Cómputo Estadístico

September 15, 2024

Godinez Bravo Diego

Tarea 1 - Modelos Lineales Generalizados

Centro de Investigación en Matemáticas

Maestría en Cómputo Estadístico

0.1 Problema 1

La siguiente tabla muestra los resultados parciales de dos encuestas que forman parte de un estudio para evaluar el desempeño del Primer Ministro del Canadá. Se tomó una muestra aleatoria de 1600 ciudadanos canadienses mayores de edad y en los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después de la primera encuesta, sólo 880 aprueban su desempeño:

Primera encuesta	Y=1, Aprueba	Y=0 Desaprueba	Total
$x = 1$, Aprueba	794	150	944
$x = 0$, Desaprueba	86	570	656
Total	880	720	1600

Sea p_1 la proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial y sea p_2 la proporción correspondiente seis meses después. Considere la hipótesis $H_0 : p_1 = p_2$, ¿Cómo puede hacerse esta prueba.

0.1.1 Solución

Sean las proporciones:

- p_1: proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial.
- p_2: proporción de ciudadanos que aprueban el desempeño del ministro seis meses después.

Se considera la siguiente hipótesis nula H_0:

$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_a : p_1 \neq p_2,$

cuando H_0 es verdadera, esperamos valores similares para p_1 y p_2.

Sea $n^* = n_{12} + n_{21}$, donde n_{12} y n_{21} son el número de éxitos y fracasos para una distribución binomial con n^* ensayos. Cuando $n^* > 10$, la distribución binomial se asemeja a una distribución normal con media $\frac{1}{2}n^*$ y desviación estándar $\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}$.

El estadístico normal estandarizado se define de la siguiente manera:

$$Z = \frac{n_{12} - (\frac{1}{2})n^*}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

El cuadrado de este estadístico posee una distribución aproximada χ^2 con $df = 1$. La prueba χ^2 para la comparación de dos proporciones dependientes es denominada **McNemar test**.

En este caso:

$$Z = \frac{150 - 86}{\sqrt{150 + 86}} = 4.166$$

Por lo tanto:

$$P(Z > |4.166|) \approx 0.0000155$$

Dado que **p-valor** < 0.05 se rechaza la hipótesis nula H_0. Es decir, no podemos afirmar que las proporciones p_1 y p_2 tengan valores similares.

0.2 Problema 2

0.2.1 Solución

Se tiene la siguiente tabla donde se eligen varios niveles de ronquidos y se ponen en relacion con una enfermedad cardiaca. Se toman como puntuaciones relativas de ronquidos los valores $\{0, 2, 4, 5\}$.

Ronquido	Sí	No	Proporción de Sí
Nunca	24	1355	0.017
Ocasional	35	603	0.055
Casi cada noche	21	192	0.99
Cada noche	30	224	0.118

Ajuste un modelo lineal generalizado **logit y probit (investigar sobre el link probit)** para analizar si existe una relacion entre los ronquidos y la posibilidad de tener una enfermedad cardiaca.

[17]:
```
heart_disease_data <- data.frame(
  Snoring_Category = c("Never", "Occasional", "Nearly_Every_Night",␣
 ↪"Every_Night"),
  Yes = c(24, 35, 21, 30),
  No = c(1355, 603, 192, 224),
  Yes_Proportion = c(0.017, 0.055, 0.099, 0.118)
) # data stored in a data frame
```

Se añade una variable numérica para la categoria de ronquidos. Se agregó una nueva variable al dataframe para mapear la columna de datos categóricos *Categoría de Ronquidos* a valores numéricos.

[3]:
```
heart_disease_data$Snoring_No_Category <-␣
 ↪ifelse(heart_disease_data$Snoring_Category == "Never", 0,␣
 ↪ifelse(heart_disease_data$Snoring_Category == "Occasional", 2,
                                    ifelse(heart_disease_data$Snoring_Category␣
 ↪== "Nearly_Every_Night", 4, ifelse(heart_disease_data$Snoring_Category ==␣
 ↪"Every_Night", 5, NA)))) # adding a numerical variable for snoring frequency
heart_disease_data
```

A data.frame: 4×5

Snoring_Category <chr>	Yes <dbl>	No <dbl>	Yes_Proportion <dbl>	Snoring_No_Category <dbl>
Never	24	1355	0.017	0
Occasional	35	603	0.055	2
Nearly_Every_Night	21	192	0.099	4
Every_Night	30	224	0.118	5

0.2.2 Logit Link Function

Ajustar un modelo de regresión logística utilizando la funcion **logit**, utilizando ambas columnas *Sí* y *No* para representar correctamente la naturaleza binomial de los datos.

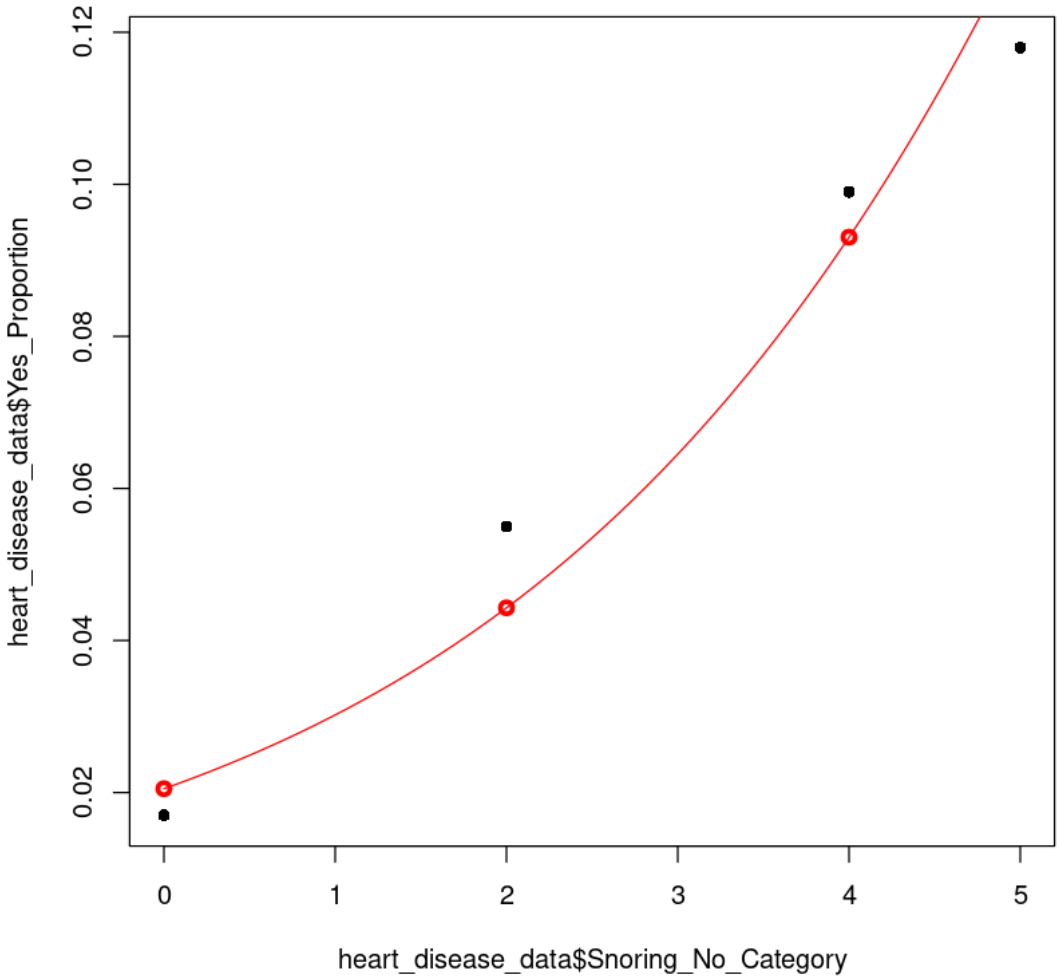
[4]:
```
logistic.model <- glm(cbind(Yes, No) ~ Snoring_No_Category, family =␣
↪'binomial'(link = logit), data = heart_disease_data)
summary(logistic.model)
```

```
Call:
glm(formula = cbind(Yes, No) ~ Snoring_No_Category, family = binomial(link =␣
↪logit),
    data = heart_disease_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.86625    0.16621 -23.261  < 2e-16 ***
Snoring_No_Category  0.39734    0.05001   7.945 1.94e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.9045  on 3  degrees of freedom
Residual deviance:  2.8089  on 2  degrees of freedom
AIC: 27.061

Number of Fisher Scoring iterations: 4
```

Categoría de Ronquidos: El coeficiente para la variable categoría de ronquidos es 0.6545 con un error estándar de 0.0825, y un **valor z** (7.931). Por otro lado, el **p-valor** ($2.18e-15$) para la variable indica que la frecuencia de ronquidos es un predictor estadísticamente significativo de la respuesta, en este caso la presencia de enfermedades cardíacas.

Nota: Los asteriscos *** indican que la variable es altamente **significativa** desde el punto de vista estadístico (**p-valor** < 0.001).

[5]:
```
plot(heart_disease_data$Snoring_No_Category, heart_disease_data$Yes_Proportion,␣
↪pch = 16, col = "black")
points(heart_disease_data$Snoring_No_Category, logistic.model$fitted.values,␣
↪col = "red", type = "p", lwd = 3)
curve(predict(logistic.model, data.frame(Snoring_No_Category = x), type =␣
↪"response"), add = TRUE, col = "red")
```



El coeficiente para la variable *categoría de ronquidos* es 0.6545, lo que significa que por cada incremento de una unidad en la *categoría de ronquidos*, los logaritmos de las probabilidades de tener una enfermedad cardíaca aumentan. En otras palabras, el valor del coeficiente sugiere que a medida que aumenta la frecuencia de ronquidos, las probabilidades de tener una enfermedad cardíaca aumentan de manera significativa. Este hallazgo apoya la idea de que la frecuencia de los ronquidos está asociada con un mayor riesgo de presentar una enfermedad cardíaca.

Dado que el p-valor es mucho menor que el nivel de significancia (0.001), hay evidencia sólida para afirmar que **existe una relación** entre la *categoría de ronquidos* y el *riesgo de tener una enfermedad cardíaca*.

0.2.3 Probit Link Function

Ajustar un modelo de regresión logística utilizando la funcion **probit**, utilizando ambas columnas *Sí* y *No* para representar correctamente la naturaleza binomial de los datos.

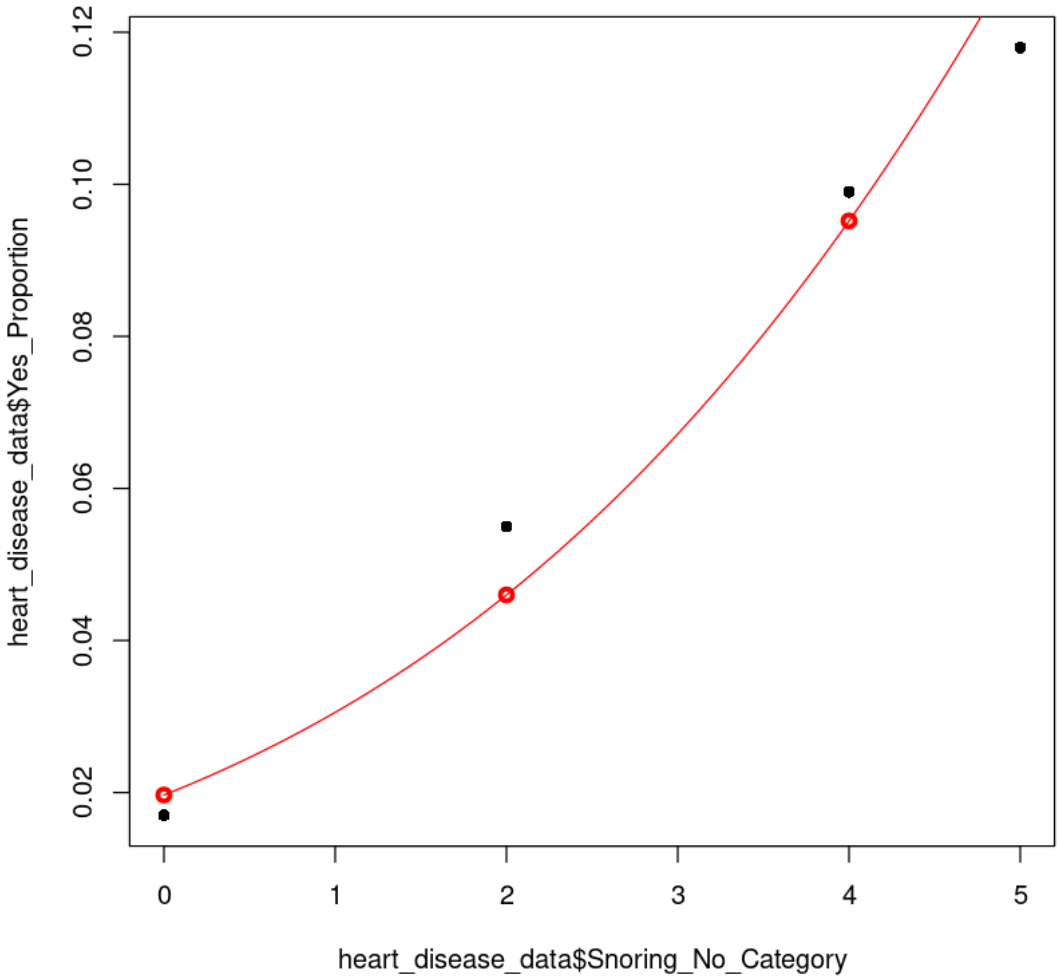
[6]:
```
logistic.model <- glm(cbind(Yes, No) ~ Snoring_No_Category, family =␣
↪'binomial'(link = probit), data = heart_disease_data)
summary(logistic.model)
```

```
Call:
glm(formula = cbind(Yes, No) ~ Snoring_No_Category, family = binomial(link =␣
↪probit),
    data = heart_disease_data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.06055    0.07017 -29.367  < 2e-16 ***
Snoring_No_Category  0.18777    0.02348   7.997 1.28e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.9045  on 3  degrees of freedom
Residual deviance:  1.8716  on 2  degrees of freedom
AIC: 26.124

Number of Fisher Scoring iterations: 4
```

