Lesson_16_Solution

MAJ Baller

October 7, 2021

```
#Load Packages
library(tidyverse)
library(ggResidpanel)
#read in data
houses = read.csv("https://raw.githubusercontent.com/danielpballer/MA376/master/Data/houses.csv", string
```

Recall Up to this point we have been using statistical models of one of two forms:

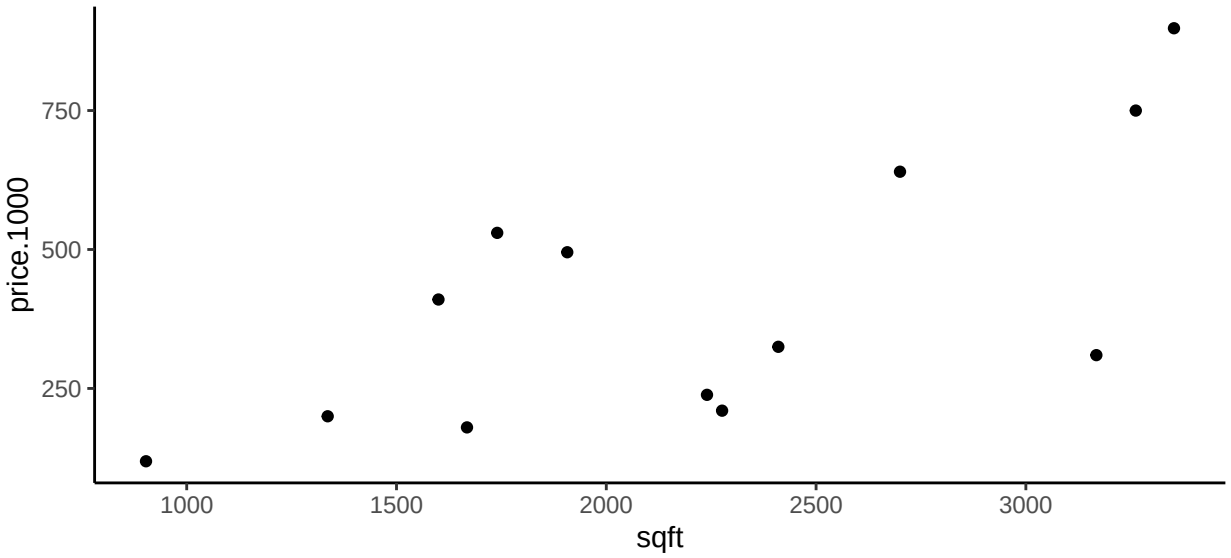
$$\hat{y}_{ij} = \mu_j \quad \text{or} \quad \hat{y}_{ij} = \mu + \alpha_j$$

In the multiple means examples we have looked at, our explanatory variable, or independent variable have been categorical. This was nice in that we could think of each of our observations as belonging to a group - one level of the categorical variable. We considered whether there was an association between the response variable and the explanatory variable.

In this section, you will continue to focus on using one explanatory variable to explain variation in a response variable; however, the explanatory variable will be quantitative and we will consider whether the means of the response variable distributions at different values of the explanatory variable tend to follow a linear pattern (i.e., a constant rate of change).

Many people say that you can predict the cost of a home by its size (square footage). Today we will look to verify this claim using a sample of homes around Lake Michigan

```
houses %>%
  ggplot(aes(x = sqft, y = price.1000)) + geom_point()+theme_classic()
```



If our explanatory variable and response variable have a linear relationship, i.e., a constant rate of change, we could use the regression approach. What is the regression approach? What do we mean by linear regression?

Simple Linear Regression is the statistical method that allows us to quantify the linear relationship between two quantitative variables with a regression.

