Interaction Models

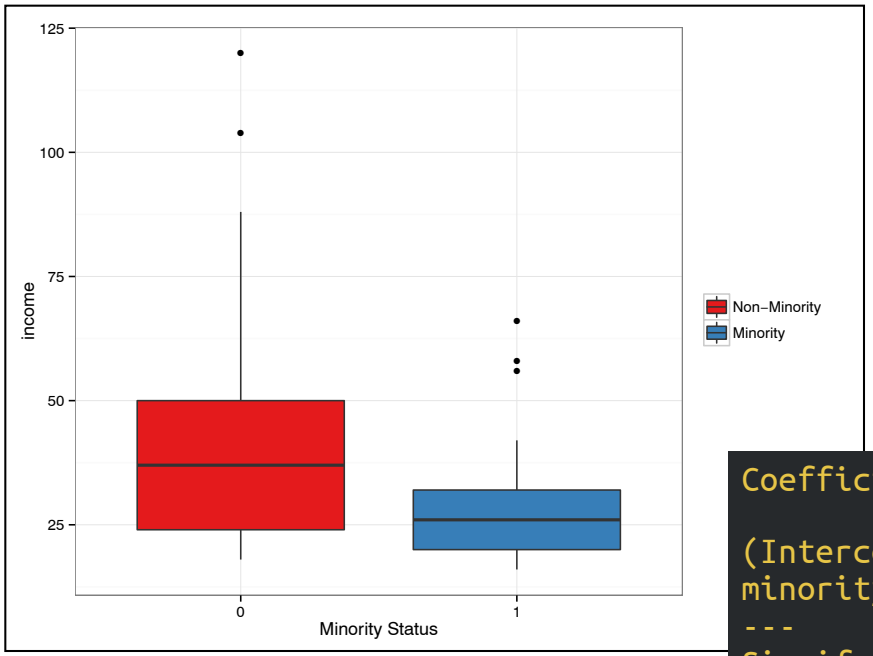
Preparation

We will use the `census-sample.csv` data. We read this into a data frame called `census`.

The data include the following variables:

- `income:` Respondent's annual income, in thousands of dollars
- `education:` Respondent's level of education, in years
- `ethnicity:` Respondent's ethnicity (black, hispanic, white)
- `black:` Is the respondent black? (1 = Yes, 0 = No)
- `hispanic:` Is the respondent hispanic? (1 = Yes, 0 = No)
- `white:` Is the respondent white? (1 = Yes, 0 = No)
- `minority:` Is the respondent a minority (non-white ethnicity)? (1 = Yes, 0 = No)

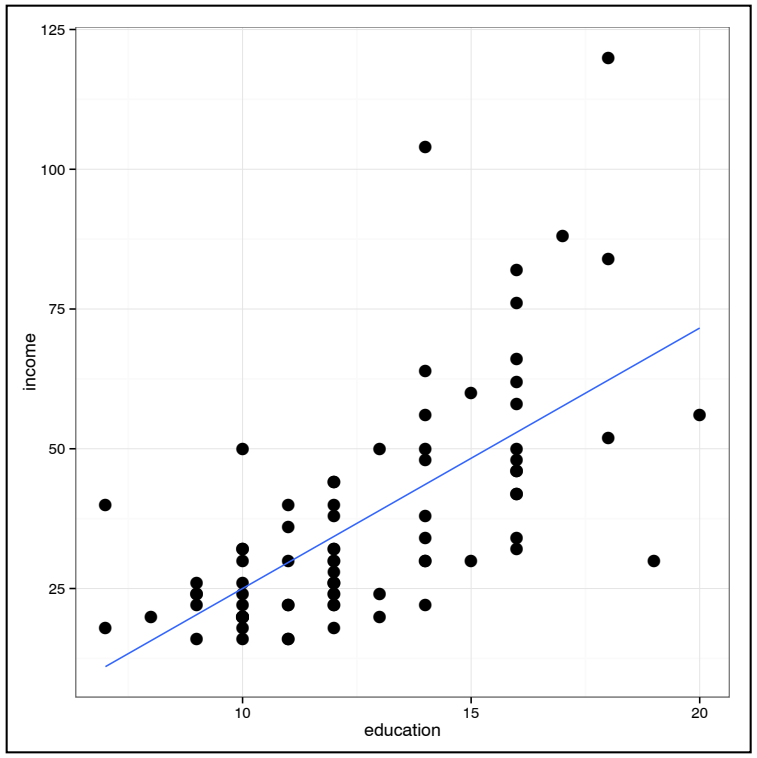
Research Question: Is there a relationship between minority status and income?




