Ejercicio 1: Regresión lineal múltiple

La base de datos Preg1B.csv contiene 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), en particular, si es posible observar que tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica. Para realizar este análisis se indica que se considere el sexo (sex: 1-hombre, 2-mujer) y la edad (age) de los pacientes, pues la presión arterial sistólica podría variar de acuerdo con estos factores.

Datos

Importamos nuestos datos y analizamos la relación de la presión arterial sislotica con cada una de las covariables.

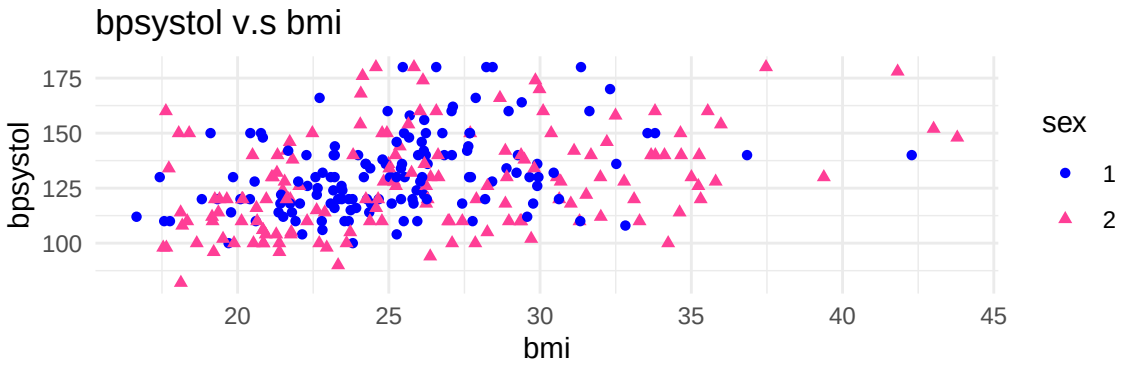

Figura 1: Relacion de los datos (bpsystol y bmi)

Analizando datos en la Figura 1 podemos ver que conforme van aumentando los valores del indice de masa corporal, van aumentando los valores de la presión arterial sistolica, sin embargo se hará un analisis aparte para argumentar a favor o en contra de es afirmación.

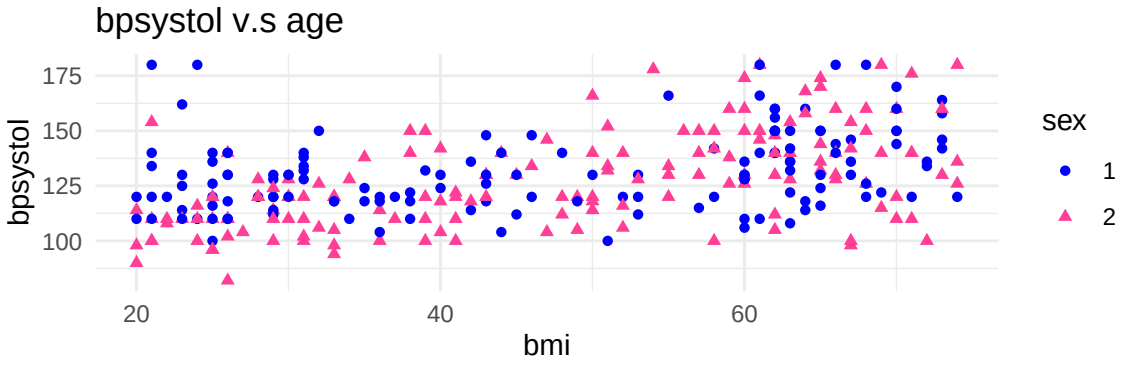

Figura 2: Relacion de los datos (bpsystol y age)

Analizando los datos en la Figura 1 , vemos que la edad tambien tiene relación con la presion arterial sistolica, ya que para ambos sexos, vemos un comportamiento creciente, es decir conforme aumenta la edad de las personas, también aumenta su presión arterial sistolica.

