STAT811

Sukhdeep Sedha

29 October 2018

# Question 1

# a) Reading the data

data<-read.csv("fish\_assn3.csv",header = T)  
head(data)

## nofish livebait camper persons child count  
## 1 1 0 0 1 0 0  
## 2 0 1 1 1 0 0  
## 3 0 1 0 1 0 0  
## 4 0 1 1 2 1 0  
## 5 0 1 0 1 0 1  
## 6 0 1 1 4 2 0

attach(data)  
library('gamlss')

## Warning: package 'gamlss' was built under R version 3.4.4

## Loading required package: splines

## Loading required package: gamlss.data

## Warning: package 'gamlss.data' was built under R version 3.4.4

## Loading required package: gamlss.dist

## Warning: package 'gamlss.dist' was built under R version 3.4.4

## Loading required package: MASS

## Loading required package: nlme

## Warning: package 'nlme' was built under R version 3.4.4

## Loading required package: parallel

## \*\*\*\*\*\*\*\*\*\* GAMLSS Version 5.1-2 \*\*\*\*\*\*\*\*\*\*

## For more on GAMLSS look at http://www.gamlss.org/

## Type gamlssNews() to see new features/changes/bug fixes.

library('plyr')

## Warning: package 'plyr' was built under R version 3.4.4

library('car')

## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.4.4

library('xtable')

## Warning: package 'xtable' was built under R version 3.4.4

library('MASS')  
library('boot')

## Warning: package 'boot' was built under R version 3.4.4

##   
## Attaching package: 'boot'

## The following object is masked from 'package:car':  
##   
## logit

## The following object is masked from 'package:gamlss.data':  
##   
## aids

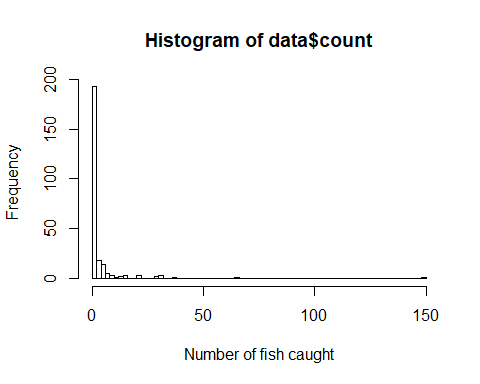
Missing value: None

Let’s look at the distribution of the response variable i.e number of fish caught (count)

count(data$count)

## x freq  
## 1 0 142  
## 2 1 31  
## 3 2 20  
## 4 3 12  
## 5 4 6  
## 6 5 10  
## 7 6 4  
## 8 7 3  
## 9 8 2  
## 10 9 2  
## 11 10 1  
## 12 11 1  
## 13 13 1  
## 14 14 1  
## 15 15 2  
## 16 16 1  
## 17 21 2  
## 18 22 1  
## 19 29 1  
## 20 30 1  
## 21 31 1  
## 22 32 2  
## 23 38 1  
## 24 65 1  
## 25 149 1

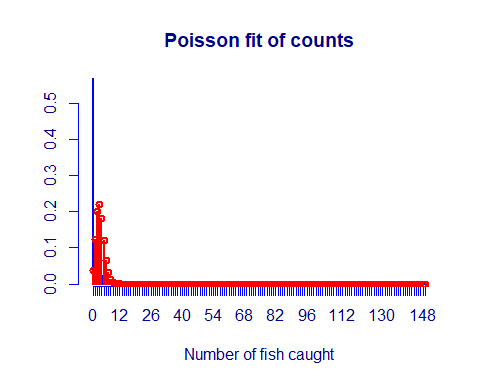
hist(data$count,breaks = 100, xlab = "Number of fish caught")



There is a very high zero frequency, so it makes sense for us to consider ZERO-INFLATED models, as poisson and negative binomial are unlikely to provide us with a satifactory fit.

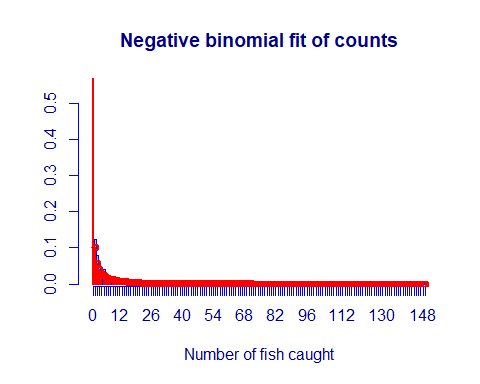
Let’s look at the fitted poisson and negative binomial distributions

histDist(data$count,family = PO, main= "Poisson fit of counts", xlab = "Number of fish caught", )



##   
## Family: c("PO", "Poisson")   
## Fitting method: "nlminb"   
##   
## Call: gamlssML(formula = data$count, family = "PO", data = sys.parent())   
##   
##   
## Mu Coefficients:  
## [1] 1.193  
##   
## Degrees of Freedom for the fit: 1 Residual Deg. of Freedom 249   
## Global Deviance: 3295.43   
## AIC: 3297.43   
## SBC: 3300.95

histDist(data$count,family = NBI, main= "Negative binomial fit of counts", xlab = "Number of fish caught")



##   
## Family: c("NBI", "Negative Binomial type I")   
## Fitting method: "nlminb"   
##   
## Call:   
## gamlssML(formula = data$count, family = "NBI", data = sys.parent())   
##   
## Mu Coefficients:  
## [1] 1.193  
## Sigma Coefficients:  
## [1] 1.694  
##   
## Degrees of Freedom for the fit: 2 Residual Deg. of Freedom 248   
## Global Deviance: 928.879   
## AIC: 932.879   
## SBC: 939.922

Looking at the above histgram it is evident that, poisson distribution does not handle the high frequency of zeroes very well, on the other hand we can see that negative binomial fit is compartively better than poisson.

