2B03 Bonus Assignment 5

Regression Modeling and Prediction (Chapters 10 & 11)

Your Name and Student ID

Due Thursday November 29 2018 (due in class prior to the start of the lecture)

Instructions: *You are to use R Markdown for generating your assignment (see the item Assignments and R Markdown on the course website for helpful tips and pointers).* This assignment is a pure bonus opportunity. It will be worth 5% of your final grade. If you choose to not attempt the assignment it will not be counted against you.

1. Define the following terms in a sentence (or *short* paragraph) and state a formula if appropriate (this question is worth 5 marks).

   i. Test of Significance

      From data, we can create hypothesis test for any variables. Test these hypothesis test can be rejected or accepted based on p-value is called test of significance.

   ii. Coefficient of Determination

      The coefficient of determination R^2 is measured how much variance explained by model from total variance.

   iii. Multiple Regression Analysis

      We can find relationship between several independent variables with a dependent variable.

   iv. Individual Prediction Interval

      Based on the variable mean and variance, we can conduct a confidence interval for this individual prediction value: Denoted individual prediction value as Y_i, for i = 1,..n where n is data size. The interval is:

      $$[Y_i - t_{\alpha/2,n-1} \cdot \sqrt{MSE + se(\hat{y_i})^2}, Y_i + t_{\alpha/2,n-1} \cdot \sqrt{MSE + se(\hat{y_i})^2}]$$

      Where MSE is the mean square error, α is the significant level, n is the data size and $se(\hat{y_i})$ is the standard error of the predicted value.

2. An economist is studying the relationship between unemployment and inflation, and has collected the following data. Inflation appears in columns, unemployment in rows (this question is worth 5 marks).

		Inflation		
Unemployment	Abated	Unchanged	Accelerated	Total
Lower	5	5	10	20
Unchanged	5	35	20	60
Higher	20	0	0	20
Total	30	40	30	100

   The data in the table above summarize the relationship between unemployment and inflation based on 100 months of data. For instance, for 35 months, inflation and unemployment were unchanged, while for 10 months inflation had accelerated and unemployment was lower.

   Using the data in the table above, conduct an appropriate hypothesis test of independence between inflation and unemployment at the 5% level of significance.

```
Unemployment = c('Lower', 'Unchanged', 'Higher')
Inflation.Abated = c(5, 5, 20)
Inflation.Unchanged = c(5, 35, 0)
Inflation.Accelerated = c(10, 20, 0)

df = as.table(cbind(Inflation.Abated, Inflation.Unchanged, Inflation.Accelerated))

dimnames(df) = list(Unemployment = Unemployment, Inflation = c('Abated', 'Unchanged', 'Accelerated'
df
```

```
##             Inflation
## Unemployment Abated Unchanged Accelerated
##    Lower          5         5          10
##    Unchanged      5        35          20
##    Higher        20         0           0
```

```
chisq.test(df)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df
## X-squared = 65.278, df = 4, p-value = 2.249e-13
```

Based on the contingency table we can conduct a chi square test between Unemployment and Inflation

H_0: Unemployment is dependent on Inflation.

H_1: Unemployment is independent on Inflation

Note that the default the significant level for chisq.test in R is 0.05. Then the p-value $= 2.249 * 10^{-13} < 0.05$ we reject the H_0 at 0.05 significant level.

1. Consider the following dataset on the final grade received in a particular course (`grade`) and attendance (`attend`, number of times present when work was handed back during the semester out of a maximum of six times). Note that R has the ability to read datafiles directly from a URL, so here (unlike the `odesi` data that you manually retrieved) you do not have to manually download the data *providing you are connected to the internet* (this question is worth 5 marks).

```
course <- read.table("https://socialsciences.mcmaster.ca/racinej/2B03/files/attend.RData")
attach(course)
```

i. Run a regression of `grade` on `attend` using the R command `lm()`. What is the impact of a 1 unit increase on `attend` on `grade` based on your model?

```
y = lm(grade ~ attend)

summary(y)
```

```
##
## Call:
## lm(formula = grade ~ attend)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.527  -9.344   2.473   9.473  31.262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   46.315      4.506  10.278 8.05e-14 ***
## attend         6.106      1.152   5.302 2.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.23 on 49 degrees of freedom
## Multiple R-squared:  0.3646, Adjusted R-squared:  0.3516
## F-statistic: 28.11 on 1 and 49 DF,  p-value: 2.724e-06
```

