

ITMR solution paper 1

Time: 2 hours
Total Marks: 60 marks

Note:

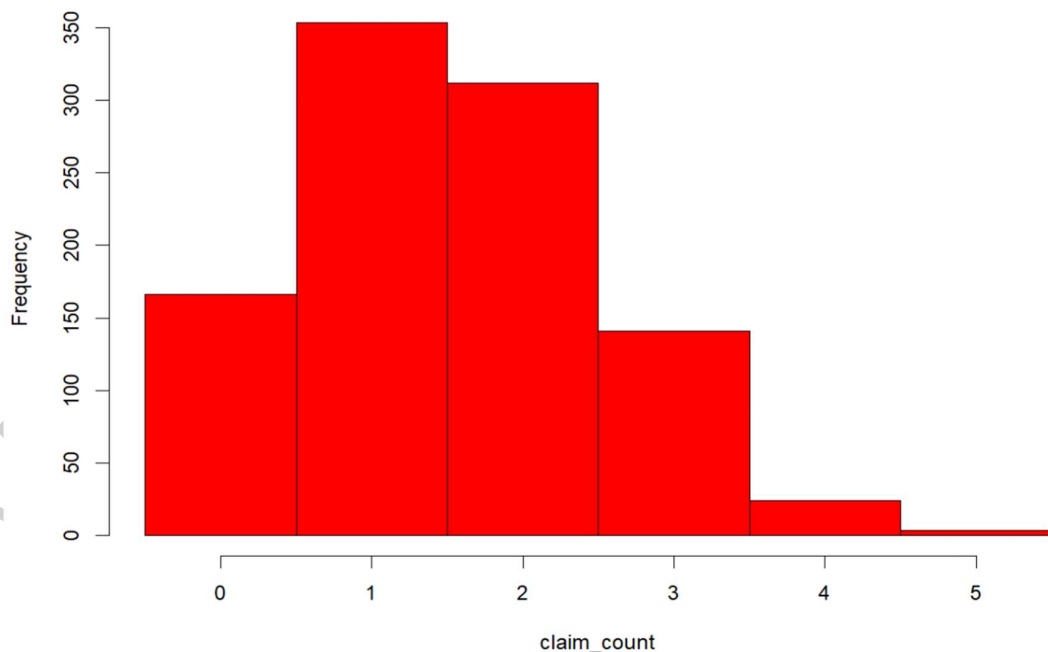
1. The candidate has option to either attempt question 4A or question 4B. Rest all questions are mandatory.
2. Numbers to the right indicate full marks.
3. The candidates will be provided with the formula sheet and graph papers (if required) for the examination.
4. Use of approved scientific calculator is allowed.

Q1 A

5 Marks

```
> #Q1
> #A
> #(i)
> set.seed(2000)
> claim_count = rbinom(1000,5,0.3)
> mean(claim_count)
[1] 1.512
> sd(claim_count)
[1] 1.022204
> #(ii)
> hist(claim_count,breaks = -0.5:5.5,col = "red", main = "Histogram of Claim Count")
```

Histogram of Claim Count



Q1 B

5 Marks

```
> #B
> mean1 = 4000
> var1 = 8000
> lambda = mean1/var1
> alpha = mean1*lambda
> #(a)
> pgamma(5700,alpha,lambda,lower = FALSE)
```

```

[1] 6.305341e-64
> # (b)
> (pgamma(6000,alpha,lambda)-pgamma(5700,alpha,lambda))/pgamma(5700,alpha,
lambda,lower = FALSE)
[1] 0
> # (c)
> qgamma(0.5,alpha,lambda)
[1] 3999.333

```

Q1 C

5 Marks

```

> # (C)
> set.seed(19932)
> claims = rnorm(10,4.2,0.3)
> t.test(claims,
+       mu = 4.5,
+       alt = "t")

```

One Sample t-test

```

data: claims
t = -1.2573, df = 9, p-value = 0.2403
alternative hypothesis: true mean is not equal to 4.5
95 percent confidence interval:
 4.101836 4.613686
sample estimates:
mean of x
 4.357761

```

Since the p-value is greater than 5%, we conclude that we do not have enough evidence to reject H₀.

