# “Modeling Real Estate Valuations and Concrete Strength”

*Emily Lu and Jason Edwards*

*May 26th, 2019*

# Abstract

This report is split into two components. The first component serves to evaluate the relationship between real estate valuation metrics and real estate prices. In particular, this report investigates how real estate valuations collected from the Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013 are affected by 6 variables: the transaction date, the house age, the distance to the nearest Metro station, the number of convenience stores within walking distance, and the latitude & longitude coordinates. The second component of this report investigates how the concrete compressive strength is affected by 8 concrete components: cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age.

# Part I: Real Estate Valuation Data

## Introduction

The main purpose of this section is to understand what real estate valuation metrics influence price the most and to illustrate what their relationship with price entails. We completed a case study on house prices in the Sindian District in New Taipei City, Taiwan and how metrics such as transaction date, house age, distance to the nearest Metro station, number of convenience stores within walking distance, and the latitude & longitude coordinates affect the price of houses within the district. This was done by building a model with the most significant valuation metrics as predictors for price. We then strived to choose the best metrics as predictors and transformed them in order to achieve the best fit for predicting and estimating price.

For this data set, our main questions of interests are:

1. What real estate valuation metrics form the best model to predict and estimate price?
2. How does each historical market data affect the real estate valuation?
3. How does changing certain valuation metrics affect real estate valuations?

To answer our questions above, we used Multiple Linear Regression and various measures of linearity such as AIC, r-squared, t-tests, multiple scatter plots and residuals vs. fitted plots. Afterwards, we transformed the predictor and response variables using appropriate lambda derivation techniques such as the BoxCox plots and power transformations to ensure we chose the best transformation.

## Data Description

The market historical data set of real estate valuation are collected from Sindian District, New Taipei City, Taiwan during the period of June 2012 to May 2013. The original data set is available on the *UC Irvine Machine Learning Repository* and was donated by Dr. I-Cheng Yeh.

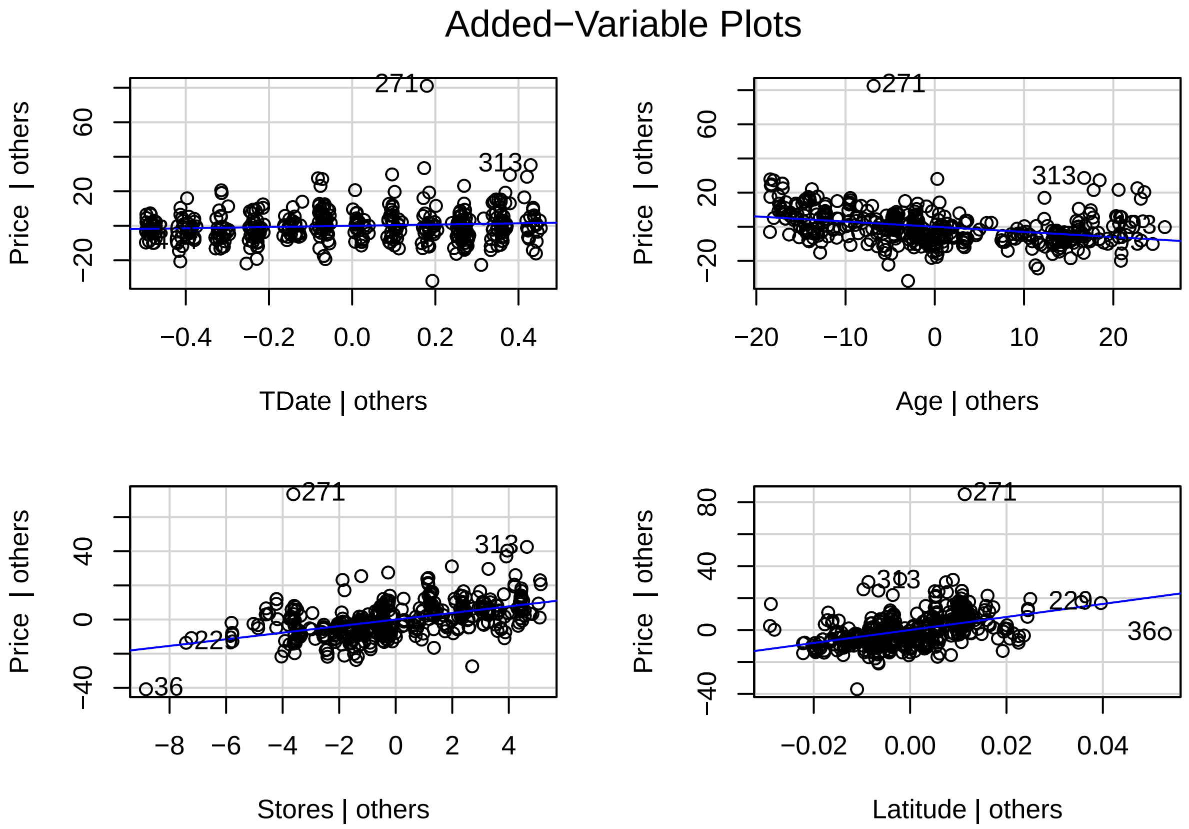
This data set serves to model the relationship between the real estate valuations and the 6 market historical data sets.

