## MECHANICS AND COMPUTATIONS

### MODEL #1

ANOVA:

|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| X1 | 1 | 1974.53 | 1974.53 | 209.8340 | <0.0001 |
| X2 | 1 | 118.8642568 | 118.8642568 | 12.6339 | 0.0007 |
| X3 | 1 | 32.47012585 | 32.47012585 | 3.4512 | 0.0676 |
| X4 | 1 | 0.435606985 | 0.435606985 | 0.0463 | 0.8303 |
| Residuals | 67 | 630.36 | 9.41 |  |  |
|  |  |  |  |  |  |
| Note: You can make the following calculations from the ANOVA table above to get Overall F statistic | | | | | |
| Model (adding 4 rows) | 4 | 2126 | 531.50 | <0.0001 | |
| Total (adding all rows) | 71 | 2756.37 | | | |

Coefficients:

|  | Estimate | Std. Error | t value | Pr(>t) |
|---|---|---|---|---|
| Intercept | 11.3303 | 1.9941 | 5.68 | <.0001 |
| X1 | 2.186 | 0.4104 |  | <.0001 |
| X2 | 8.2743 | 2.3391 | 3.54 | 0.0007 |
| X3 | 0.49182 | 0.2647 | 1.86 | 0.0676 |
| X4 | -0.49356 | 2.2943 | -0.22 | 0.8303 |

| Residual standard error: 3.06730 on 67 degrees of freedom |
|---|
| Multiple R-sqaured: 0.7713,   Adjusted R-squared: 0.7577 |
| F-statistic:     on 4 and 67 DF,  p-value < 0.0001 |

| Number of predictors | C(p) | R-square | AIC | BIC | Variables in the model |
|---|---|---|---|---|---|
| 4 | 5 | 0.7713 | 166.2129 | 168.9481 | X1 X2 X3 X4 |

(1) How many observations are in the sample data?

Observations = Total + 1 = df (67) + k (4) + 1 = **72**

(2) Write out the null and alternate hypotheses for the t-test for Beta1.

$H_o : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

(3) Compute the t- statistic for Beta1.  Conduct the hypothesis test and interpret the result.

$t_1 = \hat{B}_1 / S_{\hat{B}_1}$ = 2.186 / 0.4104 = 5.3265
t-test with 99% confidence ($\alpha$ = 0.01),
Threshold: $t_{\alpha/2, n-p-2} = t_{0.005, 66}$ = 2.6524

**Reject** $H_o$, since $|t_o| > t_{0.005, 66}$

There is insufficient evidence to accept the null hypothesis, $\beta 1 = 0$, therefore $X_1$ is a valid indicator of Y in this model and therefore should be included in the model.

*(4) Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.*

Sum of Squares due to **Regression** = 1974.53 + 118.8643 + 32.4701 + 0.4356 = 2126.3 = $SS_R$
Sum of Squared **Error** = 630.36 = $SS_E$
Sum of Squares **Total** = $SS_T = SS_R + SS_E$

$R^2 = SS_R / SS_T = 2126.3 / 2756.66 = \mathbf{0.7713}$

The total / "global" proportion of variation explained by the regression model, model 1, is 77.13%.

*(5) Compute the Adjusted R-Squared value for Model 1. Discuss why Adjusted R-squared and the R-squared values are different.*

Let,
   n = 72, $R^2 = 0.7713$, k = 4
**Adjusted $R^2$** = $1 - [(1 - R^2)(n - 1) / (n - k - 1)] = \mathbf{0.7577}$

The adjusted $R^2$ statistic penalizes the model for adding independent / predictor variables to the model that don't have relevance in predicting the response variable. The adjusted $R^2$ term is the proportion of variance explained by the relevant terms in the model.

*(6) Write out the null and alternate hypotheses for the Overall F-test.*

$H_o : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
$H_a : \beta_j \neq 0$, for at least one value of j (for j in 1, 2, 3, 4)

*(7) Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.*

Sum of Squares due to **Regression** = $SS_R$ = 1974.53 + 118.8642568 + 32.47012585 + 0.435606985
Sum of Squared **Error** = $SS_E$ = 630.36
Sum of Squares **Total** = $SS_T = SS_R + SS_E$ = 2756.66

Let,
   N = 72, p = 4

F = $[(SS_T - SS_E) / p] / [SS_E / (n - p - 1)]$ = 531.575 / 9.4084 = **56.5003** on p = 4 and 67 DF
p-value: < 0.0001

There is insufficient evidence (F = 56.5003, P < 0.001) to conclude that at least one of the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone.

