**CA Two for Statistics for Data Analysis**

**Prasad Tambe**

**10515513(A batch)**

**Question 1**

**Consider a relational dataset and specify your input and output variables, then:**

**Solution:**

A dataset cartrain.csv from Kaggle was taken in order to perform GLM.

Description: Used cars listings for sale. An exploratory analysis for the data by doing regression to understand the variable significance and selection for modelling (GLM) for price prediction.

Here, we are considering rownum, acquisition\_date, badge,body\_type,category,colour,cylinders

Economy, fuel, last\_updated, litres,location,make,model,odometer,transmission as Input Variables and Price as Output Variable.

**#Loading the dataset**

**Solution performed in R:**

mydata<- read.csv("C:/Users/Dell/Desktop/car\_train.csv")  
head(mydata)

## rownum price acquisition\_date badge body\_type category colour cylinders  
## 1 0 8560 2017-06-22 RS Hatch Used Silver 4  
## 2 3 17074 2017-06-22 2.0i Hatch Used Silver 4  
## 3 4 8526 2017-06-22 R Hatch Used Blue 4  
## 4 5 10952 2017-06-22 R Hatch Used Black 4  
## 5 6 33964 2017-06-22 WRX STI Sedan Used Grey 4  
## 6 8 18070 2017-06-22 2.0i-S Hatch Used Black 4  
## economy fuel last\_updated litres location make model odometer  
## 1 8.9 Unleaded 2017-06-22 2.0 2 Subaru Impreza 134944  
## 2 6.8 Unleaded 2017-06-22 2.0 3 Subaru Impreza 33304  
## 3 8.9 Unleaded 2017-06-22 2.0 6 Subaru Impreza 81668  
## 4 8.8 Unleaded 2017-06-22 2.0 8 Subaru Impreza 48051  
## 5 10.5 Unleaded 2017-06-22 2.5 3 Subaru Impreza 51516  
## 6 6.8 Unleaded 2017-06-22 2.0 7 Subaru Impreza 60294  
## transmission year  
## 1 Manual 2009  
## 2 Automatic 2014  
## 3 Manual 2007  
## 4 Automatic 2009  
## 5 Manual 2011  
## 6 Automatic 2012

**# Defining the variables**

**Solution performed in R:**

x1<- mydata$rownum  
x3<- mydata$body\_type  
x4<- mydata$category  
x6<- mydata$cylinders  
x7<- mydata$economy  
x8<- mydata$fuel  
x10<- mydata$litres  
x11<- mydata$location  
x12<- mydata$make  
x13<- mydata$model  
x14<- mydata$odometer  
x15<- mydata$transmission  
x16<- mydata$year  
y<- mydata$price  
dataset<- na.omit(data.frame(x1,x3,x4,x6,x7,x8,x10,x11,x12,x13,x14,x15,x16,y))  
dim(mydata)

## [1] 38506 18

**#fitting the model**

set.seed(44444)  
fit.glm<- glm(y~., dataset, family = 'gaussian')  
summary(fit.glm)

##   
## Call:  
## glm(formula = y ~ ., family = "gaussian", data = dataset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -13941 -2812 -509 2284 418036   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.121e+06 3.186e+04 -129.379 < 2e-16 \*\*\*  
## x1 -1.464e-02 1.897e-03 -7.719 1.21e-14 \*\*\*  
## x3Sedan 2.069e+03 1.092e+02 18.942 < 2e-16 \*\*\*  
## x3SUV 3.844e+03 1.057e+02 36.356 < 2e-16 \*\*\*  
## x4Other 1.523e+03 1.511e+02 10.077 < 2e-16 \*\*\*  
## x4Private -4.286e+03 1.599e+02 -26.810 < 2e-16 \*\*\*  
## x4Used -3.777e+03 1.459e+02 -25.881 < 2e-16 \*\*\*  
## x6 -3.249e+02 1.919e+02 -1.693 0.090404 .   
## x7 3.715e+03 5.125e+01 72.484 < 2e-16 \*\*\*  
## x8Unleaded -9.090e+03 1.482e+02 -61.335 < 2e-16 \*\*\*  
## x10 -3.375e+03 2.070e+02 -16.304 < 2e-16 \*\*\*  
## x11 -1.491e+02 1.293e+01 -11.530 < 2e-16 \*\*\*  
## x12Toyota -3.776e+02 7.676e+01 -4.920 8.72e-07 \*\*\*  
## x13Impreza NA NA NA NA   
## x13RAV4 NA NA NA NA   
## x14 -3.925e-02 6.665e-04 -58.889 < 2e-16 \*\*\*  
## x15Manual 3.178e+02 8.202e+01 3.875 0.000107 \*\*\*  
## x16 2.054e+03 1.572e+01 130.672 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 26998594)  
##   
## Null deviance: 4.0247e+12 on 33071 degrees of freedom  
## Residual deviance: 8.9247e+11 on 33056 degrees of freedom  
## AIC: 659777  
##   
## Number of Fisher Scoring iterations: 2

**#indexing the model and dividing it into 80/20 ratio**

n=nrow(dataset)   
  
indexes = sample(n,n\*(80/100))   
  
trainset = dataset[indexes,]   
  
testset = dataset[-indexes,]

