Applied Statitics Group Courswork

Kyriacos Xanthos 01389741
Hugo Barnett 01511877
Xiaowei 02141000
Yiheng Jia 02035952
Erwan Delorme 02114711,
We all contributed to everything

December 17, 2021

1 Exploratory Data Analysis

We consider the Real Estate evaluations dataset, which are the historical real estate valuations from a particular region. First we read the data and change the column names to shorter ones.

```
> dat <- read.csv("data.csv")
> colnames(dat) <- c("date", "age", "distance", "stores", "lat",
"lon", "price") #changed column names to make titles shorter
> head(dat)
      date  age   distance stores      lat      lon price
1 2012.917 32.0   84.87882     10 24.98298 121.5402  37.9
2 2012.917 19.5  306.59470      9 24.98034 121.5395  42.2
3 2013.583 13.3  561.98450      5 24.98746 121.5439  47.3
4 2013.500 13.3  561.98450      5 24.98746 121.5439  54.8
5 2012.833  5.0  390.56840      5 24.97937 121.5425  43.1
6 2012.667  7.1 2175.03000      3 24.96305 121.5125  32.1
```

```
> par(mfrow = c(1,2))
> boxplot(dat$price,xlab = "price")
> hist(dat$price,xlab = "price")
```

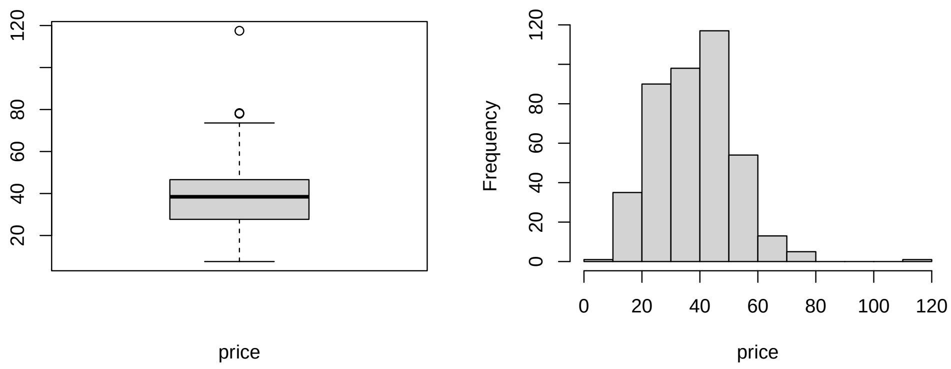

Figure 1: Boxplot and histgram for house price

We looked at the pairs plot and saw that there seemed to be a linear relationship between **price** and other variables. It is also a reasonable assumption that the response variable would be dependent on all the other variables.

```
> my_sf <- st_as_sf(dat, coords = c('lon','lat'), crs = 4326)
> ggplot(my_sf) + geom_sf(aes(color = price))
```

```
> ggplot(my_sf) + geom_sf(aes(color = stores))
```

