Midterm 1 Stat 445

Matthew Parker

9/25/20

For this Midterm project I will be using the "Real estate valuation data set" by Prof. I-Cheng Yeh from the Department of Civil Engineering, Tamkang University, Taiwan. This data set was taken from the UCI Machine Learning Repository.

Research Question

When factoring for location differences and the variability of the house age and year of transaction, does distance to the nearest MRT station (metro) cause a practically significant increase in the price of housing?

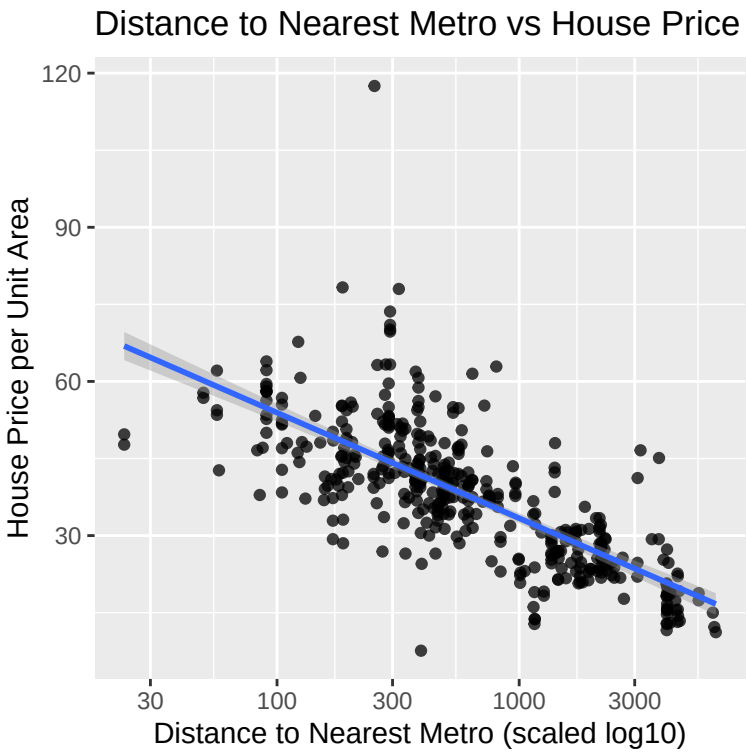
This is an interesting and relevant research question because it is clear, it has complete details on what is being factored for and what the question aims to solve. It is focused, there is a set goal and direction. It is concise, there is no extra wording. The question is also complex enough that you cannot simply answer yes or no, there is arguability in the effect of the output price, because there are other possible factors to consider, outside of the used data set. This leads to the last point, in that the question is arguable.

Exploratory Analysis

The first relationship I found interesting in my data was the relationship between the distance to the nearest MRT (metro) station and the house price per unit area. On graphing the relationship between these two, there was an obvious exponential relationship, with house price clearly being decreased with distance to MRT being increased. I then graphed the x variable as scaled by log10 and could see a clearly linear relationship between the two variables, which was exciting. My research question is "When factoring for location differences and the variability of the house age and year of transaction, does distance to the nearest MRT station (metro) cause a practically significant increase in the price of housing?", and with a linear relationship being seen, clearly there is a practically significant increase in the price of housing due to distance to MRT station.

```
ggplot(data = df,
       mapping = aes(x = X3.distance.to.the.nearest.MRT.station,
                     y = Y.house.price.of.unit.area))+
  geom_point(alpha=0.75) +
  labs(title = "Distance to Nearest Metro vs House Price",
       x = "Distance to Nearest Metro (scaled log10)",
       y = "House Price per Unit Area") +
  scale_x_log10() +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Here we can see that there is likely a linear relationship between the predictor of MRT station distance and response variable house price.

The next relationship to explore is that between X3 distance to metro and X4 number of convenience stores in the immediate area. I found through a correlation test that the two were indeed correlated (after testing many other variable relationships as well).

