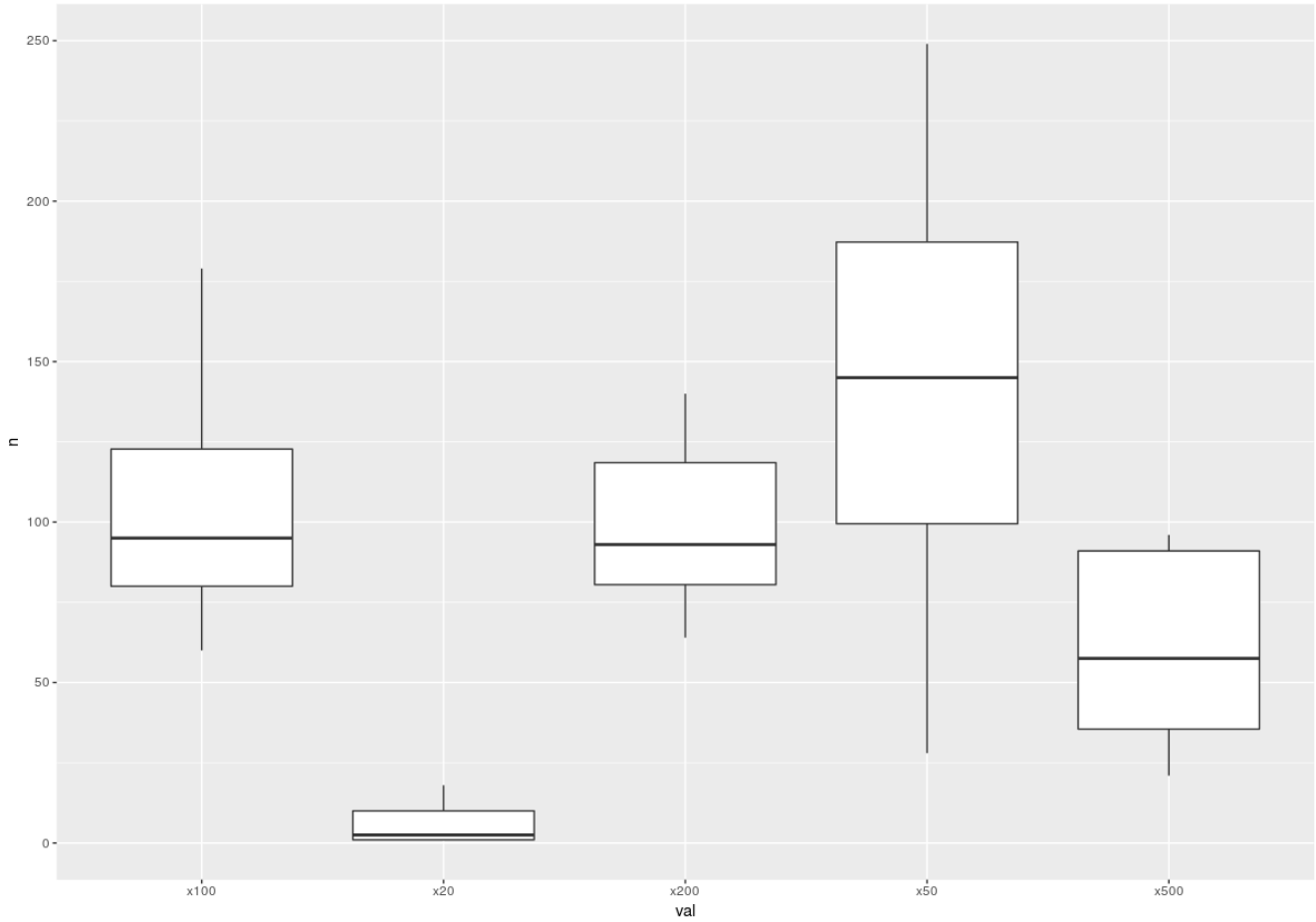
