Estimation and prediction of student apartment price around UW and WLU Area

University of Waterloo

Department of Economics

ECON 421

Instructor: Tao Chen

Group member:

Zhaoqi Jiang 20718480

Simeng Li 20734989

Menglin Shao 20725851

Jiawei Zhang 20725839

Abstract

This study examines the rent prices of apartments and on-campus residence in UW and WLU area. OLS, Two-Stage Least Squares model are used to capture potential factors that could affect the rent price. We find that there is a significantly positive or negative relationship between the number of the bathroom, bedroom, distance to campus, and surprisingly whether utility is included. The findings are robust for most cases in waterloo. Based on this, we finally provide some suggestions for students, housing company and the department of the Waterloo Residence.

Keywords: Rental Price, OLS, Two-Stage Least Squares

1. Introduction

1.1 Motivation

According to the enrollment historical data from University of Waterloo, the undergraduate enrollment has been increasing from 29, 150 to 32,644 in the past 5 years. As the enrollment continues to increase, the demand for living goes up correspondingly. Students need to make a decision between living on-campus or off-campus, which is an important decision because the living place is more than just a place to sleep, it will influence students' chances of academic success, while also introduce students to friends and lots of fun experiences. In order to provide a fully information and suggestions for students to choose suitable living place, this paper collected data of the apartment amenities and price from the website, including UWP, Sage, Accommod8u, Rez-one and KW-4 Rent Company and use OLS, Two-Stage Least Squares to exam potential factors that could affect the rent, including the number of bathroom, bedroom etc. The details will be discussed in the following sessions. In addition, this study also predicts the rent of the new building of UWP to provide advice for the department of Waterloo Residence and other living companies.

1.2 Literature Review

Students' living choice has attracted the attention of academia. Numerous empirical studies have been conducted, seeking to understand the rent prices surrounding universities. For instance, Joe Clark et al. (2012) studies what amenities and characteristics of the apartment are most important and influential to the rent price surrounding the University of Oregon using Hedonic price model. The important points from Joe Clark et al. (2012) such as the choice of the variables and the estimation method help this paper a lot. However, unlike the research did around the University of Oregon, we investigate the rental market around the University of Waterloo and the Wilfrid Laurier University because there are only a few researches in the KW area. In addition, we not only use the OLS method Clark et al. used, but also introduce the instrument variable during estimation. Xinyue Pi (2017) did a very detailed research on the housing market in KW area and find that different factors have effects on different type of people, including students, couples with children, and retired households. She also used the OLS model but she focused on the survey method to obtain the data.

The rest of this paper is structured as follows. Section 2 presents data while Section 3 introduces the econometric model used in this study. Results are discussed in Section 4. Finally, Section 5 concludes.

2. Model Introduction

In order to decompose the housing price, the variables are selected as follows:

$price$ ———— price of the suit per person

$bathroom$ ———— number of bathrooms of the unit per person

$bedroom$ ———— number of bedrooms per person

$bathroom$ ———— number of bathrooms of the unit

$area$ ———— the size of the unit per person in m^2

$distance_U$ ———— the minimum distance to campus of either UW or WLU

$distance_{mart}$ ———— the distance to the closest grocery in meters

$parking$ ———— whether includes a parking space (dummy variable)

gym ———— whether includes a gym (dummy variable)

$utility$ ———— whether includes utility (dummy variable)

$insuit_{laundry}$ ———— whether includes laundry in suit (dummy variable)

web_{value} ———— the value of the website of the company (from https://www.siteprice.org/)

