Example4_4

Kevin Cummiskey

October 24, 2019

Recall in Example 4.3, we looked at the effect of house size on price after adjusting for location. Here is the model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where y_i is the price of house i, $x_{1,i}$ is the size (sq ft) of house i, and $x_{2,i}$ is 1 if house i is lakefront and is 0 otherwise.

What assumptions does this model make?

How do we interpret β_1? β_2?

```r
houses = read.table(file = "http://www.isi-stats.com/isi2/data/housing.txt",
                    header = TRUE)

#Reverse coding of lake
houses$lake = factor(houses$lake, levels = c("notlakefront",
                                             "lakefront"))

model_withLake = lm(price.1000 ~ sqft + lake, data = houses)
summary(model_withLake)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft + lake, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.059 -48.444   3.072  38.191 140.421
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -70.1821    62.8062  -1.117 0.289933
## sqft            0.1481     0.0283   5.233 0.000383 ***
## lakelakefront 331.2235    41.8470   7.915 1.29e-05 ***
## ---
```

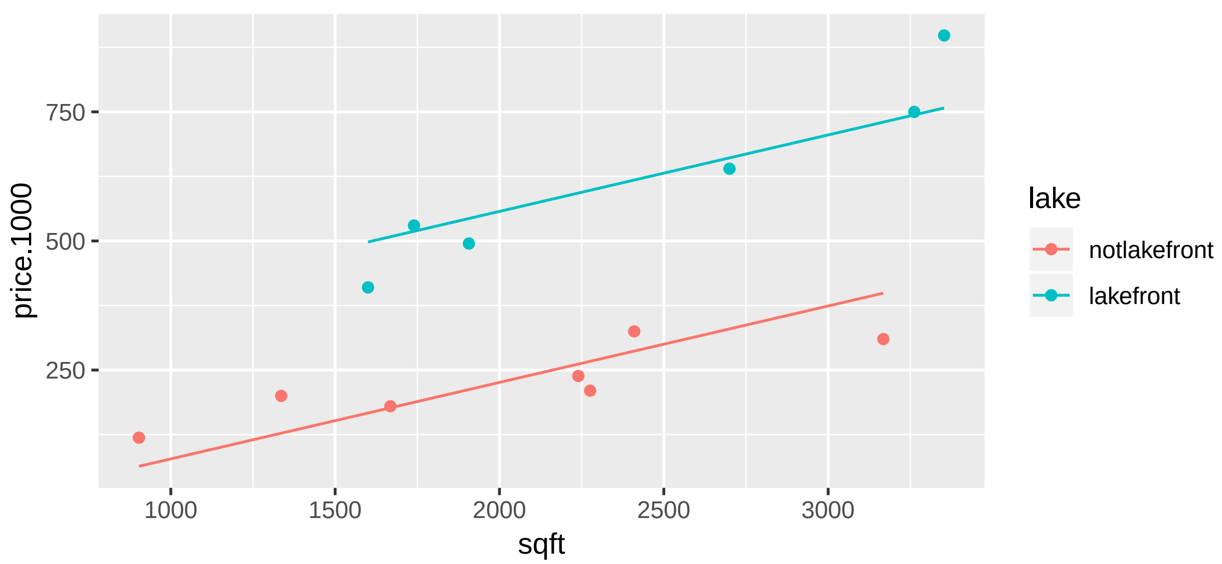
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72.02 on 10 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9106
## F-statistic: 62.15 on 2 and 10 DF,  p-value: 2.289e-06
```

```
anova(model_withLake)
```

```
## Analysis of Variance Table
##
## Response: price.1000
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sqft       1 319753  319753  61.654 1.386e-05 ***
## lake       1 324911  324911  62.649 1.293e-05 ***
## Residuals 10  51862    5186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

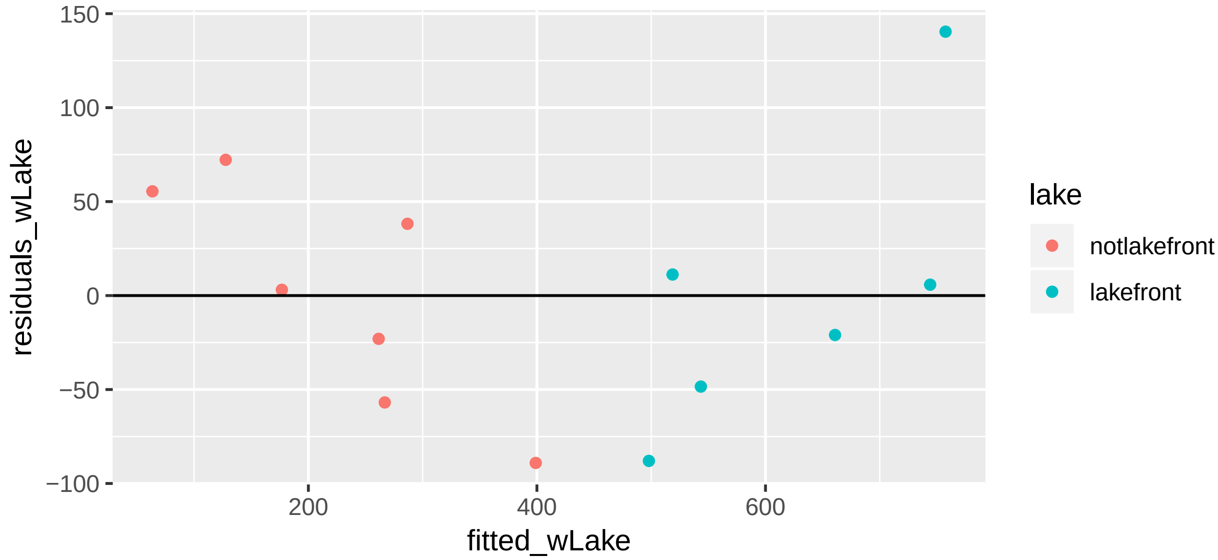
```
#add fitted values and residuals to the data set
houses = houses %>% mutate(fitted_wLake = fitted.values(model_withLake))
houses = houses %>% mutate(residuals_wLake = residuals(model_withLake))

houses %>% ggplot(aes(x = sqft, y = price.1000, color = lake)) +
  geom_point() +
  geom_line(aes(y = fitted_wLake))
```



