Clase 7
Regresión Lineal Múltiple III Inferencia Estadística

Análisis Avanzado de Datos

Gabriel Sotomayor

Evaluaciones

Tarea 2: 10 de octubre

- Regresión lineal múltiple

Informe 1: 30 de Octubre

- Regresión lineal múltiple o regresión logística
- Se subirà la pauta la pròxima semana

Recordatorio clase anterior

Introducción al Modelo de Regresión Múltiple

Un modelo de regresión múltiple examina la relación entre una **variable dependiente** y **varias variables independientes o predictores**.
La regresión múltiple permite **controlar otras variables** mientras se evalúa el efecto de una variable predictora específica.
Ejemplo: Si estudiamos la relación entre el ejercicio y la pérdida de peso, también podemos controlar la cantidad de alimentos consumidos para aislar su efecto.

Ecuación básica del modelo:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \epsilon$$

Donde Y es la variable dependiente, b_0 es la constante o intercepto, $X_1, X_2, \ldots X_k$ son las variables independientes, $b_1, b_2, \ldots b_k$ son los coeficientes de regresión, y ϵ es el error.

Objetivo de la sesión

Introducir la inferencia estadística en el contexto de Regresión Lineal Múltiple.

R^2 en Regresión Lineal Múltiple

- **Definición de R^2**:
  - El R^2 mide la proporción de la **variabilidad explicada** por el modelo en relación a la variabilidad total.
  - Se interpreta como el porcentaje de la variación en la variable dependiente que es explicado por las variables independientes.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- Donde: - $\hat{y}_i$ son los valores predichos. - y_i son los valores observados. - $\bar{y}$ es el promedio de la variable dependiente.

- **Limitaciones**:
  - El R^2 puede aumentar al agregar más predictores, incluso si no aportan significativamente al modelo.

R^2 Ajustado y su Utilidad

- El R^2 ajustado corrige la sobreestimación del R^2 al penalizar por el número de predictores en el modelo.
- Tiene en cuenta tanto el **número de predictores** como el **tamaño de la muestra**.
- **Cálculo**:
  - $$R^2_{ajustado} = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$
  - Donde:
    - n es el número de observaciones.
    - k es el número de predictores en el modelo.
  - A diferencia del R^2, el R^2 ajustado **disminuye** si se agregan predictores que no mejoran el modelo, ayudando a evitar el sobreajuste.
  - Es más útil cuando se compara la calidad de diferentes modelos con un número distinto de predictores.

Ejemplo con CASEN

Trabajaremos con CASEN utilizando como variable dependiente los ingresos del trabajo, y como variable predictora la edad y los años de escolaridad. Utilizaremos esto como base para introducir los conceptos de inferencia respecto de los coeficientes de regresión.

```
# A tibble: 6 × 3
  ytrabajocor   esc  edad
        <dbl> <dbl> <dbl>
1      411242    15    40
2      590000     5    64
3      520000    12    34
4      450000    12    30
5      160000    10    68
6      580000     8    56
```

Ejemplo con CASEN

	Model 1
(Intercept)	-742740.49***
	(14303.97)
esc	88022.20***
	(704.29)
edad	7681.28***
	(189.88)
R^2	0.15
Adj. R^2	0.15
Num. obs.	88391
***p < 0.001; **p < 0.01; *p < 0.05	

Statistical models

Estadístico vs. Parámetro

- **Estadístico**: Es un valor calculado a partir de los datos de una muestra. Ejemplos incluyen la media muestral ($\bar{x}$), la desviación estándar (s) y los coeficientes de regresión estimados (b_j).
- **Parámetro**: Es un valor que describe una característica de una población. Ejemplos incluyen la media poblacional (μ) y el coeficiente de regresión verdadero (β_j).
- Los **estadísticos** se utilizan para hacer **inferencia** sobre los **parámetros** desconocidos de la población.

Variabilidad de las Muestras

- La variabilidad de las muestras se refiere a cómo los **estadísticos** pueden cambiar de una muestra a otra. Esto se debe a la **aleatoriedad** inherente en el proceso de muestreo.
- La **distribución muestral** de un estadístico describe cómo varía dicho estadístico en múltiples muestras de una misma población.
- Ejemplo: Si tomamos varias muestras de una población, el valor del coeficiente de regresión estimado (b_j) puede ser diferente en cada muestra.

Inferencia en la Regresión Lineal

- En regresión lineal, utilizamos los **estadísticos muestrales** (por ejemplo, b_0, b_1, b_2, etc.) para hacer inferencias sobre los **parámetros poblacionales** (β_0, β_1, β_2, etc.).
- Debido a la **variabilidad muestral**, cada estimación está sujeta a **incertidumbre**. Esta incertidumbre se cuantifica a través del **error estándar** y se refleja en los **intervalos de confianza**.
- La **significancia estadística** nos ayuda a determinar si un coeficiente de regresión es diferente de cero en la población.

Estimación e Interpretación de los Betas

- **Estimación**: Los coeficientes de regresión (b_j) se estiman utilizando el método de **mínimos cuadrados ordinarios (OLS)**, que minimiza la suma de los residuos al cuadrado.
- **Interpretación**:
  - Cada **beta** (β_j) representa el cambio esperado en la variable dependiente (Y) por un cambio unitario en la variable independiente (X_j), manteniendo las demás variables constantes.
  - Ejemplo: En la regresión de los ingresos del trabajo sobre escolaridad y edad, el coeficiente de escolaridad representa el cambio en los ingresos por cada año adicional de educación, manteniendo la edad constante.

