Stat 306 Project: Predicting LTSA Filing Transactions

Xu Hao Yang(Harry), Yin Dong Xing(David), Bagk Ji Soo(Sally), Jiang Lin(Lily), and Huang Pei Xi(Terrence)

1.1.1 Abstract/summary

The goal of this project is to forecast the total number of transactions in the Land Title Registry and Surveying Authority (LTSA) for the next month. Explanatory variables being considered are Number of residence buildings sold, Number of detached homes sold, Number of apartments sold, Number of townhouses sold, Average price of homes sold, Housing Price Index (HPI), Number of building permits, Season, Prime time, Month, Number of new homes built, 5 year mortgage rate, 1 year mortgage rate, Interest rate and Guaranteed Investment Certificate (GIC). Season and Prime time are categorical variables with 4 and 2 categories respectively.

After checking the correlation between explanatory variables and transformations, insignificant variables were removed. The best regression equation, after variable selection, checking with residual plots and quadratic terms, using the criteria of the adjusted R^2, Cross-validation and out-of-sample comparisons, is (Note: HPI is in 100's, tot.perm is in 1000000's and the rest are in 1000's):

log(num.trans) = 3.3506 + 0.1182*apart + 0.2296*HPI + 0.2202*tot.perm + 0.0756*primeTime + 0.0712*season(Fall) + 0.0915*season(Summer) + 0.0509*season(Winter).

(See Appendix Table 1.4 for full model details.)

With explanatory variables other than Season held constant, expected log(num.trans) is increased by an average of 0.0712, 0.0915 and 0.0509 for Fall, Summer and Winter months respectively. This translates to an increase of a multiple of $e^{0.0712} = 1.0738$, $e^{0.0915} = 1.0958$ and $e^{0.0509} = 1.0522$ for Fall, Summer and Winter months respectively.

With other explanatory variables held fixed, a larger number of apartment sold (unit in 1000's) adds on average 0.1182 to log(num.trans) or a multiple of 1.1254 to num.trans; a higher Housing Price Index(HPI) in 100's adds on average 0.2296 to log(num.trans) or a multiple of 1.2580; a larger number of building permits of 100000 units adds on average 0.2202 to log(num.trans) or a multiple of 1.246326. PrimeTime is a binary variable that is 1 if predicting time period between May to October and 0 otherwise. If prediction occurs during prime time, on average 0.0756 is added to log(num.trans) or a multiple of 1.0785 to num.trans.

1.1.2 Description of data

In this case study, data are collected for a local technology company, LandSure(a LTSA Subsidiary). The company sought out an accurate prediction model for their future transaction quantities. We will discover which external variables will support the best prediction.

LandSure provided data on their past 34 months of transactions. The number of transactions per month is the response variable. The sample size n = 34.

Secondary research was conducted from websites of Bank of Canada, BC Statistics, BC Housing and MLS. From these sources, variables obtained are number of residential sales (including 3 types of residence homes), average price of homes, HPI, number of building permits, 1-year and 5-year mortgage rates, 5-year term interest rate and GIC. Each was collected for each month starting from April 2014 to January 2017.

