3.6 — Regression with Categorical Data

ECON 480 • Econometrics • Fall 2020

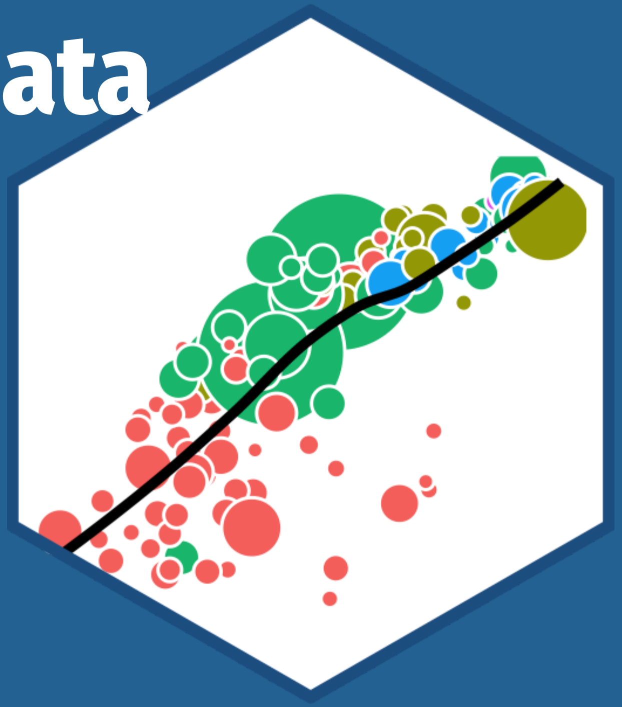
Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF20

metricsF20.classes.ryansafner.com



Outline

Categorical Data

- **Categorical data** place an individual into one of several possible *categories*
  - e.g. sex, season, political party
  - may be responses to survey questions
  - can be quantitative (e.g. age, zip code)
- `R` calls these `factors`

Question	Categories or Responses
Do you invest in the stock market?	___ Yes ___ No
What kind of advertising do you use?	___ Newspapers ___ Internet ___ Direct mailings
What is your class at school?	___ Freshman ___ Sophomore ___ Junior ___ Senior
I would recommend this course to another student.	___ Strongly Disagree ___ Slightly Disagree ___ Slightly Agree ___ Strongly Agree
How satisfied are you with this product?	___ Very Unsatisfied ___ Unsatisfied ___ Satisfied ___ Very Satisfied

Factors in R

- `factor` is a special type of `character` object class that indicates membership in a category (called a `level`)

- Suppose I have data on students:

```
students %>% head(n = 5)
```

ID	Rank	Grade
<dbl>	<chr>	<dbl>
1	Sophomore	77
2	Senior	72
3	Freshman	73
4	Senior	73
5	Junior	84

5 rows

Factors in R

- Rank is currently a `character` (`<chr>`) variable, but we can make it a `factor` variable, to indicate a student is a member of one of the possible categories: freshman, sophomore, junior, senior

```
students<-students %>%
  mutate(Rank = as.factor(Rank))
students %>% head(n = 5)
```

ID	Rank	Grade
<dbl>	<fctr>	<dbl>
1	Sophomore	77
2	Senior	72
3	Freshman	73
4	Senior	73
5	Junior	84

5 rows

- See now it's a `factor` (`<fctr>`)

Factors in R

```
# what are the categories?
students %>%
  group_by(Rank) %>%
  count()
```

Rank <fctr>	n <int>
Freshman	1
Junior	4
Senior	2
Sophomore	3

4 rows

```
# note the order is arbitrary!
```

Ordered Factors in R

- If there is a rank order you wish to preserve, you can make an `ordered factor`
  - list rankings from 1st to last

```
students<-students %>%
  mutate(Rank = ordered(Rank, levels = c("Freshman", "Sophomore", "Junior", "Senior")))
students %>% head(n = 5)
```

ID <dbl>	Rank <ord>	Grade <dbl>
1	Sophomore	77
2	Senior	72
3	Freshman	73
4	Senior	73
5	Junior	84

5 rows

Ordered Factors in R

```
students %>%
  group_by(Rank) %>%
  count()
```

Rank	n
<ord>	<int>
Freshman	1
Sophomore	3
Junior	4
Senior	2

4 rows

Example Research Question

Example: do men earn higher wages, on average, than women? If so, how much?

