Assessing Residential Building Costs with Economic Indicators in Calgary

Oct 17, 2023

DATA 602 – Final Project

Arie DeKraker, Crystal Wai, Luke Bramfield, Rozita Ghasemi

Introduction

From the statistical methods taught in class, our project has focused on modeling the relationship between various predictor variables in our dataset and our chosen response variable - “estimated project cost" for completed residential building permits in the city of Calgary. This topic is motivated by the housing affordability crisis that Calgary is currently experiencing. According to the City of Calgary 75% of Calgary households have insufficient income to buy a single-family home (City of Calgary, n.d.). Our project can be viewed as a preliminary investigation to assess what features may be important in understanding residential building costs. This type of research could be useful for Calgary policy and decision makers, real estate agencies, and Calgary residents who want to make informed decisions about their housing.

Data Preparation

Data Sources:

All datasets are from Calgary Open Data, and we have permission to use the dataset as it described on the Open Calgary website ("Open Calgary Terms of Use") where they state - "Open data means it is easily accessible to the public, free to use, re-use and redistribute without any legal, technological or social restriction."

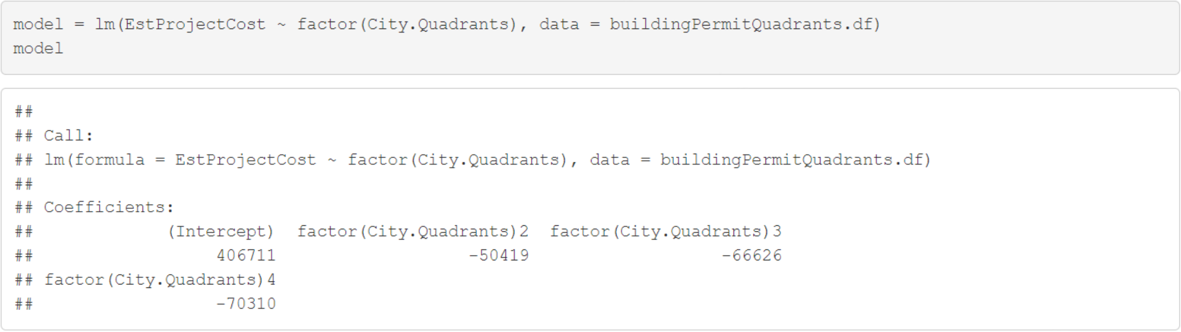
Our primary datasets are *Monthly Economic Indicators* and *Building Permits*, both found on Open Calgary (The City of Calgary, 2023). Each data set contains the features that help facilitate the investigation of the relationship between Calgary's estimated project cost and our chosen predictor variables. Both datasets contain structured data in a tabular(.CSV) format that will be read into, wrangled, and analyzed in R.

We refined our dataset to encompass the years from 2000 to 2023, omitting any incomplete records. Our analysis is centered on completed residential projects, ensuring that anomalies like white spaces, negative project costs, and negative square footage were eliminated.

Guiding Question #1 – How does the Quadrant in Calgary affect the estimated project cost?

We approached this question by developing a multiple categorical linear regression model that predicts the average estimated project cost in the city quadrant (Quadrant 1 - SW, Quadrant 2 - NW, Quadrant 3 - SE, Quadrant 4 - NE).

Compute Our Model Using R



```
model = lm(EstProjectCost ~ factor(City.Quadrants), data = buildingPermitQuadrants.df)
model
```

```
##
## Call:
## lm(formula = EstProjectCost ~ factor(City.Quadrants), data = buildingPermitQuadrants.df)
##
## Coefficients:
##            (Intercept)  factor(City.Quadrants)2  factor(City.Quadrants)3
##                 406711                   -50419                   -66626
## factor(City.Quadrants)4
##                 -70310
```

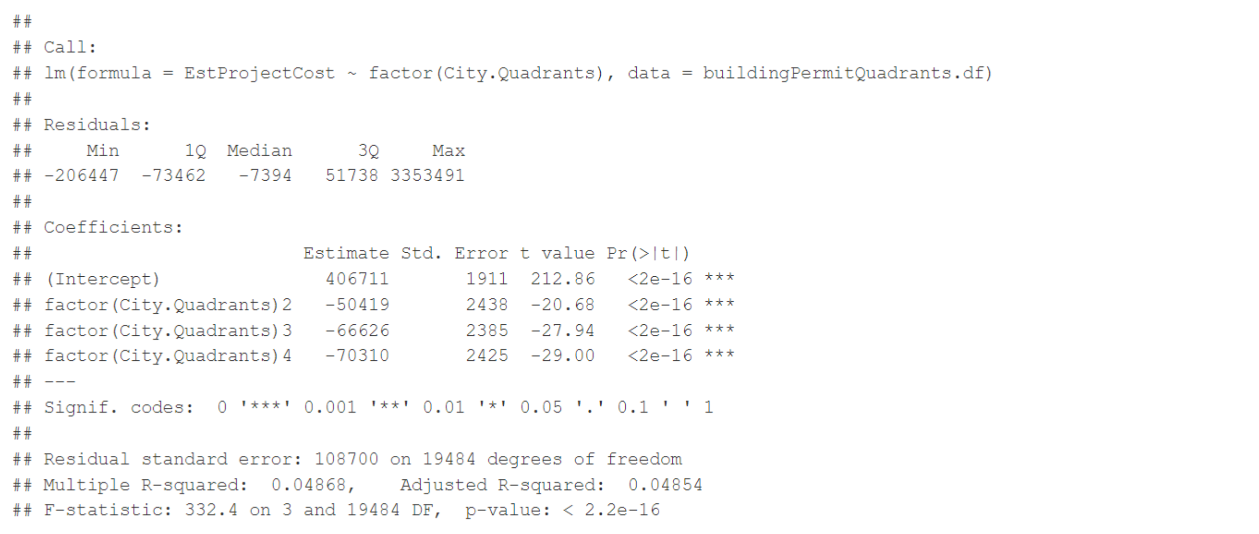
The model that we computed was:

$$Predicted\widehat{Estimated}Cost = 406710.823 + (-50418.68 * Quadrant\ 2 - 66625.82 * Quadrant\ 3 - 70310.05 * Quadrant\ 4)$$

Interpreting the model

The reference beta was set to quadrant one by default. The betas computed for quadrants 2, 3, and 4 represent the average adjustment value with respect to the reference beta. These values predict the average estimated project cost for the specified city quadrant.