ANOVA:

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| X1 | 1 | 1928.27000 | 1928.27000 | 218.8890 | <.0001 |
| X2 | 1 | 136.92075 | 136.92075 | 15.5426 | 0.0002 |
| X3 | 1 | 40.75872 | 40.75872 | 4.6267 | 0.0352 |
| X4 | 1 | 0.16736 | 0.16736 | 0.0190 | 0.8908 |
| X5 | 1 | 54.77667 | 54.77667 | 6.2180 | 0.0152 |
| X6 | 1 | 22.86647 | 22.86647 | 2.5957 | 0.112 |
| Residuals | 65 | 572.60910 | 8.80937 | | |
| | | | | | |
| Note: You can make the following calculations from the ANOVA table above to get Overall F statistic | | | | | |
| Model (adding 6 rows) | 6 | 2183.75946 | 363.96 | 41.3200 | <0.0001 |
| Total (adding all rows) | 71 | 2756.37 | | | |

Coefficients:

| | Estimate | Std. Error | t value | Pr(>t) |
|---|---|---|---|---|
| Intercept | 14.3902 | 2.89157 | 4.98 | <.0001 |
| X1 | 1.97132 | 0.43653 | 4.52 | <.0001 |
| X2 | 9.13895 | 2.30071 | 3.97 | 0.0002 |
| X3 | 0.56485 | 0.26266 | 2.15 | 0.0352 |
| X4 | 0.33371 | 2.42131 | 0.14 | 0.8908 |
| X5 | 1.90698 | 0.76459 | 2.49 | 0.0152 |
| X6 | -1.0433 | 0.64759 | -1.61 | 0.112 |

Residual standard error: 2.968 on 65 degrees of freedom

Multiple R-sqaured: 0.7923,    Adjusted R-squared: 0.7731

F-statistic: 41.32 on 6 and 65 DF,  p-value < 0.0001

| Number of predictors | C(p) | R-square | AIC | BIC | Variables in the model |
|---|---|---|---|---|---|
| 6 | 7 | 0.7923 | 163.2947 | 166.7792 | X1 X2 X3 X4 X5 X6 |

8.) *Now let's consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.*

Model 1 nests Model 2 in this situation. Model 2 has additional explanatory variables that are not considered in Model 1, and all the variables considered with Model 1 are also considered with Model 2. Model 2 is a superset of Model 1, adding variables $X_5$ and $X_6$ for evaluation.

9.) *Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.*

$H_o : \beta_5 = \beta_6 = 0$

$H_a : \beta_j \neq 0$, for at least one value of j (for j in 5,6)

10.) *Compute the F-statistic for a nested F-test using Model 1 and Model 2. Conduct the hypothesis test and interpret the results.*

$$F = (SSE_R - SSE_C) / SSE_C / [n - (k + p + 1)]$$

3

$$F = (630.36 - 572.3091) / [\, 572.3091 / [\, 72 - (6 + 1)]$$

$$= 58.30509 / 8.8048$$

$$= 6.5931$$

Critical value at 99% confidence ($\alpha$ = 0.01), % confidence, = $F_{99, 4, 65}$ = **3.6223**

The given F-statistic yielded a value of **6.5932** and at 99% confidence, we should reject the null hypothesis that the more complex, or complete, model 2 with the additional explanatory variables $\beta_5$ and $\beta_6$ is no more powerful than the reduced model 1.

