Multivariate regressions

Lecture 9

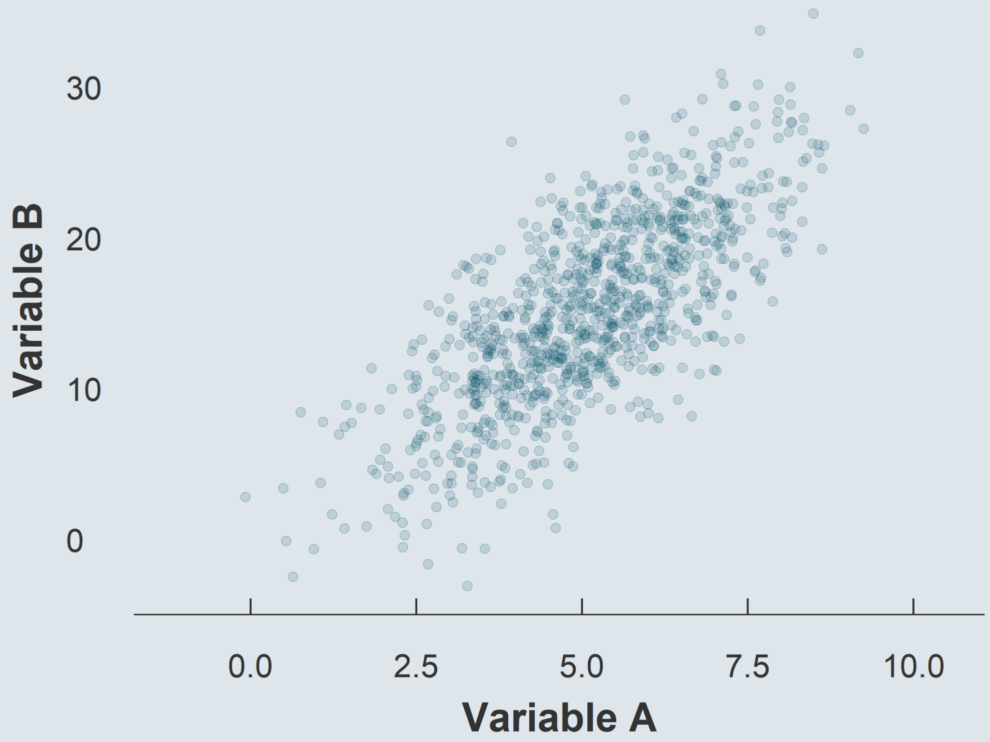
Louis SIRUGUE

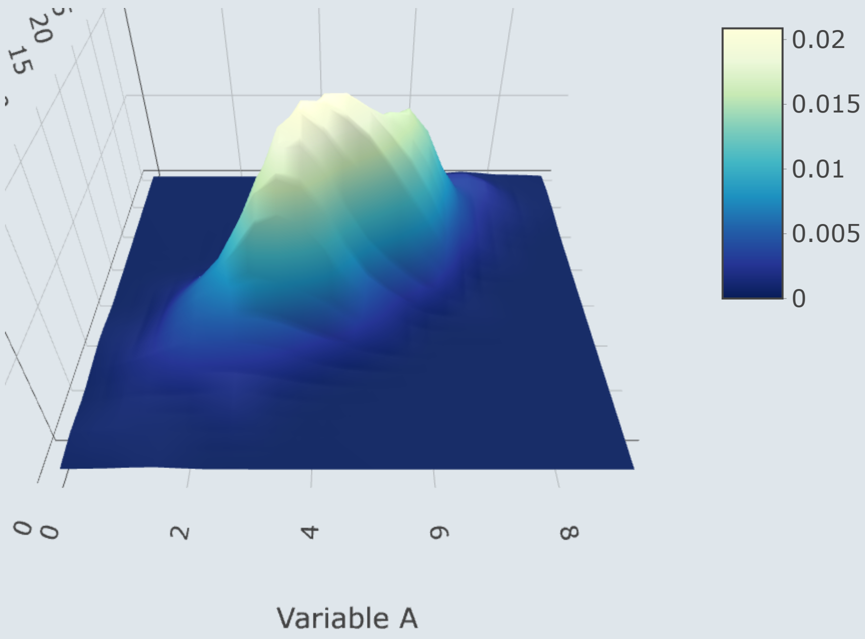
CPES 2 - Fall 2022

Last time we saw

1. Joint distribution

The **joint distribution** shows the possible **values** and associated **frequencies** for **two variables** simultaneously





Last time we saw

1. Joint distribution

➜ ***When describing a joint distribution, we're interested in the relationship between the two variables***

- The **covariance** quantifies the joint deviation of two variables from their respective mean
  - It can take values from $-\infty$ to ∞ and depends on the unit of the data

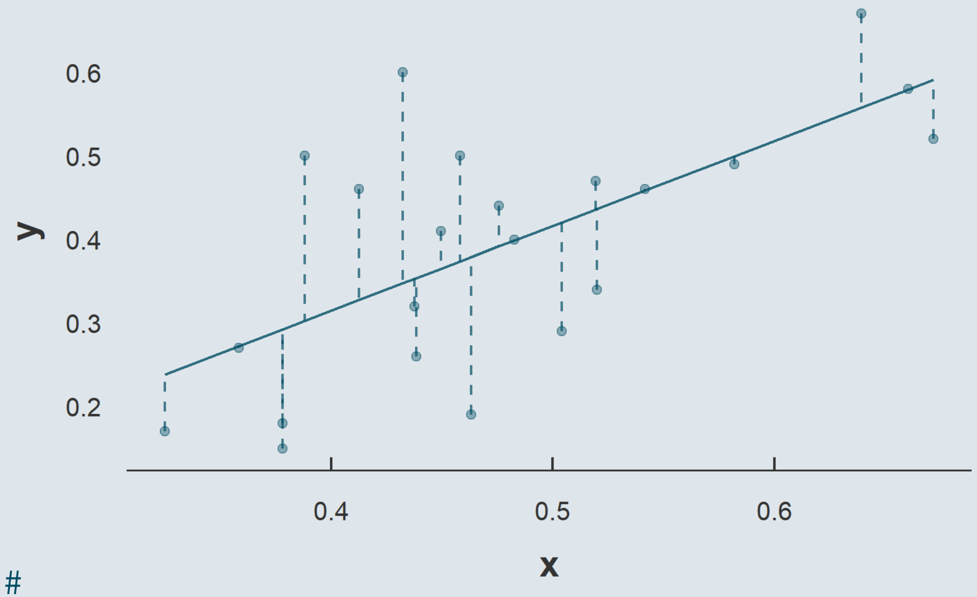
$$\text{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- The **correlation** is the covariance of two variables divided by the product of their standard deviation
  - It can take values from -1 to 1 and is independent from the unit of the data

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x) \times \text{SD}(y)}$$

Last time we saw

2. Regression




```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Coefficients:
## (Intercept)            x
##    -0.09129      1.01546
```

- This can be expressed with the **regression equation:**

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{\varepsilon}_i$$

- Where $\hat{\alpha}$ is the **intercept** and $\hat{\beta}$ the **slope** of the **line** $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, and $\hat{\varepsilon}_i$ the **distances** between the points and the line

$$\hat{\beta} = \frac{\text{Cov}(x_i, y_i)}{\text{Var}(x_i)}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \times \bar{x}$$

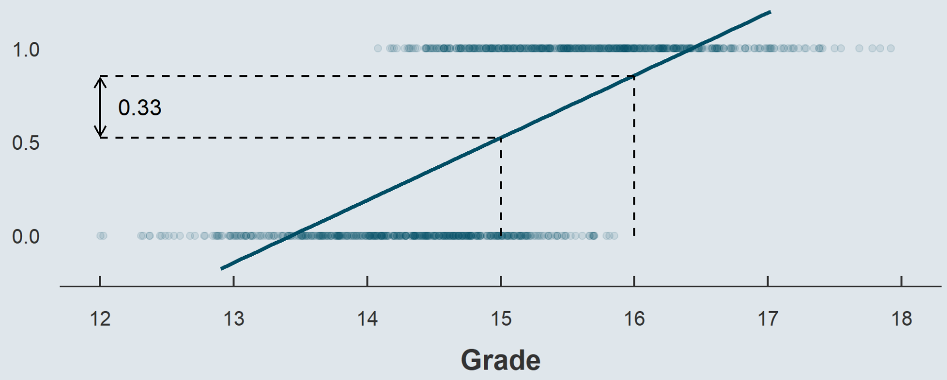
- $\hat{\alpha}$ and $\hat{\beta}$ minimize $\hat{\varepsilon}_i$

Last time we saw

3. Binary variables

Binary **dependent** variables

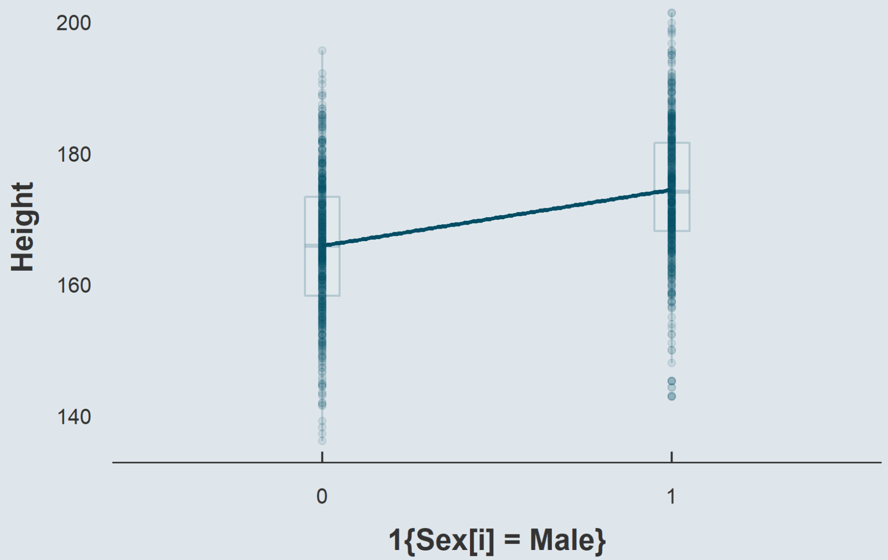
- The **fitted values** can be viewed as **probabilities**
  - $\hat{\beta}$ is the expected increase in the probability that $y = 1$ for a one unit increase in x



  - We call that a **Linear Probability model**

Binary **independent** variables

- The x variable should be viewed as a **dummy 0/1**
  - $\hat{\beta}$ is the difference between the average y for the group $x = 1$ and the group $x = 0$



Warm up practice

1) Open the `asec.csv` data containing sex, race, weekly work hours, and annual earnings ($)

2) Regress the earnings variable on the sex variable

3) Check that the slope coefficient is equal to the difference between male and female average earnings

You've got 10 minutes!

