Ejercicio 2_

Equipo

2024-03-26

2. Modelos lineales generalizados para datos continuos

Consideraremos la base de datos `Preg1B.csv` con información sobre 295 pacientes seleccionados de forma aleatoria. Se desea analizar si existe una asociación entre la presión arterial sistólica (bpsystol) y el índice de masa corporal (bmi), considerando el sexo (sex: 1-hombre, 2-mujer, con hombre como referencia) y la edad (age) de los pacientes.

i. Explorando los diferentes modelos lineales generalizados comúnmente usados cuando la variable dependiente es continua (normal, gamma, inversa gaussiana), presente un modelo que le parezca adecuado para modelar E(bpsystol; bmi, sex, age). Considere por simplicidad que no hay interacción entre las covariables del modelo. Deberá indicar con claridad cuál es la expresión matemática que se usa para modelar E(bpsystol; bmi, sex, age), así como describir el procedimiento y criterio usado para seleccionar el modelo.
ii. Repita los incisos iii) y iv) de la pregunta 1 con el modelo en i).
iii. Comparando el modelo en i) con el usado en la pregunta 1, compare las conclusiones e interpretaciones que se pueden obtener e indique qué modelo prefiere usar. Argumente con claridad su respuesta, por ejemplo, debe incluir los valores de AIC o BIC, así como ventajas y desventajas en la interpretación.

i) Explorando modelos con variable dependiente contínua.

Para presentar un modelo que parezca adecuado para modelar E(bpsystol; bmi, sex, age), exploramos una malla de los diferentes modelos lineales generalizados comúnmente usados: para el componente aleatorio cuando la variable dependiente es continua exploramos las distribuciones normal, gamma, e inversa gaussiana; empleamos distintas funciones ligas tales como la inversa, identidad, logarítmica, y 1/mu^2(solo para IG); y consideramos el componente lineal tanto de potencias (-3, -2.5, . . . , 2.5, 3) como de polinomios (grado 1 al 5). Consideramos por simplicidad que no hay interacción entre las covariables del modelo. En el siguiente Cuadro se muestraa el mejor modelo, con el menor AIC de 2484.009 (que coincide con el mejor modelo por su BIC de 2502.443), con la siguiente estructura:

glm(formula = bpsystol ~ age+sex+I(bmi^(1.5)), family = inverse.gaussian(link = identity), data = datos).

```
##
## Call:
## glm(formula = form, family = Dist(link = FunLigas[l]), data = datos)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.16891    3.90205  23.108  < 2e-16 ***
## age           0.48671    0.05732   8.491 1.06e-15 ***
## sex2         -7.05833    1.90756  -3.700 0.000258 ***
## I(bmi^(1.5))  0.15131    0.02619   5.777 1.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.0001280556)
##
##     Null deviance: 0.053716  on 294  degrees of freedom
## Residual deviance: 0.036270  on 291  degrees of freedom
## AIC: 2484
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Family: inverse.gaussian
## Link function: identity
```

```
## [1] 2484.009
```

