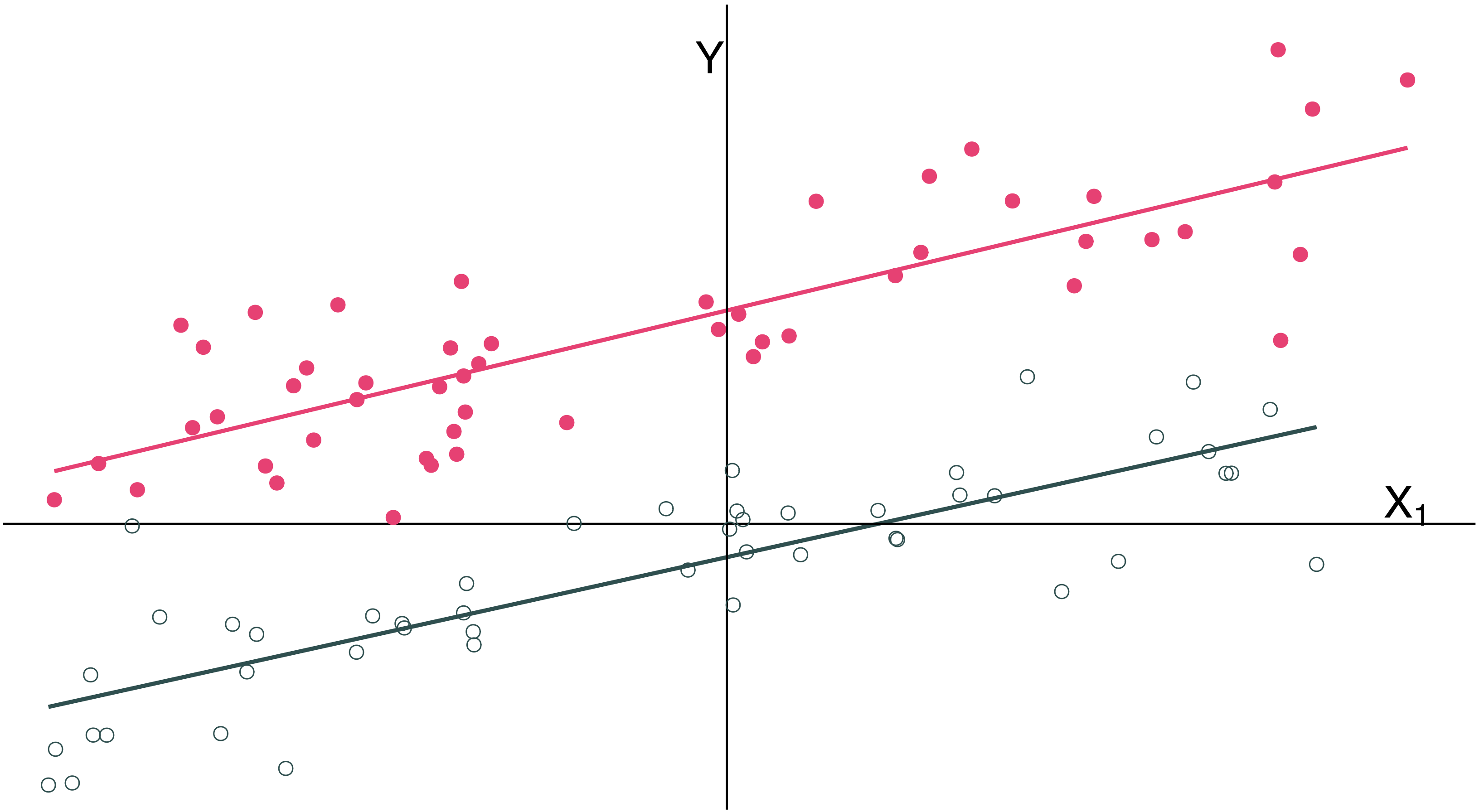
Categorical Variables

$Y_i = \beta_0 + \beta_1 X_i + u_i$ 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 × 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.

