# R Project 3:
# Multiple Regression Analysis

Data Analytics 2 – ADS523

James Sears

Applied Data Science Program, Graduate School, Longmeadow, MA

Submitted To: Professor Aja Shabana

Summer 2021

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

**Introduction**

The object of this project is to perform multiple regression analysis (using R) to find the least-squares regression line that would enable home sale price predictions based on related quantitative real estate (independent) variables such as square footage, lot size, number of bedrooms, year built, etc. The components of the least-squares multiple regression line (partial slopes and the y-intercept) will be interpreted to describe the relationship between several independent variables and a dependent variable (home sale price).

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

**Data Description & Initial Variable Selection**

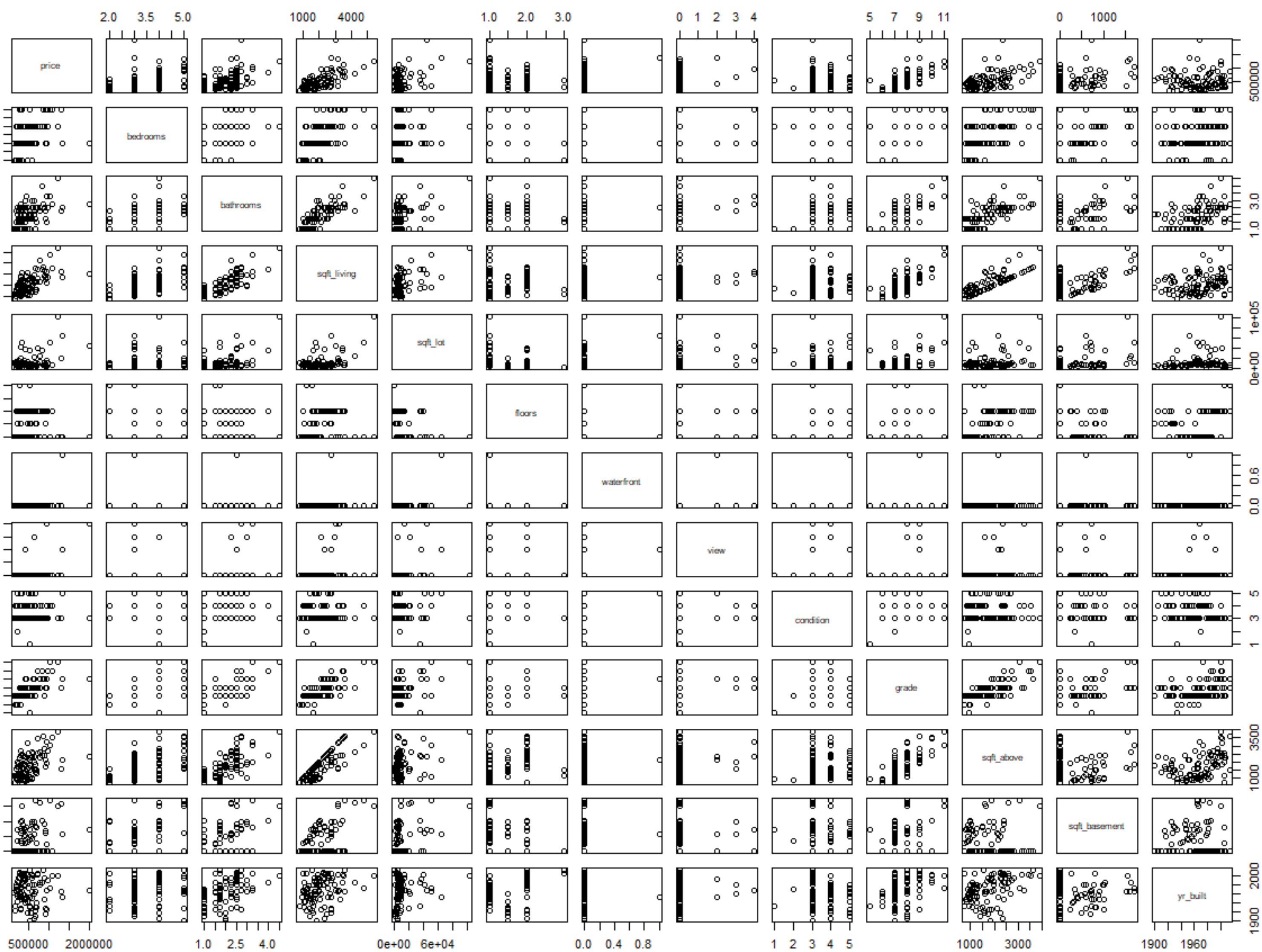
The provided data set contains 100 data points related to real estate is the greater Seattle, WA area. It includes home prices and other variables associated with each home, displayed in Table 1. Inspecting the variables reveals a number that should be removed prior to further analysis. These include nominal

