Análisis de Datos Categóricos

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

Semestre: *2017-II*

Tema: PC 3

Prof. Victor Giancarlo Sal Y Rosas

Pontificia Universidad Católica del Perú
Escuela de Posgrado
Maestría en Estadística

Ejercicio 1 (Teoría)

Modelo:

$$Y_{ij} \quad = \quad \beta_0 + \beta_1 X + b_i + \epsilon_{ij} \tag{1}$$

donde b_i y ε_{ij} son independientes y

$$b_i \sim N\left(0, \sigma_b^2\right) \quad , \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right) \tag{2}$$

a) Veamos que $Var[Y_{ij}] = \sigma^2 + \sigma_b^2$:

$$\begin{aligned} Var[Y_{ij}] &= Var[\beta_0 + \beta_1 X + b_i + \epsilon_{ij}] \\ &= Var[b_i + \varepsilon_{ij}] \\ &\overset{\text{por indep}}{=} Var[b_i] + Var[\epsilon_{ij}] \\ &= \sigma_b^2 + \sigma^2 \end{aligned} \tag{3}$$

b) Veamos que $Cor[Y_{ij}, Y_{ik}] = \dfrac{\sigma_b^2}{\sigma^2 + \sigma_b^2}$ para $j \neq k$:

$$\begin{aligned} Cor[Y_{ij}, Y_{ik}] &= \frac{Cov[Y_{ij}, Y_{ik}]}{\sqrt{Var[Y_{ij}]}\sqrt{Var[Y_{ik}]}} \\ &\overset{\text{por (3)}}{=} \frac{Cov[\beta_0 + \beta_1 X + b_i + \epsilon_{ij}, \beta_0 + \beta_1 X + b_i + \epsilon_{ik}]}{\sigma_b^2 + \sigma^2} \\ &= \frac{Cov[b_i + \epsilon_{ij}, b_i + \epsilon_{ik}]}{\sigma_b^2 + \sigma^2} \\ &= \frac{Var[b_i] + Cov(b_i, \epsilon_{ik}) + Cov(\epsilon_{ij}, b_i) + Cov(\epsilon_{ij}, \epsilon_{ik})}{\sigma_b^2 + \sigma^2} \\ &\overset{\text{por indep}}{=} \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2} \end{aligned} \tag{4}$$

Ejercicio 2 (Simulación e interpretación)

Se desea hacer un estudio donde se desea corroborar el efecto de una vacuna para prevenir la recaida en abuso de drogas. Variables:

- **X**: (1=recibir vacuna con el medicamento en evaluación, 1=recibir vacuna placebo).
- **Y**: Respuesta (1=recae en drogas, 0=no recae).

Se asume el siguiente modelo:

$$log\left[\frac{P(Y_{ij=1} \mid X_{ij})}{P(Y_{ij=0} \mid X_{ij})}\right] = \beta_0 + \beta_1 X_{ij} + b_i \tag{5}$$

donde $b_i \sim N(0, 9)$ y $(\beta_0, \beta_1) = (-1, -0.69)$

a) Interpretación de $beta_1$:
Si la persona i-ésima recibe vacuna con el medicamento en evaluación, su odds de reacer en drogas disminuye en $49.8\,\%$ ($e^{-0.69} = 0.502$) con respecto al odds de no recaer, a diferencia de que si tal persona recibiera la vacuna de placebo.

b) Se simulará 1000 bases de datos con 200 participantes cada uno, usando el modelo propuesto.

Listing 1: Simulación.

```
beta1.estimado.L = numeric()
beta1.estimado.LM = numeric()
p=0.5
n=c(1000,200,5)
beta0.media=0
beta0.s=3
beta=c(-1,-0.69)


for(i in 1:n[1]){
        X = rep(rbinom(n[2],1,p),n[3])
        dataset = data.frame(dataset,
        Y=rbinom(n[3]*n[2], size = 1,
        prob = 1/(1+exp(-(beta[0]+beta[1]*X+rep(rnorm(n[2],beta0.media,beta0.s),n
            [3])))))),
        participante=rep(1:n[2],n[3]))

        #Implementacion para calculo de beta1 estimado
        beta1.estimado.L[i] <- beta1.f.L(dataset,X,Y)
        beta1.estimado.LM[i] <- beta1.f.LM(dataset,X,Y)
}
```

c) Se estimará β_1 mediante la simulación realizada en item b)