To arrive at the final result we will need to conduct formal regression analysis to check if ZERO-INFLATED model performs better than negative binomial and poisson.

Poisson GLM Initial screening :

library('knitr')

## Warning: package 'knitr' was built under R version 3.4.4

df = read.table('table1.txt',header = TRUE)  
kable(df)

|  |  |
| --- | --- |
| Variate | P.Value |
| Person | <0.0001 |
| Child | <0.0001 |
| Camper | <0.0001 |
| All the co | variates, pass the initial screening test. |

Model selection using AIC:

library('knitr')  
df = read.table('table2.txt',header = TRUE, sep = '\t')  
kable(df)

|  |  |
| --- | --- |
| Model..Poisson.with.log.link. | AIC |
| Person | 2758.1 |
| Child | 2891.8 |
| factor(Camper) | 3115.2 |
| Person+Child | 1809.6 |
| Person+factor(Camper) | 2537.1 |
| Child+factor(Camper) | 2723.2 |
| Person+Child+factor(Camper) | 1682.1 |

AIC picks the model which compromises of Person+Child+factor(Camper)

Final model combining the selected covariates:

model<-glm(count~persons+child+factor(camper),family = poisson(link = log),data= data)  
summary(model)

##   
## Call:  
## glm(formula = count ~ persons + child + factor(camper), family = poisson(link = log),   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.8096 -1.4431 -0.9060 -0.0406 16.1417   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.98183 0.15226 -13.02 <2e-16 \*\*\*  
## persons 1.09126 0.03926 27.80 <2e-16 \*\*\*  
## child -1.68996 0.08099 -20.87 <2e-16 \*\*\*  
## factor(camper)1 0.93094 0.08909 10.45 <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: 2958.4 on 249 degrees of freedom  
## Residual deviance: 1337.1 on 246 degrees of freedom  
## AIC: 1682.1  
##   
## Number of Fisher Scoring iterations: 6

log(count) = -1.98183 + 1.09126(person) -1.68996(child)+ 0.93094[factor(camper)]

All covariates are significant. For a poisson response’ distribution, this model provides the best AIC, but we can see that there is a case of overdispersion since we see a residual deviance of 1337.1 on 246 degrees of freedom.

In order to address this we will fit a Negative binomial model.

NBI model covariate selection:

library('knitr')  
df = read.table('table3.txt',header = TRUE, sep = '\t')  
kable(df)

|  |  |
| --- | --- |
| Covariate | P.value |
| Person | <0.0001 |
| Child | <0.0001 |
| factor(Camper) | <0.0001 |
| All the considere | d covariates pass the screening test. |

Let’s look at the models:

library('knitr')  
df = read.table('table4.txt',header = TRUE, sep = '\t')  
kable(df)

|  |  |
| --- | --- |
| Model..NBI. | AIC |
| Person | 899.89 |
| Child | 894.83 |
| factor(Camper) | 923.28 |
| Person+Child | 825.04 |
| Person+factor(camper) | 893.45 |
| child+factor(Camper) | 887.42 |
| Person+Child+factor(Camper) | 820.44 |

AIC picks up the model with all the covariates as the best one.

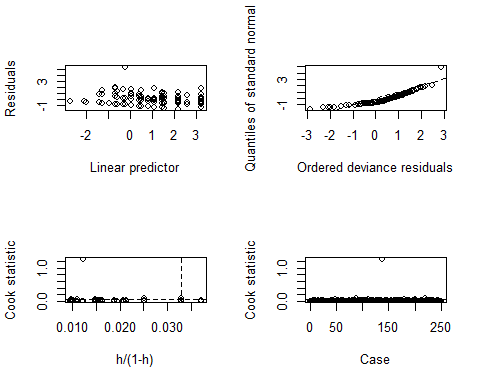
model=glm.nb(count~persons+child+factor(camper), data=data)  
summary(model)

##   
## Call:  
## glm.nb(formula = count ~ persons + child + factor(camper), data = data,   
## init.theta = 0.4635287626, link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6673 -0.9599 -0.6590 -0.0319 4.9433   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.6250 0.3304 -4.918 8.74e-07 \*\*\*  
## persons 1.0608 0.1144 9.273 < 2e-16 \*\*\*  
## child -1.7805 0.1850 -9.623 < 2e-16 \*\*\*  
## factor(camper)1 0.6211 0.2348 2.645 0.00816 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.4635) family taken to be 1)  
##   
## Null deviance: 394.25 on 249 degrees of freedom  
## Residual deviance: 210.65 on 246 degrees of freedom  
## AIC: 820.44  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.4635   
## Std. Err.: 0.0712   
##   
## 2 x log-likelihood: -810.4440

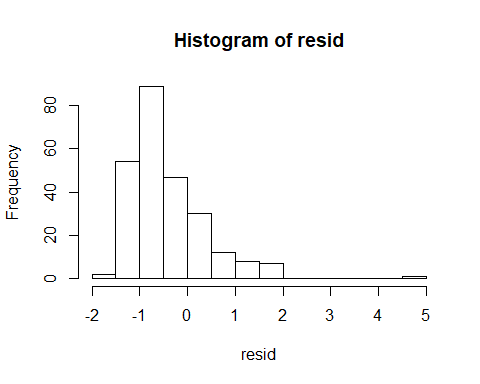
All the covariates in the model are highly significant. We can also confirm that now residual deviance looks much better than before i.e 210.65 on 246 degrees of freedom.

