Notas de Clase:
Módulo IV Aplicaciones de Machine Learning a través de R y Python, y Módulo VI Temas selectos para el análisis de datos

Benjamín Oliva [1]

Draft Febrero 2024

[1] benjov@ciencias.unam.mx y https://github.com/benjov

Documento siempre proceso de mejora.
Comentarios, dudas, etc., siempre serán
bienvenidos...

Índice general

Índice de figuras

Índice de cuadros

Introducción, motivación y alcance del documento

Queremos discutir los documentos de la bibliografia como: Abadie (2021) [Aba21], Adams (2020) [Ada20], Cameron y Trivedi (2005) [CT05]; Cunningham (2021) [Cun21]; Greene (2012) [Gre12], Hastie, Tibshirani y Friedman (2017) [HTF17], James, Witten, Hastie y Tibshirani (2013) [Jam+13], y Wooldridge (2010) [Woo10]; entre muchos más que utilicemos en el curso.

Conocimientos previos

Módulos I, II y III del Diplomado.

Recursos en línea y otros materiales

Direcciones del GitHub, recomendaciones de cursos en línea, materiales recomendados y demas autoaprendizaje.

1

Módulo IV Aplicaciones de Machine Learning a través de R y Python

1.1. Introducción: El concepto de esperanza condicional y Causalidad

El objetivo del análisis empírico en las ciencias sociales es determinar cuál es el efecto o cambio que una variable causa en otra. Por ejemplo, cuando pensamos en qué factores o variables determinan el nivel salarial promedio de las personas es posible que relacionemos el efecto que tienen los años de estudio y en particular el efecto que tiene un año adicional de estudio. Otro ejemplo puede presentarse cuando tratamos de entender el efecto que tienen las horas de estudio en las calificaciones finales de un grupo de estudiantes.

Dentro del análisis empírico se suele hacer uso de dos términos o conceptos:

1. *Ceteris paribus*
2. *Causalidad*

El primero se emplea en el análisis empírico para establecer que el efecto de una variable se sostiene siempre que asumamos que el resto de variables que pueden afectar a dicha variable que queremos explicar permanecen constantes. El segundo sirve para entender que en el análisis empírico la existencia de correlaciones no significa la presencia de causalidad. La causalidad, en un primer momento, es la condición observada entre dos variables y deriva de la construcción teórica, modelación o racionalización que hagamos de los fenómenos que queremos analizar. De forma más sofisticada, es una relación analizada bajo un enfoque conocido como *inferencia causal*.

Partamos de lo que en estadística se conoce como **esperanza condicional**. Supongamos dos variables, y y x, que tienen una distribución conjunta y para las cuales queremos estimar el efecto que tiene la segunda en la primera. Así, es posible que utilicemos una expresión de la esperanza condicional de y en x, misma que solemos representar como:

$$\mathbb{E}[y|x, \mathbf{C}] \tag{1.1}$$

Donde $\mathbf{C}$ representa un conjunto o vector de variables de control, en el sentido de que también explican la variabilidad de y por lo que no podemos omitirlas en un proceso de estimación de la esperanza condicional.

Dicho lo anterior, el análisis empírico que proponemos en este curso está basado en identificar correlaciones que pueden ser interpretadas como causalidad y que consiste en estimar a la ecuación (1.1). En la ecuación (1.1) hemos asumido que $\mathbf{C}$ contiene toda la información disponible que sirve de control, por lo que es un vector que contiene sólo información que es observable. Sin embargo, no siempre es posible observar toda la información que pudiera servir de control.

Por ejemplo, al responder qué factores determinan el salario de las personas podríamos pensar en múltiples factores cuantificables y observables como: edad, sexo, años de educación, etc. No obstante, también consideraríamos factores como las habilidades propias de cada individuo, las cuales no son necesariamente observables.

Existen tres casos de los tipos de datos que son comúnmente analizados en el análisis de regresión lineal. El primero es un análisis de datos de sección cruzada en el que las ecuaciones representativas de la ecuación (1.1) son como la siguiente:

$$ln(Salario_i) = \beta_0 + \beta_1 Educacion_i + \beta_2 Experiencia_i$$

Donde las variables de $Salario_i$, $Educacion_i$ y $Experiencia_i$ son observadas para cada uno de los individuos en la muestra es indexado por $i = 1, 2, 3, \ldots, N$.

El segundo ejemplo se trata de datos de series de tiempo como la siguiente expresión:

$$ln(PIB_t) = \beta_0 + \beta_1 Empleo_t + \beta_2 Salarios_t$$

Donde las variables PIB_t, $Empleo_t$ y $Salarios_t$ son observadas para un individuo o entidad a lo largo de una muestra del tiempo indexada por $t = 1, 2, 3, \ldots, T$.

El tercer ejemplo resulta de la combinación de los dos anteriores. De esta forma podríamos observar a una muestra de un conjunto de individuos, $i = 1, 2, 3, \ldots, N$, a lo largo del tiempo, $t = 1, 2, 3, \ldots, T$, y analizar el comportamiento de una variable através de una ecuación como la siguiente:

$$ln(Salario_{it}) = \beta_0 + \beta_1 Educacion_{it} + \beta_2 Experiencia_{it}$$

Donde las variables $Salario_{it}$, $Educacion_{it}$ y $Experiencia_{it}$ se observan para los mismos individuos a lo largo del tiempo que se indexan con los pares (i, t), $i = 1, 23, , \ldots, N$ y $t = 1, 2, 3, \ldots, T$.

1.1.1. Esperanza condicional

En primer lugar estableceremos un poco de notación. En análisis de regresión siempre partimos de una representación de una ecuación lineal como:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK} + \varepsilon_i \tag{1.2}$$

Donde cada una de las variables y_i, x_{ik} y ε_i, se observan para $i = 1, 2, 3, \ldots, N$ y $k = 1, 2, \ldots, K$, por lo que podremos utilizar una representación para cada i de la forma:

$$\begin{aligned} y_i &= \begin{pmatrix} x_{i1} & x_{i2} & \ldots & x_{iK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \varepsilon_i \\ &= \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_i \end{aligned} \tag{1.3}$$

La ecuación (1.3) se puede generalizar para el total de elementos en la muestra como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.4}$$

Donde:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{pmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \dots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

En la mayoría de los casos analizados vamos a asumir que la ecuación (1.2) tiene un término constante, por lo que $x_{i1} = 1$ para todo $i = 1, 2, \dots, N$. Bajo este escenario, representaremos la matriz $\mathbf{X}$ con una columna compuesta del número 1 (uno) en todas sus entradas, tal que, el parámetro β_1 es un término constante en las ecuaciones (1.3) y (1.4). De esta forma la matriz anteriormente mostrada se puede ver como:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1K} \\ 1 & x_{22} & \dots & x_{2K} \\ 1 & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n2} & \dots & x_{nK} \end{bmatrix}$$

De forma similar al caso de una regresión lineal que incluye un término constante, podemos construir matrices $\mathbf{X}$ para casos en que alguna de las variables explicativas es dicotómica –que toma valores de 0 y 1– y casos en los que las variables explicativas han sido interactuadas –multiplicada entre sí– o tranformadas mediante potencias o logaritmos.

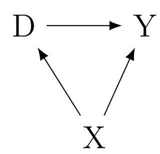


Figura 1.1: Ejemplo 1 de un DAG

1.1.2. Introducción a inferencia causal

Introducción a los Gráficos Acíclicos Dirigidos o Directed Acycle Graphical

La notación de los Gráficos Acíclicos Dirigidos (o DAG, en inglés) representa que la causalidad va en una dirección. Así, para mostrar causalidad en reversa es necesario creas múltiples nodos. La simúltaneidad, tal como en el caso de las curvas de oferta y demanda, no tiene una representación directa o inmediata con los DAG. Otra forma plantear a los DAG es que se trata de una explicación de un fenómeno en términos de contrafactuales.

Definición 1.1 *Un DAG es una representación gráfica de una cadena de efectos causales. Los efectos causales están en sí mismos basados en algún proceso subyacente no observable.*

Los efectos causales pueden observarse en dos vías:

1. Directa: $D \longrightarrow Y$

2. Indirecta, a través de una tercer variable: $D \longrightarrow X \longrightarrow Y$

En este setido, la ausencia de $\longrightarrow$ indica que no hay relación entre las variables.

Un ejemplo sencillo de DAG es el siguiente:

El DAG mostrado en la Figura (1.1) ilustra que hay una ruta directa de D a Y, lo cual representa un efecto causal. Por su parte la ruta a D tiene una ruta trasera (backdoor path) a través de la ruta $D \longleftarrow X \longrightarrow Y$. En este caso consideremos que la ruta directa es un efecto causal, pero la backdoor path es no causal. En su caso este proceso crea una correlación espúrea entre D y Y.

Pensemos el proceso de ruta trasera como una situación en la que a veces cuando D toma diferentes valores, Y toma diferentes valores debido a que X

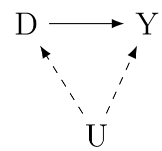


Figura 1.2: Ejemplo 2 de un DAG, confusor

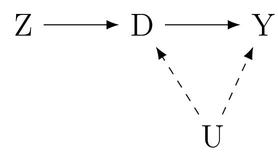


Figura 1.3: Ejemplo 4 de un DAG, IV estimator

toma diferentes valores. De esta forma, decimos que existe una correlación espuria entre D y Y.

Un segundo ejemplo de una DAG que nos permite ilustrar el concepto de confusor, el cual describe a una variable que no es observable, U, y que representamos su relación en el DAG con líneas punteadas de la forma que se muestra en la Figura (1.2).

De forma similar al caso anterior, existen dos formas para ir de D a Y. Existe la ruta directa de D a Y, lo cual representa un efecto causal. Por su parte la ruta a D tiene una ruta trasera a través de la ruta $D \longleftarrow U \longrightarrow Y$, pero con la diferencia de que la variable U es no observable.

¿Cómo solucionar o mitigar este efecto confusor? Una manera es conocida como usar variables instrumentales. La Figura 1.3 ilustra que U son un conjunto de factores no observables que impieden indenticar la causalidad entre D y Y. De esta forma, bucamos un Z que no está correlacionado con U pero si con D con el objeto de hace una estimación auxiliar.

Veamos otro ejemplo. Una pregunta clásica en economía y, a veces, en el discurso común es que la educación, en general, tiene el potencial de incrementar los ingresos laborales de las personas. Una posible teoría indicaría que la educación incrementa la productividad de las personas, de esta forma los trabajadores más educados deberían ser mejor pagados, ya que sus salarios son establecidos en función de su productividad. Así, la teoría afirma que la educación incrementa los ingresos.

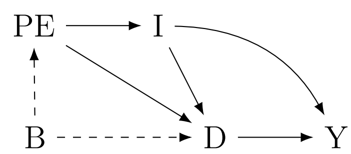


Figura 1.4: Ejemplo 3 de un DAG, ganancias de la educación

Particularicemos el ejemplo a un caso en el cual analizamos el efecto que tiene la educación secundaria (a veces denominada, educación media superior) en los ingresos. Partamos del hecho de que educarse a un nivel secundario no es un fenómeno aleatorio, puesto que existen múltiples factores que lo pueden explicar. Partamos de el siguiente DAG en el cual D será la variable de tratamiento o la variable que indica si una persona tiene formación de educación secundaria; Y es nuestra variable de interés o los ingresos; PE es la variable que indica el nivel educativo de los padres; I es una variable de ingreso familiar, y B son el background no observable de la persona (factores como la genética, habilidades individuales, ambiente familiar, etc.).

El DAG de la Figura 1.4 es en sí mismo una historia. Cada persona tiene un cierto background que normamente no se reporta en los conjuntos de datos. Medidas como el grado de inteligencia, la personalidad, la estabilidad emocional, capacidad de resilencia, dinámica familar y cualquier otro relacionado con factores ambientales o del entorno no se encuentran en los datos comúnmente. Por lo que se les denomina factores no observables.

Estos factores "ambientales" (B) están correlacionados o expresados en variables de padres e hijos. Este background causa que los padres elijan un nivel educativo o ruta de educación para los hijos y también afecta las decisiones individuales de los niños respecto de la ruta educativa que quieren seguir.

En este caso, tabién es posible notar que los DAG cuentan 2 historias. Nos dicen lo que está pasando, pero también nos dicen lo que no está pasando. Así, B no tiene un efecto directo en los ingresos laborales, excepto a través de su efecto en la elección educativa. Sin embargo, en muchas ocasiones se suele criticar este tipo de planteamientos por parecer más un supuesto que un hecho real. En estos casos, la desición es del investigador.

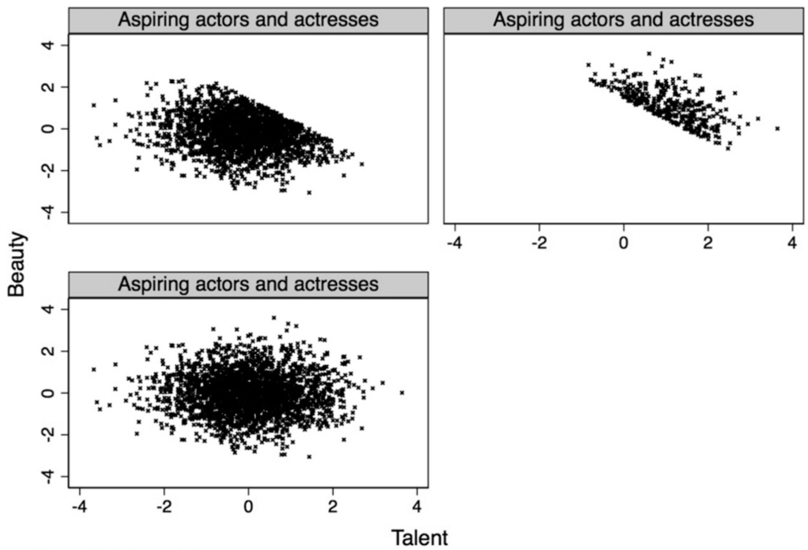
Truncamiento o cómo evitar sesgar nuestros análisis

Tomemos un ejemplo (retomado de Cunningham (2021) [Cun21]). Una publicación del blog de CNN de 2009 informó que Megan Fox, quien protagonizó la película Transformers, fue votada como la peor y la más atractiva actriz de 2009 en una encuesta sobre estrellas de cine. La publicación infería que había una implicación de que el talento y la belleza están negativamente correlacionados. ¿Pero lo son? ¿Y por qué podrían serlo? ¿Qué pasa si en realidad son independientes entre sí pero están correlacionados negativamente en una muestra de estrellas de cine debido al sesgo de cómo se seleccionó la información? ¿Es eso es posible? Es decir, ¿el talento y la belleza están negativamente correlacionados?

