Contributions

Dai - R-code, Methods, Results, Ethics, Introduction Editing
Abhigyan - R-code, Introduction, Conclusion
Uma - Poster

Introduction

With rents soaring and vacancy rates drying up, Toronto's rental market has been a tough climb for years, pushing up the financial stress on renters. The knowledge of drivers of rental prices will, therefore, be important to policymakers, landlords, and tenants in making their ways around this competitive landscape. The present study, therefore, tries to investigate how property characteristics—including the number of bedrooms and bathrooms, square footage, housing type, and furnishing status—affect rental prices in Ontario.

Previous studies have also used hedonic pricing models in analyzing the housing market. Among them, one pioneering work is "The Price Effects of Rent Control on Controlled and Uncontrolled Rental Housing in Toronto: A Hedonic Index Approach", which found that one unintended result of rent control could be to raise prices in uncontrolled sectors because of the scarcity of rent-controlled units. Komagome-Towne emphasized that physical property attributes like lot size and square footage were major drivers of housing prices. Basu and Thibodeau (1998) surmised that structural characteristics such as living area, number of rooms, and the age of a house were closely allied to geographic location as drivers of housing prices.

Building on this growing literature, we have adapted an ordinary least squares model to measure various crucial property attributes affecting Ontario rents. A linear regression framework is adopted to relate different predictor variables with a single response variable by highlighting their relationship by description and prediction. Through identifying major determinants of rent, the aim is to draw suggestions for framing housing policy decisions and market practices that could lead to a more equitable rental market.

Methods

The process of determining the final model takes a process of validating the two multiple regression assumptions, the four linear regression assumptions for the model and using goodness measures to choose the best possible subset of the original set of predictors to use. The initial step is to check that the original model satisfies assumptions through checking for easily identifiable linear or nonlinear patterns in the response vs fitted graph and linear patterns in the pairwise scatter plots. If no issues arise, we proceed with the residual analysis, making sure fanning patterns, non linear patterns and clusters do not appear which show violations of constant variance, linearity, and uncorrelated errors respectively. We also check the normal qq plot to verify that there are no notable deviations. If violations are present, we attempt to address them through power transformations which are approximations of the box cox transformation. The above process can be understood visually in a simplified flow chart. (Figure 0)