Let’s look at the diagnostic plots:

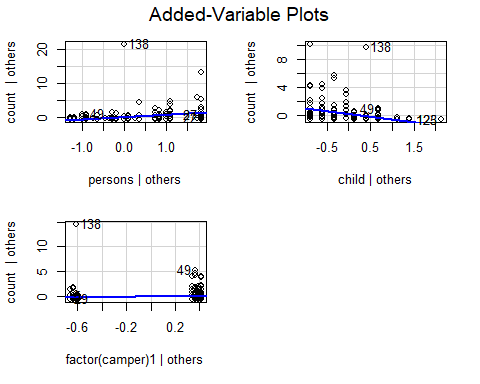
glm.diag.plots(model)



resid <- residuals(model, type="deviance")  
hist(resid, breaks=10)



avPlots(model)



Not good, diagnostics plots show us that the ordered deviance residuals are not normally distributed (refer to the QQplot and hist), we also see a observation which has high leverage and cook’s statistic. Therefore we can conclude that this model fails to meet our criteria.

Let’s look at Zero-Inflated Poisson regression approach:

model<-gamlss(count~persons+child +factor(camper),sigma.formula=~child+persons+  
factor(camper),family =ZIP,data = data)

## GAMLSS-RS iteration 1: Global Deviance = 1573.035   
## GAMLSS-RS iteration 2: Global Deviance = 1512.495   
## GAMLSS-RS iteration 3: Global Deviance = 1506.305   
## GAMLSS-RS iteration 4: Global Deviance = 1505.575   
## GAMLSS-RS iteration 5: Global Deviance = 1505.487   
## GAMLSS-RS iteration 6: Global Deviance = 1505.486

summary(model)

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("ZIP", "Poisson Zero Inflated")   
##   
## Call: gamlss(formula = count ~ persons + child + factor(camper),   
## sigma.formula = ~child + persons + factor(camper),   
## family = ZIP, data = data)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: log  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.81052 0.17097 -4.741 3.64e-06 \*\*\*  
## persons 0.83249 0.04396 18.935 < 2e-16 \*\*\*  
## child -1.14334 0.09305 -12.288 < 2e-16 \*\*\*  
## factor(camper)1 0.73022 0.09317 7.838 1.45e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: logit  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.6562 0.5162 3.208 0.00151 \*\*   
## child 1.9012 0.3263 5.826 1.80e-08 \*\*\*  
## persons -0.9208 0.1993 -4.620 6.24e-06 \*\*\*  
## factor(camper)1 -0.8304 0.3530 -2.352 0.01946 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 250   
## Degrees of Freedom for the fit: 8  
## Residual Deg. of Freedom: 242   
## at cycle: 6   
##   
## Global Deviance: 1505.487   
## AIC: 1521.487   
## SBC: 1549.658   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

We see that AIC has worsened! For the NBI model shown previously.

Let’s check what happens what happens when we consider not modelling the sigma“”

model=gamlss(count~persons+child +factor(camper),family =ZIP,data = data)

## GAMLSS-RS iteration 1: Global Deviance = 1593.971   
## GAMLSS-RS iteration 2: Global Deviance = 1557.743   
## GAMLSS-RS iteration 3: Global Deviance = 1553.545   
## GAMLSS-RS iteration 4: Global Deviance = 1553.014   
## GAMLSS-RS iteration 5: Global Deviance = 1552.958   
## GAMLSS-RS iteration 6: Global Deviance = 1552.955   
## GAMLSS-RS iteration 7: Global Deviance = 1552.954

summary(model)

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("ZIP", "Poisson Zero Inflated")   
##   
## Call: gamlss(formula = count ~ persons + child + factor(camper),   
## family = ZIP, data = data)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: log  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.01142 0.17570 -5.756 2.56e-08 \*\*\*  
## persons 0.87571 0.04497 19.474 < 2e-16 \*\*\*  
## child -1.36763 0.09407 -14.538 < 2e-16 \*\*\*  
## factor(camper)1 0.80167 0.09413 8.517 1.68e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: logit  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.3718 0.1873 -1.985 0.0483 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 250   
## Degrees of Freedom for the fit: 5  
## Residual Deg. of Freedom: 245   
## at cycle: 7   
##   
## Global Deviance: 1552.954   
## AIC: 1562.954   
## SBC: 1580.561   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

No, improvement yet.

