Microeconometría I

Maestría en Econometría

Lecture 1

Variables Dependientes Limitadas y Cualitativas

Agenda

Distintos Tipos de Variables Discretas

- En esta primera parte del curso vamos a estudiar los métodos econométricos que existen para analizar variables dependientes que son cualitativas o asumen valores discretos.
- Ejemplos de estas variables son participación laboral, la demanda de bienes durables e inmuebles, elección entre diferentes marcas del mismo producto, grado de preferencia por determinado bien, número de patentes emitidas, número de empleados en diferentes firmas de un mismo sector, etc.
- La característica de todas estas variables es que asumen valores discretos.
- Estos valores pueden representar diferentes **categorías**, como es el caso de la elección entre diferentes marcas de un mismo producto, ó los valores no representan categorías pero son discretos, como el caso de las patentes emitidas.

Distintos Tipos de Variables Discretas

- Las variables discretas que adoptan valores que representan categorías se denominan variables categóricas.
- La variable categórica más simple es la variable binaria que adopta solamente dos categorías.
- Las variables categóricas con más de dos categorias pueden clasificarse en (a) ordenadas y (b) no ordenadas.

Agenda

Clasificación de las Variables Categóricas

- Un ejemplo de variable categórica no ordenada puede ser:

  $Y = 1$ si la persona viaja a su trabajo en *tren*
  $Y = 2$ si la persona viaja a su trabajo en *colectivo*
  $Y = 3$ si la persona viaja a su trabajo en *taxi*
  $Y = 4$ si la persona viaja a su trabajo en *auto*
  $Y = 5$ si la persona viaja a su trabajo en otro medio de transporte.

- En este caso, la variable categórica es no ordenada porque los valores que adopta la variable representan categorías sin ningún orden pre-establecido.

Clasificación de las Variables Categóricas

- Un ejemplo de variable categórica ordenada puede ser:

  $Y = 1$ si la calificación (rating) de la deuda del país es AAA
  $Y = 2$ si la calificación de la deuda del país es AA
  $Y = 3$ si la calificación de la deuda del país es A
  $Y = 4$ si la calificación de la deuda del país es menor a A

- En este caso, la variable categórica adopta valores que representan categorías con un orden pre-establecido. Nosotros sabemos que un país que tiene una categoría igual a 1 es porque tiene una mejor calificación de su deuda que un país que tiene una categoría igual a 2.

Agenda

Modelo de Probabilidad Lineal

- Nuestro objetivo es explicar el comportamiento de alguna de estas variables utilizando un modelo de regresión lineal.
- Y las primeras preguntas que vamos a tener que responder son:
  - Es posible estimar este modelo usando el método de mínimos cuadrados clásicos?.
  - Si no es posible, cuál es el mejor método de estimación alternativo?.
- Para analizar qué es lo que sucede cuando queremos estimar un modelo de variable dependiente categórica por mínimos cuadrados clásicos vamos a comenzar con la más simple de estas variables.
- El ejemplo más sencillo de variable categórica es el de una variable binaria, es decir, una variable que solo tiene dos categorías y que por convención adopta dos valores que se denotan con 0 y 1.

Modelo de Probabilidad Lineal

- Como ilustración considere la decisión de un trabajador entre trabajar y no trabajar.
- Entre las variables que influyen en esta decisión están: salario ofrecido, cantidad de hijos menores a 6 años en el hogar, si el cónyuge está desocupado, si se recibe algún tipo de ayuda social, etc.
- Las variables explicativas de la decisión de trabajar o no las vamos a agrupar en el vector x.

Modelo de Probabilidad Lineal

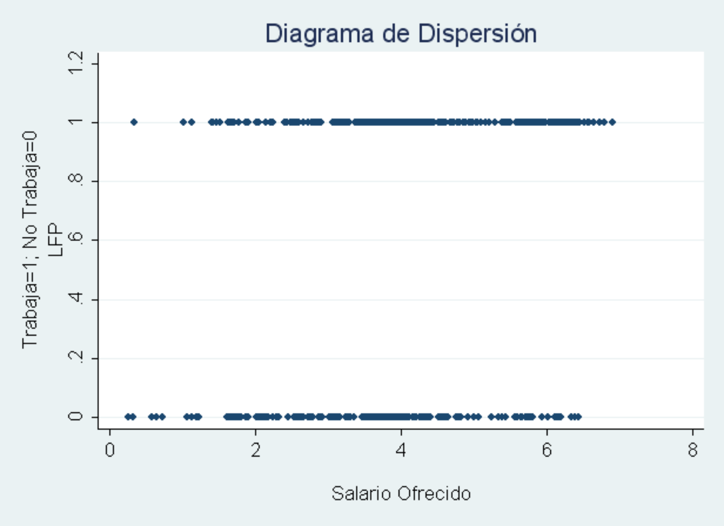
- Supongamos, para mantener el modelo simple, que x contiene una sola variable explicativa (el salario ofrecido) y definamos la siguiente ecuación de regresión:

$$y_i = \alpha + \beta\, x_i + u_i \tag{1}$$

donde $y_i = 1$ si el trabajador i decide trabajar e $y_i = 0$ en otro caso.

- La teoría económica diría que a mayor salario ofrecido, el trabajador i debería tener más incentivos a trabajar.
- Como primera aproximación al tema veamos un diagrama de dispersión de las variables.

Modelo de Probabilidad Lineal



Modelo de Probabilidad Lineal

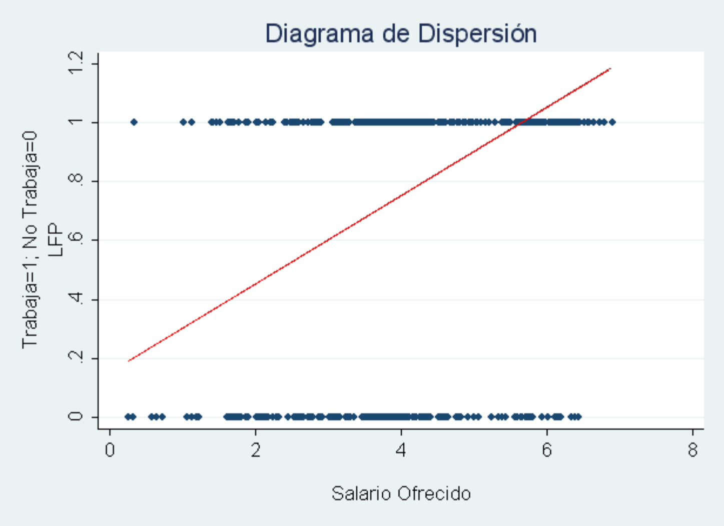
- La estimación por mínimos cuadrados clásicos del modelo es,

$$\begin{aligned} \hat{y}_i \quad = \quad & \underset{(0.063)}{0.1858} + \underset{(0.015)}{0.0968}\, x_i \end{aligned}$$

donde los números entre paréntesis son los desvíos estándar.

