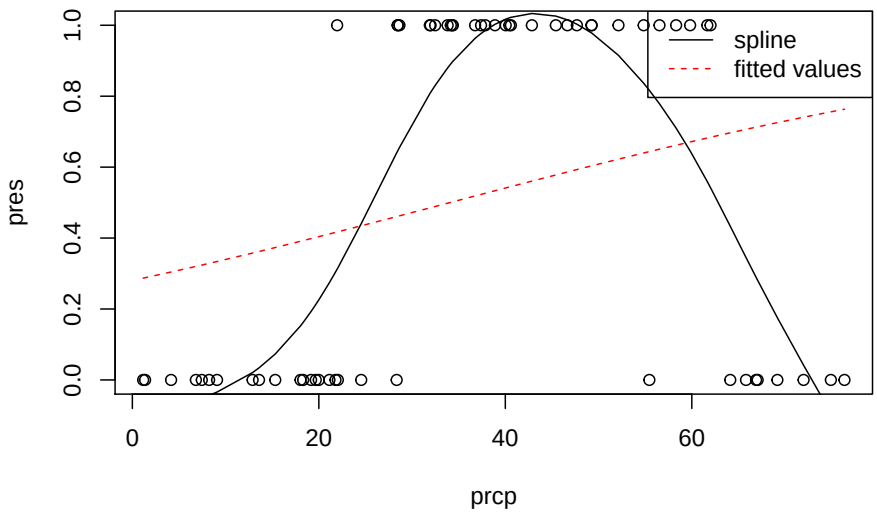
# Assignment 3

Masroor

23/10/2020

1 a)

```
Call:
glm(formula = pres ~ prcp, family = binomial, data = data.ecology)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6975  -1.0146   0.8729   1.1113   1.3223

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.94198    0.56735  -1.660   0.0969 .
prcp         0.02766    0.01384   1.999   0.0456 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.111  on 59  degrees of freedom
Residual deviance: 78.793  on 58  degrees of freedom
AIC: 82.793

Number of Fisher Scoring iterations: 4
```




Fitted curve doesn't quite get the same curvature that we see in the spline; this means that the model is not reproducing the effect of precipitation on presence of species. So the logistic regression is not suitable.

1 (b)

```
Call:
glm(formula = pres ~ prcp + I(prcp^2), family = binomial, data = data.ecology)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4502  -0.1485   0.1263   0.2371   2.0234

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.294766   6.263487  -3.559 0.000372 ***
prcp          1.238396   0.335703   3.689 0.000225 ***
I(prcp^2)    -0.014123   0.003829  -3.689 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.111  on 59  degrees of freedom
Residual deviance: 21.921  on 57  degrees of freedom
AIC: 27.921

Number of Fisher Scoring iterations: 7
```

Quadratic: $ln(\frac{pi}{1-pi}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$

$ln(\frac{pi}{1-pi}) = -22.295 + 1.238 * x_i - 0.014123 * x_i^2$

```
Call:
glm(formula = pres ~ prcp + I(prcp^2) + I(prcp^3), family = binomial,
    data = data.ecology)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4980  -0.1518   0.1229   0.2438   2.0055

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.032e+01  1.599e+01  -1.271    0.204
prcp         1.074e+00  1.278e+00   0.840    0.401
I(prcp^2)   -9.908e-03  3.200e-02  -0.310    0.757
I(prcp^3)   -3.333e-05  2.525e-04  -0.132    0.895

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.111  on 59  degrees of freedom
Residual deviance: 21.904  on 56  degrees of freedom
AIC: 29.904

Number of Fisher Scoring iterations: 8
```

Cubic:

$ln(\frac{pi}{1-pi}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$

$ln(\frac{pi}{1-pi}) = -2.032e+01 + 1.074e+00 * x_i - 9.908e-03 * x_i^2 - 3.333e-05 * x_i^3$

1 (c)

```
            df      AIC
model1       2 82.79312
model.quad   3 27.92140
model.cubic  4 29.90372

            df      BIC
model1       2 86.98180
model.quad   3 34.20443
model.cubic  4 38.28110
```