Model using Zero inflated Negative binomial distribution:

model=gamlss(count~persons+child+factor(camper),nu.formula = ~persons+child+factor(camper),family=ZINBI,data = data)

## GAMLSS-RS iteration 1: Global Deviance = 2588.319   
## GAMLSS-RS iteration 2: Global Deviance = 1558.449   
## GAMLSS-RS iteration 3: Global Deviance = 1286.903   
## GAMLSS-RS iteration 4: Global Deviance = 964.8234   
## GAMLSS-RS iteration 5: Global Deviance = 810.6044   
## GAMLSS-RS iteration 6: Global Deviance = 810.4465   
## GAMLSS-RS iteration 7: Global Deviance = 810.4455

summary(model)

## Warning in summary.gamlss(model): summary: vcov has failed, option qr is used instead

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("ZINBI", "Zero inflated negative binomial type I")   
##   
## Call:   
## gamlss(formula = count ~ persons + child + factor(camper), nu.formula = ~persons +   
## child + factor(camper), family = ZINBI, data = data)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: log  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.6293 0.3537 -4.606 6.59e-06 \*\*\*  
## persons 1.0605 0.1329 7.981 5.55e-14 \*\*\*  
## child -1.7828 0.2118 -8.417 3.20e-15 \*\*\*  
## factor(camper)1 0.6273 0.1936 3.240 0.00136 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: log  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.77105 0.09387 8.214 1.17e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Nu link function: logit   
## Nu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -3.604e+01 1.806e+04 -0.002 0.998  
## persons -6.501e-07 6.804e+03 0.000 1.000  
## child 1.050e-06 8.898e+03 0.000 1.000  
## factor(camper)1 -8.624e-07 1.287e+04 0.000 1.000  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 250   
## Degrees of Freedom for the fit: 9  
## Residual Deg. of Freedom: 241   
## at cycle: 7   
##   
## Global Deviance: 810.4455   
## AIC: 828.4455   
## SBC: 860.1387   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Let’s try a variation by modelling only mu paratmeter

model<-gamlss(count~persons+child+factor(camper),family=ZINBI,data =data)

## GAMLSS-RS iteration 1: Global Deviance = 2588.319   
## GAMLSS-RS iteration 2: Global Deviance = 1558.449   
## GAMLSS-RS iteration 3: Global Deviance = 1286.903   
## GAMLSS-RS iteration 4: Global Deviance = 964.8234   
## GAMLSS-RS iteration 5: Global Deviance = 810.6044   
## GAMLSS-RS iteration 6: Global Deviance = 810.4465   
## GAMLSS-RS iteration 7: Global Deviance = 810.4455

summary(model)

## Warning in summary.gamlss(model): summary: vcov has failed, option qr is used instead

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("ZINBI", "Zero inflated negative binomial type I")   
##   
## Call:   
## gamlss(formula = count ~ persons + child + factor(camper), family = ZINBI,   
## data = data)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: log  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.6293 0.3537 -4.606 6.59e-06 \*\*\*  
## persons 1.0605 0.1329 7.981 5.55e-14 \*\*\*  
## child -1.7828 0.2118 -8.417 3.20e-15 \*\*\*  
## factor(camper)1 0.6273 0.1936 3.240 0.00136 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: log  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.77105 0.09387 8.214 1.17e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Nu link function: logit   
## Nu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -36.04 6324.56 -0.006 0.995  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 250   
## Degrees of Freedom for the fit: 6  
## Residual Deg. of Freedom: 244   
## at cycle: 7   
##   
## Global Deviance: 810.4455   
## AIC: 822.4455   
## SBC: 843.5743   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

The above model has an impressive AIC of 822.44

In this scanerio, we choose model (using ZINBI) as the final model as its AIC was 822.44.

summary(model)

## Warning in summary.gamlss(model): summary: vcov has failed, option qr is used instead

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("ZINBI", "Zero inflated negative binomial type I")   
##   
## Call:   
## gamlss(formula = count ~ persons + child + factor(camper), family = ZINBI,   
## data = data)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: log  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.6293 0.3537 -4.606 6.59e-06 \*\*\*  
## persons 1.0605 0.1329 7.981 5.55e-14 \*\*\*  
## child -1.7828 0.2118 -8.417 3.20e-15 \*\*\*  
## factor(camper)1 0.6273 0.1936 3.240 0.00136 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: log  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.77105 0.09387 8.214 1.17e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Nu link function: logit   
## Nu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -36.04 6324.56 -0.006 0.995  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 250   
## Degrees of Freedom for the fit: 6  
## Residual Deg. of Freedom: 244   
## at cycle: 7   
##   
## Global Deviance: 810.4455   
## AIC: 822.4455   
## SBC: 843.5743   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Model equation :

‘mu’ (log link): log(mui) = -1.6293 +1.0605(persons)-1.7828(child)+0.6273[factor(camper)]

‘nu’ (logit link): log(pii / 1-pii) = -36.04

Ques2:

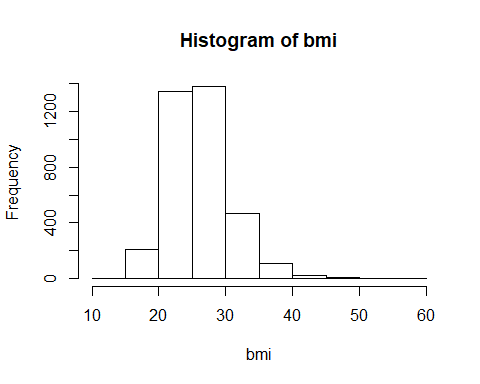
* 1. Reading the data

bmes1<-read.csv("bmes1\_assn3.csv",header=T)  
head(bmes1)

## age sex race colour burn skincol skindam bmi  
## 1 79 2 1 3 3 2 2 30.11940  
## 2 61 1 1 3 3 2 1 27.14158  
## 3 66 1 1 1 3 2 3 38.36874  
## 4 57 1 1 3 4 3 1 21.00684  
## 5 64 2 1 1 3 2 2 25.53545  
## 6 75 2 1 1 4 3 1 23.73866

