HW4

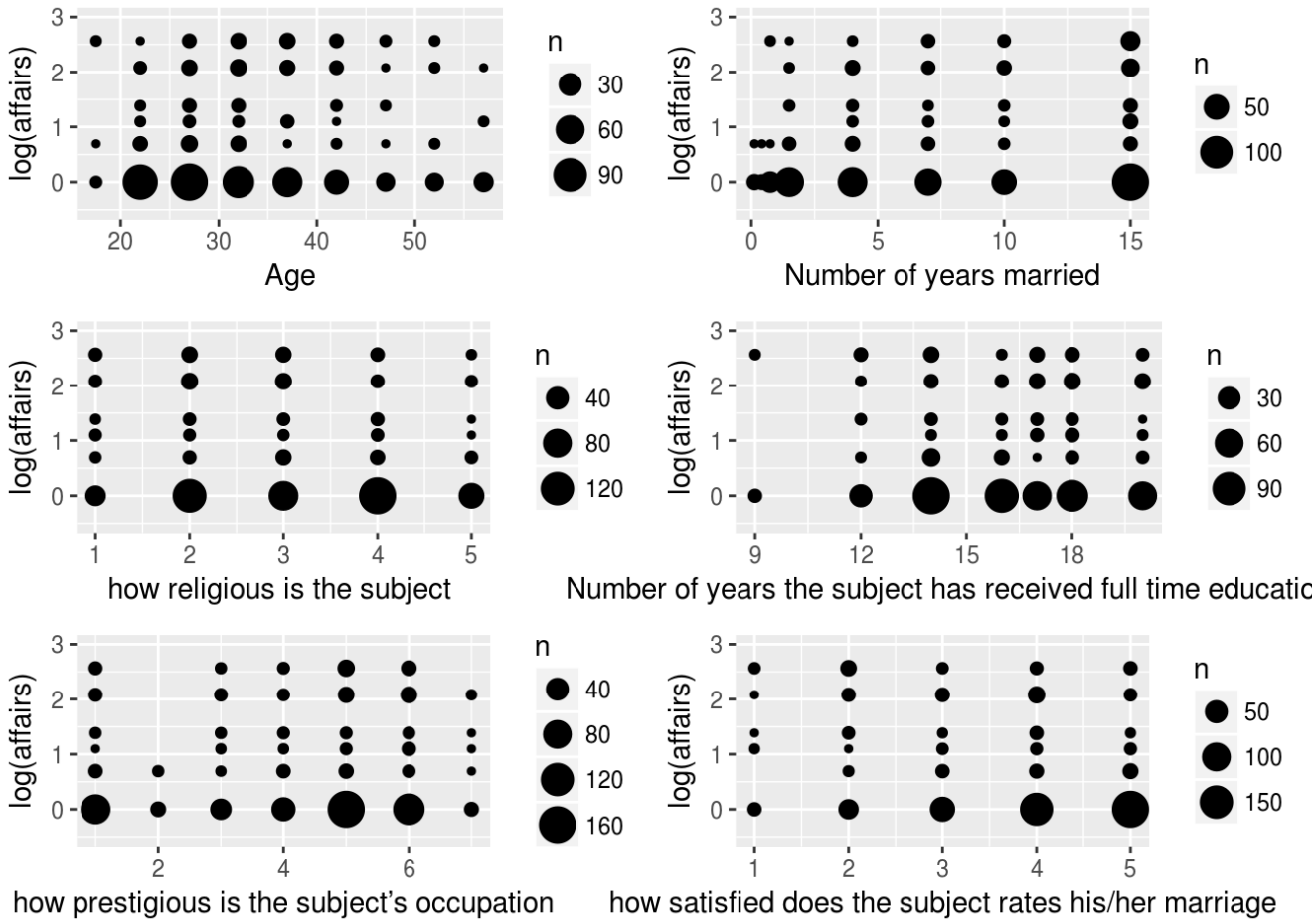
Chong Zhang

November 8, 2018

Question 1: Exploratory Data Analysis

1. Using a scatterplot describe the relationship between the log of number of extramarital affairs log(nbaffairs+1) and the numeric independent variables. Add 1 to number of affairs before log transformation to avoid zero value during transformation. Describe the general trend (direction and form). What are the values of the correlation coefficients? Please interpret.

```
p1 = ggplot(affair, aes(x=affair$age,y=log(affair$nbaffairs+1)))+geom_count()+xlab('Age')
+ylab('log(affairs)')+scale_y_continuous(limits = c(-0.5, 3))
p2 = ggplot(affair, aes(x= affair$ym,y=log(affair$nbaffairs+1)))+geom_count()+xlab('Numbe
r of years married')+ylab('log(affairs)')+scale_y_continuous(limits = c(-0.5, 3))
p3 = ggplot(affair, aes(x=affair$religious,y=log(affair$nbaffairs+1)))+geom_count()+xlab(
'how religious is the subject ')+ylab('log(affairs)')+scale_y_continuous(limits = c(-0.5,
 3))
p4 = ggplot(affair, aes(x= affair$education,y=log(affair$nbaffairs+1)))+geom_count()+xlab
