MA256 Lesson 17 & 18 - Multiple Variables and Interactions (Causality and Multiple Regression Supplement Ch3 & 4)

Many people say that you can predict the cost of a home (in thousands of USD) based on its size (square footage) because bigger homes cost more to build, so I'd like for you to verify this claim using a random sample of homes around Lake Michigan (Houses.csv).

A1.) Identify and classify the variables of interest.

Explanatory: Home size (in square feet) - Quantitative

Response: Price (in thousands of USD) - Quantitative

A2.) Define the coefficient symbols and interpret what they represent (in context) given the simple linear regression equation in the form of $y = \beta_0 + \beta_1 * sqft$.

β_0 (Intercept Coefficient): the average price of a 0 sqft home (vacant lot).

β_1 (Slope Coefficient): the population slope is the predicted increase in average price (in thousands of USD) for every one-square-foot increase in home size.

A3.) Express the null and alternate hypotheses in words & symbols.

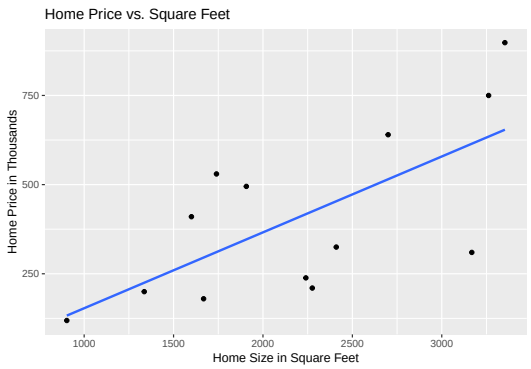
H_0 : $\beta_1 = 0$. There is no association between the size of a home and its price.

H_a : $\beta_1 \neq 0$. There is a positive association between the size of a home and its price.

A4.) Create a scatterplot of Home Size vs Price, and add a layer (`geom_smooth`) to include the best-fitting least squares regression line. Describe what you see.

```
houses %>% ggplot(aes(x = sqft, y = `price in thousands`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x="Home Size in Square Feet", y = "Home Price in Thousands",
       title = "Home Price vs. Square Feet")
```

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

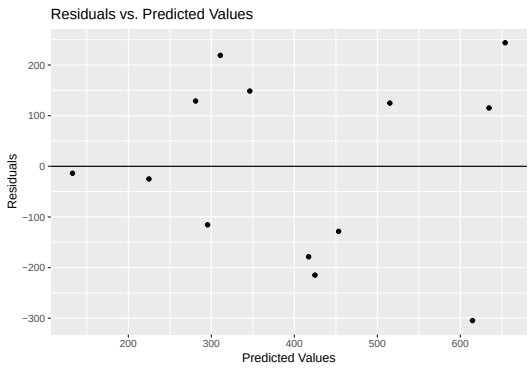


A5.) Use the Course Guide example R Code for Lesson 18-19 to find estimates of the fitted model and then write the complete equation for `house.lm` given the form of $y = \beta_0 + \beta_1 * sqft$. Are the 4 validity conditions met for us to use the theory-based approach to assess strength of evidence? Use headers to explicitly state each of them, and then clarify if they are met.

```
house.lm <- houses %>% lm(`price in thousands` ~ sqft, data = .)
summary(house.lm)
```

