Question 1: Perform Data Exploration analysis and discuss your observations from the analysis.

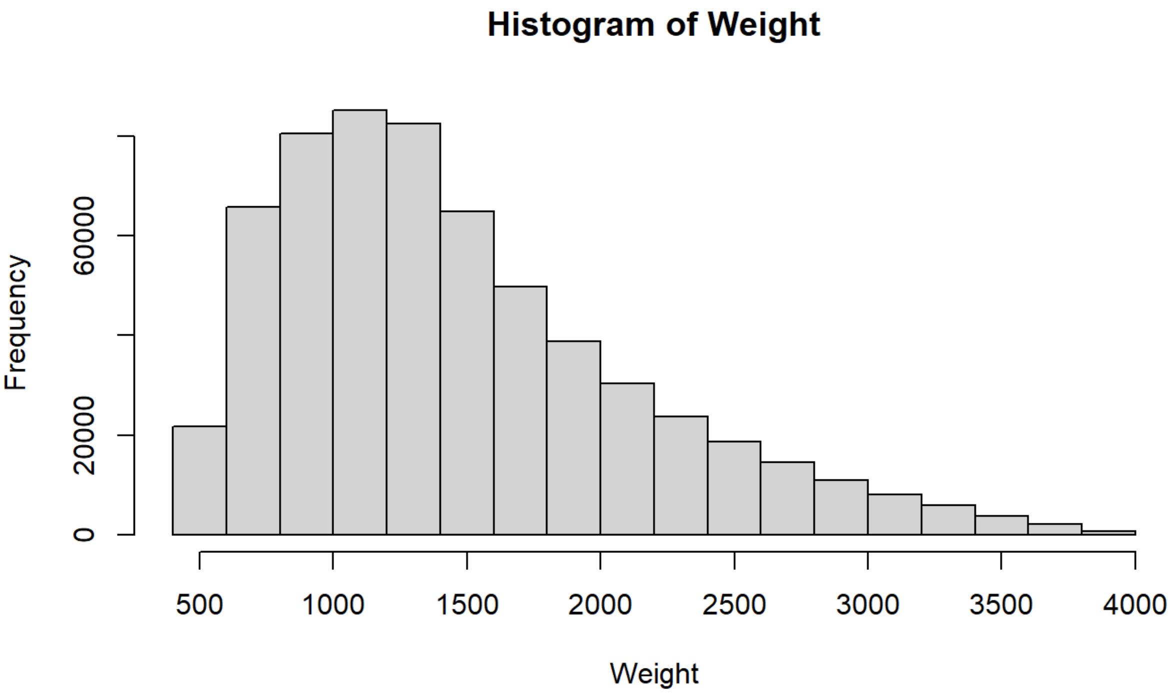
First we summarize the variables using summary() function to return the following values:

	Count	Exposure	Distance	Weight	Age	Car Age
Min	0.0000	0.8000	1.00	450	18.00	1.000
1st Quarter	0.0000	0.8500	5.00	962	35.00	3.000
Median	0.0000	0.8999	10.00	1319	46.00	5.000
Mean	0.0609	0.8999	14.85	1464	47.25	7.762
3rd Quarter	0.0000	0.9498	19.00	1826	58.00	10.000
Max	3.0000	1.0000	95.00	3994	98.00	45.000