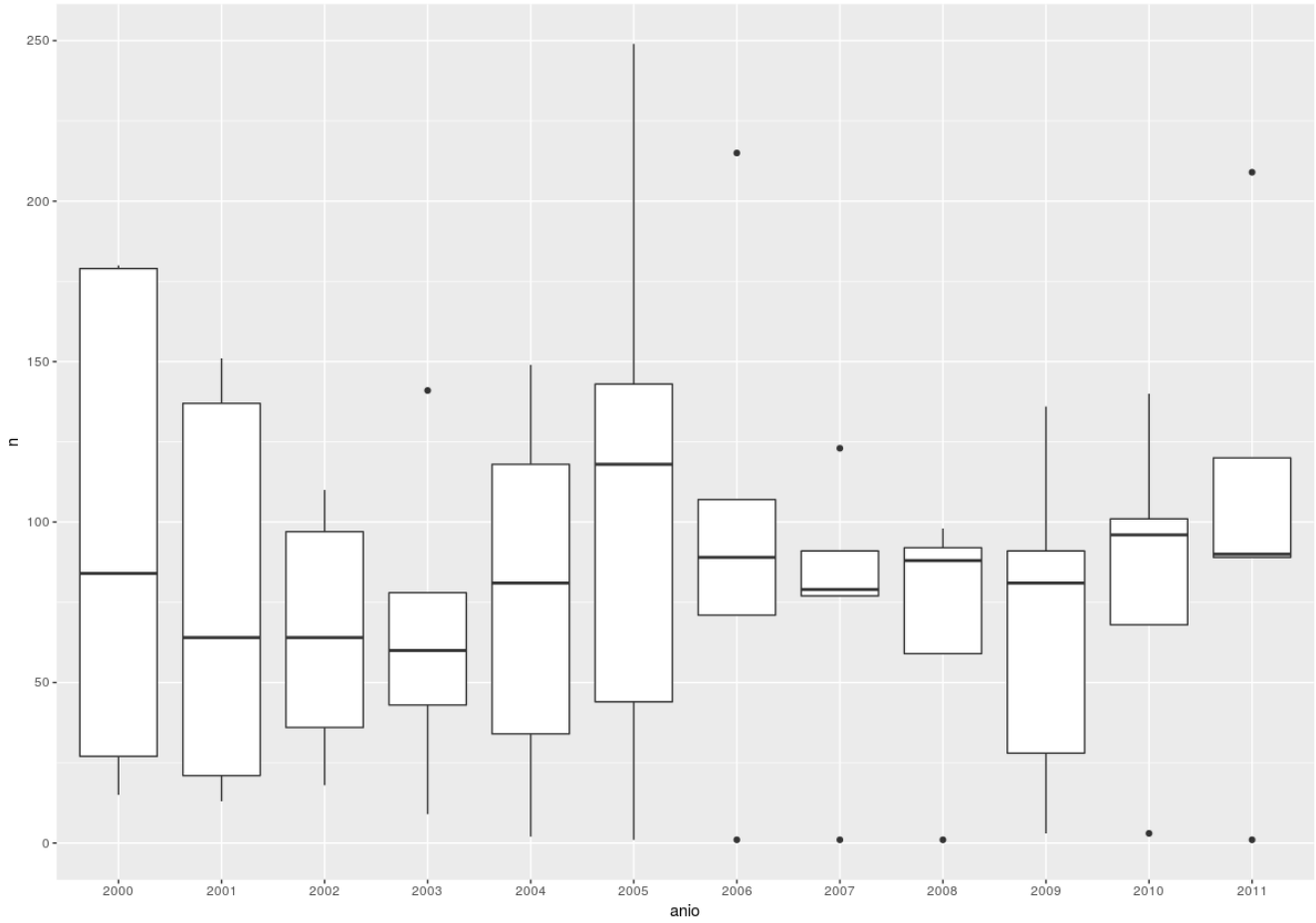
Abraham Nieto 51556