Table 1.1: Table of variables that might explain number of future transactions

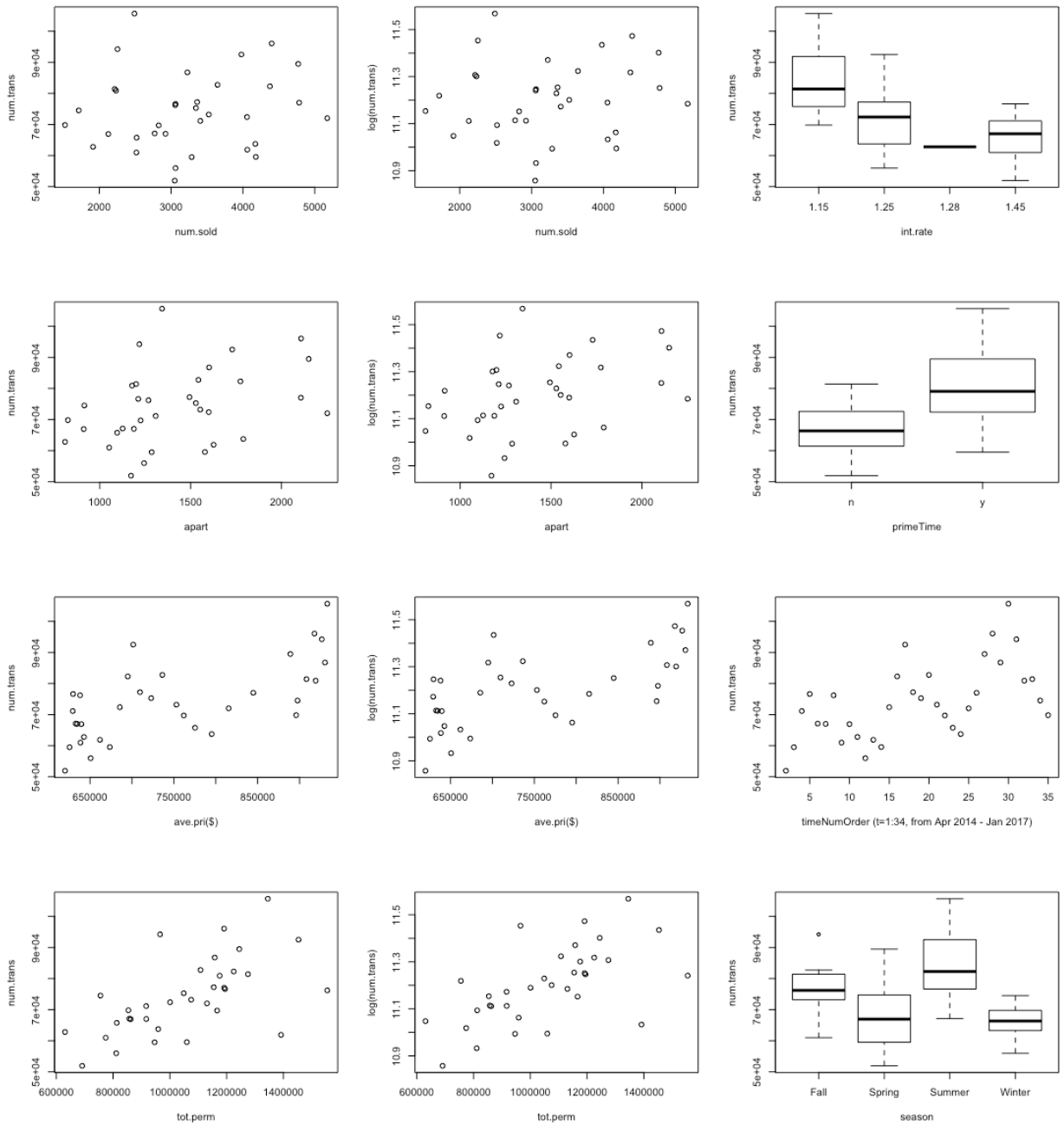
Variables	**Explanation or unit**
num.trans	Total number of transactions per month
num.sold	Total number of residence buildings sold per month
detached	Total number of detached homes sold per month
apart	Total number of apartments sold per month
tHouse	Total number of townhouses sold per month
ave.pri($)	Average price of homes sold per month (CAD)
HPI	Housing Price Index
tot.perm	Total number of building permits
season	4 categories: Spring, Summer, Fall, Winter
primeTime	2 categories: y for months May to October, n for rest
monthNum	12 categories, 1 for January, 2 for February, etc
newHomes	Total number of new homes built that month
5yr.mortg	Average 5 year mortgage rate (CAD)
1yr.mortg	Average 1 year mortgage rate (CAD)
int.rate	Interest rate (CAD)
GIC	Guaranteed Investment Certificate (CAD)

Season (spring, summer, fall, winter), and primeTime (y or n) are categorical explanatory variables.

1.1.3 Data analysis and results

Summary statistics and plots are given in Figure 1.1 and Appendix Figure 1.2. As all the variables are positive and is common to see more scatter in the higher positive values of the explanatory variable, log of the num.trans helped to better fit the model.