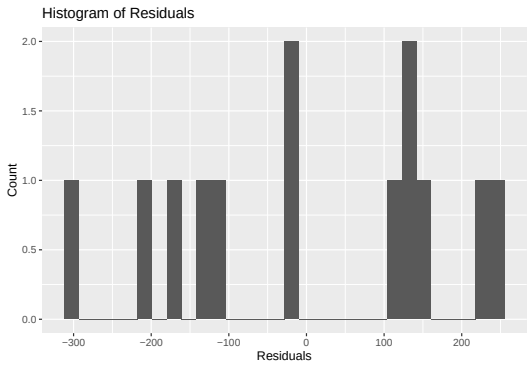
```
##
## Call:
## lm(formula = `price in thousands` ~ sqft, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.70 -128.44  -13.74  128.98  244.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.36870  161.36807  -0.368   0.7199
## 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
```

```
house.lm %>%
  fortify(house.lm$model) %>%  # fortify converts a model to a dataframe
  ggplot(aes(x = .fitted,
             y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs( x = "Predicted Values",
        y = "Residuals",
        title = "Residuals vs. Predicted Values")
```




```
house.lm %>%
  fortify(house.lm$model) %>%
  ggplot(aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals", y = "Count",
       title = "Histogram of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Price= -59.36870 + 0.21274*sqft

Linearity: The data seems to generally follow a linear trend.

Independence (New): It makes sense the home prices are independent from one another e.g. if my neighbor's house has ugly paint inside that won't influence my house price – or – if my neighbor's house only has 2 bathrooms but mine has 4 that won't effect my home price.

Normality: Equal number of observations above (6) and below (7) the line.

Equal Variance: There is an approximately equal variance throughout the entire domain, given flexibility due to small sample size.

A6.) Interpret your slope and intercept coefficients:

Slope Coefficient: The increase in average price (in thousands of USD) for every 1 sqft increase in home size is 0.21274.

Alternative Interpretation: The increase in average price for every 1 square foot increase in home size is $212.74 (0.21274 * 1,000 since price is in thousands).

Intercept Coefficient: The average price (in thousands of USD) of a 0 square foot home is -59.36870.

Alternative Interpretation: The average price of a 0 square foot home is -$59,368.70 (-59.36870 * 1,000 since price is in thousands).

A7.) Would you be comfortable using your model to predict the price of a 0 sqft home (vacant lot)? Why or why not?

Since positive numbers indicate money flowing from a buyer to a seller to purchase their property, and since the intercept coefficient is negative, it doesn't make sense that someone would pay me $59,368.70 to "purchase" a vacant lot i.e. 0 square foot home. I am not comfortable using my model in this way. WE MUST ALWAYS USE CAUTION WHEN EXTRAPOLATING WITH OUR MODEL.

A7.) If we want to predict prices for vacant lots, what should we use to create a model?

If we wanted to predict the prices of vacant lots, perhaps we should only use vacant lots to create a new least squares regression model.

A8.) Calculate the 95% confidence interval for the slope statistic and provide an interpretation of the CI.

```r
sqft.coef <- summary(house.lm)$coefficients['sqft','Estimate']
sqft.se <- summary(house.lm)$coefficients['sqft','Std. Error']
multiplier <- qt(0.975, df = 11)
c(sqft.coef - multiplier*sqft.se, sqft.coef + multiplier*sqft.se)
```

```
## [1] 0.05949022 0.36599682
```

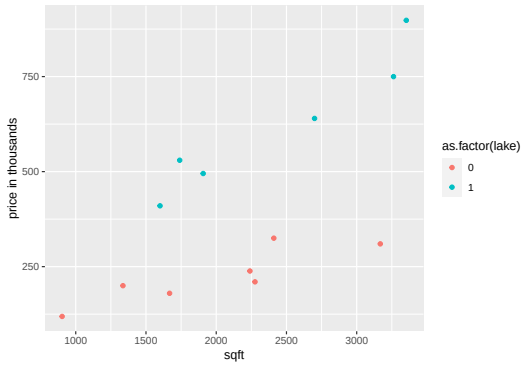
The 95% CI = Statistic ± mult*SD. Since the slope coefficient (statistic) is 0.21274, since its Standard Error (SE) is 0.06963, and the multiplier is approx 2.2, the 95% CI can be calculated as follows:
Lower Bound: 0.21274 – (2.2*0.06963) = 0.0595
Upper Bound: 0.21274 + (2.2*0.06963) = 0.3660
95% CI: (0.0595, 0.3660)
I am 95% confident that a 1 sqft increase in home size is associated with an average increase of 0.0595 thousands of USD to 0.3660 thousands of USD in home price for homes around Lake Michigan.

A9.) Briefly explain why location i.e. lakefront (=1) or non-lakefront (=0), may be another explanatory variable.

Lake houses are traditionally more expensive than non-lakefront houses. Lake impacts the response variables.

A10.) Does lakefront effect price and sqft? Create a new scatterplot that is colored based on lakefront status.