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    42.480      2.796  15.193  < 2e-16 ***
minority      -13.213      4.566  -2.894  0.00493 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.77 on 78 degrees of freedom
Multiple R-squared:  0.09696,  Adjusted R-squared:  0.08538
F-statistic: 8.375 on 1 and 78 DF,  p-value: 0.00493
```

Yes there seems to be a relationship between minority status and income. The estimated slope coefficient is –13.21, $t(78) = -2.89$, $p = .005$. Minority status explains roughly 9.7% of the variability in income.

Research Question: Is there a relationship between level of education and income?




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.5908     8.0833  -2.671   0.0092 **
education     4.6594     0.6216   7.496 8.85e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.86 on 78 degrees of freedom
Multiple R-squared:  0.4187,  Adjusted R-squared:  0.4113
F-statistic: 56.19 on 1 and 78 DF,  p-value: 8.852e-11
```

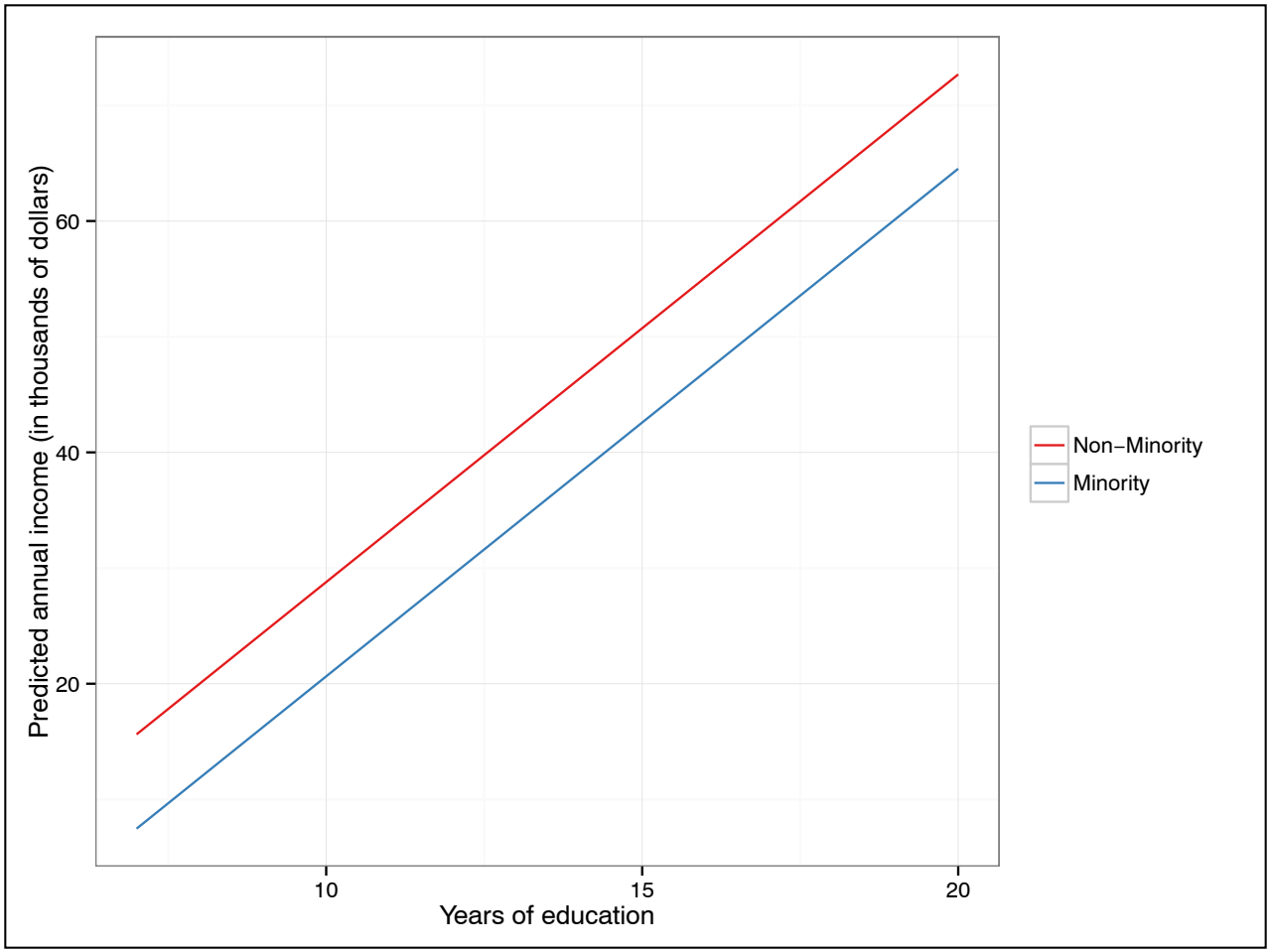
Yes there seems to be a relationship between level of education and income. The estimated slope coefficient is 4.66, $t(78) = 7.50$, $p < .001$. Level of education explains roughly 41.8% of the variability in income.

Sub-Question: Is there a still a relationship between level of education and income after controlling for minority status?

