Modelos Lineales 2

Alumno:
Huertas Quispe, Anthony Enrique ***Cod:*** *20173728*

Semestre: *2017-II*

Tema: PC 4

Prof. Enver Tarazona

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Pregunta 3

El 28 de enero de 1986, el transportador espacial challenger tuvo una falla catastrófica debido a la quema de una junta tórica(O-rings) en una articulación de uno de los cohetes propulsores de combustible sólido. Este fue el 25avo vuelo de un transbordador. Los datos de los vuelos anteriores se encuentran en el archivo orings.csv que contiene una variable respuesta falla (1= si hubo un falla, 0 en caso contrario) y una covariable temperatura (en grados Fahrenheit). Estos datos recolectan información sobre la falla de cada una de las juntas tóricas (o-rings) diseñadas para prevenir el escape de combustible muy caliente producido durante la ignición del transbordador Challenger. Había seis de estos anillos en cada lanzamiento y se han considerado los 23 vuelos previos que tuvo el transportador totalizando 138 observaciones (en un vuelo el dato de la temperatura no se registró). El objetivo del estudio es evaluar si la falla de una junta tórica están relacionadas con la temperatura.

Listing 1: Lectura de Base de datos.

```
library(R2WinBUGS)
library(coda)
library(mcmcplots)
library(pROC)
library(Epi)
library("BRugs")


setwd("C:/Users/Anthony/Documents/PUCP/2017 - 2/MODELOS LINEALES II/Datos")
bugs.dir <- "C:/Users/Anthony/Downloads/WinBUGS14/"

data <- read.csv("orings.csv")
data <- as.data.frame(data)

attach(data)
head(data)
```

```
  temperatura falla
1          53     1
2          53     1
3          53     1
4          53     1
5          53     1
6          53     0
```

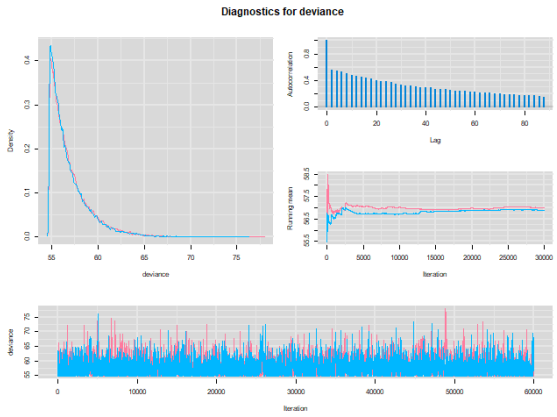
a) A partir de los datos realice un análisis de regresión completo bajo inferencia bayesiana usando BUGS para explicar la variable falla en función de la temperatura teniendo en cuenta la selección de un MLG para una variable respuesta con distribución de Bernoulli considerando los siguientes enlaces:

- **Logit:**

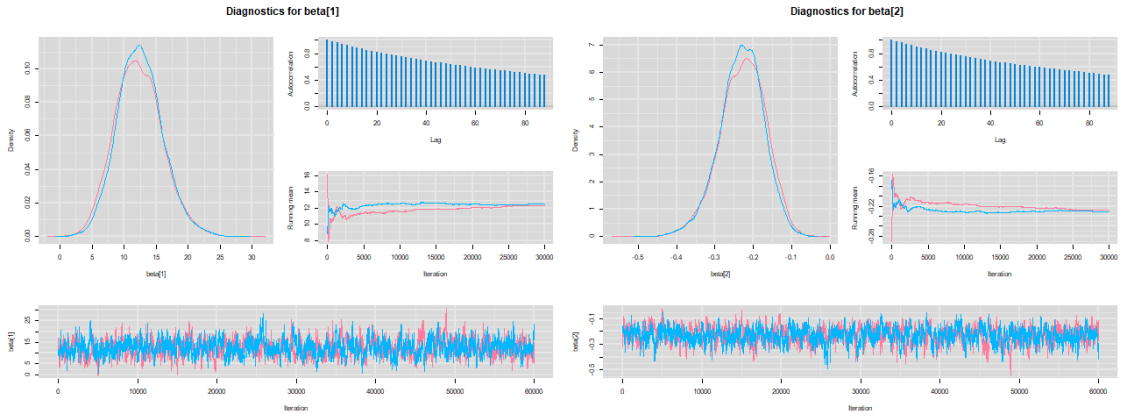
Listing 2: Logit.

```
modelo <- function(){
  for (i in 1:n) {
    y[i] ~ dbern(mu[i])
    logit(mu[i]) <- eta[i]
    eta[i] <- inprod(beta[],X[i,])
  }
```

```
  for(j in 1:p){
    beta[j] ~ dnorm(0,0.000001)
  }
}

write.model(modelo, "mod1.bug")
file.show("mod1.bug")

X <- model.matrix(~ temperatura)
parametros <- c("beta")
iniciales <-function(){list(beta=rep(rnorm(1),ncol(X)) )}

datos = list(X = X,y = falla,n=length(falla),p=ncol(X))


modelo1 <- bugs(data = datos,inits = iniciales,
                parameters.to.save = parametros,
                model.file="MB1.bug",
                n.chains=2, n.iter=100000,
                n.burnin=40000,n.thin=2,
                bugs.directory = bugs.dir,
                clearWD=TRUE, debug=FALSE)

print(modelo1,4)
mcmcplot(modelo1)
```



Plots for beta



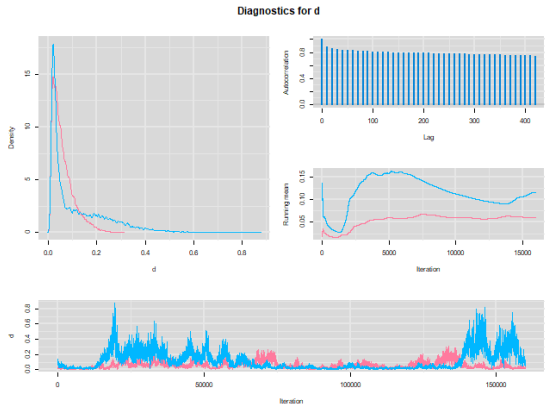
```
             mean     sd    2.5%     25%     50%     75%   97.5%   Rhat n.eff
beta[1]  12.3823 3.6500   5.626  9.8678 12.2500 14.7300 20.0103 1.0031   850
beta[2]  -0.2291 0.0591  -0.355 -0.2666 -0.2263 -0.1881 -0.1216 1.0029   910
deviance 56.9505 2.2020  54.820 55.3900 56.2700 57.7900 62.9300 1.0019  1600

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 2.1 and DIC = 59.0
DIC is an estimate of expected predictive error (lower deviance is better).
```