- Es decir, existe una relación directa entre el salario ofrecido y la decisión de trabajar. A medida que aumenta el salario ofrecido hay más incentivos a trabajar. Gráficamente:

Modelo de Probabilidad Lineal



Modelo de Probabilidad Lineal

- Como puede observarse en el gráfico anterior, un problema con esta estimación es que hay muchos valores ajustados para la variable dependiente que son MAYORES A UNO!
- Esta es una consecuencia del método de estimación elegido, ya que MCC no contiene ninguna restricción que nos diga que la estimación de la variable dependiente siempre tiene que ser cero o uno.
- El problema anterior puede plantearse de la siguiente manera. Cuál es la probabilidad de que una persona trabaje, dado el valor del salario ofrecido?
- Evidentemente, es poco probable que alguien pueda responder a esa pregunta. Sin embargo, podemos observar el valor de esta probabilidad ex-post.

Modelo de Probabilidad Lineal

- Esto es lo que observamos en la muestra. Es decir, un individuo, ex-post o trabaja (probabilidad de realización igual a uno) o no trabaja (probabilidad de realización igual a cero).
- Esto significa que lo que observamos es la realización de una variable latente (la probabilidad de trabajar).
- En realidad la pregunta relevante en este tipo de análisis es: cuáles son las características que afectan la probabilidad de trabajar?
- Es decir que a nosotros nos interesa el valor de la probabilidad ex-ante. Esto es:

$$Pr(y_i = 1 | Salario \ Ofrecido)$$

Modelo de Probabilidad Lineal

- Note que de acuerdo al modelo especificado anteriormente tenemos:

$$y_i = \alpha + \beta \; x_i + u_i,$$

- Por lo tanto,

$$\begin{aligned} E(y_i|x_i) &= \alpha + \beta \; x_i + E(u_i|x_i) \\ &= \alpha + \beta \; x_i \end{aligned} \qquad (2)$$

Modelo de Probabilidad Lineal

- Además, por definición de esperanza matemática, la esperanza matemática condicional de una variable es la suma de cada uno de los valores que adopta la variable multiplicados por su probabilidad de ocurrencia.
- En este caso:

$$
\begin{aligned}
E(y_i|x_i) &= 1 \times Pr(y_i = 1|x_i) + 0 \times Pr(y_i = 0|x_i) \\
&= Pr(y_i = 1|x_i) \qquad (3)
\end{aligned}
$$

Modelo de Probabilidad Lineal

- Igualando las ecuaciones (2) y (3) se tiene,

$$Pr(y_i = 1|x_i) = \alpha + \beta\ x_i \tag{4}$$

- Como puede observarse en la ecuación (4) la probabilidad condicional de que el evento y_i ocurra (en este caso la decisión de trabajar) dado que conocemos el valor de la variable independiente, está expresada como una relación lineal en los parámetros del modelo.
- Debido a este hecho, los modelos de variable dependiente binaria reciben el nombre de Modelos de Probabilidad Lineal (MPL).

Modelo de Probabilidad Lineal

- Dado que y_i solo puede adoptar dos valores (cero ó uno) podemos obtener su distribución de probabilidad con la siguiente tabla:

y_i	$Pr(y_i \mid \cdot)$
0	$1 - \alpha - \beta\ x_i$
1	$\alpha + \beta\ x_i$

- Dada la distribución de la variable dependiente podemos utilizar el modelo para obtener la distribución de los errores. Esto es, sabemos que cuando $y_i = 0$ entonces $u_i = -\alpha - \beta\ x_i$ y cuando $y_i = 1$ entonces $u_i = 1 - \alpha - \beta\ x_i$. Por lo tanto,

Modelo de Probabilidad Lineal

u_i	$Pr(u_i \vert \cdot)$
$1-\alpha-\beta\ x_i$	$\alpha+\beta\ x_i$
$-\alpha-\beta\ x_i$	$1-\alpha-\beta\ x_i$

- Calculemos la esperanza matemática y la varianza de los errores del modelo.
-

$$\begin{aligned} E(u_i|\cdot) &= [1-\alpha-\beta\ x_i]\times[\alpha+\beta\ x_i] \\ &+ [-\alpha-\beta\ x_i]\times[1-\alpha-\beta\ x_i] \\ &= 0 \end{aligned}$$

Modelo de Probabilidad Lineal

- $$\begin{aligned} Var(u_i|\cdot) &= E(u_i^2|\cdot) - E(u_i|\cdot)^2 \\ &= E(u_i^2|\cdot) \\ &= [1 - \alpha - \beta\ x_i] \times [\alpha + \beta\ x_i] \end{aligned}$$

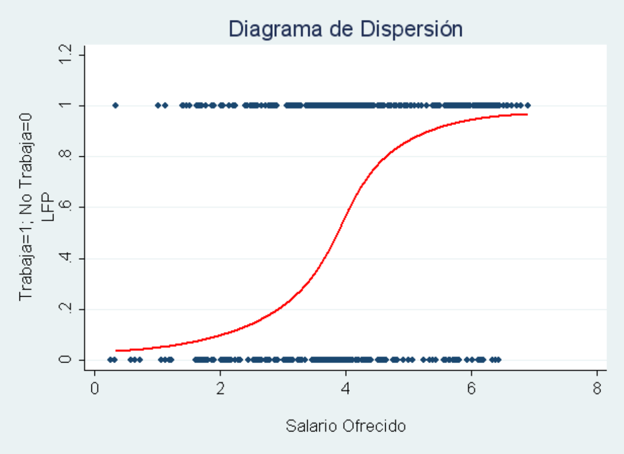
- De la ecuación anterior se desprende que, cuando la variable dependiente es binaria, los errores del modelo tienen heterocedasticidad.
- Esta característica puede generalizarse a cualquier modelo que tenga por variable dependiente una variable categórica.

Modelo de Probabilidad Lineal

- Ahora podemos resumir las características de los MPL:
  - Estos modelos reciben este nombre porque la variable dependiente puede interpretarse como una probabilidad.
  - Los valores estimados, por el método de MCC, de la variable dependiente pueden caer fuera del rango [0, 1].
  - Los errores del modelo son heterocedásticos por lo tanto MCC nos dará estimadores ineficientes.
- Otro de los problemas que sufren los MPL es el de la interpretación de los coeficientes estimados.
- En nuestro caso particular el coeficiente β mide cuánto afecta a la probabilidad de trabajar un cambio en el salario ofrecido.
- Económicamente uno pensaría que este efecto debiera ser chico para salarios pequeños (y grandes) y mayor para salarios intermedios. Es decir, el coeficiente β no debiera ser constante.