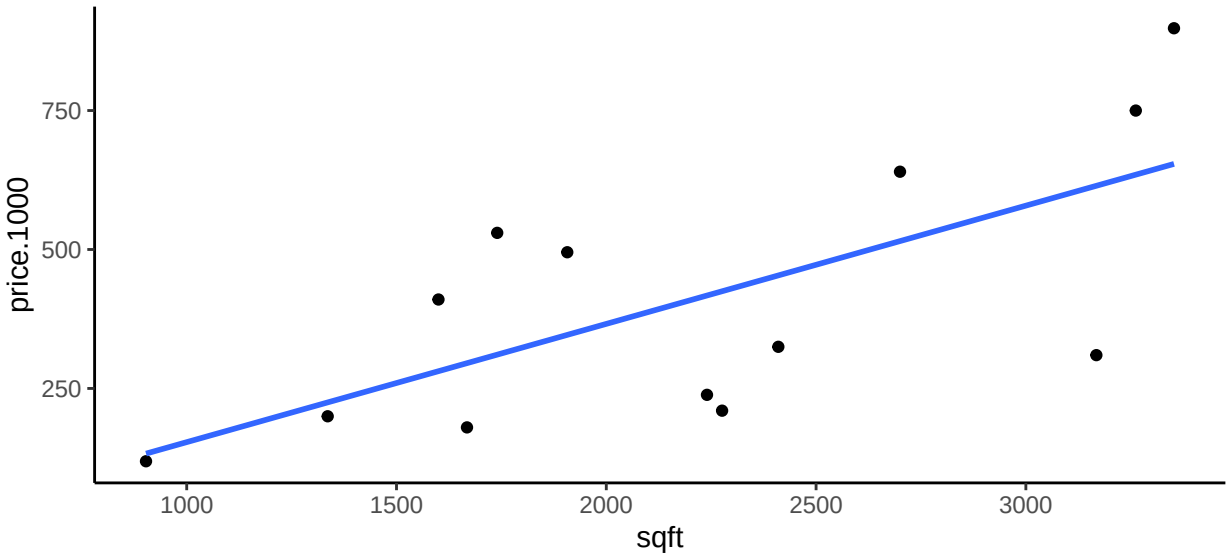
$$\hat{price}_i = \beta_0 + \beta_1 sqft_i$$

where $price_i$ is the price of house i and $sqft_i$ is the size (sq ft) of house i.

Our book uses b_0 and b_1 instead of $\hat{\beta}_0$ and $\hat{\beta}_1$. I prefer to use the beta hats, and the beta hats are standard in statistics literature, but use whatever you like- just be consistent. The more important issue is, how do we interpret these estimated parameters?

Is linear regression appropriate in this scenario? Why or why not? How are you deciding?

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



Fit a linear regression model

```
model_simple <- lm(price.1000 ~ sqft, data = houses)
summary(model_simple)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.70 -128.44  -13.74  128.98  244.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.36870  161.36807  -0.368   0.7199
## sqft          0.21274    0.06963   3.055   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.1 on 11 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4099
## F-statistic: 9.335 on 1 and 11 DF,  p-value: 0.01094
```

How do we interpret the coefficients in this example?

The yintercept $\hat{\beta}_0$ implies that the predicted price of a home is -59,370 dollars when the square footage is 0. However, we have to be very careful in extrapolating our regression line beyond the range of data in the sample. We don't know if the linear trend we observed between 903 and 3353 square feet continues outside of that range. The slope value $\hat{\beta}_1$ of 212.70 dollars indicates that the price of a home is predicted to increase by 212.70 dollars with each additional square foot.

