MSG500-MVE190
Linear Statistical Models
1 Summary of mini-analyses

Margareta Carlerös & Devosmita Chatterjee

(11 pages)

Multivariate statistical analysis of King county house prices data

1 Introduction

In this project, we have two main objectives with the given data set of King country house prices - firstly, to predict the price of the houses using multivariate linear regression and secondly, to predict the number of bedrooms using Poisson regression.
The project is classified in the following way: §2 comprises of the important aspects of the King country house prices data set. In $ 3, we build a multivariate model to predict the price of the houses using the King country house prices data set and in $ 4, we build a poisson regression model to predict the number of bedrooms for the same data set. Finally, we give our concluding remarks in §5.

2 Data set

The King country house prices data set consists of information of houses for different areas of King County, including Seattle which are sold between May 2014 and May 2015. The data set consists of the variables- id, date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, lat, long, sqft_living15 and sqft_lot15.

3 Multivariate regression model

In this section, we try to predict the price of the houses using only numerical covariates for a randomly selected 500 observations of the king country house prices data set as training data and another randomly selected 500 new observations from the remaining 21,613-500=21113 observations of the same data set as testing data . So, in this case, price is the response.

3.1 Selection of covariates

In the beginning, we have 20 covariates in total. First, we exclude the 4 covariates- waterfront, view, condition and grade as suggested. Again, we exclude another 5 covariates- id, zipcode, date, yr_built and yr_renovated with our reasoning. Next, we include 3 new covariates- age, time_since_reno and dist_hav. Although date and yr_built are excluded, it is used to calculate age of the house (age) which is yr_sold (extracted from date) - yr_built. Also, yr_renovated is excluded but it is used to calculate time since renovation (time_since_reno) which is yr_sold-yr_renovated. Haversine distance (dist_hav) from downtown Seattle is calculated using latitude and longitude coordinates by Haversine distance formula. So, we start

working with 14 covariates. We take different transformations of the response variable and each of the 14 predicted variables so that the five basic assumptions hold which are presented in table 1.

Table 1: This table presents the data transformation of the variables.

Type of variable	**Variable**	**Transformed variable**
Response	price	log10(price)
Covariates	bedrooms	bedrooms
	bathrooms	bathrooms
	sqft_living	log10(sqft_living)
	sqft_lot	log10(sqft_lot)
	floors	floors
	sqft_above	log10(sqft_above)
	sqft_basement	sqrt(sqft_basement)
	lat	lat
	long	long
	sqft_living15	log10(sqft_living15)
	sqft_lot15	log10(sqft_lot15)
	age	sqrt(age)
	time_since_reno	sqrt(time_since_reno)
	dist_hav	log10(dist_hav)

Next we check multicollinearity using the following steps:

Step 1: First, we plot a heat map which calculates distance between the transformed variables defined as 1-correlation $\in$ [0,2]. Bright red colour indicates high correlation between the transformed covariates which are seen in some parts of figure 1.

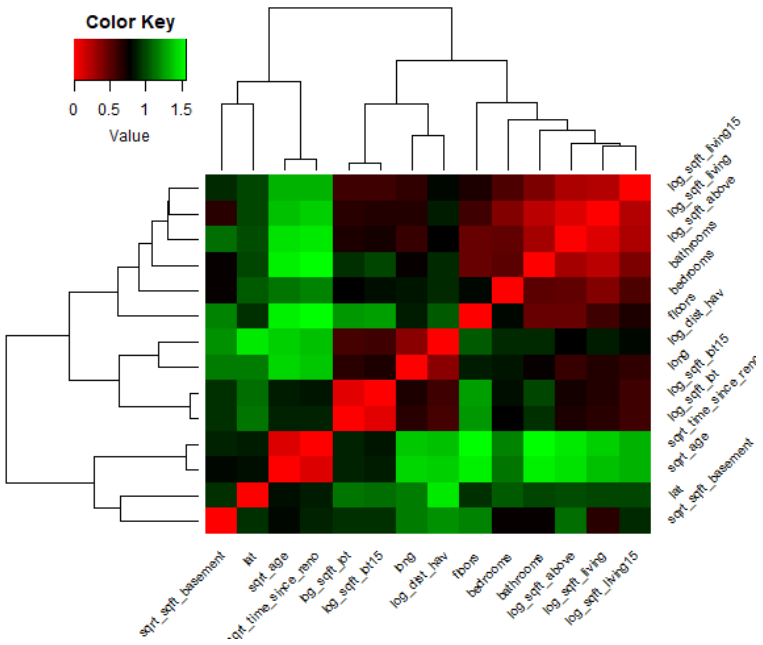

Figure 1: This figure shows the heat map of the transformed covariates.