# fitting the model using trainset

trainset.glm <- glm(trainset$y ~.,trainset, family="gaussian")  
summary(trainset.glm)

##   
## Call:  
## glm(formula = trainset$y ~ ., family = "gaussian", data = trainset)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -14206 -2831 -503 2272 418007   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.206e+06 3.640e+04 -115.569 < 2e-16 \*\*\*  
## x1 -1.436e-02 2.188e-03 -6.562 5.39e-11 \*\*\*  
## x3Sedan 2.090e+03 1.257e+02 16.628 < 2e-16 \*\*\*  
## x3SUV 3.845e+03 1.221e+02 31.499 < 2e-16 \*\*\*  
## x4Other 1.522e+03 1.750e+02 8.696 < 2e-16 \*\*\*  
## x4Private -4.373e+03 1.848e+02 -23.661 < 2e-16 \*\*\*  
## x4Used -3.815e+03 1.689e+02 -22.586 < 2e-16 \*\*\*  
## x6 -3.031e+02 2.167e+02 -1.399 0.16195   
## x7 3.726e+03 5.907e+01 63.084 < 2e-16 \*\*\*  
## x8Unleaded -9.137e+03 1.706e+02 -53.543 < 2e-16 \*\*\*  
## x10 -3.416e+03 2.383e+02 -14.333 < 2e-16 \*\*\*  
## x11 -1.533e+02 1.491e+01 -10.278 < 2e-16 \*\*\*  
## x12Toyota -4.247e+02 8.866e+01 -4.790 1.68e-06 \*\*\*  
## x13Impreza NA NA NA NA   
## x13RAV4 NA NA NA NA   
## x14 -3.645e-02 7.392e-04 -49.307 < 2e-16 \*\*\*  
## x15Manual 2.936e+02 9.481e+01 3.097 0.00196 \*\*   
## x16 2.097e+03 1.796e+01 116.720 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 28780424)  
##   
## Null deviance: 3.2705e+12 on 26456 degrees of freedom  
## Residual deviance: 7.6098e+11 on 26441 degrees of freedom  
## AIC: 529504  
##   
## Number of Fisher Scoring iterations: 2

# full model

fit = glm(trainset$y ~., data= trainset, family='gaussian')   
fit

##   
## Call: glm(formula = trainset$y ~ ., family = "gaussian", data = trainset)  
##   
## Coefficients:  
## (Intercept) x1 x3Sedan x3SUV x4Other x4Private   
## -4.206e+06 -1.436e-02 2.090e+03 3.845e+03 1.522e+03 -4.373e+03   
## x4Used x6 x7 x8Unleaded x10 x11   
## -3.815e+03 -3.031e+02 3.726e+03 -9.137e+03 -3.416e+03 -1.533e+02   
## x12Toyota x13Impreza x13RAV4 x14 x15Manual x16   
## -4.247e+02 NA NA -3.645e-02 2.936e+02 2.097e+03   
##   
## Degrees of Freedom: 26456 Total (i.e. Null); 26441 Residual  
## Null Deviance: 3.271e+12   
## Residual Deviance: 7.61e+11 AIC: 529500

**#Predicting the Model**

pred=predict(fit, testset)

pred

## 3 10 12 19 20 22   
## 10012.12233 31186.78547 29663.72236 25223.44237 33206.09295 11822.83726   
## 23 28 31 35 40 41   
## 3897.23739 24671.80468 8231.43427 14150.19443 10037.86576 26090.36530   
## 37290 37294 37300 37305 37307 37309   
## 10205.47940 14331.42265 5685.25126 13789.95375 18860.01521 19322.19571   
## 37312 37316 37318 37331 37338 37339   
## 21688.90046 8538.07938 26480.90114 16906.28128 20199.30806 39946.59801   
## 37343 37348 37349 37352 37361 37369   
## 14261.90221 23665.28962 12503.62708 39192.81741 7962.65675 24948.93298   
## 37373 37376 37377 37387 37391 37407   
## 37859.11011 41057.44484 25281.73142 14341.00450 30368.77776 23461.24874   
## 37412 37414 37420 37424 37432 37433   
## 26838.36222 18028.80249 8699.27838 456.49979 -422.17916 17436.55813   
## 37441 37453 37471 37472 37473 37475   
## 3316.64861 9022.20604 37401.59685 17980.78069 29441.16935 4933.74653   
## 37477 37483 37486 37487 37490 37495   
## 21708.00513 16049.84666 13759.77676 5937.62235 13608.96923 14680.32802   
## 37498 37503 37516 37525 37534 37553   
## 11292.67324 13649.61354 15967.59805 35876.03398 15159.10315 7025.02760   
## 37615 37646 37709 37756 37770 37786   
## 17762.61948 15331.64125 30645.99803 31930.48687 27772.90069 32226.23750   
## 37816 37832 37910 37965 38005 38109   
## 19209.83333 -731.59763 37843.69535 39507.20796 37849.91235 37851.47047   
## 38118 38150 38188 38310 38328 38343   
## 23773.19738 19917.29291 7200.38908 31668.26192 23990.25980 27652.96290   
## 38353 38358 38363 38365 38372 38378   
## -3514.92309 29098.35393 20829.76727 16975.64380 27135.38082 4001.88151   
## 38392 38395 38396 38401 38402 38404   
## 26912.64369 33843.53071 22049.77932 19469.77120 36481.99023 -1268.03003   
## 38408 38414 38417 38421 38423 38427   
## 19480.88097 28516.43078 31381.92978 31724.05562 -664.18169 31750.74702   
## 38438 38444 38447   
## 32396.10982 861.77455 1293.86838