- **Power Logit**

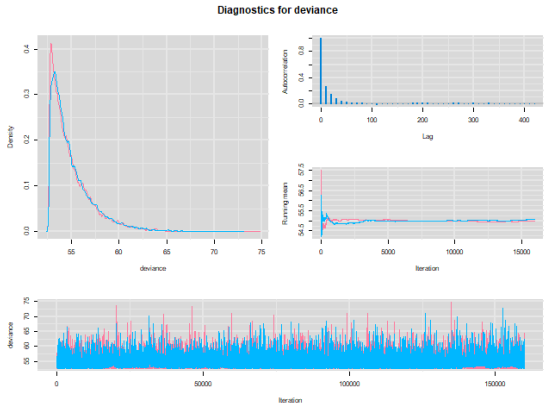
Listing 3: Logit.

```
modelo <- function(){
  for (i in 1:n) {
    y[i] ~ dbern(mu[i])
    p[i] <- exp(eta[i])/(1+exp(eta[i]))
    mu[i] <- pow(p[i],d)
    eta[i] <- inprod(beta[],X[i,])
  }
  d~dgamma(1,0.001)
  for (j in 1:k){
    beta[j] ~ dnorm(0.0,0.00001)
  }
}
write.model(modelo,"mod2.bug")
file.show("mod2.bug")

parametros <- c("beta","d")
iniciales <-function(){list(beta=c(300,-5),d=0.05)}

modelo2 <- bugs(data = datos,inits = iniciales,
  parameters.to.save = parametros, model.file="mod2.bug", n.chains=2,
  n.iter=200000, n.burnin=40000,n.thin=10, bugs.directory = bugs.dir,
  clearWD=TRUE, debug=F)
print(modelo2,4)
mcmcplot(modelo2)
```

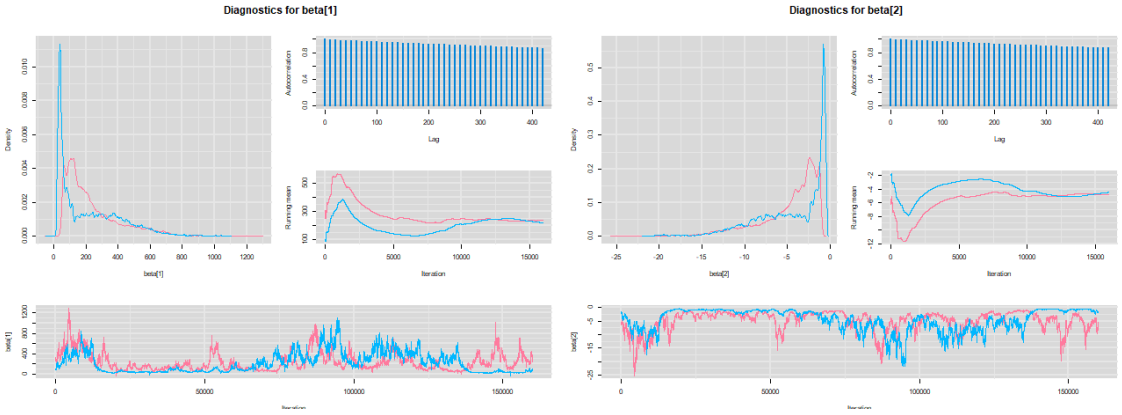
Plots for d



Plots for deviance



Plots for beta



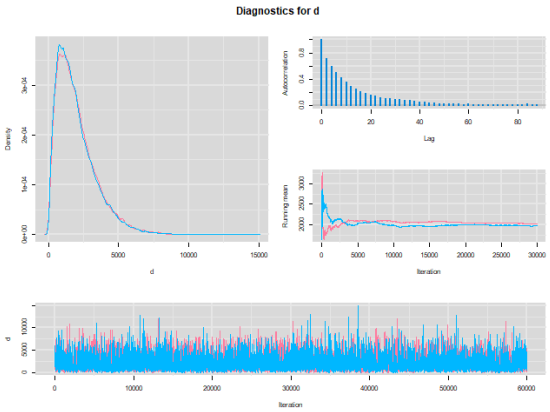

```
                mean       sd     2.5%      25%      50%      75%    97.5%   Rhat n.eff
beta[1]   230.4574 189.9538  27.9297 80.6100 173.5500 336.4000 691.4025 1.0099   410
beta[2]    -4.6810   3.8910 -14.0700 -6.8132  -3.5100  -1.6320  -0.5636 1.0096   450
d           0.0875   0.0955   0.0100  0.0254   0.0503   0.1118   0.3610 1.1079    34
deviance   55.0235   2.3087  52.7400 53.3800  54.3200  55.9500  61.2500 1.0016  2300

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = -41.7 and DIC = 13.3
DIC is an estimate of expected predictive error (lower deviance is better).
```

- **Power Logit recíproco:**

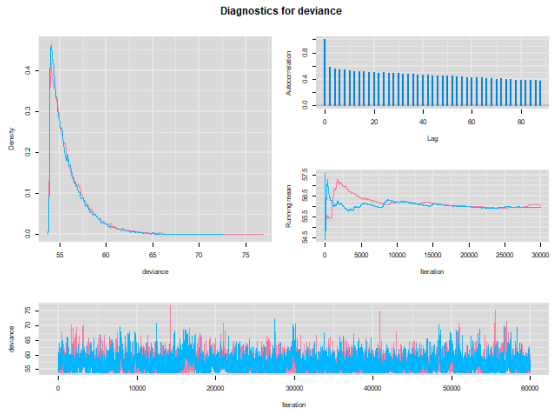
Listing 4: Power Logit recíproco.

```
modelo <- function(){
  for (i in 1:n) {
    y[i] ~ dbern(mu[i])
    p[i] <- exp(-eta[i])/(1+exp(-eta[i]))
    mu[i] <- 1-pow(p[i],d)
    eta[i] <- inprod(beta[],X[i,])
  }
  for(j in 1:k){
    beta[j]~dnorm(0,0.000001)
  }
  d~dgamma(2,0.001)
}

write.model(modelo, "mod3.bug")
file.show("mod3.bug")

parametros <- c("beta","d")
iniciales <- function(){list(beta=rep(rnorm(1),ncol(X)),d=1)}

modelo3 <- bugs(data = datos,inits = iniciales,
  parameters.to.save = parametros,model.file="mod3.bug",n.chains=2,
  n.iter=100000, n.burnin=40000,n.thin=2, bugs.directory = bugs.dir,
  clearWD=TRUE, debug=F)

print(modelo3,4)
mcmcplot(modelo3)
```

