MA256 Lesson 18 & 19 - 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:

Response:

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

β_1

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

H_0 :

H_a :

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(.....
#   labs(x="Home Size in Square Feet", y = "Home Price in Thousands",
#        title = "Home Price vs. Square Feet")
```

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 4x 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(....)
# summary(house.lm)
#
# house.lm %>%
#   fortify(....
#   labs( x = "Predicted Values",
#         y = "Residuals",
#         title = "Residuals vs. Predicted Values")
#
# house.lm %>%
#   fortify(....
#   labs(x = "Residuals", y = "Count",
#        title = "Histogram of Residuals")
```

Linearity:

Independence (New):

Normality:

Equal Variance:

A6.) Interpret your slope and intercept coefficients:

Slope Coefficient:

Alternative Interpretation:

Intercept Coefficient:

Alternative Interpretation:

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

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

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

```
# XXX
```

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

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

```
# XXX
```

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.

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

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:

β_1:

β_2:

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

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

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

A17.) Which model predicts home price more accurately?

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(....)
#   labs( x = "Predicted Values",
#         y = "Residuals",
#         title = "Residuals vs. Predicted Values")
```

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(....)
# summary(house.int.lm)
```

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

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(...
#  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")
```

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:

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

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

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.

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(....
#   labs( x = "Predicted Values",
#         y = "Residuals",
#         title = "Residuals vs. Predicted Values")
#
# house.int.lm %>%
#   fortify(...
#   labs(x = "Residuals", y = "Count",
#        title = "Histogram of Residuals")
```