Counter argument: Perhaps minorities earn less because of a differences in education.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.0993     8.4002  -1.797   0.0762 .
education     4.3887     0.6182   7.100 5.38e-10 ***
minority     -8.1517     3.6430  -2.238   0.0281 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.47 on 77 degrees of freedom
Multiple R-squared:  0.4542,   Adjusted R-squared:   0.44
F-statistic: 32.04 on 2 and 77 DF,  p-value: 7.502e-11
```

Ok…even after accounting for the differences explained by minority status, there **still** seems to be a relationship between level of education and income. The estimated **partial slope** coefficient is 4.39, $t(77) = 7.10$, $p < .001$. Level of education **and** minority status explain roughly 45.5% of the variability in income.



The model we fitted is called a main-effects model. The effect of a predictor is the same for both the groups (the lines are parallel). However, the predicted income for minority students is lower than that for non-minority students, *even if the have the same level of education*. This is because of the effect of minority status.

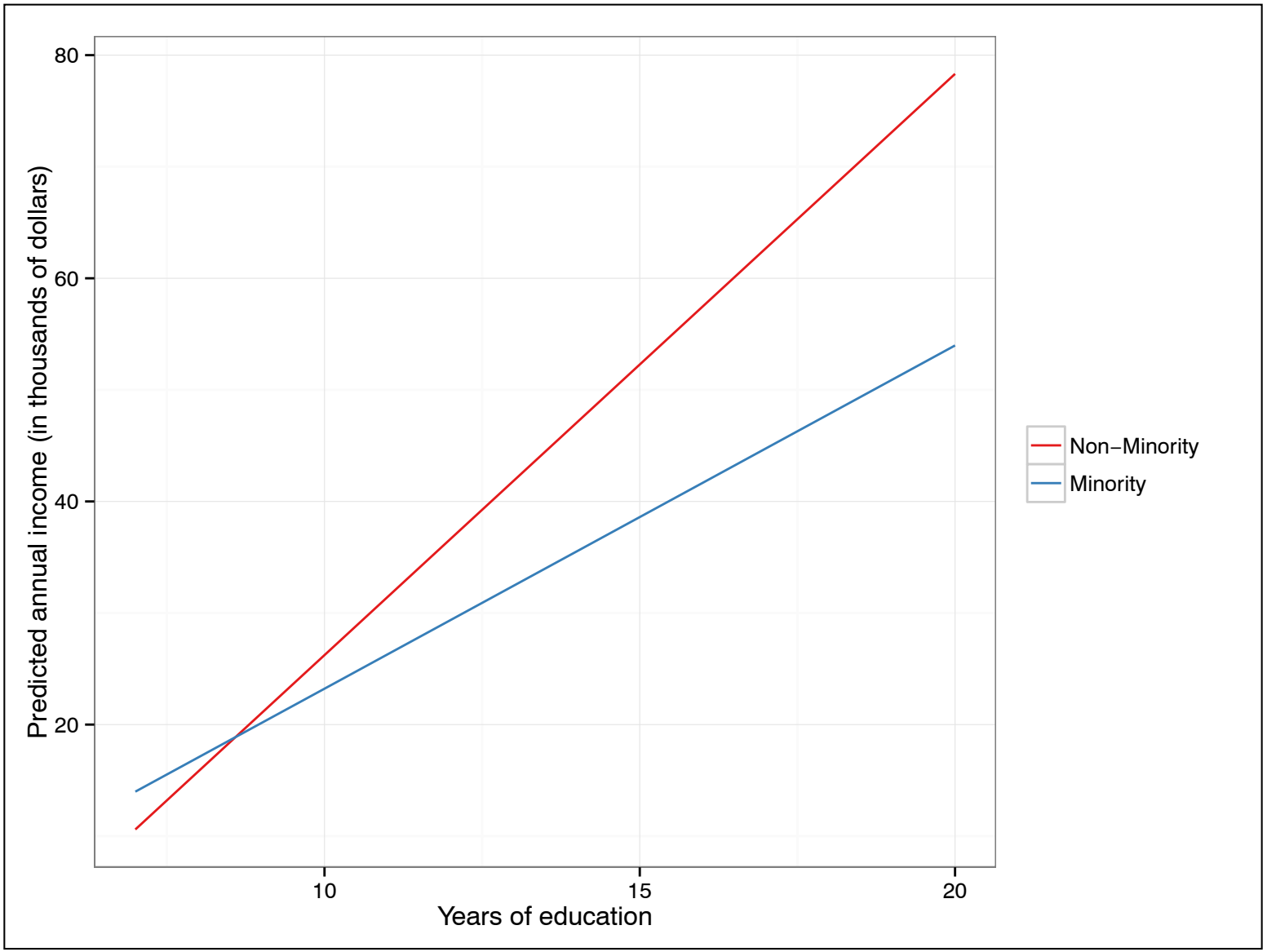
For our RQ it means we think there is a positive effect of education on income, and we believe this even after accounting for differences in minority status.

Interaction Effect: Different Effect of Education for Minority and Non-Minority Groups?

Interaction effects allow the effect of a predictor (X1) to differ across levels of another predictor (X2).

For example, the effect of education might have a larger relationship (larger slope) with income for non-minorities than it does for minorities.

Interactions allow us to examine whether there are **differential effects** of a predictor across groups.



The effect of education on income for minority groups has a smaller effect (shallower slope) than that for non-minorities. This is indicative of an **interaction** between minority status and education. In a plot, the differential effect of education shows up in the **non-parallel lines**.

Testing for an Interaction

Create a variable that is the product of the two predictors you want to examine the interaction between (in our example this would be the product of minority status and education).