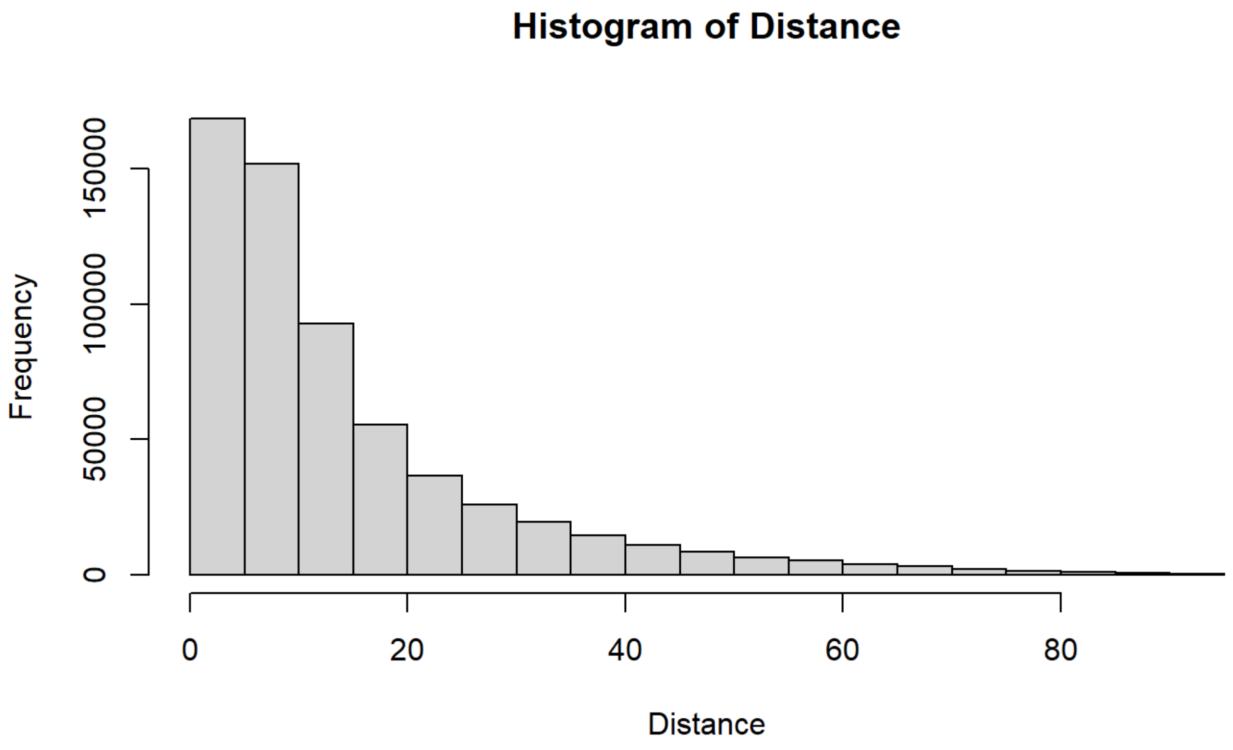
For the Categorical variables, we have:

State	Number of Counts	Gender	Number of Counts
ACT	120952	Female	243052
NSW	121441	Male	364645
QLD	121990		
SA	122092		
VIC	121222		

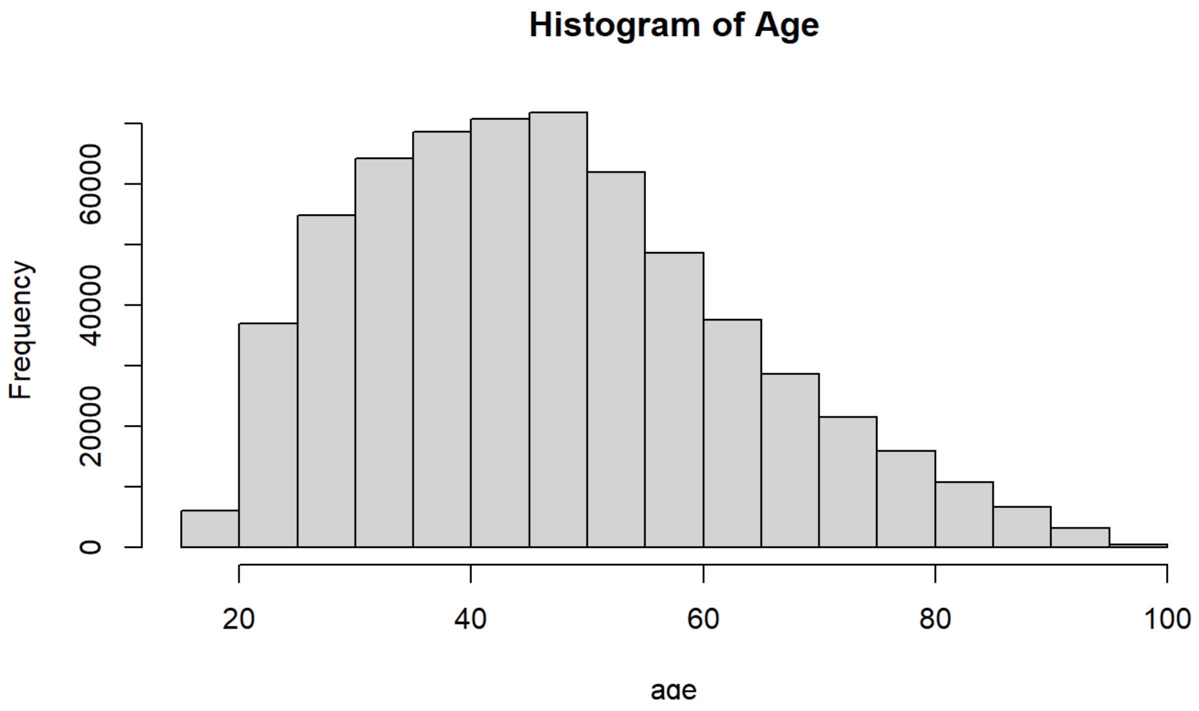
Next we look deeper to each variable through their histogram and their skewness (if applicable):



