Homework 9

Miles Tweed

4/18/2021

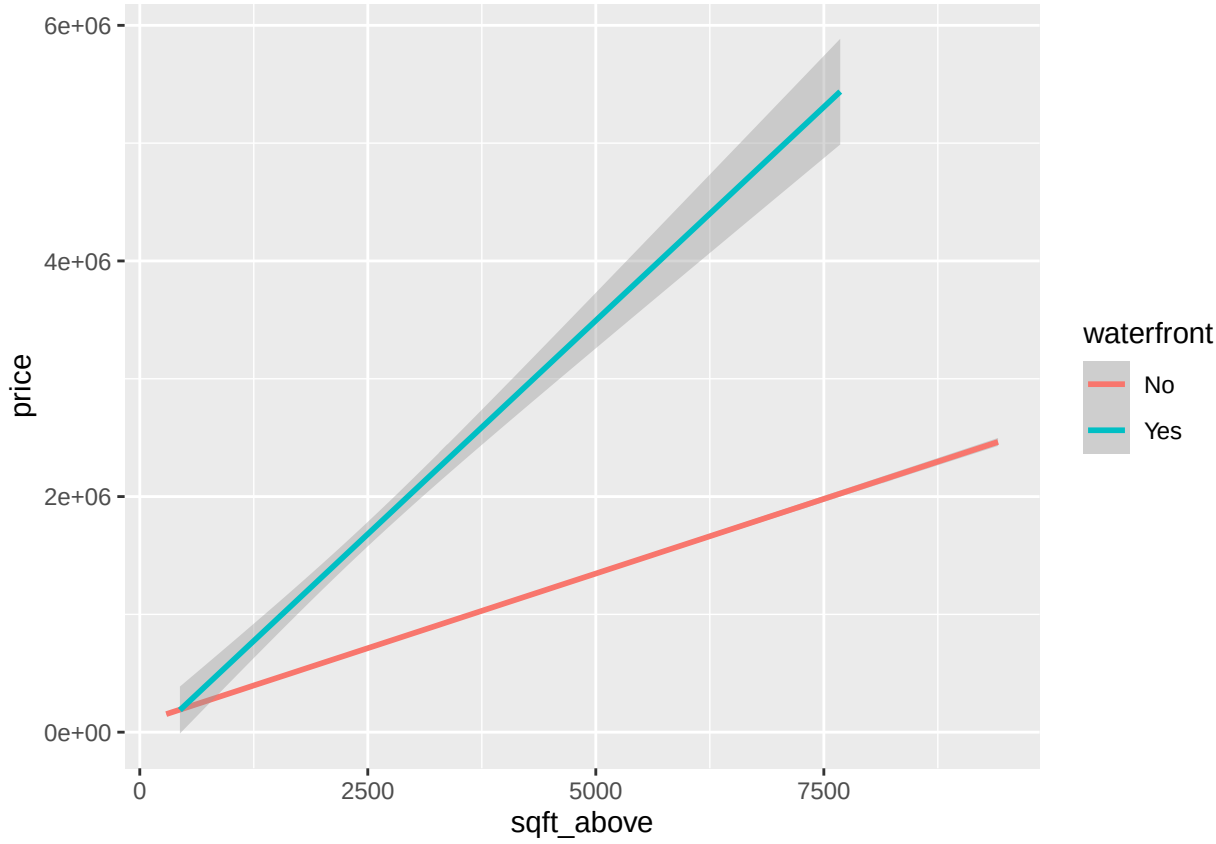
Problem 1

Part 1

I am using the King County house price data. The quantitative response is the house price and the quantitative predictor I chose is sqft_above and the categorical variable I chose was waterfront which has the values of 'yes' or 'no' indicating is the property is waterfront or not. As shown in the plot below, there is clear indication of interaction between sqft_above and waterfront indicated by the differing slopes.

```
df <- read_csv('../../Data/kc_house_data.csv')

red.df <- df %>% select(price, sqft_above, waterfront)
red.df$waterfront <- as_factor(red.df$waterfront)
levels(red.df$waterfront) <- c('No','Yes')
ggplot(data = red.df, aes(y = price, x = sqft_above, color = waterfront)) +
  geom_smooth(method = 'lm')
```



Part 2

1)