Categoría de Ronquidos: El coeficiente para la variable *categoría de ronquidos* es 0.3157 con un error estándar de 0.0397 y un **valor z** (7.948). Al igual que en el caso anterior, el **p-valor** ($1.89e-15$) para la variable indica que la frecuencia de ronquidos es un predictor estadísticamente significativo de la respuesta (enfermedad cardíaca).

Como se mencionó anteriormente, *** indica que la variable es altamente significativa desde el punto de vista estadístico, en este caso **p-valor** < 0.001.

[7]:
```
plot(heart_disease_data$Snoring_No_Category, heart_disease_data$Yes_Proportion,␣
↪pch = 16, col = "black")
points(heart_disease_data$Snoring_No_Category, logistic.model$fitted.values,␣
↪col = "red", type = "p", lwd = 3)
curve(predict(logistic.model, data.frame(Snoring_No_Category = x), type =␣
↪"response"), add = TRUE, col = "red")
```



De manera similar con el modelo *logit*, el coeficiente para la variable *categoría de ronquidos* sugiere que a medida que aumenta la frecuencia de ronquidos, las probabilidades de tener una enfermedad cardíaca también aumentan. De manera que se mantiene la idea de que la frecuencia de ronquidos está asociado con un mayor riesgo de padecer una enfermedad cardíaca.

Dado que el p-valor es mucho menor que el nivel de significancia (0.001), podemos decir que el predictor es altamente **significativo** desde el punto de vista estadístico.

0.2.4 Comparación de los Modelos

El Criterio de Información de Akaike (AIC) es una métrica para evaluar la calidad del modelo en términos del ajuste de bondad y complejidad del modelo. Un valor de AIC más bajo indica un mejor ajuste. Usando la función **probit**, el valor de AIC obtenido fue 29.026 en comparación con el valor obtenido usando la función **logit** (30.492). Esto sugiere que el modelo utilizando la función

probit es **ligeramente mejor**.

La desviación residual mide qué tan bien se ajusta el modelo a los datos; una menor desviación indica un mejor ajuste. El modelo usando la función **probit** tiene una desviación residual más baja (4.7733) en comparacion con el modelo utilizando la función **logit** (6.2398), lo que sugiere que el modelo **probit** proporciona un ajuste ligeramente mejor.

Ambos modelos muestran que el predictor (*categoría de ronquidos*) es estadísticamente significativo con p-valores menores a 0.001. La diferencia entre ellos es mínima, ambas funciones **proporcionan resultados similares**.

0.3 Problema 3

0.3.1 Solución

Entre los cangrejos se sabe que cada hembra tiene un macho en su nido, pero puede tener mas machos concubinos. Se consiera que la variable de respuesta es el numero de concubinos y las variables explicativas son: color, estado de la espina central, peso y anchura del caparazon.

Color	Spine	Width	Satellite	Weight
3	3	28.3	8	3050
4	3	22.5	0	1550
2	1	26.0	9	2300
4	3	24.8	0	2100
4	3	26.0	4	2600
3	3	23.8	0	2100
2	1	26.5	0	2350

Realizar e interpretar los resultados de ajustar un modelo lineal generalizado tipo Poisson.

Hay $\mathbf{n = 7}$ observaciones y **5** variables en el conjunto de datos. Donde la variable *Satellite* es la variable de respuesta.

[16]:
```
crabs_data <- data.frame(
  Color = c(3, 4, 2, 4, 4, 3, 2),
  Spine = c(3, 3, 1, 3, 3, 3, 1),
  Width = c(28.3, 22.5, 26.0, 24.8, 26.0, 23.8, 26.5),
  Satellite = c(8, 0, 9, 0, 4, 0, 0),
  Weight = c(3050, 1550, 2300, 2100, 2600, 2100, 2350)
) # data stored in a data frame
```

Ajustar un modelo de regresión Poisson con la variable `Satellite` como la variable de respuesta, y tanto el ancho, altura y color de la hembra (`Width`, `Weight`, `Color`) como las variables explicativas.

[107]:
```
poisson.model <- glm(Satellite ~ Width + Weight + Color, family =␣
  ↪'poisson'(link = log), data = crabs_data)
summary(poisson.model)
```

```
Call:
glm(formula = Satellite ~ Width + Weight + Color, family = poisson(link = log),
    data = crabs_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.712413  13.619198   0.419    0.675
Width       -0.432511   0.740442  -0.584    0.559
Weight       0.003794   0.003005   1.263    0.207
Color       -0.935272   0.627179  -1.491    0.136
```

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

    Null deviance: 37.770  on 6  degrees of freedom
Residual deviance: 19.536  on 3  degrees of freedom
AIC: 38.794

Number of Fisher Scoring iterations: 7
```

Ancho: El coeficiente para la variable *Width* es de -0.4325 con un error estándar de 0.7404. El **p-valor** (0.559) indica que el efecto de la variable no es estadisticamente significativo, es decir, que no podemos afirmar que la variable *Width* influye en el numero de satelites (variable *Satellite*).

Peso: El valor de coeficiente de la variable *Weight* es de 0.003794 con un error estándar de 0.0030. Al igual que para la variable *Width*, el **p-valor** nos indica que el efecto del peso no es significativo, desde el punto de vista estadistico, es decir **p-valor** > 0.05.

Color: El coeficiente para la variable *Color* es de -0.9352 con un error estándar de 0.6271. Basándonos en el **p-valor** podemos afirmar que el efecto de esta variable hacia la respuesta (*Satellite*) no es significativo.

Ajustar un modelo de regresión Poisson con la variable `Satellite` como la variable de respuesta, y el ancho y altura y de la hembra (`Width`, `Weight`) como las variables explicativas.

[108]:
```
poisson.model <- glm(Satellite ~ Width + Weight, family = 'poisson'(link =␣
 ↪log), data = crabs_data)
summary(poisson.model)
```

```
Call:
glm(formula = Satellite ~ Width + Weight, family = poisson(link = log),
    data = crabs_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.395e+01  8.480e+00  -1.645    0.100
Width        5.945e-01  4.727e-01   1.258    0.209
Weight      -2.114e-04  1.736e-03  -0.122    0.903

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37.770  on 6  degrees of freedom
Residual deviance: 22.166  on 4  degrees of freedom
AIC: 39.423

Number of Fisher Scoring iterations: 6
```

Ancho: El coeficiente para la variable *Width* es de $5.945e-01$ con un error estándar de $4.727e-01$. El **p-valor** (0.209) nos indica que, al igual que el modelo anterior, el efecto de la variable no es estadisticamente significativo (**p-valor** > 0.05).

Peso: El valor de coeficiente de la variable *Weight* es de $-2.114e-04$ con un error estándar de $1.736e-03$. Al igual que para la variable *Width*, el **p-valor** (-0.122) nos indica que el efecto de la variable no es significativo desde el punto de vista estadistico.

En este caso, considerando un modelo de regresion logistica con las variables predictoras `Width` y `Weight`, podemos decir que ninguna variables tienen un efecto significativo hacia la variable de respuesta `Satellite`, esto ya que sus **p-valores** > 0.05.

Modelo de regresión Poisson ajustado con la variable `Satellite` como la variable de respuesta y el ancho de la hembra (`width`) como la variable explicativa.

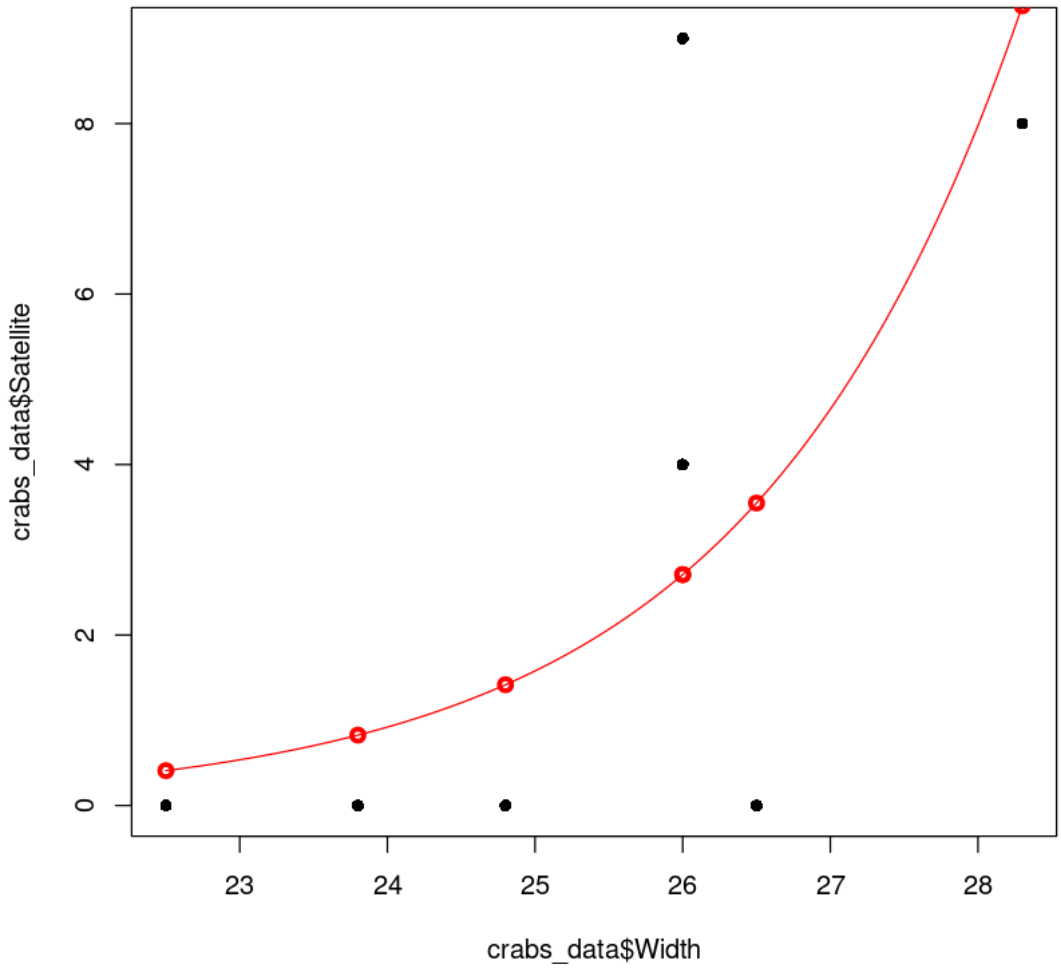
[2]:
```
poisson.model <- glm(Satellite ~ Width, family = 'poisson'(link = log), data =␣
↪crabs_data)
summary(poisson.model)
```

```
Call:
glm(formula = Satellite ~ Width, family = poisson(link = log),
    data = crabs_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.0435     3.9930  -3.267 0.001088 **
Width         0.5400     0.1483   3.640 0.000273 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37.770  on 6  degrees of freedom
Residual deviance: 22.181  on 5  degrees of freedom
AIC: 37.438