Listing 2: Función para estimación de β_1 (Modelo Logístico Simple).

```
beta1.f.L <- function(dataset,X,Y){
        #beta1 estimado por modelo logistico simple.
        mod.GLM = glm(Y ~ X, data=dataset,family=binomial)$coef["X"]
        return (mod.GLM$coef["X"])
}
```

Listing 3: Función para estimación de β_1 (Modelo Logístico de efectos mixtos).

```
beta1.f.LM <- function(dataset,X,Y){
        #beta1 estimado por modelo con efectos mixtos.
        mod.GLMER = unique(coef(glmer(Y ~ X + (1|participante),
        data=X,family=binomial))$participante$X)
        return(unique(coef(mod.GLMER)$participante$X)
}
```

Listing 4: Cálculo de $beta_1$ estimado por ambos modelos.

```
beta1.estimado=c(mean(beta1.f.L[1:n[1]]),mean(beta1.f.LM[1:n[1]]))
```

```
> beta1.estimado
[1] -0.3132479 -0.6945377
```

Figura 1: β_1 estimado

Se obtiene un $\beta_1 = -0.3132479$ estimado mediante el modelo logístico simple, a diferencia de un $\beta_1 = -0.6945377$ estimado mediante el modelo logístico de efectos mixtos siendo este último quien mejor estima al valor verdadero.

Ejercicio 3 (Análisis de datos)

Se necesita evaluar dos tratamientos orales para infección de la uña del pie. Variables:

- **trt**: Tratamiento (0=Itraconazole, 1=Terbinafine).
- **y**: Respuesta (0=No o suave, 1=moderado o severo).
- **month**: Tiempo exacto de la medidad en meses.
- **visit**: (1-7) correspondiente a las 0,4,8,12,24,36 y 48 semanas.

Listing 5: Lectura de datos.

```
library(dplyr)
library(ggplot2)
library(lme4)
library(nlme)
library(foreign)
library(geeM)

dataset <- read.dta("http://www.hsph.harvard.edu/fitzmaur/ala2e/toenail.dta")
dataset<-data.frame(dataset)
colnames(dataset)<-c("persona","grado","tratamiento","mes","visita")
head(dataset)
```

```
> head(dataset)
  persona grado tratamiento       mes visita
1       1     1           1 0.0000000      1
2       1     1           1 0.8571429      2
3       1     1           1 3.5357144      3
4       1     0           1 4.5357141      4
5       1     0           1 7.5357141      5
6       1     0           1 10.0357141     6
```

Figura 2: Datos de estudio

Como se observa de tratan de datos cuya variable respuesta ha sido medida varias veces en una misma persona.

a) Se realizará una gráfica del tiempo medido en visitas programadas versus la proporción de buena respuesta, es decir en la que el resultado fue “No” o “Suave”.

Listing 6: Número de visita vs Proproción de buena respuesta.

```
summary.toenail<- dataset %>%
                  group_by(tratamiento,visita)  %>%
                  summarize(mean.y = 1-mean(grado,na.rm=T))
summary.toenail

ggplot(summary.toenail, aes(x=visita,y=mean.y,group=as.factor(tratamiento),
       color=as.factor(tratamiento))) +
       geom_point() +
       geom_smooth(se=FALSE,method=lm)+
       labs(x = "Visitas",y=" Proporcion de respuestas No o Suave",size=18) +
       coord_cartesian(ylim = c(0, 1))  +
       theme(legend.position = "top") +
       scale_colour_discrete(breaks=c("0","1"),
       labels=c("Itraconazole","Terbinafine"),
       name="Tratamiento")
```