| **Variable Name** | Description |
|---|---|
| **TDate** | Transaction Date (e.g., 2013.250=2013 March, 2013.500=2013 June, etc.) |
| **Age** | House Age (unit: year) |
| **Metro** | Distance to the Nearest MRT Station (unit: meter) |
| **Stores** | Number of Convenience Stores in the Living Circle on Foot (integer) |
| **Latitude** | Geographic Coordinate (unit: degree) |
| **Longitude** | Geographic Coordinate (unit: degree) |
| **Price** | House Price of Unit Area (10000 New Taiwan Dollar/Ping, where Ping is a local 7 unit, 1 Ping = 3.3 $m^3$) |

## Regression Analysis, Results and Interpretation:

Before we dove deep into assessing how each of the predictor variables affect our response variable, the house price, we expected to see varying relationships among the response and each of the predictor variables. For instance, we predicted the house price to have a positive correlation with the transaction date since global real estate prices have been on the rise ever since the end of the Great Recession in June 2009. We also expected age to have a negative association with price because the Sindian district is an urban district so newer developments tend to be priced with a premium. Additionally, we expected the number of convenience stores within walking distance to have a positive correlation with the house price since people are typically more willing to pay more in exchange for convenience. As for latitude and house price, we expected an insignificant relationship due to the district being surrounded by a river at both latitudinal ends and because of waterfront property typically having much higher valuations.

```
## Loading required package: carData
```



From our exploratory analysis using the Added-Variable plots, we could see that when accounting for the other predictors: Age, Stores, and Latitude, Transaction date does not appear to be significant when included in our model. Age, Stores, and Latitude all appear to be useful.

### Global F-Test

The fitted regression line, extracted from the R-code below, is:

$$Price = -1.742e^4 + 3.613\text{TDate} - 3.020e^{-1}\text{Age} + 1.929\text{Stores} + 4.078e^2\text{Latitude}.$$

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude, data = REVals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.620  -5.601  -0.714   4.207  80.465
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.742e+04  3.524e+03  -4.944 1.12e-06 ***
## TDate        3.613e+00  1.686e+00   2.143   0.0327 *
## Age         -3.020e-01  4.178e-02  -7.227 2.44e-12 ***
## Stores       1.929e+00  1.801e-01  10.712  < 2e-16 ***
## Latitude     4.078e+02  4.278e+01   9.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.654 on 409 degrees of freedom
## Multiple R-squared:  0.5015, Adjusted R-squared:  0.4966
## F-statistic: 102.8 on 4 and 409 DF,  p-value: < 2.2e-16
```

Since the p-value for the Global F-Test is relatively small, we concluded that at least one of our predictors has a relationship with the price when taking into account the other predictors. As for marginal p-values, we could see that Age, Stores, and Latitude are significant with respect to their relationship with the price when considering the other predictors.

- If TDate, Stores, and Latitude are held constant, a one year increase in age will result in an estimated average decrease of $-3.020e^{-1}$ house price of unit area.
- If TDate, Age and Latitude are held constant, a one store increase in the number of stores will result in an estimated average increase of 1.929 house price of unit area.
- If TDate, Stores, and Age are held constant, a one-degree increase in latitude will result in an estimated average increase of $4.078e^{2}$ house price of unit area.

### Model Comparison after the Addition of Metro and Longitude

```r
fitMetro <- lm(Price~ TDate + Age + Stores + Latitude + Metro, data = REVals)
fitLongitude <- lm(Price~ TDate + Age + Stores + Latitude + Longitude, data = REVals)
fitboth <- lm(Price~ TDate + Age + Stores + Latitude + Metro + Longitude, data = REVals)
summary(fitboth)
```

```
##
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Metro +
##     Longitude, data = REVals)
##
## 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 *
## TDate        5.146e+00  1.557e+00   3.305  0.00103 **
## Age         -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## Stores       1.133e+00  1.882e-01   6.023 3.84e-09 ***
## Latitude     2.255e+02  4.457e+01   5.059 6.38e-07 ***
## Metro       -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## Longitude   -1.242e+01  4.858e+01  -0.256  0.79829
## ---
## 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
```

```
summary(fitMetro)$r.squared
```

```
## [1] 0.5823033
```

```
summary(fitLongitude)$r.squared
```

```
## [1] 0.5422915
```

```
#using AIC forward selection
m0<-lm(Price~TDate + Age + Stores + Latitude, data = REVals)
f = ~ TDate + Age + Stores + Latitude + Metro + Longitude
step(m0, f, direction = "forward")
```

```
## Start:  AIC=1882.36
## Price ~ TDate + Age + Stores + Latitude
## 
##              Df Sum of Sq   RSS    AIC
## + Metro      1    6181.8 31938 1811.1
## + Longitude  1    3122.5 34997 1849.0
## <none>                   38119 1882.4
## 
## Step:  AIC=1811.11
## Price ~ TDate + Age + Stores + Latitude + Metro
## 
##              Df Sum of Sq   RSS    AIC
## <none>                   31938 1811.1
## + Longitude  1    5.1308 31933 1813.0
```

