SLUCE2 Hedonic Regression Lab Report:

David Pan - 20581567

2019/03/20

Dr. Dawn Parker

PLAN 416: Modelling the City

University of Waterloo

95/100 Very thorough.

Questions:

1)What relationship would you expect between each of the independent variables and the dependent variable (assessed value)? (i.e., would you expect the beta coeffiecient to be positive or negative?) Would you expect to see some sort of spatial dependence or spatial autocorrelation in these data? If so, why?

I expect the assessed value should be in:

- A direct positive relationship with living area square footage: the greater the square footage, the higher the assessed value. The beta coefficient is therefore expected to be positive.
- A direct positive relationship with acres: the greater the acreage of the land, the higher the assessed value. The beta coefficient is therefore expected to be positive.
- A direct positive relationship with bedrooms: the greater the number of bedrooms, the higher the assessed value. The beta coefficient is therefore expected to be positive.
- An inverse negative relationship with YEAR_BUILT: the older a property, the lower the assessed value. The property value tends to depreciate as the property ages and declines in conditions unless the property is in a prime location. Hence, the beta coefficient is expected to be negative.
- A negative relationship with the distance to highway: the closer the distance to highway, the lower the property value. A residential neighbourhood character, especially established neighbourhood, is not fond of traffic and noise. Commercial and industrial uses are more willing to be located close to highway for the movement of goods than residential uses. The beta coefficient is therefore expected to be negative.
- A generally inverse relationship with the distance to retail: the closer the distance to retail, the higher the property value. Retail services and amenities are expected to be in the vicinity of residential neighbourhoods, serving local demands. The beta coefficient is therefore expected to be negative.

I would expect to see some sort of spatial dependence in these data. For instance, the Distance to Highway and Distance to Retail may contribute to the effect of spatial autocorrelation (clustering of residuals) because many retail developments are located close to highway for the ease of goods movement. In this sense, the location of many retail uses is spatially dependent upon the location of highways.

Living area square footage and the number of bedrooms may contribute to spatial autocorrelation. Typically, the greater the living square footage, the higher the number of bedrooms. This multicollinearity would contribute to redundancy because they would explain the same thing. Similarly, living area square footage and the acreage of the land may also contribute to spatial autocorrelation. The development potential of a residential property is circumscribed by the acreage of the land. Therefore, it is more likely that a parcel of higher acreage will yield a higher living area square footage, which means the latter is spatially dependent upon the former.

2.2) Discuss one outlier (note the point number, and show the data values. Can you explain why the value of this property might be relatively high or low, looking at the values of the independent variables?

Point 126 is an outlier in the high end. The value of this property is significantly higher than average because of at least two factors: first, it is very recently built, and secondly, it is extremely far from highways. These two favorable characteristics seldom come together. Houses that are such a distance away from highways are typically located close to pristine landscapes such as a lake or in a forested area where many affluent neighbourhoods are found. The sharp rise in the assessed value of this reasonably new home may, therefore, be the result of the effect of being a member of an upper-middle-class neighbourhood marked by high median income and other desirable natural amenities (Bucholz, 2004).

126	126	2624	270000	0.177600	3	2002	141960.828285	7192.536379

2.3) Using the "explore" menu, create scatter plots between assessed value and living square feet, and between assessed value and distance to highway. **Are the results consistent with what you would expect? Why or why not?**

The result of living area square feet versus assessed value is generally consistent with my expectation that the greater the square footage, the higher the assessed value. In the scatterplot below (see Figure 1), the assessed value and living area square feet are in a positive direct relationship, as evidenced in the positive straight line.