PATKAR-VARDE

Q2 A**5 Marks**

```
> #Q2
> #A
> #(a)
> data = read.csv("data_Q2A_PaperA.csv")
> nrow(data)
[1] 124
> #(b)
> n = length(data$Marks)
> S = sd(data$Marks)
> alpha = 0.05
> sqrt((n-1)*S^2/qchisq(c(alpha/2,1-alpha/2),n-1,lower = FALSE))
[1] 8.634889 11.097692
```

The 95% confidence interval for sigma is (8.64,11.10)

Q2 B**5 Marks**

```
> #(B)
> #(a)
> fit = lm(Sepal.Length~Petal.Length-1,data = iris)
> fit
```

```
Call:
lm(formula = Sepal.Length ~ Petal.Length - 1, data = iris)
```

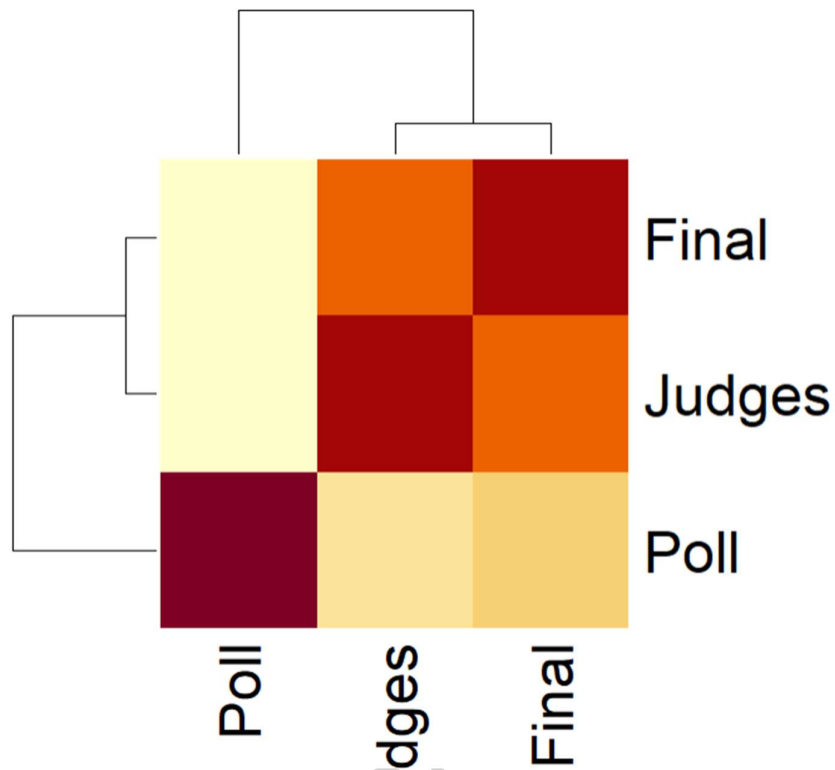
```
Coefficients:
Petal.Length
      1.349
```

```
> #(b)
> confint(fit)
                2.5 %    97.5 %
Petal.Length 1.275914 1.421842
```

Since the 95% confidence interval does not contain 0, we can conclude that Petal Length is a statistically significant variable in predicting Sepal Length.