The Pure Statistics of Comparing Group Means

- Basic statistics: can test for statistically significant difference in group means with a **t-test**[†], let:

- Y_M: average earnings of a sample of n_M men

- Y_W: average earnings of a sample of n_W women

- **Difference** in group averages: $d = \bar{Y}_M - \bar{Y}_W$

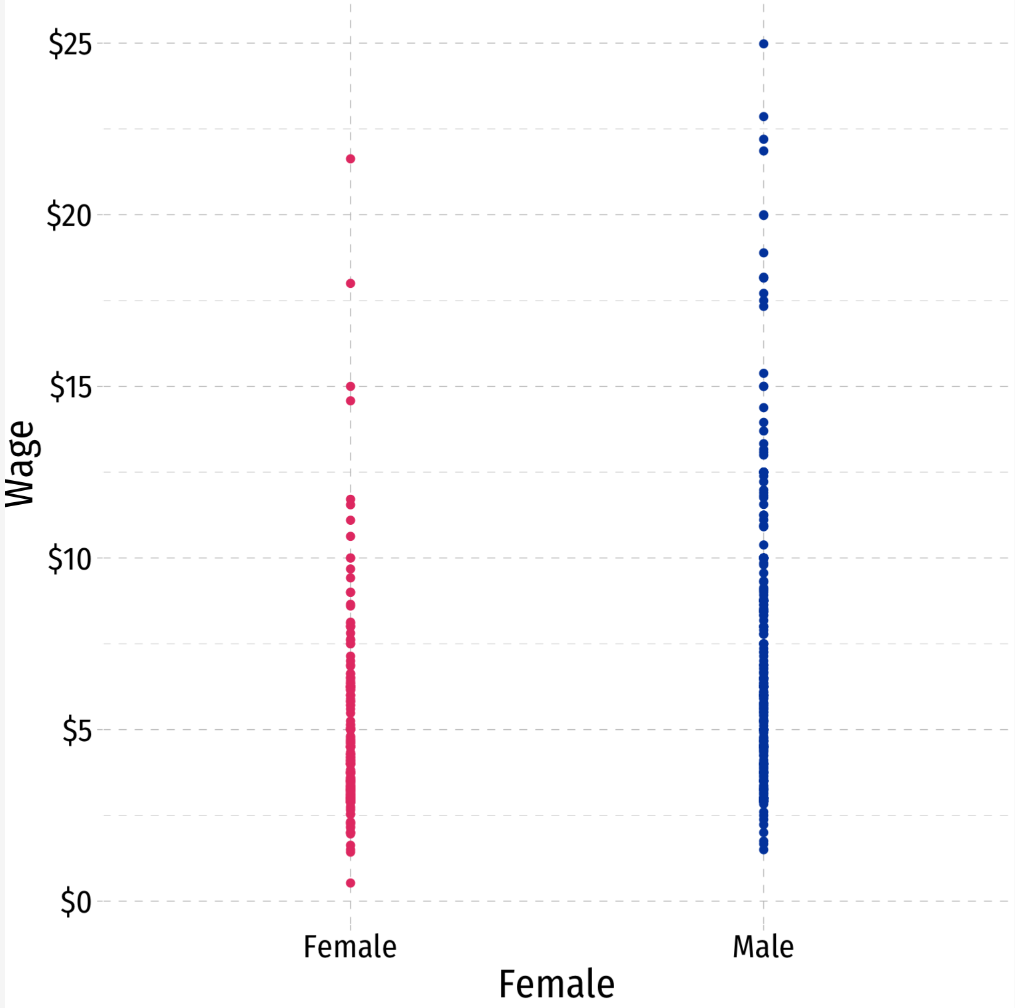
- The hypothesis test is:

  - $H_0 : d = 0$
  - $H_1 : d \neq 0$

[†] See today's class page for this example

Plotting Factors in R

- If I plot a `factor` variable, e.g. `Gender` (which is either `Male` or `Female`), the scatterplot with `wage` looks like this
  - effectively `R` treats values of a factor variable as integers
  - in this case, `"Female"` = 0, `"Male"` = 1
- Let's make this more explicit by making a **dummy variable** to stand in for Gender



Regression with Dummy Variables

Comparing Groups with Regression

- In a regression, we can easily compare across groups via a **dummy variable**[†]
- Dummy variable *only* $=0$ or $=1$, if a condition is `TRUE` vs. `FALSE`
- Signifies whether an observation belongs to a category or not

Example:

