

Análisis de Regresión Lineal Múltiple

2025-03-23

Ejercicio 1

Introducción

Este análisis tiene como objetivo evaluar la relación entre la presión arterial sistólica (`bpsystol`) y el índice de masa corporal (`bmi`), considerando el efecto de la edad y el sexo. Se ajusta un modelo de regresión lineal múltiple, se evalúan los supuestos y se realiza una transformación para mejorar su validez.

Modelo Inicial

El modelo inicial a ajustar es:

$$E[\textit{bpsystol}|\textit{bmi}, \textit{sex}, \textit{age}] = \beta_0 + \beta_1(\textit{bmi}) + \beta_2(\textit{sex}) + \beta_3(\textit{age})$$

- Variables:
 - Dependiente: `bpsystol` (presión arterial sistólica en mmHg)
 - Explicativas: `bmi` (índice de masa corporal), `sex` (factor con niveles de 1: Hombre , 2: Mujer), `age` (edad en años).

El resumen del modelo inicial (`summary(modelo)`) indica que todos los coeficientes son estadísticamente significativos (p-values < 0.05), con un R^2 ajustado de 0.2928, sugiriendo que el 29.28% de la variabilidad en `bpsystol` es explicada por las variables.

Evaluación de Supuestos

Se evaluaron los supuestos del modelo inicial:

- **Linealidad:** La prueba `residualPlots` mostró un p-value de $0.003678 < 0.05$ para `age`, indicando que la relación entre `age` y `bpsystol` no es lineal.
- **Normalidad:** Las pruebas de Shapiro-Wilk (p-value = 0.4961), Lilliefors (p-value = 0.2516) y Jarque-Bera (p-value = 0.6947) no rechazaron la normalidad de los residuos.
- **Homoscedasticidad:** Las pruebas de Breusch-Pagan (p-value = $0.01289 < 0.05$) y Score (p-value = $0.0025605 < 0.05$) indicaron heterocedasticidad.
- **Independencia:** Las pruebas de Durbin-Watson (p-value = 0.4793) y Breusch-Godfrey (p-value = 0.9615) no detectaron autocorrelación.

El p-value de `age` ($0.003678 < 0.05$) indica que la relación con `bpsystol` no es lineal.

Modelo Transformado

Para corregir la heterocedasticidad y linealidad, aplicamos una transformación logarítmica a `bpsystol` y `bmi`, además de incluir un término cúbico para `age`:

$$E[\log(bpsystol)] = \beta_0 + \beta_1 \log(bmi) + \beta_2(sex) + \beta_3(age^3)$$

- **log(bpsystol)**: La transformación logarítmica estabiliza la varianza, corrigiendo la heterocedasticidad, ya que los residuos mostraban mayor dispersión en valores altos de `bpsystol`.
- **log(bmi)**: Se aplicó una transformación logarítmica a `bmi` para corregir una posible relación no lineal con `bpsystol`.
- **age³**: El término cúbico de `age` se incluyó para corregir la no linealidad.

Se repiten las pruebas de normalidad y homocedasticidad para el nuevo modelo:

- **Linealidad**: `residualPlots` resulto p-values > 0.05 para todas las variables y el Tukey test (p-value = 0.2104), confirmando linealidad.
- **Normalidad**: Shapiro-Wilk (p-value = 0.4479), Lilliefors (p-value = 0.6279) y Jarque-Bera (p-value = 0.3555) no rechazaron la normalidad.
- **Homocedasticidad**: Breusch-Pagan (p-value = 0.5778) y Score (p-value = 0.23218) indicaron homocedasticidad.
- **Independencia**: Durbin-Watson (p-value = 0.5184) y Breusch-Godfrey (p-value = 0.9571) confirmaron independencia.

Ahora se cumple con la normalidad y la homocedasticidad, lo que valida el uso del modelo transformado.

Prueba de Hipótesis

Para evaluar la relación entre `bmi` y `bpsystol`, se plantea la siguiente prueba de hipótesis:

- $H_0 : \beta_1 \leq 0$ (No hay asociación positiva entre `bmi` y `bpsystol`)
- $H_A : \beta_1 > 0$ (Existe una asociación positiva significativa entre `bmi` y `bpsystol`)

Como el p-value es menor que 0.05, rechazamos H_0 y concluimos que existe una relación positiva significativa entre el índice de masa corporal (`bmi`) y la presión arterial sistólica (`bpsystol`).