('Number of years the subject has received full time education')+ylab('log(affairs)')+sca
le_y_continuous(limits = c(-0.5, 3))
p5 = ggplot(affair, aes(x= affair$occupation,y=log(affair$nbaffairs+1)))+geom_count()+xla
b('how prestigious is the subject’s occupation')+ylab('log(affairs)')+scale_y_continuous(
limits = c(-0.5, 3))
p6 = ggplot(affair, aes(x= affair$rate,y=log(affair$nbaffairs+1)))+geom_count()+xlab('how
 satisfied does the subject rates his/her marriage')+ylab('log(affairs)')+scale_y_continu
ous(limits = c(-0.5, 3))
grid.arrange(p1,p2,p3,p4,p5,p6)
```




```
cor1 = cor(x= affair$age,y=log(affair$nbaffairs+1))
cor2 = cor(x= affair$ym,y=log(affair$nbaffairs+1))
cor3 = cor(x= affair$religious,y=log(affair$nbaffairs+1))
cor4 = cor(x= affair$education,y=log(affair$nbaffairs+1))
cor5 = cor(x= affair$occupation,y=log(affair$nbaffairs+1))
cor6 = cor(x= affair$rate,y=log(affair$nbaffairs+1))
```

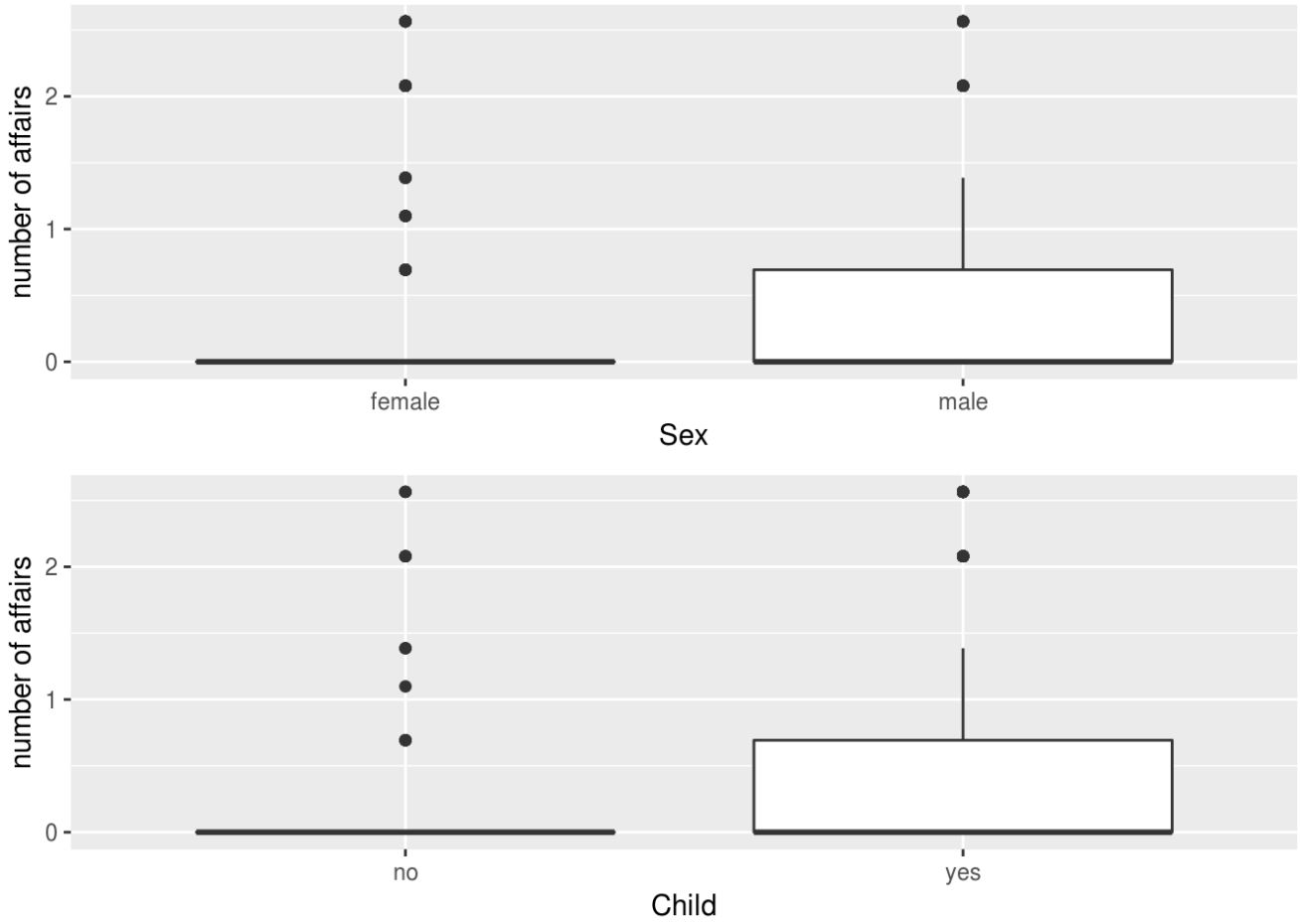
There are 6 numeric independent variables. They are **Age,Number of years married, Religous, Education, Occupation and Rate**.