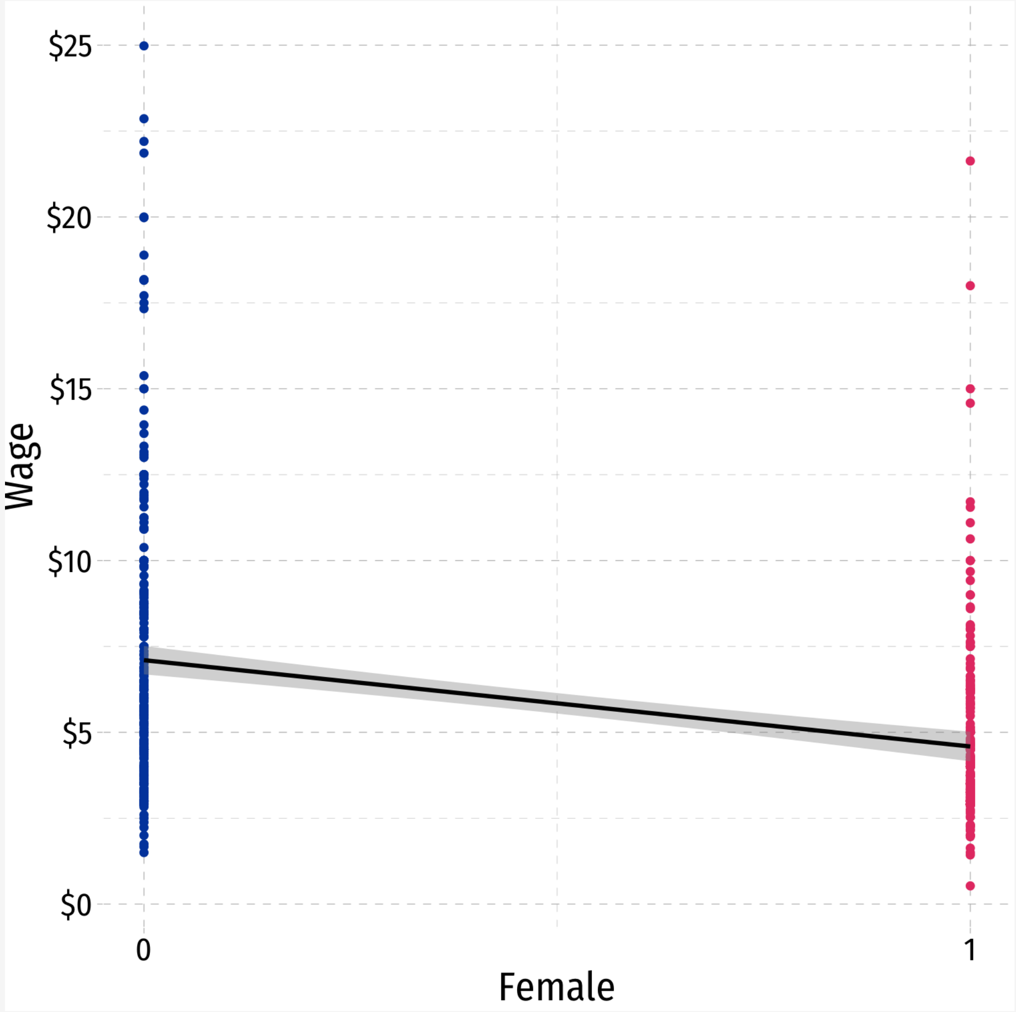
$$\widehat{Wage_i} = \hat{\beta}_0 + \hat{\beta}_1 Female_i \qquad \text{where } Female_i = \begin{cases} 1 & \text{if individual } i \text{ is } Female \\ 0 & \text{if individual } i \text{ is } Male \end{cases}$$

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

[†] Also called a **binary variable** or **dichotomous variable**

Comparing Groups in Regression: Scatterplot

- `Female` is our dummy x-variable
- Hard to see relationships because of **overplotting**



Comparing Groups in Regression: Scatterplot

- `Female` is our dummy x-variable
- Hard to see relationships because of **overplotting**
- Use `geom_jitter()` instead of `geom_point()` to *randomly* nudge points
  - *Only* for plotting purposes, does not affect actual data, regression, etc.

Comparing Groups in Regression: Scatterplot

- `Female` is our dummy x-variable
- Hard to see relationships because of **overplotting**
- Use `geom_jitter()` instead of `geom_point()` to *randomly* nudge points
    - *Only* for plotting purposes, does not affect actual data, regression, etc.