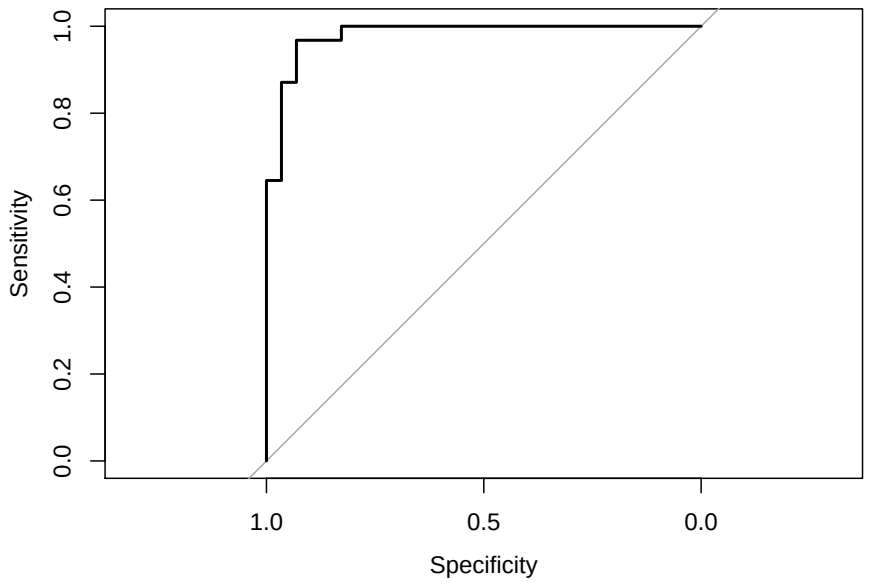
The AIC and BIC values are very small for quadratic model compared to linear and cubic model. So I would suggest to use quadratic model.

1 (d)

```
    FALSE TRUE
  0    27    2
  1     1   30
```

This means that, for the ith observation, if the fitted probability pi is less than 0.5, then a prediction of "absence" is made; if pi is greater than equal 0.5, then a prediction of "presence" is made. In the classification table, we then analyse the observed presence versus the predictions. We have 27 + 30 = 57 cases out of 60, or 95%, that were correctly predicted. Sensitivity = P(predicting presence of species|presence of species) =30/31= 96.7% Specificity = P(predicting absence of species|absence of species) =27/29=93.1%

```
Warning: package 'pROC' was built under R version 4.0.3

Type 'citation("pROC")' for a citation.


Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```




```
Call:
roc.default(response = data.ecology$pres, predictor = probs.quad,     plot = TRUE)

Data: probs.quad in 29 controls (data.ecology$pres 0) < 31 cases (data.ecology$pres 1).
Area under the curve: 0.98
```

Area under the curve is 0.98. So the model performs well in terms of predictive ability. 1 e)

```
[1] 0.01767714
```

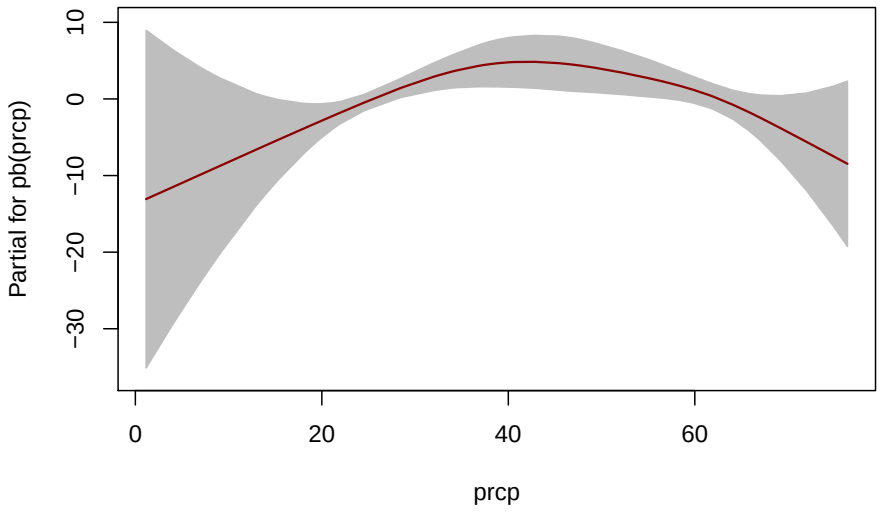
UnderH0, this will be distributed as $\chi 24$. The p-value is therefore

```
[1] 0.8942286
```

We do not reject the term that the prcp is linear as p value is greater than 0.5. 1 f)

```
Loading required package: nlme

This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.

