Modelos com Variáveis Dependentes Qualitativas

Considere, por exemplo, modelar a participação na força de trabalho. Atribuiria-se valor *zero* a não participação e *um* a participação no mercado de trabalho. Fatores tais como: idade, educação, estado civil, número de crianças e, outros, seriam relevantes em explicar se um indivíduo procura emprego ou não em um dado período de tempo.

Neste exemplo, considere todas as variáveis exógenas representadas pelo vetor X, então este vetor explica a decisão entre trabalhar ou não, i.e.,

Prob(Y = 1) = $F(\beta' X)$ $\rightarrow$ probabilidade de um indivíduo com determinadas características X trabalhar

$$\text{Prob}\left(Y = 0\right) = 1 - F\left(\beta' X\right)$$

$$E(Y) = 1.F(\beta' X) + 0.(1 - F(\beta' X)) = F(\beta' X)$$

Os parâmetros β refletem o impacto das mudanças de X na probabilidade de participação no mercado de trabalho. Podemos, por exemplo, querer saber qual é o efeito marginal do estado civil do indivíduo na probabilidade de participação no mercado de trabalho. Portanto, precisamos de um modelo (forma-funcional) para o lado direito da equação.

Uma possibilidade é um modelo linear onde,

$$F(X, \beta) = \beta' X$$

uma vez que $E(Y)=F(X, \beta)$, podemos construir o modelo de regressão.

$$Y = E(Y) + (Y - E(Y))$$

$$Y = \beta' X + \varepsilon$$

Porém, o modelo linear tem alguns problemas.

Um deles é o fato de ε ser heteroscedástico.

$$\beta'X + \varepsilon = 1 \qquad \varepsilon = 1 - \beta'X \qquad \text{com} \qquad \Pr ob(Y = 1) = F$$

$$\beta'X + \varepsilon = 0 \qquad \varepsilon = -\beta'X \qquad \text{com} \qquad \Pr ob(Y = 0) = 1 - F$$

Então,

$$Var(\varepsilon) = (1 - \beta'X)^2 \cdot F + (-\beta'X)^2 (1 - F)$$

$$Var(\varepsilon) = \left\lfloor 1 - 2\beta'X + (\beta'X)^2 \right\rfloor F + (\beta'X)^2 (1 - F)$$

Se considerarmos um modelo linear,

$$F = \beta'X$$

então,

$$Var(\varepsilon) = \beta'X\left[1 - 2\beta'X + (\beta'X)^2 + \beta'X - (\beta'X)^2\right]$$

$$Var(\varepsilon) = \beta'X(1 - \beta'X)$$

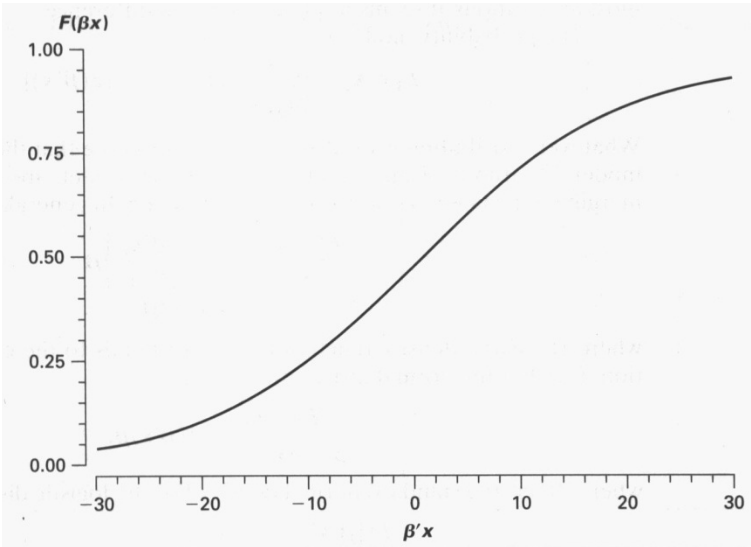
e, portanto, não é constante.

Ademais, teríamos problemas com as previsões, pois não poderíamos restringir $\beta'X$ ao intervalo 0–1.

Esperaríamos que:

$$\lim_{\beta'X \to +\infty} \text{Prob}(Y = 1) = 1$$

$$\lim_{\beta'X \to -\infty} \text{Prob}(Y = 1) = 0$$



Isto pode ser obtido com uma distribuição normal ou logística, dentre outras.