Based on scatterplots, It seems that there is no apprent relatonship between number of affairs and those three numeric independent variables.

The correlation coefficinets between log(naffairs+1) and Age,Number of years married, Religous, Education, Occupation and Rate are **`0.0895224, 0.1835569, -0.1501462, 0.0176202, 0.0530024, -0.287666`** . The correlation coefficient between number of affairs and number of years married and how religous is the subject are slightly larger than 0.15, which indicates a **very very mild** relationship. The correlation coefficient between number of affairs and how satisfied does the subject rates his/her marriage is larger than 0.25, which indicates a **mild** relationship. The rest are very close to 0, which means no correlation.

2. Describe the relationship between the log transform of the number of extramarital affairs and the categorical independent variables sex and child. Do the log transform vary with the categorical variables?

```
p1 = ggplot(affair, aes(x=affair$sex,y=log(affair$nbaffairs+1)))+geom_boxplot()+xlab('Se
x')+ylab('number of affairs')
p2 = ggplot(affair, aes(x=affair$child,y=log(affair$nbaffairs+1)))+geom_boxplot()+xlab('C
hild')+ylab('number of affairs')
grid.arrange(p1,p2)
```



It seems that the log transformed number of affairs varies a lot with the categorical variables. Majority of subjects who are female or have no child have no affairs. Whereas subjects who are male or have children tend to have more affairs.

3. Based on this exploratory analysis, is it reasonable to assume a linear relationship between the log transform of number of extramarital affairs and the predictors?

Based on this exploratory analysis, **it is not reasonable to assume a linear relationship.**

Question 2: Fitting the Regression Model.

