First Homework- Data Analysis Course of Professor Murthi

Meisam Hejazinia

1/27/2013

R-square of the correlation is 85.45 which shows that considerable percentage of the veriation is explained by these changes in demographics, such as house hold unit, multi family unit, vacant unit, small/ large unit, owner occupied, primary house usage of house, per capita income, and size of population of young-middle age versus old people changes. Adjusted R-square which takes into account the number of explanatory variables, showing whether adding the variables improves result just by chance or in reality, calculating in the form of $\bar{R}^2 = R^w - (1 - R^2)\frac{p}{n-p-1}$, where p is total number of regressors in linear model, and n is the sample size has only very small variation with R-square, and is 0.824, showing that all of the variables played considerable role in explanation, and the variation explanation is not just by chance. In summery adjusted R-square is taking care of anthropy, avoiding overfitting of model that would hinder model prediction ability. If we look for 95% confidence interval, multi family unit change, small house, vacant house change, primary gas user change, and the change in the hierarchy of population (percentage between 18-65 and percentage older than 65) played significant role in explaining the variation with p-values of 0.023, < 0.0001, < 0.0001, 0.013, 0.0001, and 0.0021 correspondingly. F-test works over total model, checking whether the variance is explained significantly or not, through checking whether the coefficients are not zero. Each of the t-tests checks whether each of the coefficients are zero or not. As a result the null hypothesis for the t-test here would be the coefficient would be zero for each of the coeffcients. Coefficient changes would be -1.26, 2.53, 2.68, -0.696, -0.33, -0.34 respectively for unit change, small house, vacant house change, primary gas user change, and the change in the hierarchy of population (percentage between 18-65 and percentage older than 65) in non-standard format, and as we use the standardized form by normalizing the variables using option STB, taking into account the interaction we would have standard coeffcients of -0.29 for multi family unit change, 0.42 for primary gas user, and $26 * 10^{-5}$ for the interaction of large homes change, and owner occupied homes.

Looking at this numbers, it seems the highest effect is by primary gas users. This was kind of clear, since as consumers do not use gas as their primary source of energy, alternative sources of energy will win the competition, resulting in reduction in gas usage, and increase consumer loss. F-value of the model is whithin 95% interval, and even lower than 0.0001 which shows model significance. The only problem could be collinearity. Variance inflation is high for housing unit , multi family unit , large homes unit , small homes unit , and vacant housing changes. There is high correlation between size of population between 18 and 65, change and size of population older than 65 change. Also vacant house and size of population between 18-65 have correlation, which I am not sure what means theoritically. Owners occupied house and per capita income also have correlation of 0.7, and all these collinearity may result in non unique answer for the estimates.

The interaction between large homes unit change and owner occupied homes change is calculated using their multiplication in the model statement of 'proc reg', and was significant with the correlation coefficient of $26 * 10^{-5}$. This says that people who have large houses, if would be home owner will

have different effect, mean there would cancel out each others effect, since the coefficient of the large homes is positive, yet if the person would be home owner, he or she would have lower churn, and all these can be translated to cancelling out, since their interaction has positive coefficient.