Coursework Project

We aim to analyse the effect that various explanatory variables have on house prices and infer the price of an average house in both Framington and Natick, Boston, USA, from which our data is sampled. Our dataset contains 138 datapoints, where each point contains the price (in $), amount of rooms, bedrooms, bathrooms, shower rooms, garages and location of select houses. We first apply a multiple linear regression model, $Y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i} + b_5x_{5i} + b_6x_{6i} + e_i$ where i ranges from 1 to 138 and $e_i \sim N(0,\sigma^2)$ where the b_js correspond to the explanatory variables ordered as above and the e_is are identically independently distributed.

Checking of Residual Plots

We first consider the standardised residuals, $r_i \sim N(0,1)$ and check for departures from normality. *Fig.1* does not particularly exhibit skewness in our data, however it does suggest that there may be high kurtosis at higher prices. *Fig.*2 shows there may be heteroscedasticity within our residuals due to the "funnel" shape of our plotted values. We see this further in *Fig.3* as the trend exhibits our standard deviation increasing with expectation. Such heteroscedasticity is a problem as it violates our assumption that the residuals are spread constantly. As such, our explanatory variables may be less reliable for prediction. We now plot the log-likelihoods of transformations of our expectation against the Box-Cox transformation parameter with a 95% confidence interval. In *Fig.4*, we see that by rounding our transformation parameter with the highest likelihood to the nearest simple rational number (-½), we may want to transform Y_i to $1/\sqrt{Y_i}$ in order to better our adherence to normality. We will now consider this altered model and check its residual plots for any signs of heteroscedasticity. *Fig.5*, *Fig.6* and *Fig.7* show no signs of heteroscedasticity - our model now better fits our assumptions. We will now continue with this altered model.

Considering Subsets of Explanatory Variables

We will now consider more parsimonious models and see if any are better suited for inference and prediction. First, let us perform a backward elimination of our explanatory variables in order to consider the Akaike's Information Criterion (AIC) of our full model to more parsimonious ones. The AIC estimates the likelihood of a model to efficiently predict future values. We aim to choose a model with an AIC that is low/near to that of the full model. We see that our full model has an AIC of -2408.27, whereas the model that does not take into account the number of rooms has an AIC of -2410.27. This value is smaller but still similar to that of the full model. Let us consider also consider this roomless model for further checking. We will now analyse the coefficients of determination (R^2). The R^2 of both our full and roomless model is 70.41%. This value lies within a 70% threshold, lending rope to the sufficiency of both our models for prediction – we can say that 70.41% of the variation in our select house prices in Framington and Natick is due to the regression of our model(s). An R^2 closest to that of the full model is most desirable, which is satisfied by our reduced model. We next take into account the residual mean squares (MSE) for our models. A model with the lowest MSE is the most desirable as it indicates a higher accuracy of our explanatory variables. The MSEs (multiplied by 1000 for ease of comparison) of our two models are 0.1584 and 0.1578 for the full and roomless model respectively. We see that the roomless model has the lower MSE. We now check for multicollinearity between rooms and the other variables, in order to see whether our full model may be overly sensitive to small changes in the data and whether our parameter estimates may have large variances. Let us consider the partial F-tests of $x_j \mid x_1$, $j \in \{2, 3, 4, 5, 6\}$. We test $H_{0j} : b_j = 0$ against $H_{1j} : b_j \neq 0$ given that x_1 (rooms) is in our model. The only significant p-value here is 97.43% from the partial F-test containing x_2 (bedrooms). We would not reject this null hypothesis at any feasible significance level, meaning that, when our model only contains the number of rooms, also including the number of bedrooms does not improve it. From this, we can see multicollinearity between the number of rooms and bedrooms. We now calculate the Variance Inflation Factor of rooms and bedrooms in our full model. The VIF measures how much of the variance of our model is due to collinearity. This yields a value of 5.127 (3 d.p) and 3.588 (3 d.p). Our value for rooms is above a 5 threshold. Together with our partial F-tests, this sufficiently evidences multicollinearity in the number of rooms and bedrooms. Considering the VIFs of the variables in the roomless model, eliminating rooms significantly reduces any cause for concern in this respect, as the highest VIF here is 2.046 (3 d.p) for bathrooms. From our analysis of our two potential models, we have seen that our roomless model has an R^2 sufficiently equivalent to that of the maximum and a lower residual mean square and AIC than that of the full model. Paired with the existence of multicollinearity in the number of rooms and bedrooms, we are left with sufficient evidence to support dismissing our full model and continuing our analysis with our roomless model.