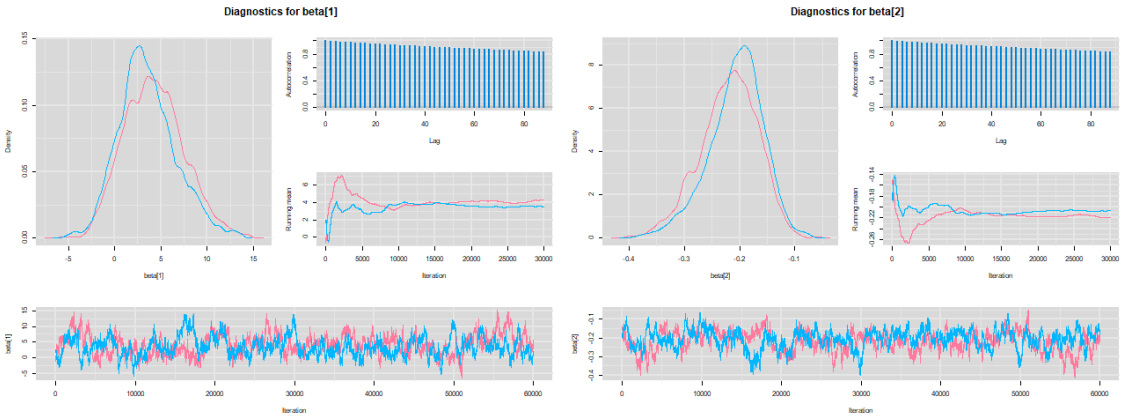
Plots for d



Plots for deviance



Plots for beta




```
                mean        sd     2.5%      25%      50%      75%    97.5%   Rhat
beta[1]       3.8997    3.2058  -1.7690   1.6580   3.6700   5.9300  10.7600 1.0239
beta[2]      -0.2132    0.0518  -0.3277  -0.2449  -0.2084  -0.1771  -0.1238 1.0277
d          1997.0615 1414.5184 248.2975 953.7000 1672.4999 2694.0000 5545.0250 1.0013
deviance     56.0197    2.1374  53.9300  54.4800  55.3500  56.8700  61.8900 1.0024
           n.eff
beta[1]       72
beta[2]       63
d           4200
deviance    1100

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = Dbar-Dhat)
pD = 1.8 and DIC = 57.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

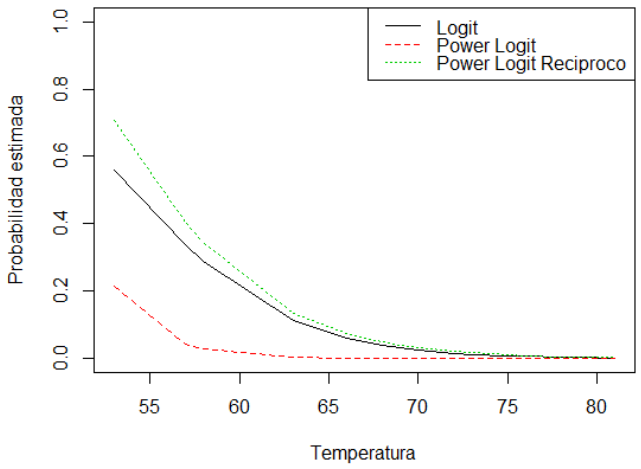
Listing 5: Indicadores (DIC, EAIC, EBIC).

```
val.DIC=c(modelo1$DIC,modelo1$DIC,modelo3$DIC)
val.EAIC=c(modelo1$DIC-modelo1$pD+4,modelo2$DIC-modelo2$pD+6,
  modelo3$DIC-modelo3$pD+6)
val.EBIC=c(modelo1$DIC-modelo1$pD+2*log(138),modelo2$DIC-modelo2$pD+3*log
    (138),modelo3$DIC-modelo3$pD+3*log(138))

indicadores=rbind(val.DIC,val.EAIC,val.EBIC)
colnames(indicadores)<-c("Logit","Power Logit"," Power Logit Reciproco")
indicadores
```

```
                Logit Power Logit  Power Logit Reciproco
val.DIC  59.04400    13.33600                57.78900
val.EAIC 60.95000    61.02300                62.02000
val.EBIC 66.80451    69.80476                70.80176
```

Listing 6: Probabilidad estimada.

```r
#Logit
b1.mod1 = modelo1$sims.matrix[,1]
b2.mod1 = modelo1$sims.matrix[,2]
n1 = nrow(modelo1$sims.matrix)
eta.mod1 = mean(b1.mod1) + mean(b2.mod1)*temperatura
mu.mod1 = 1/(1+exp(-eta.mod1))

#Power Logit
b1.mod2 = modelo2$sims.matrix[,1]
b2.mod2 = modelo2$sims.matrix[,2]
d.mod2 = modelo2$sims.matrix[,3]
n2 = nrow(modelo2$sims.matrix)
eta.mod2 = mean(b1.mod2) + mean(b2.mod2)*temperatura
mu.mod2 = (1/(1+exp(-eta.mod2)))^mean(d.mod2)

#Power Logit Reciproco
b1.mod3 = modelo3$sims.matrix[,1]
b2.mod3 = modelo3$sims.matrix[,2]
d.mod3 = modelo3$sims.matrix[,3]
n3 = nrow(modelo3$sims.matrix)
eta.mod3 = mean(b1.mod3) + mean(b2.mod3)*temperatura
mu.mod3 = 1-(1/(1+exp(eta.mod3)))^(mean(d.mod3))

#Grafica
plot(temperatura,mu.mod1,ylim=c(0,1),col=1,type="l", ylab="Probabilidad
    estimada", xlab="Temperatura")
lines(temperatura,mu.mod2,lty=2,col=2)
lines(temperatura,mu.mod3,lty=3,col=3)
legend("topright",legend=c("Logit","Power Logit","Power Logit Reciproco"),col
    =c(1:3),lty=c(1:3))
```



b) Realice un análisis comparativo del desempeño de los modelos en predecir de que ocurra una falla en un anillo de goma. Considerando un punto de corte de 0.5, presente la matriz de confusión para cada modelo y luego presente una tabla comparativa considerando como criterios de comparación el error de clasificación, la sensitividad, especificidad y el AUC (área bajo la curva ROC).

Listing 7: Matriz de confusión, punto de corte 0.5.

```
y=cbind(ifelse(mu.mod1>0.5,1,0),ifelse(mu.mod2>0.5,1,0),ifelse(mu.mod3
    >0.5,1,0))

t1 = table(y[,1],falla)
t2 = table(y[,2],falla)
t3 = table(y[,3],falla)
```

```
> table(y[,1],falla)
   falla
     0   1
0 126   6
1   1   5
> table(y[,2],falla)
   falla
     0   1
0 127  11
> table(y[,3],falla)
   falla
     0   1
0 126   6
1   1   5
```

Se visualizan las matrices de confusión, del modelo logit, power logit y logit recíproco respectivamente. A continuación se presenta una tabla comparativa usando distintos criterios.

Listing 8: Error, Sensitividad,Especificidad, AUC.

```
Error = c(mean(y[,1]!=falla),mean(y[,2]!=falla),mean(y[,3]!=falla))
Sensitividad = c(t1[2,2]/(t1[1,2]+t1[2,2]),0,t3[2,2]/(t3[1,2]+t3[2,2]))
Especificidad=c(t1[1,1]/(t1[2,1]+t1[1,1]),t2[1,1]/(0+t2[1,1]),
t3[1,1]/(t3[2,1]+t3[1,1]))
AUC1<-roc(response=falla, predictor=y[,1])
AUC2<-roc(response=falla, predictor=y[,2])
AUC3<-roc(response=falla, predictor=y[,3])
val.AUC = c(auc(AUC1),auc(AUC2),auc(AUC3))

Criterios=rbind(Error,Sensitividad,Especificidad,val.AUC)
colnames(Criterios)<-c("Logit","Power Logit","Power Logit Reciproco")