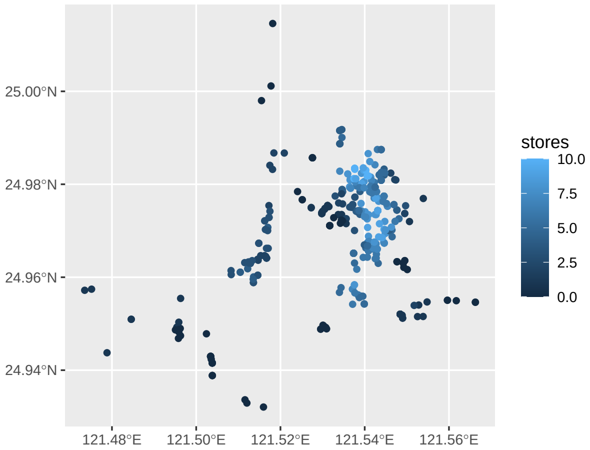

Figure 2: Longitude vs latitude : vs stores

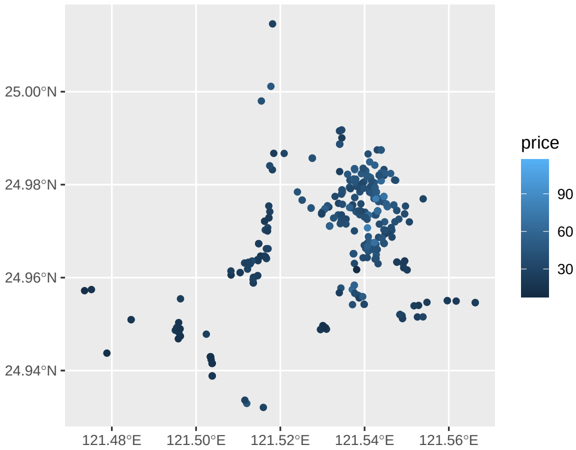

Figure 3: Longitude vs latitude : vs Price

It seems like the latitude and longitude are not directly correlated to the price or stores density but rather that there is a central location for which the distance to the house is correlated to price and stores. This central point has high prices and high density. This is located around (24.975,121.54), we thus record the distance between this location and each point to use this as a predictor.

```
>distance=function(x,y,c) sqrt((x-c[1])^2+(y-c[2])^2)
>dat$dcc=distance(dat$lat,dat$lon,c(24.975,121.54))
```

There is a good argument for either store density or price as a response variable. We decided to use **price** for the first model and **stores** for the second model.

2 Linear Model

We began by including all variables so that there was a metric in which we could compare any transformations or changes to the model. We call this our 'Null model'.

$$y_i = \beta_0 + \sum_{j=1}^{7} \beta_j x_{i,j} + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right) \tag{1}$$

where y is the response variable **Price**, β_0 is the intercept, the β_i where $i = 1, 2, \ldots, 7$ are the coefficients for the variables (which are age, distance to metro station, store density, latitude, longitude and date) and $x_{i,j}$ is the i^{th} observation for the j^{th} variable.
We debated on whether we should include date of transaction (date) in the model. The argument for inclusion of this variable were; (1) we would expect date to have an impact due to annual inflation rate, (2) upon examination in the linear model we see that it is a significant explanatory variable.

```
> model0 <- lm(price ˜ age + distance + stores + lat + lon + date, data=dat)
> summary(model0)

Call:
lm(formula = price ˜ age + distance + stores + lat + lon + date,
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-35.664  -5.410  -0.966   4.217  75.193

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.444e+04  6.776e+03  -2.131  0.03371 *
age         -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
distance    -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
stores       1.133e+00  1.882e-01   6.023 3.84e-09 ***
lat          2.255e+02  4.457e+01   5.059 6.38e-07 ***
lon         -1.242e+01  4.858e+01  -0.256  0.79829
date         5.146e+00  1.557e+00   3.305  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.858 on 407 degrees of freedom
Multiple R-squared:  0.5824,        Adjusted R-squared:  0.5762
F-statistic: 94.59 on 6 and 407 DF,  p-value: < 2.2e-16
```

We note that the `intercept`, `longitude` and `date` do not seem to be significant. We debated o whether to include or exclude the intercept as this significantly increased the R^2 value, however, it was noted that it did not make sense for the price to be equal to zero when `distance to stores`, `latitude` and `longitude` was zero and also the R^2 metric is meaningless without the intercept. We see that the Adjusted R^2 value is only 58% so there is a significant improvement that needs to be made for this model.

We can check the correlation between our variables using the variance inflation factor (VIF) [1] page 45.

```
> vif(model0)
     age distance   stores      lat      lon     date
1.014287 4.323019 1.617038 1.610234 2.926302 1.014674
```

We can see that all of the VIFs are small so we continue with the exploration of the diagnostics of this fit.

```
> par(mfrow = c(2,3))
> for(k in 1:6){
```

```
+   plot(model0,k)
+ }
```

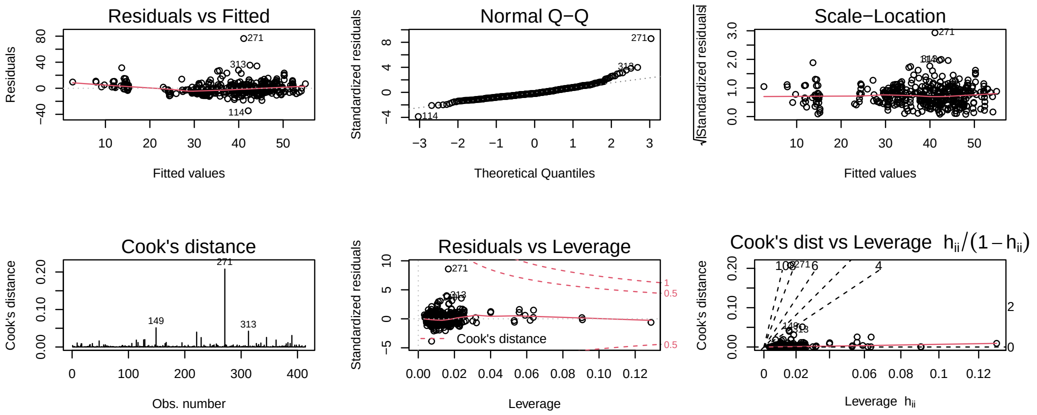

Figure 4: Diagnostic plot for model 0

Looking at the above diagnostics in Figure 4, we can see that there are some outliers in our model. We can remove these by setting a maximum amount of Cook's distance [1] page 37 and leverage [2] page 35.

```
> dat_new <- dat[(cooks.distance(model0)<=0.02)&(hatvalues(model0)<=0.05),]
```