**Table 1.** Independent variable description and initial variable selection.

| Variable | Description | In/Exclude | Reason for Exclusion |
|---|---|---|---|
| `id` | Transaction ID | Exclude | Nominal data |
| `date` | Date of Sale or Listing | Exclude | Only of interest if able to transform to days on market |
| `bedrooms` | Bedrooms, Each | Include | |
| `bathrooms` | Bathrooms, Each | Include | |
| `sqft_living` | Size of House, Square Feet | Include | |
| `sqft_lot` | Size of Lot, Square Feet | Include | |
| `floors` | Floors, Each | Include | |
| `waterfront` | On Waterfront, 0 (No) or 1 (Yes) | Include | |
| `view` | View Rating, Weighted Scale of 0, 2, 3, 4 | Include | |
| `condition` | Condition Rating, Scale of 1-5 | Include | |
| `grade` | Grade, Scale of 5-11 | Include | |
| `sqft_above` | Size of House Above Ground, Square Feet | Include | |
| `sqft_basement` | Size of Basement, Square Feet | Include | |
| `yr_built` | Year Built | Include | |
| `yr_renovated` | Year Renovated, 0 (Not Renovated) or Year | Exclude | Improper scale, 0 for Not Renovated would mean renovated in the Year 0 |
| `zipcode` | Zip Code | Exclude | Nominal data |
| `lat` | Latitude, Degrees | Exclude | Unable to transform to meaningful continuous variable |
| `long` | Longitude, Degrees | Exclude | Unable to transform to meaningful continuous variable |

variables such as the `id` (Transaction ID) and `zipcode`. There are also GPS coordinates, latitude (`lat`) and longitude (`long`) that are continuous variables, but not of any useful scale if they are not being compared to a single point (if distance from city center, for example, was of interest). The same can be said about the `date` variable; the date alone is not useful. Without knowing if it is a sale date or date of listing, nor the current date to compare them, calculating a valuable attribute, like number of days on the market, is not possible. Improper scaling is of concern for the variable `yr_renovated`, associated with the year in which the home was renovated. In cases where the home was not renovated, the value of zero was assigned, which is interpreted as the year zero; it is assumed the renovation did not occur over two thousand years ago and prior to the home's initial construction. The discrete variables `waterfront`, `view`, `condition`, and `grade` raised concerns as well, but they will be retained for their significance to the multiple regression model, explained below.