Modelo de Probabilidad Lineal

- La forma de resolver estos problemas es utilizando una función de probabilidad no lineal en los parámetros. Gráficamente,



Modelo de Probabilidad Lineal

- Las características de la curva de la figura anterior resuelve nuestros problemas ya que:
  - Empieza en cero y termina en uno. Esto es, solo adopta valores en el intervalo $[0, 1]$.
  - Tiene diferentes pendientes en distintos puntos. Para valores muy pequeños y muy grandes de la variable independiente la pendiente es chica y para valores intermedios la pendiente es más grande.
- Cualquier curva de probabilidad acumulada cumple con las características antes mencionadas. Por lo tanto, uno puede especificar el modelo utilizando la función de probabilidad acumulada de cualquier distribución.
- Las funciones más utilizadas son la distribución Normal y la distribución Logística.

Agenda

Modelo Logit

- El modelo Logit asume que la curva del gráfico anterior puede ser aproximada por la distribución logística.
- En términos matemáticos, una variable z se dice que tiene distribución logística cuando su función de distribución de probabilidad viene dada por:

$$F(z) = \frac{e^z}{1 + e^z}$$

- En economía, la forma tradicional de analizar este tipo de modelos es utilizando el concepto de variable latente.
- Supongamos que los individuos toman su decisión de trabajar o no sobre la base de la utilidad que les da el trabajo.

Modelo Logit

- Asumamos utilidades estocásticas como funciones lineales del salario ofrecido. Entonces:

  $$U_{T,i} = \alpha_T + \beta_T \ x_i + u_{T,i}$$

  es la utilidad para el individuo i de trabajar. Y

  $$U_{D,i} = \alpha_D + \beta_D \ x_i + u_{D,i}$$

  es la utilidad para el individuo i de no trabajar.
- Entonces, el individuo i trabajará si $U_{T,i} > U_{D,i}$.

Modelo Logit

- Por lo tanto, la probabilidad de trabajar para el individuo i puede expresarse como,

$$
\begin{aligned}
Pr[y_i = 1|\cdot] &= Pr[U_{T,i} > U_{D,i}|\cdot] \\
&= Pr[(\alpha_T - \alpha_D) + (\beta_T - \beta_D)\, x_i > u_{D,i} - u_{T,i}|\cdot] \\
&= Pr[u_i \leq \alpha + \beta\, x_i|\cdot]
\end{aligned}
$$

donde $u_i = u_{D,i} - u_{T,i}$, $\alpha = \alpha_T - \alpha_D$ y $\beta = \beta_T - \beta_D$.

- Ahora tenemos expresado el modelo en términos de una probabilidad acumulada hasta $\alpha + \beta\, x_i$ para la variable aleatoria u_i.

Modelo Logit

- Entonces, si u_i tiene distribución logística tenemos:

$$F(\alpha + \beta x_i) = \frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}}$$

- A diferencia de lo que ocurría con el MPL en este caso existe una relación no lineal en los parámetros del modelo y por lo tanto no puede usarse el método de mínimos cuadrados para la estimación.
- Una consecuencia del último punto es que los efectos de las variables explicativas sobre la variable dependiente no son lineales.

Modelo Logit: Interpretación de los Coeficientes

- En la función logística $\beta\, x_i$ está en el exponente del número e en el numerador y en el denominador y por lo tanto no es inmediatamente claro cuál es el efecto sobre la probabilidad de trabajar de un cambio en el salario ofrecido.
- Este efecto no lineal puede verse calculando la derivada parcial de la probabilidad de trabajar con respecto al salario ofrecido,

$$\begin{aligned}\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} &= F(\alpha + \beta x_i) \times [1 - F(\alpha + \beta x_i)] \times \beta \\ &= f(\alpha + \beta\, x_i) \times \beta\end{aligned}$$

Modelo Logit: Interpretación de los Coeficientes

- Esto muestra que el efecto, sobre la probabilidad de trabajar, de un cambio en el salario ofrecido depende no solo del valor de β sino también del valor tomado por la función de densidad de la logística.
- Recuerde que en el caso del MPL esta derivada era constante e igual a β. Esto nos permitía interpretar los coeficientes del modelo como el cambio marginal que se produce en la variable dependiente cuando cambia una de las variables independientes, manteniendo constante el resto de las variables.
- Ahora, la derivada refleja las diferentes pendientes de la curva logística. Esto significa que hay un valor para el cambio marginal de la probabilidad de trabajar para cada valor del salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- Una forma de interpretar los coeficientes del modelo Logit es calculando los efectos marginales promedio sobre la muestra de n observaciones. Hay dos formas de hacer esto,
  - evaluando la función de densidad logística en la media de las variables,

$$\overline{\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i}} = f(\alpha + \beta\,\overline{x}) \times \beta$$

  - calculando el promedio de la función de densidad logística,

$$\overline{\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i}} = \frac{1}{n}\sum_{i=1}^{n} f(\alpha + \beta\,x_i) \times \beta$$

Modelo Logit: Interpretación de los Coeficientes

- Una segunda forma de interpretar los coeficientes del modelo Logit es a través del cociente de probabilidades ("odds ratio").
- Las chances ("tasa de probabilidad") de que un evento suceda se calculan como el cociente entre la probabilidad de que un evento suceda y la probabilidad de que no suceda. Por ejemplo, en el modelo Logit, las chances de trabajar $(y_i = 1)$ son,

$$\frac{Pr(y_i = 1|x_i)}{Pr(y_i = 0|x_i)} = \frac{F(\alpha + \beta\, x_i)}{1 - F(\alpha + \beta\, x_i)} = e^{\alpha + \beta\, x_i}$$

- Fijando el salario ofrecido (x_i) en algún valor, se obtienen las chances de trabajar a ese salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- Cuando x_i es una variable continua, como en nuestro caso, el cociente de probabilidades (CP) mide el cambio en las chances de trabajar de aumentar en una unidad el salario ofrecido:

$$CP = \frac{Pr(y_i = 1|x_i = x+1)/Pr(y_i = 0|x_i = x+1)}{Pr(y_i = 1|x_i = x)/Pr(y_i = 0|x_i = x)} = e^{\beta}$$

- El cociente de probabilidades nos dice que ante un aumento de un peso en el salario ofrecido esperamos ver alrededor de $(e^{\beta} - 1)\%$ de cambio en las chances de trabajar. Este valor del cambio porcentual no depende del valor que adopte el salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- El cociente de probabilidades también se utiliza para calcular las chances de que el evento analizado suceda comparando dos grupos.
- Supongamos que $x_{i,1} = 1$ denota que el individuo i es de género femenino y $x_{i,1} = 0$ denota que i es de género masculino. Entonces, el cociente de probabilidades se define como,

$$CP = \frac{Pr(y_i = 1|x_{i,1} = 1)/Pr(y_i = 0|x_{i,1} = 1)}{Pr(y_i = 1|x_{i,1} = 0)/Pr(y_i = 0|x_{i,1} = 0)} = e^{\beta_1}$$

- El cociente de probabilidades nos dice cuantas más (menos) chances hay de que una mujer trabaje comparada con un hombre (con las mismas características).

