Estimación del Modelo Preliminar

José Ángel Carretero Montes
Ismael Sallami Moreno
Fernando José Gracia Choin

Octubre 2024

1 Estimación del Modelo Preliminar

El modelo de regresión lineal múltiple estimado se representa de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

donde:

- Y_i es la variable dependiente (Peso en este caso),
- $X_{1i}, X_{2i}, \ldots, X_{ki}$ son las variables independientes (Altura, Edad, Género, etc.),
- β_0 es el término constante,
- u_i es el término de error aleatorio.

Los resultados preliminares de la regresión OLS para este modelo se presentan en la siguiente tabla:

OLS Regression Results

Dep. Variable:	Weight	R-squared:	0.576
Model:	OLS	Adj. R-squared:	0.572
Method:	Least Squares	F-statistic:	177.5
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	07:09:09	Log-Likelihood:	-8983.6
No. Observations:	2111	AIC:	1.800e+04
Df Residuals:	2094	BIC:	1.810e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-207.9919	10.281	-20.230	0.000	-228.155	-187.829
Gender	-4.7575	1.026	-4.639	0.000	-6.769	-2.746
Age	0.7403	0.082	9.069	0.000	0.580	0.900
Height	125.2342	5.894	21.247	0.000	113.675	136.793
family_history_with_overweight	17.0101	1.115	15.259	0.000	14.824	19.196
FAVC	7.9764	1.252	6.373	0.000	5.522	10.431
FCVC	9.0480	0.750	12.070	0.000	7.578	10.518
NCP	1.0882	0.506	2.151	0.032	0.096	2.080
CAEC	8.2303	0.869	9.470	0.000	6.526	9.935
SMOKE	-0.2063	2.664	-0.077	0.938	-5.430	5.017
CH2O	1.3156	0.645	2.041	0.041	0.052	2.580
SCC	-6.9609	1.878	-3.706	0.000	-10.645	-3.277
FAF	-2.7221	0.484	-5.619	0.000	-3.672	-1.772
TUE	-1.6760	0.652	-2.571	0.010	-2.954	-0.398
CALC	-4.1794	0.764	-5.472	0.000	-5.677	-2.682
MTRANS	3.6899	0.382	9.658	0.000	2.941	4.439
NObeyesdad	2.3519	0.220	10.680	0.000	1.920	2.784

Omnibus:	37.671	Durbin-Watson:	0.741
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.704
Skew:	-0.047	Prob(JB):	1.94e-05
Kurtosis:	2.512	Cond. No.	812.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

2 Interpretación de los Estimadores

A continuación, se ofrece una interpretación de los coeficientes más relevantes del modelo:

- **Constante**: El coeficiente de la constante es -207.99, lo que indica el valor estimado del peso cuando todas las variables independientes son 0. Aunque en la práctica, esta interpretación rara vez es útil, ya que variables como la altura no pueden ser cero.
- **Género**: El coeficiente de Género es -4.757. Esto indica que, manteniendo todas las demás variables constantes, ser mujer (asumiendo que el género está codificado como 1 para mujeres y 0 para hombres) está asociado con una reducción de 4.76 unidades en el peso.
- **Edad**: El coeficiente de 0.7403 indica que por cada año adicional de edad, el peso aumenta en promedio 0.74 unidades, manteniendo las demás variables constantes.
- **Altura**: El coeficiente de altura es 125.23, lo que sugiere que, por cada unidad adicional de altura, el peso aumenta en 125.23 unidades, manteniendo todas las demás variables constantes.
- **Historial Familiar de Sobrepeso**: El coeficiente es 17.01, indicando que tener antecedentes familiares de sobrepeso está asociado con un aumento promedio de 17.01 unidades en el peso.
- **FAVC (Comida grasa frecuente)**: El coeficiente positivo de 7.976 indica que consumir frecuentemente comida grasa aumenta el peso en 7.976 unidades, manteniendo todo lo demás constante.
- **FCVC (Consumo de verduras)**: El coeficiente de 9.048 indica que un mayor consumo de verduras está asociado con un aumento de 9.05 unidades en el peso, lo cual es inesperado y podría requerir mayor análisis.
- **NCP (Número de comidas diarias)**: El coeficiente de 1.088 sugiere que, por cada comida adicional consumida al día, el peso aumenta en promedio 1.088 unidades.
- **CAEC (Comer entre comidas)**: El coeficiente de 8.230 sugiere que comer entre comidas aumenta el peso en 8.23 unidades, manteniendo el resto de factores constantes.
- **Fumar (SMOKE)**: El coeficiente es -0.206, indicando que fumar no tiene un impacto estadísticamente significativo en el peso, dado su alto valor p ($P > 0.05$).
- **Consumo de agua (CH2O)**: El coeficiente de 1.315 indica que por cada unidad adicional de agua consumida, el peso aumenta en 1.315 unidades, lo cual puede ser contraintuitivo y merece un análisis más profundo.
- **SCC (Conteo de calorías)**: El coeficiente de -6.961 sugiere que el control de calorías está asociado con una reducción del peso en 6.96 unidades, lo cual es consistente con lo esperado.
- **FAF (Actividad física)**: El coeficiente de -2.722 indica que un aumento en la actividad física reduce el peso en promedio 2.72 unidades, manteniendo constantes las demás variables.
- **TUE (Uso de tecnología)**: El coeficiente de -1.676 sugiere que más tiempo usando tecnología está asociado con una reducción del peso, lo cual puede requerir más análisis.
- **CALC (Consumo de alcohol)**: El coeficiente de -4.179 indica que consumir alcohol está asociado con una reducción del peso, lo cual es un resultado inesperado.
- **MTRANS (Medios de transporte)**: El coeficiente de 3.689 indica que el uso de ciertos medios de transporte se asocia con un aumento del peso en 3.689 unidades.
- **NObeyesdad (Nivel de obesidad)**: El coeficiente de 2.351 indica que un mayor nivel de obesidad previo está asociado con un aumento de 2.351 unidades en el peso.

