Ejercicio 2

Equipo #

2024-03-31

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.

Usando la información de la pregunta 1 realizar lo siguiente:

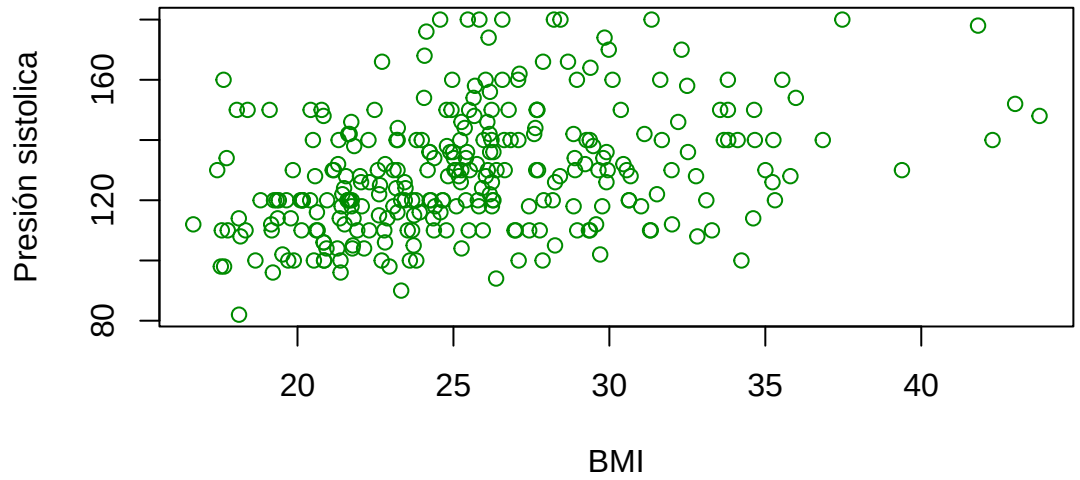
1.Explorando los diferentes modelos lineales generalizados comúnmente usados cuando la variable dependiente es continua (normal, gamma, inversa gaussiana), presente un modelo que le parezca adecuado para modelar E(bpsystol; bmi,sex, age). Considere por simplicidad que no hay interacción entre las covariables 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.

Antes de comenzar con la modelación y selección del modelo adecuado para nuestros datos veamos como están conformados los mismos.

```
##       ...1          bpsystol        bpdiast            bmi       
##  Min.   :  1.0   Min.   : 82.0   Min.   : 55.00   Min.   :16.66  
##  1st Qu.: 74.5   1st Qu.:114.0   1st Qu.: 74.00   1st Qu.:21.77  
##  Median :148.0   Median :128.0   Median : 80.00   Median :25.08  
##  Mean   :148.0   Mean   :129.3   Mean   : 82.39   Mean   :25.62  
##  3rd Qu.:221.5   3rd Qu.:140.0   3rd Qu.: 90.00   3rd Qu.:28.42  
##  Max.   :295.0   Max.   :180.0   Max.   :124.00   Max.   :43.79  
##       age             sex       
##  Min.   :20.00   Min.   :1.000  
##  1st Qu.:31.00   1st Qu.:1.000  
##  Median :48.00   Median :2.000  
##  Mean   :47.17   Mean   :1.519  
##  3rd Qu.:63.00   3rd Qu.:2.000  
##  Max.   :74.00   Max.   :2.000
```