```
> census$educMin = census$education * census$minority

> head(census)

  income education ethnicity black hispanic white minority educMin
1     16        10     black     1        0     0        1      10
2     18         7     black     1        0     0        1       7
3     26         9     black     1        0     0        1       9
4     16        11     black     1        0     0        1      11
5     34        14     black     1        0     0        1      14
6     22        12     black     1        0     0        1      12
```

Fit a model that includes both of the constituent predictors (the main-effects) and the product term (the interaction effect) as predictors of the outcome.

```
> lm.3 = head(income ~ education + minority + educMin, data = census)

> summary(lm.3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.8688    10.4417  -2.477   0.0155 *
education     5.2095     0.7786   6.691 3.34e-09 ***
minority     18.3233    15.9883   1.146   0.2554
educMin      -2.1333     1.2552  -1.700   0.0933 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.28 on 76 degrees of freedom
Multiple R-squared:  0.4742,   Adjusted R-squared:  0.4534
F-statistic: 22.85 on 3 and 76 DF,  p-value: 1.207e-10
```

There is some statistical evidence of an interaction effect ($p = 0.093$). This suggests that in the population, the effect of education on income probably differs between minorities and non-minorities.

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education}) + 18.3(\text{Minority}) - 2.1(\text{Education})(\text{Minority})$$

Non-Minorities (`Minority = 0`)

Minorities (`Minority = 1`)

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education}) + 18.3(\text{Minority}) - 2.1(\text{Education})(\text{Minority})$$

Intercept

Main-Effect of Education

Main-Effect of Minority

Interaction between Education and Minority

When X2 is a Dummy Variable

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \beta_3(X_1)(X_2) + \epsilon$$

X2 = 0

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(0) + \beta_3(X_1)(0) + \epsilon$$

$$Y = \beta_0 + \beta_1(X_1) + \epsilon$$

X2 = 1

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(1) + \beta_3(X_1)(1) + \epsilon$$

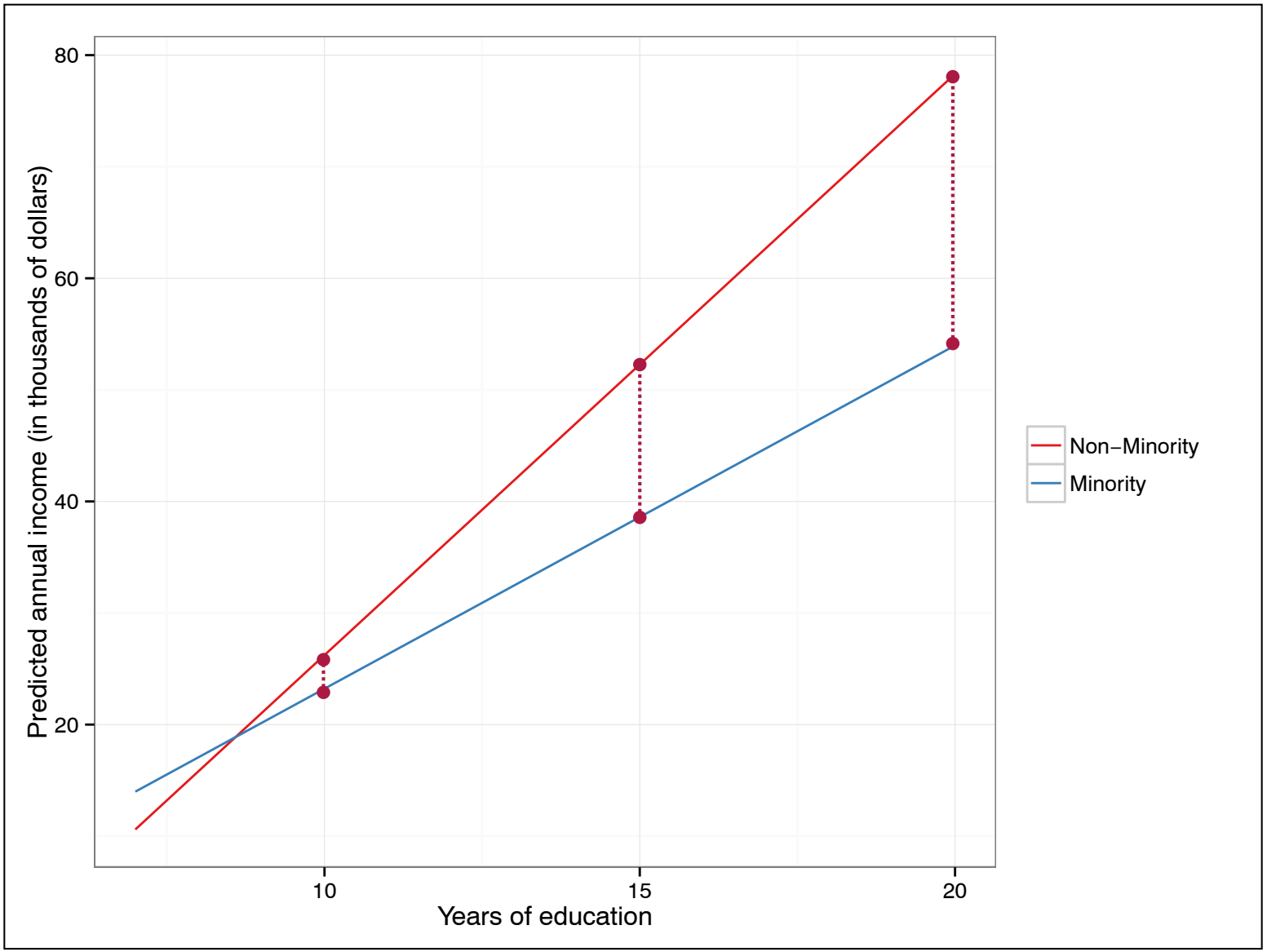
$$Y = \beta_0 + \beta_1(X_1) + \beta_2 + \beta_3(X_1) + \epsilon$$

$$Y = [\beta_0 + \beta_2] + [\beta_1 + \beta_3](X_1) + \epsilon$$

The interaction effect is the difference in the slopes for the two groups.

$$H_0 : \beta_3 = 0$$

The test of the interaction is thus a test of the difference in the effect of X1 across the different levels of X2. It allows us to examine whether the slopes are different!