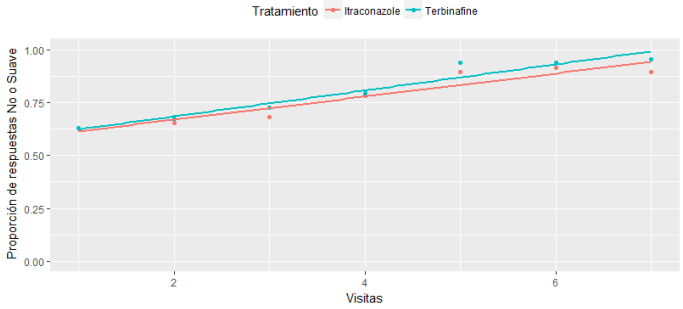

Figura 3: Visitas vs Proporción de respuestas

Se puede observar que la buena respuesta en ambos tratamientos al parecer sigue una misma tendencia lineal en el tiempo o al menos parece no ser significante como difieren éstas, por lo que podemos intuir un buen modelo sin interacción entre las variables Tratamiento y Tiempo. Además, podemos observar que las proporciones de buena respuesta por ambos tratamientos parecen ser iguales, intuyendo que no existe diferencia entre tratamientos. Lo analizaremos a continuación.

b) Se analizarán modelos y se optará por el más adecuado con la observación que debido a la correlación entre meses y visitas, solo usaremos una de aquellas variables.

Listing 7: Modelo aditivo (Sin interacción).

```
model.1a = glm(grado~tratamiento+visita,data=dataset,family=binomial(link="logit"
    ))
summary(model.1a) #Intercepto con poca significancia | AIC=1822.2

model.1b = glm(grado~0+tratamiento+visita,data=dataset,family=binomial(link="
    logit"))
summary(model.1b)  # Supuesto: beta0=0 | AIC=1821.1 <----------MENOR AIC

dataset2=dataset
dataset2$pred1 = predict.glm(model.1b,new.data=dataset,type="response")
dataset2 %>%
        group_by(tratamiento,visita)  %>%
        summarize(oprob = mean(grado,na.rm=T),
        prpob = mean(pred1,na.rm=T))
```

```
Call:
glm(formula = grado ~ 0 + tratamiento + visita, family = binomial(link = "logit$
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0288  -0.7206  -0.5192  -0.3684   2.3342

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
tratamiento  -0.1353     0.1042  -1.298    0.194
visita       -0.3602     0.0203 -17.743   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2645.0  on 1908  degrees of freedom
Residual deviance: 1817.1  on 1906  degrees of freedom
AIC: 1821.1

Number of Fisher Scoring iterations: 4
```

Figura 4: Modelo sin interacción

Listing 8: Modelo aditivo (Con interacción).

```r
model.1c = glm(grado~tratamiento+tratamiento*visita,data=dataset,family=binomial(
    link="logit"))
summary(model.1c)
# AIC=1821.6

model.1d = glm(grado~0+tratamiento+tratamiento*visita,data=dataset,family=
    binomial(link="logit"))
summary(model.1d)
# AIC=1819.6 <--------------MENOR AIC

dataset2$pred2 = predict.glm(model.1d,new.data=dataset,type="response")
dataset2 %>%
group_by(tratamiento,visita)  %>%
summarize(oprob = mean(grado,na.rm=T),
prpob1 = mean(pred1,na.rm=T),
prpob2 = mean(pred2,na.rm=T))
```

```
Call:
glm(formula = grado ~ 0 + tratamiento + tratamiento * visita,
    family = binomial(link = "logit"), data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0393  -0.7209  -0.4910  -0.3152   2.4604

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
tratamiento        0.10675    0.16635   0.642   0.5211
visita            -0.34245    0.02189 -15.643   <2e-16 ***
tratamiento:visita -0.09811    0.05347  -1.835   0.0665 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2645.0  on 1908  degrees of freedom
Residual deviance: 1813.6  on 1905  degrees of freedom
AIC: 1819.6

Number of Fisher Scoring iterations: 5
```

Figura 5: Modelo con interacción

Observamos que el coeficiente respecto a la interacción entre tratamiento y visita posee un p-valor cercano a 0.05.

```
# A tibble: 14 x 5
# Groups:   tratamiento [?]
   tratamiento visita      oprob     prpob1     prpob2
         <dbl>  <dbl>      <dbl>      <dbl>      <dbl>
 1           0      1 0.36986301 0.41091648 0.41521400
 2           0      2 0.34751773 0.32731461 0.33516769
 3           0      3 0.31884058 0.25340499 0.26359706
 4           0      4 0.21969697 0.19143513 0.20265091
 5           0      5 0.10769231 0.14174278 0.15287085
 6           0      6 0.08547009 0.10330150 0.11357716
 7           0      7 0.10526316 0.07438213 0.08338924
 8           1      1 0.37162162 0.37861502 0.41731338
 9           1      2 0.32653061 0.29825759 0.31553328
10           1      3 0.27586207 0.22867873 0.22882820
11           1      4 0.20714286 0.17136772 0.16036652
12           1      5 0.06015038 0.12607222 0.10947964
13           1      6 0.06299213 0.09142813 0.07332976
14           1      7 0.04580153 0.06558959 0.04846682
```