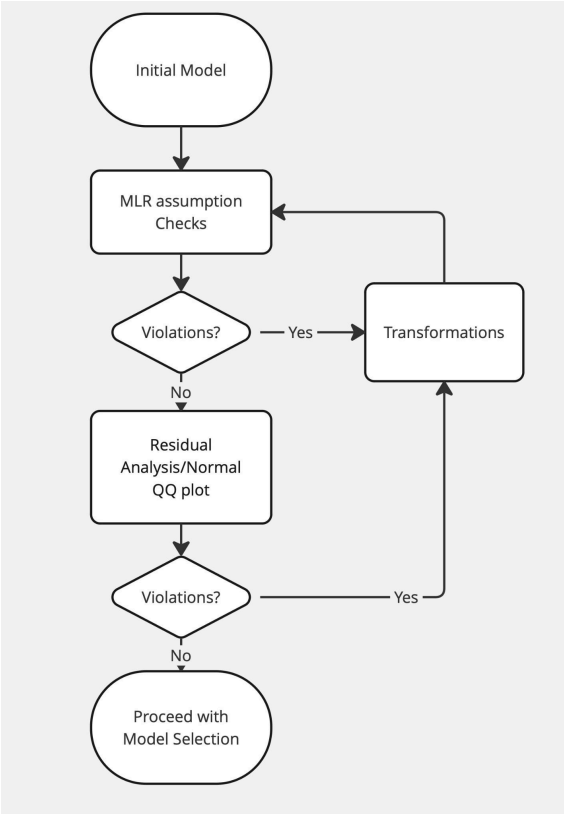

Figure 0

After these violations have been addressed, we proceed with the all possible subsets method to see what predictors create the best model. This process will output the choice of the predictors to use to attain the highest level of adjusted R^2, given a set number of predictors, for each number below 6. This results in checking through 64 possible models, which is a form of an exhaustive search. The output will give us 6 models that give the best adjusted R^2 with each containing a respective number of predictors. After calculating the R^2 adj values for all the best models for every possible number of predictors, we also introduce BIC and AIC values of each value to test the validity from multiple approaches. The model with the largest R^2 adj and smallest AIC, BIC values would be chosen as the best model. After fitting a regression with the chosen model, we verify that the model passes the ANOVA test and that each predictor passes the t test. If the model fails the ANOVA test, we would return to a form of exploratory data analysis, since at least some predictors should be related linearly. If a predictor were to fail to reject the null, we may have a case where they are not linearly related to the response. In such cases we would operate a F partial test to see whether we can take the predictor out of the model. After the above procedure has been complete, we check for multicollinearity by computing VIF values for each predictor. If we find values above 5, we recognize that multicollinearity is present in our model and note that in our limitations of the model. Once the above procedure is complete, the final model has been verified as an optimal regression model.

Results

We began with fitting a model using all 6 predictors, beds, baths, square footage, latitude, type, and furnishing with price as a response. Multiple linear regression assumptions showed no notable violation with the response v fitted graph showing a scattered linear trend and the plot wise scatter plot having all linear patterns as can be seen in Figure 1.

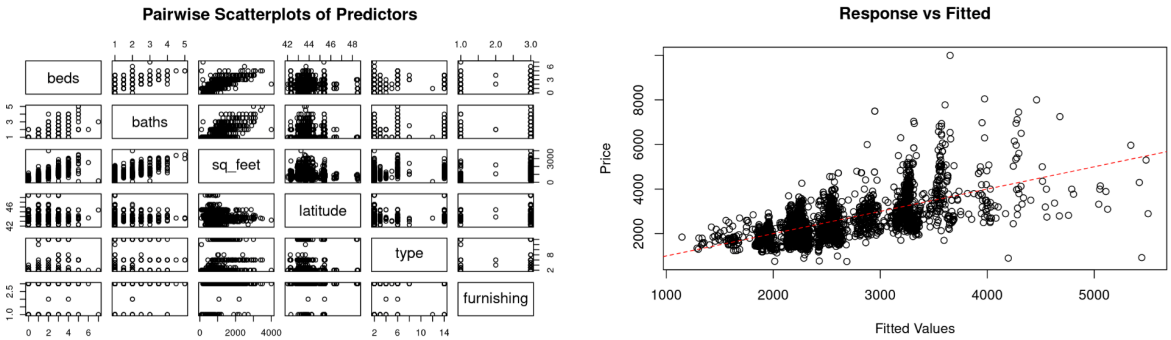

Figure 1: Checking Multiple Linear Regression Assumptions

Our normal QQ plot showed a large deviation and the residuals v square foot showed a left skew. There were also some fanning issues with our plot, hence indicating normality, constant variance and uncorrelated errors violations. We attempted to mitigate these errors through a boxcox and power transformation, though the only predictors that would fall under these transformations were square foot, latitude and price since the rest of the predictors are discrete data or categorical data.
Performing the power transformation, we found the values associated to square foot was -1.1, a term close enough to 0 to use a log transformation. The box-cox function gave us a range including 0 for the response, so we also performed a log transformation on prices as well. Furthermore, the power transform gave us -5.5 on latitude; recognizing that this was a rather absurd value, we attempted the second model assumption verification with and without this transformations. Overall, these transformations resulted in a positive result – we saw a large improvement in normality and constant variance assumptions in our residual analysis though some deviation from the diagonal on the normal QQ plot remained. (Figure 2)