Solution

1) Open the `asec.csv` data containing sex, race, weekly work hours, and annual earnings ($)

```
asec <- read.csv("asec.csv")
```

2) Regress the earnings variable on the sex variable

```
lm(Earnings ~ Sex, asec)
```

```
##
## Call:
## lm(formula = Earnings ~ Sex, data = asec)
##
## Coefficients:
## (Intercept)      SexMale
##       50915        21612
```

Solution

3) Check that the slope coefficient is equal to the difference between male and female average earnings

```
asec %>%

  # Group the data by sex
  group_by(Sex) %>%

  # Summarise mean earnings -> 2x2 dataset
  summarise(Mean = mean(Earnings)) %>%

  # Put means in columns instead of rows -> 1x2 dataset
  pivot_wider(names_from = Sex, values_from = Mean) %>%

  # Compute the difference in means
  mutate(Difference = Male - Female)
```

```
## # A tibble: 1 x 3
##   Female   Male Difference
##    <dbl>  <dbl>      <dbl>
## 1 50915. 72527.     21612.
```

Today: Multivariate regressions!

1. Adding variables

- 1.1. Continuous variables
- 1.2. Discrete variables

2. Control variables

- 2.1. Motivation
- 2.2. Discrete controls
- 2.3. Continuous controls

3. Interactions

- 3.1. Motivation
- 3.2. Discrete interactions
- 3.3. Continuous interactions

4. Wrap up!

Today: Multivariate regressions!

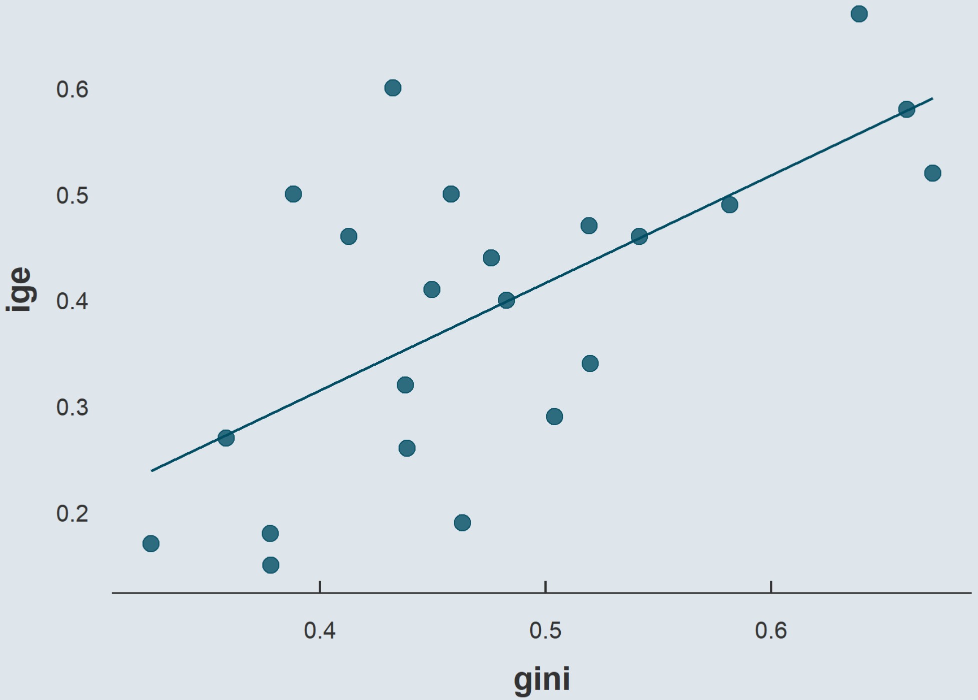
1. Adding variables

- 1.1. Continuous variables
- 1.2. Discrete variables

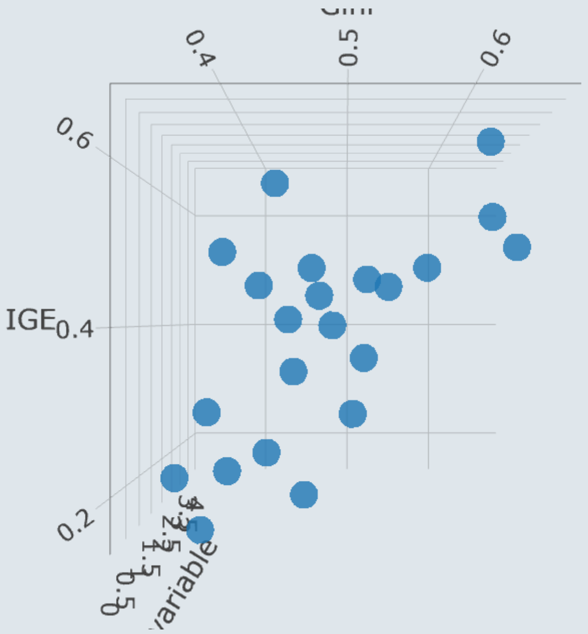
1. Adding variables

1.1. Continuous variables

- So far we focused on two-variable relationships

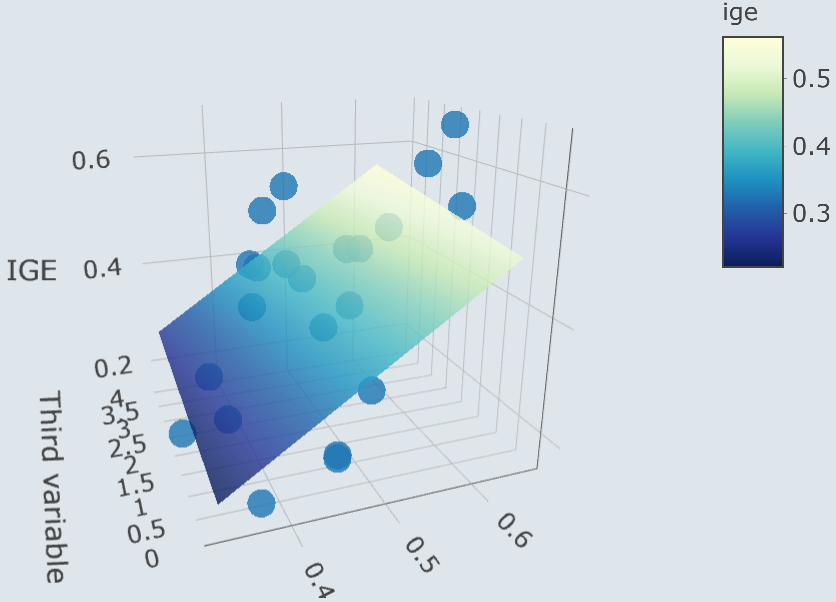


- What about three variable? *(pivot the plot)*



1. Adding variables

1.1. Continuous variables



- In this case we must fit a **plane**
  - It is characterized by **3 parameters**
  - And can be expressed as:

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \hat{\varepsilon}_i$$

- $\hat{\alpha}$ is still the **intercept**
  - The value of $\hat{y}$ (height) when $x_1 = x_2 = 0$

- And now there are **2 slopes**
  - $\hat{\beta}_1$ along the x_1 axis and $\hat{\beta}_2$ along the x_2 axis

1. Adding variables

1.1. Continuous variables

- The **same** applies with **more than 2** independent variables
  - We would fit a **hyperplane** with as many dimension as x variables
  - We would obtain one intercept and one slope per x variables

$$y_i = \hat{\alpha} + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \ldots + \hat{\beta}_k x_{k,i} + \hat{\varepsilon}_i$$

- We can estimate the parameters of these hyperplanes in **lm()**
  - **Additional variables** must be introduced after a **+ sign**

```
lm(ige ~ gini + third_variable, ggcurve)
```

```
##
## Call:
## lm(formula = ige ~ gini + third_variable, data = ggcurve)
##
## Coefficients:
##    (Intercept)            gini  third_variable
##       -0.09536         0.98153         0.01122
```

1. Adding variables

1.2. Discrete variables

- **So far** we've been working with **binary** categorical variables:
  - Accepted vs. Rejected, Male vs. Female
  - But what about discrete variables with **more than two categories?**

- Take for instance the **race variable:**

```
asec %>%
  group_by(Race) %>%
  tally()
```

```
## # A tibble: 3 x 2
##   Race       n
##   <chr> <int>
## 1 Black  6835
## 2 Other  6950
## 3 White 50551
```

How can we use this variable as an independent variable in our regression framework?

1. Adding variables

1.2. Discrete variables

- Remember how we converted our **2-category** variable into **1 dummy** variable
  - We can convert an **n-category** variable into **n-1 dummy** variables

Sex	Male		Race	Black	Other
Female	0		White	0	0
Female	0		White	0	0
Female	0		Black	1	0
Male	1		Black	1	0
Male	1		Other	0	1
Male	1		Other	0	1

➜ But why do we omit one category every time?