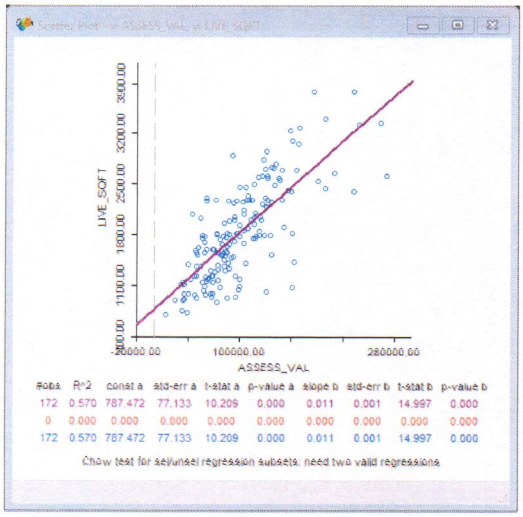

Figure 1: scatterplot of assessed value and living area

The scatterplot of assessed value and the distance to highway did not exhibit any significant correlations (see Figure 2), and contrary to my expectation, they are slightly positively correlated. The variations in assessed value likely depend on the fickle preference of the homebuyer, which can vary from one person to another. Workers, for example, may want their homes to be closer to the transportation network, whereas seniors may desire a quieter living away from the traffic. Taken together, any significant correlation has been counterbalanced, as appeared from the R-square of 0.002 which means that the assessed value is less than 1% explained by the distance to highway.

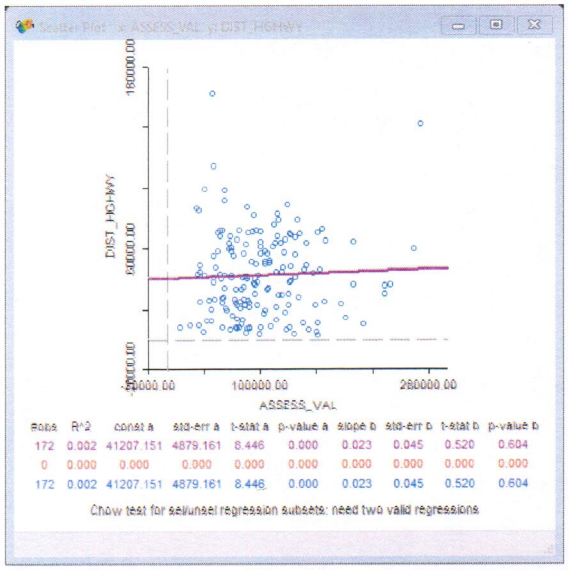

Figure 2: scatterplot of assessed value and distance to highway

3) Optional: Calculate the Moran's I value of assessed value using the spatial weight matrix (under Space >Univariate Moran's I). **Show the Moran's I output. Interpret the results.**

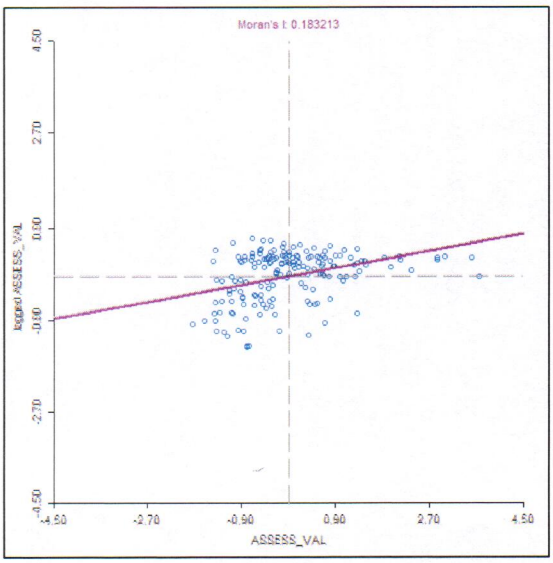

Figure 3: Moran's I output for assessed value

The Moran's I is an indicator of spatial correlation which can be observed through clustering or no clustering. A value near 1.0 indicates strong clustering, 0 no clustering, and -1 dispersion (Esri, 2018). In this case, the Moran's I value is 0.183213, which means there is a light or mild clustering, and hence some positive spatial autocorrelation in the variable "assessed value." This is another example of the neighbourhood effect. The characteristics of a neighbourhood are often factored in the assessment of a property. Residential property is assessed for higher value near higher value, more respectable neighbourhoods of high median income, and vice-versa (Bucholz, 2004). Therefore, it was observed that

homes with a high assessed value tend to be next to other homes with a high assessed value and those with a low assessed value tend to be next to other homes with a low assessed value.

4.1) First, compare the results of Model 1 and Model 2. **How do the models compare in terms of fit? (R2). Which coefficient estimates are significant in each regression? Most important, have the coefficient estimates for the variables that are in both models changed between model runs? Can you explain these differences using the concept of omitted variables?**

Note: I'm explaining using adjusted R-squared because it is a better measure that provides the percent variation explained by only independent variables that actually affect the dependent variable rather than supposing every independent variable in the model will explain the variation in the dependent variable, as in the case of R-squared.

