**High Rise Building Cost Estimation**

The subset of data that will be the basis of our analysis has been taken from a 2005 research paper aimed at developing a simple cost estimation for high-rise buildings in Hong Kong. Three variables already proven to be relevant to cost estimation have been provided; average floor area, total floor area and storey height. The aim of this report is to statistically assess the individual importance of each of these variables on adjusted cost using regressional analysis, as well as determining whether combining multiple variables into one model improves the accuracy of the model.

**Question 1**

In figure 2, bivariate scatter plots have been plotted for every combination of variables, as well as histograms for each of the individual variables. This allows us to visualise potential correlations before a more thorough analysis of regression takes place in questions 2 and 3. As can be seen from the key in figure 2, blue bars and dots represent reinforced concrete, and orange represents steel.

There is an obvious positive correlation between average floor area and total floor area, which is stronger for reinforced concrete buildings and appears to become weaker as floor area increases. For both reinforced concrete and steel buildings, there is a strong positive correlation between average floor area and adjusted construction cost, and total floor area and adjusted construction cost. There’s a weak positive correlation between floor area and storey height, and total floor area and storey height, and in both cases the correlation appears to be stronger for reinforced concrete than steel. Upon first impressions, total and average floor areas appear to be the best individual predictors of adjusted construction cost.

**Question 2**

For both reinforced concrete and steel, individual bivariate regression models were created for average floor area, total floor area and storey height against adjusted construction cost. Least squares regression is a straight line through all points where the sum of the squares of the vertical distance between the line and each point is minimum. The equation of this line takes the form:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Matlab’s regress() function was used, with the two sets of data matrices as inputs, and and as outputs. These values were input into equation 1 to find an estimate for cost given each independent variable. From this, the residual (eqn. 2) and total (eqn. 3) sum of squares can be calculated:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |
|  |  | (3) |

Where is the actual adjusted construction cost, is the estimated construction cost based on the equation, and is the mean of all values. These are important results that allow the coefficient of determination, , and adjusted to be calculated using the equations:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |
|  |  | (5) |

Where is the number of data points and is the number of predictorsthat would be present in the multivariate equation.

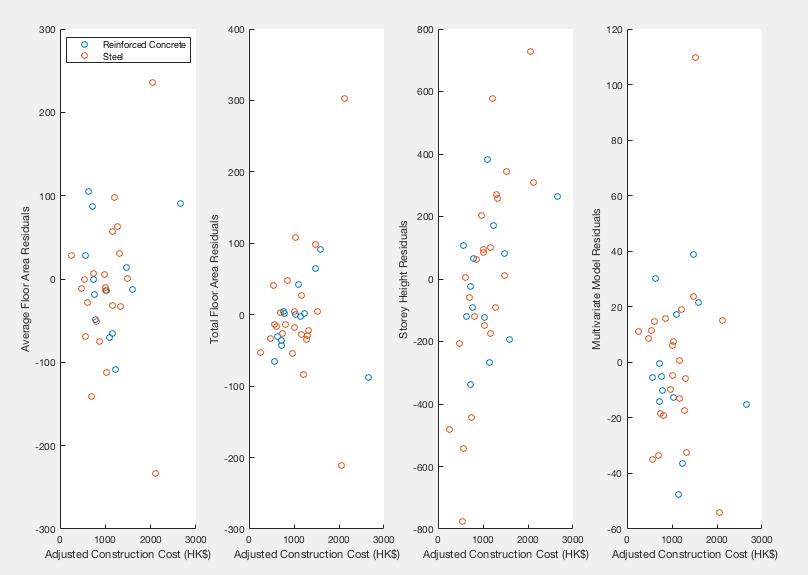
|  |  |  |
| --- | --- | --- |
|  |  |  |
| RC Avg. Floor Area | 0.9867 | 0.9843 |
| RC Tot. Floor Area | 0.9912 | 0.9896 |
| RC Storey Height | 0.8639 | 0.8391 |
| RC Multivariate | 0.9977 | 0.9973 |
| Steel Avg. Floor Area | 0.8391 | 0.9399 |
| Steel Tot. Floor Area | 0.9618 | 0.9580 |
| Steel Storey Height | 0.4220 | 0.3643 |
| Steel Multivariate | 0.9954 | 0.9949 |

is a statistical measure of how close the data are to the fitted regression line, which enables a simple assessment of correlation between two variables. However, this becomes limited when comparing multivariate models, as they tend to have a higher value simply because each new variable increases the value. adjusted takes this into account, and only increases in value if the added variable improves the predictive power of the model more than random chance. It always has a value lower than .