From the heat map, we can see that the transformed pairs- log10(sqft_living) & log10(sqft_above), log10(sqft_lot15) & log10(sqft_lot) and, sqrt(time_since_reno) & sqrt(age) are highly correlated.
Step 2: Next, we calculate Variation inflaton factor (VIF) for each of the transformed covariates to detect multicollinearity. VIF =1 when the transformed covariates are uncorrelated, 1<VIF<5 when the transformed covariates are moderately correlated and, VIF>5 when the transformed covariates are highly correlated. Variation inflation factor of each of the transformed covariates are presented in table 2.

Table 2: This table presents the variation inflation factor (VIF) of the transformed covariates.

Transformed variable	**VIF**
bedrooms	1.68483154682093
bathrooms	3.68561846173556
log10(sqft_living)	50.9223901598964
log10(sqft_lot)	6.84354243104594
floors	2.25152554940008
log10(sqft_above)	45.4446251640547
sqrt(sqft_basement)	12.1371955110074
lat	1.42531459932563
long	2.01184367609556
log10(sqft_living15)	2.79144643215606
log10(sqft_lot15)	6.77795016973106
sqrt(age)	5.26476429832872
sqrt(time_since_reno)	5.281538283318
log10(dist_hav)	3.17218918018215

Now, out of the two highly correlated transformed covariates log10(sqft_living) & log10(sqft_above), we remove log10(sqft_living) using VIF. Similarly, we exclude log10(sqft_lot) and sqrt(time_since_reno). So, now, we have 11 transformed covariates presented in table 3.

Table 3: This table presents the final 11 transformed covariates.

Final transformed covariates
sqrt(sqft_basement)
sqrt(age)
long
log10(sqft_lot15)
log10(sqft_living15)
log10(sqft_above)
log10(dist_hav)
lat
floors
bedrooms
bathrooms

3.2 Comparison of model selection methods

We build the multivariate regression model using 11 transformed covariates. Adjusted R- squared for 11 covariates model is around 0.7928 which is fairly high. The "training set" and the "test set" account for 1000 observations. We consider 20 such random subsets of data with different proportions of training and testing data- 0.5-0.5, 0.6-0.4, 0.7-0.3, 0.8-0.2 and 0.9-0.1. Model selection methods are backward search and selection based on pMSE. We compare these methods by calculating:

• predicted mean squared error using test data (pMSE),

• Model size and,

• Number of times variables were selected (Count)

which are shown in figure 2, figure 3, figure 4 and figure 5.

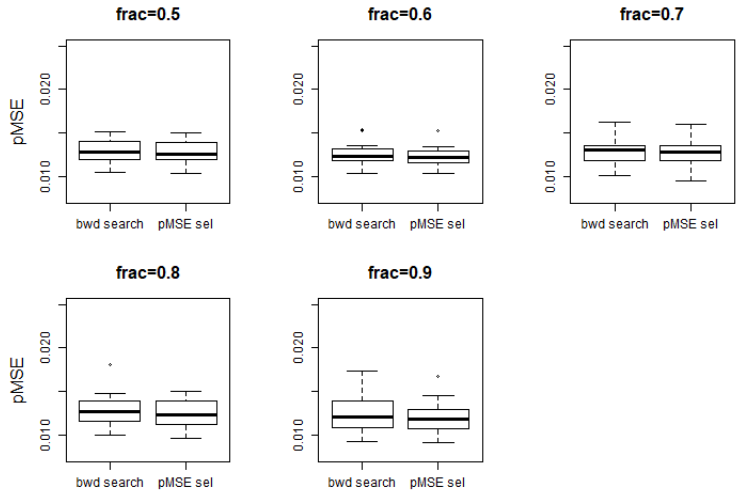

Figure 2: This figure shows pMSE calculation using backward search and selection based on pMSE.

