STA9750 – Homework 2 [Due 10/25]

All the datasets can be found in “hw2data.zip”

Exercise 1

The dataset “college.txt” has data on new matriculants in a college. It has 3 variables: financial aid given by the college (gift_aid), family income (family_income), and how much they actually paid in tuition (price_paid) [all in thousands of $].

1) Summarize the data: find statistics such as means and standard deviations, create univariate plots that describe the data well, and create plots that describe the bivariate (2D) relations between pairs of variables. Interpret the results, and identify key features, such as potential outliers.
2) Find all the pairs of correlations between the variables in the samples. Which variables are most correlated? Interpret the strength and signs of the correlations in context.
3) Run a regression where the dependent variable (“y”) is financial aid and the independent variable (i.e. the covariate, predictor, “x”) is family income.
    a. Assess model fit: look at residuals, influential observations, high leverage points, etc. Is the model adequate?
    b. Assuming the model is adequate, does financial aid significantly depend on family income? Use a statistical test to answer the question.
    c. Find a 95% interval to predict the financial aid awarded to 3 new students with family incomes equal to $30k, $70k, and $500k, respectively. Comment on the width on the intervals and their adequacy.
    d. Find a 99% interval for the financial aid needed for the population of households whose incomes are $30k, $70k, and $500k. Again, comment on the width and the adequacy of the intervals, and compare them to the ones you found in part c.
4) Run a regression where the dependent variable (“y”) is financial aid and the independent variable (i.e. the covariate, predictor, “x”) is price_paid.
    a. Assess model fit: look at residuals, influential observations, high leverage points, etc. Is the model adequate?
    b. Assuming the model is adequate, does financial aid significantly depend price_paid? Use a statistical test to answer the question.
    c. Find a 95% interval to predict the financial aid awarded to somebody who paid $0 in tuition. Comment on the results.
5) Compare the model fit of the models you ran in parts 3) and 4). Which model do you think is most likely to be useful to the college in a real scenario?

Exercise 2

The dataset “speed_gender_height” has data from a survey. It has 3 columns: the maximum speed at which the participants have driven (speed), gender, and height. A group of social scientists is interested in knowing whether the maximum speed at which one has driven depends upon gender and/or height.

1) Summarize the data: find statistics such as means and standard deviations, create univariate plots that describe the data well, and create plots that describe the bivariate (2D) relations between pairs of variables. Interpret the results, and identify key features, such as potential outliers.
2) Find a useful visualization for the social scientists which synthesizes the relationship between the 3 variables in the model.
3) Run a regression where the dependent variable is speed and the independent variable is height. How well does the model fit the data? Interpret the sign and magnitude of the regression coefficient you find, and find a 95% interval for predicting the fastest speed at which LeBron James and Kristaps Porzingis have driven.
4) A social scientist is thinking about writing a paper whose title is “NBA players shouldn’t drive.” Would you encourage or discourage the scientist to write the article? Explain your answer.
5) Describe the how “speed” depends upon height and gender using regression analysis, after finding the best model you can. For the model you chose, identify outliers, influential observations and observations with high leverage (if there are any).
6) Does the effect of height depend on gender? If so, by how much and in which direction?
7) An agency is interested in knowing what the average fastest speed is for the population of 5’6” women. Find a 95% interval for that quantity.
8) [Optional] Find a 95% interval for your predicting fastest driving speed (given your gender and height). Does it contain your actual fastest driving speed?

Exercise 3

The dataset "gpa" has 5 columns which have data on a survey on some college students. The columns are their GPA, how many hours they study each week, the number of nights that they go out, how many hours they sleep per night and their gender.

1) Summarize the data: find statistics such as means and standard deviations, create univariate plots that describe the data well, and create plots that describe the bivariate (2D) relations between pairs of variables. Interpret the results, and identify key features, such as potential outliers.
2) Find the best model that you can to predict GPA, given the variables that you have. Explain how you reached that model and provide diagnostics. Identify any remaining outliers, influential observations or high leverage points and interpret them in context.
3) Given the best model you have, describe the relationship between GPA and the variables that are in the model; that is, explain which variables seem important (and which don't), and their effects.
4) A female student who goes out once a week, sleeps 6 hours every night, and studies 20 hours a week wants to boost her GPA. She's considering 3 strategies:
    a. Don't ever go out, study 5 more hours each week, sleep 6 hours
    b. Go out once a week, study 5 more hours each week, sleep 2 more hours
    c. Go out twice a week, study 20 hours more, sleep 6 hours
    d. Go out twice a week, study 10 hours more, and sleep 2 hours more
    e. Don't change her strategy

    Which one should she choose? Justify your answer.
5) Find a 95% predictive interval for the GPA of a male student who studies 15 hours a week, sleeps 7 hours each day, and goes out 5 times a week.
6) [Optional] Find a 95% predictive interval for your undergraduate GPA, given "your" values of the covariates. Does it capture the true GPA?

Exercise 4

The dataset "starbucks.csv" has nutritional information on some items that are available at Starbucks.

1) Summarize the data: find statistics such as means and standard deviations, create univariate plots that describe the data well, and create plots that describe the bivariate (2D) relations between pairs of variables. Interpret the results, and identify key features, such as potential outliers.
2) Run a model where the dependent variable is calories and the independent variable is carbs. Is the model adequate? Identify outliers, influential observations, etc.
3) Find the best model you can to predict calories given carbs. Interpret the resulting coefficient of "carbs," and predict the calorie amount of a new item that has 10g of carbs.
4) Find the best model you can to predict calories given all the information you have in the sample. Justify your choices explaining what you did in words. Please, don't submit all the output you found after trying out different models. Simply describe what you saw, and add at most 2 plots of models that "didn't work." Provide plots with diagnostics for the model you chose.
5) With the model you found in part 4), find 95% predictive intervals for the calories of these McDonald's items:

Item	Fat	Carb	Fiber	Protein	Type
Big Mac	28	45	3	24	sandwich
Cheeseburger	11	33	2	15	sandwich
McDouble	17	34	2	21	sandwich
Chocolate Chip Cookie	7	22	1	2	bakery

6) Compare your results with the actual calories of the items, which are 520, 290, 370, and 160 kcal. Are you satisfied with how well your model performs?