Lecture Note 7: Simple Linear Regression

1. Introduction

During exploratory data analysis (EDA), if we observe a significant linear relationship between two variables, X and Y, based on scatter plots and correlation coefficients, it prompts the question of how Y would respond to a one-unit change in X. Simple linear regression addresses this inquiry. It is a statistical technique utilized to model the relationship between a dependent variable, Y, and a single independent variable, X. The objective is to determine a straight line that optimally depicts the association between the two variables, effectively delineating the linear trend line on the scatter plot mathematically.

It's important to highlight that simple linear regression naturally extends from the Pearson correlation coefficient. We'll delve deeper into this later to illustrate that simple linear regression provides an alternative method of characterizing the correlation between X and Y. Furthermore, it's essential to recognize that simple linear regression, and linear regression more broadly, solely elucidates the linear relationship between independent and dependent variables. However, it does not infer, let alone establish, causation, implying that a one-unit change in X causes Y to change by a specific amount.

2. Simple Linear Regression Model

Let's consider quantifying the linear relationship between X ('sqft_living') and Y ('price'), given their observed strong linear association during exploratory data analysis (EDA). While we can express a linear equation algebraically as Y = mX + b, this mathematical, functional representation suggests that for a specific value of X (the square footage of living space), there will be a unique corresponding price level. However, in reality, houses of the same size can vary significantly in price due to various factors. To accommodate this variability, statistics and data science introduce an error term to account for randomness or variation. As a result, the standard simple linear regression (SLR) model is typically formulated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Here, X is the independent variable or feature, 'sqft_living'; X_i is the square-footage of living space of the i^{th} house in the dataset; Y is the dependent variable or the target, 'price'; Y_i is the price of the i^{th} house; β_0 is the intercept; β_1 is the slope; ε represents random error, often called the error term; and ε_i is the realization of the random error on the i^{th} house. The coefficients β_0 and β_1 are constant regardless of i. Therefore, we also commonly see the SLR represented by the equation with subscript i being dropped.

The error term ε captures the variability in the dependent variable that cannot be explained by the independent variable(s) included in the model. It accounts for factors other than the predictors that influence the dependent variable, including measurement errors, unobserved variables, and random variability. The error term is typically assumed to follow a normal distribution with a mean of zero, and a constant standard deviation, often denoted as σ. The assumption indicates that, on average, the model predicts the true value of the dependent variable. If this condition is not met, then utilizing a linear model would not be appropriate from the outset.

Formally, the expected value of the error term is zero: $E(\varepsilon_i) = 0$. Consequently, the expected value of Y_i is: $E(Y_i) = \beta_0 + \beta_1 X_i$. In simpler terms, simple linear regression suggests that the expected value of the target variable Y ('price') increases or decreases linearly with the predictor variable X ('sqft_living').

Some argue that the normal distribution serves as a reasonable approximation of real-world randomness, providing a rationale for the normality assumption. However, in my view, the main purpose of the normality assumption is to facilitate statistical inference on estimated parameters and predicted target values. A less realistic assumption is that all error terms follow an identical normal distribution with a common variance. This assumption, known as homoscedasticity, implies that regardless of a house's size, the distribution of its price has the same variance as that of a much smaller house. Classical statistics and econometrics often focus on addressing violations of this assumption and others in linear regression. Conversely, heteroscedasticity, the opposite of homoscedasticity, is a common occurrence. In modern data science and machine learning, the emphasis is more on techniques like training set/test set splits and cross-validation to assess model validity and generalizability. We will explore these assumptions more thoroughly in the section on multiple linear regression.

2.1 Estimated SLR Equation

In constructing the simple linear regression (SLR) model, the actual values of the intercept and slope are unknown. Instead, we aim to estimate these parameters using the available data. We represent the estimated SLR equation as follows:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Here, $\hat{Y}$ is the estimated price; $\hat{\beta}_0$ is the estimated intercept; and $\hat{\beta}_1$ is the estimated slope. You might notice the absence of the error term in the estimated SLR equation. This is because the error term is assumed to have a mean of zero. Any systematic deviation from this mean would either be absorbed into the intercept term or indicate bias in our model.

However, the estimated SLR equation still leaves us wondering how exactly we derive the estimates for the intercept and the slope. Ultimately, we seek to quantify these values and assign specific numerical values to them. We could choose any values for the intercept and slope, resulting in a unique linear trend line for each combination. However, the data play a crucial role in guiding our selection process. We aim to ensure that the trend line, or algebraically our estimated SLR equation, best fits the data points. However, this raises the question of what we mean by "best fit" or how we should define it. To address this, we introduce a well-known solution: the least squares method, also known as ordinary least squares (OLS).