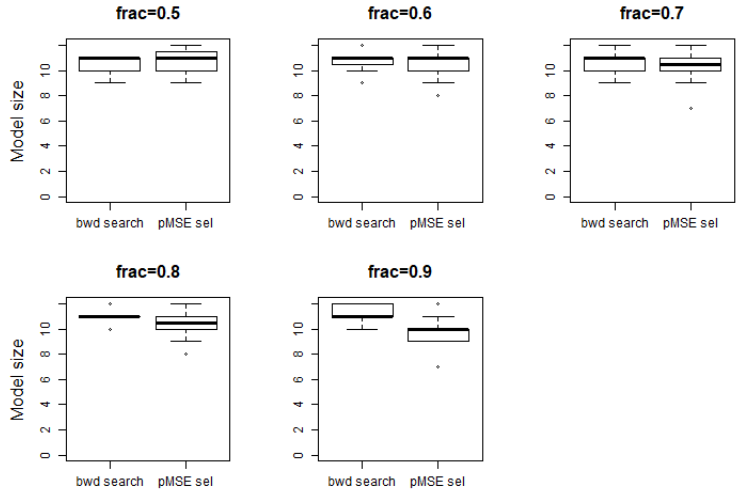

Figure 3: This figure shows model size calculation using backward search and selection based on pMSE.

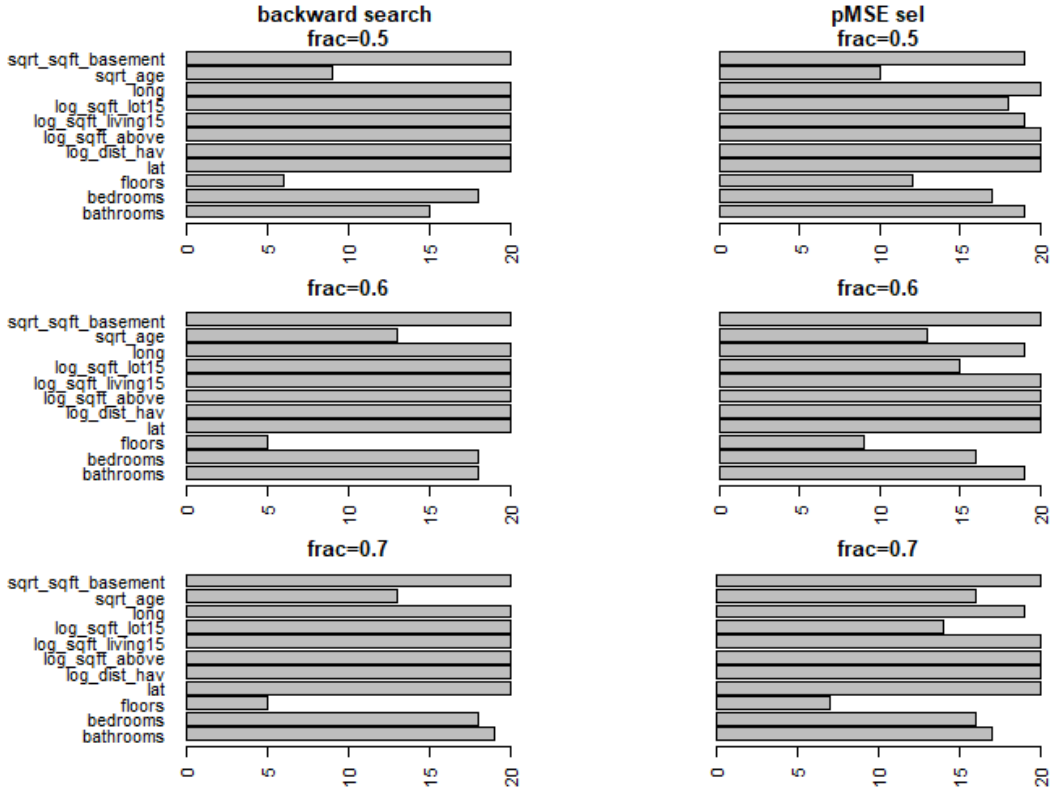

Figure 4: This figure shows count calculation using backward search and selection based on pMSE.

