1. Regresión lineal múltiple

Tarea-examen 1: Aprendizaje estadístico supervisado

Carlos Iván Canto Varela, 315649888

Los datos proporcionados tienen la visualización siguiente:

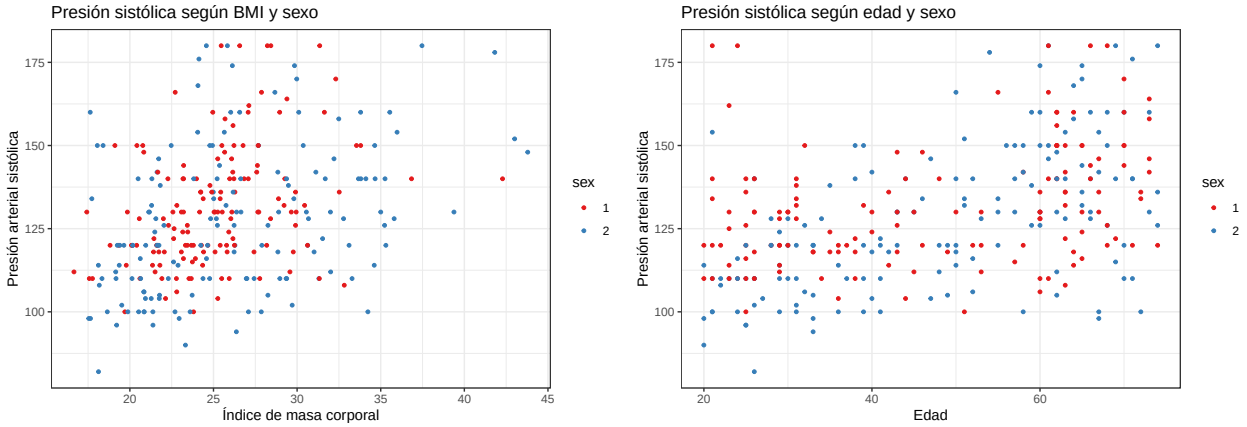

Figura 1: Diagramas de dispersión por covariable

La intención es ajustar un modelo de regresión lineal múltiple sin interacciones.

Con este, un interés primario es verificar el supuesto de linealidad:

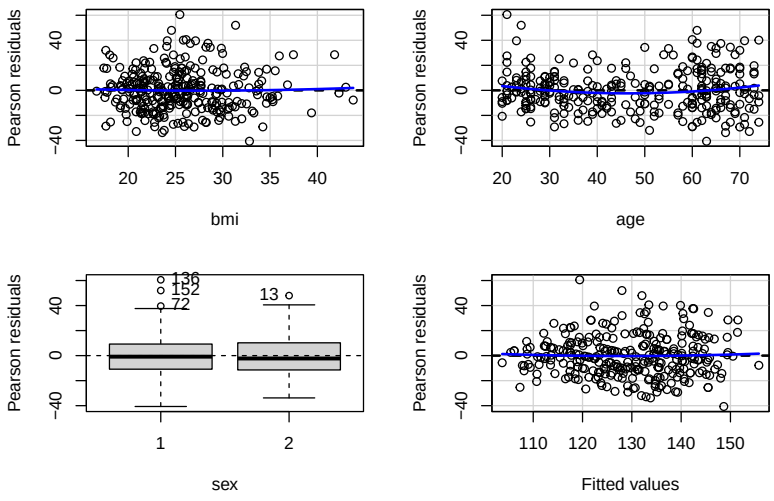

Figura 2: Gráficas de residuales para el supuesto de linealidad

Se puede apreciar que se cumple. Sin embargo, otro tipo de modelo lineal podría ser de mayor utilidad pues los diagramas de dispersión parecen apuntar hacia la no homocedasticidad.

