Institute of Actuaries of India

Subject CS1-Paper B – Actuarial Statistics

June 2019 Examination

INDICATIVE SOLUTION

Introduction

The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiner have given credit for any alternative approach or interpretation which they consider to be reasonable.

Solution 1:

i)

```
state_mean<-aggregate(PAID~STATE,data = AutoClaims,FUN = mean)
names(state_mean)<-c("State","Mean")
state_sd<-aggregate(PAID~STATE,data = AutoClaims,FUN = sd)
names(state_sd)<-c("State","SD")
state_summary<-merge(state_mean,state_sd)
state_summary$CV<-state_summary$SD/state_summary$Mean
#Mean, Standard Deviation and Coefficient of Variance for each state
state_summary                                                   [4]
```

```
##       State      Mean        SD        CV
## 1  STATE 01 10235.800 10932.877 1.0681018
## 2  STATE 02  7055.078  6327.473 0.8968678
## 3  STATE 03  8714.932  6494.346 0.7451976
## 4  STATE 04  8152.759  6985.210 0.8567910
## 5  STATE 06  8786.739 10749.517 1.2233796
## 6  STATE 07  4960.479  3065.092 0.6179023
## 7  STATE 10 12340.643 14599.291 1.1830251
## 8  STATE 12  6893.705  8634.955 1.2525856
## 9  STATE 14 10399.313  8406.388 0.8083599
## 10 STATE 15  3321.449  3364.269 1.0128920
## 11 STATE 17  7886.282  7831.913 0.9931059
```

Note: One mark is awarded for each correct column

```
# a. The State with minimum coefficient of variance is
state_summary$State[state_summary$CV==min(state_summary$CV)]
```

```
## [1] STATE 07                                                   [0.5]
## 11 Levels: STATE 01 STATE 02 STATE 03 STATE 04 STATE 06 ... STATE 17
```

```
# b. The State with maximum coefficient of variance is
state_summary$State[state_summary$CV==max(state_summary$CV)]
```

```
## [1] STATE 12                                                   [0.5]
## 11 Levels: STATE 01 STATE 02 STATE 03 STATE 04 STATE 06 ... STATE 17
```

Note: Even if the code is not present for finding the state with max/min CV and the student has written the answer by observation, mark can be awarded

```
Class_mean<-aggregate(PAID~CLASS,data = AutoClaims,FUN = mean)
names(Class_mean)<-c("Class","Mean")
Class_sd<-aggregate(PAID~CLASS,data = AutoClaims,FUN = sd)
names(Class_sd)<-c("Class","SD")
Class_summary<-merge(Class_mean,Class_sd)
Class_summary$CV<-Class_summary$SD/Class_summary$Mean
#Mean, Standard Deviation and Coefficient of variance for each rating class
Class_summary                                                   [3]
```

```
##    Class      Mean       SD        CV
## 1     C1 17464.484 9613.323 0.5504499
## 2    C11  5887.049 3434.371 0.5833774
## 3   C6   2367.215 1533.807 0.6479373
## 4   F6  17434.533 9188.423 0.5270243
```

```
# Arranging the rating class in ascending order of CV
Class_summary[order(Class_summary$CV),]
```

[2]

[10]

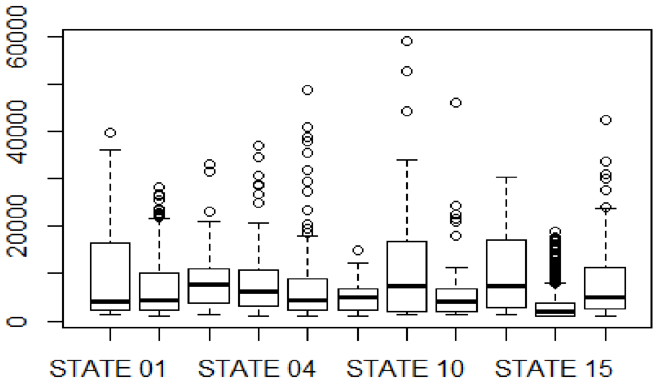
```
##   Class     Mean       SD       CV
## 4   F6  17434.533 9188.423 0.5270243
## 1   C1  17464.484 9613.323 0.5504499
## 2   C11  5887.049 3434.371 0.5833774
## 3   C6   2367.215 1533.807 0.6479373
```