Alternatively, instead of looking at the slopes, we can look at the vertical distance between the lines. Remember this is the visual representation of the effect of the minority predictor. The difference between the lines is different depending on which value of education we look at.

Interpretation of an Interaction Effect

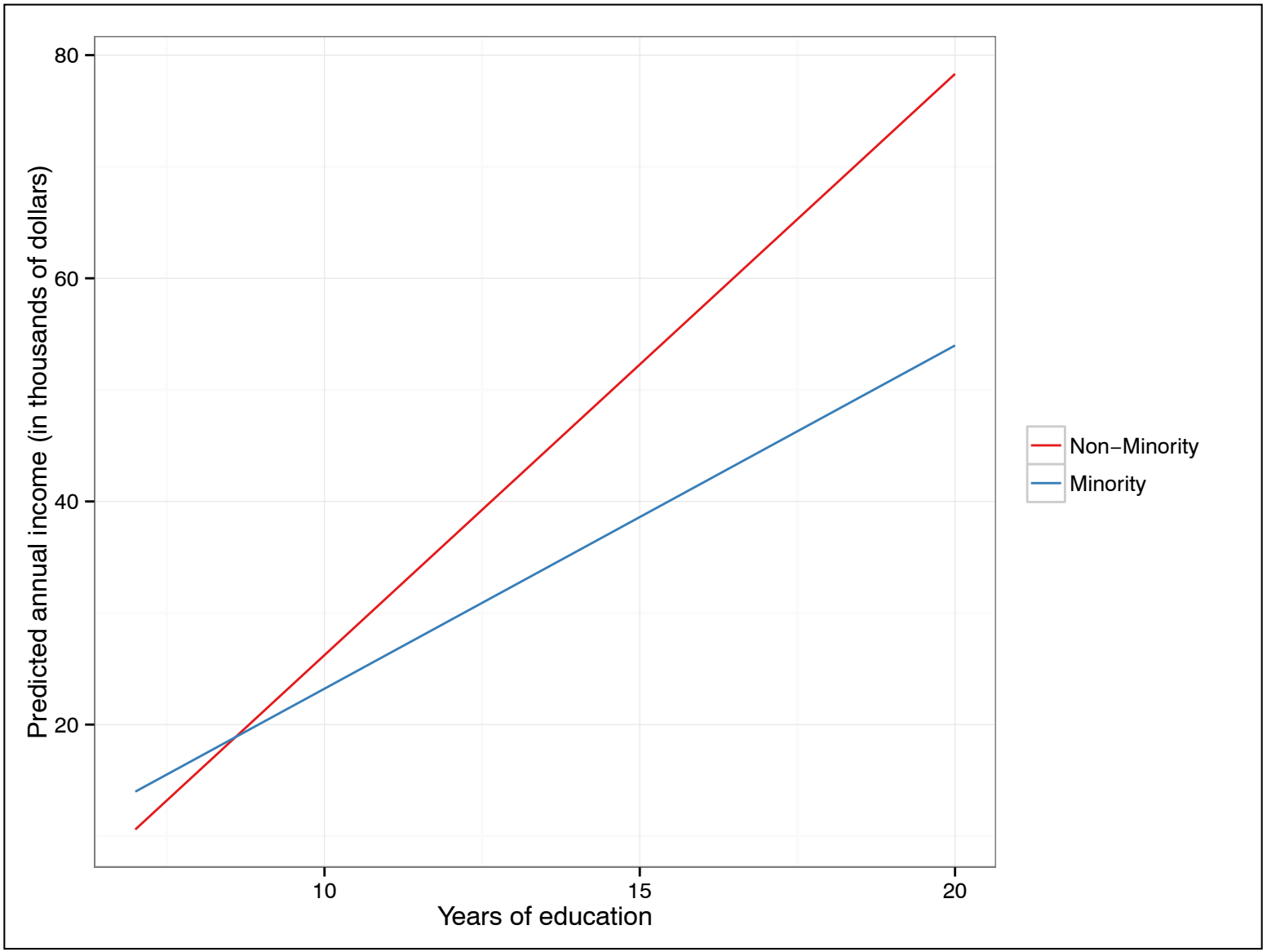
Interaction effects (between two predictors) can always be interpreted two different ways…

The effect of education on income differs between minorities and non-minorities.

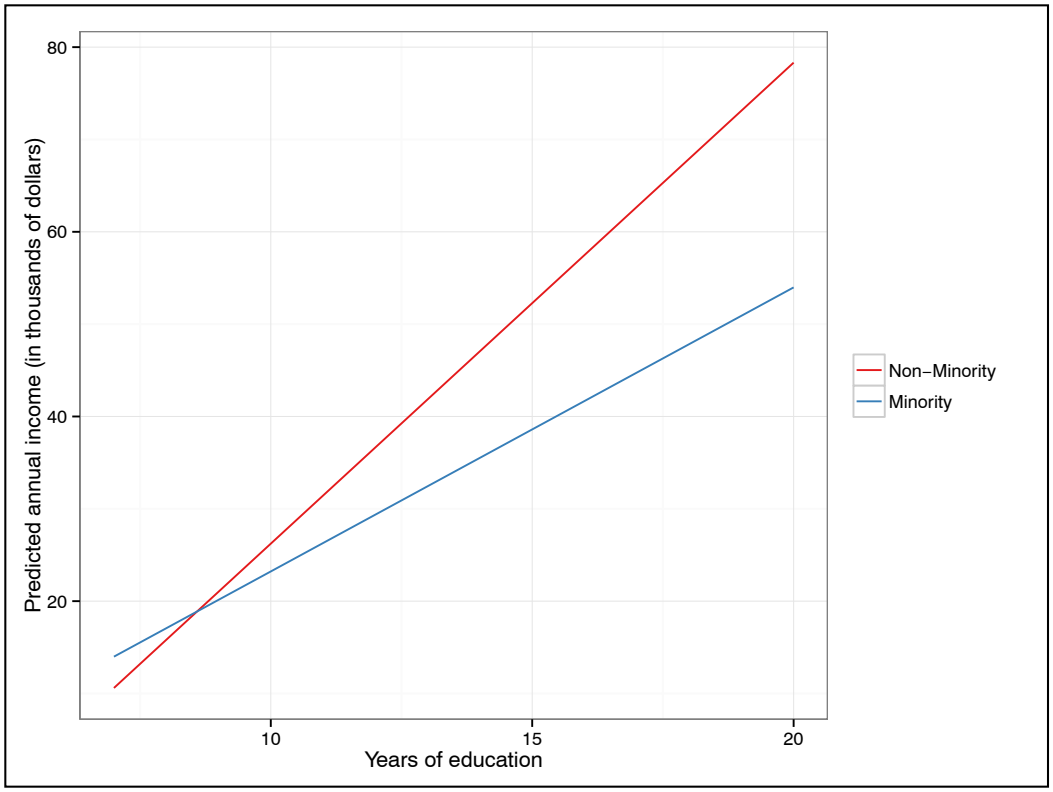
The effect of minority status on income differs across levels of education.

The effect of X1 on Y differs across levels of X2.

The effect of X2 on Y differs across levels of X1.



When the lines cross within the range of data we have, we sometimes refer to the interaction as a **disordinal** interaction. (If they do not cross within the range of data we have, it is referred to as an **ordinal** interaction.)



Disordinal interactions require a more elaborate interpretation if you are making predictions.

The effect of education on income is complicated. In general, for the same level of education, the average income for minorities is lower than the average income for non-minorities. Moreover for higher levels of education, this differential increases. However, for people with education levels below 8th-grade, .

You can fit an interaction in the `lm()` function without explicitly creating a product by using the colon (`:`) operator.