Figure 1.1: Plots of num.trans and log(num.trans) versus some of the numerical explanatory variables, also box plots of num.trans by season and primeTime. Since the data was obtained sequentially in time, graph with respect to time (from April 2014 - Jan 2017) is included.



Based on the plots, there is an increasing relation of num.trans to num.sold, apart, tHouse, and tot.perm. There is a very weak positive linear relation of num.trans with detached and new.homes. Furthermore, the relation of num.trans to ave.pri and HPI is a positive near sinusoidal relation. However, as the relation does not perfectly resemble a sinusoidal function, transforming the variable in many ways did not help with increasing the correlation between HPI and num.trans. X5yr.mortg and X1yr.mortg do not show a relationship with num.trans, and int.rate shows a weak negative relationship with num.trans. The box plots of season and primeTime show that there is a clear difference between a group of spring and winter months versus summer and fall months, and using ANOVA test, with multiple comparisons from the Tukey HSD test, the summer season has a significantly different num.trans compared to winter and spring seasons (Appendix Table 1.5). The primeTime is also recognized to be from May to October of each year. Based on the sample correlation matrix and summary statistics (Appendix Figure 1.2 and Figure 1.3), there is a high correlation of apart with num.sold, detached with num.sold, tHouse with num.sold, apart with detached, tHouse with detached, HPI with ave.pri (correlation of nearly 1), int.rate with ave.pri, int.rate with HPI, int.rate with X5yr.mortg.

From the plots and sample correlation table together, we are able to suggest that detached, new.homes, X5yr.mortg, X1yr.mortg and int.rate will most likely not be used in the prediction mode, and the summary statistics indicates that many variables should be linearly transformed to have the a beta range between 0.01 and 10. Num.trans, num.sold, ave.price and apart are scaled to units of 1,000's , HPI to 100's, and tot.permits in 10,000's.

Since this analysis is partly forecasting, and data were collected sequentially in time, a multiple regression model with *lag 1* and *lag 2* with different explanatory variables is used. In the case for *lag 1*, the numerical explanatory variables such as num.sold, detached, apart, and tHouse were manually shifted down 1 month because these factors *take a month to process before effecting the transactions requested*. Residual plots show heteroscedasticity when the response variable is log(num.trans) or num.trans. The adjusted R^2 is 0.8336 and 0.8006 respectively.

To obtain a better model, variable selection with the "exhaustive" method with *lag1* data gave the lowest cp(0.9075195) with 5 variables as shown in Table 1.2 below: apart + HPI + tot.perm + season + primeTime. The multiple regression is re-run with "Spring" as the baseline category for season, as shown on the right side. It gives an adjusted R^2 is 0.8629. "Spring" as the baseline shows that its num.trans is significantly different from all other seasons, whereas Fall transactions are only significantly different from Spring ones. Fall is from September to November, which is considered prime time (September and October).

