Advanced Data Analysis I

Interactions and Qualitative Predictors

PA 541 Week 7

Michael D. Siciliano

Department of Public Administration

College of Urban Planning and Public Affairs

Admin Stuff

- **Optional Lab** on the 26th at 3pm.
- **Homework 2**: Due on March 1st at 3pm.
- On the horizon:
  - **Week 8**: Non-linear relationships. Checking the validity of regressions assumptions.
  - **Week 9**: Midterm
  - **Week 10**: Model specification and data issues. Log Models. Data screening and cleaning. Methods for handling outliers and missing data.
  - **Week 11**: Spring Break

This week's lecture

- Introduction to interactions and ways to visualize interaction effects
  - Note: I will briefly review quadratics next week and log models in week 10.
- Regression modeling and interpretation with qualitative predictors
- Interpreting interactions with qualitative predictors

Wooldridge also discussed binary dependent variables in ch. 7 – this is a topic we will tackle under generalized linear models and logistic regression after the midterm.

MODEL INTERPRETATION AND INTERACTION EFFECTS

Start with a simple model

- Let's start with just a null model, a model with no predictors.

```
lm(formula = wage ~ 1, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3714 -2.5714 -1.2514  0.9786 19.0786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9014     0.1612    36.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.695 on 524 degrees of freedom
```

How do we interpret the intercept?

No R-squared is printed here, why? What do you suspect the value to be?

- Let's add a single predictor variable to examine the main effect of education on wages.

```
Call:
lm(formula = wage ~ educ, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3392 -2.1550 -0.9738  1.1950 16.6044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8916     0.6858   -1.30    0.194
educ          0.5406     0.0533   10.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.381 on 523 degrees of freedom
Multiple R-squared:  0.1644,   Adjusted R-squared:  0.1628
F-statistic: 102.9 on 1 and 523 DF,  p-value: < 2.2e-16
```

How do we interpret the intercept now? Why is it negative?

How do we interpret education?

- Let's also look at the main effect of experience, shown here in model 3.

	Model 1	Model 2	Model 3
(Intercept)	5.901 (0.161)***	-0.892 (0.686)	-3.380 (0.769)***
educ		0.541 (0.053)***	0.644 (0.054)***
exper			0.070 (0.011)***
R^2	0.000	0.164	0.224
Adj. R^2	0.000	0.163	0.221
Num. obs.	525	525	525

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

How do we interpret the intercept in model 3?
Why did the effect of education change?
How much of an increase in explanatory power did adding experience have?

Centering variables

- Given the lack of an interpretable intercept, let's center our two predictor variables.
- For each variable, what makes sense as a centering point?
- Let's construct two new variables and rerun the models:

```
> wage1$educ12 = wage1$educ - 12
> wage1$exper10 = wage1$exper - 10
> head(wage1)
   wage educ exper tenure nonwhite white female
1  3.24   12    22      2        0     1      1
2  3.00   11     2      0        0     1      0
3  6.00    8    44     28        0     1      0
4  5.30   12     7      2        0     1      0
5  8.75   16     9      8        0     1      0
6 11.25   18    15      7        0     1      0
  tenursq      RES_1   resid_sq educ12 exper10
1       4 -2.3556033 5.54886673      0      12
2       0 -2.0550030 4.22303720     -1      -8
3     784  2.5667979 6.58845152     -4      34
4       4 -0.2956033 0.08738129      0      -3
5      64  0.9919956 0.98405520      4      -1
```

	Model 1	Model 2	Model 3
(Intercept)	5.901 (0.161)***	5.596 (0.151)***	5.044 (0.169)***
educ12		0.541 (0.053)***	0.644 (0.054)***
exper10			0.070 (0.011)***
R^2	0.000	0.164	0.224
Adj. R^2	0.000	0.163	0.221
Num. obs.	525	525	525

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

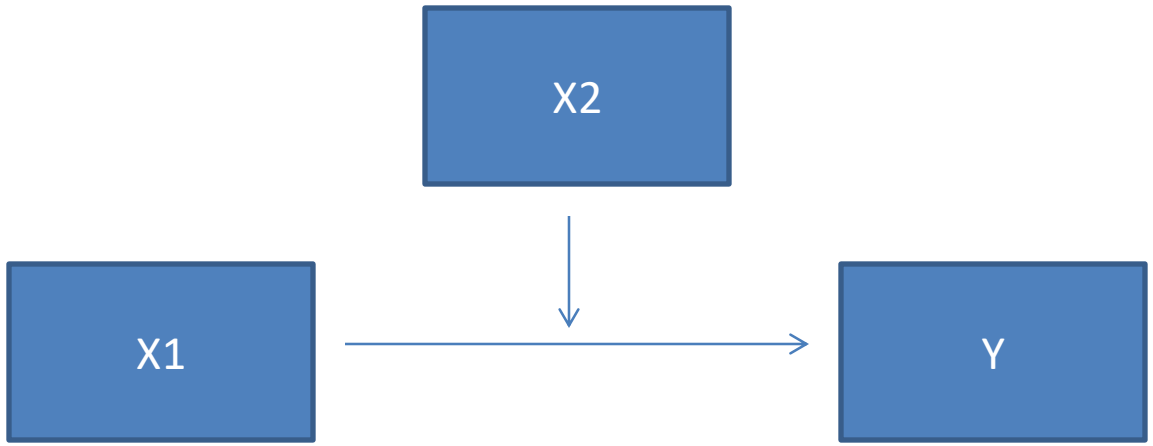
- Why didn't the slope coefficients change for education and experience?
- How do we interpret the intercept now?

Interaction terms

- In the previous examples, we were only looking at the main effects of each of our predictors.
- This implicitly assumes that the effect of education on wages is the same across all levels of experience (or vice versa).
- But what if the effect of education varies at different levels of experience?
- This idea that the effect of one predictor may be dependent on another predictor is often termed moderation and can be explored through interaction terms.

Examples of interactions in your area of study? Or one's you have read about in other research?

- Moderator: name for a variable that moderates or changes the effect of another variable
  - Paul Johnson calls this a ‘Moderated Slope Model’



An interaction effect is said to exist when the effect of the independent variable on the dependent variable differs depending on the value of a third variable, called the moderator variable (Jaccard and Turrisi, 2003).

- Let’s write out our standard regression model with two variables.

$$Y_i = B_0 + B_1X1_i + B_2X2_i + e_i$$

- Now add an interaction.

$$Y_i = B_0 + B_1X1_i + B_2X2_i + B_3X1_iX2_i + e_i$$

- Let’s regroup the terms and we see X1’s slope depends on X2

$$Y_i = B_0 + (B_1 + B_3X2_i)X1_i + B_2X2_i + e_i$$

$$Y_i = B_0 + B_1X1_i + B_2X2_i + B_3X1_iX2_i + e_i$$

- B_3 shows us the interaction effect

- B_1 and B_2 often referred to as the simple main effects or the conditional main effects of x1 and x2.
  - This term is used because the effects of say x1 is only interpretable when the value of x2 = 0. Thus, we see another reason centering is important, especially with interaction effects.

