Australian National University

Research School of Finance Actuarial Studies & Statistics
ANU College of Business and Economics
Australian National University
Canberra ACT 2601
https://www.rsfas.anu.edu.au/

RSFAS ASSIGNMENT COVER SHEET

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. If you have elected to submit in hard copy rather than Turnitin, you must provide copies of all references included in the assessment item.

Student ID: u7031432

Course Code and Name: STAT3015/STAT7030 Generalised Linear Model

Assignment Number: Assignment 1

Assignment Due Date: Aug 31, 2020

Lecturer: Professor Andy Wood

Tutor:

Tutorial number, day and time:

Word Count:

I declare that this work:

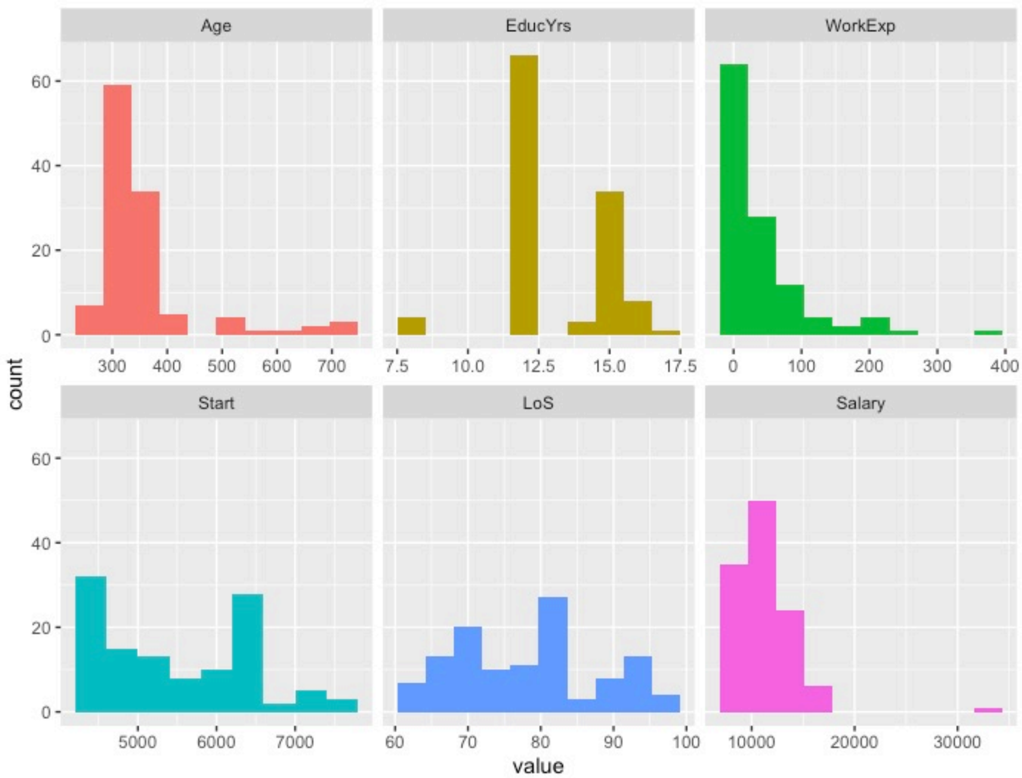
- upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Students University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle 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.

