Lesson 17 Michigan Housing Prices (4.3-4.4)

by LTC J. K. Starling

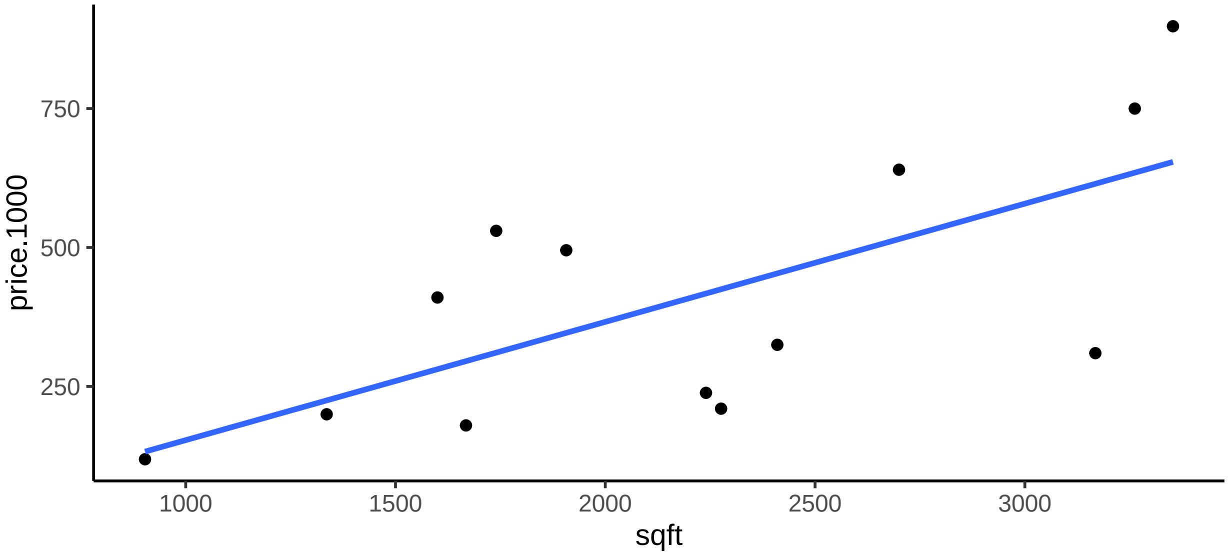
```
#Load Packages
library(tidyverse)
library(ggResidpanel)
library(car)
#read in data
houses = read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/houses.csv", stringsAsFactor
```

Background: 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.

1) What are the observational units? What is the response variable and which is the explanatory variable? Are they quantitative or categorical?

2) Plot the data and include a regression line. What do we learn from the regression line? Does a linear model appear to be reasonable?

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



3) Write out the appropriate linear model and fit the model in R. How do we interpret these estimated parameters? Are our validity conditions met?