**Figure 1.** Scatterplot matrix after intial variable selection.

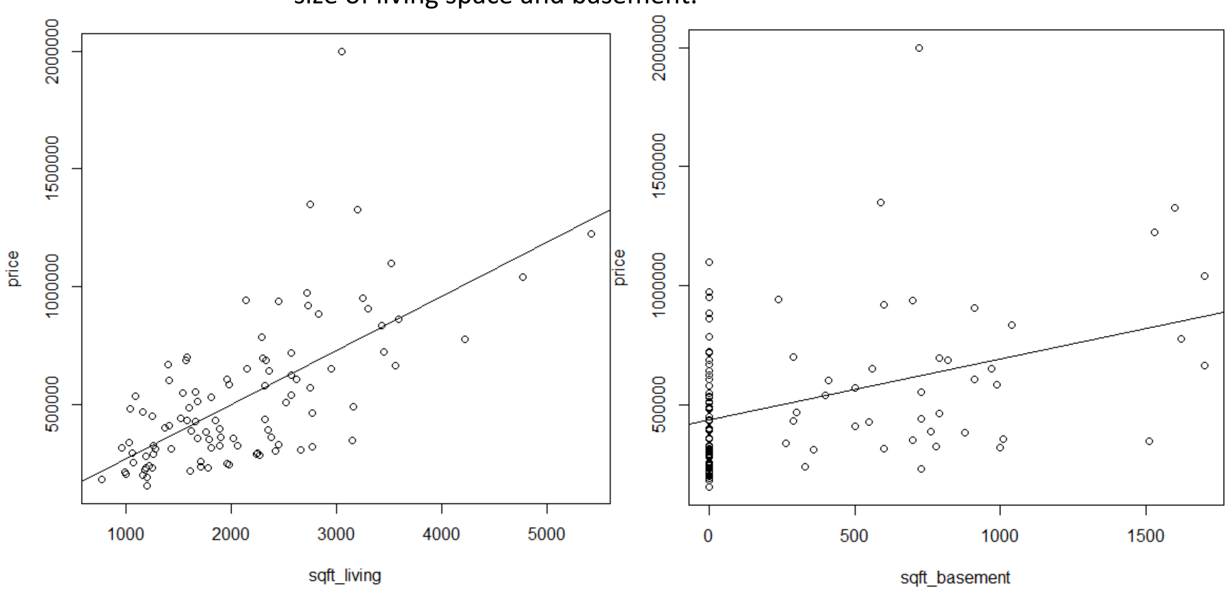


**Bivariate Analysis: Linear Correlation and Regression**

With initial varibale selection complete, scatterplot (Figure 1) and correlation (Table 2) matricies of the remainig variables were constructed to aid in bivariate anaylsis. The scatter plot matrix plots the dependent variable (`price`) against each independent variable - seen in the first colummn or first row of Figure 1 - as well as each possible independent varibale pair.

A scatterplot with data points clustered together in a linear fashion indicates a linear correlation between the two variables; a plot of more dispersed data points indicates a lack of linear correlation. For example, the scatterplot (also see Figure 2) that compares the dependent (response) variable `price` and the independent (explanatory) variable for living space (`sqft_living`) somewhat resembles a diagonal line, suggesting that the two variables are potentially linearly correlated. Further, the two variables are positively correlated as an increase in the size of the house is associated with an increase in price. Conversely, the plots that compare `price` and the number of `floors` (also see Figure 2) does not appear to resemble a line. This indicates that the number of floors does not have a linear association with the price of a home.

**Figure 2.** Scatterplots (with least-squares regression lines) of price associated with the size of living space and basement.




A more precise measure of the strength and direction of the linear relation between two quantitative variables may be achieved by calculating their linear correlation coefficient, *r*. The formula for a sample's linear correlation coefficient (*r*) is the sum of the products of the *z*-scores for the explanatory and response variables (divided by the degrees of freedom (sample size minus one). As the numerator is a standardized measurement of each variables' relationship to its respective mean, their product is a measurement of the relationship between the two variables once it is normalaized by dividing it by the degrees of freedom (which may strengthen or weaken the relationship depending on how large or small, respectively, the sample size is) (Sullivan, page 188).

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

The linear correlation coefficient (*r*) is on a scale from -1 to +1, with values at those extremes representing perfect negative and postive linear relationships, respectively. A value of zero signifies that a linear relationship between the two variables does not exist (however correlation may still exist). Values closer to zero (*r* $-0.3 \le r \le +0.3$) represent weak relationships, and those close to the extremes ($r \le -0.7$ or $r \ge +0.7$) represent strong ones. The sign of the coefficient represents the nature of the relationship. Positive values result when the response variable increases with the explanatory variable, or they both decrease. A negative value results when the response variable decreases with an increase in the explanatory variable, or when the response variable increases with a decrease in the explanatory variable.