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{sqft\_above}_i + \beta_2 \cdot D_{\text{waterfront},i} + \beta_3 \cdot \text{sqft\_above}_i \cdot D_{\text{waterfront},i} + \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma^2)$$

2)

$$\text{price}_i = \begin{cases} \beta_0 + \beta_1 \cdot \text{sqft\_above}_i + \epsilon_i & , \ i^{th} \text{ is not waterfront} \\ (\beta_0 + \beta_2) + (\beta_1 \cdot + \beta_3) \cdot \text{sqft\_above}_i + \epsilon_i & , \ i^{th} \text{ is waterfront} \end{cases}$$

The two equations have different intercepts and slopes. In the equation for houses that are not waterfront the intercept is β_0 whereas in the equation for houses that are waterfront the intercept is $(\beta_0 + \beta_2)$. In the equation for houses that are not waterfront the slope is β_1 whereas in the equation for houses that are waterfront the slope is $(\beta_1 + \beta_3)$.

Fill in the blank: Interaction terms allow for difference in linear regression slope across the categories.

Part 3

```
lm.obj <- lm(price ~ sqft_above*waterfront, data=red.df)
summary(lm.obj)
```

```
##
## Call:
## lm(formula = price ~ sqft_above * waterfront, data = red.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1713255  -160823   -38149   108170  5449048
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.981e+04  4.500e+03  17.735  < 2e-16 ***
## sqft_above                 2.533e+02  2.292e+00 110.546  < 2e-16 ***
## waterfrontYes             -2.112e+05  4.853e+04  -4.352 1.35e-05 ***
## sqft_above:waterfrontYes   4.718e+02  1.762e+01  26.768  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 275900 on 21609 degrees of freedom
## Multiple R-squared:  0.4353, Adjusted R-squared:  0.4352
## F-statistic:  5552 on 3 and 21609 DF,  p-value: < 2.2e-16
```

The extremely low p-value for the interaction term (<2e-16) indicated that the interaction is statistically significant.