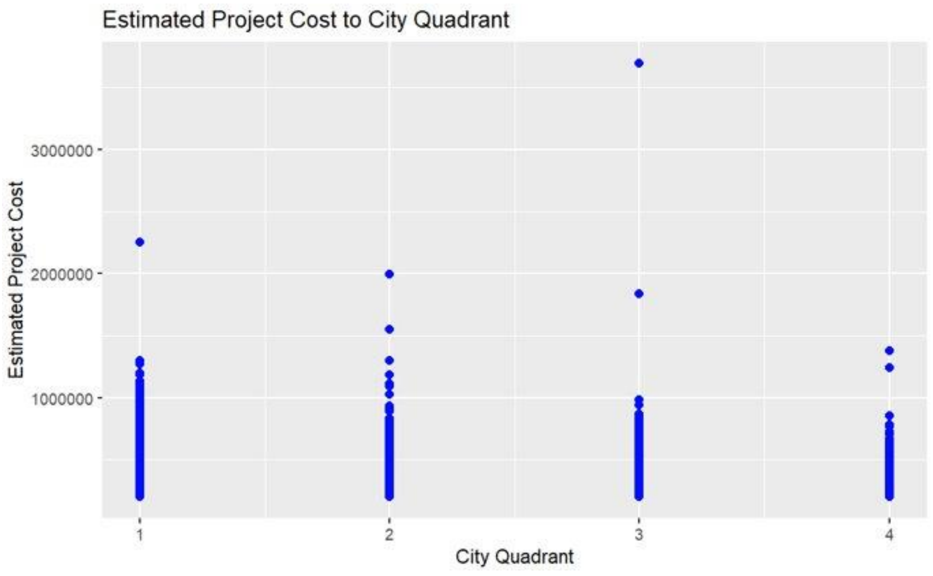
Estimating the model in R



```
##
## Call:
## lm(formula = EstProjectCost ~ factor(City.Quadrants), data = buildingPermitQuadrants.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206447  -73462   -7394   51738 3353491
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               406711       1911  212.86   <2e-16 ***
## factor(City.Quadrants)2   -50419       2438  -20.68   <2e-16 ***
## factor(City.Quadrants)3   -66626       2385  -27.94   <2e-16 ***
## factor(City.Quadrants)4   -70310       2425  -29.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108700 on 19484 degrees of freedom
## Multiple R-squared:  0.04868,    Adjusted R-squared:  0.04854
## F-statistic: 332.4 on 3 and 19484 DF,  p-value: < 2.2e-16
```

Evaluating model significance

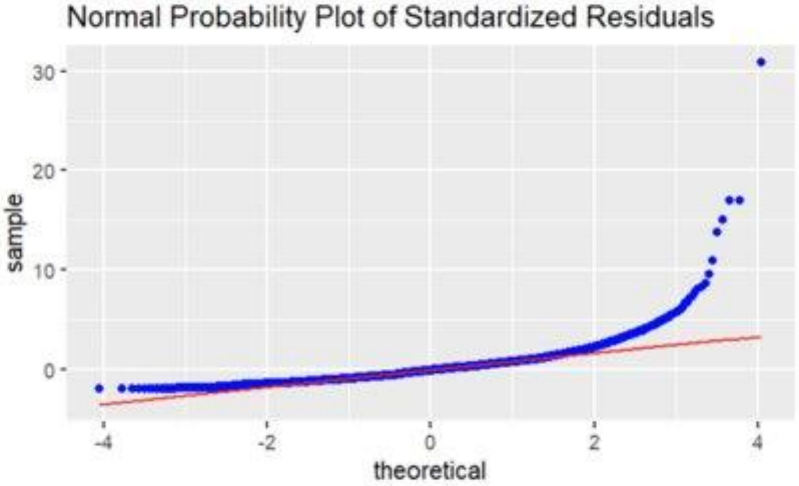
To evaluate our linear model computed, it is important to note our hypothesis test. Which is H0: A = 0, HA: A/=0 and H0: B = 0, HA: B/=0 (for all beta terms). This model explains 4.86% of the variability in estimated project cost. Although, our r squared value is low, since the p-value is less than 0.05, we know that the estimated project cost is related to the city quadrant. From our computed p-value(2e-16), we also found for quadrants 1 to 4 that there is statistically significant relationship with estimated project cost and city quadrant. We can infer that because our p-value our city quadrants are statistically different in comparison to the reference quadrant (Q1).



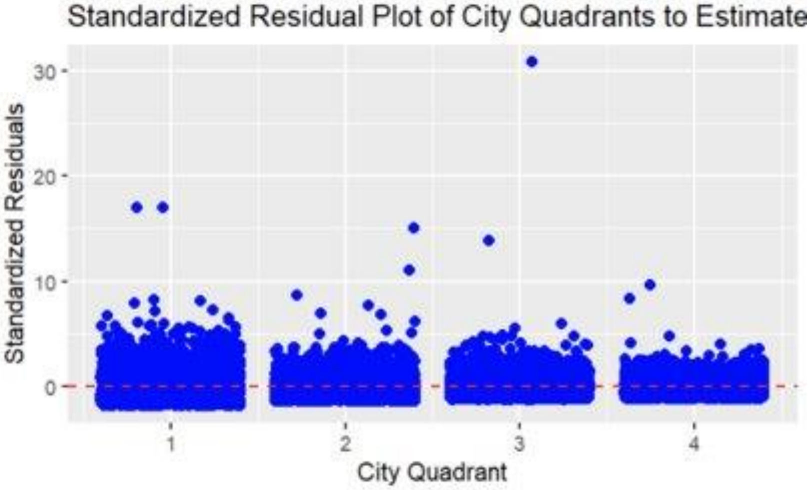
Here we visualized the data to summarize the estimated project cost to the city quadrant.

Model Validation

Condition - Normality of the residuals
From the normal probability plot of standardized residuals, we can observe that the residuals follow the normal distribution, with a few outliers. An assumption made is that there will be larger projects included in the dataset that are outside of the range of an average project cost.



Condition - Homoscedasticity
From the residual plot computed below we can observe that there is a "rectangular" distribution, therefore, we know that common variance holds centered at a residual common spread. Which meets our condition of homoscedasticity.



Guiding Question #2 How is a residential building Permits Estimated project cost affected by the contractor it is awarded to?

To answer this question, we developed a multiple categorical linear regression model that predicts the average estimated project cost if that particular contractor is awarded the bid. Because there were over 800 contractors in our data set, we focused on the top 7 contractors that had more than 1,000 contracts. These 7 contractors make up 44.47% of the proportion of all contracts in the data set. Outside of the top 10 contractors you see the number of contracts start to taper off quickly, indicating that these 7 (with the addition of a few others) are the dominant homebuilders in the Calgary housing market.