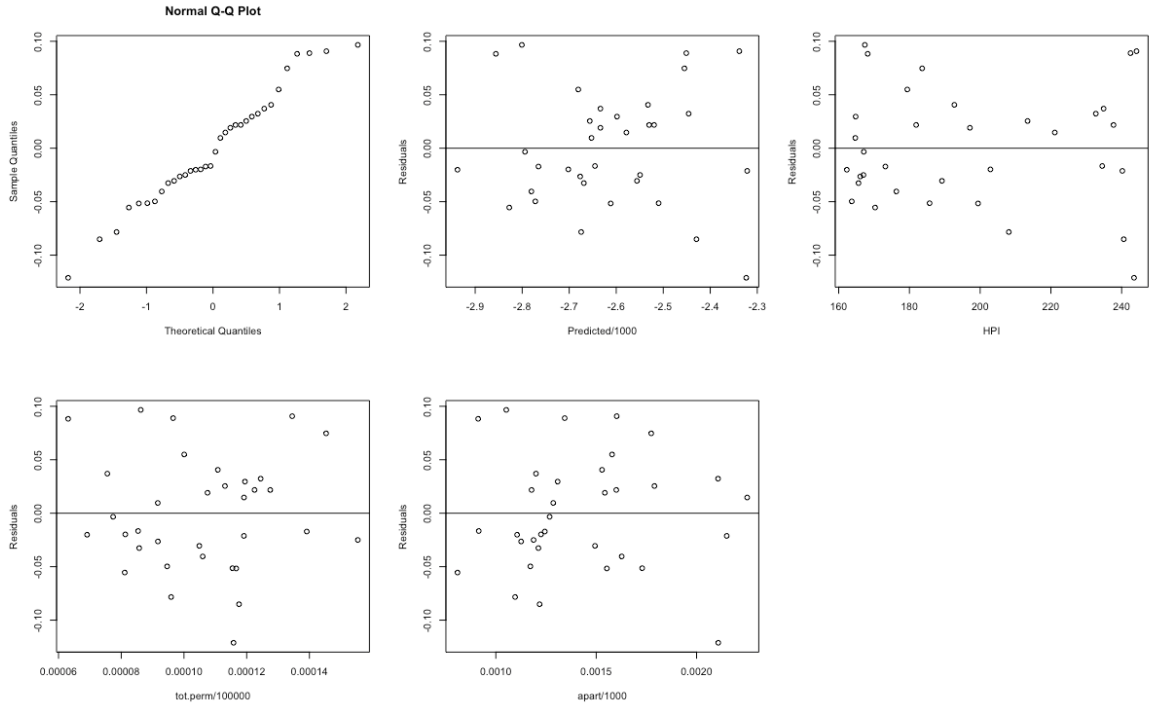
Table 1.2: Multiple regression summaries with explanatory variables from “exhaustive” method with “Fall” and “Spring” as the baseline

```
Call:
lm(formula = log(num.trans) ~ apart + HPI + tot.perm + season +
    primeTime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.121161 -0.032087 -0.009966  0.031623  0.096633

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4152285  0.0867222  39.381  < 2e-16 ***
apart         0.1014140  0.0459529   2.207  0.03635 *
HPI           0.0025477  0.0004382   5.814 3.99e-06 ***
tot.perm      0.0186279  0.0060711   3.068  0.00498 **
seasonSpring -0.0994875  0.0345901  -2.876  0.00793 **
seasonSummer  0.0114692  0.0337737   0.340  0.73689
seasonWinter -0.0019687  0.0385738  -0.051  0.95969
primeTimey    0.1077200  0.0307264   3.506  0.00167 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06037 on 26 degrees of freedom
Multiple R-squared:  0.892,  Adjusted R-squared:  0.8629
F-statistic: 30.66 on 7 and 26 DF,  p-value: 5.264e-11
```

```
Call:
lm(formula = log(num.trans) ~ apart + HPI + tot.perm + season1 +
    primeTime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.121161 -0.032087 -0.009966  0.031623  0.096633

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3157410  0.0838139  39.561  < 2e-16 ***
apart         0.1014140  0.0459529   2.207  0.03635 *
HPI           0.0025477  0.0004382   5.814 3.99e-06 ***
tot.perm      0.0186279  0.0060711   3.068  0.00498 **
season1Fall   0.0994875  0.0345901   2.876  0.00793 **
season1Summer 0.1109567  0.0352701   3.146  0.00412 **
season1Winter 0.0975188  0.0415291   2.348  0.02675 *
primeTimey    0.1077200  0.0307264   3.506  0.00167 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06037 on 26 degrees of freedom
Multiple R-squared:  0.892,  Adjusted R-squared:  0.8629
F-statistic: 30.66 on 7 and 26 DF,  p-value: 5.264e-11
```

The residuals of the fit plotted against each of the explanatory variables show a fairly patternless plot, and there is homoscedasticity. One could also interpret that there is a slightly quadratic pattern for HPI and tot.perm. However, adding any quadratic terms for HPI and tot.perm does not increase the adjusted R^2 (0.8581), and none of the quadratic *betas* are significant either.

Figure 5.2: QQ plot and residual plots for the better model with five explanatory variables



To cross-validate, all four different regressions with *Lag 1* and *Lag 2* (lag explanatory variables 2 month down) are compared with leave-one-out and training/holdout RMSE. For the training/holdout set, we randomly selected half of the data as training set, and make the rest holdout set. The results are shown in Table 1.3 below. Overall, the lag 2 model with 5 explanatory variables has the smallest residual SD. The Durbin-Watson test checks for serially uncorrelated residuals, which they are, as all of the values are close to 2. In addition, a check on influential observations was made with Cook's distance and dfits from ls.diag in R. The plots do not suggest anything abnormally influential.