```
## 
## Call:
## lm(formula = Price ~ TDate + Age + Stores + Latitude + Metro, 
##     data = REVals)
## 
## Coefficients:
## (Intercept)        TDate          Age       Stores     Latitude  
##  -1.596e+04    5.135e+00   -2.694e-01    1.136e+00    2.269e+02  
##       Metro  
##  -4.353e-03
```

The t-value for Longitude is -0.256. Since the p-value (0.7983) for Longitude is relatively large when Metro is included, we failed to reject the null hypothesis, i.e. longitude does not have a relationship with Price when also accounting for Metro in the model. From this, we concluded that including both Metro and Longitude is not useful in our model. However, since both predictors are significant on their own, we used the fact that the model with Metro included has a higher r-squared value (0.582) than the model with Longitude (0.542) to choose it as the better model. This was also confirmed by the AIC of the model with Metro being lower than the AIC of the model with Longitude when starting with the four original predictors followed by the model with Longitude not having a lower AIC than *none* when Metro is included in the model. This means that not adding any more predictors produces a better AIC than adding Longitude after Metro.

### Model Comparisons between the Addition of Metro and Addition of Stores

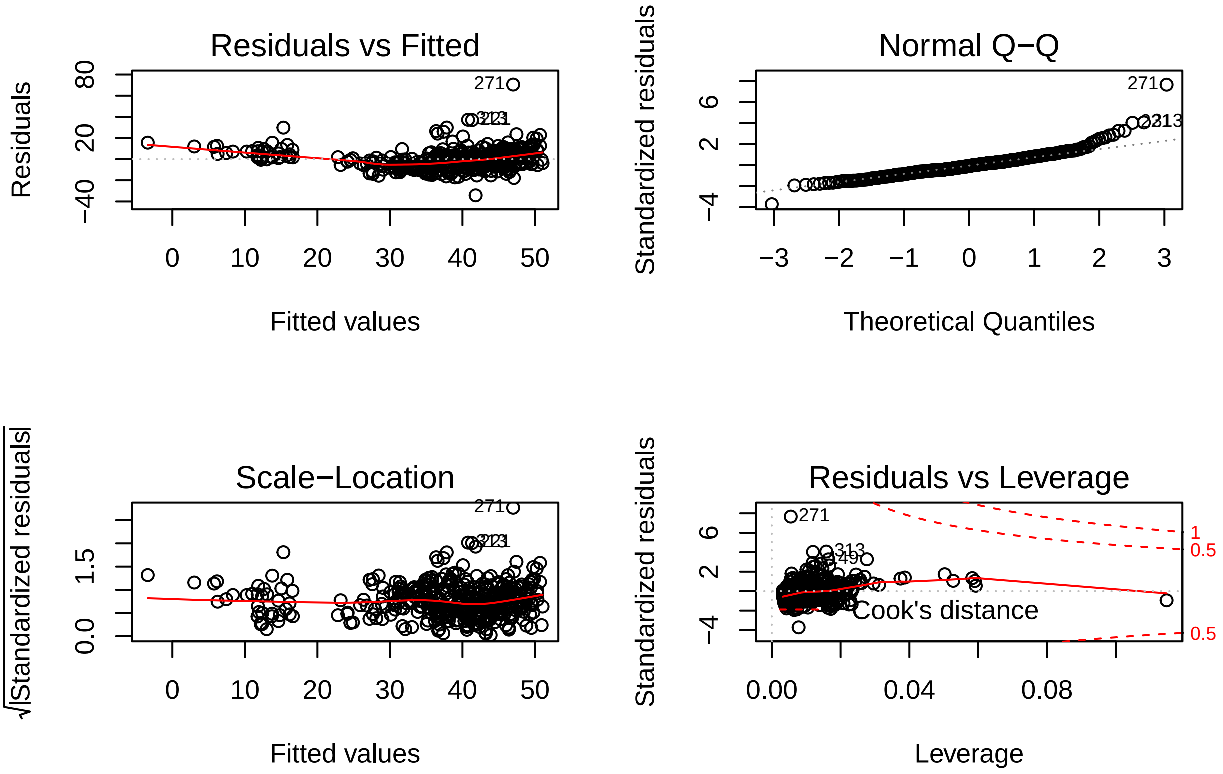
```
fitstores <- lm(Price ~ TDate + Age + Stores + Latitude, data = REVals)
fitmetro <- lm(Price ~ TDate + Age + Metro + Latitude, data = REVals)
summary(fitmetro)$r.squared # Model (1)
```

```
## [1] 0.5447599
```

```
summary(fitstores)$r.squared # Model (2)
```

```
## [1] 0.5014543
```

```
par(mfrow=c(2,2))
plot(fitmetro)
```



Between the models (1) and (2), we preferred model (2) since the addition of Metro yields a higher R-Squared value, 0.545. This means that 54.5% of the variability is explained by the model compared to the 50.15% of variability that is explained by the model using Stores. We used R-Squared as our deciding metric because (1) and (2) are not submodels of each other and contain the same number of predictors.

### Transforming Price and Metro

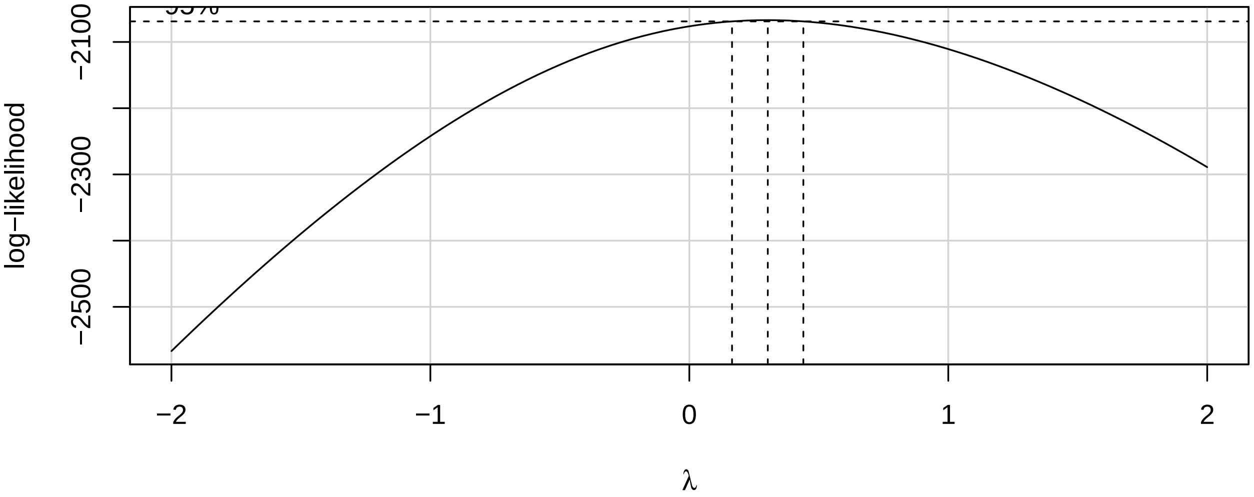
```
fitlog <- lm(Price ~ TDate + Age + log(Metro) + Latitude, data = REVals)
summary(fitlog)$r.squared
```