2.2 Ordinary Least Squares

OLS provides an intuitive method for defining the best fit. Regardless of the values chosen for the intercept and slope, we can make predictions (computing $\hat{Y}$) for any given value of X, or square footage of living space. We also have the actual target values, prices. The difference between the estimate and the actual price, often called residual (realization of random error, denoted as e), is a natural way of measuring how good a fit it is. However, summing up all residuals presents a problem. Naturally, some are positive while others are negative. If we merely sum them, they would cancel each other out. A pair of residuals such as (-100, 100) would be considered as good as a pair of residuals like (-1, 1). To tackle these issues, we square all errors or residuals, resulting in a metric for best fit known as the sum of squared errors (SSE).

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2$$

Here, n represents the sample size or the number of data points (observations). For our housing price dataset, n would be 21,613. The quest to determine the "best fit," which is inherently vague, transforms into a precisely defined optimization problem: Select the values for the intercept and slope judiciously to minimize the sum of squared errors (SST). The technique of optimizing the sum of squared errors (SSE) to identify the best-fit linear regression model is known as the least squares method or ordinary least squares (OLS).

The optimal values of the estimated intercept and slope resulted from optimizing SSE are commonly denoted as b_0 and b_1. There are two methods for determining the actual values of b_0 and b_1: mathematical solutions and computational solutions relying on algorithms such as gradient descent. Below are the formulas for calculating the optimal values of the estimated slope (b_1) and intercept (b_0):

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

In the formulas, $\bar{X}$ and $\bar{Y}$ represent the mean or average of X and Y calculated from the dataset. Therefore, the resulting optimal estimated Simple Linear Regression (SLR) equation is as follows:

$$\hat{Y} = b_0 + b_1 X$$

b_0 represents the intercept of the regression line or the trend line, which is the value of the dependent variable (Y) when the independent variable (X) is zero. It indicates the starting point of the regression line on the Y-axis. On the other hand, b_1 represents the slope of the regression line, which reflects the change in the dependent variable (Y) for a one-unit change in the independent variable (X). It quantifies the rate of change of Y with respect to X and determines the direction and steepness of the regression line.

Readers interested in the derivation of the mathematical formulas for b_0 and b_1 can readily find resources on the Internet. Our focus, however, is to provide a deeper understanding of these two entities, particularly b_1. To delve deeper into this, let's rewrite the formula for b_1 as follows:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$$= \frac{COV(X,Y)}{VAR(X)}$$

In other words, the optimal estimated slope of the best-fit linear trend line is merely the ratio of the covariance between X and Y and the variance of X. Recall that the Pearson correlation coefficient is

computed as the ratio of the covariance to the product of the standard deviations of X and Y. We can further express the formula for b_1 as follows:

$$b_1 = \frac{COV(X,Y)}{VAR(X)}$$

$$= \frac{COV(X,Y)}{[SD(X)]^2} \frac{SD(Y)}{SD(Y)}$$

$$= \frac{COV(X,Y)}{SD(X)SD(Y)} \cdot \frac{SD(Y)}{SD(X)}$$

$$= r \cdot \frac{SD(Y)}{SD(X)}$$

The fact that b_1 is equal to Pearson correlation coefficient r times the standard deviation of Y divided by the standard deviation of X highlights the relationship between the estimated slope coefficient, the Pearson correlation coefficient r, and the variability of the variables X and Y. The Pearson correlation coefficient r quantifies the strength and direction of the linear relationship between *X* and *Y*. The standard deviation measures the spread or dispersion of the data points around the mean. Dividing the standard deviation of *Y* by the standard deviation of *X* normalizes the variability in both variables. If one variable has a larger spread than the other, the slope coefficient adjusts accordingly to account for this difference in variability. Multiplying *r* by the ratio of standard deviations scales the correlation coefficient appropriately based on the variability of *X* and *Y*. This ensures that the estimated slope b_1 captures the change in *Y* per unit change in *X* in a standardized manner, reflecting the strength of the linear relationship relative to the variability of the variables.

With the results above, we can redefine the optimal estimated Simple Linear Regression (SLR) equation as follows:

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = b_0 + r \cdot SD(Y) \cdot \frac{X}{SD(X)}$$

In other words, for each standard deviation increase in the independent variable X, the dependent variable Y is expected to change by a magnitude of r standard deviations of Y, where r represents the Pearson correlation coefficient between X and Y.