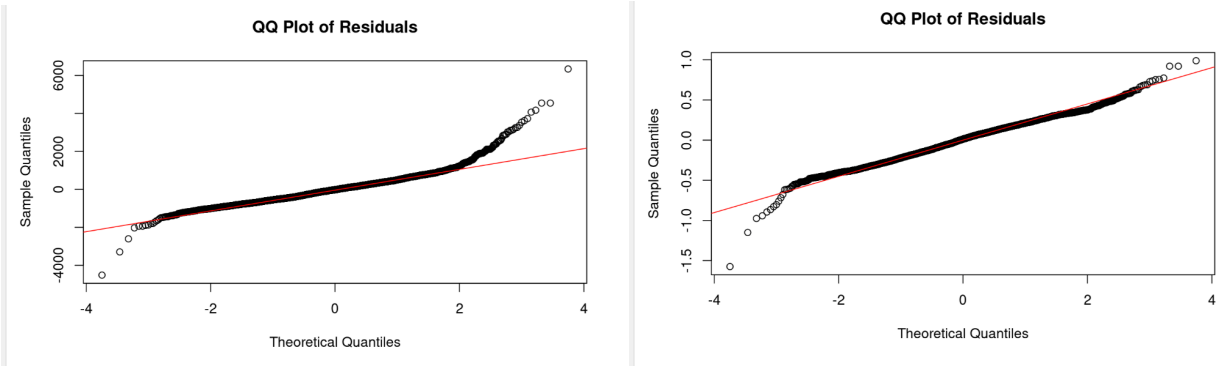

Figure 2: Normal QQ Plot for initial model and transformed model

Though we attempted to improve violations of assumptions, we found that the transformation on the latitude had nearly no noticeable effect on the violations, and thus we disregarded that transformation moving forward.

Given these improvements with linear regression assumptions, we proceeded to the model selection with these transformations. Using the all possible subsets method on R, we found 6 models that each have their own number of predictors which result in the greatest R^2 adj. Given these models, we explicitly calculate their R^2 adj, BIC, and AIC values to see which model performs best amongst the models chosen for each number of predictors. From our derivations, the model with five predictors excluding latitude turned out to have the greatest R^2 and lowest AIC,BIC values. (Figure 3)

AIC/BIC/R^2 Values for Models

Model_Type	R2_adj	AIC	BIC
Model 1	0.3162	−367.1218	−347.2458
Model 2	0.3669	−795.2110	−768.7097
Model 3	0.4099	−1,180.0070	−1,107.1280
Model 4	0.4218	−1,292.8700	−1,213.3660
Model 5	**0.4243**	**−1,314.6100**	**−1,221.8550**
Model 6	0.4242	−1,313.4780	−1,214.0980

Figure 3: AIC, BIC, Adjusted R^2 for models

We choose this as our most preferred model, and continue with verifying regression assumptions. As before, we check MLR assumptions and complete a residual analysis. With our final model, the results are nearly identical to the residual plots that have been curated prior with six models, so we accepted this model to be sufficient, with some limitations of violated assumptions, which can be seen in deviations on the Normal QQ plot and some fanning in residuals. Attempts to mitigate violations would involve further transforming price and square feet, or performing transformations on discrete or categorical predictors, which does not make much sense; hence we suspected there could be some multicollinearity present that would explain the reasons for some persistent violations. Checking VIF for each predictor, we see that all

predictors have VIF greater than 1, and log square feet, beds at 3. (Figure 4) This informs us that there is at least some multicollinearity present, which we denote as a limitation to our interpretation of this model.

Predictors	VIF
lsqft	2.978
baths	1.701
beds	3.079
type	1.338
furnishing	1.083

Figure 4: VIF for each predictor

Producing a summary for this final model, we see that the p value for the ANOVA test is 2.2e^(-16), so we reject null successfully – there are predictors in the model that is linearly related to the response. Furthermore, all predictors have very low p values for the t test, and hence we conclude that every predictor is linearly related to the response. (Figure 5) For the categorical variables type and furnishing, we use the partial f test method to ensure that they are related to the response significantly. They both have very small p-values, thus rejecting null and ensuring that they are indeed significantly related to the response, like the other predictors in the model.