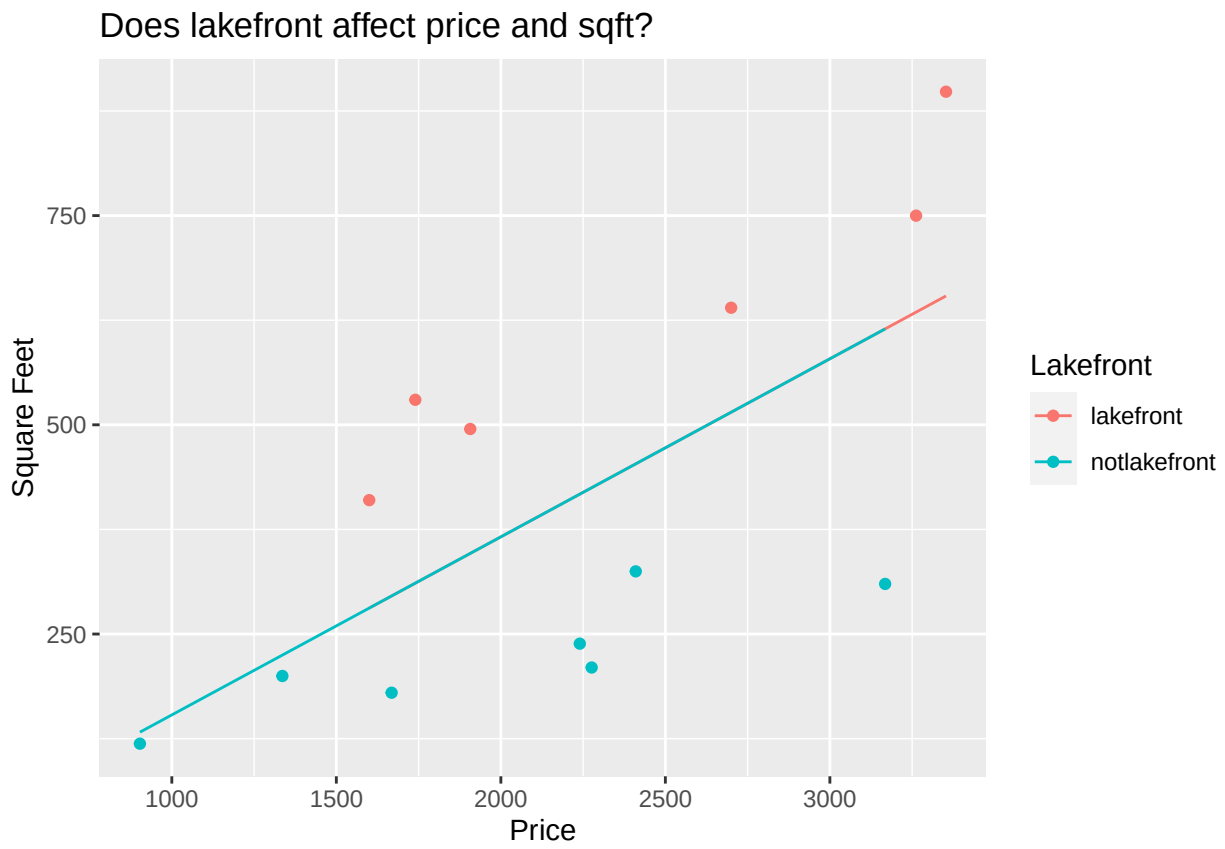
We have another variable in the data set - location. We can consider adding this variable to the regression model. What is the classification of the location variable? What do we hope to accomplish by adding this variable to the regression model?

Categorical. We can explain variation in house prices AND we can explain variation in house prices not already explained by the house size variable.

Briefly explain why location (lake or nonlake) may be a confounding variable.

Lake houses are traditionally more expensive and bigger than non lake houses. Lake impacts both the explanatory and response variables.

```
houses %>%
  ggplot(aes(x = sqft, y = price.1000, color = lake)) +
  geom_point() +
  geom_line(aes(x = houses$sqft, y = model_simple$fitted.values))+
  labs(title = "Does lakefront affect price and sqft?",
       x = "Price", y = "Square Feet", color = "Lakefront")
```



We can adjust for a confounding variable by including it in a multiple regression model. In this model, we estimate the price per square foot when we hold the location of the house constant.

$$\hat{price}_i = \alpha_0 + \alpha_1 sqft_i + \alpha_2 lake_i$$

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?

The slope for both the lake houses and nonlake houses are the same.

How do we interpret α_0? α_1? α_2?

α_0 = the average price of a home that is 0 square feet after and not on the lake.

α_1 = the change in the average price of a home for an increase of one square foot after adjusting for location.

α_2 = the change in the average price of a home for a change of location to lake after adjusting for size.

In general, when will α_1 from this model equal β_1 in the simple model?