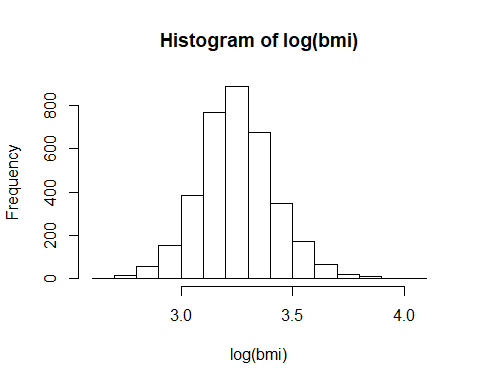
attach(bmes1)

hist(bmi,breaks = 10)



We can see a slight right skewness in the distribution for BMI, we can use log tranformation to fix that.

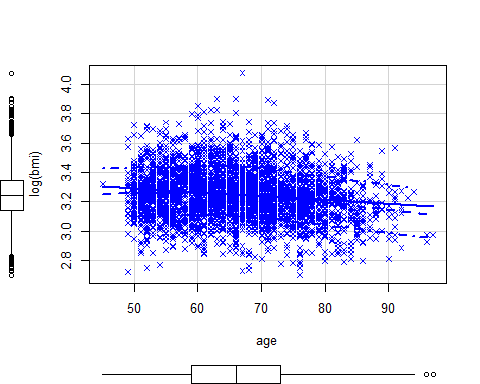
hist(log(bmi),breaks = 10)



As expected the log transformed histogram is fairly normal.

Checking the relationship b/w age and bmi

scatterplot(age,log(bmi),pch = 4)



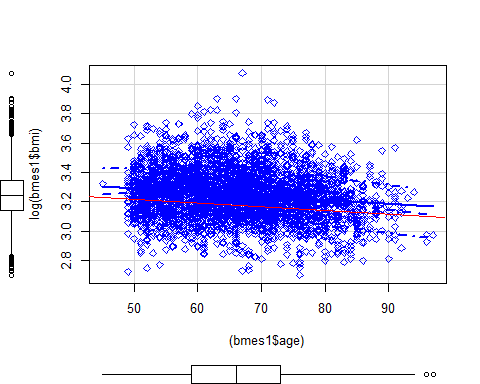
No significant linear trend is observed within the response (BMI) and age.

Building a normal model using age as a covariate:

model=lm(log(bmi)~age,data=bmes1)  
summary(model)

##   
## Call:  
## lm(formula = log(bmi) ~ age, data = bmes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56871 -0.10913 -0.00552 0.09859 0.82730   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4150093 0.0193867 176.152 <2e-16 \*\*\*  
## age -0.0025177 0.0002909 -8.656 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1667 on 3542 degrees of freedom  
## (110 observations deleted due to missingness)  
## Multiple R-squared: 0.02072, Adjusted R-squared: 0.02044   
## F-statistic: 74.93 on 1 and 3542 DF, p-value: < 2.2e-16

scatterplot((bmes1$age),log(bmes1$bmi),pch = 5)  
abline(model,col="red")



We can see that there is weak negative relationship b/w response and age. Fit of the model is very poor, which can be confirmed by looking at the multiple adjusted R-sqaured’s value i.e 0.02072

Checking the effect of ‘age’ on BMI for both sexes and all races.

count(sex)

## x freq  
## 1 1 2072  
## 2 2 1582

Both the genders are balanced so no need to combine any levels. We know from theory that if a continous variable exhitbits different values for different levels of a co-varitate, effect of interaction is said be present.

Checking for interaction within ‘age’ and ‘sex’

model<-lm(log(bmi)~(age)\*factor(sex),data=bmes1)  
summary(model)

##   
## Call:  
## lm(formula = log(bmi) ~ (age) \* factor(sex), data = bmes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.57005 -0.10989 -0.00604 0.09769 0.83123   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4306898 0.0255395 134.329 < 2e-16 \*\*\*  
## age -0.0028104 0.0003823 -7.352 2.41e-13 \*\*\*  
## factor(sex)2 -0.0385355 0.0392301 -0.982 0.326   
## age:factor(sex)2 0.0007124 0.0005890 1.209 0.227   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1666 on 3540 degrees of freedom  
## (110 observations deleted due to missingness)  
## Multiple R-squared: 0.02173, Adjusted R-squared: 0.02091   
## F-statistic: 26.22 on 3 and 3540 DF, p-value: < 2.2e-16

Since the interaction term is insignificant in the model (p-value 0.227), we can confirm that change of BMI with age is same for both the sexes.

Checking for interaction between age and race, first let’s chekc the frequency for each level.

count(bmes1$race)

## x freq  
## 1 1 3626  
## 2 2 6  
## 3 3 2  
## 4 5 3  
## 5 6 8  
## 6 7 6  
## 7 NA 3

We can discern that the categories are not well balanced, we have a large number of observations in the category 1 (whites), it makes sense to club the rest of levels into single ‘Non-White’ group or ‘others’.