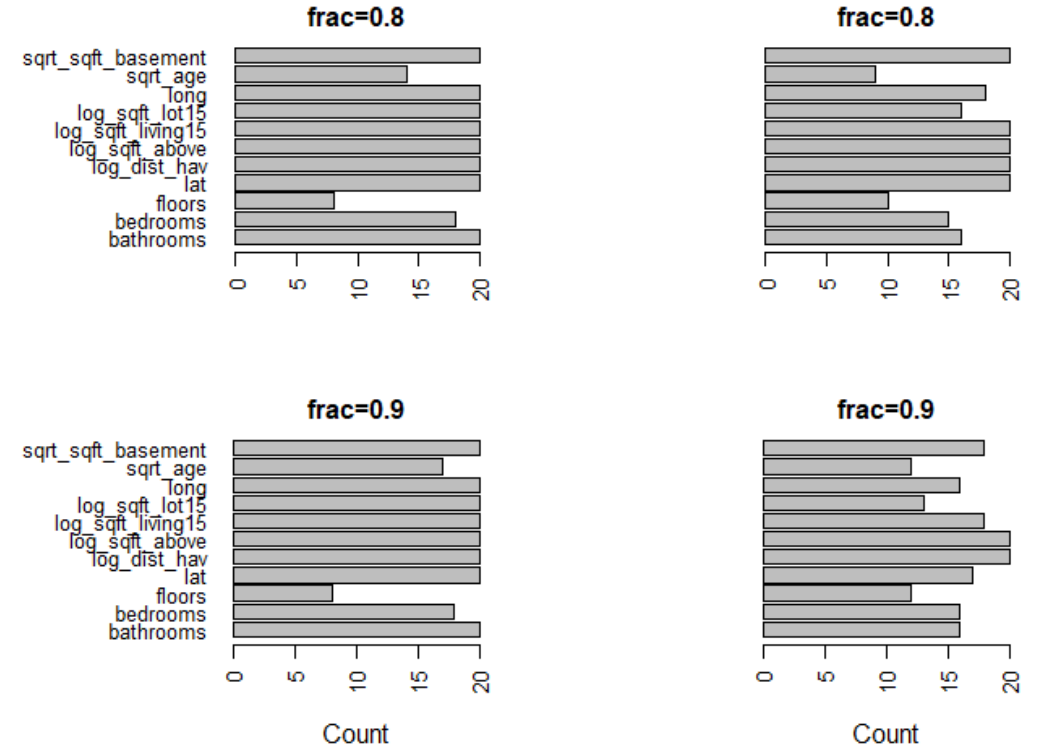

Figure 5: This figure shows count calculation using backward search and selection based on pMSE.

4 Poisson regression model

In this section, we use Poisson regression model to predict the number of bedrooms for the King country house data set using both numerical and categorical covariates. So, bedrooms is the response in this case.

4.1 Selection of covariates

At the starting, we have 20 covariates. 2 new covariates- dist_hav (Haversian distance from downtown Seattle) & age (age of the house) are included. Covariates- grade, condition, view, lat, long & dist_hav are recoded as categorical covariates which are presented in table 4 and table 5.

Table 4: This table presents the categorical covariates- grade, condition and view.

Grade_cat	**Grade**	**Condition_cat**	**Condition**	**View_cat**	**View**
average (baseline)	7	ok (baseline)	3	ok (baseline)	2-3
poor	1-3	poor	1-2	poor	0-1
lower	4-6	good	4-5	amazing	4
better	8-9				
higher	10-13				

Table 5: This table presents the categorical covariates- lat, long and dist_hav.

Lat_cat	**Lat**	**Long_cat**	**Long**	**Dist_hav_cat**	**Dist_hav**
low	47.2-47.5	low	-122.5-(-122.3)	low	981-12170
medium	47.5-47.6	medium	-122.3-(-122.2)	medium	12170-21897
high	47.6-47.8	high	-122.2-(-121.3)	high	21897-77178

The excluded and included covariates are presented in table 6.

Table 6: This table presents the excluded covariates and included covariates.