Warning: package 'gamlss' was built under R version 4.0.3

Loading required package: splines

Loading required package: gamlss.data

Warning: package 'gamlss.data' was built under R version 4.0.3


Attaching package: 'gamlss.data'

The following object is masked from 'package:datasets':

    sleep

Loading required package: gamlss.dist

Warning: package 'gamlss.dist' was built under R version 4.0.3

Loading required package: MASS

Loading required package: parallel

 **********   GAMLSS Version 5.2-0  **********

For more on GAMLSS look at https://www.gamlss.com/

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

*********************************************************************
Family:  c("BI", "Binomial")

Call:  gamlss(formula = pres ~ pb(prcp), family = BI, data = data.ecology,
    trace = F)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.78774    1.40066  -1.276    0.207
pb(prcp)     0.04767    0.03025   1.576    0.121
```

```
-----------------------------------------------------------------
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:  60
Degrees of Freedom for the fit:  4.222611
      Residual Deg. of Freedom:  55.77739
                       at cycle:  2

Global Deviance:     21.4781
            AIC:     29.92332
            SBC:     38.76693
*****************************************************************
```




```
[1] 29.92332
```

```
[1] 27.9214
```

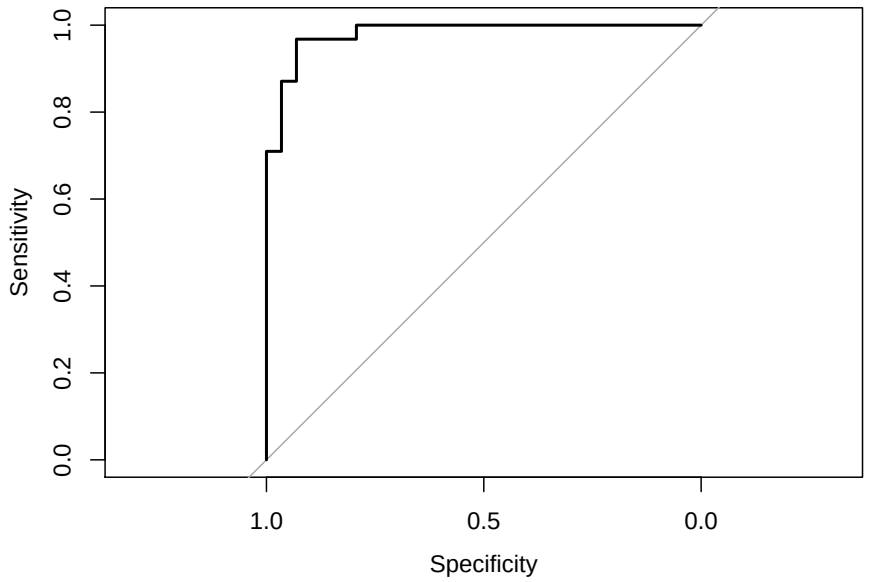
```
[1] 38.76693
```

```
[1] 34.20443
```

The AIC and BIC is less for quadratic model than generalized additive model,but we can see from the classification table, specificity, sensitivity and roc that the performance of additive model is same as quadratic model as shown below.

```
   FALSE TRUE
0     27    2
1      1   30
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```




```
Call:
roc.default(response = data.ecology$pres, predictor = probs.gam1,     plot = TRUE)

Data: probs.gam1 in 29 controls (data.ecology$pres 0) < 31 cases (data.ecology$pres 1).
Area under the curve: 0.9811
```

The area under the curve is 0.98. $27 + 30 = 57$ cases out of 60, or 95%, that were correctly predicted.