Fit a Poisson regression to evaluate the relationship between the number of extramarital affairs and all the predictors. (i) What are the model parameters and what are their estimates?

```
method = glm(data = affair,nbaffairs~., family = 'poisson')
summary(method)
```

```
##
## Call:
## glm(formula = nbaffairs ~ ., family = "poisson", data = affair)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5331  -1.5813  -1.1597  -0.7084   8.3386
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.5528720  0.2877313   8.872  < 2e-16 ***
## sexmale      0.0577932  0.0816503   0.708   0.4791
## age         -0.0330294  0.0059571  -5.545 2.95e-08 ***
## ym           0.1169683  0.0107798  10.851  < 2e-16 ***
## childyes    -0.0026631  0.1027267  -0.026   0.9793
## religious   -0.3547250  0.0309683 -11.454  < 2e-16 ***
## education    0.0006042  0.0169084   0.036   0.9715
## occupation   0.0717169  0.0247803   2.894   0.0038 **
## rate        -0.4105613  0.0279314 -14.699  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2925.5  on 600  degrees of freedom
## Residual deviance: 2359.6  on 592  degrees of freedom
## AIC: 2871.5
##
## Number of Fisher Scoring iterations: 7
```

Model parameters include: **sexmale, age, ym, childyes, religous, education, occupation, rate and intercept**. The estimates of each parameters are shown above.

ii. Write down the equation for the estimated log transformation of number of affairs given the predicting variables

$$Log(naffairs) = 2.55 + 0.058 * sexmale - 0.033 * age + 0.117 * ym - 0.003 * childyes - 0.355 * religous + 0.0006 * education + 0.072 * occupation - 0.411 * rate$$

iii. Interpret the estimated value of the parameters corresponding with sex and age in the context of the problem.

For **sex**, it means that the expected number of affairs per year for male vs female is exp(0.058)= `1.059715` , holding all other variables constant.

For **age**, it means that the expected number of affairs per year for one unit increase in age is exp(0.033)= `0.9675386` , holding all other variables constant.

Question 3: Testing the significance of the linear relationship observed in the data.

i. Find a 99% confidence interval for the parameters corresponding to all predictors plus the intercept.

```
confint(object = method, level = 0.99)
```

```
##                   0.5 %      99.5 %
## (Intercept)  1.80794350  3.29050918
## sexmale     -0.15209887  0.26872946
## age         -0.04862963 -0.01792676
## ym           0.08924910  0.14479004
## childyes    -0.26318945  0.26641958
## religious   -0.43473946 -0.27515410
## education   -0.04270046  0.04442813
## occupation   0.00812714  0.13584429
## rate        -0.48238171 -0.33844153
```

99% confidence interval for the parameters corresponding to all predictors plus the intercept are shown above.

ii. Which variables are significant at the significance level α = 0.01? Give the p-value for any variable that is not significant.

Intercept, age, ym, religous, occupation, rate are significant at the level α = 0.01. For variables that are not significant, **Sexmale: p = 0.4791, childyes: p = 0.9793, education: p = 0.9715**.

Question 4: Goodness of Fit.