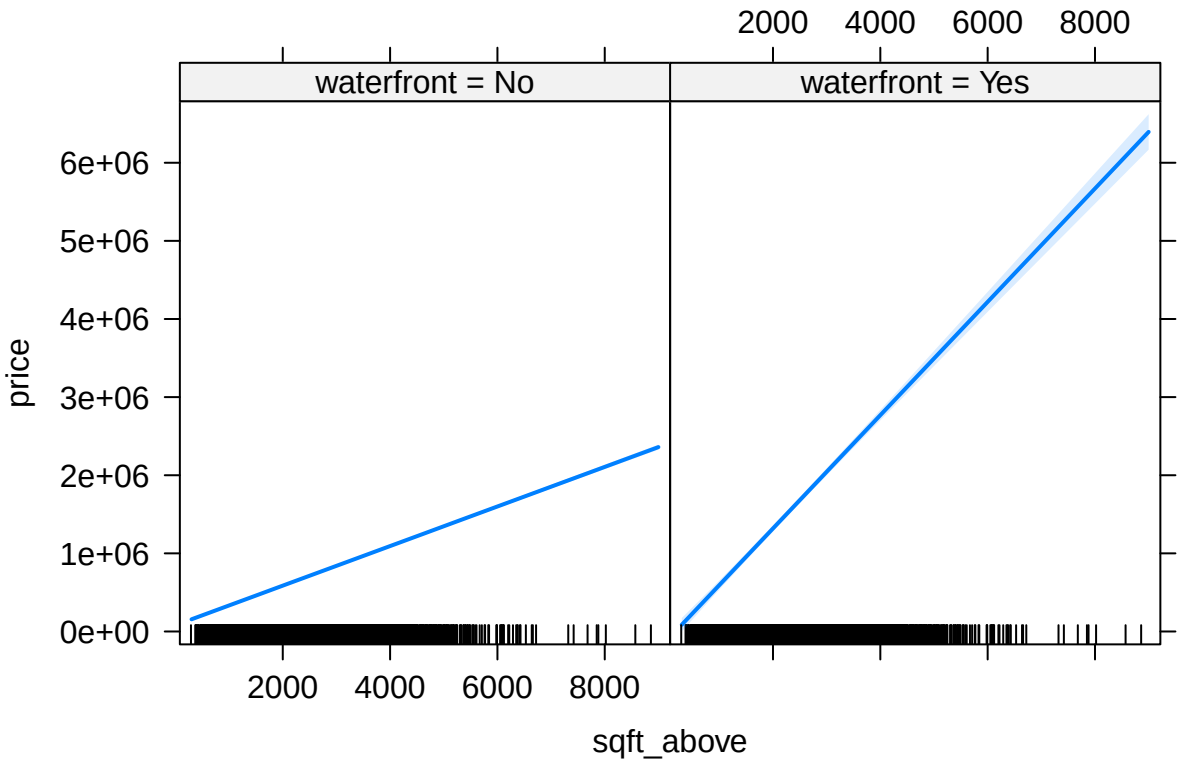
Part 4

a.

$$\hat{\text{price}} = \begin{cases} 7.98e4 + 2.53e2 \cdot \text{sqft\_above} & , \ i^{th} \text{ is not waterfront} \\ (7.98e4 - 2.11e5) + (2.53e2 + 4.72e2) \cdot \text{sqft\_above} & , \ i^{th} \text{ is waterfront} \end{cases}$$

b.

```
library(effects)
plot(effect("sqft_above:waterfront", lm.obj))
```



In the effect display, the slope is much steeper for houses where waterfront is 'Yes'. This indicates that, on average, an increase in sqft_above results in a greater increase in price if the house is waterfront.

Interpretation

For slope: On average, a one square foot increase results in a $253 increase in home price for houses that are not waterfront. This average per unit change in home price is increased to $253 +$472 = $725 for houses that are waterfront.

For intercept: For houses that are zero square feet (empty lots), the base price will be $79,800 when it is not waterfront and -$131,200 if it is waterfront. Obviously, this parameter has a nonsensical interpretation.

Part 5

On average, waterfront houses that are zero square feet in size will be priced $211,200 less than houses that are not waterfront. This interpretation sounds absurd and certainly not "all-encompassing". This interpretation only focuses on situations where the home is zero square feet (empty lots) which is overly specific and it is very doubtful that waterfront lots would be priced lower that non-waterfront lots on average. We should never rely on interpretations or the significance of the main effects in models with strong interactions because these effects are considered to be marginal to the interaction.