2.3 Statistical Performance Metrics for SLR

Next, we will assess the performance of our simple linear regression (SLR) model by introducing several quantitative performance metrics.

2.3.1

Simple linear regression quantifies how Y changes around its mean when X changes around its mean. SSE plays a pivotal role in simple linear regression (SLR). Here, we introduce two similar terms to SSE. One is the sum of the total squared variation of Y around its mean $\bar{Y}$ (SST); the other, the sum of the

squared variation of $\hat{Y}$ around $\bar{Y}$. The latter term is frequently referred to as the sum of squares due to regression, abbreviated as SSR.

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

Interestingly, the sum of squared total variation of Y around its mean equals the sum of squares due to regression plus the sum of squared errors, expressed as SST = SSR + SSE (mathematical proof omitted). Intuitively, the larger the proportion of SSR over SST, the better the SLR model. This leads us to define our first performance metric: the coefficient of determination, more famously known as r-squared.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

R-squared effectively quantifies the proportion of the total variation of Y around its mean that can be explained by the regression. A higher r-squared indicates a better-fitting model, with values ranging between 0 and 1, inclusive.

You might be curious about the relationship between the "r" in r-squared and the Pearson correlation coefficient. In simple linear regression (SLR), the coefficient of determination, or r-squared, equals the square of the Pearson correlation coefficient between X and Y. The proof of this result is quite straightforward, leveraging two earlier results: $b_0 = \bar{Y} - b_1\bar{X}$ and $b_1 = r \cdot \frac{SD(Y)}{SD(X)}$.

But why not just use the Pearson correlation coefficient to assess the SLR model? Well, Pearson correlation coefficient can be negative, so its magnitude is a more informative metric. In fact, the absolute value of the Pearson correlation coefficient, or the positive square root of the coefficient of determination, known as the multiple correlation coefficient (or multiple r), serves as a second performance metric for SLR.

2.3.2 MSE and Standard Error

Residuals (e)represent realizations of the random errors (ε). The variance of the residuals can serve as an estimate of the variance of the error term, commonly referred to as the mean squared error (MSE). The mean squared error (MSE) is calculated as the sum of squared errors (SSE) divided by the degrees of freedom of residuals (n-2), where n represents the sample size. The term "degrees of freedom" can be difficult to explain concisely. Instead in data science and machine learning, people often calculate the mean squared error (MSE) as the sum of squared errors (SSE) divided by the sample size (n), avoiding the need to delve into the concept of degrees of freedom. Similarly, the standard deviation of the residuals, often termed the standard error, provides an estimate of the standard deviation of the error term.

Both MSE and standard error are frequently utilized as indicators of the quality of linear regression. Unlike coefficient of determination and multiple r, MSE and standard error are not normalized, making them difficult to use for comparisons.

2.4 Standardization and SLR

We previously explored standardization as a significant feature scaling technique. Now, let's delve into why standardization holds particular appeal for data analysts and scientists.

Let's represent the z-scores resulting from the standardization of X and Y as z^X and z^Y. Their means and standard deviations are 0 and 1, respectively.

$$z^X = \frac{X-\bar{X}}{SD(X)} \text{ and } z^Y = \frac{Y-\bar{Y}}{SD(Y)}$$

$$\bar{z}^X = \bar{z}^Y = 0 \text{ and } SD(z^X) = SD(z^Y) = 1$$

The information provided, combined with the definition of Pearson correlation coefficient, allows us to determine that the correlation coefficient between z^X and z^Y, denoted as r^Z is identical to the correlation coefficient between X and Y, denoted as r. If you're interested, you're welcome to derive these results more formally or mathematically.

Now, let's explore a simple linear regression (SLR) model between z^X and z^Y. Here is the estimated SLR equation:

$$\hat{z}^Y = b_0^z + b_1^z z^X$$

We know from earlier that $b_1 = r \cdot \frac{SD(Y)}{SD(X)}$. Therefore, $b_1^z = r^Z \cdot \frac{SD(z^Y)}{SD(z^X)} = r^Z = r$. And similarly, $b_0^z = \bar{z}^Y - b_1^z \bar{z}^X = 0$.

I highly recommend plotting the scatter plot of z^X against z^Y and adding the linear trend line. You'll observe that the trend line or the regression line passes through the origin, and its slope corresponds to the Pearson correlation coefficient of X and Y.