Note: Full marks (5) are awarded if the student directly shows the table in ascending of the CVs

ii)

```
boxplot(PAID~STATE,data = AutoClaims)
```

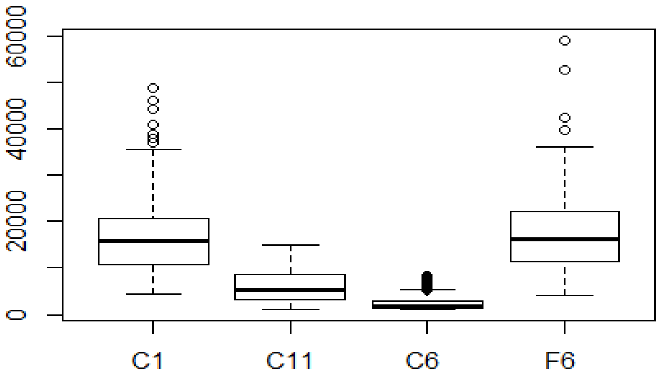
[1.5]



States with no outliers are **State 14.** [1]

```
boxplot(PAID~CLASS,data = AutoClaims)
```

[1.5]



Class with no outlier is **C11.** [1]

[5]

iii)

```
#Contingency Table for the distribution of Gender across rating classes
State_Class_Freq<-table(AutoClaims$GENDER,AutoClaims$CLASS)                    [2]
#Contingency Table in terms of proportion of Rating classes for each gender
State_Class_RowProp<-prop.table(State_Class_Freq,margin = 1)
State_Class_RowProp                                                            [2]
```

```
##
##           C1         C11        C6         F6
##   F 0.07312925 0.30102041 0.55272109 0.07312925
##   M 0.07840772 0.34620024 0.46200241 0.11338963
```

```
#Chi-Square test for independence of two variables
chisq.test(AutoClaims$GENDER,AutoClaims$CLASS)                                 [3]
```

```
##
##  Pearson's Chi-squared test
##
## data:  AutoClaims$GENDER and AutoClaims$CLASS
## X-squared = 13.704, df = 3, p-value = 0.003338
```

Interpretation: Chi-Squared values is 13.704 with 3 degrees of freedom. P-Value for the chi square test is 0.0033 < 0.05. Hence the null hypothesis of independence between the gender and rating class is rejected at 95% confidence level. [2]

Looking at the proportions in the table, it is evident that more than 55% of the females are in rating class C6 whereas as only 46% of the males are in that rating class. Similarly 34.6% of the males are in C11 as against only 30% of females in that class. Same is the case with F6. Hence the difference in proportions [1]

[10]

iv)

```
#Testing for mean amount paid between Male and Female
t.test(AutoClaims$PAID[AutoClaims$GENDER=="M"],AutoClaims$PAID[AutoClaims$GENDE
R=="F"])                                                                       [1.5]
```

```
##
##  Welch Two Sample t-test
##
## data:  AutoClaims$PAID[AutoClaims$GENDER == "M"] and
AutoClaims$PAID[AutoClaims$GENDER == "F"]
## t = -1.0808, df = 1165.9, p-value = 0.28
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1176.7935   340.7988
## sample estimates:
## mean of x mean of y
##  5953.770  6371.767
```

P-value = 0.28 > 0.05 → Null hypothesis of "Mean claim paid is same between males and females" cannot be rejected at 95% confidence level [1]

```
#Testing for variance of amount paid between Male and Female
var.test(AutoClaims$PAID[AutoClaims$GENDER=="M"],AutoClaims$PAID[AutoClaims$GEND
ER=="F"])                                                                      [1.5]
```

```
##
##  F test to compare two variances
##
## data:  AutoClaims$PAID[AutoClaims$GENDER == "M"] and
AutoClaims$PAID[AutoClaims$GENDER == "F"]
## F = 0.78416, num df = 828, denom df = 587, p-value = 0.001337
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6744929 0.9098948
## sample estimates:
## ratio of variances
##         0.7841591
```

P-value = 0.0013 < 0.05 → Null hypothesis of "Variance of claim paid is same between males and females" is rejected at 95% confidence level. Claims paid among females is more volatile compared to that of males [1]

[5]

[30 Marks]

Part (i) was well answered by the students who were successful. Students struggled with computing the coefficient of variance and interpreting based on that. In part (ii), the students were able to use the R Code to generate the box plot. Only a few of them were able to identify outliers successfully. In Part (iii), many students were able to perform Chi square test and do the interpretation successfully but majority of them failed in preparing contingency tables based on proportions. Also majority of them failed in making detailed interpretation based on the results of the chi-square test. Part (iv) was hardly attempted. Very few students were able to perform the correct tests and provide right interpretation. A number of students had difficulty in identifying the correct test itself.