Covariates excluded	Covariates included
id	price
date	bathrooms
yr_built	sqft_living
yr_renovated	sqft_lot
zipcode	floors
	waterfront
	view_cat
	condition_cat
	grade_cat
	sqft_above
	sqft_basement
	lat_cat
	long_cat
	sqft_living15
	sqft_lot15
	age
	dist_hav_cat

At first, we analyse the included covariates by plotting boxplots and we remove sqft_lot, waterfront, lat_cat, long_cat, sqft_lot15, age and dist_hav_cat. Next, to check multicollinearity, we plot the correlation graph of the remaining numerical covariates which is shown in figure 6. The diagonal matrix shows the distribution graphs of the covariates.

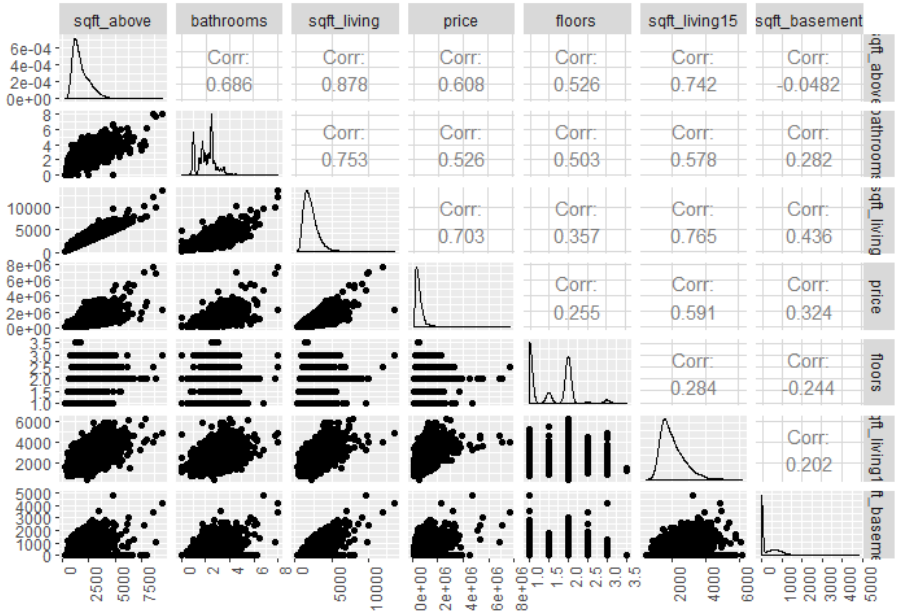

Figure 6: This figure shows the correlation graph of the most highly correlated covariates.

Using figure 6, we remove the covariates- sqft_living, sqft_living15 and sqft_basement. So, now, we have

7 covariates which is presented in table 7.

Table 7: This table presents the final 7 covariates.

Numerical covariates	**Categorical covariates**
sqft_above	view_cat
price	condition_cat
bathrooms	grade_cat
floors	

4.2 Model building

For model building, we use the final 7 covariates. We use 70% training data to actually build the model and 30% testing data to evaluate the final model at the end. First, we fit a standard poisson regression model. We get highly significant underdispersion with a value of 0.1626492 ($p<2.2e-16$). Then we fit using a quasipoisson model. We get the same parameter estimates for the models but the standard error is reduced.

We initially build a model using grade_cat and we used forward search and the likelihood ratio test to add additional variables. Also we investigated the effect of the initial variable on the model that was selected. These models are compared in table 8. In table 8, -view_cat means that the covariate was not selected by forward selection.

Table 8: This table presents the different models.

Initial variable	grade_cat	bathrooms	sqft_above	price	floors	condition_cat	view_cat
Variables added by forward selection	bathrooms sqft_above floors condition_cat price -view_cat	sqft_above grade_cat floors condition_cat price -view_cat	bathrooms grade_cat floors condition_cat price -view_cat	bathrooms sqft_above grade_cat floors condition_cat -view_cat	bathrooms sqft_above grade_cat condition_cat price -view_cat	bathrooms sqft_above grade_cat floors price -view_cat	bathrooms sqft_above grade_cat floors condition_cat price

This indicates that if we start with a good covariate, we converge to a good model. Finally, we get the best model as the one including all 7 covariates except view_cat.

