THE AUSTRALIAN NATIONAL UNIVERSITY

RESEARCH SCHOOL OF FINANCE,
ACTUARIAL STUDIES AND STATISTICS

STAT3008/STAT7001
APPLIED STATISTICS

Assignment 1 Solution

Lecturer: Dr Tao Zou

Last Updated: Sat Sep 23 21:27:48 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.

(The indicator variables selected can be different in this question. But the result should be the same.)

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.

Solution: The estimated coefficient is

```
EdCode
0.1121
```

If EdCode is increased by one unit, the **estimated mean** of **log(WeeklyEarnings)** will increase 0.1121 units, **with other variables held constant**. Based on the explanations in the display on page 2 of this assignment, EdCode corresponds roughly to increasing education categories. Hence the interpretation of this estimated coefficient also implies that if Education is increased, the **estimated mean** of **log(WeeklyEarnings)** will also increase.

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?

Solution:

```
Call:
lm(formula = Y ~ Age + IMale + IMarried + EdCode)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-7.8730 -0.3196  0.0103  0.3412  1.9651

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5210142  0.0928107   16.39   <2e-16 ***
Age         0.0085773  0.0004691   18.29   <2e-16 ***
IMale       0.2327844  0.0114581   20.32   <2e-16 ***
IMarried    0.1290131  0.0119685   10.78   <2e-16 ***
EdCode      0.1120721  0.0022187   50.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5614 on 9830 degrees of freedom
Multiple R-squared:  0.2682,    Adjusted R-squared:  0.2679
F-statistic: 900.6 on 4 and 9830 DF,  p-value: < 2.2e-16
```

H_0 : The coefficients of all the variables, except for the intercept, are zeros $\leftrightarrow$

H_a : at least one of the coefficients is not zero.

The p-value is smaller than 0.05. Hence we reject the null hypothesis and conclude that at least one of the considered explanatory variables is useful in explaining the mean of log(WeeklyEarnings).

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?

Solution: All four variables. R stops at the backward step for four variables, which corresponds to the start model for backward elimination. Hence, R returns nothing in the output.

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

Solution:

```r
rm(list=ls())
library('Sleuth3')
data=ex1225

#Q1 a)
attach(data)
Y=log(WeeklyEarnings)
IMale=ifelse(Sex=="Male",1,0)
IMarried=ifelse(MaritalStatus=="Married",1,0)
fit=lm(Y~Age+IMale+IMarried+EdCode)
round(fit$coef[5],4)

#Q1 b)
summary(fit)

#Q1 c)
X=cbind(Age,IMale,IMarried,EdCode)
#install.packages('wle')
library(wle) #need to load this library!
###
mle.stepwise(Y~X,f.in=4,f.out=4,type="Backward")
detach(data)
```

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.

(The indicator variables selected can be different in this question. But the result should be the same.)

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

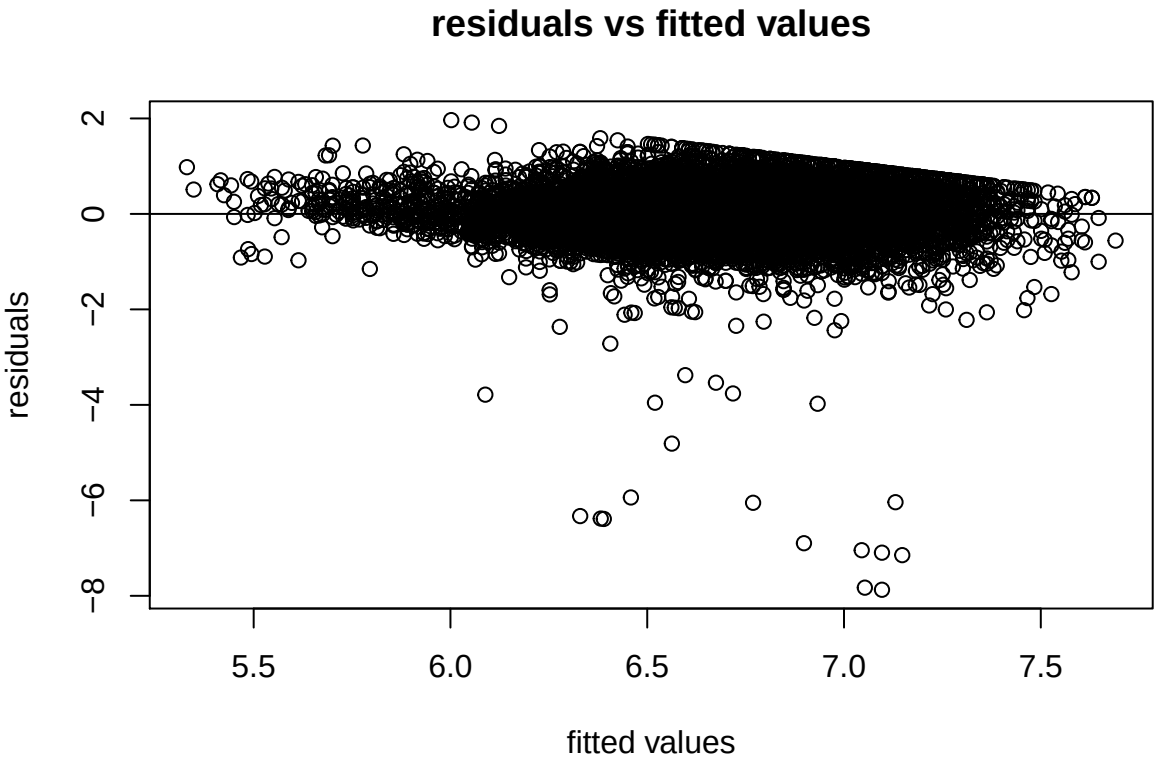
Solution: The R-squared is

```
[1] 0.2681771
```