Predictors	p-value
lsqft	2e^(-16)
baths	2e^(-16)
beds	2e^(-16)
Type (Partial F test)	2.2e^(-16)
Furnishing (Partial F test)	2.653^(-16)
ANOVA (Overall)	2.2e^(-16)

Figure 5: Summary of the final model

Conclusion and Limitations

We can conclude from our final model that the log of rental price (log(price)) is positively related to the log of square footage (lsqft), as well as to the number of bathrooms (baths) and bedrooms (beds). For instance, holding other factors constant, adding one bathroom corresponds to an increase of about 24% in the expected rent (since exp(0.217056) ≈ 1.24). Similarly, an increase in lsqft leads to about an 18.5% increase in expected rent (exp(0.170397) ≈ 1.185), and each additional bedroom raises the expected rent by roughly 6.5% (exp(0.062927) ≈ 1.065). These effects indicate that bathroom count and square footage have particularly strong associations with higher rental prices compared to other attributes.

The model's categorical predictors also influence rent. Relative to the baseline property type (likely "Apartment"), categories such as "House" or "Basement" reduce the expected rental price, with "House" having around a 31% lower expected rent than the baseline (exp(-0.368825) ≈ 0.69). Furnishing status similarly matters, as unfurnished units tend to have about 10% lower expected rents than the baseline category (exp(-0.108050) ≈ 0.90). These findings are consistent with previous literature, which emphasizes that both physical attributes and qualitative features of a dwelling shape its market value.

This model, however, is not without limitations. The residual analysis suggests unresolved violations of constant variance assumptions and imperfect normality, evidenced by deviations in the lower portion of the normal QQ plot and lingering patterns in the residual plots. Such issues may be partially explained by the presence of outliers, a highly skewed price distribution, or omitted variables that influence rental prices. The analysis also demonstrated some multicollinearity among the variables. While the model provides valuable insights, addressing these assumption violations—perhaps through more sophisticated transformations, robust regression techniques, or additional covariates—could yield a more accurate and reliable representation of the rental market.

Ethics

In the research, we used the all subsets regression method to identify the models with the highest R^2 adjusted, while also comparing BIC and AIC values. Given that there were only 6 predictors, this approach was both feasible and effective, being able to compare all possible 64 models, a power set of 6 predictors. Though calculating all 64 models using AIC, BIC, R^2 adj would be very tedious and error-prone, the automated element of the all subsets regression method allows us to quickly deduce the best models for each number of predictors. This reduces chance for potential human error, and hence more ethical than calculating all information by hand.

This method provided an advantage over all other methods of automated selection such as stepwise selection, since this approach was more thorough in its comparisons. Every model possible would at least have a lower R^2 adjusted score than the selected final model without ambiguity. Further evidencing the validity of the final model using BIC and AIC values, we can be rather certain that this method is better than having just done a fully automated approach as that would have only compared BIC or AIC scores. There is a distinct ethical issue with using stepwise or other automated selection methods when we know already that the method of all subsets selection would yield a more thorough comparison and computationally feasible. Hence from both a practical and ethical perspective, the all subsets method made a lot of sense.

Bibliography

Fallis, G., & Smith, L. B. (1985). Price Effects of Rent Control on Controlled and Uncontrolled Rental Housing in Toronto: A Hedonic Index Approach. The Canadian Journal of Economics, 18(3), 652–659. https://doi.org/10.2307/135026

Zietz, J., Emily Norman Zietz, & G. Stacy Sirmans. (2007). Determinants of House Prices: A Quantile Regression Approach. The Journal of Real Estate Finance and Economics, 37(4), 317–333. https://doi.org/10.1007/s11146-007-9053-7

Basu, S., & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. Journal of Real Estate Finance and Economics, 17(1), 61-85. https://doi.org/10.1023/A:1007703229507