actual=testset$y   
actual

## [1] 8526 30216 30277 28741 25084 16022 11023 18509 13034 15006  
## [11] 33087 29977 15589 27934 38089 32058 27998 17500 13003 25060  
## [21] 11532 28047 18016 31544 15087 23537 10968 11888 30031 26023  
## [31] 6014 15006 25988 28893 24042 13990 8198 11995 25057 18898  
## [41] 12034 36084 29530 26000 30037 7003 7862 29002 20530 33032  
## [51] 15051 6991 36024 25994 23592 31054 17538 23082 13679 20015  
## [61] 8030 24055 22064 34524 29015 9517 15028 12518 26061 17606  
## [71] 27592 30025 9039 24069 10762 9263 22335 10971 11554 23494  
## [81] 18327 13065 12007 9071 17562 6723 25027 10501 19764 16494  
## [91] 6506 12555 12045 6995 9792 9880 5555 7527 29984 10033  
## [101] 27986 10348 28554 7059 19084 11053 13098 8973 6938 15047  
## [111] 6559 5584 8076 8030 18066 7890 6053 17562 28009 25894  
## [121] 28268 6782 21039 13529 33048 41594 45517 28792 36001 36083  
## [131] 36047 25860 33966 42589 37614 13064 43885 35992 36075 33833  
## [141] 43039 25017 50012 29901 36978 22983 51594 38634 35688 36902  
## [151] 18848 17750 39999 39064 19916 21528 17523 16514 52980 34888  
## [161] 43735 39837 6514 38338 28044 6016 2396 43036 14995 14056  
7734  
## [5461] 12956 8010 44727 32507 9068 24084 7996 4976 10592 40058  
## [5471] 8512 10821 8238 10022 9001 13002 26002 1814 28289 4546  
## [5481] 14007 13072 8521 26505 12028 28238 25215 20544 7035 3081  
## [5491] 29010 27731 16997 27915 7592 7654 20923 35908 34911 34965  
## [5501] 9041 16564 35501 4008 4538 18992 13557 37066 20993 24931  
## [5511] 28004 22663 40773 10534 23067 28058 10910 21927 29928 18511  
## [5521] 27053 21816 30920 32964 12570 23003 23936 26965 26040 24959  
## [5531] 25540 25986 16411 24990 34052 45035 26886 23974 33961 12888  
## [5541] 26028 20075 24534 24963 11062 25459 46055 31013 24683 25990  
## [6251] 19517 29553 4088 44034 4307 5531 5085 7996 5092 13509  
## [6261] 28068 15020 35003 16933 32031 32056 11556 29037 37002 30996  
## [6271] 30928 39075 13009 29993 26005 14886 5052 16010 22037 34065  
## [6281] 19089 12905 18503 23978 24996 9073 10088 12568 28274 34984  
## [6581] 31896 27383 16940 3612 36012 40942 36013 35944 25008 27008  
## [6591] 8823 32944 24022 36055 3844 32023 19034 10006 22982 7032  
## [6601] 28892 24674 17071 16583 39978 2567 16067 25945 28011 34044  
## [6611] 5078 26927 35059 3849 5096

rmse=sqrt((sum((pred-actual)^2))/nrow(testset))

# reduced model

library(MASS)   
fit\_red = stepAIC(fit)

## Start: AIC=529504.1  
## trainset$y ~ x1 + x3 + x4 + x6 + x7 + x8 + x10 + x11 + x12 +   
## x13 + x14 + x15 + x16  
##   
##   
## Step: AIC=529504.1  
## trainset$y ~ x1 + x3 + x4 + x6 + x7 + x8 + x10 + x11 + x12 +   
## x14 + x15 + x16  
##   
## Df Deviance AIC  
## - x6 1 7.6104e+11 529504  
## <none> 7.6098e+11 529504  
## - x15 1 7.6126e+11 529512  
## - x12 1 7.6164e+11 529525  
## - x1 1 7.6222e+11 529545  
## - x11 1 7.6402e+11 529608  
## - x10 1 7.6690e+11 529707  
## - x3 2 7.8964e+11 530478  
## - x4 3 8.1641e+11 531358  
## - x14 1 8.3095e+11 531829  
## - x8 1 8.4349e+11 532226  
## - x7 1 8.7552e+11 533212  
## - x16 1 1.1531e+12 540497  
##   
## Step: AIC=529504.1  
## trainset$y ~ x1 + x3 + x4 + x7 + x8 + x10 + x11 + x12 + x14 +   
## x15 + x16  
##   
## Df Deviance AIC  
## <none> 7.6104e+11 529504  
## - x15 1 7.6134e+11 529513  
## - x12 1 7.6180e+11 529528  
## - x1 1 7.6228e+11 529545  
## - x11 1 7.6409e+11 529608  
## - x10 1 7.6982e+11 529806  
## - x3 2 7.9330e+11 530598  
## - x4 3 8.1642e+11 531356  
## - x14 1 8.3102e+11 531829  
## - x8 1 8.4349e+11 532224  
## - x7 1 8.8190e+11 533402  
## - x16 1 1.1613e+12 540684