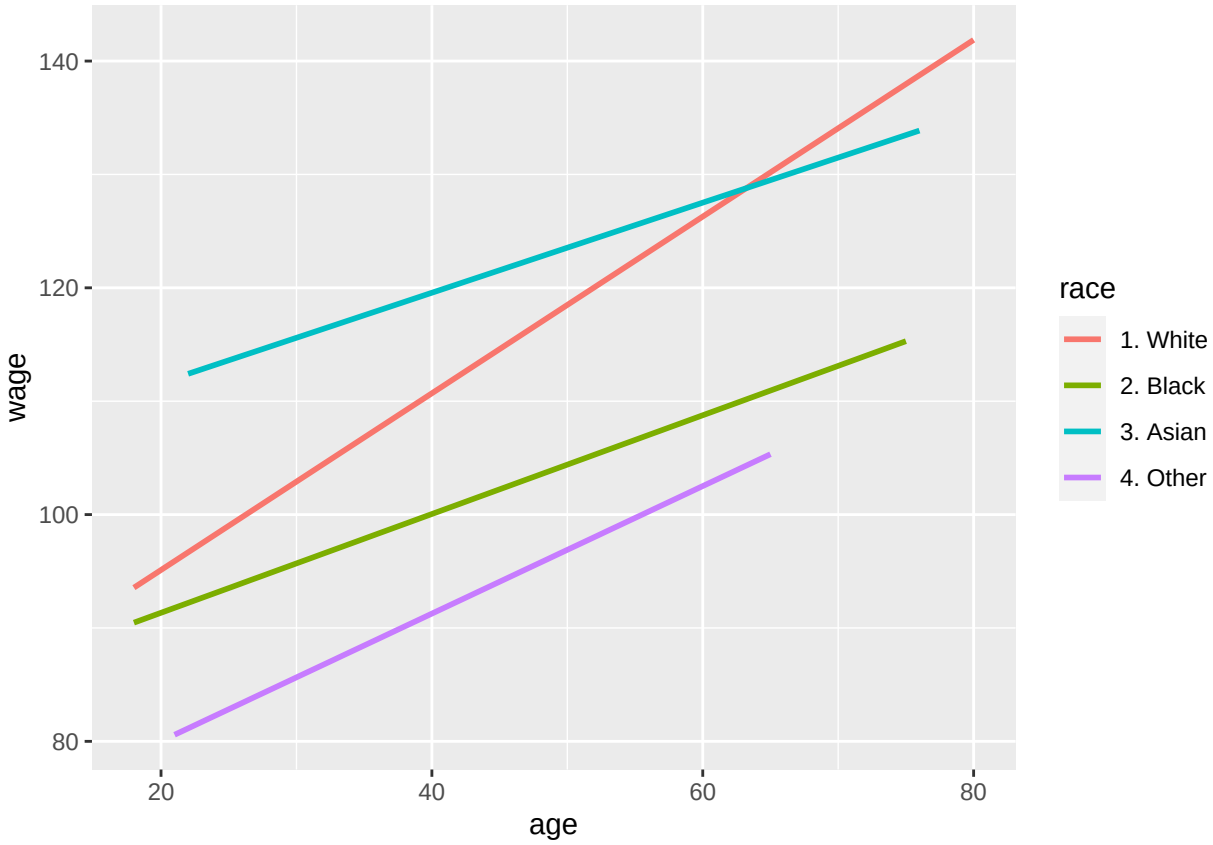
Problem 2

Part 1

```
library(ISLR)
fix(Wage)

ggplot(data = Wage, aes(x = age, y = wage, color = race)) +
  geom_smooth(method = 'lm', se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Since the slopes of the fitted line of age and wage are different for the various categories of race we can conclude that there is evidence of interaction between race and age on wage.

Part 2

```
lm.obj.wage = lm(wage~age*race, data=Wage)
summary(lm.obj.wage)
```

```
##
## Call:
## lm(formula = wage ~ age * race, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.549  -23.994   -6.003   16.389  215.680
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       79.53003    3.15470  25.210   <2e-16 ***
## age                0.77919    0.07188  10.841   <2e-16 ***
## race2. Black       3.09726    8.89338   0.348   0.7277
## race3. Asian      24.14253   11.84459   2.038   0.0416 *
## race4. Other     -10.76045   23.32390  -0.461   0.6446
## age:race2. Black  -0.34364    0.19653  -1.749   0.0805 .
## age:race3. Asian  -0.38189    0.27333  -1.397   0.1625
```

```
## age:race4. Other  -0.21680    0.59111  -0.367   0.7138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.67 on 2992 degrees of freedom
## Multiple R-squared:  0.05214,    Adjusted R-squared:  0.04992
## F-statistic: 23.51 on 7 and 2992 DF,  p-value: < 2.2e-16
```

1) $\text{wage}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot D_{\text{Black},i} + \beta_3 \cdot D_{\text{Asian},i} + \beta_4 \cdot D_{\text{Other},i} + \beta_5 \cdot \text{age} \cdot D_{\text{Black},i} + \beta_6 \cdot \text{age} \cdot D_{\text{Asian},i} + \beta_7 \cdot \text{age} \cdot D_{\text{Other},i} + \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma^2)$

2)

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$
$$H_a : \{\text{at least one } \beta_j \neq 0\}, j = 5, 6, 7$$

3)

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot D_{\text{Black},i} + \beta_3 \cdot D_{\text{Asian},i} + \beta_4 \cdot D_{\text{Other},i} \epsilon_i,\ \epsilon_i \sim \mathbb{N}(0, \sigma^2)$$

4)

```
lm.obj.null <- lm(wage~age + race, data = Wage)

anova(lm.obj.null, lm.obj.wage)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ age + race
## Model 2: wage ~ age * race
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   2995 4957481
## 2   2992 4949793  3    7687.7 1.549 0.1998
```

5) The value of the F statistic was 1.549 which led to an insignificant p-value of 0.1998. This result leads us to conclude that we should fail to reject the null hypothesis which states that the coefficients of interaction terms between race categories and age should be zero. This indicates that there is likely not a significant interaction between age and race that affects an individual's wage on average.