- Because it would be redundant
- We only need 2 dummies for 3 groups:
  - **White:** Black = **0** & Other = **0**
  - **Black:** Black = **1** & Other = **0**
  - **Other:** Black = **0** & Other = **1**

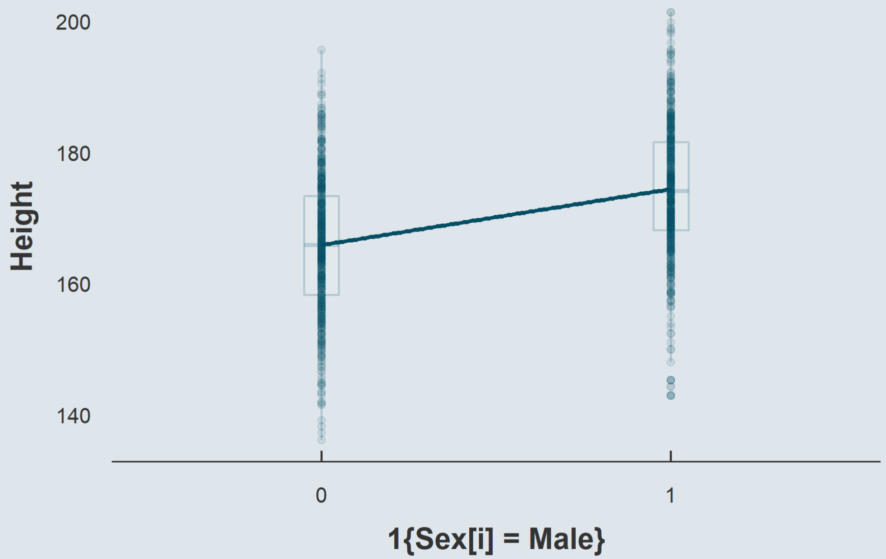
$\hat{\alpha}$ is the expected $\hat{y}$ when $x_k = 0 \, \forall k$

- Thus is does the job for the omitted groups!
- This group is called the **reference group**
- $\hat{\beta}_k$ are interpreted **relative** to that group

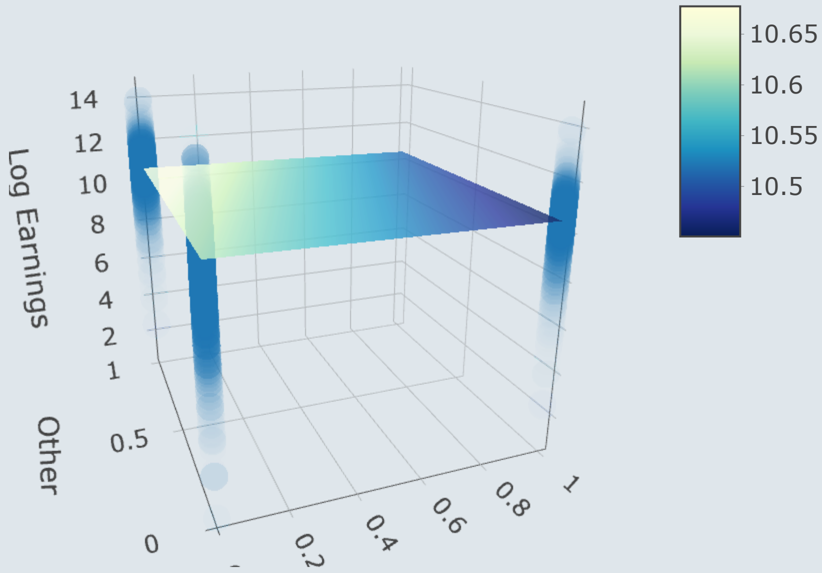
1. Adding variables

1.2. Discrete variables

2-category variable



3-category variable



1. Adding variables

1.2. Discrete variables

- This **plane** can be expressed as:

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1 1\{\text{Race}_i = \text{Other}\} + \hat{\beta}_2 1\{\text{Race}_i = \text{White}\} + \hat{\varepsilon}_i$$

- And the **average** incomes for each group equal:
  - **Black:** $\hat{\alpha} + 0\hat{\beta}_1 + 0\hat{\beta}_2 = \hat{\alpha}$
  - **Other:** $\hat{\alpha} + 1\hat{\beta}_1 + 0\hat{\beta}_2 = \hat{\alpha} + \hat{\beta}_1$
  - **White:** $\hat{\alpha} + 0\hat{\beta}_1 + 1\hat{\beta}_2 = \hat{\alpha} + \hat{\beta}_2$

```
##
## Call:
## lm(formula = Earnings ~ Race, data = asec)
##
## Coefficients:
## (Intercept)    RaceOther    RaceWhite
##       50577        17477        12303
```

Average by group

Race	Mean earnings
Black	50577.49
Other	68054.63
White	62880.49

1. Adding variables

1.2. Discrete variables

- By **default**, lm() sorts categories by **alphabetical** order
  - So every coefficient should be **interpreted relative** to the group which is first alphabetically

- But usually this is **not** the most **intuitive**
  - You may want everything to be relative to the **majority group**
  - Or to any group that has reasons to be the **reference**

- The **relevel()** function allows you to **change the reference** category
  - But it works **only on factor** variables

```r
asec <- asec %>%
  mutate(Race_fct = relevel(as.factor(Race),
                            "White"))

lm(Earnings ~ Race_fct, asec)
```

```
##
## Call:
## lm(formula = Earnings ~ Race_fct, data = asec)
##
## Coefficients:
##   (Intercept)  Race_fctBlack  Race_fctOther
##         62880         -12303           5174
```

1. Adding variables

1.2. Discrete variables

- What you can also do is **create the dummies yourself:**

```
asec <- asec %>%
  mutate(Black = as.numeric(Race == "Black"),
         Other = as.numeric(Race == "Other"))
```

```
lm(Earnings ~ Black + Other, asec)
```

```
##
## Call:
## lm(formula = Earnings ~ Black + Other, data = asec)
##
## Coefficients:
## (Intercept)        Black        Other
##       62880       -12303         5174
```

→ *This might be the **safest** option*

1. Adding variables

1.2. Discrete variables

- But a **categorical** variable must **not** be introduced **as numeric** in lm()

```
asec <- asec %>%
  mutate(num_cat = case_when(Race == "White" ~ 0,
                             Race == "Black" ~ 1,
                             Race == "Other" ~ 3))
```

```
lm(Earnings ~ num_cat, asec)
```

```
##
## Call:
## lm(formula = Earnings ~ num_cat, data = asec)
##
## Coefficients:
## (Intercept)      num_cat
##     61799.2        774.3
```

*→ lm() used our **categorical** variable as a **continuous** variable*

1. Adding variables

1.2. Discrete variables

- Use the **factor** class

```
asec <- asec %>%
  mutate(fac_cat = as.factor(num_cat))
```

```
lm(Earnings ~ fac_cat, asec)
```

```
##
## Call:
## lm(formula = Earnings ~ fac_cat, data = asec)
##
## Coefficients:
## (Intercept)     fac_cat1     fac_cat3
##       62880       -12303         5174
```

→ ***Converting*** *all your* ***categorical*** *variables into* ***factors*** *is also a* ***safe*** *option*

Overview

Overview

1. Adding variables ✓

- 1.1. Continuous variables
- 1.2. Discrete variables

2. Control variables

- 2.1. Motivation
- 2.2. Discrete controls
- 2.3. Continuous controls

2. Control variables

2.1. Motivation

- But **why** would we include **additional variables** in our regressions?
  - The main reason is to **control** for potential **confounders**

- Consider estimating the **relationship** between **income** and exposure air **pollution** in the Paris region

$$\text{Pollution}_i = \hat{\alpha}_1 + \hat{\beta}_1 \text{Income}_i + \hat{\varepsilon}_i$$

- You would probably expect that $\hat{\beta}_1 < 0$
  - Meaning that **higher income** earners live in **less polluted** areas
  - But the closer from **Paris** the higher the **rents** and the **ring-road**
  - This phenomenon might counteract this effect and pull $\hat{\beta}_1$ towards 0

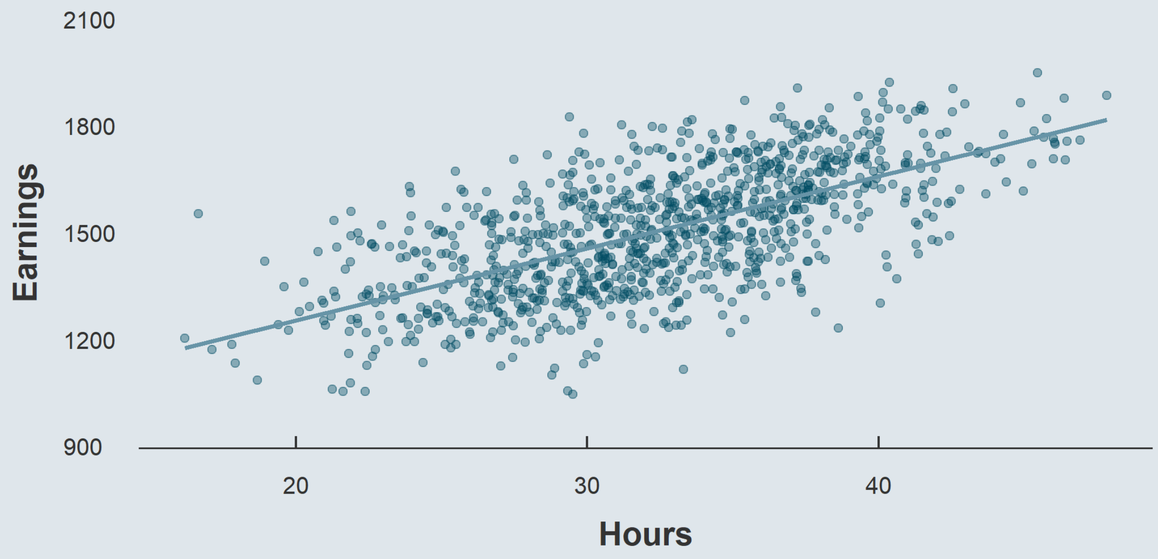
- But how to **remove** the **impact** that **distance** from Paris has on the relationship?
  - **Including it** in the regression would make the corresponding coefficient **absorb the confounding effect**
  - In that case we would call distance a ***control* variable**

$$\text{Pollution}_i = \hat{\alpha}_2 + \hat{\beta}_2 \text{Income}_i + \hat{\beta}_3 \text{Distance}_i + \hat{\epsilon}_i$$