# fitting reduced model

pred\_r=predict(fit\_red, testset)   
  
rmse\_r=sqrt((sum((pred\_r - actual)^2))/nrow(testset))

1. **Since we have a linear model, we do not have a confusion matrix**

# Accuracy

RMSE=c(0,0)   
RMSE=RMSE+c(rmse,rmse\_r)

# To find the probability of correctness of prediction is accuracy

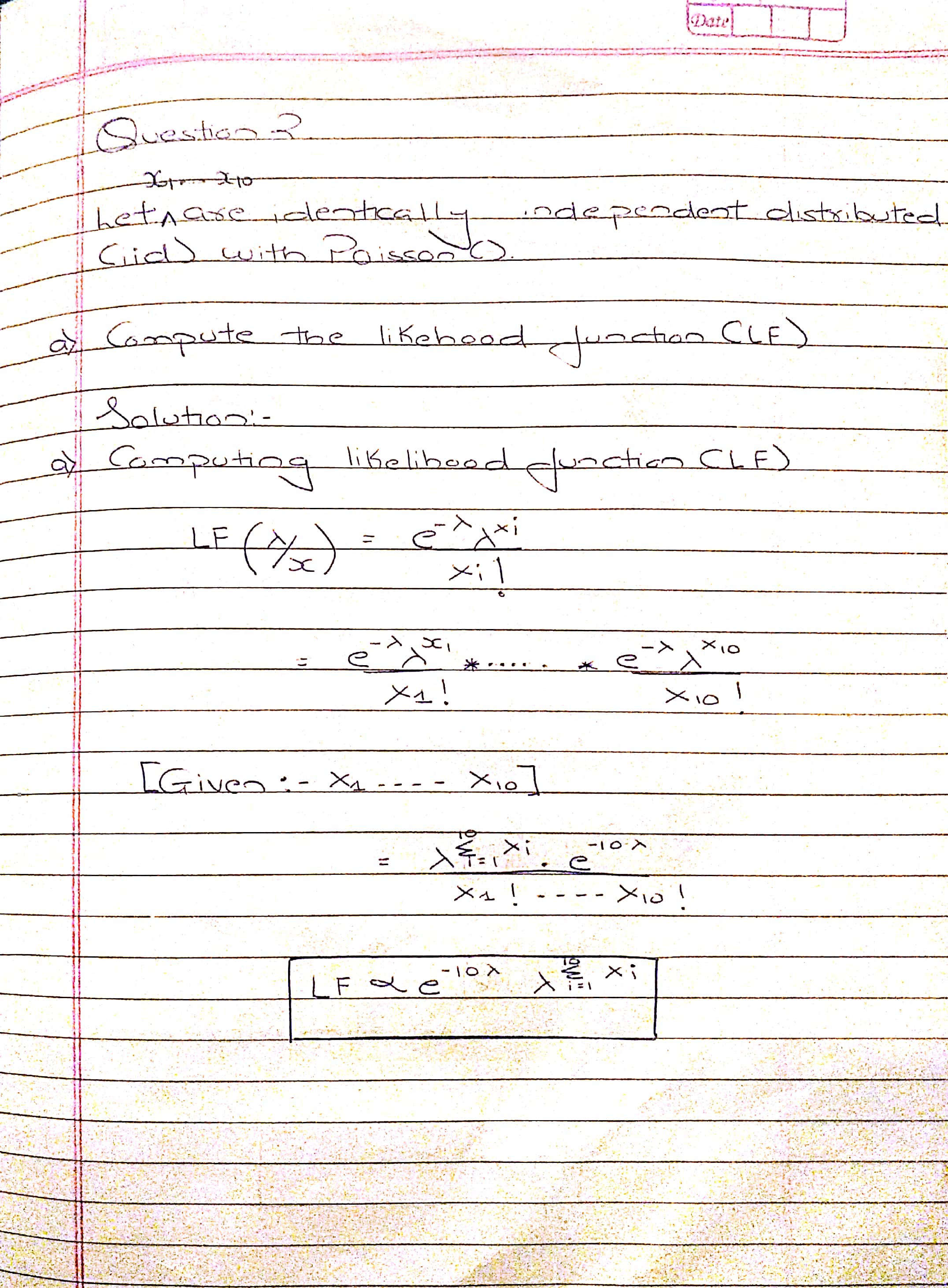
14

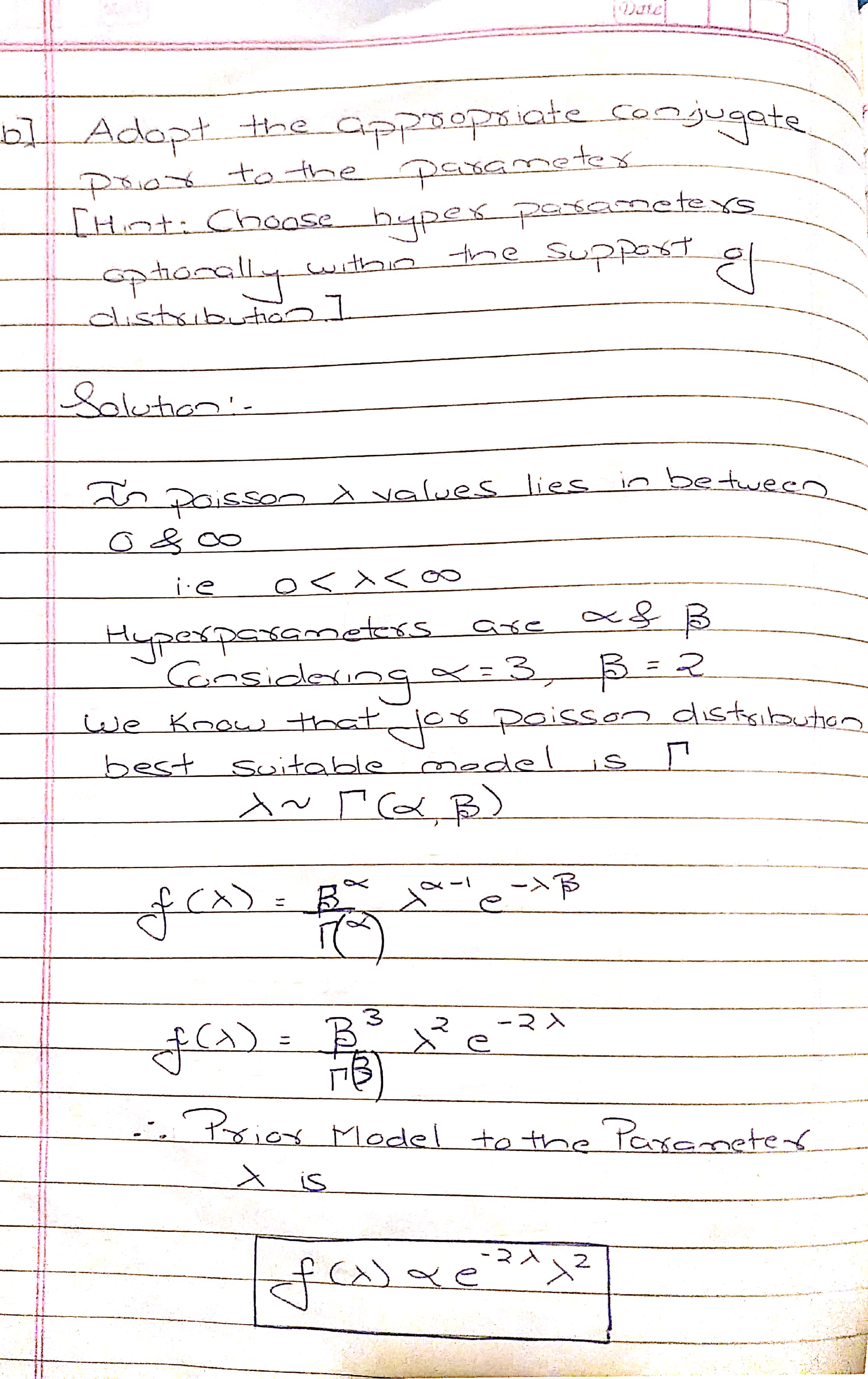