Solution 2:

i)

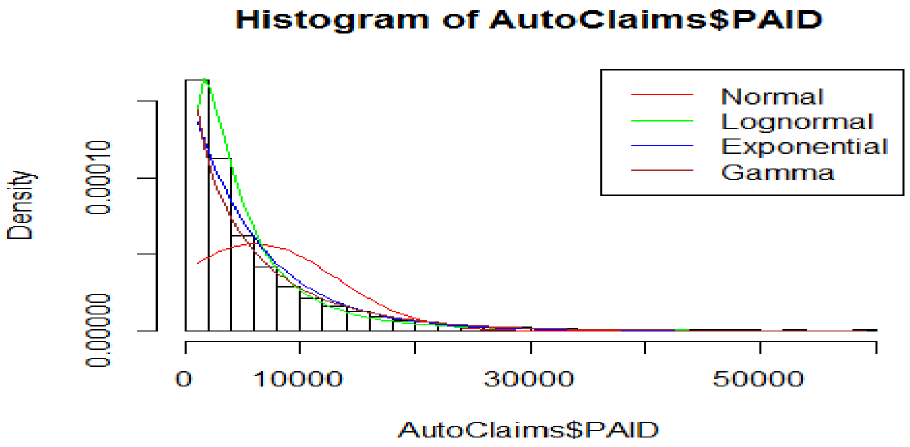
```
sampleMean<-mean(AutoClaims$PAID)
sampleVariance<-var(AutoClaims$PAID)

#Method of Moments Estimates - Normal distribution
Normalmu <- sampleMean                                              [1]
Normalsigma <- sqrt(sampleVariance)                                 [1]
Normalmu
## [1] 6127.222

Normalsigma

## [1] 7027.434

#Method of Moments Estimates - Log Normal distribution
LNsigma<- sqrt(log(1+sampleVariance/sampleMean^2))                  [1]
LNmu<-log(sampleMean)-LNsigma^2/2                                   [1]
LNsigma

## [1] 0.9162933

LNmu

## [1] 8.3007
```

```
#Method of Moments Estimates - Exponential distribution
Exprate <- 1/sampleMean                                                        [2]
Exprate

## [1] 0.0001632061

#Method of Moments Estimates - Gamma distribution
GammaBeta<-sampleMean/sampleVariance                                           [1]
GammaAlpha<-GammaBeta*sampleMean                                               [1]
GammaBeta

## [1] 0.0001240709

GammaAlpha

## [1] 0.7602103
```

[8]

ii)

```
#Histogram of Paid Claims Data
hist(AutoClaims$PAID,breaks = 30,freq = FALSE)                                 [2]
#Superimposing a Normal distribution over the histogram
curve(dnorm(x,mean = Normalmu,sd = Normalsigma),from = min(AutoClaims$PAID), to = max(AutoClaims$PAID), add = TRUE, col= "red")      [1.5]
#Superimposing a Log Normal distribution over the histogram
curve(dlnorm(x,meanlog = LNmu,sdlog = LNsigma),from = min(AutoClaims$PAID), to = max(AutoClaims$PAID), add = TRUE, col= "green")     [1.5]
#Superimposing a Exponential distribution over the histogram
curve(dexp(x,rate = Exprate),from = min(AutoClaims$PAID), to = max(AutoClaims$PAID), add = TRUE, col= "blue")                        [1.5]
#Superimposing a Gamma distribution over the histogram
curve(dgamma(x,shape = GammaAlpha,rate = GammaBeta),from = min(AutoClaims$PAID), to = max(AutoClaims$PAID), add = TRUE, col= "brown") [1.5]

legend("topright",legend = c("Normal", "Lognormal", "Exponential", "Gamma"),lty = 1, col = c("red","green","blue","brown"))
```