3 Medidas de Bondad de Ajuste

Para evaluar la calidad del modelo estimado, utilizamos diversas medidas de bondad de ajuste que nos permiten juzgar cómo de bien el modelo ajusta a los datos observados. Las medidas que se consideran son el coeficiente de determinación R^2, el coeficiente de determinación ajustado R^2_{ajustado}, y los criterios de información de Akaike (AIC) y Schwarz (BIC).

3.1 Coeficiente de Determinación R^2

El coeficiente de determinación R^2 mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo. En otras palabras, indica qué tan bien las variables independientes explican las variaciones en la variable dependiente.

$$R^2 = 1 - \frac{\text{Suma de Cuadrados de los Residuos (SCR)}}{\text{Suma Total de Cuadrados (SCT)}}$$

Donde:

- La **Suma Total de Cuadrados (SCT)** mide la variabilidad total de la variable dependiente en torno a su media:

$$\text{SCT} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- La **Suma de Cuadrados de los Residuos (SCR)** mide la variabilidad que no ha sido explicada por el modelo:

$$\text{SCR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Para este modelo, se ha obtenido un R^2 de **0.576**, lo que significa que el **57.6%** de la variabilidad del peso puede explicarse por las variables independientes incluidas en el modelo. Aunque este valor indica que el modelo captura más de la mitad de la variación en el peso, hay un **42.4%** de la variabilidad que no está siendo explicada, lo que sugiere que existen otros factores que no se han tenido en cuenta o que pueden requerir un tratamiento más sofisticado.

3.2 Coeficiente de Determinación Ajustado R^2_{ajustado}

El coeficiente de determinación ajustado R^2_{ajustado} ajusta el R^2 para el número de predictores en el modelo. Esto es importante porque el R^2 tiende a aumentar al agregar más variables, incluso si esas variables no tienen una relación significativa con la variable dependiente. El R^2_{ajustado} penaliza el uso de variables adicionales, lo que permite obtener una mejor medida del poder explicativo real del modelo.

$$R^2_{\text{ajustado}} = 1 - \frac{\frac{\text{SCR}}{n-k}}{\frac{\text{SCT}}{n-1}}$$

Donde:

- n es el número de observaciones,
- k es el número de variables explicativas incluidas en el modelo.

El valor obtenido para el R^2_{ajustado} es de **0.572**, lo cual es muy cercano al valor de R^2. Esto indica que el modelo no está sobreajustado (es decir, no ha añadido variables innecesarias), ya que el R^2_{ajustado} no disminuye considerablemente respecto al R^2 original.

3.3 Criterio de Información de Akaike (AIC)

El criterio de información de Akaike (AIC) es una medida que penaliza la complejidad del modelo. Es útil cuando se comparan varios modelos, ya que tiene en cuenta tanto la bondad del ajuste (medida por la suma de cuadrados de los residuos) como el número de parámetros estimados. Cuanto menor sea el valor del AIC, mejor es el modelo.

$$AIC = \ln\left(\frac{\text{SCR}}{n}\right) + \frac{2k}{n}$$

Donde:

- n es el número de observaciones,
- k es el número de parámetros estimados (incluyendo la constante).

En este modelo, el valor de **AIC** es $\mathbf{1.800 \times 10^4}$. Para comparar diferentes modelos, se elegiría aquel que minimice este valor. Un menor AIC indica que el modelo no solo ajusta mejor los datos, sino que también evita la inclusión innecesaria de parámetros.

3.4 Criterio de Información de Schwarz (BIC)

El criterio de información bayesiano de Schwarz (BIC) es otra medida que, como el AIC, penaliza la complejidad del modelo, pero de manera más estricta. El BIC introduce una penalización más fuerte por la inclusión de más variables en el modelo, lo que lo hace útil para evitar el sobreajuste.

$$BIC = \ln\left(\frac{\text{SCR}}{n}\right) + \frac{k}{n} \cdot \ln(n)$$

En este modelo, el valor del **BIC** es $\mathbf{1.810 \times 10^4}$. Al igual que con el AIC, un menor valor de BIC es preferible, y su objetivo es identificar el modelo que ofrece el mejor ajuste sin sobrecargarlo con demasiadas variables.

3.5 Comparación y Selección del Mejor Modelo

Las medidas de bondad de ajuste nos indican diferentes aspectos de la calidad del modelo. En resumen:

- Un R^2 de 0.576 indica que el modelo es razonablemente bueno para explicar la variabilidad del peso.
- El R^2_{ajustado} de 0.572 sugiere que no hay un sobreajuste significativo.
- Los criterios de información AIC y BIC nos permiten comparar este modelo con otros posibles modelos. Si generamos otros modelos, escogeremos aquel con los valores más bajos de AIC y BIC.

En función de estas medidas, este modelo tiene un ajuste aceptable, pero se puede mejorar probando otros modelos que optimicen las medidas de AIC y BIC.

4 Conclusiones y Selección del Modelo

El modelo preliminar parece explicar bien la variabilidad del peso, con un R^2 razonable y estimadores significativos. Para mejorar el modelo, podemos considerar variables adicionales o la eliminación de aquellas que no resulten significativas en otros ensayos.

Además, el uso de AIC y BIC será clave para la selección del mejor modelo en función de los diferentes criterios de ajuste que hemos estudiado.