I. Ajuste el modelo de regresión lineal múltiple para E(bpsystol; bmi,sex, age)

Tenemos que ajustar un modelo de regresión lineal multilple de la siguiente forma

$$\mathbb{E}(y; bmi, age, sex) = \beta_0 + \beta_1 \cdot (bmi) + \beta_2 \cdot (age) + \beta_3 \cdot (sex)$$

Por lo que la expresión del modelo ajustado de regresión lineal multiple para E(bpsystol; bmi,sex, age) es

$$\mathbb{E}(bpsystol; bmi, sex, age) = 84.1598805 + 1.2082201 \cdot (bmi) + 0.4840862 \cdot (age) - 5.6637222 \cdot (sex)$$

De acuerdo a los datos obtenidos, observamos que la prueba de hipotesis referente a la tabla ANOVA, H_0 se rechaza, ya que el $p-vale = 2.2e-16 < 0.05$ es decir que el modelo tiene sentido y que al menos una de las variables explicativas (bmi, age o sex) aporta información a nuestro modelo.

Ahora como nuestro objetivo es la estimación, es necesario realizar la verificación de supuestos

Realizando la verificación de supuestos de nuestro primer modelo notamos que:

Linealidad

Para verificar el supuesto de linealidad, se usaron dos metodos, uno mediante un metodo grafico, donde y mediante pruebas de hipotesis. Con base a estos metodos concluimos que no se encontró evidencia en contra del supuesto de linealidad, analizando las graficas podemos ver que a simple vista no se encontró algun patron que nos de indicio a pensar que el supuesto de linealidad no se cumple, además en las pruebas de hipotesis en todas (tanto en la linealidad general como en la linealidad por cada covariable), el p-value es mayor a la significancia, por lo que no encontramos evidencia en contra del supuesto de linealidad.

Homocedasticidad

Hacemos la prueba de hipotesis donde se contrasta: H_0: la varianza es constante vs H_a: la varianza depende

Para la prueba de hipotesis global o general de modelo, se obtuvo un $p-value = 0.016013 < 0.05$ por lo que se rechaza H_0 Para las pruebas de hipotesis individuales, tenemos que para la covariable "age", $p - value = 0.0073463 < 0.05$ por lo que tambien encontramos evidencia en contra de este supuesto para esta covariable. Para el supuesto de homocedasticidad tenemos que encontramos evidencia en contra del supuesto de homocedasticidad, es decir que la varianza depende y no es constante, además las pruebas nos dicen que la variable que está asociada con la varianza es la varible "age".

Normalidad

Para verificar el supuesto de normalidad, realizaremos pruebas de hipotesis donde se contraste lo siguiente:

H_0: los datos provienen de la distribucion normal vs H_a: los datos **NO** provienen de la distribución normal

Al realizar las dos pruebas de hipotesis, primero usando "Shapiro-Wilk normality test", se obtiene un $p - value = 0.0009943 < 0.05$, ahora usando "Lilliefors (Kolmogorov-Smirnov) normality test" se obtiene un $p-value = 0.01641 < 0.05$ Con base a los resultados de las pruebas de hipotesis, el p-value es menor a la significancia por lo que encontramos evidencia en contra del supuesto de normalidad, de igual forma con graficamente, los datos parecen no ajustarse a la recta.

Algunas conclusiones de la verificación de supuestos

Se encontró evidencia en contra del supuesto de homocedasticidad y de normalidad por lo que el modelo no es adecuado y optaremos por realizar transformaciones a las variables o realizar una regresión ponderada.

II. Tranformación de variables

Dado que no se cumple el supuesto de homocedasticidad, es necesario hacer una tranformación en las variables, un camino rapido seria transformar la variable "bpsystol", sin embargo tendriamos un tema con la interpretación si es que lo que queremos hacer es en terminos de la media, por lo que tomaremos el camino de la regresión ponderada.

Con base a lo que se obtuvo en la regresión ponderada, el modelo con menor AIC, fue

$$\mathbb{E}(bpsystol; bmi, age, sex) = \beta_0 + \beta_1 \cdot (bmi) + \beta_2 \cdot (age) + \beta_3 \cdot (sex)$$

con peso: $1/age^{0.5}$

Ahora es necesario saber si este modelo cumple con los supuestos

Linealidad

Vemos de acuerdo a los metodos para verificar el supuesto de linealidad (graficas y pruebas de hipotesis) se puede concluir que el supuesto de linealidad se sigue cumpliendo de forma global y por cada covariable.