[8]

iii)

```
#Computed quantiles of the actual data as well as that of different distributions
quantile(AutoClaims$PAID,c(0.05,0.25,0.5,0.75,0.95))                           [1]
```

```
##       5%       25%       50%       75%       95%
##  1116.379  1610.660  3395.368  7774.382 20405.050
```

```
qnorm(c(0.05,0.25,0.5,0.75,0.95),mean = Normalmu,sd = Normalsigma)                    [1]
```

```
## [1] -5431.878  1387.290  6127.222 10867.155 17686.323
```

```
qlnorm(c(0.05,0.25,0.5,0.75,0.95),meanlog = LNmu,sdlog = LNsigma)                     [1]
```

```
## [1]   892.0584  2170.4062  4026.6904  7470.5996 18176.2032
```

```
qexp(c(0.05,0.25,0.5,0.75,0.95),rate = Exprate)                                       [1]
```

```
## [1]   314.2854  1762.6920  4247.0670  8494.1339 18355.5180
```

```
qgamma(c(0.05,0.25,0.5,0.75,0.95),shape = GammaAlpha,rate = GammaBeta)                [1]
```

```
## [1]   142.0733  1276.6130  3737.9364  8453.3371 20244.3595
```

[5]

iv)

#Comment based on (ii) and (iii)

From the histogram and the superimposed plots, it is clear that normal distribution does not fit the data well. [1]

The other three curves are getting superimposed more or less similarly to the data. Even from the quantiles we observe that lower values are aptly modeled using lognormal distribution (5th percentile of lognormal being close to actual values) whereas gamma distribution is modeling the higher values more appropriately (95th percentile). [2]

Hence, just by looking at (ii) and (iii), best fitting distribution among Lognormal, Gamma and Exponential distributions cannot be concluded. It requires additional analysis in the form of other statistical tests to confirm the best fit

[4]

[25 Marks]

Part (i) was well attempted though many students had difficulty in arriving at the parameters corresponding to lognormal distribution and Gamma distribution. All those who attempted Part (i) successfully were able to attempt Part (ii) and Part (iii) as well, but the failure in fitting a few distributions in part (i) resulted in only part answers for part (ii) and (iii). Many students were able to plot the histogram and the distributions decently well but the labelling through appropriate legends was not done well. Part (iv) was not attempted by many students and among them who attempted, complete interpretation was lagging. Very few wrote a detailed interpretation of the result.

Solution 3:

i)

```
#Fitting a Linear Regression Model
model1<-lm(PAID~.,data = AutoClaims)
summary(model1)                                                                       [5]
```

```
##
## Call:
## lm(formula = PAID ~ ., data = AutoClaims)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10462  -2276    119   1611  36377
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    19818.12    1391.58  14.242  < 2e-16 ***
## STATESTATE 02  -2306.41     658.69  -3.502 0.000477 ***
## STATESTATE 03   -580.09     761.36  -0.762 0.446242
## STATESTATE 04   -689.08     702.04  -0.982 0.326495
## STATESTATE 06    440.79     752.27   0.586 0.558010
## STATESTATE 07  -1254.29     837.22  -1.498 0.134318
## STATESTATE 10   2275.25     885.44   2.570 0.010284 *
## STATESTATE 12   -752.99     850.10  -0.886 0.375897
## STATESTATE 14   -404.69     842.90  -0.480 0.631216
## STATESTATE 15  -4791.86     623.56  -7.685 2.87e-14 ***
## STATESTATE 17   -883.67     704.58  -1.254 0.209982
## CLASSC11      -11743.95     430.60 -27.274  < 2e-16 ***
## CLASSC6       -14833.37     410.84 -36.105  < 2e-16 ***
## CLASSF6         -225.16     517.68  -0.435 0.663670
## GENDERM        -1193.01     215.50  -5.536 3.69e-08 ***
## AGE               15.76      24.10   0.654 0.513418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3943 on 1401 degrees of freedom
## Multiple R-squared:  0.6886, Adjusted R-squared:  0.6852
## F-statistic: 206.5 on 15 and 1401 DF,  p-value: < 2.2e-16
```