Sensitivity = P(predicting presence of species|presence of species) =30/31= 96.7% Specificity = P(predicting absence of species|absence of species) =27/29=93.1% Moreover, if we compare the term plot and the summary table, we see that

```
Call:
glm(formula = pres ~ prcp + I(prcp^2), family = binomial, data = data.ecology)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4502  -0.1485   0.1263   0.2371   2.0234

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.294766   6.263487  -3.559 0.000372 ***
prcp          1.238396   0.335703   3.689 0.000225 ***
I(prcp^2)    -0.014123   0.003829  -3.689 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.111  on 59  degrees of freedom
Residual deviance: 21.921  on 57  degrees of freedom
AIC: 27.921

Number of Fisher Scoring iterations: 7
```

Neither the linear part nor the smooth term of prcp looks significant in the term plot of generalized additive model. On the other hand, we can see that the prcp looks highly significant in quadratic model if we look at the summary table. So quadratic model performs better than generalized additive model.

2 a)

```
  AIScode Freq
1       1   60
2       2  148
3       3   72
4       4   30
5       5   16
6       6   14
7    <NA>   12
```

There are 12 missing values in AIScode. As 5 and 6 are less than 20, we will join these two AIScode values.

```
  AIScode Freq
1       1   60
2       2  148
3       3   72
4       4   30
5       6   30
```

2 b)

```
Loading required package: stats4


Attaching package: 'VGAM'

The following object is masked from 'package:mgcv':

    s


Call:
vglm(formula = ordered(AIScode) ~ factor(dp) + weight, family = cumulative(parallel = TRUE),
    data = crash)

Pearson residuals:
                       Min      1Q  Median      3Q   Max
logitlink(P[Y<=1]) -0.8734 -0.5862 -0.2924 -0.1376 3.568
```

```
logitlink(P[Y<=2]) -2.5892 -0.5848  0.3131  0.7642 1.298
logitlink(P[Y<=3]) -4.2447  0.1546  0.2178  0.3250 1.030
logitlink(P[Y<=4]) -4.8403  0.1192  0.1530  0.2211 1.566

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept):1    -0.7838649  0.4898795  -1.600  0.10957
(Intercept):2     1.3287791  0.4910795   2.706  0.00681 **
(Intercept):3     2.4911537  0.5055276   4.928 8.31e-07 ***
(Intercept):4     3.3134829  0.5250469   6.311 2.78e-10 ***
factor(dp)Passen  0.8685650  0.2045149   4.247 2.17e-05 ***
weight           -0.0004361  0.0001597  -2.730  0.00633 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
logitlink(P[Y<=3]), logitlink(P[Y<=4])

Residual deviance: 943.1875 on 1354 degrees of freedom

Log-likelihood: -471.5937 on 1354 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
factor(dp)Passen           weight
        2.383488         0.999564
```

The estimated model equations are j= 1, $\ln\left(\frac{\gamma 1}{1-\gamma 1_i}\right) = -0.7838649 + 0.8685650xi1 - 0.0004361xi2$

Interpretation of parameters

dp:

$\beta_1$

The ratio of the cumulative odds of a driver versus passenger,of the same weight is

$\frac{\exp[\alpha_j+\beta_1+\beta_2 xi2]}{\exp[\alpha_j+\beta_2 xi2]} = \exp\beta_1$

$\exp(\beta_1) = \exp 0.8685650 = 2.383$

This means that, for any AIS code groupj= 1,2,3,4, the odds of being in group less than equal j for driver is 2.383 times the odds of being in group less than equal j for passengers,for subjects of the same weight. This is an increase in the odds of 138.3%. For example, the odds of having no or AIS code for driver is 2.383 times that of the odds for passengers, for subjects of the same weight. This confirms that drivers have better AIS code than passengers.

Weight:

$\beta_2$ The ratio of the cumulative odds of AIS,of a person of weight x+1,compared with a person of weight x, of the same dp, is

$\frac{\exp[\alpha_j+\beta_1 xi1+\beta_2(x+1)]}{\exp[\alpha_j+\beta_1 xi1+\beta_2 x]} = \exp\beta_2$

$\exp \beta_2 = \exp -0.0004361 = 0.99956$ For any ais code groupj=1,2,3 and 4, the odds of being in group less than equal j for a person of weight x+ 1 is 0.99956 times the odds of being in group less than equal j for a person of weight x,for subjects of the same driver or passenger. In other words, for every increasing year of weight, there is a 0.044% reduction in the odds of being in AIS group less than equal j,for j= 1,2,3,4. This is indicating that ais code is worsening with increasing weight.