i. Perform a goodness of fit test for the model? Does the model fit the data well?

```
with(method, cbind(res.deviance = deviance, df = df.residual,
               p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

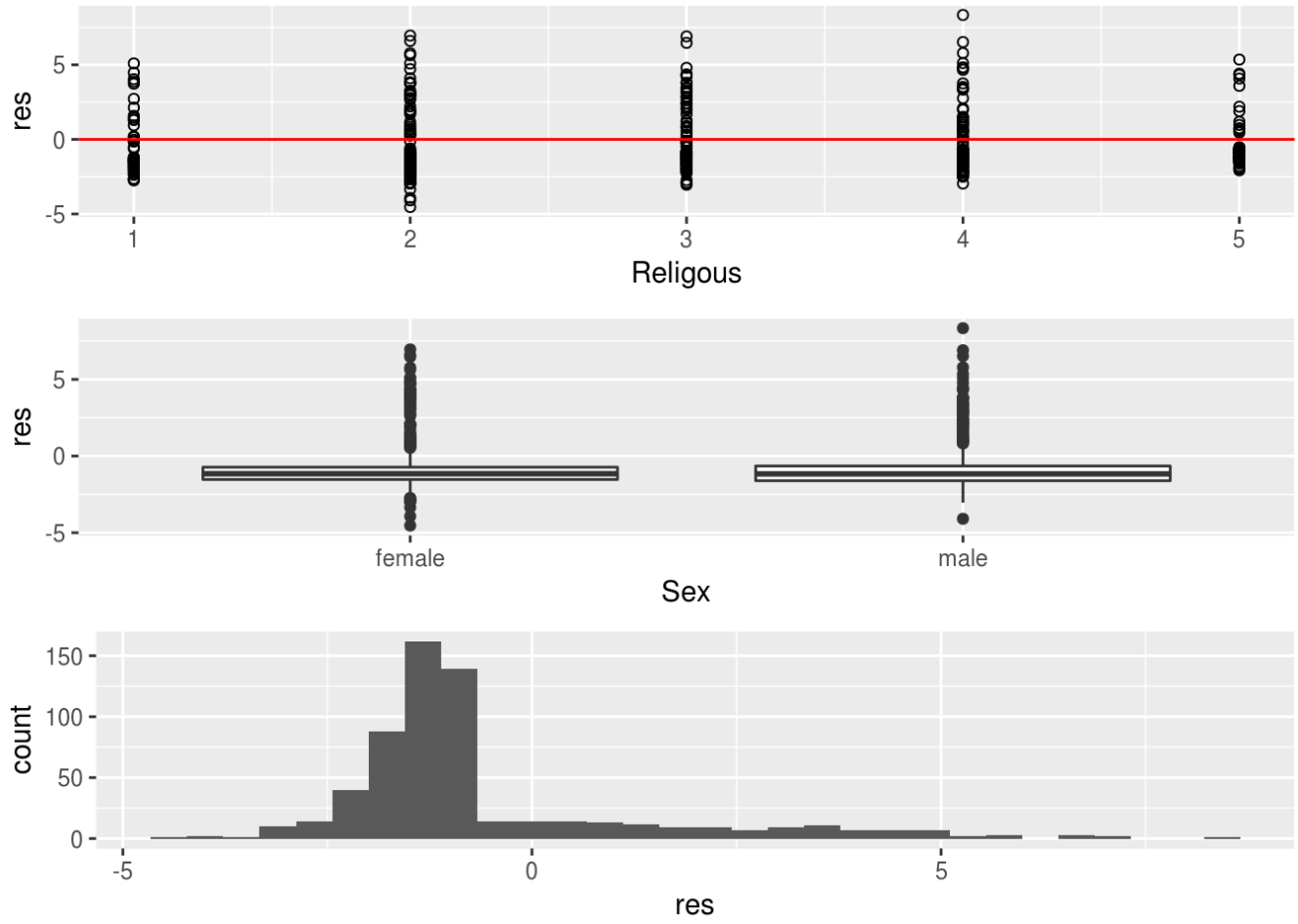
```
##      res.deviance  df             p
## [1,]     2359.551 592 6.621396e-209
```

Since the **p value is close to 0**, it indicates that the model is **not a good fit**.

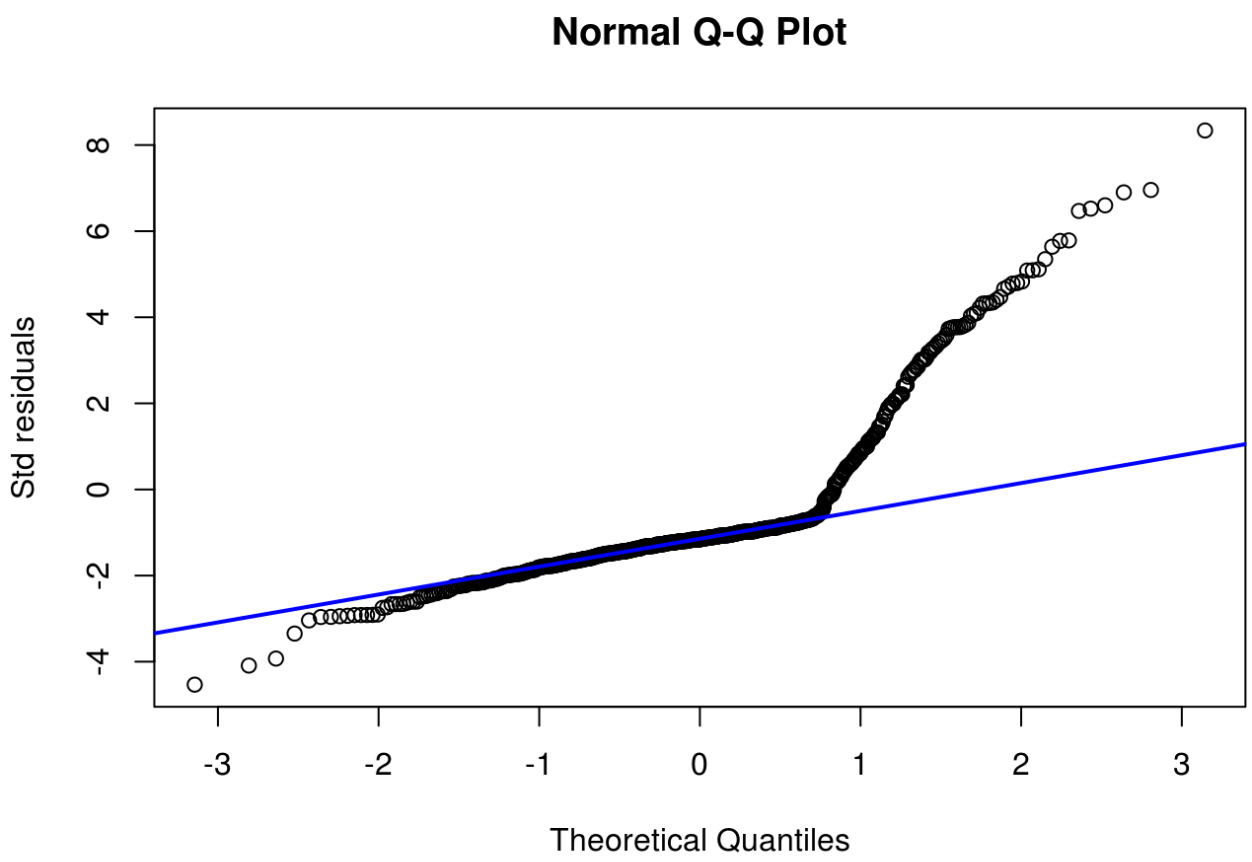
ii. Residual Analysis

Produce the following deviance residual plots: 1) Scatterplot of religious against the residuals, 2) Boxplot of the residuals by sex, 3) QQPlot of the residuals, 4) Histogram of the residuals. Comment on the plots.