En la figura 1.5 muestra una simulación de la situación. El panel inferior izquierdo muestra el diagrama de dispersión entre talento y belleza. Observe que las dos variables son extracciones aleatorias e independientes de la distribución normal estándar, lo que crea una nube de datos alargada. Pero debido a que .estrella de cine"se encuentra en el percentil 85 superior de la distribución de una combinación lineal de talento y belleza, la muestra consta de personas cuya puntuación combinada se encuentra en la parte superior derecha de la distribución conjunta. Esta frontera tiene una pendiente negativa y se encuentra en la parte superior derecha de la nube de datos, lo que crea una correlación negativa entre las observaciones en la muestra de estrellas de cine.Sin embargo, sabemos que, de hecho, no hay relación entre las dos variables. Este tipo de selección de muestras crea correlaciones espurias. Una muestra aleatoria de la población completa sería suficiente para mostrar que no existe relación entre las dos variables, pero al dividir la muestra en estrellas de cine únicamente, introducimos correlaciones espurias entre las dos variables de interés.

El mensaje final es que debemos tener cuidado de no estar analizando una muestra trunca de la información y, más aún, pretender hacer inferencias generales con muestras truncas de la población.

Figura 1.5: Relación entre belleza y talento, retomado de [Cun21].



1.2. 4.1 Aplicación de modelos de regresión lineal.

1.2.1. El estimador de Mínimos Cuadrados Ordinarios (MCO) y el análisis clásico de regresión

El concepto de regresión se puede ilustrar en las Figuras 1.6, 1.7 y 1.8.

Si partimos de las ecuaciones (1.3) u (1.4), podemos establecer que el término de error estará dado por:

$$\varepsilon_i = y_i - \mathbf{X}_i'\boldsymbol{\beta} \tag{1.5}$$

Donde $i = 1, 2, \ldots, N$. De forma similar podremos decir que un estimador de éste término de error será aquel que resulte de:

$$e_i = y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}} \tag{1.6}$$

Donde $\hat{\boldsymbol{\beta}}$ es un vector de estimadores de los parámetros $\boldsymbol{\beta}$. De lo dicho hasta ahora es fácil ver que la siguiente ecuación es cierta $\forall i$:

$$y_i = \mathbf{X}_i'\boldsymbol{\beta} + \varepsilon_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + e_i \tag{1.7}$$

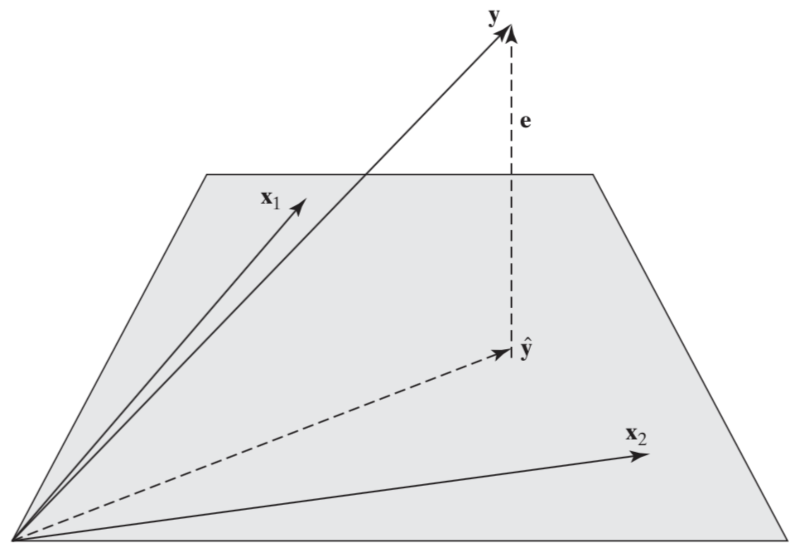


Figura 1.6: Ilustración de la poroyección de y en x_1 y x_2 y el significado del error e, retomado de Greene (2012, pp. 20) [Gre12]

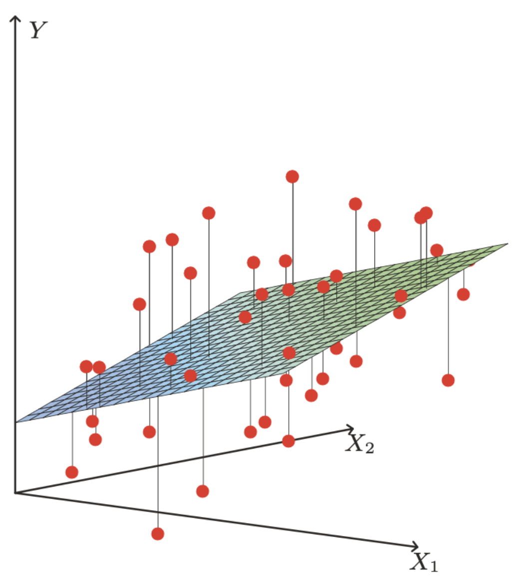


Figura 1.7: Ilustración del hiperplano generado dee regresar X_1 y X_2 en Y, retomado de Hastie, Tibshirani y Friedman (2009, pp. 45) [HTF17]

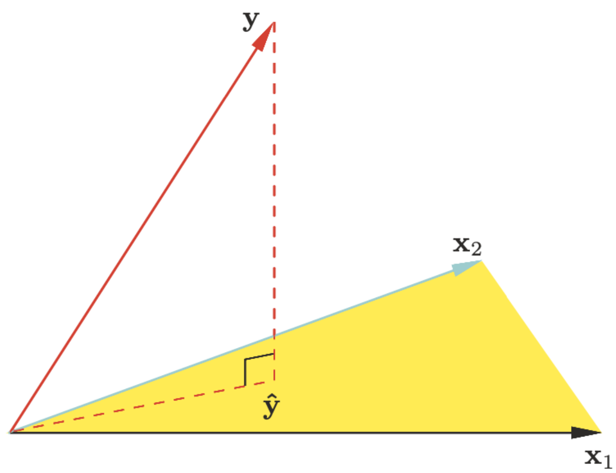


Figura 1.8: Ilustración de la poroyección de y en x_1 y x_2 y el significado del error e, retomado de Hastie, Tibshirani y Friedman (2009, pp. 46) [HTF17]

Intuitivamente, la ecuación (1.7) significa que siempre que poseamos una muestra de los elementos de la población, podremos explicar una parte de la variable dependiente, no su totalidad. En este sentido, el análisis de regresión consiste en un proceso de ajuste a la variable dependiente. Está es la idea que da origen al R^2 y otras medidas de bondad de ajuste, mismas que se analizan en textos convencionales de análisis de regresión.

El método de MCO, en consecuencia, resulta en encontrar la combinación de valores de los estimadores de los parámetros $\hat{\boldsymbol{\beta}}$ que permita minimizar la suma de los residuales (estimadores de los términos de erro $\boldsymbol{\varepsilon}$) al cuadrado dada por:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}})^2 \tag{1.8}$$

Donde $\hat{\boldsymbol{\beta}}$ denota el vector de estimadores $\hat{\beta}_1, \ldots, \hat{\beta}_K$. En términos matriciales, dado que $(e_1, e_2, \ldots, e_n)'(e_1, e_2, \ldots, e_n) = \mathbf{e}'\mathbf{e}$, el problema del método de MCO consiste en resolver el problema de óptimización:

$$\begin{aligned} Minimizar_{\hat{\boldsymbol{\beta}}} S(\hat{\boldsymbol{\beta}}) &= Minimizar_{\hat{\boldsymbol{\beta}}} \mathbf{e}'\mathbf{e} \\ &= Minimizar_{\hat{\boldsymbol{\beta}}} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \end{aligned}$$

Expandiendo la expresión $\mathbf{e}'\mathbf{e}$ obtenemos:

$$\mathbf{e}'\mathbf{e} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \tag{1.9}$$

De esta forma obtenemos que las condiciones necesarias de un mínimo son:

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0} \tag{1.10}$$

De ecuación anterior obtenemos para la solución del problema del mínimo a las ecuaciones siguientes conocidas como *ecuaciones normales* dadas por:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y} \tag{1.11}$$

Notemos que dichas ecuaciones normales son en realidad un sistema de ecuaciones de K variables o incógnitas. Por un lado, recordemos que $\mathbf{X}$ es una matriz de dimensión $N \times K$, con lo cual $\mathbf{X}'$ es de dimensión $K \times N$. Así, el producto $\mathbf{X}'\mathbf{X}$ dará como resultado una matriz cuadrada de dimensión $K \times K$. Por otro lado, sabemos que $\mathbf{Y}$ es un vector de tamaño $N \times 1$, con lo cual el producto $\mathbf{X}'\mathbf{Y}$ da como resultado un vector de dimención $K \times 1$. En conclusión, el sistema de ecuaciones normales consiste en K ecuaciones con K incógnitas $(\hat{\beta}_1, \ldots, \hat{\beta}_K)$. Ante este hecho, existen múltiples formas mediante las cuales se puede solucionar dicho sistema, sin embargo en nuesto caso seguiremos el siguiente procedimiento de operaciones matriciales.

Si las matriz $\mathbf{X}$ es de rango completo por columnas, entonces la inversa de la matriz $\mathbf{X}'\mathbf{X}$ existe. De esta forma, la solución esta dada por la siguiente expresión:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.12}$$

Finalmente, para que esta solución dada para el procedimiento de MCO y mostrada en la ecuación (1.12) sea un mínimo, debemos buscar las condiciones de segundo orden:

$$\frac{\partial^2 S(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}} \partial \hat{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X} \tag{1.13}$$

Donde la matriz $\mathbf{X}'\mathbf{X}$ debe ser positiva definida para que la solución de MCO sea un mínimo. Sea $q = \mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c}$ para algún vector $\mathbf{c}$ distinto de cero. Entonces:

$$q = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2, \text{ donde } \mathbf{v} = \mathbf{X}\mathbf{c}$$

Así, q es positivo. Si $\mathbf{v}$ fuera cero, entonces existe una combinación lineal de las columnas de $\mathbf{X}$ que da como resultado cero, lo cual contradice el supuesto de que $\mathbf{X}$ es de rango completo. En todos los casos, si $\mathbf{X}$ es de rango completo, entonces la solución del método de MCO, $\hat{\boldsymbol{\beta}}$, es la única que mínimiza la suma de los residuales al cuadrado.

Bondad de ajuste

Una vez planteada la solución de MCO plantearemos una medida para determinar en qué grado los datos estimados, definidos como $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, se ajustan al valor real de $\mathbf{Y}$. La medida o métrica es el R^2, la cual contrasta el análisis de regresión respecto de hacer una simple estimación de una media para dar un pronóstico de $\mathbf{Y}$. El coeficente R^2 está montado en el supuesto de que el modelo incluye un término constate. Si el modelo no incluye una constante, no es posible hacer una interpretación del R^2, en los siguientes parráfos abundaremos al respecto.

Antes de iniciar el desarrollo del R^2, partamos de que la suma de residuales es igual a cero ($\sum_{i=1}^{n} e_i = 0$), si y sólo si, el modelo tiene un término constante. Un corolario de este hecho es que el valor promedio de los residuales es cero, sí y solo si, el modelo tiene un término constante, esto es, la única forma en que se cumpla la siguiente expresión es que la suma de residuales sea cero:

$$\begin{aligned}
\overline{e} &= \frac{\sum_{i=1}^{n} e_i}{n} \\
&= \frac{1}{n} \cdot \begin{bmatrix} 1, & 1, & 1, & \ldots & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \\
&= 0
\end{aligned}$$

Condición que resulta de las ecuaciones normales de MCO, recordemos que ellas se derivan de que $\mathbf{X}'\mathbf{e} = \mathbf{0}$, y que el primer producto punto de vectores implica a la columna de la constante en la matriz $\mathbf{X}$ y que está definido por $\begin{bmatrix} 1, & 1, & 1, & \ldots & 1 \end{bmatrix}$.

El R^2 descompone la variación total en dos tramos: la variación originada por la regresión y la variación originada por el término de error. Como primer paso definamos la variación total observada respecto de la media ($\overline{Y}$) como:

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\overline{Y})^2 &= \begin{bmatrix} y_1-\overline{Y}, & y_2-\overline{Y}, & y_3-\overline{Y}, & \ldots & y_n-\overline{Y} \end{bmatrix} \begin{bmatrix} y_1-\overline{Y} \\ y_2-\overline{Y} \\ y_3-\overline{Y} \\ \vdots \\ y_n-\overline{Y} \end{bmatrix} \\
&= (\mathbf{Y}-\overline{\mathbf{Y}})'(\mathbf{Y}-\overline{\mathbf{Y}}) \\
&= \left[\mathbf{Y}-\frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}\mathbf{Y}\right]' \left[\mathbf{Y}-\frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}\mathbf{Y}\right] \\
&= \mathbf{Y}'\left[\mathbb{I}_n-\frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}\right]' \left[\mathbb{I}_n-\frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}\right]\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{M}^0\mathbf{Y} \qquad (1.14)
\end{aligned}
$$

En la expresión (1.14) a $\mathbf{M}^0$ se puede interpretar como una matriz que resta a una matriz o vector su promedio, notemos que como resultado dará las desviaciones respecto de la media. Asimismo, $\mathbf{M}^0$ tiene un par de propiedades que son fácilmente demostrables: idémpotencia y simetría, es decir, i) $\mathbf{M}^0 = \mathbf{M}^{0'}$ y ii) $\mathbf{M}^0\mathbf{M}^0 = \mathbf{M}^0$. Dicho lo anterior, retomemos a (1.14) para mostrar que:

$$
\begin{aligned}
\mathbf{Y}'\mathbf{M}^0\mathbf{Y} &= \mathbf{Y}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{Y} \\
&= (\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e})'\mathbf{M}^{0'}\mathbf{M}^0(\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}) \\
&= (\hat{\boldsymbol{\beta}}'\mathbf{X}'+\mathbf{e}')\mathbf{M}^{0'}\mathbf{M}^0(\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}) \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{e}+\mathbf{e}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e}+\mathbf{e}'\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^{0'}\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}'\mathbf{e} \qquad (1.15)
\end{aligned}
$$

No debe pasar desapercibido que el desarrollo algebraíco para llegar a la ecuación (1.15) sólo es posible si el promedio de los residuales es cero, es decir:

$$\mathbf{M}^0\mathbf{e} = \mathbf{e} - \begin{bmatrix} \bar{e} \\ \bar{e} \\ \bar{e} \\ \vdots \\ \bar{e} \end{bmatrix} = \mathbf{e} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{e} \tag{1.16}$$

De otra forma, no se puede concluir la expresión (1.15). Finalmente, (1.15) la expresaremos así:

$$\mathbf{Y}'\mathbf{M}^0\mathbf{Y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}'\mathbf{e} \tag{1.17}$$

Así, como mencionamos anteriormente, (1.17) se puede interpretar como que la variación total respecto de la media se puede descomponer en dos variaciones, una que se origina de la regresión, $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}$, y otra de los residuales, $\mathbf{e}'\mathbf{e}$. De esta forma planteamos que R^2 es una metrica que cuantifica cuánto de la variación total es explicada por la regresión y cuánto es explicada por los residuales (es decir, por la información no observable):

$$\begin{aligned} R^2 &= \frac{Var.Regresion}{Var.Total} \\ &= \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{M}^0\mathbf{X}\hat{\boldsymbol{\beta}}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \\ &= \frac{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \\ &= 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \end{aligned} \tag{1.18}$$

Existe una expresión más que se le denomina como un R^2 ajustado o $R^2 Adj.$, el cual castiga por un uso excesivo de variables independientes:

$$R^2 Adj. = 1 - \frac{n-1}{n-K}\frac{\mathbf{e}'\mathbf{e}}{\mathbf{Y}'\mathbf{M}^0\mathbf{Y}} \tag{1.19}$$

Así, es obvio que la siguiente relación siempre es cierta: $R^2 \geq R^2 Adj.$.