Dummy Variables as Group Means

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

- 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:

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

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

$$E[Wage|Female = 0] = \hat{\beta}_0$$

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

$$E[Wage|Female = 0] = \hat{\beta}_0$$

- Mean wage for women:

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

$$E[Wage|Female = 0] = \hat{\beta}_0$$

- Mean wage for women:

$$E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$$

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

$$E[Wage|Female = 0] = \hat{\beta}_0$$

- Mean wage for women:

$$E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$$

- Difference in wage between men & women:

Dummy Variables as Group Means: Our Example

Example:

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

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

- Mean wage for men:

$$E[Wage|Female = 0] = \hat{\beta}_0$$

- Mean wage for women:

$$E[Wage|Female = 1] = \hat{\beta}_0 + \hat{\beta}_1$$

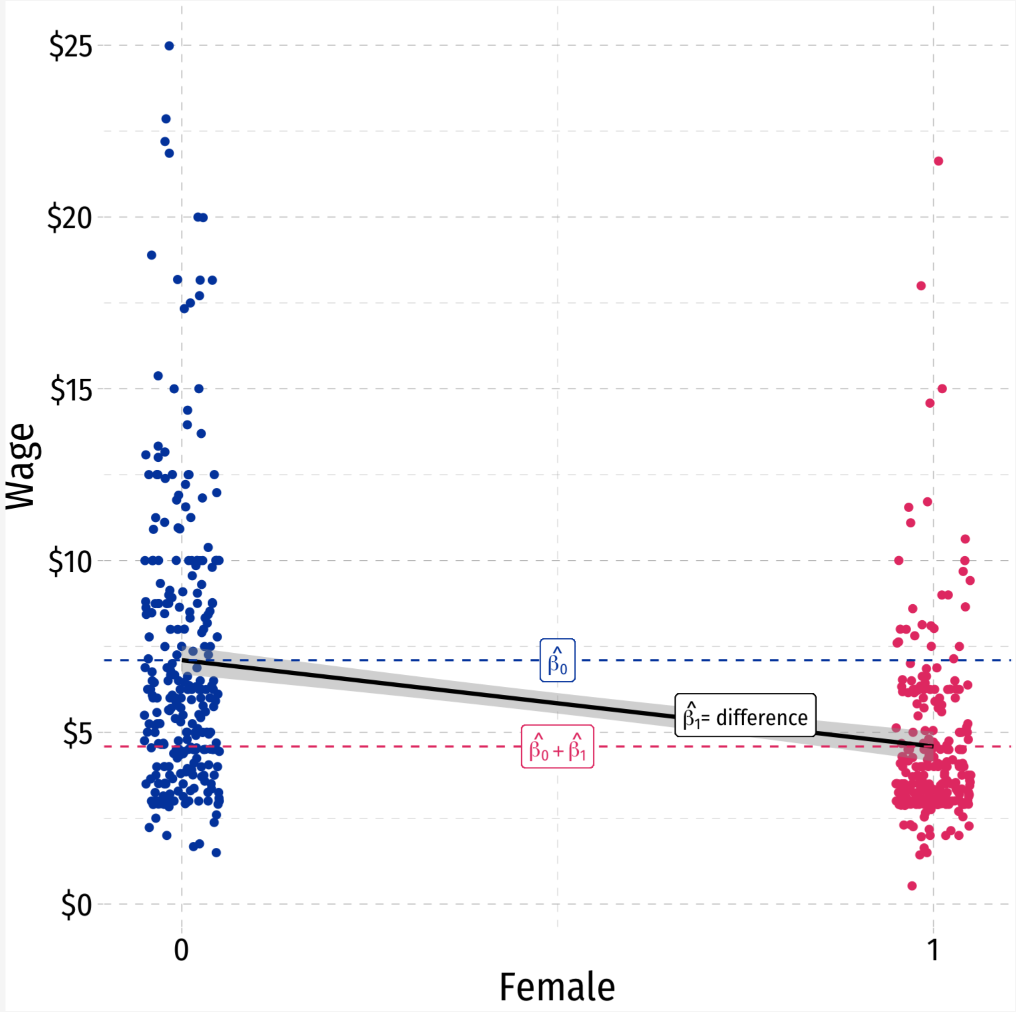
- Difference in wage between men & women:

$$d = \hat{\beta}_1$$

Comparing Groups in Regression: Scatterplot

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

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



The Data

```r
# from wooldridge package
library(wooldridge)

# save as a dataframe
wages<-wooldridge::wage1
```

```r
wages
```

wage	educ	exper	tenure	nonwhite	female	married	numdep	smsa	northcen
<dbl>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
3.10	11	2	0	0	1	0	2	1	0
3.24	12	22	2	0	1	1	3	1	0
3.00	11	2	0	0	0	0	2	0	0
6.00	8	44	28	0	0	1	0	1	0
5.30	12	7	2	0	0	1	1	0	0
8.75	16	9	8	0	0	1	0	1	0
11.25	18	15	7	0	0	0	0	1	0
5.00	12	5	3	0	1	0	0	1	0
3.60	12	26	4	0	1	0	2	1	0

