Multiple Linear Regression (MLR) Report

1. Model 1 Mechanisms and Computation

1.1 Sample Data Observations

Based on the provided ANOVA table of Model 1, the sample data contains 72 observations. This is calculated from the model's 71 total degrees of freedom using the formula below:

$$\textbf{Total DF} = N - 1$$
$$\textbf{71} = N - 1 \textbf{ (add 1 on both sides)}$$
$$N = 72$$

1.2 β_1 Hypotheses and *t*-Test Calculation

To determine the statistical significance of a linear regression model, the slope of the predictor variable, Beta 1 (β_1), is tested using a Student's *t*-test. The null hypothesis (H_0) for the Student's *t*-test of β_1 states that the slope of the coefficient is equal to 0; while the alternative hypothesis (H_1) for the Student's *t*-test of β_1 states that the slope of the coefficient is not equal to 0. The hypotheses can be written as:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

The Student's *t*-test statistic can be calculated using the following formula:

$$t = b/SE_b$$

where b is the coefficient estimate of β_1, and SE_b is the standard error of the coefficient estimate. From provided summary table, the coefficient estimate for β_1 is 2.186, and the standard error of the coefficient estimate is 0.4104. Plugging these values into the formula above, we can calculate the Student's *t*-test statistic as:

$$t = 2.186/0.4104$$
$$t = 5.3265$$

The Student's t-test statistic was found to be 5.3265, which was statistically significant ($p < 0.001$) based on provided summary table. This indicates that we *reject* the null hypothesis and determine that variable X1 is a strong predictor of the response variable in Model 1. Since the coefficient estimate for X1 is positive, for each one-unit increase in X1, the response variable will increase by 2.186 units, holding all other variables constant.

1.3 Multiple R^2 and Adjusted R^2 Calculation

The multiple R^2 coefficient represents the proportion of variation in the response variable that is accounted for by the explanatory variables in the model. The multiple R^2 coefficient can be calculated using the following formula:

$$R^2 = 1 - (SSR/SST)$$

where *SSR* represents the residual sum of squares, and *SST* represents the total sum of squares. From the provided ANOVA table, we can enter the values of SSR and SST into the formula above to calculate the multiple R^2 coefficient:

$$R^2 = 1 - (630.36/2756.37)$$

$$\boldsymbol{R^2 = 0.7713}$$

The multiple R^2 coefficient for Model 1 was 0.7713, indicating that the explanatory variables X1, X2, X3, and X4 account for 77.13% of the variability of the response variable in Model 1. This suggests that only around 23% of the variation in the model remains unexplained and the explanatory variables capture the main drivers of the response variable with variables X1 and X2 being statistically significant. The adjusted R^2 coefficient adjusts the multiple R^2 value based on the number of predictors in the model by accounting for the degrees of freedom and penalizes the addition of irrelevant predictors. This means that the adjusted R^2 coefficient can decrease if a new predictor does not sufficiently improve the model. The adjusted R^2 coefficient can be calculated using the formula:

$$\boldsymbol{Adjusted\ R^2 = 1 - (1 - R^2) \times (n - 1/n - k - 1)}$$

where R^2 is the original R^2 coefficient, n is the number of observations, and k is the number of predictors in the model. From the previous section, we know the original R^2 value, and from section 1.1, we know the number of observations in the sample data; therefore, we can insert these into the formula above to calculate the adjusted R^2 coefficient:

$$\boldsymbol{Adjusted\ R^2 = 1 - (1 - 0.7713) \times (72 - 1/72 - 4 - 1)}$$

$$\boldsymbol{Adjusted\ R^2 = 0.7576}$$

The adjusted R^2 coefficient for Model 1 was 0.7576, indicating that 75.76% of the variation in the response variable is explained by the model after accounting for the number of predictors used (i.e., X1, X2, X3, and X4). This suggests strong model fit, with a slight dip from the original R^2 coefficient of 0.7713. Furthermore, most predictors in Model 1 contribute useful information and Model 1 is not being inflated by unnecessary variables.