```
res3 <- cor.test(df$X3.distance.to.the.nearest.MRT.station,
                 df$X4.number.of.convenience.stores, method="pearson")
res3
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$X3.distance.to.the.nearest.MRT.station and df$X4.number.of.convenience.stores
## t = -15.324, df = 412, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6605398 -0.5373447
## sample estimates:
##        cor
## -0.6025191
```

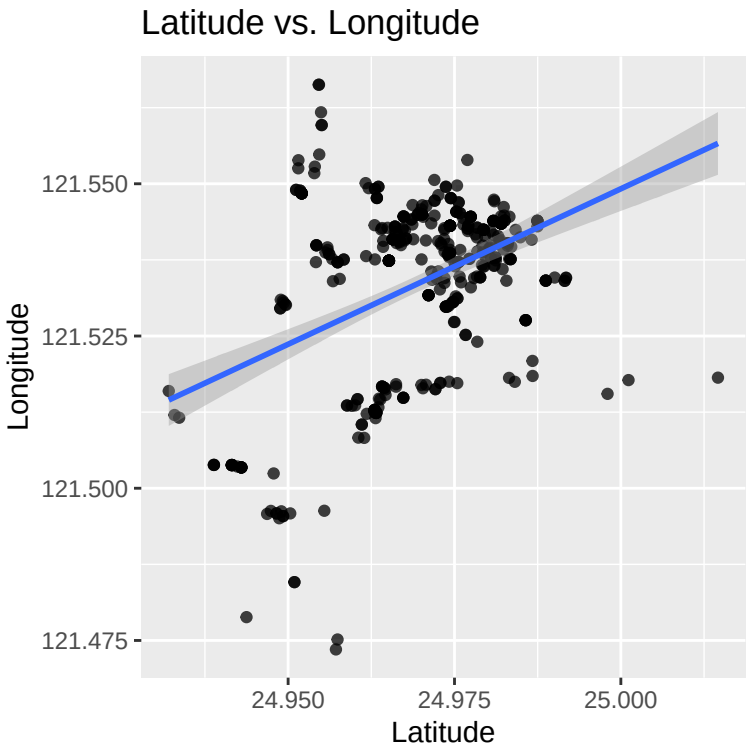
The important takeaway here is the p-value of 2.2e-16, which is obviously extremely low, meaning that there is a practically significant correlation between the two variables.

Another important correlation is between latitude and longitude, the two location variables.

```
ggplot(data = df,
       mapping = aes(x = X5.latitude,
                     y = X6.longitude))+
  geom_point(alpha=0.75) +
```

```
  labs(title = "Latitude vs. Longitude",
       x = "Latitude",
       y = "Longitude") +
  geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

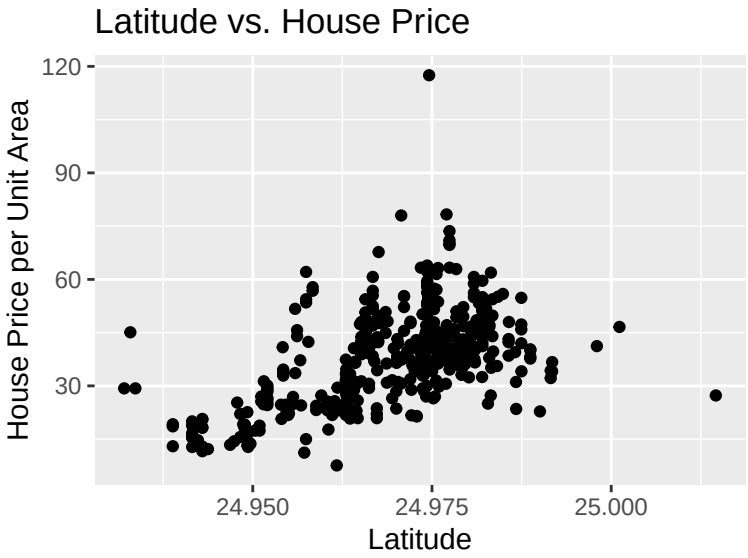



```
res4 <- cor.test(df$X5.latitude, df$X6.longitude, method="pearson")
res4
```