Q2 C**5 Marks**

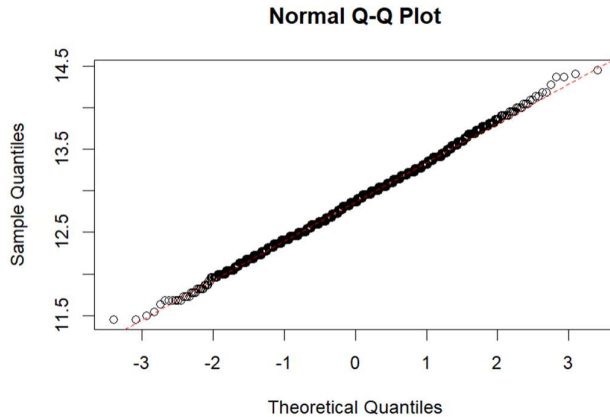
```
> #(C)
> dance = read.csv("dance.csv")
> #(a)
> cor(dance,method = "s")
      Judges      Poll      Final
Judges 1.00000000 -0.02579756 0.7821864
Poll   -0.02579756 1.00000000 0.1087621
Final  0.78218638 0.10876210 1.0000000
> #(b)
> heatmap(cor(dance,method = "s"))
```



Judges have a highly positive correlation co-efficient with the Final score whereas the correlation between Poll and Final is weaker in nature. Infact , Judges and Audience scores have a slightly negative correlation meaning Judges and Audiences like different kinds of performance.

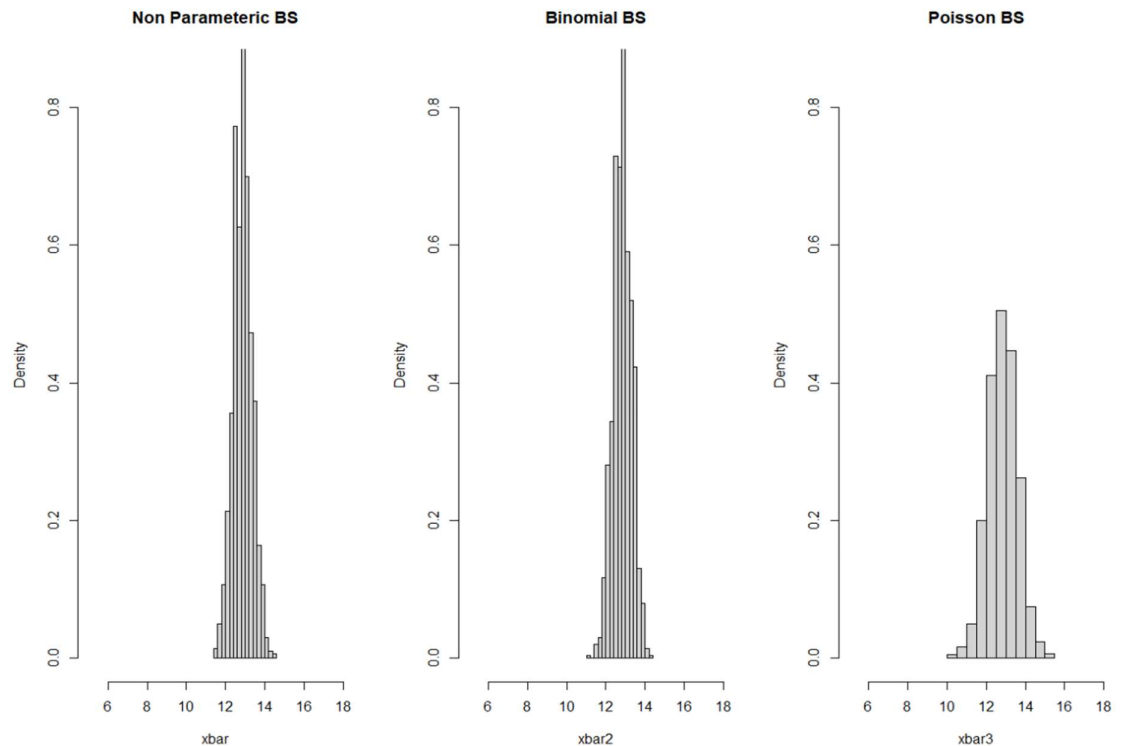
Q3 A**30 Marks**

```
> #Q3
> #A
> #(a)
> x = c(12,14,11,12,13,7,13,12,12,14,13,13,11,20,13,14,13,14,12,14,12,14)
> set.seed(2024)
> xbar = replicate(1500,
+                 mean(sample(x,replace = TRUE)))
> #(b)
> qqnorm(xbar)
> qqline(xbar,col = "red",lty = 2)
```