1.4 F-test Hypothesis and Calculation

To determine overall significance of the model, the overall omnibus F-test is calculated. The F-statistic determines whether a full model with predictors is statistically significantly better than a model with no predictors. The null hypothesis (H_0) states all the coefficients are equal to zero, while the alternative hypothesis (H_1) states that at least one of coefficients is not equal to zero. The hypotheses can be written as:

$$\boldsymbol{H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0}$$

$$\boldsymbol{H_1: \text{At least one of } \beta_1, \beta_2, \beta_3, \beta_4 \neq 0}$$

The omnibus F-statistic can be calculated using the following formula:

$$\boldsymbol{F = MSR/MSE}$$

where *MSR* is the mean squared regression, which is variance explained by the regression, and *MSE* is the mean squared error, which is the unexplained variance. From the Model 1 ANOVA table, the MSR is 531.36, and the MSE is 9.41; plugging these into the formula above we get:

$$\boldsymbol{F = 531.36/9.41}$$

$$\boldsymbol{F = 56.4676}$$

The resulting F-statistic was 56.4676, which is statistically significant ($p < 0.001$) from the provided table. This indicates that we can *reject* the null hypothesis, and the regression model is

highly statistically significant overall. Additionally, this means that at least one predictor has a meaningful relationship with the response variable, specifically X1 and X2, and Model 1 has better fit than a model with no predictors.

2. Model 2 Mechanisms and Computation

Model 1 is a nested model within Model 2, as Model 2 has all the predictor variables used in Model 1 (X1-X4) in addition to variables X5 and X6. This means that Model 1 is a special case of Model 2, so Model 2 is the full or *unrestricted* model, while Model 1 is a smaller or *restricted* model. Since these models are nested, a partial F-test can be calculated to determine whether adding variables X5 and X6 significantly improves the model fit to the data. This can be seen in the provided summary table of Model 2, where the R^2 is increased to 0.7923, though a nested F-test can determine if this is statistically significant.

The null hypothesis (H_0) of a partial F-test states the coefficients for the additional predictors in the full model are equal to zero, meaning that the reduced model is sufficient. The alternative hypothesis (H_1) of a partial F-test states that at least one of the coefficients is not zero, meaning that the full model significantly improves model fitness and should be used. The hypotheses can be written as:

$$\mathbf{H_0: \beta_5 = \beta_6 = 0}$$
$$\mathbf{H_1: \text{At least one of } \beta_5, \beta_6 \neq 0}$$

To calculate the partial F-test for the nested model, the following formula is used:

$$\boldsymbol{F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)}}$$

where SSR_R is the sum of squared residuals from the restricted model, SSR_U is the sum of squared residuals from the unrestricted model, q is number of restrictions, n is the sample size, and k is the number of predictors in the unrestricted model. From section 1.3, the restricted model (Model 1) SSR_R was 630.36; from the provided ANOVA table of Model 2, the unrestricted model SSR_U is 572.6091. Inserting these variables into the formula above, the partial F-test is:

$$\boldsymbol{F = \frac{(630.36 - 572.6091)/2}{572.6091/(72-6-1)}}$$
$$\boldsymbol{F = 3.2778}$$

The resulting partial F-test is 3.2778, with $\alpha = 0.05$, is statistically significant based on the standard F-distribution table that has an approximate critical value of $F_{(2, 65)} \approx 3.14$. Since 3.2778 is greater than the critical value of 3.14, we can *reject* the null hypothesis, meaning that Model 2 provides significant improvement over Model 1. Furthermore, this indicates that additional variables of X5 and X6 improve the model fit beyond variables X1-X4, and either X5 or X6 contribute additional explanatory power for the response variable.

3. *Ames Housing* Data MLR Model

3.1 Variable Selection

Based on previous EDA of the *Ames Housing* dataset, 10 variables were selected that contained important attributes of home sales price in Ames, Iowa. These variables were: *GrLivArea*, *LotFrontage*, *LotArea, TotalBsmtSF*, *GarageArea*, *OverallQual*, *FullBath*, *BedroomAbvGr*, *YearBuilt*, and *Neighborhood*. Four sets of at least two variables were created to capture different aspects of transaction prices. The first set of variables contains *GrLivArea*, *TotalBsmtSF*, and *GarageArea* that describe the overall scale of a home and are associated with higher sales price. The second set of variables contains *LotArea* and *LotFrontage* that provide information on the parcel of land the home is on and add spatial context about the property. The third set of variables contains *BedroomAbvGr*, *FullBath*, and *OverallQual* that characterize the interior quality of a home which can directly influence buyer preferences and market value. The final set of variables contains *YearBuilt* and *Neighborhood* that represent the age and location of the home.