Get Group Averages & Std. Devs.

```
# Summarize for Men

wages %>%
  filter(female==0) %>%
  summarize(mean = mean(wage),
            sd = sd(wage))
```

mean	sd
<dbl>	<dbl>
7.099489	4.160858

1 row

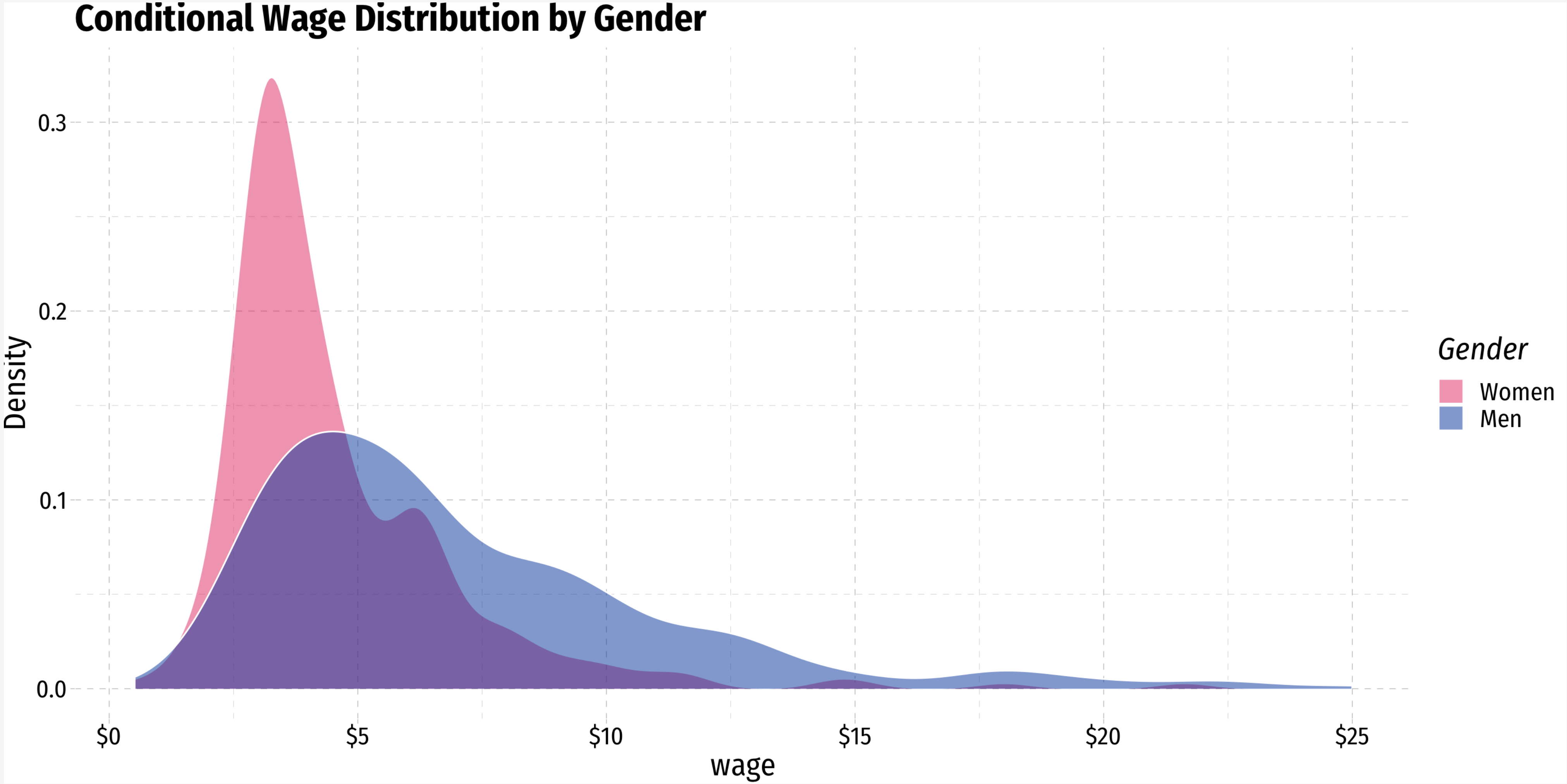
```
# Summarize for Women

wages %>%
  filter(female==1) %>%
  summarize(mean = mean(wage),
            sd = sd(wage))
```

mean	sd
<dbl>	<dbl>
4.587659	2.529363

1 row

Visualize Differences



The Regression I

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

```
##
## 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
```

```
library(broom)
tidy(femalereg)
```

term <chr>	estimate <dbl>	std.error <dbl>
(Intercept)	7.099489	0.2100082
female	-2.511830	0.3034092

