MSBA7002 Business Statistics - HW 1

Name: ________________________ Student ID: ____________________________

21 October 2024

Overview / Instructions

This homework will be *due on 30 October 2024 by 11:55 PM* via Moodle.

You are required to submit 1) an original R Markdown file and 2) a knitted HTML or PDF file. Please provide comments for the R code wherever you see appropriate. Nice formatting of the assignment will have extra points.

In general, be as concise as possible while giving a fully complete answer. All necessary data are available in Moodle.

Remember that the Class Policy strictly applies to homework. You are encouraged to discuss this with fellow students. However, each student has to know how to answer the questions on her/his own. Note that the final exam is individually based.

Question 0

Review the lectures.

Question 1: Confidence intervals

(a) An electrical firm manufactures light bulls that have a length of life that is approximately normally distributed with a standard deviation of 36 hours. If a sample of 45 bulbs has an average life of 920 hours, find a 90% confidence interval for the population mean of all bulbs produced by this firm.

(b) This Exercise is a "what-if" analysis designed to determine what happens to the interval estimate when the confidence level, sample size, and standard deviation change. A statistics practitioner took a random sample of 60 observations from a population with a standard deviation of 32 and computed the sample mean to be 120.

i. Estimate the population mean with 90% confidence.

ii. Repeat part (a) using a 95% confidence level.

iii. Repeat part (a) using a 99% confidence level.

iv. Describe the effect on the confidence interval estimate of increasing the confidence level.

(c) A vitamin company has developed a new diet program intended to help customers to lose weight. Before marketing the diet program, the company would like to get an estimate of the population proportion of users of this diet program who lose weight. They decide to select a random sample of 120 customers with weight control problems and determine the proportion who successfully lose weight. If 90 of the 120 customers indeed lost weight, determine a 95% confidence interval for the population proportion of users of this diet program who will lose weight.

Question 2: Hypothesis Testing

(a) Formulate the null and alternative hypotheses in each case:

i. Five years ago, the average Hong Kong resident drank 2 cups of coffee per day. You suspect that the amount has changed since then.

ii. An admission officer at UC Berkeley speculates that the average SAT score for entering freshmen is higher than the national average of 1560.

(b) Suppose a delivery company states that their packages arrive within two days or less on average. You want to find out whether the actual average delivery time is longer than this. You conduct a hypothesis test.

i. State the null and alternative hypotheses

ii. Suppose you conclude that the delivery company's average time to delivery is longer than two days, but your conclusion is wrong. What type of error did you. commit? Please explain your answer.

iii. Suppose you conclude that the delivery company's average time to delivery is less than two days, and your conclusion is correct. How does this correspond to Type I error or Type II error?

(c) The owner of a local nightclub has recently surveyed a simple random sample of n = 260 customers of the club. She would now like to determine whether or not the mean age of her customers is over 42. If so, she plans to alter the entertainment to appeal to an older crowd. If not, no entertainment changes will be made. Suppose she found that the sample mean was 42.6 years, and the population standard deviation was 5 years.

i. State the null and alternative hypotheses.

ii. Using R, what is the p-value associated with the test statistic? What is the function you used?

iii. Draw a conclusion at 5% significance level using the p-value approach.

iv. Draw a conclusion at 1% significance level using the p-value approach.

Question 3: Dummy Variable

The goal of creating dummy variables is to transform categorical variables into numeric values. However, dummy variables do not have to only take values in 0 and 1.

We discussed this example in class using the following dummy variable

Origin[Internal]=1, if Origin=``Internal"; =0, otherwise,

and considered the following interaction model:

$$Rating = \beta_0 + \beta_1 Origin[Internal] + \beta_2 Salary + \beta_3 Origin[Internal] * Salary + \varepsilon.$$

Now define another dummy variable

Origin[External]=1, if Origin=`` External"; Origin[External]=-1, otherwise,

And consider the following model

$$Rating = \alpha_0 + \alpha_1 Origin[\text{External}] + \alpha_2 Salary + \alpha_3 Origin[\text{External}] * Salary + \varepsilon.$$

Please derive the relationships between $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ and $\{\beta_0, \beta_1, \beta_2, \beta_3\}$.

Question 4: Categorical Variable in Linear Regression

ProdTime.dat contains information about 20 production runs supervised by each of the three managers. Each observation gives the time (in minutes) to complete the task, Time for Run, as well as the number of units produced, Run Size, and the manager involved, Manager. Our goal is to investigate which manager performs the best.

(a) Which variable should be the response? And which are the predictors?

(b) When including the categorical variable in the regression model, R will automatically create dummy variables for you. What are the two dummy variables created in this question?

(c) Fit a linear regression with only the main effects. How do you interpret the coefficients of the dummy variables? Are they significant? State the reasons.

(d) Now include the interaction terms in the linear regression. How do you interpret the coefficients of the interaction terms? Are they significant? State the reasons.

(e) What is the R^2 and the adjusted R^2 for your model in (c) and (d)? Which model should you choose? State the reasons.

Question 5: Linear Regression

The original data contains 392 observations about cars. To get the data, first install and load the packages ISLR and ISLR2. The data Auto should be loaded automatically. We use this case to go through methods learned so far.

Get familiar with the data first. You can use the code

?ISLR::Auto

to view a description of the data.

(a) Explore the data, with a particular focus on pairwise plots and summary statistics. Briefly summarize your findings and any peculiarities in the data.

(b) What effect does time have on MPG?

(c) Start with a simple regression of mpg vs. year and report R's `summary` output. Is year a significant variable at the .05 level? State what effect year has on mpg, if any, according to this model.

(d) Add horsepower on top of the variable year. Is year still a significant variable at the .05 level? Give a precise interpretation of the year effect found here. Include diagnostic plots with a particular focus on the model residuals and diagnoses.

(e) The two 95% CI's for the coefficient of the year differ between i) and ii). How would you explain the difference to a non-statistician?

(f) Construct a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at the 0.05 level? Explain the year effect (if any).

Note that the same variable can play different roles! Take a quick look at the variable `cylinders`, and try to use this variable in the following analyses wisely. We all agree that a larger number of cylinders will lower mpg. However, we can interpret `cylinders` as either a continuous (numeric) variable or a categorical variable.

(g) Fit a model, that treats `cylinders` as a continuous/numeric variable: `lm(mpg ~ horsepower + cylinders, ISLR::Auto)`. Is `cylinders` significant at the 0.01 level? What effect does the variable `cylinders` play in this model?

(h) Fit a model that treats `cylinders` as a categorical/factor variable: `lm(mpg ~ horsepower + as.factor(cylinders), ISLR::Auto)`. Is `cylinders` significant at the .01 level? What is the effect of `cylinders` in this model?

(i) What are the fundamental differences between treating `cylinders` as a numeric or a factor? Use adjusted R^2 for model selection. Explain their difference.

Question 6: Model Selection

We will continue using the *Auto* data in the previous question. We are interested in selecting the variables that have significant effects on the response `mpg`. Perform subset selection and use adjusted R^2, Mallow's Cp, and BIC to select the best model. State the reasons for choosing the best model.