```r
houses %>% ggplot(aes(x = sqft, y = `price in thousands`, color = as.factor(lake))) +
  geom_point()
```



Based on the scatterplot results, it seems like even the smallest lakefront homes are more expensive than the largest non-lakefront homes that are twice the size (in sqft).

A11.) We can adjust for the additional explanatory 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.

$price_i = \beta_0 + \beta_1 * sqft_i + \beta_2 * lake_i$

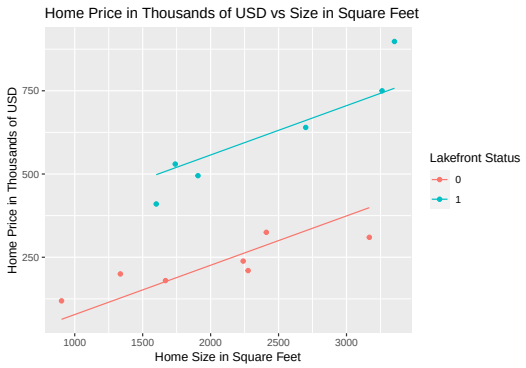
where $price_i$ is the price of house i, $sqft_i$ is the size (sq ft) of house i, and $lake_i$ is 1 if house i is lakefront and 0 if not lakefront. It may help to think of "i" as a counter from house #1 to #13

Create a new model with lake as an additional variable to create a scatter plot to view the new model.

```r
house.lake.lm <- houses %>% lm(`price in thousands` ~ sqft + lake, data = .)

houses <- houses %>% mutate(price.predict = predict(house.lake.lm, newdata = .),
             residuals = `price in thousands` - price.predict)

houses %>% ggplot(aes(x = sqft, y = `price in thousands`, color = as.factor(lake))) +
  geom_line(aes(x=sqft, y=house.lake.lm$fitted.values)) +
  geom_point(aes(y = `price in thousands`)) +
  labs(x="Home Size in Square Feet", y="Home Price in Thousands of USD",
            color="Lakefront Status",
            title="Home Price in Thousands of USD vs Size in Square Feet")
```



A12.) Based on the slopes of the two lines, what assumption does this model make about the price per square foot?

The slope for both the lakefront and non-lakefront homes is the same i.e. the price per square foot is the same regardless of whether a home is on the lake (no "interaction").

A13.) How do we interpret $\beta_0, \beta_1, \beta_2$ given the regression line equation in the form of $price_i = \beta_0 + \beta_1 * sqft_i + \beta_2 * lake_i$?

β_0: The price of a 0 square foot home (vacant lot) that is not on the lake.

β_1: The change in average price of a home for every 1 square foot change in home size after adjusting for location (lakefront or not lakefront).

β_2: The increase in average price because a home is lakefront after adjusting for size.

A14.) What is our new regression model with lakefront status as an additional variable in the form of $price_i = \beta_0 + \beta_1 * sqft_i + \beta_2 * lake_i$?

```r
summary(house.lake.lm)
```

```
##
## Call:
## lm(formula = `price in thousands` ~ sqft + lake, data = .)
##
## 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 ***
## lake        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
```

$Price_i = -70.1821 + 0.1481 * sqft_i + 331.2235 * lake_i$

A15.) Calculate Prices of 2500 sqft houses on the lake and not on the lake.