Some notifications about the data: first, they are collected from the five major housing companies: Accommod8u, Sage Living, KW4rent, Icon and Rez-one in the Waterloo region and from the UW Place. Most of the parameters are chosen after Joe Clark (2012) in which they found significant impact of these variables on the housing rent. Second, the room type "One bedroom on den" has been counted as two bedrooms and can accommodate two people. Third, the area includes private room and common area. Fourth, for UWP, $distance_U$ is just the distance to UW. Fifth, $distance_{mart}$ has been counted as the distance to a large grocery like Walmart, Sobeys, or Food basics. A small mart in plaza isn't been counted. Sixth, $parking$ is initially thought to be number of parking spots available in each building. However, since the number of parking spots was usually not found on the website, we made it a dummy variable and found to be not helpful for estimation. Seventh, dummy variable $utility$ is one if all utilities: water, electricity, heat and internet. The last but not the least, the web value was initially developed by us to catch the effect of the company's reputations on its housing price. But it was later in Section 4 found to be correlated with the residual, which we interpreted as that the web value itself is correlated with the company's reputation which we cannot estimate directly and is attributed to residual.
After all information about data are set, the OLS model reads as follows:

$$\begin{aligned} price = \alpha_0 &+ \alpha_1 bathroom + \alpha_2 bedroom + \alpha_3 area + \alpha_4 distance_U + \alpha_5 distance_{mart} \\ &+ \alpha_6 parking + \alpha_7 gym + \alpha_8 utility + \alpha_9 insuit_{laundry} + \alpha_{10} web_{value} + \mathcal{E} \end{aligned}$$

As explained later in Section 3 in the graph, the dummy variable $parking$ is of no help for our estimation, hence, excluding $parking$ in this model yields:

$$\begin{aligned} price = b_0 &+ b_1 bathroom + b_2 bedroom + b_3 area + b_4 distance_U + b_5 distance_{mart} \\ &+ b_6 gym + b_7 utility + b_8 insuit_{laundry} + b_9 web_{value} + e \end{aligned}$$

Now, this model would give us best linear unbiased estimator (BLUE) if it satisfies the four conditions for OLS by Gaussian-Markov Theorem. However, in section 4.1, variable web_{value} is found to be correlated with $\hat{e}$. Hence condition four is violated. And we need to find some instrument variable to model the exogeneity of web_{value}. Some candidates are number of years of foundation of the company ($year$), number of building constructed ($building$) and the daily hit of the company approximated by siteprice.org (hit). They need to be correlated with web_{value} ($cor(iv, web_{value}) \neq 0$) and not correlated with $\hat{e}$ ($cor(iv, \hat{e}\,) = 0$). After the correlation test in Section 4.2.1, we found hit and $building$ is suitable for instrument variables. Hence the first-stage model would be constructed as:

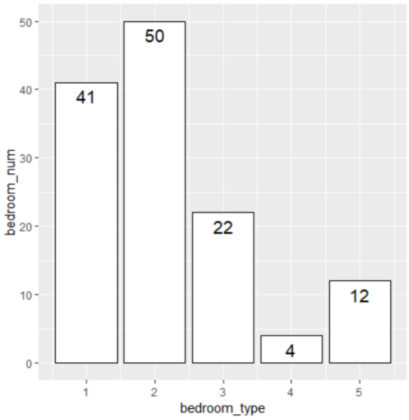
$$web\_value = \pi_0 + \pi_1 hit + \pi_2 building + u_1$$

Note that the $\widehat{web\_value}$ in first-stage is per company. Hence the building under same company would have the same $\widehat{web\_value}$. And the second-stage would then become:

$$\begin{aligned} price = \beta_0 + \beta_1 bathroom + \beta_2 bedroom + \beta_3 area + \beta_4 distance_U + \beta_5 distance_{mart} \\ + \beta_6 gym + \beta_7 utility + \beta_8 insuit_{laundry} + \beta_9 \widehat{web_{value}} + u_2 \end{aligned}$$

3. Data Analysis

We collected 129 different data points, and their distribution along with the number of the bedrooms is shown in Graph 1. Most of the room types we have collected are single or double bedrooms. There are only four buildings provides the suite has four bedrooms.