Since the points, fall on the red line we can say the values in $\bar{x}$ follow a normal distribution. We can conclude that Central Limit theorem is applicable.

```
> #(c)
> p = mean(x)/20
> p
[1] 0.6431818
> #(d)
> set.seed(303)
> xbar2 = replicate(1500,
+                 mean(rbinom(22,20,p)))
> #(e)
> lambda = mean(x)
> lambda
[1] 12.86364
> #(f)
> xbar3 = replicate(1500,
+                 mean(rpois(22,lambda)))
> #(g)
> par(mfrow = c(1,3))
> hist(xbar,xlim = c(5,18),freq = FALSE, ylim = c(0,0.85),main = "Non Par
+ ameteric BS")
> hist(xbar2,xlim = c(5,18),freq = FALSE, ylim = c(0,0.85),main = "Binomia
+ l BS")
> hist(xbar3,xlim = c(5,18),freq = FALSE, ylim = c(0,0.85), main = "Poisso
+ n BS")
```



The shape of the Binomial bootstrapping is closer to the non-parametric bootstrapping and hence Binomial seems like the appropriate distribution to be used to model the number of deaths.

```
> #(i)
> cumsum(22*dbinom(0:20,20,p))
 [1] 2.462420e-08 9.123503e-07 1.611395e-05 1.805236e-04 1.440038e-03
 [6] 8.705109e-03 4.144420e-02 1.594718e-01 5.051910e-01 1.336092e+00
[11] 2.983604e+00 5.683348e+00 9.333162e+00 1.338176e+01 1.703065e+01
[16] 1.966157e+01 2.114355e+01 2.177211e+01 2.196094e+01 2.199677e+01
[21] 2.200000e+01
> #Based on the above <=11 is one group, 12,13,>=14 is another group
> table(x)
x
 7 11 12 13 14 20
 1  2  6  6  6  1
> obs = c(3,6,6,7)
> exp = 22*c(pbinom(11,20,p),dbinom(12:13,20,p),pbinom(13,20,p,lower = FALSE))
> chisq_cal = sum((obs-exp)^2/exp)
> df = length(obs) - 1 - 1 #-1 as p is estimated
> pchisq(chisq_cal,df, lower = FALSE)
[1] 0.1336755
```

Since the p-value is greater than 5%, we accept H_0 . The number of deaths follow a Binomial distribution with $n = 20$.

OR

Q3 B

> #Q3

```
> #B
> #(a)
> motor = read.csv("MotorClaims.csv")
> fit1 = glm(CLAIM~AGE+GENDER, data = motor,family = Gamma("log"))
> fit1
```

```
Call: glm(formula = CLAIM ~ AGE + GENDER, family = Gamma("log"), data = motor)
```

```
Coefficients:
(Intercept)      AGE      GENDERM
  8.594639    0.003332   -0.068636
```

```
Degrees of Freedom: 1416 Total (i.e. Null); 1414 Residual
```

```
Null Deviance: 1353
```

```
Residual Deviance: 1352 AIC: 27540
```

```
> summary(fit1)
```

```
Call:
glm(formula = CLAIM ~ AGE + GENDER, family = Gamma("log"), data = motor)
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.594639	0.345344	24.887	<2e-16 ***
AGE	0.003332	0.006932	0.481	0.631
GENDERM	-0.068636	0.061818	-1.110	0.267

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Gamma family taken to be 1.310975)
```

```
Null deviance: 1353.4 on 1416 degrees of freedom
```

```
Residual deviance: 1351.5 on 1414 degrees of freedom
```

```
AIC: 27542
```

```
Number of Fisher Scoring iterations: 5
```

Interpretation : The Residual deviance is just slightly lower than the NULL deviance which would signal that the variables used in modelling are not statistically significant in prediction of claim amount. Similarly, AIC value is also too high. In comparing models, lower the AIC the better and so this model does not seem like a good fit.