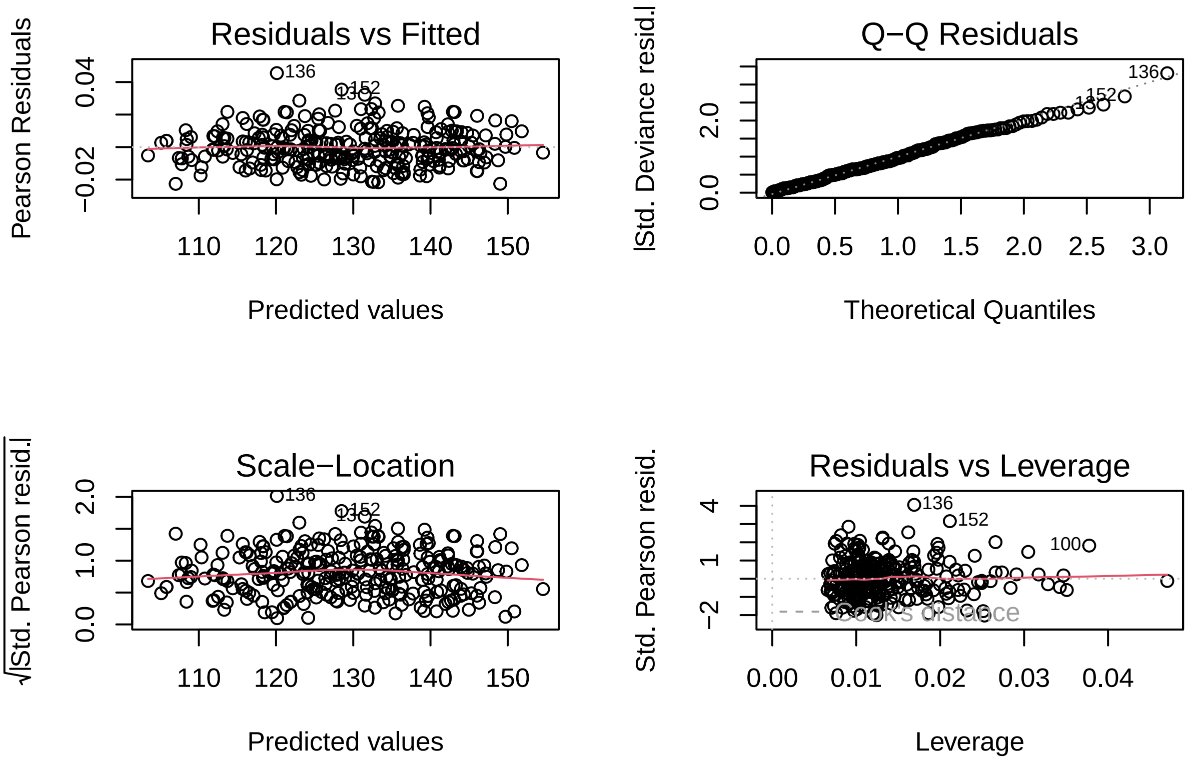
```
## [1] 2502.443
```

```
## [1] "bpsystol ~ age+sex+I(bmi^(1.5))"
```

Sin embargo, se elige el modelo más simple o parsimonioso sin el exponente de 1.5 para la variable bmi, pues al considerar bmi sin modificación se obtiene un AIC de 2484.1, el cual no parece ser muy diferente a 2484.009. En el siguiente Cuadro se muestra el modelo final elegido, cuya expresión matemática es xx.

```
##
## Call:
## glm(formula = bpsystol ~ age + sex + bmi, family = inverse.gaussian(link = identity),
##     data = datos)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.02163    5.25361  15.232  < 2e-16 ***
## age          0.48269    0.05744   8.403 1.95e-15 ***
## sex2        -6.88649    1.90588  -3.613 0.000356 ***
## bmi          1.17620    0.20261   5.805 1.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.0001279792)
##
##     Null deviance: 0.053716  on 294  degrees of freedom
## Residual deviance: 0.036286  on 291  degrees of freedom
## AIC: 2484.1
##
## Number of Fisher Scoring iterations: 5
```

En las siguientes gráficas podemos observar en **Residuals vs Fitted** que se conserva la linealidad, en **Q-Q Residuals** se observa un buen comportamiento de la normalidad de los errores, en **Scale-Location** se observa que hay homoscedasticidad, y en **Residuals vs Leverage** parece no haber outliers influyentes.




```
##
## Attaching package: 'flextable'

## The following objects are masked from 'package:kableExtra':
##
##     as_image, footnote

## The following object is masked from 'package:purrr':
##
##     compose

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine

## Suggested APA citation: Thériault, R. (2023). rempsyc: Convenience functions for psychology.
## Journal of Open Source Software, 8(87), 5466. https://doi.org/10.21105/joss.05466
```

	1
Normality (Shapiro-Wilk)	0.001
Homoscedasticity (Breusch-Pagan)	0.095
Autocorrelation of residuals (Durbin-Watson)	0.981

ii) Asociación entre masa corporal y presión arterial sistólica, y estimación puntual.

En esta sección describiremos por una parte la asociación entre masa corporal y presión arterial sistólica y la prueba de hipótesis de esta relación. Por otra presentaremos una gráfica resumen con la estimación puntual de la relación bpsystol y bmi, considerando edades de 30, 50 y 64, así como la diferenciación entre mujeres y hombres.

iii) Comparativo modelo de regresión lineal múltiple contra modelo lineal generalizado.

En esta sección se compara modelo de regresión lineal múltiple del ejercicio anterior (ejercicio 1) contre el modelo lineal generalizado con base en sus AIC. Además, compararemos las conslusiones e interpretaciones de ambos modelos, para indicar cuál nos parece más adecuado.