**Table 2.** Correlation matrix after initial variable selection.

| | price | bedrooms | bathrooms | sqft_living | sqft_lot | floors | waterfront | view | condition | grade | sqft_above | sqft_basement | yr_built |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| price | 1.000 | | | | | | | | | | | | |
| bedrooms | 0.354 | 1.000 | | | | | | | | | | | |
| bathrooms | 0.465 | 0.486 | 1.000 | | | | | | | | | | |
| sqft_living | 0.651 | 0.560 | 0.790 | 1.000 | | | | | | | | | |
| sqft_lot | 0.452 | 0.119 | 0.372 | 0.498 | 1.000 | | | | | | | | |
| floors | 0.072 | 0.142 | 0.373 | 0.224 | -0.231 | 1.000 | | | | | | | |
| waterfront | 0.276 | -0.051 | 0.070 | 0.078 | 0.364 | -0.093 | 1.000 | | | | | | |
| view | 0.443 | 0.065 | 0.258 | 0.206 | 0.283 | 0.014 | 0.247 | 1.000 | | | | | |
| condition | -0.007 | -0.004 | 0.018 | 0.011 | 0.092 | -0.114 | 0.218 | 0.061 | 1.000 | | | | |
| grade | 0.654 | 0.362 | 0.650 | 0.732 | 0.438 | 0.318 | 0.135 | 0.235 | 0.010 | 1.000 | | | |
| sqft_above | 0.516 | 0.541 | 0.729 | 0.838 | 0.333 | 0.511 | 0.056 | 0.192 | -0.003 | 0.726 | 1.000 | | |
| sqft_basement | 0.396 | 0.192 | 0.323 | 0.540 | 0.398 | -0.377 | 0.057 | 0.081 | 0.025 | 0.221 | -0.008 | 1.000 | |
| yr_built | -0.034 | 0.094 | 0.505 | 0.282 | 0.124 | 0.424 | -0.049 | 0.021 | -0.264 | 0.335 | 0.417 | -0.127 | 1.000 |

The line that is “seen” when viewing a scatter plot can be calculated formally as the least-squares regression line, noted in Figue 2. It represents a line that minimizes the sum of the squared errors (or residuals), which is the distanse between the observed values of the response variable and those predicted by the least-squares regression line. The equation for this line is given by Formula 1:

$$\hat{y} = b_1 x + b_0$$

where $b_1 = r \cdot \frac{s_y}{s_x}$ is the regression, or slope coefficient of the least-squares reggression line,

and $b_0 = \bar{y} - b_1 \bar{x}$ is the y-intercept of the least-squares reggression line.

It is noted in the equations above that $\bar{x}$ is the sample mean and $s_x$ is the sample standard deviation of the explanatory variable $x$; $\bar{y}$ is the sample mean and $s_y$ is the sample standard deviation of the response variable $y$ (Sullivan, page 205). The linear correlation coefficient (*r*) can be thought of as a representation of how closely the data points (of the scatter plot) adhere to the least-squares regression line. More significant *r* values (closer to -1 or +1) means there is a more significant linear relationship that results from a tighter scatter plot.

The correlation matrix Table 2 supplies the correlation coefficients that correspond to the scatter plot matrix of Figure 1. Applying the concepts related to correlation from above enables confirmation of the previously noted relationships derived from visual inspection of the scatterplot matrix Figure 1. The noted positive linear correlation between the dependent response variable `price` and the independent explanatory variable for living space (`sqft_living`) is confirmed with a relatively strong *r* value of 0.651. This is reasonable, as when all other variables are held constant, larger houses are typically more expensive than smaller ones.