Number of Fisher Scoring iterations: 6
```

Ancho: El coeficiente para la variable *Width* es de 0.5400 con un error estándar de 0.1483. En este caso el valor z (3.640) y el **p-valor** (0.000273) indican que el coeficiente es altamente significativo, es decir, que la variable *Width* es un fuerte predictor de la variable *Satellite*.

Nota: Los asteriscos *** indican que la variable es altamente **significativa** desde el punto de vista estadístico (**p-valor** < 0.001).

[3]:
```
plot(crabs_data$Width, crabs_data$Satellite, pch = 16, col = "black")
points(crabs_data$Width, poisson.model$fitted.values, col = "red", type = "p",␣
↪lwd = 3)
curve(predict(poisson.model, data.frame(Width = x), type = "response"), add =␣
↪TRUE, col = "red")
```



El coeficiente para la variable *Width* es de 0.5400, lo que significa que por cada incremento de una unidad en la variable *Width*, el valor esperado del logaritmo de la variable *Satellite* aumenta. El p-valor (0.000273) indica que este coeficiente es altamente significativo desde el punto de vista estadístico (p-valor < 0.001).

De manera general, dado que el p-valor es mucho menor que el nivel de significancia (0.001), podemos afirmar que existe una **relación significativa** entre la variable *Width* y el número de la variable *Satellite*. El coeficiente positivo (0.5400) sugiere que a medida que aumenta la variable

Width, también aumenta el recuento esperado de la variable *Satellite*.

0.3.2 Comparación de los Modelos

De todos los modelos propuestos: Modelo 1 - `Satelittle ~ Width + Weight + Color`, Modelo 2 - `Satelittle ~ Width + Weight`, y Modelo 3 - `Satelittle ~ Width`; el que tiene un mejor ajuste es el **Modelo 3**.

Este modelo muestra que el efecto de la variable `Width` sobre la variable `Satellite` es **estadísticamente significativo**. Además, al comparar los valores de AIC entre los modelos, observamos que el más bajo corresponde al modelo 3. Esto respalda lo mencionado anteriormente: el modelo 3, que utiliza únicamente la variable Width como predictor, ofrece el mejor ajuste. Los otros modelos presentan un incremento en el AIC, lo que sugiere un aumento innecesario en la complejidad del modelo.

0.4 Problema 4

Suponga $(x_1, y_1), \ldots, (x_n, y_n)$ observaciones independientes de variables aleatorias definidas como sigue:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p), \quad i = 1, \cdots, n \\ X_i \mid \{Y_i = 1\} &\sim N(\mu_1, \sigma^2) \\ X_i \mid \{Y_i = 0\} &\sim N(\mu_0, \sigma^2) \end{aligned}$$

Usando el Teorema de Bayes, muestre que $\mathrm{P}(Y_i = 1 \mid X_i)$ satisface el modelo de regresión logística, esto es

$$\text{logit}\left(P(Y_i = 1 \mid X_i)\right) = \beta_0 + \beta_i X_i$$

con $\beta_i = (\mu_1 - \mu_0)/\sigma^2$.

0.4.1 Solución

Demostrar que la probabilidad condicional ($\mathrm{P}(Y_i = 1 \mid X_i)$ sigue un modelo de **regresión logística**, i.e.

$$\text{logit}\left(\mathrm{P}(Y_i = 1 \mid X_i)\right) = \beta_0 + \beta_i X_i,$$

con $\beta_i = \frac{\mu_1 - \mu_0}{\sigma^2}$.

Recordando el **Teorema de Bayes**:

$P(Y_i = 1|X_i) = \frac{P(X_i|Y_i=1)P(Y_i)=1}{P(X_i)}$ para $Y_i = 1$

y

$P(Y_i = 0|X_i) = \frac{P(X_i|Y_i=0)P(Y_i)=0}{P(X_i)}$ para $Y_i = 0$

Por lo tanto:

$$\frac{P(Y_i = 1 \mid X_i)}{P(Y_i = 0 \mid X_i)} = \frac{P(X_i \mid Y_i = 1)P(Y_i = 1)}{P(X_i \mid Y_i = 0)P(Y_i = 0)}$$

Sea $P(Y_i = 1) = p$ y $P(Y_i = 0) = 1 - p$, entonces:

$$\frac{P(Y_i = 1 \mid X_i)}{P(Y_i = 0 \mid X_i)} = \frac{P(X_i \mid Y_i = 1) \cdot p}{P(X_i \mid Y_i = 0)(1 - p)}$$

Dado que:

$X_i|Y_i = 1 \sim N(\mu_1, \sigma^2)$, y $X_i|Y_i = 0 \sim N(\mu_0, \sigma^2)$

Entonces:

$P(X_i|Y_i = 1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu_1)^2}{2\sigma^2}\right)$,

y

$P(X_i|Y_i = 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i-\mu_0)^2}{2\sigma^2}\right)$

De manera que:

$$\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu_1)^2}{2\sigma^2}\right)p}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(X_i-\mu_0)^2}{2\sigma^2}\right)(1-p)}$$

Por propiedades de los exponentes:

$$\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)} = \frac{p}{(1-p)}\exp\left(-\frac{(X_i-\mu_1)^2-(X_i-\mu_0)^2}{2\sigma^2}\right)$$

Aplicando logaritmo:

$$\begin{aligned}
\log\left(\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}\right) &= \log\left(\frac{p}{1-p}\cdot\exp\left(-\frac{(X_i-\mu_1)^2-(X_i-\mu_0)^2}{2\sigma^2}\right)\right) \\
&= \log\left(\frac{p}{1-p}\right) + \log\left(e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}}\right) - \log\left(e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}}\right) \\
&= \log\left(\frac{p}{1-p}\right) - \frac{(x_i-\mu_1)^2}{2\sigma^2} + \frac{(x_i-\mu_0)^2}{2\sigma^2} \\
&= \log\left(\frac{p}{1-p}\right) - \frac{1}{2}\left(\frac{x_i^2-2x_i\mu_1+\mu_1^2}{\sigma^2}\right) + \frac{1}{2}\left(\frac{x_i^2-2x_i\mu_0+\mu_0^2}{\sigma^2}\right) \\
&= \log\left(\frac{p}{1-p}\right) - \frac{x_i^2}{2\sigma^2} + \frac{x_i\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \frac{x_i^2}{2\sigma^2} - \frac{x_i\mu_0}{\sigma^2} + \frac{\mu_0^2}{2\sigma^2} \\
&= \log\left(\frac{p}{1-p}\right) + \frac{x_i\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} - \frac{x_i\mu_0}{\sigma^2} + \frac{\mu_0^2}{2\sigma^2} \\
&= \log\left(\frac{p}{1-p}\right) + \frac{1}{2}\left(\frac{\mu_0^2-\mu_1^2}{\sigma^2}\right) + \frac{(\mu_1-\mu_0)}{\sigma^2}x_i
\end{aligned}$$

Sea $\beta_0 = \log\left(\frac{p}{1-p}\right) + \frac{1}{2}\left(\frac{\mu_0^2-\mu_1^2}{\sigma^2}\right)$ y $\beta_i = \frac{\mu_1-\mu_0}{\sigma^2}$, finalmente:

$$= \beta_0 + \beta_i x_i$$

0.5 Problema 5

Cuando usamos un modelo de regresion logistica para clasificacion, tenemos que definir el umbral p, a partir del cual declaramos un positivo.

Las curvas ROC grafican las tasas TPR vs FPR para diferentes umbrales p.

$TPR = True\ Positive\ Rate = \frac{TP}{P} =$ sensitividad

$FPR = False\ Positive\ Rate = \frac{FP}{N} = 1 -$ especificidad

- La grafica de TPR vs FPR puede interpretarse como una grafica de "poder" vs. "error tipo I".
- Idealmente, una regla de decision estaria en el punto (0, 1).
- El area bajo la curva, AUC, puede verse, es la probabilidad de que un individuo de los positivos, tomado al azar, tenga un riesgo estimado mayor que un individuo de los negativos, tomado al azar.
- El estadistico J de Youden, es una medida que, con un solo numero, trata de capturar el desempeño de una prueba de diagnostico. Es la maxima distancia vertical, entre la diagonal y la curva ROC, o equivalentemente:

$J =$ sensitividad $- (1 -$ especificidad$)$

Construyan la curva ROC para el problema de daño coronario y su relacion con la edad.

0.5.1 Solución

[23]:
```
library(pROC) # libraries for drawing ROC curves
library(ROCR)
```

[24]:
```
library(ggplot2)
library(caret)
```

Datos correspondientes a la edad y su relacion con el daño coronario.

[13]:
```
age <- c(20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32,␣
↪32, 33, 33,
       34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39,␣
↪40, 40, 41,
       41, 42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47,␣
↪47, 47, 48,
       48, 48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56,␣
↪56, 56, 57,
       57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64,␣
↪64, 65, 69)

coro <-␣
↪c(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,0,1,0,1,0,
       ␣
↪0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,0,1,1,0,0,1,0,1,0,0,1,1,1,1,0,1,1,1,1,1,0,
        0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,0,1,1,1)
```

Modelo de regresión logistica ajustado con la variable `coro` como la variable de respuesta y la edad (`age`) como la variable explicativa.

[26]:
```
logistic.model <- glm(coro ~ age, family = 'binomial'(link = probit))
summary(logistic.model)
```

```
Call:
glm(formula = coro ~ age, family = binomial(link = probit))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.14573    0.62460  -5.036 4.74e-07 ***
age          0.06580    0.01335   4.930 8.20e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.50  on 98  degrees of freedom
AIC: 111.5

Number of Fisher Scoring iterations: 4
```

Edad: El coeficiente de la variable *age* es de 0.06580, y dado que el **p-valor** es pequeño ($8.20e-07$), podemos afirmar que la variable es estadisticamente significativa al momento de predecir la respuesta (i.e. problema de daño coronario).

Nota: Los asteriscos *** indican que la variable es estadisticamente **significativa** considerando un nivel de significancia del 0.001 (**p-valor** < 0.001).

Considerando un nivel de significancia del 0.001 (p-valor < 0.001), podemos concluir que la variable *age* es estadisticamente significativa. Por lo tanto, la variable *age* tiene una **relacion estadisticamente significativa** con la variable de respuesta, es decir, la edad tiene relación directa con la presencia de daño coronario.

[27]:
```
predictions <- predict(logistic.model, type = 'response')
pred <- prediction(predictions, coro) # prediction object
pred
```

```
A prediction instance
  with 100 data points
```

[28]:
```
perf <- performance(pred, "tpr", "fpr") # performance object; performance␣
 ↪measures for the evaluation: 1) true positive rate (y-axis), and 2) false␣
 ↪positive rate (x-axis)
perf
```

```
A performance instance
  'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
  with 44 data points
```

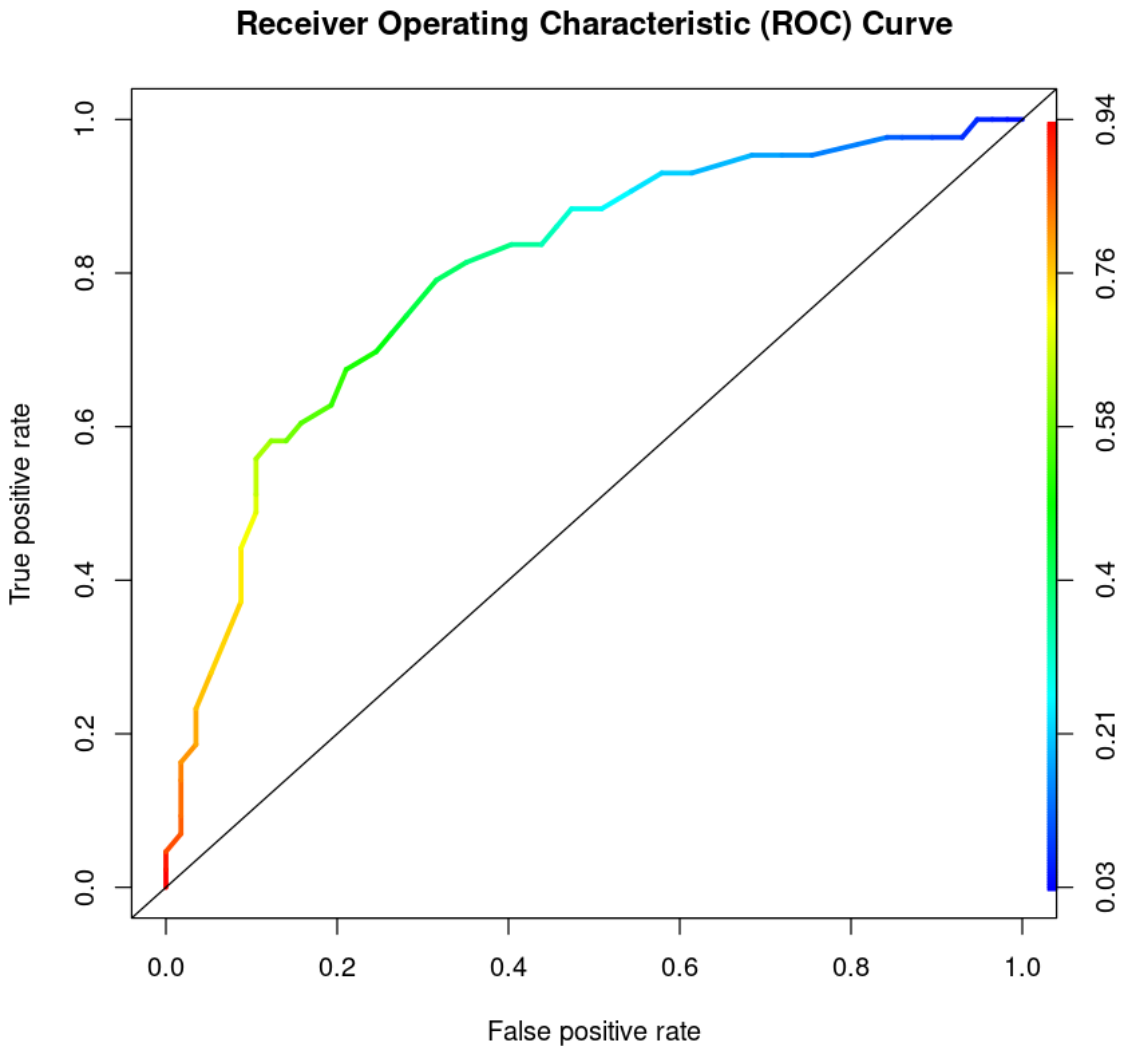
El **área bajo la curva ROC** es una medida que representa la habilidad del modelo para discriminar entre dos clases. Un modelo con un AUC de 1.0 indicaria una perfecta distinción entre las clases, mientras que un valor de 0.5 representa un 50% de probabilidad de clasificar una instancia de manera correcta.

En este caso se obtuvo un valor AUC del `0.799`, lo cual resalta la habilidad del modelo para distinguir entre dos clases.

[29]:
```r
AUC <- performance(pred, "auc") # area under the curve
AUC_value <- AUC@y.values
cat("Area under the ROC cuerve: ", AUC_value[[1]])
```