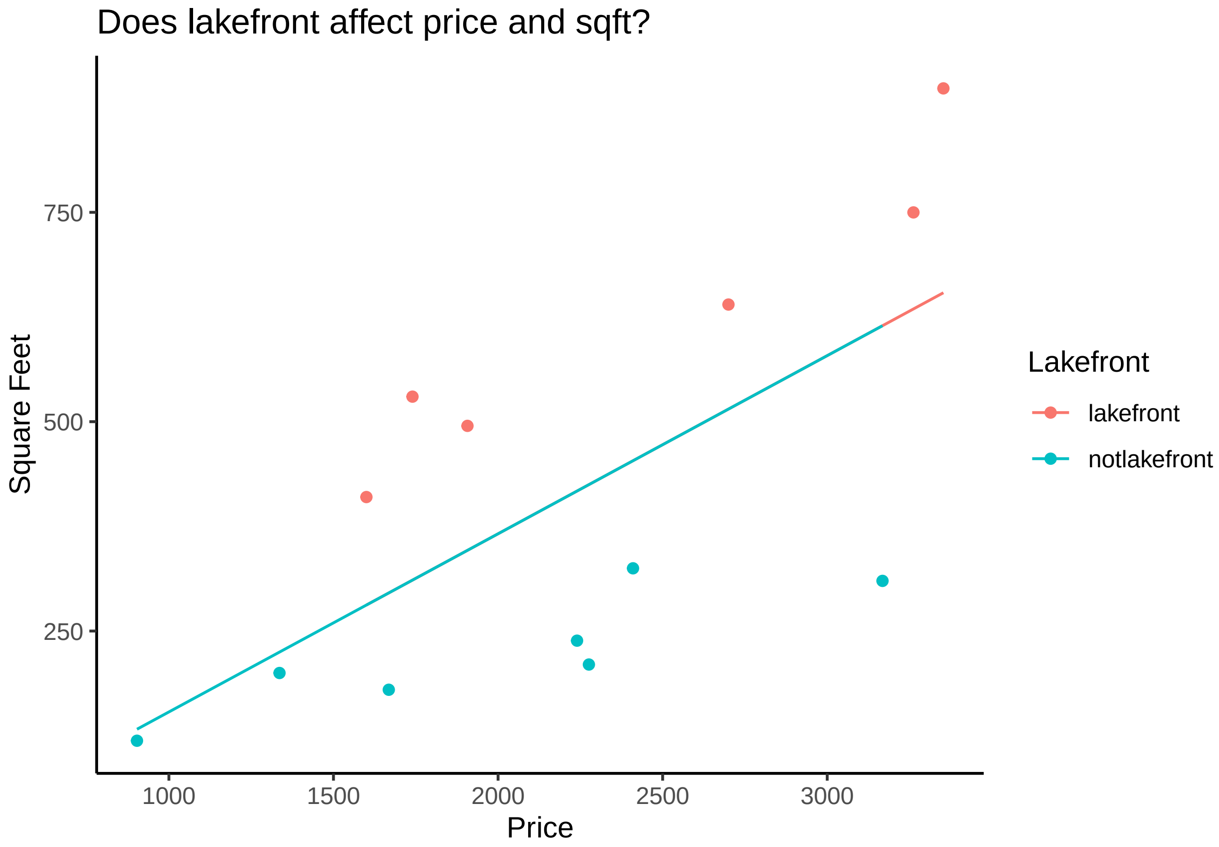
```r
model_simple <- houses %>% lm(price.1000 ~ sqft, data = .)
summary(model_simple)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = .)
##
## 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
```

4) 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?

Take a look at the figure. What do we notice? What would we hope to accomplish by adding this`lake` variable to the regression model?

```r
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") +
  theme_classic()
```



5) We can adjust for this categorical 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. Write out the updated regression model.

6) How do we interpret the coefficients? What assumptions does this model make? In general, when will β_1 from this model equal β_1 in the simple model?

