Lecture 13: Dummy Variables

ECON 480 - Econometrics - Fall 2018

Ryan Safner

November 7, 2018

Dummy Variables

Recoding Dummies

Dummy Variables

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

Overview of the Remainder of the Course

- You now have the “minimal toolkit” for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:
    1. “Data Wrangling”

HOOD COLLEGE

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

  1. "Data Wrangling"

     - Altering variables or data for useful analysis

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

  1. "Data Wrangling"
     - Altering variables or data for useful analysis
     - Dummy variables for categorical data (R calls them `factors`)

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data
2. Advanced identification strategies and unique problems

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data
2. Advanced identification strategies and unique problems
    - Time Series data

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data
2. Advanced identification strategies and unique problems
    - Time Series data
    - Panel data

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data
2. Advanced identification strategies and unique problems
    - Time Series data
    - Panel data
    - Fixed effects and random effects models

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

  1. "Data Wrangling"
     - Altering variables or data for useful analysis
     - Dummy variables for categorical data (R calls them `factors`)
     - Transforming variable scales or models to fit nonlinear data
  2. Advanced identification strategies and unique problems
     - Time Series data
     - Panel data
     - Fixed effects and random effects models
     - Difference-in-difference models

Overview of the Remainder of the Course

- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

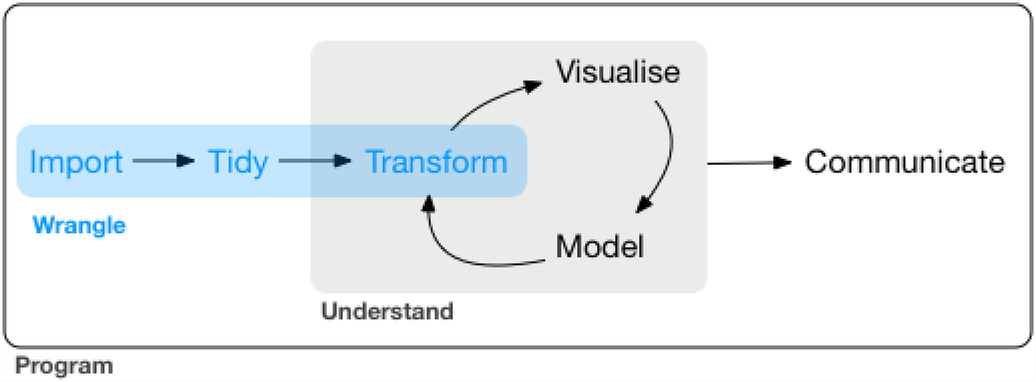
  1. "Data Wrangling"
     - Altering variables or data for useful analysis
     - Dummy variables for categorical data (R calls them `factors`)
     - Transforming variable scales or models to fit nonlinear data
  2. Advanced identification strategies and unique problems
     - Time Series data
     - Panel data
     - Fixed effects and random effects models
     - Difference-in-difference models
     - Instrumental variables models

Overview of the Remainder of the Course

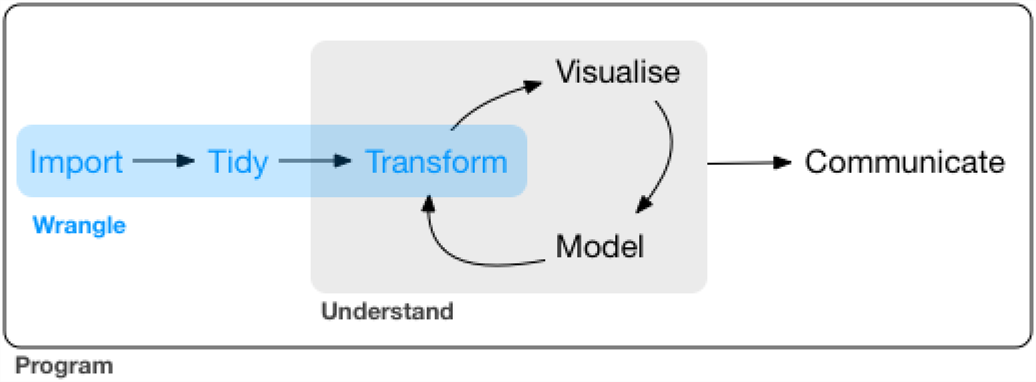
- You now have the "minimal toolkit" for data analysis with multivariate OLS regression
- The remainder of the course is about two types of extensions to the toolkit:

1. "Data Wrangling"
    - Altering variables or data for useful analysis
    - Dummy variables for categorical data (R calls them `factors`)
    - Transforming variable scales or models to fit nonlinear data
2. Advanced identification strategies and unique problems
    - Time Series data
    - Panel data
    - Fixed effects and random effects models
    - Difference-in-difference models
    - Instrumental variables models
    - Linear probability, logit, and probit models

- "Data wrangling" is a term for altering and cleaning data from raw form (often unusable) to a form that is useful for analysis (e.g. plotting and regressions)



- "Data wrangling" is a term for altering and cleaning data from raw form (often unusable) to a form that is useful for analysis (e.g. plotting and regressions)
- A **significant portion** of data analysis is initial data wrangling



- Recall categorial variables place an individual into one of several possible categories

Categorical Variables

- Recall categorical variables place an individual into one of several possible categories
  - e.g. sex, season, political party

Categorical Variables

- Recall categorial variables place an individual into one of several possible categories
  - e.g. sex, season, political party
  - may be responses to questions

- Recall categorial variables place an individual into one of several possible categories
  - e.g. sex, season, political party
  - may be responses to questions
  - can be quantitative (e.g. age, zip code)

Cut	Fair	Good	Very Good	Premium	Ideal
Count	1610	4906	12082	13791	21551
Proportion	0.030	0.091	0.224	0.256	0.400

Cut characteristics of 53,940 diamonds

- Recall categorial variables place an individual into one of several possible categories
  - e.g. sex, season, political party
  - may be responses to questions
  - can be quantitative (e.g. age, zip code)

Cut	Fair	Good	Very Good	Premium	Ideal
Count	1610	4906	12082	13791	21551
Proportion	0.030	0.091	0.224	0.256	0.400

Cut characteristics of 53,940 diamonds

- Also recall R calls this type of data a `factor`

Example

Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a t-**test**[1]

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a *t*-**test**[1]
- Let:

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a *t*-**test**[1]
- Let:
    - $\overline{Y}_M$ the average earnings of a sample of n_M men

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a t-**test**[1]
- Let:
  - $\overline{Y}_M$ the average earnings of a sample of n_M men
  - $\overline{Y}_W$ the average earnings of a sample of n_W women

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a *t*-**test**[1]
- Let:
  - $\overline{Y}_M$ the average earnings of a sample of n_M men
  - $\overline{Y}_W$ the average earnings of a sample of n_W women
  - Difference in group averages $d = \overline{Y}_M - \overline{Y}_W$

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a t-**test**[1]
- Let:
  - $\overline{Y}_M$ the average earnings of a sample of n_M men
  - $\overline{Y}_W$ the average earnings of a sample of n_W women
  - Difference in group averages $d = \overline{Y}_M - \overline{Y}_W$
- The hypothesis test is:

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a t-**test**[1]
- Let:
  - $\overline{Y}_M$ the average earnings of a sample of n_M men
  - $\overline{Y}_W$ the average earnings of a sample of n_W women
  - Difference in group averages $d = \overline{Y}_M - \overline{Y}_W$
- The hypothesis test is:
- $H_0 : d = 0$

[1] See the **Handout** on Blackboard for this example.

Example
Do men earn higher wages on average than women?

- Using basic statistics, we can test for a statistically significant difference in group means with a *t*-**test**[1]
- Let:
  - $\overline{Y}_M$ the average earnings of a sample of n_M men
  - $\overline{Y}_W$ the average earnings of a sample of n_W women
  - Difference in group averages $d = \overline{Y}_M - \overline{Y}_W$
- The hypothesis test is:
- $H_0 : d = 0$
- $H_1 : d \neq 0$

[1] See the **Handout** on Blackboard for this example.