Agenda

Modelo Logit: Estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
- Entonces, en términos matemáticos la función de verosimilitud es,

$$L(\alpha, \beta; x_i) = \prod_{i=1}^{n} \left[\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right]^{y_i} \times \left[1 - \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right]^{1-y_i}$$

Modelo Logit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\alpha,\beta;x_i)=\sum_{i=1}^{n}\left[y_i ln\{\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\}+(1-y_i)ln\{1-\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\}\right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$\frac{\partial l(\cdot)}{\partial\hat{\alpha}} = \sum_{i=1}^{n}\left[y_i-\frac{e^{\hat{\alpha}+\hat{\beta}x_i}}{1+e^{\hat{\alpha}+\hat{\beta}x_i}}\right]=0$$

$$\frac{\partial l(\cdot)}{\partial\hat{\beta}} = \sum_{i=1}^{n}\left[y_i-\frac{e^{\hat{\alpha}+\hat{\beta}x_i}}{1+e^{\hat{\alpha}+\hat{\beta}x_i}}\right]x_i=0$$

Modelo Logit: Estimación

- Como se puede observar en las condiciones de primer grado las incógnitas de ambas ecuaciones $(\hat{\alpha}, \hat{\beta})$ entran en forma no lineal y por lo tanto no pueden resolverse por métodos lineales.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Logit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen.
- Las condiciones de segundo orden vienen dadas por las siguientes expresiones,

$$\frac{\partial^2 l(\cdot)}{\partial \hat{\alpha}^2} = -\sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) \times \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right)$$

- $$\begin{aligned}\frac{\partial^2 l(\cdot)}{\partial\hat{\alpha}\partial\hat{\beta}} &= -\sum_{i=1}^{n}\left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)\times\left(1-\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)x_i \\ \frac{\partial^2 l(\cdot)}{\partial\hat{\beta}^2} &= -\sum_{i=1}^{n}\left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)\times\left(1-\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)x_i^2\end{aligned}$$
- Llamando P_i a $\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$ para abreviar notación, estas condiciones de segundo orden pueden agruparse en la matriz Hesiana (la matriz de las segundas derivadas).

$$H(\hat{\alpha},\,\hat{\beta}) = -\begin{bmatrix} \sum_{i=1}^{n} P_i(1-P_i) & \sum_{i=1}^{n} P_i(1-P_i)x_i \\ \sum_{i=1}^{n} P_i(1-P_i)x_i & \sum_{i=1}^{n} P_i(1-P_i)x_i^2 \end{bmatrix}$$

Modelo Logit: Estimación

- Los estimadores de máxima verosimilitud del modelo Logit son insesgados, consistentes y eficientes.
- La distribución asintótica es Normal,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \ [E\{-H(\hat{\alpha}, \hat{\beta})\}]^{-1}\right]$$

- La matriz de varianzas y covarianzas de los estimadores del modelo está dada por la inversa de la matriz Hesiana anterior con el signo opuesto.
- En la diagonal principal de esa matriz tenemos las varianzas de los coeficientes mientras que fuera de la diagonal principal tenemos las covarianzas entre los coeficientes.
- Volvamos a la modelización de la probabilidad de trabajar y ahora estimemos el modelo utilizando el Stata.

Agenda

Modelo Logit: Evaluación

- Una vez estimado el modelo, es importante (como sucedía en el caso del modelo de regresión lineal) chequear el modelo estimado utilizando medidas de bondad del ajuste y los contrastes de hipótesis habituales.
- Sin embargo, para los modelos de variable dependiente categórica no se puede calcular la medida de bondad del ajuste conocida como R^2.
- Una medida alternativa de bondad del ajuste se conoce como R^2 de McFadden y se define como:

$$R^2_{MF} = 1 - \frac{l(\hat{\alpha}, \hat{\beta})}{l(\hat{\alpha}_0)}$$

Modelo Logit: Evaluación

- Donde, $l(\hat{\alpha}, \hat{\beta})$ es el valor del logaritmo de la función de verosimilitud evaluada en los estimadores de MV y $l(\hat{\alpha}_0)$ es el valor del logaritmo de la función de verosimilitud de un modelo que tiene solo una constante.
- El R^2_{MF} tiene la misma interpretación que el R^2 común. Es decir, nos dice que porcentaje de la variabilidad de la variable dependiente está explicado por la regresión.
- El R^2_{MF} tiene la misma característica que el R^2 común: no sirve para comparar el ajuste de diferentes modelos con distinto número de variables explicativas.
- La medida de bondad del ajuste que se utiliza para comparar ajustes es el R^2_{MF} ajustado.

Modelo Logit: Evaluación

- El R^2_{MF} ajustado se define como:

$$\bar{R}^2_{MF} = 1 - \frac{l(\hat{\alpha}, \hat{\beta}) - k}{l(\hat{\alpha}_0)}$$

donde k es el número de parámetros a estimar.

- Al igual que el R^2 ajustado en la regresión lineal, el $\bar{R}^2_{MF}$ puede ser negativo.
- Otra forma de medir la bondad del ajuste es observar como clasifica a las observaciones el modelo en comparación con los datos realmente observados.
- En nuestro ejemplo, sería preguntarse cuán bien clasifica el modelo estimado a las personas que trabajan?