Similarmente, la normalidad no pinta bien con el diagrama Q-Q en las colas:

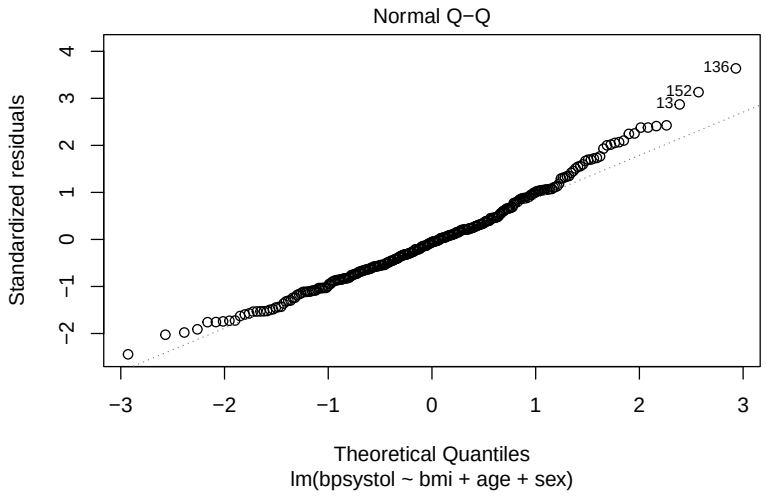

Figura 3: Gráfica Q-Q para el supuesto de normalidad

Para arreglar los supuestos erróneos, se procede a hacer una tranformación Box-Cox.

Con el nuevo modelo ajustado, los supuestos fallidos regresan a ser válidos con significancia de 0.05,

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8359438, Df = 1, p = 0.36056
```

```
## Non-constant Variance Score Test
## Variance formula: ~ bmi
## Chisquare = 0.001760642, Df = 1, p = 0.96653
```

```
## Non-constant Variance Score Test
## Variance formula: ~ age
## Chisquare = 2.87007, Df = 1, p = 0.090241
```