3.2 Model 3 Analysis & Interpretation

For the first model using the *Ames Housing* dataset, the first variable set containing *GrLivArea*, *TotalBsmtSF*, and *GarageArea* was chosen to develop a multiple linear regression (MLR) to determine the relationship between the response variable *SalePrice*. This variable set was chosen because in previous reporting on the Ames Housing dataset, larger-sized homes were indicative of a higher sale price.

A restricted MLR model, deemed Model 3, was fitted to the first set of variables to determine their linear relationship with *SalePrice*. The null hypothesis (H_0) of the MLR model states that each coefficient in the model has no linear effect on *SalePrice*, while the alternative hypothesis (H_1) of the MLR model states each coefficient in the model has a nonzero effect on *SalePrice*. These hypotheses can be written as:

$$\mathbf{H_0: \beta_j = 0 \ \forall \ j = 1, 2, 3}$$
$$\mathbf{H_1: \beta_j \neq 0 \ \forall \ j = 1, 2, 3}$$

When Model 3 was executed, the intercept ($t = -10.43, p < 0.001$) and all three predictors—GrLivArea ($t = 35.01, p < 0.001$), TotalBsmtSF ($t = 24.18, p < 0.001$), and GarageArea ($t = 22.19, p < 0.001$)—were statistically significant, therefore we would *reject* the null hypothesis, and all variables have a nonzero linear effect on *SalePrice*. The significant intercept indicates that when all variables are 0, *SalePrice* decreases by -$10.43, which is unrealistic to have a home with zero square footage. However, each variable had a large positive coefficient, indicating a strong linear relationship with *SalePrice*. *GarageArea* demonstrated the largest marginal increase in *SalePrice* with a coefficient of 101.09, meaning for every 1 square foot increase in *GarageArea*, *SalePrice* increased by $101. Finally, the R^2 coefficient for the MLR model was 0.6795, indicating that 67.95% of the variability in SalePrice is explained by these three variables.

An omnibus Overall F-test was performed to determine if the overall model is significant by testing whether a model with no predictors is worse than the fully fitted model. The null hypothesis (H_0) of the F-test states that all coefficient slopes are zero and the model has no

explanatory power, while the alternative hypothesis states that at least one of the coefficient slopes has a nonzero effect, meaning that it has explanatory power. These hypotheses can be written as:

$$\mathbf{H_0: \beta_1 = \beta_2 = \beta_3 = 0}$$
$$\mathbf{H_1: \text{At least one of } \beta_1, \beta_2, \beta_3 \neq 0}$$

When the omnibus Overall F-test was implemented, the model was statistically significant ($F = 2066$, $p < 0.001$), and the predictors explain a significant portion of the variability in *SalePrice*. This suggests that we can *reject* the null hypothesis and determine that the full model explains a statistically significant amount of variance in *SalePrice*. Moreover, each variable in the model has a meaningful independent effect with *SalePrice*, with GrLivArea ($F = 4556.85$, $p < 0.001$) being the strongest predictor and explains a substantial amount of the variance in *SalePrice*.

3.3 Model 4 Analysis & Interpretation

The third variable set containing *BedroomAbvGr*, *FullBath*, and *OverallQual* was chosen to add to Model 3 to create an unrestricted model, labeled Model 4. This variable set was selected because it added interior quality criteria of the homes, which can directly influence buyer preferences. Like Model 3, the null hypothesis (H_0) of model coefficients states that each of the six coefficients in the model has no linear effect on *SalePrice*, while the alternative hypothesis (H_1) of the MLR model states each of the six coefficients in the model has a nonzero effect on *SalePrice*. The hypotheses can be written as follows:

$$\mathbf{H_0: \beta_j = 0 \ \forall \ j = 1, 2, 3, 4, 5, 6}$$
$$\mathbf{H_1: \beta_j \neq 0 \ \forall \ j = 1, 2, 3, 4, 5, 6}$$