**Table 1**: and adjusted values for given variables against cost, where RC stands for reinforced concrete

It can be seen in table 1 that the individual variable with the highest value of against cost of construction is total floor area for both reinforced concrete and steel buildings. The range of values of for the three variables is much greater for steel, with the lowest value being 0.42 for the storey height model, compared to 0.86 for the same model applied to reinforced concrete. In steel buildings storey height appears to be much less correlated with cost than the other two variables, whereas all three variables are well correlated in reinforced concrete buildings. For , the multivariate equations containing all three variables cannot be directly compared with the bivariate values. However, , which enables a direct comparison, demonstrates that adding all three variables to the regression model vastly improves the accuracy of cost prediction. values of 0.9973 and 0.9949 are seen for the concrete and steel structures’ multivariate models, while the highest values for bivariate models are for total floor area, with values 0.9896 and 0.9580 respectively.

**Question 3**



**Figure 1**: Residual plots for the 8 regression cost models

A residual is the difference between the observed independent variable and the one predicted by a given regression model:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

In figure 1, the residuals of the 8 regression cost models analysed in this report have been plotted, with the dependent variable on the x-axis. In order for a regression model to be valid, points must have a random distribution. A healthy residual plot shouldn’t contain predictive information.

The three bivariate residual plots in figure 1 appear to be close to randomly distributed around zero for both reinforced concrete and steel, suggesting that a linear regression model would be appropriate in these cases. The average floor area and total floor area residuals are mostly evenly distributed to within 100 of the regression line, however in both cases the steel datasets contain significant outliers both above and below the main data cluster. These can affect the accuracy of the regression line by skewing it away from the main body of data, and can render the model almost useless in extreme cases. The residual plot for storey height contains by far the largest range of values, but has no significant outliers and the data is well distributed above and below the regression line. In this case a linear model is highly appropriate.

The multivariate residual plot for reinforced concrete is tightly and evenly distributed to within 50 of the regression line and contains no outliers, making it a good fit for our linear model. However, the multivariate plot for steel paints a very different picture. While the number of data points above and below zero is approximately equal, there is a greater spread of values to the negative that will skew the regression line. There is also an extreme positive outlier almost three times farther from the line than the next farthest point. which could also render a linear model inaccurate. These issues could be resolved by removing outlying data and potentially using a different transform, such as a log transform, for the skewed y-axis data. Alternatively a skewed y-axis may suggest that a variable is missing from the model.

**References**

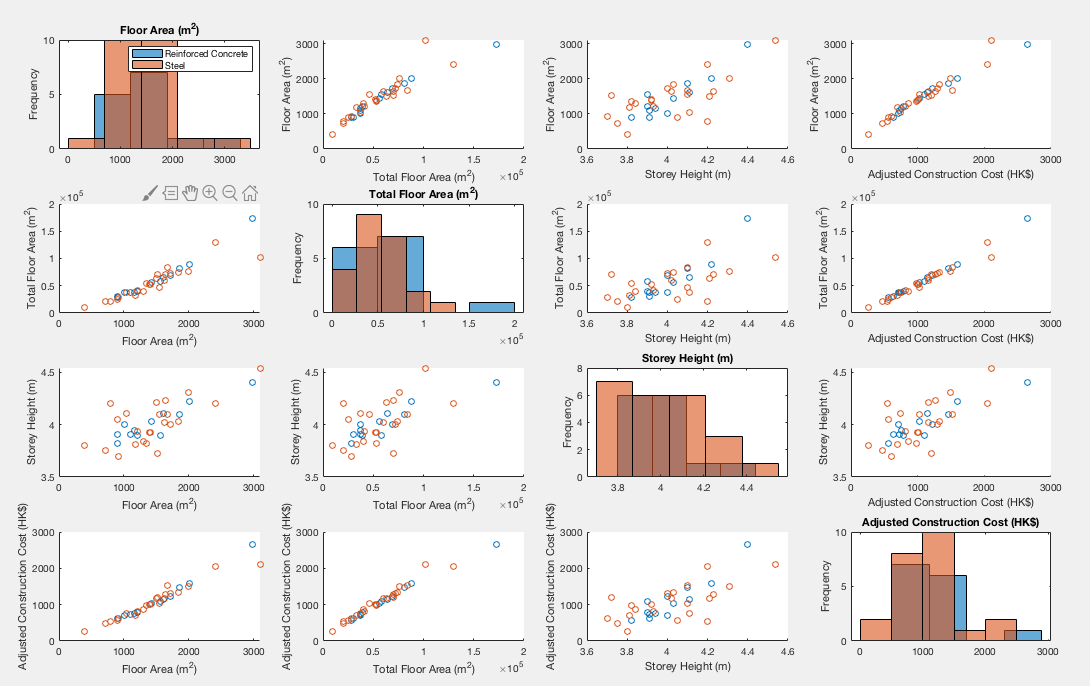
[1] Cameron, C. (2006). *Review of Bivariate Regression*. [online] Cameron.econ.ucdavis.edu. Available at: http://cameron.econ.ucdavis.edu/e240a/reviewbivariate.pdf [Accessed 15 Jan. 2019].