Signed: ZIHAO LI

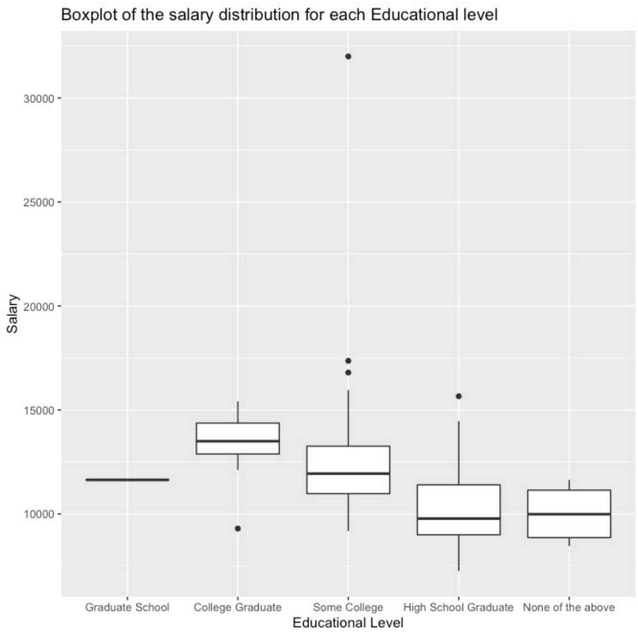
Dated Submitted: Aug 28 2020

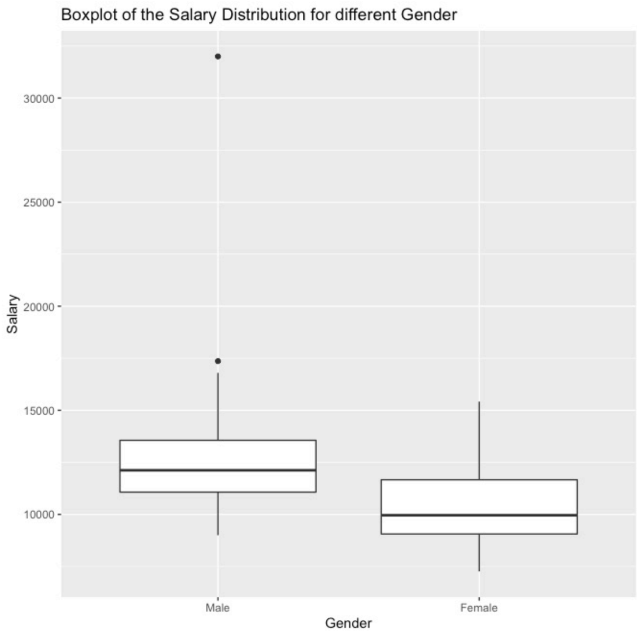
1(a)

My first step of Exploratory Data Analysis on Law Firm Dataset is to generate a visualized plot for all variables exclude the first column X and another two factors, EducLvl and Gender. From the output given below, we can find clear evidence of positive skewness in histogram of Age, WorkExp and Salary. That tells us the right tail of the distribution is longer and majority is most likely to occur on the left hand side of the distribution. Another point we can notice is the histograms for Starting Salary and Length of Service are bimodal, there are also no gaps occur on their x-axes. So far, we still have not found any associations between variables.

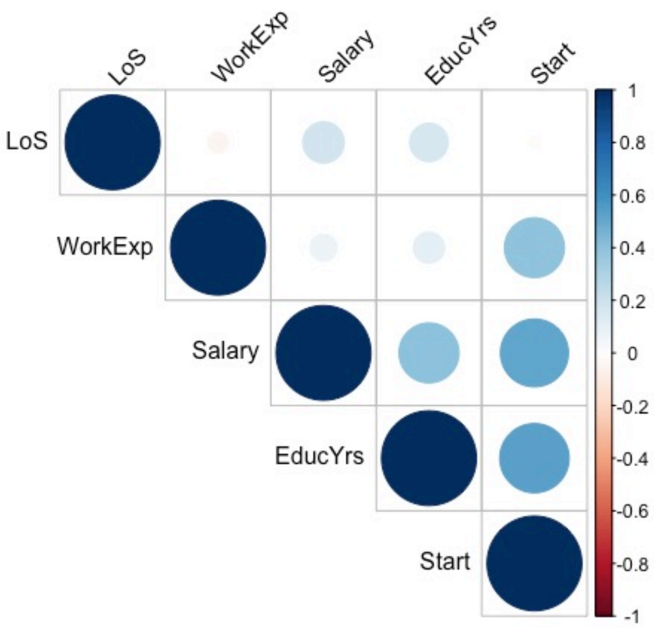


Then, I generate boxplots for distribution of Salary based on the different Genders and multiple levels of Education. From the output below, I found College Graduate has highest median value of Salary, but its range or spread is not wide comparing to Some College or High School Graduate. And Male seems to have higher Salary than Female. We can observe the outliers in both Salary distributions for Males and people from Some College.