Let's add an interaction term to our model

- Note: I'm using tenure now instead of experience as the interaction effects were more pronounced. For this example, let's assume education is the moderator.

```
lm(formula = wage ~ educ12 + tenure + educ12 * tenure, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1301 -1.7374 -0.6706  1.2572 14.7244

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.661486   0.167150  27.888  < 2e-16 ***
educ12        0.425891   0.061113   6.969 9.69e-12 ***
tenure        0.187605   0.018501  10.140  < 2e-16 ***
educ12:tenure 0.022542   0.005919   3.808 0.000157 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.055 on 521 degrees of freedom
Multiple R-squared:  0.3201,   Adjusted R-squared:  0.3162
F-statistic: 81.78 on 3 and 521 DF,  p-value: < 2.2e-16
```

How do we interpret each of the parameters?

- As noted by Hoffman (2015), some textbooks have suggested the main effects should not be interpreted when they are included in an interaction.
- However, this is not true, they can and should be interpreted, especially because they were included in an interaction.
- The trick is to interpret the main effects correctly and the correct way is to do so conditionally on their interacting predictor.
- Thus, the effects are simple (conditional) main effects that apply only when the interacting predictor is 0.
  - This is why we need/want to have a meaningful zero value for our predictors.

$$Y_i = B_0 + B_1X1_i + B_2X2_i + B_3X1_iX2_i + e_i$$

$$Y_i = B_0 + (B_1 + B_3X2_i)X1_i + B_2X2_i + e_i$$

Recall, we can regroup our variables to show that the effect of X1 on Y depends on the value of X2. In the regrouped from, the simple main effect is easier to see.

- **What is the effect of tenure at the following levels of education**
  - **High School Only**
  - **College Degree**

- To answer this question, we first need to specify the value of education for which we want understand the relationship between tenure and wages.

- Let's look high school only and add it into our equation:
  - Wages = 4.66 + .43*educ12* + .19*tenure* + .02*edu12*tenure*
  - Wages = 4.66 + .43(0) + .19*tenure* + .02(0)*tenure*
  - Wages = 4.66 + ***.19tenure***

- What about for those with a college degree
  - Wages = 4.66 + .43*educ12* + .19*tenure* + .02*educ12*tenure*
  - Wages = 4.66 + .43(4) + .19*tenure* + .02(4)*tenure*
  - Wages = 6.38 + ***.27tenure***

- It is clear here that education level moderates the size of the tenure effect. Tenure has a larger impact on wages for those with higher levels of education.

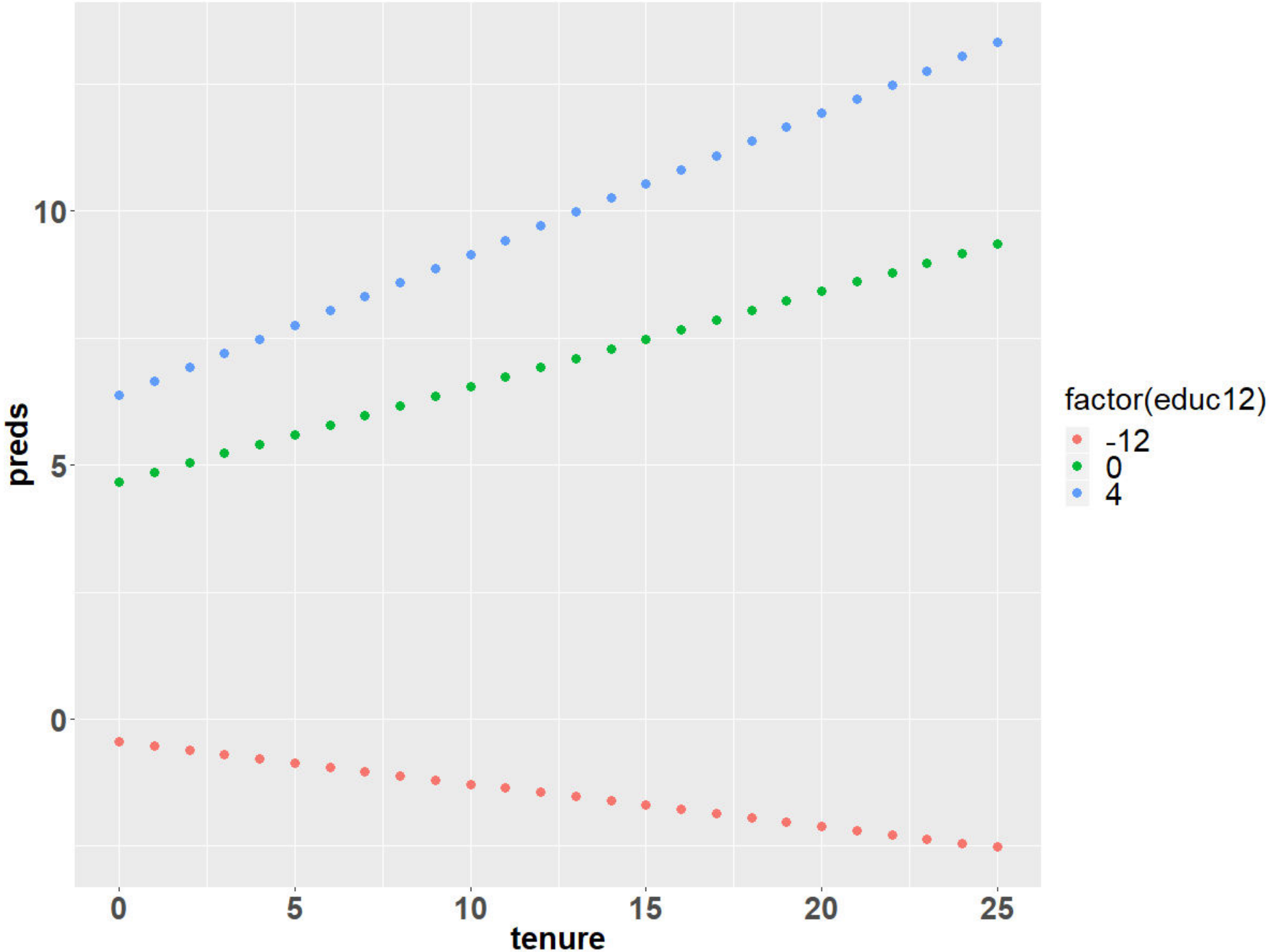
Graphing interactions

- It is often useful, for both you and your audience, to see a graphical display of interaction effects.
- How might you do this for a continuous by continuous interaction, such as the one we explored above?
- Can you draw a diagram of what the output may look like?

- One approach, which we will use, is to plot three regression lines for the regression of wages on tenure at three levels of education (low, medium, high).
- Choosing the levels is up to you. One useful approach is to use the mean as well as 1 standard deviation above and below the mean. But since we are working with education it may make more sense to look at those with no education, a high school degree, and a college degree.