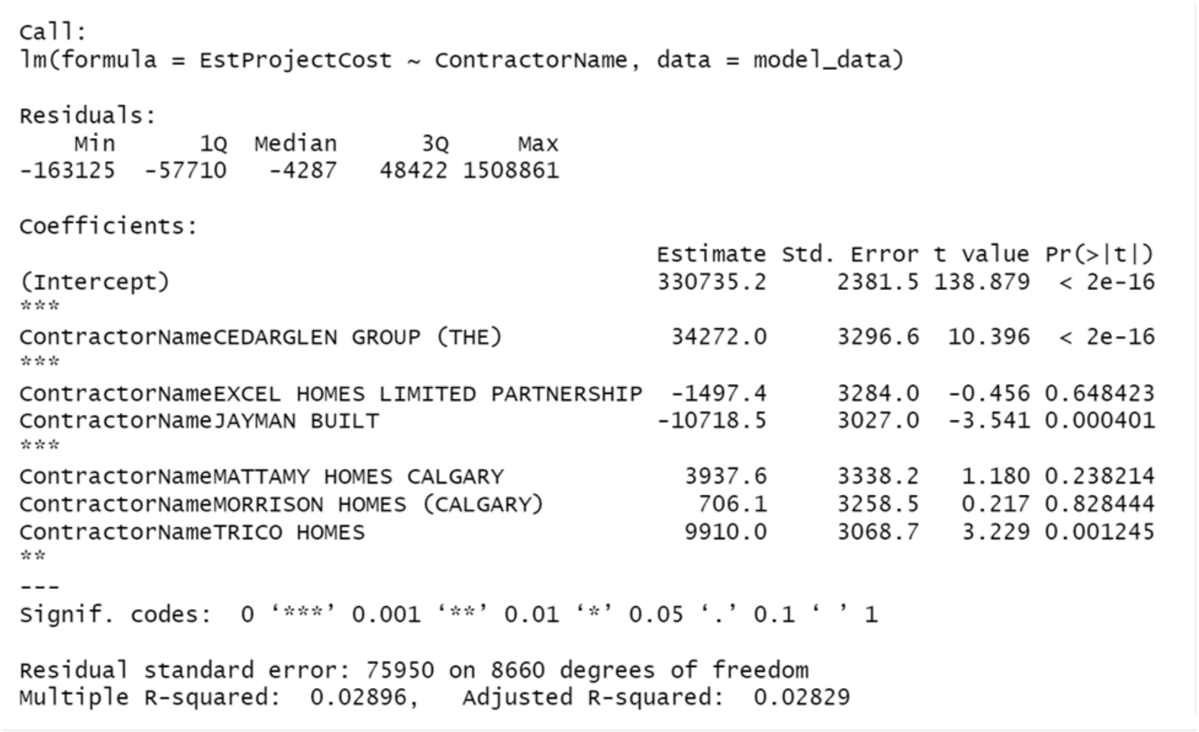
The model we will be creating can be expressed by the following equation:

$$\text{EstProjectCost} = 330735 + 34272 \times CG - 1497.4 \times EH - 10718.5 \times JB + 3937.6 \times MatH + 706.1 \times MorH + 9910.0 \times TH$$

Interpreting the model

The equation and the output below represent a categorical multiple linear regression model that predicts what the average estimated project cost for a building permit is, given which of the 7 contractors is awarded the contract. Because we are dealing with multiple categorical variables in our regression, we must introduce dummy variables to represent the presence (coded as a 1) or absence (coded as a 0) of that contractor. The reference variable for the model is set to "BROOKFIELD RESIDENTIAL" which represents the y int and the average estimated project cost when the contractor selected is Brookfield (the contractor chosen by default). The other Betas (contractors) in the model indicate the average adjustment that needs to be made to the average Brookfield project cost if that contractor is selected instead, which then predicts the average estimated project cost for that specific contractor.

Estimating the model in R:

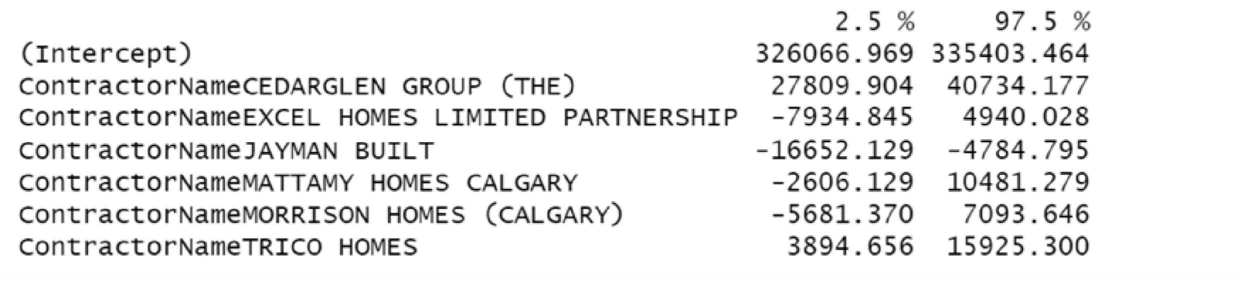


```
Call:
lm(formula = EstProjectCost ~ ContractorName, data = model_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-163125  -57710   -4287   48422 1508861 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               330735.2     2381.5 138.879  < 2e-16
***
ContractorNameCEDARGLEN GROUP (THE)        34272.0     3296.6  10.396  < 2e-16
***
ContractorNameEXCEL HOMES LIMITED PARTNERSHIP  -1497.4     3284.0  -0.456 0.648423
ContractorNameJAYMAN BUILT                -10718.5     3027.0  -3.541 0.000401
***
ContractorNameMATTAMY HOMES CALGARY         3937.6     3338.2   1.180 0.238214
ContractorNameMORRISON HOMES (CALGARY)       706.1     3258.5   0.217 0.828444
ContractorNameTRICO HOMES                   9910.0     3068.7   3.229 0.001245
**
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75950 on 8660 degrees of freedom
Multiple R-squared:  0.02896,	Adjusted R-squared:  0.02829 
```

Evaluating model significance:

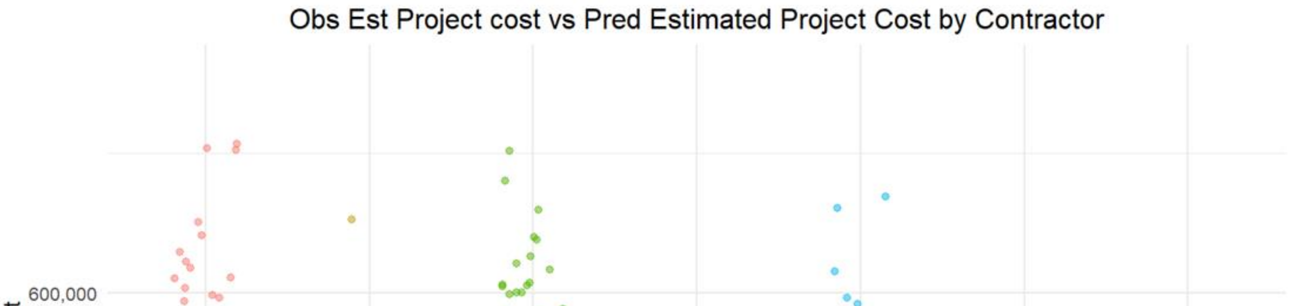
Further evaluating the model lets us look at the test statistics and p-values of our contractor variables. Please note that all hypothesis tests completed by R are testing H0: A = 0, HA: A =/0 and H0: B = 0, HA: B =/0 (for all contr. With this in mind, we can determine the significance of the model starting with Brookfield Residential the y int, and the reference group. We can see the model estimates a Brookfield Residential project cost as $330,735. The t statistic is reported as 138.88 which equates to a p-value extremely close to zero, therefore we can reject the null and conclude our estimate as statistically significant. Now the rest of the contractors compared pairwise to this reference group to see if there is a statistically significant difference in the estimated project cost of that specific contractor and Brookfield residential. Here we can see Cedarglen group, Jayman Built, and Trico homes report p-values of ~0, 0,.0004, and .0012. The interpretation of these p-values can all be described as the likelihood of obtaining another random sample with at least as strong evidence against the null as what we currently have. The inference is that we can statistically say that these contractors do have a different estimated project cost compared to Brookfield. If our beta terms are positive, it indicates that those contractors on average have more expensive projects than Brookfield and if negative less expensive projects than Brookfield. The other contractors included in the model - Excel, Mattamy, and Morrison have p values greater than our typical significance level of .05 and therefore we cannot reject the null B = 0 or that the estimated project cost is different than that of Brookfield.