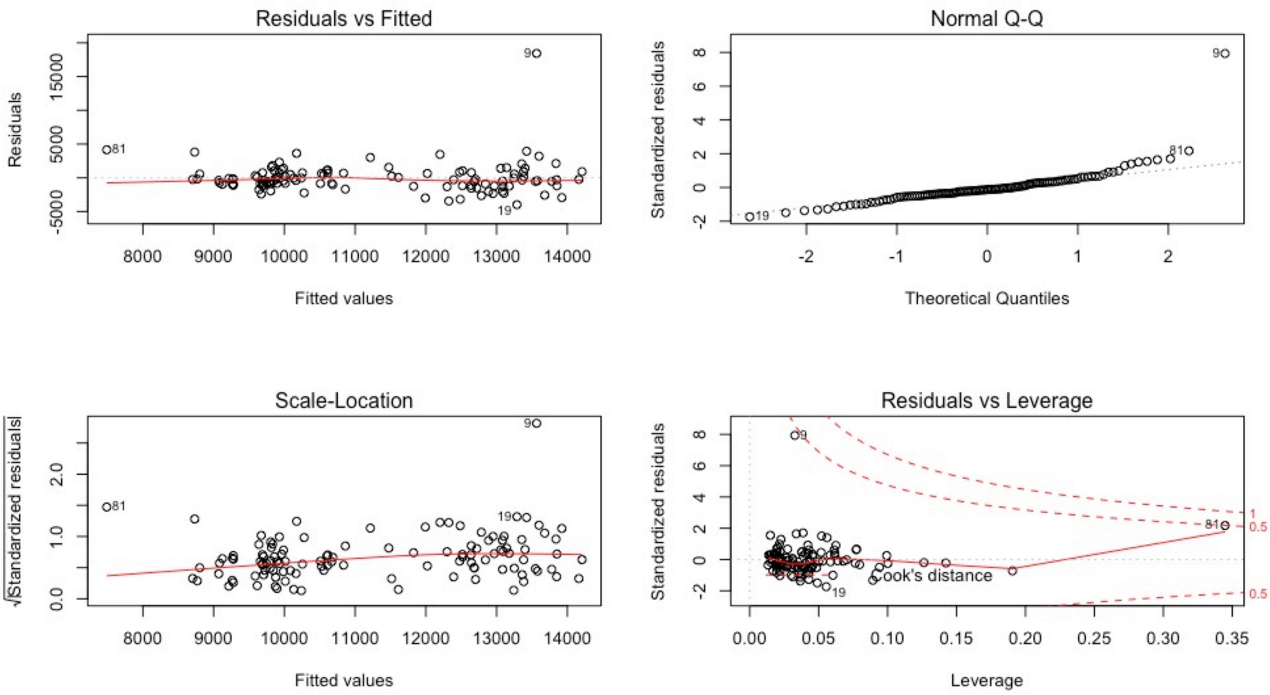


In order to investigate the relationships between numerical variables and our target variable, I generate a plot of correlation for our dataset. As you can see from the image below, years of education and Starting salary tend to have stronger relationship with Salary. Numerically, correlation coefficient between Starting salary and Salary is 0.51. That suggests there is a moderate positive relationship between Starting salary and Salary. At this stage, I possibly think Starting Salary would most likely become the determinant of Salary.



1(b)

After applying AIC, the output shows that the ‘maximal’ model should be Salary = 1.427*Start + 51.467*LoS – 1308.262*Gender (Female) – 7.603*WorkExp + 614.055. However, after examining our ‘maximal’ model with summary statistics, I found that WorkExp seems not that statistically significant. The reason is that its p-value is 0.0553, which is greater than 0.05. So I try to test if my concern stays when I remove the WorkExp from the model. The output shows the new model without WorkExp will decrease the R-Squared value. Based on that, I decide to keep the best model as same as the result from StepAIC. Then, I generate the diagnostic plots for the ‘maximal’ model, and result is listed below.



As we can see in the plot of Residuals vs. Fitted Value, there is no obvious pattern for the residual points spread around the horizontal regression line. That checks the assumption of linear relationship. Then, we can see majority of the residual points are lying on the red dash line, that also checks the assumption of residuals are normally distributed. In the third plot, horizontal line with equally spread points tells that it’s a good indication of assumption of homoscedasticity. On the plot of Residuals vs. Leverage, observation #81 is on the dash line of 0.5 as well as observation #9 is very close to it. The most of the residuals are clustered on the left and some of them have passed the dash line of Cook’s distance. After further check with the plot of Cook’s distance, it has identified two influential observations as #81 and #9.

1(c)

Based on the result from previous part, our optimal model suggested by stepAIC function is Salary ~ Start + LoS + Gender + WorkExp. I will call this model as the base model in my following presentation about adding new interaction terms to our final model.