Criterios
```

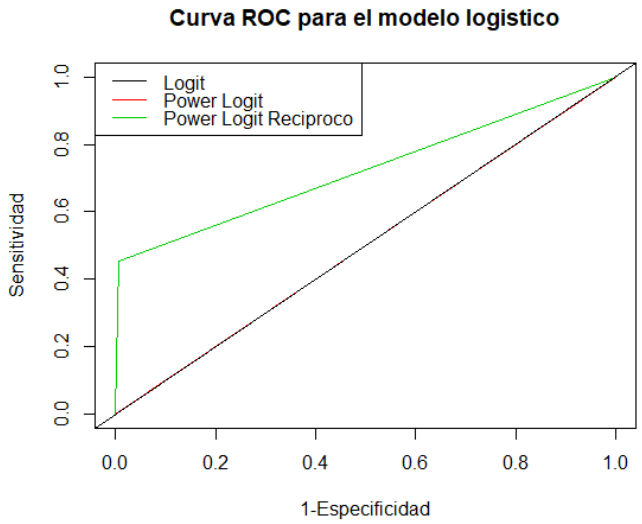
```
                Logit Power Logit Power Logit Reciproco
Error       0.05072464  0.07971014            0.05072464
Sensitividad 0.45454545 0.00000000            0.45454545
Especificidad 0.99212598 1.00000000           0.99212598
val.AUC     0.72333572  0.50000000            0.72333572
```

Se optará por el modelo Power Logit recíproco, dado que si bien presenta una tabla comparativa semejante al del modelo logit, en el siguiente item se evaluará que este modelo puede ser ajustado de tal modo que su área bajo la curva ROC sea mayor en comparación con los demás modelos.

c) Presente una gráfica comparativa que muestre las curvas ROC para los modelos considerados en (a).

Listing 9: Curva ROC.

```
plot(1-AUC1$specificities,AUC1$sensitivities,type="l",ylab="Sensitividad",
    xlab="1-Especificidad",col=3,lwd=1, main = "Curva ROC para el modelo
    logistico")
points(1-AUC2$specificities,AUC2$sensitivities,type="l",col=2,lwd=1)
points(1-AUC3$specificities,AUC3$sensitivities,type="l",col=3,lwd=1)
abline(a=0,b=1)
legend("topleft",legend=c("Logit","Power Logit","Power Logit Reciproco"),col=
    c(1:3),lty=c(1))
```



Como observamos el área bajo la curva ROC del modelo power logit recíproco es mayor, dando efectiva la elección del modelo.

d) El día del accidente del transbordador había una temperatura de 31 grados Fahrenheit. Considerando el mejor modelo encontrado en b) realice una estimación puntual y por intervalo de la probabilidad de que ocurra una falla en un anillo de goma.

Listing 10: Estimación puntual y de intervalo para temperatura de 31.

```
eta.new = b1.mod3 + b2.mod3*31
mu.new = 1-(1/(1+exp(eta.new)))^(mean(d.mod3))

summary(mu.new)
quantile(mu.new,probs=c(0.025,0.975))
```

2.5 %	mean	97.5 %
0.9972927	0.9984	1

Pregunta 4

El conjunto de datos College de la librería ISRL contiene estadísticas sobre características demográficas, matrícula, etc. De un gran número de universidades estadounidenses de la edición de 1995 de US News and World Report. Divida los datos en dos partes: Seleccione al azar 600 observaciones y asígnelas a una muestra de entrenamiento y las restantes a una muestra de evaluación (use 281117 con valor de semilla para realizar la selección).

Listing 11: Lectura de Base de datos.

```
library(ISLR)

# Datos
attach(College)
Private = factor(Private)
head(College)
```

```
                            Private Apps Accept Enroll Top10perc Top25perc F.Undergrad
Abilene Christian University     Yes 1660   1232    721        23        52        2885
Adelphi University               Yes 2186   1924    512        16        29        2683
Adrian College                   Yes 1428   1097    336        22        50        1036
Agnes Scott College              Yes  417    349    137        60        89         510
Alaska Pacific University        Yes  193    146     55        16        44         249
Albertson College                Yes  587    479    158        38        62         678
                            P.Undergrad Outstate Room.Board Books Personal PhD Terminal
Abilene Christian University         537     7440       3300   450     2200  70       78
Adelphi University                  1227    12280       6450   750     1500  29       30
Adrian College                        99    11250       3750   400     1165  53       66
Agnes Scott College                   63    12960       5450   450      875  92       97
Alaska Pacific University            869     7560       4120   800     1500  76       72
Albertson College                     41    13500       3335   500      675  67       73
                            S.F.Ratio perc.alumni Expend Grad.Rate
Abilene Christian University      18.1          12   7041        60
Adelphi University                12.2          16  10527        56
Adrian College                    12.9          30   8735        54
Agnes Scott College                7.7          37  19016        59
Alaska Pacific University         11.9           2  10922        15
Albertson College                  9.4          11   9727        55
```

Listing 12: Selección de muestras de entrenamiento y evaluacion.

```
#Semilla
set.seed(281117)

n = sample(nrow(College), 600)

# Muestra de entrenamiento (600 observaciones)
College.training = College[n,]

# Muestra de evaluacion
College.testing = College[-n,]
```

Observación: Con lo que respecta a los modelos con ciertos grados de libertad, tal grado en el spline será aumentado en uno, para que el modelo haga el balance correcto entre la estimación paramétrica y no paramétrica.

a) Con la muestra de entrenamiento, use la variable Outstate como variable respuesta y el resto como predictoras y realice el ajuste de los siguientes modelos:

- Modelo 1 : MLG asumiendo que la variable respuesta tiene una distribución normal y enlace identidad.

Listing 13: Selección de Modelo (Criterio AIC).

```
library(ISLR)

mod1 = glm(Outstate ~ ., data=entrenamiento, family=gaussian(link=identity))

stepAIC(mod1)
```

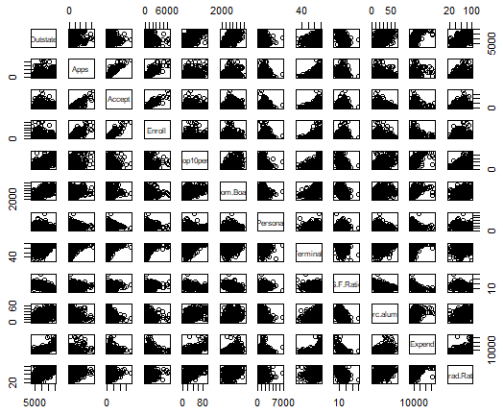
```
Call:  glm(formula = Outstate ~ Private + Apps + Accept + Enroll + Top10perc +
    Room.Board + Personal + Terminal + S.F.Ratio + perc.alumni +
    Expend + Grad.Rate, family = gaussian(link = identity), data = entrenamiento)

Coefficients:
(Intercept)   PrivateYes         Apps       Accept       Enroll    Top10perc
 -2291.0086    2454.5137      -0.3032       0.7895      -0.9280      11.1834
 Room.Board     Personal     Terminal    S.F.Ratio  perc.alumni       Expend
     0.8419      -0.2384      38.9056     -51.7516      37.3612       0.2585
  Grad.Rate
    23.9400

Degrees of Freedom: 599 Total (i.e. Null);  587 Residual
Null Deviance:      9.398e+09
Residual Deviance: 2.183e+09    AIC: 10790
```

Listing 14: Modelos MLG.

```
# Modelo optimo bajo Criterio AIC
mod1.AIC = glm(Outstate ~ Private + Apps + Accept + Enroll + Top10perc +
                Room.Board + Personal + Terminal + S.F.Ratio + perc.alumni +
                Expend + Grad.Rate, data = College.training,
                family=gaussian(link=identity))
```

Listing 15: Gráfica de dispersión de variables escogidas.