$\beta_0 =$

$\beta_1 =$

$\beta_2 =$

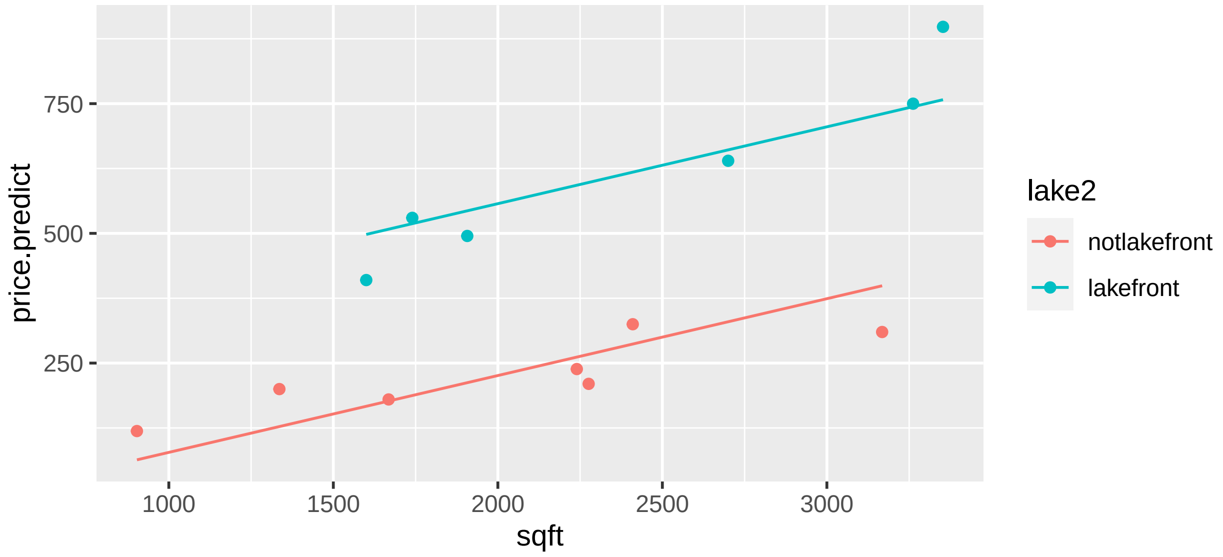
7) Consider the updated model below and the updated figure. How did the estimate of the price per square footage change when we held location constant?

```
#Reverse coding of lake so 1 is lake and 0 is not lake
houses$lake2 = factor(houses$lake, levels = c("notlakefront",
                                             "lakefront"))

model_withLake = lm(price.1000 ~ sqft + lake2, data = houses)

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

houses %>% ggplot(aes(x = sqft, y = price.predict, color = lake2)) +
  geom_line() +
  geom_point(aes(y = price.1000))
```



```
summary(model_withLake)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft + lake2, 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 ***
## lake2lakefront 331.2235    41.8470   7.915 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

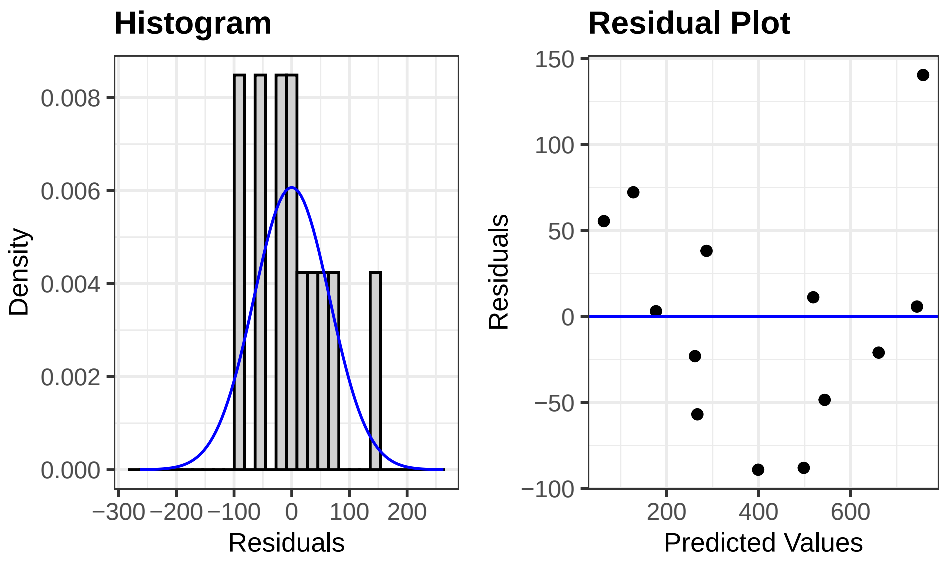
```