Inferencia asintótica bajo MCO

En estadística es común que porpongamos como estimador de σ^2 a:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} \tag{1.20}$$

El cual es un estimador insesgado de la varianza. Además, queda pendiente en esta sección es demostrar que el estimador de MCO ($\hat{\boldsymbol{\beta}}$) alcanza la cota inferior Cramér-Rao. De esta forma podríamos afirmar que el estimador es de mínima varianza o, equivalentemente, es el más eficiente.

Dicho lo anterior podemos construir las dos pruebas: t y F. Previo a dicha construcción debemos recordar que una prueba t se construlle por la relación de una función normal con media cero y variaza 1 (uno), y la raíz cuadrada de una función chi-cuadrado. Supongamos que una variable Z se distribuye de forma normal con media cero y variaza 1 (uno), y que la variable χ^2_m se distribulle de forma chi-cuadrado con m grados de libertad, entonces:

$$t = \frac{Z}{\sqrt{\frac{\chi^2}{m}}} \sim t_m \tag{1.21}$$

Donde t_m tal y como se describe en (1.21) es una distribución t de Student con m grados de libertad. Por otro lado, cuando se tiene una variable Z con distribución normal con media cero y varianza $\sigma^2 = 1$, entonces lo siguiente es cierto:

$$Z^2 \sim \chi^2_1$$

Así, la suma de variables chi-cuadrado es una chi-cuadaro en los siguientes términos:

$$\sum_{i=1}^{m} Z^2 \sim \chi^2_m$$

Dada la distribución $\hat{\boldsymbol{\beta}}$, y que la varianza de cada uno de los elementos de $\hat{\boldsymbol{\beta}}$, digamos $\hat{\beta}_k$, es: $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}$. En este punto denotaremos con $\hat{\beta}_k$ al elemento k-ésimo dentro de $\hat{\boldsymbol{\beta}}$. Adicionalmente, con $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$ al elemento en la fila k y la columna k en la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Derivado de lo anterior, lo siguiente es cierto:

$$Z = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim \mathcal{N}(0, 1)$$

Considerando la distribución de ε_i y de que e_i es un estimador de ε_i, podemos afirmar que:

$$Z = \frac{e_i - 0}{\sigma} = \frac{e_i}{\sigma} \sim \mathcal{N}(0,1)$$

De esta forma:

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} \sim \chi_n^2$$

Dada la distribución t de Student en la ecuación (1.21) y los elementos antes mencionados podemos llegar a las siguientes conclusiones:

$$\begin{aligned}
t &= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (\mathbf{X'X})^{-1}_{kk}}}}{\sqrt{\sum_{i=1}^{n} \frac{e_i^2}{\sigma^2} / (n-K)}} \\
&= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (\mathbf{X'X})^{-1}_{kk}}}}{\sqrt{\sum_{i=1}^{n} \frac{e_i^2}{(n-K)\sigma^2}}} \\
&= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (\mathbf{X'X})^{-1}_{kk}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \\
&= \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2}\sqrt{(\mathbf{X'X})^{-1}_{kk}}}}{\frac{\sqrt{\hat{\sigma}^2}}{\sqrt{\sigma^2}}} \\
&= \frac{(\hat{\beta}_k - \beta_k)\sqrt{\sigma^2}}{\sqrt{\hat{\sigma}^2}\sqrt{\sigma^2}\sqrt{(\mathbf{X'X})^{-1}_{kk}}} \\
&= \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (\mathbf{X'X})^{-1}_{kk}}}
\end{aligned}$$

De esta forma:

$$t = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (\mathbf{X'X})^{-1}_{kk}}} \sim t_{n-k} \qquad (1.22)$$

Construida esta expresión (1.22) podemos establecer el siguiente intervalo de confianza para cada uno de nuestros estimadores $\hat{\beta}_k$:

$$-t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2 (\mathbf{X'X})^{-1}_{kk}} < \hat{\beta}_k - \beta_k < t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2 (\mathbf{X'X})^{-1}_{kk}}$$

$$\hat{\beta}_k - t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}} < \beta_k < \hat{\beta}_k + t_{\alpha/2,n-K}\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}$$

Donde $t_{\alpha/2,n-K}$ es el valor crítico de tablas t-Student. Así, podemos enunciar que el estimador de MCO de la ecuación (1.12) se sujetará a la siguiente prueba de hipotésis general:

$$\begin{aligned} H_0 : \beta_k &= r \\ H_a : \beta_k &\neq r \end{aligned}$$

Donde $k = 1, 2, \ldots, K$ y r es un valor respecto del cual se desea comparar a $\hat{\beta}_k$. Dicha prueba se distribuye como una t_{n-K} y su especificación más común en el análisis de regresión es la siguiente:

$$\begin{aligned} H_0 : \beta_k &= 0 \\ H_a : \beta_k &\neq 0 \end{aligned}$$

Lo que en términos de una prueba t es la siguiente:

$$t = \frac{\hat{\beta}_k - 0}{\sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim t_{n-k} \tag{1.23}$$

Dicha prueba la podemos interpretar como sí es posible que el estimador $\hat{\beta}_k$ sea igual a 0, o en térmninos económicos, si existe evidencia estadística de que el efecto de la variable x_k en y es nulo.

Como se mostró anteriormente, sólo se hemos hablado de pruebas de hipotésis cuándo se desea comprobar si alguno de los estimadores $\hat{\beta}_k$ de MCO es estadísticamente igual a algún valor determinado. No obstante, en ciertas condiciones puede ser interesante cuestionar si en conjunto todos los estimadores cumplen una cierta restricción. A este tipo de situaciones se les suele conocer como pruebas de hipótesis globales. Sin pérdida de generalidad podemos enunciar a una prueba global como:

$$\begin{aligned} H_0 : \mathbf{R}\boldsymbol{\beta} &= \mathbf{r} \\ H_a : \mathbf{R}\boldsymbol{\beta} &\neq \mathbf{r} \end{aligned}$$

Donde, $\boldsymbol{\beta}$ es el vector de parámetros del MCO de la ecuación (1.12), $\mathbf{R}$ y $\mathbf{r}$ son de la siguiente manera:

$$\mathbf{R} = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1K} \\ R_{21} & R_{22} & \dots & R_{2K} \\ R_{31} & R_{32} & \dots & R_{3K} \\ \vdots & \vdots & \dots & \vdots \\ R_{J1} & R_{J2} & \dots & R_{JK} \end{bmatrix}$$

En $\mathbf{R}$ cada una de las R_{jk} toma el valor de 1 o 0, según sea la estructura de la restricción. Con $j = 1, 2, \dots, J$. En el caso de $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_J \end{bmatrix}$$

Donde cada uno de los r_j representa el valor de la restricción que es evaluada. No debe pasar desapercibido que en este caso existen J restricciones $(J \leq K)$, mismas que se evaluan de forma conjunta en una sola prueba. Esa es la ventaja de la prueba que se enuncia a continuación.

Definamos a F de Fisher como la razón de dos pruebas t de Student, la primera para los valores de una combinación lineal de $\hat{\boldsymbol{\beta}}$, y la segunda para el estimador de la varianza $\hat{\sigma}^2$:

$$\begin{aligned} \mathbf{F} &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/J}{\hat{\sigma}^2} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \\ &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}Var(\hat{\boldsymbol{\beta}})\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \end{aligned}$$

Por lo tanto la estadística F de prueba será:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{J} \sim F_{J,n-K} \tag{1.24}$$

Así, la prueba de hipótesis que es la más común en el análisis de regresión y que se le conoce como prueba global. Dicha prueba asume una forma de la matriz $\mathbf{R}$ definida así:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Y un vector $\mathbf{r}$:

$$\mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

De esta forma la hipotésis nula de una prueba global se puede escribir como:

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \ldots = \beta_K = 0 \\ H_a: &\quad NoH_0 \end{aligned}$$

Esta prueba se le conoce como prueba global, ya que prueba si en conjunto todas las variables independientes tienen un efecto nulo en $\mathbf{Y}$.

En este caso nos referimos a casos como:

$$H_0 : \mathbf{R}\boldsymbol{\beta} = r$$

Donde $\mathbf{R}$ es una matriz de restricciones de dimensión $Q \times K$, con $Q \leq K$, r es un vector de dimensión $Q \times 1$. Así en casos de heterocedasticidad utilizaremos una prueba de Wald dada por:

$$\mathbf{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - r)'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - r)$$

Donde $\hat{\mathbf{V}} = \widehat{AVar(\hat{\boldsymbol{\beta}})}$, por lo que la distribución de $\mathbf{W}$ será como:

$$\mathbf{W} \sim \chi^2_Q$$

Lo que implica que la formulación $F_{Q,N-K}$ no es de aplicación.

Otro tipo de pruebas que pueden realizarse son aquellas basadas en Multiplicadores de Lagrange. Supongamos un modelo lineal particionado de la siguiente manera:

$$y_i = \mathbf{x}_{i1}\boldsymbol{\beta}_1 + \mathbf{x}_{i2}\boldsymbol{\beta}_2 + \varepsilon \tag{1.25}$$

Supongamos que desemos probar la hipótesis de que:

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

Sea $\widetilde{\boldsymbol{\beta}}_1$ esl estimador de $\boldsymbol{\beta}_1$ bajo la hipótesis nula anterior, digamos que este estimador corresponde al modelo restringido. Definamos los residuales:

$$e_i = y_i - \mathbf{x}_{i1}\widetilde{\boldsymbol{\beta}}_1$$

Bajo la hiótesis nula, $\mathbf{x}_{i2}$ debería ser una varible no correlacionada con e_i. En este caso la prueba estadística es obtenida como sigue:

1. Corremos una regresión de e_i en $\mathbf{x}_{i1}$ y $\mathbf{x}_{i2}$

2. Asumiendo que $\mathbf{x}_{i1}$ contiene término constante, digamos que R_E^2 denota al R^2 asociado a la regresión.

3. Entonces propongamos que la estadística de Multiplicadores de Lagrange es:

$$LM = N \cdot R_E^2$$

   Para llgar a la expresión anterior se requiere de solucionar un problema de optimización.

4. Bajo la H_0, tenemos que:

$$LM \sim \chi_{K_2}^2$$

   Donde K_2 es el número de restricciones que hemos probado. De esta forma, si $N \cdot R_E^2$ es suficientemente grande, entonces e_i estaba significativamente correlacionado con $\mathbf{x}_{i2}$, por lo que la hipótesis debió ser rechazada.

En este punto debe quedar claro que R^2 requiere que $\sum_{i=1}^{N} e_i = 0$, lo que implica que el modelo tiene un término constante.

1.2.2. Modelos no lineales

¿Qué pasa cuando la relación observada entre las variables no es lineal? Casi siempre pasa eso. A continuación discutiremos tres casos de ecuaciones que nos ayudaran a describir una relación no lineal mediante una transformación que hace la relaciones lineales. Cada uno responde a condiciones o características que la experiencia del investigador tiene que solventar.

Cada una de las ecuaciones se distingue por la tasa de cambio de la variable y en relación a los cambios de la variable x. Partamos de que la forma más simple de una relación lineal está dada por la ecuación (1.26). Note que hemos eliminado los subíndices para indicar que la ecuacion se cumple para cada uno de los elementos en la muestra, $i = 1, 2, \ldots, n$. Sin pérdida de generalidad digamos que en lugar de β_1 y β_2 empleamos a y b como constantes, como una forma de ilustrar cada uno de los fenómenos que discutimos a continuación.

$$y = a + bx \tag{1.26}$$

Esta ecuación (1.26) tiene la característica de que tiene una tasa de variación constante para y en relación a x, de esta forma:

$$\frac{dy}{dx} = b \iff dy = bdx \iff \int dy = \int bdx \iff y = a + bx$$

Los siguientes modelos no lineales tienen una motivación similar, dependen de la forma en que se asuma la variación de y en función de la variación de x.

Regresión exponencial

Supongamos que y depende de x y que los cambios en y darivados de x son proporcionales a y, es decir:

$$\frac{dy}{dx} = by$$

Donde b es una constante. De esta forma podemos encontar que:

$$\frac{dy}{dx} = by \iff \frac{dy}{y} = bdx \iff \int \frac{dy}{y} = \int bdx \iff ln(y) = a + bx$$

De esta forma, tenemos que:

$$y = e^{bx}e^{a} \iff y = ke^{bx}$$

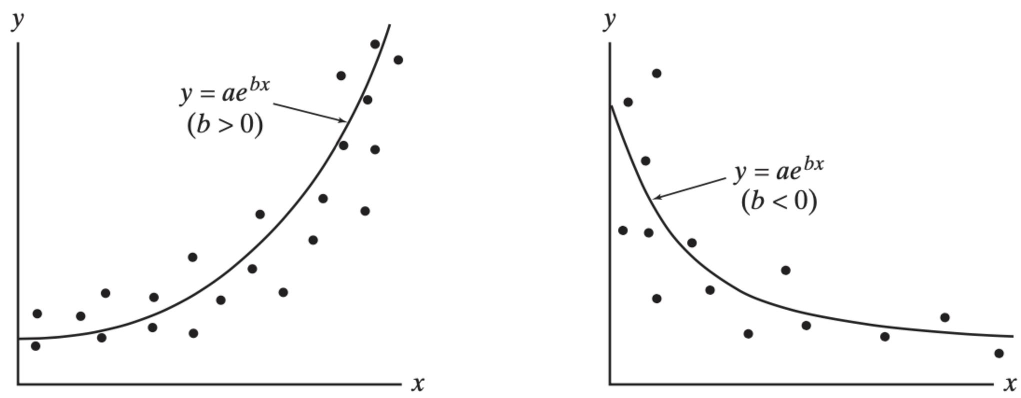


Figura 1.9: Funciones exponenciales, Retomado de Larsen y Marx (2012, p. 532) [LM12]

Por lo tanto, cuando estimamos una relación: $ln(y) = a + bx$ estamos asumiendo que existe la sigueinte relación: $\frac{dy}{dx} = by$. A este tipo de ecuaciones se les conoce como log-lineales. En la Figura 1.9 se ilustra la relación log-lineal - exponencial.