The regression model can explain 26.82% of the total response variation.

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?

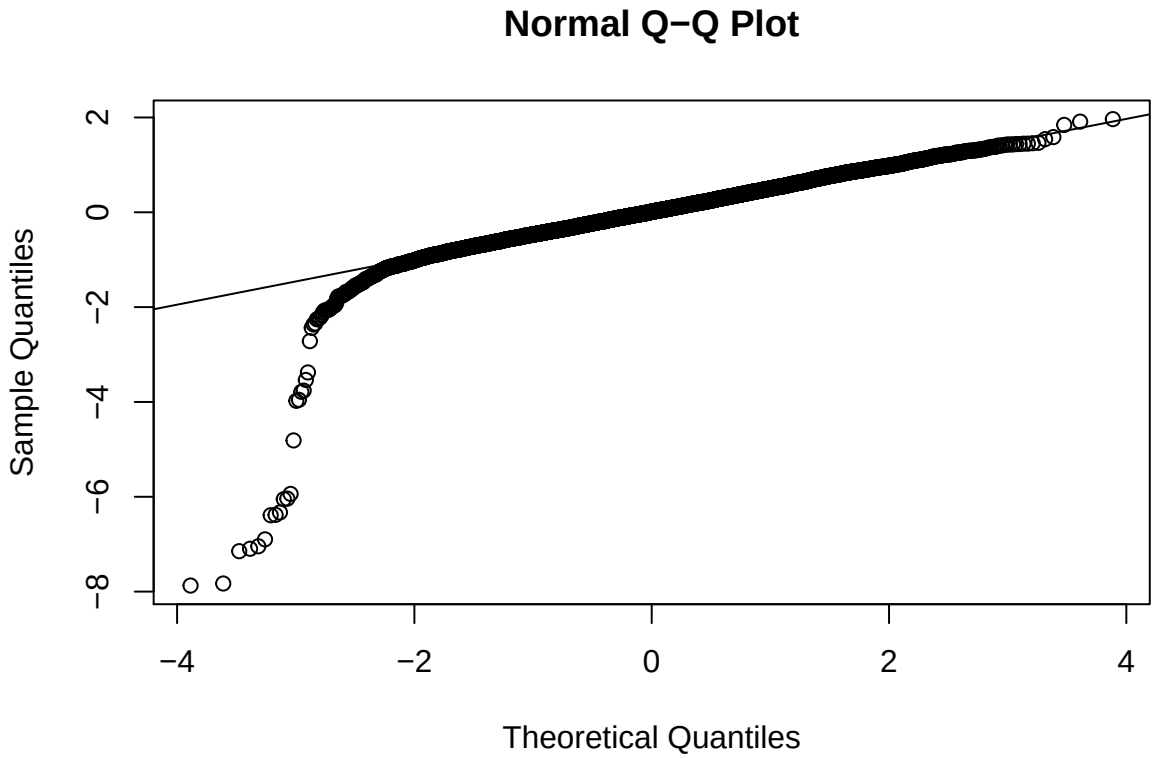
Solution:



Yes. The assumptions in the multiple linear regression model are 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?