## Residual standard error: 72.02 on 10 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9106
## F-statistic: 62.15 on 2 and 10 DF,  p-value: 2.289e-06
```

8) What is our new regression model? 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. How do we interpret β_0?

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

9) Let's take a look at the residuals vs the predicted (fitted) values. Do we meet validity conditions?

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

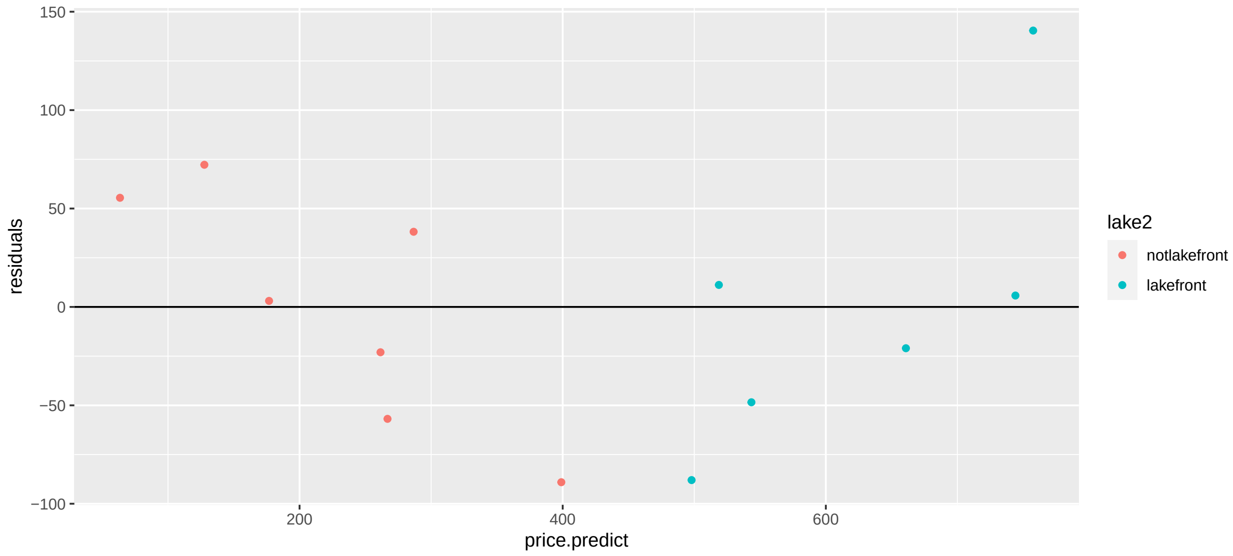


Histogram

Residual Plot

10) Do we see evidence of an interaction here?

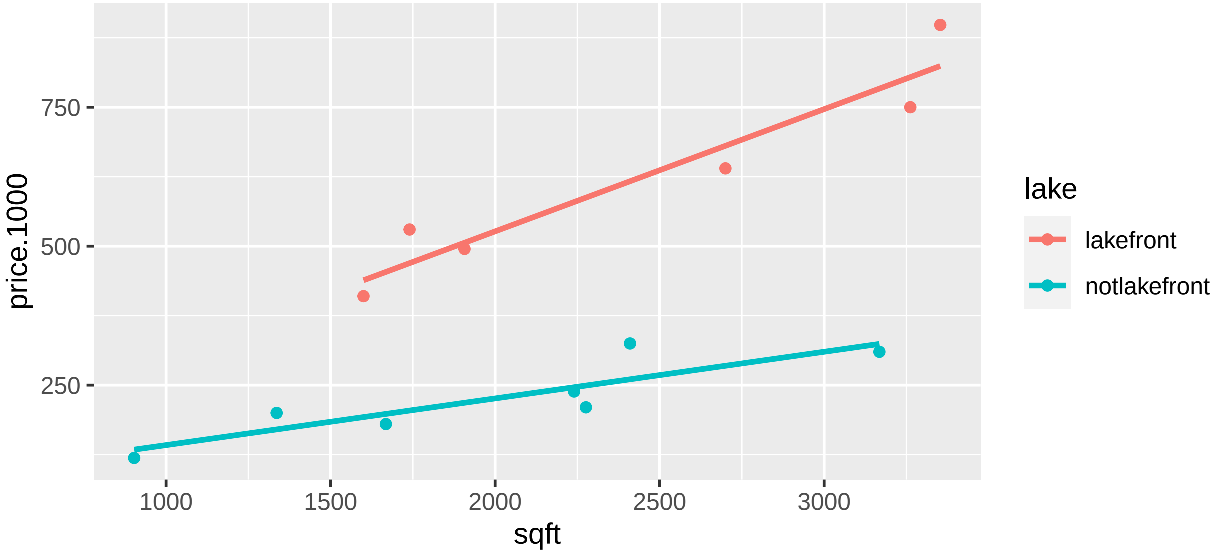
```
houses %>% ggplot(aes(x = price.predict,
                      y = residuals,
                      color = lake2)) +
  geom_point() + geom_hline(yintercept = 0)
```



Two models

11) If we suspect the price per square foot differs for lake front and nonlake front, we could fit two separate simple models. How do we interpret our two slope coefficients?