- **Here's one approach for doing this in R:**

```
> newdata2 = expand.grid(educ12= c(-12, 0, 4),
+                        tenure = c(1:25) )
> #expand grid is nice function to create a new dataset.  What we have done is
> #develop a dataset for the particular set of individuals we want to examine
> #the effect of tenure on given the interaction in the model.
> head(newdata2)
  educ12 tenure
1    -12      1
2      0      1
3      4      1
4    -12      2
5      0      2
6      4      2
>
>
> newdata2$preds = predict(mod4, new=newdata2, type="response")
>
> ggplot(newdata2, aes(x=tenure, y=preds, group=educ12)) +
+   geom_point(aes(colour=as.factor(educ12)))
```


10
5
0
preds
0
5
10
15
20
25
tenure
factor(educ12)
-12
0
4

ggeffects Package

- In the script for today, I also show you how to use the ggeffects package for making effects plots.

- Basically same approach, but helpful when you have more complex models, need to control for other variables, etc…

QUALITATIVE PREDICTORS

Incorporating Qualitative Data

- Up until this point our work with regression has solely focused on continuous variables as IVs and DVs. For instance, we used a CEO's tenure to predict salary.
- Our analyses have been constrained by our use of continuous variables and many real-world applications will involve discrete variables.
- Regression can easily incorporate discrete or categorical values as regressors (i.e. ordinal and nominal level data).

For example…

- You may want to know if there are significant differences in earnings based on gender?
- You may want to know if jail sentences for drug addicts differ by race?
- You may want to know if organizations in different sectors are more likely to engage in collaborative relationships.
- You may want to know if public service motivation differs by level of government.

- In each of the examples above, the key independent variable can only take on only a limited number of values.

Let's take the first example on gender and earnings

- Individuals in our dataset will be placed into two categories: male or female (data was collected prior to considerations of gender identity).
- If we created a variable that only took on these two values in a regression, it is termed a **dummy variable**. Dummy variables are also known as indicator variables, binary variables, dichotomous variables, or qualitative variables.

Descriptive Stats

```
> summary(wages$wage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.530   3.330   4.650   5.901   6.880  24.980
> summary(wages$Male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.5219  1.0000  1.0000
```

- How do we interpret the min and max values for earnings and male?
- How do we interpret the mean?

Consider the following model:

$$y_i = \alpha_0 + \alpha_1 D_i + \varepsilon_i$$

- Where:
  - y_i = hourly wage of a worker
  - D_i = 1 if the worker is a man
  - D_i = 0 otherwise (worker is a woman)
- To understand what the model tells us we need to examine the expected value of hourly wages conditional on the value of D_i.

- Mean earnings for women: $E(y_i|D=0) = \alpha_0 + \alpha_1 D_i$
  $= \alpha_0$

- Mean earnings for men: $E(y_i|D=1) = \alpha_0 + \alpha_1 D_i$
  $= \alpha_0 + \alpha_1$

Interpreting the model

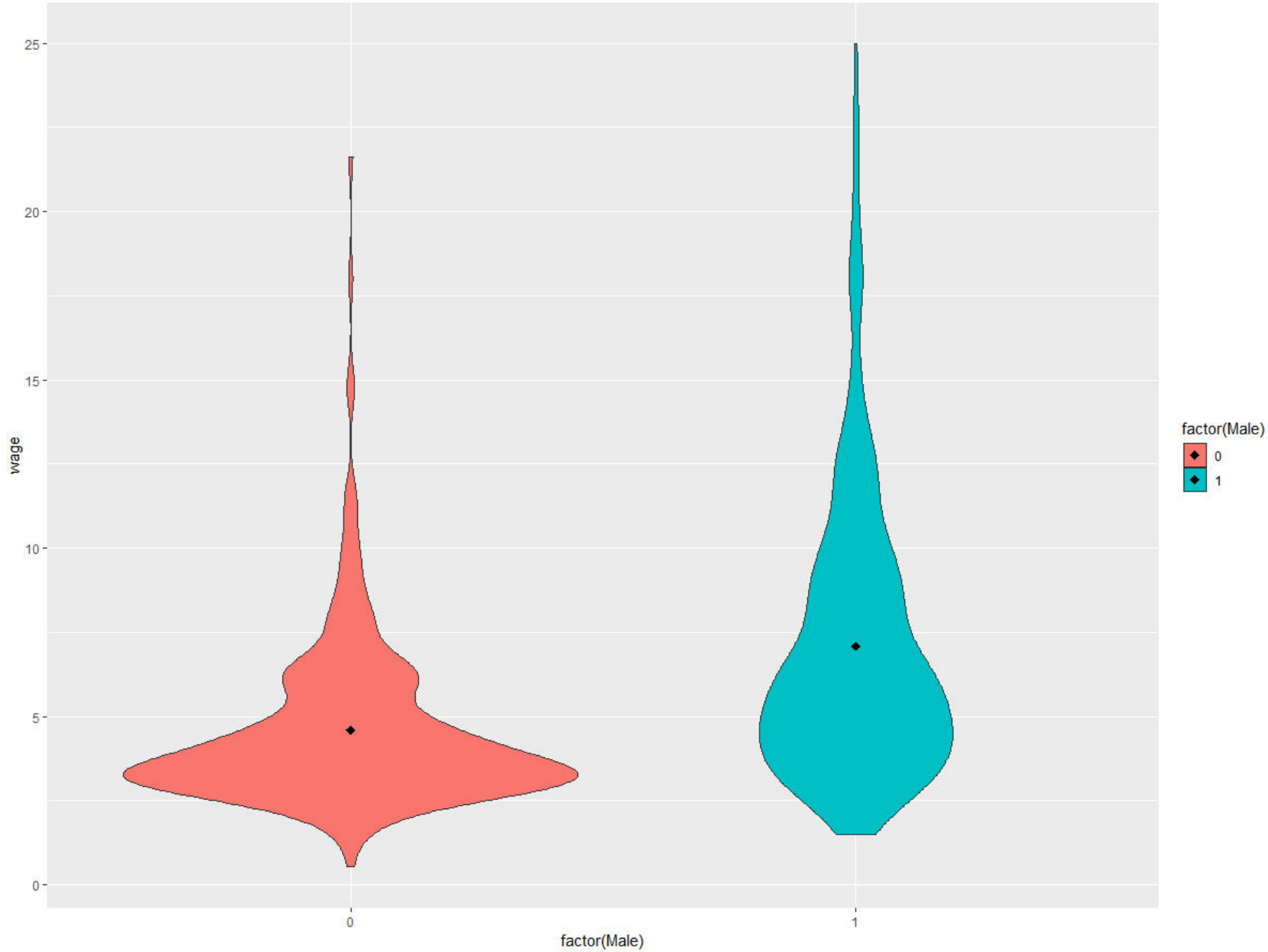
$$y_i = \alpha_0 + \alpha_1 D_i + \varepsilon_i$$