95% Confidence Interval Estimation:



```
                                                2.5 %     97.5 %
(Intercept)                                326066.969 335403.464
ContractorNameCEDARGLEN GROUP (THE)         27809.904  40734.177
ContractorNameEXCEL HOMES LIMITED PARTNERSHIP  -7934.845   4940.028
ContractorNameJAYMAN BUILT                 -16652.129  -4784.795
ContractorNameMATTAMY HOMES CALGARY         -2606.129  10481.279
ContractorNameMORRISON HOMES (CALGARY)      -5681.370   7093.646
ContractorNameTRICO HOMES                    3894.656  15925.300
```

Examining the 95% confidence intervals of the terms of the model supports the previous conclusion of what contractors yield significant results vs those who do not. For example, we can see the insignificant terms confidence intervals contain 0.

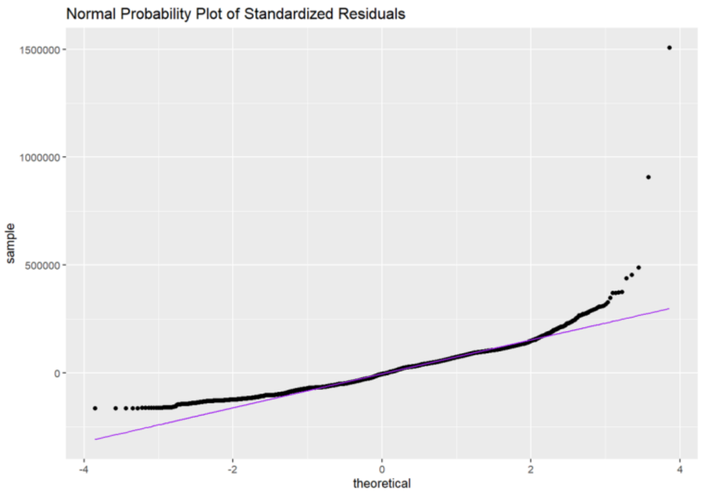


This image beautifully summarizes and depicts exactly what is happening in the model. Each of the building permits estimated project costs are grouped by their respective contractor. The red line indicates the estimated project cost determined by the regression. Please note that some substantial outliers were excluded from the visualization due to the scaling of the y axis but were not excluded from the derivation of the model. This model also portrays that the regression would fail to accurately predict many project costs as the range of project costs is so wide.

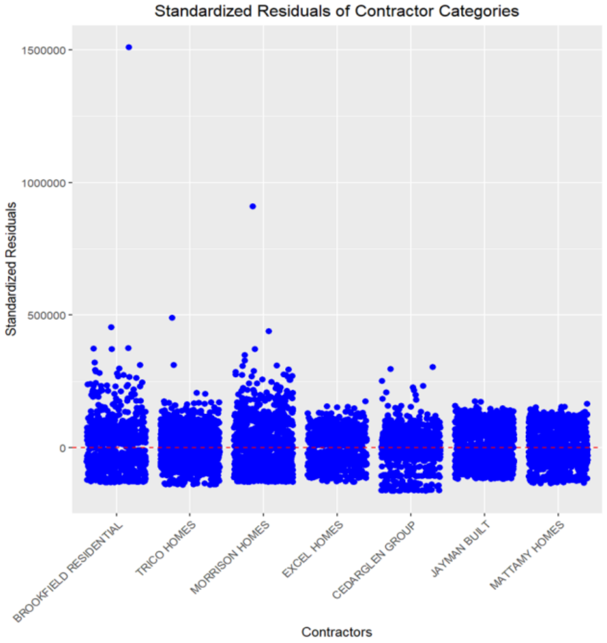
Model Validation

Condition 1 – Normality of the residuals

From the normality of the residuals plot it is evident that the residuals approximately follow a normal distribution apart from a few extreme values due to outliers in the dataset (very high estimated project costs) outside of the typical range of project costs.



Condition 2 – Homoscedasticity



Plotting the residuals of each contractor we can see the error terms are centered around 0 and the distribution errors appear to be consistent among each contractor. Therefore, we can conclude the condition of Homoscedasticity is met.

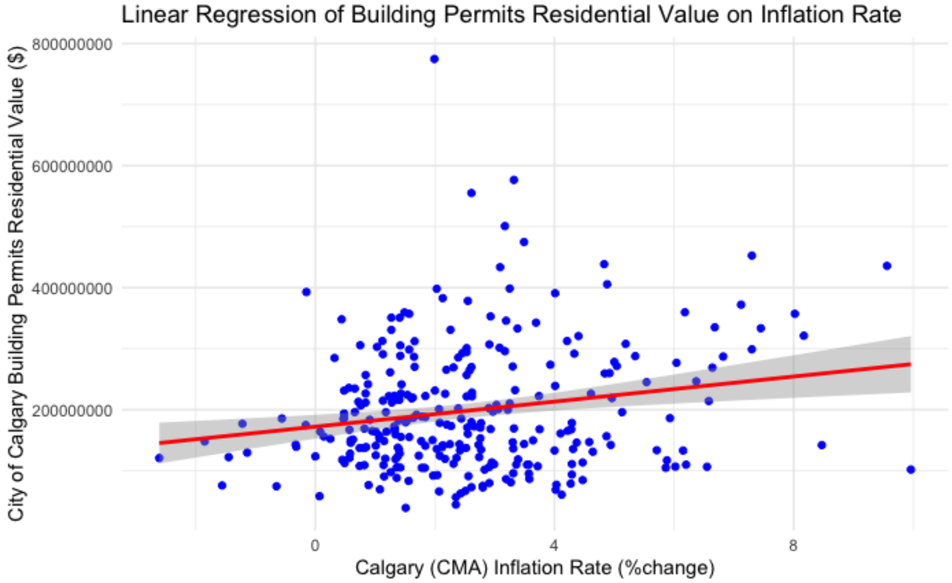
Guiding Questions #3 Is there a correlation between residential permit values and inflation rates?

The reason behind this statistical test is we wanted to see if there is a correlation expressed between the total value of residential building permits per month and the inflation rate expressed that month. To evaluate this relationship, a testable hypothesis was created and provided below.

$$\mathrm{H}_0 : \beta_1 = 0 (\textit{The total residential building permit value : is not positively related to the inflation rate per month.})$$
$$\mathrm{H}_A : \beta_1 > 0 (\textit{The total residential building permit value : is positively related to the inflation rate per month.})$$

Creating the model

We created a linear model where the inflation rate during that month is on the x axis and the total residential building permit value is on the y axis.