Modelo Logit: Evaluación

- Para responder a esta pregunta necesitamos saber cuándo nuestro modelo predice que $y_i = 1$ (es decir, cuándo $\hat{y}_i = 1$).
- Como nosotros tenemos una estimación de la probabilidad de trabajar una regla de clasificación sencilla podría ser: Siempre que la estimación de la probabilidad de trabajar sea mayor a un valor c (por ejemplo 0.5) entonces clasifique a esa persona como trabajando. En términos matemáticos:

$$Si\ F(\hat{\alpha} + \hat{\beta}\, x_i) = \frac{e^{\hat{\alpha}+\hat{\beta}\, x_i}}{1 - e^{\hat{\alpha}+\hat{\beta}\, x_i}} > c \Longrightarrow \hat{y}_i = 1$$

Modelo Logit: Evaluación

- En la práctica uno puede elegir cualquier valor para c. Una posibilidad es emplear la probabilidad empírica:

$$c = \frac{\# y_i = 1}{n}$$

- Una vez que clasificadas las observaciones se puede construir lo que se denomina la tabla de predicción-realización:

	Predicción	
Realización	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	p_{00}	p_{10}
$y_i = 1$	p_{01}	p_{11}

Modelo Logit: Evaluación

- La fracción $p_{00} + p_{11}$ se denomina tasa de aciertos. Cuanto mayor es la tasa de aciertos mejor es el ajuste del modelo.
- La tasa de aciertos se define como la fracción de predicciones correctas en la muestra. Formalmente, si definimos a la variable aleatoria w_i como la indicadora de una predicción correcta (esto es $w_i = 1$ si $y_i = \hat{y}_i$ y $w_i = 0$ si $y_i \neq \hat{y}_i$) entonces la tasa de aciertos se define como $h = \frac{1}{n} \sum_{i=1}^{n} w_i$
- En la población, la proporción de unos es p. Supongamos que nosotros hacemos una predicción completamente aleatoria. Esto es, predecimos un 1 con probabilidad p y predecimos un 0 con probabilidad $(1 - p)$. En este escenario, la probabilidad de hacer una predicción correcta es $q = p^2 + (1 - p)^2$.

Modelo Logit: Evaluación

- Usando las propiedades de la distribución binomial para el número de predicciones (hechas en forma aleatoria) correctas, la tasa de aciertos "aleatoria" (h_a) tiene $E(h_a) = q$ y $Var(h_a) = q(1-q)/n$. Entonces la habilidad predictiva del modelo estimado puede evaluarse comparándola con esta tasa de aciertos "aleatoria".
- La idea es hacer un test de hipótesis cuya hipótesis nula sea que las predicciones del modelo estimado no son mejores que las predicciones hechas en forma aleatoria y cuya hipótesis alternativa sea que las predicciones del modelo son mejores que las hechas aleatoriamente.

Modelo Logit: Evaluación

- Bajo la hipótesis nula, la tasa de aciertos, h se distribuye normalmente con media q y varianza $q(1-q)/n$. Por lo tanto el estadístico de contraste es:

$$z = \frac{h-q}{\sqrt{q(1-q)/n}} = \frac{nh-nq}{\sqrt{nq(1-q)}}$$

  En la práctica $q = p^2 + (1-p)^2$ se estima con $\hat{p}^2 + (1-\hat{p})^2$, con $\hat{p}$ siendo la proporción de unos en la muestra.

- La regla de decisión es rechazar la hipótesis nula siempre que el valor de probabilidad de z sea menor al nivel de significación del test ó siempre que z sea mayor al valor crítico de la normal estándar.

Modelo Logit: Evaluación

- Otros índices de interés son los denominados sensitividad y especificidad
- La sensitividad es la probabilidad de predecir un "éxito" entre los "éxitos": $Pr(\hat{y}_i = 1 | y_i = 1)$
- La especificidad es la probabilidad de predecir un "fracaso" entre los "fracasos": $Pr(\hat{y}_i = 0 | y_i = 0)$
- La probabilidad de predecir un "éxito falso" ó "falso positivo" es uno menos la especificidad:
  $Pr(\hat{y}_i = 1 | y_i = 0) = 1 - Pr(\hat{y}_i = 0 | y_i = 0)$
- Los falsos positivos corresponden a lo que llamamos error de tipo I.
- Es claro que una mejor bondad de ajuste se obtiene con una alta sensitividad y especificidad.

Modelo Logit: Evaluación

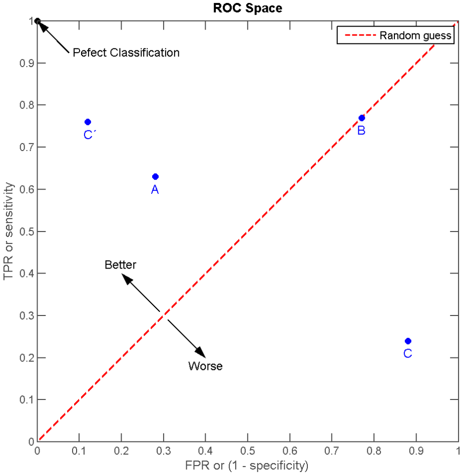
- En una tabla de contingencia (similar a la de predicción-realización) se pueden ver estas definiciones:

Realización	Predicción	
	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	$Pr(\hat{y}_i = 0 \vert y_i = 0)$ Verdadero Negativo	$Pr(\hat{y}_i = 1 \vert y_i = 0)$ Falso Positivo
$y_i = 1$	$Pr(\hat{y}_i = 0 \vert y_i = 1)$ Falso Negativo	$Pr(\hat{y}_i = 1 \vert y_i = 1)$ Verdadero Positivo

- Una forma de resumir la bondad del ajuste con estas dos medidas es graficando la curva ROC (Relative (Reciever) Operating Characteristic).
- La curva ROC es una representación gráfica de la sensibilidad frente a $(1 -$ especificidad$)$ para un sistema clasificador binario según se varía el umbral de clasificación.

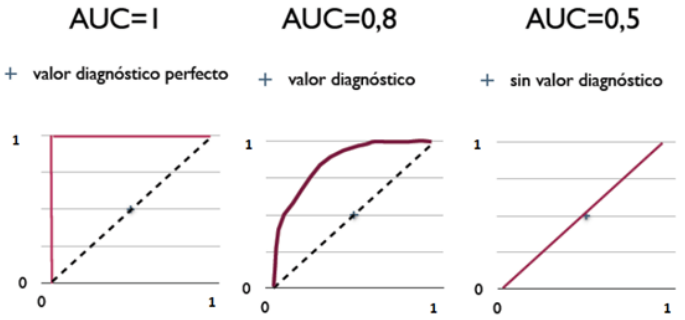
Modelo Logit: Evaluación

- Gráficamente:



Modelo Logit: Evaluación

- Una forma de resumir la curva ROC es calcular el área bajo la curva (AUC)
- Diferentes curvas ROC y sus AUC



Agenda

Modelo Probit

- Ahora asumimos que los errores del modelo en lugar de describirse con la función de distribución logística, pueden describirse con la función de distribución Normal estándar.
- Cuando representamos los errores con la distribución Normal, el modelo resultante recibe el nombre de Modelo Probit.
- Una variable z se dice que tiene distribución normal cuando su función de distribución de probabilidad acumulada tiene la siguiente forma:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