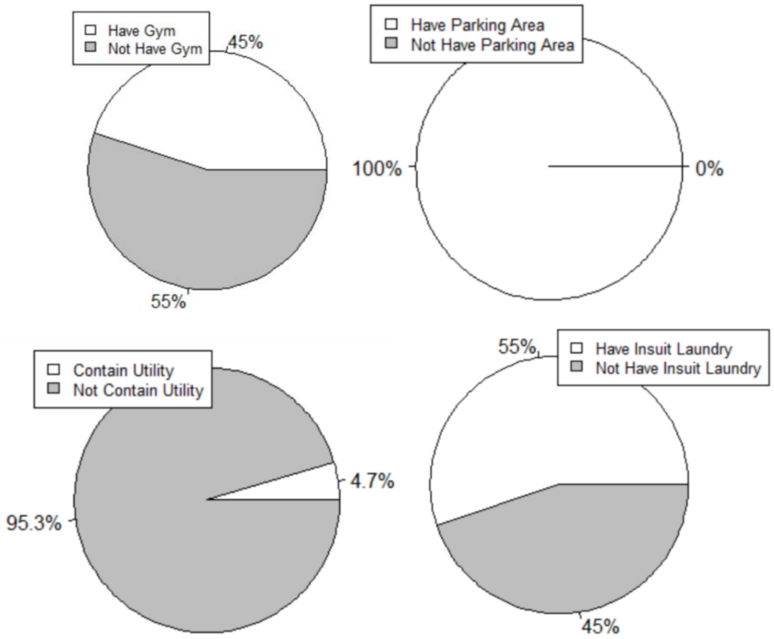
Graph 1: Data size for different number of bedrooms

Then, we calculate the maximum, minimum, average and median of some variables by the data. We calculated the prices of different types of suites by distinguishing the number of bedrooms the suite contains. Table 1 presents the calculation results for the exogenous variables. As the number of bedrooms increases, the rent (price) per room decreases. However, the suite which has four bedrooms is an exception. One possible reason is that we just have 4 data points for the four-bedroom suite and half of them are from Rez-one. The pricing strategy of a particular company may affect the result substantially.

Table 1 Descriptive Statistics for One Bedroom Suite

Variables	Minimum	Maximum	Mean	Median
Price (1 bedroom)	822	1525	1306.27	1280
Price (2 bedrooms)	725	950	830.08	842.5
Price (3 bedrooms)	700	910	763.41	747.5
Price (4 bedrooms)	704	890	807.25	817.5
Price (5 bedrooms)	600	844	660.33	637.5
Bathroom per person	0.33	1	0.76	1
Area per person	15.1	60.1	35.74	34.5
Distance to DC Library	550	2100	1009.69	820
Distance to Laurier Library	42	1300	769.17	850
Distance to Market	1346	2500	1897.22	1914

Graph 2 shows that the apartments with gyms are almost as numerous as apartments without gyms. And we also find that all apartments have parking area. Because of this, the dummy variable, having parking area or not, is not significant in our model. Besides, the rent of most apartment does not contain the fee of utility.



Graph 2： Pie Charts for Dummy Variables

4. Model evaluation

4.1 OLS

After finalizing our regression model based on the variables, we had obtained information on during our data collection, our investigation produced the following regression outputs:

$$price = \alpha_0 + \alpha_1 bathroom + \alpha_2 bedroom + \alpha_3 area + \alpha_4 distance_U + \alpha_5 distance_{mart} + \alpha_6 parking + \alpha_7 gym + \alpha_8 utility + \alpha_9 insuit_{laundry} + \alpha_{10} web_{value} + \mathcal{E}$$

Table 2. OLS for apartment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	758.35867	172.56370	4.395	2.42e-05 ***
bathroom	141.38796	65.08020	2.173	0.031799 *
bedroom	-166.88160	21.74118	-7.676	5.05e-12 ***
area	9.60341	1.71864	5.588	1.48e-07 ***
distance_U	0.18450	0.08197	2.251	0.026229 *
distance_mart	-0.13776	0.09904	-1.391	0.166832 .
parking	NA	NA	NA	NA
gym	7.71108	53.72966	0.144	0.886125
utility	430.58141	88.55833	4.862	3.59e-06 ***
insuit_laundry	155.58156	42.04864	3.700	0.000328 ***
web_value	0.08778	0.01762	4.981	2.17e-06 ***