RMSE

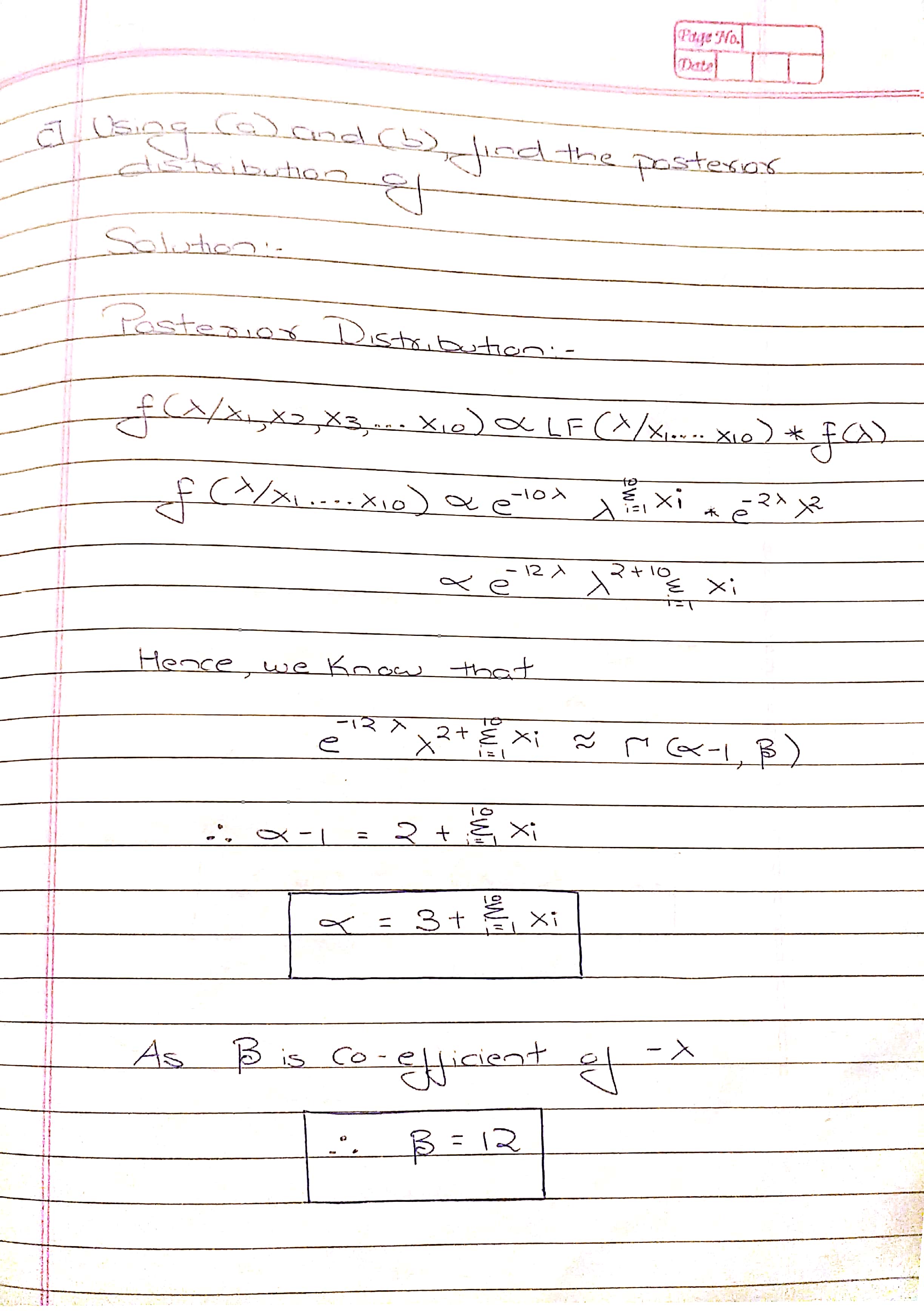
## [1] 4467.313 4467.667

**Question 2**

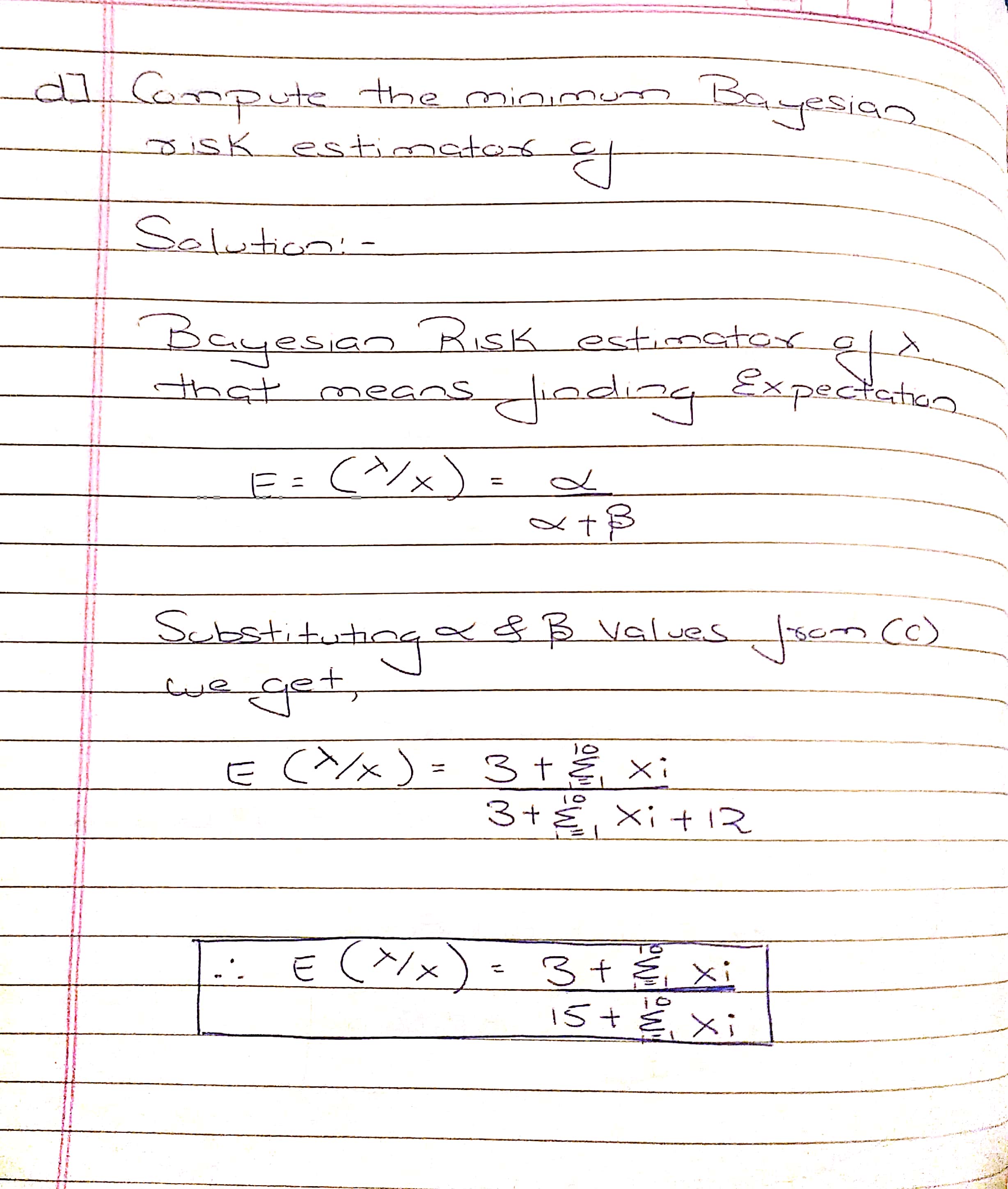
Let are identically independently distributed (iid) with Poisson ().