Now looking at the price response variable against the distance in Figure 5, it seems like there is a exponential relationship between them.

```
> plot(dat$distance, dat$price, cex=0.5)
```

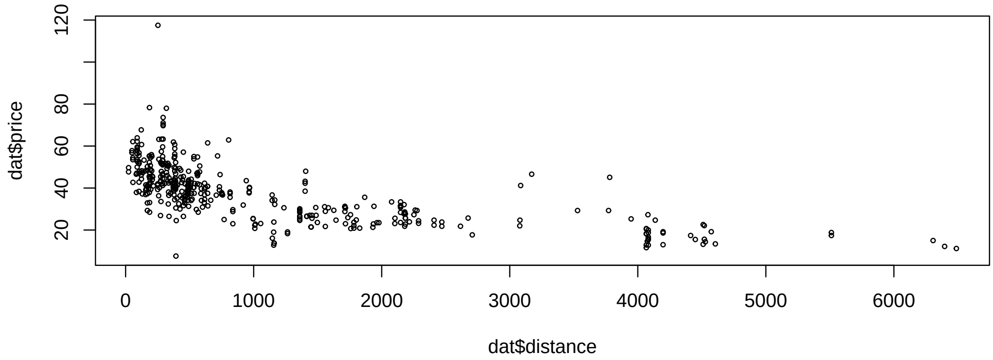

Figure 5: Scatter plot of price against distance

This suggests that we can transform the distance variable to the logarithm of the distance. Moreover, we can look at a Box-Cox plot to assess whether our response's needs any transformations.

```
> boxcox(model0)
```

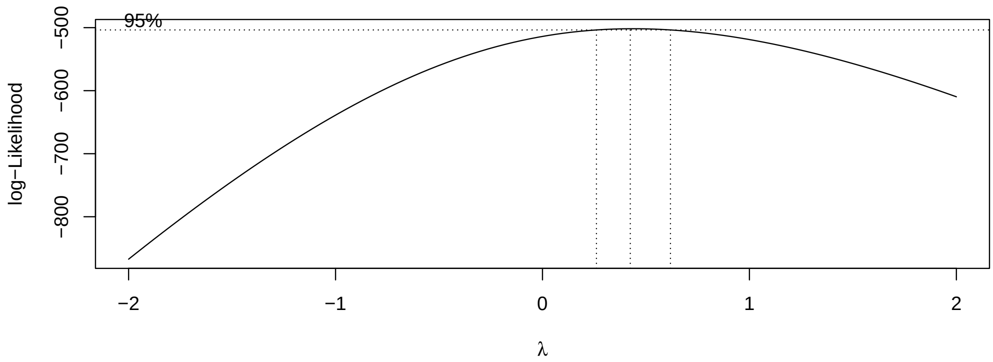

Figure 6: Box-cox: log-likelihood vs y-transformations

Looking at the box-cox plot, we see that the 95% confidence interval for λ does not include the value of 1 which suggests we need to do a transformation to the response variable price. We do this transformation and we re-run the model including all of the above changes.

```
> distance=function(x,y,c){
+   sqrt((x-c[1])^2+(y-c[2])^2)
+ }
> dat_new$dcc=distance(dat_new$lat,dat_new$lon,c(24.975, 121.54))
> model2 <- lm(price^0.5 ~ 1+age + log(distance) + stores + lat + lon +date
+dcc, data=dat_new)
> summary(model2)

Call:
lm(formula = price^0.5 ~ 1 + age + log(distance) + stores + lat +
    lon + date + dcc, data = dat_new)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42904 -0.29004 -0.00328  0.28339  2.15477

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.018e+03  5.009e+02  -2.032  0.04282 *
age           -2.299e-02  2.319e-03  -9.912  < 2e-16 ***
log(distance) -3.984e-01  4.632e-02  -8.602  < 2e-16 ***
stores         3.984e-02  1.216e-02   3.277  0.00114 **
lat            1.784e+01  3.804e+00   4.690 3.79e-06 ***
lon           -2.077e+00  3.487e+00  -0.596  0.55168
date           4.143e-01  9.089e-02   4.558 6.91e-06 ***
dcc           -2.065e+01  6.289e+00  -3.284  0.00112 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.502 on 389 degrees of freedom
Multiple R-squared:  0.776,        Adjusted R-squared:  0.772
F-statistic: 192.6 on 7 and 389 DF,  p-value: < 2.2e-16
```

We can see that there is a significant improvement on the Adjusted R^2 metric which is 77% at the moment. We can also see that the new `dcc` variable is significant at the 99% significance level so it makes sense to include it but the Longitute variable is not significant at all.