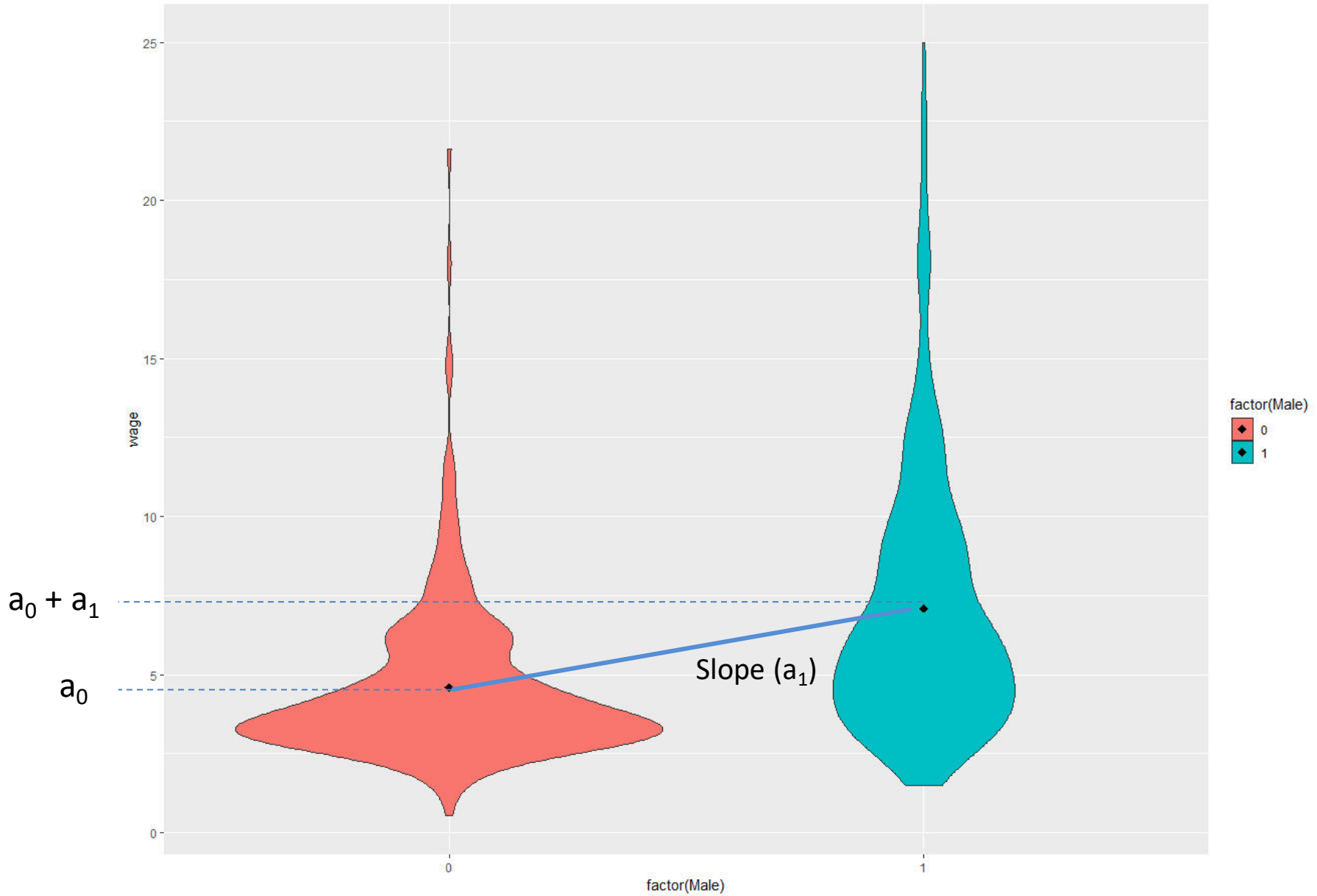
- The intercept term, α_0, is the mean hourly earnings of female workers.
- The slope term, α_1, is the difference between the mean hourly wage of men and the mean hourly wage of women.
- Thus, the sum of $\alpha_0 + \alpha_1$ is the mean hourly wage of male workers.
- A test for significance of the difference between male and female wages is a test of the null hypothesis that $\alpha_1 = 0$.

Regression Output

```
> wreg1 = lm(wages$wage ~ Male, data=wages)
> summary(wreg1)

Call:
lm(formula = wages$wage ~ Male, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5995 -1.8495 -0.9936  1.4305 17.8805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5936     0.2196  20.919  < 2e-16 ***
Male          2.5059     0.3040   8.244 1.35e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.479 on 523 degrees of freedom
Multiple R-squared: 0.115,    Adjusted R-squared: 0.1133
F-statistic: 67.97 on 1 and 523 DF,  p-value: 1.351e-15
```

- What is the estimated earnings for female workers? For male workers? Is the difference significant?
- Also, it is important to consider what dummy variables do not tell us. They do not tell us the reasons for the differences we find.


25
20
15
10
5
0
wage
0
1
factor(Male)
factor(Male)
0
1


25
20
15
10
5
0
wage
0
1
factor(Male)
factor(Male)
0
1
$a_0 + a_1$
a_0
Slope (a_1)

What happens if we switch the dummy variable to indicate female?

```
> wreg1 = lm(wages$wage ~ female, data=wages)
> summary(wreg1)

Call:
lm(formula = wages$wage ~ female, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5995 -1.8495 -0.9936  1.4305 17.8805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.0995     0.2102  33.779  < 2e-16 ***
female       -2.5059     0.3040  -8.244 1.35e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.479 on 523 degrees of freedom
Multiple R-squared: 0.115,    Adjusted R-squared: 0.1133
F-statistic: 67.97 on 1 and 523 DF,  p-value: 1.351e-15
```

Base Category

- When creating dummy variables for gender, the category that is assigned the value of zero is termed the base category (also termed the omitted or reference category).
- Any interpretation of the meaning of the coefficient on the dummy variable in the regression must be done relative to the base category. So, choose your base category wisely, especially when the you have more than two categories.
  - If you make a new variable for each category, you just exclude the variable you want as the base from your model.
    - Ex. For gender, you would make a single indicator for male.
  - If you are using factors, then you can set your base as follows: relevel("varname", ref = "base")

Adding a continuous variable to our model – years of experience

$$y_i = \alpha_0 + \alpha_1 D_i + \beta x_i + \varepsilon_i$$

- How do we interpret this model?
  - Mean earnings of women: $E(y_i|x_i, D=0) = \alpha_0 + \beta x_i$
  - Mean earnings of men: $E(y_i|x_i, D=1) = (\alpha_0 + \alpha_1) + \beta x_i$
- So, just as before, the average difference in earnings of men and women is α_1.
- Also, note that the slope of years of work experience is the same for both males and females. (The model specifies that it is).