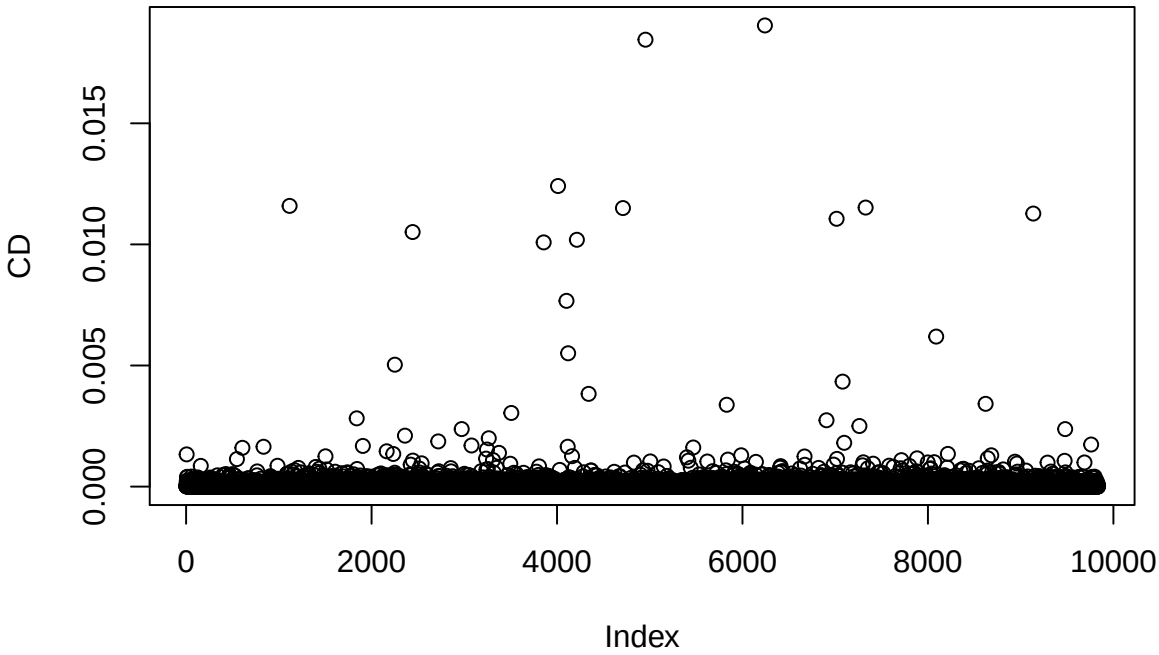
Solution:



The distribution of the residuals is heavy-tailed.

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?

Solution:



The answer can be either of the following.

No, there are no influential observations. Because a rough "rule of thumb" cut-off for Cook's distance is 1. In the plot, there are no Cook's distances larger than 1.

Yes, there are influential observations. Since another option is to identify an influential observation that has a relatively large value of Cook's distance. Hence, we can identify, for instance, the observation with the largest Cook's distance as the influential observation.

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?

Solution:

The observation

6242

has the largest Cook's distance. Its studentized residual is

```
-14.08946
```

which is smaller than the cut-off -1.96 (or -2). Hence this observation is an outlier. We should omit the observation and proceed to use the left observations for model fitting.

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?

Solution:

The observation

```
6242
```

has the largest Cook's distance. Its leverage is

```
0.0004889895
```

which is smaller than the cut-off $2(k+1)/n$, namely,

```
[1] 0.001016777
```

Hence this observation does not have distant explanatory variable values.

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

Solution:

```r
#Q2 a)
attach(data)
summary(fit)$r.squared

#Q2 b)
plot(fit$fitted.values,fit$residuals,main="residuals vs fitted values"
     ,xlab="fitted values",ylab="residuals")
abline(h=0)
```

```r
#Q2 c)
qqnorm(fit$residuals)
qqline(fit$residuals)
```

```r
#Q2 d)
CD=cooks.distance(fit) #Cook's distance
plot(CD)
abline(h=1,col='red')
```

```r
#Q2 e)
result=which(CD==max(CD))
names(result)=''
result
Rstd<-rstudent(fit) #Studentized residuals
result=Rstd[which(CD==max(CD))]
names(result)=''
result

#Q2 f)
X=cbind(Age,IMale,IMarried,EdCode)
lev=hat(X) #Leverage
result=lev[which(CD==max(CD))]
names(result)=''
result

n=length(Y)
k=4
2*(k+1)/n
detach(data)
```

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.

