ASSIGNMENT COVER SHEET

Research School of Finance, Actuarial Studies and Statistics
ANU College of Business and Economics
Australian National University
Canberra ACT 0200 Australia
www.anu.edu.au

Submission and assessment is anonymous where appropriate and possible. Please do not write your name on this coversheet.

This coversheet must be attached to the front of your assessment when submitted in hard copy.

All assessment items submitted in hard copy are due by 12:00 pm.

Student ID For group assignments, list each student's ID			
Course Code	STAT7001		
Course Name	Applied Statistics		
Assignment number	1		
Assignment Topic			
Lecturer	Dr. Tao Zou		
Tutor	Ziren Chen		
Tutorial (day and time)	Friday 10-11 am		
Word count		Due Date: Sep 27	
Date Submitted		Extension Granted	

I declare that this work:

- ☐ upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Student Academic Integrity;
- ☐ is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;
- ☐ is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- ☐ gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- ☐ in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

Initials

For group assignments, each student must initial.

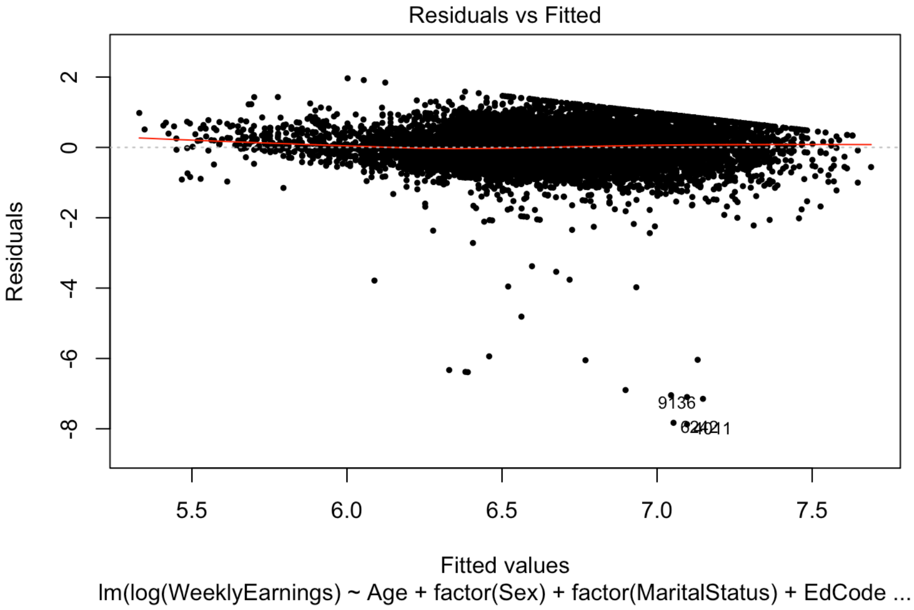
ANSWER SHEET

Research School of Finance, Actuarial Studies and Statistics
ANU College of Business and Economics
Australian National University
Canberra ACT 0200 Australia
www.anu.edu.au

Please input your answers of the questions in Assignment 1 on the right side of the table.

<table>
<tr><th colspan="2">Question 1 (2.0 points)</th></tr>
<tr><td>a</td><td>The least square estimate for the coefficient of EdCode is 0.1121.
It means when holding other independent variables constant, if we increase EdCode by 1, the dependent variable, i.e. the logarithm of WeeklyEarnings will increase by 0.1121.</td></tr>
<tr><td>b</td><td>Null hypothesis: The estimated coefficients ($\beta_{Age}, \beta_{Sex}, \beta_{MaritalStatus}, \beta_{EdCode}$) are all 0s.
Alternative hypothesis: At least one of the estimated coefficients above is not 0.
Since p-value is $2.2\times10^{-16} < 0.05$, so reject the null hypothesis, we believe at least one of the coefficients is significantly different from 0.</td></tr>
<tr><td>c</td><td>In fact, we did not drop any term in our fitted model. So we keep Age, Sex, MaritalStatus and EdCode to predict the logarithm of WeeklyEarnings via backward elimination.</td></tr>
<tr><td>d</td><td><pre>library(Sleuth3)
library(wle)
data <- ex1225
head(data)
attach(data)
mlr <- lm(log(WeeklyEarnings)~Age+factor(Sex)+factor(MaritalStatus)+EdCode)
(a)
summary(mlr)
(c)
lm.full <- mlr
lm.null <- lm(log(WeeklyEarnings)~1)
drop1(lm.full,test="F")
or
mle.stepwise(log(WeeklyEarnings)~Age+factor(Sex)+factor(MaritalStatus)+EdCode,
type="Backward")</pre></td></tr>
<tr><th colspan="2">Question 2 (3.5 points)</th></tr>
<tr><td>a</td><td>R-squared indicates that only 26.82% of variation can be explained by the model.</td></tr>
</table>