```
plot(~Outstate+Apps+Accept+Enroll+Top10perc+Room.Board+Personal+Terminal +
S.F.Ratio +perc.alumni+Expend+Grad.Rate)
```



- Modelo 2 : GAM considerando splines de suavizamiento con df=1.

Listing 16: Modelos 2 - Splines de suavizamiento df =1.

```
mod2.df1 = gam(Outstate ~ Private + s(Apps,2) + s(Accept,2) + s(Enroll,2) +
  s(Top10perc,2) + s(Room.Board,2) + s(Personal,2) + s(Terminal,2) +
  s(S.F.Ratio,2) + s(perc.alumni,2) + s(Expend,2) + s(Grad.Rate,2),
  data=College.training)

summary(mod2.df1)
```

```
Call: gam(formula = Outstate ~ Private + s(Apps, 2) + s(Accept, 2)
    s(Enroll, 2) + s(Top10perc, 2) + s(Room.Board, 2) + s(Personal
    2) + s(Terminal, 2) + s(S.F.Ratio, 2) + s(perc.alumni, 2) +
    s(Expend, 2) + s(Grad.Rate, 2), data = College.training)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-4777.37 -1179.53    40.56  1183.24  8318.15

(Dispersion Parameter for gaussian family taken to be 3257299)

    Null Deviance: 9398382681 on 599 degrees of freedom
Residual Deviance: 1876204155 on 575.9999 degrees of freedom
AIC: 10726.08

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
                Df     Sum Sq    Mean Sq F value    Pr(>F)
Private          1 2812605582 2812605582 863.478 < 2.2e-16 ***
s(Apps, 2)       1  975724172  975724172 299.550 < 2.2e-16 ***
s(Accept, 2)     1   50556464   50556464  15.521 9.161e-05 ***
s(Enroll, 2)     1   75397009   75397009  23.147 1.919e-06 ***
s(Top10perc, 2)  1 1036759680 1036759680 318.288 < 2.2e-16 ***
s(Room.Board, 2) 1  807946893  807946893 248.042 < 2.2e-16 ***
s(Personal, 2)   1   36527587   36527587  11.214 0.0008648 ***
s(Terminal, 2)   1  214176729  214176729  65.753 3.088e-15 ***
s(S.F.Ratio, 2)  1  157636566  157636566  48.395 9.500e-12 ***
s(perc.alumni, 2)1  128549642  128549642  39.465 6.588e-10 ***
s(Expend, 2)     1  405265479  405265479 124.418 < 2.2e-16 ***
s(Grad.Rate, 2)  1   62256888   62256888  19.113 1.462e-05 ***
Residuals      576 1876204155    3257299
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                Npar Df Npar F     Pr(F)
(Intercept)
Private
s(Apps, 2)            1  5.183   0.02317 *
s(Accept, 2)          1  5.480   0.01958 *
s(Enroll, 2)          1  0.763   0.38266
s(Top10perc, 2)       1  0.103   0.74845
s(Room.Board, 2)      1  2.613   0.10654
s(Personal, 2)        1  4.709   0.03041 *
s(Terminal, 2)        1  3.803   0.05164 .
s(S.F.Ratio, 2)       1  5.542   0.01890 *
s(perc.alumni, 2)     1  0.660   0.41686
s(Expend, 2)          1 57.472 1.383e-13 ***
s(Grad.Rate, 2)       1  1.988   0.15904
---
```

- Modelo 3 : GAM considerando splines de suavizamiento con df=4.

Listing 17: Modelos 3 - Splines de suavizamiento df =4.

```
mod3.df4 = gam(Outstate ~ Private + s(Apps,5) + s(Accept,5) + s(Enroll,5) +
  s(Top10perc,5) + s(Room.Board,5) + s(Personal,5) + s(Terminal,5)
  + s(S.F.Ratio,5) + s(perc.alumni,5) + s(Expend,5) + s(Grad.Rate,5),
  data=College.training)

summary(mod3.df4)
```

```
Call: gam(formula = Outstate ~ Private + s(Apps, 5) + s(Accept, 5) +
    s(Enroll, 5) + s(Top10perc, 5) + s(Room.Board, 5) + s(Personal,
    5) + s(Terminal, 5) + s(S.F.Ratio, 5) + s(perc.alumni, 5) +
    s(Expend, 5) + s(Grad.Rate, 5), data = College.training)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-5008.19 -1065.77    49.37  1090.75  7045.77

(Dispersion Parameter for gaussian family taken to be 2872001)

    Null Deviance: 9398382681 on 599 degrees of freedom
Residual Deviance: 1559494837 on 542.9994 degrees of freedom
AIC: 10681.15

Number of Local Scoring Iterations: 8

Anova for Parametric Effects
                Df     Sum Sq    Mean Sq   F value    Pr(>F)
Private          1 2918311105 2918311105 1016.1247 < 2.2e-16 ***
s(Apps, 5)       1 1016977876 1016977876  354.1008 < 2.2e-16 ***
s(Accept, 5)     1   73911513   73911513   25.7352 5.385e-07 ***
s(Enroll, 5)     1  135396519  135396519   47.1436 1.809e-11 ***
s(Top10perc, 5)  1  911217821  911217821  317.2763 < 2.2e-16 ***
s(Room.Board, 5) 1  698205094  698205094  243.1075 < 2.2e-16 ***
s(Personal, 5)   1   23945188   23945188    8.3375  0.004038 **
s(Terminal, 5)   1  166929416  166929416   58.1230 1.107e-13 ***
s(S.F.Ratio, 5)  1  178178572  178178572   62.0399 1.845e-14 ***
s(perc.alumni, 5)1  114527182  114527182   39.8771 5.630e-10 ***
s(Expend, 5)     1  498469825  498469825  173.5619 < 2.2e-16 ***
s(Grad.Rate, 5)  1   56953121   56953121   19.8305 1.028e-05 ***
Residuals      543 1559494837    2872001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                Npar Df  Npar F     Pr(F)
(Intercept)
Private
s(Apps, 5)            4  1.9663  0.098272 .
s(Accept, 5)          4 14.3230 4.001e-11 ***
s(Enroll, 5)          4  2.1766  0.070376 .
s(Top10perc, 5)       4  0.5095  0.728753
s(Room.Board, 5)      4  3.4809  0.008054 **
s(Personal, 5)        4  3.6857  0.005675 **
s(Terminal, 5)        4  4.5420  0.001289 **
s(S.F.Ratio, 5)       4  2.9062  0.021277 *
s(perc.alumni, 5)     4  1.3760  0.240943
s(Expend, 5)          4 17.9508 7.416e-14 ***
s(Grad.Rate, 5)       4  3.3129  0.010718 *
---
```

- Modelo 4 : GAM considerando regresión local y span = 0.25.

Listing 18: Modelos 4 - Regresión local sapn =0.25.