1. El Banco de México es el responsable de emitir los billetes que circulan en la economía mexicana. Se cuenta con la información del número de billetes en circulación (C) y la cantidad de billetes falsos (Y), ambas en millones de piezas, para los años de 2000 a 2011. Para identificar la denominación del billete definimos variables indicadoras x20, x50, x100, x200 y x500. La base de datos se encuentra en http://allman.rhon.itam.mx/~lnieto/index_archivos/BillsMXc.csv

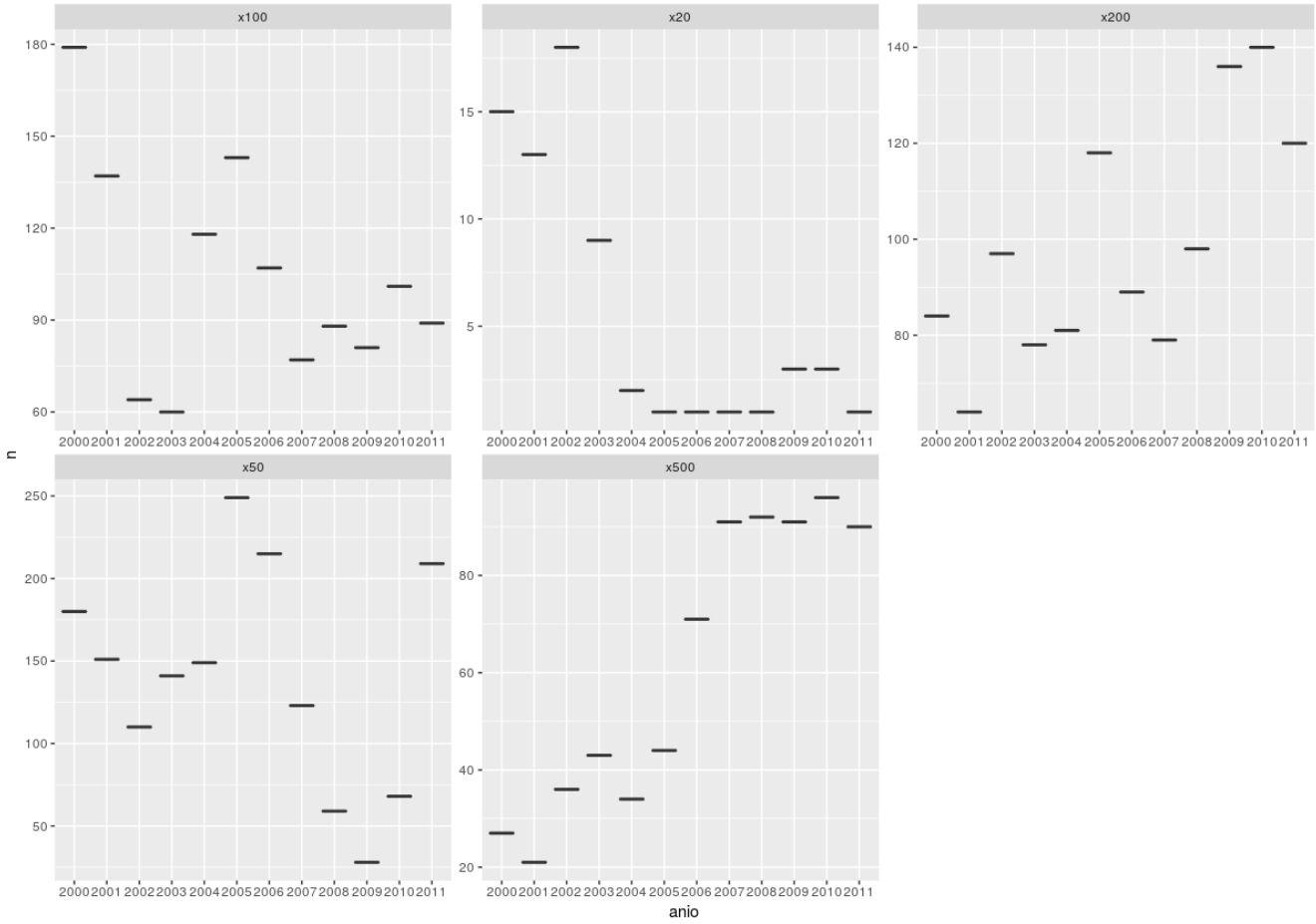
a)Realiza un análisis exploratorio de los datos. Crea las gráficas y encuentra las estadísticas que mejor describan la información y coméntalas. Obtén conclusiones por tipo de denominación.



Se hizo una gráfica de caja y brazos donde se muetra en ele eje de las xs la denominación de los billetes vs.el número de billetes falsos cada caja representa la distribción del número de billetes falsos por año, podemos observar que el billete más falsificado es el de 50 ya que podemos ver claramente que su mediana es de casi 150 millones de billetes falsos, y al menos 75% de los años se han falsificado más de 100 millones de billetes de 50, el que menos se falsifica es el billete de 20, el seundo menos falsificado es el de 500 con mediana arriba de 50 millones de piezas peron ningún año pasó los 100 millonesde piezas ,el de 200 y el de 100 son parecidos en sus medianas cerca de 100 millones.



Ahora vemos la distribución del número de falsificaciones por año de cualquier denominación, en este gráfico podemos ver que el año donde hubo mayor falsificación de billetes fue en el 2005 ya que al menos el 50% de las denominaciones tuvo mas de 100 millones de piezas falsas, el a;o donde pudo haber menos piezas falsas fue 2003 y 2007.
La tendencia de falsificación parece decrecer a a partir de 2005 aunque 2011 parece un ligero repunte.



También podemos ver por tipo de billete en que años fueron más falsificados por ejemplo los billetes de 100 se falsficaron más en el 2000 y menos en 2003, los de 20 en 2002 más, los de 200 en 2010, los de 50 en 2005(250 mill de pzas.) y menos en 2009(menos de 50 mill de pzas.) y los de 500 en 2010 fueron más flasificados.

b)Ajuste el modelo de regresión binomial con liga logística, i.e., como distribuciones iniciales ~ 0,0.001 para 1, ... ,5. Calcula los indicadores de ajuste DIC y pseudoR2. Encuentra los estimadores puntuales y por intervalo de los parámetros del modelo, interprétalos y comenta qué tan bueno es el modelo.