2.5 Concluding Remarks

We conclude this section with two important remarks. Firstly, it's essential to note that ordinary least squares (OLS) regression, the method used to find the optimal estimated slope and intercept in simple linear regression (SLR), does not strictly rely on the normality assumption. Even without this assumption, OLS can still derive these estimates. However, without normality assumptions, our analyses are primarily limited to point estimates. It's the normality assumption that enables us to make various statistical inferences, such as hypothesis testing and constructing confidence intervals.

Secondly, it's crucial to understand that SLR provides a quantitative estimation of the expected change in Y around its mean when X changes around its mean. However, it does not imply causality. In other words, if we were to switch the roles of X and Y, we would obtain similar quantitative results. Only when we establish that X influences Y can we interpret SLR as quantifying the effect of X on Y. For instance, in our housing price example, SLR quantifies the effect of square footage of living space on housing price, assuming this relationship holds true. Similar to correlation, SLR merely quantifies the linear association between X and Y. Indeed, SLR and correlation are closely related, with SLR providing a different perspective on the linear association between the variables.

3. Interpretations with Example

Now, let's examine our SLR model, where X represents 'sqft_living' and Y represents 'price'. Below is the regression output generated from Excel. We'll concentrate on interpreting some of the key results from this output.

Regression Statistics						
Multiple R	0.70203505					
R Square	0.49285322					
Adjusted R Square	0.49282975					
Standard Error	261452.888					
Observations	21613					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	1.4356E+15	1.4356E+15	21001.9096	0	
Residual	21611	1.4773E+15	6.8358E+10			
Total	21612	2.9129E+15				
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-43580.743	4402.68969	-9.8986634	4.7234E-23	-52210.33964	-34951.14655
sqft_living	280.623568	1.93639855	144.920356	0	276.8280839	284.4190519

Firstly, let's examine the coefficient of 'sqft_living', represented by b_1 in the SLR model. Its value is 280.62. That is, we anticipate that for every additional square foot in a house's living area, the price is expected to increase by $280.62. This reflects an average expectation. It's important to note that while each individual house may not increase in price by exactly $280.62 for each additional square foot, the average increase across all houses would be around $280.62.

Once again, it's crucial to emphasize the linear assumption. Due to the linear nature of the equation, the incremental impact of 'sqft_living' on 'price' remains constant. However, intuitively, we understand that adding 100 square feet to a 1000 square foot house would have a more substantial impact than adding the same to a 10000 square foot house. This underscores the presence of some non-linear relationships. Therefore, to ensure the effectiveness of our SLR model, we must initially verify that a linear equation adequately approximates the true relationship between 'sqft_living' and 'price'.

The intercept value of -43580 implies that according to the model, a property with zero square feet of living space would have a predicted price of negative $43,580. This interpretation seems nonsensical because even a property without living space could still hold some value, such as in the case of land. Additionally, it's worth noting that the dataset contains no data points with living space below 290 square feet. Therefore, using the model to predict the prices of properties with little to no living space may not be reliable, as the relationship between price and living space may not be linear in that range, or it may follow a different linear pattern with distinct slope and intercept values.

The ANOVA table within the regression output provides several key values, including SSR (Sum of Squares Regression), SSE (Sum of Squares Error), SST (Total Sum of Squares), MSR (Mean Square Regression), and MSE (Mean Square Error). MSE serves as an estimate for the variance of the error

term. The standard error, the square root of MSE, is the estimated standard deviation of the error term. In this specific instance, the standard error is calculated at $271,452.

This standard deviation appears considerable, particularly when juxtaposed with the average price of $540,088. Such a large standard deviation suggests that relying solely on this simple linear regression model may not be advisable. Typically, there are myriad other factors at play in real-world scenarios, rendering a simplistic linear model insufficient. This concern is further substantiated by the relatively low R-squared value of 0.4928, indicating that only 49.28% of the squared variation in price around its mean can potentially be attributed to changes in square footage of living space.

The multiple correlation coefficient, denoted as multiple r, represents the positive square root of the coefficient of determination (R^2). In this instance, it stands at 0.7, signifying the strength of the linear relationship between 'sqft_living' (and consequently the predicted price) and the actual price. Notably, multiple r is always non-negative.

In this specific case, the multiple correlation coefficient equals the Pearson correlation coefficient between 'sqft_living' and 'price' due to their positive correlation, resulting in a positive Pearson correlation coefficient.