```
## [1] 0.6492648
```

```
summary(fitmetro)$r.squared
```

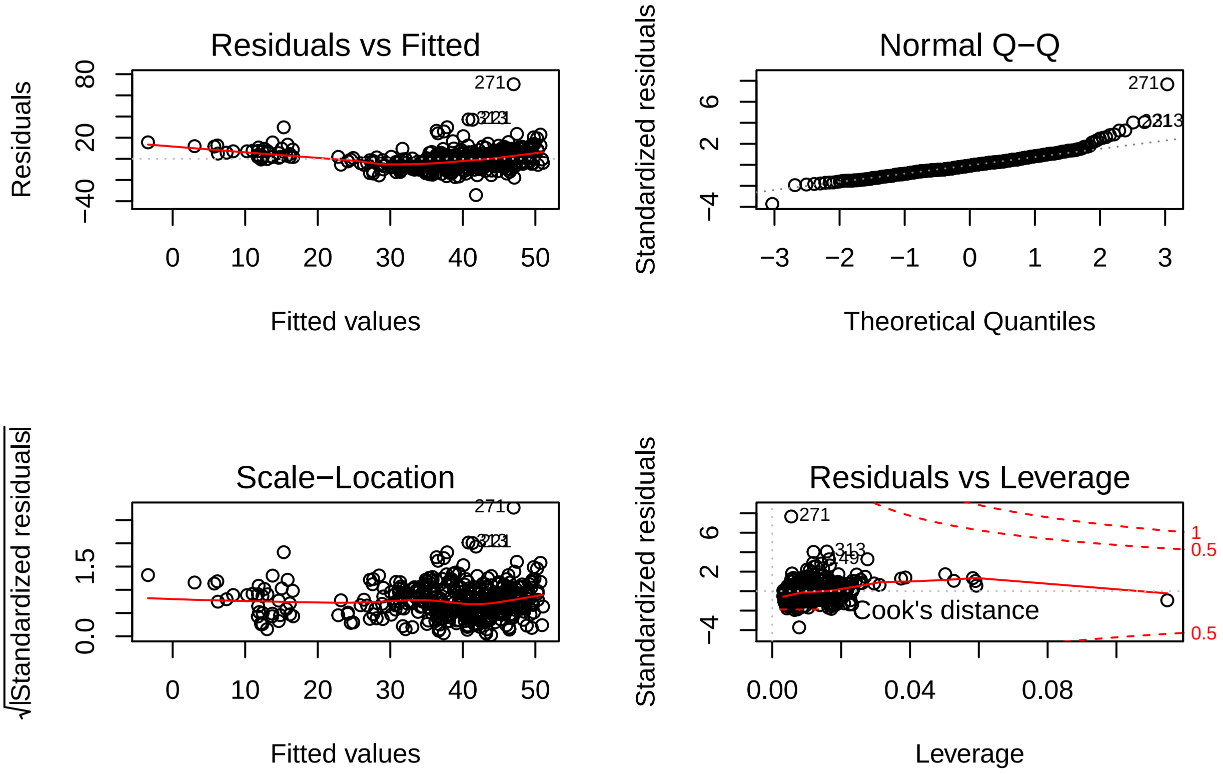
```
## [1] 0.5447599
```

As you can see, transforming Metro leads to a significantly better r-squared value, improving the linearity of our model. Next, we transformed the response. However, to determine which transformation should be applied, we used the Box-Cox transformation.

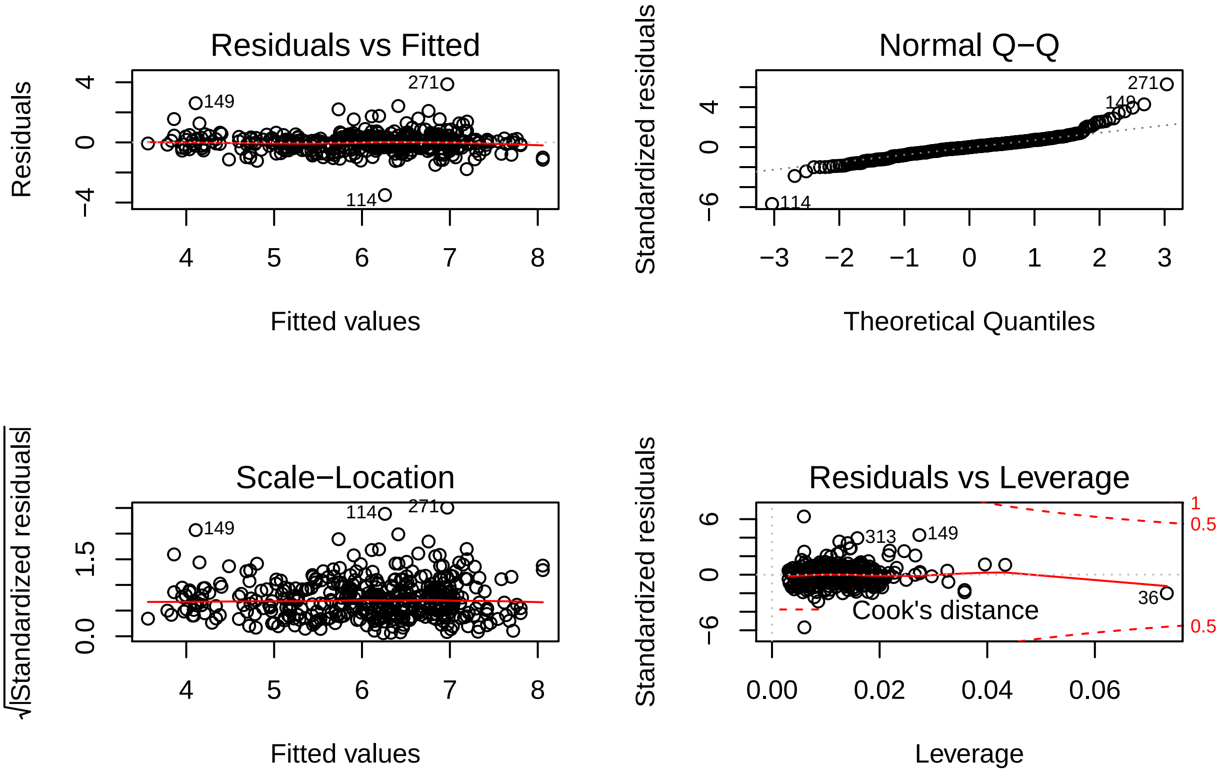


