Base de datos: datosP
En todas las preguntas utilice un error de 0.05

Pontificia Universidad Javeriana
Departamento de Matemáticas
Análisis de Regresión. Parcial 3

Nota: (1) Las respuestas deben estar argumentadas, pregunta sin procedimiento no tendrá validez (2) Se pueden utilizar apuntes, pero no teléfonos, tabletas o computadores personales (3) Escriba los comandos de R con los que obtuvo sus resultados

Nombre:

Pregunta 1 **(2.5 puntos)** Una persona está interesada en construir un modelo que le permita identificar las variables que influyen en la admisión (o no) de un estudiante a una universidad X. Para eso realiza un estudio en el que recoge información de un grupo de personas que se presentaron al proceso de admisión de la universidad
Variables

admit: 1 si fue admitido , 0 si no
exam: examen de ingreso a la Universidad
prom: promedio de calificaciones en los estudios de bachillerato
rango: categoría de la escuela donde estudio, según su prestigio. Variable categórica ordinal, valores 1,2,3,4. Donde 1 representa el mayor prestigio y 4 el menor

(a) Usted decide ajustar un modelo de regresión para analizar el problema, ¿qué tipo de modelo de regresión ajustaría? Justifique su respuesta

Respuesta (a) : La variable de respuesta, admit, tiene distribución Bernoulli por lo que se aplica la regresión logística

Para las preguntas (b), (c) y (d) hay que hacer un procesamiento común
(b) Construya el modelo que propone, escriba los coeficientes estimados en la tabla que se adjunta.
(c) Analice el efecto de las variables de pronóstico sobre la variable de respuesta del modelo. Escriba los valores-p para analizar cada variable en la tabla que se adjunta.
(d) Analice el ajuste del modelo. Represente la concordancia de las estimaciones con la realidad en la tabla que se adjunta

Lo primero es construir un factor para ingresar al modelo de regresión la variable cualitativa rango. Para una mayor facilidad en la interpretación poner como valor de referencia el código 4 que corresponde con la escuela de menor prestigio

```
> attach(adm)
> rangoF=factor(rango)
> rangoF=relevel(rangoF,ref="4")
> contrasts(rangoF)
  1 2 3
4 0 0 0
1 1 0 0
2 0 1 0
3 0 0 1
```

Después ajustamos (construimos) el modelo de regresión logística

```
> r=glm(admit ~ exam + prom + rangoF,family="binomial")
> summary(r)
```

Base de datos: datosP
En todas las preguntas utilice un error de 0.05

Respuesta pregunta (b)

```
Coefficients:
             Estimate
exam         0.002264
prom         0.804038
rangoF1      1.551464
rangoF2      0.876021
rangoF3      0.211260
```

Respuesta pregunta (c)
Primeramente se hace la prueba de las desviaciones:

H_0: las variables de pronóstico no influyen sobre el ser o no admitido
H_A: algunas de las variables influyen

```
    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
```

```
> #desviaciones
> d0=499.98
> dR=458.52
> #valor-prueba desviaciones
> 1-pchisq(d0-dR,5)
[1] 7.574756e-08
```

Valor-p: 0.0000 < 0.05, se rechaza H_0 por lo que se observan evidencias de que al menos una de las variables puede mostrar influencia sobre el ser admitido a la Universidad

Para el análisis individual se aplica la prueba de Wald, para cada coeficiente (i)

$$H_0: \beta_i = 0 \qquad H_A: \beta_i \neq 0$$

Valores-p:

variable	Valor-p
exam	0.038465
prom	0.015388
rangoF1	0.000205
rangoF2	0.016908
rangoF3	0.590748

Las variables que muestran influencia, valor-p < 0.05, son: exam, prom, rangoF1, rango F2. La variable rangoF3, que su coeficiente compara las escuelas de categoría 3 con las de categoría 4 no muestra efecto.

Respuesta pregunta (d)

Para el ajuste, se analiza el coeficiente de determinación de McFadden

Base de datos: datosP

En todas las preguntas utilice un error de 0.05

```
> r0=glm(admit~1,family="binomial")
> #calculo del coeficiente
> 1-(logLik(r)/logLik(r0))
'log Lik.' 0.08292194 (df=6)
```

Conjuntamente se elabora la table con las concordancias del modelo con la realidad al predecir el comportamiento de las admisiones

```
      clasif
admit   0   1
    0 254  19
    1  97  30
```

De los 400 casos considerados en (254+30) = 284 fueron bien clasificados (71%) y el resto 400-280= 116 (29%, cerca de la tercera parte) fueron clasificados erróneamente al aplicar el modelo de regresión logística construido.

Teniendo en cuenta la cantidad de errores al clasificar y que el coeficiente de McFadden es bajo se puede afirmar que el ajuste del modelo no es bueno.

(e) Utilice su modelo para estimar si una persona con una puntuación de 400 en el examen de ingreso, con un promedio de 3 en los estudios de bachillerato, y que proviene de una escuela de categoría 4 sería admitida en la universidad X. ¿Será confiable esta estimación que usted está haciendo?