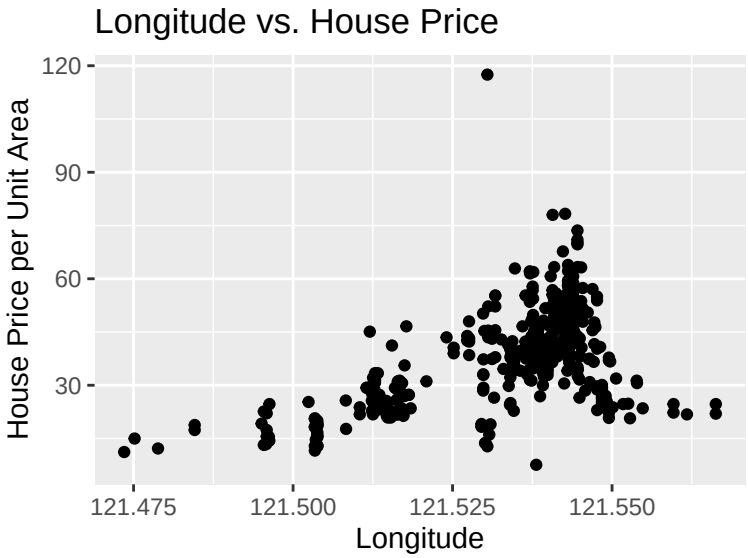
```
##
##  Pearson's product-moment correlation
##
## data:  df$X5.latitude and df$X6.longitude
## t = 9.2026, df = 412, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3296657 0.4898090
## sample estimates:
##       cor
## 0.4129239
```

Here we again have a very low p-value of 2.2e-16, leading to a conclusion that latitude and longitude variables will be highly correlated in the linear regression.

```
ggplot(data=df, mapping = aes(x = X5.latitude, y=Y.house.price.of.unit.area)) +
  geom_point() +
  labs(title = "Latitude vs. House Price",
       x = "Latitude",
       y = "House Price per Unit Area")
```




```
ggplot(data=df, mapping = aes(x = X6.longitude, y=Y.house.price.of.unit.area)) +
  geom_point() +
  labs(title = "Longitude vs. House Price",
       x = "Longitude",
       y = "House Price per Unit Area")
```



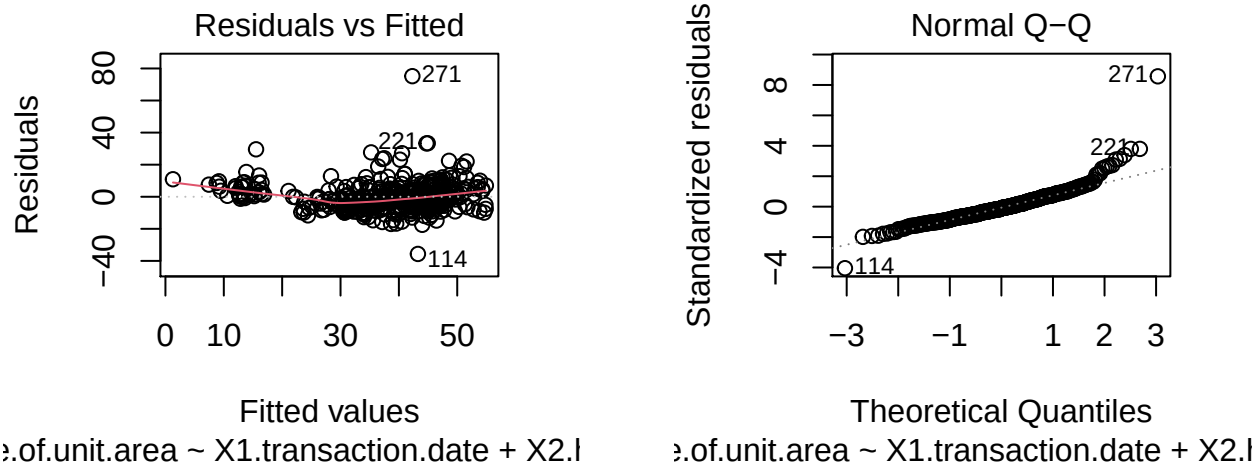
If we graph latitude and longitude vs house price, we can see some definitive cluster points, assumed around downtown areas that would be more expensive. This would make sense as to why the two variables might be correlated, because there is a clustering of expensive housing, and only one of the two variables being necessary in the regression would make sense.