The linear regression formula based on the R summary is

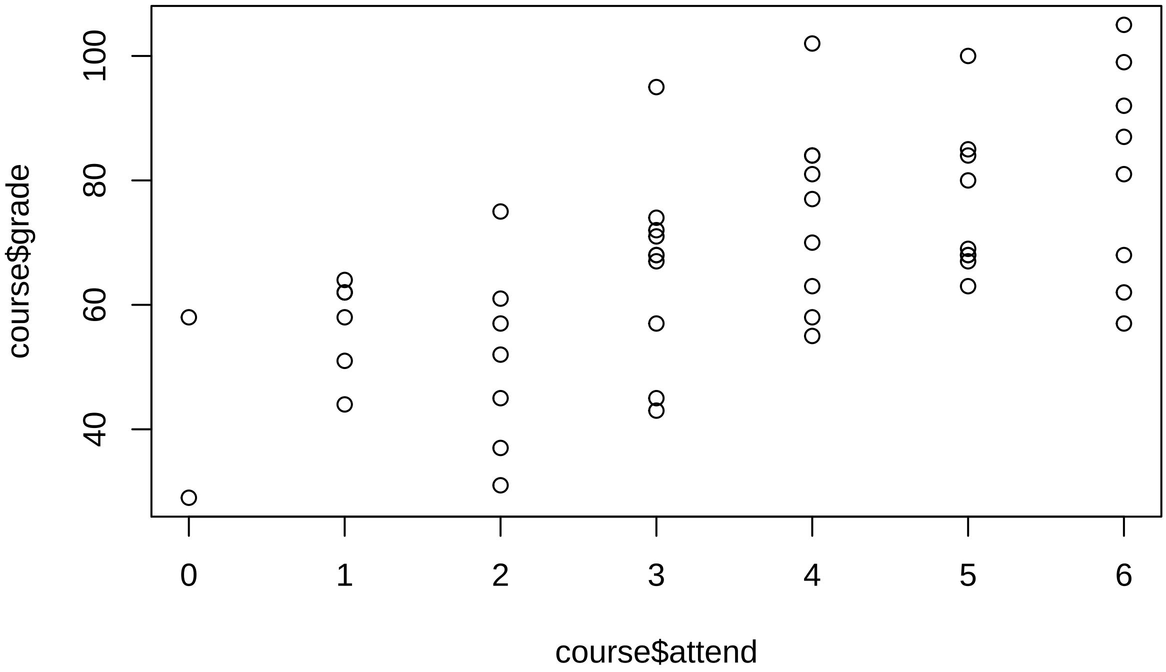
$$grade = 6.106attend + 46.315$$

Hence 1 unit increase on attend will increase grade by $6.106 \cdot 1 = 6.106$

i. In class we distinguished between correlation and causation and cautioned against inferring causation from statistical correlation. Do your results suggest that an individual who increased their attendance by 1 unit would also experience an increase in their expected grade? Why or why not? Explain the roles of *confounders* in this context (e.g. along the lines of Sir. R. A. Fisher's concerns).

Since the correlation describes the relationship between two or more variables and the causation is a event will cause another event. Also we can look at the plot from crouse data that grade against attend.

```
plot(course$attend, course$grade)
```



Although, we can find based on the linear regression, there is a positive relationship between attend and grade but for an individual we have no information can show that his attend may lead an increase or not in his grade. The attendence may only give a chance for student to study which is not guarentee the grade will rise. In addition, the original data is for different student's grade against their attendence which can not gives us clear information about the individual's grade against attend.