Considere então a distribuição normal:

$$\text{Prob}(Y = 1) = F(\beta' X) = \int_{-\infty}^{\beta' X} \phi(t)\,dt = \Phi(\beta' X)$$

$$\phi(t) = (2\pi)^{-1/2}\; e^{-t^2/2}$$

e a logística:

$$P(Y = 1) = F(\beta' X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}} = \Lambda(\beta' X)$$

O modelo de probabilidade é uma regressão:

$$E(Y) = 0 \cdot [1 - F(\beta' X)] + 1 \cdot [F(\beta' X)]$$

ou

$$E(Y) = F(\beta' X)$$

Observe que os parâmetros do modelo não são necessariamente os efeitos marginais que estávamos acostumados a analisar.

Na regressão linear:

$$Y = \alpha + \beta X + \varepsilon \qquad \rightarrow \qquad \frac{\partial Y}{\partial X} = \beta$$

Neste caso

$$\frac{\partial E(Y)}{\partial X} = \left\{ \frac{\partial F(\beta' X)}{\partial(\beta' X)} \right\} \cdot \beta = f(\beta' X) \cdot \beta$$

onde *F* é a função de distribuição e *f* é a função densidade de probab.

Modelo Lógite

$$E(Y) = P(Y=1) = F(\beta'X) = \frac{e^{\beta'X}}{1+e^{\beta'X}}$$

$$P(Y=1) = \frac{\frac{e^{\beta'X}}{e^{\beta'X}}}{\frac{1+e^{\beta'X}}{e^{\beta'X}}} = \frac{1}{\frac{1}{e^{\beta'X}}+1} \qquad \rightarrow$$

$$P(Y=1)\cdot\left(\frac{1}{e^{\beta'X}}+1\right) = 1 \qquad \rightarrow \qquad \frac{P(Y=1)}{e^{\beta'X}} = 1 - P(Y=1) \qquad \rightarrow$$

$$e^{\beta'X} = \frac{P(Y=1)}{1-P(Y=1)} \qquad \rightarrow \qquad \log\frac{P(Y=1)}{P(Y=0)} = \beta'X$$

Para saber o efeito que uma mudança em X causa na probabilidade de Y_i=1, faz-se

$$\frac{\partial E(Y_i)}{\partial X} = e^{\beta' X_i} \cdot \beta \left(1 + e^{\beta' X_i}\right)^{-1} + -\left(1 + e^{\beta' X_i}\right)^{-2} e^{\beta' X} \beta \left(e^{\beta' X}\right)$$

$$= \frac{e^{\beta' X_i} \cdot \beta \left(1 + e^{\beta' X_i}\right) - e^{\beta' X_i} \beta \cdot e^{\beta' X_i}}{\left(1 + e^{\beta' X_i}\right)^2}$$

$$= \frac{e^{\beta' X_i} \cdot \beta \left(1 + e^{\beta' X_i} - e^{\beta' X_i}\right)}{\left(1 + e^{\beta' X_i}\right)^2}$$

$$\frac{\partial P(Y_i = 1)}{\partial X} = \beta \cdot \underbrace{\frac{e^{\beta' X_i}}{\left(1 + e^{\beta' X_i}\right)^2}}_{f.densidade}$$

Observe que $\partial E(Y_i)/\partial X$ não é β como na regressão linear

$$\frac{\partial E(Y)}{\partial X} = \beta \cdot F(\beta' X)(1 - F(\beta' X))$$

onde

$$F(\beta' X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}$$

Estimação e Inferência

A estimação é baseada no método de máxima verossimilhança. Cada observação é tratada como sendo retirada de uma distribuição de Bernoulli. O modelo então será:

$$L = \text{Prob}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \prod_{i=1}^{n} [F(\beta' X_i)]^{y_i} [1 - F(\beta' X_i)]^{1 - y_i}$$

Aplicando logaritmo,

$$\ln L = \sum_i \left\{ y_i \ln F\left(\beta' X_i\right) + \left(1 - y_i\right) \ln\left[1 - F\left(\beta' X_i\right)\right] \right\}$$

A condição de primeira ordem para o máximo requer:

$$\frac{\partial \ln L}{\partial \beta} = \sum \left[\frac{y_i f_i}{F_i} + \left(1 - y_i\right) \frac{-f_i}{\left(1 - F_i\right)} \right] X_i = 0$$

A não ser que estejamos usando o modelo linear de probabilidade, a equação acima será não linear e necessitará de um método iterativo para a solução.

