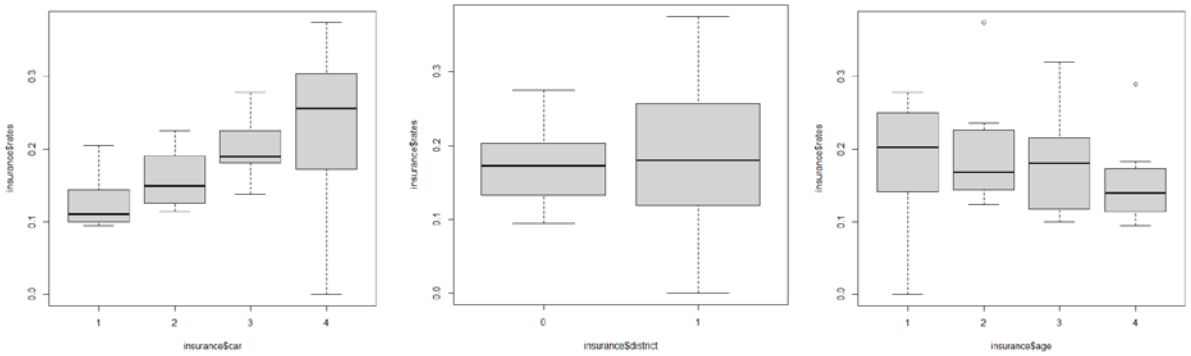
Q3) we can see from the plots below that rates increase with *CAR*, decrease with *AGE* and are higher for *DIST* = 1.




Here is the model with main effects treated as categorical.

```
glm_car.age.district.factor <-glm(y ~ factor(car) + factor(age) + factor(district),
                                  offset = log(n +1), family = "poisson", data=insurance)
```

**add1()** function is used for adding interactions to model above. We see that neither of the missing two-factor interactions is significant by itself at the conventional five percent level

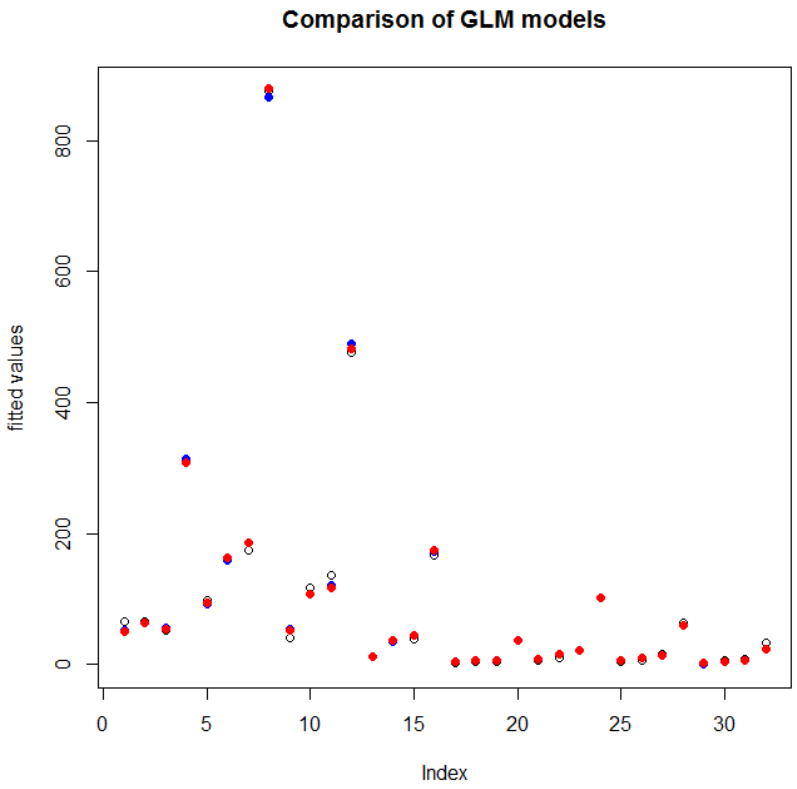
```
Model:
y ~ factor(car) + factor(age) + factor(district)
                             Df Deviance    AIC     LRT Pr(>Chi)
<none>                           24.702 209.06
factor(car):factor(age)       9   13.955 216.31 10.7467   0.2935
factor(car):factor(district)  3   20.659 211.02  4.0430   0.2569
factor(age):factor(district)  3   20.172 210.53  4.5296   0.2097
```

**step()** function is used for stepwise comparison . we see that age and car are the most significant varibles based on deviance and AIC.

```
Start:  AIC=209.06
y ~ factor(car) + factor(age) + factor(district)

                               Df Deviance    AIC
<none>                              24.702 209.06
+ factor(age):factor(district)  3   20.172 210.53
+ factor(car):factor(district)  3   20.659 211.02
+ factor(car):factor(age)       9   13.955 216.31
- factor(district)              1   36.756 219.12
- factor(age)                   3  106.775 285.14
- factor(car)                   3  112.818 291.18
```

Part c) So we can use only age and car type and treat them as continuous. Plot below is comparis on of models in part b and part c. we can see part c model fits the data fairly well. (Black = actual data, Blue = part b, red= part c)




## Appendix: R code

```
require(dobson)

require(ggplot2)


#preparation of data

insurance <- dobson::insurance

insurance$rates <- insurance$y/insurance$n

#plots
```

```
boxplot(insurance$rates ~ insurance$age)

boxplot(insurance$rates ~ insurance$district)

boxplot(insurance$rates ~ insurance$car)


########################## #models

glm_car.age <- glm(y ~ car + age,

        offset = log(n +1), family = "poisson", data=insurance)


#part c model - AGE and CAR treated as continuous variables.

glm_car.age.district <- glm(y ~ car + age + factor(district),

        offset = log(n +1), family = "poisson", data=insurance)


glm_car.age.district.factor <-glm(y ~ factor(car) + factor(age) + factor(district),

                offset = log(n +1), family = "poisson", data=insurance)



#######################################

add1(glm_car.age.district.factor, ~ .^2, test = "Chisq")

search = step(glm_car.age.district.factor, ~.^2)

######################################## plots

plot( insurance$y, col="black", pch=21,

    main="Comparison of GLM models", ylab = "fitted values")

# points(glm_car.age$fitted, col="red", pch=19)

points(glm_car.age.district.factor$fitted , col="blue", pch=19)

points(glm_car.age.district$fitted, col="red", pch=19) #part c model
```