Example4_4

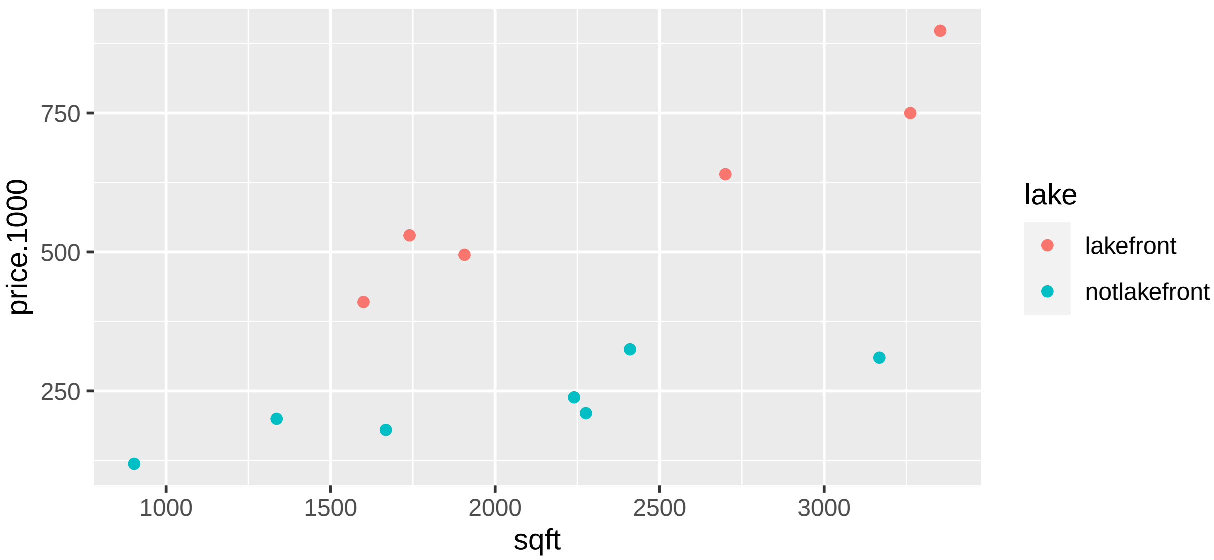
Kevin Cummiskey

October 24, 2019

Review

In Chapter 4, we've been looking at housing prices in Michigan.

```r
houses = read.table(file = "http://www.isi-stats.com/isi2/data/housing.txt",
                    header = TRUE)
houses %>%
  ggplot(aes(x = sqft, y = price.1000, color = lake)) +
  geom_point()
```



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

$$price_i = \alpha_0 + \alpha_1 sqft_i + \alpha_2 lake_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where $price_i$ is the price of house i, $sqft_i$ is the size (sq ft) of house i, and $lake_i$ is 1 if house i is lakefront and is 0 otherwise.

What assumptions does this model make?

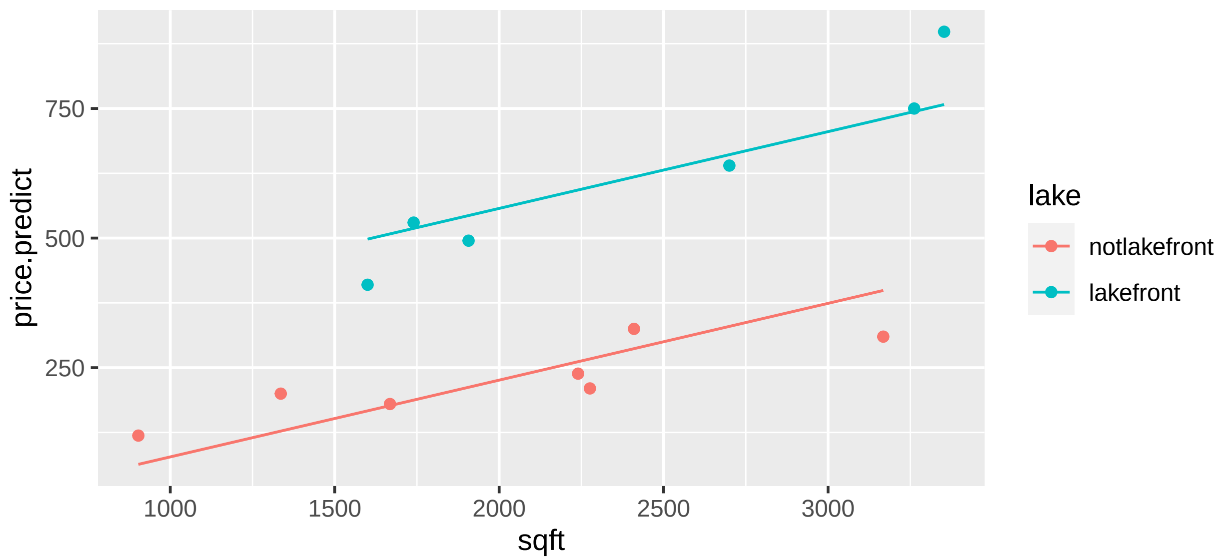
How do we interpret α_1? α_2?

```r
#Reverse coding of lake so 1 is lake and 0 is not lake
houses$lake = factor(houses$lake, levels = c("notlakefront",
                                             "lakefront"))

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

houses %>%
  mutate(price.predict = predict(model_withLake, newdata = .),
         residuals = price.1000 - price.predict) -> houses

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