```
## bcPower Transformation to Normality
##              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## REVals$Price    0.5894         0.5       0.3888         0.79
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 33.41698  1 7.4372e-09
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 16.14636  1 5.8631e-05
```

Since $\lambda = 0.5$ is within the confidence interval provided by Box-Cox and Power Transformation methods, we used a square root transformation.



The variable **age**, the house age in years, and the variable **TDate**, the transaction date, are measures of time so a transformation will not be very useful for understanding our model. We tried a log transformation on metro since it is a measure of distance and cannot be less than or equal to 0. Metro also has a wide range, so a log transformation is likely to be appropriate. After doing a likelihood ratio test on Metro, we could see that the p-value after doing a log transformation is 0.63, which is relatively large compared to any standard significance level, allowing us to confidently transform Metro.



After transforming both price and metro, we could see significant improvement in linearity since the residuals vs fits plot looks more like a cloud of points with no clear pattern after the transformation. We could also see an improvement in r-squared (0.699) and constant variance based on a more even spread of points from left to right across the scale-location plot. Before the transformations, there was a skew of points to the right on the scale-location plot. As for normality, the Q-Q plots seem to stay relatively constant before and after the transformations.

## Conclusion

Starting with the first model using predictors TDate, Age, Stores and Latitude, we noticed that TDate was not as related to price as we had expected. However, after replacing Stores with Metro, TDate showed a significant relationship as we had initially expected. We expected this to be due to a correlation between TDate and Stores, most likely because as time passes (transaction date increases), more stores develop within the Sindian District, making TDate less useful when already considering the number of stores. Another aspect we thought was interesting was the significance of the relationship between price and latitude before and after changes were made to the model. Latitude showed a consistent positive relationship with real estate prices throughout the introduction of new predictors and the transformation of others. This is perhaps explained by how houses with higher latitudes are closer to the city center of New Taipei City and thus, typically closer to where the Yamsui River runs through the Sindian District. Since waterfront properties are typically deemed more valuable, the housing prices were much higher. Finally, the housing prices declines as the distance between the house and the nearest MRT Train Station increases. This gave us insight into the housing location preferences of the citizens living in the Sindian District are since their interest in the proximity of a nearby MRT station is reflected in the housing prices.