Figura 6: Comparación entre lo observado y lo predicho por los modelos sin interacción y con interacción

Se puede observar que ambos modelos producen valores muy cercanos a la probabilidad de respuesta buena, y esto se debe a causa de la poca interacción analizada en la gráfica de visita vs respuesta buena. Sin embargo, se optará por el modelo con interacción a causa de tener un AIC menor.

Interpretación (Modelo aditivo con interacción):

- El odds de tener una respuesta "No o suave" a la n-visita fue $e^{-0.10675+0.09811*n}$ veces el odds de tener una respuesta "Moderado o severo" en aquellos que recibieron el tratamiento de Terbinafine que aquellos que recibieron Itraconazole. Se puede analizar que tales efectos tienden a ser mayores a 1 desde la visita 2; lo cual deja conforme debido a que según la gráfica el tratamiento de Terbinafine es relativamente mejor.

Se diseñará un modelo para el caso multivariado por cada cluster (persona).

Listing 9: Modelo GEE.

```
model.2b = geem(grado ~ tratamiento+tratamiento*visita,id=persona, data=as.data.
    frame(dataset),
family=binomial(link="logit"), corstr="exchangeable" )
summary(model.2b)
```

```
                   Estimates Model SE Robust SE    wald      p
(Intercept)         -0.02743  0.14890   0.21400 -0.1282 0.8980
tratamiento          0.17370  0.20760   0.31690  0.5480 0.5837
visita              -0.33010  0.03456   0.04724 -6.9880 0.0000
tratamiento:visita  -0.10630  0.05307   0.07225 -1.4720 0.1411

 Estimated Correlation Parameter:  0.44
 Correlation Structure:  exchangeable
 Est. Scale Parameter:  0.9652

 Number of GEE iterations: 4
 Number of Clusters:  294    Maximum Cluster Size:  7
 Number of observations with nonzero weight:  1908
```

Figura 7: Modelo GEE

Interpretación (Modelo GEE):

- El odds de que una persona tenga una respuesta "No o suave" a la n-visita fue $e^{-0.17370+0.10630*n}$ veces el odds de tener una respuesta "Moderado o severo" en aquellos que recibieron el tratamiento de Terbinafine que aquellos que recibieron Itraconazole. Se puede analizar que tales efectos tienden a ser mayores a 1 desde la visita 2.
- Se observa que no todas las observaciones son independientes pues se encuentra una correlación indistinta de cero, de 0.44.

Se diseñará un modelo mixto generalizado

Listing 10: Modelo GMLER.

```
model.3 = glmer(grado ~ tratamiento+tratamiento*visita+(1|persona), data=as.data.
    frame(dataset),
family=binomial(link="logit"))
summary(model.3)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: grado ~ tratamiento + tratamiento * visita + (1 | persona)
   Data: as.data.frame(dataset)

     AIC      BIC   logLik deviance df.resid
  1258.8   1286.5   -624.4   1248.8     1903

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2199 -0.1739 -0.0529 -0.0077 23.4810

Random effects:
 Groups  Name        Variance Std.Dev.
 persona (Intercept) 22.39    4.732
Number of obs: 1908, groups:  persona, 294

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.42519    0.86564  -1.646   0.0997 .
tratamiento        -0.01012    0.78365  -0.013   0.9897
visita             -0.80444    0.08923  -9.015   <2e-16 ***
tratamiento:visita -0.23512    0.12612  -1.864   0.0623 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) trtmnt visita
tratamiento -0.384
visita       0.046  0.329
tratmnt:vst  0.280 -0.455 -0.543
```

Figura 8: Modelo GLMER

Interpretación (Modelo GLMER):