The *r* value for `price` and the number of `floors` is weak at 0.072, confirming that the number of floors does not likely have a linear association with the price of a home and is therfore removed from the multiple regression model (noted on Table 4). The variables `condition` and `yr_built` also have *r* values near zero, but they will be retained for their significance to the multiple regression model, explained below.

The scatterplot (Figure 1) and correlation (Table 2) matricies also compare each possible independent varibale pair. When building a multiple regression model, care is taken to avoid independent variables that exhibit multicollinearity. Multicollinearity exhists when two independent variables are highly correlated ($r \leq -0.7$ or $r \geq +0.7$), suggesting that they represent the same thing, or perhaos are the result of, or associatred with, the same thing.

For example, the linear correlation coefficient (*r*) that represents the relationship between living space (`sqft_living`) and `bathrooms` is 0.790. This strong positive relationship is reasonable as the number of bathrooms in a house is expected to be greater for a larger house. Similarly, the number of square feet above ground (`sqft_above`) is also highly positively, and reasonably, correlated with living space (`sqft_living`) with an *r* value of 0.838. Therefore, due to concerns for multicollinearity, the variables `bathrooms` and `sqft_above` will be removed from the multiple regression model (noted on Table 4) as their association with the response variable `price` ($r = 0.465$ and $r = 0.516$, respectively) is weaker than that of `sqft_living` ($r = 0.651$).

Removing independednt variables due to multicollinearity is not always practiced. For example, there is a strong relationship ($r = 0.732$) between the independent variables `sqft_living` and `grade` that would suggest multicollinearity and the subsequent removal of `grade` due to its weaker relationship with the response variable `price`. However, the variables are not necessarily collinear as size and rating are not obviously related. Further, `grade` will be retained for its significance to the multiple regression model, explained below.

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

**Multiple Regression Analysis, Part One**

As linear regression enables bivariate analysis of the dependent response variable and a single independent explanatory variable, multiple regression expands on that concept to describe the relationship when more than one explanatory variable is present. While linear regression describes a sloped line on a plane, with additional variables, multiple regression describes a plane with multiple partial slopes in a multi-dimensional space. Its equation is like that of Formula 1, but with the additional variables and their partial slope coefficients, given by Formula 2:

$$\boldsymbol{\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k}$$

where $b_0$ is the intercept when all *x*'s are zero; $b_k$ refers to the sample regression coefficients - or partial slopes - for *k* independent variables; and $x_k$ refers to the value of the various independent variables. In the home sale price scenario, multiple regression analysis will be applied to predict the dependent response variable `price` ($\hat{y}$) based on a selection of (*k*) independent variables and their derived coefficients ($b_k$).

The process begins by determining whether it is possible that all the independent variables have zero as regression coefficients. In other words, if all the regression coefficients equal zero, then the multiple regression equation is reduced to only the y-intercept ($b_0$); none of the selected independent variables are associated with the response variable. The null and alternative hypotheses are therefore:

$H_0$: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$
$H_1$: Not all $\beta_i$'s are 0

The *F*-distribution is used to find the *P*-value and determine the significance of regression coefficients of the selected variables. Technology (R or Excel) was used to perform Analysis of Variance (ANOVA) to calculate the *F*-test statistic and subsequent *P*-value, based on a 0.05 level of significance ($\alpha$), the results are shown in the middle section of Table 3. As the *P*-value of $6\text{x}10^{-18}$ is very low, the null hypothesis is rejected; at least one of the selected variable's regression coefficient is not equal to zero.

As the null hypothesis is rejected, indicating that at least one of the variable's regression coefficient is not equal to zero, the next step is calculating (also with technology, R or Excel) the regression coefficient of each independent variable. A test statistic (based on the Student's t-distribution) and subsequent *P*-value are also calculated for each variable to determine which coefficients differ significantly from zero (*P*-value < $\alpha$). The variables that have regression coefficients that are not significant (differ from zero) are usually excluded from the analysis as they signify a lack of association with the response variable.