(The indicator variables selected for "Sex" and "MaritalStatus" can be different in this question. But the result should be mostly the same.)

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?

Solution: The baseline level should be "Private". The indicator variables we should choose are

```
IFedGov=ifelse(JobClass=="FedGov",1,0)
IStateGov=ifelse(JobClass=="StateGov",1,0)
ILocalGov=ifelse(JobClass=="LocalGov",1,0)
```

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"?

Solution:

```
Call:
lm(formula = Y ~ Age + IMale + IMarried + EdCode + IMidwest +
    INortheast + ISouth + IMetropolitan + INotMetropolitan +
    IFedGov + ILocalGov + IStateGov)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8845 -0.3144  0.0105  0.3337  1.8625

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.6808935  0.1146721  14.658  < 2e-16 ***
Age               0.0084347  0.0004697  17.959  < 2e-16 ***
IMale             0.2279893  0.0114137  19.975  < 2e-16 ***
IMarried          0.1359792  0.0119130  11.414  < 2e-16 ***
EdCode            0.1096774  0.0022599  48.532  < 2e-16 ***
IMidwest         -0.0169560  0.0163922  -1.034  0.30098
INortheast        0.0681199  0.0171160   3.980 6.94e-05 ***
ISouth           -0.0456402  0.0154390  -2.956  0.00312 **
IMetropolitan    -0.0410380  0.0683425  -0.600  0.54820
INotMetropolitan -0.1402097  0.0694120  -2.020  0.04341 *
IFedGov           0.2148419  0.0298767   7.191 6.90e-13 ***
ILocalGov         0.0075609  0.0202282   0.374  0.70858
IStateGov        -0.0420323  0.0243692  -1.725  0.08459 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5571 on 9822 degrees of freedom
Multiple R-squared:  0.2798,    Adjusted R-squared:  0.2789
F-statistic:   318 on 12 and 9822 DF,  p-value: < 2.2e-16
```

The mean of log(WeeklyEarnings) in the category of "FedGov", is significantly higher than that in the category of "Private".

The mean of log(WeeklyEarnings) in each category of "StateGov" and "LocalGov", is not 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?

Solution: Based on the following R output,

```
Analysis of Variance Table

Model 1: Y ~ Age + IMale + IMarried + EdCode + IMidwest + INortheast +
    ISouth + IMetropolitan + INotMetropolitan
Model 2: Y ~ Age + IMale + IMarried + EdCode + IMidwest + INortheast +
    ISouth + IMetropolitan + INotMetropolitan + IFedGov + ILocalGov +
    IStateGov
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   9825 3066.4
2   9822 3048.9  3    17.563 18.86 3.419e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

the test statistic is 18.86 and the corresponding p-value is 3.419×10^{-12}. Hence we reject the null hypothesis (please write the null and alternative hypotheses by yourselves), and conclude that at least one of categories of "FedGov", "StateGov" and "LocalGov" has a sigfinicantly different level of the mean of log(WeeklyEarnings), compared to the category of "Private", when other variables are held constant.

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?

Solution: The SSEs are

```
[1] 3048.872
```

```
[1] 3043.887
```

respectively. The model with the interaction term has a smaller SSE.

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?

Solution: First we have the following R output

```
Call:
lm(formula = Y ~ Age + IMale + IMarried + EdCode + IMidwest +
    INortheast + ISouth + IMetropolitan + INotMetropolitan +
    IFedGov + ILocalGov + IStateGov + IMale * IMarried)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8990 -0.3108  0.0098  0.3345  1.8453