2 c)

```
Call:
vglm(formula = AIScode ~ factor(dp) + weight, family = multinomial(refLevel = 1),
    data = crash)

Pearson residuals:
                      Min      1Q  Median       3Q   Max
log(mu[,2]/mu[,1]) -1.788 -0.6926 -0.4412  1.01738 1.618
log(mu[,3]/mu[,1]) -1.636 -0.4411 -0.3184 -0.16392 4.230
log(mu[,4]/mu[,1]) -1.566 -0.3290 -0.1473 -0.06002 7.886
log(mu[,5]/mu[,1]) -1.328 -0.2384 -0.1535 -0.10225 5.358

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept):1       2.1204281  0.8024382   2.642  0.00823 **
(Intercept):2       1.5774830  0.9275440   1.701  0.08900 .
(Intercept):3      -2.4722408  1.1429361  -2.163  0.03054 *
(Intercept):4      -2.5128223  1.1162273  -2.251  0.02437 *
factor(dp)Passen:1 -0.5010468  0.3204467  -1.564  0.11791
factor(dp)Passen:2 -1.5123195  0.3750566  -4.032 5.52e-05 ***
factor(dp)Passen:3 -2.1005086  0.5387538  -3.899 9.67e-05 ***
factor(dp)Passen:4 -0.5739474  0.4609891  -1.245  0.21312
weight:1           -0.0003182  0.0002633  -1.208  0.22695
weight:2           -0.0002296  0.0003079  -0.746  0.45583
weight:3            0.0008602  0.0003553   2.421  0.01546 *
weight:4            0.0007064  0.0003442   2.053  0.04010 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]),
log(mu[,4]/mu[,1]), log(mu[,5]/mu[,1])

Residual deviance: 916.8306 on 1348 degrees of freedom

Log-likelihood: -458.4153 on 1348 degrees of freedom

Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates


Reference group is level  1  of the response
```

Outcome 2: The model for AIS code=2 relative to AIS code=1 is:

$\ln [\frac{\pi_2 i}{\pi_1 i}] = 2.1204281 - 0.5010468 xi1 - 0.0003182 xi2$

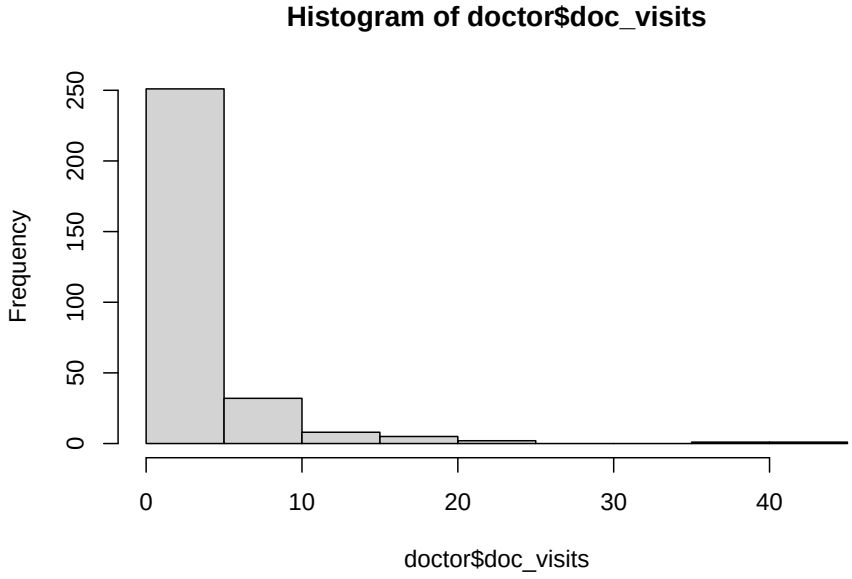
Interpretation:

outcome 2:

$\exp(-0.5010468) = 0.606$ The effect of a passenger seat as opposed to driver's seat on the odds of having a head injury with AIS code 2 relative to head injury with AIS code 1 in terms of a car crash is 0.606, i.e. the odds of having a head injury resulting from car crash decreases by 39.4% when sitting in the passenger seat.

