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:

$$y = \beta_0 + \beta_1 x \tag{1}$$

Matlab's regress() function was used, with the two sets of data matrices as inputs, and β_0 and β_1 as outputs. These values were input into equation 1 to find an estimate for cost y given each independent variable x. From this, the residual (eqn. 2) and total (eqn. 3) sum of squares can be calculated:

$$SS_E = \Sigma(y_i - \hat{y_i})^2 \tag{2}$$
$$SS_T = \Sigma(y_i - \bar{y_i})^2 \tag{3}$$

Where y_i is the actual adjusted construction cost, $\hat{y_i}$ is the estimated construction cost based on the equation, and $\bar{y_i}$ is the mean of all y_i values. These are important results that allow the coefficient of determination, R^2, and R^2 adjusted to be calculated using the equations:

$$R^2 = 1 - \frac{SS_E}{SS_T} \tag{4}$$

$$R^2_{adj} = 1 - \frac{SS_E/(n-1)}{SS_T/(n-p)} \tag{5}$$

Where n is the number of data points and p is the number of predictors that would be present in the multivariate equation.

R^2 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 R^2 value simply because each new variable increases the value. R^2 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 R^2.

Table 1: R^2 and R^2 adjusted values for given variables against cost, where RC stands for reinforced concrete

	R^2	R^2_{adj}
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

It can be seen in table 1 that the individual variable with the highest value of R^2 against cost of construction is total floor area for both reinforced concrete and steel buildings. The range of values of R^2 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 R^2, the multivariate equations containing all three variables cannot be directly compared with the bivariate values. However, R^2_{adj}, which enables a direct comparison, demonstrates that adding all three variables to the regression model vastly improves the accuracy of cost prediction. R^2_{adj} values of 0.9973 and 0.9949 are seen for the concrete and steel structures' multivariate models, while the highest R^2_{adj}values for bivariate models are for total floor area, with values 0.9896 and 0.9580 respectively.

Question 3

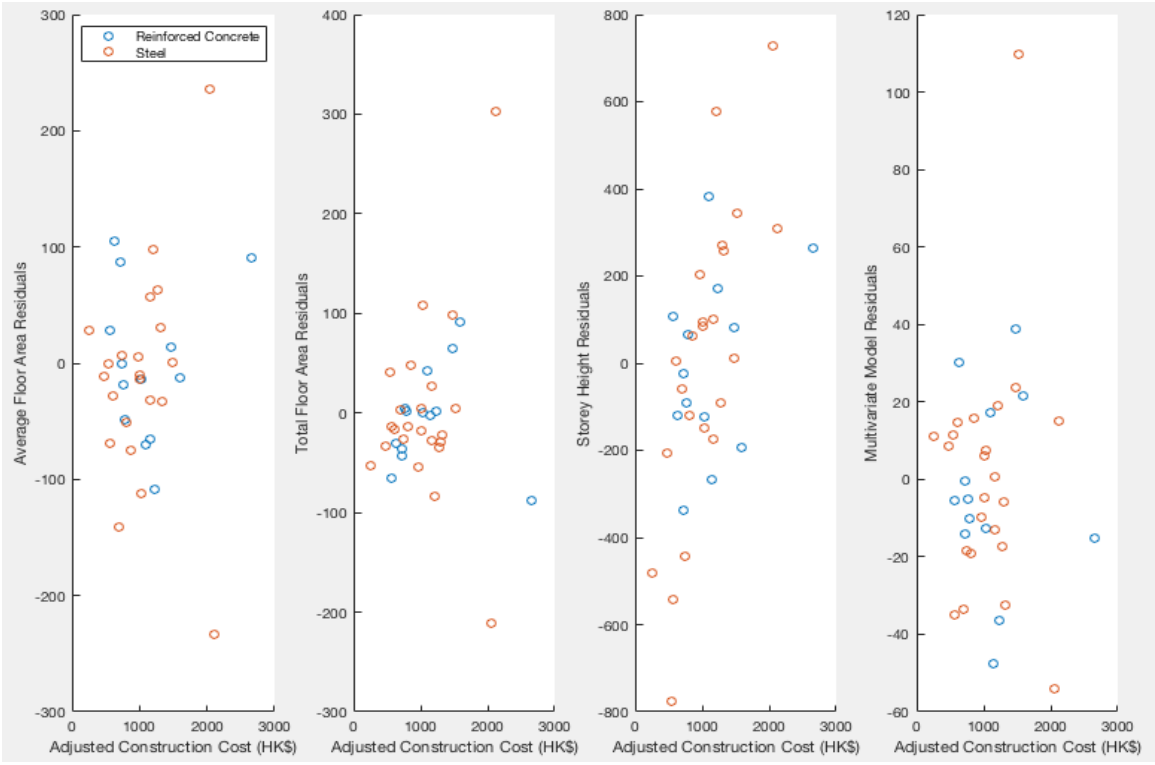

Figure 1: Residual plots for the 8 regression cost models

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

$$Residual, e = y_i - \hat{y}_i \tag{6}$$

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].