```
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
```

Interactions

12) Let’s look at a model with an interaction.

$$price_i = \beta_0 + \beta_1 sqft_i + \beta_2 lake_i + \beta_3 sqft_i lake_i + \epsilon_i$$

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

$\beta_0 =$

$\beta_1 =$

$\beta_2 =$

$\beta_3 =$

$\beta_1 + \beta_3 =$

```
model_interaction = lm(price.1000 ~ sqft * lake2, data = houses)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft * lake2, 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 *
## lake2lakefront        28.65098   91.32560   0.314  0.76088
## sqft:lake2lakefront    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
```

```
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 *
## lake2         241.0  1  0.0984 0.760881
## sqft:lake2  29829.0  1 12.1844 0.006824 **
## Residuals   22033.2  9
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

13) Write the equation for the price of a lake front house as a function of size and write the equation for the price of a nonlake house as a function of size.

14) How do these equations compare to the two models above? (#11)

15) Is there evidence the price per square foot is different for lakefront and nonlakefront homes? How do you know?“,‘blue’)‘

16) 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.

Comparing Models

17) Certainly our R^2 increased, but if we know square footage do we gain anything by knowing if the house is by the water or not? That is, we want to do a single test that evaluates whether the variable square footage is significant or not.

(Note that if we have multiple categories, if we test each category separately we are inflating the Type I error.)

Is the interaction model preferable?

18) We may also want to use something called the partial-F test to see if the addition of additional variables is beneficial. For this we go to the ANOVA to answer the question to ‘Is the interaction model preferable?'' First we will compare themodel_simpleto themodel_withLakeand then themodel_simpleto themodel_interaction‘ :

```
Anova(model_simple, type="3")
```

```
## Anova Table (Type III tests)
##
## Response: price.1000
##             Sum Sq Df F value  Pr(>F)
## (Intercept)   4636  1  0.1354 0.71992
## sqft        319753  1  9.3353 0.01094 *
## Residuals   376773 11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_withLake, type="3")
```

```
## Anova Table (Type III tests)
##
## Response: price.1000
##             Sum Sq Df F value    Pr(>F)
## (Intercept)   6476  1  1.2487 0.2899334
## sqft        142022  1 27.3845 0.0003826 ***
## lake2       324911  1 62.6489 1.293e-05 ***
## Residuals    51862 10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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 *
## lake2         241.0  1  0.0984 0.760881
## sqft:lake2  29829.0  1 12.1844 0.006824 **
## Residuals   22033.2  9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_simple, model_withLake)
```

```
## Analysis of Variance Table
##
## Model 1: price.1000 ~ sqft
## Model 2: price.1000 ~ sqft + lake2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     11 376773
## 2     10  51862  1    324911 62.649 1.293e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
anova(model_simple, model_interaction)
```