Regresión logarítmica

En este caso suponemos que los cambios en y por causa de cambios en x son proporcionales a la razón que guardan y y x, es decir:

$$\frac{dy}{dx} = b\frac{y}{x}$$

Por lo que podemos establecer:

$$\frac{dy}{dx} = b\frac{y}{x} \Longleftrightarrow \frac{dy}{y} = b\frac{dx}{x} \Longleftrightarrow \int \frac{dy}{y} = \int b\frac{dx}{x} \Longleftrightarrow ln(y) = a + ln(x)$$

Así, podemos establecer que la forma funcional de la relación de y y x es como sigue y como se muestra en la Figura 1.10.

$$y = kx^b$$

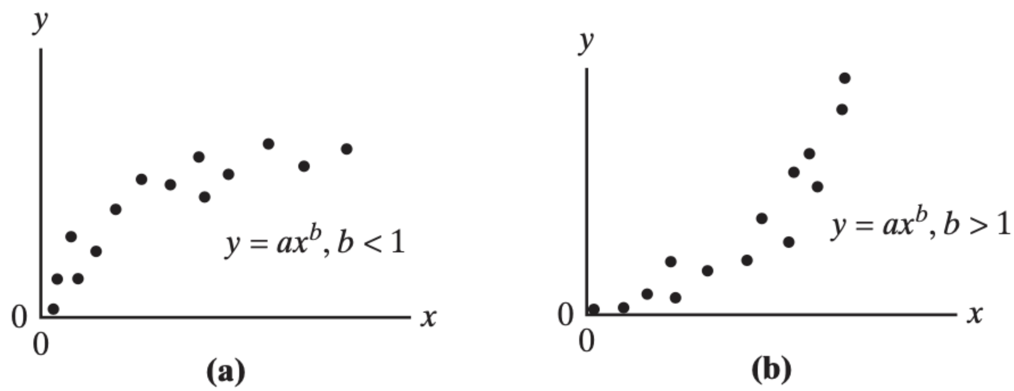


Figura 1.10: Funciones logarítmicas, Retomado de Larsen y Marx (2012, p. 536) [LM12]

Funciones logísticas

Este tipo de ecuaciones tienen la característica de que permiten modelar crecimientos poblacionales, aceptación de políticas públicas, aceptación de tecnologías, epidemias, etc. En este caso suponemos que enfrentamos que los cambios en y dados los cambios en x son proporcionales a y y a las distancia que y tiene respecto de un factor de saturación L (límite superior o punto de saturación poblacional, cobertura universal, infecciones en el total de la población, etc.).

En este caso suponemos que la variación estará dada por:

$$\frac{dy}{dx} = ky(L - y)$$

Donde k y L son constantes. Una vez solucionada la ecuación diferencial anterior, encontraremos que, si $L = 1$, entonces l a solución será como sigue, la Figura 1.11 ilustra ecuaciones como la siguiente:

$$y = \frac{1}{1 + e^{a+bx}} \tag{1.27}$$

Ahora mostremos como llegar a la solución de la ecuación logística:

$$\frac{dy}{dx} = ky(L - y) \Longleftrightarrow \frac{dy}{y(L - y)} = kdx \Longleftrightarrow \left(\frac{\frac{1}{L}}{y} + \frac{\frac{1}{L}}{(L - y)} \right) dy = kdx$$

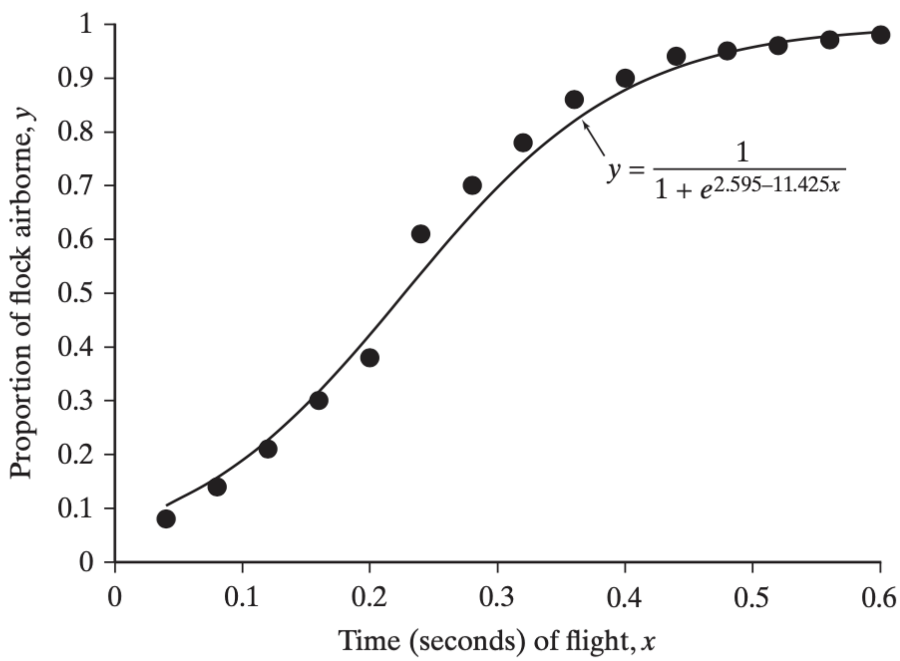


Figura 1.11: Funciones logísticas, Retomado de Larsen y Marx (2012, p. 540) [LM12]

$$\Longleftrightarrow \frac{1}{L}\left(\frac{dy}{y}+\frac{dy}{(L-y)}\right)=kdx \Longleftrightarrow \int \frac{1}{L}\left(\frac{dy}{y}+\frac{dy}{(L-y)}\right)=\int kdx$$

$$\Longleftrightarrow \left(\int \frac{dy}{y}+\int \frac{dy}{(L-y)}\right)=\int Lkdx \Longleftrightarrow (ln(y)-ln(L-y))=Lkx+C$$

$$\Longleftrightarrow (ln(L-y)-ln(y))=-LC-Lkx \Longleftrightarrow ln\left(\frac{L-y}{y}\right)=-LC-Lkx$$

$$\Longleftrightarrow \left(\frac{L-y}{y}\right)=e^{-LC-Lkx} \Longleftrightarrow y=\frac{L}{1+e^{-LC-Lkx}}$$

Por lo tanto, la solución general de la ecucación será:

$$y=\frac{L}{1+e^{a+bx}}$$

Donde asumimos que la forma lineal de la ecuación logística es:

$$ln\left(\frac{L-y}{y}\right)=a+bx$$

Funciones polinomiales

Otro tipo de ecuaciones o funciones se asocian con una relación entre y y x del tipo no lineal, pero que puede describirse bajo una formulación de un polinomio de algún grado l. Este tipo de ecuaciones tienen la característica de que permiten modelar crecimientos poblacionales, aceptación de políticas públicas, aceptación de tecnologías, epidemias, etc., mediante una especificación menos compleja en términos matemáticos.

En muchas aplicaciones, una simple transformación logaritmica de la variable x ya no puede ser suficiente para lograr un ajuste satisfactorio. Una forma sencilla de obtener más flexibilidad es utilizar la regresión polinomial. Normalmente, se prefieren polinomios de bajo grado l. En la práctica, rara vez utilizamos polinomios de grado superior a $l=3$, ya que las estimaciones pueden volverse bastante inestables con una alta variabilidad en los límites

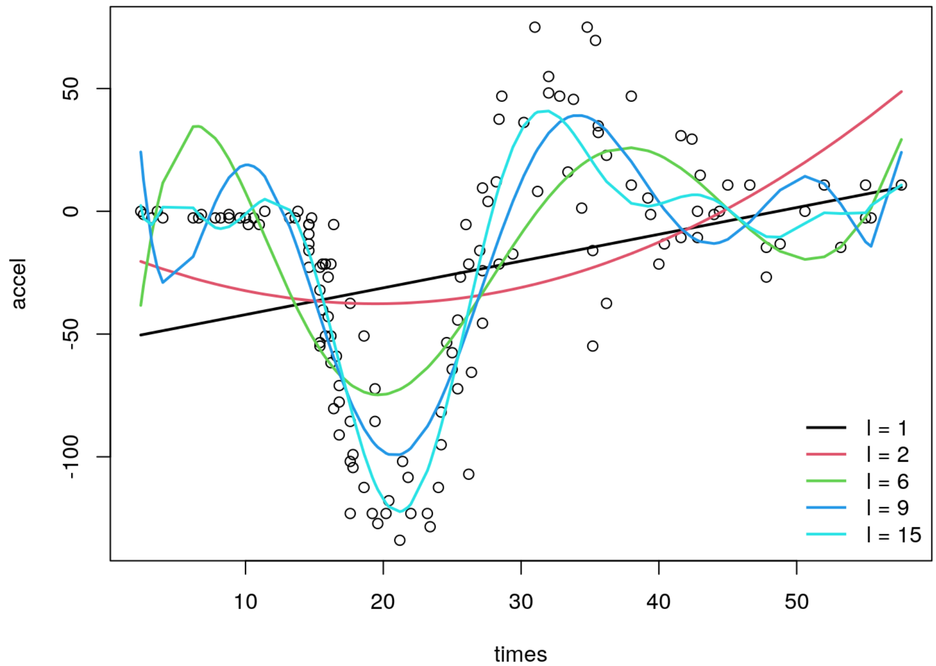


Figura 1.12: Funciones polinomiales, retomado de `https://discdown.org/flexregression/smoothreg.html`

del dominio de variables. Además, pueden surgir inestabilidades numéricas en grados muy altos, en cuyo caso se requiere de re-escalar los datos.

En general partamos de ecuaciones como:

$$y = a + b_1 x^1 + b_2 x^2 + \ldots + b_l x^l \tag{1.28}$$

La figura 1.12 ilustra diferentes escenarios de ecuaciones como 1.28con diferentes grados l.

1.2.3. Introducción al Aprendizaje Estadístico

El aprendizaje estadístico juega un rol esencial en muchas áreas de la ciencia, finanzas y la industtria. Algunos ejemplos son:

1. Predecir si un paciente–que se encuetra hospitalizado debido a un ataque al corazón–tenndrá un segundo ataque. La predicción estará basada en métricas demográficas, de la dieta y de regitros clínicos.

2. Predecir el precio de una acción en los siguienntes 6 meses; considerando la base de las medidas de desempeño de la compañía y de otros datos económicos.

3. Identificar los números en la digitalización de formas escritas a mano.

4. Identificar los factores de riesgo para el cáncer de próstata, basados en datos clínicos y de otras variables demográficas.

El aprendizaje estadístico comprende a un conjunto de herramientas para modelar y entender conjuntos de datos complejos. También se le conoce como Machine Learning (ML), el cual conjuga el desarrollo reciente en el área de la estadística junto con el crecimiento en paralelo de la computación.

El aprendizaje estadístico considera muchos métodos convencionales y de uso amplio como análisis de regresión, clasificación, árboles de decisión, etc. También se refiere a una amplia gama de herramientas para entender o interpretar datos clasificadas como supervisadas y no supervisadas.

- El aprendizaje estadístico supervisado involucra la construcción de un modelo estadístico para predecir o estimar un resultado (o variable que se pede supervisar) basado en uno o más variables explicativas.

- El aprendizaje no supervisado considera variables explicativas, pero los resultados observados no son una variable explicada que sea suceptible de supervisión.

Ahora introduzcamos un poco de notación. Denotaremos a una variable independiente o explicativa con x_{ij}, si dicha variable es un conjunto de variables acomodadas en un vector utilizaremos $\mathbf{X}_i$; en estos casos denotaremos a un elemento o variable del vector $\mathbf{X}_i$ como x_{ik}, donde $i = 1, 2, \ldots, n$ denota a los individuos en la muestra y $k = 1, 2, \ldots, K$ denota al n úmero de variables. Por convención diremos que $x_{i1} = 1$ para todo $i = 1, 2, \ldots, n$, ya que en dicha variable consideraremos al término constante en la regresión.

Por su parte, los resultados, variables dependientes o variables de respuesta se denotarán como:

- y_i denotará una respuesta que es una cantidad continua

- g_i denotará una respuesta cualitativa, discreta o de grupo

Así, con **X**, y **Y** y **G** denotaremos al a la matriz y vector columna que contiene a todos los valores de las variables dependientes y de respuesta apiladas para cada uno de los elementos en la muestra indexados con i.

Usaremos mayúsculas como X, Y o G para representar a los aspectos genéricos de las variables. De esta forma, debe ser claro que las letras minúsculas seran empleadas para representar a valores observados de las variables, así el valor observado de la variable k-ésima para el elemento de la muestra i-ésimo en **X** será representado como x_{ij}. Finalmente, con la notación $\hat{Y}$ o $\hat{G}$ representaremos a los valores estimados o predecidos.

El aprendizaje estadístico parte del establecimiento de que una variable dependiente Y es una función de un conjunto de variables explicativas $\mathbf{X} = [X_1, X_2, \ldots, X_p]$. De esta forma plateraremos:

$$Y = f(\mathbf{X}) + \varepsilon \tag{1.29}$$

Así, el aprendizaje estadístico se trata de un conjunto de aproximaciones para f. ¿Para qué estimar f? La respuesta es para hacer predicciones y para hacer inferencia. La más común de ambas razones es la predicción, con una predicción de Y podríamos establecer:

$$\hat{Y} = \hat{f}(\mathbf{X}) \tag{1.30}$$

Ajuste y separación del conjunto de datos

Supomgamos la variable objetivo Y, un vector de variables explicativas o variables 'input' X y un modelo predictivo $\hat{f}(X)$ que es estimado a partir de un conjunto de entrenamiento τ.

Definiremos la función de pérdida derivada de la estimación y capturada por los errores entre Y y $\hat{f}(X)$ estará dada por:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{error cuadrático} \\ |Y - \hat{f}(X)| & \text{error absoluto} \end{cases} \tag{1.31}$$

De esta forma podemos establecer un error cuadrático como:

$$\begin{aligned} \mathbb{E}[Y - \hat{Y}]^2 &= \mathbb{E}[f(X) + \varepsilon - \hat{f}(X)]^2 \\ &= \mathbb{E}[f(X) - \hat{f}(X)]^2 + Var(\varepsilon) \end{aligned}$$

Donde $\mathbb{E}[f(X) - \hat{f}(X)]^2$ es el único componente reducible. Para hacer predicciones requerimos de un connjunto de datos de entrenamiento y otro

Figura 1.13: División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) [HTF17]

más de prueba – en el primero estimamos $f(\cdot)$ y enn el segundo hacemos predicciones –.