When Model 4 was performed, the intercept ($t = -21.47$, $p < 0.001$) and all six predictors—GrLivArea ($t = 25.12$, $p < 0.001$), TotalBsmtSF ($t = 16.14$, $p < 0.001$), GarageArea ($t = 13.48$, $p < 0.001$), BedroomAbvGr ($t = -8.79$, $p < 0.001$), FullBath ($t = 1.97$, $p = 0.0489$), and OverallQual ($t = 32.26$, $p < 0.001$)—were statistically significant. This indicates that we can *reject* the null hypothesis and conclude that, when holding the other variables constant, each has a nonzero linear relationship with *SalePrice*. *OverallQual* has a very large and the highest estimated effect on *SalePrice*: one-unit increase in *OverallQual* is associated with an average increase in $23,191 in *SalePrice*. *BedroomAbvGr* had a significant negative coefficient (-$9,220.43 per additional bedroom), suggesting that, controlling for total living area and quality, houses with more bedrooms above-ground, tend to sell for less. This suggests a diminishing returns effect of subdividing living space into more bedrooms rather than increasing overall space or quality, consistent with previous findings reported for the Ames Housing data. *FullBath* was only marginally significant ($p = 0.0489$), and a moderate estimated effect ($3,302.65 per addition full bathroom) compared to other predictors. This suggests that, while the number of full bathrooms has a statistically detectable association with SalePrice, it is comparatively weak predictor once size and quality are considered. The R^2 coefficient was 0.7848, indicating that around 78.48% of the variation in *SalePrice* was explained by Model 4. This represents a large

improvement over Model 3 ($R^2 = 0.6795$), with an increase of 0.1053, showing that the additional variables in Model 4 provide a meaningful gain in explanatory power, though a partial F-test will determine if this is statistically significant.

An omnibus Overall F-test was completed to determine if the overall model is significant by testing whether a model with no predictors is worse than the fully fitted model. The null hypothesis (H_0) of the F-test states that all coefficient slopes are zero and the model has no explanatory power, while the alternative hypothesis states that at least one of the coefficient slopes has a nonzero effect, meaning that it has explanatory power. The hypothesis can be written as:

$$\mathbf{H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0}$$
$$\mathbf{H_1: \text{At least one of } \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \neq 0}$$

When the omnibus Overall F-test was employed, the model was statistically significant ($F = 1775$, $p < 0.001$), and the predictors explain a significant portion of the variability in *SalePrice*. This suggests that we can *reject* the null hypothesis and determine that the full model explains a statistically significant amount of variance in *SalePrice*. Moreover, each variable in the model has a meaningful independent effect with *SalePrice*, with GrLivArea ($F = 6779.48$, $p < 0.001$) by far the strongest predictor and explains a substantial amount of the variance in *SalePrice*. A distant second strongest predictor is TotalBsmtSF ($F = 1709.56$, $p < 0.001$), followed by one of the newly added variables, OverallQual ($F = 1040.80$, $p < 0.001$), indicating more moderate explanatory power of the variance in *SalePrice*. The other newly added variables, BedroomAbvGr ($F = 280.33$, $p < 0.001$) and FullBath ($F = 108.07$, $p < 0.001$), were both statistically significant but had the lowest F-test, suggesting weak explanatory power of the variance in *SalePrice*.

3.4 Partial F-test Analysis & Interpretation

Model 3 and Model 4 are nested MLR models, with Model 3 being the restricted or smaller model and Model 4 is the unrestricted or full model. Nested models can determine if the additional variables in the full model provide a statistical improvement in explaining the response variable using a partial F-test. The null hypothesis (H_0) for a partial F-test states that the additional predictors in the full model do not significantly improve the fitness of the model. The alternative hypothesis (H_1) for a partial F-test states that the additional predictors in the full model do significantly improve the fitness of the model. These hypotheses can be written as:

$$\mathbf{H_0: \beta_4 = \beta_5 = \beta_6 = 0}$$
$$\mathbf{H_1: \text{At least one of } \beta_4, \beta_5, \beta_6 \neq 0}$$

When the partial F-test was performed, the full model was significantly significant ($F = 476.4$, $p < 0.001$). This advises us to *reject* the null hypothesis and conclude that, when taken together, the additional variables—*OverallQual*, *BedroomAbvGr*, and *FullBath*—significantly improve the fit of the model. Moreover, we should use the full model instead of the restricted model to characterize the linear relationship with *SalePrice*.