2. Modelos lineales generalizados para datos continuos

Tarea-examen 1: Aprendizaje estadístico supervisado

Carlos Iván Canto Varela 315649888

En este caso se quiere representar los datos mediante un modelo lineal generalizado para variable continua.

Para esto se revisaron 7840 combinaciones distintas, en las cuales se experimentaron con 28 transformaciones (potencia, logaritmo y polinomial ortogonal) para cada variable independiente continua, tres distribuciones de variable continua (gaussiana, gamma e inversa gaussiana) y sus funciones liga disponibles (identidad, logarítmica, inversa y $1/\mu^2$).

De todos los modelos, el mejor en tanto a AIC (2481.45) y BIC (2499.885) fue aquel de distribución inversa gaussiana y liga identidad con transformaciones en potencia para las variables continuas. El modelo tendrá la siguiente forma:

$$
\begin{aligned}
y \sim \mathrm{IG}(\mu, \lambda) &\approx \mathrm{IG}(\mu, 7909.2) : \\
\mu &= \beta_0 + \beta_1 \mathrm{bmi}^{3/2} + \beta_2 \mathrm{age}^2 + \beta_3 \mathrm{sex}_2, \\
\mathrm{V}(y_i) &= \frac{\mu_i^3}{\lambda} \approx \frac{\mu_i^3}{7909.2}.
\end{aligned}
$$

Adicionalmente, los supuestos del componente aleatorio y función liga se aprueban satisfactoriamente.

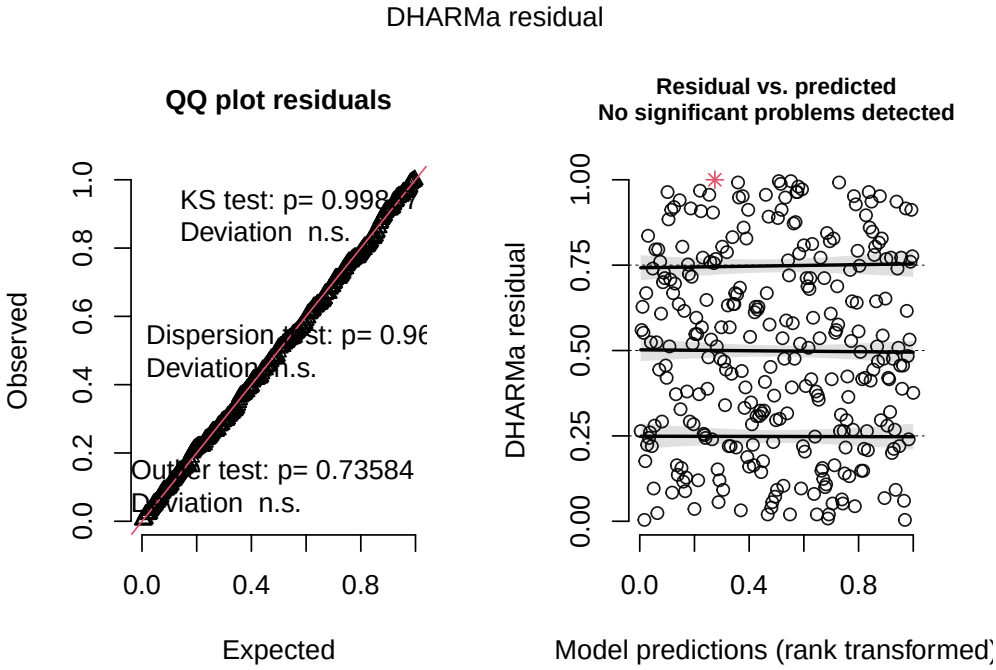

Figura 1: Gráficas de residuales simulados para los supuestos del modelo

La prueba a utilizar para revisar el argumento indicado es una Wald:

$$H_0: \quad \beta_1 \leq 0 \quad \text{vs.} \quad H_a: \quad \beta_1 > 0.$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = bpsystol ~ I(bmi^(1.5)) + I(age^(2)) + sex, family = inverse.gaussian(link = iden
##     data = Data)
##
## Linear Hypotheses:
##        Estimate Std. Error z value   Pr(>z)
## 1 <= 0  0.15299    0.02593   5.901 1.81e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Nuevamente hay suficiente evidencia para rechazar la hipótesis nula con una significancia del 0.05 y se puede asumir que -para una persona de cierta edad y sexo- tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica.

Para apoyar el entendimiento del modelo, se pueden ver las gráficas siguientes para tres edades particulares:

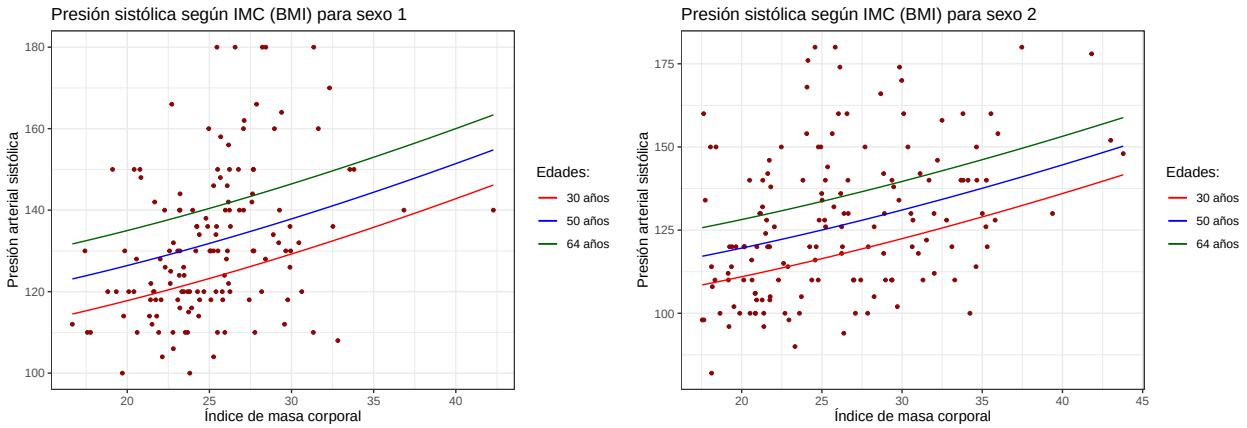

Figura 2: Modelo relacional de presión arterial sistólica con IMC para edades particulares (modelo GLM)

Se encuentra que las funciones evaluadas en los valores de las edades son crecientes, lo que es congruente con la prueba de hipótesis. Nuevamente, hay una tendencia no confirmada entre mayor presión sistólica y edad más avanzada en ambos sexos. Nótese que las regresiones graficadas no son más que un pequeño subconjunto de todas las edades.

En comparación con el modelo de regresión lineal múltiple de la pregunta anterior, en este se tratará a la media y esto puede ser un inconveniente en la práctica multidisciplinaria. Sin embargo, este modelo promete una interpretación más simple en cuanto a la variable dependiente dado que no fue transformada; además que llegó a un mejor puntaje AIC:

Puntaje AIC por modelo	
Modelo	**Puntaje AIC**
RLS con transf. Box-Cox	2485.747
RG dist. inversa gaussiana	**2481.45**

Con esto, se puede argumentar que el modelo más conveniente es el generalizado.