3.1 Residual Analysis - Normality and Homoscedasticity

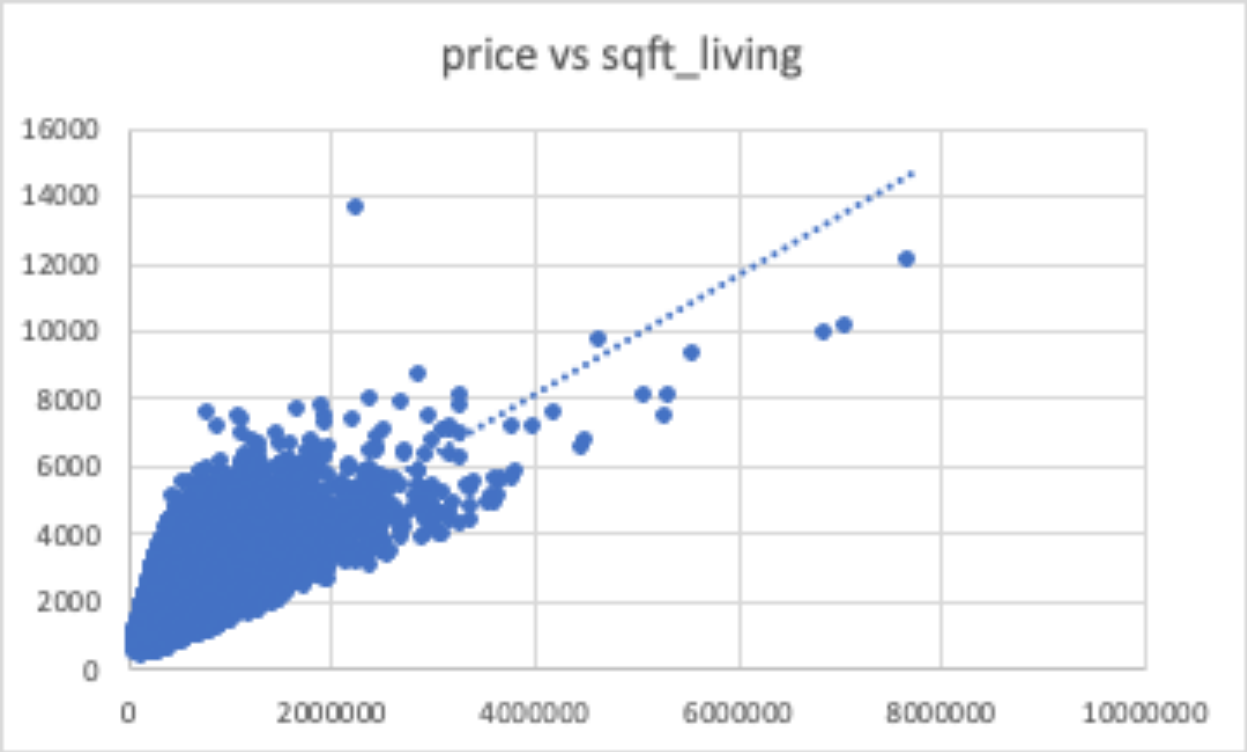
Residual analysis in linear regression involves examining the differences between observed values and the values predicted by the regression model. These differences, known as residuals, are used to evaluate the model's assumptions and performance. Residual analysis helps assess whether the residuals exhibit patterns or systematic deviations from randomness, which can indicate issues such as nonlinearity, heteroscedasticity, or outliers in the data. By analyzing residuals, researchers can diagnose problems with the regression model and make adjustments to improve its validity and predictive accuracy. Testing for normality and homoscedasticity are fundamental steps in validating linear regression models. Residual analysis serves as a reliable method to assess the reasonableness of these assumptions.

A test of normality based on residuals is used to assess whether the errors in a regression model follow a normal distribution. This test examines if the residuals, which are the differences between the observed and predicted values, are normally distributed. It helps validate one of the key assumptions of linear regression, ensuring that the model accurately captures the underlying randomness in the data. If the residuals significantly deviate from a normal distribution, it may indicate that the model's assumptions are violated, potentially affecting the reliability of the regression analysis.