2. Control variables

2.2. Discrete

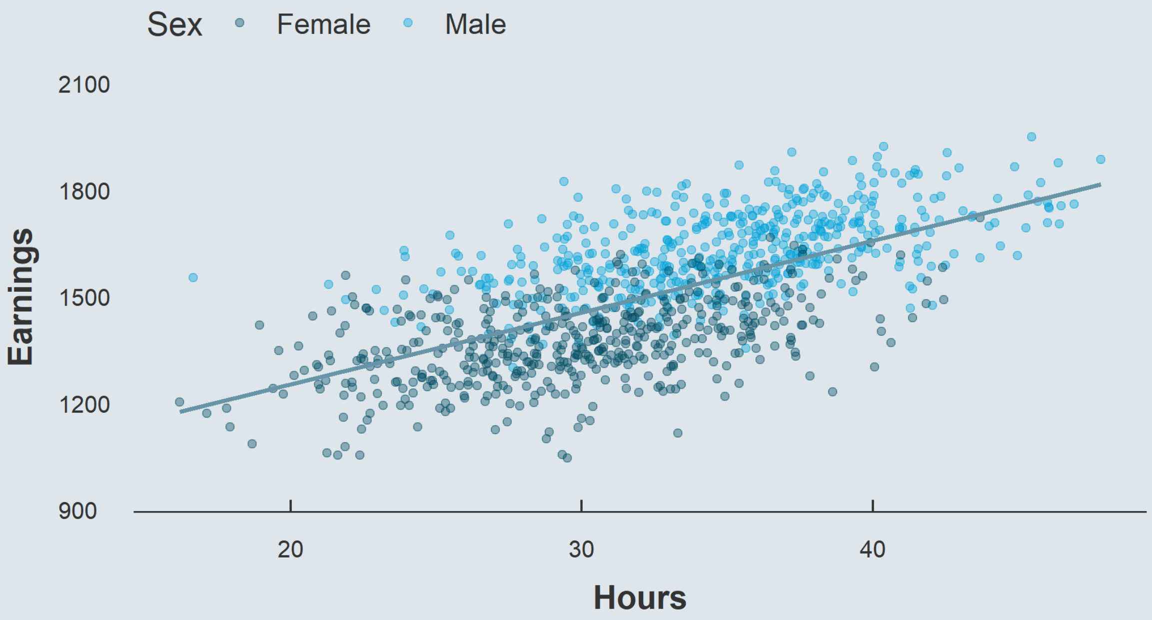
- The most **common control** variable is probably **sex/gender**
    - It may play a role in the **relationship** between **earnings** and **hours worked** for instance
    - The fact that **women** work **part time** more often and **earn less** contribute to the relationship
    - Just like distance did in the previous example



2. Control variables

2.2. Discrete

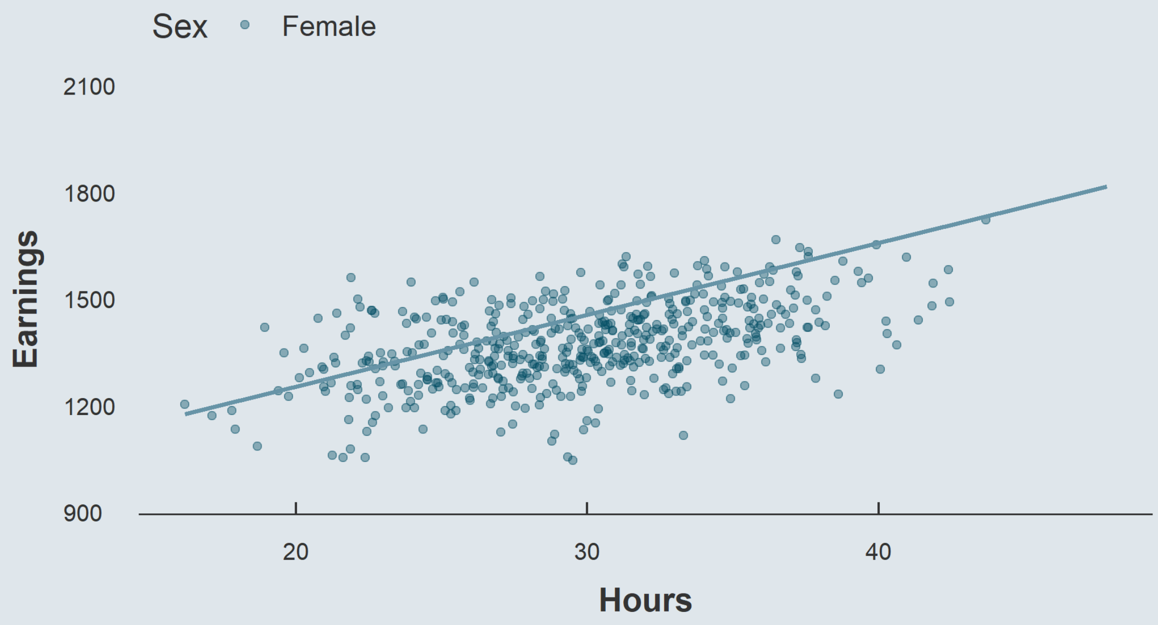
- The most **common control** variable is probably **sex/gender**
  - It may play a role in the **relationship** between **earnings** and **hours worked** for instance
  - The fact that **women** work **part time** more often and **earn less** contribute to the relationship
  - Just like distance did in the previous example



2. Control variables

2.2. Discrete

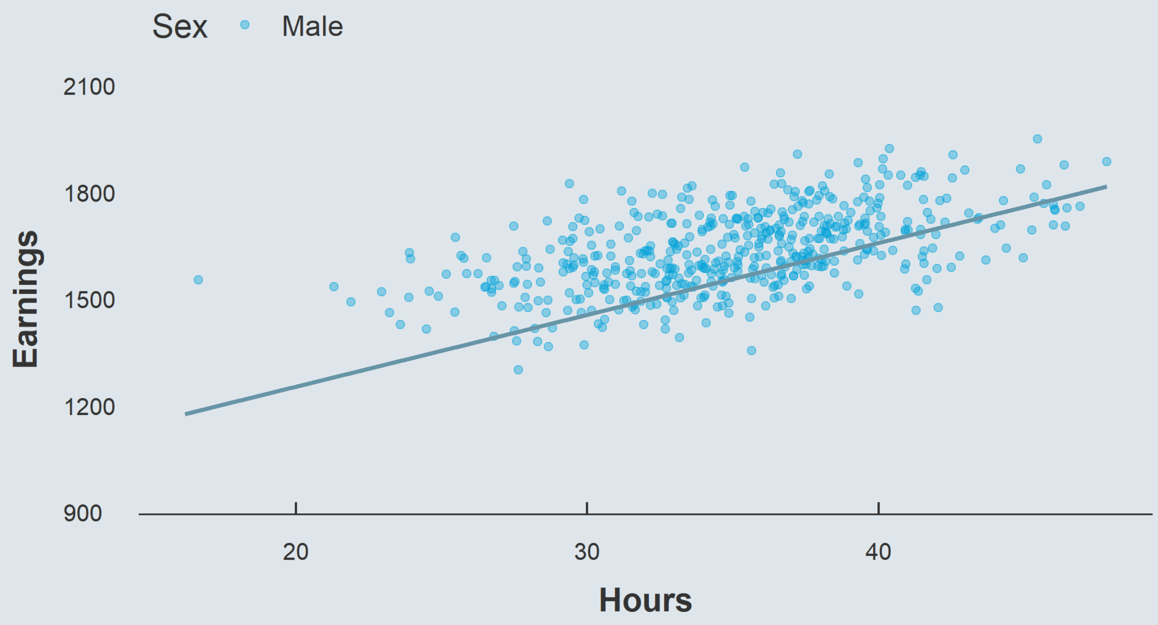
- The most **common control** variable is probably **sex/gender**
  - It may play a role in the **relationship** between **earnings** and **hours worked** for instance
  - The fact that **women** work **part time** more often and **earn less** contribute to the relationship
  - Just like distance did in the previous example



2. Control variables

2.2. Discrete

- The most **common control** variable is probably **sex/gender**
  - It may play a role in the **relationship** between **earnings** and **hours worked** for instance
  - The fact that **women** work **part time** more often and **earn less** contribute to the relationship
  - Just like distance did in the previous example



2. Control variables

2.2. Discrete

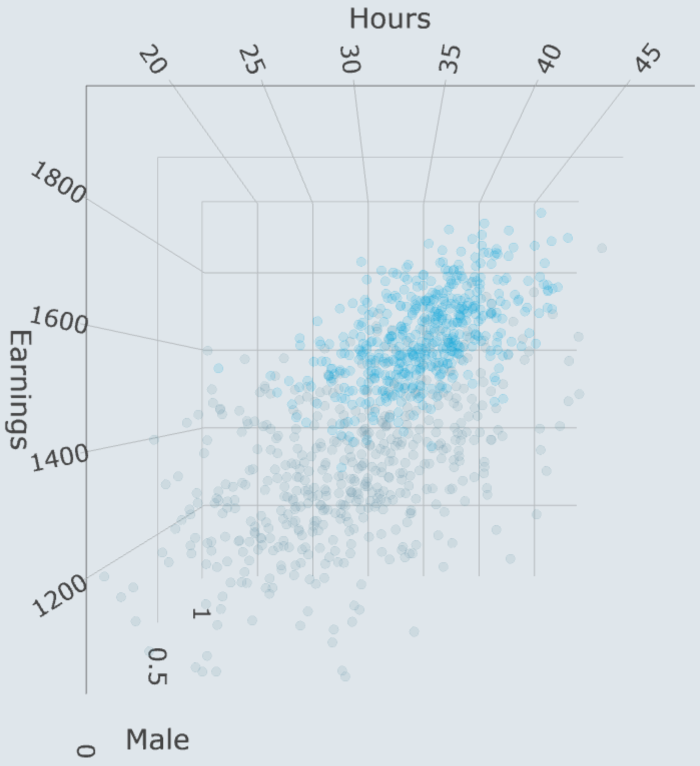
→ The **relationship** is indeed **inflated** by the sex variable

- Because being a **male** is positively **correlated** with **both x and y**

- **Controlling** for sex would **solve that problem** by absorbing this effect

- Controlling for a **discrete** variable amounts to allow **one intercept per category**

- Giving **two parallel fitted lines** which are the intersections of the plane and the scatterplots