- In a regression, we can easily compare across groups via a dummy variable[2]

[2] Also called a binary variable or dichotomous variable

- In a regression, we can easily compare across groups via a dummy variable[2]
  - Dummy variable *only* $= 0$ or $= 1$, depending on if a condition is met

[2] Also called a binary variable or dichotomous variable

- In a regression, we can easily compare across groups via a dummy variable[2]
  - Dummy variable *only* $= 0$ or $= 1$, depending on if a condition is met
  - Signifies whether an observation belongs to a category or not

[2] Also called a binary variable or dichotomous variable

- In a regression, we can easily compare across groups via a dummy variable[2]
  - Dummy variable *only* $= 0$ or $= 1$, depending on if a condition is met
  - Signifies whether an observation belongs to a category or not

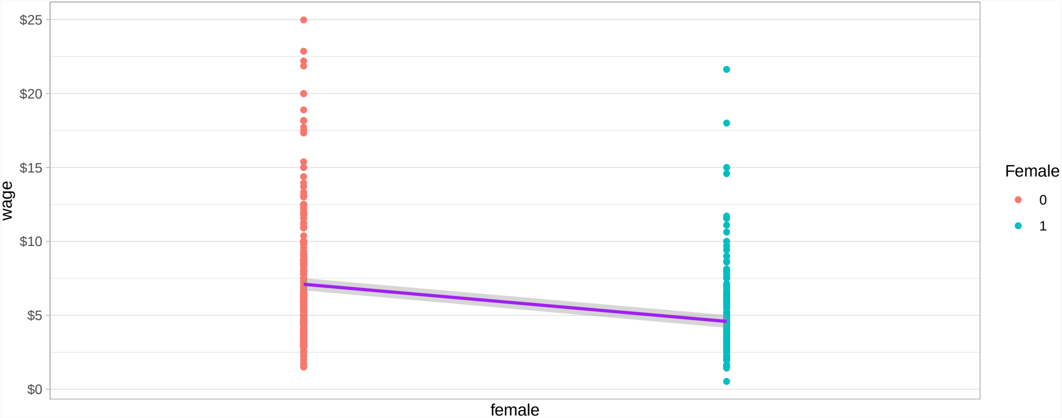
Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i \quad \text{where } Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

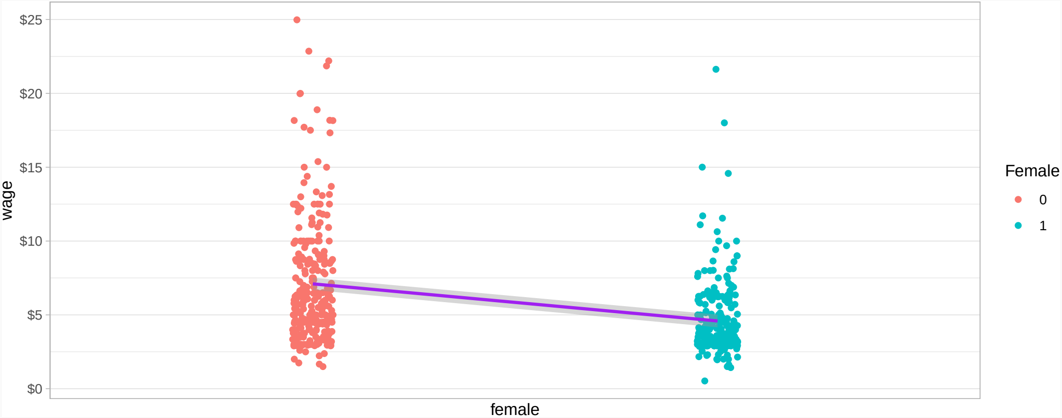
- Again, $\hat{\beta}_1$ makes less sense as the "slope" of a line in this context