The bottom section of Table 3 displays the calculated regression coefficient and subsequent *P*-value of each independent variable in the home sale price scenario. The variables whose calculated regression coefficient are determined as insignificant ($\alpha$ = 0.05) in terms of their *P*-value include: bedrooms (0.5458), bathrooms (0.5784), floors (0.7014), sqft_lot (0.7748), sqft_above (error) and sqft_basement (error). Therefore, these variables will also be excluded from the multiple regression model, noted on Table 4.

It was noted in the linear regression section above that certain variables that likely should be removed from the multiple regression model for various reasons, but were exempted due to their significance (*P*-value < $\alpha$ of 0.05) to the initial model. These include the discrete variables `waterfront`, `view`,

**Table 3.** Multiple regression analysis after initial variable selection.

| *Regression Statistics* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Multiple R | 0.824 | | | | | | | |
| R Square | 0.679 | | | | | | | |
| Adjusted R Square | 0.6275 | | | | | | | |
| Standard Error | 182997 | | | | | | | |
| Observations | 100 | | | | | | | |
| **ANOVA** | | | | | | | | |
| | *df* | *SS* | *MS* | ***F*** | ***Signifi-cance F*** | | | |
| Regression | 12 | 6E+12 | 5E+11 | **16.919** | **6E-18** | | | |
| Residual | 88 | 3E+12 | 3E+10 | | | | | |
| Total | 100 | 9E+12 | | | | | | |
| | | | | | | | | |
| | ***Coefficients*** | *Standard Error* | *t Stat* | ***P-value*** | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | **5,253,699.004** | 2E+06 | 2.9928 | **0.0036** | 2E+06 | 9E+06 | 2E+06 | 9E+06 |
| grade | **122,991.845** | 28271 | 4.3505 | **4E-05** | 66809 | 179175 | 66809 | 179175 |
| view | **105,439.198** | 27056 | 3.897 | **0.0002** | 51670 | 159208 | 51670 | 159208 |
| yr_built | **-2,947.521** | 887.98 | -3.319 | **0.0013** | -4712 | -1183 | -4712 | -1183 |
| waterfront | **433,964.582** | 206823 | 2.0982 | **0.0388** | 22947 | 844982 | 22947 | 844982 |
| sqft_living | **111.974** | 54.234 | 2.0646 | **0.0419** | 4.1948 | 219.75 | 4.1948 | 219.75 |
| condition | **-55,836.396** | 27526 | -2.029 | **0.0455** | -1E+05 | -1134 | -1E+05 | -1134 |
| bedrooms | **18,419.305** | 30377 | 0.6064 | **0.5458** | -41949 | 78788 | -41949 | 78788 |
| bathrooms | **-28,402.248** | 50921 | -0.558 | **0.5784** | -1E+05 | 72793 | -1E+05 | 72793 |
| floors | **20,995.862** | 54574 | 0.3847 | **0.7014** | -87458 | 129450 | -87458 | 129450 |
| sqft_lot | **0.515** | 1.7937 | 0.2869 | **0.7748** | -3.05 | 4.0792 | -3.05 | 4.0792 |
| sqft_above | **0.000** | 0 | 65535 | **#NUM!** | 0 | 0 | 0 | 0 |
| sqft_basement | **56.553** | 63.201 | 0.8948 | **#NUM!** | -69.04 | 182.15 | -69.04 | 182.15 |

`condition`, and `grade` which have the potential that their scales are arbitrary and not proportional or linear. And the variables `condition` and `yr_built` have a weak relationship with the response variable `price`. Also, the variable `grade` had the potential for multicollinearity with another independent explanatory variable `sqft_living`, yet both variables were retained. In each case, their *P*-values, see Table 3 indicated that their inclusion was significant to the model.

In addition to these variables, sqft_living is considered significant with a *P*-value of 0.0419 less than the 0.05 level of significance ($\alpha$).