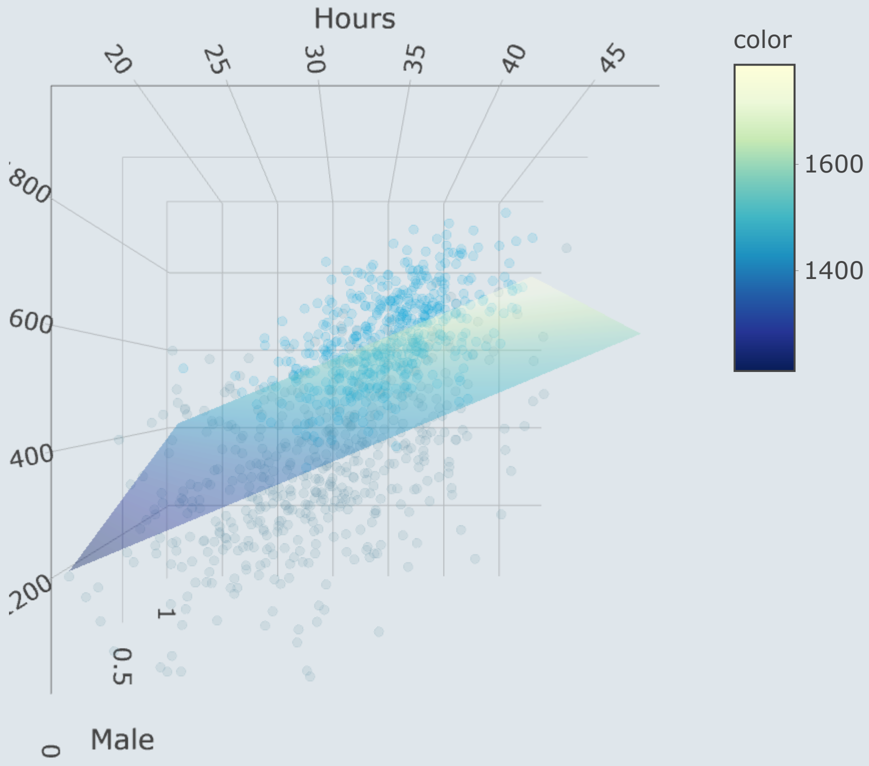


2. Control variables

2.2. Discrete

→ The **relationship** is indeed **inflated** by the sex variable

- Because being a **male** is positively **correlated** with **both x and y**
- **Controlling** for sex would **solve that problem** by absorbing this effect
- Controlling for a **discrete** variable amounts to allow **one intercept per category**
- Giving **two parallel fitted lines** which are the intersections of the plane and the scatterplots

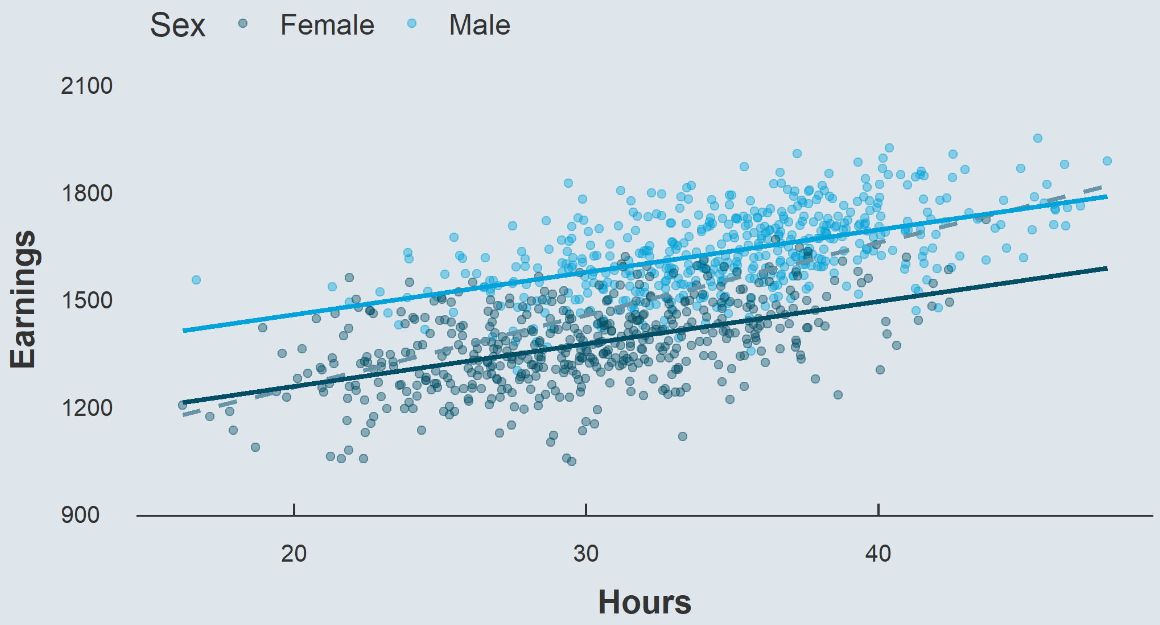


2. Control variables

2.2. Discrete

$$\text{Earnings}_i = \hat{\alpha} + \hat{\beta}_1 \text{Hours}_i + \hat{\beta}_2 1\{\text{Sex}_i = \text{Male}\} + \hat{\varepsilon}_i$$

```
## (Intercept)       Hours     SexMale
##  1019.34269    11.86326   200.98782
```



Graphical counterpart

$\hat{\alpha}$: Intercept of the reference group

$\hat{\beta}_1$: Common slope

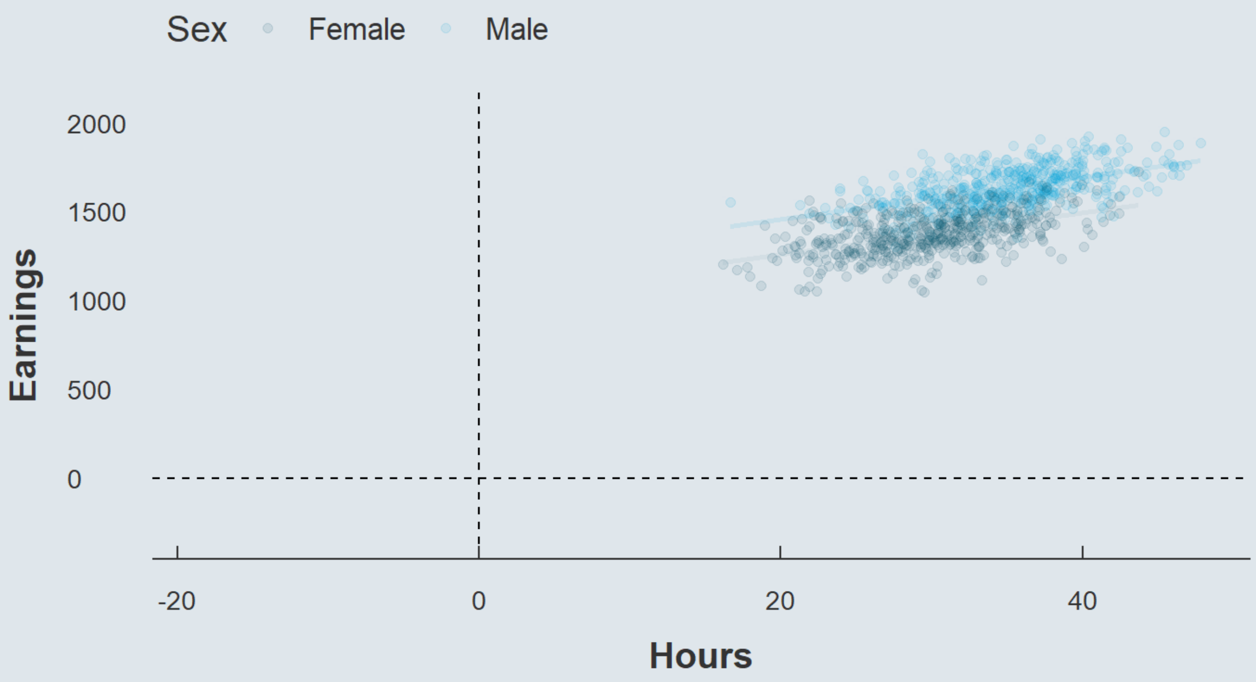
$\hat{\beta}_2$: Gap between the two lines

$\hat{\alpha} + \hat{\beta}_2$: Intercept of the other group

2. Control variables

2.2. Discrete

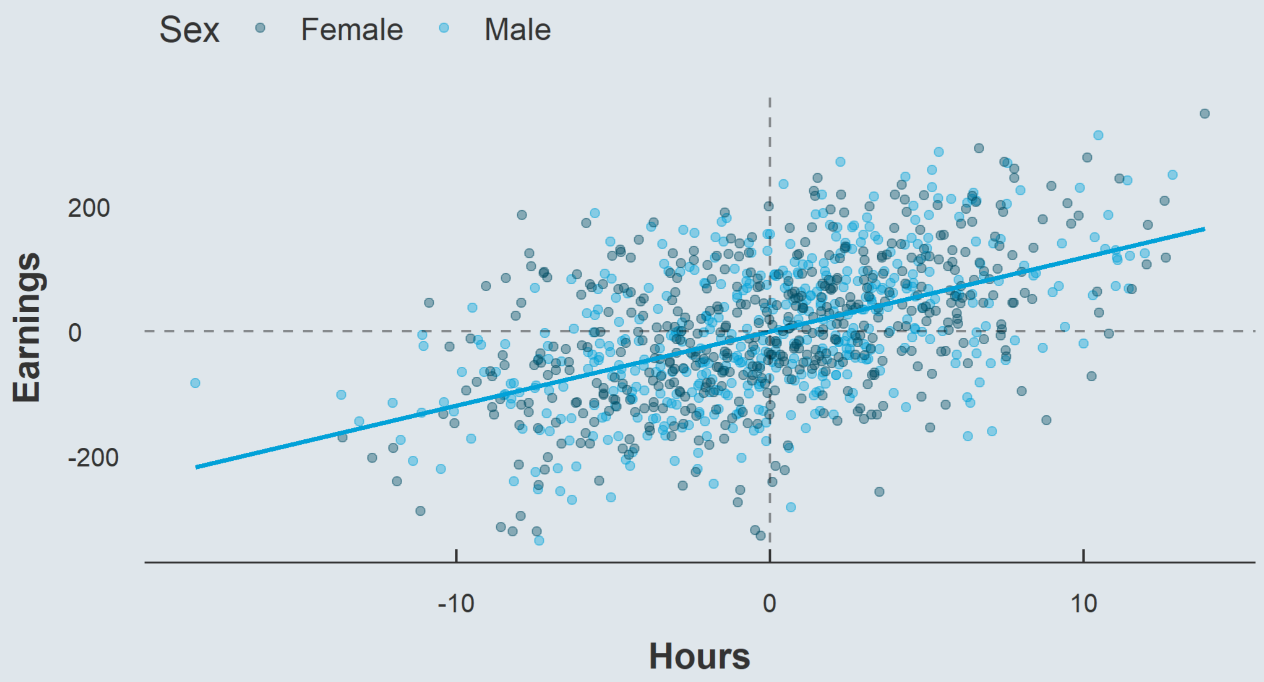
- We can **obtain** this common **slope** by:
  1. **Demeaning** earnings and hours by group
  2. **Regressing** the demeaned earnings on the hours