Residual standard error: 126.6 on 119 degrees of freedom

Multiple R-squared: 0.7903, Adjusted R-squared: 0.7745

F-statistic: 49.84 on 9 and 119 DF, p-value: < 2.2e-16

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the regression outputs in Table 2, since all apartments has parking place, the coefficient for variable parking is not defined because of singularities. Thus, a new regression based on AIC selection algorithm is used to drop the undefined variables. The formula below illustrated the new model and Table 3 shows the new regression output. In the new model, the variable parking is dropped because of its singularities.

$$price = \beta_0 + \beta_1 bathroom + \beta_2 bedroom + \beta_3 area + \beta_4 distance_U + \beta_5 distance_{mart} + \beta_6 gym + \beta_7 utility + \beta_8 insuit_{laundry} + \beta_9 web_{value} + e$$

Table 3. AIC selection OLS for apartment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	758.35867	172.56370	4.395	2.42e-05 ***
bathroom	141.38796	65.08020	2.173	0.031799 *
bedroom	-166.88160	21.74118	-7.676	5.05e-12 ***
area	9.60341	1.71864	5.588	1.48e-07 ***
distance_U	0.18450	0.08197	2.251	0.026229 *
distance_mart	-0.13776	0.09904	-1.391	0.166832 .
gym	7.71108	53.72966	0.144	0.886125
utility	430.58141	88.55833	4.862	3.59e-06 ***
insuit_laundry	155.58156	42.04864	3.700	0.000328 ***
web_value	0.08778	0.01762	4.981	2.17e-06 ***

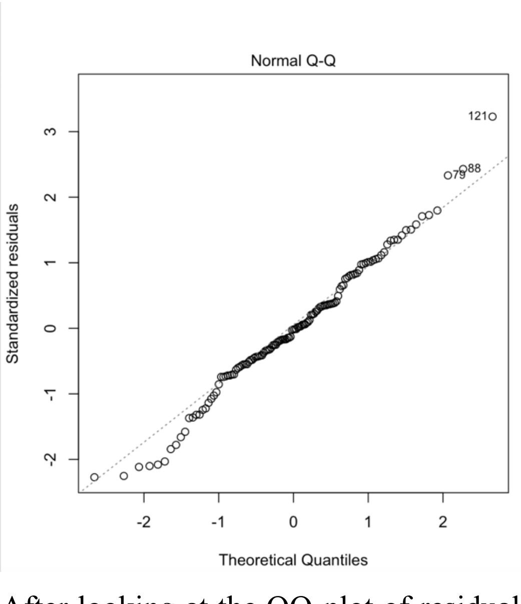
Residual standard error: 126.6 on 119 degrees of freedom

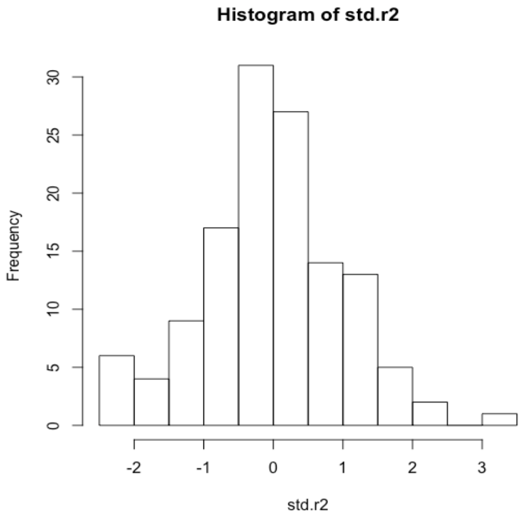
Multiple R-squared: 0.7903, Adjusted R-squared: 0.7745

F-statistic: 49.84 on 9 and 119 DF, p-value: < 2.2e-16

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Despite the variable gym is insignificant, we decided to leave it in the model since the p-value for this new regression model is smaller than 0.001 which means the model works great.