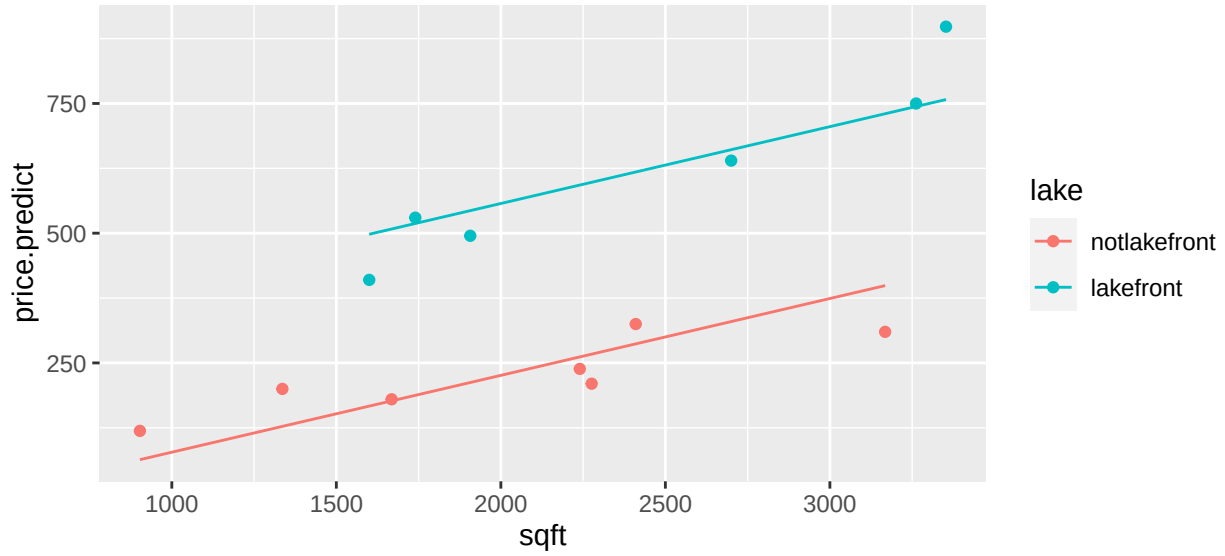
When there is no impact from lake

```
#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))
```




```
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
```

How did the estimate of the price per square footage change when we held location constant?

The price per square foot has decreased to $148.10

What is our new regression model?

$$\hat{price}_i = -70.1821 + (0.1481 \times sqft_i) + (331.2235 \times lake_i)$$

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

```
-70.1821+(0.1481*2500)+(331.2235*0)
```

```
## [1] 300.0679
```

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

```
-70.1821+(0.1481*2500)+(331.2235*1)
```

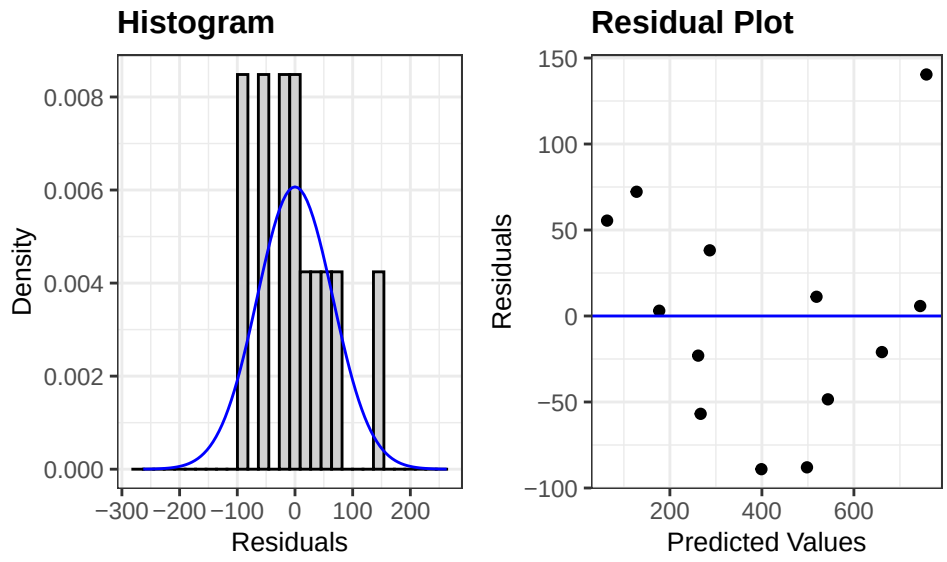
```
## [1] 631.2914
```

How do we interpret α_0?