- El odds de que una persona tenga una respuesta "No o suave" a la n-visita fue $e^{0.01012+0.23512*n}$ veces el odds de tener una respuesta "Moderado o severo" en aquellos que recibieron el tratamiento de Terbinafine que aquellos que recibieron Itraconazole. Se puede analizar que tales efectos tienden a ser mayores a 1 desde la visita 1. Lo cual indica un mejor tratamiento el de Terbinafine

c) Realizando análisis en varios modelos, se encuentra un p-valor¿0.05 ya sea en los coeficientes para el tratamiento o en el de interacción. por lo que tenemos que entender que se puede aceptar la hipótesis nula de $\beta_1 = 0$ (efecto de tratamiento) estableciendo un efecto nulo en los tratamientos.

- Modelo aditivo sin interacción: $p - valor = 0.194 > 0.05$ para β_1.
- Modelo aditivo con interacción: $p - valor = 0.5211 > 0.05$ para β_1, además de un efecto de interacción con $p - valor = 0.0665 > 0.05$.
- Modelo GEE: $p - valor = 0.5837 > 0.05$ para β_1, además de un efecto de interacción con $p - valor = 0.1411 > 0.05$.
- Modelo con Efectos Mixtos: $p - valor = 0.9897 > 0.05$ para β_1, además de un efecto de interacción con $p - valor = 0.0623 > 0.05$.

d) Si se tendrá que implementar el tiempo, en este caso nos referimos a la variable meses, en este caso la interacción con respecto a la variable visita es en un contexto distinto dado que existe correlación lineal entre ambas, dado que los meses son más exactos y las visitas representan intervalos en teoría. Por lo que se tendrá que generar un modelo que haga uso de la varianza generada entre aquellas variables y observar si se puede determinar un modelo adecuado.

e) La mejor forma de ingresar el tiempo será el siguiente

Listing 11: Modelo 2 GLMER.

```
model.4 = glmer(grado ~ tratamiento+tratamiento*mes+(visita|persona), data=as.
    data.frame(dataset),
family=binomial(link="logit"))
summary(model.4)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: grado ~ tratamiento + tratamiento * mes + (visita | persona)
   Data: as.data.frame(dataset)

     AIC      BIC   logLik deviance df.resid
   980.6   1019.5   -483.3    966.6     1901

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6162 -0.0017 -0.0001  0.0000  7.0046

Random effects:
 Groups  Name        Variance Std.Dev. Corr
 persona (Intercept) 2624.14  51.226
         visita        54.93   7.412   -0.90
Number of obs: 1908, groups:  persona, 294

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -11.0500     1.0783 -10.248  < 2e-16 ***
tratamiento      -0.3230     1.3656  -0.237  0.81301
mes              -1.3575     0.3611  -3.759  0.00017 ***
tratamiento:mes  -0.6990     0.3515  -1.989  0.04675 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) trtmnt mes
tratamiento -0.499
mes          0.144  0.182
tratamnt:ms  0.102 -0.249 -0.379
```

Figura 9: Modelo 2 GLMER

La implementación de meses adecua un mejor modelo con efectos mixtos, además de la adición de (visitas—personas) indicando que existirá cierta variabilidad intrínseca entre la variable visitas, dado que en efecto estamos tomando como si fuesen representaciones de intervalos de tiempo. El AIC es menor que todos los modelos antes vistos, además que logra determinarse un efecto de interacción no significativo ($p - valor = 0.04675 < 0.05$).

Pregunta 4 (Aplicación)

Estudio de esquizofrenia. Variables:

- **id**: Identificador de persona.
- **y**: Indicador de síntomas.
- **month**: Meses de hospitalización.
- **ages**: 0(<20 años) y 1($\geq$ 20 años).
- **sex**: 0(hombre) y 1(mujer).