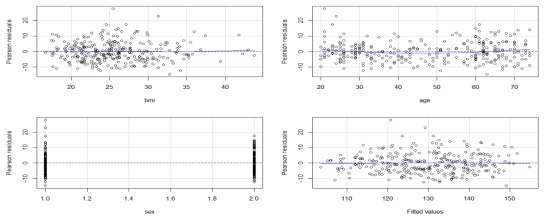

Figura 3: Grafica para verificar supuesto de linealidad

Con base en la Figura 3 a simple vista no se encontró algun patron que nos de indicio a pensar que el supuesto de liealidad no se cumple

Homocedasticidad Hacemos la prueba de hipotesis donde se contrasta: H_0: la varianza es constante vs H_a: la varianza depende

Realizando las pruebas de hipotesis, tanto para la prueba de hipotesis global o general de modelo, como las pruebas de hipotesis individuales de cada covariable se obtuvo que el p-value es mayor a la significancia por lo que sería plaucible decir que no se encontró evidencia en contra del supuesto de homocedasticidad.

Normalidad

Para verificar el supuesto de Normalidad, tambien podemos hacerlo mediante la QQ-plot

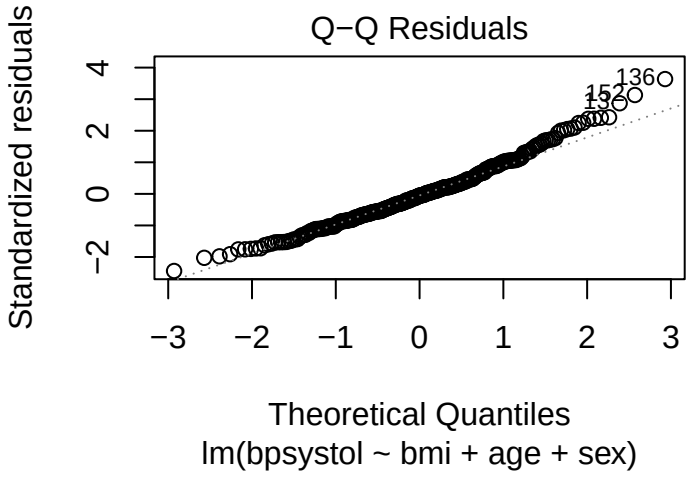

Figura 4: QQ-plot para verificar supuesto de normalidad

Con base en la Figura 4 la mayoria de los datos parecen formar una linea recta, por lo que podriamos decir que se cumple el supuesto de normalidad

En conclusión este modelo de regresión lineal multiple ponderada cumple con los supuestos.

Por lo que la expresión del modelo ajustado de regresión lineal multiple ponderada para E(bpsystol; bmi,sex, age) es

$$\mathbb{E}(bpsystol; bmi, sex, age) = 87.1032817 + 1.2110833 \cdot (bmi) + 0.4663687 \cdot (age) - 7.0790096 \cdot (sex)$$

III. ¿Se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica?

Lo que quiere decir es que para cualquier persona de cualquier sexo, entre más aumente la masa corporal, la persión arterial sistolica tambien aumentará, lo que lo podemos traducir a

si fijamos la edad a un valor x y al sexo a una variable tipo k y sea $a > b$ entonces se debe cumplir:

$$\mathbb{E}(bpsystol; bmi = a, age = x, sex = k) > \mathbb{E}(bpsystol; bmi = b, age = x, sex = k)$$

$$\beta_0 + \beta_1 \cdot (a) + \beta_2 \cdot (x) + \beta_3 \cdot (k) > \beta_0 + \beta_1 \cdot (b) + \beta_2 \cdot (x) + \beta_3 \cdot (k)$$
$$\beta_1 \cdot (a) > \beta_1 \cdot (b) \Longrightarrow \beta_1 \cdot (a) - \beta_1 \cdot (b) > 0 \Longrightarrow \beta_1 \cdot (a - b) > 0$$

Como $a > b \Longrightarrow a - b > 0$

Entonces para que se cumpla la afirmación "tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica", necesitamos que $b_1 > 0$

Entonces se realizara la prueba de hipotesis (Ponemos lo que nos interesa en la hipotesis alternativa para acotar el error tipo 2)

$$H_0 : \beta_1 \leq 0 \quad v.s \quad H_a : \beta_1 > 0$$

Podemos ver que el $p - value = 1.14e - 09 < 0.05$ es menor a la significancia, por lo que se rechaza H_0, es decir encontramos evidencia en contra de $\beta_1 \leq 0$, por lo que sería plaucible decir que $\beta_1 > 0$, es decir que la afirmación "tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica" es cierta.