Modelo Probit

- En términos de nuestro modelo,

$$\Phi(\hat{\alpha} + \hat{\beta}\, x_i) = \int_{-\infty}^{\hat{\alpha}+\hat{\beta}\, x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

- Al igual de lo que ocurría con el modelo Logit en este caso existe una relación no lineal en los parámetros del modelo y por lo tanto, el método de mínimos cuadrados no puede usarse para estimar el modelo Probit.
- Esta relación no lineal puede ser vista calculando la derivada parcial de la probabilidad de trabajar con respecto al salatio ofrecido. Por ejemplo,

$$\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} = \phi(\hat{\alpha} + \hat{\beta}\, x_i) \times \beta$$

Modelo Probit

- La ecuación anterior muestra que el efecto de un cambio en el salario ofrecido sobre la probabilidad de trabajar depende no solo del valor de β sino también del valor tomado por la función de densidad de la Normal estándar (ϕ).
- La derivada parcial refleja las diferentes pendientes de la curva de distribución acumulada de la normal estándar. Esto significa que hay un valor para el cambio marginal de la probabilidad de trabajar para cada valor del salario ofrecido.
- La forma de interpretar los coeficientes del modelo Probit es la misma que en el modelo Logit. Hay que calcular los efectos marginales promedio sobre la muestra de n observaciones con cualquiera de las dos formas descriptas anteriormente.

Agenda

Modelo Probit: Estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y sabemos que tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
- En términos matemáticos la función de verosimilitud del modelo Probit es,

$$L(\hat{\alpha}, \hat{\beta}; x_i) = \prod_{i=1}^{n} \left[\Phi(\hat{\alpha} + \hat{\beta}\, x_i)\right]^{y_i} \times \left[1 - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)\right]^{1-y_i}$$

Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\alpha},\hat{\beta};x_i)=\sum_{i=1}^{n}\left[y_i ln\{\Phi(\hat{\alpha}+\hat{\beta}x_i)\}+(1-y_i)ln\{1-\Phi(\hat{\alpha}+\hat{\beta}x_i)\}\right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$\frac{\partial l(\cdot)}{\partial\hat{\alpha}} = \sum_{i=1}^{n}\left[\frac{y_i-\Phi(\hat{\alpha}+\hat{\beta}x_i)}{\Phi(\hat{\alpha}+\hat{\beta}x_i)[1-\Phi(\hat{\alpha}+\hat{\beta}x_i)]}\phi(\hat{\alpha}+\hat{\beta}x_i)\right]=0$$

$$\frac{\partial l(\cdot)}{\partial\hat{\beta}} = \sum_{i=1}^{n}\left[\frac{y_i-\Phi(\hat{\alpha}+\hat{\beta}x_i)}{\Phi(\hat{\alpha}+\hat{\beta}x_i)[1-\Phi(\hat{\alpha}+\hat{\beta}x_i)]}\phi(\hat{\alpha}+\hat{\beta}x_i)\right]x_i=0$$

Modelo Probit: Estimación

- Como se puede observar en las condiciones de primer orden, las incógnitas de ambas ecuaciones $(\hat{\alpha}, \hat{\beta})$ entran en forma no lineal y por lo tanto no pueden resolverse por métodos lineales.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen.
- Los estimadores de máxima verosimilitud del modelo Probit son insesgados, consistentes y eficientes.
- La matriz de varianzas y covarianzas de los estimadores del modelo está dada por la inversa de la matriz de las condiciones de segundo orden con el signo opuesto.
- En la diagonal principal de esta matriz tenemos las varianzas de los coeficientes mientras que fuera de la diagonal principal tenemos las covarianzas entre los coeficientes.

Modelo Probit: Estimación

- Una vez que los parámetros del modelo han sido estimados, pueden realizarse los contrastes de hipótesis habituales.
- También pueden utilizarse todas las medidas de bondad del ajuste mencionadas para el caso del modelo Logit.
- Como existe una relación entre la distribución Normal y la distribución Logística, también existe una relación entre los coeficientes estimados del Probit y del Logit.
- Amemiya (1981) sugiere la siguiente relación entre las estimaciones del modelo Probit y Logit:

$$\hat{\beta}_{Probit} = 0.625\hat{\beta}_{Logit}$$

Análisis Discriminante

- Al igual que el método para variables categóricas, el análisis discriminante tiene el objetivo de asignar nuevos objetos (observaciones) a grupos previamente definidos.
- Para fijar ideas, supongamos que estamos interesados en asignar un nuevo objeto a una de dos clases. Vamos a llamar a estas clases π_1 y π_2.
- Los objetos son clasificados o separados sobre la base de observar, por ejemplo, p variables $X_1, X_2, \ldots, X_p$.
- Los valores de las X's difieren en alguna medida entre las dos clases y por lo tanto uno podría pensar en diferenciar a las dos clases en función de la probabilidad de pertenecer a cada población.

Análisis Discriminante

- Esto es, si $f_1(X)$ y $f_2(X)$ representan a las probabilidades de pertenecer a las clases π_1 y π_2, respectivamente, uno puede hablar de clasificar a los datos como provenientes de dos poblaciones diferentes, utilizando estas funciones de probabilidad.
- Específicamente, el conjunto de datos se divide en dos regiones R_1 y R_2, tal que si una nueva observación cae en R_1 se asigna a la población π_1 y si cae en R_2 se asigna a π_2.
- Sin embargo, las reglas de clasificación no están exentas de error. Esto puede deberse a que no hay una clara distinción entre las características medidas de ambas poblaciones y por lo tanto los grupos pueden superponerse.
- Es posible, entonces, que uno pueda clasificar incorrectamente un objeto de la población π_2 en π_1 y viceversa.

Análisis Discriminante

- Un buen procedimiento de clasificación debería resultar en pocas equivocaciones. En otras palabras, la probabilidad de clasificar incorrectamente un objeto debe ser pequeña.
- El análisis discriminante utiliza la misma regla de clasificación que la mencionada arriba para el modelo logit. Si X_0 es una nueva observación y si $f_1(X_0)/f_2(X_0) > 1$ debemos asignar X_0 a π_1. Por otro lado, si $f_1(X_0)/f_2(X_0) < 1$ debemos asignar X_0 a π_2.
- Asumamos que conocemos que $f_1(X)$ y $f_2(X)$ son funciones de densidad normales multivariantes, la primera con vector de medias μ_1 y matriz de varianzas y covarianzas Σ_1, y la segunda con vector de medias μ_2 y matriz de varianzas y covarianzas Σ_2.