```r
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
```

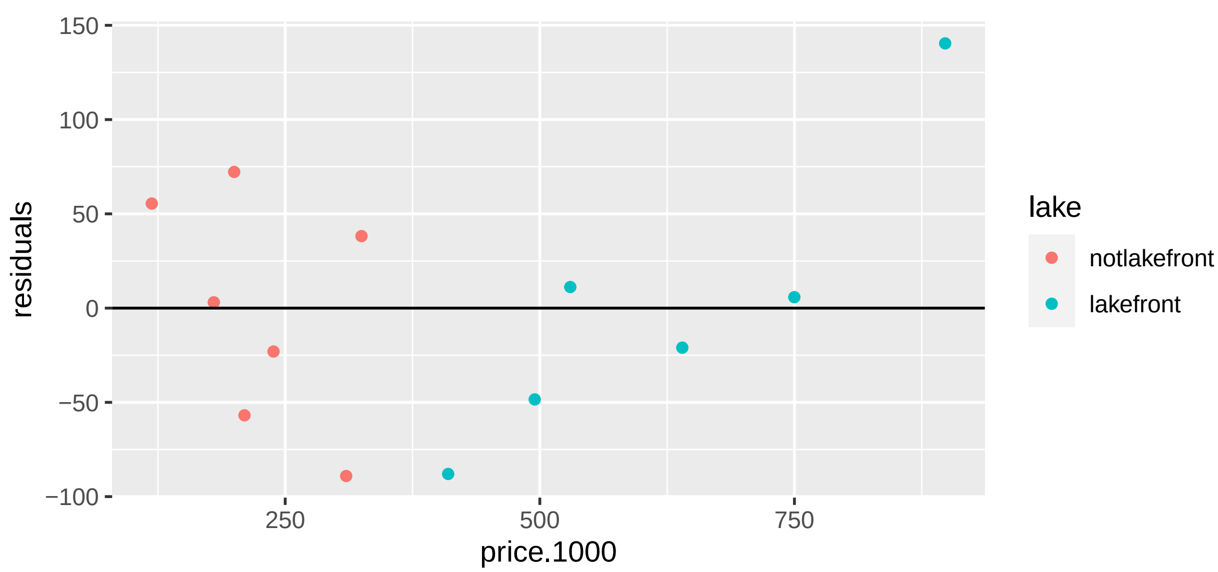
Report estimates of model parameters here.

Calculate the expected price of a 2500 sq ft house that is not on the lake front.

Calculate the expected price of a 2500 sq ft house on the lake front.

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

```r
houses %>% ggplot(aes(x = price.1000,
                      y = residuals,
                      color = lake)) +
  geom_point() + geom_hline(yintercept = 0)
```



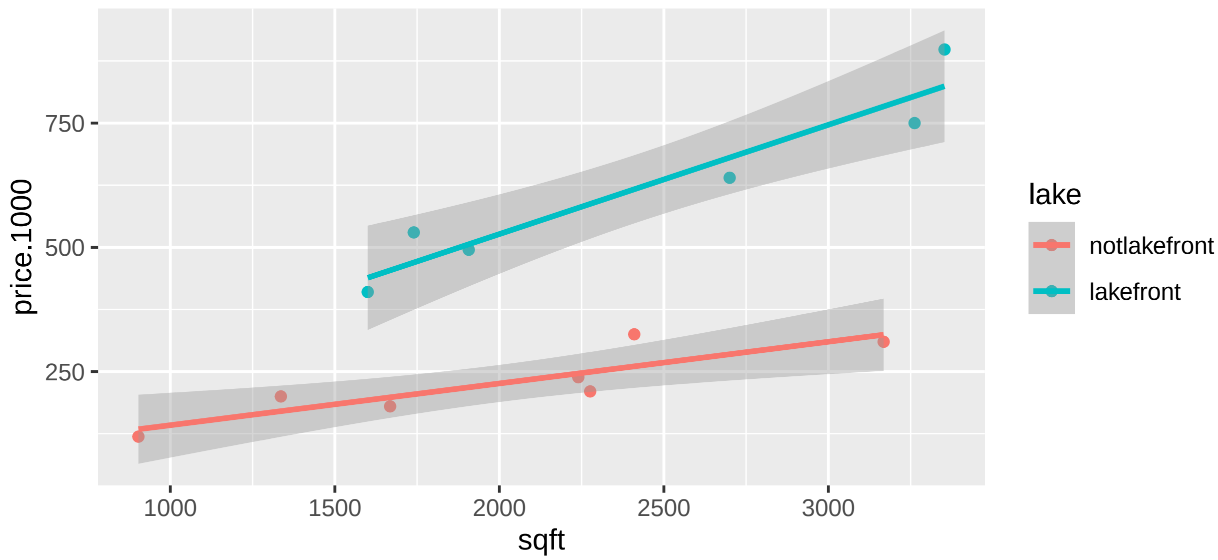
Do we see evidence of an interaction here?

Two models

If we suspect the price per square foot differs for lake front and nonlake front, we could fit two separate simple models.

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

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




```
lake_model <- lm(price.1000 ~ sqft, data = houses %>% filter(lake == "lakefront"))
coef(lake_model)
```

```
## (Intercept)        sqft
##  86.7643819   0.2198952
```

```
nonlake_model <- lm(price.1000 ~ sqft, data = houses %>% filter(lake == "notlakefront"))
coef(nonlake_model)
```

```
## (Intercept)        sqft
## 58.11340672  0.08394444
```

Interactions

Let's look at a model with an interaction.

$$price_i = \beta_0 + \beta_1 sqft_i + \beta_2 lake_i + \beta_3 sqft_i lake_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

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

```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
```

Write the equation for the price of a lake front house as a function of size.

Write the equation for the price of a nonlake house as a function of size.

How do these equations compare to the two models above? Why is the interaction model preferable?

Is there evidence the price per square foot is different for lakefront and nonlakefront homes? How do you know?

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 in the model with an interaction. Should we conclude the location of the house is not associated with price? Justify your answer.

```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
```

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.`

```r
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
```