Finalmente, la condición de inferencia nos permite construir pruebas de hipótesis, estimadores que cumple ciertas propiedades, intervalos de confianza, etc. Así, dividiremos al conjunto de datos conforme se describe en la Figura 1.13.

Para proporcionar una comprensión precisa de la generalización de nuestro modelo óptimo final, se suele dividir nuestros datos sólo en dos conjuntos de datos: de entrenamiento y de prueba

- Conjunto de entrenamiento: estos datos se utilizan para desarrollar conjuntos de características, entrenar nuestros algoritmos, ajustar hiperparámetros, comparar modelos y todas las demás actividades necesarias para elegir un modelo final (por ejemplo, el modelo que queremos poner en producción).

- Conjunto de pruebas: una vez elegido un modelo final, estos datos se utilizan para estimar una evaluación insesgada del rendimiento del modelo, a lo que nos referimos como error de generalización.

1.2.4. 4.1.1 Regresión lineal múltiple con datos de publicidad en medios de comunicación.

Ver RMarkdown: 01_Ejemplo_Datos_Publicidad.Rmd

1.2.5. Método de regresiones restringidas (Shrinkage methods).

Este tipo de métodos permiten (1) restringir los posibles valores de las estimacionnes del parámetro $\boldsymbol{\beta}$ a un subconjunto y (2) seleccionar variables o regresores dentro de un conjunto de candidatos. Veámos dos casos particulaes

i) una regresión de cresta destringida (ridge regression) y ii) la regresión Lasso (least absolute shrinkage and selection operator).

Regresión Ridge.

La regresión de ridge restringe los coeficientes de la regreesión mediante la imposición de una penalización en su magnitud. Los coeficientes esttimados por este método resultan de resolvel el problema de minimizar los residuales al cuadrado sujeto a que los coeficientes sumen un valor dado. Es decir:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (y_i - \mathbf{X}_i' \boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} \beta_k^2 \right] \tag{1.32}$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión, el cual no se optimiza.

La idea de la penalización de los paramétros se deriva de que la forma equivalente del problema en la ecuación (1.32) es:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 \right]$$

Sujeto a:

$$\sum_{k=2}^{K} \beta_k^2 \leq t$$

Note que la penalización no aplica al término constante, ¿por qué?, para garantizar que la estimación del hiperplano asociado pasa por la media de Y y no por el cero (0).

¿Qué implicaciones tiene para la estimación? Es posible estimar el modelo de la ecuación (1.32) pero ajustando las variables incluidas con la resta de la media. Esto, ya que el término constante se puede estimar considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Para los restantes coeficientes los determinaremos mediante un procedimiento dado por:

$$\min_{\boldsymbol{\beta}_R, \lambda} [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_R) + \lambda \boldsymbol{\beta}_R' \boldsymbol{\beta}_R] \tag{1.33}$$

Donde $\boldsymbol{\beta}_R$ contiene sólo las pendientes. Pare determinar un valor estimado debemos resolver el problema descrito en la ecuación (1.33):

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_R} S(\boldsymbol{\beta}_R) &= \frac{\partial}{\partial \boldsymbol{\beta}_R}((\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_R)'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_R)+\lambda \boldsymbol{\beta}'_R\boldsymbol{\beta}_R) \\
&= \frac{\partial}{\partial \boldsymbol{\beta}_R}(\mathbf{Y}'\mathbf{Y}-2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta}_R+\boldsymbol{\beta}'_R\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R+\lambda\boldsymbol{\beta}'_R\boldsymbol{\beta}_R) \\
&= -2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}_R+2\lambda\boldsymbol{\beta}_R
\end{aligned}$$

Determinando el mínimo:

$$\begin{aligned}
-2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_R^{Ridge}+2\lambda\hat{\boldsymbol{\beta}}_R^{Ridge} &= 0 \\
-\mathbf{X}'\mathbf{Y}+(\mathbf{X}'\mathbf{X}+\lambda\mathbb{I}_{K-1})\hat{\boldsymbol{\beta}}_R^{Ridge} &= 0
\end{aligned}$$

Por lo tanto, el estimador estará dado por:

$$\hat{\boldsymbol{\beta}}_R^{Ridge} = (\mathbf{X}'\mathbf{X}+\lambda\mathbb{I}_{K-1})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.34}$$

Así, el problema de la estimación es que tiene 1 grado de libertad, λ, que es un valor de penalización y que resultta arbitrario–determinado de forma iterativa hasta alcanzar uno que minimice la ecuación (1.32).

Regresión Lasso (Least Absolute Shrinkage and Selection Operator).

La regresión Lasso se define por la solución al problema:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}}\left[\sum_{i=1}^{n}(y_i-\mathbf{X}_i\boldsymbol{\beta})^2\right] \tag{1.35}$$

Sujeto a:

$$\sum_{k=2}^{K}|\beta_k| \leq t$$

Donde $\lambda \geq 0$ y suponemos que β_1 es el término constante de la regresión. El término constante se estima considerando que este toma el valor de la media de Y dado por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

La idea de la penalización de los paramétros se deriva de la forma equivalente del problema en la ecuación (1.35) es:

$$\hat{\boldsymbol{\beta}}^{Lasso} = \min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} (y_i - \mathbf{X}_i \boldsymbol{\beta})^2 + \lambda \sum_{k=2}^{K} |\beta_k| \right] \tag{1.36}$$

1.2.6. 4.1.2 Regresión lineal con métodos de restricción (Shrinkage Methods) y regularización de los parámetros.

Ver Python Notebook: 02_Ejemplo_Datos_Inmuebles24.ipynb

1.3. 4.2 Aplicación de modelos de regresión de respuesta binaria.

1.3.1. Métodos de estimación basados en verosimilitud

Introducción

En esta sección discutiremos el tratamiento general para la estimación por máxima verosimilitud. La diferencia significativa en estos casos es que asumiremos una distribución subyacente. Noten que los metodos utilizados anteriormente no requieren de que asumamos una distribución particular para realizar la estimación.

El planteamiento general del método de máxima verisimilitud considera un conjunto de observaciones independientes e idénticamente distribuidas descritas por la familia: $\{y_i, \mathbf{x}_i : y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^K\}$.

En estos casos también supondremos que buscamos etimar la regresión de y_i es $\mathbf{x}_i$ y no el caso contario. Por lo tanto, el método requiere de la función de densidad condicional de y_i en $\mathbf{x}_i$, es decir:

$$f(y_i|\mathbf{x}_i)$$

Así, al método en realidad se le conoce como método de máxima verosimilitud condicional. Su aplicación es amplia en casos como:

1. Modelos lineales: regresión de una sola ecuación, sistemas de ecuaciones, modelos de datos panel y modelos generalizados con heterocedásticidad y autocorrelación

2. Modelos no lineales: repuesta binaria, respuesta categórica, respuesta multinivel

3. Modelos de conteo y duración

Ejemplo. Sea y_i^* una variable aleatoria latente que se pueda expresar como:

$$y_i^* = \mathbf{x}_i\boldsymbol{\theta} + \varepsilon_i$$

Donde ε_i es independiente de $\mathbf{x}_i$, el cual es un vector de $1 \times K$ que contienen un término constante. $\boldsymbol{\theta}$ es un vector de $K \times 1$ parámetros a estimar. Finalmente, asumiremos que:

$$\varepsilon_i \sim Normal(0, 1)$$

Supongamos que en lugar de y_i^* nosotros sólo podemos observar la variable binaria indicadora:

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \leq 0 \end{cases}$$

En este caso, la pregunta que pretendemos responder sería: ¿cuál es la probabilidad de que $y_i = 1$ condicional en que tenemos el vector $\mathbf{x}_i$?

$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0|\mathbf{x}_i) \\ &= P(\mathbf{x}_i\boldsymbol{\theta} + \varepsilon_i > 0|\mathbf{x}_i) \\ &= P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) \end{aligned}$$

Dada la distribución ε_i, entonces:

$$P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = \int_{-\mathbf{x}_i\boldsymbol{\theta}}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i$$

Sabemos que toda función de densidad cumple con:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \int_{-\infty}^{-\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i + \int_{-\mathbf{x}_i\boldsymbol{\theta}}^{\infty} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \int_{-\infty}^{-\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i + \int_{-\infty}^{\mathbf{x}_i\boldsymbol{\theta}} f(\varepsilon_i|\mathbf{x}_i)d\varepsilon_i \\ &= \Phi(-\mathbf{x}_i\boldsymbol{\theta}) + \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ &= (1 - \Phi(\mathbf{x}_i\boldsymbol{\theta})) + \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ &= P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) + P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) \end{aligned}$$

De esta forma tenemos dos ecuaciones:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = \Phi(\mathbf{x}_i\boldsymbol{\theta}) \\ P(y_i = 0|\mathbf{x}_i) &= P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = 1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}) \end{aligned}$$

Estas dos funciones de probabilidad son independientes. De esta forma, la función de densidad condicional de y_i será:

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= P(y_i = 1|\mathbf{x}_i)^{y_i} \cdot P(y_i = 0|\mathbf{x}_i)^{1-y_i} \\ &= \Phi(\mathbf{x}_i\boldsymbol{\theta})^{y_i} \cdot (1 - \Phi(\mathbf{x}_i\boldsymbol{\theta}))^{1-y_i} \end{aligned}$$

Marco general de estimación por máxima verosimilitud

Sea $f(y_i|\mathbf{x}_i)$ la función de densidad condicional de y_i dado $\mathbf{x}_i$. Sea $\boldsymbol{\theta}$ un conjunto de parámetros de la función. Entonces la función de densidad conjunta de variables aleatorias independientes $\{y_i : y_i \in \mathbb{R}\}$ dados los valores $\{\mathbf{x}_i : \mathbf{x}_i \in \mathbb{R}^K\}$ estará dada por:

$$\Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = f(y_1, y_2, \ldots, y_n|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n;\boldsymbol{\theta}) = L(\boldsymbol{\theta}) \tag{1.37}$$

A la ecuación (1.37) se le conoce como ecuación de verosimilitud. El problema de máxima verosimilitud entonces será:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}) \tag{1.38}$$

Dado que el logaritmo natural es una transformación monotona, podemos decir que el problema de la ecuación (1.38) es equivalente a:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} lnL(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ln\Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{n} lnf(y_i|\mathbf{x}_i;\boldsymbol{\theta}) \tag{1.39}$$

Para solucionnar el problema se tiene que determinar las condicones de primer y segundo orden, las cuales serán:

$$\frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) = \nabla lnL(\boldsymbol{\theta}) \tag{1.40}$$

$$\frac{\partial^2}{\partial^2\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}) \cdot \frac{\partial}{\partial\boldsymbol{\theta}} lnL(\boldsymbol{\theta}') = H(\boldsymbol{\theta}) \tag{1.41}$$

La solución estará dada por aquel valor de $\hat{\boldsymbol{\theta}}$ que hace:

$$\frac{\partial}{\partial \boldsymbol{\theta}} lnL(\hat{\boldsymbol{\theta}}) = 0$$

A su vez, la varianza será aquella que resulta de:

$$Var[\hat{\boldsymbol{\theta}}|\mathbf{X}] = (-\mathbb{E}_{\hat{\boldsymbol{\theta}}}[H(\boldsymbol{\theta})])^{-1}$$

Pruebas de hipótesis

1. Para el caso de pruebas de hipótesis sobre cada uno de los coeficientes estimados $\hat{\theta}_i$, dada la construcción de la función de verosimilitud y que se trata de grandes muestras, la función de distribución de ese coeficiente sera una normal, porlo que las hipótesis planteadas serán asumiendo:
$$Z = \frac{\hat{\theta}_i - \theta_i}{\sqrt{Var(\hat{\theta}_i)}} \sim N(0, Var(\hat{\theta}_i)) \tag{1.42}$$

2. Para pruebas globales utilizaremos la prueba de Wald:
$$W = c(\hat{\boldsymbol{\theta}})'(\hat{C}Var(\hat{\boldsymbol{\theta}})\hat{C}')^{-1}c(\hat{\boldsymbol{\theta}}) \sim \chi^2_{(Q)} \tag{1.43}$$

   Donde $c(\hat{\theta})$ es una función de los coeficiennntes, $\hat{C}$ es el jacobiano de $c(\hat{\theta})$, Q es el número de restricciones.

3. Alternativamente, para pruebas sobre los coeficientes podremos utilizar pruebas de razón de verosimilitud:
$$LR = 2[lnL(\hat{\boldsymbol{\theta}}_0) - lnL(\hat{\boldsymbol{\theta}}_R)] \sim \chi^2_{(Q)} \tag{1.44}$$

1.3.2. Estimación de modelos no lineales

Modelos de respuesta binaria

Planteamiento general. En el caso de modelos de respuesta binaria, asumiremos que y_i es una variable aleatoria que toma sólo valores de 0 y 1. Los ejemplos pueden ser amplios. Decisiones sobre consumir o no, pagar o no, ir a una escuela determianda, viajar por tierra o aire. En general diremos que 1 es éxito y 0 es fracaso.

Como en el caso de modelos lineales diremos que y_i es la variable dependiente y $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$ un veector de variables independientes o explicativas. Así, en los modelos no lineales de respuesta binaria estaremos interesados en la probabilidades:

$$P(y_i = 1|\mathbf{x}_i) = P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{iK})$$

Dicho lo anterior y de forma similar al de la regresión lineal, podemos establecer el efecto marginal para el caso de una variable independiente x_j continua:

$$EMg_j = \frac{\partial}{\partial x_j} P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{iK})$$

Cuando la variable independiente es una variable dummy:

$$\begin{aligned} EMg_j &= P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ij} = 1, \ldots, x_{iK}) - \\ &\quad P(y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ij} = 0, \ldots, x_{iK}) \end{aligned}$$

Este tipo de modelos se basan en el modelo de Bernoulli, cuya función de densidad de probabilidad condicional y podemos platearlo como:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= p(x) \\ P(y_i = 0|\mathbf{x}_i) &= 1 - p(x) \\ Var(y_i|\mathbf{x}_i) = p(x) \cdot (1 - p(x)) \end{aligned}$$

Algunos planteamientos adicionales

Cuando la variable dependiente es binaria (0 y 1) no se puede implementar una ecuación lineal de la forma tradicional, ya que no es posible determinar una ecuación:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

Por otro la, tambien enfrentaremos el mismo problema con modelos de probabilidad lineal:

$$P(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

Y:

$$P(y_i = 0|\mathbf{x}_i) = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

En ambos casos es dificil lograr la estimación.