Modelo Lógite:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta' X_i}}{1 + e^{\beta' X_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\beta' X_i}} \right)^{1 - y_i}$$

$$\ln L(\beta) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} \right) + (1-y_i)\ln\left(\frac{1}{1+e^{\beta' X_i}} \right) \right\}$$

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left\{ y_i \left(\frac{1+e^{\beta' X_i}}{e^{\beta' X_i}} \right) \left[\frac{e^{\beta' X_i} \cdot X_i \left(1+e^{\beta' X_i}\right) - e^{\beta' X_i} \cdot e^{\beta' X_i} \cdot X_i}{\left(1+e^{\beta' X_i}\right)^2} \right] + (1-y_i)\left(1+e^{\beta' X_i}\right) \left(\frac{-e^{\beta' X_i} \cdot X_i}{\left(1+e^{\beta' X_i}\right)^2} \right) \right\} = 0$$

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i \cdot X_i \left(1+e^{\beta' X_i}\right) - y_i e^{\beta' X_i} X_i - e^{\beta' X_i} \cdot X_i + y_i e^{\beta' x_i} \cdot X_i}{\left(1+e^{\beta' X_i}\right)} \right] = 0$$

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left(y_i X_i - \frac{e^{\beta' X_i}}{1+e^{\beta' X_i}} X_i \right) = \sum_{i=1}^{n} \left(y_i - \Lambda_i \right) X_i = 0$$

se

$$\hat{\Lambda}_i = \frac{e^{\hat{\beta}' X_i}}{1+e^{\hat{\beta}' X_i}} = \hat{p}_i \quad \text{então} \qquad \Sigma y_i X_i = \Sigma \hat{p}_i X_i$$

Portanto, se X inclui constante, a soma das probabilidades estimadas é igual a soma dos Y_i ou ao número de observações da amostra em que Y_i=1.

$$\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = \sum_{i=1}^{n} -\left[\frac{e^{\beta' X_i} X_i X_i' \left(1+e^{\beta' X_i}\right) - e^{\beta' X_i} X_i e^{\beta' X_i} X_i'}{\left(1+e^{\beta' X_i}\right)^2}\right]$$

$$\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = \sum_{i=1}^{n} -\left[\frac{e^{\beta' X_i} X_i X_i' + e^{\beta' X_i} X_i X_i' e^{\beta' X_i} - e^{\beta' X_i} X_i X_i' e^{\beta' X_i}}{\left(1+e^{\beta_i' X_i}\right)^2}\right]$$

$$\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \left(\frac{e^{\beta' X_i}}{\left(1+e^{\beta' X_i}\right)} \cdot \frac{1}{\left(1+e^{\beta' X_i}\right)}\right) X_i X_i' = -\sum_{i=1}^{n} \Lambda\left(1-\Lambda\right) X_i X_i'$$

$$Var\left(\hat{\beta}\right) = \left(-\frac{\partial^2 \ln L}{\partial\beta\partial\beta'}\right)^{-1} = \left[\sum_{i=1}^{n} \frac{e^{\beta' X_i}}{\left(1+e^{\beta' X_i}\right)^2} X_i X_i'\right]^{-1}$$

Modelo Próbite

$$L = \prod_{i=1}^{n} (\Phi_i)^{y_i} (1 - \Phi_i)^{1 - y_i}$$

$$\ln L = \sum_{i=1}^{n} y_i \ln \Phi_i + \sum_{i=1}^{n} (1 - y_i) \ln (1 - \Phi_i)$$

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} y_i \frac{\phi}{\Phi} X_i - \sum_{i=1}^{n} (1 - y_i) \frac{\phi}{1 - \Phi} X_i$$

Testes de hipóteses.

Para testar uma única restrição, a forma mais simples é utilizar o teste *t* usual com o desvio-padrão retirado da matriz de informação e a tabela da distribuição normal padronizada para obter o valor crítico.

Para restrições mais complicadas é usual utilizar o teste de Wald.

Um teste semelhante ao F que testa se todas as inclinações da regressão são zero é a razão de verossimilhança.

$$LR = -2[\ln \hat{L}_r - \ln \hat{L}]$$

onde $\hat{L}_r$ e $\hat{L}$ são funções de verossimilhança no logaritmo avaliadas nos valores restritos (tudo 0 exceto o intercepto) e não restritos (estimativas).