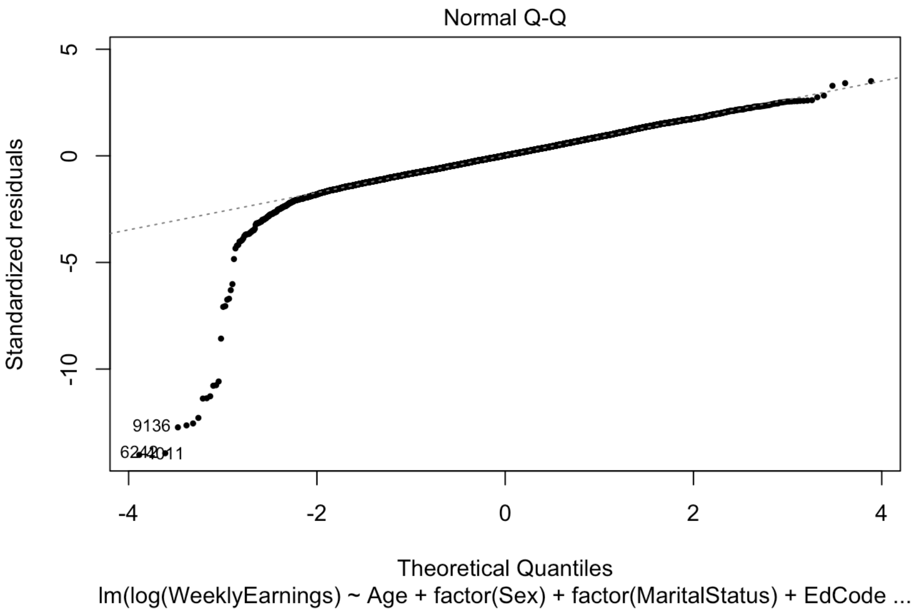
b



It seems that our residuals vs. fitted plot violates the assumption of homoscedasticity as some data in the middle quantile have relatively small residuals.

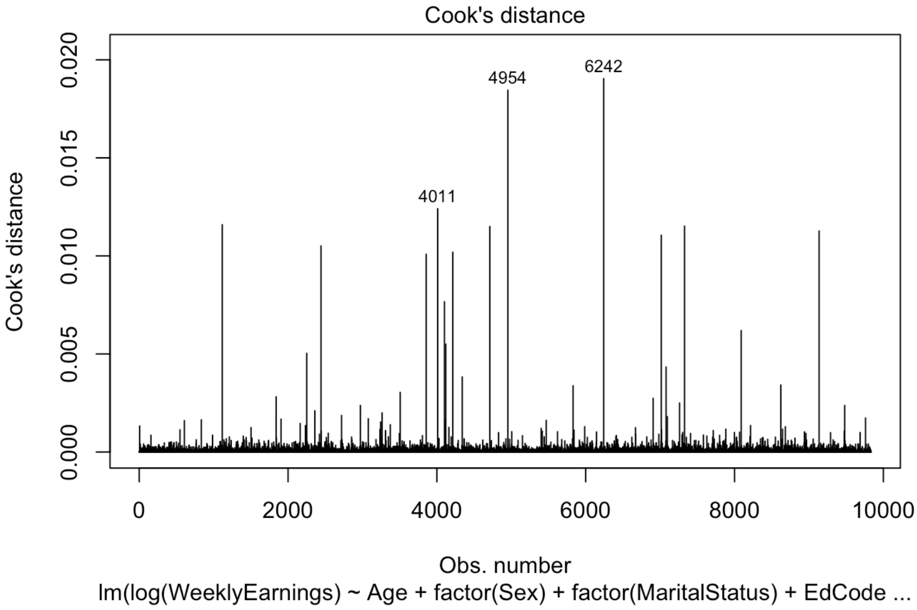
Also, the upper quantile seems to have a linear cutoff boundary which suspiciously could violate the assumption that all data are independent.