IV. Estimación puntual

La expresión del modelo ajustado de regresión lineal multiple ponderada para E(bpsystol; bmi,sex, age) es

$$\mathbb{E}(bpsystol; bmi, sex, age) = 87.1032817 + 1.2110833 \cdot (bmi) + 0.4663687 \cdot (age) - 7.0790096 \cdot (sex)$$

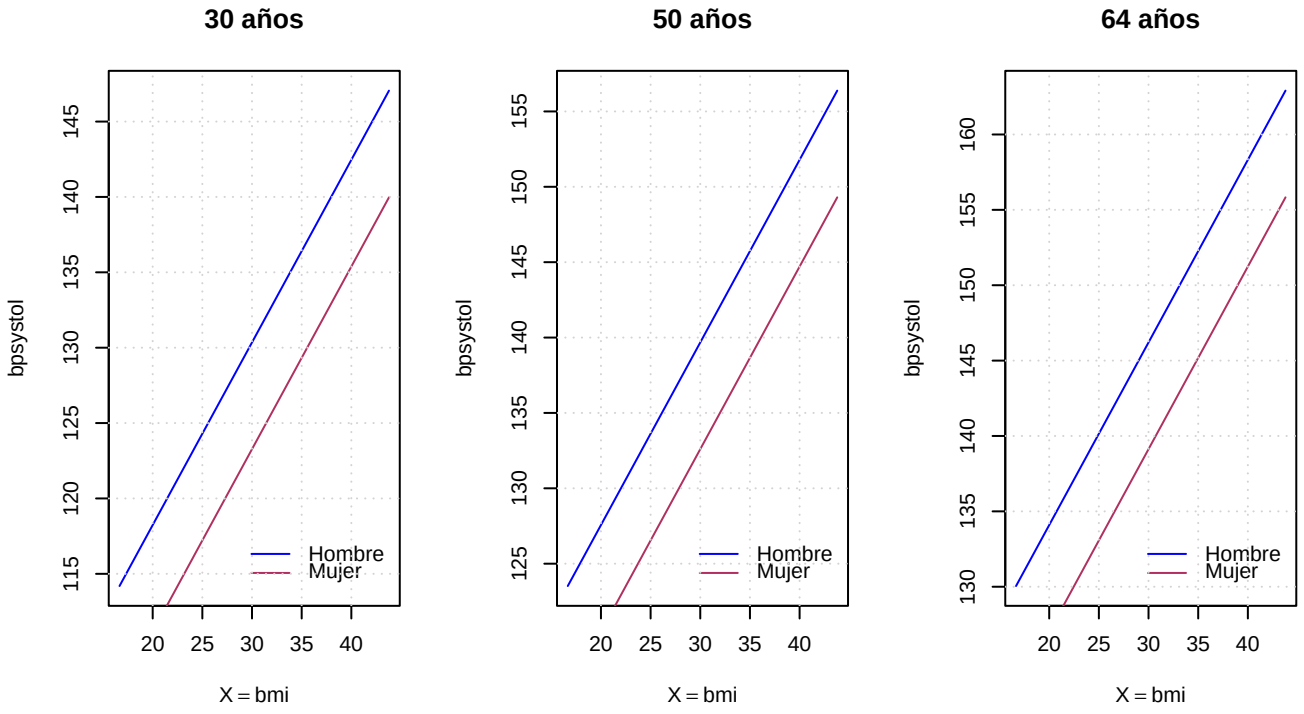

Figura 5: Estimación puntual asociada a la relación entre bpsystol y bmi, para edades 30, 50 y 64

En la Figura 5 se puede observar que para las edades observadas (30, 50 y 64) la presión arterial sistólica va creciendo conforme el índice de masa corporal (bmi) va aumentando, este comportamiento sucede para ambos sexos (Femenino y masculino). Ademas es evidente en las 3 graficas que la presión arterial sistólica es mayor en hombres que en mujeres ya que la recta de los hombres (color azul) siempre está por arriba de la recta de las mujeres (color rosa).

Finalmente podemos concluir que al ajustar un modelo de regresión lineal multiple ponderado, pudimos verificar por medio de pruebas de hipotesis que a medida que aumenta el indice de masa corporal de cualquier persona (de cualquer sexo y cualquier edad), la presion arterial sistolica igual aumenta, tambien graficamente se observó que la la presión arterial sistólica es mayor en hombres que en mujeres, además la presión arterial sistolica, va aumentando conforme va aumentando la edad de la persona.