Note: 1 Mark can be deducted if the output is not pasted

```
anova(model1)
Analysis of Variance Table

Response: PAID
            Df     Sum Sq    Mean Sq  F value    Pr(>F)
STATE       10 9.8027e+09 9.8027e+08  63.0629 < 2.2e-16 ***
CLASS        3 3.7869e+10 1.2623e+10 812.0763 < 2.2e-16 ***
GENDER       1 4.7271e+08 4.7271e+08  30.4106 4.158e-08 ***
AGE          1 6.6423e+06 6.6423e+06   0.4273    0.5134
Residuals 1401 2.1778e+10 1.5544e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation

R-Squared: 68.86% of the variation in the claims paid is explained by state, rating class, gender and age [1]

Adjusted R-Squared: 68.52% is used to compare with other models, adjusts for the number of terms in the model. We Use adjusted R-squared to compare the goodness-of-fit for regression models that contain differing numbers of independent variables. [1]

p-value of the model is <2.2*E-16 which is less than 0.05 and hence the null hypothesis of "There is no significant linear relationship between the given independent variables X and a dependent variable Y" is rejected at 5% level of significance. Using this model to predict the DV is better than simply using the expected value of the DV as a predictor for the DV [2]

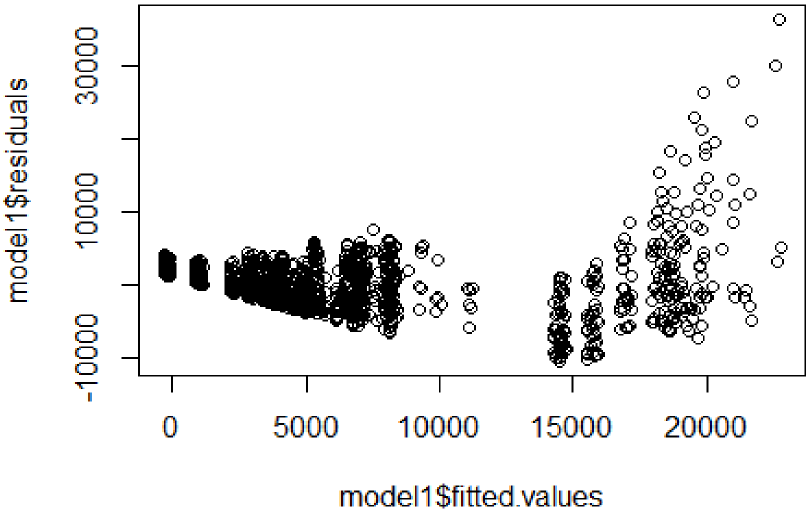
p-value of the coefficients: While the model is overall significant, some of the variables may be insignificant. As state 1, Rating class C1 and Gender female are taken as based states and their coefficients are clubbed in the intercept itself, we observe that coefficients of State 2 and state 15 (Negative) and State 10 (Positive) are significantly different from state 1 (At 95% Confidence level). Similarly rating classes C11 and C6 have significantly negative coefficients compared to C1 indicating that the claim paid for those two rating classes is significantly lesser compared to that of C1. Males have significantly lesser claim paid compared to females at 95% confidence level [2]

From the ANOVA table, we can infer that except Age, all other variables are significant in prediction of claims paid [1]

[12]

ii)

```
#Plot of residuals vs. Fitted Values
plot(model1$fitted.values,model1$residuals)
```
[3]



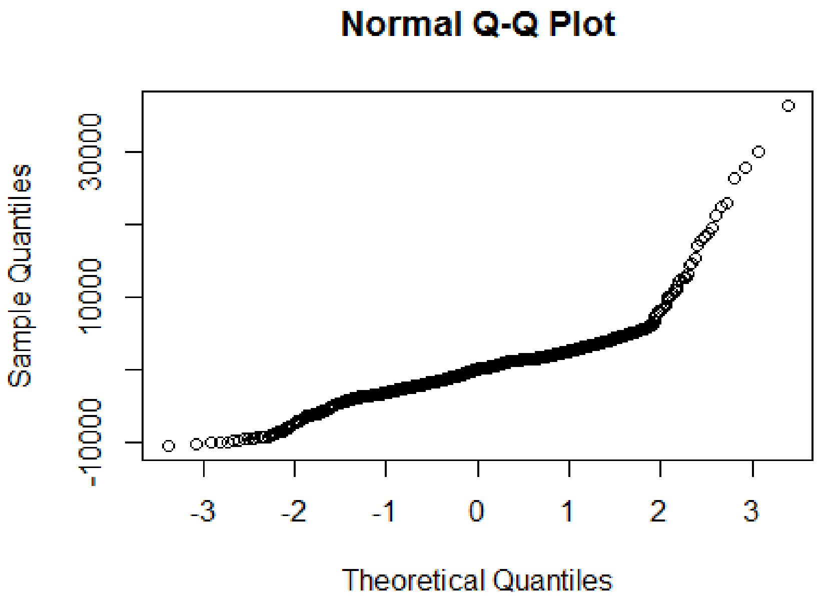
The plot is used to detect non-linearity, unequal error variances, and outliers.

The residuals "do not bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is not reasonable.

The residuals do not form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are not equal and exhibit heteroscedasticity

A few residuals "stands out" from the basic random pattern of residuals. This suggests that there are outliers. [2]