4.3 Wald test

Wald test is used to see that the parameter estimates are significant i.e., significantly different from zero but for categorical variables the interpretation is different - the parameter estimates are significantly different from the baseline level. For the Wald test, we consider the model including all 7 covariates without view_cat. The parameter estimates and the p-values of the poisson regression model obtained by Wald test are presented in table 9.

Table 9: This table presents the poisson regression model obtained by Wald test.

Parameter	Estimate	p-value
(Intercept)	8.405e-01	<2e-16
grade_catbetter	-2.498e-02	5.53e-08
grade_cathigher	-1.380e-01	<2e-16
grade_catlower	-9.099e-02	<2e-16
grade_catpoor	-1.269e+00	8.39e-06
bathrooms	1.432e-01	<2e-16
sqft_above	1.158e-04	<2e-16
floors	-7.547e-02	<2e-16
condition_catgood	5.314e-02	<2e-16
condition_catpoor	-2.967e-02	0.149
price	-3.923e-08	4.87e-11

We interpret the parameter estimates as

- 15% more bedrooms for each bathroom added.
- Same number of bedrooms if sqft_above increases by 1 sqft.
- Same number of bedrooms if price increases by 1 US dollar.
- 7% fewer bedrooms for each extra floor added.

Compared to average grade:

- 2% fewer bedrooms if grade is better.
- 13% fewer bedrooms if grade is higher.
- 9% fewer bedrooms if grade is lower.
- 72% fewer bedrooms if grade is poor.

We see the predicted vs. observed number of bedrooms plot in figure 7.

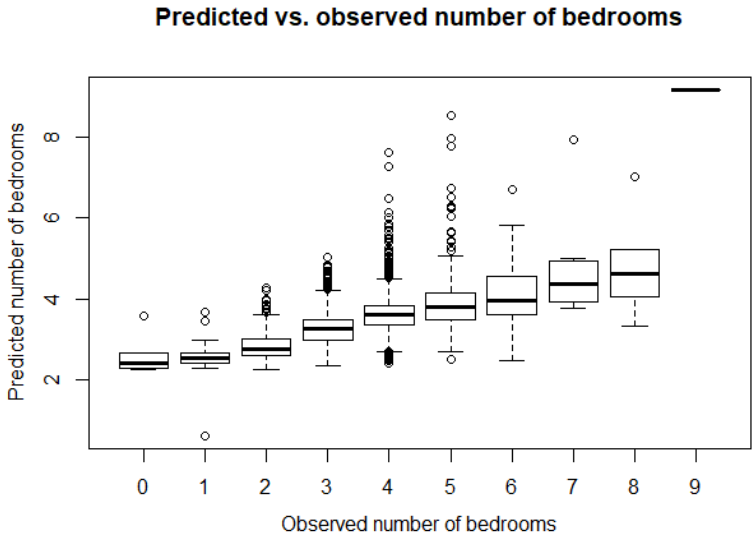

Figure 7: This figure shows the Predicted vs. observed number of bedrooms plot.

5 Concluding Remarks

A real data set is studied in the project. In the project, we get the following two results-
1. We get the best multivariate regression model to predict the price of the house using 11 transformed covariates- sqrt(sqft_basement), sqrt(age), long, log10(sqft_lot15), log10(sqft_living15), log10(sqft_above), log10(dist_hav), lat, floors, bedrooms and bathrooms. The model selection method-selection based on pMSE is better than backward search since pMSE calculated is slightly less in the former.
2. We get the best poisson regression model to predict the number of bedrooms of the house using 6 covariates- grade_cat, bathrooms, sqft_above, floors, condition_cat and price. Using the parameter estimates of the poisson regression model obtained by Wald test, we infer that there are 15% more bedrooms for each bathroom added, same number of bedrooms if sqft_above increases by 1 sqft, same number of bedrooms if price increases by 1 US dollar, 7% fewer bedrooms for each extra floor added and compared to average grade, there are 2% fewer bedrooms if grade is better, 13% fewer bedrooms if grade is higher, 9% fewer bedrooms if grade is lower & 72% fewer bedrooms if grade is poor.

References

[1] John O. Rawlings, Sastry G. Pantula and David A. Dickey(1998). Applied Regression Analysis: A Research Tool. Springer.