Análisis Discriminante

- Además, supongamos que $\Sigma_1 = \Sigma_2 = \Sigma$. Entonces, las funciones de probabilidad conjunta de $X' = [X_1, X_2, ..., X_p]$ para las poblaciones $\pi_1 y \pi_2$, vienen dadas por:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)]}, \quad i = 1,\ 2.$$

- Utilizando la regla de clasificación anterior, asignamos una nueva observación a la población π_1 si

$$e^{[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)]+[\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)]} > 1,\ \text{ó}$$

$$[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)] + [\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)] > 0$$

Análisis Discriminante

- Reordenando los términos de la última expresión,

$$[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)]+[\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)]=$$

$$(\mu_1-\mu_2)'\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)>0 \tag{5}$$

- En la práctica, los parámetros poblacionales μ_1, μ_2 y Σ son desconocidos y deben ser reemplazados por sus estimaciones muestrales.
- Supongamos que tenemos n_1 observaciones muestrales de las variables $X' = [X_1, X_2, \ldots, X_p]$ provenientes de π_1 y n_2 observaciones de las mismas variables provenientes de π_2.
- Entonces los respectivos datos de ambas poblaciones son,

Análisis Discriminante

- $$X_1 = \begin{bmatrix} x'_{11} \\ x'_{12} \\ \vdots \\ x'_{1n_1} \end{bmatrix} \quad X_2 = \begin{bmatrix} x'_{21} \\ x'_{22} \\ \vdots \\ x'_{2n_2} \end{bmatrix}$$

  con estos datos muestrales, se pueden calcular los vectores de medias y las matrices de varianzas y covarianzas.

- $$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)'$$
  $$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}, \quad S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)'$$

Análisis Discriminante

- Las matrices de varianzas y covarianzas muestrales S_1 y S_2 deben combinarse para obtener una estimación de la varianza conjunta. En particular, el promedio ponderado es un estimador insesgado de Σ si las muestras son aleatorias.

$$S = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} S_2$$

- Sustituyendo en (1), μ_1 por $\bar{x}_1$, μ_2 por $\bar{x}_2$ y Σ por S obtenemos la regla de clasificación muestral del análisis discriminante para dos poblaciones: Asigne X_0 a π_1 si

$$(\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 - \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) > 0$$

Análisis Discriminante

- Note que la regla de clasificación implica comparar dos números,

$$\hat{y} = (\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 = \hat{\alpha}' X_0,$$

- con

$$\begin{aligned} \hat{m} &= \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) \\ &= \frac{1}{2}(\bar{\hat{y}}_1 + \bar{\hat{y}}_2) \end{aligned}$$

- donde

$$\bar{\hat{y}}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1 = \hat{\alpha}' \bar{x}_1,$$

$$\bar{\hat{y}}_2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_2 = \hat{\alpha}' \bar{x}_2.$$

Análisis Discriminante

- Entonces, la regla de asignación del análisis discriminante para dos poblaciones normales consiste en crear dos grupos utilizando los valores de $\hat{y}$, calculado a través de una combinación lineal apropiada de las observaciones de las muestras de las poblaciones π_1 y π_2, y luego asignar una nueva observación, X_0, a π_1 ó π_2, dependiendo de si $\hat{y} = \hat{\alpha}' X_0$ cae a la derecha o a la izquierda del punto medio entre las medias de $\hat{y}$ en los dos grupos, $\hat{m} = \frac{1}{2}(\bar{\hat{y}}_1 + \bar{\hat{y}}_2)$.
- Los coeficientes del vector $\hat{\alpha} = S^{-1}(\bar{x}_1 - \bar{x}_2)$ se denominan coeficientes discriminantes.
- Estos coeficientes no son únicos, ya que cualquier múltiplo de los mismos también sirve para discriminar. Esto es, para cualquier $c \neq 0$, el vector $c\hat{\alpha}$ discrimina entre dos poblaciones de la misma manera que lo hace $\hat{\alpha}$.

Análisis Discriminante

- El vector $\hat{\alpha}$ usualmente se "normaliza" para facilitar su interpretación. Las dos formas más comunes de normalización son las siguientes:
  - Defina
$$\hat{\alpha}^* = c\hat{\alpha} = \frac{\hat{\alpha}}{\sqrt{\hat{\alpha}'\hat{\alpha}}}$$
de forma tal que $\hat{\alpha}^*$ tenga largo unitario.
  - Defina
$$\hat{\alpha}^* = c\hat{\alpha} = \frac{\hat{\alpha}}{\hat{\alpha}_1}$$
de forma tal que el primer elemento del nuevo vector $\hat{\alpha}^*$ sea 1.
- Las magnitudes de $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \ldots, \hat{\alpha}_p^*$ en la primera normalización caen en el intervalo $[-1, 1]$. En la segunda normalización, $\hat{\alpha}_1^* = 1$ y $\hat{\alpha}_2^*, \ldots, \hat{\alpha}_p^*$ están expresados como múltiplos de $\hat{\alpha}_1^*$.

Análisis Discriminante

- Restringir a que los $\hat{\alpha}_i^*$ caigan en el intervalo $[-1, 1]$ facilita la comparación entre los coeficientes.
- Similarmente, expresar a los coeficientes como múltiplos de $\hat{\alpha}_1^*$ permite evaluar la importancia relativa de las variables $X_2, X_3, \ldots, X_p$ como discriminantes.
- Qué se puede hacer cuando no se conoce la distribución de probabilidad de las variables?
- Una alternativa consiste en construir una combinación lineal de las variables (una suma ponderada) de forma tal que esta combinación discrimine de la mejor manera a los grupos.
- Podemos luego comparar como difieren los grupos con respecto a esta combinación lineal y también observar los pesos relativos de cada variable para determinar su importancia relativa en la discriminación.

Análisis Discriminante

- La función discriminante de Fisher es el método por el cual se determina la combinación lineal.
- La idea de Fisher es transformar las observaciones multivariantes de X en observaciones univariantes de $\hat{y}$, tal que las $\hat{y}$'s derivadas de las poblaciones π_1 y π_2 estuvieran tan separadas como fuera posible.
- Fisher sugirió tomar combinaciones lineales de X para crear $\hat{y}$. Esto es,

$$
\begin{aligned}
\hat{y}_{1i} &= \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_p X_{pi} \quad i = 1, 2, \ldots, n_1 \\
\hat{y}_{2i} &= \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_p X_{pi} \quad i = 1, 2, \ldots, n_2
\end{aligned}
$$

- La separación de estos dos conjuntos de valores de $\hat{y}$ se establece en función de la diferencia entre $\bar{\hat{y}}_1$ e $\bar{\hat{y}}_2$ expresada en unidades de desvíos estándar.