## APPLICATION

## MODEL 3:

11.) *Based on your EDA from Modeling Assignment #1, focus on 10 of the continuous quantitative variables that you though/think might be good explanatory variables for SALESPRICE. Is there a way to logically group those variables into 2 or more sets of explanatory variables? For example, some variables might be strictly about size while others might be about quality. Separate the 10 explanatory variables into at least 2 sets of variables. Describe why you created this separation. A set must contain at least 2 variables.*

Quality variables:

- Overall Quality
- Quality Index

High Value Features:

- Garage Cars
- Garage Area
- Full Bath + Half Bath (Total Bath)
- Mas Vnr Area
- Fireplaces

Temporal:

- House Age
- Year Remodel
- Year Built

Housing Lot

- Lot Area
- Lot Frontage

This grouping of variables provides a way to consider the way a given set of features should relatively perform against the sale price of the home. For example, in the high value feature set we look at mostly discrete values (the exception being Garage Area) that should all have a positive correlation to the sale price of a given home. Likewise, for the temporal values, we would intuitively expect

*12.) Pick one of the sets of explanatory variables.   Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y).   Call this Model 3.   Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:*

   *a)  all model coefficients individually*

   *Let* $\beta_1$ = House Age

   $H_o : \beta_1 = 0$
   $H_a : \beta_1 \neq 0$

   $t_1 = \hat{B}_1 / S_{\hat{B}_1} = -2291.54 / 1002.40$ **= -2.2861**

   t-test with 95% confidence ($\alpha$ = 0.05),

   Threshold: $t_{\alpha/2, n-k-2} = t_{0.025, 2420} = 1.9609$

   *abs(T)* = 2.2861 > 1.9609

There is not significant evidence here to suggest that $\beta_1$, House Age, has no impact on explaining the variance in sale prices amongst homes.

   *Let* $\beta_2$ = Year Built

   $H_o : \beta_2 = 0$
   $H_a : \beta_2 \neq 0$

   $t_2 = \hat{B}_2 / S_{\hat{B}_2} = -1172.34 / 1006.12 =$ **-1.1652**

   t-test with 95% confidence ($\alpha$ = 0.05),

   Threshold: $t_{\alpha/2, n-k-2} = t_{0.025, 2420} = 1.9609$

   *abs(T)* = 1.1652 < 1.9609

There is evidence here that supports the claim that $\beta_2$ is indeed zero when used with the house age variable, therefore we can exclude the year built variable from the model.

   *Let* $\beta_3$ = Year Remodel

   $H_o : \beta_3 = 0$
   $H_a : \beta_3 \neq 0$

   $t_3 = \hat{B}_3 / S_{\hat{B}_3} = 1108.88 / 76.64 =$ **14.4687**

   t-test with 95% confidence ($\alpha$ = 0.05),

   Threshold: $t_{\alpha/2, n-k-2} = t_{0.025, 2420} = 1.9609$

   *abs(T)* = 14.4687 > 1.9609

There is insufficient evidence here to supports the claim that $\beta_3$ is indeed zero when used with the house age variable, therefore should continue to include the year remodel variable in the model as it explains further variance of sale price.

*b) the Omnibus Overall F-test*

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$H_a : \beta_j \neq 0$, for at least one value of j (for j in 1, 2, 3)

**F-Statistic:**

Sum of Squares due to **Regression** = $SS_R$ = 6,595,944,447,461
Sum of Squared **Error** = $SS_E$ = 10,031,364,633,591
Sum of Squares **Total** = $SS_T$ = $SS_R$ + $SS_E$ = 16,627,309,081,052

Let,

　　　N = 2425, p = 3

$F = [(SS_T - SS_E) / p] / [SS_E / (n - p - 1)]$ = 2,198,648,149,154 / 4,143,479,816 = **530.6284** on p = 3 and 2421 DF
p-value: < 0.0001

```
Call:
lm(formula = SalePrice ~ HouseAge + YearBuilt + YearRemodel,
    data = data.m3)

Residuals:
    Min      1Q  Median      3Q     Max
-169517  -39487  -10854   22995  528903

Coefficients:
              Estimate Std. Error t value          Pr(>|t|)
(Intercept)  381312.30 2011929.83   0.190            0.8497
HouseAge       -2291.54    1002.40  -2.286            0.0223 *
YearBuilt      -1172.34    1006.12  -1.165            0.2440
YearRemodel     1108.88      76.64  14.469 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64370 on 2421 degrees of freedom
Multiple R-squared:  0.3967,  Adjusted R-squared:  0.3959
F-statistic: 530.6 on 3 and 2421 DF,  p-value: < 0.00000000000000022
```

There is insufficient evidence (F = 530.6284, P < 0.001) to conclude that at least one of the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone.