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:

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

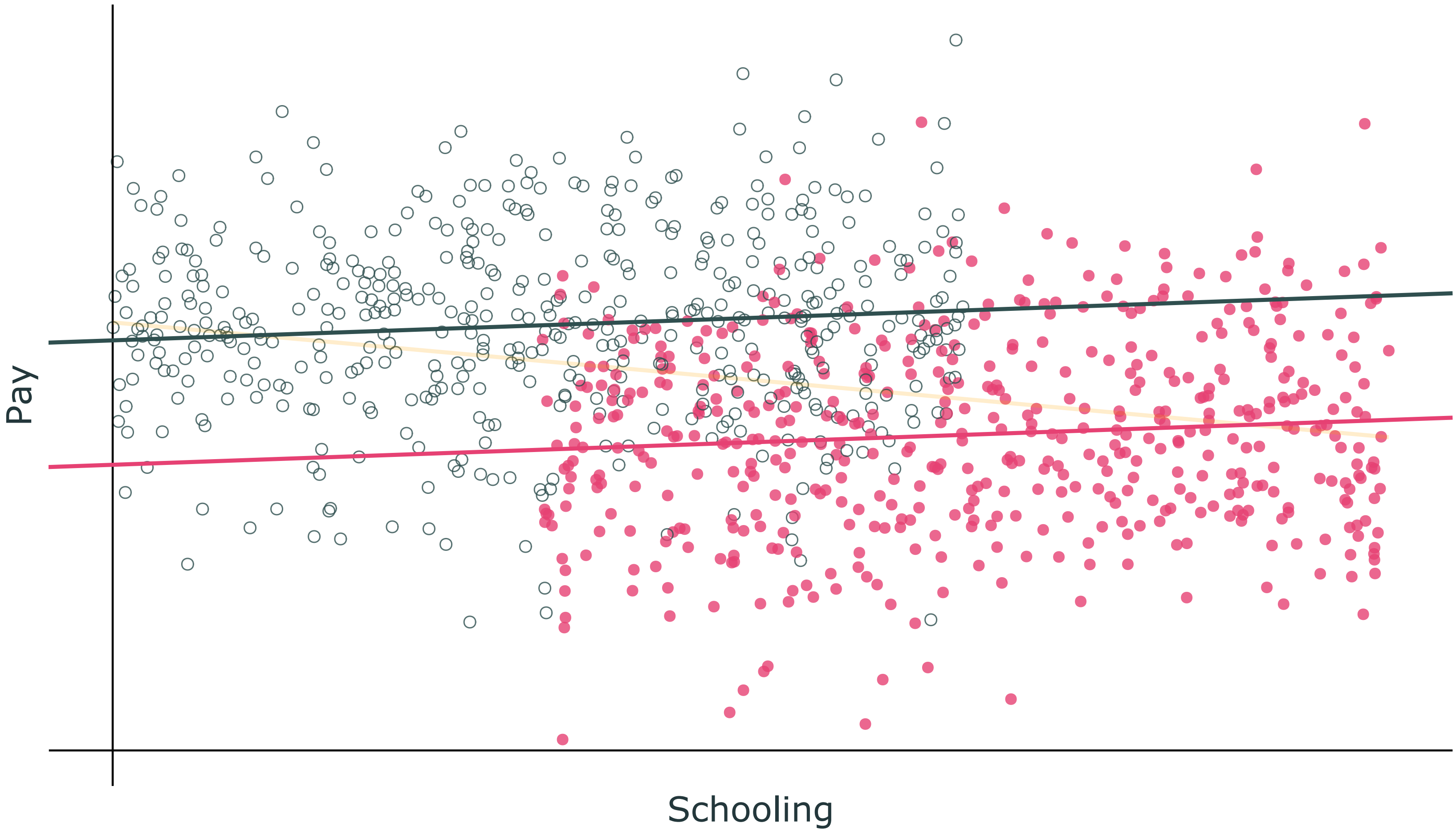
The regression model that suffers from omitted-variable bias:

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

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 × 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 × 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 × 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
```