2. Control variables

2.2. Discrete

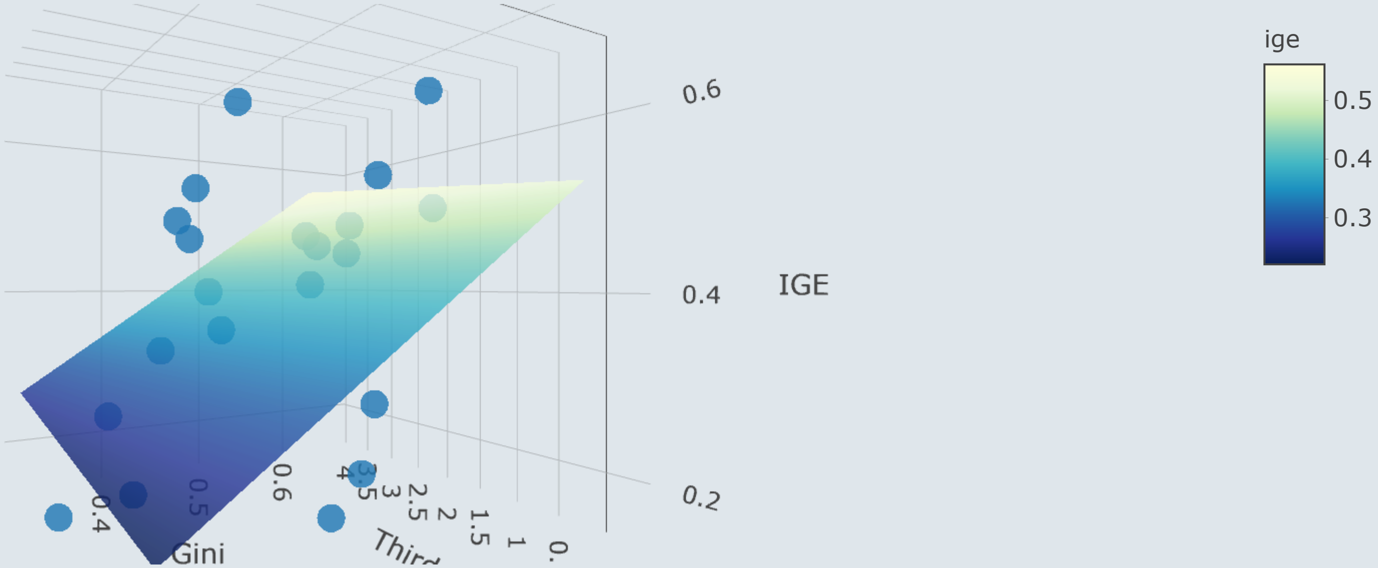
- Note that once we **control** for third variable
    1. As we move along the x axis, this **third variable remains constant**
    2. Here, as the number of **hours increases** the probability to be a **male does not** increase anymore



2. Control variables

2.3. Continuous

- The **same** idea apply when we control for **continuous** variables
  - Including it in the regression allows to **account for another dimension**
  - Such that when x moves this variable **remains constant**
  - This **nets out** the relationship of x and y from the potential **confounding effect** of this variable
  - This is why we call it **controlling for something**



Practice

1) Using the `asec` data, regress (yearly) earnings on (weekly) hours worked

2) Regress earnings on hours worked controlling for sex

3) Interpret the difference between the results from 1) and 2)

You've got 5 minutes!

Solution

1) Using the `asec` data, regress (yearly) earnings on (weekly) hours worked

```
lm(Earnings ~ Hours, asec)$coefficients
```

```
## (Intercept)       Hours
##   -20038.85     2077.79
```

2) Regress earnings on hours worked controlling for sex

```
lm(Earnings ~ Hours + Sex, asec)$coefficients
```

```
## (Intercept)       Hours     SexMale
##  -22296.150    1953.829   13794.385
```

Solution

3) Interpret the difference between the results from 1) and 2)

- The **slope** is still positive **less steep**
  - In the **first regression** as the number of **hours increases** the probability to be a **male does increase** as well
  - Because **males** tend to **earn more** this **contributes** to the positive **relationship** between Hours and Earnings
  - In the **second regression**, **controlling** for sex allows to maintain the probability to be a **male constant** along the hour axis to **remove this effect**

Overview

Overview

1. Adding variables ✓

- 1.1. Continuous variables
- 1.2. Discrete variables

2. Control variables ✓

- 2.1. Motivation
- 2.2. Discrete controls
- 2.3. Continuous controls

3. Interactions

- 3.1. Motivation
- 3.2. Discrete interactions
- 3.3. Continuous interactions

3. Interactions

3.1. Motivation

- Now we know how to **remove** the **confounding effect** of a third variable by **controlling** for it
  - But what if the main **relationship varies** depending on the value of the **third variable?**

- Let's get back to the previous example

$$\text{Pollution}_i = \hat{\alpha} + \hat{\beta}_1 \text{Income}_i + \hat{\beta}_2 \text{Distance}_i + \hat{\epsilon}_i$$

- The **equation imposes** that the **effect** of income on pollution is **constant:** $\hat{\beta}_2$
  - But what if the relationship was actually not the same close to Paris than further away?
  - Maybe that the closer from Paris the larger the effect (higher segregation, ...)

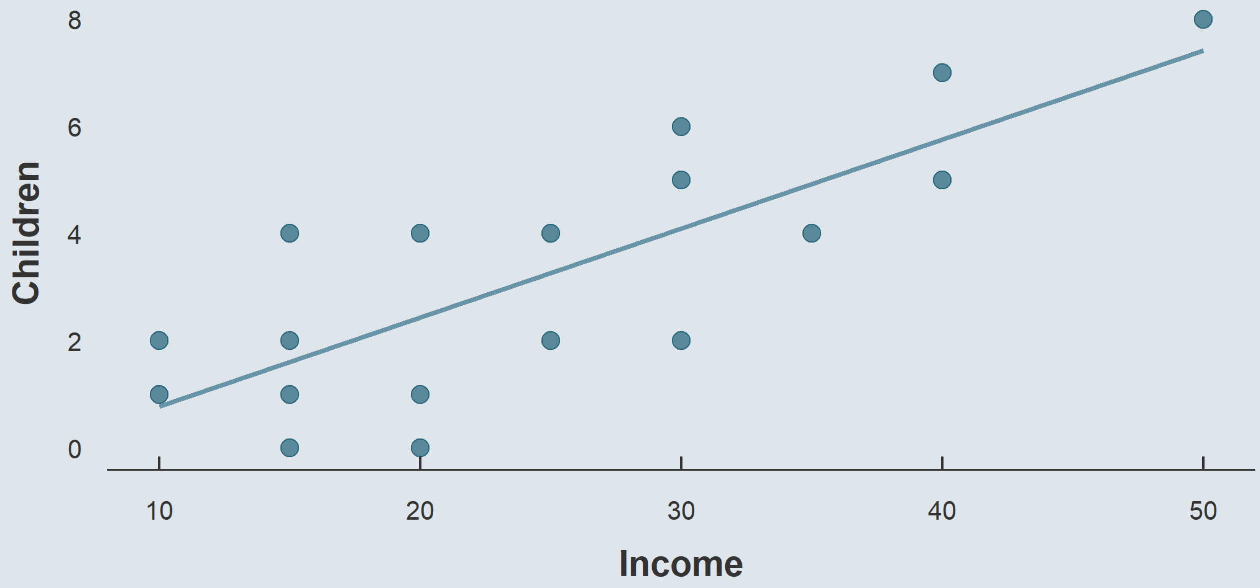
- But how to **capture how the relationship** between income and pollution varies with distance?
  - We should allow for it in the equation!
  - By **adding a term** that depends both on income and distance
  - What we use is their **product**, and we call that an **interaction**

$$\text{Pollution}_i = \hat{\alpha_2} + \hat{\beta}_3 \text{Income}_i + \hat{\beta}_4 \text{Distance}_i + \hat{\beta}_5 (\text{Distance}_i \times \text{Income}_i) + \hat{\epsilon}_i$$

3. Interactions

3.2. Discrete

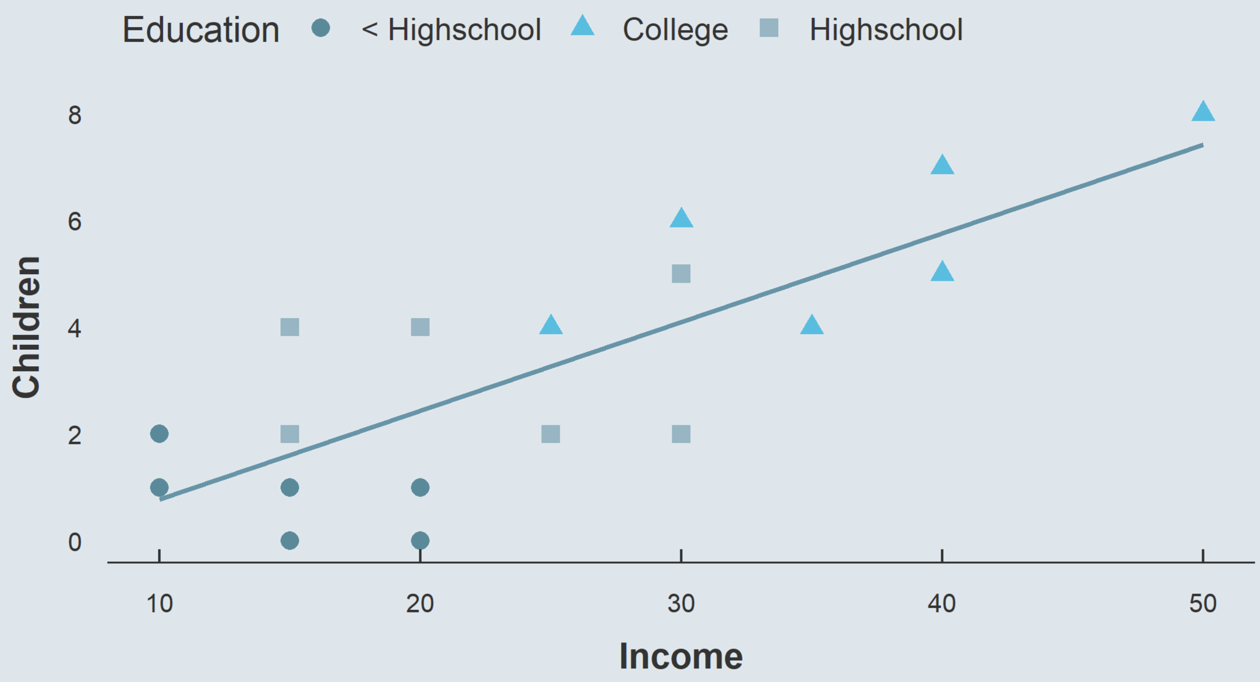
- Take for instance the following **relationship** between **household income** and the **number of children**