**Table 4.** Further variable selection from linearity, multicollinearity, and significance checks.

| Variable | Description | In/Exclude | Reason for Exclusion |
|---|---|---|---|
| `bedrooms` | Bedrooms, Each | Exclude | p-Value > α |
| `bathrooms` | Bathrooms, Each | Exclude | Multicollinearity with sqft_living (r = 0.790), p-Value > α |
| `sqft_living` | Size of House, Square Feet | Include | |
| `sqft_lot` | Size of Lot, Square Feet | Exclude | p-Value > α |
| `floors` | Floors, Each | Exclude | No correlation with price (r = 0.072), p-Value > α |
| `waterfront` | On Waterfront, 0 (No) or 1 (Yes) | Include | |
| `view` | View Rating, Weighted Scale of 0, 2, 3, 4 | Include | |
| `condition` | Condition Rating, Scale of 1-5 | Include | No correlation with price (r = -0.007), p-Value < α |
| `grade` | Grade (?), Scale of 5-11 | Include | |
| `sqft_above` | Size of House Above Ground, Square Feet | Exclude | Multicollinearity with sqft_living (r = 0.838), error calculating p-Value |
| `sqft_basement` | Size of Basement, Square Feet | Exclude | error calculating p-Value |
| `yr_built` | Year Built | Include | No correlation with price (r = -0.034), p-Value < α |

**<u>Multiple Regression Analysis, Part Two</u>**

As the initial multiple regression model included insignificant (*P*-value > $\alpha$ of 0.05) variable coefficients, they are removed, and multiple regression analysis is run again on the remaining significant (*P*-value < $\alpha$ of 0.05) variables. The results are displayed in Table 5.

The process repeats by first determining whether it is possible that all the independent variables have regression coefficients of zero. The null and alternative hypotheses are therefore:

$H_0$: $b_1 = b_2 = b_3 = \cdots = b_k = 0$
$H_1$: Not all $b_i$'s are 0

The *F*-test statistic is calculated, as is the subsequently low (based on a $\alpha$ = 0.05 level of significance) *P*-value of $1\text{x}10^{-20}$. The null hypothesis is once again rejected, meaning that at least one of the selected variable's regression coefficient is not equal to zero.

The regression coefficient of each independent variable is then calculated. A test statistic (based on the Student's t-distribution) and subsequent *P*-value are also calculated for each variable to determine which coefficients differ significantly from zero (*P*-value < $\alpha$ of 0.05 level of significance). The results are displayed in the bottom section of Table 5. It is noted that every variable's regression coefficient is deemed significant with *P*-values less than the 0.05 level of significance. Therefore, the multiple regression model is created by substituting the y-intercept, variables, and their coefficients into Formula 2:

`price` = 6,121,222.19 + 136.12 x `sqft_living` + 445,690.40 • `waterfront` + 102,040.90 • `view` – 60,784.97 • `condition` + 116,865.80 • `grade` – 3,350.53 • `yr_built`

This model can be interpreted in terms of the y-intercept and the effect of each independent variable on the dependent response variable `price`. A y-intercept of $6,121,222.19 indicates that when all independent variables are equal to zero (x = 0), the value of the home is predicted to be $6,121,222.19. As these values are outside of the scope of the data, it is unsupported by the model, and not practical as a home with zero square feet of living space is not a home.

Each independent variable's coefficient indicates the change in `price` that results from an incremental change (increase or decrease) of that variable, <u>provided that all other variables are held constant in each case</u>. For example, if the living space (`sqft_living`) increases by one square foot, the prediction of the price of the home increases by $136.12. Property categorized as being `waterfront` (1 = yes, 0 = no) will gain $445,690.40 in value, and those that are, will not change in value. The coefficients for these variables are positive terms, indicating that `price` will increase as they do. The same is the case for `view` and `grade` which increase the `price` of the home by $102,040.90 and $116,865.80 with each increase of those variables, respectively. These are reasonable as houses are more valuable when they are larger, are on water, have better views, and/or are graded/rated higher.