```
predict(house.lake.lm, data.frame(sqft = 2500, lake = 1))
```

```
##        1
## 631.2604
```

```
predict(house.lake.lm, data.frame(sqft = 2500, lake = 0))
```

```
##        1
## 300.0369
```

2500 sqft Lakefront: -70.1821 + 0.1481*(2500) + 331.2235*(1) = $631,260.40

2500 sqft Non Lakefront: -70.1821 + 0.1481*(2500) + 331.2235*(0) = $300,036.90

A16.) How did the price per square foot (slope) change compared to when we held location constant?

The price per square foot decreased from $212.74 (0.21274) to $148.10 (0.1481) because the majority of the price difference was due to the lakefront premium.

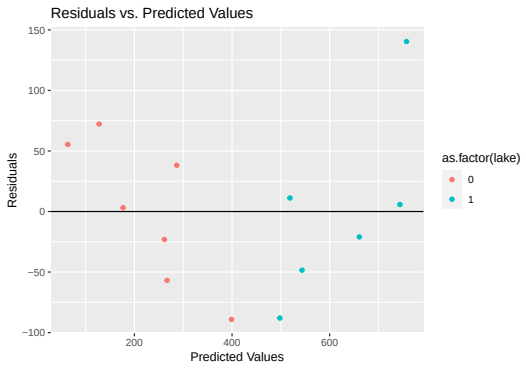
A17.) Which model predicts home price more accurately?

The first model had a coefficient of determination (R Squared) of 0.4591 so it accounted for 45.91% of the variance in Price due to size. The second model has an R Squared of 0.9255 so it accounted for 92.55% in home price due to size and location.

A18.) Use the Course Guide example R Code for Lesson 18-19 as a template to plot the residuals vs predicted values to determine whether there is evidence of an interaction?

```
house.lake.lm %>%
  fortify(house.lake.lm$model) %>%  # fortify converts a model to a dataframe
  ggplot(aes(x = .fitted,
             y = .resid,
             color = as.factor(lake))) +
  geom_point() +
```

```
  geom_hline(yintercept = 0) +
  labs( x = "Predicted Values",
        y = "Residuals",
        title = "Residuals vs. Predicted Values")
```



For non-lakefront homes, the trend in the residuals (Predicted – Actual) shows that the slope of the least-squares regression line overestimates home prices initially (lower sqft homes), and then underestimates them for larger sqft homes.
For lakefront homes, the trend in residuals underestimates home prices initially (lower sqft homes), and overestimates prices for larger sqft homes.
Therefore, adding an interaction term will allow the slopes to be individually tailored to each set of data rather than assuming the price per square foot would be the same regardless of location. This is further validated by looking at this following diagram, and recognizing that the best fit line for blue dots by themselves would have a more positive slope, while the best fit line for red dots would be less positive.

A19.) Create and provide the new model that includes an interaction between lake and sqft and takes the form $price_i = \beta_0 + \beta_1 * sqft_i + \beta_2 * lake_i + \beta_3 * sqft_i * lake_i$?

```
house.int.lm <- houses %>% lm(`price in thousands` ~ sqft * lake, data = .)
summary(house.int.lm)
```

```
##
## Call:
## lm(formula = `price in thousands` ~ sqft * lake, data = .)
##
## 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 *
## lake        28.65098   91.32560   0.314  0.76088
## sqft:lake    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
```

$Price_i = 58.11341 + 0.08394 * sqft_i + 28.65098 * lake_i + 0.13595 * sqft_i * lake_i$.

A20.) Which of the 3 models had the highest accuracy?

The interaction model had an R-Squared (Coefficient of Determination) value of 0.9684 so it accounts for 96.84% of the variance in home price that is attributed to sqft and location.

A21.) Create a scatterplot that shows the predicted linear models for those houses on a lake and those away from it using the interaction model.