	Model Name	AIC Value	Model Structure
A	fitlm1	2138.08	BaseModel
B	fitlm2	2138.141	BaseModel+EducYrs
C	fitlm3	2139.472	BaseModel+EducLvl
D	fitlm4	2140.065	BaseModel+Age
E	fitlm5	2141.45	BaseModel+EducYrs+EducLvl
F	fitlm6	2140.14	BaseModel+EducYrs+Age
G	fitlm7	2143.281	BaseModel+EducYrs+EducLvl+Age

In the output above, we can see our base model has the lowest AIC value comparing to other transformations of the base model. Noticeably, the second best model could be the base model adding EducYrs as its extra interaction term. It has the second lowest AIC value, and that's why I start to test if extra interaction terms contribute to the 'fitlm2' model. Data clearly shows us that adding new interaction terms won't help on decreasing the AIC value. So my response for this part is it possibly be the case that including EducYrs in the final model can generate low AIC value, but it's still not perfect as our base model.

1(d)

In this part, I decide to explore more on the model selection procedure other than AIC. After some searching on Google, I learn how to use 'leaps' package to apply some methods on model selection. The main method I used to generate the optimal model is using exhaustive search based on Bayesian Information Criterion, which determines the best model based on minimum BIC value. The output generated by BIC method suggests the optimal model only includes two variables, 'Start' and 'LoS'. Equation of the model is:

Salary = -1790.025 + 1.55 * Start + 57.55 * LoS.

The reason for the different output is that BIC is similar to AIC, but has larger penalty and it generally picks a smaller model than AIC, when a reasonable sample size is given.

1(e)

In the last section, I decide to standardize all the numeric variables excluding the Gender from the optimal model suggested in part(c). Then, I try to fit all the scaled variables into a new linear model. And then summarize it to seek the maximum of absolute value of standardized variables. In another word, we try to transform all the variables with the same unit scale and compare them directly. The variable with the biggest absolute number (standardized coefficient) will have the most impact on our dependent variable (Salary).

```
Call:
lm(formula = scale(Salary) ~ g + scale(Start) + scale(LoS) +
    scale(WorkExp), data = LawSalData)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3856 -0.3598 -0.0997  0.2538  6.4056

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.31744    0.15579   2.038   0.0440 *
g2             -0.45460    0.19459  -2.336   0.0213 *
scale(Start)    0.47304    0.09353   5.057 1.69e-06 ***
scale(LoS)      0.17461    0.07688   2.271   0.0251 *
scale(WorkExp) -0.16276    0.08405  -1.937   0.0553 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8207 on 111 degrees of freedom
Multiple R-squared:  0.3498,    Adjusted R-squared:  0.3264
F-statistic: 14.93 on 4 and 111 DF,  p-value: 8.576e-10
```

The output above tells us the Starting Salary contribute the most impacts on the current Salary. After that, I try to validate this answer by checking adding which variable will contribute the biggest increment on the R-squared value. The data shows adding the Starting Salary to the model can lead the biggest increment to the R-squared value. And Gender has the second biggest contribution on increasing R-squared value. To be more clearly, it also means adding Starting Salary and Gender to the model will provide a better fit and explain more information on the change of the model. In conclusion, I think the gender and Starting Salary are the most important determinants of Salary.