# Part II: Concrete Compressive Strength Data Set

## Introduction

High Performance Concrete (HPC) is a type of concrete designed to exceed the properties and constructability of conventional concrete. It is highly imperative towards the construction industry and is primarily used in bridges, tall buildings, and tunnels due to its strength, durability, and high elastic modulus. Due to the nature of its usage, HPC must be made with only the highest quality ingredients and carefully optimized mixture designs, thus making it an exceedingly complex material to model the behaviors of its components for strength performance. However, this report aims to utilize Multiple Linear Regression Analyses to model which component(s) of the 8 concrete components are directly correlated to the strength of HPC. Finding which ingredients are most influential towards the strength of HPC will provide great insights to HPC producers and in turn, ensure the safety of all constructions made with HPC.

For this data set, the two main questions of interest are:

1. Which concrete components are actively associated with the concrete compressive strength?
2. How does removing one or more of the 8 concrete components affect the concrete compressive strength?

To answer those questions above, we used the following regression analysis methods:

I. Forward Selection Algorithm using BIC
   - Diagnostic Checks:
     - Linear Regression Assumptions (Linearity, Normality, Constant Variance)
     - Leverage and Cooks' Distance for Influential Observations

II. Confidence Interval for Mean Response

III. Prediction Interval for Individual Response

IV. Backward Elimination Algorithm using BIC
   - Diagnostic Checks:
     - Linear Regression Assumptions (Linearity, Normality, Constant Variance)
     - Leverage and Cooks' Distance for Influential Observations.

## Data Description

The data set was collected for the paper, *Modeling of Strength of High Performance Concrete Using Artificial Neural Networks*, by Dr. I-Cheng Yeh. This data set is used to model the relationship between the concrete compressive strength and the 8 concrete components.

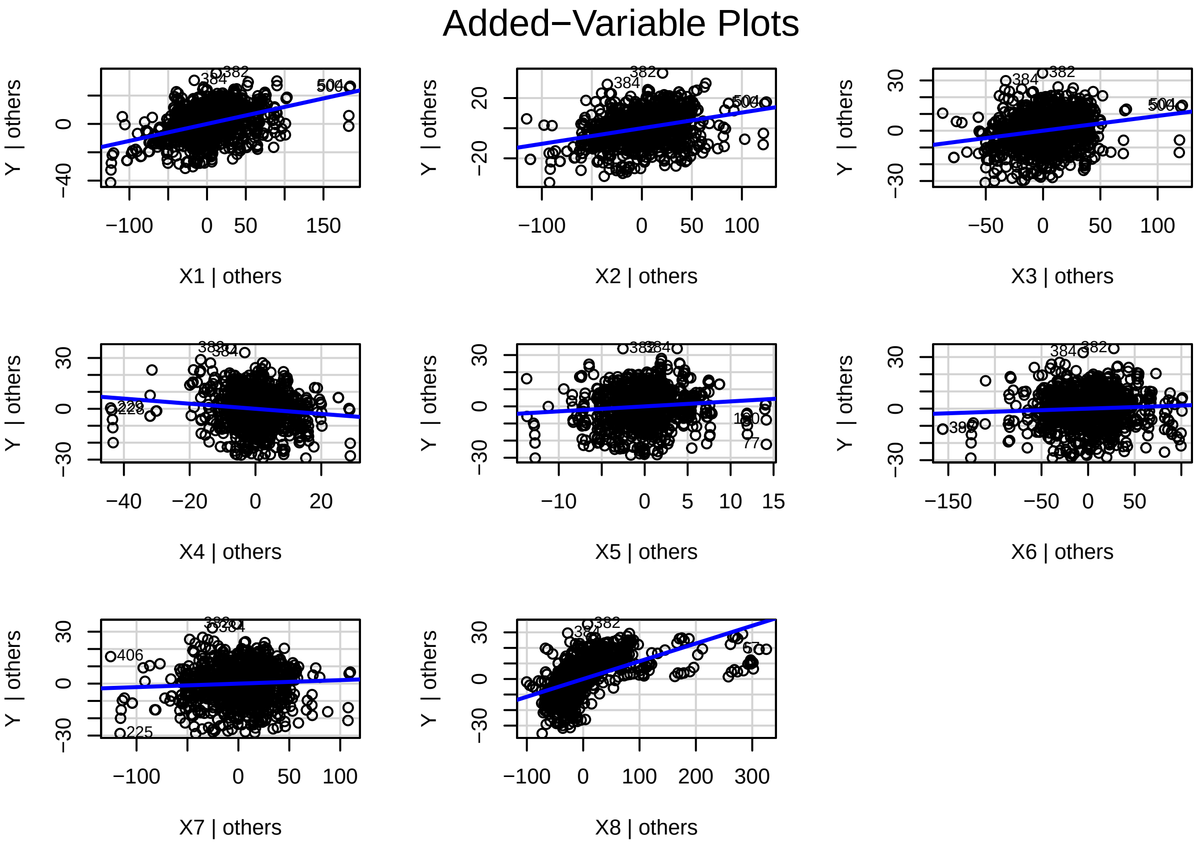
| **Variable Name** | Description |
|---|---|
| **X1** | Cement (component 1, unit $kg/m^3$) |
| **X2** | Blast Furnace Slag (component 2, unit $kg/m^3$) |
| **X3** | Fly Ash (component 3, unit $kg/m^3$) |
| **X4** | Water (component 4, unit $kg/m^3$) |
| **X5** | Superplasticizer (component 5, unit $kg/m^3$) |
| **X6** | Coarse Aggregate (component 6, unit $kg/m^3$) |
| **X7** | Fine Aggregate (component 7, unit $kg/m^3$) |
| **X8** | Age (Day, 1 ~ 36) |
| **Y** | Concrete compressive strength (MPa) |

## Regression Analysis, Results and Interpretation

Before we began with our regression analysis, we found the Added Variable Plots of the full model,

$$Y = \beta_0 + \beta_1\text{X1} + \beta_2\text{X2} + \beta_3\text{X3} - \beta_4\text{X4} + \beta_5\text{X5} + \beta_6\text{X6} + \beta_7\text{X7} + \beta_8\text{X8}$$

to assess the relationship between Y and $X_i$ after removing the effects of the other predictors (concrete components).



Looking at the AV-Plots, we could see that some predictors seem to have a significant relationship more than others. This is useful in seeing which concrete components are more directly associated with the concrete compressive strength; however, it doesn't tell us which combination of the concrete components would give us the best concrete compressive strength result. Thus to do so, we applied the Forward Model Selection using the Bayesian Information Criterion below and performed diagnostic check.

### Forward Selection Using BIC