Table 1.3: Cross-validation and out-of-sample comparisons of three regressions, (1) lag 1 with 5 explanatory variables; (2) lag 1 with 7 explanatory variables; (3) lag 2 with 5 explanatory variables; (4) lag 2 with 7 explanatory variables

Statistics/Model	**1**	**2**	**3**	**4**
Adjusted R^2	0.8629	0.8608	0.867	0.8662
Residual SD	0.06037	0.06083	0.05945	0.05963
rmsepred(leave-one-out)	0.0680032	0.07175141	0.06670109	0.0697715
rmsepred(train/holdout)	0.07863102	0.08199172	0.07539569	0.08976891
Durbin-Watson statistics	2.1101	2.2533	2.0845	2.1797

Note that the models 1 to 4 involved a logged response variable, hence the small magnitude of residual SD and RMSE. The RMSEs observed here are considered reasonably small, as the predicted values (log of transactions) are between 10 to 12. We can see that the third model, "lag 2 with 5 explanatory variables", gives the best RMSE in both the leave-one-out cross validation and the train/holdout prediction.

1.1.4 Brief discussion

Despite a small sample size n = 34, considering the data is from the past three years, a long time period may cause the prediction to be imprecise. We chose only the number of apartments sold to be in the prediction model rather than detached homes and townhomes, because the amount of apartments sold per month is much larger than the other two types.

In conclusion, we have found model 3 (*lag 2* with 5 variables), is the best model with the highest adjusted R^2 and the least Residual SD and RMSE. A possible explanation with *lag 2* being the best model does make intuitive sense in reality: the total transactions from the current month arise from homes that were being sold 2 months ago, this is due to the business timing lag when documents and titles are registered at the LSTA. In addition, the LTSA's prime time (the busiest time) is from May to October, yet the real estate industry is usually busier over the summer holidays. Therefore, by taking back 2 months

back (*lag 2*), it becomes March to August: when weather is nice and warm, when families decide to purchase and move to new places, as well as a popular time when new buildings are developed and constructed.

1.1.5 Team Contribution

It started with an idea Harry had: to connect a local company's business needs and implement it in the Stat 306 team project. Through hard-work and contributions from everyone in the group, we were able to deliver a well thought-out report:

Sally lead many group project meetings and assigned team member tasks and goals. In the project, she generated different regression models, and wrote the beginning sections of the report.

Lily led cross-validating and comparing different models. She also helped with external data research in obtaining all the interest/mortgage/GIC rates.

Terrance prepared raw data for David, organized everyone's R code, and wrote the abstract and some results/discussion.

David processed and summarized LTSA's 34 months of transactions data(millions of transactions) into a single CSV sheet. He also helped with the overall statistical analysis in R.

Harry was the overall leader in liaising with the LTSA to obtain raw data, setting up business team meetings with the firm. He came up with different ways to approach and analyse the project at hand, and worked with every member to ensure the quality and consistency of the report. He assembled and formatted the final report, edited major parts of the analysis, the appendix, including plots and R output.