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.7156946  0.1149122  14.930  < 2e-16 ***
Age               0.0083201  0.0004702  17.696  < 2e-16 ***
IMale             0.1732413  0.0177882   9.739  < 2e-16 ***
IMarried          0.0874730  0.0169700   5.155 2.59e-07 ***
EdCode            0.1095660  0.0022583  48.516  < 2e-16 ***
IMidwest         -0.0161110  0.0163809  -0.984  0.32537
INortheast        0.0686626  0.0171034   4.015 6.00e-05 ***
ISouth           -0.0447409  0.0154288  -2.900  0.00374 **
IMetropolitan    -0.0405314  0.0682902  -0.594  0.55285
INotMetropolitan -0.1394636  0.0693590  -2.011  0.04438 *
IFedGov           0.2147524  0.0298538   7.193 6.78e-13 ***
ILocalGov         0.0090446  0.0202160   0.447  0.65460
IStateGov        -0.0412285  0.0243513  -1.693  0.09047 .
IMale:IMarried    0.0925851  0.0230853   4.011 6.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5567 on 9821 degrees of freedom
Multiple R-squared:  0.281, Adjusted R-squared:   0.28
F-statistic: 295.2 on 13 and 9821 DF,  p-value: < 2.2e-16
```

We analyse the interaction as follows. The regression model is

$$\mu\{Y|X\} = \beta_0 + \beta_1 \text{IMale} + \beta_2 \text{IMarried} + \beta_3 \text{IMale} \times \text{IMarried} + \text{other variables}.$$

For different categories, we have the following sub-models:

Categories	(IMale, IMarried)	model
(Female, Not Married)	(0,0)	$\mu\{Y\|X\} = \beta_0 + \cdots$
(Male, Not Married)	(1,0)	$\mu\{Y\|X\} = \beta_0 + \beta_1 + \cdots$
(Female, Married)	(0,1)	$\mu\{Y\|X\} = \beta_0 + \beta_2 + \cdots$
(Male, Married)	(1,1)	$\mu\{Y\|X\} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \cdots$

Then we have $\hat{\beta}_1 = 0.1732$, $\hat{\beta}_2 = 0.0875$ and $\hat{\beta}_3 = 0.0926$. So the interpretation of $\hat{\beta}_3 = 0.0926$ is: **with other variables held constant**, the **estimated mean** of **log(WeeklyEarnings)** for category (Male, Married) is $0.1732 + 0.0926$ higher than that for category (Female, Married), or the **estimated mean** of **log(WeeklyEarnings)** for category (Male, Married) is $0.0875 + 0.0926$ higher than that for category (Male, Not Married). Since the p-value of the interaction term is smaller than 0.05, the interaction is significant in the model.

Note: The indicator variables for "Sex" and "MaritalStatus" can be set differently. The above interpretation can be a little bit different based on different settings. But it should be very similar.

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

Solution:

```
#Q3 a)
attach(data)
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)

IFedGov=ifelse(JobClass=="FedGov",1,0)
ILocalGov=ifelse(JobClass=="LocalGov",1,0)
IStateGov=ifelse(JobClass=="StateGov",1,0)

#Q3 b)
fit=lm(Y~Age+IMale+IMarried+EdCode+IMidwest+INortheast+ISouth+
```

```
IMetropolitan+INotMetropolitan+IFedGov+ILocalGov+IStateGov)
summary(fit)

#Q3 c)
fitr=lm(Y~Age+IMale+IMarried+EdCode+IMidwest+INortheast+ISouth+
IMetropolitan+INotMetropolitan)
#extra-sums-of-squares test
anova(fitr,fit,test='F')

#Q3 d)
fitNew=lm(Y~Age+IMale+IMarried+EdCode+IMidwest+INortheast+ISouth+
IMetropolitan+INotMetropolitan+IFedGov+ILocalGov+IStateGov
+IMale*IMarried)
deviance(fit)
deviance(fitNew)

#Q3 e)
summary(fitNew)
detach(data)
```

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 β_0, β_1 and β_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 β_0, β_1 and β_2. 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.)

Solution:

```r
#Q4
rm(list=ls())
beta0=2;beta1=1;beta2=-1
X1=1:100
n=length(X1)
set.seed(1)
X2=rt(n,3)
Y=rep(0,n)
numsamp=1000
CIl=rep(0,numsamp)
CIu=rep(0,numsamp)
x0=data.frame(X1=2.5,X2=0)
for(i in 1:numsamp) {
  errors=rnorm(n)
  Y=beta0+beta1*X1+beta2*X2+errors
  MLRfit=lm(Y~X1+X2)
  CI=predict(MLRfit,x0,interval='confidence',level=0.95)
  CIl[i]=CI[2]
  CIu[i]=CI[3]
}
MeanResponse=beta0+beta1*x0$X1+beta2*x0$X2
Count=ifelse(CIl<=MeanResponse & CIu>=MeanResponse, 1,0)
sum(Count)
```

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.

Solution: The number is

```r
sum(Count)
```

```
[1] 947
```

(This number can be different by using different "set.seed()" but should be close to 950.)

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.

Solution:

The 95% confidence interval for the mean of response means that if we use R to obtain 1,000 confidence intervals for 1,000 repeated samples, then around 950 confidence intervals will cover the mean of response.

(This is the reason why we can only say the mean of response lies in the confidence interval with confidence 95%, instead of probability 95%.)