```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = Concrete)
##
## Coefficients:
## (Intercept)           X1           X5           X8           X2
##    29.03022      0.10543      0.23900      0.11349      0.08649
##          X4           X3
##    -0.21829      0.06871
```

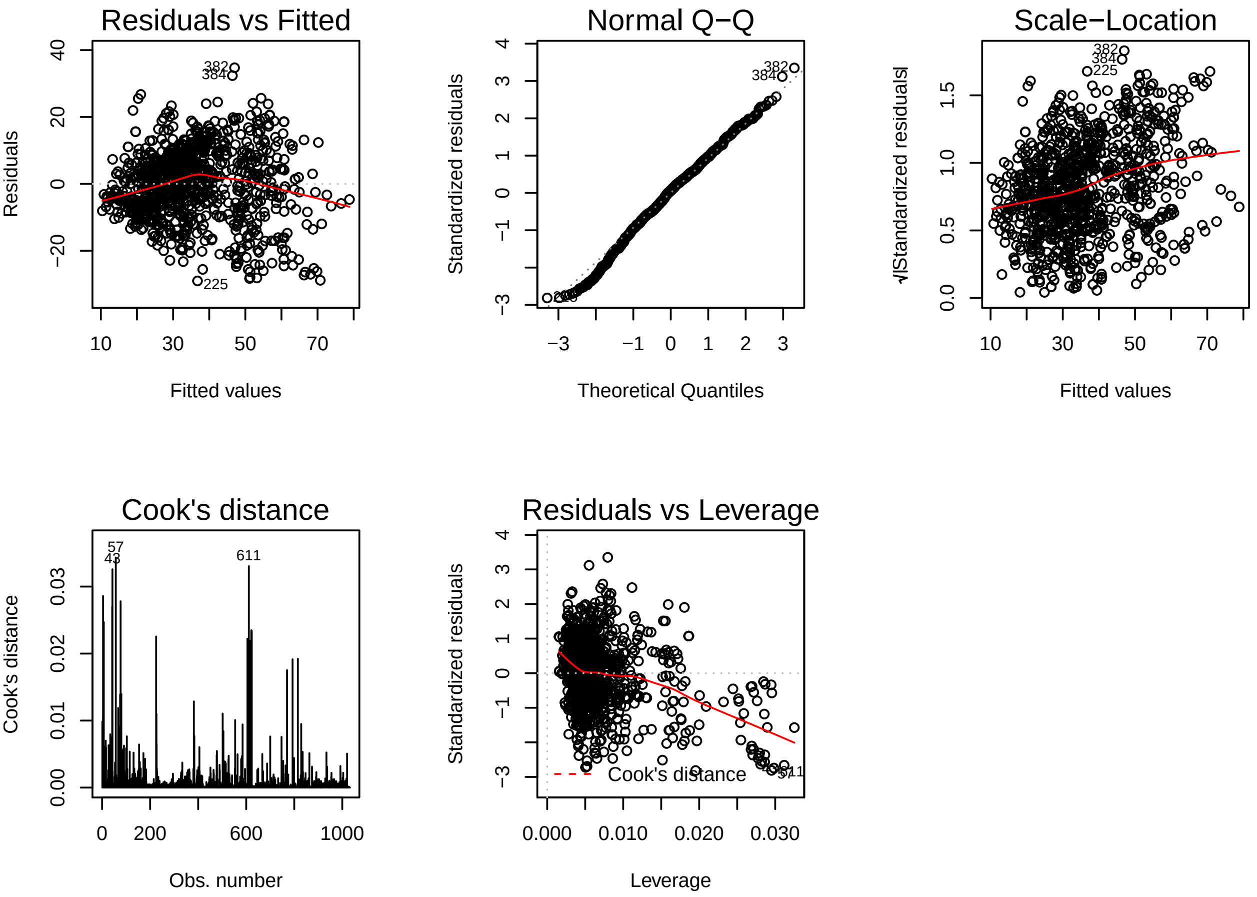
The results gave us a combination of

$$Y = 29.03 + 0.105X1 + 0.239X5 + 0.113X8 + 0.086X2 - 0.218X4 + 0.069X3.$$

However, before we could assume these predictors will produce the best combination, we must perform diagnostic checks for influential cases and linear regression assumptions (linearity, independence, normality, and constant variance).

### Diagnostic Checks for Linear Regression Assumptions

To perform diagnostic checks for linear regression assumptions, we plotted $Y \sim X1+X5+X8+X2+X4+X3$ using the Residual vs. Fitted plot to check the linear relationship assumption, Normal Q-Q plot to check if the residuals are normally distributed, Scale-Location plot to check for constant variance of residuals, and Residuals vs. Leverage & Cook's distance plots to identify influential cases.



Our diagnostic plots show that our fitted model satisfies all the model assumptions of linearity, normality and constant variance. According to the Cook's Distance plot, the most influential points are 611, 57, and 43. To explain why they're influential, we calculated the leverage and standardized residuals of each one.

### Influential Observations using Leverage and Cooks' Distance

