Statistics Final Exam Spring 2016

Name: ____________________________ Department: ________________________

Student ID: ________________________

Data 在壓縮檔。如有無法自動計算的問題，至公式→計算選項→手動改成自動

1. (34%) Psychologists are interested in how much influences does the media, especially reality TV programs, have one's decision to undergo cosmetic surgery. In a study, 170 college students answered question about their impressions of reality TV shows featuring cosmetic surgery, level of self-esteem, satisfaction with one's own body, and desire to have cosmetic surgery to alter one's body. The variations analyzed in the study were measured as follows:

DESIRE: scale ranging from 5 to 25, where the higher the value, the greater the interest in having cosmetic surgery

Gender: 1 if male; 0 if female

SELFESTM: scale ranging from 4 to 40, where the higher the value, the greater level of self-esteem

BODYSAT: scale ranging from 1 to 9, where the higher the value, the greater satisfaction with one's own body

IMPREAL: scale ranging from 1 to 7, where the higher the value, the more one believes reality TV shows featuring cosmetic surgery are realistic

The data of study are saved in the file of **BDYIMG**. The psychologists used multiple regression to model desire to have cosmetic surgery (y) as a function of gender (x_1), self-esteem (x_2), body satisfaction (x_3), and impression of reality TV (x_4).

Part I: first-order model ($E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$)

a. Fit the first-order regression model to the data using the method of least square. (3%)

b. Interpret the β estimates in the words of the problem. (4%)

c. Is the overall model statistically useful for predicting desire to have cosmetic surgery? Using $\alpha = 0.01$. (3%)

d. Which statistic, R^2 or R^2-adj, is the preferred measure of model fit? Practically interpret this statistic. (2%)

e. Conduct a test to determine whether desire to have cosmetic surgery decreases linearly as level of body satisfaction increases. Use $\alpha = 0.05$. (2%)

Part II: Interaction model ($E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_4 + \varepsilon$)

f. Fit the interaction model to the data using the method of least square. (3%)

g. Find the predicted level of desire for a male college student with an impression-of-reality-TV-scale score of 5. (2%)

h. Conduct a test of overall model adequacy. Use $\alpha = 0.10$. (3%)

i. Conduct a test at $\alpha = 0.10$ to determine if gender and impression of reality TV show interact in the prediction of level of desire for cosmetic surgery. (2%)

j. Find an estimate of the change in desire for every 1-point increase in impression of reality TV shows for female students and male students, respectively. (2%)

Part II: The psychologists theorized that one's impression of reality TV will "moderate" the impact that the first three independent variables has on one's desire to have cosmetic surgery. If so, then x_4 will interact with each of the other independent variables.

k. Write down the equation of the model for $\mathrm{E}(y)$ that matches the above theory. (2%)

l. Fit the model, part **k**, to the data in the file. Evaluate the overall model. Use $\alpha = 0.01$ (3%)

m. Setup the null hypothesis for testing the psychologists' theory. (2%)

n. Conduct a partial-F test to test theory. Use $\alpha = 0.05$. (3%)

2. (9%) In general, before an academic publisher agrees to publish a book, each manuscript is thoroughly reviewed by university professors. Suppose that the Duxbury Publishing Company has recently received two manuscripts for statistics books. To help them decide which one to publish both are sent to 30 professors of statistics who rate the manuscripts to judge which one is better. Suppose that 10 Professors rate manuscript 1 better and 20 rate manuscript 2 better.

a. Which kind of nonparametric tests is appropriate for the station? (2%)

b. Can Duxbury conclude at the 5% significance level that manuscript 2 is not highly rated than manuscript 1? (4%)

c. What is the p-value of this test? (3%)

3. (22%) To determine whether extra personnel are needed for the day, the owners of a water adventure park would like to find a model that would allow them to predict the day's attendance each morning before opening based on the day of the week and weather conditions. The model is of the form

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where y = daily admission, $x_1 = \begin{cases} 1 & \text{if weekend} \\ 0 & \text{otherwise} \end{cases}$,

$x_2 = \begin{cases} 1 & \text{if sunny} \\ 0 & \text{otherwise} \end{cases}$, x_3 = predicted daily high temperature (°F)