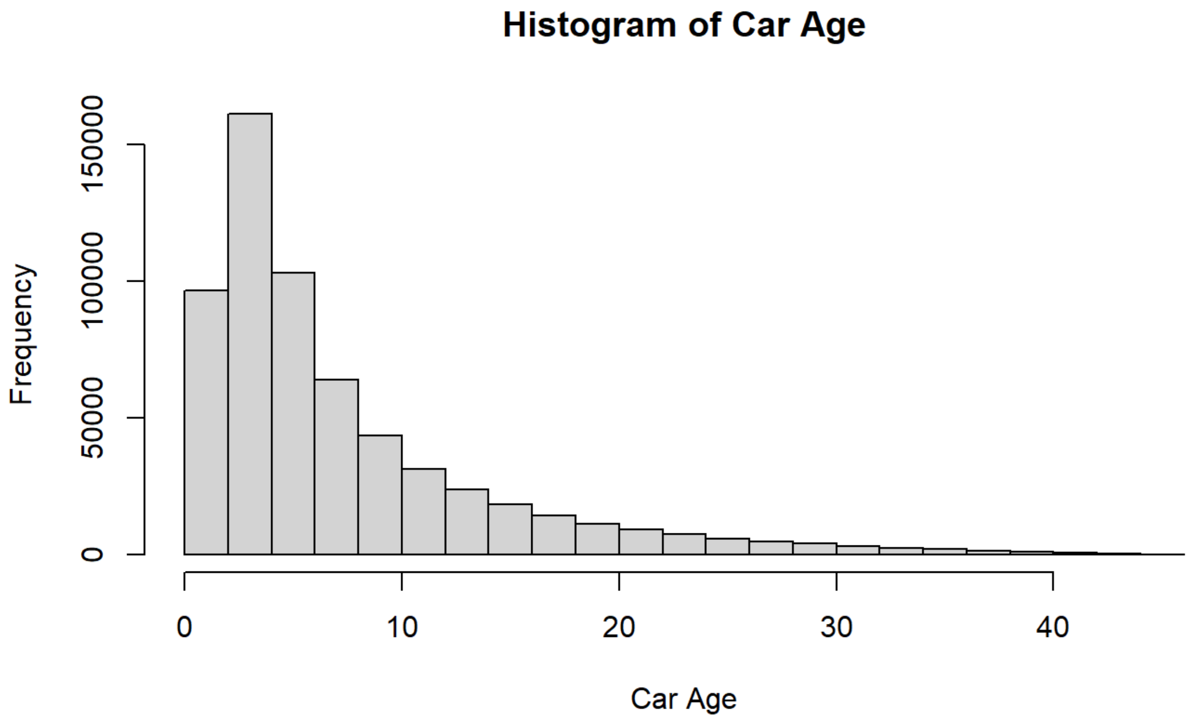
For variable Weight, based on the Histogram and its skewness being 0.9709, with its Mean (1464) > Median (1319), we can say that Weight is rightly skewed.