c



Clearly, the lower quantile data deviates from the line in Q-Q plot, looks like a "heavy tail", thus our assumption of normality is violated, i.e. the data is not normally distributed.

d



The "rule of thumb" cut-off for Cook's distance is 1, our largest Cook's distance is still less than 0.020. Also, it is not relatively larger than others. So we claim that there are no influential observations here.

e

The observation 6242 has the largest Cook's distance.

Since the studentized residual of observation 6242 is -14.09 < -2, so we believe it is an outlier. We usually delete the observation from the original dataset and refit the model.

f

The leverage of observation 6242 is **0.0004889895**, while the cutoff value is 0.001016777.

So the leverage of observation 6242 is less than the "rule of thumb" cut-off therefore, it does not have distant explanatory variable values.

g

```
# (a)
summary(mlr)
# (b)
plot(mlr, which=1, pch=16, cex=0.6)
# (c)
plot(mlr, which=2, pch=16, cex=0.6)
# (d)
plot(mlr, which=4, pch=16, cex=0.6)
# (e)
which.max(cooks.distance(mlr))
rstudent(mlr)[6242]
# (f)
lev <- hat(cbind(Age, factor(Sex), factor(MaritalStatus), EdCode))
lev[6242]
(lev.cutoff <- 2*(4+1)/nrow(data))
```

Question 3 (3.0 points)