```
res = resid(method,type="deviance")
p1 =  ggplot(data = affair, aes(x=affair$religious, y=res))+geom_point(shape =1)+xlab('Re
ligous')+geom_hline(yintercept=0,color ='red')
p2 = ggplot(data = affair, aes(x=affair$sex, y=res))+geom_boxplot()+xlab('Sex')
p3 = ggplot(data = affair, aes(x=res))+geom_histogram()
grid.arrange(p1, p2, p3)
```




```
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
```



For the scatterplot, the residues shift slightly towards above 0, which means that the residues are not evenly scattered around 0. Thus the linearity assumption does not hold. For the boxplot, the meadian of residuals are smaller than 0, approximately around -1, and there are a lot outliers. Thus the linearity assumption does not hold. For the histograme of the residuals, marjority of the residuals are smaller than 0. It is not normally distributed. Thus the normality assumption does not hold. For the qqplot, there are significant departures at both end. Thus the normality assumption does not hold.

Question 5: Prediction.

Provide the estimated number of extramarital affairs engaged for the past year and its 99% confidence interval of a 45-year-old man with 15 years of marriage with a child. The man is somewhat religious (religious = 4), has a bachelor degree (education = 16), is a sales clerk (occupation = 5) and has ambivalent feeling towards his marriage (rate = 3)

```
newdata = data.frame('sex'='male', 'age'=45, 'ym'=15, 'child'='yes', 'religious'=4, 'educ
ation'=16, 'occupation'=5,'rate'=3)

#obtain the prediction on the scale of linear predictors
response = predict(object = method, newdata=newdata, type = 'link', se.fit = T)

#Get the upper and lower ends in the scale of linear predictors
upper = response$fit+2.576*response$se.fit
lower = response$fit-2.576*response$se.fit
fit = response$fit

#apply the link function "log" to response, upper and lower to get the results in the sca
le of response
fit = method$family$linkinv(fit)
upper = method$family$linkinv(upper)
lower = method$family$linkinv(lower)
```

The prediction is **1.8109872** , with upper end **2.1720252** , lower end **1.5099616** for 99% confidence interval.