α_0 = the price of a house that is not on the lake and has 0 square feet

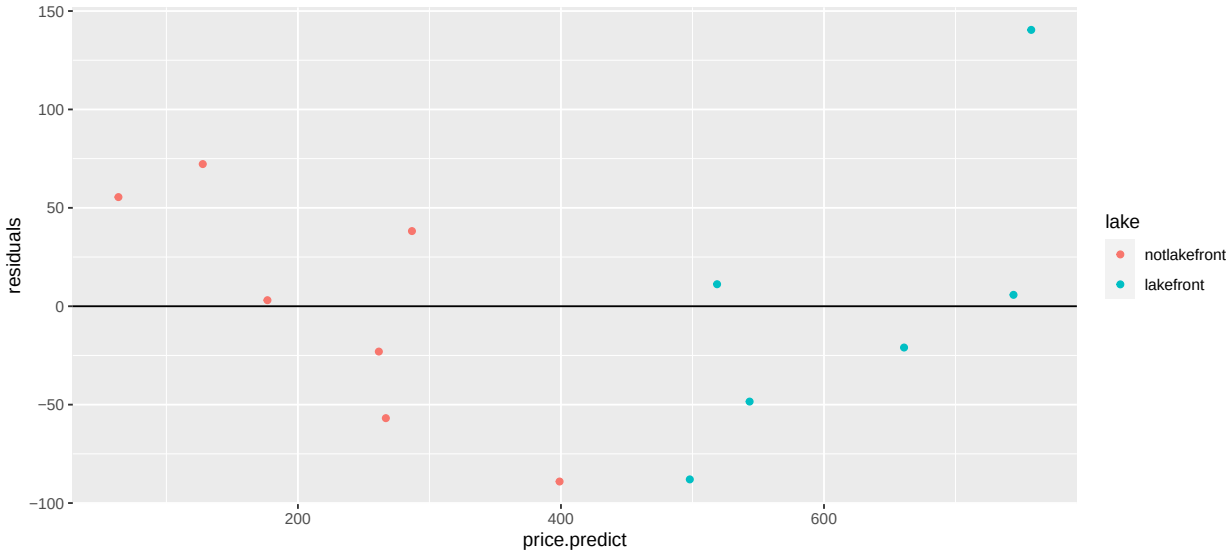
Let's take a look at the residuals vs the predicted (fitted) values.

```
resid_panel(model_withLake, plots = c('hist', 'resid'))
```



Do we see evidence of an interaction here?