After looking at the QQ-plot of residuals in figure 1 above, we notice a “S-shape” which means the distribution of residual in this model has a heavy tail than Normal distribution, and thus the simple linear regression (OLS) is not a good fit. Moreover, the result in the histogram of residual in figure 2 above also shows there are some outliers on the left. Thus, we need to check whether the dependent variable’s error terms are correlated with the independent variables. The correlation test of all independent variables with the residual are shown in table 4 below

Table 4. Correlation test

	Correlation	Pr(>|t|)
bathroom	-0.004156130	0.9627162
bedroom	-0.005555528	0.9501766
area	0.002673792	0.9760089
distance_U	-0.005681310	0.9490501
distance_mart	0.004220044	0.9621433
gym	-0.005852824	0.9475141
utility	-0.013415506	0.8800589
insuit_laundry	-0.005527504	0.9504276
web_value	0.212304429	0.0157159*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since the null hypothesis for correlation test is “true correlation is equal to 0”. From all of the p-value in table 4, only the assumption for “web_value” is rejected. And other variables have no evidence reject the null hypothesis, which indicate that they have little correlation with the dependent variable’s error term. Thus, “web_value” is endogenous in this mode. The OLS estimation gives inconsistent estimates because “web_value” is correlated with the dependent variable’s error terms. In this case, according to what we have learnt in class, Two-Stage Least Squares (2SLS) is needed.

4.2 Two-Stage Least Squares (2SLS)

As we have discussed in OLS model in section 4.1, we need to construct a 2SLS model since the variable “web_value” is endogenous.

4.2.1 First-Stage regression

In the first-stage, we first need to find the instrument variables (IV) such that they are relevant (correlated with “web_value”) and exogenous (i.e. cov (IV, residuals) = 0). From the information we collected, we consider three potential instruments: “hit”, “year”, and “building”. By testing their correlation with error terms in the OLS above, we choose “hit” and “building” to be the instrument variables since they are exogenous to the error terms. However, for the variable “year”, it shows a significant p-value that reject the null hypothesis in correlation test which indicates that it relevant to the residuals in OLS. The result for correlation test is shown in table 5 below.

Table 5. Correlation test

	Correlation	Pr(>|t|)
year	0.2275877	0.00949018**
building	-0.1732541	0.04958688.
hit	-0.2038852	0.02047374.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Hence, we construct the first-stage regression model:

$$web\_value = \pi_0 + \pi_1 hit + \pi_2 building + u_1$$

Although we cannot compute ture "web_value" because we do not know the parameters π_j, we can consistently estimate them by OLS. The first-stage regression output is shown in table 6 below. The p-value implies that this model is optimal. Thus, from this first-stage regression to the endogenous variable "web_value", we compute the estimated "$\widehat{\text{web\_value}}$" and use it in the second-stage to regress the dependent variable on all exogenous regressors and the prediction $\widehat{\text{web\_value}}$.

Table 6. First-stage regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1903.191	665.567	2.860	0.00495 **
hit	5.407	1.836	2.946	0.00383 **
building	-183.874	101.142	-1.818	0.07139 .
Residual standard error: 1206 on 129 degrees of freedom				
Multiple R-squared: 0.8936, Adjusted R-squared: 0.7524				
F-statistic: 6.329 on 2 and 129 DF, p-value: 0.002388				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2.2 Second-Stage regression

The second-stage regression model is shown below,

$$price = \beta_0 + \beta_1 bathroom + \beta_2 bedroom + \beta_3 area + \beta_4 distance_U + \beta_5 distance_{mart} + \beta_6 gym + \beta_7 utility + \beta_8 insuit_{laundry} + \beta_9 \widehat{web_{value}} + u_2$$

The Multiple R-squared and Adjusted R-squared are all greater than the OLS regression which indicates that in the new model, the dependent variable "price" has stronger linear relationship with all independent variables. Moreover, the residual standard error in 2SLS is 123.6 which is smaller than in OLS, and residual sum of squared (RSS) is also smaller in 2SLS. In addition, the result in Hausman-test for 2SLS shows that all exogenous variables are uncorrelated with all disturbance terms. Hence, the 2SLS has a better fitted model then OLS. Table 7 shows the regression output for 2SLS.