Respuesta pregunta (e)

Para el pronóstico:

```
> a=data.frame( exam=400, prom=3, rangoF="4")
> predict(r,a,type="response")
         1
0.09765467
```

La probabilidad de ser admitido se estima sería del 9.7%

Esta estimación no es confiable porque el modelo no tiene un buen ajuste

Base de datos: adm

Base de datos: datosP

En todas las preguntas utilice un error de 0.05

Pregunta 1 **(2.5 puntos)** Un equipo de investigación pedagógica de un país X desea identificar las variables que influyen en el número de premios que reciben los estudiantes en el bachillerato (variable num), específicamente desean saber si las variables:

prog: tipo de programa en el que el estudiante matricula. Categorías: 1 (general), 2 (académico), 3 (vocacional)

mat: puntaje que caracteriza su desempeño en matemáticas

influyen sobre el número de premios. En caso afirmativo desean construir un modelo de regresión que represente la relación

(a) Usted decide ajustar un modelo de regresión para analizar el problema, ¿qué tipo de modelo de regresión ajustaría? Justifique su respuesta

Respuesta pregunta (a)

La variable de respuesta es un conteo por lo que se aplica la regresión de Poisson

Para las preguntas (b), (c), (d) y (e) hay que hacer un procesamiento común

(b) Construya el modelo que propone, escriba los coeficientes estimados en la tabla que se adjunta.

(c) Analice el efecto de las variables de pronóstico sobre la variable de respuesta del modelo. Escriba los valores-p para analizar cada variable en la tabla que se adjunta.

(d) ¿Cuál sería el cambio esperado en el número de premios si se incrementa la nota en matemáticas en 5 puntos? ¿cuál sería el cambio esperado en el número de premios si compara una persona que pertenece al programa académico con una del programa general? ¿y si compara una persona del programa vocacional con una del programa general?

(e) Analice el ajuste del modelo

Lo primero, construir un factor, para ingresar la variable cualitativa (nominal) prog.

```
> attach(premios)
> progF=factor(prog)
```

Después ajustamos (construimos) el modelo de regresión logística

```
> r=glm(num ~ progF + mat,family="poisson" )
> summary(r)
```

Respuesta pregunta (b)

```
Coefficients:
             Estimate

progF2        1.08386
progF3        0.36981
mat           0.07015
---
```

Respuesta pregunta (c)

Primeramente se hace la prueba de las desviaciones:

H_0: las variables de pronóstico no influyen sobre el ser o no admitido

H_A: algunas de las variables influyen

Base de datos: datosP

En todas las preguntas utilice un error de 0.05

```
    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
```

```
> #prueba desviaciones
> 1-pchisq(287.67-189.45,3)
[1] 0
```

Valor-p: 0.0000 < 0.05, se rechaza H_0 por lo que se observan evidencias de que al menos una de las variables puede mostrar influencia sobre el ser admitido a la Universidad

Para el análisis individual se aplica la prueba de Wald, para cada coeficiente (i)

$$H_0: \beta_i = 0 \qquad H_A: \beta_i \neq 0$$

Valores-p:

variable	Valor-p
progF2	0.00248
progF3	0.40179
mat	3.63e-11

Las variables que muestran influencia, valor-p < 0.05, son: mat y progF2, la variable progF3 (cuyo coeficiente compara la matricula vocacional con la general) no muestra influencia

Respuesta pregunta (d)

Cambio esperado:

$$e^{\beta_1 \Delta X}$$

¿Cuál sería el cambio esperado en el número de premios si se incrementa la nota en matemáticas en 5 puntos?

$$e^{0.07*5} = 1.41$$

Al incrementar la nota en matemáticas 5 puntos el número esperado de premios incremente 1.41 veces

¿cuál sería el cambio esperado en el número de premios si compara una persona que pertenece al programa académico con una del programa general?

$$e^{1.08} = 2.94$$

Al comparar el programa académico con el general el número esperado de premios se incrementaría 2.94 veces

¿y si compara una persona del programa vocacional con una del programa general?

Como el coeficiente del programa vocacional no puede asumirse que sea diferente de 0, por ese motivo el programa vocacional no muestra diferencias con el general. Es decir, el número medio de premios no cambia al compararlos (manteniendo constante la puntuación en matemáticas)

Respuesta pregunta (e)

Base de datos: datosP

En todas las preguntas utilice un error de 0.05

```
> #modelo con intercepto
> r0=glm(num~1,family=poisson)
> 1-(logLik(r)/logLik(r0))
'log Lik.' 0.2118112 (df=4)
```

El coeficiente McFadden está en el valor más bajo del umbral de aceptabilidad. El ajuste no es muy bueno

(f) Estime la probabilidad de que gane dos premios una persona que matricula el programa académico y que tenga una nota en matemáticas de 43. ¿Será confiable esta estimación que usted está haciendo?

```
> a=data.frame(progF="2",mat=43)
> p=predict(r,a,type="response")
> #valor de la probabilidad
> p
        1
0.3176795
> #prob de recibir 2 premios
> dpois(2,p)
[1] 0.03672671
```

La estimación no es muy confiable. El ajuste es aceptable pero en el nivel más bajo

Base de datos: premios