B.S.: Techniques & Model Project

Introduction

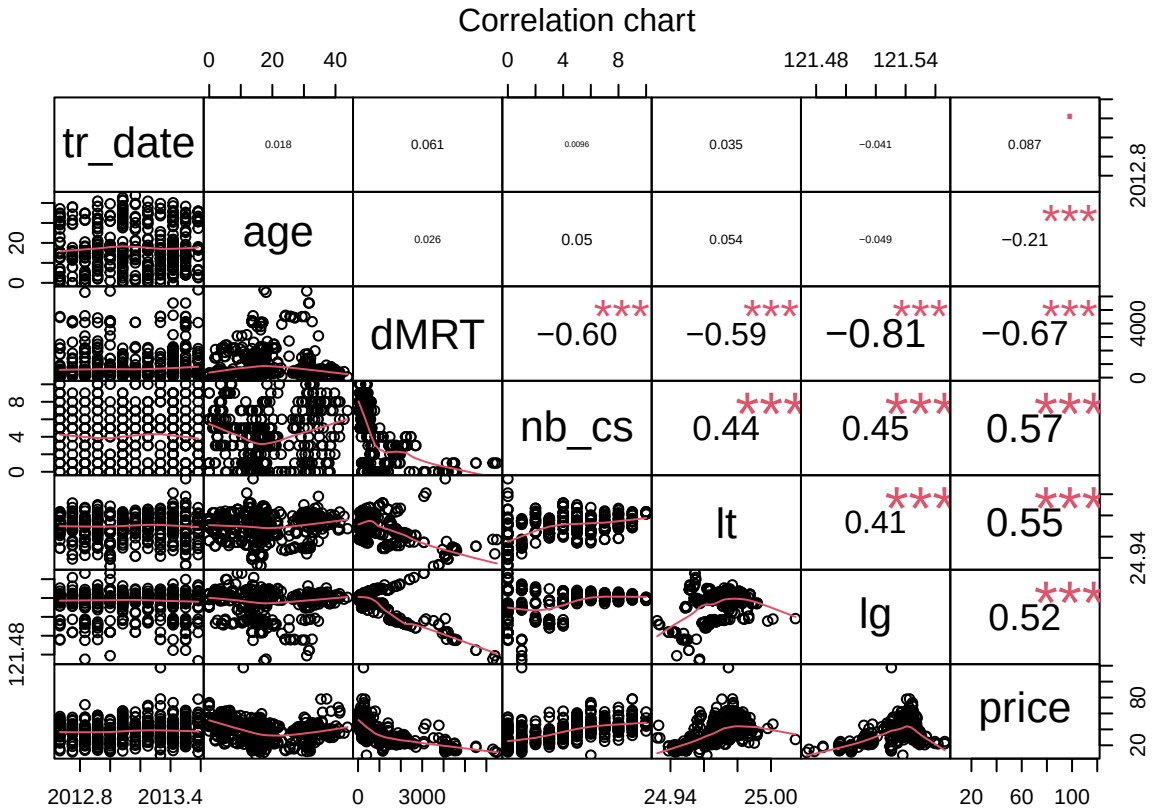
We will be studying a housing dataset that combines a set of assets that were sold between 2013 and 2014 as well as their following parameters: transaction date (tr_date), house age (age), distance to the nearest MRT station (dMRT), amount of convenience stores (nb_cs), latitude (lt) and longitude (lg).

Our aim will be to model the price according to the former parameters, assuming that these are relevant.