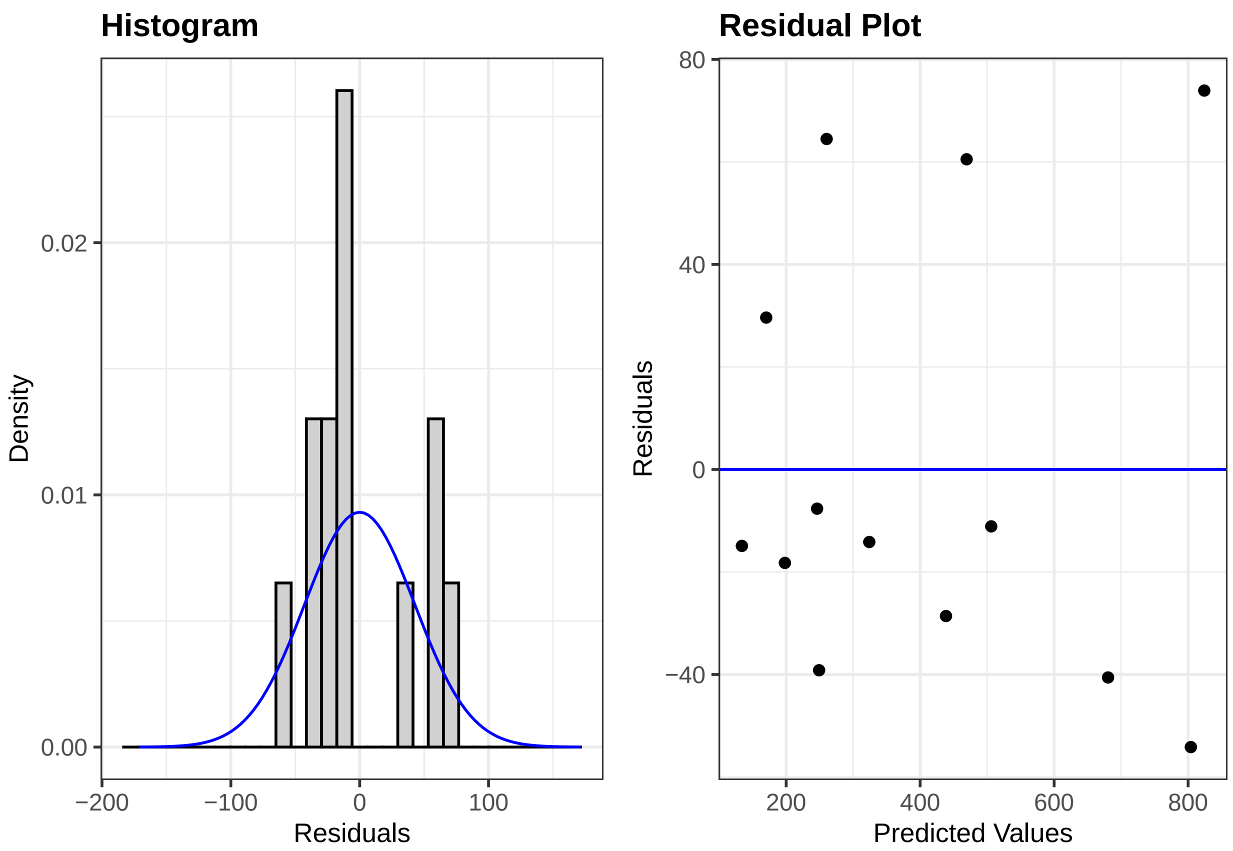
```
## Analysis of Variance Table
##
## Model 1: price.1000 ~ sqft
## Model 2: price.1000 ~ sqft * lake2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     11 376773
## 2      9  22033  2    354740 72.451 2.828e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
anova(model_withLake, model_interaction)
```