Model 1 has an adjusted R-squared of 0.125497 compared to Model 2's 0.175669 (see Figure 4 and 5). 12.5 percent of the dependent variable ASSESS_VAL has been explained by the independent variables in Model 1 versus 17.6 percent in Model 2. Model 2 is, therefore, better able to explain the variations in the assessed value variable, hence fits the set of observations better. There are fewer discrepancies between the values observed and the values predicted in the model (less residual error), which means that the addition of the variable "Bedrooms" is significant to explaining the assessed value. This is verified by its p-value of 0.00097 which is < 0.05 (statistically significant at the 95 percent confidence level).

AIC is a tool for model selection used to compare and rank multiple competing models using the same dependent variable. If it decreases when more variables are added then the new model is better (lower the score the better). In this case, the AIC index dropped more than 5 points after running Model 2 from 4165.51 to 4156.33, showing that the latter model is the better one by measure of fit.

The coefficient of each independent (explanatory) variable determines the strength and type of relationship. In Model 1, the coefficients for the two variables ACRES and YEAR_BUILT are 1161.63 and 552.03 respectively. Neither of the coefficient is 0 or close to zero, which is a good sign because if it were zero then they are not helping the model. The value of the coefficient shows the magnitude effect of each independent variable. For example, when ACRES increase by one, the dependent variable ASSESS_VAL increases by 1161.6. The direction of that effect is based on the sign of the coefficient which is + positive for both variables in this case.

In Model 1, ACRES is above the significance level of 0.05 with a p-value of 0.11298 and a t-value of 1.5932, which is closer to 0 relative to the other variable, therefore not that significant. In contrast, YEAR_BUILT is statistically significant. It is below the significance level of 0.05 with a p-value of 0, and a t-value of 4.89763 (*>0) which means that there is greater evidence that there is a significant difference since t is significantly above 0. A p-value less than 0.05 indicates that the coefficient cannot be 0, and that the variable is helping the model (statistically significant at the 95 percent confidence level).

However, the coefficient changed in Model 2 with the inclusion of the third independent variable BEDROOMS. ACRES went up from 1161.63 to 1310.21. The positive effect it has on assessed value has been increased and become even more significant: p-value dropped from 0.11 to 0.06 and t-value increased from 1.5932 to 1.84726 (the farther from 0, the more evidence that there is statistical significance). The coefficient of YEAR_BUILT also went down from 552.03 to 490.181, showing a reduced positive effect. The p-value slightly increased from 0 to 0.00002, and t-value decreased from 4.89763 to 4.41723, which means that the variable has become somewhat less significant after the run.

In the article of Filatova, Van Der Veen, & Parker (2009), the potential effects of the omitted variable bias in the land price regression analysis have been examined. It was observed that the experiment with a third independent variable included explains more of the variability in transaction prices than the experiment without that variable. Coefficient estimates for the other independent variables changed when the third variable is added. When this variable is omitted, the regression coefficients (their effects) are close to zero, and the statistical significance is low. When the variable is included, the regression coefficient is significantly negative. Therefore, the observed variation in the dependent variable may be a cause of "uniformly distributed unobserved" characteristics (spatial autocorrelation) rather than unbiased random error. It follows that the inclusion of the new variable can either increase or reduce the coefficient value of the other variables, depending on their correlations.