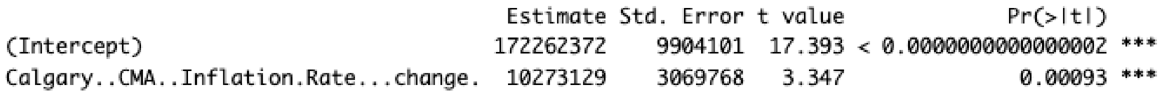


The line of best fit indicates a mild upward trend, though there is noticeable variation in the inflation rates throughout the model.

By running further analysis on the data by our two variables of interest, we are able to get a B0 value of $172,262,372 and a B1 value of $10,273,129. With this information we are able to create an estimation model which is shown below.

$$y_i = \beta_0 + \beta_1 \text{x(inflation rate)} + e_i$$

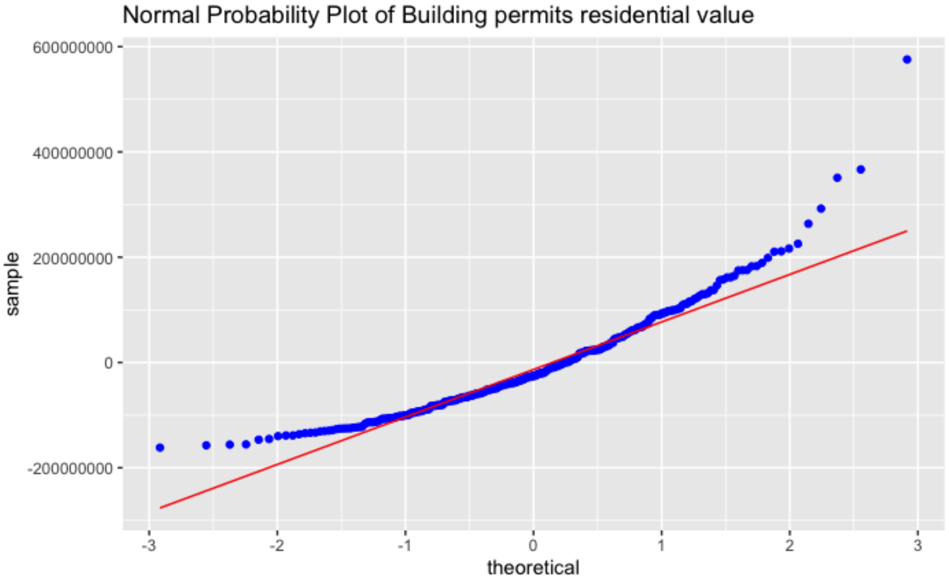
$$\widehat{y}_i = 172262372 + 10273129\widehat{x}_i$$



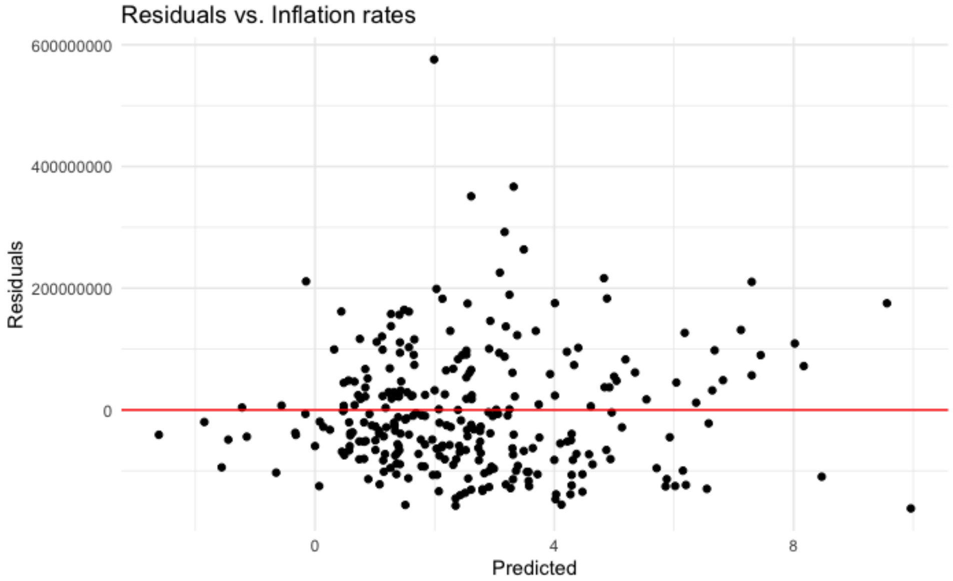
```
                                Estimate Std. Error t value              Pr(>|t|)    
(Intercept)                    172262372    9904101  17.393 < 0.0000000000000002 ***
Calgary..CMA..Inflation.Rate...change.  10273129    3069768   3.347              0.00093 ***
```

Model Validation

As part of linear regression testing there are a few conditions that need to be satisfied before we can decide that the conclusions that we come to are accurate. These conditions include linearity, normality, and homoscedasticity. The linearity has already been confirmed by the linear regression model provided above. To determine if these conditions are met, we will examine the plots below.

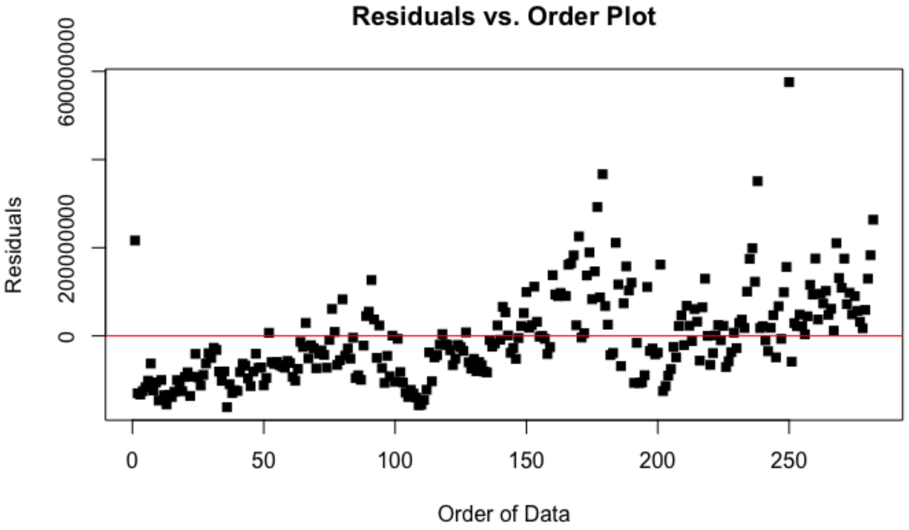


The purpose of running a normal probability plot with our residuals is to make sure that all points fall relatively close to the theoretical normal distribution. Since our points all fall close to the line (except a few major outliers expressed at the end of the plot), we can assume that the data is approximately normally distributed. Since the data expresses normality, we can accept the outcomes of our linear regression if the other conditions are met.



The next condition that needs to be met is homoscedasticity. We are able to represent this through a scatterplot that has inflation rates on the x-axis and residuals on the y-axis. What we are looking for here are any shapes that might violate the condition of homoscedasticity, such as a wedge or funnel shape. If we were to see one of these shapes, we would conclude that the model expresses heteroscedasticity, and we would not be able to rely on any outcomes found within the linear regression.