```
mod4.span025 = gam(Outstate~ Private+lo(Apps,span=0.25)+ lo(Accept,span=0.25)
+ lo(Enroll,span=0.25) + lo(Top10perc,span=0.25) + lo(Room.Board,span=0.25) +
  lo(Personal,span=0.25) + lo(Terminal,span=0.25) +lo(S.F.Ratio,span=0.25)+
  lo(perc.alumni,span=0.25)+ lo(Expend,span=0.25) + lo(Grad.Rate,span=0.25),
  data= College.training)

summary(mod4.span025)
```

```
Call: gam(formula = Outstate ~ Private + lo(Apps, span = 0.25) + lo(Accept,
    span = 0.25) + lo(Enroll, span = 0.25) + lo(Top10perc, span = 0.25) +
    lo(Room.Board, span = 0.25) + lo(Personal, span = 0.25) +
    lo(Terminal, span = 0.25) + lo(S.F.Ratio, span = 0.25) +
    lo(perc.alumni, span = 0.25) + lo(Expend, span = 0.25) +
    lo(Grad.Rate, span = 0.25), data = College.training)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-5100.53 -1015.49    75.64   995.25  6913.97

(Dispersion Parameter for gaussian family taken to be 2784948)

    Null Deviance: 9398382681 on 599 degrees of freedom
Residual Deviance: 1412544876 on 507.207 degrees of freedom
AIC: 10693.35

Anova for Parametric Effects
                          Df     Sum Sq    Mean Sq   F value    Pr(>F)
Private                 1.00 2968156170 2968156170 1065.7853 < 2.2e-16 ***
lo(Apps, span = 0.25)   1.00 1010001778 1010001778  362.6646 < 2.2e-16 ***
lo(Accept, span = 0.25) 1.00   25065124   25065124    9.0002  0.002832 **
lo(Enroll, span = 0.25) 1.00  181143924  181143924   65.0439 5.318e-15 ***
lo(Top10perc, span = 0.25) 1.00  985638588  985638588  353.9164 < 2.2e-16 ***
lo(Room.Board, span = 0.25) 1.00  728343828  728343828  261.5287 < 2.2e-16 ***
lo(Personal, span = 0.25) 1.00   19940649   19940649    7.1602  0.007695 **
lo(Terminal, span = 0.25) 1.00  141369820  141369820   50.7621 3.598e-12 ***
lo(S.F.Ratio, span = 0.25) 1.00  181062717  181062717   65.0148 5.388e-15 ***
lo(perc.alumni, span = 0.25) 1.00   97921572   97921572   35.1610 5.618e-09 ***
lo(Expend, span = 0.25) 1.00  436964229  436964229  156.9021 < 2.2e-16 ***
lo(Grad.Rate, span = 0.25) 1.00   62091077   62091077   22.2952 3.030e-06 ***
Residuals             507.21 1412544876    2784948
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                          Npar Df Npar F     Pr(F)
(Intercept)
Private
lo(Apps, span = 0.25)         7.8  6.122 1.929e-07 ***
lo(Accept, span = 0.25)       7.7 34.791 < 2.2e-16 ***
lo(Enroll, span = 0.25)       7.5  9.494 1.153e-11 ***
lo(Top10perc, span = 0.25)    6.4  0.785 0.5900980
lo(Room.Board, span = 0.25)   6.9  3.711 0.0006971 ***
lo(Personal, span = 0.25)     7.5  2.609 0.0100025 *
lo(Terminal, span = 0.25)     7.0  3.723 0.0005974 ***
lo(S.F.Ratio, span = 0.25)    7.1  2.133 0.0378235 *
lo(perc.alumni, span = 0.25)  6.5  0.948 0.4652915
lo(Expend, span = 0.25)       7.9  9.411 4.953e-12 ***
lo(Grad.Rate, span = 0.25)    7.4  1.821 0.0767429 .
---
```

- Modelo 5 : GAM considerando regresión local y span = 0.75.

Listing 19: Modelos 4 - Regresión local sapn =0.25.

```
mod5.span075 = gam(Outstate~ Private+lo(Apps,span=0.75)+ lo(Accept,span=0.75)
  + lo(Enroll,span=0.75) + lo(Top10perc,span=0.75) + lo(Room.Board,span=0.75)
       +
  lo(Personal,span=0.75) + lo(Terminal,span=0.75) +lo(S.F.Ratio,span=0.75)+
  lo(perc.alumni,span=0.75)+ lo(Expend,span=0.75) + lo(Grad.Rate,span=0.75),
  data= College.training)

summary(mod5.span075)
```

```
Call: gam(formula = Outstate ~ Private + lo(Apps, span = 0.75) + lo(Accept,
    span = 0.75) + lo(Enroll, span = 0.75) + lo(Top10perc, span = 0.75) +
    lo(Room.Board, span = 0.75) + lo(Personal, span = 0.75) +
    lo(Terminal, span = 0.75) + lo(S.F.Ratio, span = 0.75) +
    lo(perc.alumni, span = 0.75) + lo(Expend, span = 0.75) +
    lo(Grad.Rate, span = 0.75), data = College.training)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-5132.65 -1084.18    71.57  1190.19  7599.69

(Dispersion Parameter for gaussian family taken to be 3047220)

    Null Deviance: 9398382681 on 599 degrees of freedom
Residual Deviance: 1729374788 on 567.5255 degrees of freedom
AIC: 10694.13

Number of Local Scoring Iterations: 3

Anova for Parametric Effects
                          Df     Sum Sq    Mean Sq  F value    Pr(>F)
Private                 1.00 2868372989 2868372989 941.3083 < 2.2e-16 ***
lo(Apps, span = 0.75)   1.00  982208359  982208359 322.3294 < 2.2e-16 ***
lo(Accept, span = 0.75) 1.00   46496957   46496957  15.2588  0.000105 ***
lo(Enroll, span = 0.75) 1.00  126007613  126007613  41.3517 2.701e-10 ***
lo(Top10perc, span = 0.75) 1.00  928022214  928022214 304.5472 < 2.2e-16 ***
lo(Room.Board, span = 0.75) 1.00  701850726  701850726 230.3250 < 2.2e-16 ***
lo(Personal, span = 0.75) 1.00   29436648   29436648   9.6602  0.001977 **
lo(Terminal, span = 0.75) 1.00  173583797  173583797  56.9646 1.778e-13 ***
lo(S.F.Ratio, span = 0.75) 1.00  169264110  169264110  55.5471 3.425e-13 ***
lo(perc.alumni, span = 0.75) 1.00  111436274  111436274  36.5698 2.673e-09 ***
lo(Expend, span = 0.75) 1.00  435727732  435727732 142.9919 < 2.2e-16 ***
lo(Grad.Rate, span = 0.75) 1.00   63900485   63900485  20.9701 5.737e-06 ***
Residuals             567.53 1729374788    3047220
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                          Npar Df  Npar F     Pr(F)
(Intercept)
Private
lo(Apps, span = 0.75)         2.5  2.0165   0.12275
lo(Accept, span = 0.75)       2.4 27.3298 7.705e-14 ***
lo(Enroll, span = 0.75)       2.3 12.6128 1.165e-06 ***
lo(Top10perc, span = 0.75)    1.4  0.2366   0.70440
lo(Room.Board, span = 0.75)   1.6  2.8016   0.07495 .
lo(Personal, span = 0.75)     2.0  4.6053   0.01007 *
lo(Terminal, span = 0.75)     1.2  1.5122   0.22265
lo(S.F.Ratio, span = 0.75)    1.5  4.8752   0.01487 *
lo(perc.alumni, span = 0.75)  1.1  1.6247   0.20379
lo(Expend, span = 0.75)       2.4 30.1377 4.330e-15 ***
lo(Grad.Rate, span = 0.75)    1.1  2.0889   0.14629
---
```

Comente sus hallazgos y compare los resultados seleccionando un modelo usando como criterio el AIC y delta AIC (ver Burnham y Anderson, 2004).
Nota: Para determinar las variables a usar en los modelos GAM puede usar la siguiente estrategia, realice una selección de variables usando el método stepwise en el modelo 1, y utilice las variables predictoras que fueron seleccionadas para el ajuste de los siguientes modelos.

Listing 20: Selección del modelo (CRITERIO AIC).