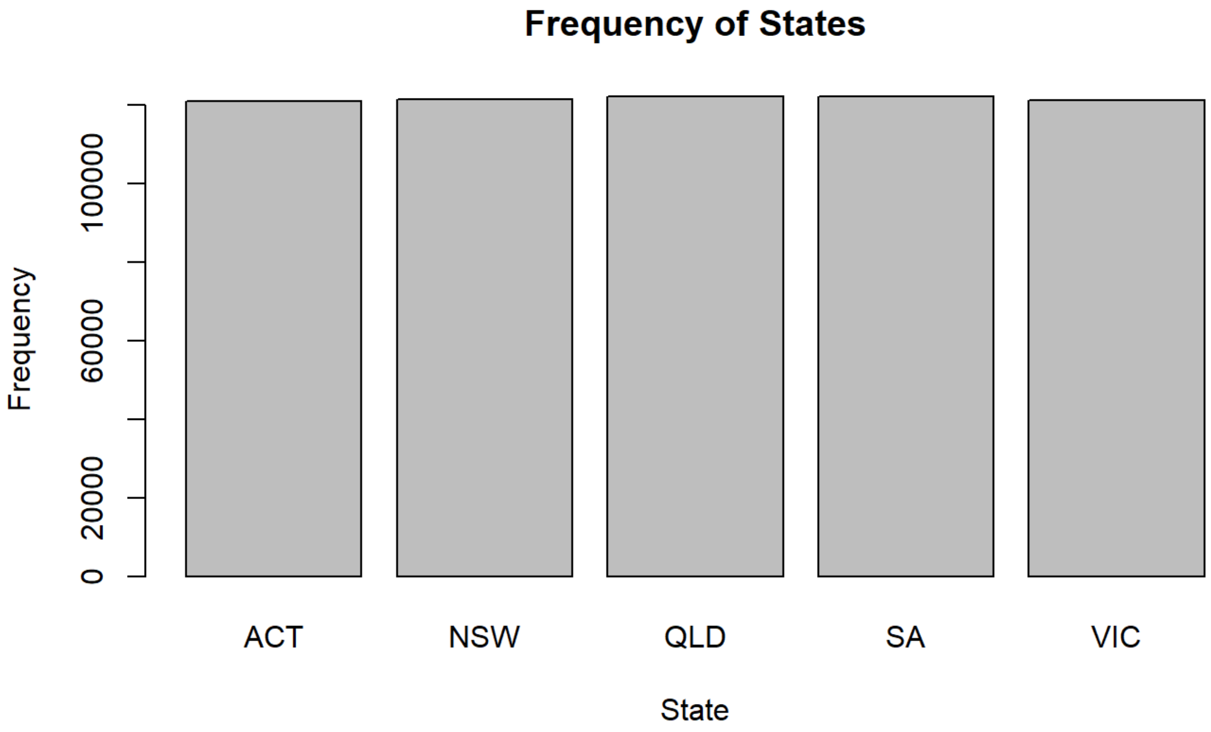
For variable Distance, with its skewness being 1.9439 and its Mean (14.85) > Median (10.00) and evidently from the histogram, we can say that Distance is rightly skewed.