When looking at the graph it seems like most points are randomly scattered and we see no identifiable shape that would cause us to reject the homoscedasticity condition. We are then able to conclude that the model expresses homoscedasticity and if all other conditions are met, we can accept the outcomes of the linear regression.



The final condition that we are looking to confirm is whether or not the variables express independence. One way we can check this is by running an order plot of the residuals. What we are looking for in this model is whether or not there are any inherent patterns. If we are to see a pattern, we would come to the conclusion that there is some resemblance of dependency expressed by both variables. When looking at the graph it seems that there is no pattern expressed. The points appear to be randomly scattered with most values falling near the line of zero. Since we notice no pattern, we can conclude that the variables do not express independence.

Since all model conditions are satisfied, we can accept the outcomes from our linear regression.

Additional analysis

With these conditions met, we can look further into the results from our data. First, we should answer the original hypothesis on whether our data expresses a positive or negative correlation. We can determine this by calculating p and f values.

P-value = 0.0009302
F-value = 3.347
R^2 -value = 0.03846

With the p value being below 0.05, we reject the null hypothesis that there is no association between the two variables. We also found out B1 value was positive. From this, we conclude that there is a positive correlation between the total residential building permit value and the inflation rate per month.

Expected values and predictions

To find a good estimation of how permit values will increase with a rising inflation rate we calculated the 95% confidence interval for B1. From our results, we can see that the lower bound for the confidence interval is equivalent to 4,244,687 and the upper bound is equivalent to 16,301,570, which indicates how much the value would increase with a 1% increase in the inflation rate. Since the values of both the upper and lower bound are well above 0, we can conclude that B1 is statistically significant.

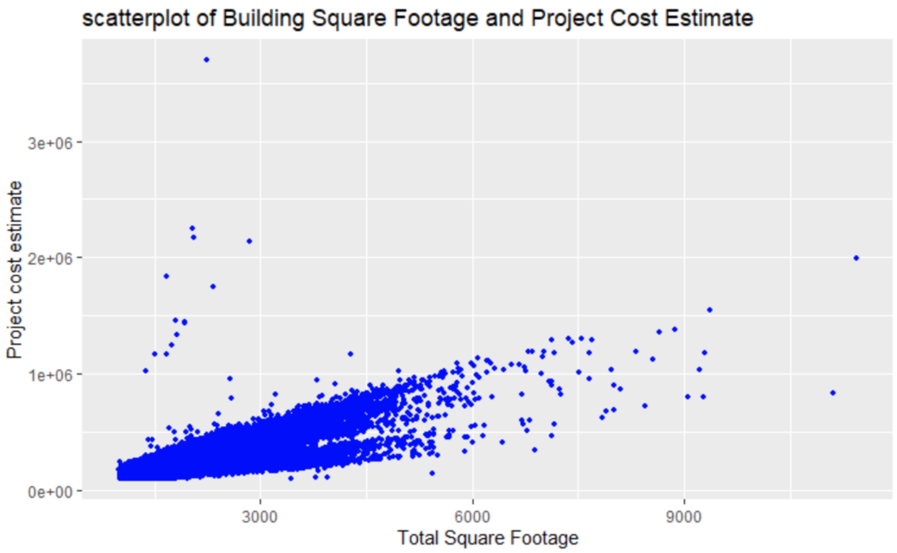
We can apply our model to predict what the total residential building permit value will be at certain inflation rate changes. For example, we calculated the 95% confidence interval for what the values would be at 2% and got the lower bound equivalent to 180,509,204 and the upper bound equivalent to 205,108,055.

Guiding Question #4 Is there a correlation between square footage of the project and the estimated project cost?

We employed linear regression to examine the relationship between square footage and estimated project cost. Hypothesis tests were conducted, the model was established, and the necessary conditions were checked for validation.

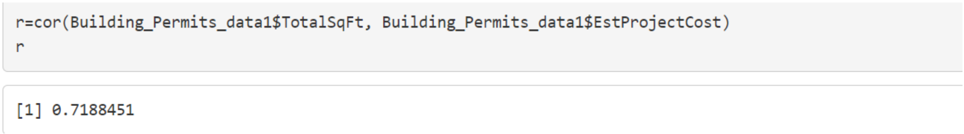
Regression Analysis – Relationship Between Square Footage and Project Estimated Cost

We would like to investigate the relationship between square footage and the estimated project cost. Firstly, to gain insight into the variables and estimate the existence of a relationship between them, we created a scatter plot of the two variables:



The distribution of these two variables shows that there is a linear relationship between them. This linear relationship can be shown by quantifying the relationship. That is, with the correlation coefficient:

Determining the Coefficient of determination in R

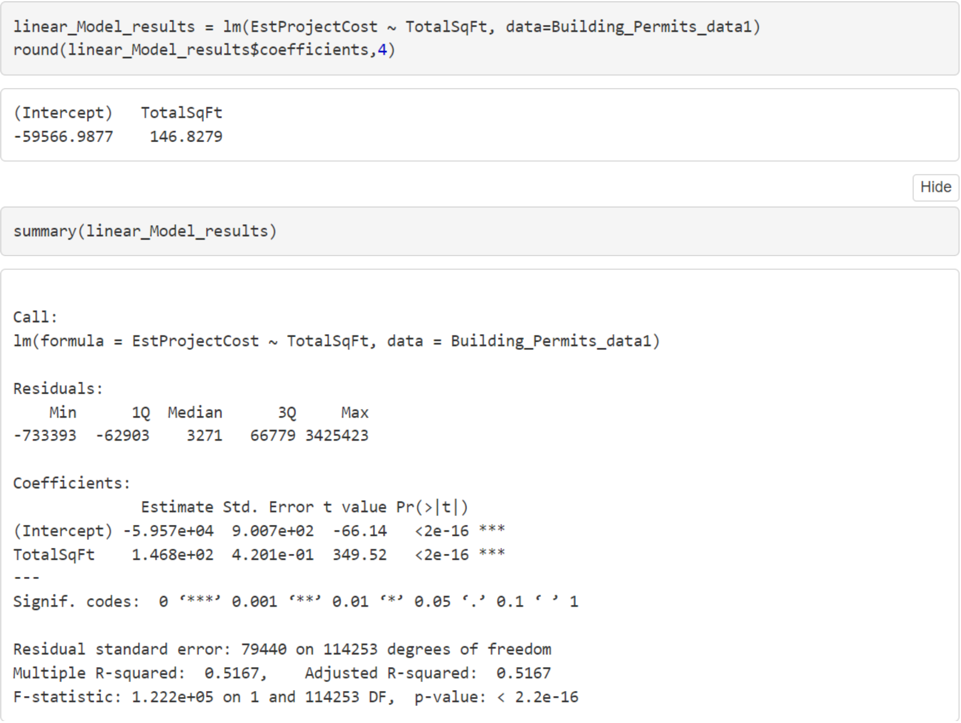


```
r=cor(Building_Permits_data1$TotalSqFt, Building_Permits_data1$EstProjectCost)
r
```

```
[1] 0.7188451
```