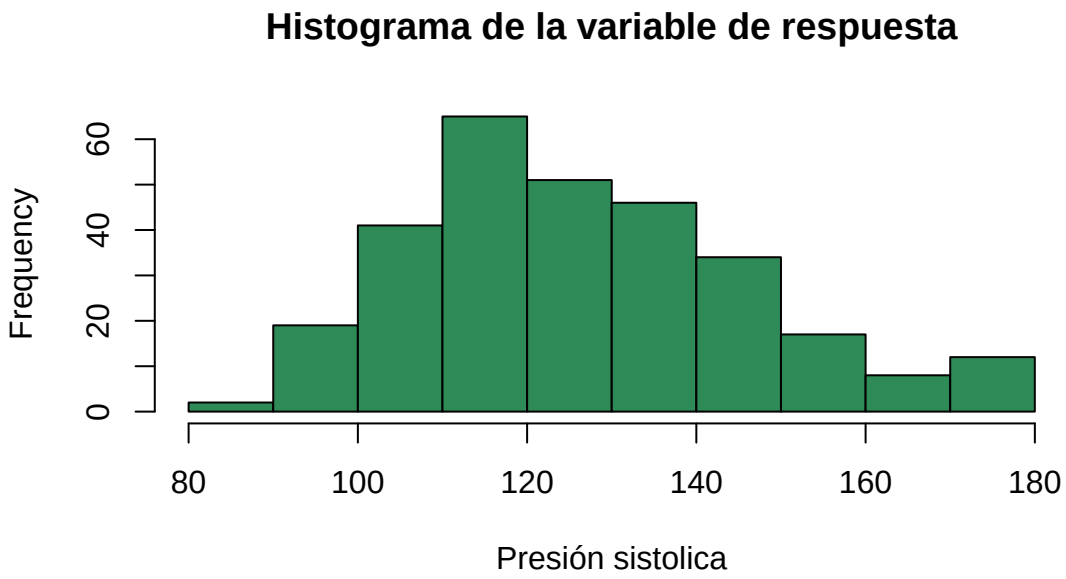
De la información que tenemos asumiremos que la variable bmi es un factor que se puede controlar lo cual nos llevaría a deducir que el bmi influye de manera directa en la presión sistolica de una persona, de lo contrario no podríamos hacer ninguna afirmación que avale que esto en verdad sucede.

Entonces veamos como se ven los datos.



De la gráfica podemos observar que la variable de la presión sistolica siempre es positiva y en general no se tiene una varianza con muchos cambios, es decir, los datos parecen agruparse con una varianza constante, ademas estos se encuentran mas agrupados del lado izquierdo, entre un rango de 10 a 3o bmi. Dado la información que obtuvimos consideraremos ajustar un GLM con familia inversa gaussiana y funcion liga identidad dado que los datos tienen una cola pesada del lado izquierdo.

Para estar más seguros de esto haremos un histograma de la variable de respuesta.



Ya que verificamos que la variable respuesta tiene una cola pesada procederemos a ajustas el GLM con la funcion liga mencionada anteriormente.

```
data1$sex <- as.factor(data1$sex)
```

```
##
## Call:
## glm(formula = bpsystol ~ bmi + sex + age, family = inverse.gaussian(link = "identity"),
##     data = data1)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.02163    5.25361  15.232  < 2e-16 ***
## bmi          1.17620    0.20261   5.805 1.68e-08 ***
```

```
## sex2        -6.88649    1.90588  -3.613 0.000356 ***
## age          0.48269    0.05744   8.403 1.95e-15 ***
## ---
## 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
```

Ya que ajustamos el modelo escribiremos de forma matemática como se ve el modelo:

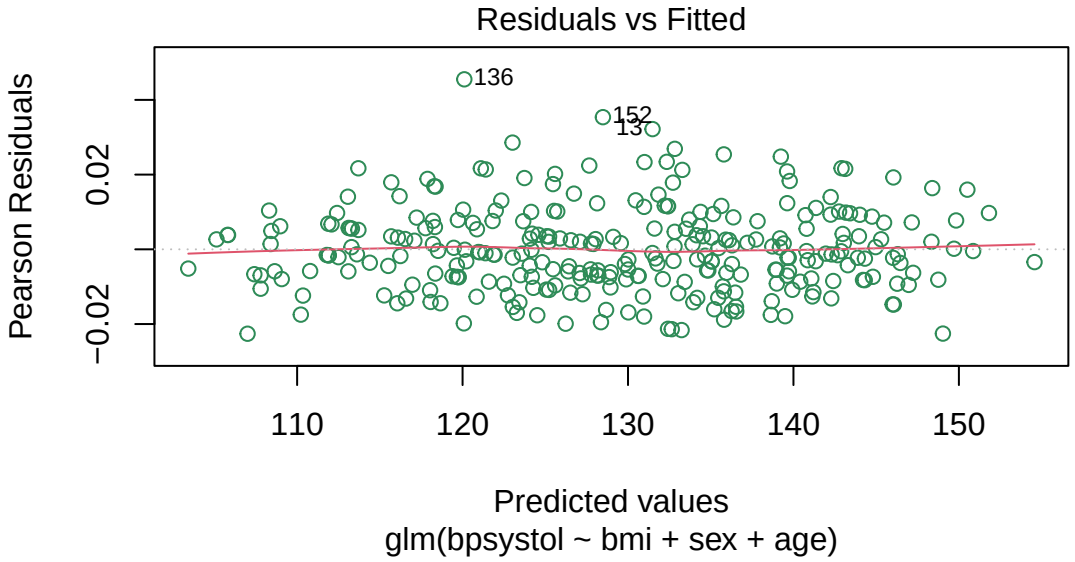
$$bpsystol_i = \beta_0 + \beta_1 BMI_i + \text{sex} + \text{age} \text{ con función liga: } \mu_i = g(\theta) = 1/\theta_i \text{ donde } \theta = \text{bpsystol}.$$

El criterio que usamos para usar este modelo se baso en la verificación de la distribución de la variable dependiente y así poder escoger que función liga seria mejor para ajustar el modelo, ademas se comparo el AIC de este modelo con otros 2 modelo en los que se cambiaba la función liga, de dicha comparación fue este modelo el que mejor AIC obtuvo de los 3.