```
##
##  Shapiro-Wilk normality test
##
## data:  BCData$.std.resid
## W = 0.9957, p-value = 0.5949
```

y se mantiene la linealidad.

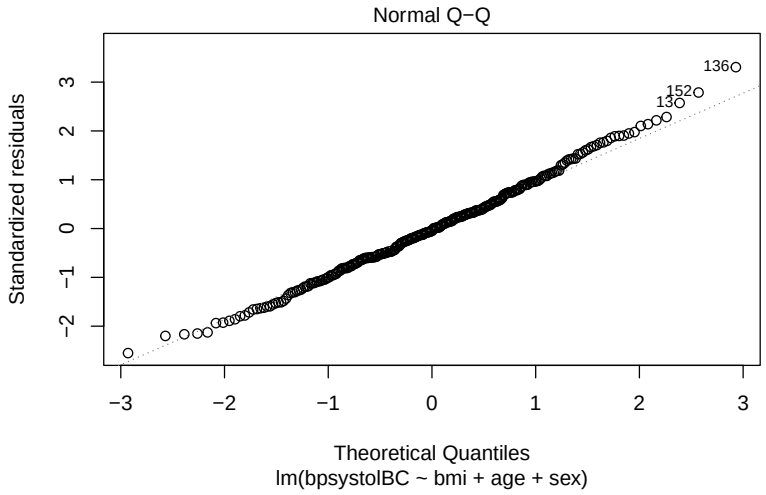

Figura 4: Gráfica Q-Q para el supuesto de normalidad en el modelo transformado

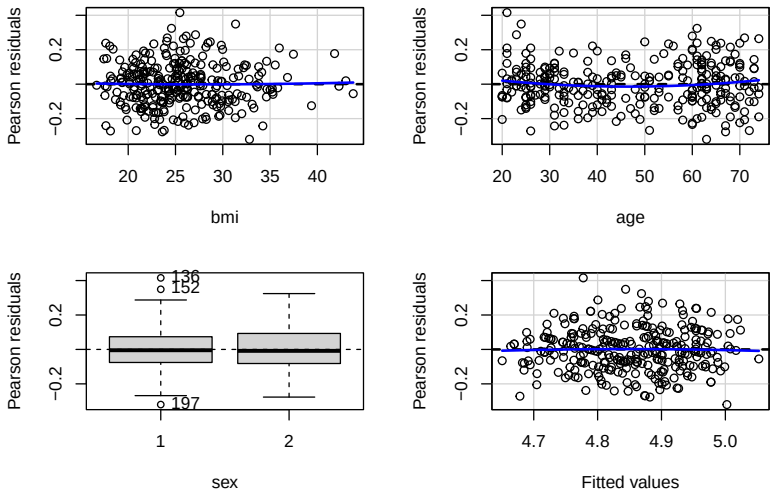

Figura 5: Gráficas de residuales para el supuesto de linealidad en el modelo transformado

Para encontrar una asociación entre alta presión arterial sistólica y un índice de masa corporal elevado, es necesario realizar una prueba de hipótesis. Sea $y^* = \ln(y)$, el modelo escogido tendrá la forma

$$\mathrm{E}\left[y^*\right] = \beta_0 + \beta_1 \mathrm{bmi} + \beta_2 \mathrm{age} + \beta_3 \mathrm{sex}_2$$

donde el nivel base para la variable categórica sex es el 1. Así la prueba buscada es

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

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = bpsystolBC ~ bmi + age + sex, data = Data)
##
## Linear Hypotheses:
##        Estimate Std. Error t value  Pr(>t)
## 1 <= 0 0.009365   0.001520   6.161 1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Dado que el valor p encontrado es menor a una significancia del 5 %, hay suficiente evidencia para rechazar la hipótesis nula y asumir que tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica.

Como complemento -para las edades de 30, 50 y 64 años- se asocian las siguientes curvas:

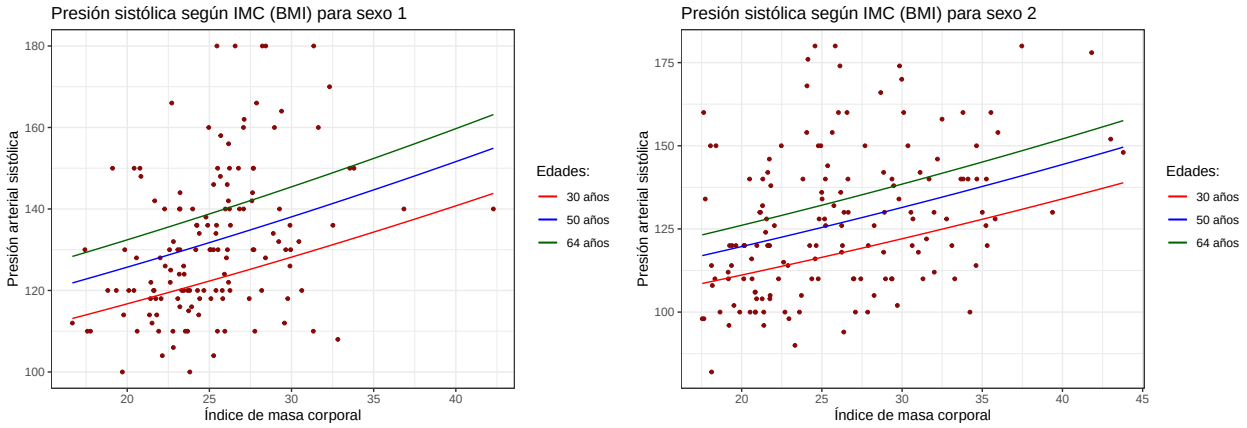

Figura 6: Modelo relacional de presión arterial sistólica con IMC para tres edades particulares

A primera instancia, los puntos no parecen estar explicados por las rectas en la escala original. Sin embargo, es importante recordar que éstas sólo representan tres de todas las edades muestreadas.

Todas las rectas son crecientes; esto apoya gráficamente a la prueba de hipótesis realizada previamente. Además se percibe un paralelismo entre las rectas para cada edad hallada, pero con valores más altos conforme a mayores edades en ambos sexos. Esto es un indicador para una hipótesis análoga a la probada, pero con las edades.