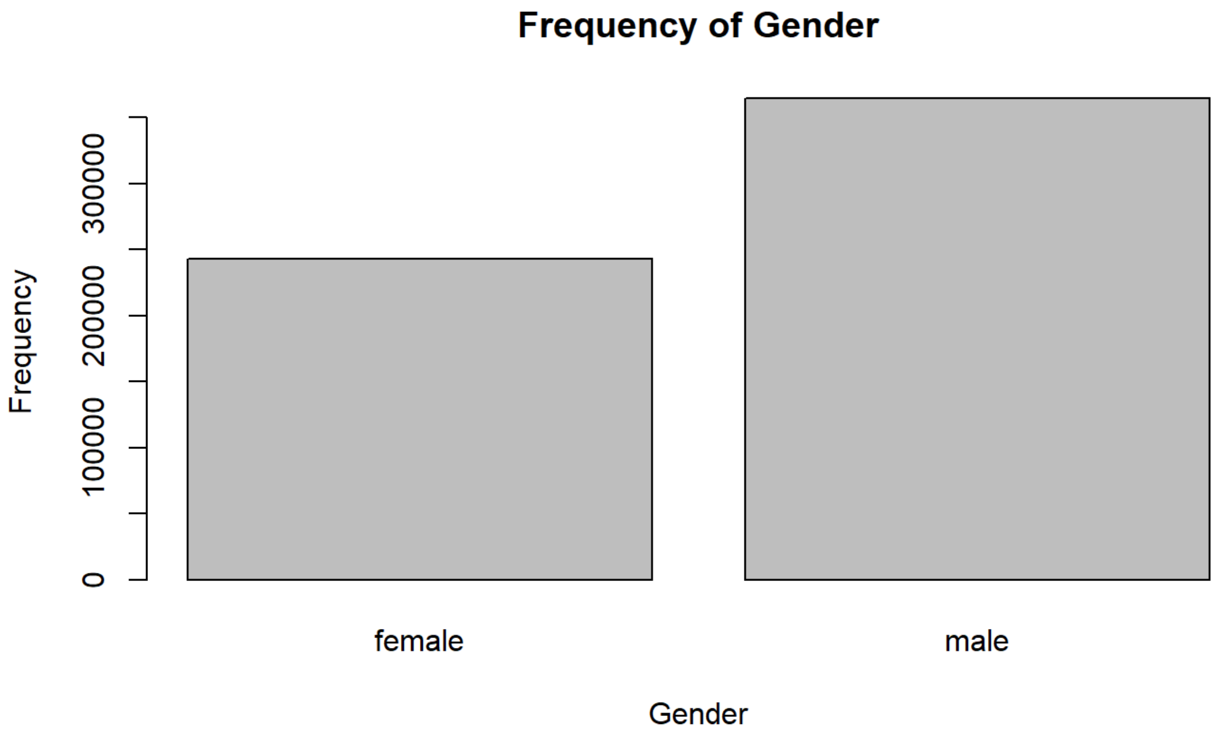
For variable Age, with its Mean (47.25) > Median (46), we can say that it is rightly skewed.



For variable Car Age, based on the Mean (7.762) > Median (5) and its skewness being 1.91, we can say that it is rightly skewed



For variable State, based on the table summarized above and the graph, we can say that the number of car insurance policies between different states are similar.



For variable Gender, we can see that there are more male drivers (364645) comparing to the number of female drivers (243052).

Question 2:

Using full dataset, estimate $\lambda(x)$ using linear predictor with linear terms x_1, x_2, x_3, x_4, x_5, x_6

To estimate $\lambda(\boldsymbol{x})$ using the GLM regression model with canonical link function, we run the GLM model with all linear terms included in the model x_1, x_2, x_3, x_4, x_5, x_6 and name it as "Model1_full", with the offset as log(exposure) and distribution family being Poisson.

Perform hypothesis testing to decide if variable state should be removed from the model. Call the resulting model as Model1. Report corresponding regression coefficients.

To decide if variable state should be removed from the model, we use Likelihood Ratio test. We will create a reduced model, named "Model1_red" with 5 covariates (x_1, x_2, x_3, x_4, x_5), without x_6 (state). We can use the Likelihood Ratio test because Model1_red and

Model1_full are nested models. As both models use the same link function and the same underlying distributional assumption for the dependent variable, we can use the Likelihood Ratio test to test between models. The Hypothesis H0 and Alternate Hypothesis (H1) is:

H0: $\beta(x_6) = 0$ given $\beta(x_1)$, $\beta(x_2)$, $\beta(x_3)$, $\beta(x_4)$, $\beta(x_5)$ are included in the model

H1: $\beta(x_6) \neq 0$ given $\beta(x_1)$, $\beta(x_2)$, $\beta(x_3)$, $\beta(x_4)$, $\beta(x_5)$ are included in the model

The formula for Likelihood Ratio test is:

$$\mathrm{LR} = D^*_{red}(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) - D^*_{full}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})).$$

With D* being Scaled Deviance, which is equal to (Deviance/ Dispersion Parameter). For Poisson distribution, the Dispersion Parameter is 1, therefore Scaled Deviance = Deviance under Poisson distribution.