a	We should use "**Private**" as the baseline level for the categorical variable "**JobClass**". Consequently, we select "**IFedGov**", "**ILocalGov**" and "**IStateGov**" as indicator variables, where - **IFedGov** is 1 if JobClass is **FedGov**, 0 otherwise. - **ILocalGov** is 1 if JobClass is **LocalGov**, 0 otherwise. - **IStateGov** is 1 if JobClass is **StateGov**, 0 otherwise.
b	The p-value of "**IFedGov**" is less than 0.05, so it is significantly different from the category of "Private", but the p-values of "**ILocalGov**" and "**IStateGove**" are greater than 0.05 so that these two categories are not significantly different from "**Private**" category.
c	F-statistic is **18.86** while p-value is less than 0.05. As a result, we suggest that we should reject null hypothesis and at least one category has a different level of the mean of **log(WeeklyEarnings)** compared to the category of "**Private**".
d	The fitted model from Q3b has **SSE=3048.87**, while the model with an extra interaction term has **SSE=3043.89**. So the second model has smaller SSE, i.e. less unexplained variation.
e	Suppose the regression model is $\mu(Y\|X) = \beta_0 + \beta_1\text{Age} + \beta_2\text{Sex} + \beta_3\text{MaritalStatus} + \beta_4\text{EdCode} + \beta_5\text{IMidwest} + \beta_6\text{INortheast} + \beta_7\text{ISouth} + \beta_8\text{IMetropolitan} + \beta_9\text{INotMetropolitan} + \beta_{10}\text{IFedGov} + \beta_{11}\text{ILocalGov} + \beta_{12}\text{IStateGov} + \beta_{13}\text{Sex: MaritalStatus}$ Thus the baseline level model for Female and Not Married: $\mu(Y\|X) = \beta_0 + \beta_1\text{Age} + \beta_4\text{EdCode} + \cdots + \beta_{12}\text{IStateGov}$ Model for Male and Not Married: $\mu(Y\|X) = (\beta_0 + \beta_2) + \beta_1\text{Age} + \beta_4\text{EdCode} + \cdots + \beta_{12}\text{IStateGov}$ Model for Female and Married: $\mu(Y\|X) = (\beta_0 + \beta_3) + \beta_1\text{Age} + \beta_4\text{EdCode} + \cdots + \beta_{12}\text{IStateGov}$ Model for Male and Married: $\mu(Y\|X) = (\beta_0 + \beta_2 + \beta_3 + \beta_{13}) + \beta_1\text{Age} + \beta_4\text{EdCode} + \cdots + \beta_{12}\text{IStateGov}$ Note that $\beta_2 = 0.2658, \beta_3 = -0.0875, \beta_{13} = -0.0926$. In this case, suppose all the other terms hold constant, then we will have the following: 1. If we only change the **Sex** of baseline (from Female to Male), the **logarithm of WeeklyEarnings** will increase by **0.2658.** 2. If we only change the **MaritalStatus** of baseline (from Married to Not Married), the **logarithm of WeeklyEarnings** will decrease by **0.0875.** 3. If we change the **Sex** from Female (baseline level) to Male and change the **MaritalStatus** from Married to Not Married at the same time, the **logarithm of WeeklyEarnings** will increase by **-0.0926+0.2658-0.0875=0.0857.** The interaction term is **significant** as its p-value (6.10×10^{-15}) is strictly less than 0.05.
f	`# (a)` `levels(Region)` `levels(MetropolitanStatus)` `IMidwest=ifelse(Region=="Midwest",1,0)`

```r
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)
levels(JobClass)
IFedGov <- ifelse(JobClass=="FedGov",1,0)
ILocalGov <- ifelse(JobClass=="LocalGov",1,0)
IStateGov <- ifelse(JobClass=="StateGov",1,0)
# (b)
mlr2 <- lm(log(WeeklyEarnings)~Age+factor(Sex)+factor(MaritalStatus)+EdCode+
             IMidwest+INortheast+ISouth+
             IMetropolitan+INotMetropolitan+
             IFedGov+ILocalGov+IStateGov)
summary(mlr2)
# (c)
mlr2.reduce <- 
lm(log(WeeklyEarnings)~Age+factor(Sex)+factor(MaritalStatus)+EdCode+
                    IMidwest+INortheast+ISouth+
                    IMetropolitan+INotMetropolitan)
anova(mlr2.reduce, mlr2, test="F")
# (d)
anova(mlr2)
mlr2.inter <- 
lm(log(WeeklyEarnings)~Age+factor(Sex)+factor(MaritalStatus)+EdCode+
                   IMidwest+INortheast+ISouth+
                   IMetropolitan+INotMetropolitan+
                   IFedGov+ILocalGov+IStateGov+
                   factor(Sex)*factor(MaritalStatus))
anova(mlr2.inter)
# (e)
summary(mlr2.inter)
```

Question 4 (1.5 points)

a

```r
set.seed(7001)
beta0 <- 2
beta1 <- 1
beta2 <- -1
n <- 100
R <- 1000
hatbeta0 <- rep(0,R)
```

	`` hatbeta1 <- rep(0,R) hatbeta2 <- rep(0,R) responses <- rep(0,R) x0 <- data.frame(X1=2.5,X2=0) CIs <- NULL X2 <- rt(n,3) for (r in 1:R){ X1 <- 1:n errors <- rnorm(n) Y <- beta0+beta1*X1+beta2*X2+errors sim.mlr <- lm(Y~X1+X2) hatbeta0[r] <- sim.mlr$coef[1] hatbeta1[r] <- sim.mlr$coef[2] hatbeta2[r] <- sim.mlr$coef[3] CIs <- rbind(CIs,predict(sim.mlr,x0,interval='confidence',level=0.95)) responses[r] <- sim.mlr$coef[1] + sim.mlr$coef[2]*2.5 + sim.mlr$coef[3]*0 } mean(hatbeta0) mean(hatbeta1) mean(hatbeta2) mean(responses) # 4.502998 ```
b	`theo.response <- 2+1*2.5+(-1)*0` `sum(theo.response > CIs[,2] & theo.response < CIs[,3])` The answer is 948.
c	Based on the previous information, if we resample, we could approximately find out that 95% of the intervals would contain the population mean.