[2] Also called a binary variable or dichotomous variable

Comparing Groups in Regression: Scatterplot

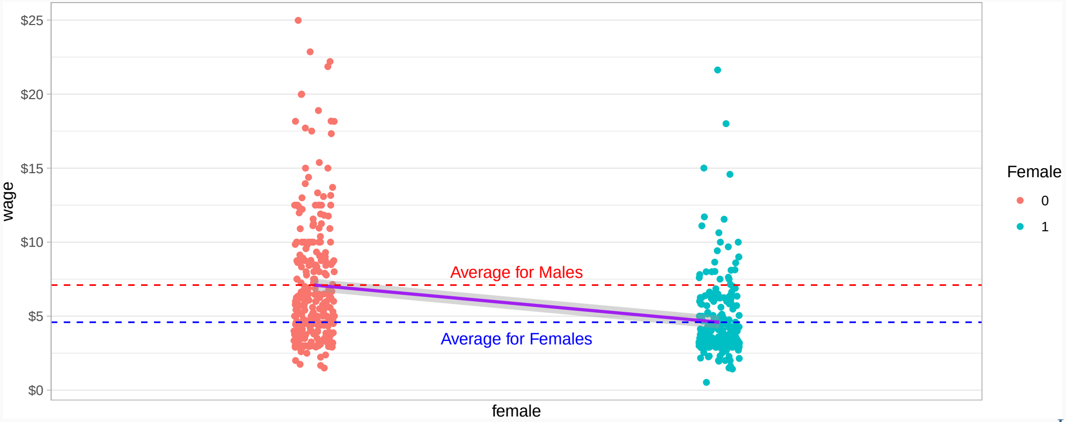


Comparing Groups in Regression: Scatterplot with Jittering



- use `geom_jitter()` instead of `geom_point()` to "jitter" the data to avoid overplotting

Comparing Groups in Regression: Scatterplot with Jittering II



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
    - $\hat{Y}_i = \hat{\beta}_0$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
    - $\hat{Y}_i = \hat{\beta}_0$
    - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
  - $\hat{Y}_i = \hat{\beta}_0$
  - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$
- When $D_i = 1$ (Treatment group):

Dummy Variables as Group Means

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
  - $\hat{Y}_i = \hat{\beta}_0$
  - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$
- When $D_i = 1$ (Treatment group):
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$

Dummy Variables as Group Means

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
  - $\hat{Y}_i = \hat{\beta}_0$
  - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$
- When $D_i = 1$ (Treatment group):
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$
  - $E[Y|D_i = 1] = \hat{\beta}_0 + \hat{\beta}_1 \iff$ the mean of Y when $D_i = 1$

Dummy Variables as Group Means

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
  - $\hat{Y}_i = \hat{\beta}_0$
  - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$
- When $D_i = 1$ (Treatment group):
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$
  - $E[Y|D_i = 1] = \hat{\beta}_0 + \hat{\beta}_1 \iff$ the mean of Y when $D_i = 1$
- So the *difference* in group means:

Dummy Variables as Group Means

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i \quad \text{where } D_i = \{0, 1\}$$

- When $D_i = 0$ (Control group):
  - $\hat{Y}_i = \hat{\beta}_0$
  - $E[Y|D_i = 0] = \hat{\beta}_0 \iff$ the mean of Y when $D_i = 0$
- When $D_i = 1$ (Treatment group):
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$
  - $E[Y|D_i = 1] = \hat{\beta}_0 + \hat{\beta}_1 \iff$ the mean of Y when $D_i = 1$
- So the *difference* in group means:

$$\begin{aligned} &= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &= (\hat{\beta}_0 + \hat{\beta}_1) - (\hat{\beta}_0) \\ &= \hat{\beta}_1 \end{aligned}$$

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } \textit{Female} \\ 0 & \text{If } i \text{ is } \textit{Male} \end{cases}$$

Example

$$\widehat{Wage}_i = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

- Mean wage for males:

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

- Mean wage for males: $E[Wage|Female = 0] = \hat{\beta}_0$

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

- Mean wage for males: $E[Wage|Female = 0] = \hat{\beta}_0$
- Mean wage for females:

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

- Mean wage for males: $E[Wage|Female = 0] = \hat{\beta}_0$
- Mean wage for females: $E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

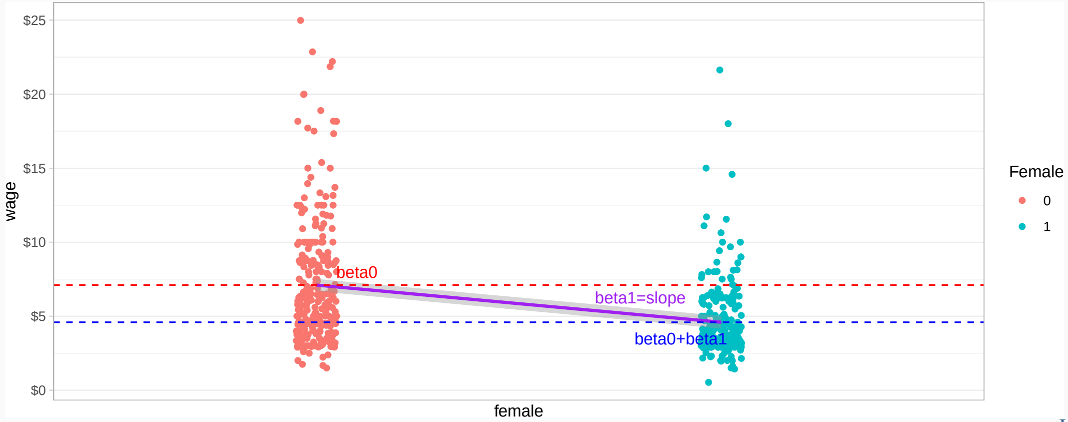
- Mean wage for males: $E[Wage|Female = 0] = \hat{\beta}_0$
- Mean wage for females: $E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$
- Difference in wage between males & females:

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i$$

$$Female_i = \begin{cases} 1 & \text{if } i \text{ is } Female \\ 0 & \text{If } i \text{ is } Male \end{cases}$$

- Mean wage for males: $E[Wage|Female = 0] = \hat{\beta}_0$
- Mean wage for females: $E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$
- Difference in wage between males & females: $\hat{\beta}_1$


$25
$20
$15
$10
$5
$0
wage
female
beta0
beta1=slope
beta0+beta1
Female
0
1

- OLS Regression: $\widehat{Wage_i} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\, Female_i$

Dummy Regression vs. Group Means

- OLS Regression: $$\widehat{\text{Wage}_i} = 7.10 - 2.51\,Female_i$$ $$\quad\quad(0.21)\quad(0.30)$$
- Simple tabulation of group means:

Sex	Avg. Wage ($\bar{Y}$)	SE(avg) (s_Y)	n
Female	4.59	0.16	252
Male	7.10	0.21	274
Difference	-2.51	(0.30)	-

- OLS Regression: $\widehat{\text{Wage}_i} = 7.10 - 2.51\, Female_i$
  $\qquad (0.21) \quad (0.30)$
- Simple tabulation of group means:

Sex	Avg. Wage ($\bar{Y}$)	SE(avg) (s_Y)	n
Female	4.59	0.16	252
Male	7.10	0.21	274
Difference	-2.51	(0.30)	-

- Differences in means: $\overline{Y_F} - \overline{Y_M} = 4.59 - 7.10 = -2.51$

Dummy Regression vs. Group Means

- OLS Regression: $\widehat{\text{Wage}_i} = 7.10 - 2.51\, Female_i$

  $\qquad\qquad\qquad (0.21) \quad (0.30)$

- Simple tabulation of group means:

Sex	Avg. Wage ($\bar{Y}$)	SE(avg) (s_Y)	n
Female	4.59	0.16	252
Male	7.10	0.21	274
Difference	-2.51	(0.30)	-

- Differences in means: $\overline{Y_F} - \overline{Y_M} = 4.59 - 7.10 = -2.51$
- $SE(\overline{Y_F} - \overline{Y_M}) = \sqrt{\frac{s_M^2}{n_M} + \frac{s_F^2}{n_F}} = \sqrt{\frac{0.21^2}{274} + \frac{0.16^2}{252}} \approx 0.30$

```r
# Our data comes from WAGE1.dta which you can find in Blackboard under data

# Load WAGE1 as wages
library("foreign") # to load .dta Stata files
wages<-read.dta("../Data/WAGE1.dta")

# there's a lot of variables in wages, let's only look at wage and female for no
wages<-subset(wages, select=c("wage","female"))
```

Looking at the Data *Conditionally* II

```r
# just get a sense of the data

head(wages)
```

```
##   wage female
## 1 3.10      1
## 2 3.24      1
## 3 3.00      0
## 4 6.00      0
## 5 5.30      0
## 6 8.75      0
```

- We want to look at the data under certain **conditions**

[3] Later, I will show you how to do this in `dplyr`, a popular package that makes data wrangling easier

- We want to look at the data under certain **conditions**
- Can do this in base R by **subsetting** data using square brackets `[ ]`[3]

[3] Later, I will show you how to do this in `dplyr`, a popular package that makes data wrangling easier

Looking at the Data *Conditionally* III

- We want to look at the data under certain **conditions**
- Can do this in base R by **subsetting** data using square brackets `[ ]`[3]
- Syntax: `data[df$variable condition]` where `condition` is likely:

[3] Later, I will show you how to do this in `dplyr`, a popular package that makes data wrangling easier

LOOKING AT THE DATA *CONDITIONALLY* III

- We want to look at the data under certain **conditions**
- Can do this in base R by **subsetting** data using square brackets `[ ]`[3]
- Syntax: `data[df$variable condition]` where `condition` is likely:
  - A logical test, i.e. `>`, `<`, `!=`, `<=`, `>=`, `==` some value