**Compute the likelihood function (LF).** 

**b**) **Adopt the appropriate conjugate prior to the parameter (Hint: Choose hyperparameters optionally within the support of distribution).** 

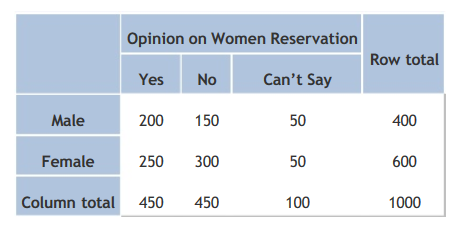
**c) Using (a) and (b), find the posterior distribution of .** 

**d) Compute the minimum Bayesian risk estimator of .**



**Question 3**

An opinion poll surveyed a simple random sample of 1000 students. Respondents were classified by gender (male or female) and by opinion (Reservation for women, No Reservation, or No Opinion). Results are shown in the observed contingency table below.

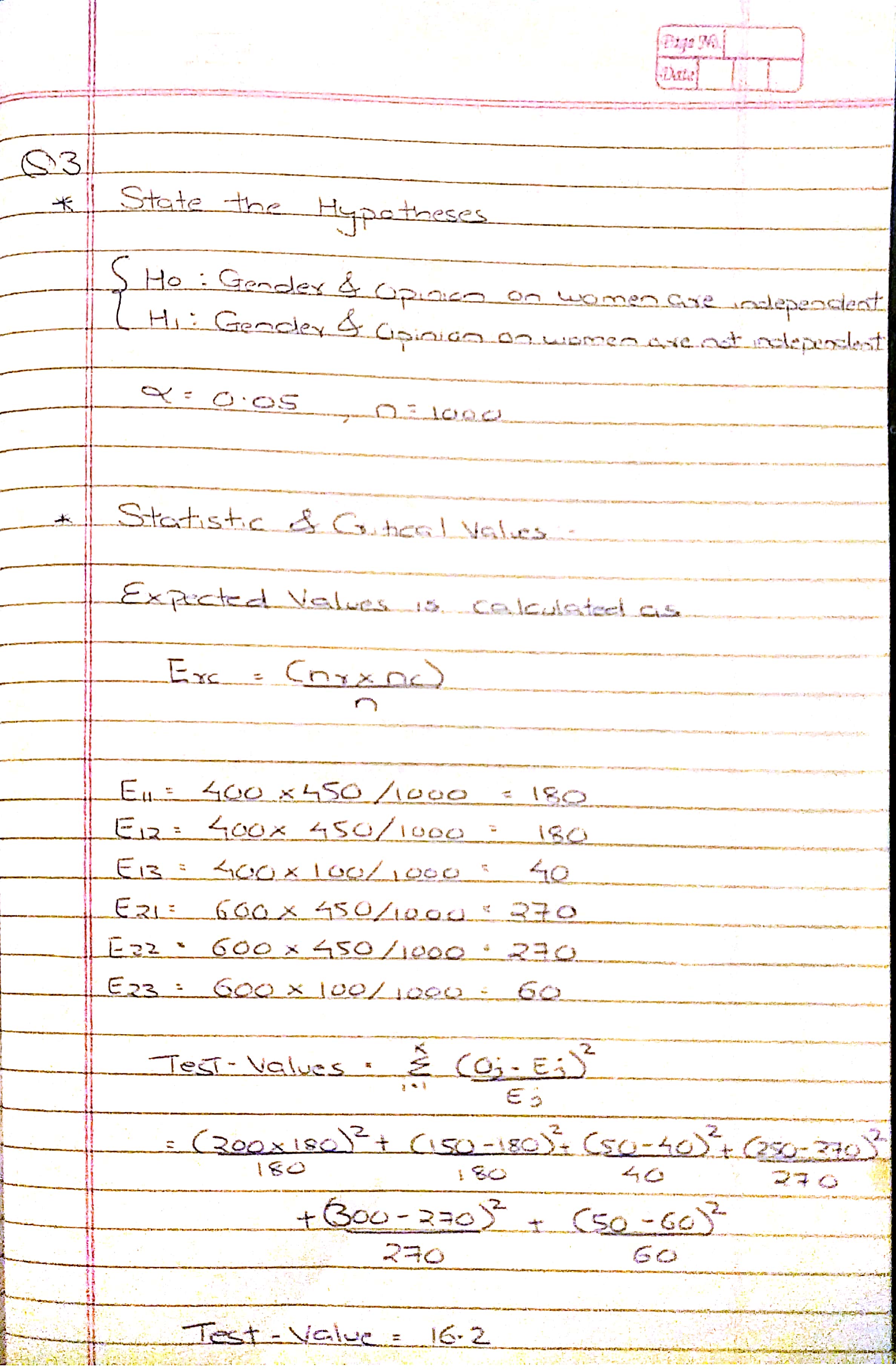


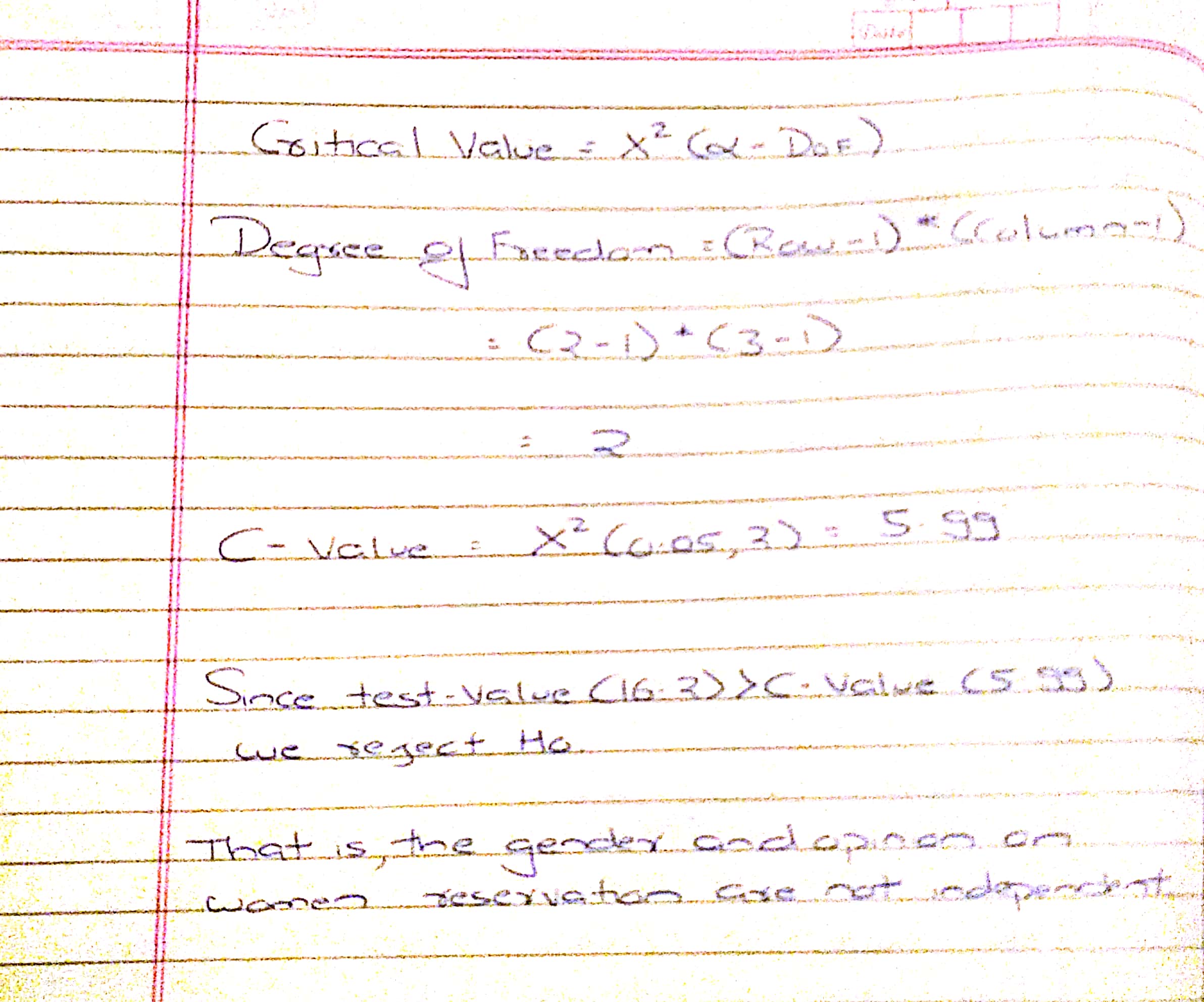
Does the gender and opinion on women reservation are independent? Use a 0.05 level of significance. To do so,

1. State the hypotheses.
2. Find the statistic and critical values.

c)Explain your decision and Interpret results

Manual Solution:





R Solution:

data = data.frame("Yes"=c(200,250),  
 "No"=c(150,300),  
 "Can't Say"= c(50,50))  
chisq.test(data)

##   
## Pearson's Chi-squared test  
##   
## data: data  
## X-squared = 16.204, df = 2, p-value = 0.000303