Modelos Logit y Probit

Este tipo de modelos suponene que tenemos que existe una variable latente que se puedes expresar como una ecuación lineal dada por:

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$$

Donde ε_i es una variable aleatoria con función de densidad con media cero y distribución simetrica al rededor de cero. Dado lo anterio, para nosotros sólo es visible que:

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* < 0 \end{cases}$$

De esta forma tenemos una estrucutura de la probabilidad dada por:

$$\begin{aligned} P(y_i = 1|\mathbf{x}_i) &= P(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i > 0) = P(\varepsilon_i > -\mathbf{x}_i\boldsymbol{\beta}) = G(\mathbf{x}_i\boldsymbol{\beta}) \\ P(y_i = 0|\mathbf{x}_i) &= P(\mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i < 0) = P(\varepsilon_i < -\mathbf{x}_i\boldsymbol{\beta}) = 1 - G(\mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

Donde $\mathbf{x}_i$ es un vector de dimensión $K \times 1$ que contiene al menos el término constante y $\boldsymbol{\beta}$ es un vector de parámetros a estimar, de forma que asumiremos:

$$\mathbf{x}_i\boldsymbol{\beta} = \beta_1 + x_{i2}\beta_2 + \ldots + x_{iK}\beta_K$$

Asumiremos que $G(\cdot)$ es uan función de densidad acumulada de forma que:

$$0 < G(\mathbf{x}_i\boldsymbol{\beta}) < 1 \ , \ \forall \mathbf{x}_i\boldsymbol{\beta} \in \mathbb{R}$$

En este caso utilizaremos dos modelos que dependen de la forma funcional de $G(\cdot)$ que está determinada por la distribución de ε_i. De esta forma tendremos dos modelos: Probit y Logit:

1. **Modelo Probit**, cuando

$$G(\mathbf{x}_i\boldsymbol{\beta}) = \Phi(\mathbf{x}_i\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}_i\boldsymbol{\beta}} \phi(\nu)d\nu$$

Donde $\phi(\cdot)$ es la densidad de probabilidad normal estándar:

$$\phi(\mathbf{x}_i\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi}} e^{-1/2\phi(\mathbf{x}_i\boldsymbol{\beta})}$$

2. **Modelo Logit**, cuando

$$G(\mathbf{x}_i\boldsymbol{\beta}) = \Lambda(\mathbf{x}_i\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i\boldsymbol{\beta}}}$$

Donde $\lambda(\cdot)$ es la función de densidad acumulada logística.

Sin importar el modelo que estemos ocupando, la forma de interpretar el modelo es mediante el efecto marginal, cuando x_j sea una variable continua:

$$EMg_j = \frac{\partial}{\partial x_j}P(y=1|\mathbf{x}_i) = \frac{\partial}{\partial x_j}G(\mathbf{x}_i\boldsymbol{\beta}) = g(\mathbf{x}_i\boldsymbol{\beta})\beta_j$$

Por su parte, cuando x_j sea una variable dicotómica:

$$\begin{aligned} EMg_j &= P(y=1|\mathbf{x}_i, x_j=1) - P(y=1|\mathbf{x}_i, x_j=0) \\ &= G(\mathbf{x}_i\boldsymbol{\beta}|x_j=1) - G(\mathbf{x}_i\boldsymbol{\beta}|x_j=0) \\ &= G(\beta_1 + x_2\beta_2 + \ldots + \beta_j + \ldots + x_K\beta_K) \\ &\quad -G(\beta_1 + x_2\beta_2 + \ldots + 0 + \ldots + x_K\beta_K) \end{aligned}$$

Estimación

Supongamos n observaciones de una variables aleatoria independientes e idénticamente distribuidas. En estos casos la función de densidad para cada y_i, $i = 1, 2, \ldots, n$, estará dada por:

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1-G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}, \text{ donde } y_i = 0, 1$$

En estos términnos la función de verosimilitud estará dada por:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \Pi_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) \\ &= \Pi_{i=1}^{n}[G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1-G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i} \end{aligned}$$

En versión logaritmica:

$$\begin{aligned} lnL(\boldsymbol{\beta}) &= \sum_{i=1}^{n} ln(f(y_i|\mathbf{x}_i;\boldsymbol{\beta})) \\ &= \sum_{i=1}^{n} y_i ln[G(\mathbf{x}_i\boldsymbol{\beta})] + \sum_{i=1}^{n}(1-y_i)ln[1-G(\mathbf{x}_i\boldsymbol{\beta})] \end{aligned}$$

Las condiciones de primer orden serán:

$$\frac{\partial}{\partial \boldsymbol{\beta}} lnL(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \frac{g(\mathbf{x}_i\boldsymbol{\beta})}{G(\mathbf{x}_i\boldsymbol{\beta})}\mathbf{x}_i' - \sum_{i=1}^{n}(1-y_i)\frac{g(\mathbf{x}_i\boldsymbol{\beta})}{1-G(\mathbf{x}_i\boldsymbol{\beta})}\mathbf{x}_i'$$

Finalmente, las condiciones de segundo orden y varianza estará dada por:

$$\begin{aligned} \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} lnL(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \frac{y_i g(\mathbf{x}_i\boldsymbol{\beta})^2}{G(\mathbf{x}_i\boldsymbol{\beta})[1-G(\mathbf{x}_i\boldsymbol{\beta})]}\mathbf{x}_i'\mathbf{x}_i \\ &= H(\boldsymbol{\beta}) \end{aligned}$$

Por lo tanto:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= [-\mathbb{E}[H(\boldsymbol{\beta})]]^{-1} \\ &= \left[\sum_{i=1}^{n} \frac{y_i g(\mathbf{x}_i\boldsymbol{\beta})^2}{G(\mathbf{x}_i\boldsymbol{\beta})[1-G(\mathbf{x}_i\boldsymbol{\beta})]}\mathbf{x}_i'\mathbf{x}_i\right]^{-1} \end{aligned}$$

1.3.3. Modelos de respuesta multinomial y ordenada

En esta sección analizaremos modelos de respuesta discreta con más de 2 resultados posibles, los cuales se dividen:

1. Modelos de respuesta no ordenada, son modelos conocidos como de respuesta nominal, donde los valores de los distintos resultados son arbitrarios y no tienen un efecto en la estimación. Ejemplos de estos modelos son casos de selección de seguro, lugar de hospedaje, etc.

2. Modelos de respuesta ordenada, en estos modelos se asigna a cada un valor no arbitrario a la respuesta. Por ejemplo, modelos de calificación de crédito, modelos de preferencia de bienes, etc.

Modelos de respuesta multinomial

Logit multinomial

Este primer modelo aplica en situaciones en las que la unidad de respuesta o elección depende de las características individuales de los elementos de la muestra, pero no de la atribuciones de la elección.

Dado lo anterior, definiremos el modelo en términos de variables aleatorias que representan a la población subyacente. Sea y_i una variable aleatoria que toma valores en un conjunto $\{0, 1, 2, \ldots, J\}$. Sea $\mathbf{x}_i$ un conjunto de valores o regresores para $i = 1, 2, \ldots, n$.

Como en el modelo de respuesta binaria, estimaremos el efecto de una variable en la probabilidad de ocurrencia de uno de los valores de respuesta. Dichas variables de respuesta estarán dados por:

$$P(y_i = j|\mathbf{x}_i), \text{ para } j = 1, 2, \ldots, J$$

Dado que la suma de probabilidades debe ser 1, se sule tomar como práctica estimar $P(y_i = 0|\mathbf{x}_i)$ como diferencia del resto de los casos $j = \{1, 2, \ldots, J\}$. Así, sea $\mathbf{x}_i$ un vector de dimensión $1 \times K$. De esta forma, el modelo Logit tendrá como probabilidad de cada una de las respuestas a:

$$P(y_i = j|\mathbf{x}_i) = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_j}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}}$$

Donde cada una de las $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$ es de dimensión $K \times 1$. Es facil observar que si la suma de probabilidades es 1, entonces, la probabilidad de $y = 0|\mathbf{x}_i$ estará dada por:

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= \frac{e^0}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &= \frac{1}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \end{aligned}$$

De esta forma tenemos:

$$\begin{aligned} 1 &= \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i) \\ &= \frac{1}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &\quad + \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_1}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \\ &\quad + \ldots + \\ &\quad + \frac{e^{\mathbf{x}_i\boldsymbol{\beta}_J}}{1 + e^{\mathbf{x}_i\boldsymbol{\beta}_1} + e^{\mathbf{x}_i\boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i\boldsymbol{\beta}_J}} \end{aligned}$$

Para el proceso de estimación consideraremos la función de verosimilitud dada por:

$$\begin{aligned} L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) &= \Pi_{i=1}^{n} \big[(P(j=0|\mathbf{x}_i))^{I(0)} \cdot (P(j=1|\mathbf{x}_i))^{I(1)} \\ &\quad \cdots (P(j=J|\mathbf{x}_i))^{I(J)} \big] \end{aligned}$$

Donde $I(j)$ es la función indicadora que toma el valor de 1 o 0 dependiendo de cual de las opciones posibles $j = 0, 1, 2, \ldots, J$ es cierta para cada individuo $i = 1, 2, \ldots, n$. De forma similar al caso de respuesta binaria en los modelos Logit Multinomiales se interpretan en sus efectos marginales y no de forma directa en los coeficientes $\boldsymbol{\beta}_j$.

Para una variable x_{ik} continua y para cada opción $j = 0, 1, 2, \ldots, J$ el efecto marginal será:

$$\begin{aligned} EMg_k &= \frac{\partial P(y_i = j|\mathbf{x}_i)}{\partial x_{ik}} \\ &= P(y_i = j|\mathbf{x}_i) \cdot \left[\beta_{jk} - \frac{\sum_{h=1}^{J} \beta_{hk} \cdot e^{\mathbf{x}_i \boldsymbol{\beta}_h}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}_1} + e^{\mathbf{x}_i \boldsymbol{\beta}_2} + \ldots + e^{\mathbf{x}_i \boldsymbol{\beta}_J}} \right] \end{aligned}$$

Donde β_{hk} es el k-èsimo elemento del vector $\boldsymbol{\beta}_h$. En el caso de que x_{ik} se una variable dicotómica (o categórica) y para cada opción $j = 0, 1, 2, \ldots, J$ el efecto marginal será:

$$EMg_k = P(y_i = j|\mathbf{x}_i, x_{ik} = 1) - P(y_i = j|\mathbf{x}_i, x_{ik} = 0)$$

Modelos de respuesta ordenada: Logit y Probit Ordinal

Sea y_i una variable que representa ordenada que toma los valores de $\{0, 1, 2, \ldots, J\}$. En estos casos en que importa el órden de la respuesta existen dos mecanismos de estimación: el Modelo Probit y el Modelo Logit.

Modelo Probit Ordinal. Este modelo se puede derivar de forma similar al modelo probit de dos respuestas 0 y 1, probit binario. Así, partimos de una variable latente y_i^* definida como:

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i;\ \boldsymbol{\varepsilon}_i \sim N(0, 1)$$

Donde $\boldsymbol{\beta}$ es un vector $(K-1) \times 1$ ya que en este caso asumiremos que no se considera un término constante, por las razones que mostramos más adelante. Sean $\alpha_1, \alpha_2, \ldots, \alpha_J$ un conjunto de umbrales que particionan la probabilidad de la siguiente forma:

$$
\begin{array}{rcl}
y_i = 0 & \text{si} & y_i^* \leq \alpha_1 \\
y_i = 1 & \text{si} & \alpha_1 < y_i^* \leq \alpha_2 \\
 & \vdots & \\
y_i = j & \text{si} & \alpha_j < y_i^* \leq \alpha_{j+1} \\
 & \vdots & \\
y_i = J & \text{si} & \alpha_J < y_i^*
\end{array}
$$

Dado que $\boldsymbol{\varepsilon}_i$ tiene una distribución normal estándar, para casa uno de los casos anteriores podemos establecer:

$$
\begin{array}{rcl}
P(y_i = 0|\mathbf{x}_i) & = & P(y_i^* \leq \alpha_1|\mathbf{x}_i) = P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\
P(y_i = 1|\mathbf{x}_i) & = & P(\alpha_1 < y_i^* \leq \alpha_2|\mathbf{x}_i) = P(\alpha_1 < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_2|\mathbf{x}_i) \\
 & \vdots & \\
P(y_i = j|\mathbf{x}_i) & = & P(\alpha_j < y_i^* \leq \alpha_{j+1}|\mathbf{x}_i) = P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\
 & \vdots & \\
P(y_i = J|\mathbf{x}_i) & = & P(\alpha_J < y_i^*|\mathbf{x}_i) = P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i)
\end{array}
$$

Desarrollando las ecuaciones anteriores tenemos para el caso de $y_i = 0$:

$$
\begin{array}{rcl}
P(y_i = 0|\mathbf{x}_i) & = & P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\
 & = & P(\boldsymbol{\varepsilon}_i \leq \alpha_1 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
 & = & \Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})
\end{array}
$$

En el caso de $y_i = j$, $j = 1, 2, \ldots, J-1$:

$$
\begin{array}{rcl}
P(y_i = j|\mathbf{x}_i) & = & P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\
 & = & P(\alpha_j - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
 & = & P(\boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) - (\boldsymbol{\varepsilon}_i < \alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
 & = & \Phi(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})
\end{array}
$$

Finalmente, para el caso $y_i = J$:

$$\begin{aligned} P(y_i = J|\mathbf{x}_i) &= P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\ &= P(\alpha_J - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\ &= 1 - P(\boldsymbol{\varepsilon}_i < \alpha_J - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= 1 - \Phi(\alpha_J - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

Es sencillo observar y analizar que:

$$1 = \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i)$$

Notemos que en este modelo los cambios en la probabilidad no serán determinados por $\mathbf{x}_i$, ya que lo que determina el cambio es el umbral α_j, $j = 0, 1, \ldots, J$. También en este caso los coeficientes no se interpretan de forma directa, en cambio se utiliza el efecto marginal que tiene expresiones similares al del probit bivariado como más adelante mostraremos.