From the summary table of the Model1_full and Model1_red, we can obtain the Deviance for each model, which is Scaled Deviance as well. Therefore, we can calculate the Likelihood Ratio = Deviance of Reduced model – Deviance of Full model = 7.184.

We then compare LR to the Critical value of Chi-square distribution given by R with p=0.05 (significance threshold for corresponding p-value), Degrees of Freedom=4 as degrees of freedom equal to the difference in number of parameters included in the full and reduced model, which is 4, as the difference between Model1_full and Model1_red is variable state, which has 5 levels and therefore 4 dummy variables. R returns the critical value at 9.487, which is larger than LR.

LR (7.184) < 9.847-> Fail to reject H0. When x_1, x_2, x_3, x_4, x_5 are included in the model, $x6$ (state) does not contribute significantly to the model.

Corresponding regression coefficients for model1 are:

Intercept	x_1	x_2	x_3	x_4	x_5 (male)
-2.6982	0.0001	0.0039	-0.0089	0.0207	0.0207

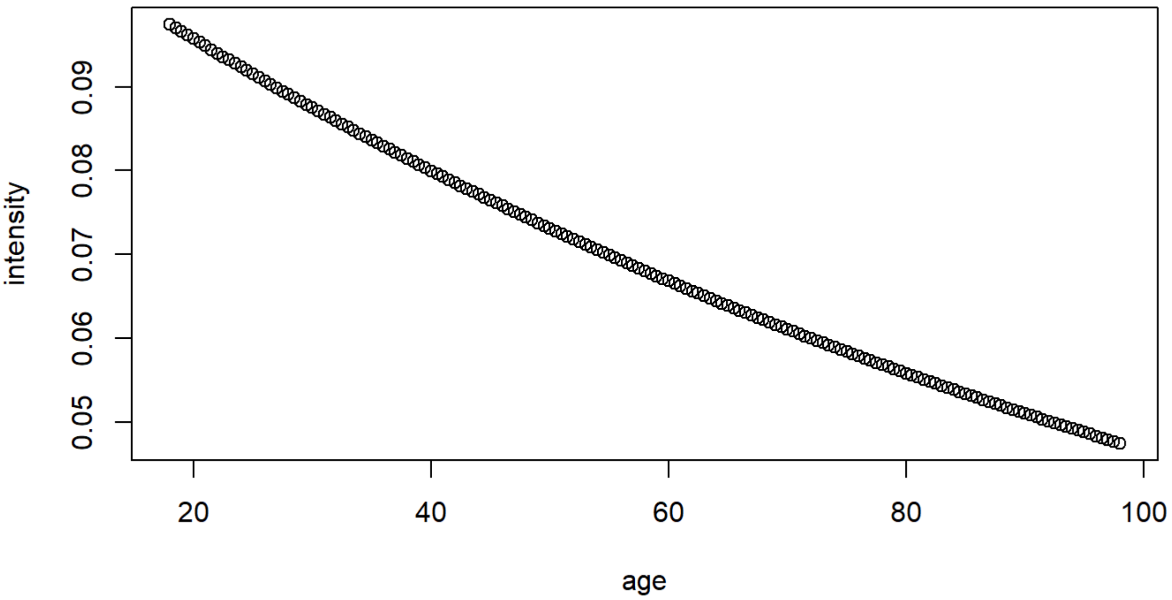
As a benchmark, report the estimate of $\lambda(x)$ for x_1= 2000, x_2= 15, x_3=30, x_4= 4, x_5 = female, x_6 = NSW.