Combining the levels

bmes1$race[bmes1$race==3]=2  
bmes1$race[bmes1$race==4]=2  
bmes1$race[bmes1$race==5]=2  
bmes1$race[bmes1$race==6]=2  
bmes1$race[bmes1$race==7]=2  
table(bmes1$race)

##   
## 1 2   
## 3626 25

Checking for interaction:

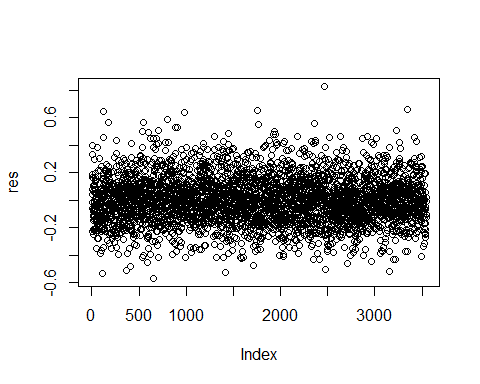
model<-lm(log(bmi)~(age)\*factor(race),data=bmes1)  
summary(model)

##   
## Call:  
## lm(formula = log(bmi) ~ (age) \* factor(race), data = bmes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56885 -0.10881 -0.00550 0.09808 0.82739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4157766 0.0194490 175.627 <2e-16 \*\*\*  
## age -0.0025305 0.0002917 -8.675 <2e-16 \*\*\*  
## factor(race)2 -0.2294215 0.2765815 -0.829 0.407   
## age:factor(race)2 0.0037188 0.0043991 0.845 0.398   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1667 on 3537 degrees of freedom  
## (113 observations deleted due to missingness)  
## Multiple R-squared: 0.02089, Adjusted R-squared: 0.02006   
## F-statistic: 25.15 on 3 and 3537 DF, p-value: 4.275e-16

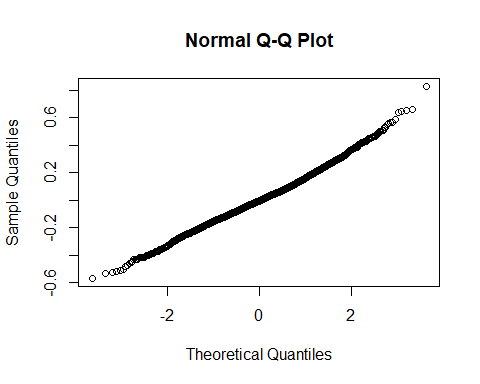
We can confirm that the interaction term for the above model is insignificant, thus we can say that change of BMI with age is same for all races.

let’s look at the Diagnostics for Model log(bmi)~age:

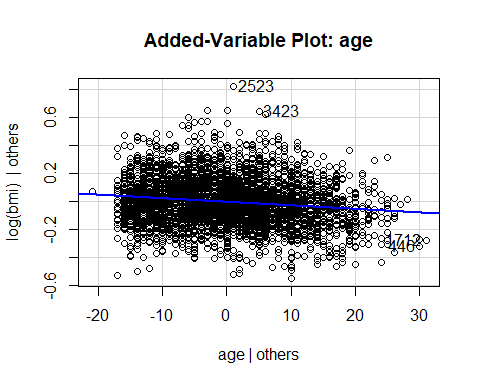
model=lm(log(bmi)~age)  
res<-residuals.lm(model)  
plot(res)



qqnorm(res)



par(mfrow=c(1,1))  
avPlot(model,variable="age")



We can see that our model assumptions for normal model are met, residuals are evenly distributed across the x=0 line with random scatter, no dicernable pattern is observed.

Looking at the qq-plot of the residuals we can confirm that the residuals are fairly normally distributed.

Let’s look at a fully fitted model :

model=lm(log(bmi)~age+factor(race)+factor(sex),data=bmes1)  
summary(model)

##   
## Call:  
## lm(formula = log(bmi) ~ age + factor(race) + factor(sex), data = bmes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56484 -0.11065 -0.00600 0.09723 0.83097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.410610 0.019603 173.982 <2e-16 \*\*\*  
## age -0.002507 0.000291 -8.613 <2e-16 \*\*\*  
## factor(race)2 0.003131 0.034147 0.092 0.927   
## factor(sex)2 0.008266 0.005648 1.463 0.143   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1666 on 3537 degrees of freedom  
## (113 observations deleted due to missingness)  
## Multiple R-squared: 0.02128, Adjusted R-squared: 0.02045   
## F-statistic: 25.63 on 3 and 3537 DF, p-value: < 2.2e-16

In the fully fitted model, we can see that race and sex are both insignificant, let’s look at a model by dropping the race

Model with age and factor(sex)

model=lm(log(bmi)~age+factor(sex),data=bmes1)  
summary(model)

##   
## Call:  
## lm(formula = log(bmi) ~ age + factor(sex), data = bmes1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56492 -0.11063 -0.00582 0.09740 0.83096   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4108585 0.0195822 174.182 <2e-16 \*\*\*  
## age -0.0025104 0.0002908 -8.631 <2e-16 \*\*\*  
## factor(sex)2 0.0084182 0.0056456 1.491 0.136   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1666 on 3541 degrees of freedom  
## (110 observations deleted due to missingness)  
## Multiple R-squared: 0.02133, Adjusted R-squared: 0.02078   
## F-statistic: 38.59 on 2 and 3541 DF, p-value: < 2.2e-16

We can see from the above output that factor(sex) is still insignificant, therefore the best explanatory model will be log(bmi)~age

Final model

log(bmi) = 3.4150093 -0.0025177(age)

Let’s try to fit generalized additive model.