Regression coefficients with negative signs indicate a negative linear relationship with `price`, meaning that, <u>provided that all other variables are held constant,</u> as the independent variable increases, the `price` of the home decreases. For example, the negative regression coefficient for the variable `condition` indicates a price decrease of -$60,784.97 for each increase on the variables discrete scale of 1 through 5. This likely means that the scale is inverted where a 1 indicates the best condition and a 5 indicates the worst; the interpretation of the coefficient is reasonable with this evaluation.

However, the variable `yr_built` has a negative regression coefficient of -3,350.53 that appears counterintuitive. The variable is reported in the actual year (e.g., 1950) the home was built and not the age of the house, meaning the model calculates that there is a decrease of -$3,350.53 for every year *newer* the house is. In other words, more value is placed on older houses, and less value is placed on newer ones.

**Table 5.** Multiple regression analysis after initial variable selection.

| *Regression Statistics* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Multiple R | 0.8213237 | | | | | | | |
| R Square | 0.67457262 | | | | | | | |
| Adjusted R Square | **0.653577305** | | | | | | | |
| Standard Error | 179222.6432 | | | | | | | |
| Observations | 100 | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | ***F*** | ***Significance F*** | | | |
| Regression | 6 | 6E+12 | 1E+12 | **32.12967397** | **1E-20** | | | |
| Residual | 93 | 3E+12 | 3E+10 | | | | | |
| Total | 99 | 9E+12 | | | | | | |
| | | | | | | | | |
| | ***Coefficients*** | *Standard Error* | *t Stat* | ***P-value*** | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | **6,121,222.187** | 1E+06 | 4.4605 | **2.28655E-05** | 3E+06 | 9E+06 | 3E+06 | 9E+06 |
| sqft_living | **136.117** | 30.839 | 4.4138 | **2.73423E-05** | 74.877 | 197.36 | 74.877 | 197.36 |
| waterfront | **445,690.403** | 191064 | 2.3327 | **0.021822239** | 66274 | 825106 | 66274 | 825106 |
| view | **102,040.904** | 25635 | 3.9805 | **0.000136239** | 51135 | 152947 | 51135 | 152947 |
| condition | **-60,784.970** | 26333 | -2.308 | **0.02319588** | -1E+05 | -8494 | -1E+05 | -8494 |
| grade | **116,865.797** | 26235 | 4.4546 | **2.33825E-05** | 64769 | 168963 | 64769 | 168963 |
| yr_built | **-3,350.533** | 702.78 | -4.768 | **6.87397E-06** | -4746 | -1955 | -4746 | -1955 |

**Conclusions**

Upon review of the final multiple regression model, it is noted that some components are dubious:

- The negative regression coefficient for `yr_built` is counterintuitive.
- The `waterfront` variable is zero for all data points except one; its inclusion is questionable.
- Explanations for the variables `grade` and `condition` are not provided, and though their *r* value is close to zero, multicollinearity is suspected as it is reasonable to assume they both describe the same concept: the state of the home.

Potential improper initial variable evaluation may be responsible for the model's mediocre Adjusted $R^2$ value of 0.6536. The Adjusted $R^2$ value is a coefficient of determination that measures the percentage of total variation in the response variable that is explained by the least-squares regression lines of the multiple regression model (Sullivan, page 717) and adjusted based on the number of degrees of freedom. Adjusted $R^2$ is a good measure of the model's predictive accuracy.

A model that is 65.36% accurate in predicting the price of a home is more reasonable than not but should be considered mediocre at best. If transformation of the existing variables - or additional ones - were available, it is possible that a more accurate model could be derived. The geography of the homes suggests that they are suburban, so the variables of interest would be tied to what is of interest to that demographic and housing market. These could include distance to the city (Seattle) center as an evaluation of commuting distance, rating of school systems, and rating of public safety.

Jim Sears
ADS5232 – Data Analytics 2
R Project 2

**References**

Sullivan III, Michael. Statistics: Informed Decisions Using Data; 6th Edition, ISBN-13: 9780135780121.