This is also true in the case at hand. The variation seen in the variable assessed value between model runs is a result of spatial autocorrelation, an issue of the omitted variable bias. Omitting the variable BEDROOMS have masked the effect of the spatial autocorrelation between the three variables, which means that these correlations between the variables had not been captured in Model 1; therefore the coefficients and statistical significance will change accordingly in Model 2 to account for the newly captured relationships.

```
>>03/13/19 12:00:35
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  hedonic_data
Dependent Variable  :  ASSESS_VAL  Number of Observations:  172
Mean dependent var  :     98209.3  Number of Variables   :    3
S.D. dependent var  :     46425.3  Degrees of Freedom    :  169

R-squared           :    0.135726  F-statistic           :     13.2698
Adjusted R-squared  :    0.125497  Prob(F-statistic)     :4.43715e-006
Sum squared residual:3.20399e+011  Log likelihood        :    -2079.76
Sigma-square        :1.89585e+009  Akaike info criterion :     4165.51
S.E. of regression  :     43541.4  Schwarz criterion     :     4174.96
Sigma-square ML     :1.86278e+009
S.E of regression ML:       43160

------------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
      CONSTANT       -990191          222077       -4.45877      0.00001
         ACRES        1161.63          729.119        1.5932      0.11298
    YEAR_BUILT         552.03          112.714       4.89763      0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   139.940865
                                (Extreme Multicollinearity)
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2          85.5085          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       2          0.4024           0.81776
Koenker-Bassett test     2          0.1928           0.90811

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : hedonic_data_outputweights
   (row-standardized weights)
TEST                          MI/DF        VALUE           PROB
Moran's I (error)             0.1183        3.5389         0.00040
Lagrange Multiplier (lag)          1        7.1057         0.00768
Robust LM (lag)                    1        0.0100         0.92036
Lagrange Multiplier (error)        1        9.2102         0.00241
Robust LM (error)                  1        2.1145         0.14591
Lagrange Multiplier (SARMA)        2        9.2201         0.00998
================================ END OF REPORT ================================
```

Figure 4: Model 1 output (Assessed Value= Acres + Yr Built)

```
>>03/13/19 12:06:21
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  hedonic_data
Dependent Variable  :  ASSESS_VAL  Number of Observations:  172
Mean dependent var  :     99209.3  Number of Variables   :    4
S.D. dependent var  :     46425.3  Degrees of Freedom    :  168

R-squared           :    0.190131  F-statistic           :      13.147
Adjusted R-squared  :    0.175669  Prob(F-statistic)     :9.40589e-008
Sum squared residual: 3.0023e+011  Log likelihood        :    -2074.16
Sigma-square        :1.78708e+009  Akaike info criterion :     4156.33
S.E. of regression  :     42273.9  Schwarz criterion     :     4169.92
Sigma-square ML     :1.74552e+009
S.E of regression ML:     41779.4

------------------------------------------------------------------------------
      Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------
      CONSTANT       -921269        216586        -4.25359       0.00003
         ACRES        1310.21        709.275        1.84726       0.06647
      BEDROOMS        16687.2        4967.2         3.35948       0.00097
    YEAR_BUILT        490.101        110.97         4.41723       0.00002
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   169.687789
                                  (Extreme Multicollinearity)
TEST ON NORMALITY OF ERRORS
TEST                 DF          VALUE              PROB
Jarque-Bera           2          112.9584           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                 DF          VALUE              PROB
Breusch-Pagan test    3            0.2267           0.97316
Koenker-Bassett test  3            0.0959           0.99233

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : hedonic_data_outputweights
   (row-standardized weights)
TEST                           MI/DF        VALUE           PROB
Moran's I (error)              0.1431        4.2355         0.00002
Lagrange Multiplier (lag)         1          7.4582         0.00631
Robust LM (lag)                   1          0.5401         0.46238
Lagrange Multiplier (error)       1         13.4736         0.00024
Robust LM (error)                 1          6.5554         0.01046
Lagrange Multiplier (SARMA)       2         14.0138         0.00091
================================ END OF REPORT ================================
```

Figure 5: Model 2 output (Acres + Bdrms + Yr Built)

4.2) Next, compare the results of Model 2 and Model 3. **Again, how do the models compare in terms of goodness of fit, significance of variables, and the values of the estimated coefficients?** Make a scatter plot of bedrooms and living area. **How are the variables related? From your analysis, which variable do you think is the more important determinant of housing value, the number of bedrooms or the size of the house?** (Note: More advanced students may choose to discuss the issue of multicollinearity and its effects, but this is not required.)

Goodness of fit

→ why are you using Adjusted R^2, not R^2?

Model 2 has an adjusted R-squared of 0.175669 compared to Model 3's 0.588049. 17.5 percent of the dependent variables have been explained by the independent variables in Model 2 versus 58.8 percent in Model 3. Model 3 is, therefore, better able to explain the variations in the assessed value variable, hence fits the set of observations better. The addition of the variable "LIVE_SQFT" is clearly significant in explaining the assessed value. This is verified by its p-value of 0 which is < 0.05 (statistically significant at the 95 percent confidence level).