*(13) Pick the other set (or one of the other sets) of explanatory variables. Add this set of variables to those in Model 3. In other words, Model 3 should be nested within Model 4. Run a multiple regression model using the explanatory variables from this set to predict SALEPRICE(Y). Conduct and interpret the following hypothesis tests, being sure you clearly state the null and alternative hypotheses in each case:*

    *a) all model coefficients individually*

*Let* $\beta_1$ = House Age

$$H_o : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

$t_1 = \hat{B}_1 / S_{\hat{B}_1} = $ -1172.86 / 733.18 **= -1.5997**

t-test with 95% confidence ($\alpha$ = 0.05),

Threshold: $t_{\alpha/2, \, n-k-2} = t_{0.025, \, 2420} = 1.9609$

*abs(T) = 1.5997 < 1.9609*

There is significant evidence here to suggest that House Age has no impact on explaining the variance in sale prices amongst homes, therefore we cannot reject the null hypothesis that the coefficient in question is zero.

*Let* $\beta_2$ = Year Built

$$H_o : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

$t_2 = \hat{B}_2 / S_{\hat{B}_2} = $ -1172.34 / 1006.12 = **-1.1652**

t-test with 95% confidence ($\alpha$ = 0.05),

Threshold: $t_{\alpha/2, \, n-k-2} = t_{0.025, \, 2420} = 1.9609$

*abs(T) = 1.1652 < 1.9609*

There is evidence here that supports the claim that $\beta2$ is indeed zero when used with the house age variable, therefore we can exclude the year built variable from the model.

*Let* $\beta_3$ = Year Remodel

$$H_o : \beta_3 = 0$$
$$H_a : \beta_3 \neq 0$$

$t_3 = \hat{B}_3 / S_{\hat{B}_3} = $ 287.52 / 62.53 = **4.5981**

t-test with 95% confidence ($\alpha$ = 0.05),

Threshold: $t_{\alpha/2, \, n-k-2} = t_{0.025, \, 2420} = 1.9609$

*abs(T) = 4.5981 > 1.9609*

There is insufficient evidence here to supports the claim that $\beta_3$ is indeed zero when used with the house age and year-built variables, therefore should continue to include the year remodel variable in the model as it explains further variance of sale price.

*Let* $\beta_4$ = Overall Quality

$$H_o : \beta_4 = 0$$
$$H_a : \beta_4 \neq 0$$

$$t_4 = \hat{B}_4 / S_{\hat{B}_4} = 41471.70 / 1317.60 = \mathbf{31.4752}$$

t-test with 95% confidence ($\alpha$ = 0.05),

Threshold: $t_{\alpha/2, n-k-2} = t_{0.025, 2420} = 1.9609$

*abs(T)* = 31.4752 > 1.9609

There is insufficient evidence here to support the claim that $\beta_4$ is indeed zero. The high t-value in this test suggest that this variable does in fact help explain the variance in the data as it relates to sale price.

*Let* $\beta_5$ = Quality Index

$$H_o : \beta_5 = 0$$
$$H_a : \beta_5 \neq 0$$

$$t_5 = \hat{B}_5 / S_{\hat{B}_5} = -144.15 / 172.51 = \mathbf{-0.8356}$$

t-test with 95% confidence ($\alpha$ = 0.05),

Threshold: $t_{\alpha/2, n-k-2} = t_{0.025, 2420} = 1.9609$

*abs(T)* = 0.8356 < 1.9609

There is sufficient evidence here to supports the claim that $\beta 5$ is indeed zero, therefore should exclude include the quality index variable in the model as it does not explain further variance of sale price.