1. This question requires you to download data obtained from Statistics Canada. If you are working on campus go to www.odesi.ca (off campus users must first sign into the McMaster library via libaccess at library.mcmaster.ca/libaccess, search for odesi via the library search facilities then select odesi from these search results). Next, select the "Find data" field in odesi and search for "Labour Force Survey June, 2018", then scroll down and select the *Labour Force Survey, June 2018 [Canada]*. Next click on the "Explore & Download" icon, then click on the download icon (i.e., the *diskette* icon, square, along the upper right of the browser pane) and then click on "Select Data Format" then scroll down and select "Comma Separated Value file" (csv) which, after a brief pause, will download the data to your hard drive (you may have to extract the file from a zip archive depending on which operating system you are using). Finally, make sure that you place this csv file in the same directory as your R code file (this file ought to have the name *lfs-71M0001-E-2018-june_F1.csv*, and in RStudio select the menu item Session -> Set Working Directory -> To Source File Location). There will be another file with (almost) the same name but with the extension .pdf that is the pdf documentation that describes the variables in this data set. Note that it would be prudent to retain this file as we will use it in future assignments (this question is worth 10 marks).

Next, open RStudio, make sure this csv file and your R Markdown script are in the same directory (in RStudio open the Files tab (lower right pane by default) and refresh the file listing if necessary). Then read the file as follows:

```r
lfp <- read.csv("LFS-71M0001-E-2019-June_F1.csv")
```

This data set contains some interesting variables on the labour force status of a random subset of Canadians. We will focus on the variable `HRLYEARN` (hourly earnings) described on page 24 of the pdf file lfs-71M0001-E-2018-june.pdf. We will also consider other variables so that we can conduct multiple regression analysis.

i. Following assignment 1, consider hourly earnings and highest educational attainment for people in the survey and consider both high school graduates (`EDUC==2`) and those holding a bachelors degree (`EDUC==5`). To construct these subsets we can use the R command `subset` as follows (the ampersand is the logical operator *and* - see `?subset` for details on the `subset` command):

```r
hs <- subset(lfp, FTPTMAIN==1 & EDUC==2 & HRLYEARN > 0)$HRLYEARN
ba <- subset(lfp, FTPTMAIN==1 & EDUC==5 & HRLYEARN > 0)$HRLYEARN
```

These commands simply tell R to take a subset of the data frame `lfp` for full-time workers having either a high school diploma or university bachelors degree for those reporting positive earnings, and then retain only the variable `HRLYEARN` and store these in the variables named `hs` (hourly earnings for high-school graduates) or `ba` (hourly earnings for university graduates).

Conduct a *t*-test of the hypothesis that the average wage is equal for the two groups using the R function `t.test()` (see `?t.test()` for details).

```r
t.test(hs, ba, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  hs and ba
## t = -54.525, df = 15313, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.05528 -10.28801
## sample estimates:
## mean of x mean of y
##  23.50020  34.17184
```

H_0: two groups have same average wage

H_1: two groups do not have same average wage

Since p-value $2.2 * 10^{-16} < 0.05$, then we can reject H_0 at 0.05 significant level.

i. Using the PDF that accompanies the data titled "Canada Labour Statistics Division, Statistics Canada Labour Force Survey, June 2018 [Canada], Study Documentation, present a description of each variable that we select below (i.e., `HRLYEARN, EDUC, SEX, AGE_12, MARSTAT, UNION`; see pages 15 and on for details).

```r
foo <- subset(lfp, FTPTMAIN==1 & HRLYEARN > 0,
              select = c(HRLYEARN,EDUC,SEX,AGE_12,MARSTAT,UNION))
```

HRLYEARN represents for Usual hourly wages and is postive continuous value

EDUC represents for Highest educational attainment and is numeric discrete value from 0 to 6 means 0 to 8 years, Some high school, High school graduate, Some postsecondary, Postsecondary certificate or diploma, Bachelor's degree and Above bachelor's degree

SEX is category discrete value 1 and 2 represents for male and female

AGE_12 is discrete value from 1 to 12 represents for the 15 to 19 years, 20 to 24 years, .., 65 to 69 years, 70 & over.