Linear Regression

Now to construct the linear regression model.

```
df <- read.csv("Data.csv")
modl = lm(formula = Y.house.price.of.unit.area ~ X1.transaction.date + X2.house.age +
            X3.distance.to.the.nearest.MRT.station + X4.number.of.convenience.stores +
```

```
            X5.latitude + X6.longitude  , df)
plot(modl)
```



1. Plotting the response variable vs the predictor variables leads to a fairly linear looking relationship.
2. Each predictor is correlated with the response with great significance less than 5%, except for longitude. We know from before that X3 MRT and X4 convenience store are correlated with great significance, as well as X5 latitude and X6 longitude.

```
summary(modl)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X1.transaction.date +
##     X2.house.age + X3.distance.to.the.nearest.MRT.station + X4.number.of.convenience.stores +
##     X5.latitude + X6.longitude, data = df)
##
## 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
## X1.transaction.date                      5.146e+00  1.557e+00   3.305  0.00103
## X2.house.age                            -2.697e-01  3.853e-02  -7.000 1.06e-11
## X3.distance.to.the.nearest.MRT.station -4.488e-03  7.180e-04  -6.250 1.04e-09
## X4.number.of.convenience.stores          1.133e+00  1.882e-01   6.023 3.84e-09
## X5.latitude                              2.255e+02  4.457e+01   5.059 6.38e-07
## X6.longitude                            -1.242e+01  4.858e+01  -0.256  0.79829
##
## (Intercept)                             *
## X1.transaction.date                     **
## X2.house.age                            ***
## X3.distance.to.the.nearest.MRT.station ***
## X4.number.of.convenience.stores         ***
## X5.latitude                             ***
## X6.longitude
## ---
## 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
```

3. Each parameter is significant , other than longitude. This variable is strongly correlated with latitude. Now to remove longitude to create a better regression.

```
modl = lm(formula = Y.house.price.of.unit.area ~ X1.transaction.date + X2.house.age +
            X3.distance.to.the.nearest.MRT.station + X4.number.of.convenience.stores +
            X5.latitude  , df)
summary(modl)
```

```
##
## Call:
## lm(formula = Y.house.price.of.unit.area ~ X1.transaction.date +
##     X2.house.age + X3.distance.to.the.nearest.MRT.station + X4.number.of.convenience.stores +
##     X5.latitude, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.623  -5.371  -1.020   4.244  75.346
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -1.596e+04  3.233e+03  -4.936 1.17e-06
## X1.transaction.date                      5.135e+00  1.555e+00   3.303  0.00104
## X2.house.age                            -2.694e-01  3.847e-02  -7.003 1.04e-11
## X3.distance.to.the.nearest.MRT.station -4.353e-03  4.899e-04  -8.887  < 2e-16
## X4.number.of.convenience.stores          1.136e+00  1.876e-01   6.056 3.17e-09
## X5.latitude                              2.269e+02  4.417e+01   5.136 4.36e-07
##
## (Intercept)                             ***
```

```
## X1.transaction.date                     **
## X2.house.age                            ***
## X3.distance.to.the.nearest.MRT.station ***
## X4.number.of.convenience.stores         ***
## X5.latitude                             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.848 on 408 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5772
## F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

4. Most estimated values stayed the same, except for that of latitude which increased greatly and actually changed signs.
5. The p-values of most variables stayed about the same. The only variables greatly affected were the intercept and X3 distance to MRT. Each of these decreased significantly in p-value, with MRT now becoming the most significant variable with a p-value five powers of ten lower than each other variable.
6. A confidence interval including zero effectively means that the null hypothesis is included in the confidence interval, so long as this is not true leads to a low p-value. A confidence interval including zero leads to a high p-value for the estimated value.
7. R2 is interpreted at a high level as the amount or variance that can be explained by the model. In this case a little over 50% of the variance is contained in the model, which means the model is above average.

Conclusion

Through the regression there has been found an extremely significant relationship between all five predictors and the resultant variable of house price per unit area. It is interesting that latitude and longitude are correlated together, but not surprising. Also not surprising is that MRT station distance and convenience stores are correlated variables, since the two likely go hand in hand for house locations within the major downtown areas. Through regression, the p-value of X3 MRT station gave a very strong answer to the research question. In conclusion, I have found that there is indeed a practically significant relationship between distance to nearest MRT station and House Price per Unit Area. Given more time to study the problem, I would try to sort out any error in the model by finding a way to include exact location of housing by combining the latitude and longitude variables. I would also test whether the correlation between MRT station and convenience stores is effecting the results of the model. There were also variables of house age and transaction date that were left fairly bare, and those should be examined further, as they are clearly significant factors for house price.

Works Cited

UCI Machine Learning Repository

Data Camp Quick R

Rstudio ggplot cheat sheet

Data Camp linear regression