- En términos matemáticos,

$$\text{Separación} = \frac{|\bar{\hat{y}}_1 - \bar{\hat{y}}_2|}{s_{\hat{y}}}$$

- Donde,

$$s_{\hat{y}}^2 = \frac{\sum_{i=1}^{n_1}(\hat{y}_{1i} - \bar{\hat{y}}_1)^2 + \sum_{i=1}^{n_2}(\hat{y}_{2i} - \bar{\hat{y}}_2)^2}{n_1 + n_2 - 2}$$

- El objetivo es seleccionar la combinación lineal de las X's para alcanzar la máxima separación entre las medias muestrales $\bar{\hat{y}}_1$ e $\bar{\hat{y}}_2$. Para ello, maximizar la ecuación anterior es lo mismo que maximizar,

$$\frac{\text{Distancia entre las medias de } \hat{y} \text{ al cuadrado}}{\text{Varianza de } \hat{y}} = \frac{(\bar{\hat{y}}_1 - \bar{\hat{y}}_2)^2}{s_{\hat{y}}^2}$$

En la expresión anterior, note que,

$$
\begin{aligned}
s_{\hat{y}}^2 &= \frac{\sum_{j=1}^{n_1}(\hat{y}_{1j} - \bar{\hat{y}}_1)^2 + \sum_{j=1}^{n_2}(\hat{y}_{2j} - \bar{\hat{y}}_2)^2}{n_1 + n_2 - 2} \\
&= \frac{\sum_{j=1}^{n_1}(\hat{\alpha}' x_{1j} - \hat{\alpha}'\bar{x}_1)(x_{1j}'\hat{\alpha} - \bar{x}_1'\hat{\alpha})}{n_1 + n_2 - 2} \\
&+ \frac{\sum_{j=1}^{n_2}(\hat{\alpha}' x_{2j} - \hat{\alpha}'\bar{x}_2)(x_{2j}'\hat{\alpha} - \bar{x}_2'\hat{\alpha})}{n_1 + n_2 - 2} \\
&= \frac{\hat{\alpha}'[\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)']\hat{\alpha} + \hat{\alpha}'[\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)']\hat{\alpha}}{n_1 + n_2 - 2} \\
&= \frac{\hat{\alpha}'\{[\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)'] + [\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)']\}\hat{\alpha}}{n_1 + n_2 - 2} \\
&= \frac{\hat{\alpha}'[(n_1 - 1)S_1 + (n_2 - 1)S_2]\hat{\alpha}}{n_1 + n_2 - 2} \\
&= \hat{\alpha}' S \hat{\alpha}
\end{aligned}
$$

- y que,

$$(\bar{\hat{y}}_1 - \bar{\hat{y}}_2)^2 = (\hat{\alpha}'\bar{x}_1 - \hat{\alpha}'\bar{x}_2)^2 = [\hat{\alpha}'(\bar{x}_1 - \bar{x}_2)]^2 = [\hat{\alpha}'d]^2,$$

donde $d = (\bar{x}_1 - \bar{x}_2)$.

- Por lo tanto el problema se reduce a encontrar los coeficientes $\hat{\alpha}$ que maximicen,

$$\frac{(\text{Distancia entre las medias muestrales de } \hat{y} \text{ al cuadrado})}{(\text{Varianza muestral de } \hat{y})} = \frac{[\hat{\alpha}'d]^2}{\hat{\alpha}'S\hat{\alpha}}$$

- Realizando la maximización sobre todos los posibles $\hat{\alpha}$ se puede mostrar que el máximo se alcanza con los coeficientes $\hat{\alpha} = (\bar{x}_1 - \bar{x}_2)'S^{-1}$ que son los mismos coeficientes que determinamos anteriormente para dos poblaciones normales.

Análisis Discriminante

- Note que el procedimiento de Fisher no asume que las poblaciones son normales, sin embargo si asume implicitamente que las matrices de varianzas y covarianzas de las dos poblaciones son iguales.
- Con estos supuestos Fisher llega a la misma regla de clasificación que asumiendo poblaciones normales. Esto es, Asigne X_0 a π_1 si

$$(\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2) > 0$$

Logit versus Discriminante

- Para ver la relación entre los coeficientes discriminantes y los del modelo Logit escribamos $Pr[y_i = 1|x]$ usando la definición de la probabilidad condicional. Esto es,

$$\begin{aligned} Pr[y_i = 1|x] &= \frac{Pr[y_i = 1 \ \wedge \ x]}{Pr[x]} \\ &= \frac{Pr[x|y_i = 1]Pr[y_i = 1]}{Pr[x]} \end{aligned}$$

de la misma manera,

$$Pr[y_i = 0|x] = \frac{Pr[x|y_i = 0]Pr[y_i = 0]}{Pr[x]}$$

- Ahora calculemos la tasa de probabilidad,

$$\frac{Pr[y_i = 1|x]}{Pr[y_i = 0|x]} = \frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]} \times \frac{Pr[y_i = 1]}{Pr[y_i = 0]}$$

Logit versus Discriminante

- Asumiendo que las probabilidades iniciales son iguales (i.e. $Pr[y_i = 1] = Pr[y_i = 0]$) y tomando logaritmos naturales en ambos miembros, la ecuación anterior queda,

$$log\left\{\frac{Pr[y_i = 1|x]}{Pr[y_i = 0|x]}\right\} = log\left\{\frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]}\right\}$$

- Reemplazando el lado izquierdo por el logaritmo de la tasa de probabilidad del modelo Logit, tenemos

$$\beta_0 + \beta_1 x = log\left\{\frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]}\right\}$$

- Note que el lado derecho de la ecuación anterior es el cociente de las funciones de probabilidad de las variables explicativas para los dos grupos bajo análisis. Esto es lo que llamamos $f_1(x)$ y $f_2(x)$ en el análisis discriminante.

Logit versus Discriminante

- Si las distribuciones de probabilidad de las variables explicativas son Normales Multivariantes con la misma matriz de varianzas y covarianzas en ambos grupos, el lado derecho de la ecuación anterior puede reemplazarse por el lado izquierdo de la ecuación (5) para obtener,

$$\beta_0 + \beta_1 x = -\frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) + (\mu_1 - \mu_2)'\Sigma^{-1}x$$

- Igualando los coeficientes de ambos lados de la ecuación se obtiene,

$$\beta_0 = -\frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

y

$$\beta_1 = (\mu_1 - \mu_2)'\Sigma^{-1}$$

- Entonces las estimaciones de los coeficientes del modelo Logit vienen dadas por,

$$\hat{\beta}_0 = -\frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2) = -\frac{1}{2}\hat{\alpha}(\bar{x}_1 + \bar{x}_2)$$

y

$$\hat{\beta}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1}$$