[3] Later, I will show you how to do this in `dplyr`, a popular package that makes data wrangling easier

Looking at the Data *Conditionally* IV

```r
# look at average wage for men
summary(wages$wage[wages$female==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.500   4.143   6.000   7.099   8.765  24.980
```

```r
sd(wages$wage[wages$female==0]) # get sd
```

```
## [1] 4.160858
```

```r
# look at average wage for women
summary(wages$wage[wages$female==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.530   3.000   3.750   4.588   5.510  21.630
```

```r
sd(wages$wage[wages$female==1]) # get sd
```

```
## [1] 2.529363
```

```r
dummyreg<-lm(wage~female, data=wages)
summary(dummyreg)
```

```
##
## Call:
## lm(formula = wage ~ female, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0995     0.2100  33.806  < 2e-16 ***
## female       -2.5118     0.3034  -8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
```

$$\widehat{Wage_i} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\, Female_i$$

The Dummy Regression: Just Checking!

$$\widehat{\text{Wage}_i} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\, Female_i$$

- Does this mean we've accurately measured the gender-wage gap as \$2.51/hr?

The Dummy Regression: Just Checking!

$$\widehat{Wage_i} = 7.10 - 2.51\,Female_i$$
$$(0.21) \quad (0.30)$$

- Does this mean we've accurately measured the gender-wage gap as \$2.51/hr?
- Are there variables for which the following is true?

$$corr(wage, Z) \neq 0$$
$$corr(female, Z) \neq 0$$

The Dummy Regression: Just Checking!

$$\underset{(0.21)}{\widehat{\text{Wage}_i}} = 7.10 - 2.51\, Female_i$$

$$\widehat{\text{Wage}_i} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\, Female_i$$

- Does this mean we've accurately measured the gender-wage gap as \$2.51/hr?
- Are there variables for which the following is true?

$$corr(wage, Z) \neq 0$$
$$corr(female, Z) \neq 0$$

- `female` is probably endogenous, must include other control variables

Recoding Dummies

- What if instead of *female* we had used:

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Male_i$$

$$Male_i = \begin{cases} 1 & \text{if } i \text{ is } Male \\ 0 & \text{If } i \text{ is } Female \end{cases}$$

- What if instead of *female* we had used:

Example

$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Male_i$$

$$Male_i = \begin{cases} 1 & \text{if } i \text{ is } Male \\ 0 & \text{If } i \text{ is } Female \end{cases}$$

- *female* is a variable already in the data, we need to generate the *male* variable

- Again, a very useful R function is

```
ifelse(conditions, do.this.if.true, do.this.if.false)
```

Recoding Dummies

- Again, a very useful R function is

```r
ifelse(conditions, do.this.if.true, do.this.if.false)
```

- So let's create a `male` variable in our `wages` dataframe that we define as `1` if `female==0` and `0` otherwise (i.e. if `female==1`)

- Again, a very useful R function is

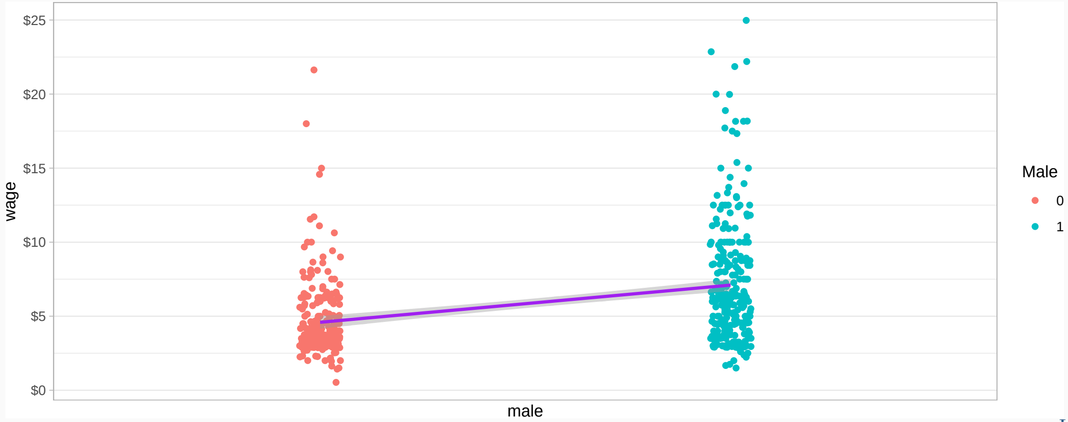
```
ifelse(conditions, do.this.if.true, do.this.if.false)
```