```
> lm.4 = head(income ~ education + minority + education:minority,
     data = census)

> summary(lm.4)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -25.8688    10.4417  -2.477   0.0155 *
education           5.2095     0.7786   6.691 3.34e-09 ***
minority           18.3233    15.9883   1.146   0.2554
education:minority  -2.1333     1.2552  -1.700   0.0933 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.28 on 76 degrees of freedom
Multiple R-squared:  0.4742,   Adjusted R-squared:  0.4534
F-statistic: 22.85 on 3 and 76 DF,  p-value: 1.207e-10
```

For plotting the model, it is easier if you use the colon operator.

Interactions with Group Variables Consisting of More Than One Dummy Variable

This time to examine the effect of ethnicity, we will use the three dummy variables (black, hispanic, and white) rather than the minority variable.

```
> head(census)

  income education ethnicity black hispanic white minority educMin
1     16        10     black     1        0     0        1      10
2     18         7     black     1        0     0        1       7
3     26         9     black     1        0     0        1       9
4     16        11     black     1        0     0        1      11
5     34        14     black     1        0     0        1      14
6     22        12     black     1        0     0        1      12
```

Even though we might be interested ultimately in the interaction between ethnicity and education, we should start with a main-effects model. In this model, the reference group is white.

```
> lm.5 = lm(income ~ education + black + hispanic,
     data = census)

> summary(lm.5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.6635     8.4121  -1.862   0.0665 .
education     4.4317     0.6191   7.158 4.42e-10 ***
black       -10.8744     4.4730  -2.431   0.0174 *
hispanic     -4.9338     4.7632  -1.036   0.3036
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.46 on 76 degrees of freedom
Multiple R-squared:  0.462,    Adjusted R-squared:  0.4408
F-statistic: 21.75 on 3 and 76 DF,  p-value: 2.853e-10
```

- Differences in ethnicity and education level explain variation in income, $F(3, 76) = 21.75$, $p < .001$, $R^2 = 0.462$.
- Education is an important predictor of the variation in income, even after controlling for differences in ethnicity, $t(76) = 7.16$, $p < .001$.
- There are differences in the average income between whites and blacks, after controlling for differences in education, $t(76) = –2.43$, $p = .017$
- There are not differences in the average income between whites and hispanics, after controlling for differences in education, $t(76) = –1.04$, $p = ..304$

To fit the interaction model we will use the colon operator. We have to create an interaction term between each component of the two predictors we are interested in the interaction between.

Ethnicity x Education

black x education

hispanic x education

```
> lm.6 = lm(income ~ education + black + hispanic +
     education:black + education:hispanic, data = census)

> summary(lm.6)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -25.8688    10.4982  -2.464   0.0161 *
education           5.2095     0.7828   6.655  4.3e-09 ***
black              19.3333    18.2928   1.057   0.2940
hispanic            9.2640    24.2797   0.382   0.7039
education:black    -2.4107     1.4177  -1.700   0.0933 .
education:hispanic -1.1208     2.0060  -0.559   0.5781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.37 on 74 degrees of freedom
Multiple R-squared:  0.4825,   Adjusted R-squared:  0.4475
F-statistic:  13.8 on 5 and 74 DF,  p-value: 1.618e-09
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -25.8688    10.4982  -2.464   0.0161 *
education           5.2095     0.7828   6.655  4.3e-09 ***
black              19.3333    18.2928   1.057   0.2940
hispanic            9.2640    24.2797   0.382   0.7039
education:black    -2.4107     1.4177  -1.700   0.0933 .
education:hispanic -1.1208     2.0060  -0.559   0.5781
---
```

Interactions

- There is some evidence to suggest that the effect of education for whites is different than the effect of education for blacks, $t(74) = -1.70$, $p = .093$.
- There is no evidence to suggest that the effect of education for whites is different than the effect of education for hispanics, $t(74) = -0.56$, $p = .578$.

Are there differences in the effect of education between blacks and hispanics? To find out we need to fit another interaction model with one of them as the reference group.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -25.8688    10.4982  -2.464   0.0161 *
education           5.2095     0.7828   6.655  4.3e-09 ***
black              19.3333    18.2928   1.057   0.2940
hispanic            9.2640    24.2797   0.382   0.7039
education:black    -2.4107     1.4177  -1.700   0.0933 .
education:hispanic -1.1208     2.0060  -0.559   0.5781
---
```

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education}) + 19.3(\text{Black}) + 9.3(\text{Hispanic}) - 2.4(\text{Education})(\text{Black}) - 1.1(\text{Education})(\text{Hispanic})$$

Whites (`black = 0`, `hispanic = 0`)

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education}) + 19.3(0) + 9.3(0) - 2.4(\text{Education})(0) - 1.1(\text{Education})(0)$$

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education})$$

$$\hat{\text{Income}} = -25.9 + 5.2(\text{Education}) + 19.3(\text{Black}) + 9.3(\text{Hispanic}) - 2.4(\text{Education})(\text{Black}) - 1.1(\text{Education})(\text{Hispanic})$$