Para el modelo lineal generalizado con liga llogistica tenemos los siguientes resultados:

DIC=[1] 1297
R2= 0.5542991

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
b[1]	-6.23353189	1.218929e-01	-6.479000	-6.313000	-6.231000	-6.151e+00	-6.002e+00	1.000963	18000

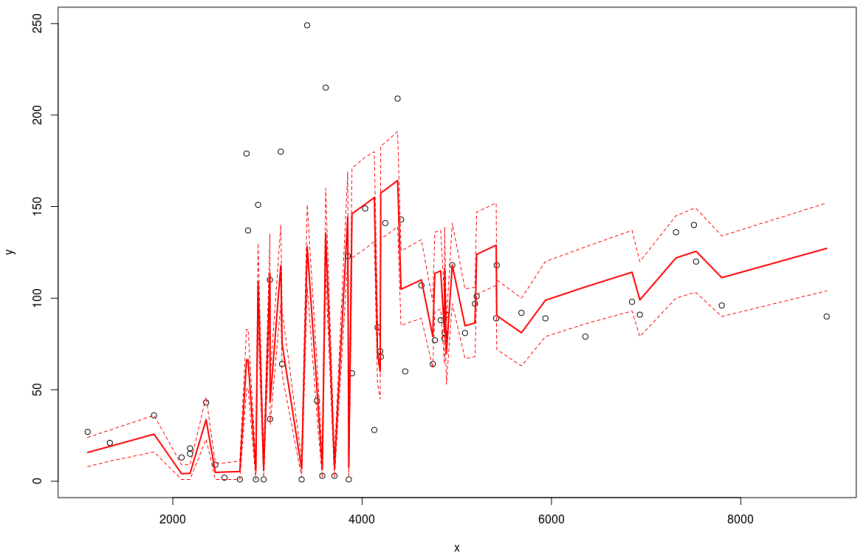
```
b[2]         2.98884361 1.243028e-01     2.754000     2.904000     2.986000   3.071e+00   3.239e+00
1.000961 18000
b[3]         2.51902294 1.249656e-01     2.279975     2.435000     2.517000   2.601e+00   2.771e+00
1.000987 18000
b[4]         2.15625694 1.247832e-01     1.918975     2.072000     2.153500   2.238e+00   2.405e+00
1.000948 18000
b[5]         1.99839500 1.274431e-01     1.754975     1.912000     1.996000   2.082e+00   2.252e+00
1.000953 18000
```

los valores de las betas se pueden ver en amarillo donde los intervalos de cfza an del percentil 2.5% al 97.5%. y la estimación es la media, en cada uno de los coficientes podemos observar que significativos dado que los intervalos de confianza no c**ontienen al 0.**

Dado que no incluimos la variable x20 en el modelo este será tomada en cuenta como la base, en tonces las b[j]'s con j=2,...5 representarán la variación en el término Logit (logaritmo neperiano del cociente de probabilidades) originada por una variación unitaria en la variable X[j] (suponiendo constantes el resto de variables explicativas) con respecto a la base.

para interpretar los coeficientes debemos observar lo siguiente primero b[1] que representa el intercepto hace referencia al efecto de la variable x20 ya que es la base dado que no se incluyo, entonces el significado es que un aumento unitario en la circulación de billetes de 20 aumentaría el momio de que fuera falso en exp{-6.23353189}, para el caso del resto de x's se explicara en el siguiente inciso para responderlo.

podemos ver que el ajuste del modelo no es muy bueno dado que la r2 es muy baja entonces el modelo explica poca variabilidad de los datos además en la gráfica se ve que muchas predicciones están fuera del intervalo de confianza.



c)En el modelo de regresión binomial con liga logística, ¿cuál es la interpretación del coeficiente b1 en modelo?, ¿cómo interpretas la suma b1+bj para
2, ... ,5?

b1 representa el valor del coeficiente para la variable de billetes de 20 que es nuestra base dado que no está en el modelo explícita, y como lo mencioné en el inciso anterior un aumento unitario en la circulación de billetes de 20 aumentaría el momio de que fuera falso en exp{-6.23353189}, entonces b1 representa el log-momio de la proporción de billestes falsosde 20 vs. El resto, para el caso de b1+bj la interpretación es la siguiente, dado que como se vio en el inciso anterior cada bj representa variaciones entonces cuando le sumamos b1 lo que obtenemos es el efecto completo del coeficiente sobre la variable xj, entonces b1+bj representa el log momio de la proporción de billetes flasos de j$ vs. El resto., entonces para interpretar un cambio unitario en la circulación de billetes de denominación j podemos ver que el amento del momio de que fuera falso aumenta en exp{b1+bj}.
Por ejemplo para el modelo del inciso b) un aumento unitario en los billetes de 50 aumenta el momio de que sea falso en
exp{b1+b[2]}=exp{ -6.23353189+2.98884361}=0.03898071.
exp{b1+b[3]}=exp{ -6.23353189+2.51902294}=0.0243674.
exp{b1+b[4]}=exp{-6.23353189+2.15625694}=0.0169536.
exp{b1+b[5]}=exp{ -6.23353189+1.99839500}=0.01447783.

d) En el modelo de regresión binomial con liga logística define "la tasa de billetes falsos por mil circulando" estima estas tasas mediante un intervalo de 95% de probabilidad y coméntalas.