As shown by Residual deviance, the p-value of Age and Gender is greater than 5% means their co-efficients seems to be close to 0. Gender is a comparatively significant variable.

```
> #(b)
> fit2 = update(fit1,~.+LOCATION)
> fit2
```

```
Call: glm(formula = CLAIM ~ AGE + GENDER + LOCATION, family = Gamma("log"),
data = motor)
```

```
Coefficients:
```

	(Intercept)	AGE	GENDE
RM	7.56051	0.01200	-0.095
95			
OA	LOCATIONBIHAR	LOCATIONDELHI	LOCATIONG
19	1.16361	0.97175	0.770
KA	LOCATIONHARYANA	LOCATIONHIMACHAL PRADESH	LOCATIONKARNATA
98	0.98579	0.72748	0.397
AB	LOCATIONKERELA	LOCATIONMAHARASTRA	LOCATIONPUNJ

```

95
1.32440
1.15493
0.902
LOCATIONWEST BENGAL
0.86502

```

```

Degrees of Freedom: 1416 Total (i.e. Null); 1404 Residual
Null Deviance: 1353
Residual Deviance: 1065 AIC: 27180

```

```

> #(c)
> confint(fit2, level = 0.90)
Waiting for profiling to be done...

```

	5 %	95 %
(Intercept)	7.058574744	8.061665582
AGE	0.002133374	0.021922096
GENDERM	-0.183169787	-0.009132635
LOCATIONBIHAR	0.921460965	1.423548177
LOCATIONDELHI	0.780540036	1.172854407
LOCATIONGOA	0.644006475	0.898777430
LOCATIONHARYANA	0.798421991	1.182458510
LOCATIONHIMACHAL PRADESH	0.485647159	0.987145598
LOCATIONKARNATAKA	0.163474729	0.649130186
LOCATIONKERELA	1.063104242	1.607160197
LOCATIONMAHARASTRA	0.913126668	1.414552078
LOCATIONPUNJAB	0.748533903	1.062677715
LOCATIONWEST BENGAL	0.705414970	1.030502775

```

> #(d)
> AIC(fit1)
[1] 27542.3
> AIC(fit2)
[1] 27179.88
> best_fit1 = fit2

```

Since lower the AIC is better, fit2 is a better model.

```

> #(e)
> fit3 = update(fit2, ~. + TYPE)
> anova(fit2, fit3, test = "Chi")
Analysis of Deviance Table

```

```

Model 1: CLAIM ~ AGE + GENDER + LOCATION
Model 2: CLAIM ~ AGE + GENDER + LOCATION + TYPE
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 1404 1065.14
2 1401 320.29 3 744.85 < 2.2e-16 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

> #Type is significant
> fit4a = update(fit3, ~. + AGE:GENDER + AGE:LOCATION + AGE:TYPE + GENDER:LOCATION
+ GENDER:TYPE + LOCATION:TYPE)
> summary(fit4a)

```

```

Call:
glm(formula = CLAIM ~ AGE + GENDER + LOCATION + TYPE + AGE:GENDER +
AGE:LOCATION + AGE:TYPE + GENDER:LOCATION + GENDER:TYPE +
LOCATION:TYPE, family = Gamma("log"), data = motor)

```

Coefficients: (3 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.6005282	0.3247718	23.403	< 2e-16 **
AGE	-0.0021283	0.0065234	-0.326	0.744284
GENDERM	-0.1957531	0.3170027	-0.618	0.537001
LOCATIONBIHAR	-0.8016250	0.9803117	-0.818	0.413659
LOCATIONDELHI	0.0903156	0.6982075	0.129	0.897097
LOCATIONGOA	1.2743323	0.4638056	2.748	0.006084 **
LOCATIONHARYANA	-0.0196791	0.6895871	-0.029	0.977238
LOCATIONHIMACHAL PRADESH	0.5933535	1.0351744	0.573	0.566610