Appendix

- Code for applying the StepAIC() function and its result in part(b)

```
> full<-lm(Salary ~ ., data = LawSalData)
>
> stepAIC(full,direction = "backward")
Start:  AIC=1813.84
Salary ~ X + Gender + Age + EducYrs + EducLvl + WorkExp + Start +
    LoS

          Df Sum of Sq       RSS    AIC
- EducLvl  4  24983592 607735343 1810.7
- EducYrs  1     78520 582830271 1811.9
- Age      1    800978 583552729 1812.0
- X        1   1227524 583979275 1812.1
<none>                 582751751 1813.8
- WorkExp  1  13840700 596592451 1814.6
- LoS      1  18942671 601694422 1815.5
- Gender   1  21976949 604728700 1816.1
- Start    1  62819790 645571541 1823.7

Step:  AIC=1810.71
Salary ~ X + Gender + Age + EducYrs + WorkExp + Start + LoS

          Df Sum of Sq       RSS    AIC
- Age      1      7329 607742673 1808.7
- X        1   1223948 608959292 1809.0
- WorkExp  1   4748355 612483698 1809.6
- EducYrs  1  10028759 617764102 1810.6
<none>                 607735343 1810.7
- LoS      1  19036638 626771981 1812.3
- Gender   1  26471801 634207144 1813.7
- Start    1  63162549 670897892 1820.2

Step:  AIC=1808.71
Salary ~ X + Gender + EducYrs + WorkExp + Start + LoS

          Df Sum of Sq       RSS    AIC
- X        1   1225866 608968538 1807.0
- EducYrs  1  10107698 617850371 1808.6
<none>                 607742673 1808.7
- WorkExp  1  17110892 624853565 1809.9
- LoS      1  21075618 628818291 1810.7
- Gender   1  31831277 639573950 1812.6
- Start    1  72406215 680148887 1819.8

Step:  AIC=1806.95
Salary ~ Gender + EducYrs + WorkExp + Start + LoS

          Df Sum of Sq       RSS    AIC
- EducYrs  1  10262395 619230934 1806.9
<none>                 608968538 1807.0
- WorkExp  1  17021871 625990409 1808.2
- LoS      1  21341708 630310246 1808.9
- Gender   1  31794195 640762733 1810.8
- Start    1  73831203 682799741 1818.2

Step:  AIC=1806.89
```

```
Salary ~ Gender + WorkExp + Start + LoS

          Df Sum of Sq       RSS    AIC
<none>                 619230934 1806.9
- WorkExp  1  20920468 640151402 1808.7
- LoS      1  28777168 648008102 1810.2
- Gender   1  30446717 649677651 1810.5
- Start    1 142685654 761916587 1828.9

Call:
lm(formula = Salary ~ Gender + WorkExp + Start + LoS, data = LawSalDat
a)

Coefficients:
(Intercept)      Gender2      WorkExp        Start          LoS
    614.055    -1308.262       -7.603        1.427       51.467
```

- Code for applying AIC() function and generate the output table in part(c)

```
> obj<-matrix(c("fitlm1",2138.08,"BaseModel"))
> obj
     [,1]
[1,] "fitlm1"
[2,] "2138.08"
[3,] "BaseModel"
> obj<-matrix(c("fitlm1",2138.08,"BaseModel","fitlm2",2138.141,"BaseMod
el+EducYrs","fitlm3",2139.472,"BaseModel+EducLvl","fitlm4",2140.065,"Ba
seModel+Age","fitlm5",2141.45,"BaseModel+EducYrs+EducLvl","fitlm6",214
0.14,"BaseModel+EducYrs+Age","fitlm7",2143.281,"BaseModel+EducYrs+EducL
vl+Age"),ncol=3,byrow=TRUE)
> obj
     [,1]     [,2]       [,3]
[1,] "fitlm1" "2138.08"  "BaseModel"
[2,] "fitlm2" "2138.141" "BaseModel+EducYrs"
[3,] "fitlm3" "2139.472" "BaseModel+EducLvl"
[4,] "fitlm4" "2140.065" "BaseModel+Age"
[5,] "fitlm5" "2141.45"  "BaseModel+EducYrs+EducLvl"
[6,] "fitlm6" "2140.14"  "BaseModel+EducYrs+Age"
[7,] "fitlm7" "2143.281" "BaseModel+EducYrs+EducLvl+Age"
> colnames(obj)<-c("Model Name","AIC Value","Model Structure")
> obj<-as.table(obj)
```

- Code for alternative of model selection in part (d), mine is to use exhaustive search method and determine the best model based on Bayesian Information Criterion.

```
> library(leaps)
> best_subset<-regsubsets(Salary~.,LawSalData)
> results<-summary(best_subset)
> best_subset
Subset selection object
Call: regsubsets.formula(Salary ~ ., LawSalData)
11 Variables  (and intercept)
         Forced in Forced out
X            FALSE      FALSE
Gender2      FALSE      FALSE
Age          FALSE      FALSE
EducYrs      FALSE      FALSE
EducLvl2     FALSE      FALSE
EducLvl3     FALSE      FALSE
EducLvl4     FALSE      FALSE
EducLvl5     FALSE      FALSE
WorkExp      FALSE      FALSE
Start        FALSE      FALSE
LoS          FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
> BIC=results$bic
> which.min(results$bic)
[1] 2
> coef(best)
Error in coef(best) : object 'best' not found
> coef(best_subset,2)
(Intercept)       Start         LoS
-1790.02480     1.55330    57.55438
> best_subset<-regsubsets(Salary~.,LawSalData,method = "exhaustive")
> results<-summary(best_subset)

> BIC=results$bic
> which.min(results$bic)
[1] 2
> coef(best_subset,2)
(Intercept)       Start         LoS
-1790.02480     1.55330    57.55438
```