3 Second Model

In the section, we choose a second variable 'Number of convenience stores' as response variable and fit the Poisson regression GLM to these data using the canonical link function, assuming the presence of an intercept term. The reason why we choose Poisson regression GLM is that the 'Number of convenience stores' is a discrete variable, which makes it possible to use Poisson GLM.

To be more specific, let Y_i be the i th observation of variable **Stores**.

$$E(Y_i) = \exp(\beta_0 + \sum_{j=1}^{6} \beta_i x_{i,j})$$

where $i = 1, 2, \ldots, 6$ are the coefficients for the variables (which are date, age, distance to metro station, latitude, longitude and price) and $x_{i,j}$ is the i^{th} observation for the j^{th} variable.

```
> model3 <- glm(stores˜date+age+log(distance)+lat+lon+price,
data = dat,family = "poisson")
> summary(model3)

Call:
glm(formula = stores ˜ date + age + log(distance) + lat + lon +
    price, family = "poisson", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3263  -0.8347   0.0639   0.6716   2.0939

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.690e+03  3.960e+02  -4.267 1.98e-05 ***
date           1.396e-01  9.051e-02   1.543  0.12291
age            5.742e-03  2.064e-03   2.782  0.00540 **
log(distance) -3.715e-01  3.389e-02 -10.961  < 2e-16 ***
lat            1.661e+01  2.752e+00   6.034 1.60e-09 ***
lon            8.209e+00  2.941e+00   2.791  0.00525 **
price          4.002e-03  2.701e-03   1.481  0.13848
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1100.90  on 413  degrees of freedom
Residual deviance:  592.95  on 407  degrees of freedom
AIC: 1748.2

Number of Fisher Scoring iterations: 5
```

From the summary we can see the convariate variables 'Transaction date' and 'House price' are not significant. Thus we remove these two variables to simplify the structure of the model and also remove the outlier.

We attempt to remove all of the variables with non significant p values to see how it affects our model diagnostics. We also remove the outliers from the normal linear model. We suppose that this will help us improve our model as residuals with high leverage will have high leverage regardless of the model.

```
dat_new <- dat[(cooks.distance(model2)<=0.02)&(hatvalues(model2)<=0.05),]
model4 <- glm(stores˜age+log(distance)+lat+lon,
data = dat_new,family = "poisson")
summary(model4)
Call:
glm(formula = stores ~ age + log(distance) + lat + lon, family = "poisson",
    data = dat_new)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2971  -0.7416   0.0456   0.7615   1.9904

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.277e+03  3.582e+02  -3.566 0.000363 ***
age            4.104e-03  2.086e-03   1.967 0.049195 *
log(distance) -4.176e-01  3.108e-02 -13.435  < 2e-16 ***
lat            1.887e+01  2.712e+00   6.958 3.45e-12 ***
lon            6.663e+00  2.983e+00   2.234 0.025492 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 1066.60  on 402  degrees of freedom
Residual deviance:  572.45  on 398  degrees of freedom
AIC: 1688.7
```

The model seems to perform better as it has a much lower AIC and all p values are significant at least at a 0.05 significance level. The null deviance and residual deviance is also slightly reduced. We therefore keep this new model.

4 Summary

We find that our final linear model is a good fit. Not only the adjusted R^2 rose up to 77%, all explanatory variables also have low p values and therefore seem to be significant enough to be included in the model. The price as a response variable makes intuitive sense since we can use this linear model to predict the estate evaluation price given a number of factors. If we had more time we could test this by splitting the data into training and testing sets and try to predict the estate price given the variables.

Further, our final general linear model is also a good fit. We removed variables and outliers to improve the AIC and make sure we have significant p values.

Comparing the two models we find that fitting a model to price is easier. Our diagnostics for the linear model are excellent. This may be due to the explanatory variables in the dataset. These indeed seem to be tailored to price rather than counting stores.

References

[1] Din-Houn Lau. Applied statistics lecture notes. *Imperial College London (Lecture)*, October 2020.

[2] Nick Heard. Normal linear model. *Imperial College London (Lecture)*, October 2021.