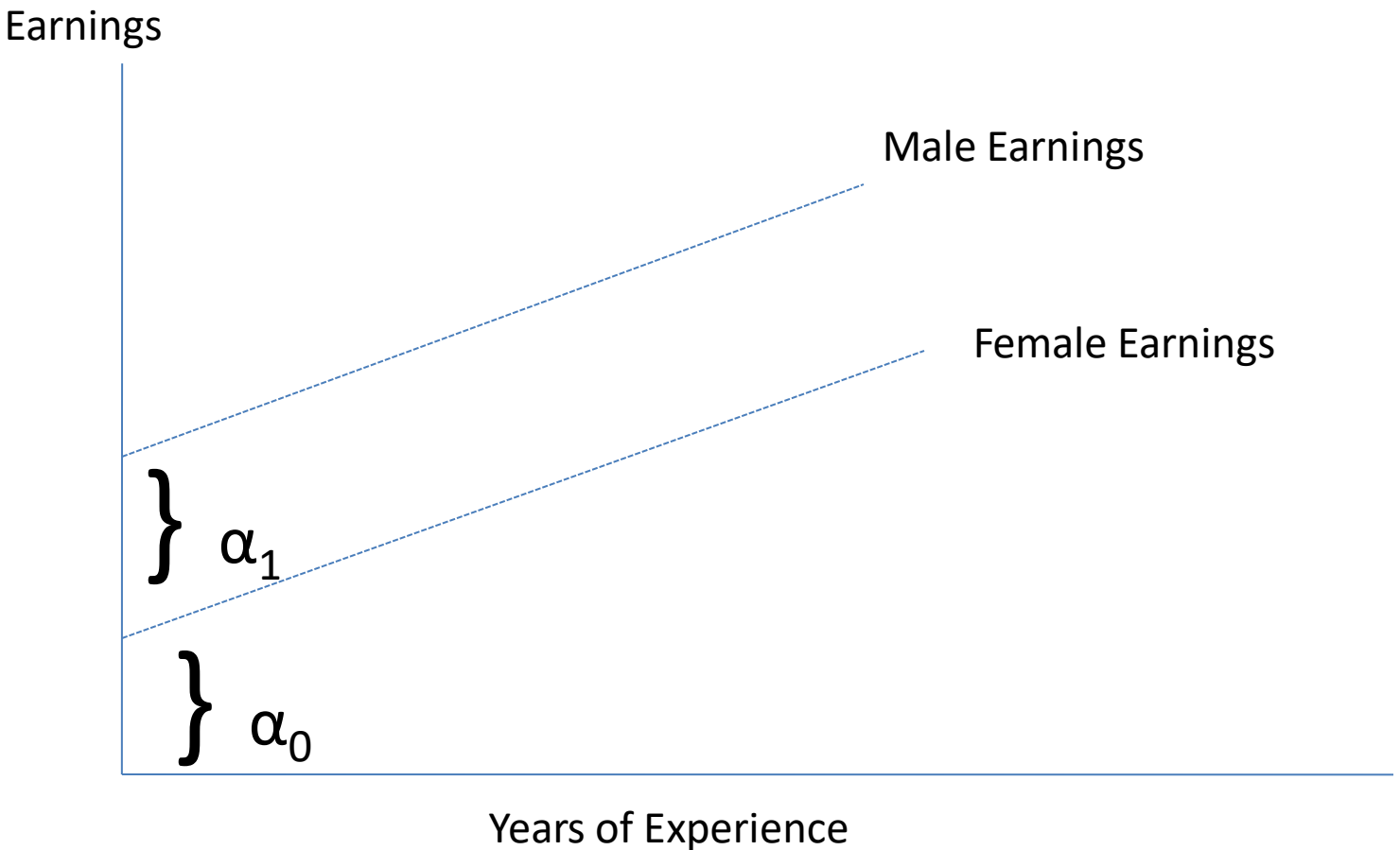
Running the regression

```
> wreg2 = lm(wages$wage ~ Male + exper, data=wages)
> summary(wreg2)

Call:
lm(formula = wages$wage ~ Male + exper, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9591 -2.0354 -0.9003  1.4554 17.5744

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.15257    0.28571  14.534  < 2e-16 ***
Male         2.47722    0.30283   8.180 2.17e-15 ***
exper        0.02675    0.01116   2.397   0.0169 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret the intercept? How do we interpret the coefficient on male?
- Is there a significant difference between male and female earnings when controlling for years of experience?

Looking at this relationship graphically

- Because α_1 does not equal zero, there is a difference in the intercept term between males and females. The increase in earnings based on experience is the same for both male and female and so the lines are parallel.



The Dummy Variable Trap

- The process to create dummy variables in R is fairly easy. We simply take the categorical variable and set it to a factor. R will automatically incorporate the factor into your model and create the appropriate number of dummy variables.
  - You could also do this by hand and create a new binary variable for each category that indicates if that observation is of that type. The examples we use follow this approach as it is more transparent when first learning to use qualitative variables.
  - Categories for dummy variables should be mutually exclusive

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- The dummy variable trap occurs when we attempt to run a regression with both a dummy variable indicating *female* and a dummy variable indicating *male*. While this may make sense at first, it creates a situation of exact multicollinearity. For example, if D1 is a male indicator and D2 is female indicator:

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

  - Then for every observation in the dataset: $D_{2i} = 1 - D_{1i}$
- As a general rule of thumb, if a dummy variable has m categories, you can have at most m-1 dummy variables in your regression. This forces one category to be the base category. In our previous example, female was the base category (more on this later).

Running a regression with dummy variable trap

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

Where D_1 is an indicator for male and D_2 is an indicator for female

```
> wreg3 = lm(wages$wage ~ Male + female, data=wages)
> summary(wreg3)

Call:
lm(formula = wages$wage ~ Male + female, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5995 -1.8495 -0.9936  1.4305 17.8805

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.5936     0.2196  20.919  < 2e-16 ***
Male          2.5059     0.3040   8.244 1.35e-15 ***
female            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R drops one of the regressors and runs the model with only one dummy variable

Incorporating Qualitative Variables with More than Two Categories

- Many categorical variables have more than just two categories.
- Consider our earnings example; suppose we wanted to split observations based on region in the United States.
- The impact of region on hourly earnings could be modeled as.