Table 7. AIC selection OLS for apartment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1717.74101	235.97686	7.279	3.93e-11 ***
bathroom	126.90668	63.73583	1.991	0.048757 *
bedroom	-139.26850	17.99276	-7.740	3.61e-12***
area	10.56895	1.66115	6.362	3.85e-09 ***
distance_U	0.16334	0.08051	-2.213	0.028824 *
distance_mart	-0.21309	0.09630	-1.391	0.166832 .
gym	-201.15335	54.99578	-3.658	0.000381 ***
utility	611.30084	99.87247	6.1217	1.23e-08 ***
insuit_laundry	-16.66986	46.93209	-0.355	0.723075
$\widehat{web\_value}$	-0.39602	0.07026	-5.637	1.18e-07 ***
Residual standard error: 123.6 on 119 degrees of freedom				
Multiple R-squared: 0.8, Adjusted R-squared: 0.7849				
F-statistic: 52.9 on 9 and 119 DF, p-value: < 2.2e-16				

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Our final regression appears as the following:

$$\begin{aligned}\widehat{price} = 1717.74101 + 126.90668 * bathroom - 139.26850 * bedroom + 10.56895 * area + 0.16334 \\ * distance_U - 0.21309 * distance_{mart} - 201.15335 * gym + 611.30084 * utility \\ - 16.66986 * insuit_{laundry} - 0.39602 * \widehat{web\_value}\end{aligned}$$

Note now we could happily say that, based on our estimation, the marginal effect of increasing one bathroom per person or one bedroom is of around \$127 and \$ -139 money value per month. This would concrete the impact of each variable in money value and would give we some insight and expectation when we choose different room types from different companies. And surprisingly, statistical significantly, including all utilities in rent is worth \$611 per month of the rent, which would potentially be much higher than just paying the monthly fees for all utilities. This is abnormal and after we turn to our data, we find that most of the data points that have the utilities included is under company Rez-one. This also coincide our belief in section 3 which says the pricing strategy of a particular company may affect the result substantially. Note that $distance_U$ and $distance_{mart}$ are in meters so in kilometers, one kilometer farer away from one of the universities and groceries would give a \$163 higher rent and \$213 lower rent respectively. The coefficient of the $distance_U$ is negative might partially because our data are all close to universities.

5. Comparison for UWP monthly payment

Basing on the 2SLS regression model we get above, we try to predict the appropriate monthly payment for UWP. Our model estimation tells us that a double-bedroom room of UWP should have a price of $895.80, a three-bed room should have a price of $734.96 and a four-bed room should have a price a $612.82. The current price for these type of rooms are $810.25 per month, $766.9 per month and $707.58 respectively. Hence base on our estimation, two-bedroom in UWP is relatively cheap than housing companies. Three-bed room and four-bed room are higher than outside campus. But all of the rooms in UWP have much smaller area than housing companies, because they have other great advantages such as closer to campus and other unmeasurable advantages like access to a network of your peers and mentors who can help you gain valuable skills.

6. Conclusion & Further Consideration

In this paper, we concreted impact of different variables on the housing price in money value and also we compared the suits outside campus with inside campus (UWP). This would give us guide expectations and when newcomers who doesn’t know about the pricing want to choose between living on campus and outside campus or choose among different room types and companies. This would also give insights for university and housing agent on what amenities to construct and how to pricing their newly built buildings. For example, including utilities is a better policy for pricing. And they would considerate more on distance to grocery rather than distance to campus and so on. We hope that by using our models and other information, the city planners, housing agents along with the other interested community groups, will be able to continue the with the planning in the UW and WLU area with the most efficient processes, and end up with a design that is the most beneficial to the students, neighborhood and the city.