MARSTAT represents for Marital status of respondent and is discrete value from range 1 to 6 represent for Married, Living in common-law, Widowed, Separated, Divorced and Single, never married

UNION represents for Union status is the discrete numeric value from range 1 to 3 means for Union member, Not a member but covered by a union contract or collective and Non-unionized

i. Estimate the multivariate linear regression model and produce a model summary via `summary(model)`.

```
model <- lm(HRLYEARN~EDUC+AGE_12+factor(SEX)+factor(MARSTAT)+factor(UNION),data=foo)
summary(model)
```

```
##
## Call:
## lm(formula = HRLYEARN ~ EDUC + AGE_12 + factor(SEX) + factor(MARSTAT) +
##     factor(UNION), data = foo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.845  -7.894  -1.962   5.828  78.869
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      18.45172    0.28938  63.763  < 2e-16 ***
## EDUC              3.45415    0.04157  83.092  < 2e-16 ***
## AGE_12            0.61311    0.02584  23.723  < 2e-16 ***
## factor(SEX)2     -4.77566    0.11703 -40.809  < 2e-16 ***
## factor(MARSTAT)2 -1.56908    0.16584  -9.461  < 2e-16 ***
## factor(MARSTAT)3 -3.46673    0.62011  -5.590 2.28e-08 ***
## factor(MARSTAT)4 -2.26681    0.34882  -6.498 8.20e-11 ***
## factor(MARSTAT)5 -2.01411    0.28192  -7.144 9.19e-13 ***
## factor(MARSTAT)6 -4.61922    0.15915 -29.024  < 2e-16 ***
## factor(UNION)2   -0.09259    0.40589  -0.228     0.82
## factor(UNION)3   -3.36436    0.12565 -26.776  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.06 on 43887 degrees of freedom
## Multiple R-squared:  0.2257, Adjusted R-squared:  0.2256
## F-statistic:  1280 on 10 and 43887 DF,  p-value: < 2.2e-16
```

What is the coefficient of determination? Interpret this result.

Coefficient of determination Multiple R-squared: 0.2257 which show the porprotion of the variance explained by the model over total variance. And this value is not close to the 1 menas the HRLYEARN do not have a strong linear relationship with other independent variables.

Interpret the coefficient estimates. Note that `factor(SEX)2` is the dummy variable that represents the difference between wages between group 2 and the base group group 1, other things equal (i.e., 2=Female, 1=Male). Also, note that AGE_12 contains five year ranges, so the coefficient is the change in expected wages for an additional five years of age, other things equal.

$$
\begin{aligned}
HRLYEARN = {} & 3.45415 * EDUC + 0.61311 * AGE_{12} - 4.77566 * Female - 1.56908 * Living\_in\_common\_law \\
& - 3.46673 * Widowed - 2.26681 * Separated - 2.01411 * Divorced\_and\_Single \\
& - 4.61922 * never\_married - 0.09259 * Not\_member\_covered\_by\_union \\
& - 3.36436 * Non\_unionized + 18.45172
\end{aligned}
$$

What is the impact of either being covered but not unionized or being non-unionized relative to unionized workers? Are they both significant? Does this make sense?

Since the p-value for factor(UNION)2 (either being covered but not unionized workers) is 0.82 which is larger than 0.05. Hence we do not reject the H_0: Coefficient estimate for either being covered but not unionized workers is equal to 0 at 0.05 significant level.

Same for factor(UNION)3 (non-unionized workers) p-value is $2*10^{-16} < 0.05$, then we reject H_0: Coefficient estimate for non-unionized workers is equal to 0 at 0.05 significant level.

Since the Coefficients estimate for factor(UNION)2 is -0.09259 which is very less compare to the change of non-unionized workers, then factor(UNION)2 is not significant at 0.05 level makes sense.

Conduct a test of significance for the variable EDUC. Show all steps and interpret the results.

Conduct a test of significance for the variable EDUC:

$$
\begin{aligned}
H_0 &: \text{coefficient estimate for EDUC is 0} \\
H_1 &: \text{coefficient estimate for EDUC is not 0}
\end{aligned}
$$

Since the p-value for EDUC is $2*10^{-16}$ in the model summary table and it is less than 0.05, we can conclude that we reject H_0 at 0.05 significant level.