THE AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE,
ACTUARIAL STUDIES AND STATISTICS

STAT3008/STAT7001
APPLIED STATISTICS

# Assignment 1

*Lecturer: Dr Tao Zou*

*Last Updated: Sun Sep 17 10:29:02 2017*

This assignment is due at 12:00 pm, Sep 27, 2017.

This assignment is worth 10% of your final grade but is optional and redeemable. Maximum points: 10.0. You cannot get partially correct for all the questions, since each question is only worth 0.5 points. **Assignments can only be submitted via the physical assignment box at the front of the reception on Level 4, CBE Building (26C). Hard copy submission is required.** Late submission will not be accepted and the weight will roll over to your final exam. Identical submissions are treated as cheating.

Please **exactly follow the instructions of questions** and write down the answers of the following questions in the **answer sheet** file on the Wattle. Note that you do not need to copy the questions in the answer sheet. Please only submit your finished answer sheet and do not paste any unrelated results. The data used in this assignment are in the R package "Sleuth3", whose instruction manual is on the Wattle.

The significance level for all the questions is set to be 0.05.

**Gender Differences in Wages** (Revised based on ex 25 of Chapter 12 in "The Statistical Sleuth"). Display 12.21 is a partial listing of a data set with weekly earnings for 9,835 Americans surveyed in the March 2011 Current Population Survey (CPS). The dataset is stored in the object "ex1225" of the R library "Sleuth3". What evidence is there from these data that males tend to receive higher earnings than females with the same values of the other variables? Note that there might be an interaction between Sex and Marital Status. (Data from U.S. Bureau of Labor Statistics and U.S. Bureau of the Census: Current Population Survey, March 2011 http://www.bls.census.gov/cps ftp.html#cpsbasic; accessed July 25, 2011.)

**DISPLAY 12.21** Region of the United States (Northeast, Midwest, South, or West) where individual worked, Metropolitan Status (Metropolitan, Not Metropolitan, or Not Identified), Age (years), Sex (Male or Female), Marital Status (Married or Not Married), EdCode (corresponding roughly to increasing education categories), Education (16 categories), Job Class (Private, Federal Government, State Government, Local Government, or Private), and Weekly Earnings (in U.S. dollars) for 9,835 individuals surveyed in the March 2011 Current Population Survey; first 5 of 9,835 rows

| Region | MetropolitanStatus | Age | Sex | MaritalStatus | Edcode | Education | JobClass | WeeklyEarnings |
|---|---|---|---|---|---|---|---|---|
| Northeast | Not Metropolitan | 20 | Male | Not Married | 39 | HighSchoolDiploma | Private | 467.50 |
| West | Metropolitan | 59 | Male | Married | 43 | BachelorsDegree | Private | 1,269.00 |
| West | Metropolitan | 62 | Male | Married | 34 | SeventhOrEighthGrade | Private | 1,222.00 |
| West | Metropolitan | 39 | Male | Married | 39 | HighSchoolDiploma | Private | 276.92 |
| South | Not Metropolitan | 60 | Female | Married | 36 | TenthGrade | Private | 426.30 |

Display taken from class text: "The Statistical Sleuth".

In order to investigate the above problem, please use R to answer the following Questions 1 – 3 in the answer sheet.

**Question 1** (Multiple Linear Regression and Variable Selection, 2.0 points)

Consider the multiple linear regression model to regress the logarithm of "WeeklyEarnings" on variables "Age", "Sex", "MaritalStatus" and "EdCode" (please do not consider the interaction terms for now).

Please answer the following questions in the answer sheet.

a) (0.5 points) Please use R to obtain the fitted model based on the above variables. What is the least squares estimate for the coefficient of "EdCode" (rounded to four decimal places)? Please also interpret this estimated coefficient.

b) (0.5 points) Based on the "summary" function output of this fitted model, what are the null hypothesis and the alternative hypothesis for the "F-statistic" in the "summary" function output? What conclusion can you obtain for this $F$-test?

c) (0.5 points) Please use R to perform the backward elimination based on $F$-statistic. Which variables should we choose to predict the logarithm of "WeeklyEarnings" by using this variable selection method?

d) (0.5 points) Please paste the R codes for all the above analyses of Question 1 in the answer sheet.

**Question 2** (Model Diagnostics, 3.5 points)

Consider the multiple linear regression model in Quesiton 1 a). Please answer the following questions in the answer sheet.

a) (0.5 points) Based on the "summary" function output of the fitted model in Quesiton 1 a), please interpret the R-squared.

b) (0.5 points) Please paste the residuals versus fitted values plot of the fitted model in Quesiton 1 a) in the answer sheet. Are the assumptions in the multiple linear regression model violated based on this plot?

c) (0.5 points) Please paste the Q-Q plot of the residuals based on the fitted model in Quesiton 1 a) in the answer sheet. What conclusions can you obtain via the Q-Q plot?

d) (0.5 points) Please paste the Cook's distance plot of the fitted model in Quesiton 1 a) in the answer sheet. Based on the criterion introduced in lectures, are there any influential observations? Why or why not?

e) (0.5 points) Please find the observation with the largest Cook's distance. (Hint: use "which" function in R.) Based on the "rule of thumb" cut-offs for the studentized residual, is this observation an outlier? How to deal with this suspected influential observation?

f) (0.5 points) We have found the observation with the largest Cook's distance in e). Based on the "rule of thumb" cut-off for the leverage, does this observation have distant explanatory variable values? Why or why not?

g) (0.5 points) Please paste the R codes for all the above analyses of Question 2 in the answer sheet.