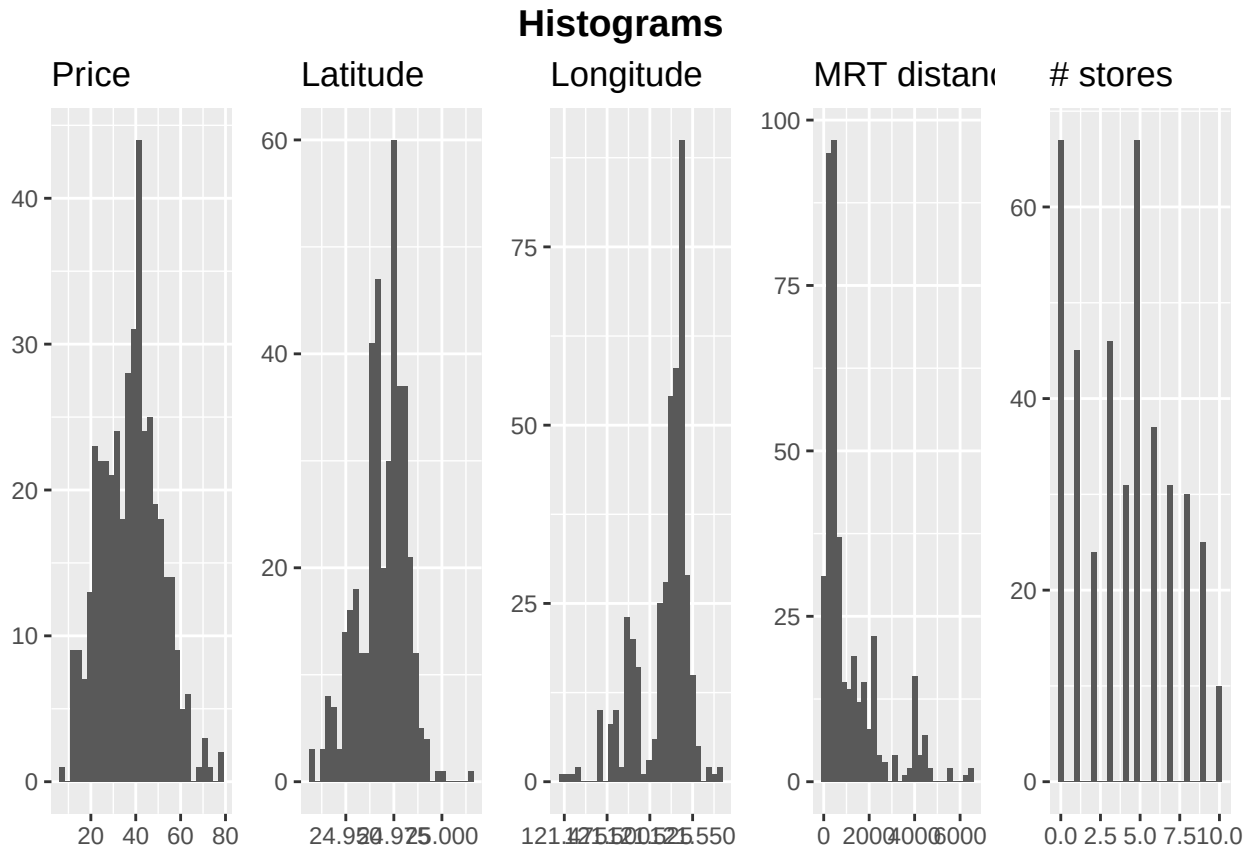
We first aimed to fit a regression model to predict the price according to lt, lg, dMRT and nb_cs. This was compared with another model were nb_cs was used as a category to build groups, which were then used to create localised regression models with the same parameters as before. Both models showed a good performance in terms of residuals, even though they exhibited fat tailed distributions. However, they both suffered from convergence issues (high Gelman factors for several parameters). Notwithstanding its likelihood improvements, the second model displayed a noteworthy overfitting.

Exploratory data analysis

A closer look to the relationship between parameters will help us to select the ones that are relevant



The transaction date and house age are not really correlated with the price. We will thus discard them in our analysis. After removing some obvious outliers that may appear with boxplots, the distribution of the remaining variables are as follows:



Modeling

1^{st} model: linear regression

Model choice: We will thus use the log values of price and the MRT distance and try to fit the following model:

$$\forall i,\ \ln(\text{price}_i) = \beta_1 + \beta_2 \ln(\text{dMRT}_i) + \beta_3 \text{lt}_i + \beta_4 \text{lg}_i + \beta_5 \text{nbcs}_i + \epsilon_i, \epsilon_i \sim N(0, \sigma)$$

This choice is mainly motivated by the linear relationships that can be shown trough scatter plots. Without an a priori knowledge of the topic, we will use non informative priors for our 6 parameters (normal distribution for βs, inverse gamma for σ).