Verificacion de supuestos

Linealidad

Para empezar hagamos una gráfica con la que veremos como se comporta la linealidad de nuestro modelo y obtengamos el valor promedio de los residuos.



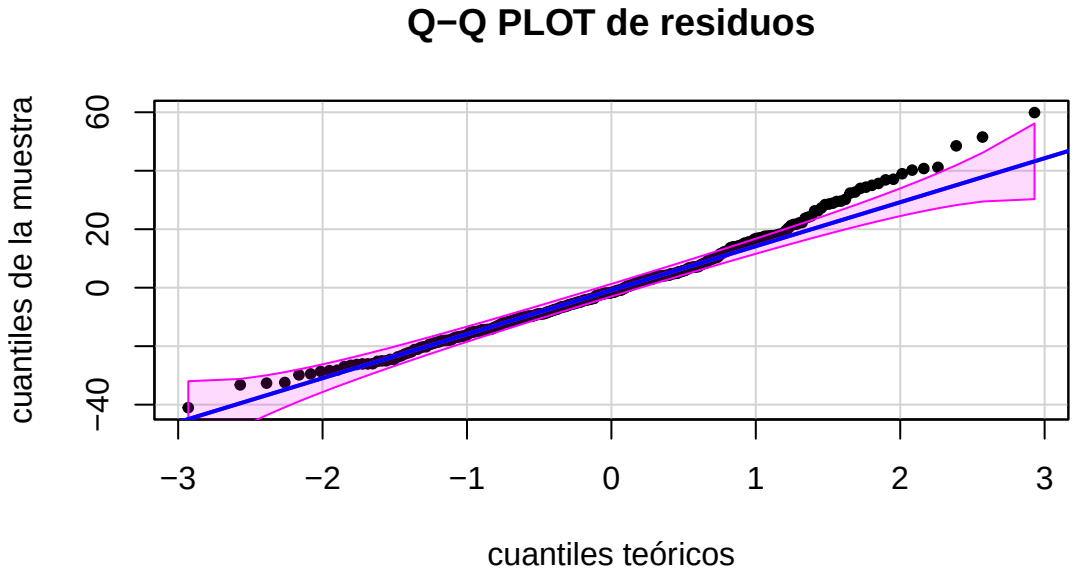

```
## [1] -0.005171727
```

De la gráfica podemos decir que no tenemos problemas con la normalidad del modelo, ademas obtuvimos un valor de media de los residuos muy cercano a 0, lo cual también nos indica que no hay ningún problema con este supuesto.

Normalidad

Para verificar si el modelo ajustado cumple con este supuesto realizaremos la prueba de Shapiro-Wilk con una alfa de 5%.

```
##
##  Shapiro-Wilk normality test
##
## data:  glm_fit$residuals
## W = 0.98169, p-value = 0.0007955
```



Del resultado del test y con la gráfica tenemos que lamentablemente el modelo no paso el supuesto de normalidad por lo que mas adelante veremos como arreglar este supuesto.

Homocedasticidad

```
##
##  studentized Breusch-Pagan test
##
## data:  glm_fit
## BP = 6.3699, df = 3, p-value = 0.09494
```

De los resultados obtenidos del test, podemos decir que el modelo no viola el supuesto de homocedasticidad.