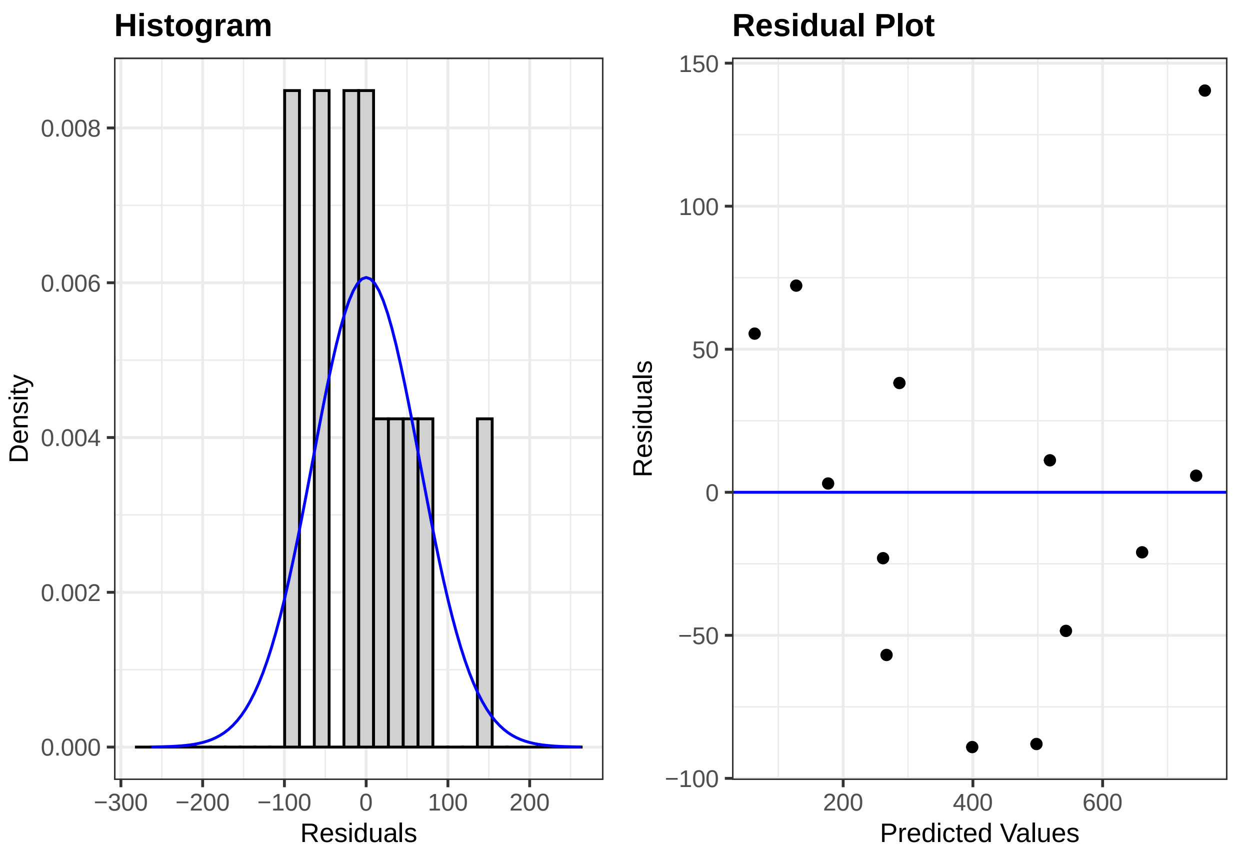
```
## Analysis of Variance Table
##
## Model 1: price.1000 ~ sqft + lake2
## Model 2: price.1000 ~ sqft * lake2
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     10 51862
## 2      9 22033  1     29829 12.184 0.006824 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

19) Now that we have a final model what should we check before we can trust our conclusions?

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



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


Histogram
Density
0.008
0.006
0.004
0.002
0.000
−300
−200
−100
0
100
200
Residuals
Residual Plot
Residuals
150
100
50
0
−50
−100
200
400
600
Predicted Values