Fitting: Fitting was made with 3 chains, each having 1K burn-in iterations out of 100K in total. Final estimation of our parameters is given below. Despite the high number of iterations, Gelman factor exhibits some convergence issues (fine-tuning was tried but no real improvements were noticed).

```
##             b[1]       b[2]      b[3]       b[4]        b[5]        sig2
## mean   -2.644711 -0.2259611 0.2026948 0.02060695 0.021392804 0.108277796
## sd      7.823372  0.0201253 0.2622679 0.02914555 0.007597293 0.007565374
## Gelman  9.432010  1.0109082 6.3580253 3.29456694 1.000818473 1.009973217
```

2^{nd} model: linear regression with groups based on the amount of convenience stores

Model choice: this second model is merely a location dependent adaptation of the first one, based on the number of convenience store. Instead of embedding this parameter directly onto the linear regression, we will use it as a categorical variable. Hence:

$$\forall i, \#cs_j,\ \ln(\text{price}_i) = \beta_1^{\#cs_j} + \beta_2^{\#cs_j} \ln(\text{dMRT}_i) + \beta_3^{\#cs_j} \text{lt}_i + \beta_4^{\#cs_j} \text{lg}_i + \epsilon_i, \epsilon_i \sim N(0, \sigma)$$

Fitting: Given that we have eleven distinct categories in terms of convenience stores amounts, we will retrieve a 11×4 matrix of βs in addition to σ (we have arbitrarily chosen a unique sigma across the different groups given the limited scope of this project). Means of all the components are given below. Gelman indicators are similar across the different parameters families as before

```
##    # Stores # Obs b[1] (mean) b[2] (mean) b[3] (mean)  b[4] (mean)
## 1         0    67   -12.85311  -0.2780466   0.4791522   0.05058122
## 2         1    45    7.754248   -0.264615  -0.1966961   0.01903738
## 3         2    24   -7.391444  -0.3531812   0.2829174   0.05148846
## 4         3    46   -8.942583  -0.3193195   0.3022182   0.05839657
## 5         4    31    11.99384  -0.2272015  -0.2455805 -0.006061617
## 6         5    67   -5.452578   -0.151109   0.1126773   0.06006757
## 7         6    37   -12.02469  -0.2866494 -0.01585576    0.1467783
## 8         7    31   -1.613014  -0.1247553   0.3899914  -0.03039671
## 9         8    30   -9.448089  -0.1828977   0.1901907   0.07818992
## 10        9    25    19.37577     -0.1774  -0.1399434   -0.0908898
## 11       10    10    16.04852  0.09408184  -0.2835863  -0.04596285
```

```
##        sig2
## 1 0.1057057
```