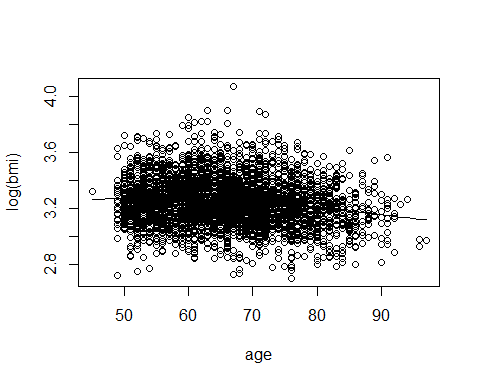
Selecting the subset of the data frame with no missing values:

collist=c('age','sex','race','colour','burn','skincol','skindam','bmi')  
newdat=bmes1[complete.cases(bmes1[collist]),collist]

In order to fit the GAM, we require subset which has no missing values to therefore above procedure was conducted to ensure that there are no missing values

GAM Model

plot(log(bmi)~age,data=newdat)  
lines(smooth.spline(log(newdat$bmi)~newdat$age,df=3))



Looking at the above graph, we can see slight curavature in the trend, the slope remains fairly constant initially, hwever we see a bigger dip after 70 years

Looking the linear term for age

model=gamlss(bmi~(age)+factor(sex)+factor(race),family = 'NO', data=newdat)

## GAMLSS-RS iteration 1: Global Deviance = 20459.42   
## GAMLSS-RS iteration 2: Global Deviance = 20459.42

summary(model)

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("NO", "Normal")   
##   
## Call: gamlss(formula = bmi ~ (age) + factor(sex) + factor(race),   
## family = "NO", data = newdat)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: identity  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.269833 0.532979 56.794 < 2e-16 \*\*\*  
## age -0.062933 0.007911 -7.955 2.4e-15 \*\*\*  
## factor(sex)2 0.019519 0.153531 0.127 0.899   
## factor(race)2 0.322675 0.922768 0.350 0.727   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: log  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.50467 0.01195 125.9 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## No. of observations in the fit: 3499   
## Degrees of Freedom for the fit: 5  
## Residual Deg. of Freedom: 3494   
## at cycle: 2   
##   
## Global Deviance: 20459.42   
## AIC: 20469.42   
## SBC: 20500.22   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

par(mfrow=c(1,1))

Model with spline(age)

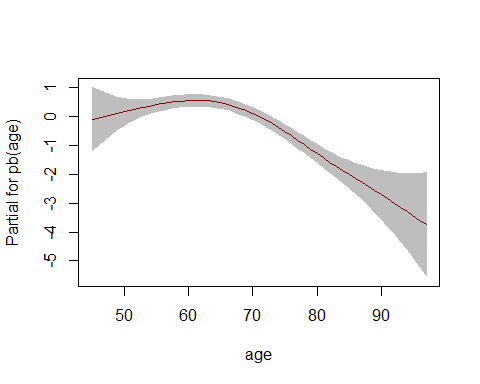
gam<-gamlss(bmi~pb(age), family = NO, data=newdat)

## GAMLSS-RS iteration 1: Global Deviance = 20423.79   
## GAMLSS-RS iteration 2: Global Deviance = 20423.79

summary(gam)

## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
## Family: c("NO", "Normal")   
##   
## Call: gamlss(formula = bmi ~ pb(age), family = NO, data = newdat)   
##   
## Fitting method: RS()   
##   
## ------------------------------------------------------------------  
## Mu link function: identity  
## Mu Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.287371 0.524317 57.765 < 2e-16 \*\*\*  
## pb(age) -0.063036 0.007866 -8.014 1.5e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## Sigma link function: log  
## Sigma Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.49958 0.01195 125.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## ------------------------------------------------------------------  
## NOTE: Additive smoothing terms exist in the formulas:   
## i) Std. Error for smoothers are for the linear effect only.   
## ii) Std. Error for the linear terms maybe are not accurate.   
## ------------------------------------------------------------------  
## No. of observations in the fit: 3499   
## Degrees of Freedom for the fit: 5.419859  
## Residual Deg. of Freedom: 3493.58   
## at cycle: 2   
##   
## Global Deviance: 20423.79   
## AIC: 20434.63   
## SBC: 20468.02   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

term.plot(gam)



We cans see that AIC is improved

Model equation:

Yi ~ N(mu, sigma) Mu = 30.29 - 0.06(age)+ S1(age)

Creating a new variable ‘bmi\_cat’:

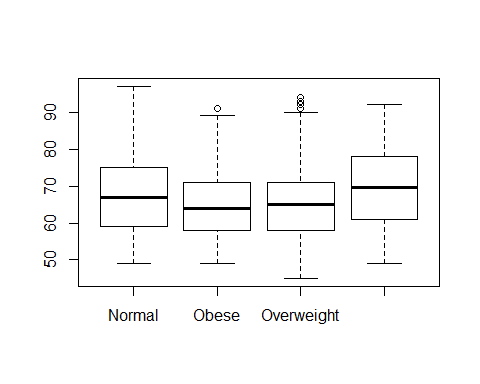
newdat$bmi\_cat[newdat$bmi<18.5]<-"Underweight"  
newdat$bmi\_cat[newdat$bmi>=18.5 & newdat$bmi<25]<-"Normal"  
newdat$bmi\_cat[newdat$bmi>=25 & newdat$bmi<30]<-"Overweight"  
newdat$bmi\_cat[newdat$bmi>=30]<-"Obese"  
  
head(newdat)

## age sex race colour burn skincol skindam bmi bmi\_cat  
## 1 79 2 1 3 3 2 2 30.11940 Obese  
## 2 61 1 1 3 3 2 1 27.14158 Overweight  
## 3 66 1 1 1 3 2 3 38.36874 Obese  
## 4 57 1 1 3 4 3 1 21.00684 Normal  
## 5 64 2 1 1 3 2 2 25.53545 Overweight  
## 6 75 2 1 1 4 3 1 23.73866 Normal

We can see that above output to confirm our new variable has indeed been added to the data frame.