```
# QQ Plot of the residuals
qqnorm(model1$residuals)
```
[3]



A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here it is not, indicating deviance of the residuals from normality. Thus linear regression may not be a better fit to the data [2]

[10]

iii)

```
#Reason for better model
#Checking for the normality of Auto Claims Paid vs. Logairthm of Auto Claims Paid
#Writing Functions for Skewness and Kurtosis
skew<-function(x)mean((x-mean(x))^3)/sd(x)^3                              [2]
kurt<-function(x)(mean((x-mean(x))^4)/sd(x)^4)-3                          [2]
skew(AutoClaims$PAID)                                                     [0.5]

## [1] 2.619422

kurt(AutoClaims$PAID)                                                     [0.5]

## [1] 9.20876

skew(log(AutoClaims$PAID))                                                [0.5]

## [1] 0.4528057

kurt(log(AutoClaims$PAID))                                                [0.5]

## [1] -0.787689
```

Skewness and Kurtosis of Log (Claims) are more close to Zero compared to those of actual claims paid, thus indicating the possibility of using linear regression with this dependent variable [1]

[7]

iv)

```
# Using Natural Logarithm of the claims paid
model2<-lm(log(PAID)~.,data = AutoClaims)
summary(model2)                                                           [3]
```

```
anova(model2)
Analysis of Variance Table

Response: log(PAID)
            Df Sum Sq Mean Sq   F value    Pr(>F)
STATE       10 246.26  24.626  107.6969 < 2.2e-16 ***
CLASS        3 690.09 230.031 1006.0090 < 2.2e-16 ***
GENDER       1  10.88  10.882   47.5908 7.927e-12 ***
AGE          1   0.12   0.123    0.5384    0.4632
Residuals 1401 320.35   0.229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Call:
## lm(formula = log(PAID) ~ ., data = AutoClaims)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96098 -0.34264 -0.05047  0.36828  1.08237
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.896308   0.168777  58.635  < 2e-16 ***
## STATESTATE 02  -0.154804   0.079889  -1.938   0.0529 .
## STATESTATE 03   0.110585   0.092342   1.198   0.2313
## STATESTATE 04   0.049554   0.085147   0.582   0.5607
## STATESTATE 06   0.116190   0.091239   1.273   0.2031
## STATESTATE 07   0.142721   0.101543   1.406   0.1601
## STATESTATE 10   0.098014   0.107391   0.913   0.3616
## STATESTATE 12   0.027982   0.103105   0.271   0.7861
## STATESTATE 14   0.090316   0.102231   0.883   0.3771
## STATESTATE 15  -0.645918   0.075628  -8.541  < 2e-16 ***
## STATESTATE 17   0.004611   0.085455   0.054   0.9570
## CLASSC11       -1.203098   0.052225 -23.037  < 2e-16 ***
## CLASSC6        -1.988743   0.049829 -39.911  < 2e-16 ***
## CLASSF6        -0.034909   0.062787  -0.556   0.5783
## GENDERM        -0.180923   0.026136  -6.922 6.75e-12 ***
## AGE             0.002145   0.002923   0.734   0.4632
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4782 on 1401 degrees of freedom
## Multiple R-squared:  0.7473, Adjusted R-squared:  0.7446
## F-statistic: 276.2 on 15 and 1401 DF,  p-value: < 2.2e-16
```

Key Differences

1. R-Squared and Adjusted R-Squared improved and hence the model is a better fit compared to the initial model [1.5]

2. While a the significance level of a few factor coefficients when compared with the base categories changed, the overall significant variables did not change which can be inferred from the ANOVA table [1.5]

[6]

v)