First, we create a new data frame with x_1= 2000, x_2= 15, x_3=30, x_4= 4, x_5 = female, x_6 = NSW. Then we use the predict() function to predict λ(x) for our new data frame, using the type=”response” to specify that we want to obtain the probabilities as the output.

$\lambda(x) = 0.0865$

Plot $\lambda(x)$ versus x_3, when other predictors are for x_1= 2000, x_2= 15, x_4= 4, x_5 = female, x_6 = NSW

First we create “xage_1” variable, using the seq() function to order the age variable from the smallest to the largest value. Next, we create “yage_1” variable which is prediction for $\lambda(x)$ using model1, given values of x_1= 2000, x_2= 15, x_4= 4, x_5 = female, x_6 = NSW and exposure=1. The graph is given below. From the graph, we can see from the graph that $\lambda(x)$ goes down as age goes up, indicating the number of claims go down as age of driver increases. The relationship between λ(x) and age is negative linear correlation.



Question 3: Using full dataset, estimate $\lambda(x)$ using linear predictor with linear terms x_1, x_2, x_3, x_4, x_5, x_6 and quadratic terms x_1^2, x_2^2, x_3^2, x_4^2.

Similar to what we did for Question 2, we fit a GLM model (model2_full) with all linear terms included in the model x_1, x_2, x_3, x_4, x_5, x_6 and quadratic terms x_1^2, x_2^2, x_3^2, x_4^2 with offset being the log(exposure) and the distribution family being Poisson.

Select the best model using the backward variable selection procedure (i.e. identify what terms should be kept in the model and what terms can be dropped). Report regression coefficients corresponding to this model (call it Model2).

We run the command step(model2_full, direction=”backward”) for R to remove the least important variables as measured by in-sample error and we can find the optimized model. x_1^2 and x_6 are excluded from our model in order, letting our final model includes x_1, x_2, x_3, x_4, x_5 and x_2^2, x_3^2, x_4^2. One note is that when we run the summary() command for model2, the p-value of x_4^2 (0.0509)is slightly over the 0.05 significance threshold.

Corresponding regression coefficients for model2 are:

Intercept	x_1	x_2	x_3	x_4	x_5 (male)	x_2^2	x_3^2	x_4^2
-2.3256	0.0001	0.0019	0.0019	0.0245	-0.1684	0.00003	0.00016	-0.0001

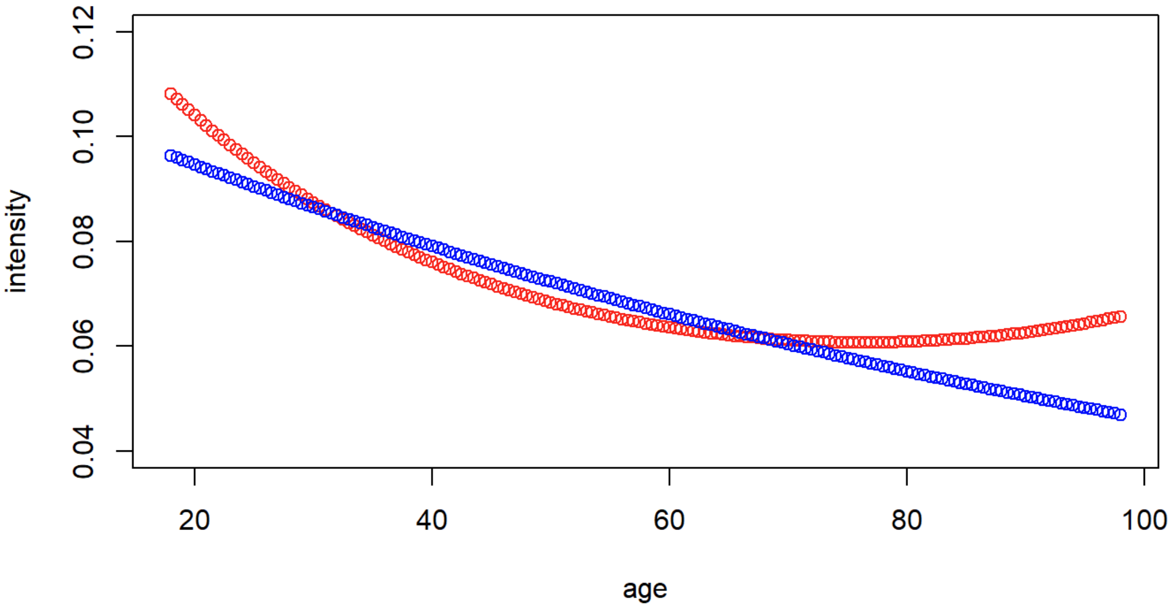
As a benchmark, report the estimate of λ(x) for x_1= 2000, x_2= 15, x_3= 30, x_4= 4, x_5 = female, x_6 = NSW.