LOCATIONKARNATAKA	0.0246337	0.7820638	0.031	0.974877
LOCATIONKERELA	0.0800726	0.9804678	0.082	0.934923
LOCATIONMAHARASTRA	0.7565433	0.8084926	0.936	0.349572
LOCATIONPUNJAB	1.0391061	0.5580167	1.862	0.062801 .
LOCATIONWEST BENGAL	0.9394247	0.5838331	1.609	0.107837
TYPEPC	0.3673381	0.3597420	1.021	0.307383
TYPEPCV	1.2172693	0.5484883	2.219	0.026631 *
TYPETW	1.1038699	0.6464689	1.708	0.087953 .
AGE:GENDERM	0.0020607	0.0063249	0.326	0.744623
AGE:LOCATIONBIHAR	0.0289017	0.0188478	1.533	0.125407
AGE:LOCATIONDELHI	0.0106956	0.0142275	0.752	0.452332
AGE:LOCATIONGOA	-0.0173922	0.0094252	-1.845	0.065214 .
AGE:LOCATIONHARYANA	0.0151637	0.0143401	1.057	0.290506
AGE:LOCATIONHIMACHAL PRADESH	-0.0006159	0.0206592	-0.030	0.976220
AGE:LOCATIONKARNATAKA	0.0145009	0.0157652	0.920	0.357840
AGE:LOCATIONKERELA	0.0082060	0.0195273	0.420	0.674381
AGE:LOCATIONMAHARASTRA	-0.0037231	0.0166102	-0.224	0.822679
AGE:LOCATIONPUNJAB	-0.0116899	0.0113556	-1.029	0.303462
AGE:LOCATIONWEST BENGAL	-0.0087952	0.0116584	-0.754	0.450736
AGE:TYPEPC	0.0070992	0.0071310	0.996	0.319650
AGE:TYPEPCV	0.0100834	0.0105568	0.955	0.339668
AGE:TYPETW	0.0145956	0.0127453	1.145	0.252341
GENDERM:LOCATIONBIHAR	0.0396868	0.1912009	0.208	0.835599
GENDERM:LOCATIONDELHI	-0.0556349	0.1244039	-0.447	0.654794
GENDERM:LOCATIONGOA	0.0295705	0.0813582	0.363	0.716318
GENDERM:LOCATIONHARYANA	-0.0573998	0.1227255	-0.468	0.640069
GENDERM:LOCATIONHIMACHAL PRADESH	0.1859144	0.1557788	1.193	0.232903
GENDERM:LOCATIONKARNATAKA	0.0933436	0.1507996	0.619	0.536027
GENDERM:LOCATIONKERELA	-0.3076543	0.1768562	-1.740	0.082162 .
GENDERM:LOCATIONMAHARASTRA	-0.1703471	0.1850993	-0.920	0.357580
GENDERM:LOCATIONPUNJAB	-0.1550539	0.1044269	-1.485	0.137829
GENDERM:LOCATIONWEST BENGAL	-0.0308026	0.1016093	-0.303	0.761824
GENDERM:TYPEPC	-0.1253258	0.0635396	-1.972	0.048768 *
GENDERM:TYPEPCV	-0.0943229	0.1047591	-0.900	0.368080
GENDERM:TYPETW	-0.0831610	0.1090584	-0.763	0.445874
LOCATIONBIHAR:TYPEPC	0.1935016	0.2044075	0.947	0.343989
LOCATIONDELHI:TYPEPC	0.2648917	0.1481124	1.788	0.073928 .
LOCATIONGOA:TYPEPC	0.1662012	0.0891823	1.864	0.062594 .
LOCATIONHARYANA:TYPEPC	0.1776337	0.1615027	1.100	0.271581
LOCATIONHIMACHAL PRADESH:TYPEPC	0.0472376	0.2324178	0.203	0.838975
LOCATIONKARNATAKA:TYPEPC	0.0220435	0.1613746	0.137	0.891369
LOCATIONKERELA:TYPEPC	0.6510932	0.1977340	3.293	0.001018 **