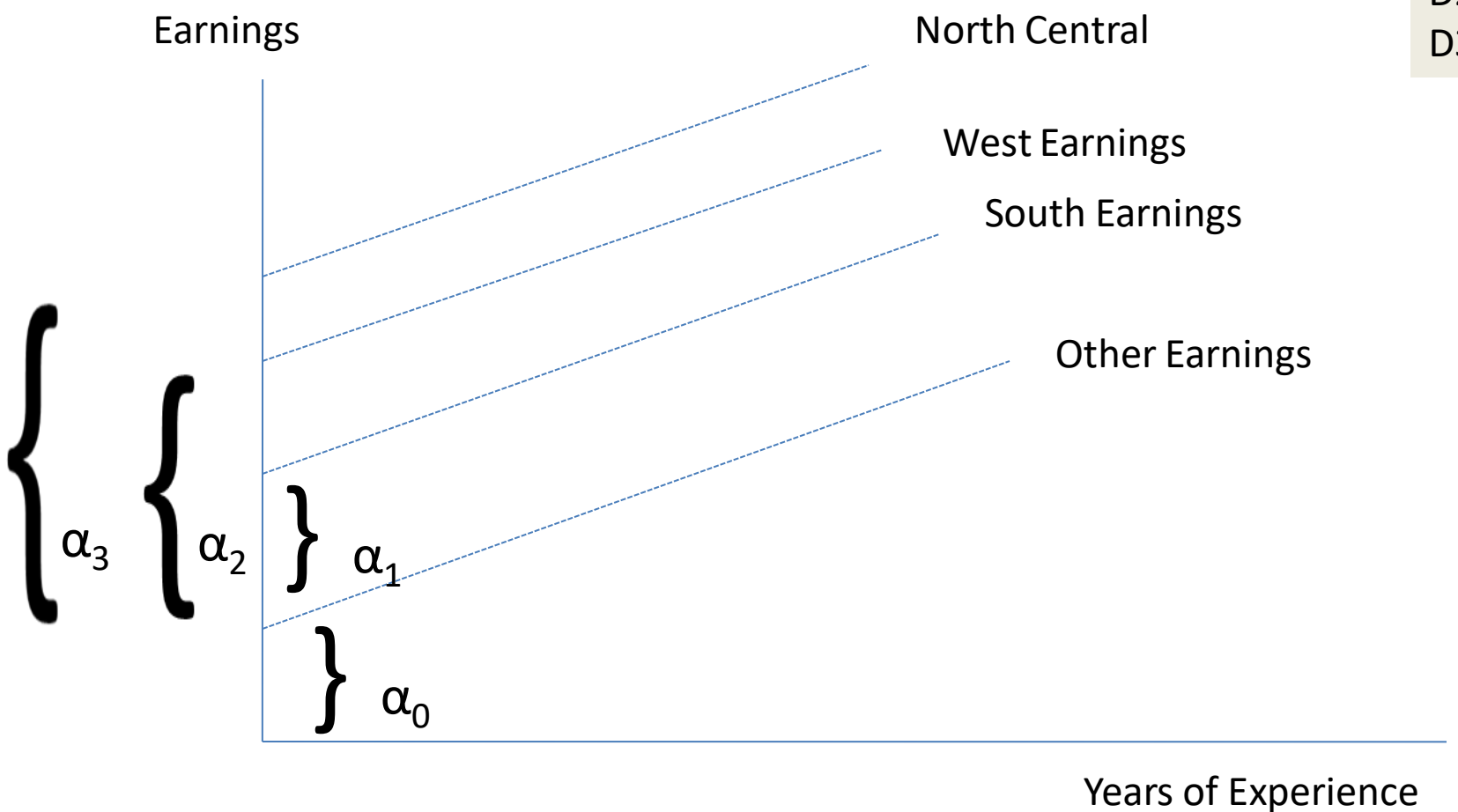
$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta x_i + \varepsilon_i$$

- Where:
  - y_i = average hourly wages
  - X_i = years of work experience
  - D_{1i} = 1 if lives in the south, 0 otherwise
  - D_{2i} = 1 if lives in west, 0 otherwise
  - D_{3i} = 1 if lives in north-central, 0 otherwise
  - Other is the base category (i.e. north-east)
- Note that we had 4 region categories, so we included 3 dummy variables

Visual Representation of Region Model

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta x_i + \varepsilon_i$$

D1i = 1 if lives in the south
D2i = 1 if lives in west
D3i = 1 if lives in north-central



Visual not matched to actual regression results.

Expected Values of Region Model

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta x_i + \varepsilon_i$$

D1i = 1 if lives in the south
D2i = 1 if lives in west
D3i = 1 if lives in north-central

- Mean Hourly Wage of **other**:
  - $E(y_i | x_i, D_1=0, D_2=0, D_3=0) = \alpha_0 + \beta x_i$
- Mean Hourly Wage of **south**:
  - $E(y_i | x_i, D_1=1, D_2=0, D_3=0) = (\alpha_0 + \alpha_1) + \beta x_i$
- Mean Hourly Wage of **west**:
  - $E(y_i | x_i, D_1=0, D_2=1, D_3=0) = (\alpha_0 + \alpha_2) + \beta x_i$
- Mean Hourly Wage of **north-central**:
  - $E(y_i | x_i, D_1=0, D_2=0, D_3=1) = (\alpha_0 + \alpha_3) + \beta x_i$