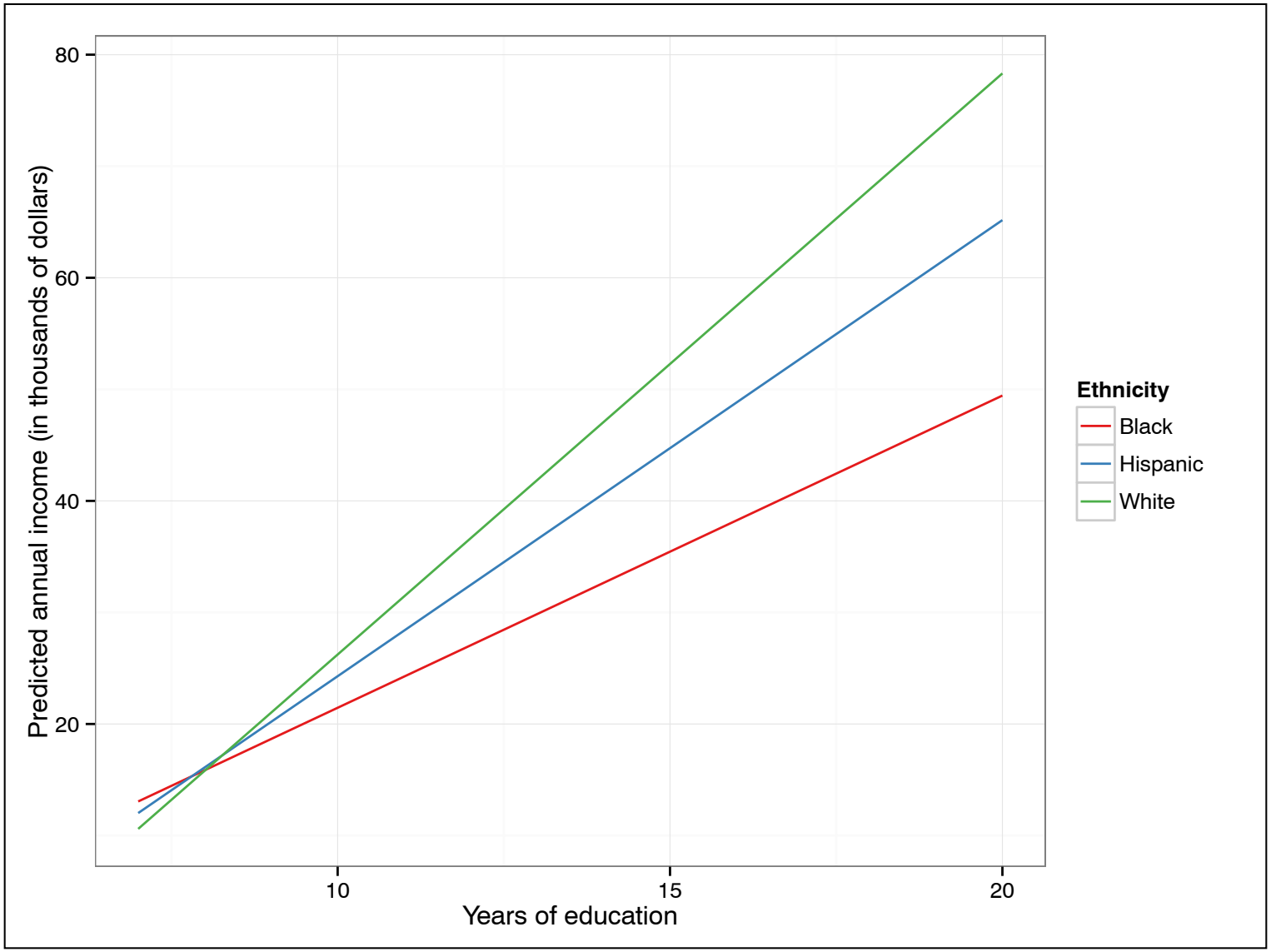
Blacks (`black = 1, hispanic = 0`)

Hispanics (`black = 0, hispanic = 1`)

```
> lm.7 = lm(income ~ education + white + hispanic +
     education:white + education:hispanic, data = census)

> summary(lm.6)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         -6.536     14.980  -0.436   0.6639
education            2.799      1.182   2.368   0.0205 *
white              -19.333     18.293  -1.057   0.2940
hispanic           -10.069     26.527  -0.380   0.7053
education:white      2.411      1.418   1.700   0.0933 .
education:hispanic   1.290      2.193   0.588   0.5582
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.37 on 74 degrees of freedom
Multiple R-squared:  0.4825,   Adjusted R-squared:  0.4475
F-statistic:  13.8 on 5 and 74 DF,  p-value: 1.618e-09
```

Interactions

- There is some evidence to suggest that the effect of education for blacks is different than the effect of education for whites, $t(74) = 1.70$, $p = .093$. (*Same is in the previous model.*)
- There is no evidence to suggest that the effect of education for blacks is different than the effect of education for hispanics, $t(74) = 0.59$, $p = .558$.



Based on our examination of the results, Blacks and Hispanics do not have statistically different slopes (the red and blue lines are parallel in the population). Also, Hispanics and Whites have the same slope (blue and green lines are parallel….But, there is some evidence that Blacks and Whites have different effects of education (red and green lines are not parallel)!!!??