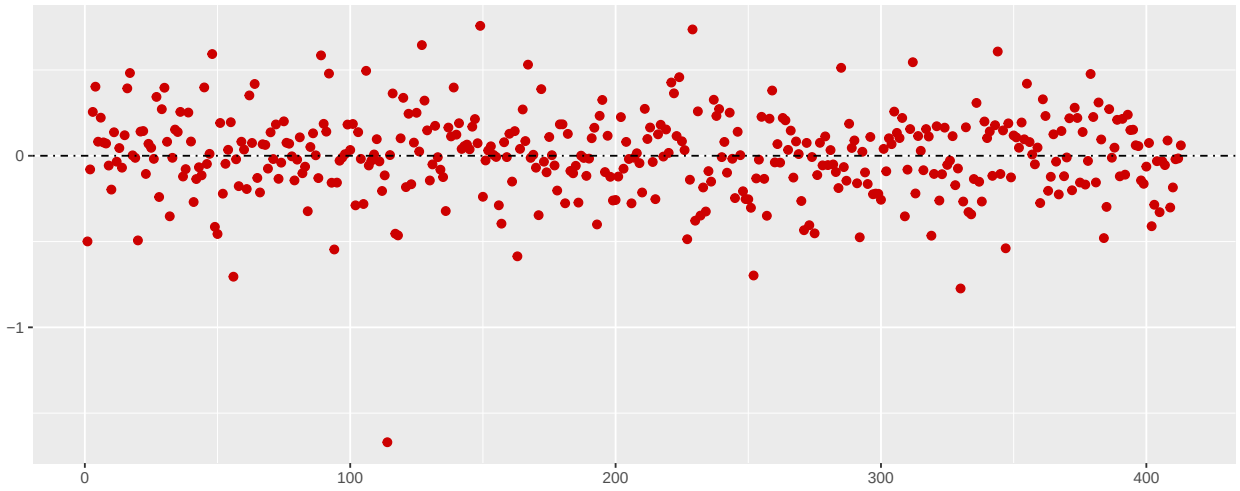
Results

- Overall, the relationships between data properties are reproduced in the posteriors: most of β_2s are negative, which follows the negative correlation between dMRT and price. The same applies for β_5 in the first model. Latitude/longitude β parameters in the second model may indicate that houses with the same numbers of convenience stores nearby are part of similar districts. Distance to the nearest MRT seems to be the major factor that drives housing prices
- In terms of residuals, no pattern is noticeable (see figures thereafter). Q-Q plots exhibit a residuals distribution with fat tails for both models, this may be a possible future improvement.
- DIC results indicates that our second model suffers from overfitting: despite increasing the likelihood, the significant penalty makes the overall variance worse than the simpler one.

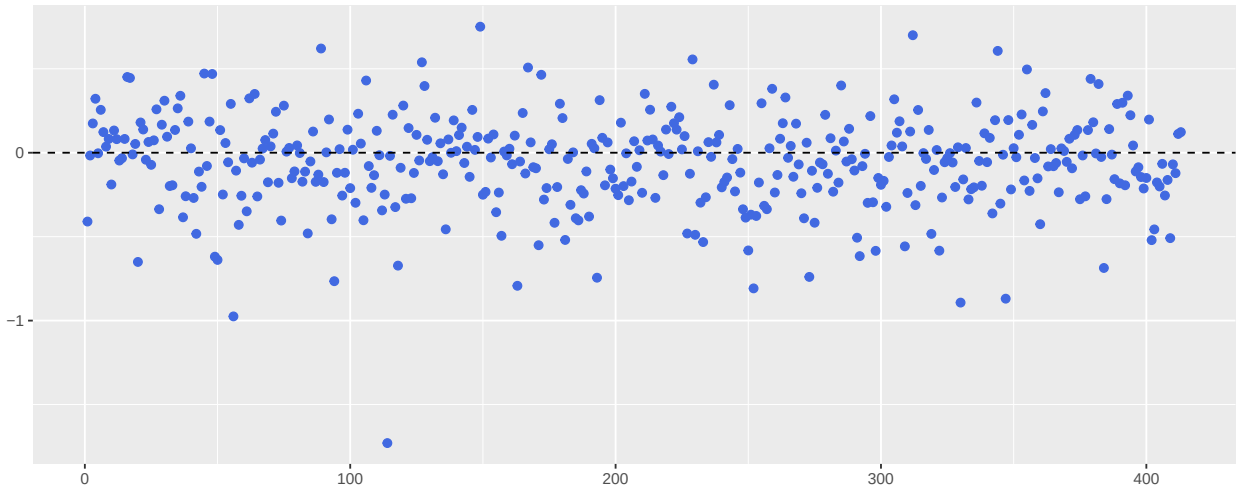
```
##                    first_model second_model
## Mean deviance           72.210        61.05
## Penalty                  4.237        24.89
## Penalized variance      76.440        85.85
```

Residuals and QQplots

1st model: residuals



2nd model: residuals



QQplots: 1st model (red) and 2nd model (blue)