Output from Region Model

```
> wreg4 = lm(wages$wage ~ south + west + northcen + exper, data=wages)
> summary(wreg4)

Call:
lm(formula = wages$wage ~ south + west + northcen + exper, data = wages)

Residuals:
   Min     1Q Median     3Q    Max
-5.680 -2.425 -1.097  1.253 18.177

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.86049    0.38320  15.293  < 2e-16 ***
south       -1.05619    0.42974  -2.458  0.01431 *
west         0.25168    0.51399   0.490  0.62458
northcen    -0.71085    0.46258  -1.537  0.12498
exper        0.03248    0.01177   2.760  0.00599 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret each parameter estimate? How do we interpret the hypothesis tests when we use a qualitative variable with more than two categories? In other words, what is the null hypothesis for ‘south’?

Quick Exercise: Create Model, Graph, and Expected Values

- Take out a piece of paper and write a model to predict hourly wages based on education (some highshcool, highschool grad, college) and experience (a continuous variable).
- Write the regression model, draw a graph representing your model (and the hypothesized effects), and then develop the expected value equations for each category.

Allowing Slopes to Vary by Group

- The previous slides have focused on allowing the intercept term to vary across different groups. This is interesting from both a policy and theoretical standpoint. But, as you may have noticed, we have forced the coefficients on our continuous variable to be the same for each group.
- Going back to our original gender model, we may want to test the hypothesis that not only does the intercept term differ by gender, but also the returns to earnings from work experience.

Varying Slopes Model

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_1 x_i + \beta_2 (D_i x_i) + \varepsilon_i$$

- Where:
  - y_i = hourly wage of a worker
  - D_i = 1 if the worker is a man; D_i = 0 otherwise (worker is a woman)
  - X_i = years of experience
- The term $D_i x_i$ means "D_i times X_i"; this is the interaction.
- If the worker is a women, this product is always zero.
- If the worker is a man, then $D_i = 1$, which means that $D_i x_i = x_i$. So basically, is there an additional effect of being a male on the slope of experience.

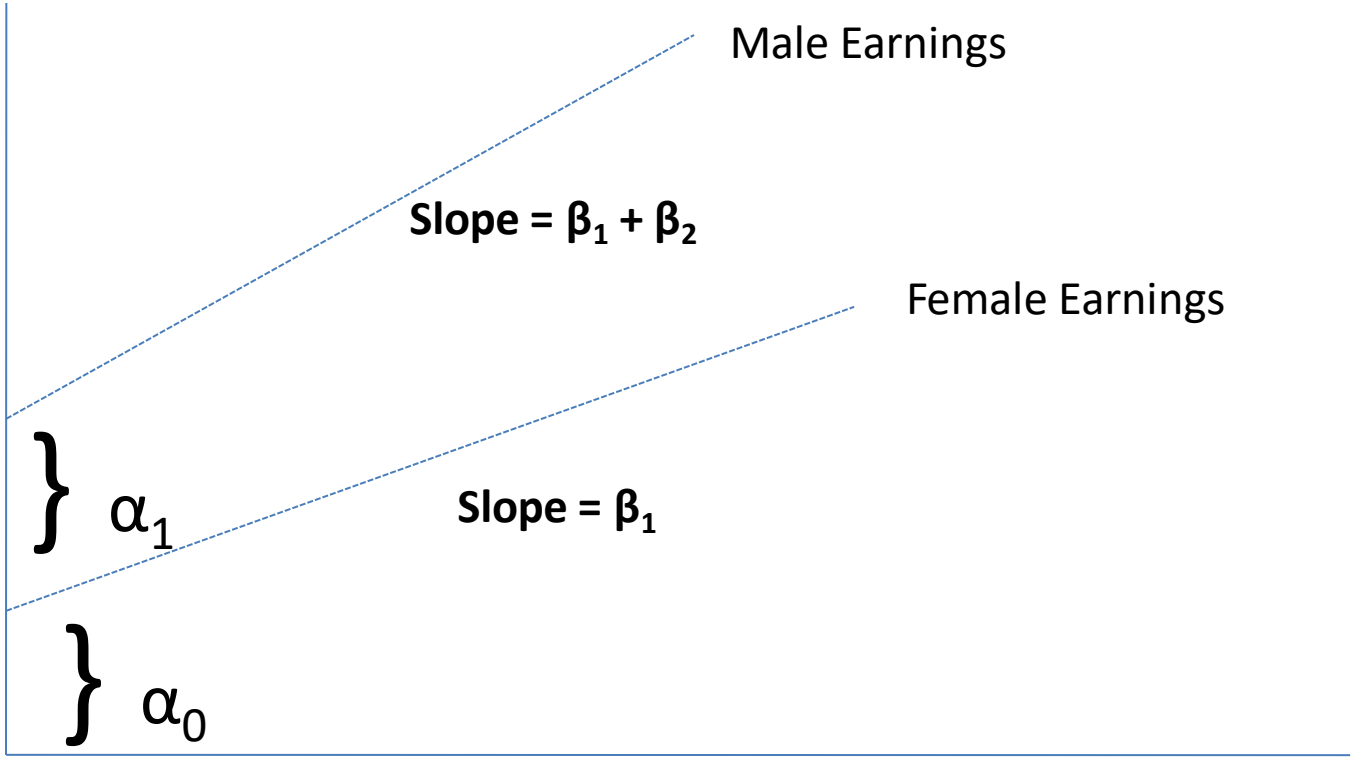
Varying Slopes Model Cont…

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_1 x_i + \beta_2 (D_i x_i) + \varepsilon_i$$

- What are the expected values for our new model and how does the interaction affect our interpretation?
- Mean Earnings of **Women**:
  - $E(y_i | x_i, D = 0) = \alpha_0 + \beta_1 x_i$
- Mean Earnings of **Men**:
  - $E(y_i | x_i, D=1) = (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2) x_i$
- In this model, the difference in intercept term between men and women is α_1.
- The difference in the slope of work experience between men and women is β_2.

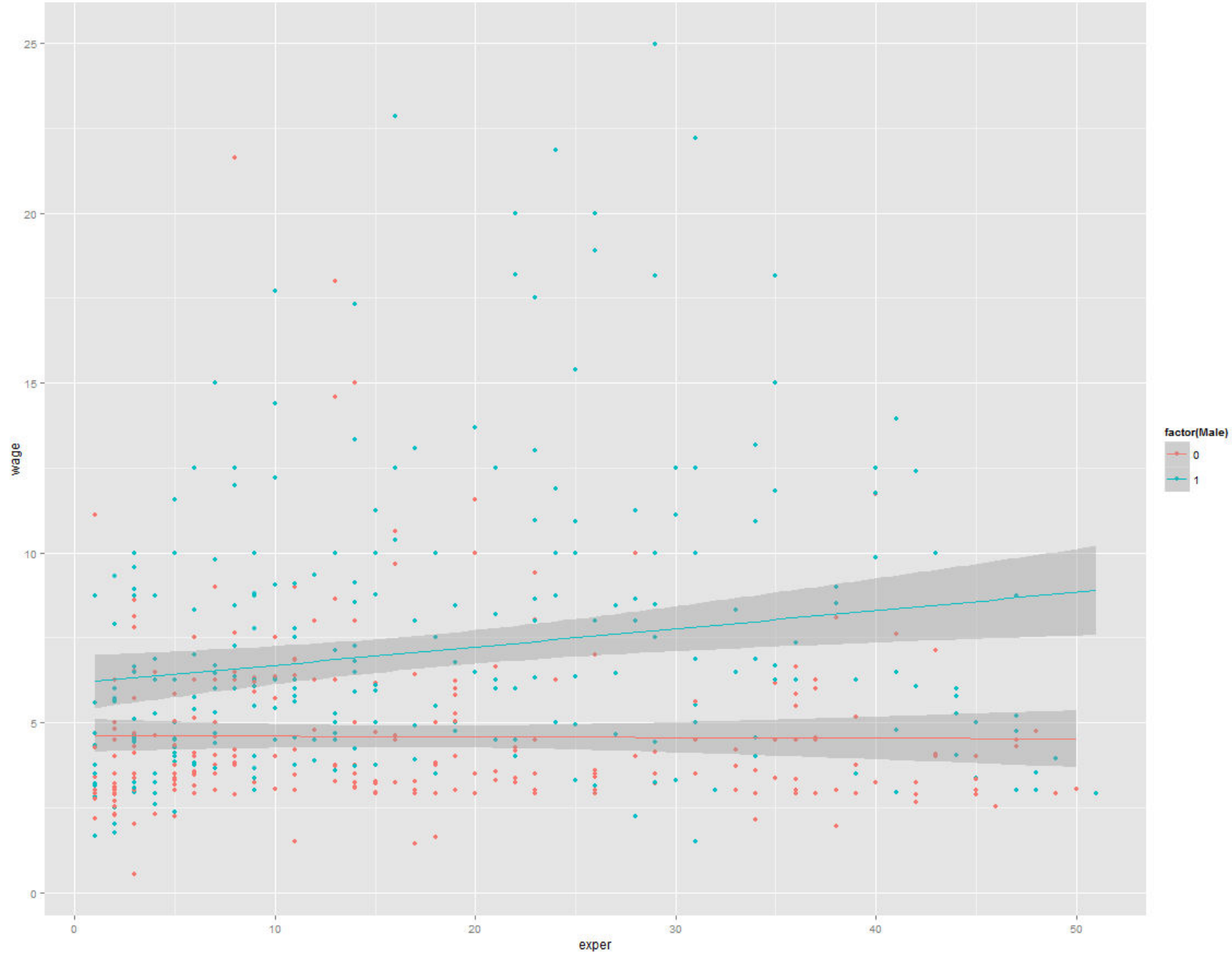
Graphic Representation of Model

- If both $\alpha_1 > 0$ and $\beta_2 > 0$ then our model could be visualized as below



Plotting the actual data:

```
ggplot(data=wages, aes(x=exper, y=wage, colour=factor(Male))) +
geom_point() +
stat_smooth(method="lm")
```



Output for Varying Slopes Model

```
> wreg5 = lm(wage ~ Male + exper + Male*exper, data=wages)
> summary(wreg5)

Call:
lm(formula = wage ~ Male + exper + Male * exper, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3200 -1.8191 -0.9679  1.4113 17.2672

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.625476   0.341474  13.546   <2e-16 ***
Male         1.532799   0.483246   3.172   0.0016 **
exper       -0.001934   0.015967  -0.121   0.9036
Male:exper   0.055539   0.022218   2.500   0.0127 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- How do we interpret the coefficients for each variable?
- What do we conclude about the interaction?
- What is the predicted wage for a male with 10 years of experience?

Using Two Categorical Variables

- In a previous example, we developed models that included qualitative variables that had two or more categories.
- It is possible to develop a model that has multiple qualitative variables.
- We could expand our gender model to include a variable that includes both the **gender** of and the **race** of each observation.

Two Categorical Variable Model

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- Where
  - y_i = hourly wage of a worker
  - D_{1i} = 1 if the worker is a man, 0 otherwise
  - D_{2i} = 1 if white, 0 otherwise
  - X_i = years of experience
- Based on the dummy variables there are four possible groups: white males, non-white males, white females, and non-white females. Each group will have a different intercept term.
- Which group is our base group?

Conditional Expectations for Each Group

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- Mean earnings of non-white women
  - $E(y_i | x_i, D_1=0, D_2=0) = \alpha_0 + \beta x_i$
- Mean earnings of non-white men
  - $E(y_i | x_i, D_1=1, D_2=0) = (\alpha_0 + \alpha_1) + \beta x_i$
- Mean earnings of white women
  - $E(y_i | x_i, D_1=0, D_2=1) = (\alpha_0 + \alpha_2) + \beta x_i$
- Mean earnings of white men
  - $E(y_i | x_i, D_1=1, D_2=1) = (\alpha_0 + \alpha_1 + \alpha_2) + \beta x_i$

Remember:
D1 is dummy for gender (1 = male)
D2 is dummy for race (1 = white)

- What are the null hypotheses being tested in this model?

Output from Two Categorical Variable Model

```
> wreg8 = lm(wage ~ Male + white + exper, data=wages)
> summary(wreg8)

Call:
lm(formula = wage ~ Male + white + exper, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0176 -2.0839 -0.9145  1.4012 17.5163

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.67047    0.53340   6.881 1.71e-11 ***
Male         2.48037    0.30280   8.191 2.01e-15 ***
white        0.53253    0.49758   1.070   0.2850
exper        0.02691    0.01116   2.412   0.0162 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.463 on 521 degrees of freedom
Multiple R-squared: 0.1266,    Adjusted R-squared: 0.1215
F-statistic: 25.17 on 3 and 521 DF,  p-value: 3.222e-15
```

Interaction Effects in a Two Categorical Variable Model

Original Model $$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + \varepsilon_i$$

- Where
  - y_i = hourly wage of a worker
  - D_{1i} = 1 if the worker is a man, 0 otherwise
  - D_{2i} = 1 if white, 0 otherwise
  - X_i = years of experience
- In this model we assume that the impact on earnings for being a male, α1, is the same for both whites and non-whites. We may believe however, that the impact may be larger for non-whites than for whites.
- We also assume that the impact on earnings for being white, α2, is the same for both men and women. We may believe however, that the impact may be larger for women than for men.
- Thus, we have implicitly imposed some constraints on our model based on its specification. In a sense, this is the reverse problem of the dummy variable trap. Here we have two terms to represent four groups.
- How can we remove such constraints? We can incorporate an interaction term.

Interaction Term

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{1i} D_{2i} + \beta x_i + \varepsilon_i$$

- The interaction term we added to the model is D_{1i} times D_{2i}.
- We now have a different interpretation for white men since that is the only group where $D_{1i}D_{2i} = 1$. Now we have three terms in our model and can say something about each specific group.

- Let's write out the expected earnings for each group.

Conditional Expectations for Each Group

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{1i} D_{2i} + \beta x_i + \varepsilon_i$$

- Mean earnings of non-white women
  - E(
- Mean earnings of non-white men
  - E(
- Mean earnings of white women
  - E(
- Mean earnings of white men
  - E(

Remember:
D1 is dummy for gender (1 = male)
D2 is dummy for race (1 = white)

- What are the null hypotheses being tested in this model?

Output from Categorical Interaction Model

```
> wreg8 = lm(wage ~ Male + white + exper + Male*white, data=wages)
> summary(wreg8)


Call:
lm(formula = wage ~ Male + white + exper + Male * white, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0293 -2.0716 -0.9034  1.4091 17.5045

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.78426    0.71825   5.269 2.02e-07 ***
Male         2.26812    0.94604   2.397   0.0169 *
white        0.40595    0.73054   0.556   0.5787
exper        0.02692    0.01117   2.410   0.0163 *
Male:white   0.23647    0.99847   0.237   0.8129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.466 on 520 degrees of freedom
Multiple R-squared: 0.1267,    Adjusted R-squared: 0.1199
F-statistic: 18.85 on 4 and 520 DF,  p-value: 1.73e-14
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.78426    0.71825   5.269 2.02e-07 ***
Male         2.26812    0.94604   2.397   0.0169 *
white        0.40595    0.73054   0.556   0.5787
exper        0.02692    0.01117   2.410   0.0163 *
Male:white   0.23647    0.99847   0.237   0.8129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We interpret the coefficient on ‘***male***’ as the impact on earnings for being male relative to being a non-white female. Thus, it is the difference between non-white women and non-white men
- We interpret the coefficient on ‘***white***’ as the impact on earnings for being a white female relative to being a non-white female. Here the difference in earnings between white women and non-white women is 0.406.
- What about the difference in earnings between white males and non-white females? The difference is the sum of the coefficients on ‘male’, ‘white’, and ‘Male:white’.
- Lastly, what is the difference in earnings between white males and non-white males?

IN-CLASS EXERCISES

- Using the '**wages**' dataset we used above, run a model predicting wage, by educ, exper, white, male, and tenure.
  - Interpret the results.
  - What is the expected wage of a white, female, with 8 years of educ, and 0 years of exper and tenure?
  - What is the expected wage of a white, female, with 9 years of educ, and 0 years of exper and tenure?
- Create an interaction effect between tenure and white.
  - Interpret the results.
  - Draw a diagram by hand displaying the results of the interaction (regardless of whether or not the interaction is significant).