```
houses%>%
 ggplot(aes(x=sqft,y=`price in thousands`, color=as.factor(lake)))+
 geom_point()+
 geom_smooth(method = "lm")+
 labs(x="Home Size in Square Feet", y="Home Price in Thousands of USD",
    title="Home Price in Thousands of USD vs Size in Square Feet")
```

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

A22.) How do we interpret the two slope coefficients associated with the size of a home i.e. sqft and sqft*lake given the least squared regression equation of:

$price_i = 58.11341 + 0.08394 * sqft_i + 28.65098 * lake_i + 0.13595 * sqft_i * lake_i.$

The price per square foot for non-lakefront homes is simply 0.08394*sqft = \$83.94 per sqft because lake = 0 which eliminates the second sqft term. However, when lake=1 (lakefront homes), the price per square foot becomes (0.08394 + 0.13595)*sqft = 0.21989*sqft = \$219.89 per sqft.

A23.) Is there evidence that the price per square foot is different for lakefront and nonlakefront homes? How do you know?

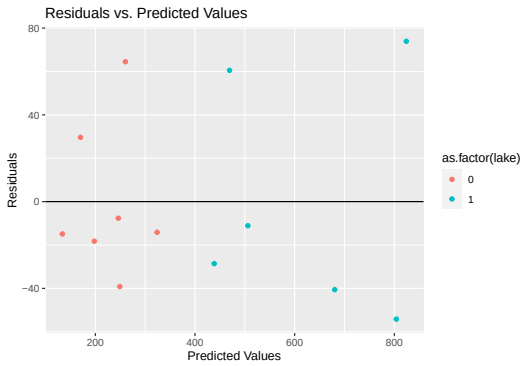
Yes. The p-value associated with the interaction term is 0.00682, which is statistically significant because it provides very strong evidence against the null hypothesis of no interaction term, as long as the validity conditions are met for this model.

A24.) The main effect of location (by itself) is not significant in the model with an interaction. Should we conclude that the location of the house is not associated with price? Justify your answer.

No. Location is significant in that it modifies the effect of size through the interaction.

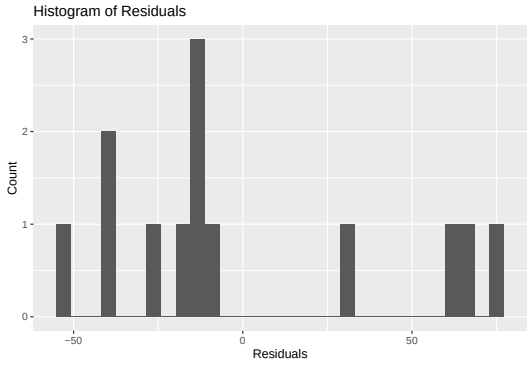
A25.) Use the Course Guide example R Code for Lesson 18-19 as a template to assess whether the model that includes lakefront status meets the four validity conditions (LINE):

```
house.int.lm %>%
  fortify(house.int.lm$model) %>%  # fortify converts a model to a dataframe
  ggplot(aes(x = .fitted,
             y = .resid,
             color = as.factor(lake))) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs( x = "Predicted Values",
        y = "Residuals",
        title = "Residuals vs. Predicted Values")
```




```
house.int.lm %>%
  fortify(house.int.lm$model) %>%
  ggplot(aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals", y = "Count",
       title = "Histogram of Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Line & Equal Variance assessed in one step by comparing predicted values & residuals: There are no trends in the residuals vs predicted values data so we meet the linear validity condition criteria. Also, the variance appears to be approximately even above and below the line for the entire domain, so we meet that validity condition also.

Independence assessed via context of the problem:
We use context of the problem to identify if random sampling or random assignment occurred. When there is no "random" anything, with some critical thinking, it will become apparent whether the observations are independent from one another.

Normality assessed via histogram of residuals: The data appears to be centered just below data with wide tails on both sides of the center, which means it is approximately normal since there are no significant outliers strongly skewing the histogram.