Medida de ajustamento.

Índice da razão de verossimilhança = LRI = 1 – ln L/ln L_r

Também é conhecido como pseudo R^2

Total de 0's e 1's previstos corretamente no total de dados.

Modelo Próbite Bivariado.

$$z^* = \mathbf{w}'\boldsymbol{\alpha} + \varepsilon_1, \qquad z = 1 \text{ se } z^* > 0, \ 0 \text{ em caso contrário}$$

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon_2, \qquad y = 1 \text{ se } y^* > 0, \ 0 \text{ em caso contrário}$$

Considera-se:
z = 1 quando crianças freqüentam escola, e z = 0 em caso contrário.
y = 1 quando crianças trabalham, e y = 0 em caso contrário.

Pressuposições:

$$E[\varepsilon_1 | \mathbf{w}, \mathbf{x}] = E[\varepsilon_2 | \mathbf{w}, \mathbf{x}] = 0,$$

$$\text{Var}[\varepsilon_1 | \mathbf{w}, \mathbf{x}] = \text{Var}[\varepsilon_2 | \mathbf{w}, \mathbf{x}] = 1,$$

$$\text{Cov}[\varepsilon_1, \varepsilon_2 | \mathbf{w}, \mathbf{x}] = \rho.$$

ε_1 e ε_2 tem distribuição normal bivariada (BVN)

$$\text{Prob}[y=1, z=1] = \text{BVN}(\mathbf{w}'\boldsymbol{\alpha}, \mathbf{x}'\boldsymbol{\beta}, \rho),$$

$$\text{Prob}[y=1, z=0] = \text{BVN}(-\mathbf{w}'\boldsymbol{\alpha}, \mathbf{x}'\boldsymbol{\beta}, -\rho),$$

$$\text{Prob}[y=0, z=1] = \text{BVN}(\mathbf{w}'\boldsymbol{\alpha}, -\mathbf{x}'\boldsymbol{\beta}, -\rho),$$

$$\text{Prob}[y=0, z=0] = \text{BVN}(-\mathbf{w}'\boldsymbol{\alpha}, -\mathbf{x}'\boldsymbol{\beta}, \rho),$$

A função de verossimilhança será:

$$\begin{aligned}\log \text{L} = &\sum_{i \text{ para } y=1, z=1} \ln \text{BVN}(\mathbf{w}'\boldsymbol{\alpha}, \mathbf{x}'\boldsymbol{\beta}, \rho) + \sum_{i \text{ para } y=1, z=0} \ln \text{BVN}(-\mathbf{w}'\boldsymbol{\alpha}, \mathbf{x}'\boldsymbol{\beta}, -\rho) \\ &+ \sum_{i \text{ para } y=0, z=1} \ln \text{BVN}(\mathbf{w}'\boldsymbol{\alpha}, -\mathbf{x}'\boldsymbol{\beta}, -\rho) + \sum_{i \text{ para } y=0, z=0} \ln \text{BVN}(-\mathbf{w}'\boldsymbol{\alpha}, -\mathbf{x}'\boldsymbol{\beta}, \rho)\end{aligned}$$

Efeito marginal:

$$E(y|z=1,\mathbf{w},\mathbf{x}) = \text{Prob}[y=1|z=1,\mathbf{w},\mathbf{x}] = \frac{\text{Prob}[y=1,z=1|\mathbf{w},\mathbf{x}]}{\text{Prob}[z=1|\mathbf{w}]}$$

$$= \frac{\text{BVN}(\mathbf{w}'\boldsymbol{\alpha},\mathbf{x}'\boldsymbol{\beta},\rho)}{\Phi(\mathbf{w}'\boldsymbol{\alpha})}$$

$$\frac{\partial E(y|z=1,\mathbf{w},\mathbf{x})}{\partial x_h}$$

Modelo Lógite Multinomial

Quando a decisão é feita entre mais de duas alternativas. Ex: ir ao trabalho de ônibus, carro, bicicleta ou trem. Trabalhar no mercado formal, informal ou não trabalhar.