AIC is a tool for model selection used to compare and rank multiple competing models using the same dependent variable. If it decreases when more variables are added then the new model is better than the previous one (lower the score the better). In this case, the AIC index dropped more than 10 points after running Model 3 from 4156.33 to 4042.06, showing that the latter model is the better one by measure of fit.

Significance of variables

In Model 2, BEDROOMS and YEAR_BUILT are statistically significant at the 95 percent confidence level and are clearly helping the model. BEDROOMS has a p-value of 0.00097 (<0.05) and a t-value of 3.35948 (*>0). YEAR_BUILT has a p-value of 0.00002 (<0.05) and a t-value of 4.41723 (*>0). On the other hand, the variable ACRES is above the 0.05 significance level at 0.06647 and has a t-value closer to 0 than the other variables, which means it is not helping the model significantly and can be removed if necessary.

In Model 3, LIVE_SQFT and YEAR_BUILT are statistically significant at the 95 percent confidence level and are clearly helping the model. LIVE_SQFT has a p-value of 0 (<0.05) and a t-value of 12.7008(*>0). YEAR_BUILT has a p-value of 0.02338 (<0.05) and a t-value of 2.28822 (*>0). On the other hand, variables ACRES and BEDROOMS are above the 0.05 significance level at 0.14589 and 0.65667 respectively, which means that they are not helping the model significantly (especially the latter) and can be removed if necessary (it is necessary to remove BEDROOMS in this case because the extremely high p-value shows that there is a problem of multicollinearity and redundancy).

YEAR_BUILT and ACRES are less significant than in Model 2 because the model was adjusted for the additional effect of spatial autocorrelation after the inclusion of a fourth independent variable.

Values of the Estimated Coefficient

In Model 2 the coefficients are 1310.21 for ACRES, 16687.2 for BEDROOMS, and 490.181 for YEAR_BUILT. In Model 3 the coefficients of ACRES, BEDROOMS, and YEAR_BUILT changed into 744.14, -1708.68, and 189.545 respectively.

The positive effect ACRES had on assessed value has been decreased. The effect of BEDROOMS on assessed value has been reversed so that when the number of bedrooms increases by 1, the assessed value decreases by 1708.68, which is unreasonable from a land market perspective. The extremely high p-value noted above, and the instance here together show that there is a problem of multicollinearity and redundancy in the model variables. The positive effect YEAR_BUILT has on assessed value has also been decreased with the additional effect of spatial autocorrelation being adjusted for.

The new variable LIVE_SQFT has a coefficient of 51.549, which is above 0 and has a positive effect on assessed value. When living area increases by 1 square footage, the assessed value increases by $51.55.

Based on the analysis, the size of the house (LIVE_SQFT) is the more important determinant of housing value as seen in their p-values and t-values (bedroom variable has an astronomical p-value and a t-value close to 0, which rendered it statistically insignificant). As previously noted, the number of bedrooms or the size of the house explains the same thing. The greater the living square footage, the higher the number of bedrooms. This collinear relationship between two explanatory variables will contribute to redundancy.

The scatterplot of bedrooms and living area shows that they are in a positive direct relationship (see Figure 7). The presence of multicollinearity in a model means that there are more than one variables (in this case, BEDROOMS) that explain the same thing, which contributes to redundancy. Redundancy is revealed when a duplicate variable has an extremely high p-value. Alternatively, each variable can be tested using a variable inflation factor. If a variable is has a VIF greater than 7.5 then it should be removed from the model (Esri, 2018).

```
>>03/13/19 12:20:36
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  hedonic_data
Dependent Variable  :  ASSESS_VAL  Number of Observations:  172
Mean dependent var  :     99209.3  Number of Variables   :    5
S.D. dependent var  :     46425.3  Degrees of Freedom    :  167

R-squared           :    0.588049  F-statistic           :      59.597
Adjusted R-squared  :    0.578182  Prob(F-statistic)     :3.46317e-031
Sum squared residual:1.52716e+011  Log likelihood        :    -2016.03
Sigma-square        :3.14467e+008  Akaike info criterion :     4042.06
S.E. of regression  :     30240.2  Schwarz criterion     :      4057.8
Sigma-square ML     :2.97883e+008
S.E of regression ML:     28797.4

--------------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------------
      CONSTANT       -364614          161013       -2.26451      0.02489
         ACRES        744.14          509.326        1.46103      0.14589
      BEDROOMS      -1708.69          3837.1       -0.445306      0.65667
    YEAR_BUILT        189.548         92.0353        2.28822      0.02338
     LIVE_SQFT        81.849         4.05871        12.7008      0.00000
--------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   199.750055
TEST ON NORMALITY OF ERRORS
TEST                 DF          VALUE            PROB
Jarque-Bera           2          151.2401         0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                 DF          VALUE            PROB
Breusch-Pagan test    4           26.9898         0.00002
Koenker-Bassett test  4            9.8322         0.04335

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : hedonic_data_outputweights
   (row-standardized weights)
TEST                          MI/DF        VALUE           PROB
Moran's I (error)             0.1122        3.4069         0.00066
Lagrange Multiplier (lag)        1          3.1176         0.07745
Robust LM (lag)                  1          0.0024         0.96092
Lagrange Multiplier (error)      1          8.2867         0.00399
Robust LM (error)                1          5.1715         0.02296
Lagrange Multiplier (SARMA)      2          8.2891         0.01585
================================ END OF REPORT ================================
```