*b) the Omnibus Overall F-test*

Ho : $\beta_1 = \beta_2 = \beta_3 = B_4 = B_5 = 0$
Ha : $\beta j \neq 0$, for at least one value of j (for j in 1, 2, 3, 4, 5)

**F-Statistic:**

Sum of Squares due to **Regression** = $SS_R$ = 11,276,768,796,962
Sum of Squared **Error** = $SS_E$ = 5,350,540,284,090
Sum of Squares **Total** = $SS_T$ = $SS_R$ + $SS_E$ = 16,627,309,081,052

Let,
	N = 2425, p = 5

F = [($SS_T$ - $SS_E$) / p] / [$SS_E$ / (n − p − 1)] = 2,255,353,759,392 / 2,211,881,060= **1019.654**  on p = 5 and 2419 DF
p-value: < 0.0001

```
Call:
lm(formula = SalePrice ~ HouseAge + YearBuilt + YearRemodel +
    OverallQual + QualityIndex, data = data.m4)

Residuals:
    Min      1Q   Median      3Q     Max
-203460  -27267    -5706   19106  400005

Coefficients:
                Estimate Std. Error t value   Pr(>|t|)
(Intercept)  1094218.47 1470592.68   0.744      0.457
HouseAge       -1172.86     733.18  -1.600      0.110
YearBuilt       -854.39     735.14  -1.162      0.245
YearRemodel      287.52      62.53   4.598 0.00000448 ***
OverallQual    41471.70    1317.60  31.475    < 2e-16 ***
QualityIndex    -144.15     172.51  -0.836      0.403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47030 on 2419 degrees of freedom
Multiple R-squared:  0.6782,   Adjusted R-squared:  0.6775
F-statistic:  1020 on 5 and 2419 DF,  p-value: < 2.2e-16
```

There is insufficient evidence (F = 1020, P < 0.001) to conclude that at least one of the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone.

**Nested Model:**

*(14)  Write out the null and alternate hypotheses for a nested F-test using Model 3 and Model 4, to determine if the Model 4 variables, as a set, are useful for predicting SALEPRICE or not.  Compute the F-statistic for this nested F-test and interpret the results.*

$H_o : \beta_4 = \beta_5 = 0$

$H_a : \beta_j \neq 0$, for at least one value of j (for j in 4, 5)

$$F = (SSE_R - SSE_C) / SSE_C / [n - (k + p + 1)]$$

$$F = (10{,}031{,}364{,}633{,}591 - 5{,}350{,}540{,}284{,}090) / [\, 5{,}350{,}540{,}284{,}090 / [\, 2{,}425 - (5 + 1)]$$

$$= 4{,}680{,}824{,}349{,}501 / 2{,}211{,}881{,}060$$

$$= \textbf{2116.219}$$

Critical value at 99% confidence ($\alpha$ = 0.01), % confidence, = $F_{99,\, 5,\, 2419}$ = **3.3267**

The given F-statistic yielded a value of **2116.219** and at 99% confidence, we should reject the null hypothesis that the more complex, or complete, model 4 with the additional explanatory variables $\beta_4$ and $\beta_5$ is no more powerful than the reduced model 3.

## CONCLUSION

In this lab I learned how to deep-dive into an ANOVA table for a multivariate linear regression model, and how to make statistical inferences based on the analysis of the coefficients and residual variance. Specifically, performing single variable t-tests on regression coefficients, how to formulate a hypothesis about the overall fit of the model using both $R^2$ and adjusted $R^2$, how to calculate statistics long-hand. The difference between the standard R2 metric and the adjusted R2 metric is especially useful when attempting to assess the model accuracy vs complexity tradeoff, which is a fundamental aspect of statistical modeling.

The most valuable part of this lab was the formulation of hypothesis around testing the validity of individual components (beta coefficients) of a given model, performing t-tests on individual parameters to assert the validity of including additional variables in a model, and formulating an overall f-statistic that is indictive of all model parameters. The overall F-test is a useful tool for assessing model's performance, and especially useful is the ability to use this statistic to assess the added explained variance by more complicated models. The formulation and evaluation of nested models was a particularly useful exercise, as it further solidified my understanding of both the F-test statistic and comparing models that live in the same family regarding the set of explanatory variables they are constructed upon.

The application part of this exercise was particularly useful in reflecting on the models built in previous exercises and enhancing them with the addition of various additional categorizations of variables from the dataset. The construction of model 3 and model 4, where model 3 is nested inside model 4 proved to be particularly insightful given the analysis of each individual variable with their respective t-test, as wells as knowing the underlying mathematics that makes up every piece of the summary(lm) and anova(lm) functions, was insightful in that illuminated a rigorous procedure for model parameter evaluation. The overall F-test of the two models from a practical example will be particularly invaluable as we look to improve on our evaluation and formulation of future models.