# Regresión lineal múltiple y generalizada.

MTRA. MELODY TREVIÑO RODRIGUEZ

# Regresión lineal múltiple

En la regresión lineal simple, se tiene una variable predictora que se relaciona con una variable de respuesta mediante una línea recta. En cambio, la regresión lineal múltiple considera múltiples variables predictoras para modelar la relación con una variable de respuesta.

Regresión lineal múltiple permite analizar la relación entre una variable dependiente y dos o más variables independientes.

El modelo matemático para la regresión lineal múltiple puede expresarse como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

donde:

$Y$ es la variable de respuesta.

$\beta_0$ es la intersección en el eje Y (término constante).

$\beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes de regresión asociados a las variables predictoras $X_1, X_2, \ldots, X_n$.

$\varepsilon$ es el término de error, que representa la variabilidad no explicada por el modelo.

# Regresión lineal múltiple

El objetivo de la regresión lineal múltiple es estimar los coeficientes $\beta_0$, $\beta_1$, …, $\beta_n$ de manera que el modelo se ajuste de la mejor manera posible a los datos observados.

Esto facilita la predicción de la variable de respuesta en función de los valores de las variables predictoras.

La calidad del ajuste del modelo se evalúa mediante medidas como el ción ($R^2$) y otras pruebas estadísticas.

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \rho^2$$

Donde:

- $\sigma_{XY}$ es la covarianza de $(X, Y)$
- $\sigma_X^2$ es la varianza de la variable $X$
- $\sigma_Y^2$ es la varianza de la variable $Y$

Nota: En el caso particular de una regresión lineal, $R^2$ coincide con el cuadrado del coeficiente de correl $\rho^2$ n de Pearson

# Regresión lineal múltiple

## Condiciones para la regresión lineal múltiple

**No colinialidad o multicolinialidad**:

En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinialidad entre ellos. La colinialidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores.

No existe un método estadístico para determinar la existencia de colinialidad o multicolinialidad entre los predictores de un modelo de regresión, sin embargo, se han desarrollado numerosas reglas prácticas:

- Si el coeficiente de determinación R2 es alto, pero ninguno de los predictores resulta significativo.
- Calcular una matriz de correlación en la que se estudia la relación lineal entre cada par de predictores (Nota: a pesar de no obtenerse ningún coeficiente de correlación alto, no está asegurado que no exista multicolinialidad. Se puede dar el caso de tener una relación lineal casi perfecta entre tres o más variables y que las correlaciones simples entre pares de estas mismas variables no sean mayores que 0.5).
- Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el coeficiente de determinación R2 es alto, estaría señalando a una posible colinialidad.

# Regresión lineal múltiple

## Condiciones para la regresión lineal múltiple

**Parsimonia:**

Este término hace referencia a que el mejor modelo es aquel capaz de explicar con mayor precisión la variabilidad observada en la variable respuesta empleando el menor número de predictores, por lo tanto, con menos asunciones.

**Relación lineal entre los predictores numéricos y la variable respuesta:**

Cada predictor numérico tiene que estar linealmente relacionado con la variable respuesta y mientras los demás predictores se mantienen constantes, de lo contrario no se puede introducir en el modelo. La forma más recomendable de comprobarlo es representando los residuos del modelo frente a cada uno de los predictores. Si la relación es lineal, los residuos se distribuyen de forma aleatoria entorno a cero. Estos análisis son solo aproximados, ya que no hay forma de saber si realmente la relación es lineal cuando el resto de predictores se mantienen constantes.

# Regresión lineal múltiple

## Condiciones para la regresión lineal múltiple

**Distribución normal de los residuos:**

Los residuos se deben distribuir de forma normal con media cero. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a test de hipótesis de normalidad.

**Variabilidad constante de los residuos (homocedasticidad):**

La varianza de los residuos debe de ser constante en todo el rango de observaciones. Para comprobarlo se representan los residuos. Si la varianza es constante, se distribuyen de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico. Una distribución cónica es un claro identificador de falta de homocedasticidad. También se puede recurrir a contrastes de homocedasticidad como el test de Breusch-Pagan.

# Regresión lineal múltiple

## Condiciones para la regresión lineal múltiple

**No autocorrelación (Independencia):**

Los valores de cada observación son independientes de los otros, esto es especialmente importante de comprobar cuando se trabaja con mediciones temporales. Se recomienda representar los residuos ordenados acorde al tiempo de registro de las observaciones, si existe un cierto patrón hay indicios de autocorrelación. También se puede emplear el test de hipótesis de Durbin-Watson.

**Valores atípicos, con alto leverage o influyentes:**

Es importante identificar observaciones que sean atípicas o que puedan estar influenciando al modelo. La forma más fácil de detectarlas es a través de los residuos.

# Regresión Lineal Generalizada (GLM)

John Nelder y Robert Wedderburn formularon modelos lineales generalizados como una forma de unificar otros modelos estadísticos, se utiliza para describir un enfoque más amplio que abarca modelos de regresión estándar (lineales) y otros modelos más flexibles, como el modelo logístico para respuestas binarias o el modelo de Poisson para datos de conteo

Modelo lineal generalizado (GLM) es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal.

# Regresión Lineal Generalizada (GLM)

El Modelo Lineal Generalizado es una extensión poderosa y flexible del modelo de regresión lineal clásico, diseñada para manejar una variedad de situaciones en las que los supuestos de normalidad y homocedasticidad no se cumplen.

Algunas características del GLM son:

**Función de Vinculación (Link Function):** El MLG utiliza una función de vinculación para relacionar la media de la variable de respuesta con una combinación lineal de las variables predictoras. La elección de la función de vinculación depende de la naturaleza de los datos y del tipo de modelo que se esté utilizando. Ejemplos comunes incluyen la función logit, la función logaritmo y la función identidad.

**Distribución de la Variable de Respuesta:** A diferencia del modelo de regresión lineal clásico, el MLG no asume una distribución normal para la variable de respuesta. Puede trabajar con diversas distribuciones, como la binomial, la Poisson, la gamma, entre otras. La elección de la distribución depende de la naturaleza de los datos y del tipo de modelo que se está ajustando.

# Regresión Lineal Generalizada (GLM)

**Varianza No Constante (Dispersion):** El MLG permite modelar la varianza de la variable de respuesta como una función de la media. Esto es útil cuando se enfrenta a datos con varianza no constante, lo que no está permitido en el modelo de regresión lineal clásico.

**Estimación de Parámetros:** La estimación de los parámetros del modelo se realiza mediante métodos de máxima verosimilitud. Este enfoque busca encontrar los valores de los parámetros que maximizan la probabilidad de observar los datos dados los parámetros.

**Aplicaciones Versátiles:** El MLG es extremadamente versátil y se aplica a una amplia gama de problemas. Puede utilizarse para modelar respuestas binarias (regresión logística), contar de eventos raros (regresión de Poisson), respuestas continuas positivas (regresión gamma), entre otros.

**Inferencia Estadística:** La inferencia estadística en el MLG se basa en distribuciones asintóticas, lo que significa que a medida que el tamaño de la muestra aumenta, las distribuciones de los estimadores se aproximan a distribuciones normales. Esto facilita la construcción de intervalos de confianza y pruebas de hipótesis.

# Tarea (Opcional para puntos extras)

1) Investigar y describir un ejercicio de GLM:

Investigar un ejemplo paso a paso que se realice en R o Python y describir lo que se hace en cada paso y aportar las conclusiones personales a las que llegaste con el resultado obtenido.