```
Area under the ROC cuerve:  0.7998776
```

[30]:
```r
plot(perf, colorize = TRUE, lwd = 3, type = 'l', main = "Receiver Operating␣
↪Characteristic (ROC) Curve")
abline(a = 0, b = 1)
```



Como se mencionó anteriormente, el **valor AUC** indica el buen rendimiento del modelo y por ende de su capacidad para distinguir dos clases. Esto se reafirma con el grafico de la curva ROC generado, donde el valor `0.799` coincide con lo observado.

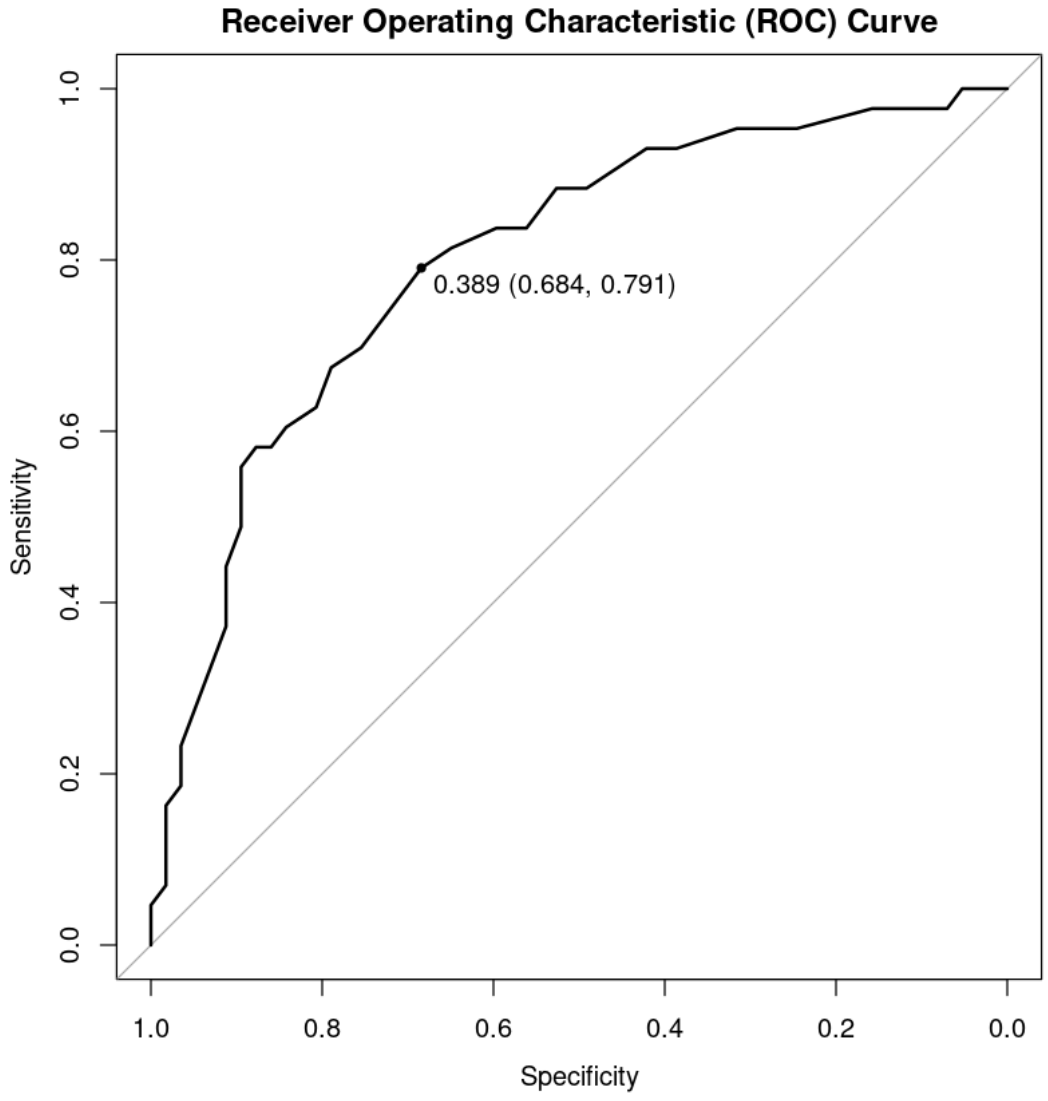
[31]:
```
roc_object <- roc(coro, predictions, auc = TRUE, ci = TRUE)
```

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

Setting direction: controls < cases
```

[32]:
```
roc_object
```

```
Call:
roc.default(response = coro, predictor = predictions, auc = TRUE,     ci = TRUE)

Data: predictions in 57 controls (coro 0) < 43 cases (coro 1).
Area under the curve: 0.7999
95% CI: 0.7114-0.8884 (DeLong)
```

[33]:
```
plot.roc(roc_object, print.thres = 'best', main = "Receiver Operating␣
↪Characteristic (ROC) Curve")
```



Elementos de la matriz de confusión:

- True Positives (TP) es el número de observaciones clasificadas correctamente como positivo, en este caso `coro = 1`.

- False Positives (FP) es el número de observaciones clasificadas de manera incorrecta como positiva (`coro = 1`) cuando en realidad es negativo (`coro = 0`).
- True Negative (TN) es el número de observaciones clasificadas correctamente como negativo, en este caso `coro = 0`.
- False Negative (FN) es el número de observaciones clasificadas de manera incorrecta como negativo (`coro = 0`) cuando en realidad es positivo (`coro = 1`).

Teniendo en cuenta las definiciones anteriores se realiza el cálculo de cada uno de los elementos.

[34]:
```
pred <- ifelse(predictions > 0.389, 1, 0)
TP <- sum(pred == 1 & coro == 1)
FP <- sum(pred == 1 & coro == 0)
FN <- sum(pred == 0 & coro == 1)
TN <- sum(pred == 0 & coro == 0)
```

[46]:
```
cat("True Positives (TP):", TP, "\nFalse Positives (FP):", FP, "\nTrue␣
↪Negatives (TN):", TN, "\nFalse Negatives (FN):", FN)
```

```
True Positives (TP): 34
False Positives (FP): 18
True Negatives (TN): 39
False Negatives (FN): 9
```

Basándonos en los cálculos anteriores se obtienen los valores de `sensitividad` y `especificidad` de acuerdo con las siguientes definiciones:

$TPR = True\ Positive\ Rate = \frac{TP}{P} =$ sensitividad

$FPR = False\ Positive\ Rate = \frac{FP}{N} = 1 -$ especificidad

[62]:
```
sensitivity <- TP / (TP + FN) # true positive rate
ratio <- FP / (FP + TN) # false positive rate
specificity <- 1 - ratio # true negative rate; also calculated as TN / (TN + FP)
```

[63]:
```
cat("Sensitivity (true positive rate):", sensitivity)
```

```
Sensitivity (true positive rate): 0.7906977
```

[64]:
```
cat("Ratio (false positive rate):", ratio)
```

```
Ratio (false positive rate): 0.3157895
```

[65]:
```
cat("Specificity (true negative rate):", specificity)
```

```
Specificity (true negative rate): 0.6842105
```

Calculo del estadistico de `Youden`:

$J =$ sensitividad $-$ (1 $-$ especificidad)

[66]:
```
J <- sensitivity - (1 - specificity)
```

[67]:
```
cat("Youden's J statistic:", J)
```

```
Youden's J statistic: 0.4749082
```

El **estadístico J** de Youden es una medida de la eficacia de una prueba, resume ambas metricas, la `sensibilidad` y la `especificidad`. Valores grandes del estadístico indican que ambas metricas poseen valores altos, lo que resulta en una buena eficiencia a la hora de identificar de manera correcta instancias positivas y negativas.

Un valor de 1.0 indica que el modelo distingue de manera perfecta ambos casos (i.e. 100% de sensibilidad, 100% de especificidad). Por otro lado, valores negativos indican que el desempeño del modelo es peor que suponer clasificaciones aleatorias.

En este caso, el valor del estadístico J de Youden obtenido es de aproximadamente `0.4749`, lo que indica que el modelo tiene una eficacia moderada para distinguir entre las clases positivas y negativas. Esto coincide con los valores obtenidos de cada uno de los elementos (e.g. True Positives = 34; False Positives = 18)

0.6 Problema 6

La siguiente tabla muestra conteos de células T_4 por mm^3 en muestras de sangre de 20 pacientes (en remisión) con enfermedad de Hodgkin, así como conteos en 20 pacientes en remisión de otras enfermedades. Una cuestión de interés es si existen diferencias en las distribuciones de conteos en ambos grupos.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
H	396	568	1212	171	554	1104	257	435	295	397
No-H	375	375	752	208	151	116	736	192	315	1252
H	288	1004	431	795	1621	1378	902	958	1283	2415
No-H	675	700	440	771	688	426	410	979	377	503

- Haga una comparación gráfica exploratoria de estos datos.
- Ajuste un modelo de Poisson apropiado.
- Usando la normalidad asintótica de los estimadores de máxima verosimilitud, dé un intervalo del 90% de confianza para la diferencia en medias. ¿Hay evidencia de diferencias en los dos grupos en cuanto a las medias de los conteos?.