Figure 6: Model 3 (Acres + Bdrms + Yr Built + Living Area)

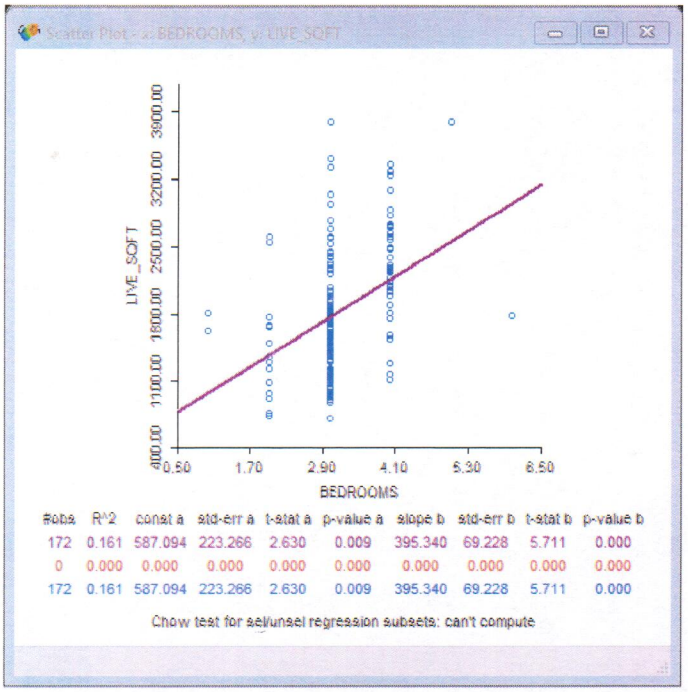

Figure 7: Scatterplot of bedrooms and living area

4.3) Model 4 is essentially the most complete model that we can run with our current set of independent variables. Run this model, first using OLS, then using a spatial lag model. **How do the models compare in terms of goodness of fit, significance of variables, and the values of the estimated coefficients? Do the results of the spatial lag regression differ substantially from the results of the spatial model that does not correct for spatial autocorrelation? What factors do you think might account for these results?**

Goodness of fit

Model 4 (OLS) has an adjusted R-squared of 0.586531, the highest of all previous models (see Figure 8). 58.7% of the dependent variables have been explained by the independent variables in Model 4, versus 57.8% in Model 3, 17.6% in Model 2, and 12.5% in Model 1. Model 4 is, therefore, better able to explain the variations in the assessed value variable, hence fits the set of observations better relative to the other OLS models. The addition of the variable "DIST_RTL" is clearly significant in explaining the assessed value. This is verified by its p-value of 0.03704 which is < 0.05 (statistically significant at the 95 percent confidence level). The addition of "DIST_HGHWY," however, is not, as further explained in the next section. The AIC index is the lowest of all previous OLS models, 4039.59, showing that the last OLS model is the best one by measure of fit.

The spatial lag model (see Figure 9), on the other hand, has a higher R-squared of 0.612105. 61.2% of the dependent variables have been explained by the independent variables in Model 4 (spatial lag), versus 58.7% in Model 4 (OLS). Model 4 (spatial lag) is, therefore, better able to explain the variations in the assessed value variable, hence fits the set of observations better. The AIC index of the spatial lag model is 4036.47, which is even lower than that of simple OLS, showing that the spatial lag model for Model 4 is the best one by measure of fit.