```
VAL.AIC=AIC(mod1.GLM,mod2.df1,mod3.df4,mod4.span025,
  mod5.span075)
delta.AIC=VAL.AIC[,2]-min(VAL.AIC[,2])
VAL.AIC=cbind(VAL.AIC,delta.AIC)

VAL.AIC
```

```
                df      AIC delta.AIC
mod1.GLM        14 10794.83 113.68748
mod2.df1        14 10726.08  44.93227
mod3.df4        14 10681.15   0.00000
mod4.span025    14 10693.35  12.20348
mod5.span075    14 10694.13  12.98668
```

De acuerdo al Criterio AIC, el mejor modelo será el que presente menor AIC, siendo el correspondiente al **modelo 3 GAM - splines de suavizamiento con df = 1**. Tomandose como referencia el delta AIC < 2.6 (por Burnham y Anderson, 2004). Aún el modelo optado es el único cumpliendo el criterio.

b) Evalúe los modelos obtenidos usando la muestra de evaluación. Explique los resultados obtenidos y compárelos con los de la pregunta a).

Listing 21: Diseño de tabla de predicción por los modelos sobre los datos de evaluación.

```
pred1 = predict(mod1.GLM,newdata=College.testing)
pred2 = predict(mod2.df1,newdata=College.testing)
pred3 = predict(mod3.df4,newdata=College.testing)
pred4 = predict(mod4.span025,newdata=College.testing)
pred5 = predict(mod5.span075,newdata=College.testing)

Tabla.pred = cbind(data.frame(College.testing$Outstate),pred1,pred2,pred3,pred4)
```

Listing 22: Errores Medios Cuadráticos.

```
Tabla.dif=cbind((Tabla.pred[,1]-Tabla.pred[,2])^2,
  (Tabla.pred[,1]-Tabla.pred[,3])^2,(Tabla.pred[,1]-Tabla.pred[,4])^2,
  (Tabla.pred[,1]-Tabla.pred[,5])^2,(Tabla.pred[,1]-Tabla.pred[,5])^2)

ECM=apply(Tabla.dif,2,mean)
```

	modelo 1	modelo 2	modelo 3	modelo 4	modelo 5
ECM	4514790	3804333	3971827	4669766	4669766

Como observamos, los modelos 2 y modelos 3 presentan los menores errores medios cuadráticos. En el item a) se escogío el modelo 3 y pues parece que el ajuste es aceptable estadísticamente analizando su ECM. Con respecto al modelo 2, pues aunque presente el menor ECM, vimos que su delta AIC es muy grande a comparación del criterio delta, por lo que no creemos sea un buen ajuste.

c) ¿Para qué variables (si fuera el caso) hay evidencia de una relación no lineal con la variable respuesta? Sustente su respuesta usando gráficas y pruebas estadísticas.

Primero realicemos un análisis de varianza para evidenciar que modelo mediante este análisis es estadísticamente aceptable.

Listing 23: ANOVA.

```
anova(mod1.GLM,mod2.df1,mod3.df4,mod4.span025,mod5.span075,test="F")
```

```
  Resid. Df Resid. Dev      Df   Deviance       F    Pr(>F)
1    587.00 2182584356
2    576.00 1876204155  11.000  306380201 10.0011 < 2.2e-16 ***
3    543.00 1559494837  33.000  316709318  3.4461  1.55e-09 ***
4    507.21 1412544876  35.792  146949961  1.4742 0.0403232 *
5    567.53 1729374788 -60.318 -316829911  1.8861 0.0001453 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 1: Modelo 1

Observamos que la igualdad de varianzas se mantiene estadísticamente significativa hasta el modelo 3 (el modelo optado). A partir del modelo 4 la significancia crece, es claro que el p-valor para el modelo 4 es menor a 0.05; puede deberse a que en los análisis previos se toma un criterio de significancia un poco menor. Concluimos que le modelo 3 es el de mejor ajuste y dado que toma splines de suavizamiento

Listing 24: Gráfico de Modelos.

```
par(mfrow=c(3,4))
plot.gam(mod1.GLM, se=TRUE, col="red")
plot.gam(mod2.df1, se=TRUE, col="red")
plot.gam(mod3.df4, se=TRUE, col="red")
plot.gam(mod4.span025, se=TRUE, col="red")
plot.gam(mod5.span075, se=TRUE, col="red")
```