Resultados Numéricos

```
##
## Call:
## lm(formula = log(bpsystol) ~ log(bmi) + sex + I(age^3), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37683 -0.07462  0.00037  0.08043  0.28981
```

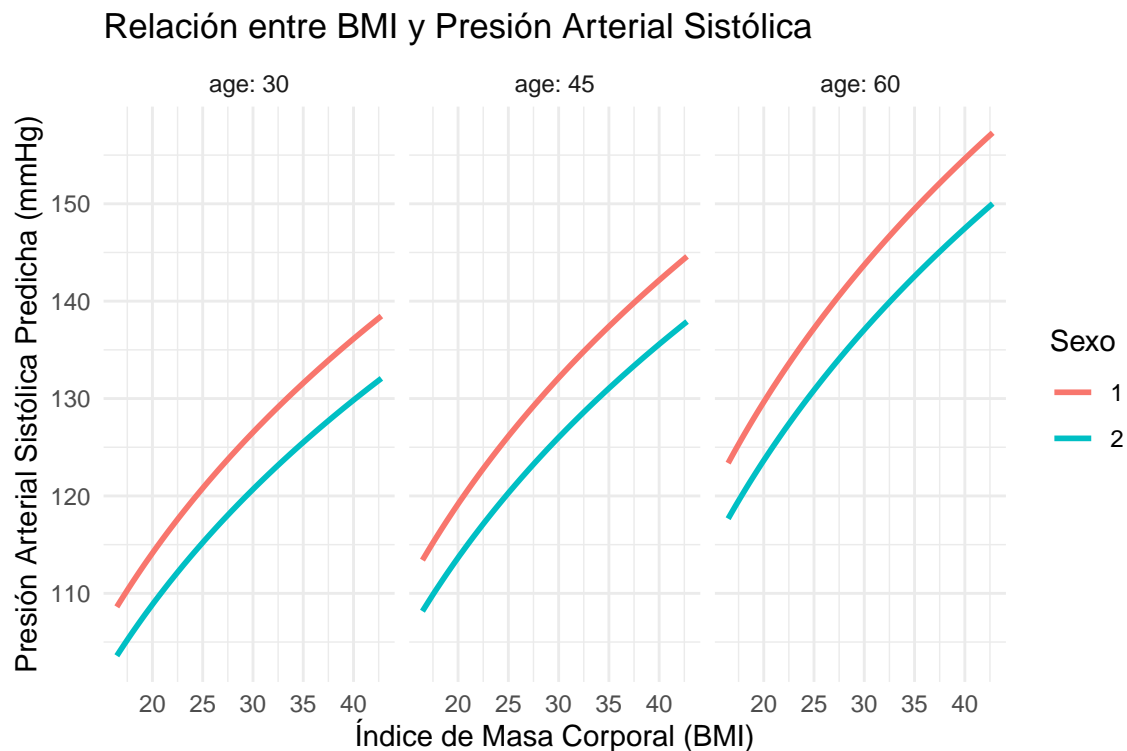
```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.957e+00  9.352e-02  42.313  < 2e-16 ***
## log(bmi)     2.544e-01  2.905e-02   8.754  < 2e-16 ***
## sex2        -4.735e-02  1.083e-02  -4.372  1.54e-05 ***
## I(age^3)     6.741e-07  7.861e-08   8.575  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1126 on 434 degrees of freedom
## Multiple R-squared:  0.3081, Adjusted R-squared:  0.3033
## F-statistic: 64.41 on 3 and 434 DF,  p-value: < 2.2e-16
```

Todos los coeficientes son significativos (p -values < 0.05). El coeficiente de $\log(bmi)$ (0.2544) indica una relación positiva con `bpsystol`, mientras que `sex2` (-0.04735) sugiere que las mujeres tienen una presión arterial menor que los hombres, y age^3 ($6.741e-07$) refleja un efecto creciente de la edad.

Visualización de Resultados

Nota: La edad máxima en el conjunto de datos es 62 años, por lo que se elige 60 en lugar de 65.

La gráfica consta de tres partes (una por edad: 30, 45, 60), con rectas que representan `bpsystol_pred` en función de `bmi`, diferenciadas por sexo (Hombre: 1, Mujer: 2). Las rectas ascendentes reflejan la relación positiva entre `bmi` y `bpsystol` ($B_1 = 0.2544$), con un efecto más notable a mayor edad (60 años) debido al término cuadrático age^3 , y en hombres, debido al coeficiente negativo de `sex2`.



Conclusiones

- Se encontró una relación significativa entre *bmi* y *bpsystol*, confirmada por la prueba de hipótesis.
- La transformación del modelo corrigió problemas de heterocedasticidad y linealidad, mejorando su validez.
- A mayor *bmi*, mayor presión arterial sistólica, con diferencias según edad y sexo.

Referencias

- Documentación de R: `lm()`, `bptest()`, `shapiro.test()`, `ggplot2`.
- Modelos lineales en regresión aplicada.
- Notas vistas en clase.