Significance of variables

In Model 4 (OLS), LIVE_SQFT, YEAR_BUILT, and DIST_RTL (distance to retail) are all statistically significant at the 95 percent confidence level and are clearly helping the model. LIVE_SQFT has a p-value of 0 (<0.05) and a t-value of 13.8287 (*>0). YEAR_BUILT has a p-value of 0.04922 (<0.05) and a t-value of 1.98122 (*>0- somewhat above 0, therefore statistically significant). On the other hand, the variables ACRES, DIST_HGWY are not statistically significant. ACRES has a p-value of 0.06495 (>0.05) and a t-value of 1.85796 (closer to 0 than other variables). DIST_HGHWY has a p-value of 0.99646 and a t-value of 0.00390176 (=~0), which means that it is not helping the model significantly and can be removed if necessary. In this case, I think it is necessary to remove not only the BEDROOMS (as previously discussed) but also DIST_HGHWY. The extremely high p-value of DIST_HGHWY shows that there is another problem of multicollinearity and redundancy. Distance to highway and distance to retail may contribute to spatial autocorrelation (clustering of residuals) because many retail developments are located close to highway for the ease of goods movement. In this sense, the location of many retail uses is spatially dependent upon the location of highways. To reduce redundancy in the model, the DIST_HGHWY should be removed.

The spatial lag model, on the other hand, has adjusted the significance of certain variables. ACRES, in particular, has become statistically significant. In Model 4 (spatial lag), ACRES has a p-value of 0.02553 (<0.05) versus 0.06495 in the OLS model, and a z-value of 2.23332 which is higher than the t-value in the OLS model before (z-value is used instead of t because the population standard deviation is known in the spatial lag model; however, both serve the same purpose). This is because the spatial lag model has adjusted for the spatial dependencies of independent variables, particularly the correlation of acres and living area. As noted before, the living area square footage and the acreage of the land may also contribute to spatial autocorrelation. The development potential of a residential property is circumscribed by the acreage of the land. Therefore it is more likely that a parcel of higher acreage will lead to a higher living area square footage, which means the latter is spatially dependent upon the former. When this is adjusted or corrected ACRES has become a significant variable alongside LIVE_SQFT.

Values of the Estimated Coefficient

In Model 4 (OLS), the coefficients are 953.993 for ACRES, 52.3826 for LIVE_SQFT, 164.195 for YEAR_BUILT, 0.000340484 for DIST_HGHWY and -0.662144 for DIST_RTL. In Model 4 (spatial lag), the coefficients of ACRES, LIVE_SQFT, YEAR_BUILT, DIST_HGHWY, and DIST_RTL are changed into 1118.46, 50.8065, 128.788, 0.0409183, and -0.654272 respectively.

In Model 4 (spatial lag), ACRES has a stronger effect on assessed value, while LIVE_SQFT and YEAR_BUILT have a reduced effect as a result of spatial autocorrelation adjustments. No major changes have been observed between the two models since all variables have a positive effect on assessed value except for distance to retail (an increase in the distance to retail will decrease the assessed value by 0.65 to 0.66.).

The other new variable DIST_HGHWY has a coefficient of 0.000340484 in the OLS model, which is nearly 0 and has an insignificant positive effect on assessed value, meaning that it is not helping the model. This is corroborated by the extremely high p-value. However, this coefficient is improved to 0.0409183 in the spatial lag model after the effect of spatial autocorrelation has been adjusted and corrected.

5) What spatial and aspatial variables do you believe would be important determinants of housing value that are not included in the data set provided? **List the variables and explain your reasoning**.

The socio-economic characteristics of a neighborhood would be important determinants of housing value that are not included in the data set provided. In the case of Chicago, residential property in the urban ghetto known as the Black Belt are undervalued regardless of their acreage, living area, and distance to highway/retail (Hirsch, 2009). This is because of social disorganization and local issues such as crime and safety, access to health care, and school reputation that have not been addressed and yet have a significant impact on property value. Variables such as *crime rates* (negatively correlated with assessed value), *access to hospital/clinic* (positively correlated), and the general character of a neighbourhood (*i.e., median income*) should, therefore, be taken into account when determining housing value.