Modelo Logit Ordinal. Este modelo se puede derivar de forma similar al modelo Probit Ordinal con una variable latente y_i^*, también con $\boldsymbol{\beta}$ es un vector $(K-1) \times 1$ ya que en este caso asumiremos que no se considera un término constante, por las razones que hemos mostrado anteriormente. Igualmente, sean $\alpha_1, \alpha_2, \ldots, \alpha_J$ un conjunto de umbrales que particionan la probabilidad considerando una función logistica $\Lambda(\cdot)$ de la siguiente forma:

$$\boldsymbol{\varepsilon}_i \sim \Lambda(\cdot)$$

$$P(y_i = J|\mathbf{x}_i) = P(\alpha_J < y_i^*|\mathbf{x}_i) = P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i)$$

Desarrollando las ecuaciones de probabilidad de forma similar al Probit Multinomial tenemos para el caso de $y_i = 0$:

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= P(y_i^* \leq \alpha_1|\mathbf{x}_i) \\ &= P(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_1|\mathbf{x}_i) \\ &= P(\boldsymbol{\varepsilon}_i \leq \alpha_1 - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\ &= \Lambda(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

En el caso de $y_i = j$, $j = 1, 2, \ldots, J-1$:

$$\begin{aligned}
P(y_i = j|\mathbf{x}_i) &= P(\alpha_j < y_i^* \leq \alpha_{j+1}|\mathbf{x}_i) \\
&= P(\alpha_j < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \leq \alpha_{j+1}|\mathbf{x}_i) \\
&= P(\alpha_j - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= P(\boldsymbol{\varepsilon}_i \leq \alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) - (\boldsymbol{\varepsilon}_i < \alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= \Lambda(\alpha_{j+1} - \mathbf{x}_i\boldsymbol{\beta}) - \Lambda(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})
\end{aligned}$$

Finalmente, para el caso $y_i = J$:

$$\begin{aligned}
P(y_i = J|\mathbf{x}_i) &= P(\alpha_J < y_i^*|\mathbf{x}_i) \\
&= P(\alpha_J < \mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\
&= P(\alpha_J - \mathbf{x}_i\boldsymbol{\beta} < \boldsymbol{\varepsilon}_i|\mathbf{x}_i) \\
&= 1 - P(\boldsymbol{\varepsilon}_i < \alpha_J - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i) \\
&= 1 - \Lambda(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})
\end{aligned}$$

También, es sencillo observar y analizar que en los Logit Multinomiales:

$$1 = \sum_{j=0}^{J} P(y_i = j|\mathbf{x}_i)$$

Notemos que en este modelo los cambios en la probabilidad no serán determinados por $\mathbf{x}_i$, ya que lo que determina el cambio es el umbral α_j, $j = 0, 1, \ldots, J$. También en este caso los coeficientes no se interpretan de forma directa, en cambio se utiliza el efecto marginal que tiene expresiones similares al del probit bivariado como más adelante mostraremos.

Estimación de modelos Probit y Logit Ordinal. Para la estimación implementaremos un proceso de maximización de de la siguiente función de verosimilitud:

$$L(\alpha, \boldsymbol{\beta}) = \Pi_{i=1}^{n} P(y_i = 0|\mathbf{x}_i)^{I(0)} \cdot P(y_i = 1|\mathbf{x}_i)^{I(1)} \cdots P(y_i = J|\mathbf{x}_i)^{I(J)}$$

Donde $I(j)$ es la función indicadora que toma el valor de 0 y 1 si el individuo i-ésimo tiene como respuesta alguno de los posibles valores de $j = 0, 1, \ldots, J$. La función es válida sin importar la función $\Phi(\cdot)$ o $\Lambda(\cdot)$ que utilicemos para determinar la $P(y_i = j|\mathbf{x}_i)$, para $j = 1, 2, \ldots, J$. Por lo cual simbolizaremos indistintamente estas funciones como $G(\cdot)$.

Figura 1.14: División del conjunto de datos, retomado de Hastie, Tibshirani, y Friedman (2017, p. 222) [HTF17]

Finalmente, sin importar el modelo que estemos ocupando, Probit o Logit ordinal, la forma de interpretar el modelo es mediante el efecto marginal, cuando x_k sea una variable continua:

$$EMg_k = \frac{\partial}{\partial x_k} P(y_i = j|\mathbf{x}_i) = \frac{\partial}{\partial x_k} G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}) = -\beta_k \cdot g(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})$$

Por su parte, cuando x_j sea una variable dicotómica:

$$\begin{aligned} EMg_k &= P(y_i = j|\mathbf{x}_i, x_{ik} = 1) - P(y_i = j|\mathbf{x}_i, x_{ik} = 0) \\ &= G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i, x_{ik} = 1) - G(\alpha_j - \mathbf{x}_i\boldsymbol{\beta}|\mathbf{x}_i, x_{ik} = 0) \\ &= G(\alpha_j - (\beta_1 + x_2\beta_2 + \ldots + \beta_k + \ldots + x_K\beta_K)) \\ &\quad -G(\alpha_j - (\beta_1 + x_2\beta_2 + \ldots + 0 + \ldots + x_K\beta_K)) \end{aligned}$$

También es posible hacer un efecto marginal intra respuestas, es decir, para cambios entre una respuesta j y una $j + h$, la cual será:

$$EMg_{j-h} = \beta_k \cdot [g(\alpha_h - \mathbf{x}_i\boldsymbol{\beta}) - g(\alpha_j - \mathbf{x}_i\boldsymbol{\beta})], \text{ con } h < j$$

Donde $k = 1, 2, \ldots, K - 1$ y $j = 0, 1, \ldots, J$.

Ajuste y separación del conjunto de datos

De forma similar al caso de regresión, dividiremos al conjunto de datos conforme se describe en la Figura 1.14.

Evaluando los modelos de aprendizaje supervisado

Matriz de confusión

Una vez que se entrena un problema de aprendizaje automatizado supervisado en un conjunto de datos históricos, se prueba el modelo obtenido mediante el uso de datos del conjunto de entrenamiento. De esta forma, es posible comparar las predicciones del modelo entrenado con los valores reales de la variable sujeta de análisis. La matriz de confusión proporciona un medio para evaluar el éxito de un problema de clasificación y dónde se cometen errores (es decir, dónde se vuelve 'confuso').

En el Cuadro 1.1 se muestra un ejemplo de la forma en que se suele mostrar una matriz de confusión.

		Predicciones	
		Positivas (1)	Negativas (0)
Real	Positivas (1)	True Positive (TP)	False Negative (FN)
	Negativas (0)	False Positive (FP)	True Negative (TN)

Cuadro 1.1: Matriz de Confusión

Con base en la matriz de confusión en el Cuadro 1.1, se pueden construir métricas que se calculan de la siguiente manera:

$$Precision = \frac{TP}{TP + FP} \tag{1.45}$$

$$Recuperacion = \frac{TP}{TP + FN} \tag{1.46}$$

$$F1 = 2 \times \frac{Precision \times Recuperacion}{Precision + Recuperacion} \tag{1.47}$$

La medida $F1$ o $F1 - score$ combina precisión y recuperación mediante la media armónica de los dos valores. De esta forma, siempre se ubicará entre ambos indicadores.

1.3.4. 4.2.2 Regresión logística para el estudio de equidad de violencia.

Ver "05_Ejemplo_Logit_Delitos" en GitHub.

1.3.5. 4.2.1 Regresión logística con información de videojuegos.

Planteamiento del caso

Supuestos y necesidades del caso:

1. Determinar cuáles son las features/características/steam tags que se asocian con videojuegos para PC existosos.

2. Hay varias tiendas de videojuegos (Steam, itch, etc.) que utilizan estas tags para clasificar a sus juegos.

Propuesta: Usar un modelo de respuesta ordenada

Sea y una variable que representa una respuesta ordenada que toma valores $\{0, 1, 2, \cdots, J\} \in \mathbb{Z}$. El modelo de repuesta ordenada (conocido como Probit o Logit ordenado) para y (condicional en un vector de variables expicativas $\mathbf{X}$–features, características, tags, etc.–) es derivado a partir de un modelo de varible latente y^*, el cual se puede escribir como:

$$y^* = \mathbf{X}\boldsymbol{\beta} + \varepsilon,\ \varepsilon|\mathbf{X} \sim f(\varepsilon) \tag{1.48}$$

Donde $f(\cdot)$ es una función de densidad de probabilidad simétrica cuya función de densidad acumuada será denotada por $F(\cdot)$, $\boldsymbol{\beta}$ es un vector de dimensión $K \times 1$ y, por razones que más adelante explicamos, $\mathbf{X}$ no tiene término constante. Por otro lado, sean $\mu_1 < \mu_2 < \cdots < \mu_J$ puntos de corte o paramétros umbrales que definen lo siguiente:

$$\begin{array}{rcl} y = 0 & \text{si} & y^* \leq \mu_1 \\ y = 1 & \text{si} & \mu_1 < y^* \leq \mu_2 \\ & \vdots & \\ y = J & \text{si} & y^* > \mu_J \end{array} \tag{1.49}$$

De esta forma y considerando el conjunto de desigualdades en la ecuación (1.49) y la ecuación (1.48), podemos determinar las siguientes probabilidades a partir de una partición de una función de distribución dada:

$$\begin{aligned} \mathbf{P}(y = 0|\mathbf{X}) &= \mathbf{P}(y^* \leq \mu_1|\mathbf{X}) \\ &= \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_1|\mathbf{X}) \\ &= \mathbf{P}(\varepsilon \leq \mu_1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\ &= F(\mu_1 - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$
\begin{aligned}
\mathbf{P}(y=1|\mathbf{X}) &= \mathbf{P}(\mu_1 < y^* \leq \mu_2|\mathbf{X}) \\
&= \mathbf{P}(\mu_1 < \mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_2|\mathbf{X}) \\
&= \mathbf{P}(\mu_1 - \mathbf{X}\boldsymbol{\beta} < \varepsilon \leq \mu_2 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\
&= F(\mu_2 - \mathbf{X}\boldsymbol{\beta}) - F(\mu_1 - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

$$\vdots$$

$$
\begin{aligned}
\mathbf{P}(y=J-1|\mathbf{X}) &= \mathbf{P}(\mu_{J-1} < y^* \leq \mu_J|\mathbf{X}) \\
&= \mathbf{P}(\mu_{J-1} < \mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_J|\mathbf{X}) \\
&= \mathbf{P}(\mu_{J-1} - \mathbf{X}\boldsymbol{\beta} < \varepsilon \leq \mu_J - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\
&= F(\mu_J - \mathbf{X}\boldsymbol{\beta}) - F(\mu_{J-1} - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{P}(y=J|\mathbf{X}) &= \mathbf{P}(y^* > \mu_J|\mathbf{X}) \\
&= \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \varepsilon > \mu_J|\mathbf{X}) \\
&= \mathbf{P}(\varepsilon > \mu_J - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\
&= 1 - F(\mu_J - \mathbf{X}\boldsymbol{\beta})
\end{aligned}
$$

Note que en este caso asumimos que $\mathbf{X}$ no tiene término constante, ya que la incorporamos mediante los umbrales μ_j, $j = 1, 2, \vdots, J$. Un caso particular de esta modelación son los modelos de clasificación binarios en los que $J = 1$. En la figura 1.15 ilustramos la partición de la función de densidad $f(\cdot)$ que resultaría de las ecuaciiones anteriore.

De esta manera, los resutados de la estimación se pueden interpretar a través de sus efectos marginales dados por:

$$\frac{\partial \mathbf{P}(y=j|\mathbf{X})}{\partial x_k} = -\beta_k(f(\mu_j - \mathbf{X}\boldsymbol{\beta}) - f(\mu_{j-1} - \mathbf{X}\boldsymbol{\beta})) \text{ para } 0 < j < J \quad (1.50)$$

Así, este efecto se interpreta como la contribución que tiene la variable x_k a la probabilidad de que la variable y tome el valor de j

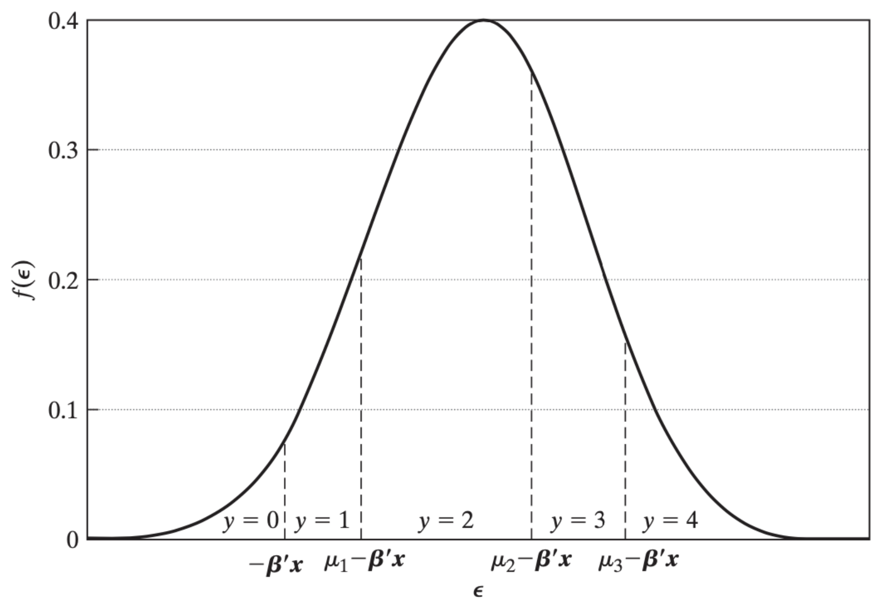


Figura 1.15: Probabilidades en un modelo de respuesta ordenada, considerando $J = 4$ (retomado de Greene (2012, 788) [Gre12])

Aplicación al caso de videojuegos

En el caso de los videojuegos, podemos aplicar el modelo de respuesta ordenada asumiendo algunas cosas. Primero, F tendrá una forma de función logística, para facilitar el proceso computacional. Segundo, la variable y será particionada en rangos de ingresos obtenidos por los videojuegos en su primer año (y^*)–año de lanzamiento–, por ejemplo:

$$y^* = \mathbf{X}\boldsymbol{\beta} + \varepsilon,\ \varepsilon|\mathbf{X} \sim f(\varepsilon)$$

Where y^* is the revenue per year, $\mathbf{X}$ are are the characteristics and other variables of video games, $\boldsymbol{\beta}$ are the coefficients and ε is the error term.