Boxplot of bmi\_cat according to age :

boxplot(age~bmi\_cat,data= newdat)



table(newdat$sex,newdat$bmi\_cat)

##   
## Normal Obese Overweight Underweight  
## 1 851 383 678 61  
## 2 595 211 699 21

frequency of underweight is lesser than other categories

Chisq test of association.

chisq.test(newdat$sex,newdat$bmi\_cat)

##   
## Pearson's Chi-squared test  
##   
## data: newdat$sex and newdat$bmi\_cat  
## X-squared = 58.815, df = 3, p-value = 1.053e-12

P-value < 0.05 implies that there is a strong association between sex and bmi\_cat.

Let’s look at other co-variates

table(newdat$race,newdat$bmi\_cat)

##   
## Normal Obese Overweight Underweight  
## 1 1437 589 1368 81  
## 2 9 5 9 1

We see that there is Just 1 observaion in underweight category who is a ‘Non-white’. Most of the observation are on ‘race’=1 (White)

chisq test for bmi\_cat and race

chisq.test(newdat$bmi\_cat,newdat$race)

## Warning in chisq.test(newdat$bmi\_cat, newdat$race): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: newdat$bmi\_cat and newdat$race  
## X-squared = 0.66123, df = 3, p-value = 0.8823

There strong evidence for no association between sex and bmi\_cat.

Set the refrence level to “Normal” BMI and fitting the cummulative proportion model

library('VGAM')

## Warning: package 'VGAM' was built under R version 3.4.4

## Loading required package: stats4

##   
## Attaching package: 'VGAM'

## The following objects are masked from 'package:boot':  
##   
## logit, simplex

## The following object is masked from 'package:car':  
##   
## logit

newdat$bmi\_cat = relevel(factor(newdat$bmi\_cat),ref="Normal")  
model=vglm(ordered(bmi\_cat)~age+factor(sex),family=cumulative(parallel=TRUE),data=newdat)  
summary(model)

##   
## Call:  
## vglm(formula = ordered(bmi\_cat) ~ age + factor(sex), family = cumulative(parallel = TRUE),   
## data = newdat)  
##   
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max  
## logit(P[Y<=1]) -1.794 -0.4923 -0.4231 1.0554 1.3550  
## logit(P[Y<=2]) -1.485 -1.0759 0.4138 0.4785 1.7287  
## logit(P[Y<=3]) -8.454 0.1055 0.1291 0.2104 0.2495  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.285250 0.222564 -5.775 7.71e-09 \*\*\*  
## (Intercept):2 -0.591420 0.221712 -2.668 0.00764 \*\*   
## (Intercept):3 2.817109 0.243887 11.551 < 2e-16 \*\*\*  
## age 0.015806 0.003297 4.793 1.64e-06 \*\*\*  
## factor(sex)2 -0.248012 0.063747 -3.891 0.00010 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Number of linear predictors: 3   
##   
## Names of linear predictors:   
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])  
##   
## Residual deviance: 7808.356 on 10492 degrees of freedom  
##   
## Log-likelihood: -3904.178 on 10492 degrees of freedom  
##   
## Number of iterations: 4   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
## Exponentiated coefficients:  
## age factor(sex)2   
## 1.0159315 0.7803502

AIC(model)

## [1] 7818.356

model2= vglm(ordered(bmi\_cat)~age+factor(race),family=cumulative(parallel=TRUE),data=newdat)  
summary(model2)

##   
## Call:  
## vglm(formula = ordered(bmi\_cat) ~ age + factor(race), family = cumulative(parallel = TRUE),   
## data = newdat)  
##   
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max  
## logit(P[Y<=1]) -1.741 -0.4878 -0.4344 1.0601 1.2742  
## logit(P[Y<=2]) -1.420 -1.0915 0.4188 0.4784 1.6622  
## logit(P[Y<=3]) -7.984 0.1057 0.1319 0.2114 0.2387  
##   
## Coefficients:   
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept):1 -1.396164 0.220515 -6.331 2.43e-10 \*\*\*  
## (Intercept):2 -0.705660 0.219554 -3.214 0.00131 \*\*   
## (Intercept):3 2.693901 0.241615 11.150 < 2e-16 \*\*\*  
## age 0.015903 0.003295 4.827 1.39e-06 \*\*\*  
## factor(race)2 -0.054985 0.382165 -0.144 0.88560   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Number of linear predictors: 3   
##   
## Names of linear predictors:   
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])  
##   
## Residual deviance: 7823.444 on 10492 degrees of freedom  
##   
## Log-likelihood: -3911.722 on 10492 degrees of freedom  
##   
## Number of iterations: 4   
##   
## No Hauck-Donner effect found in any of the estimates  
##   
## Exponentiated coefficients:  
## age factor(race)2   
## 1.0160303 0.9464997

AIC(model2)

## [1] 7833.444

Looking the above outputs we can see that model number 1, i.e (bmi\_cat)~age+factor(sex) is the best option