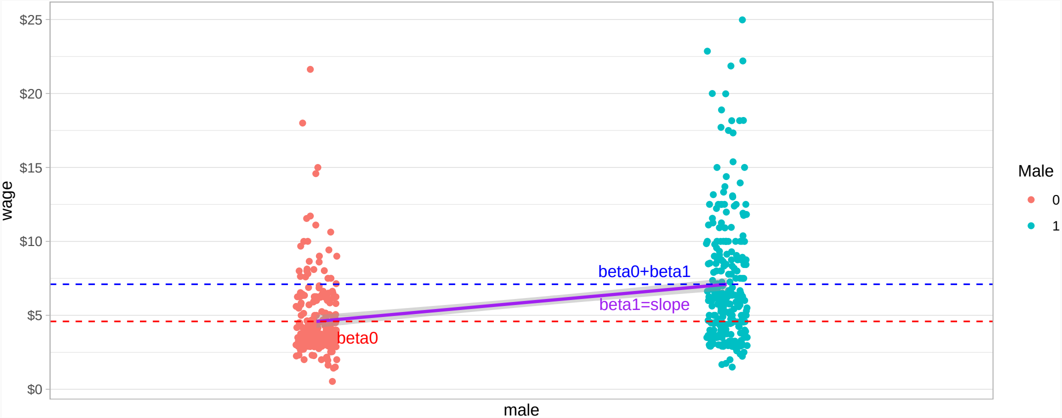
- So let's create a `male` variable in our `wages` dataframe that we define as `1` if `female==0` and `0` otherwise (i.e. if `female==1`)

```
wages$male<-ifelse(wages$female==0,1,0)
head(wages) # verify that it worked
```

```
##   wage female male
## 1 3.10      1    0
## 2 3.24      1    0
## 3 3.00      0    1
## 4 6.00      0    1
## 5 5.30      0    1
## 6 8.75      0    1
```

Scatterplot with Male




$25
$20
$15
$10
$5
$0
wage
male
beta0+beta1
beta1=slope
beta0
Male
0
1

The Dummy Regression with Male

```r
mreg<-lm(wage~male, data=wages)
summary(mreg)
```

```
##
## Call:
## lm(formula = wage ~ male, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5995 -1.8495 -0.9877  1.4260 17.8805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5877     0.2190  20.950  < 2e-16 ***
## male          2.5118     0.3034   8.279 1.04e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 524 degrees of freedom
```

$$\widehat{Wage_i} = \underset{(0.21)}{4.59} + \underset{(0.30)}{2.51}\, Male_i$$

The Dummy Regression: Male or Female

```r
library("stargazer")
stargazer(dummyreg, mreg, type="latex",
          header=FALSE, float=FALSE)
```

	Dependent variable:	
	wage	
	(1)	(2)
female	-2.512^{***}	
	(0.303)	
male		2.512^{***}
		(0.303)
Constant	7.099^{***}	4.588^{***}
	(0.210)	(0.219)
Observations	526	526
R^2	0.116	0.116
Adjusted R^2	0.114	0.114
Residual Std. Error (df = 524)	3.476	3.476
F Statistic (df = 1; 524)	68.537^{***}	68.537^{***}
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$	

- Note it doesn't matter if we use `male` or `female`, males always earn \$2.51 more than females

The Dummy Regression: Male or Female

```r
library("stargazer")
stargazer(dummyreg, mreg, type="latex",
          header=FALSE, float=FALSE)
```

	Dependent variable:	
	wage	
	(1)	(2)
female	-2.512^{***}	
	(0.303)	
male		2.512^{***}
		(0.303)
Constant	7.099^{***}	4.588^{***}
	(0.210)	(0.219)
Observations	526	526
R^2	0.116	0.116
Adjusted R^2	0.114	0.114
Residual Std. Error (df = 524)	3.476	3.476
F Statistic (df = 1; 524)	68.537^{***}	68.537^{***}
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

- Note it doesn't matter if we use `male` or `female`, males always earn $2.51 more than females
- Compare the constant (mean for the `D=0` group)