2 rows | 1-3 of 5 columns

Dummy Regression vs. Group Means

From tabulation of group means

Gender	Avg. Wage	Std. Dev.	n
Female	4.59	2.33	252
Male	7.10	4.16	274
Difference	2.51	0.30	–

From t-test of difference in group means

term	estimate	std.error
<chr>	<dbl>	<dbl>
(Intercept)	7.099489	0.2100082
female	-2.511830	0.3034092

2 rows | 1-3 of 5 columns

$$\widehat{\text{Wages}}_i = 7.10 - 2.51\,\text{Female}_i$$

Recoding Dummies

Recoding Dummies

Example:

- Suppose instead of $female$ we had used:

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

Recoding Dummies with Data

```r
wages<-wages %>%
  mutate(male = ifelse(female == 0, # condition: is female equal to 0?
                       1, # if true: code as "1"
                       0)) # if false: code as "0"

# verify it worked
wages %>%
  select(wage, female, male) %>%
  head()
```

	wage	female	male
	<dbl>	<int>	<dbl>
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

6 rows

Scatterplot with Male

Dummy Variables as Group Means: With Male

Example:

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

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

- Mean wage for men:

$$E[Wage|Male = 1] = \hat{\beta}_0 + \hat{\beta}_1$$

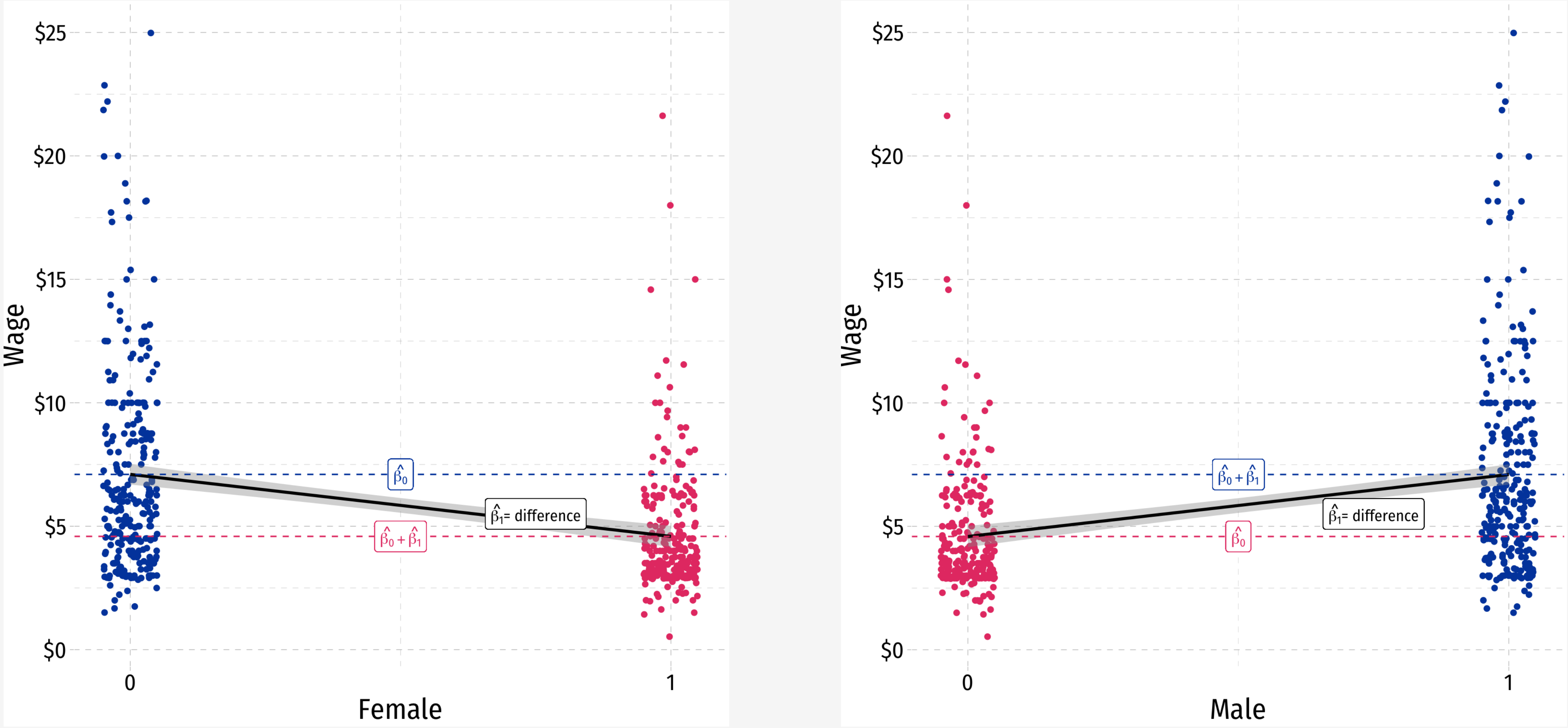
- Mean wage for women:

$$E[Wage|Male = 0] = \hat{\beta}_0$$

- Difference in wage between men & women:

$$d = \hat{\beta}_1$$

Scatterplot with Male



The Regression with Male I

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

```
##
## 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
```

```r
library(broom)
tidy(malereg)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>
(Intercept)	4.587659	0.2189834	20.949802
male	2.511830	0.3034092	8.278688

2 rows | 1-4 of 5 columns

The Dummy Regression: Male or Female

	(1)	(2)
Constant	4.59 ***	7.10 ***
	(0.22)	(0.21)
Female		-2.51 ***
		(0.30)
Male	2.51 ***	
	(0.30)	
N	526	526
R-Squared	0.12	0.12
SER	3.48	3.48

*** p < 0.001; ** p < 0.01; * p < 0.05.

- Note it doesn't matter if we use `male` or `female`, males always earn $2.51 more than females
- Compare the constant (average for the $D = 0$ group)
- Should you use `male` AND `female`? We'll come to that...

Categorical Variables (More than 2 Categories)

Categorical Variables with More than 2 Categories

- A **categorical variable** expresses membership in a category, where there is no ranking or hierarchy of the categories
  - We've looked at categorical variables with 2 categories only
  - e.g. Male/Female, Spring/Summer/Fall/Winter, Democratic/Republican/Independent

- Might be an **ordinal variable** expresses rank or an ordering of data, but not necessarily their relative magnitude
  - e.g. Order of finalists in a competition (1st, 2nd, 3rd)
  - e.g. Highest education attained (1=elementary school, 2=high school, 3=bachelor's degree, 4=graduate degree)

Using Categorical Variables in Regression I

Example: How do wages vary by region of the country? Let $Region_i = \{Northeast, Midwest, South, West\}$

- Can we run the following regression?

$$\widehat{Wages_i} = \hat{\beta_0} + \hat{\beta_1} Region_i$$

Using Categorical Variables in Regression II

Example: How do wages vary by region of the country?

Code region numerically:

$$Region_i = \begin{cases} 1 & \text{if } i \text{ is in } Northeast \\ 2 & \text{if } i \text{ is in } Midwest \\ 3 & \text{if } i \text{ is in } South \\ 4 & \text{if } i \text{ is in } West \end{cases}$$

- Can we run the following regression?

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Region_i$$

Using Categorical Variables in Regression III

Example: How do wages vary by region of the country?

Create a dummy variable for *each* region:

- $Northeast_i = 1$ if i is in Northeast, otherwise $= 0$
- $Midwest_i = 1$ if i is in Midwest, otherwise $= 0$
- $South_i = 1$ if i is in South, otherwise $= 0$
- $West_i = 1$ if i is in West, otherwise $= 0$

- Can we run the following regression?

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i + \hat{\beta}_4 West_i$$

- For every $i: \; Northeast_i + Midwest_i + South_i + West_i = 1$!

The Dummy Variable Trap

Example: $\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i + \hat{\beta}_4 West_i$

- If we include *all* possible categories, they are **perfectly multicollinear**, an exact linear function of one another:

$$Northeast_i + Midwest_i + South_i + West_i = 1 \quad \forall i$$

- This is known as the **dummy variable trap**, a common source of perfect multicollinearity

The Reference Category

- To avoid the dummy variable trap, always omit one category from the regression, known as the **"reference category"**

- It does not matter which category we omit!

- **Coefficients on each dummy variable measure the *difference* between the *reference* category and each category dummy**

The Reference Category: Example

Example: $\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i$

- $West_i$ is omitted (arbitrarily chosen)
- $\hat{\beta}_0$:

The Reference Category: Example

Example: $\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i$

- $West_i$ is omitted (arbitrarily chosen)
- $\hat{\beta}_0$: average wage for i in the West
- $\hat{\beta}_1$:

The Reference Category: Example

Example: $\widehat{Wages}_i = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i$

- $West_i$ is omitted (arbitrarily chosen)
- $\hat{\beta}_0$: average wage for i in the West
- $\hat{\beta}_1$: difference between West and Northeast
- $\hat{\beta}_2$:

The Reference Category: Example

Example: $\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i$

- $West_i$ is omitted (arbitrarily chosen)
- $\hat{\beta}_0$: average wage for i in the West
- $\hat{\beta}_1$: difference between West and Northeast
- $\hat{\beta}_2$: difference between West and Midwest
- $\hat{\beta}_3$:

The Reference Category: Example

Example: $\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Northeast_i + \hat{\beta}_2 Midwest_i + \hat{\beta}_3 South_i$

- $West_i$ is omitted (arbitrarily chosen)
- $\hat{\beta}_0$: average wage for i in the West
- $\hat{\beta}_1$: difference between West and Northeast
- $\hat{\beta}_2$: difference between West and Midwest
- $\hat{\beta}_3$: difference between West and South

Dummy Variable Trap in R