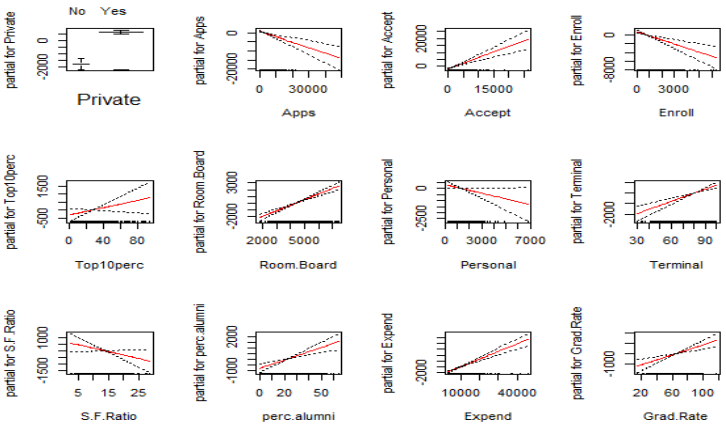


Figura 2: Modelo 1

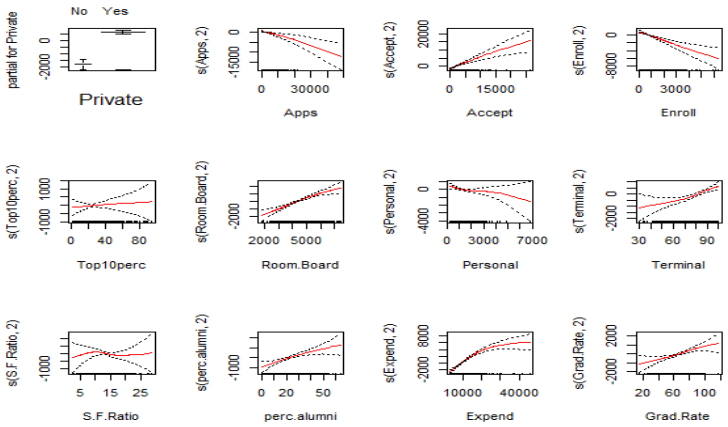


Figura 3: Modelo 2

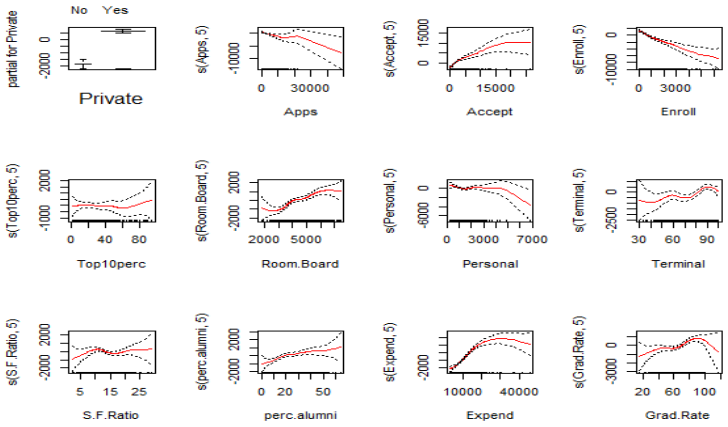


Figura 4: Modelo 3

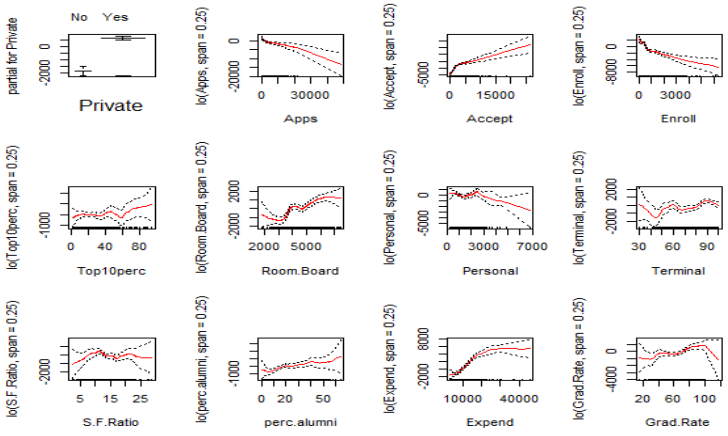


Figura 5: Modelo 4

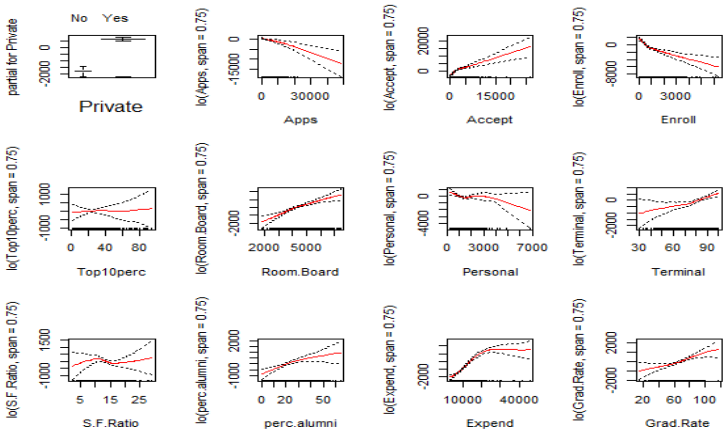


Figura 6: Modelo 5

La matriz de dispersión generada en el item a), hace presencia de una posible relación no lineal de la variable respuesta con la variable Expend. Es estadísticamente significativa según el modelo, sin embargo veamos mediante un gráfico

Listing 25: Gráfico de Modelos Splines, sobre variable Expend.

```
Expendlims=range(Expend)
Expend.grid=seq(from=Expendlims[1],to=Expendlims[2])
plot(College.testing$Expend,College.testing$Outstate,xlim=Expendlims,cex=.5,col="
    darkgrey")

m2=smooth.spline(College.training$Expend,College.training$Outstate,df=2)
m3=smooth.spline(College.training$Expend,College.training$Outstate,df=5)
lines(m2,col="red",lwd=2)
lines(m3,col="blue",lwd=2)
title("Smoothing Spline")
legend("topright",legend=c("1 DF","4 DF"),col=c("red","blue"),lty=1,lwd=2,cex=.8)
```

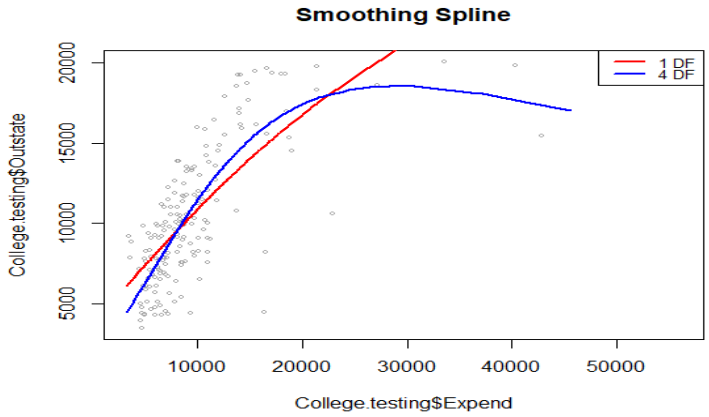


Figura 7: Modelo 1

Según la gráfica y la significancia del modelo, podemos concluir que efectivamente existe evidencia estadística de que la variable Expend mantiene una relación no lineal con la variable respuesta.

Podemos observar que será necesaria ajustar mejor los modelos dado que si bien el modelo, por ejemplo considera a la variable Grade.Rate en una relación no lineal veamos que gráficamente podría no ser un ajuste óptimo.

Listing 26: Gráfico de Modelos Splines, sobre variable Grade.Rate.

```
Grad.Ratelims=range(Grad.Rate)
Grad.Rate.grid=seq(from=Grad.Ratelims[1],to=Grad.Ratelims[2])
plot(College.testing$Grad.Rate,College.testing$Outstate,xlim=Grad.Ratelims,cex
    =.5,col="darkgrey")

m2=smooth.spline(College.training$Grad.Rate,College.training$Outstate,df=2)
m3=smooth.spline(College.training$Grad.Rate,College.training$Outstate,df=5)
lines(m2,col="red",lwd=2)
lines(m3,col="blue",lwd=2)
title("Smoothing Spline")
legend("topright",legend=c("1 DF","4 DF"),col=c("red","blue"),lty=1,lwd=2,cex=.8)
```

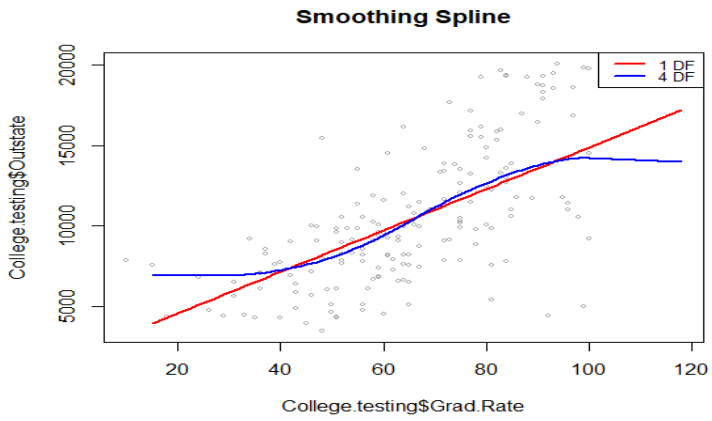


Figura 8: Modelo 1