Let's take a look at the residuals vs the predicted (fitted) values.

```
houses %>% ggplot(aes(x = fitted_wLake,
                      y = residuals_wLake,
                      color = lake)) +
  geom_point() + geom_hline(yintercept = 0)
```



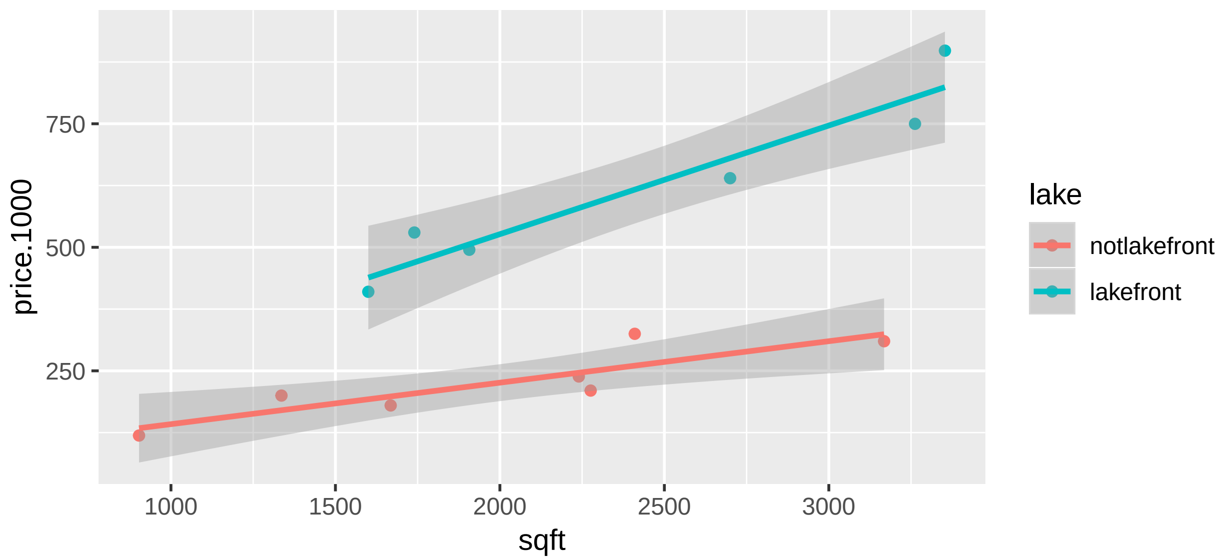
Do we see evidence of an interaction here?

Let's look at a model with an interaction.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

How do we interpret β_0?β_1? β_2? β_3? $\beta_1 + \beta_3$?

```
houses %>% ggplot(aes(x = sqft, y = price.1000, color = lake)) +
  geom_point() + geom_smooth(method = "lm")
```




```r
model_interaction = lm(price.1000 ~ sqft * lake, data = houses)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft * lake, data = houses)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54.16 -28.60 -14.15  29.64  73.93
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        58.11341   56.68270   1.025  0.33202
## sqft                0.08394    0.02675   3.138  0.01197 *
## lakelakefront      28.65098   91.32560   0.314  0.76088
## sqft:lakelakefront  0.13595    0.03895   3.491  0.00682 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.48 on 9 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9578
## F-statistic: 91.84 on 3 and 9 DF,  p-value: 4.547e-07
```

```r
anova(model_interaction)
```

```
## Analysis of Variance Table
##
## Response: price.1000
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sqft       1 319753  319753 130.611 1.166e-06 ***
## lake       1 324911  324911 132.718 1.089e-06 ***
## sqft:lake  1  29829   29829  12.184  0.006824 **
## Residuals  9  22033    2448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What would you conclude from these results? Your answer should include discussion of effect sizes, significance, and overall predictive capability of the model.

The main effect of location is not significant. Should we conclude there is no effect of location on price?

Note: The ANOVA table from R will not match the textbook because there are different ways to calculate sums of squares when the data is unbalanced (as in observational studies). The textbook reports Type III ANOVA tables in this article https://mcfromnz.wordpress.com/2011/03/02/anova-type-iiiiii-ss-explained/

For Type III ANOVA, you can use the `car package in R. Note the capital "A" in the function below.`

```
library(car)
Anova(model_interaction, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: price.1000
##              Sum Sq Df F value   Pr(>F)
## (Intercept)  2573.3  1  1.0511 0.332015
## sqft        24104.0  1  9.8459 0.011970 *
## lake          241.0  1  0.0984 0.760881
## sqft:lake   29829.0  1 12.1844 0.006824 **
## Residuals   22033.2  9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```