Part I:

These data were recorded for a random sample of 30 days, and a linear regression model was fitted to the data. The least squares analysis produced the following results:

$\hat{y} = -105 + 25x_1 + 100x_2 + 10x_3$ with $s_{b1} = 10, s_{b2} = 30, s_{b3} = 4, R^2 = 0.65$

a. Interpret the estimated model coefficients. (3%)

b. Is there sufficient evidence to conclude that this model is useful for the prediction of daily attendance? Use $\alpha = 0.05$ [Hint: think about how to calculate F value if R^2 is known] (3%)

c. Is there sufficient evidence to conclude that the mean attendance increases on weekend? Use $\alpha = 0.05$ (3%)

d. Use the model to predict the attendance on a sunny weekday with a predicted high temperature of 95°F. (2%)

e. Suppose that 90% prediction interval for part **d** is (645, 1245). Interpret this interval. (2%)

Part II: The owners of the water adventure park are advised that the prediction model could probably be improved if interaction terms were added. In particular, it is thought that the rate at which mean attendance increases as predicted high temperature increases will be greater on weekend than on weekdays. The following model is therefore proposed: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3$
The same 30 days of data in part **a** are used to obtain the least squares model
$\hat{y} = 250 - 700x_1 + 100x_2 + 5x_3 + 15x_1 x_3$ with $s_{b4} = 3.0, R^2 = 0.96$

f. Do the data indicate that the interaction term is a useful addition to the model? Use $\alpha = 0.05$ (3%)

g. Use the model to predict the attendance for a sunny weekday with a predicted high temperature of 95°F. (2%)

h. Suppose that 90% prediction interval for part **g** is (800, 850). Compare this result with the prediction model in part **e**. Do the relative widths of the confidence interval support or refute your conclusion about the effectiveness of interaction term (part **f**)? (2%)

i. The owners, noting that the coefficient $b_1 = -700$, conclude the model is ridiculous because it seems to imply that the mean attendance will be 700 less on weekends than weekdays. Does their argument make sense? If yes, how could you solve the problem? If no, state your reason. (3%)

4. (9%) A movie critic wanted to determine whether or not moviegoers of different age groups evaluate a movie differently. With this objective, he commissioned a survey that asked people their ratings of the most recently watched movies. The

rating categories where: 1 = terrible, 2 = fair, 3 = good, 4 = excellent. Each respondent was also asked to categorize his or her age as either: 1 = teenager, 2 = young adult (20-34), 3 = middle age (35-50), 4 = over 50. The result are shown below.

Movie Ratings

Teenager	Young Adult	Middle Age	Senior
3	2	3	3
4	3	2	4
3	3	1	4
3	2	2	3
3	2	2	3
4	1	3	4
2	3	1	4
4	2	4	3

a. Which test can the movie critic use in this situation? (3 points)

b. Does this data provide sufficient evidence to infer at the 5% significance level that there are differences in ratings among the different age categories? (6 points)

5. (10%) Consider the 2011 monthly closing prices (i.e., closing prices on the last day of each month) given in file **HITECH** for IBM company stocks listed on the New York Stock Exchange,

a. Use the exponentially smoothed series (with w = .5) from January to September 2011 to forecast the monthly values of the IBM stock price from October to December 2011. Calculate the forecast errors. (4%)

b. Use a simple linear regression model fit to the IBM stock prices from January to September 2011. Let time *t* range from 1 to 9, representing the 9 months in the sample. Interpret the least squares estimates. (3%)

c. Compare the exponential smoothing forecasts, part **a**, to the regression forecasts, part **b**, using MAD and SSE. (3%)

6. (16%) The revenue (in $thousands) of a chain of fast food stores are listed for each quarter during the previous 5 years in file **Fastfood**.

   a. Use the regression analysis to determine the trend line. (3%)

   b. Determine the seasonal indexes. (4%)

c. Using the seasonal indexes and trend line to forecast the next 4 quarters. (4%)

d. Discuss whether exponential smoothing is an appropriate forecasting tool in this problem. State your reasons. (3%)