Outliers and Influential Points

From the standardised residuals of our roomless model, we have one possible outlier with a standardised residual value of 2.793 (3.d.p). This residual corresponds to the 19th observation with a fitted value of 0.0008423657 from an original of 0.001154778. This observation is the house in Framington with a price of $749,900 with 26 rooms, 11 bedrooms, 7 bathrooms, 2 shower rooms and 0 garages. We now calculate the Cook's distances for our observations,

in order to identify any points that have a significant influence on our model. Such points are problematic as they play a disproportionate role in determining our explanatory variables. Observation 19 has a Cook's distance of 1.275 (3 d.p). This value is significantly higher than that of all other observations. Since this value is greater than 1, this datapoint should be considered highly influential. Considering our dataset, we can see that this observation has a significantly higher amount of bedrooms and bathrooms than other observations around that price level. The leverage of this observation is 0.495 (3 d.p). This value is significantly large and lends support to this datapoint having a disproportionate effect on our model. We see that the 19th observation is both an outlier and an influential datapoint.

Exploratory Analysis

We now conduct t-tests to test hypotheses H_{0j}: $b_j = 0$ against H_{1j}: $b_j \neq 0$ for all explanatory variables for the model with and without the 19th observation. For both models, we have that the p-values for the t-tests for $b_3,\ldots,b_6$ are all 0.00…%, meaning that we reject the null hypotheses at any feasible significance level and conclude these variables are all correlated with price. However, considering b_2 (the bedrooms parameter) we have that, for our model with the outlier, our p-value is 12.5%, meaning that we wouldn't reject the null hypothesis at a significance level of 5%, 1% etc. but with our other model, our p-value is 0.861%, meaning that we would reject the null hypothesis at a 1% significance level. For the sake of reliable prediction, we will only consider our model without the outlier. Its parameter estimates are: -5.526443e-05, -1.139000e-04, -1.236128e-04, -1.033580e-04, -1.597765e-04 respective to our previously stated order of variables. From this, we see that, with respect to our sample, average house prices in Natick are higher than in Framington. Moreover, an increase in amount of garages lead to the highest increase in price, followed by bathrooms, shower rooms and, finally, bedrooms.

Prediction

We conclude by using our model to predict the average house price in both Framington and Natick using the dataset median values (i.e. 4 bedrooms, 2 bathrooms, 1 shower room, 2 garages) along with a 99% prediction interval and achieve, for Framington, a prediction of $504,742.90 with an interval of [995127.4, 304307.8] and, for Natick, a prediction of $642,225.20 with an interval of [1412926, 365364.5]. Contextually, we can say that a new datapoint with such median values has a 99% chance to lie within these intervals. Our prediction intervals here are significantly large since there are 25 datapoints that match these median values that range in price from $899,900 to $384,900. This could be due to nour model is missing other qualitative/quantitative information that could help to distinguish between these houses (e.g. quality of interior/exterior, whether the house has a balcony/garden etc.). Moreover, the interval for Natick is larger than Framington possibly since our dataset contains only 54 datapoints from Natick compared to 83 from Framington. Overall, we can infer that the price of an average house is higher in Natick than in Framington.