Las estimaciones para cada p[i] son las siguientes:

	mean	2.5%	97.5%
prob[1]	1.97313	1.533	2.467
prob[2]	37.52851	35.810	39.320
prob[3]	23.79721	22.490	25.150
prob[4]	16.67761	15.750	17.620
prob[5]	14.28078	13.260	15.320

p[1]=1.97 es la tasa estimada de billetes falsos por mil de 20 con intervalo de confianza al 95% (1.533,2.467)
p[2]=37.5 es la tasa estimada de billetes falsos por mil de 20 con intervalo de confianza al 95% (35.810, 39.320)
p[3]=23.79 es la tasa estimada de billetes falsos por mil de 100 con intervalo de confianza al 95% (22.490, 25.150)
p[4]=16.67 es la tasa estimada de billetes falsos por mil de 200 con intervalo de confianza al 95% (15.750, 17.620)
p[5]=14.28 es la tasa estimada de billetes falsos por mil de 500 con intervalo de confianza al 95% (13.260,15.320)

podemos ver que la mayor tasa de billetes falsos por mil es la de los billetes de 50, seguido de los billetes de 100 y la menor es la de los billetes de 20.

e)Ajuste el modelo de regresión binomial con liga complementaria log-log, i.e., 1
y usa como distribuciones iniciales ~ 0,0.001 para 1, ... ,5. Calcula los indicadores de ajuste DIC y pseudoR2. Encuentra los estimadores puntuales y por intervalo de los parámetros del modelo, interpétalos y comenta qué tan bueno es el modelo.

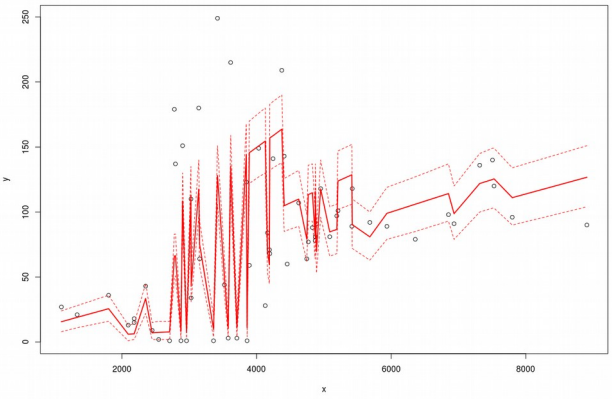
Los resultados del modelo con liga cloglog son los siguientes:

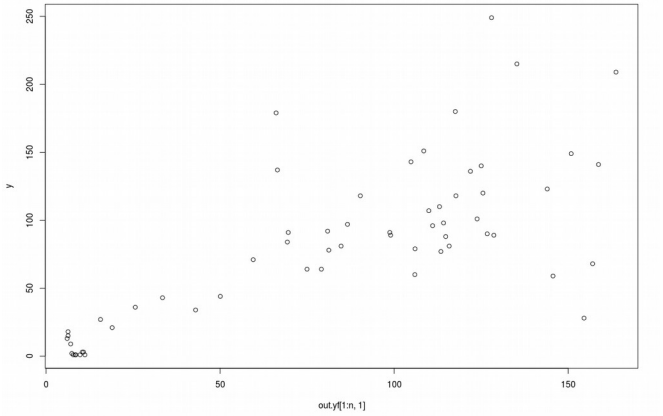
DIC=1296
R2= 0.5545616

```
             mean   2.5%  97.5%
b[1]    -6.242012 -6.463 -6.030
b[2]     2.978136  2.762  3.205
b[3]     2.515380  2.299  2.742
b[4]     2.156488  1.935  2.382
b[5]     2.000020  1.770  2.232
```

los valores de las betas se muestran arriba los intervalos de cfza an del percentil 2.5% al 97.5%. y la estimación es la media, en cada uno de los coeficientes podemos observar que significativos dado que los intervalos de confianza no c**ontienen al 0.**

En general, el modelo no es muy bueno con respecto a su pseudoR2 .55 explica sólo el 55% de la varianza el DIC tampoco es muy bueno.
Podemos ver también que el ajuste deja muchos puntos fuera del intervalo de confianza.
Por otro lado viendo la gráfica de y ajustada vs. Y parece que en los billetes de 100 en adelante hay mucho error igual en los de 20 pero ahondaremos en el inciso h)





f)En el modelo de regresión binomial con liga complementaria log-log, ¿cuál es la

interpretación del coeficiente b1 en modelo?, ¿cómo interpretas la suma b1+bj para
2, ... ,5?