```
##            ei        hii        ri         Di        ti
## 43  -27.27104 0.03118460 -2.661580 0.03257467 -2.669538
## 57  -28.82860 0.02951080 -2.811166 0.03432931 -2.820707
## 611 -28.12314 0.02982989 -2.742826 0.03304473 -2.751621
```

In general, we compared the leverage values, hii to $\frac{14}{1030}$ to identify high leverage points. Based on the leverage and standardized residual calculations of each observations above, we could see clearly why those data points are influential.

Removing influential points could potentially improve our model fit and linear regression assumptions. Thus, we tested below to see if whether removing these influential data points will make a difference or not.

**Summary of the Original Forward Fitted Model**

```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = Concrete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.014  -6.474   0.650   6.546  34.726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.030224   4.212476   6.891 9.64e-12 ***
## X1           0.105427   0.004248  24.821  < 2e-16 ***
## X5           0.239003   0.084586   2.826  0.00481 **
## X8           0.113495   0.005408  20.987  < 2e-16 ***
## X2           0.086494   0.004975  17.386  < 2e-16 ***
## X4          -0.218292   0.021128 -10.332  < 2e-16 ***
## X3           0.068708   0.007736   8.881  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 1023 degrees of freedom
## Multiple R-squared:  0.614,  Adjusted R-squared:  0.6117
## F-statistic: 271.2 on 6 and 1023 DF,  p-value: < 2.2e-16
```

**Summary of New Model without the Influential Points**

```
##
## Call:
## lm(formula = Y ~ X1 + X5 + X8 + X2 + X4 + X3, data = concrete.rmv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.413  -6.396   0.721   6.566  34.790
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.541262   4.180688   6.827 1.49e-11 ***
## X1           0.105495   0.004213  25.038  < 2e-16 ***
## X5           0.239689   0.083727   2.863  0.00429 **
## X8           0.120535   0.005542  21.751  < 2e-16 ***
## X2           0.086786   0.004937  17.580  < 2e-16 ***
## X4          -0.217171   0.020937 -10.373  < 2e-16 ***
## X3           0.068850   0.007666   8.981  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.3 on 1020 degrees of freedom
## Multiple R-squared:  0.6227, Adjusted R-squared:  0.6205
## F-statistic: 280.6 on 6 and 1020 DF,  p-value: < 2.2e-16
```

Comparing the two summaries above, there is not a signficant enough difference to remove the influential points. Therefore, we kept the original model to move forward.

### Confidence Interval for Mean Response

We predicted each concrete components to be:

```
##    X1 X5 X8 X2 X4 X3
## 1 380  4 28 95  2  0
```

```
predict(forward.fit, newdata = new, interval = "confidence", level = 0.95,
        type = "response")[1,]
```

```
##      fit      lwr      upr
## 81.00684 73.29586 88.71783
```

From our confidence interval calculation, the estimated average concrete compressive strength is 81.00684. Additionally, we are 95% confident that the actual average of the concrete compressive strength is in between the interval of (73.29586, 88.71783).

### Prediction Interval for Response

Using the same predictor values as for our Confidence Interval, we computed the 95% predictor for the individual response.

```
predict(forward.fit, newdata = new, interval = "prediction", level = 0.95,
        type = "response")[1,]
```

```
##       fit       lwr       upr
##  81.00684  59.17292 102.84077
```

From the results above, it appears that the predicted value for the concrete compressive strength given the values of the selected concrete components is 81.00684. We are 95% confident that the concrete compressive strength is in between the interval of (59.17292, 102.84077).

### Backward Elimination Using BIC

```
step(mod.full, scope = list(lower = mod.0, upper = mod.full), direction = 'backward',
     k = log(length(Concrete$X1)), trace = 0)
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X8, data = Concrete)
## 
## Coefficients:
## (Intercept)           X1           X2           X3           X4  
##    29.03022      0.10543      0.08649      0.06871     -0.21829  
##          X5           X8  
##     0.23900      0.11349
```

The backward elimination algorithm gave us the same combination of predictor variables; thus, it appears that X1, X2, X3, X4, X5, and X8 are the active predictor variables in determining the response variable.

### Diagnostic Checks

Since the backward elimination algorithm using BIC gave us the same model as the forward selection algorithm, the linear regression assumptions, influential points, and model quality will also be the same.

## Conclusion

Through our Regression Analysis process, we found that only 5 out of the 8 concrete components are strongly associated with the concrete compressive strength. Among these 5 concrete components, each of them, with the exception of the water (X4), have a positive linear relationship with the compressive strength of concrete. Water has a negative relationship with the compressive strength of concrete, which we did not find surprising. Interestingly, we found the age (X3) of the concrete to have a greater linear relationship with the concrete compressive strength than the concrete component (X1) has with the concrete compressive strength.

Although our findings strongly lean towards 5 out of the 8 concrete components being actively associated with the concrete compressive strength, we acknowledge that it may still lack reliability. Since we are not experts in the production of concrete, we may have left out possible interaction terms between the concrete components. For instance, the Age component (X8) may have been inversely related the Water content of the concrete (X4) due to evaporation over time. If that were the case, then our fitted model obtained from the forward & backward selections using BIC would not be the best model selection. However, disregarding the potential of a non-parallel model, we found our model to be best in displaying the linear relationship between the response and predictor variables since it satisfied all model assumptions.

# References

Kosmatka, S. H.; Kerkhoff, B.; and Panarese, W. C., *Design and Control of Concrete Mixtures*, 14th edition, Portland Cement Association, Skokie, IL, 2002, pp. 299-300.