LOCATIONMAHARASTRA:TYPEPC	0.3095461	0.2478975	1.249	0.211997
LOCATIONPUNJAB:TYPEPC	0.4875099	0.1154661	4.222	2.58e-05 **
*				
LOCATIONWEST BENGAL:TYPEPC	0.3214184	0.1122168	2.864	0.004245 **
LOCATIONBIHAR:TYPEPCV	0.1921842	0.2102846	0.914	0.360920
LOCATIONDELHI:TYPEPCV	0.2805475	0.2455484	1.143	0.253435
LOCATIONGOA:TYPEPCV	0.2723722	0.1366999	1.992	0.046520 *
LOCATIONHARYANA:TYPEPCV	-0.3910537	0.5075842	-0.770	0.441185
LOCATIONHIMACHAL PRADESH:TYPEPCV	NA	NA	NA	NA
LOCATIONKARNATAKA:TYPEPCV	NA	NA	NA	NA
LOCATIONKERELA:TYPEPCV	0.9160172	0.2647610	3.460	0.000557 **
*				
LOCATIONMAHARASTRA:TYPEPCV	0.5656367	0.1867777	3.028	0.002505 **
LOCATIONPUNJAB:TYPEPCV	0.5045768	0.1874295	2.692	0.007189 **
LOCATIONWEST BENGAL:TYPEPCV	0.4733259	0.1624041	2.914	0.003622 **
LOCATIONBIHAR:TYPETW	0.0617971	0.2435741	0.254	0.799759
LOCATIONDELHI:TYPETW	0.1774479	0.2441399	0.727	0.467457
LOCATIONGOA:TYPETW	0.0581700	0.1580778	0.368	0.712944
LOCATIONHARYANA:TYPETW	0.3687807	0.1668317	2.210	0.027238 *
LOCATIONHIMACHAL PRADESH:TYPETW	0.2556054	0.2340277	1.092	0.274940
LOCATIONKARNATAKA:TYPETW	NA	NA	NA	NA
LOCATIONKERELA:TYPETW	0.9219041	0.2773740	3.324	0.000912 **
*				
LOCATIONMAHARASTRA:TYPETW	0.2806532	0.5227910	0.537	0.591469
LOCATIONPUNJAB:TYPETW	0.7509063	0.2759812	2.721	0.006595 **
LOCATIONWEST BENGAL:TYPETW	0.2019377	0.3015793	0.670	0.503227

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.2375342)

Null deviance: 1353.40 on 1416 degrees of freedom

Residual deviance: 300.94 on 1347 degrees of freedom

AIC: 25378

Number of Fisher Scoring iterations: 6

> #All age and gender interactions are insignificant.

> fit4b = update(fit3, ~.+LOCATION:TYPE)

> anova(fit3, fit4b, test = "Chi")

Analysis of Deviance Table

Model 1: CLAIM ~ AGE + GENDER + LOCATION + TYPE

Model 2: CLAIM ~ AGE + GENDER + LOCATION + TYPE + LOCATION:TYPE

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 1401 320.29

2 1374 307.42 27 12.868 0.001418 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> #(f)

> check_fit = glm(CLAIM~AGE+GENDER+LOCATION+TYPE+LOCATION:TYPE, data = moto
r, family = Gamma())

> AIC(check_fit)

[1] 25369.71

> AIC(fit4b)

[1] 25355.63

Since the AIC of fit4b is lower, we will opt for that model as final_selected.

> #(g)

> newdata = data.frame(AGE = 30,

+ GENDER = "M",

+ LOCATION = "MAHARASTRA",

+ TYPE = "PC")

> exp(predict(final_selected, newdata))

1
6133.219