$\exp(-0.0003182) = 0.9997$

The effect of a weight x+1 as opposed to weight x on the odds of having a head injury with AIS code 2 relative to head injury with AIS code 1 in terms of a car crash is 0.9997, i.e. a heavier person has 0.03% less chance in facing an accident than a lightweight person. 3 a)




There is a very high zero frequency and the Poisson and negative binomial distributions are unlikely to provide a good fit to the data. So zero inflated models should be considered as there are lots of zeroes in the model.

3 b) ZINB model is:

```
******************************************************************
Family:  c("ZINBI", "Zero inflated negative binomial type I")

Call:  gamlss(formula = doc_visits ~ factor(age50) + health +
    schooling, family = ZINBI, data = doctor, trace = F)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.584354   0.424776   6.084 3.64e-09 ***
factor(age50)TRUE  0.405424   0.156234   2.595  0.00993 **
health            -0.254116   0.035348  -7.189 5.42e-12 ***
schooling          0.002509   0.029351   0.085  0.93194
---
```

```
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) -0.09708    0.25123  -0.386    0.699

------------------------------------------------------------------
Nu link function:  logit
Nu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.0537     0.7203  -2.851  0.00467 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

------------------------------------------------------------------
No. of observations in the fit:  300
Degrees of Freedom for the fit:  6
      Residual Deg. of Freedom:  294
                      at cycle:  11

Global Deviance:     1249.986
            AIC:     1261.986
            SBC:     1284.209
******************************************************************
```

We see that the schooling is insignificant as it is greater than 0.05. So we will remove schooling from the model So our next model for zinbi is:

```
******************************************************************
Family:  c("ZINBI", "Zero inflated negative binomial type I")

Call:  gamlss(formula = doc_visits ~ factor(age50) + health,
    family = ZINBI, data = doctor, trace = F)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.61397    0.24598  10.627  < 2e-16 ***
factor(age50)TRUE  0.40527    0.15624   2.594  0.00996 **
health            -0.25410    0.03534  -7.190 5.36e-12 ***
---
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) -0.09695    0.25098  -0.386      0.7
```

```
-------------------------------------------------------------------
Nu link function:  logit
Nu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.0537     0.7193  -2.855  0.00461 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-------------------------------------------------------------------
No. of observations in the fit:  300
Degrees of Freedom for the fit:  5
      Residual Deg. of Freedom:  295
                      at cycle:  11

Global Deviance:     1249.993
            AIC:     1259.993
            SBC:     1278.512
*******************************************************************
```

We see that all the variables are significant in the model. So this is our final model For model selection, now, we will compare the two models using AIC and BIC.

```
             df      AIC
doctor.zinb1  5 1259.993
doctor.zinb   6 1261.986

             df      BIC
doctor.zinb   6 1284.209
doctor.zinb1  5 1278.512
```

ZINB second version appears to have less AIC and BIC compared to ZINB first version. So we will choose ZINB second version. So our final model equation is:

$Yi \sim ZINB(\mu_i, \sigma, \pi_i)$

The fitted model is independently for i= 1,...,n

$\log(\mu_i) = 2.61397 + 0.40527 * xi_1 - 0.25410 * xi_2$ $\sigma = \exp -0.09695 = 0.908$ $\log(\frac{\pi_i 1-\pi_i}{1-\pi_i}) = -2.0537$

Health:

$\exp(-0.25410) = 0.776$

Age50:

$\exp(0.40527) = 1.4997$

interpretation:

$\beta_1$

The effect of age50 on the mean number of doctor visits is 1.4997. Age50 increases the expected number of doctor visits by 49.97%, amongst doctors that have visited.