Relación entre el Beta y el Error Estándar

- **Coeficiente Beta (b_j)**: Representa la magnitud y dirección del efecto de la variable independiente (X_j) sobre la variable dependiente (Y).
- **Error Estándar (SE)**: Mide la precisión de la estimación del coeficiente. Un error estándar más pequeño indica una estimación más precisa del coeficiente.
- **Prueba de Hipótesis**: Para determinar si un coeficiente (b_j) es significativamente diferente de cero, realizamos una prueba de hipótesis. La hipótesis nula (H_0) establece que el coeficiente es igual a cero ($b_j = 0$), es decir, que no hay efecto de la variable X_j sobre Y.

Ejemplo con el modelo de ingresos

- Para el coeficiente de **escolaridad (esc)**:
  - $b_j = 88022.20$
  - $SE = 704.29$
  - Valor t:

$$t = \frac{88022.20}{704.29} = 124.99$$

  - Valor p: Con un valor t tan alto, el valor p será menor a 0.001, indicando que el coeficiente es **altamente significativo**.

Ejemplo con el modelo de ingresos

- Para el coeficiente de **edad**:
  - $b_j = 7681.28$
  - $SE = 189.88$
  - Valor t:

$$t = \frac{7681.28}{189.88} = 40.45$$

  - Valor p: Dado que el valor t es muy alto, el valor p también será menor que 0.001, indicando que el coeficiente es **altamente significativo**.

Relación entre el Beta y el Error Estándar

Interpretación de los Coeficientes Beta en Regresión Lineal Múltiple

- **Tamaño, Dirección del Efecto y Significancia Estadística**:
  - Cada coeficiente beta (β) indica el **cambio en la variable dependiente** por **cada unidad de cambio en la variable independiente** correspondiente.
  - **Dirección**:
    - $\beta > 0$: Indica un **efecto positivo** (la variable dependiente aumenta).
    - $\beta < 0$: Indica un **efecto negativo** (la variable dependiente disminuye).
  - **Significancia Estadística**: Indica si el efecto observado es **estadísticamente diferente de cero** en la población. Un valor p pequeño (por ejemplo < 0.05) indica que es poco probable que el coeficiente sea igual a cero.

Interpretación de los Coeficientes Beta en Regresión Lineal Múltiple

- **Controlando por las demás variables del modelo**:
  - El valor de β refleja el **efecto neto** de la variable independiente, es decir, **ajustado por todas las otras variables** incluidas en el modelo.
  - Permite evaluar el efecto **aislado** de cada variable independiente mientras se **controlan** los posibles efectos de las demás.
- **Tamaño del Efecto vs. Significancia Estadística**:
  - **Tamaño del Efecto**: Se refiere a la magnitud del impacto que tiene una variable independiente sobre la variable dependiente. Un beta grande implica un cambio considerable en Y por cada unidad de X.
  - **Significancia Estadística**: Nos dice si el efecto observado es diferente de cero, pero no necesariamente cuán grande o importante es el efecto. Un coeficiente puede ser estadísticamente significativo, pero con un tamaño de efecto pequeño.

Interpretación de Coeficientes Beta para Variables Categóricas

- **Diferencia con la Categoría de Referencia**:
  - El coeficiente beta para una variable categórica representa la **diferencia promedio** en la variable dependiente entre el grupo correspondiente y la **categoría de referencia**.
  - Si β es positivo, indica que el grupo en cuestión tiene una **mayor** media en comparación con la categoría de referencia. Si β es negativo, la media es **menor**.
- **Controlando por las demás variables del modelo**:
  - Al igual que en las variables continuas, los efectos están **ajustados** por las demás variables independientes, lo que permite interpretar el efecto de la categoría como si las otras variables permanecieran **constantes**.

Ejemplo con CASEN

	Model 1
(Intercept)	-742740.49***
	(14303.97)
esc	88022.20***
	(704.29)
edad	7681.28***
	(189.88)
R^2	0.15
Adj. R^2	0.15
Num. obs.	88391
***p < 0.001; **p < 0.01; *p < 0.05	

Statistical models

Ejemplo con CASEN

Intercepto (b_0):
El intercepto de -742,740.49 indica el valor esperado de los ingresos cuando tanto la escolaridad como la edad son cero, lo cual no tiene una interpretación práctica debido a la falta de sentido de esta situación en el contexto real. Este valor es estadísticamente significativo ($p < 0.001$).

Escolaridad (esc):
El coeficiente de la escolaridad es 88,022.20, indicando que, por cada año adicional de escolaridad, los ingresos del trabajo aumentan en promedio 88,022.20 pesos, manteniendo constante la edad. Este coeficiente es estadísticamente significativo ($p < 0.001$).

Edad:
El coeficiente de la edad es 7,681.28, lo que implica que cada año adicional de edad se asocia con un aumento promedio de 7,681.28 pesos en los ingresos, manteniendo constante la escolaridad. Este coeficiente también es estadísticamente significativo ($p < 0.001$).