0.6.1 Solución

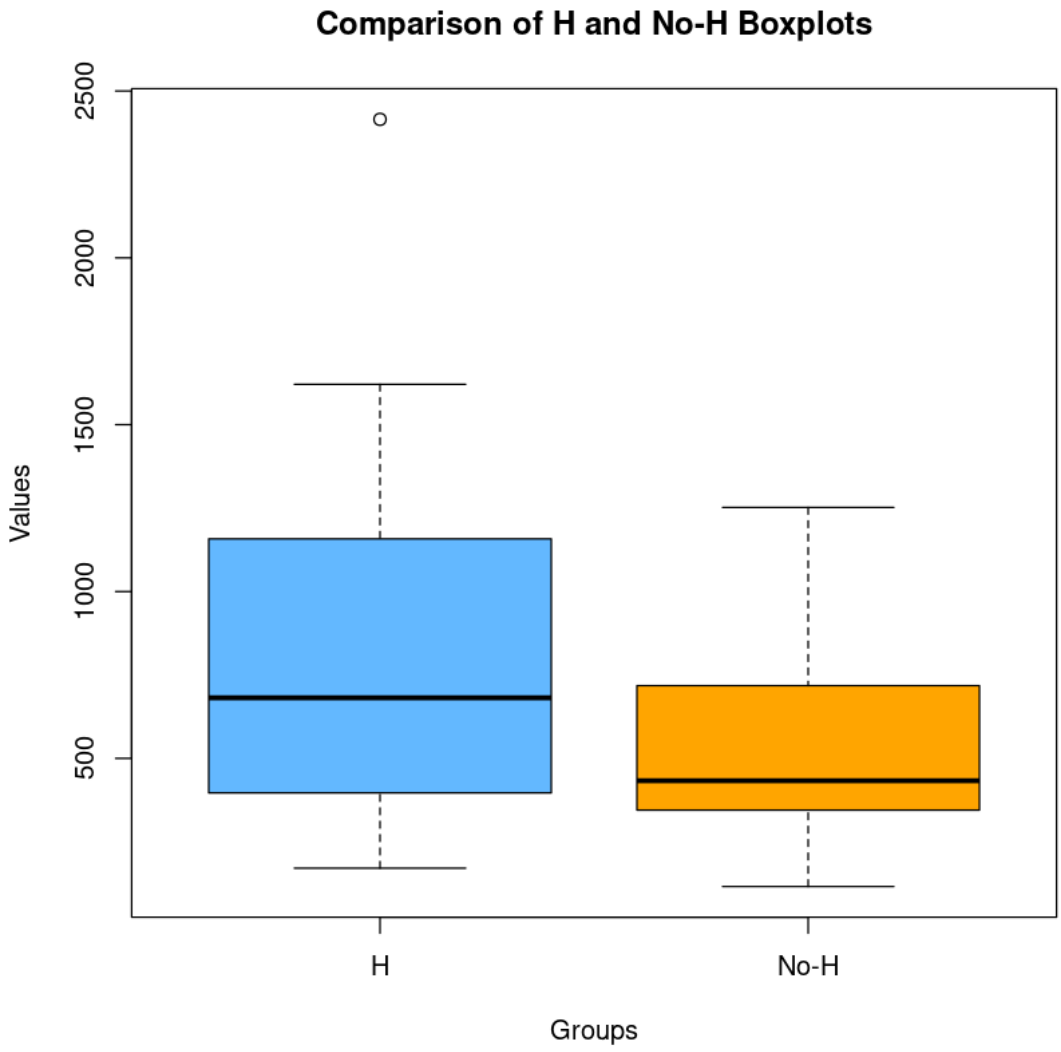
[106]:
```
library(ggplot2) # loading libraries
library(tidyr)
library(dplyr)
```

Datos correspondientes al conteo celular por mm^3 en dos grupos, cada uno conformado por 20 pacientes donde la diferencia radica en el tipo de enfermedad que estos padecen (i.e. enfermedad de Hodgkin).

[107]:
```
hodgkin_data <- data.frame(
  V1 = c(396, 375), V2 = c(568, 375), V3 = c(1212, 752), V4 = c(171, 208),
  V5 = c(554, 151), V6 = c(1104, 116), V7 = c(257, 736), V8 = c(435, 192),
  V9 = c(295, 315), V10 = c(397, 1252), V11 = c(288, 675), V12 = c(1004, 700),
  V13 = c(431, 440), V14 = c(795, 771), V15 = c(1621, 688), V16 = c(1378, 426),
  V17 = c(902, 410), V18 = c(958, 979), V19 = c(1283, 377), V20 = c(2415, 503)
) # data stored in a data frame

rownames(hodgkin_data) <- c("H", "No-H") # row names; H and No-H
```

0.6.2 Comparación de las Distribuciones de los Grupos mediante Boxplots

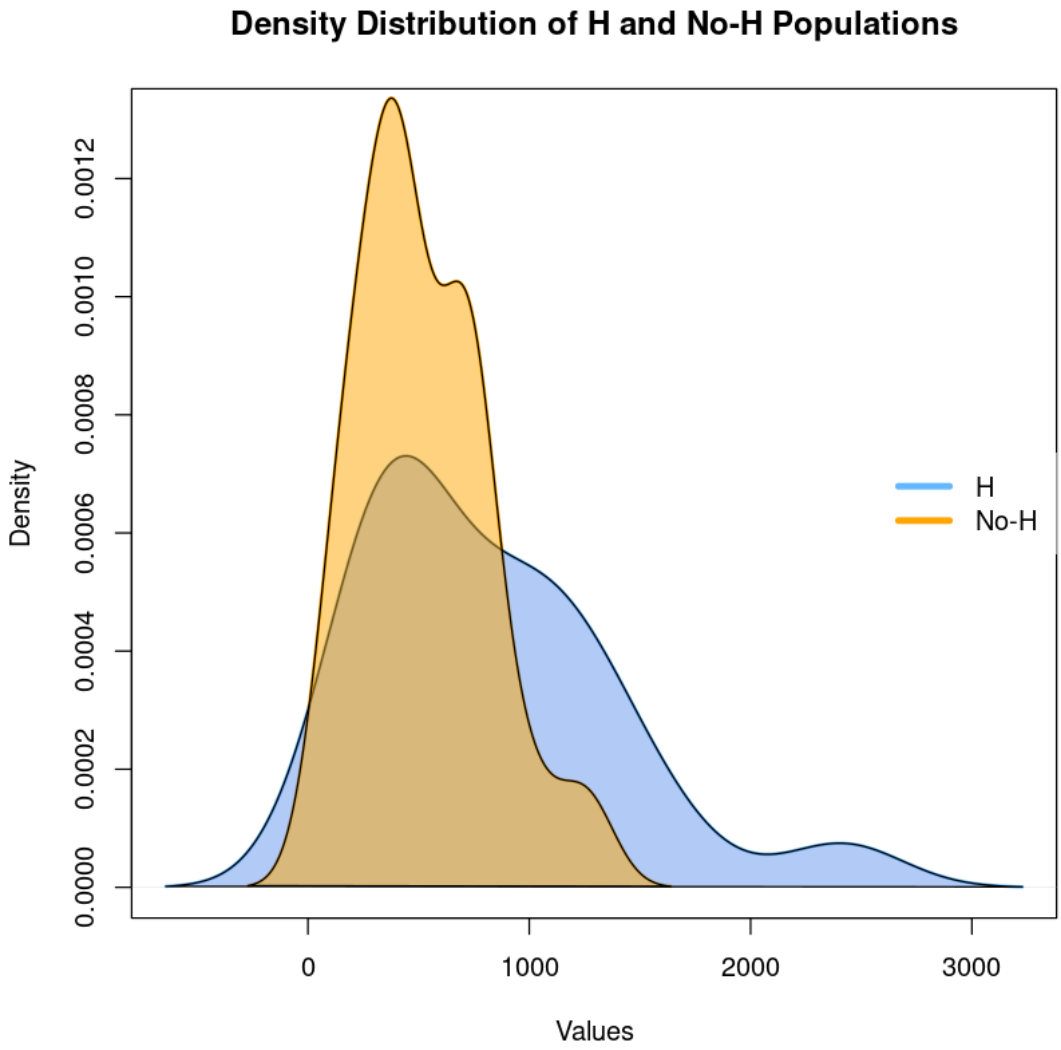
[122]:
```r
transposed_data <- t(hodgkin_data) # transpose data

boxplot(transposed_data[, "H"], transposed_data[, "No-H"], names = c("H","No-H"),
        main = "Comparison of H and No-H Boxplots", xlab = "Groups", ylab = "Values",
        col = c("Steel Blue 1", "Orange"))
```



En el gráfico generado se muestra un resumen de los principales estadísticos de ambos grupos. Se observa la presencia de un **valor atípico** en el grupo `H`, lo que podría afectar la media, ya que esta es sensible a valores extremos. Además, el grupo `H` tiene una mediana superior en comparación con el grupo `No-H`, lo que sugiere que el grupo `H` tiende a tener **valores más altos**. Por último, el grupo `H` presenta una mayor variabilidad en los datos en comparación con el grupo `No-H`.

0.6.3 Comparación de las Distribuciones de Densidad entre Grupos

[131]:
```r
H_density <- density(transposed_data[, "H"])
NoH_density <- density(transposed_data[, "No-H"])

plot(H_density, col = "Steel Blue 1", lwd = 2, main = "Density Distribution of␣
 ↪H and No-H Populations",
     xlab = "Values", ylab = "Density", xlim = range(c(H_density$x,␣
 ↪H_density$x)), ylim = range(c(0, 0.0013)))
lines(density_NoH1, col = "orange", lwd = 2) # plot density distribution for␣
 ↪both H and No-H populations

polygon(H_density, col = rgb(100/255, 149/255, 237/255, 0.5), border = "Black")␣
 ↪# fill area under the density curve
polygon(NoH_density, col = rgb(255/255, 165/255, 0/255, 0.5), border = "Black")
legend("right", legend = c("H", "No-H"),
       col = c("Steel Blue 1", "orange"), box.lty=0, lwd = 4) # add legend
```



Como se observa en el gráfico anterior, el grupo `H` presenta una mayor variabilidad en los datos. Esto indica que los valores en el grupo `H` están más dispersos en comparación con el grupo `No-H`. Además, se puede inferir que la **media del grupo `H` es mayor** a la del grupo `No-H`, ya que este valor medio se ve afectado por la presencia de un **valor atípico presente en el grupo `H`**. A pesar de esta diferencia, una porción significativa de ambas distribuciones se solapa, lo que sugiere que las dos poblaciones comparten ciertas similitudes en sus rangos de valores.

Adicionalmente, se puede observar que la distribución del grupo `No-H` es **bimodal**, lo que indica la presencia de dos picos o subgrupos dentro de esta población. Por otro lado, la distribución del grupo `H` es **unimodal**, con un solo pico, lo que sugiere una tendencia más uniforme en la distribución de los datos. Esta diferencia en la forma de las distribuciones podría proporcionar información útil sobre las características subyacentes de cada grupo y cómo se comparan entre sí.

0.6.4 Ajuste del Modelo de Regresión de Poisson

Crear un data frame con dimensiones $40x2$ (i.e. 40 filas, 2 columnas), donde las columnas corresponden al conteo celular T_4 y a su respectivo grupo perteneciente (`H` o `No-H`), mientras que las filas indican el conteo celular del paciente asi al grupo que este pertenece. A partir de este data frame se ajusta el modelo de regression Poisson, con la variable `Counts` como la variable de respuesta y la variable `Group` como la variable explicativa.

[132]:
```
Counts <- c(396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397,
            288, 1004, 431, 795, 1621, 1378, 902, 958, 1283, 2415,
            375, 375, 752, 208, 151, 116, 736, 192, 315, 1252,
            675, 700, 440, 771, 688, 426, 410, 979, 377, 503)

Group <- factor(rep(c("H", "No-H"), each = 20))

hodgkin_data <- data.frame(Counts = Counts, Group = Group)
```

Ajustar un modelo de regresión Poisson con la variable `Counts` como la variable de respuesta, y la variable `Group` como la variable explicativa.

[133]:
```
poisson.model <- glm(Counts ~ Group, family = poisson, data = hodgkin_data)
summary(poisson.model)
```

```
Call:
glm(formula = Counts ~ Group, family = poisson, data = hodgkin_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.713199   0.007793   861.4   <2e-16 ***
GroupNo-H   -0.455436   0.012511   -36.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11325  on 39  degrees of freedom
Residual deviance:  9965  on 38  degrees of freedom
AIC: 10294

Number of Fisher Scoring iterations: 5
```

Grupo: El coeficiente de la variable *Group* es de -0.4554, y dado que el **p-valor** es pequeño ($< 2e-16$), podemos afirmar que la variable es estadisticamente significativa al momento de predecir la respuesta (i.e. conteo celular T_4).

Nota: Los asteriscos *** indican que la variable es estadisticamente **significativa** considerando un nivel de significancia del 0.001 (**p-valor** < 0.001).

Considerando un nivel de significancia del 0.001, podemos concluir que la variable *Group* es estadisticamente significativa. Por lo tanto, el efecto que tiene la variable *Group* sobre la variable de respuesta es **estadisticamente significativo**.

0.7 Problema 7

Los datos de la tabla en la siguiente hoja son números, n, de pólizas de seguros y los correspondientes números, y, de reclamos (esto es, número de accidentes en los que se pidió el amparo de la póliza). La variable CAR es una codificación de varias clases de carros, EDAD es la edad del titular de la póliza y DIST es el distrito donde vive el titular.

- Calcule la tasa de reclamos, y/n, para cada categoría y grafique estas tasas contra las diferentes variables para tener una idea de los efectos principales.
- Use regresión logística para estimar los efectos principales (cada variable tratada como categórica y modelada usando variables indicadoras) así como sus interacciones.

CAR	EDAD	DIST = 0 (y)	DIST = 0 (n)	DIST = 1 (y)	DIST = 1 (n)
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

0.7.1 Solución

[29]:
```
library(ggplot2) # loading libraries
```

Datos correspondientes al números de pólizas de seguros n, y sus respectivos números de reclamos y.