```
# Using Interaction effects in the model
model3<-lm(PAID~.+STATE:CLASS+STATE:GENDER+CLASS:GENDER,data = AutoClaims)
summary(model3)                                                              [5]
##
## Call:
## lm(formula = PAID ~ . + STATE:CLASS + STATE:GENDER + CLASS:GENDER,
##     data = AutoClaims)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13110.3  -1475.9   -377.5   1250.5  20442.8
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              24373.09    3236.06   7.532 9.08e-14 ***
## STATESTATE 02            -7868.94    3269.51  -2.407 0.016227 *
## STATESTATE 03            -3695.87    3422.76  -1.080 0.280427
## STATESTATE 04             6883.00    3545.51   1.941 0.052425 .
## STATESTATE 06             5428.58    3262.57   1.664 0.096363 .
## STATESTATE 07             -979.80    1184.59  -0.827 0.408314
## STATESTATE 10             7340.08    3546.35   2.070 0.038664 *
## STATESTATE 12             1048.26    3382.21   0.310 0.756659
## STATESTATE 14            -2796.37    3538.40  -0.790 0.429494
## STATESTATE 15           -14038.72    3164.04  -4.437 9.86e-06 ***
## STATESTATE 17            -4266.43    3640.98  -1.172 0.241490
## CLASSC11                -15321.38    3534.31  -4.335 1.56e-05 ***
## CLASSC6                 -20741.82    3262.53  -6.358 2.79e-10 ***
## CLASSF6                   4075.85    3386.90   1.203 0.229026
## GENDERM                  -5508.80    1305.95  -4.218 2.62e-05 ***
## AGE                         23.06      19.35   1.192 0.233581
## STATESTATE 02:CLASSC11    4114.13    3666.38   1.122 0.262008
## STATESTATE 03:CLASSC11    2022.51    3802.06   0.532 0.594847
## STATESTATE 04:CLASSC11   -7719.04    3904.17  -1.977 0.048229 *
## STATESTATE 06:CLASSC11   -6209.18    3743.46  -1.659 0.097412 .
## STATESTATE 07:CLASSC11   -1049.15    1827.01  -0.574 0.565899
## STATESTATE 10:CLASSC11   -6795.03    3956.71  -1.717 0.086144 .
## STATESTATE 12:CLASSC11   -3756.14    3939.41  -0.953 0.340517
## STATESTATE 14:CLASSC11    1668.66    3927.62   0.425 0.671012
## STATESTATE 15:CLASSC11    7776.13    3581.01   2.171 0.030066 *
## STATESTATE 17:CLASSC11    2227.20    4011.07   0.555 0.578806
## STATESTATE 02:CLASSC6     5677.35    3416.32   1.662 0.096777 .
## STATESTATE 03:CLASSC6     2122.55    3605.64   0.589 0.556177
## STATESTATE 04:CLASSC6    -8231.17    3692.27  -2.229 0.025957 *
## STATESTATE 06:CLASSC6    -6151.60    3417.36  -1.800 0.072066 .
## STATESTATE 07:CLASSC6          NA         NA      NA       NA
## STATESTATE 10:CLASSC6    -8117.07    3687.43  -2.201 0.027884 *
## STATESTATE 12:CLASSC6   -2090.95    3561.06  -0.587 0.557187
## STATESTATE 14:CLASSC6    1711.83    3617.52   0.473 0.636143
## STATESTATE 15:CLASSC6   10891.63    3315.76   3.285 0.001046 **
## STATESTATE 17:CLASSC6    2376.31    3776.77   0.629 0.529330
## STATESTATE 02:CLASSF6   -1213.53    3542.37  -0.343 0.731971
## STATESTATE 03:CLASSF6   -2272.17    3835.79  -0.592 0.553708
## STATESTATE 04:CLASSF6  -11888.76    3838.95  -3.097 0.001996 **
## STATESTATE 06:CLASSF6  -17556.04    4687.92  -3.745 0.000188 ***
## STATESTATE 07:CLASSF6         NA         NA      NA       NA
## STATESTATE 10:CLASSF6    1619.22    3953.37   0.410 0.682178
## STATESTATE 12:CLASSF6         NA         NA      NA       NA
## STATESTATE 14:CLASSF6   -2570.84    3769.07  -0.682 0.495299
## STATESTATE 15:CLASSF6   -3790.79    3441.11  -1.102 0.270822
## STATESTATE 17:CLASSF6     362.96    3914.01   0.093 0.926130
## STATESTATE 02:GENDERM    3404.19    1199.24   2.839 0.004598 **
## STATESTATE 03:GENDERM    3061.44    1347.32   2.272 0.023227 *
## STATESTATE 04:GENDERM    1514.18    1279.13   1.184 0.236716
## STATESTATE 06:GENDERM    2052.42    1337.41   1.535 0.125108
## STATESTATE 07:GENDERM    3094.31    1450.30   2.134 0.033057 *
## STATESTATE 10:GENDERM     -73.99    1540.53  -0.048 0.961699
## STATESTATE 12:GENDERM    2476.44    1474.81   1.679 0.093351 .
## STATESTATE 14:GENDERM    2684.73    1624.58   1.653 0.098650 .
## STATESTATE 15:GENDERM    3280.15    1148.55   2.856 0.004357 **
## STATESTATE 17:GENDERM    3033.26    1261.58   2.404 0.016335 *
## CLASSC11:GENDERM         1171.02     730.27   1.604 0.109047
## CLASSC6 :GENDERM         2372.86     692.55   3.426 0.000630 ***
## CLASSF6 :GENDERM        -1786.40     905.11  -1.974 0.048620 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3111 on 1361 degrees of freedom
## Multiple R-squared:  0.8116, Adjusted R-squared:  0.804
## F-statistic: 106.6 on 55 and 1361 DF,  p-value: < 2.2e-16
```

Interpretation **[5]**