En primer lugar, debemos decir igual que en el inciso c) que b1 represenat el efecto de billetes de 20 y que b1+bj representa el efecto de la variable xj en el modelo, en el inciso c) se podía hacer una interpretación del efecto de cambio en las bj's por el lad del impacto en el momio pero en este caso recordamos que p1=(1-exp(-exp(b[1]))) entonces un cambio unitario en los billetes de 20 no tiene una interpretación directa en cuanto a log de los momios, lo mismo sucede con p[j]=(1-exp(-exp(b[1]+b[j]))) para cambios unitarios en los billetes de j$ no exist interpetación directa sobre el log de los momios.
Para la función cloglog las betas se pueden interpretar como aumento de probabilidad de que el billete sea falso entonces b1 es el aumento en la proba de que un billete de 20 sea falso y b1+bj el aumento de la proba de que un billete de j$ lo sea.

g)En el modelo de regresión binomial con liga complementaria log-log, define "las tasas de billetes falsos por mil circulando" para cada denominación, ,
1, ... ,5
y estímalas por intervalo al 95%. NOTA que estas nuevas tasas no se definen igual que en el caso de la liga logística. Encuentra tú la definición correcta para la liga que estas usando.

Los resultados son los siguientes:
Primero muestro las expresiones correspondientes a cada p[j] para j=1,...5:
dado log(-log(1-pi))=b1 se iguala a b1 y entonces la expresión de p1=1000*(1-exp(-exp(b[1]))) tasas por mil.
Y para b[j] log(-log(1-pi))=b1+bj entonces pj=1000*(1-exp(-exp(b[1]+b[j])))
y dado esto las estiaciones son las siguientes

```
             mean  2.5%  97.5%
prob[1]  1.956437  1.56  2.403
prob[2] 37.528211 35.82 39.270
prob[3] 23.795395 22.54 25.120
prob[4] 16.680519 15.76 17.600
prob[5] 14.285443 13.29 15.310
```

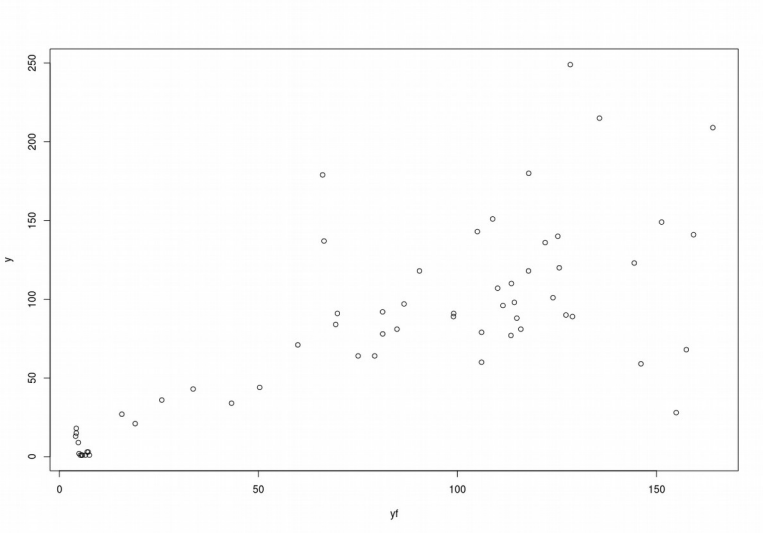
p[1]=1.95 es la tasa estimada de billetes falsos por mil de 20 con intervalo de confianza al 95% (`1.56, 2.403`)
p[2]=37.5 es la tasa estimada de billetes falsos por mil de 20 con intervalo de confianza al 95% (`35.82, 39.270`)
p[3]=23.79 es la tasa estimada de billetes falsos por mil de 100 con intervalo de confianza al 95% (`22.54 25.120`)
p[4]=16.68 es la tasa estimada de billetes falsos por mil de 200 con intervalo de confianza al 95% (`13.29,15.310`)
p[5]=14.28 es la tasa estimada de billetes falsos por mil de 500 con intervalo de confianza al 95% (`13.29,15.310`)

h)Compara los modelos de regresión binomial con las dos ligas, logística y complementaria log-log. De acuerdo con sus medidas de ajuste determina cuál de los dos es el mejor. Con el mejor modelo realiza una gráfica de predicción del número de billetes falsos y compáralo con los datos observados. Comenta los puntos importantes de esta gráfica. En particular comenta sobre los billetes de $20 y de $50.

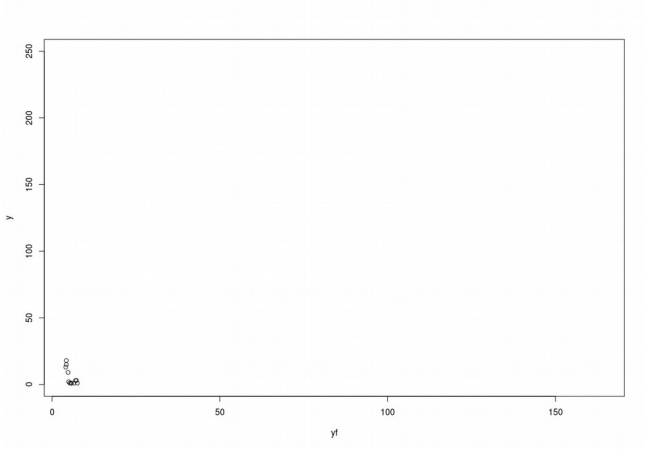
Modelo	DIC	pseudoR2
b) con liga logistica	`1297`	`0.5542991`
e) con liga cloglog	`1296`	`0.5545616`

Podemos ver que el DIC del modelo e) es menor al del inciso b) y además la pseudoR2 del modelo e) es ligeramente mejor que la del modelo con liga logistica, por tanto el mejor modelo el el de la liga clog-log, pero en realidad los modelos parecen iguales en temas de ajuste.