[48]:
```
car_insurance_policy_data <- data.frame(
  CAR = c(1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4),
  AGE = c(1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4),
  DIST_0_y = c(65, 65, 52, 310, 98, 159, 175, 877, 41, 117, 137, 477, 11, 35,␣
 ↪39, 167),
  DIST_0_n = c(317, 476, 486, 3259, 486, 1004, 1355, 7660, 223, 539, 697, 3442,␣
 ↪40, 148, 214, 1019),
  DIST_1_y = c(2, 5, 4, 36, 7, 10, 22, 102, 5, 7, 16, 63, 0, 6, 8, 33),
  DIST_1_n = c(20, 33, 40, 316, 31, 81, 122, 724, 18, 39, 68, 344, 3, 16, 25,␣
 ↪114)
) # data stored in a data frame
```

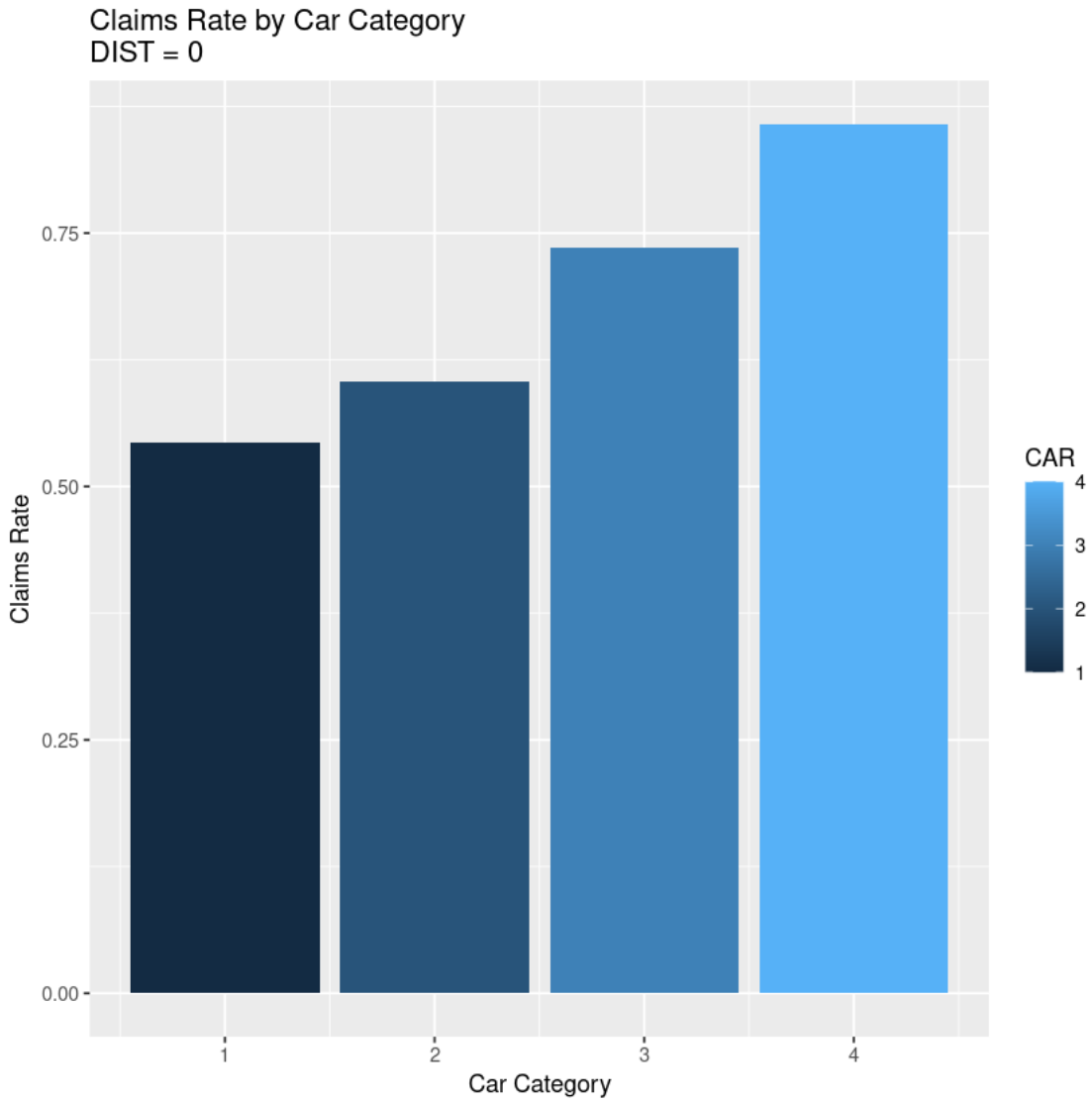
0.7.2 Tasas de Reclamos en Función de las Variables por Categorías: DIST = 0 y DIST = 1

Cálculo de las tasas de reclamos como la relación y_i/n_i para cada una de las categorías.

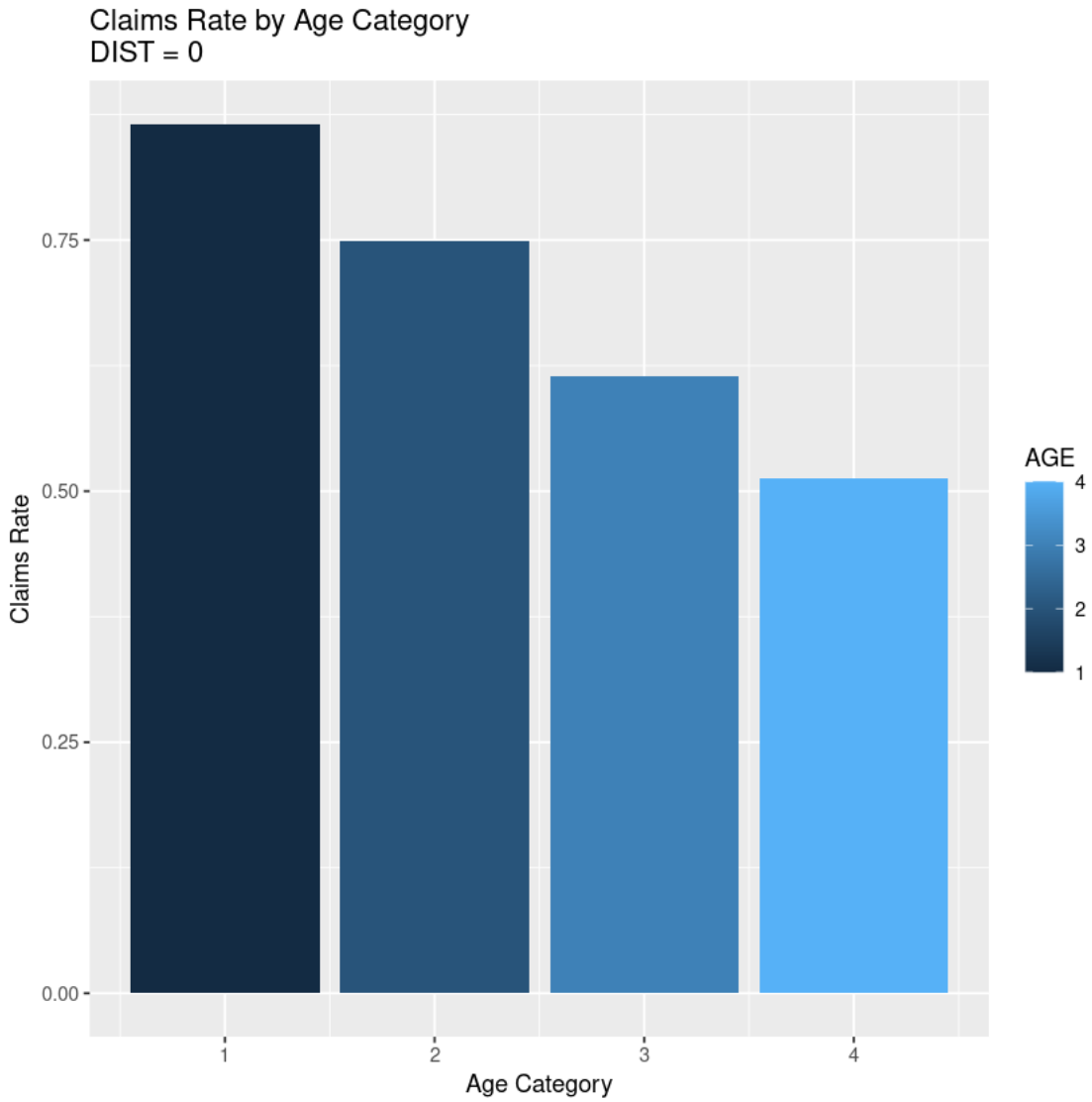
[50]:
```
car_insurance_policy_data$CLAIMS_RATE_0 <- car_insurance_policy_data$DIST_0_y/
 ↪car_insurance_policy_data$DIST_0_n # category 0
car_insurance_policy_data$CLAIMS_RATE_1 <- car_insurance_policy_data$DIST_1_y/
 ↪car_insurance_policy_data$DIST_1_n # category 1
```

Tasa de Reclamos en Función de las Variables: Distrito 0

[67]:
```
ggplot(car_insurance_policy_data, aes(x = CAR, y = CLAIMS_RATE_0, fill = CAR)) +
  geom_bar(stat = "identity") +
  labs(title = "Claims Rate by Car Category\nDIST = 0",
       x = "Car Category",
       y = "Claims Rate") # claims rate vs car category
```



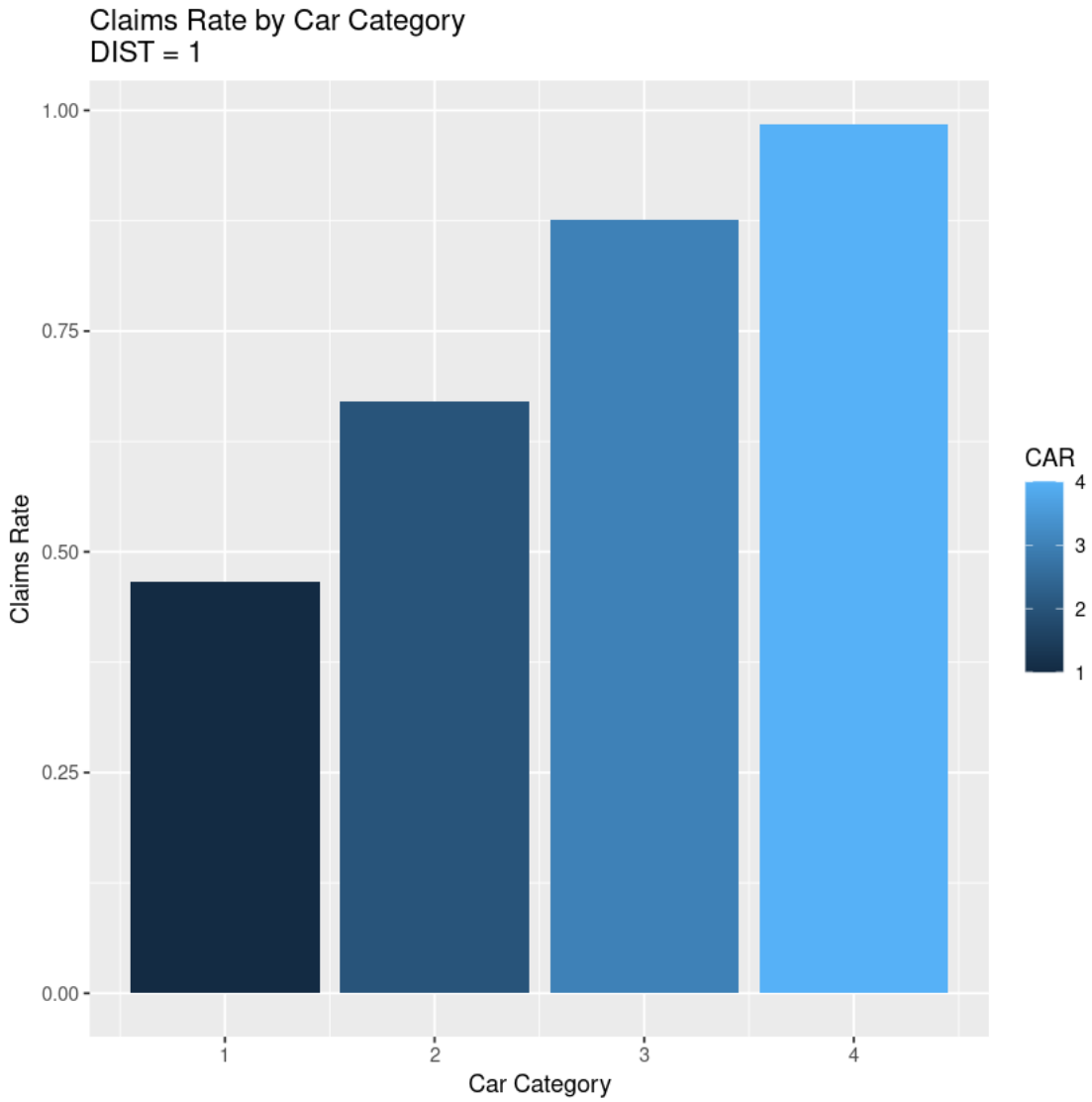
[68]:
```
ggplot(car_insurance_policy_data, aes(x = AGE, y = CLAIMS_RATE_0, fill = AGE)) +
  geom_bar(stat = "identity") +
  labs(title = "Claims Rate by Age Category\nDIST = 0",
       x = "Age Category",
       y = "Claims Rate") # claims rate vs age category
```