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

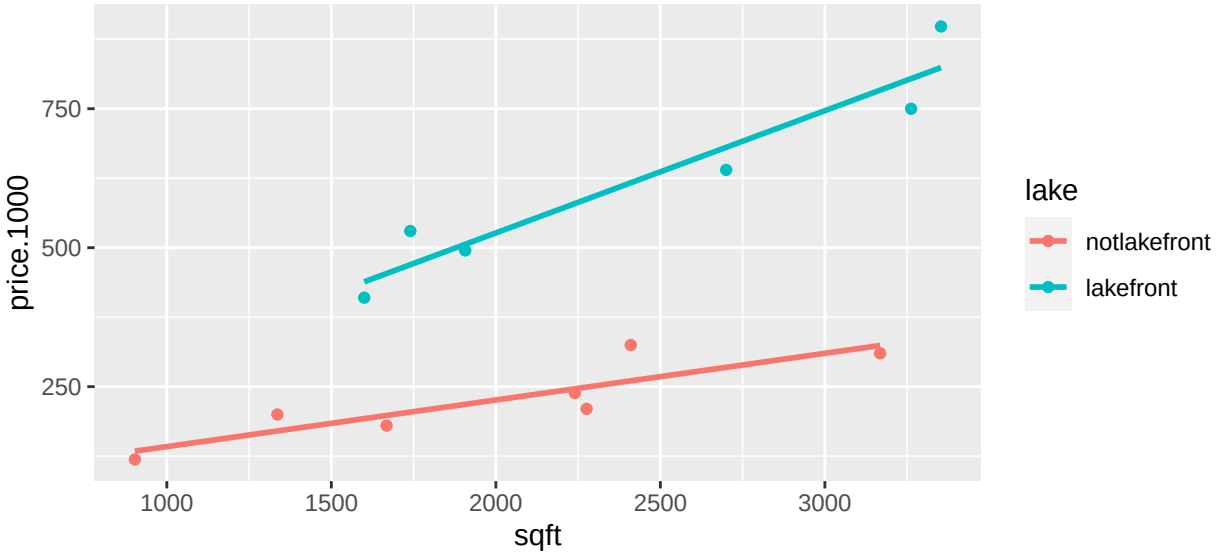


Yes, we tend to over underpredict smaller non lake homes and overpredict larger non lake homes while we overpredict smaller lake homes and underpredict larger lake homes.

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", se = F)
```




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

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = houses %>% filter(lake ==
##     "lakefront"))
##
## Residuals:
##      1      2      3      4      5      6
## -40.58  73.93 -28.60  60.52 -11.10 -54.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.76438   87.58078   0.991  0.37792
## sqft         0.21990    0.03462   6.352  0.00315 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.52 on 4 degrees of freedom
## Multiple R-squared:  0.9098, Adjusted R-squared:  0.8872
## F-statistic: 40.34 on 1 and 4 DF,  p-value: 0.003148
```

```
coef(lake_model)
```

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

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

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = houses %>% filter(lake ==
##     "notlakefront"))
##
## Residuals:
##       1       2       3       4       5       6       7
##  29.637  64.480 -14.915 -14.149 -18.233 -39.171  -7.649
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.11341   44.02459    1.32  0.24403
## sqft         0.08394    0.02078    4.04  0.00992 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.43 on 5 degrees of freedom
## Multiple R-squared:  0.7655, Adjusted R-squared:  0.7186
## F-statistic: 16.32 on 1 and 5 DF,  p-value: 0.009923
```

```
coef(nonlake_model)
```

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

How do we interpret our two slope coefficients?

The price per square foot for lakefront homes is \$219.89 and the price per square foot for non-lakefront homes is \$83.94

Interactions

Let's look at a model with an interaction.

$$\hat{price}_i = \beta_0 + \beta_1 sqft_i + \beta_2 lake_i + \beta_3 sqft_i lake_i$$

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

β_1 = the price per square foot of a non lakefront home

β_2 = the difference in price between a lakefront and non lakefront home with 0 square feet

β_3 = the difference in price per square foot between non lakefront and lakefront homes

$\beta_1 + \beta_3$ = the price per square foot of a lakefront home

β_0 = the average price for a non lakefront home with 0 square feet

```
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
```

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

$$\hat{price}_i = (58.11341 + 28.65098) + (0.08394 + 0.13595) \times sqft_i$$

$$=$$

$$\hat{price}_i = 86.76439 + (0.21989 \times sqft_i)$$

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

$$\hat{price}_i = 58.11341 + (0.08394 \times sqft_i)$$

How do these equations compare to the two models above?

The two equations are the same as the two individual models that we created earlier.

Why is the interaction model preferable?

If you look at the R^2 of the interaction model it is 0.9684 while the R^2 values of the two individual models are 0.9098 and 0.7655. We account for more of the variability in the response variable with the interaction model.

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

Yes, the interaction term in the model is significant.

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

We have a relatively good predictive model since we are accounting for almost 97% of the price of the homes. Both the square footage and interaction of square footage and location are significant in 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.

No, location is significant in that it modifies the affect of size.

Let's check our validity conditions for the interaction model

```
resid_panel(model_interaction, plots = c('hist', 'resid'))
```