[2] Docs.statwing.com. (2019). *Interpreting residual plots to improve your regression*. [online] Available at: http://docs.statwing.com/interpreting-residual-plots-to-improve-your-regression/ [Accessed 15 Jan. 2019].

[3] Blog.minitab.com. (2019). *Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables*. [online] Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables [Accessed 14 Jan. 2019].

[4] Blog.minitab.com. (2019). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*. [online] Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit [Accessed 14 Jan. 2019].

[5] Blog.minitab.com. (2019). *Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err is Human, To Err Randomly is Statistically Divine*. [online] Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/why-you-need-to-check-your-residual-plots-for-regression-analysis [Accessed 16 Jan. 2019].



**Figure 2**: Histograms and scatter plots of all variable combinations for both reinforced concrete and steel buildings

**Demographic Electricity Consumption**

This smart meter data has been taken from two demographic user groups, Mosaic G and Mosaic I. These were obtained from the Customer Revolution Network smart meter trial as it recorded not only the energy consumption of each participant, but also broke them down into separate demographics based on household type. For each demographic, electricity consumption has been recorded for a 30 minute period beginning at 4 am and 10 am on the same day; Monday the 9th January 2012. The data will be statistically analysed using Matlab and conclusions drawn.

**Question 1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean (kWh) | Standard Deviation | Confidence level (kWh) |
| Mosaic G 4am | 0.214 | 0.184 | 0.214 +/- 0.022 |
| Mosaic G 10am | 0.468 | 0.600 | 0.468 +/- 0.071 |
| Mosaic I 4am | 0.181 | 0.176 | 0.181 +/- 0.011 |
| Mosaic I 10am | 0.483 | 0.546 | 0.483 +/- 0.036 |

**Table 1:** Mean, Standard Deviation and Confidence level for both demographics at 4am and 10am

**Mean**

The mean is calculated by dividing the sum of all energy consumptions in a particular data set by the number of elements in that set. This can be achieved in Matlab using the mean() function. From the data in Table 1 we can see that while Mosaic G has a higher energy consumption in the early morning, by mid morning Mosaic I were consuming more electricity on average. However, both saw a sharp increase by a factor of 2.2 and 2.7 respectively as people began to wake up and make use of appliances. Based on the mean there doesn’t appear to be a dramatic difference in energy consumptions between the two Mosaics.

**Standard Deviation**

Standard deviation is a measure of the variation in a set of values. It is calculated using the std() function in matlab, which implements the equation:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where

It can be seen that for both Mosaic G and I, standard deviation is approximately equal in the early morning, and in both cases is over three times higher at 10am than at 4am. This is likely because very few people are awake at 4am, and the only energy being consumed is by appliances running overnight. At 10am on a Monday some people will be at work and some, such as retirees or stay-at-home parents, might be at home, and in some homes washing machines or dishwashers may be running. Hence there is far more variation in potential energy consumption by mid morning. Mosaic I’s standard deviation at 10am is slightly less at 0.55 as opposed to 0.6, potentially suggesting that this demographic’s energy habits are more similar.

**Confidence Interval**

The confidence level is the range in which 95% of values lie. Since there are more than 30 values in the dataset, central limit theorem can be used. The matlab norminv(0.975) function returns the confidence level, which can be used to calculate the confidence interval:

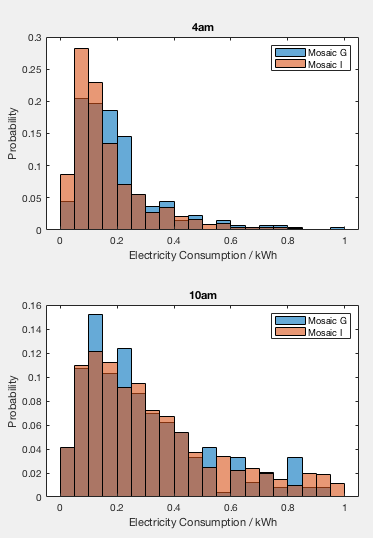
|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where = standard deviation and = confidence level

For both mosaics, the absolute confidence interval was much greater at 10am than 4am. For mosaic G the confidence interval a factor of 3.2 greater at 10am and 3.3 times greater for Mosaic I, suggesting a much greater variation in usage at 10am which is to be expected since the range of activity should be less at 4am as almost everyone is asleep.