Independencia

Para verificar este supuesto realizaremos el test Durbin-Watson

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1140265       2.22097   0.054
##  Alternative hypothesis: rho != 0
```

De los resultados obtenidos tenemos un valor de 2.22097 muy cercano a 2, es decir, no tenemos problemas con la independencia en el modelo generado.

2.¿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?Argumente su respuesta, indicando con claridad la prueba o pruebas de hipótesis usadas y las hipótesis que se están contrastando.

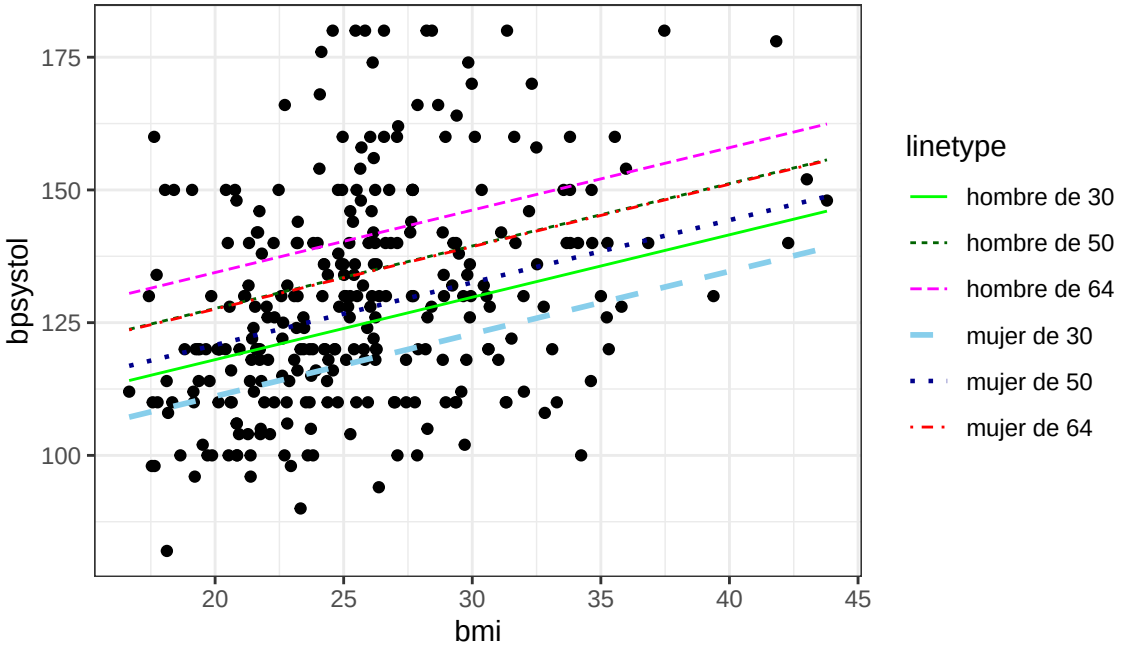
Usando la información que obtuvimos del modelo podemos afirmar que el sexo y edad de una persona con un IMC alto va a afectar directamente la presion sistolica de la misma, resultado que podriamos decir era el esperado.

Dicho resultado lo podemos denotar con las siguientes pruebas de hipotesis:

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

Tomando una alfa de 0.05 y dado que obtuvimos un p-value menor a esta alfa, entonces rechazamos la hipotesis nula, por lo tanto el parametro β_1 es estadisticamente significativo.

3.Para complementar la interpretación de los resultados del inciso iii), presente una gráfica resumen con la estimación puntual asociada a la relación entre bpsystol y bmi. Para esto considere sólo tres posibles edades: 30, 50 y 64, así como la diferenciación entre mujeres y hombres. Comente e interprete lo que observa en la gráfica, indicando con claridad a qué parámetro corresponde la curva/recta.



De los resultados es evidente que la presión sistólica de los hombres es mayor que la de las mujeres para los 3 grupos,por lo que desde ahi nos damos cuenta que el sexo ya es un factor que influye de manera directa en la presión sistólica de una persona, aunado a eso tenemos el factor de la edad donde es claro que a mayor edad la presion de una persona también sera mayor, algo que se mantiene para ambos casos es el factor del IMC donde se tiene que si este aumenta también crece la presion arterial sistólica.

4.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.

Si bien ambos modelos obtenidos dieron buenos resultados y ademas similares, en el sentido de que ambos nos fueron de ayuda para explicar como las variables IMC,Edad,Sexo afectan la presion arterial sistolica de una persona es claro que la complejidad de un modelo fue mayor a la otra, para el caso del modelo usado en el ejercicio 1 usamos un modelo lineal simple con una transformación de log, lo cual lo hace mas fácil de interpretar que el modelo usado en este ejercicio el cual requirió de ajustar un GLM con familia inversa gaussiana y función liga identidad.

Por otro lado también hay que considerar que los modelos dieron distintos AIC, en el caso que estos 2 sean muy parecidos, es decir, la diferencia de uno y otro sea pequeña lo que haremos sera tomar el modelo mas fácil que seria el usado en el ejercicio 1, por el contrario si la diferencia es mayor nos quedaremos con el que tenga un menor AIC.