Suponha que existam *m* categorias. Deixe P_1, P_2, ..., P_m serem as probabilidades associadas a essas *m* categorias. Então, a idéia é expressar estas probabilidades na forma binária.

$$\frac{P_1}{P_1 + P_m} = F\left(\beta_1' X\right) \qquad \frac{P_2}{P_2 + P_m} = F\left(\beta_2' X\right) \qquad \frac{P_{m-1}}{P_{m-1} + P_m} = F\left(\beta_{m-1}' X\right)$$

Então,

$$\frac{P_j}{P_j + P_m} = \frac{\dfrac{P_j}{P_m}}{\dfrac{P_j}{P_m} + 1} = F\left(\beta'_j X\right) \Rightarrow \frac{P_j}{P_m} = F\left(\beta'_j X\right)\frac{P_j}{P_m} + F\left(\beta'_j X\right)$$

Portanto,

$$\frac{P_j}{P_m} = \frac{F\left(\beta'_j X\right)}{1 - F\left(\beta'_j X\right)} = G\left(\beta'_j X\right)$$

Uma vez que

$$\sum_{j=1}^{m-1} \frac{P_j}{P_m} = \frac{1 - P_m}{P_m} = \frac{1}{P_m} - 1$$

$$\sum_{j=1}^{m-1} \frac{P_j}{P_m} = \sum_{j=1}^{m-1} G\left(\beta'_j X\right) = \frac{1}{P_m} - 1$$

Portanto,

$$\frac{1}{P_m} = 1 + \sum_{j=1}^{m-1} G\left(\beta'_j X\right) \quad \rightarrow \quad P_m = \left[1 + \sum_{j=1}^{m-1} G\left(\beta'_j X\right)\right]^{-1}$$

como
$$P_j = G\left(\beta'_j X\right) \cdot P_m$$

temos
$$P_j = \frac{G\left(\beta'_j X\right)}{1 + \sum_{j=1}^{m-1} G\left(\beta'_j X\right)}$$

Considerando que $F(\beta_j{}'X)$ é uma função logística, então

$$G\left(\beta'_j X\right) = \frac{F}{1-F} = \frac{\dfrac{e^{\beta' X}}{1+e^{\beta' X}}}{1 - \dfrac{e^{\beta' X}}{1+e^{\beta' X}}} = \frac{\dfrac{e^{\beta' X}}{1+e^{\beta' X}}}{\dfrac{1}{1+e^{\beta' X}}} = e^{\beta'_j X}$$

$$P_j = \frac{e^{\beta'_j X}}{1 + \sum_{K=1}^{m-1} e^{\beta'_K X}}$$

$$P_m = \frac{1}{1 + \sum_{K=1}^{m-1} e^{\beta'_K X}} = \frac{1}{1 + e^{\beta'_1 X} + e^{\beta'_2 X} + \ldots + e^{\beta'_{m-1} X}}$$

y_{ij} = 1 se o i-ésimo indivíduo se enquadra na j-ésima categoria, então a função de verossimilhança para o modelo lógite multinominal é :

$$L = \prod_{i=1}^{n} P_{i1}^{y_{i1}} \cdot P_{i2}^{y_{i2}} \ldots P_{im}^{y_{im}}$$

$$\log L = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log P_{ij}$$

Para obter a solução:

$$\frac{\partial \ln L}{\partial \beta_K} = \sum_{i=1}^{n} \left(y_{iK} - P_{iK} \right) X_i = 0 \qquad p/ \quad K = 1, \ldots, m$$

Efeito marginal:

$$\frac{\partial P_K}{\partial X_i} = P_K \left[\beta_K - \sum_{j=1}^{m} P_j \beta_j \right]$$

Modelos de Resposta Ordenada

Temos mais de duas alternativas e elas são ordenadas. Por exemplo, o estado de saúde de um paciente é excelente, bom, regular ou ruim. As crianças têm distorção idade-série igual a 0, 1, 2, 3 etc.

Considere um modelo probite ordenado com $J + 1$ alternativas. Considere também a variável latente, y^* tal que:

$$y^* = \mathbf{x}'\boldsymbol{\beta} + e \qquad e|\mathbf{x} \sim N(0,1)$$

Seja $\alpha_1 < \alpha_2 < \cdots < \alpha_J$ pontos de cortes conhecidos (ou parâmetros dos limites) e defina:

$$\begin{aligned}
&y = 0 \quad \text{if} \quad y^* \le \alpha_1, \\
&y = 1 \quad \text{if} \quad \alpha_1 \le y^* \le \alpha_2, \\
&\vdots \\
&y = J \quad \text{if} \quad y^* > \alpha_J.
\end{aligned}$$