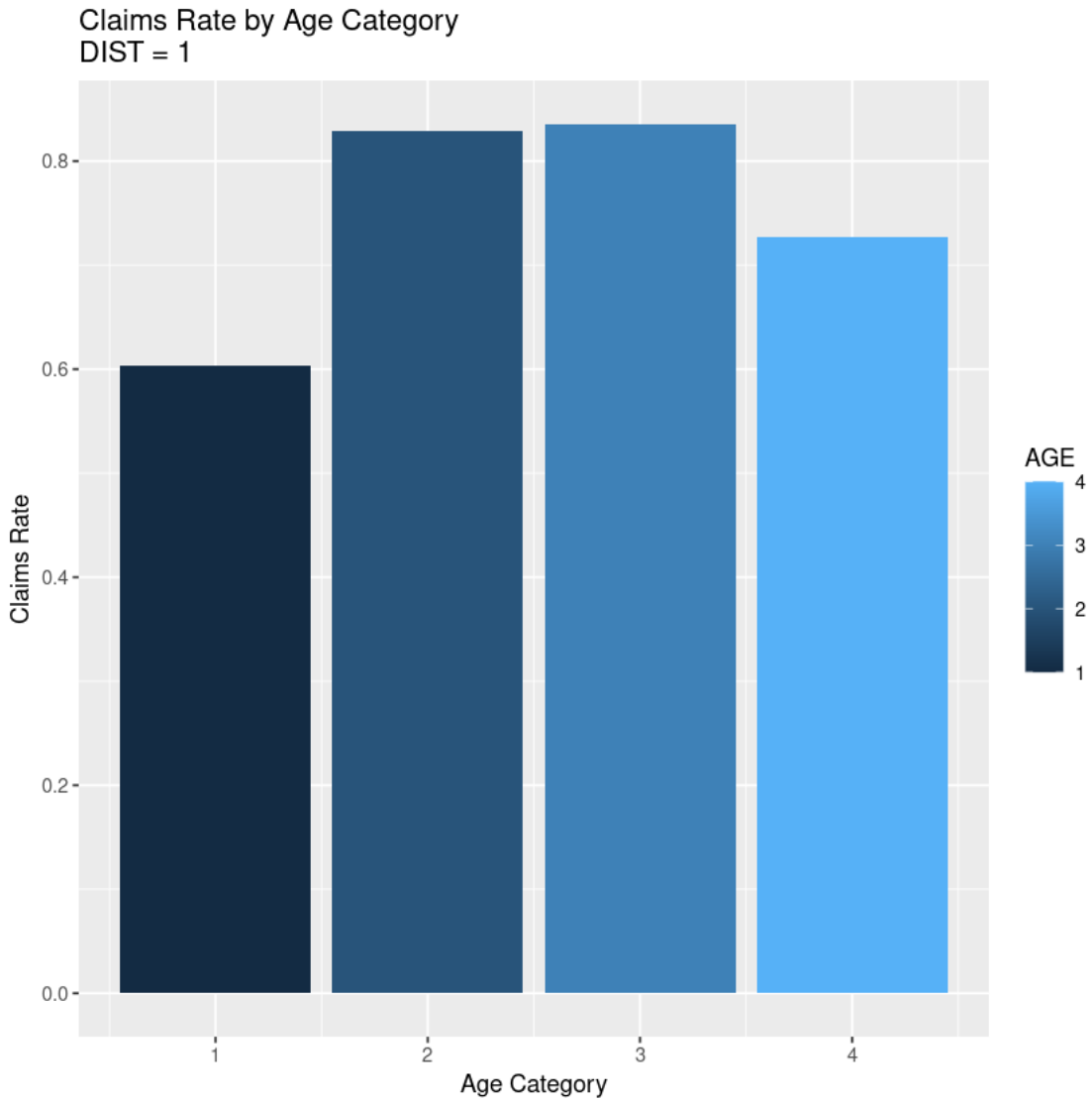
En términos generales, observamos que la tasa acumulada de reclamos tiende a aumentar a medida que avanzamos por las distintas categorías de la variable `CAR`. Por otro lado, para la variable `AGE` se observa el fenómeno opuesto, la tasa disminuye progresivamente a través de las diferentes categorías de edad.

Tasa de Reclamos en Función de las Variables: Distrito 1

[71]:
```
ggplot(car_insurance_policy_data, aes(x = CAR, y = CLAIMS_RATE_1, fill = CAR)) +
  geom_bar(stat = "identity") +
  labs(title = "Claims Rate by Car Category\nDIST = 1",
       x = "Car Category",
       y = "Claims Rate") # claims rate vs car category
```



[72]:
```
ggplot(car_insurance_policy_data, aes(x = AGE, y = CLAIMS_RATE_1, fill = AGE)) +
  geom_bar(stat = "identity") +
  labs(title = "Claims Rate by Age Category\nDIST = 1",
       x = "Age Category",
       y = "Claims Rate") # claims rate vs age category
```



Al igual que en el caso anterior, la tasa de reclamos frente a la variable `CAR` muestra un incremento a medida que se avanza en sus diferentes categorías. Sin embargo, para la variable `AGE`, no se observa un patrón claro entre las distintas categorías. Lo único relevante es que la tasa de reclamos en la categoría 1 es significativamente menor en comparación con las demás categorías.

0.7.3 Ajuste del Modelo de Regresión Logistica

Crear un data frame con dimensiones $32x5$ (i.e. 35 filas, 5 columnas), donde las columnas corresponden a las distintas variables: `CAR` corresponde a las distintas clases de carros; `AGE` corresponde a la edad del titular de la póliza; y `DIST` al distrito donde vive el titular.

[91]:
```
car_insurance_policy_data <- data.frame(
  CAR = factor(rep(c(1, 2, 3, 4), each = 4)),
  AGE = factor(rep(c(1, 2, 3, 4), times = 4)),
```

```
  DIST = factor(rep(c(0, 1), each = 16)),
  y = c(65, 65, 52, 310, 98, 159, 175, 877, 41, 117, 137, 477, 11, 35, 39, 167,
        2, 5, 4, 36, 7, 10, 22, 102, 5, 7, 16, 63, 0, 6, 8, 33),
  n = c(317, 476, 486, 3259, 486, 1004, 1355, 7660, 223, 539, 697, 3442, 40,␣
 ↪148, 214, 1019,
        20, 33, 40, 316, 31, 81, 122, 724, 18, 39, 68, 344, 3, 16, 25, 114)
)
```

Ajustar un modelo de regresión logística utilizando la variable **reclamos** como la variable de respuesta y las variables `CAR`, `AGE` y `DIST` como los predictores. En este caso, y representa el número de reclamos y $n-y$ el número de pólizas menos el número de reclamos, es decir el número de pólizas que no han pedido amparo.

[92]:
```
logistic.model <- glm(cbind(y, n - y) ~ CAR * AGE * DIST, family =␣
 ↪binomial(link = "logit"), data = car_insurance_policy_data)
summary(logistic.model)
```

```
Call:
glm(formula = cbind(y, n - y) ~ CAR * AGE * DIST, family = binomial(link =␣
 ↪"logit"),
    data = car_insurance_policy_data)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.3550     0.1391  -9.740  < 2e-16 ***
CAR2              -0.0210     0.1793  -0.117 0.906760
CAR3              -0.1354     0.2219  -0.610 0.541753
CAR4               0.3856     0.3805   1.014 0.310756
AGE2              -0.4892     0.1928  -2.537 0.011174 *
AGE3              -0.7668     0.2022  -3.792 0.000149 ***
AGE4              -0.8976     0.1514  -5.929 3.04e-09 ***
DIST1             -0.8422     0.7582  -1.111 0.266686
CAR2:AGE2          0.1948     0.2396   0.813 0.416346
CAR3:AGE2          0.6968     0.2792   2.495 0.012586 *
CAR4:AGE2          0.2865     0.4472   0.641 0.521707
CAR2:AGE3          0.2343     0.2454   0.955 0.339707
CAR3:AGE3          0.8492     0.2826   3.005 0.002654 **
CAR4:AGE3          0.2349     0.4446   0.528 0.597177
CAR2:AGE4          0.2280     0.1923   1.185 0.235856
CAR3:AGE4          0.5609     0.2350   2.387 0.017001 *
CAR4:AGE4          0.2374     0.3943   0.602 0.547094
CAR2:DIST1         0.9861     0.8788   1.122 0.261808
CAR3:DIST1         1.3771     0.9390   1.467 0.142493
CAR4:DIST1       -21.3479 37437.8581  -0.001 0.999545
AGE2:DIST1         0.9636     0.9102   1.059 0.289733
AGE3:DIST1         0.7668     0.9350   0.820 0.412179
```

```
AGE4:DIST1          1.0436     0.7809   1.336 0.181439
CAR2:AGE2:DIST1    -1.3972     1.0711  -1.304 0.192095
CAR3:AGE2:DIST1    -1.7355     1.1490  -1.510 0.130932
CAR4:AGE2:DIST1    21.8877 37437.8581   0.001 0.999534
CAR2:AGE3:DIST1    -0.5163     1.0647  -0.485 0.627725
CAR3:AGE3:DIST1    -1.0724     1.1278  -0.951 0.341658
CAR4:AGE3:DIST1    22.1708 37437.8581   0.001 0.999527
CAR2:AGE4:DIST1    -0.9497     0.9054  -1.049 0.294206
CAR3:AGE4:DIST1    -1.2466     0.9688  -1.287 0.198168
CAR4:AGE4:DIST1    21.8782 37437.8581   0.001 0.999534
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.4433e+02  on 31  degrees of freedom
Residual deviance: 5.2608e-10  on  0  degrees of freedom
AIC: 225.92

Number of Fisher Scoring iterations: 21
```

Los coeficientes de las variables `AGE` son negativos y **estadísticamente significativos**. En particular, para `AGE3` y `AGE4`, los **p-valores** son menores al nivel de significancia 0.001, mientras que para `AGE2`, el **p-valor** es menor a 0.05. Esto indica que los efectos de `AGE` son significativos desde un punto de vista estadístico. Dado que los coeficientes son negativos, se puede concluir que, a medida que aumenta la edad, el número de reclamos disminuye. Esta observación coincide con los resultados mostrados en los gráficos anteriores.

Algunas interacciones entre las variables `CAR` y `AGE` resultan ser **significativas**, especialmente aquellas que involucran a `CAR3`. Esto sugiere que el impacto de `CAR` en la variable de respuesta varía según la categoría de `AGE`.

Las interacciones significativas indican que, en estos casos, no es adecuado interpretar el efecto de `CAR` o `AGE` de manera aislada, ya que es **la combinación de ambas variables lo que genera un efecto significativo sobre la variable de respuesta** (reclamos).

Por último, las interacciones con la variable `DIST1` **no resultan significativas**, ya que los **p-valores** son mayores al nivel de significancia 0.001.