3. Interactions

3.2. Discrete

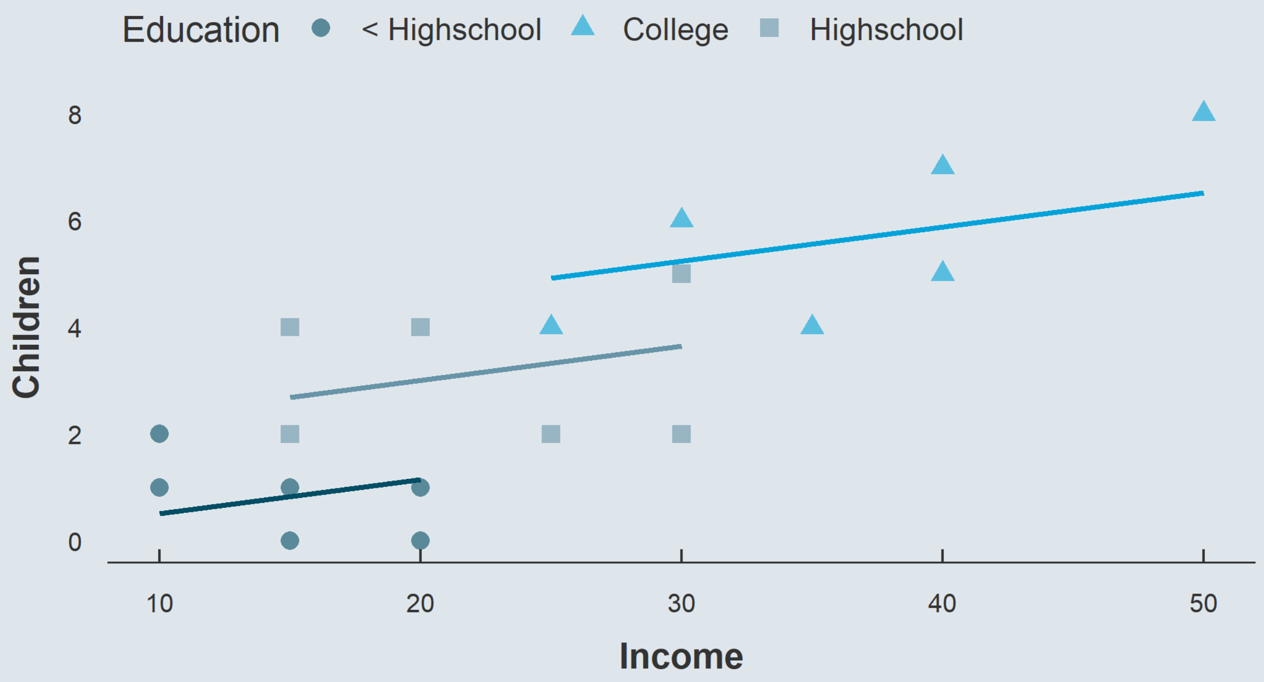
- Take for instance the following **relationship** between **household income** and the **number of children**
  - The level of **education** seems to **play a role** in the relationship



3. Interactions

3.2. Discrete

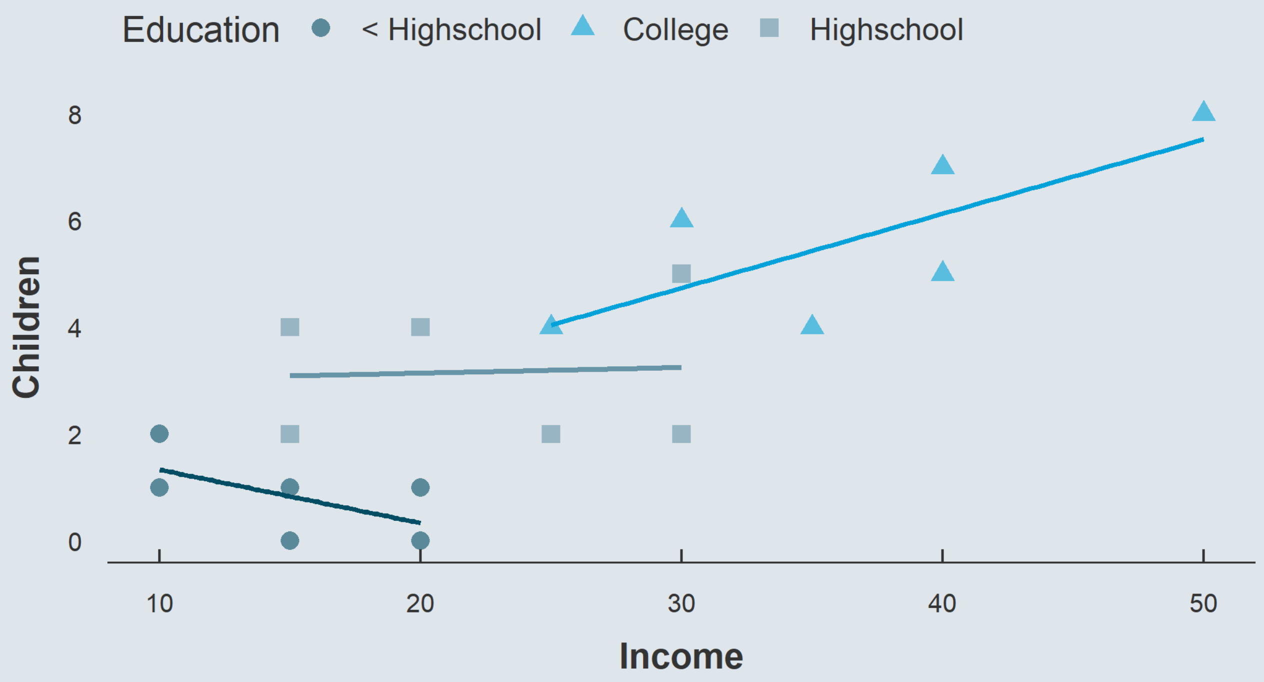
- Take for instance the following **relationship** between **household income** and the **number of children**
  - The level of **education** seems to **play a role** in the relationship
  - But simply **controlling** for education does **not** seem **sufficient**



3. Interactions

3.2. Discrete

- This is because the **relationship** between income and children **varies with education**
  - **Interacting** income with education allows to **account for that**
  - Like **controlling** allows for **different intercepts**, **interacting** allows for **different slopes**



3. Interactions

3.2. Discrete

→ *It is clearly* ***equivalent to regressing*** *children on income separately* ***per education group***

$$\begin{aligned}
\text{Income}_i = {} & \hat{\alpha} + \hat{\beta}_1 \text{Children}_i + && \text{Baseline equation} \\
& \hat{\beta}_2 (<\text{Highschool})_i + \hat{\beta}_3 \text{Highschool}_i + \hat{\beta}_4 \text{College}_i + && \text{Allow for} \neq \text{slopes} \\
& \text{Children}_i \times \left[\hat{\beta}_5 (<\text{Highschool})_i + \hat{\beta}_6 \text{Highschool}_i + \hat{\beta}_7 \text{College}_i \right] + \hat{\varepsilon}_i && \text{Allow for} \neq \text{intercepts}
\end{aligned}$$

3. Interactions

3.2. Discrete

→ *It is clearly* ***equivalent to regressing*** *children on income separately* ***per education group***

$$\begin{aligned} \text{Income}_i = {} & \hat{\alpha} + \hat{\beta}_1 \text{Children}_i + && \text{Baseline equation} \\ & \hat{\beta}_2 \underbrace{(<\text{Highschool})_i}_{1} + \hat{\beta}_3 \underbrace{\text{Highschool}_i}_{0} + \hat{\beta}_4 \underbrace{\text{College}_i}_{0} + && \text{Allow for} \neq \text{slopes} \\ & \text{Children}_i \times \left[\hat{\beta}_5 \underbrace{(<\text{Highschool})_i}_{1} + \hat{\beta}_6 \underbrace{\text{Highschool}_i}_{0} + \hat{\beta}_7 \underbrace{\text{College}_i}_{0} \right] + \hat{\varepsilon}_i && \text{Allow for} \neq \text{intercepts} \end{aligned}$$

< Highschool: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_5)\text{Children}_i + \hat{\varepsilon}_i$

3. Interactions

3.2. Discrete

→ *It is clearly* ***equivalent to regressing*** *children on income separately* ***per education group***

$$\text{Income}_i = \hat{\alpha} + \hat{\beta}_1 \text{Children}_i + \qquad \text{Baseline equation}$$

$$\hat{\beta}_2 \underbrace{(<\text{Highschool})_i}_{0} + \hat{\beta}_3 \underbrace{\text{Highschool}_i}_{1} + \hat{\beta}_4 \underbrace{\text{College}_i}_{0} + \qquad \text{Allow for} \neq \text{slopes}$$

$$\text{Children}_i \times \left[\hat{\beta}_5 \underbrace{(<\text{Highschool})_i}_{0} + \hat{\beta}_6 \underbrace{\text{Highschool}_i}_{1} + \hat{\beta}_7 \underbrace{\text{College}_i}_{0} \right] + \hat{\varepsilon}_i \qquad \text{Allow for} \neq \text{intercepts}$$