Model Development in R:

Now we will implement our model as below.



```
linear_Model_results = lm(EstProjectCost ~ TotalSqFt, data=Building_Permits_data1)
round(linear_Model_results$coefficients,4)
```

```
(Intercept)   TotalSqFt 
-59566.9877    146.8279 
```

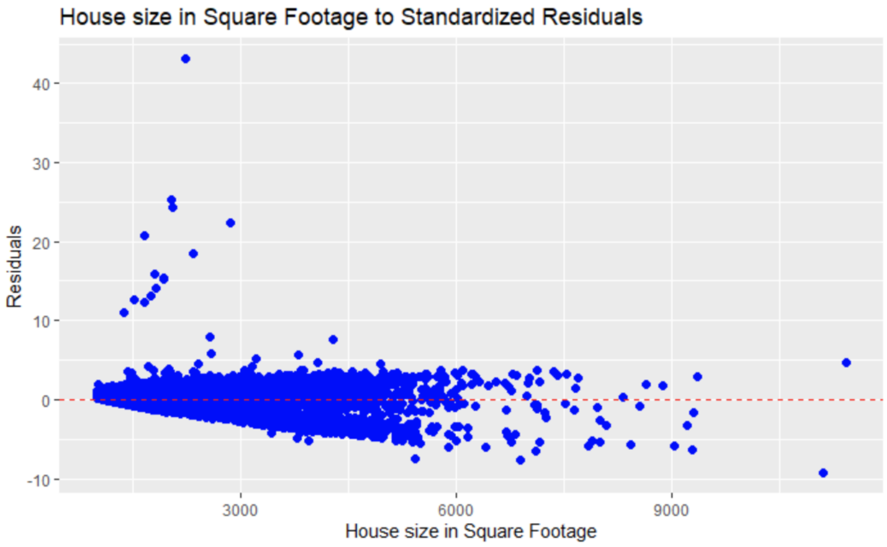
```
summary(linear_Model_results)
```

```

Call:
lm(formula = EstProjectCost ~ TotalSqFt, data = Building_Permits_data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-733393  -62903    3271   66779 3425423 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.957e+04  9.007e+02  -66.14   <2e-16 ***
TotalSqFt    1.468e+02  4.201e-01  349.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79440 on 114253 degrees of freedom
Multiple R-squared:  0.5167,	Adjusted R-squared:  0.5167 
F-statistic: 1.222e+05 on 1 and 114253 DF,  p-value: < 2.2e-16
```

Model Validation:

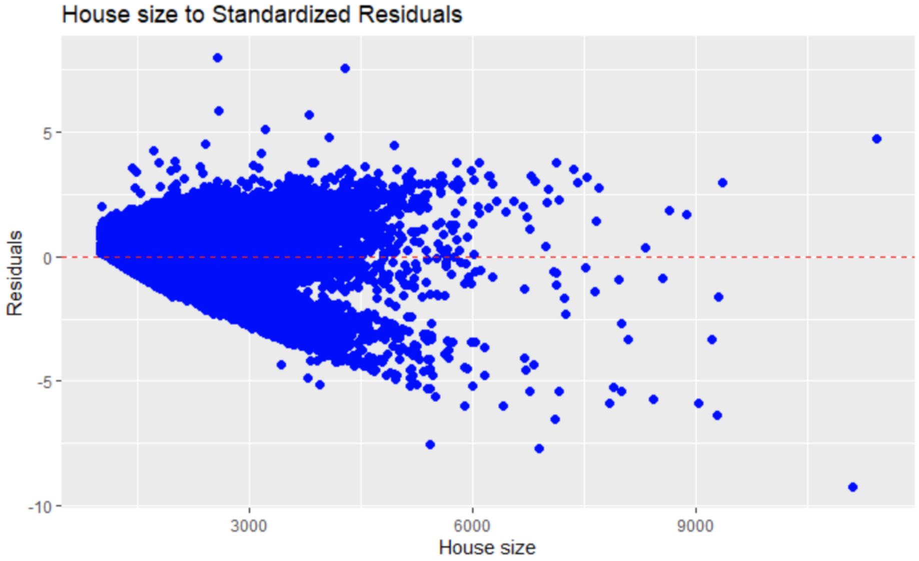
The model we have fitted is based on theoretical assumptions. To validate the linear model, we need to check the two conditions normality and homoscedasticity of the residuals.

Homoscedasticity:

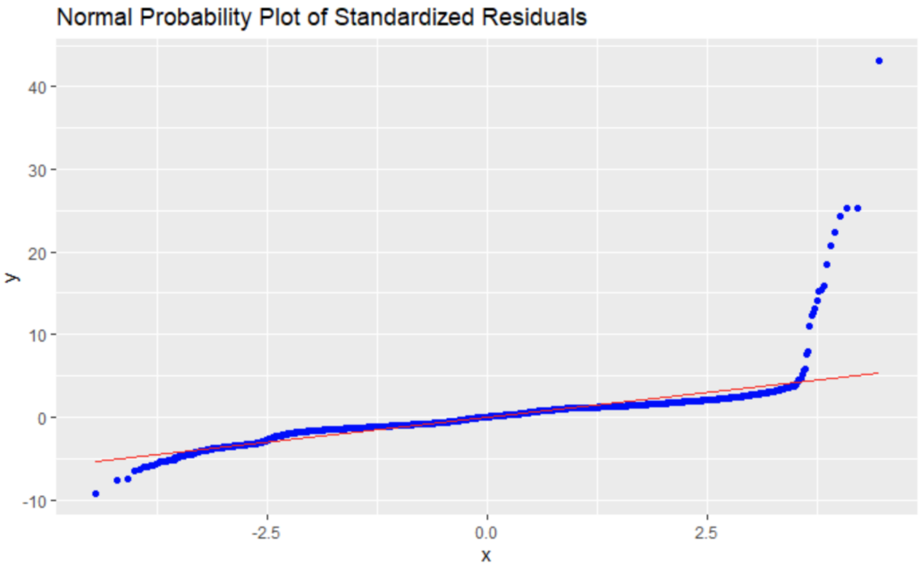


Aside from a few exceptions, when observing the distribution, we can infer that as house size in square footage increases, the variation also expands. This can be rationalized by understanding that as homes get larger, the projected construction costs rise. In some instances, the homes are more luxurious, leading to a significant surge in building expenses.

Now we will remove some of the outliers to reassess the homoscedasticity. ￼



By removing the few points that had larger residuals, the correlation between the residuals and "Actual Total SqFt" becomes more apparent. For enhanced visualization, we graphed a histogram of the residuals as well. Notably, there are pronounced spikes primarily in 2018, 2020, and 2022. It's well understood that housing prices began to increase in 2018 and escalated post-pandemic. Other factors involved may include a rise in migration to Calgary from other provinces, supply chain disruptions, and economic instability. These factors could be explored in future studies.