**Question 2**

Density histograms were plotted in order to visually compare the energy consumption distributions of the 4 datasets. The data was reduced to only include energy use below one kilowatt hour, as a few extreme outliers would skew the x axis and make the main body of data less visible.

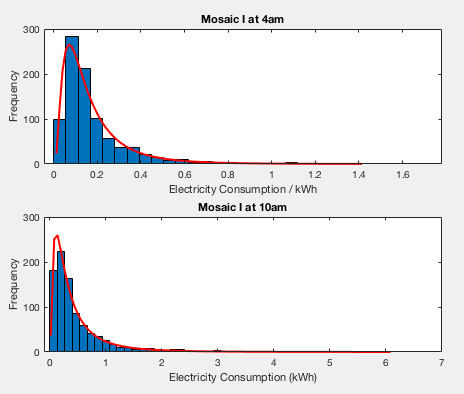
Mosaic G and Mosaic I were plotted on the same axes for 4am and 10am. The number of bins would normally be determined by taking the square root of n, but since we’re plotting multiple histograms on a single axis a standard number has been used to make comparisons more clear.

Distributions were similar for both mosaics. At 4am mosaic G has a slightly longer tail and appears to have a higher probability of energy consumption around 0.2kWh than Mosaic I, with a lower, longer distribution. This supports the findings in part 1 suggesting they have similar standard deviations, with Mosaic G having a higher mean electricity consumption.

Distributions for 10am were much lower and longer than the 4am distribution. Both mosaics peak at a similar electricity consumption than before, but have far thicker tails, with consumptions being much more varied than before. The distributions for the two Mosaics are very similar, however mosaic I has a greater proportion of higher consumption customers whereas G seems to have more at the lower end around 0.2kWh. Again these findings support the results from part 1, with very similar means, and Mosaic I having a larger standard deviation to account for its slightly more even spread of probability.

**Figure 1:** Histograms for the 4 data sets

**Question 3**



**Figure 2**: Histograms fitted with lognormal curves

In figure 2, the sample data for mosaic I at 4am and 10am has been represented as a histogram, with a lognormal model fitted over the top as a means of comparison. This was achieved using the histfit() function in matlab. As a first impression, the data appears to fit the model well in both cases, with the exception of the peak frequency value being slightly too low at 10am.

To find out whether the lognormal distribution could accurately be used to model and make assumptions these two datasets, the number of customers predicted by the model to use over 1 kilowatt hour of energy in the 30 minutes after 4am and 10am was compared to the actual number of people who used over 1 kilowatt hour in each case.

The predicted values were arrived at by using the lognfit() function in matlab which returns the mean and variance of the fitted lognormal distribution. These can be used to find the z values using the equation:

**Table 2**: Comparison between predicted number of customers using

over 1kWh vs actual number

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

|  |  |  |
| --- | --- | --- |
|  | 4am | 10am |
| Number of customers predicted to use more than 1kWh | 5 | 104 |
| Number of customers who actually used more than 1kWh | 7 | 106 |

With z, the number of customers with consumption above 1 can be found by using the normcdf() with z as the input variable, subtracting the answer from 1 and multiplying this by the number of customers in that dataset.

As can be seen in table 2, the predictions are impressively close to the true values. Both were 2 below the real value. With a larger sample size this prediction would likely become more accurate.

**Appendix**

[1] En.wikipedia.org. (2019). *Log-normal distribution*. [online] Available at: https://en.wikipedia.org/wiki/Log-normal\_distribution [Accessed 15 Jan. 2019].

[2] Mathsisfun.com. (2019). *Standard Deviation Formulas*. [online] Available at: https://www.mathsisfun.com/data/standard-deviation-formulas.html [Accessed 13 Jan. 2019].

[3] Stat.yale.edu. (1997). *Confidence Intervals*. [online] Available at: http://www.stat.yale.edu/Courses/1997-98/101/confint.htm [Accessed 13 Jan. 2019].