0.8 Problema 8

A lo largo del curso hemos enfatizado el uso del método de Máxima Verosimilitud para todo lo relacionado con estimación. Consideremos ahora una alternativa: El método de la **mínima Ji-cuadrada**. Suponga que las celdas de una multinomial están parametrizadas en términos de un vector $\theta = (\theta_1, \ldots, \theta_s)^T$. El método de la mínima Ji-cuadrada consiste en estimar θ mediante aquel valor que minimice el estadístico de Pearson:

$$\chi^2 = \sum \frac{(\text{obs} - \text{esp})^2}{\text{esp}} = \sum_{j=1}^{K} \frac{(y_j - n\pi_j(\theta))^2}{n\pi_j(\theta)}$$

Considere el siguiente problema. Suponga una población muy grande de objetos que pueden clasificarse en tres categorías: A, B y C. Para estimar las proporciones π_1, π_2, π_3 correspondientes a cada una de esas categorías, se efectuó un estudio; se obtuvieron tres muestras de tamaños $n_1, n_2, , n_3$ tomadas de la población global, sin embargo, en vez de registrar la frecuencia observada de A's, B's y C's de cada muestra, lo que se hizo fue anotar:

- Número de A's en la muestra de tamaño $n_1 = y_1$
- Número de B's en la muestra de tamaño $n_2 = y_2$
- Número de A's en la muestra de tamaño $n_3 = y_3$

Estime π_1, π_2, π_3 usando el método de la **mínima ji-cuadrada**: suponga que $n_1 = 100, y_1 = 22, n_2 = 150, y_2 = 52, n_3 = 200, y_3 = 77$. Esto es, encuentre π_1, π_2, π_3 que minimicen:

$$\frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1} + \frac{[(n_1 - y_1) - n_1(1-\pi_1)]^2}{n_1(1-\pi_1)} + \ldots + \frac{(y_3 - n_3\pi_3)^2}{n_3\pi_3} + \frac{[(n_3 - y_3) - n_3(1-\pi_3)]^2}{n_3(1-\pi_3)}$$

con la restricción $\pi_3 = 1 - \pi_1 - \pi_2$. (Sugerimos usar directamente `nlminb` de R).

0.8.1 Solución

[76]:
```
library(stats) # loading stats library
```

[18]:
```
n1 <- 100 # initialize samples and categories sizes
y1 <- 22

n2 <- 150
y2 <- 52

n3 <- 200
y3 <- 77
```

[19]:
```
objective_function <- function(parameters){ # function to be minimized
    pi1 <- parameters[1]
    pi2 <- parameters[2]
    pi3 <- 1 - pi1 - pi2 # constraint
```

```
    A <- ((y1 - n1 * pi1)^2 / (n1 * pi1)) + (((n1 - y1) - (n1 * (1 - pi1))^2) /␣
 ↪(n1 * (1 - pi1))) # category A

    B <- ((y2 - n2 * pi2)^2 / (n2 * pi2)) + (((n2 - y2) - (n2 * (1 - pi2))^2) /␣
 ↪(n2 * (1 - pi2))) # category B

    C <- ((y3 - n3 * pi3)^2 / (n3 * pi3)) + (((n3 - y3) - (n3 * (1 - pi3))^2) /␣
 ↪(n3 * (1 - pi3))) # category C

    chi_squared <- A + B + C

    return(chi_squared) # return chi-squared value
}
```

[20]:
```
estimated_parameters <- nlminb(start = c(0.33, 0.33), objective =␣
 ↪objective_function, lower = c(0, 0), upper = c(1, 1))
estimated_parameters$par # estimated parameters pi1 and pi2
```

1. 0.309614190725317 2. 0.346226208296422

[98]:
```
pi1_estimated <- estimated_parameters$par[1]
pi2_estimated <- estimated_parameters$par[2]
pi3_estimated <- 1 - pi1_estimated - pi2_estimated # constraint
```

Estimaciones de π_1, π_2 y π_3 dada la restricción $\pi_3 = 1 - \pi_1 - \pi_2$.

[100]:
```
cat("Valor estimado de pi1:", pi1_estimated)
cat("\nValor estimado de pi2:", pi2_estimated)
cat("\nValor estimado de pi3:", pi3_estimated)
```

```
Valor estimado de pi1: 0.3096142
Valor estimado de pi2: 0.3462262
Valor estimado de pi3: 0.3441596
```

0.9 Problema 9

Se toman los datos relacionados con el hundimiento del Titanic en abril 1912. El resultado se puede expresar en la tabla de dimensión 4.

Las variables son:

- **Class** de los pasajeros (1, 2, 3, Tripulación)
- **Sex** de los pasajeros (Female, Male)
- **Age** de los pasajeros (Adult, Child)
- **Survived** si los pasajeros sobrevivieron o no (Yes, No)

Usar la librería en R `titanic` y los datos se encuentran en la variable `Titanic`.

Considerar entonces un modelo log-lineal para analizar los posibles efectos.

0.9.1 Solución

Se carga el conjunto de datos `Titanic`.

[8]:
```
library(titanic)
library(MASS)
```

[9]:
```
titanic_data <- as.data.frame(Titanic) # data stored in a data frame
head(titanic_data)
```

A data.frame: 6 × 5

	Class <fct>	Sex <fct>	Age <fct>	Survived <fct>	Freq <dbl>
1	1st	Male	Child	No	0
2	2nd	Male	Child	No	0
3	3rd	Male	Child	No	35
4	Crew	Male	Child	No	0
5	1st	Female	Child	No	0
6	2nd	Female	Child	No	0

Ajuste del modelo **log-lineal** inicial sin interacciones adicionales.

[28]:
```
loglinear.model <- loglm(Freq ~ Class + Age + Sex + Survived, data =␣
 ↪titanic_data, param = TRUE, fit = TRUE)
loglinear.model
```

```
Call:
loglm(formula = Freq ~ Class + Age + Sex + Survived, data = titanic_data,
    param = TRUE, fit = TRUE)

Statistics:
                     X^2 df P(> X^2)
Likelihood Ratio 1243.663 25        0
Pearson          1637.445 25        0
```

Likelihood Ratio Test: muestra que el modelo log-lineal es significativamente mejor que un modelo nulo sin efectos.

Pearson Chi-Square Test: indica que hay discrepancias significativas entre las frecuencias observadas y las esperadas, lo que podría sugerir que el modelo no está capturando completamente las relaciones en los datos.

Los resultados sugieren que, aunque el modelo ajustado es significativamente mejor que el modelo nulo, podría ser útil considerar modelos más complejos (mayor no. de interacciones) para capturar mejor la variabilidad en los datos.

Ajuste del modelo **log-lineal** inicial considerando una interacción adicional entre las variables `Class` y `Sex`.

[29]:
```
loglinear.model <- loglm(Freq ~ Class + Age + Sex + Survived + Class*Sex, data
 ↪= titanic_data, param = TRUE, fit = TRUE)
loglinear.model
```

```
Call:
loglm(formula = Freq ~ Class + Age + Sex + Survived + Class *
    Sex, data = titanic_data, param = TRUE, fit = TRUE)

Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 831.0620 22        0
Pearson          833.8984 22        0
```

Likelihood Ratio Test: El valor del estadistico indica que el modelo con `Class`, `Age`, `Sex`, `Survived`, y la interacción `Class * Sex` es significativamente mejor que el modelo nulo.

Pearson Chi-Square Test: Al igual que el modelo anterior, indica una discrepancia significativa entre las frecuencias observadas y las esperadas, sugiriendo que el modelo podría no estar ajustando perfectamente los datos.

Ajuste del modelo **log-lineal** inicial considerando una interacción adicional entre las variables `Class`, `Sex` y `Age`.

[30]:
```
loglinear.model <- loglm(Freq ~ Class + Age + Sex + Survived + Class*Sex*Age,
 ↪data = titanic_data, param = TRUE, fit = TRUE)
loglinear.model
```

```
Call:
loglm(formula = Freq ~ Class + Age + Sex + Survived + Class *
    Sex * Age, data = titanic_data, param = TRUE, fit = TRUE)

Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 671.9622 15        0
Pearson               NaN 15      NaN
```

Likelihood Ratio Test: Indica que el modelo con las interacciones `Class * Sex * Age` proporciona una explicación significativamente mejor de los datos en comparación con un modelo nulo.

Pearson Chi-Square Test: El valor `NaN` sugiere que hubo un problema con el cálculo de esta

estadística, lo que podría indicar problemas con la calidad del ajuste del modelo o con los datos.

Ajuste del modelo **log-lineal** inicial considerando una interacción adicional entre las variables `Class`, `Sex`, `Age` y `Survived`.

[31]:
```
loglinear.model <- loglm(Freq ~ Class + Age + Sex + Survived +␣
 ↪Class*Sex*Age*Survived, data = titanic_data, param = TRUE, fit = TRUE)
loglinear.model
```

```
Call:
loglm(formula = Freq ~ Class + Age + Sex + Survived + Class *
    Sex * Age * Survived, data = titanic_data, param = TRUE,
    fit = TRUE)

Statistics:
                  X^2 df P(> X^2)
Likelihood Ratio    0  0        1
Pearson           NaN  0        1
```

Likelihood Ratio Test: El valor de 0 y los grados de libertad de 0 sugieren que el modelo con todas las interacciones posibles (`Class * Sex * Age * Survived`) no mejora el ajuste respecto al modelo nulo (i.e. como si el modelo ajustado y el modelo nulo fueran equivalentes en términos de ajuste).

Pearson Chi-Square Test: El valor `NaN` y los grados de libertad de 0 indican que no se puede calcular esta estadística, posiblemente debido a una mala especificación del modelo o problemas con los datos.

El modelo con todas las interacciones posibles puede ser **demasiado complejo** para los datos disponibles, resultando en un ajuste redundante o no informativo.

0.9.2 Comparación de los Modelos

Los modelos con interacciones parciales, ya sean dobles o triples (`Class * Sex` y `Class * Sex * Age`), muestran **mejoras significativas** sobre el modelo nulo, pero también presentan **discrepancias** en el ajuste.

El modelo con todas las interacciones posibles (`Class * Sex * Age * Survived`) muestra problemas de ajuste, evidenciados por los valores 0 en la razón de verosimilitudes y valores `NaN` en la estadístico χ^2, indicando un posible sobreajuste o especificación incorrecta.

Es necesario considerar cada una de las interacciones dobles y triples a fin de poder determinar el modelo que brinde el mejor ajuste con los datos y a su vez que capture la mayor cantidad de información.

0.10 Problema 10

Se ha realizado un análisis sobre el valor terapéutico del ácido ascórbico (vitamina C) en relación a su efecto sobre la gripe común. Se tiene una tabla 2×2 con los recuentos correspondientes para una muestra de 279 personas.

	Gripe	No Gripe	Total
Placebo	31	109	140
Acido Ascorbico	17	122	139
Total	48	231	279

Aplicar un modelo lineal para determinar si existe evidencia suficiente para asegurar que el ácido ascórbico ayuda a tener menos gripe.

0.10.1 Solución

[21]:
```
ascorbic_acid_data <- data.frame(
  Treatment = c("Placebo", "Ascorbic Acid", "Total"),
  Cold = c(31, 17, 48),
  No_Cold = c(109, 122, 231),
  Total = c(140, 139, 279)
) # data stored in a data frame
```

Se modifica el data frame para crear una nueva columna con el objetivo de mapear la variable *Treatment*. Los valores '0' y '1' indican la presencia y ausencia del tratamiento de ácido ascórbico, respectivamente.

[2]:
```
ascorbic_acid_data <- ascorbic_acid_data[-c(3), ]
ascorbic_acid_data$Treatment_No_Category <- ifelse(ascorbic_acid_data$Treatment␣
↪== "Placebo", 0, ifelse(ascorbic_acid_data$Treatment == "Ascorbic Acid", 1,␣
↪NA)) # mapping the treatment variable
ascorbic_acid_data
```

A data.frame: 2×5

	Treatment <chr>	Cold <dbl>	No_Cold <dbl>	Total <dbl>	Treatment_No_Category <dbl>
1	Placebo	31	109	140	0
2	Ascorbic Acid	17	122	139	1

Ajustar un modelo de **regresion logistica**, utilizando ambas columnas *Cold* y *No Cold* para representar de manera correcta la naturaleza binaria del conjunto de datos.

[3]:
```
logistic.model <- glm(cbind(Cold, No_Cold) ~ Treatment_No_Category, family =␣
↪'binomial'(link = logit), data = ascorbic_acid_data)
summary(logistic.model)
```

```
Call:
glm(formula = cbind(Cold, No_Cold) ~ Treatment_No_Category, family =␣
↪binomial(link = logit),
```

```
    data = ascorbic_acid_data)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.2574     0.2035  -6.177 6.53e-10 ***
Treatment_No_Category  -0.7134     0.3293  -2.166   0.0303 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4.8717e+00  on 1  degrees of freedom
Residual deviance: 7.5495e-15  on 0  degrees of freedom
AIC: 13.578

Number of Fisher Scoring iterations: 3
```

Tratamiento: Dado que el coeficiente para la variable `treatment category` toma el valor de -0.7134 con un **p-valor** pequeño (0.0303), podemos decir que la variable es estadisticamente significativa al momento de predecir la respuesta, en este caso el efecto sobre la gripe común.

Nota: Los asteriscos * indican que la variable es estadisticamente **significativa** considerando un nivel de significancia del 0.05 (**p-valor** < 0.05).

Considerando un nivel de significancia del 0.05 (**p-valor** < 0.05), podemos afirmar que la variable `treatment` es estadisticamente significativa. En otras palabras, la variable `treatment` guarda una **relacion significativa** (estadisticamente hablando) con la variable de respuesta, es decir, **el ácido ascórbico tiene un efecto sobre la gripe común**.