Hacemos la gráfica de predicción:

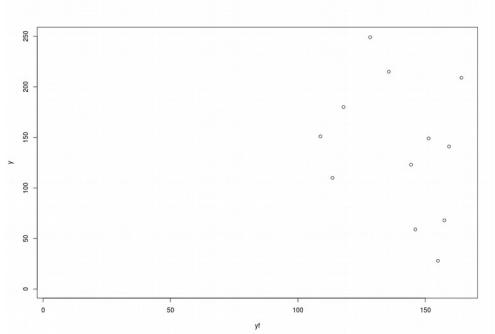


vemos que el las predicciones (yf) y los datos observados (y) tienen mucha dispersión o error se corrobora lo que se concluyó en el inciso b) del no tan buen ajuste del modelo con liga logistica.
En particular para los billetes de 20:



las predicciones y los datos observados del número de billetes falsos ajustan muy mal, es decir las predicciones con este modelo para los billetes de 20 son poco confiables.

Para los billetes de 50:



Existe muchísima varianza en las predicciones y las observaciones de billetes falsos de 50 por tanto las predicciones no son buenas, en conclusión las predicciones para ambos casos son deficientes.

i)Con el mejor modelo, compara las estimaciones de "las tasas de billetes falsos por mil circulando" para las cinco denominaciones. Determina cuales de ellas son estadísticamente diferentes justificando tu respuesta con las estimaciones obtenidas.

Mostramos las pi s del modelo e) con liga clog-log:

	mean	2.5%	97.5%
prob[1]	1.956437	1.56	2.403
prob[2]	37.528211	35.82	39.270
prob[3]	23.795395	22.54	25.120
prob[4]	16.680519	15.76	17.600
prob[5]	14.285443	13.29	15.310

primero notar que el intervalo de confianza de p1=prob[1]=(1.56,2.403) que no se intersecta con ninguno de los intervalos de p[j] (35.82,39.270),(22.54,25.120),(15.76,17.600),(13.29,15.310) intervalosde p2 a p5 respectivamente de hecho se puede ver que ninguno de los intervalos se intersecta por tanto son estadístcamente distintas.
Podemos ver que los billetes de mayor tasa de flasificación son los de 50 (37.5) seguido por los 100(23.79) después los de 200(16.68), los de 500(14.28) y los menos falsificados son los de 20(1.95).

Anexo codigo bugs:

ejea.txt

```
model{
#Likelihood
for (i in 1:n){
y[i]~dbin(pi[i],e[i])

logit(pi[i])<-b[1]+b[2]*x50[i]+b[3]*x100[i]+b[4]*x200[i]+b[5]*x500[i]


}


p[1]<-1000*exp(b[1])/(1+exp(b[1]))

for (k in 2:5){p[k]<-1000*exp(b[1]+b[k])/(1+exp(b[1]+b[k]))}

#Priors

for (j in 1:5){
b[j]~dnorm(0,0.001)

}
```

```
for (i in 1:n) { yf[i] ~ dbin(pi[i],e[i])}

prob[1]<-1000*exp(b[1])/(1+exp(b[1]))

for (k in 2:5){prob[k]<-1000*exp(b[1]+b[k])/(1+exp(b[1]+b[k]))}

}
```

eje.txt

```
model{
#Likelihood
for (i in 1:n){
y[i]~dbin(pi[i],e[i])

logit(pi[i])<-b[1]+b[2]*x50[i]+b[3]*x100[i]+b[4]*x200[i]+b[5]*x500[i]


}


p[1]<-1000*exp(b[1])/(1+exp(b[1]))

for (k in 2:5){p[k]<-1000*exp(b[1]+b[k])/(1+exp(b[1]+b[k]))}

#Priors

for (j in 1:5){
b[j]~dnorm(0,0.001)

}


for (i in 1:n) { yf[i] ~ dbin(pi[i],e[i])}

prob[1]<-1000*exp(b[1])/(1+exp(b[1]))

for (k in 2:5){prob[k]<-1000*exp(b[1]+b[k])/(1+exp(b[1]+b[k]))}

}
```