Similar to what we did in question 2, we will also run the predict() function using model2. As the data frame in Question 3 is similar to the data frame in Question 2, we use the data frame we created in Question 2 to predict λ(x) for Question 3.

λ(x) = 0.0874

Plot λ(x) for this model versus x_3, when other predictors for x_1= 2000, x_2= 15, x_4= 4, x_5 = female, x_6 = NSW (plot this curve with the λ(x) curve obtained in (2) in the same figure)

We also create “xage_2” variable, using the seq() function to order the age variable from the smallest to the largest value, xage_2 is similar to xage. Next, we create “yage” variable which is prediction for *λ(x)* using model2, given values of x_1= 2000, x_2= 15, x_4= 4, x_5 = female, x_6 = NSW and exposure=1. The graph is given below, with the red line is prediction of λ(x) using model2 and the blue line is prediction of λ(x) using model1. We can see from the graph that under model 1, λ(x) is predicted to decrease continuously as age goes up. Under model 2, λ(x)’s pattern given age is different. Comparing to model1, drivers from 20-30 years old are predicted to get higher λ(x), while drivers from 30-70 years old are predicted to get lower λ(x) and for drivers from 70-100 years old, λ(x) is predicted to be higher.



Question 4: Compare Model1 and Model2 obtained in (2) and (3) using the likelihood ratio test and select the best model

We can use the Likelihood Ratio test because model1 and model2 are nested models. As both models use the same link function and the same underlying distributional assumption for the dependent variable, we can use Likelihood Ratio test to test between models. The Hypothesis H0 and Alternate Hypothesis (H1) are:

H0: $\beta(x_2^2)$, $\beta(x_3^2)$, $\beta(x_4^2) =0$ given x_1, x_2, x_3, x_4, x_5 are already included in the model

H1: $\beta(x_2^2)$, $\beta(x_3^2)$, $\beta(x_4^2) \neq 0$ given x_1, x_2, x_3, x_4, x_5 are already included in the model

We calculate Likelihood Ratio = Deviance of Reduced model (model1) – Deviance of Full model (model2) = 106.629.

We then compare LR to the Critical value of Chi-square distribution given by R with p=0.05 (significance threshold for corresponding p-value), Degrees of Freedom=3 as degrees of freedom equal to the difference in number of parameters included in the model1 and model2, which are 3 (x_2^2, x_3^2, x_4^2). R returns the critical value at 7.814, which is smaller than LR.

We conclude that H0 is rejected. When $\boldsymbol{x_1}$**,** $\boldsymbol{x_2}$**,** $\boldsymbol{x_3}$**,** $\boldsymbol{x_4}$**,** $\boldsymbol{x_5}$ are already included in the model x_2^2, x_3^2, x_4^2 contributed significantly to the model.

Question 5: Compare Model1 and Model2 using the 10-fold cross validation error to select the best model. Compare your findings with findings in (4) and state which model you will use to predict frequency of claims of the portfolio policies.

We compare model1 and model2 by calculating their 10-fold cross-validation error to select the model. We use the cv.glm() function and set K=10 . We want to compare the raw LOOCV of each model, and raw LOOCV of model1 is 0.06072492; while raw LOOCV of model2 is 0.06071381. As the 10-fold cross validation error for model2 is smaller than that of model1, we conclude that model2 should be used to predict frequency of claims of the portfolio policies.