1.1.6 Appendix

Figure 1.2: Sample Correlation of all numeric explanatory variables

	num.trans	num.sold	detached	apart	tHouse
num.trans	1.00000000	0.40512975	0.213862869	0.59413190	0.354053749
num.sold	0.40512975	1.00000000	0.960168490	0.93890370	0.956330145
detached	0.21386287	0.96016849	1.000000000	0.81076949	0.921446537
apart	0.59413190	0.93890370	0.810769485	1.00000000	0.860858752
tHouse	0.35405375	0.95633015	0.921446537	0.86085875	1.000000000
ave.pri	0.66388179	0.10340953	-0.140136177	0.41391271	-0.009708620
HPI	0.66390488	0.10348688	-0.140058415	0.41398553	-0.009645386
tot.perm	0.60217055	0.50073757	0.435930122	0.52727406	0.464242052
monthNum	0.31834207	-0.03315484	-0.139779242	0.05623808	0.061361523
newHomes	0.42906912	-0.13311505	-0.283950077	0.08132506	-0.197353782
X5yr.mortg	-0.38389084	-0.30705836	-0.167329914	-0.44923600	-0.255721095
X1yr.mortg	0.04039561	-0.22268322	-0.269341273	-0.11064713	-0.302039287
int.rate	-0.53694945	-0.21010140	-0.002459294	-0.45783038	-0.118533455
GIC	-0.27707377	-0.11900719	0.011357272	-0.26721619	-0.084358584
timeNumOrder	0.58420302	0.04760577	-0.183980845	0.34525449	-0.050004003

	ave.pri	HPI	tot.perm	monthNum	newHomes
num.trans	0.66388179	0.663904880	0.60217055	0.31834207	0.42906912
num.sold	0.10340953	0.103486879	0.50073757	-0.03315484	-0.13311505
detached	-0.14013618	-0.140058415	0.43593012	-0.13977924	-0.28395008
apart	0.41391271	0.413985525	0.52727406	0.05623808	0.08132506
tHouse	-0.00970862	-0.009645386	0.46424205	0.06136152	-0.19735378
ave.pri	1.00000000	0.999999969	0.24220039	0.08429110	0.61323011
HPI	0.99999997	1.000000000	0.24224560	0.08426243	0.61323144
tot.perm	0.24220039	0.242245595	1.00000000	0.21320573	0.08715076
monthNum	0.08429110	0.084262426	0.21320573	1.00000000	-0.07213516
newHomes	0.61323011	0.613231439	0.08715076	-0.07213516	1.00000000
X5yr.mortg	-0.61021015	-0.610234867	-0.26402415	0.01931874	-0.31462760
X1yr.mortg	0.30904356	0.308999178	-0.22946491	0.12949815	0.08052333
int.rate	-0.85648868	-0.856510196	-0.27711167	0.14274680	-0.54180350
GIC	-0.53349712	-0.533515775	-0.19099376	0.19350219	-0.26063329
timeNumOrder	0.95723182	0.957224415	0.22578020	0.07611812	0.58614383

	X5yr.mortg	X1yr.mortg	int.rate	GIC	timeNumOrder
num.trans	-0.38389084	0.04039561	-0.536949450	-0.27707377	0.58420302
num.sold	-0.30705836	-0.22268322	-0.210101400	-0.11900719	0.04760577
detached	-0.16732991	-0.26934127	-0.002459294	0.01135727	-0.18398085
apart	-0.44923600	-0.11064713	-0.457830381	-0.26721619	0.34525449
tHouse	-0.25572110	-0.30203929	-0.118533455	-0.08435858	-0.05000400
ave.pri	-0.61021015	0.30904356	-0.856488684	-0.53349712	0.95723182
HPI	-0.61023487	0.30899918	-0.856510196	-0.53351578	0.95722441
tot.perm	-0.26402415	-0.22946491	-0.277111669	-0.19099376	0.22578020
monthNum	0.01931874	0.12949815	0.142746799	0.19350219	0.07611812
newHomes	-0.31462760	0.08052333	-0.541803503	-0.26063329	0.58614383
X5yr.mortg	1.00000000	0.29549428	0.794777020	0.71112715	-0.74567859
X1yr.mortg	0.29549428	1.00000000	0.111332227	0.35268472	0.15289416
int.rate	0.79477702	0.11133223	1.000000000	0.75382220	-0.91652194
GIC	0.71112715	0.35268472	0.753822200	1.00000000	-0.62237910
timeNumOrder	-0.74567859	0.15289416	-0.916521943	-0.62237910	1.00000000