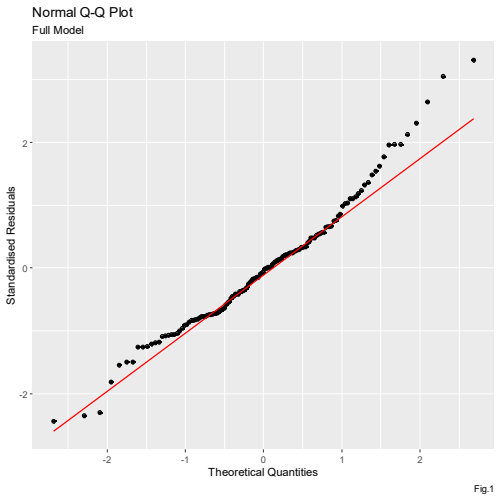

Fig.1

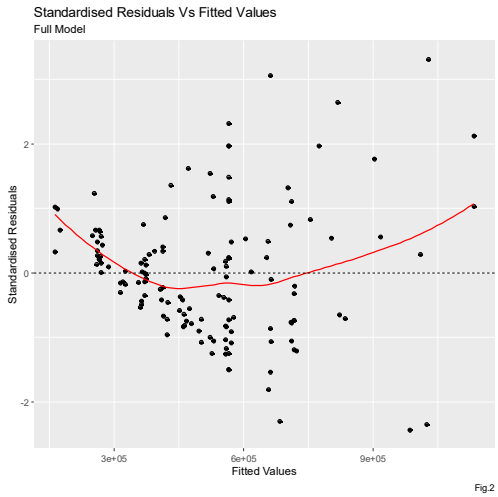

Fig.2

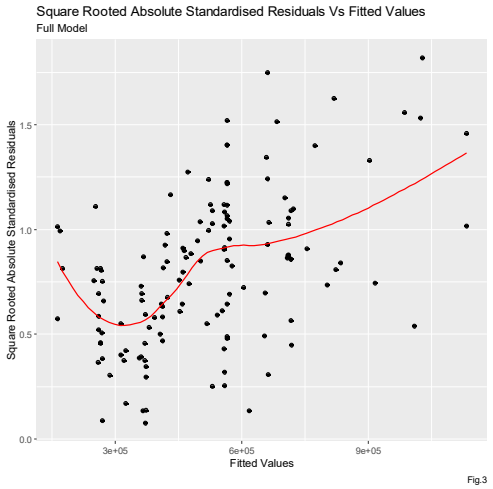

Fig.3

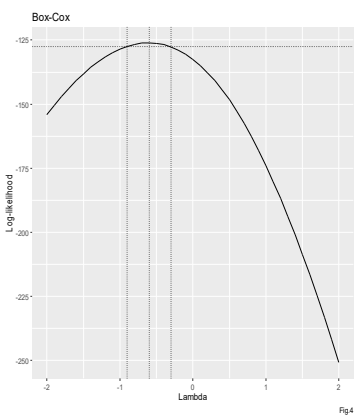

Fig4

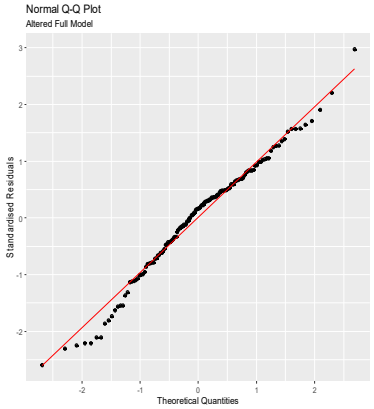

Fig.5

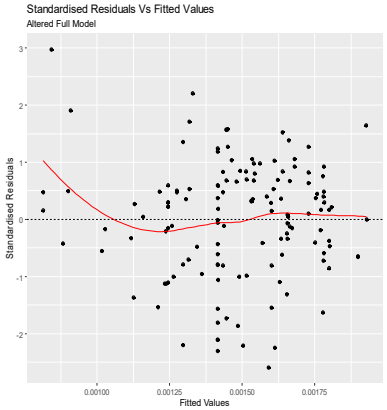

Fig.6

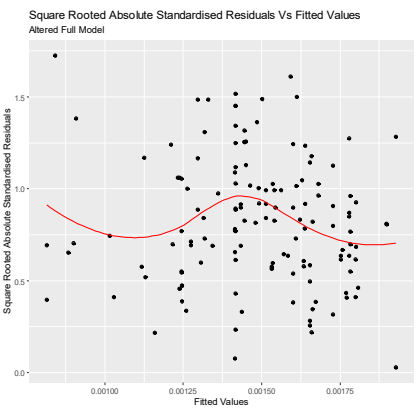

Fig.7