A probabilidade de y tomar cada um dos valores é expresso como:

$$P[y=0|\mathbf{x}] = P[y^* \le \alpha_1|\mathbf{x}] = P[\mathbf{x}'\boldsymbol{\beta} + e \le \alpha_1|\mathbf{x}] = \Phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta})$$

$$P[y=1|\mathbf{x}] = P[\alpha_1 \le y^* \le \alpha_2|\mathbf{x}] = \Phi(\alpha_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta})$$

$$\vdots$$

$$P[y=J-1|\mathbf{x}] = P[\alpha_{J-1} \le y^* \le \alpha_J|\mathbf{x}] = \Phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\alpha_{J-1} - \mathbf{x}'\boldsymbol{\beta})$$

$$P[y=J|\mathbf{x}] = P[y^* > \alpha_J|\mathbf{x}] = 1 - \Phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta})$$

Quando J = 1, obtemos o modelo probit.

A função de verossimilhança é:

$$\begin{aligned} lnL = (y_i = 0)\ln[\Phi(\alpha_1 - X'\beta)] \\ + (y_i = 1)\ln[\Phi(\alpha_2 - X'\beta) - \Phi(\alpha_3 - X'\beta)] + \cdots \\ + (y_i = J)\ln\left[1 - \Phi\left(\alpha_J - X'\beta\right)\right] \end{aligned}$$

Os efeitos marginais são:

$$\frac{\partial P[y=0|\mathbf{x}]}{\partial x_k} = -\beta_k \phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta}), \qquad \frac{\partial P[y=J|\mathbf{x}]}{\partial x_k} = \beta_k \phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta})$$

$$\frac{\partial P[y=j|\mathbf{x}]}{\partial x_k} = \beta_k \left[\phi(\alpha_{j-1} - \mathbf{x}'\boldsymbol{\beta}) - \phi(\alpha_j - \mathbf{x}'\boldsymbol{\beta})\right], \qquad 0 < j < J$$

A grande crítica aos modelos logit multinomial e condicional é que as probabilidades relativas de duas escolhas não são afetadas pelas outras escolhas disponíveis. Isso é chamado de independência das alternativas irrelevantes (IIA).

$$\frac{P[y_i = j|\mathbf{x}_i]}{P[y_i = k|\mathbf{x}_i]} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}$$

Adicionando outra alternativa não afeta a relação entre as alternativas j e k

Indivíduos escolhem entre duas alternativas com igual probabilidade, por exemplo, ir de carro para o trabalho com probabilidade 0,5 e ir de ônibus azul com probabilidade 0,5. Suponha que um ônibus vermelho é incluído na escolha. Se os indivíduos forem indiferentes quanto a cor, espera-se que eles escolham entre o ônibus azul e o vermelho com igual probabilidade. Mas, de acordo com IIA, cada probabilidade ficará agora 1/3, inclusive a de ir de carro, o que não é muito realístico.

Amostras Truncadas e Censuradas

Quando dados amostrais são retirados de um subconjunto de uma grande população, dizemos que a amostra é truncada.

Ex: Somente indivíduos acima da linha de pobreza são amostrados. Um estudo sobre renda de toda a população com base na amostra descrita não teria utilidade.

Se ao invés de não observável, os dados dos indivíduos abaixo da linha de pobreza fossem coletados, mas a renda só fosse reportada como sendo igual a linha de pobreza, teríamos uma amostra censurada.

Modelo de regressão truncado.

No modelo de regressão:

$$y_i = \beta' X_i + \varepsilon_i \qquad\qquad \varepsilon_i \sim N(0, \sigma^2)$$

estamos interessados na distribuição de y_i dado que y_i é maior do que um ponto a, isto é, $E[y_i \mid y_i > a]$.

Propriedades do Estimador de Mínimos Quadrados no modelo truncado

$$E\left(y_i \middle| y_i > a\right) = \beta' X_i + E\left(\varepsilon_i \middle| \varepsilon_i > a - \beta' X_i\right)$$

$$E(\varepsilon_i | \varepsilon_i > a - \beta' X_i) = \frac{\int_{a-\beta' X_i}^{\infty} \varepsilon_i \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}\varepsilon^2} d\varepsilon}{1 - \Phi\left(\frac{a - \beta' X_i}{\sigma}\right)} = \frac{\sigma\phi}{1-\Phi}$$

$$\frac{1}{\sqrt{2\pi}} \sigma \int_{a-\beta' X_i}^{\infty} \frac{\varepsilon \cdot e^{-\frac{1}{2\sigma^2}\varepsilon^2}}{\sigma \cdot \sigma} d\varepsilon = \frac{\sigma}{\sqrt{2\pi}} \int \frac{\varepsilon \cdot e^{-\frac{1}{2\sigma^2}\varepsilon^2}}{\sigma^2} d\varepsilon = \left[\frac{-\sigma}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon^2}\right]_{a-\beta' X_i}^{\infty}$$