Listing 12: Lectura de datos.

```
datos <-read.table("http://faculty.washington.edu/heagerty/Books/AnalysisLongitudinal/
    madras.data")
datos <- datos[,1:5]
names(datos) <- c("id","y","month","age","sex")
head(datos)
```

```
> head(datos)
  id y month age sex
1  1 1     0   0   0
2  1 1     1   0   0
3  1 1     2   0   0
4  1 1     3   0   0
5  1 1     4   0   0
6  1 0     5   0   0
```

Figura 10: Datos de estudio

a) Graficando evolución del paciente a lo largo del tiempo

Listing 13: Meses vs Proporción de mejora (Agrupados por Edad).

```
summary.datos<-datos %>%
                        group_by(age,month) %>%
                        summarize(mean.y=1-mean(y,na.rm=T))

summary.datos
ggplot(summary.datos,aes(x=month,y=mean.y,group=as.factor(age),
       color=as.factor(age))) +
       geom_point() +
       labs(x = "Meses",y=" Proporcion de mejora",size=18) +
       coord_cartesian(ylim = c(0, 1))  +
       theme(legend.position = "top") +
       scale_colour_discrete(breaks=c("0","1"),
       labels=c("<20 anos","=>20 anos"),
       name="Edad")
```

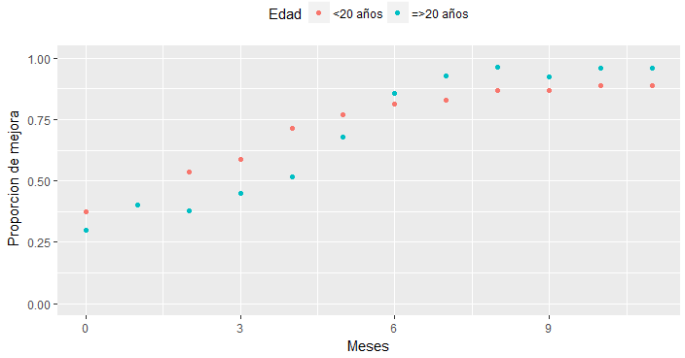

Figura 11: Meses vs Proporción de mejora (Agrupados por Edad)

Listing 14: Meses vs Proporción de mejora (Agrupados por Sexo).

```
summary.datos2<-datos %>%
        group_by(sex,month) %>%
        summarize(mean.y=1-mean(y,na.rm=T))

summary.datos2
ggplot(summary.datos2,aes(x=month,y=mean.y,group=as.factor(sex),
       color=as.factor(sex))) +
       geom_point() +
       labs(x = "Meses",y=" Proporcion de mejora",size=18) +
       coord_cartesian(ylim = c(0, 1))  +
       theme(legend.position = "top") +
       scale_colour_discrete(breaks=c("0","1"),
       labels=c("Hombre","Mujer"),
       name="Sexo")
```

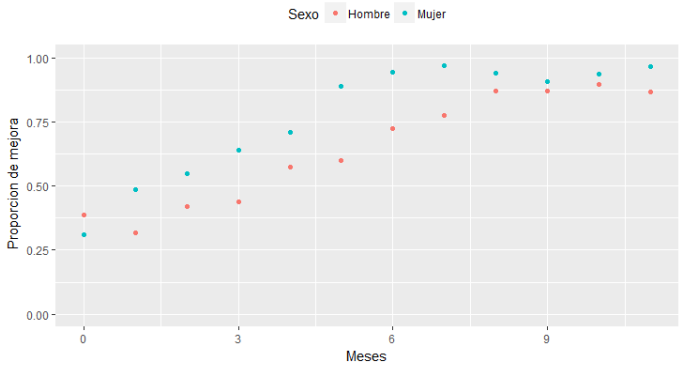

Figura 12: Meses vs Proporción de mejora (Agrupados por Sexo)

b) Analizando la gráfica en la que el agrupamiento es conforme a la edad, se observa que desde el mes inicial la proporción de mejora de los pacientes con edades menores que 20 son más altas que para las que tienen mayor o igual que 20, hasta el sexto mes en el que los papeles se invierten hasta el final. Esta es una evidencia fuerte de que pueden existir tendencias lineales distintas con respecto a los grupos de edades, dichas tendencias tendrían pendientes distintas para que cause el efecto del cambio mencionado.

Analizando la gráfica en la que el agrupamiento es conforme el sexo, no se observa una tendencia creciente aunque no tan lineal como se esperaría, por lo que tendría que evaluarse la interacción para ver la significancia respectiva; aunque si se puede observar que la proporción de mejora de los pacientes es por lo general más alta en las mujeres que en los hombres.