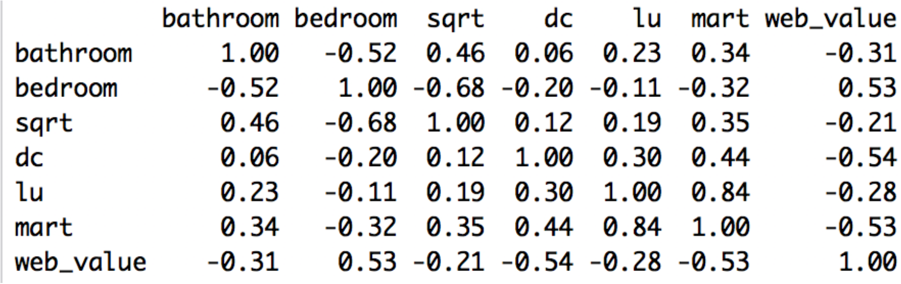
However, our data only analyzed apartment rooms and doesn’t give insight for room in houses. So further studies can integrate house price to our model and give suggestions on living in house or apartment. Also, for each room type, we only count one data point. Further studies would also investigate the amount of rooms of each type and put weights on each room type to introduce discrepancies among different room types. Also, we didn’t consider selling strategies of different companies, such as the early-bird contract of Sage Living, and discount price of KW4rent. Integrating these policies into estimation could better explained the estimated higher price for the two-bedroom of UWP.

7. Reference

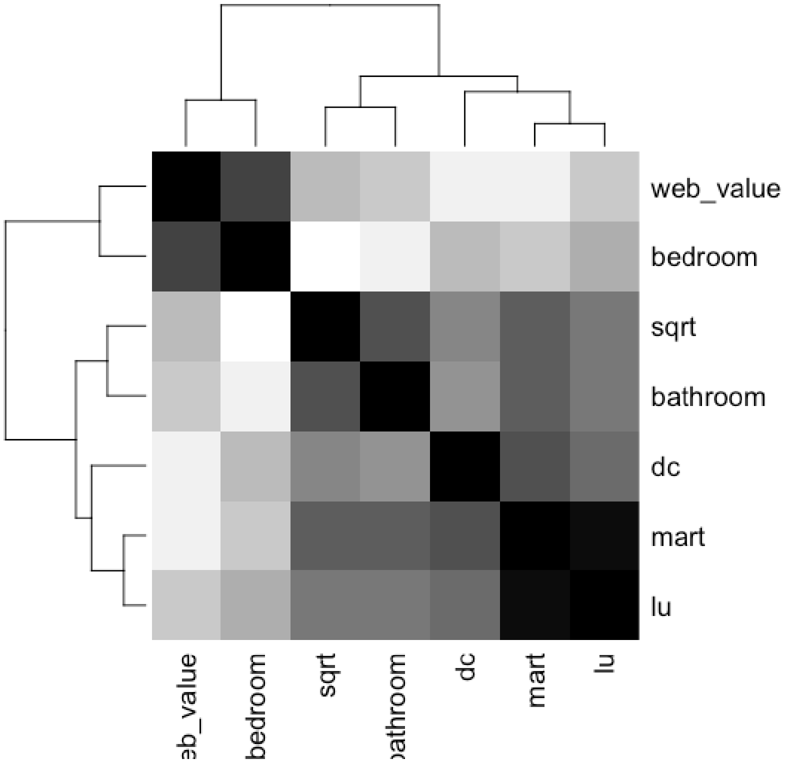
Friedman, J., Clark, J., & Stone, J. (2012). Student Apartment Price Models for the Glenwood Riverfront Development.

Pi, X., Parker, D., & University of Waterloo. School of Planning. (2017). Exploring rental housing market in Kitchener-Waterloo, Ontario. Waterloo, Ontario, Canada: University of Waterloo.

