

Tutorial 3
Regression Analysis

Question 1

Fanfare International, Inc., designs, distributes, and markets ceiling fans and lighting fixtures. The company's product line includes 120 basic models of ceiling fans and 138 compatible fan light kits and table lamps. These products are marketed to over 1000 lighting showrooms and electrical wholesales that supply the remodeling and new construction markets. The product line is distributed by a sales organization of 58 independent sales representatives.

In the summer of 2013, Fanfare decided it needed to develop forecasts of future sales to help determine future salesforce needs, capital expenditures, and so on. The worksheet named Q1 contains the year and month the data being collected, monthly sales data, as well as three additional variables for the period July 2009 through April 2013. The variables are defined as follows

SALES = total monthly sales in thousands of dollars
ADEX = advertising expense in thousands of dollars
MTGRATE = mortgage rate for 30-year loans (%)
HSTARTS = housing starts in thousands of units

Questions:

- As a consultant to Fanfare, your job is to try each of the three variables in a simple linear regression, decide and justify which is the best variable to create a forecasting model for Fanfare.
- If you are allowed to develop a multiple regression model, decide which variables you should include in the equation and which should be omitted with justification. How well do you believe the developed model will do at forecasting future sales?
- What problems do you see with using each of the three possible variables to help forecast sales?

Solutions:

- From the t-test results of the three simple linear regressions, it is found that each of the three variables has significant linear relationship with SALES (p-value < 0.0001). Comparatively, HSTARTS seems to be the best single variable to forecast Fanfare's future sales, as HSTARTS explains more of the variation in SALES ($r^2 = 0.6843$).
- Based on the t-test results, ADEX would be judged unimportant as it has a p-value of 0.4602, which is much higher than the typical levels of significance (i.e. 0.05). Hence, it is suggested to remove ADEX from the regression model, and only include MTGRATE and HSTARTS. By doing so, the model has better predictive power, as its adjusted r^2 is slightly higher than the model with ADEX.

CB2203 Data-Driven Business Modeling

- c. Note that whichever variable is chosen, one problem will be determining future values of it. As without a given X -value, the model can never generate forecast of the Y -variable.

Question 2

The ABX Company is interested in conducting a study to identify factors that affect absenteeism among its production employees. After reviewing the literatures on absenteeism and interviewing several production supervisors and a number of employees, the researcher in charge of the project defined the variables shown as below

ABSENT = the number of distinct occasions that the worker was absent during the year
PAY = basic hourly pay rate in dollars
AGE = age of the employee
COMPLX = a measure of job complexity, ranging from 0 to 100
SENIOR = number of complete years with the company
DEPEND = employee response to the question "how many individuals other than yourself depend on you for most of their financial support?"

A sample of 77 employees was randomly selected, the data was contained in the worksheet named Q2. The dependent variable is absenteeism. The other variables are considered possible explanatory variables.

Identify the factors that may relate to absenteeism at a 5% level of significance. Check to see if your choice of variables and the estimated coefficients make intuitive sense. How much variation in absenteeism has been explained? What does this tell you? Does your model give you some sense of which employees might be absent most often? If so, which ones? What might be done to reduce absenteeism?

Solutions:

The initial model with all five possible explanatory variables indicating that PAY is the most insignificant variable as it has the largest p-value of 0.7885, thus it is chosen to be removed. A new regression model is run, and DEPEND is found to be insignificant at 5% level, as a result, DEPEND also being removed. The process continues until all the variables in the model as found to be significant at 5% level. Finally, the process stopped when only COMPLX and AGE are in the model.

The signs of the estimated coefficients suggest that absenteeism will decrease with increases in either job complexity or age. Management might consider encouraging employees working on more complex jobs, or the younger employees for further studies, so that they can handle more complex jobs in the future and improve their sense of achievements.

However, we have to note that only 17.85% of the variation in absenteeism has been explained by the final regression model, indicating that the model might not be good to provide prediction. Other factors should be identified so as to develop a better model for future prediction.