**Question 3** (Multiple Linear Regression for Continuous and Categorical Explanatory Variables, 3.0 points)

Consider the multiple linear regression model in Quesiton 1 a), but we would like to add more explanatory variables. Please answer the following questions in the answer sheet.

a) (0.5 points) We first use the following R codes to generate the indicator variables for categorical variables "Region" and "MetropolitanStatus", respectively:

```
IMidwest=ifelse(Region=="Midwest",1,0)
INortheast=ifelse(Region=="Northeast",1,0)
ISouth=ifelse(Region=="South",1,0)

IMetropolitan=ifelse(MetropolitanStatus=="Metropolitan",1,0)
INotMetropolitan=ifelse(MetropolitanStatus=="Not Metropolitan",1,0)
```

If we are also interested to show whether or not the mean of log(WeeklyEarnings) in each category of "FedGov", "StateGov" and "LocalGov", is significantly different from that in the category of "Private", directly via the R output, which category should we choose as the baseline level for the categorical variable "JobClass"? Which indicator variables of "JobClass" should we select for model fitting to realise the above purpose?

b) (0.5 points) Please use R to obtain the fitted model based on all the variables involved in Question 1 a) and Question 3 a). Still please do not consider the interaction terms for now. Based on the "summary" function output of this fitted model, if we control the other variables, is the mean of log(WeeklyEarnings) in each category of "FedGov", "StateGov" and "LocalGov", is significantly different from that in the category of "Private"?

c) (0.5 points) Based on the fitted model in Question 3 b), now we are interested in testing whether or not at least one of categories of "FedGov", "StateGov" and "LocalGov" has a different level of the mean of log(WeeklyEarnings), compared to the category of "Private", when other variables are held constant. Please use R to obtain an appropriate test statistic and the corresponding $p$-value. What conclusion can you obtain based on the result?

d) (0.5 points) Consider the model and the variables in Question 3 b). But now we add an interaction between Sex and Marital Status and obtain a new model. Compute and show the sum of squared errors (SSE) for these two fitted models. Which one is smaller?

Compute ? manually or use ANOVA table.

e) (0.5 points) Consider the model with the interaction in Question 3 d). What are the explanations of the estimated coefficient of the interaction term? Is the interaction between Sex and Marital Status significant? Why or why not?

f) (0.5 points) Please paste the R codes for all the above analyses of Question 3 in the answer sheet.

**Question 4** (Simulation for Multiple Linear Regression, 1.5 points)

Consider the multiple linear regression model $\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ for the observations $\{Y_i, X_{1,i}, X_{2,i}\}_{i=1}^n$, and the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for the coefficients $\beta_0$, $\beta_1$ and $\beta_2$ can be obtained.

Lily wants to use R to generate random samples based on the multiple linear regression model assumptions. She follows the steps below.

STEP 1: Specify $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = -1$,

STEP 2: Suppose the observations $X_{1,1}, \cdots, X_{1,n}$ are $1, 2, \cdots, 100$, so the number of observations $n = 100$.

STEP 3: Generate $X_{2,1}, \cdots, X_{2,n}$ from the $t_3$ distribution. (Hint: similar to the codes on page 18 of Lecture Notes 3.)

STEP 4: Generate $\mathcal{E}_1, \cdots, \mathcal{E}_n$ from the standard normal distribution [N(0,1) with mean 0 and variance 1].

STEP 5: Generate $Y_i = \mu\{Y_i|X_{1,i}, X_{2,i}\} + \mathcal{E}_i$, $i = 1, \cdots, n$.

STEP 6: Repeat Step 4 – Step 5 1,000 times and obtain 1,000 different datasets of $\{Y_i, X_{1,i}, X_{2,i}\}_{i=1}^n$.

Lei Li is a friend of Lily. Lily hands over the above 1,000 datasets to him but she does not tell him the true values of $\beta_0$, $\beta_1$ and $\beta_3$. Based on each dataset, Lei Li computes the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ as well as the 95% confidence interval for the mean of response given $X_1 = 2.5$ and $X_2 = 0$. Ultimately, he obtains 1,000 different confidence intervals.

Then Lily computes the mean of response $\mu\{Y|X_1 = 2.5, X_2 = 0\}$ and tells Lei Li this information. Lei Li counts the number of the confidence intervals that cover $\mu\{Y|X_1 = 2.5, X_2 = 0\}$.

Please answer the following questions in the answer sheet.

a) (0.5 points) Suppose you play both roles of Lily and Lei Li and realise the above steps in R. Please paste the complete R codes for all the above procedures in the answer sheet. (Hint: similar to the codes on page 7 of Lecture Notes 2.)

b) (0.5 points) What is the number of the confidence intervals that cover $\mu\{Y|X_1 = 2.5, X_2 = 0\}$ based on the above steps? Please answer this question in the answer sheet.

c) (0.5 points) Based on the result of b), interpret the 95% confidence interval for the mean of response. Please answer this question in the answer sheet.