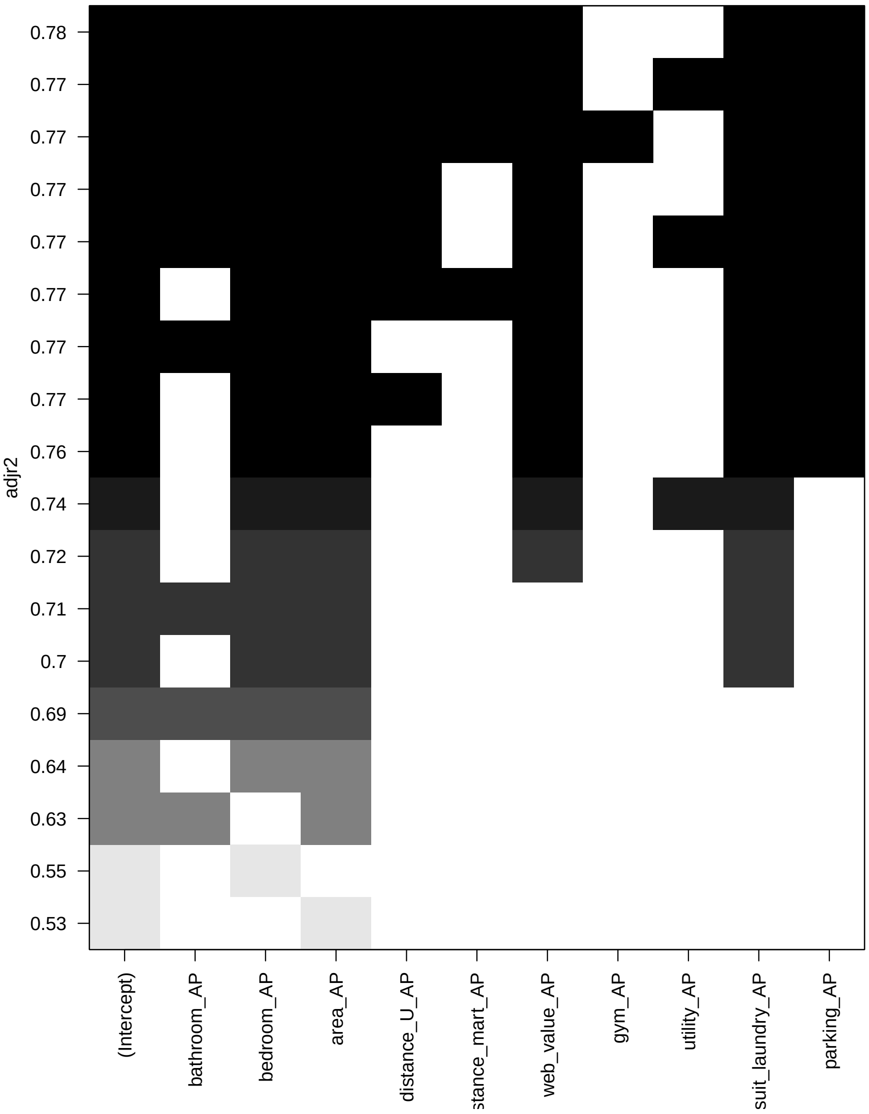
Appendix 1 Correlation between independent variables



```
          bathroom bedroom  sqrt    dc    lu  mart web_value
bathroom      1.00   -0.52  0.46  0.06  0.23  0.34     -0.31
bedroom      -0.52    1.00 -0.68 -0.20 -0.11 -0.32      0.53
sqrt          0.46   -0.68  1.00  0.12  0.19  0.35     -0.21
dc            0.06   -0.20  0.12  1.00  0.30  0.44     -0.54
lu            0.23   -0.11  0.19  0.30  1.00  0.84     -0.28
mart          0.34   -0.32  0.35  0.44  0.84  1.00     -0.53
web_value    -0.31    0.53 -0.21 -0.54 -0.28 -0.53      1.00
```



Appendix 2: Subset Selection between independent variables


adjr2
(Intercept)
bathroom_AP
bedroom_AP
area_AP
distance_U_AP
stance_mart_AP
web_value_AP
gym_AP
utility_AP
suit_laundry_AP
parking_AP