$\beta_2$

The effect of health on the mean number of doctor visits is 0.776. Health decreases the expected number of doctor visits by 22.4%, amongst doctors that have visited.

```r
#1a
load("data_ecology.RData")

#1a
model1 <- glm(pres ~ prcp, family=binomial, data=data.ecology)
summary(model1)
#1a
plot(pres~prcp,data=data.ecology)
lines(smooth.spline(data.ecology$pres~data.ecology$prcp,df=5))
lines(fitted(model1)[order(data.ecology$prcp)]~sort(data.ecology$prcp),lty=2,col="red")
legend("topright",legend=c("spline","fitted values"),lty=1:2,col=c("black","red"))
#1b
model2 <- glm(pres ~ prcp + I(prcp^2), family=binomial, data=data.ecology)
summary(model2)
#1b
model3 <- glm(pres ~ prcp + I(prcp^2) + I(prcp^3), family=binomial, data=data.ecology)
summary(model3)
#1c
model.quad <- glm(pres ~ poly(prcp,2), family=binomial, data=data.ecology)
model.cubic <- glm(pres ~ poly(prcp,3), family=binomial, data=data.ecology)
#1c
AIC(model1, model.quad, model.cubic)
#1c
BIC(model1, model.quad, model.cubic)
#1d
probs.quad <- fitted(model.quad)
table(data.ecology$pres, probs.quad>=0.5)
#1d
#install.packages("pROC")
library(pROC)
roc(data.ecology$pres,probs.quad, plot=TRUE)
#1e
LR <- deviance(model2)-deviance(model3)
LR
#1e
df<-3-2
1-pchisq(LR, df)
#1f
#install.packages("mgcv")
library(mgcv)
#install.packages("gamlss")
library(gamlss)
gam1 <- gamlss(pres ~ pb(prcp), family=BI, data=data.ecology, trace=F)
summary(gam1)
term.plot(gam1,pages=1)
AIC(gam1)
AIC(model2)
BIC(gam1)
BIC(model2)
probs.gam1<- fitted(gam1)
table(data.ecology$pres, probs.gam1>=0.5)
library(pROC)
roc(data.ecology$pres,probs.gam1, plot=TRUE)
```

```r
summary(model2)
#2a
crash=read.csv("crash.csv",header=TRUE)

#2a
crash$AIScode=crash$head_ic
crash$AIScode[crash$AIScode >=135 & crash$AIScode <=519] = 1
crash$AIScode[crash$AIScode >=520 & crash$AIScode <=899] = 2
crash$AIScode[crash$AIScode >=900 & crash$AIScode <=1254] = 3
crash$AIScode[crash$AIScode >=1255 & crash$AIScode <=1574] = 4
crash$AIScode[crash$AIScode >=1575 & crash$AIScode <=1859] = 5
crash$AIScode[crash$AIScode >1860] = 6
w = table(crash$AIScode,useNA="ifany")
t = as.data.frame(w)
names(t)[1] = 'AIScode'
t
#2a
crash$AIScode[crash$AIScode==5] <- 6
w = table(crash$AIScode)
t = as.data.frame(w)
names(t)[1] = 'AIScode'
t
#2b
#install.packages("VGAM")
library(VGAM)
## Loading required package:
#install.packages("stats4")
library(stats4)## Loading required package:
#install.packages("splines")
library(splines)
driver <- vglm(ordered(AIScode)~factor(dp) + weight,family=cumulative(parallel=TRUE),data=crash)
summary(driver)
#2c
#install.packages("VGAM")
library(VGAM)
## Loading required package:
#install.packages("stats4")
library(stats4)## Loading required package:
#install.packages("splines")
library(splines)

crash1 <- vglm(AIScode~factor(dp) + weight,family=multinomial(refLevel=1),data=crash)
summary(crash1)
#3a
doctor=read.csv("doc_visits.csv",header=TRUE)

hist(doctor$doc_visits)
summary(doctor.zinb <- gamlss(doc_visits ~ factor(age50)+health+schooling,family=ZINBI, data=doctor,tra
summary(doctor.zinb1 <- gamlss(doc_visits ~ factor(age50)+health,family=ZINBI, data=doctor,trace=F))
AIC(doctor.zinb,doctor.zinb1)
BIC(doctor.zinb,doctor.zinb1)
```