# STA371G Homework Assignment 6

**(50 Points. Group homework.) Please write down the NAME and EID of each group member. Each group consists of up to three members.**

## Problem 1: Housing Price Structure (10 points)

The file **MidCity.csv**, available in the course website, contains data on 128 recent sales of houses in a town. For each sale, the file shows the neighborhood in which the house is located, the number of offers made on the house, the square footage, whether the house is made out of brick, the number of bathrooms, the number of bedrooms, and the selling price. Neighborhoods 1 and 2 are more traditional whereas 3 is a more modern, newer and more prestigious part of town. Let $N_2$ be 1 if the house is in neighborhood 2 and be 0 otherwise, and let $N_3$ be 1 if the house is in neighborhood 3 and be 0 otherwise.

To estimate the house pricing structure in this town, we consider a regression model as

$$\begin{aligned} Y = \beta_0 + \beta_1\, Brick + \beta_2\, N_2 + \beta_3\, N_3 + \beta_4\, Bids \\ + \beta_5\, SqFt + \beta_6\, Bed + \beta_7 Bath + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

Consider, in particular, the following questions and be specific in your answers:

(a) Is there a premium for brick houses?

(b) Is there a premium for houses in neighborhood 3?

(c) For the purposes of prediction could you combine the neighborhoods 1 and 2 into a single "older" neighborhood?

(d) Is there an extra premium for brick houses in neighborhood 3?

(Hint: include $N_3 * Brick$ into your regression)

(e) Based on this model, explain the relationship between the selling price of a house and the number of offers made on the house. Does it make sense to use this model to predict the selling price of a house before putting it on the market? If no, provide a suggestion to improve the model.

## Problem 2 (10 points)

The data file **Profits.csv** is available in the course website. It contains information on 18 projects developed at a firm. The variables included in the file are:

- **Profit**: profit of the project in thousands of dollars
- **RD**: expenditure on research and development for the project in thousands of dollars
- **Risk**: a measure of risk assigned at the outset of the project

**(a)** Regress **Profit** on **RD** and **Risk**. Does there appear to be a relationship between **Profit** and **RD** after the risk of the project has been controlled for?

**(b)** Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

**(c)** Plot residuals versus **RD** and residuals versus **Risk**. Based on these plots suggest a correction for any violation detected in (b)? Try implementing your suggested correction. Does your new model appear to be an improvement over the original model? Justify your answer.

**(d)** What does your new model suggest is the expected change in **Profit** when **RD** changes? What is the expected change in **Profit** when **Risk** changes?

## Problem 3 (10 points)

The data file **AutoMPG.csv** is available in the course website. It contains information on **MPG** (miles per gallon) and **Weight** (in pounds) of 392 cars.

**(a)** Regress **MPG** on **Weight**. What does your model suggest is the expected change in **MPG** when **Weight** changes?

**(b)** Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

**(c)** Regress **log(MPG)** on **log(Weight)**. What does your model suggest is the expected change in **MPG** when **Weight** changes?

**(d)** Plot residuals versus fitted values. Do any of the regression assumptions appear to be violated? If yes, state which assumptions and justify your answer.

## Problem 4 (10 points)

Read the "Oakland A's (A)" case in the course pack. The data is available in the course website. The tab in the spreadsheet labeled Full Data Set contains the data in Exhibit 1 of the case while the tab labeled Nobel Data contains the attendance figures for the games Nobel pitched in and those he did not pitch in.

(a) Compute the descriptive statistics (sample means and sample standard deviations) for the attendance at the games Nobel pitched in and those he did not pitch in. What is the difference in the average attendance for these two sets of games? Does this provide meaningful evidence that Nobel should be paid more because attendance was higher in the games he pitched in?

(b) Plot Ticket against Time (i.e. create a time series plot of Ticket). Do you see any patterns in the data?

(c) Run the regression

$$Ticket_t = \beta_0 + \beta_1 Nobel_t + \epsilon_t$$

where $Nobel$ is a dummy variable that takes the value 1 when Nobel starts on day $t$.

What are the estimates of $\beta_0$ and $\beta_1$? How do these relate to the average attendance figures computed in part (a)?

(d) Do the residuals from the regression in part (c) appear to be independent? Why or why not? If they are not independent, what factors might explain the pattern?

(e) Run the regression

$$Ticket_t = \beta_0 + \beta_1 Pos_t + \beta_2 GB_t + \beta_3 Temp_t + \beta_4 Prec_t + \beta_5 TOG_t + \beta_6 TV_t + \\ + \beta_7 Promo_t + \beta_8 Nobel_t + \beta_9 Yanks_t + \beta_{10} Weekend_t + \beta_{11} OD_t + \beta_{12} DH_t + \epsilon_t$$

Do the residuals from this regression appear to be independent? (It is a close call but assume they are independent.) Why would these residuals be independent while the residuals from the model in part (c) are dependent?

(f) What evidence is there about Nobel pitching in a game being related to the attendance at the game? Do you have more confidence in drawing a conclusion from the model in part (c) or the model in part (e) to answer this question? Why?

(g) Do you think Nobel's agent has a legitimate case that Nobel should be paid more because he brings fans to the games?

## Problem 5 (10 points)

Read the "Oakland A's (B)" case in the course packet. The data file is available in the course website.

(a) Run a regression of Attendance against Wins. What is the interpretation of the coefficient associated with Wins? What is the interpretation of $R^2$ in this regression? What is the practical problem associated with using this model to forecast Attendance for the next season (i.e. to forecast attendance in the 1981 season)?

(b) Now run a regression of Attendance against Roddey's forecast of the number of wins for that season. Why is the $R^2$ value obtained from this regression so much lower than the $R^2$ obtained from the regression in part (a)?

(c) Why is it more appropriate to use the model in part (b) for forecasting Attendance than the model in part (a)?

(d) Before the 1981 season starts Roddey forecasts 95 wins for the season. Using the model from part (b), what is the prediction for attendance in the 1981 season? What is the standard deviation associated with the prediction?

(e) Using the prediction and standard deviation for the prediction from the model in part (b), what is the probability associated with a bonus to Nobel of \$0, \$50,000, \$100,000 and \$150,000? What is the mean of this distribution?

(f) Using the probability distribution from part (e), what is the expected cost if the lump-sum incentive plan is used?