As an illustration, we could employ the residuals obtained from our 'sqft_living' vs. 'price' simple linear regression (SLR) model to conduct a normality test. The test of normality, akin to the test of independence, involves binarizing or discretizing the residuals, typically into 10 deciles. Each decile represents a portion of the distribution, allowing us to assess the normality assumption. By using a normal distribution as a reference, we establish threshold values for each decile. Next, we count the number of occurrences of residuals within each decile. Ideally, we anticipate approximately 10% of the total residuals, or 2161.3 observations, in each decile. To evaluate the goodness of fit, we compute the Chi-Square test statistic and its corresponding p-value. This statistical test helps determine whether the observed distribution of residuals significantly deviates from a normal distribution.

To check the homoscedasticity assumption in linear regression, you can use visual inspection methods or statistical tests. Plot the residuals (the differences between observed and predicted values) against the predicted values or the independent variable. Look for patterns in the plot. Ideally, the plot should show a random scatter of points with no discernible pattern. If the spread of residuals appears to change systematically with the predicted values or independent variable, it suggests heteroscedasticity. Common statistical test for homoscedasticity is the Breusch-Pagan test or White's test. We won't delve into the specifics of how these tests operate.

Occasionally, we may also notice heteroscedasticity in scatter plots. Take the scatter plot of 'sqft_living' vs. 'price' below as an example, where it seems that the variability in 'price' increases as 'sqft_living' increases, resulting in a funnel shape. This observation suggests that we should further investigate whether the homoscedasticity assumption holds true. However, we place less emphasis on these assumptions and tests. Instead, as mentioned earlier, we rely more on data science and machine learning techniques like dataset splitting and cross-validation to assess the models' validity and generalizability.



3.2 'view' vs. 'price'

Since I have a specific interest in 'view', I conducted a simple linear regression analysis with 'view' as the independent variable and 'price' as the dependent variable. I treated 'view' as a numerical variable, ranging from 0 to 4. Here are the results of the regression analysis:

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.39729349			
R Square	0.15784212			
Adjusted R S	0.15780315			
Standard Err	336917.341			
Observations	21613			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	4.5978E+14	4.5978E+14	4050.45898
Residual	21611	2.4531E+15	1.1351E+11	
Total	21612	2.9129E+15		
	Coefficients	*tandard Erro*	*t Stat*	*P-value*
Intercept	495491.941	2396.47563	206.758597	0
view	190335.248	2990.66042	63.6432163	0

The findings reveal that the model using 'view' as the predictor performs poorly compared to the 'sqft_living' model. With an R-square value of only 0.1578 and an exceptionally high standard error of $336,917, the predictive capability of the model is limited. The slope coefficient for 'view' is $190,335, suggesting that for every unit increase in 'view' level, the predicted price increases by $190,335. However, this implies that the difference in price between 'view' levels of 1 and 0 is the same as the difference between levels 4 and 3, which is illogical. The table provided below summarizes the predicted price solely based on the 'view' variable.

view	price predict
0	$495,492
1	$685,827
2	$876,162
3	$1,066,498
4	$1,256,833

Some might consider applying one-hot encoding to the 'view' variable and utilizing four binary variables in a linear regression model. This approach is indeed promising. By implementing such a model, we can create binary variables such as 'view1' to 'view4', each indicating whether a house's view holds a particular value. The coefficients of these binary variables can be interpreted straightforwardly. For instance, the coefficient of 'view1' would represent the average price of houses with a view value of 1,

and similarly for the other binary variables. The average prices corresponding to different view values are as follows:

view	count	%	Ave. Price
0	19489	90.173%	$496,564
1	332	1.536%	$812,281
2	963	4.456%	$792,401
3	510	2.360%	$971,965
4	319	1.476%	$1,463,711
total	21613	1	

Surprisingly or not, we observe an anomaly: the average price of houses with a view value of 1 is higher than that of houses with a view value of 2. This observation may be attributed to the subjective evaluation of the view, but it could also suggest the presence of Simpson's paradox. For example, houses with a view value of 2 might be situated in less affluent neighborhoods, warranting further exploration. It's worth considering employing a more sophisticated model to discern or possibly combine view values 1 and 2. Ultimately, data science and analytics, along with their applications, entail more art than science.

4. Limitation of SLR

Simple Linear Regression (SLR) tends to oversimplify by exclusively spotlighting one feature, attributing observed effects solely to that variable. In doing so, it risks overlooking the nuanced contributions of other factors. As illustrated by Simpson’s paradox, there are often additional variables influencing the target variable, either directly or indirectly. Consequently, relying solely on SLR can be inadequate for comprehensively understanding the relationship between X and Y. Therefore, while SLR offers insights into the impact of a single feature, it's crucial to recognize its limitations and consider more comprehensive approaches when studying complex relationships.