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  hedonic_data
Dependent Variable  :  ASSESS_VAL  Number of Observations:  172
Mean dependent var  :     99209.3  Number of Variables   :    6
S.D. dependent var  :     46425.3  Degrees of Freedom    :  166

R-squared           :    0.598621  F-statistic           :     49.5140
Adjusted R-squared  :    0.586531  Prob(F-statistic)     :3.39388e-031
Sum squared residual:1.48797e+011  Log likelihood        :     -2013.8
Sigma-square        :8.96366e+008  Akaike info criterion :     4039.59
S.E. of regression  :     29939.4  Schwarz criterion     :     4059.49
Sigma-square ML     :8.65098e+008
S.E of regression ML:     29412.5

--------------------------------------------------------------------------------
      Variable      Coefficient      Std.Error    t-Statistic   Probability
--------------------------------------------------------------------------------
      CONSTANT       -316863          161187       -1.96581      0.05099
         ACRES       953.993          513.463        1.85796      0.06495
     LIVE_SQFT       52.3826          3.78796        13.8287      0.00000
    YEAR_BUILT       164.195          82.8755        1.98122      0.04922
    DIST_HGHWY   0.000340484       0.0872644     0.00390176      0.99646
      DIST_RTL     -0.662144        0.314963       -2.10229      0.03704
--------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   212.360938
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE              PROB
Jarque-Bera             2           145.1577           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF           VALUE              PROB
Breusch-Pagan test      5           64.1632           0.00000
Koenker-Bassett test    5           23.3928           0.00028

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : hedonic_data_outputweights
   (row-standardized weights)
TEST                           MI/DF        VALUE           PROB
Moran's I (error)              0.1038        3.3075         0.00094
Lagrange Multiplier (lag)         1          3.0385         0.08131
Robust LM (lag)                   1          0.0027         0.95831
Lagrange Multiplier (error)       1          7.0955         0.00775
Robust LM (error)                 1          4.0597         0.04408
Lagrange Multiplier (SARMA)       2          7.0922         0.02884
================================ END OF REPORT ================================
```

Figure 8: Model 4 OLS output (Acres + Living Area + Yr Built + Dist to Highway + Dist. to Retail)

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : hedonic_data
Spatial Weight      : hedonic_data_outputweights
Dependent Variable  :  ASSESS_VAL  Number of Observations:  172
Mean dependent var  :     99209.3  Number of Variables   :    7
S.D. dependent var  :     46425.3  Degrees of Freedom    :  165
Lag coeff.   (Rho)  :    0.196687

R-squared           :    0.612105  Log likelihood        :   -2011.23
Sq. Correlation     : -            Akaike info criterion :    4036.47
Sigma-square        :8.36035e+008  Schwarz criterion     :    4058.5
S.E of regression   :     28914.3

------------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value     Probability
------------------------------------------------------------------------------
     W_ASSESS_VAL       0.196687      0.0870379        2.25979      0.02383
         CONSTANT        -268578         187264       -1.60075      0.09127
            ACRES        1118.46         500.807        2.23332      0.02553
        LIVE_SQFT        60.8065        3.74319         13.573      0.00000
       YEAR_BUILT        128.788         81.4746        1.58071      0.11394
       DIST_HGHWY      0.0409183      0.0853033       0.47968      0.63146
         DIST_RTL      -0.654272       0.304501       -2.14867      0.03166
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF       VALUE          PROB
Breusch-Pagan test                      5         84.5231       0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : hedonic_data_outputweights
TEST                                    DF       VALUE          PROB
Likelihood Ratio Test                   1          5.1217       0.02363
================================ END OF REPORT =================================
```

Figure 9: Model 4 – spatial lag regression (Acres + Living Area + Yr Built + Dist to Highway + Dist. to Retail)

References:

Bucholz, S. (2004) *A Brief Introduction to Spatial Econometrics*

Filatova, T., Van Der Veen, A., & Parker, D. C. (2009). Land Market Interactions between Heterogeneous Agents in a Heterogeneous Landscape—Tracing the Macro-Scale Effects of Individual Trade-Offs between Environmental Amenities and Disamenities. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, *57*(4), 431-457.

Hirsch, A. R. (2009). *Making the second ghetto: Race and housing in Chicago 1940-1960*. University of Chicago Press.