A derivada de,

$$e^{-\frac{1}{2\sigma^2}\varepsilon^2} = e^{-\frac{1}{2\sigma^2}\varepsilon^2} \cdot -\frac{2\varepsilon}{2\sigma^2}$$

$$= 0 - (-\sigma)\phi\left(\frac{a - \beta' X_i}{\sigma}\right) = \sigma\phi$$

e

$$\Pr\left(\varepsilon_i > a - \beta' X_i\right) = 1 - \Pr(\varepsilon_i < a - \beta' X_i) = 1 - \Phi\left(\frac{a - \beta' X_i}{\sigma}\right)$$

porque a normal é simétrica

Consequentemente, no modelo de regressão, se $y_i > a$

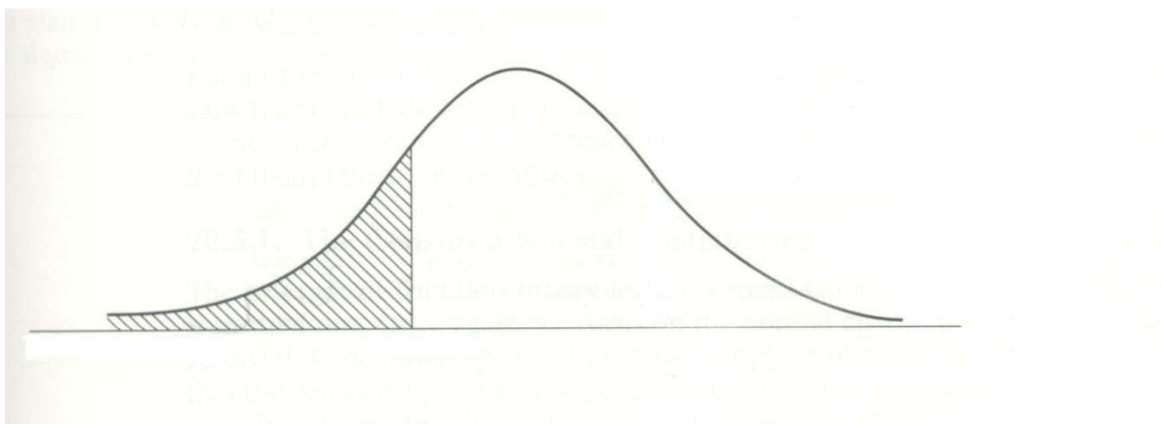
$$y_i \middle| y_i > a = \beta' X_i + \frac{\sigma \phi_i}{1 - \Phi_i} + u_i$$

e aplicando MQO estaríamos omitindo o termo $\frac{\sigma \phi_i}{1 - \Phi_i}$, que não é independente de X_i , causando tendenciosidade e inconsistência das estimativas dos parâmetros.

O fato de a esperança condicional do erro ser positiva, isto é,

$$E\left(\varepsilon_i \middle| \varepsilon_i > a - \beta' X_i\right) = \frac{\sigma\phi_i}{1-\Phi_i} > 0$$

pode ser visualizada no gráfico. Se $\varepsilon_i \sim N(0,\sigma^2)$, sua função de densidade pode ser representada como na figura. O valor médio de ε_i é zero porque a função de densidade é simétrica no ponto 0. Se, entretanto, os valores de ε forem restritos a serem maiores do que $a - \beta' X_i$ então a função de densidade não estaria mais centrada no ponto 0 nem seria simétrica, e a esperança matemática seria positiva.



Estimação por Máxima Verossimilhança.