```
lm(wage ~ noreast + northcen + south + west, data = wages) %>% summary()
```

```
##
## Call:
## lm(formula = wage ~ noreast + northcen + south + west, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.083 -2.387 -1.097  1.157 18.610
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6134     0.3891  16.995  < 2e-16 ***
## noreast      -0.2436     0.5154  -0.473  0.63664
## northcen     -0.9029     0.5035  -1.793  0.07352 .
## south        -1.2265     0.4728  -2.594  0.00974 **
## west              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.671 on 522 degrees of freedom
## Multiple R-squared:  0.0175,    Adjusted R-squared:  0.01185
## F-statistic: 3.099 on 3 and 522 DF,  p-value: 0.02646
```

Using Different Reference Categories in R

```r
# let's run 4 regressions, each one we omit a different region
no_noreast_reg <- lm(wage ~ northcen + south + west, data = wages)
no_northcen_reg <- lm(wage ~ noreast + south + west, data = wages)
no_south_reg <- lm(wage ~ noreast + northcen + west, data = wages)
no_west_reg <- lm(wage ~ noreast + northcen + south, data = wages)

# now make an output table
library(huxtable)
huxreg(no_noreast_reg,
       no_northcen_reg,
       no_south_reg,
       no_west_reg,
       coefs = c("Constant" = "(Intercept)",
                 "Northeast" = "noreast",
                 "Midwest" = "northcen",
                 "South" = "south",
                 "West" = "west"),
       statistics = c("N" = "nobs",
                      "R-Squared" = "r.squared",
                      "SER" = "sigma"),
       number_format = 3)
```

Using Different Reference Categories in R II

	(1)	(2)	(3)	(4)
Constant	6.370 ***	5.710 ***	5.387 ***	6.613 ***
	(0.338)	(0.320)	(0.268)	(0.389)
Northeast		0.659	0.983 *	-0.244
		(0.465)	(0.432)	(0.515)
Midwest	-0.659		0.324	-0.903
	(0.465)		(0.417)	(0.504)
South	-0.983 *	-0.324		-1.226 **
	(0.432)	(0.417)		(0.473)
West	0.244	0.903	1.226 **	
	(0.515)	(0.504)	(0.473)	
N	526	526	526	526
R-Squared	0.017	0.017	0.017	0.017
SER	3.671	3.671	3.671	3.671

- Constant is alsways average wage for reference (omitted) region
- Compare coefficients between Midwest in (1) and Northeast in (2)...
- Compare coefficients between West in (3) and South in (4)...
- Does not matter which region we omit!
  - Same R^2, SER, coefficients give same results

Dummy *Dependent* (Y) Variables

- In many contexts, we will want to have our *dependent* (Y) variable be a dummy variable

Example:

$$\widehat{Admitted}_i = \hat{\beta}_0 + \hat{\beta}_1 GPA_i \quad \text{where } Admitted_i = \begin{cases} 1 & \text{if } i \text{ is Admitted} \\ 0 & \text{if } i \text{ is Not Admitted} \end{cases}$$

- A model where Y is a dummy is called a **linear probability model**, as it measures the **probability of Y occuring $(=1)$ given the X's, i.e. $P(Y_i = 1 | X_1, \cdots, X_k)$**
    - e.g. the probability person i is Admitted to a program with a given GPA
- Requires special tools to properly interpret and extend this (**logit**, **probit**, etc)
- Feel free to write papers that have dummy Y variables (but you may have to ask me some more questions)!