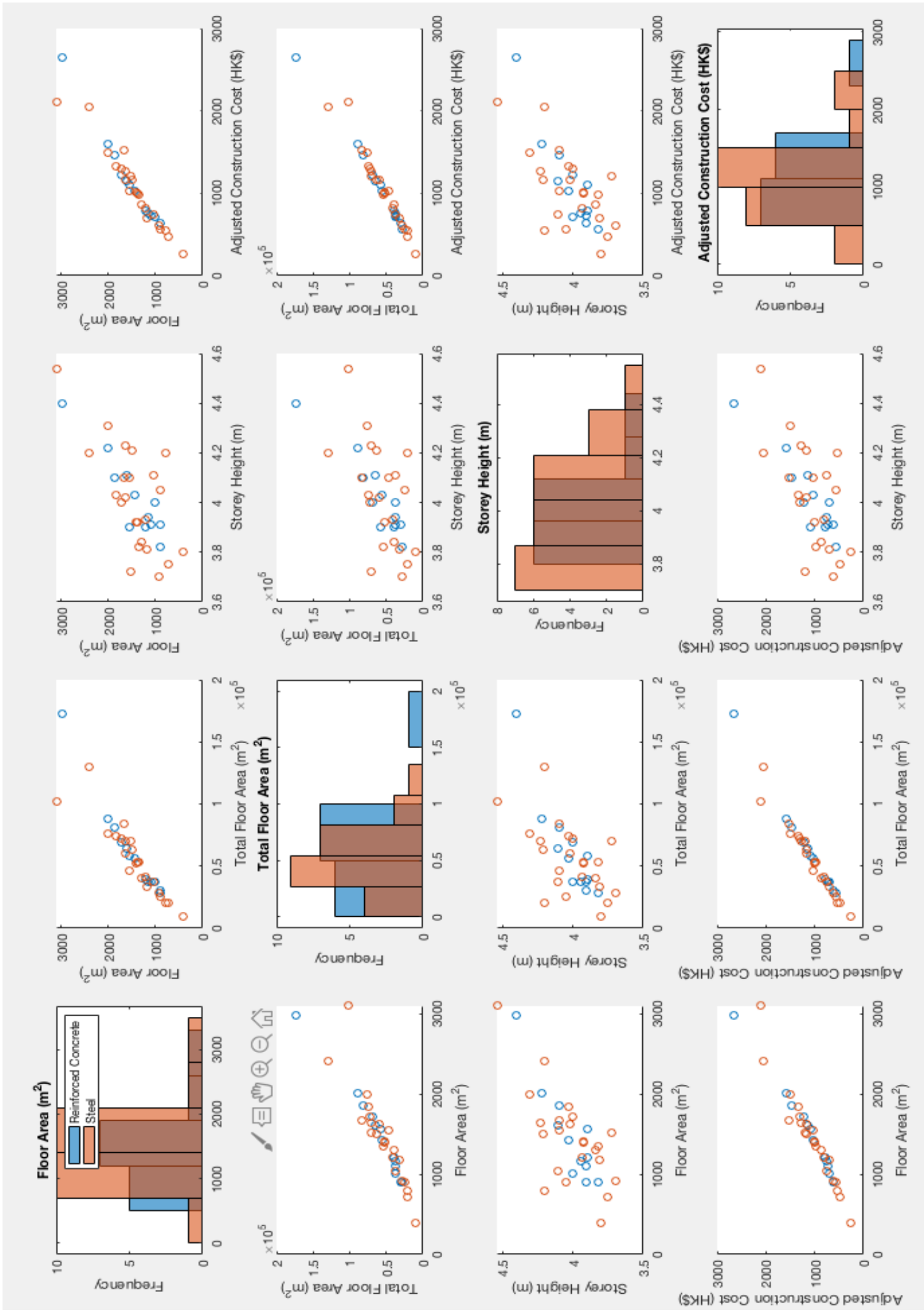

Figure 2: Histograms and scatter plots of all variable combinations for both reinforced concrete and steel buildings