Figure 1.3: Summary statistics of the data set

```
  num.trans         num.sold       detached        apart
Min.   : 51935   Min.   :1714   Min.   : 541   Min.   : 809
1st Qu.: 66054   1st Qu.:2550   1st Qu.:1054   1st Qu.:1180
Median : 72794   Median :3144   Median :1293   Median :1297
Mean   : 74131   Mean   :3268   Mean   :1299   Mean   :1420
3rd Qu.: 81287   3rd Qu.:4036   3rd Qu.:1561   3rd Qu.:1602
Max.   :105671   Max.   :5173   Max.   :2135   Max.   :2252
    tHouse          ave.pri            HPI           tot.perm
Min.   :258.0   Min.   :620100   Min.   :162.3   Min.   : 630705
1st Qu.:436.2   1st Qu.:640375   1st Qu.:167.6   1st Qu.: 875793
Median :543.0   Median :716250   Median :187.4   Median :1067412
Mean   :548.5   Mean   :750935   Mean   :196.5   Mean   :1055789
3rd Qu.:677.2   3rd Qu.:878025   3rd Qu.:229.8   3rd Qu.:1191298
Max.   :786.0   Max.   :933100   Max.   :244.2   Max.   :1554411
   season  primeTime    monthNum         newHomes       X5yr.mortg
Fall  :9   n:16      Min.   : 1.000   Min.   : 617   Min.   :4.640
Spring:8   y:18      1st Qu.: 4.000   1st Qu.:1186   1st Qu.:4.640
Summer:9             Median : 7.000   Median :1630   Median :4.640
Winter:8             Mean   : 6.735   Mean   :1843   Mean   :4.696
                     3rd Qu.: 9.750   3rd Qu.:2371   3rd Qu.:4.790
                     Max.   :12.000   Max.   :3811   Max.   :4.790
  X1yr.mortg       int.rate          GIC
Min.   :2.890   Min.   :1.150   Min.   :0.7300
1st Qu.:2.953   1st Qu.:1.150   1st Qu.:0.7800
Median :3.140   Median :1.250   Median :0.8500
Mean   :3.074   Mean   :1.271   Mean   :0.9138
3rd Qu.:3.140   3rd Qu.:1.407   3rd Qu.:0.8800
Max.   :3.140   Max.   :1.450   Max.   :1.3000
```

Table 1.4: Summary result for *lag 2*, the best model

```
Call:
lm(formula = log(num.trans) ~ apart + HPI + tot.perm + season +
    primeTime)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115349 -0.040872 -0.002642  0.036243  0.090873

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.35064    0.08169  41.015  < 2e-16 ***
apart         0.11824    0.04241   2.788 0.009778 **
HPI           0.22960    0.04555   5.041 3.02e-05 ***
tot.perm      0.22022    0.05757   3.825 0.000736 ***
seasonFall    0.07119    0.03130   2.275 0.031400 *
seasonSummer  0.09154    0.03489   2.624 0.014360 *
seasonWinter  0.05090    0.03488   1.459 0.156507
primeTimey    0.07556    0.03247   2.327 0.027992 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05945 on 26 degrees of freedom
Multiple R-squared:  0.8952,    Adjusted R-squared:  0.867
F-statistic: 31.74 on 7 and 26 DF,  p-value: 3.561e-11
```

Table 1.5: Multiple comparisons of log(num.trans) between different seasons

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log(num.trans) ~ season)

$season
                     diff         lwr         upr     p adj
Spring-Fall   -0.12963543 -0.31134939  0.05207853 0.2334287
Summer-Fall    0.08500361 -0.09128482  0.26129205 0.5632518
Winter-Fall   -0.14643324 -0.32814720  0.03528072 0.1488201
Summer-Spring  0.21463905  0.03292508  0.39635301 0.0157245
Winter-Spring -0.01679781 -0.20377993  0.17018432 0.9947650
Winter-Summer -0.23143685 -0.41315081 -0.04972289 0.0083833
```

1.1.7 Source

BC Statistics:
http://www.bcstats.gov.bc.ca/StatisticsBySubject/Economy/BuildingPermitsHousingStartsandSales.aspx
MLS Sales:
http://www.rebgv.org/home-price-index?region=all&type=all&date=2017-01-01
HPI:
http://www.crea.ca/housing-market-stats/mls-home-price-index/hpi-tool/
Interest Rates:
http://www.bankofcanada.ca/rates/interest-rates/canadian-interest-rates/
New Homes(multi-units):
https://www.bchousing.org/research-centre/housing-data/new-homes-data