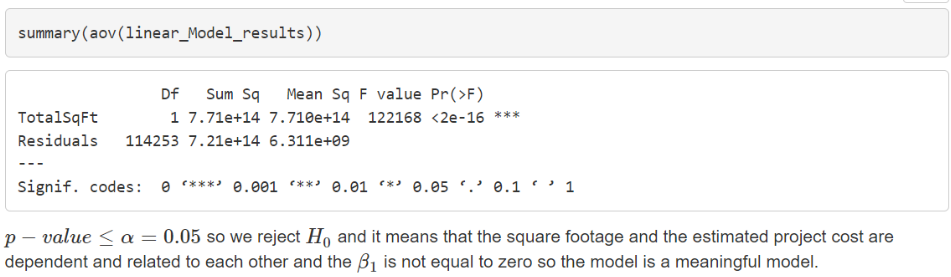
By using a normal probability plot, the model residuals were inspected for a normal distribution, and we can say its distribution is normal, with a skewness to the right. The majority of data points exhibit normal distribution, thus satisfying the condition of normality.

F-Test Validation:

Since our model has met the conditions of normality and homoscedasticity with our justifications above, the statistical hypothesis was analyzed against the model using an F-value which produced the following results:

$$\mathrm{H}_0 : \beta_1 = 0 \qquad\qquad \mathrm{H}_A : \beta_1 \neq 0 (The\ model\ is\ meaningful)$$

If the null hypothesis holds true, then the average predicted project cost (ŷ) is equal to β0 regardless of the value of x̂. This would imply that changes in x̂ have no impact on ŷ, making the model ineffective. On the contrary, the alternative hypothesis suggests that variations in x̂ are correlated to changes in ŷ. We will examine this with an ANOVA table.



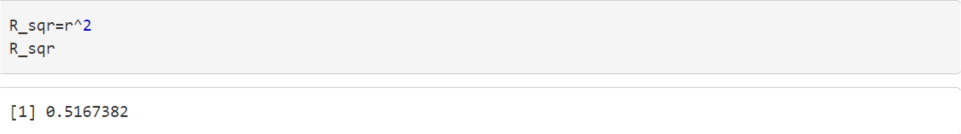
```
summary(aov(linear_Model_results))
```

```
                Df   Sum Sq   Mean Sq F value Pr(>F)    
TotalSqFt        1 7.71e+14 7.710e+14  122168 <2e-16 ***
Residuals   114253 7.21e+14 6.311e+09                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$p - value \leq \alpha = 0.05$ so we reject H_0 and it means that the square footage and the estimated project cost are dependent and related to each other and the β_1 is not equal to zero so the model is a meaningful model.

From the ANOVA table, we observe an extremely low p-value. This leads us to reject H0. It indicates a relationship between square footage and the estimated project cost, suggesting that B1 is not zero, confirming the model is significant.

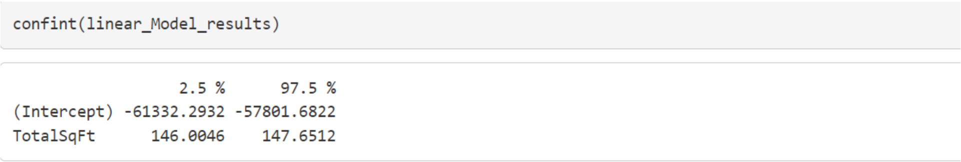
Calculating the Coefficient of Determination.



```
R_sqr=r^2
R_sqr
```

```
[1] 0.5167382
```

Determine Confidence Interval for Beta Using R

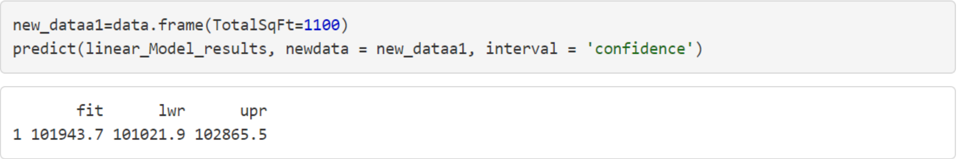


```
confint(linear_Model_results)
```

```
                  2.5 %      97.5 %
(Intercept) -61332.2932 -57801.6822
TotalSqFt      146.0046    147.6512
```

The interval (146.0046, 147.6512) represents the estimated average increase in the project cost for every additional square foot of house size, with 95% confidence. This means that if we increase the house size by one square foot, we can be 95% confident that the project cost will rise by an amount between $146.0046 and $147.6512.

Using the Model to make a Prediction



```
new_dataa1=data.frame(TotalSqFt=1100)
predict(linear_Model_results, newdata = new_dataa1, interval = 'confidence')
```

```
       fit      lwr      upr
1 101943.7 101021.9 102865.5
```

As an example, where the square footage is 1100 square feet the 95% confidence interval for mean “estimated project cost” is between

$$101943.7 \leq \mu_{EstProjectCost|TotalSqFt=1100} \leq 102865.5$$

Comparing the average predicted estimated project cost of 1100 SqFt, $101,943.70, with the actual average Estimated cost of projects which is $102,143, we can conclude that the fitted made an accurate estimate.

Conclusion

All four linear models created statistically indicated an association with the response variable of residential building permit estimated project costs. In the case of inflation rates, contractors, and city quadrants these variables produced very weak R2 indicating that they explain a small proportion of the variability in the dependent variable. However, the regressions still reported significant coefficients allowing us to make inferences on the impact of these predictor variables on the response variable. The square footage told another story indicating a strong positive correlation to the estimated project cost. While there did appear to be a potential violation of the condition of Homoscedasticity, we found external factors that justify why we see this pattern, and therefore our model may still be valid.

An area of opportunity would be exploring other variables that may offer a better explanation in terms of what drives the estimated project cost of residential building permits. For example, what about population density in a city? How about the cost of lumber? On and on we could go. Eventually, we could combine some of these variables into a multiple linear regression that might have better predictability but of course, the tradeoff of accurate predictions and overfitting becomes a concern. As well as other issues to look out for like multi-collinearity. Nevertheless, an undertaking of this kind of analysis would be beneficial for understanding what drives residential project costs and subsequently housing prices. If we could map and understand these relationships, we could use this information and these models to make decisions and tackle difficult problems like the City of Calgary housing crisis. We look forward to learning the tools necessary in our continued studies of the program to undertake a project of this rigor.

References:

Affordable Housing Facts. (n.d.). https://www.calgary.ca. https://www.calgary.ca/social-services/low-income/affordable-housing-facts.html

Jarad Niemi. (2020, October 27). *Regression with Categorical Explanatory Variables* [Video]. YouTube. https://www.youtube.com/watch?v=fDFyG_q2xQs

The City of Calgary (2023). Monthly Economic Indicators. https://data.calgary.ca/Business-and-Economic-Activity/Monthly-Economic-Indicators/7cvb-8ame [Retrieved September 18, 2023]

The City of Calgary (2023). Building Permits.https://data.calgary.ca/Business-and-Economic-Activity/Building-Permits/c2es-76ed. [Retrieved September 16, 2023]

Aldrich, J. (2023, February 8). Residential construction costs in Calgary jumped 14 per cent in 2022. Calgary herald.https://calgaryherald.com/business/calgary-residential-construction-costs