Highschool: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_6)\text{Children}_i + \hat{\varepsilon}_i$

3. Interactions

3.2. Discrete

→ *It is clearly* ***equivalent to regressing*** *children on income separately* ***per education group***

$$
\begin{aligned}
\text{Income}_i = \hat{\alpha} + \hat{\beta}_1 \text{Children}_i + & & \text{Baseline equation} \\
\hat{\beta}_2 \underbrace{(<\text{Highschool})_i}_{0} + \hat{\beta}_3 \underbrace{\text{Highschool}_i}_{0} + \hat{\beta}_4 \underbrace{\text{College}_i}_{1} + & & \text{Allow for} \neq \text{slopes} \\
\text{Children}_i \times \left[\hat{\beta}_5 \underbrace{(<\text{Highschool})_i}_{0} + \hat{\beta}_6 \underbrace{\text{Highschool}_i}_{0} + \hat{\beta}_7 \underbrace{\text{College}_i}_{1} \right] + \hat{\varepsilon}_i & & \text{Allow for} \neq \text{intercepts}
\end{aligned}
$$

College: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_4) + (\hat{\beta}_1 + \hat{\beta}_7)\text{Children}_i + \hat{\varepsilon}_i$

3. Interactions

3.2. Discrete

→ *It is clearly **equivalent to regressing** children on income separately **per education group***

$$\begin{aligned}
\text{Income}_i = {} & \hat{\alpha} + \hat{\beta}_1 \text{Children}_i + && \text{Baseline equation} \\
& \hat{\beta}_2 (<\text{Highschool})_i + \hat{\beta}_3 \text{Highschool}_i + \hat{\beta}_4 \text{College}_i + && \text{Allow for} \neq \text{slopes} \\
& \text{Children}_i \times \left[\hat{\beta}_5 (<\text{Highschool})_i + \hat{\beta}_6 \text{Highschool}_i + \hat{\beta}_7 \text{College}_i \right] + \hat{\varepsilon}_i && \text{Allow for} \neq \text{intercepts}
\end{aligned}$$

< Highschool: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_5)\text{Children}_i + \hat{\varepsilon}_i$

Highschool: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_6)\text{Children}_i + \hat{\varepsilon}_i$

College: $\text{Income}_i = (\hat{\alpha} + \hat{\beta}_4) + (\hat{\beta}_1 + \hat{\beta}_7)\text{Children}_i + \hat{\varepsilon}_i$

3. Interactions

3.3. Continuous

- The **same principle** applies to **continuous variables:**

$$\text{Pollution}_i = \hat{\alpha} + \hat{\beta}_1\text{Income}_i + \hat{\beta}_2\text{Distance}_i + \hat{\beta}_3(\text{Distance}_i \times \text{Income}_i) + \hat{\epsilon}_i$$

- What is the **effect of** 1-unit increase in **income here?**

$$\hat{\beta}_1 + \hat{\beta}_3\text{Distance}_i$$

- The **coefficient** associated with the **interaction**, $\hat{\beta}_3$, indicates:
  - By how the **effect** of a one unit increase in **income** on pollution **varies with distance**
  - When **distance = 0** the effect of income is $\hat{\beta}_1$
  - For every **additional unit** of distance, the effect of income on pollution **increases by** $\hat{\beta}_3$

➜ Don't omit to include your interaction variable as a control in the regression

Overview

1. Adding variables ✓

- 1.1. Continuous variables
- 1.2. Discrete variables

2. Control variables ✓

- 2.1. Motivation
- 2.2. Discrete controls
- 2.3. Continuous controls

3. Interactions ✓

- 3.1. Motivation
- 3.2. Discrete interactions
- 3.3. Continuous interactions

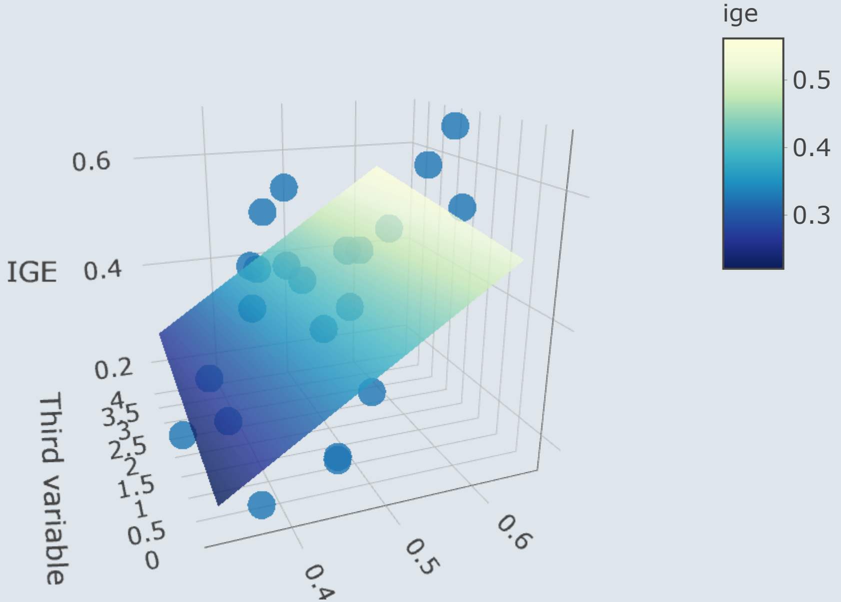
4. Wrap up!

4. Wrap up!

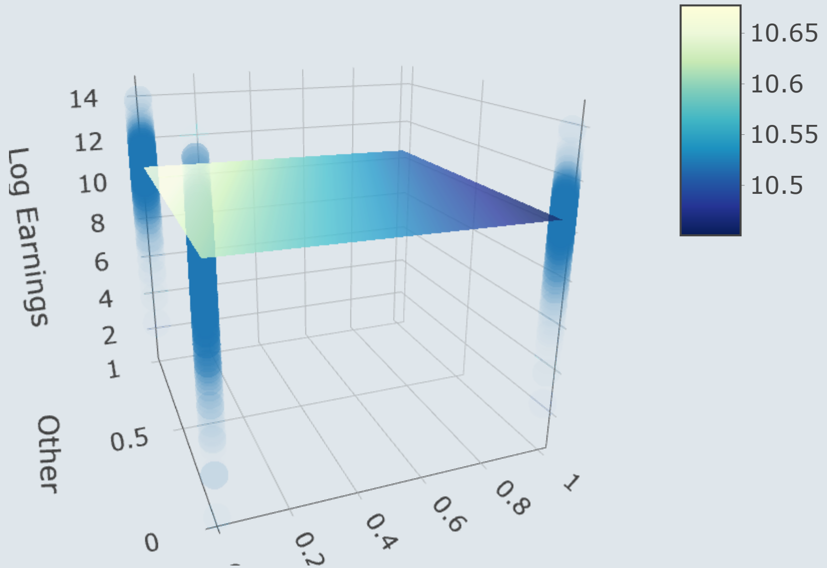
1. Multivariate regressions

- **Adding** a second independent **variable** in the regression amounts to **fitting a plane** instead of a line
  - Adding a third variable would fit an hyperplane of dimension 3 and so on

Adding a continuous variable



Adding a discrete variable

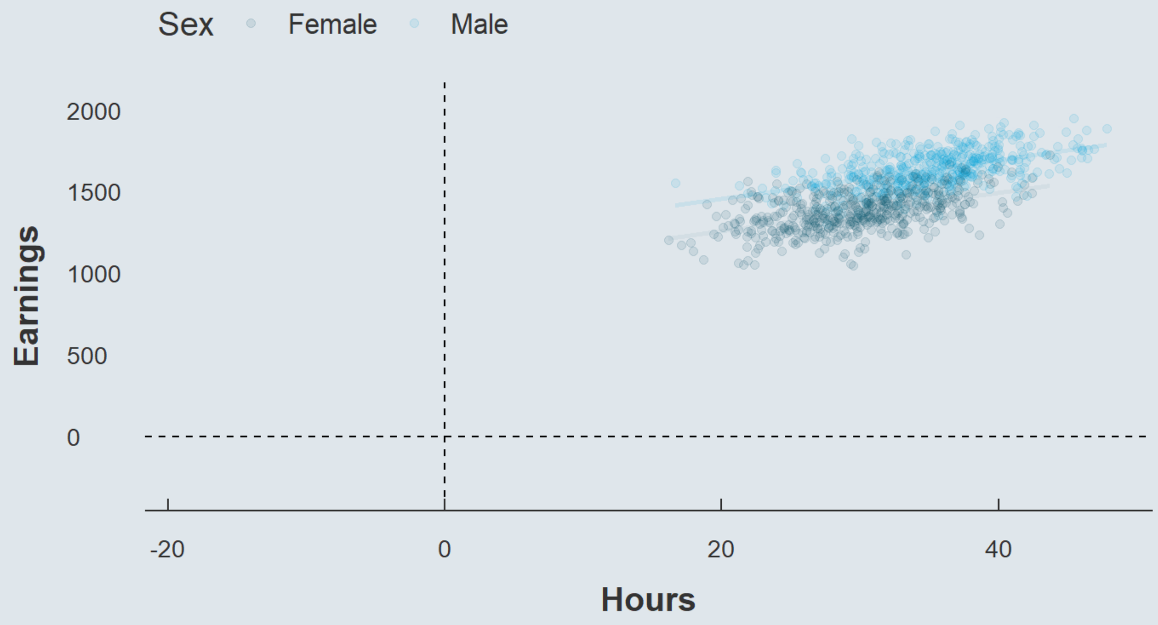


4. Wrap up!

2. Control variables

- Adding a third variable z **removes** its potential **confounding effect** from the relationship between x and y
  - As we move along the x axis, the **third variable remains constant**

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\varepsilon}_i$$



4. Wrap up!

3. Interactions

- Adding an **interaction** term with z allows to see **how the effect** of x on y **varies** with z
  - If z is **discrete**, it amounts to **regressing** y on x **separately** for each z group

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_3 (x \times z) + \hat{\varepsilon}_i$$