```
anova(model3)
Analysis of Variance Table

Response: PAID
               Df     Sum Sq    Mean Sq   F value    Pr(>F)
STATE          10 9.8027e+09 9.8027e+08  101.2664 < 2.2e-16 ***
CLASS           3 3.7869e+10 1.2623e+10 1304.0318 < 2.2e-16 ***
GENDER          1 4.7271e+08 4.7271e+08   48.8334 4.350e-12 ***
AGE             1 6.6423e+06 6.6423e+06    0.6862  0.407612
STATE:CLASS    27 7.8659e+09 2.9133e+08   30.0960 < 2.2e-16 ***
STATE:GENDER   10 2.7570e+08 2.7570e+07    2.8482  0.001634 **
CLASS:GENDER    3 4.6128e+08 1.5376e+08   15.8841 3.733e-10 ***
Residuals    1361 1.3175e+10 9.6801e+06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. R-Squared and Adjusted R-Squared increased to above 80% and hence the model is a better fit compared to the earlier models [2]

2. Interaction effect between a few classes and states emerged out to be very significant (State 6 and Class F6 came out to be significantly negative). Though State 15 came out to be significantly negative when main effects alone were considered, the interaction effects compensated that negative significantly when interacted with class C6 and Class 11 whereas the interaction coefficient is not significant between State 15 and Class F6 indicating that the claims paid is significantly lesser when the state is 15 and class is F6 compared to other rating classes. Digging deeper into the relationships is possible with the interaction effect. Similarly main effect of Gender is significantly negative compared to females but that is offset to some extent for some states (2,3,7,15) and for some rating classes (C6) whereas it is further negative in case of F6. So the differences can be magnified by considering the interaction effects, improving the predictability of the model [1]

3. ANOVA table for the model suggests that except the age all the main effects and their interaction effects are significant at 5% significance level indicating their contribution to the predictability of the model [2]

[10]

[45 Marks]

Part (i) and Part (ii) were attempted by maximum number of students. Many of them were successful in writing the code to fit a linear regression model but only successful students were able to provide a good interpretation of the results. Likewise many students were able to write the code for plotting but very few of them ended up in writing the interpretations based on the graphs. Part (iii) was not attempted by many students and those who attempted also ended up in writing wrong functions. Overall, this question was very poorly done. Part (iv) was attempted well, code was written well but a few ended up in writing correct interpretations. Part (v) was attempted by a few students only and many of them failed to provide appropriate interpretation.

- *Performance of students in CS1B varied drastically. Only a few questions were answered successfully by majority of the participants*
- *R code was used well by majority of the participants but they failed to make interpretations based on the results of the code*
- *Topics wise, students showed a good understanding of the regression models, data analysis and visualizations but the topics on distributions, Hypothesis testing were not addressed well.*
- *The level of interpretation and the comments provided next to the R Code varied significantly among the students.*
- *A good number of students failed in submitting/pasting the output for a few R Functions which resulted in them losing a few marks.*

****