$$\begin{array}{lll} y=0 & \text{if} & y^* \leq 1\text{M USD} \\ y=1 & \text{if} & 1 \text{ MM USD} < y^* \leq \ 5 \text{ MM USD} \\ y=2 & \text{if} & 5 \text{ MM USD} < y^* \leq \ 20 \text{ MM USD} \\ y=3 & \text{if} & y* > \ 20 \text{ MM USD} \end{array}$$

$$\begin{aligned} \mathbf{P}(y=0|\mathbf{X}) &= \mathbf{P}(y^* \leq \mu_1|\mathbf{X}) \\ &= \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_1|\mathbf{X}) \\ &= \mathbf{P}(\varepsilon \leq \mu_1 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\ &= F(\mu_1 - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} \mathbf{P}(y=1|\mathbf{X}) &= \mathbf{P}(\mu_1 < y^* \leq \mu_2|\mathbf{X}) \\ &= \mathbf{P}(\mu_1 < \mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_2|\mathbf{X}) \\ &= \mathbf{P}(\mu_1 - \mathbf{X}\boldsymbol{\beta} < \varepsilon \leq \mu_2 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\ &= F(\mu_2 - \mathbf{X}\boldsymbol{\beta}) - F(\mu_1 - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned} \mathbf{P}(y=2|\mathbf{X}) &= \mathbf{P}(\mu_2 < y^* \leq \mu_3|\mathbf{X}) \\ &= \mathbf{P}(\mu_2 < \mathbf{X}\boldsymbol{\beta} + \varepsilon \leq \mu_3|\mathbf{X}) \\ &= \mathbf{P}(\mu_2 - \mathbf{X}\boldsymbol{\beta} < \varepsilon \leq \mu_3 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\ &= F(\mu_3 - \mathbf{X}\boldsymbol{\beta}) - F(\mu_2 - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

$$\begin{aligned}\mathbf{P}(y=3|\mathbf{X}) &= \mathbf{P}(y^* > \mu_3|\mathbf{X}) \\ &= \mathbf{P}(\mathbf{X}\boldsymbol{\beta} + \varepsilon > \mu_3|\mathbf{X}) \\ &= \mathbf{P}(\varepsilon > \mu_3 - \mathbf{X}\boldsymbol{\beta}|\mathbf{X}) \\ &= 1 - F(\mu_3 - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

$$\mathbf{P}(y=0|\mathbf{X}) = \mu_1 - \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{P}(y=1|\mathbf{X}) = \mu_2 - \mathbf{X}\boldsymbol{\beta}$$

$$\mathbf{P}(y=2|\mathbf{X}) = \mu_3 - \mathbf{X}\boldsymbol{\beta}$$

$$\varepsilon$$

Supongamos que tenemos n observaciones o, en este caso, videojuegos, los cuales son indexados con $i = 1, 2, 3, \ldots, n$. Así, indexaremos a la variable de respuesta y como y_i, por lo que diremos que buscamos estimar las siguientes probabilidades:

$$\begin{aligned}\mathbf{P}(\text{Revenue}_i \leq \text{1M USD}|\mathbf{X}_i) &= \mathbf{P}(\varepsilon_i \leq \mu_1 - \mathbf{X}_i\boldsymbol{\beta}|\mathbf{X}_i) \\ &= F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})\end{aligned}$$

$$\begin{aligned}\mathbf{P}(\text{ 1 MM USD} < \text{Revenue}_i \leq \text{ 5 MM USD}|\mathbf{X}_i) &= \mathbf{P}(\mu_1 - \mathbf{X}_i\boldsymbol{\beta} < \varepsilon_i \leq \mu_2 - \mathbf{X}_i\boldsymbol{\beta}|\mathbf{X}_i) \\ &= F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})\end{aligned}$$

$$\begin{aligned}\mathbf{P}(\text{ 5 MM USD} < \text{Revenue}_i \leq \text{ 20 MM USD}|\mathbf{X}_i) &= \mathbf{P}(\mu_2 - \mathbf{X}_i\boldsymbol{\beta} < \varepsilon_i \leq \mu_3 - \mathbf{X}_i\boldsymbol{\beta}|\mathbf{X}_i) \\ &= F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta})\end{aligned}$$

$$
\begin{aligned}
\mathbf{P}(\text{Revenue}_i > \ 20 \text{ MM USD}|\mathbf{X}_i) &= \mathbf{P}(\varepsilon_i > \mu_3 - \mathbf{X}_i\boldsymbol{\beta}|\mathbf{X}_i) \\
&= 1 - F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta})
\end{aligned}
$$

Donde, particularmente, F es una función logistica. Nuestro objetivo es máximizar la función de verosimilitud (función que maximiza el valor de los parámetros μ_1, μ_2, μ_3 y $\boldsymbol{\beta}$ dada la información disponible):

$$
\begin{aligned}
\mathbf{L}(\boldsymbol{\theta}|\text{Data}) &= \prod_{y_i=0} \mathbf{P}(\text{Revenue}_i \leq 1\text{M USD}|\mathbf{X}_i) \\
&\quad \times \prod_{y_i=1} \mathbf{P}(\ 1 \text{ MM USD} < \text{Revenue}_i \leq \ 5 \text{ MM USD}|\mathbf{X}_i) \\
&\quad \times \prod_{y_i=2} \mathbf{P}(\ 5 \text{ MM USD} < \text{Revenue}_i \leq \ 20 \text{ MM USD}|\mathbf{X}_i) \\
&\quad \times \prod_{y_i=3} \mathbf{P}(\text{Revenue}_i > \ 20 \text{ MM USD}|\mathbf{X}_i) \\
&= \prod_{y_i=0} (F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})) \times \prod_{y_i=1} (F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})) \\
&\quad \times \prod_{y_i=2} (F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta})) \times \prod_{y_i=3} (1 - F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta}))
\end{aligned}
$$

where:

$$
\theta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \boldsymbol{\beta} \end{bmatrix}
$$

Así, el problema consiste en resolver:

$$
\begin{aligned}
\max_{\theta} \mathbf{L}(\boldsymbol{\theta}|\text{Data}) &= \prod_{\text{Revenue}_i \leq 1\text{M USD}} (F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})) \\
&\quad \times \prod_{1 \text{ MM USD} < \text{Revenue}_i \leq \ 5 \text{ MM USD}|\mathbf{X}_i} (F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_1 - \mathbf{X}_i\boldsymbol{\beta})) \\
&\quad \times \prod_{5 \text{ MM USD} < \text{Revenue}_i \leq \ 20 \text{ MM USD}} (F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta}) - F(\mu_2 - \mathbf{X}_i\boldsymbol{\beta})) \\
&\quad \times \prod_{\text{Revenue}_i > \ 20 \text{ MM USD}} (1 - F(\mu_3 - \mathbf{X}_i\boldsymbol{\beta}))
\end{aligned}
$$

Tags	Juegos	Tags	Juegos	Tags	Juegos	Tags	Juegos	Tags	Juegos
Indie	36,782	Combat	3,866	Multiple Endings	2,065	Historical	1,313	Music	910
Action	25,650	Female Protagonist	3,857	Roguelike	2,054	Walking Simulator	1,312	Competitive	882
Adventure	23,865	Violent	3,796	Puzzle Platformer	2,047	Resource Management	1,293	Dark Humor	881
Casual	23,618	0	3,677	War	1,994	Interactive Fiction	1,292	Aliens	880
Singleplayer	23,397	Open World	3,604	Shoot 'Em Up	1,933	Memes	1,252	Clicker	872
2D	12,837	Sexual Content	3,598	Linear	1,849	Dungeon Crawler	1,240	Nature	866
Strategy	11,854	Visual Novel	3,502	Character Customization	1,813	Score Attack	1,229	Driving	862
Simulation	11,569	Comedy	3,398	Hand-drawn	1,804	Dating Sim	1,225	Emotional	862
RPG	11,097	Turn-Based	3,359	Old School	1,794	Surreal	1,212	Flight	857
Puzzle	9,354	Action-Adventure	3,342	Massively Multiplayer	1,738	Top-Down Shooter	1,212	Arena Shooter	846
Multiplayer	8,185	Co-op	3,295	Procedural Generation	1,738	Experimental	1,204	Detective	846
Atmospheric	8,117	Top-Down	3,289	Action RPG	1,735	Post-apocalyptic	1,201	Party	845
Early Access	7,604	Management	3,287	Turn-Based Strategy	1,734	Mod	1,181	1980s	837
Platformer	7,103	Third Person	3,240	Replay Value	1,706	RTS	1,167	Perma Death	836
Story Rich	6,711	Cartoon	3,210	Classic	1,672	Isometric	1,151	Immersive Sim	824
3D	6,584	Gore	3,205	Local Co-Op	1,651	Stealth	1,146	Investigation	820
Shooter	6,181	FPS	3,100	Online Co-Op	1,641	Dark Fantasy	1,129	Precision Platformer	815
Pixel Graphics	6,174	Sports	2,943	Survival Horror	1,598	Education	1,127	Economy	809
Fantasy	5,910	Mystery	2,704	Hidden Object	1,597	Narration	1,125	Beat 'em up	805
Colorful	5,629	Point & Click	2,671	Logic	1,586	1990's	1,112	NSFW	781
Arcade	5,564	Psychological	2,660	Mature	1,582	Drama	1,104	Time Management	782
Horror	5,504	Stylized	2,659	Magic	1,575	Text-Based	1,095	Beautiful	767
Anime	5,424	Space	2,611	Roguelite	1,558	Base Building	1,089	Cinematic	728
Cute	5,417	Minimalist	2,591	3D Platformer	1,519	Card Game	1,077	Metroidvania	727
First-Person	5,351	Physics	2,573	Futuristic	1,518	Abstract	1,072	Team-Based	726
Funny	5,160	PvP	2,494	Fast-Paced	1,480	Hentai	1,051	City Builder	718
Soundtrack	4,886	Choices Matter	2,464	Turn-Based Combat	1,467	Fighting	1,033	Real Time Tactics	702
VR	4,862	Building	2,456	PvE	1,466	Third-Person Shooter	1,029	Strategy RPG	696
Exploration	4,806	Sandbox	2,415	Crafting	1,451	4 Player Local	1,005	Loot	677
Retro	4,734	Cartoony	2,389	Bullet Hell	1,447	Tower Defense	997	Psychedelic	672
Sci-fi	4,601	2D Platformer	2,378	Short	1,428	2.5D	992	Character Action Game	669
Great Soundtrack	4,580	Psychological Horror	2,323	Medieval	1,424	Cyberpunk	986	Twin Stick Shooter	662
Free to Play	4,473	Tactical	2,317	JRPG	1,396	Military	979	Conversation	657
Difficult	4,445	Side Scroller	2,243	Romance	1,373	Robots	963	Tabletop	653
Survival	4,410	Realistic	2,220	Hack and Slash	1,365	Action Roguelike	942	Level Editor	651
Family Friendly	4,231	Racing	2,198	Choose Your Own Adventure	1,325	Immersive	934	Runner	647
Dark	3,940	Controller	2,148	Turn-Based Tactics	1,325	Board Game	916	Life Sim	645
Relaxing	3,924	Local Multiplayer	2,136	RPGMaker	1,317	LGBTQ+	913	Wargame	636

Tags	Juegos
Demons	626
Destruction	616
Tactical RPG	616
Party-Based RPG	594
Real-Time	589
2D Fighter	583
Automobile Sim	582
Match 3	581
Rhythm	571
Moddable	565
Nonlinear	553
Modern	551
Tutorial	543
Dystopian	541
Parkour	539
Comic Book	534
Thriller	534
Alternate History	515
Artificial Intelligence	496
CRPG	490
Grand Strategy	489
Crime	487
Grid-Based Movement	484
Space Sim	483
Lore-Rich	471
Mouse only	468
Supernatural	465
MMORPG	454
Addictive	451
Souls-like	448
Trading	448
Split Screen	436
Dark Comedy	431
Open World Survival Craft	429
Science	426
Mythology	425
Inventory Management	413
Philosophical	413

There is a large quantity of tags found in the database, thus it makes complicated to classify a game into a single category. However, tag's value lies in the fact that users are the ones who tag a game. It tells us how a user perceives a game.

There are 417 tags in the database with a particular category for each one in which a game can be.

Figura 1.16: En la tabla de datos ubicamos 417 tags distintas

Datos

Usamos como fuente: `https://games-stats.com/`, una descarga de datos de 2021. La tabla de datos originalmente descargada contenía información de 53,650 videojuegos para computadora de los últimos 40 años, acumulando 18.5 billones de dólares.

La tabla de datos de games-stats.com se utilizó como base de datos central debido a su tamaño y variables disponibles. Fue validado con otros datos encontrados en línea, y tenían un tamaño similar entre ellos. Se agregaron más variables con el propósito de enriquecer el análisis.

Hay una gran cantidad de etiquetas (tags) encontradas en la tabla de datos, por lo que resulta complicado clasificar un juego en una sola categoría. Sin embargo, el valor de la etiqueta radica en el hecho de que son los usuarios quienes etiquetan un juego. Esto nos dice cómo un usuario percibe un juego. En total ubicamos 417 tags distintas.

En este caso, requerimos una reducción de dimensionalidad. El modelo debía evaluar más de 450 etiquetas diferentes, algo inviable. Tener 450 etiquetas para evaluar reduce la simplicidad e interpretabilidad del modelo. Así, aplicamos un par de metodologías:

- Extreme Random Decision Tree method: Que es una metodología de Machine Learning que ayudará a consider menos variables dada la relevancia para los ingresos.

- Spearman Correlation: que muestra la correlación con los ingresos de determinadas variables y muestra también la significación estadística basada en una prueba t. Así, la etiqueta de un jugador tiene la correlación significativa más alta con los ingresos.

El código del ejemplo se encuentra en GitHub con el nombre:
"06_Ejemplo_Logit_Ordenado" en GitHub.

1.4. 4.3 Aplicación de modelos clasificación mediante clustering.

1.4.1. 4.3.1 Método de Componentes Principales con datos de seguridad pública.

1.4.2. 4.3.2 Método de K-means con datos de epidemiológicos.

1.4.3. 4.3.3. Análisis de paridad y violencia de género.

1.5. 4.4 Aplicación de modelos de análisis de imágenes y texto.

1.5.1. 4.4.1 Clasificación de imágenes de dígitos escritos a mano.

1.5.2. 4.4.3 Minado de texto con perspectiva de género.

Bibliografía

[Aba21] Alberto Abadie. «Using synthetic controls: Feasibility, data requirements, and metodological acpects». En: *Journal of Econometric Literature* (2021).

[Ada20] Christopher P. Adams. *Learning Microeconometrics with R*. Estados Unidos: CRC Press, 2020.

[CT05] Colin Cameron y Pravin K. Trivedi. *Microeconometrics: Methods and Applications*. Estados Unidos: Cambridge University Press, 2005.

[Cun21] Scott Cunningham. *Causal inference: The mixtape*. Estados Unidos: Yale University Press, 2021.

[Gre12] William Greene. *Econometric Analysis*. Estados Unidos: Prentice Hall, 2012.

[HTF17] Trevor Hastie, Robert Tibshirani y Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. LLC: Springer, 2017.

[Jam+13] Gareth James et al. *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013.

[LM12] Richard J. Larsen y Morris L. Marx. *An Introduction to Mathematical Statistics and its Applications*. Estados Unidos: Prentice Hall, 2012.

[Woo10] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Estados Unidos: The MIT Pres, 2010.