Anexo codigo R:

```
library(R2OpenBUGS)
library(R2jags)

wdir<-"/home/abraham/RA2018/ejemex/ex2018"
setwd(wdir)

#--- Funciones utiles ---
prob<-function(x){
  out<-min(length(x[x>0])/length(x),length(x[x<0])/length(x))
  out
}

#Leemos datos
bill<-read.table("http://allman.rhon.itam.mx/~lnieto/index_archivos/BillsMXc.csv",sep=",",header=TRUE)
bill
x<-bill$C#/1000
y<-bill$Y#/1000

library(tidyverse)
library(ggplot2)


#a)*********************************************************************************************

datan<-gather(bill,val,I,x20:x500) %>%filter(.,I==1) %>%mutate(.,proporcion_falsos=Y/C)

explo <- ggplot(datan,aes(x=val,y=Y))+geom_boxplot()+xlab("val")+ylab("n")
exploanio <- ggplot(datan,aes(x=as.factor(Year),y=Y))+geom_boxplot()+xlab("anio")+ylab("n")
explogpo <- ggplot(datan,aes(x=val,y=Y))+geom_boxplot()+xlab("val")+ylab("n")+
  facet_wrap(~as.factor(Year), scale="free")



#plot(x,y,type = "p")
#plot(bill$Year,y)
#hist(bill$Year)
#cor(x,y)
```

```
#b)*****************************************************************************************
n<-nrow(bill)
#-Defining data-
#data<-list(x1=sal$X1,x2=sal$X2,x3=sal$X3)
data<-list("n"=n,"y"=y,"e"=bill$C,"x50"=bill$x50,"x100"=bill$x100,"x200"=bill$x200,"x500"=bill$x500)
#-Defining inits-
inits<-function(){list(b=rep(0,5),yf=rep(0,n))}
parameters<-c("b","yf","prob")

#OpenBUGS
ex1<-bugs(data,inits,parameters,model.file="eja.txt",
            n.iter=10000,n.chains=2,n.burnin=1000)




res<-function(func,out){
  ex1.sim<-func
  #out<-ex1.sim$sims.list


 for(i in 1:5){
  b<-out$b[,i]
  par(mfrow=c(1,1))
  plot(b,type="l",xlab = "beta")
  plot(cumsum(b)/(1:length(b)),type="l",xlab = "beta")
  hist(b,freq=FALSE, main = "beta")
  acf(b)
 }
  plot(out$b)

  #out.sum<-ex1.sim$summary
  #return(ex1.sim)

}

rex<-function(fun){return(fun)}

out<-rex(ex1)$sims.list

out.sum<-rex(ex1)$summary


res(ex1,out)


#Predictions
adj<-function(out.sum,y,a){
  out.yf<-out.sum[grep(a,rownames(out.sum)),]
  or<-order(x)
  ymin<-min(y,out.yf[,c(1,3,7)])
  ymax<-max(y,out.yf[,c(1,3,7)])
  par(mfrow=c(1,1))
  plot(x,y,ylim=c(ymin,ymax))
  lines(x[or],out.yf[or,1],lwd=2,col=2)
  lines(x[or],out.yf[or,3],lty=2,col=2)
  lines(x[or],out.yf[or,7],lty=2,col=2)

  plot(out.yf[1:n,1],y,type ="p")

}

adj(out.sum,y,"yf")

#DIC
met<-function(ex1.sim,out.sum){
  out.dic<-ex1.sim$DIC
  #out.dic<-ej4.sim$BUGSoutput$DIC
  print(out.dic)
  print(out.sum)
}

met(ex1,out.sum)

r2<-function(y,out.sum,a){
  out.yf<-out.sum[grep(a,rownames(out.sum)),]
  R2<-(cor(y,out.yf[1:n,1]))^2
  print(R2)
}

r2(y,out.sum,"yf")
```

```
#d)*************************************************************************************************
out.sum[grep("prob",rownames(out.sum)),c(1,3,7)]



#e)*************************************************************************************************

parameters<-c("b","yf","prob")
#OpenBUGS
ex2<-bugs(data,inits,parameters,model.file="eje.txt",
        n.iter=100000,n.chains=2,n.burnin=10000)
traceplot(ex2)


out2<-rex(ex2)$sims.list

out2.sum<-rex(ex2)$summary


res(ex2,out2)
#ajustes
adj(out2.sum,y,"yf")
#estimaciones y DIC
met(ex2,out2.sum)
#R2
r2(y,out2.sum,"yf")

out2.sum[grep("b",rownames(out2.sum)),c(1,3,7)]

#f)*************************************************************************************************


#g)*************************************************************************************************
out2.sum[grep("prob",rownames(out2.sum)),c(1,3,7)]


#h)*************************************************************************************************
#El DIC de la logistica es 1297 vs. 1317 de la liga cloglog  y sus R2 se parecen
#adj(out.sum,y,"yf")

out2.yf<-out2.sum[grep("yf",rownames(out2.sum)),]
yf<-out2.yf[1:n,1]
plot(yf,y,type ="p")
plot(yf,y,type ="p",col=bill$x20)
plot(yf,y,type ="p",col=bill$x50)
#i)*************************************************************************************************
out2.sum[grep("prob",rownames(out2.sum)),c(1,3,7)]
```