A função de densidade de y é:

$$f(y_i | y_i > a) = \frac{f(y_i)}{1-\Phi(\alpha)} = \frac{\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2}\left(\frac{y_i - \beta' X_i}{\sigma}\right)^2}}{1-\Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi(\frac{y_i - \beta' X_i}{\sigma})}{1-\Phi(\alpha)}$$

onde

$$\alpha = \frac{a - \beta' X_i}{\sigma}$$

Então, o logaritmo da função de verossimilhança é:

$$\log L = -\frac{n}{2}[\log(2\pi) + \log \sigma^2] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta' X_i)^2 - \sum_{i=1}^{n}\log\left[1 - \Phi\left(\frac{a - \beta' X_i}{\sigma}\right)\right]$$

Efeito Marginal:

$$E\left(y_i \middle| y_i > a\right) = \beta' X_i + \sigma \frac{\phi\left(\dfrac{a-\beta' X_i}{\sigma}\right)}{1-\Phi\left(\dfrac{a-\beta' X_i}{\sigma}\right)}$$

$$\frac{\partial E\left(y_i \middle| y_i > a\right)}{\partial X_i} = \beta + \sigma\left[\frac{-\alpha\phi(1-\Phi)-\phi(-\phi)}{(1-\Phi)^2}\right]\frac{-\beta}{\sigma} \qquad \alpha = \frac{a-\beta' X_i}{\sigma}$$

$$\frac{\partial E\left(y_i \middle| y_i > a\right)}{\partial X_i} = \beta - \beta\left[\frac{-\alpha\phi}{(1-\Phi)} + \frac{\phi^2}{(1-\Phi)^2}\right]$$

$$\frac{\partial E\left(y_i \middle| y_i > a\right)}{\partial X_i} = \beta + \beta\alpha\lambda - \beta\lambda^2 = \beta(1+\alpha\lambda-\lambda^2)$$

onde $\quad \lambda = \dfrac{\phi}{1-\Phi}$

Modelo de regressão censurado ou Modelo Tobite

Ex: gasto domiciliar com bens duráveis

Número de horas trabalhadas no mercado de trabalho pela mulher

Neste caso a variável dependente é censurada.

$$y_i^* = \beta' X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

onde X_i é um vetor de variáveis explanatórias e $\varepsilon_i \sim N(0,\sigma^2)$ e é independente de outros erros.

$$y_i = 0 \quad \text{se} \quad y_i^* \leq 0$$

$$y_i = y_i^* \quad \text{se} \quad y_i^* > 0$$

$$E\left(y_i \middle| X_i\right) = \Pr\left(y_i = 0\right) \cdot E\left(y_i \middle| y_i = 0\right) + \Pr\left(y_i > 0\right) \cdot E\left(y_i \middle| y_i > 0\right)$$

$$= \Pr\left(y_i^* \le 0\right) \cdot 0 + \Pr\left(y_i^* > 0\right) \cdot E\left(y_i^* \middle| y_i^* > 0\right)$$

$$= \Pr\left(\beta' X_i + \varepsilon_i > 0\right) \cdot E\left(\beta' X_i + \varepsilon_i \middle| \beta' X_i + \varepsilon_i > 0\right)$$

$$= \beta' X_i \Pr\left(\beta' X_i + \varepsilon_i > 0\right) + E\left(\varepsilon_i \middle| \beta' X_i + \varepsilon_i > 0\right) \Pr\left(\beta' X_i + \varepsilon_i > 0\right)$$

$$= \beta' X_i \Pr\left(\varepsilon_i > -\beta' X_i\right) + E\left(\varepsilon_i \middle| \varepsilon_i > -\beta' X_i\right) \Pr\left(\varepsilon_i > -\beta' X_i\right)$$

$$= \beta' X_i \cdot \Phi\left(\frac{\beta' X_i}{\sigma}\right) + \frac{\sigma \phi\left(\frac{-\beta' X_i}{\sigma}\right)}{\Phi\left(\frac{\beta' X_i}{\sigma}\right)} \cdot \Phi\left(\frac{\beta' X_i}{\sigma}\right)$$

$$E\left(y_i \middle| X_i\right) = \beta' X_i \cdot \Phi + \sigma \phi \qquad \therefore \quad E\left(y_i \middle| X_i\right) \neq \beta' X_i$$

Estimação por Máxima Verossimilhança.

Para $y_i = 0$

$$f\left(0 \middle| X_i\right) = \Pr\left(y_i = 0 \middle| X_i\right) = \Pr\left(y_i^* \leq 0\right) = \Pr\left(\varepsilon_i < -\beta' X_i\right) = 1 - \Phi\left(\frac{\beta' X_i}{\sigma}\right)$$

Para $y_i > 0$

$$f\left(y_i \middle| X_i\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - \beta' X_i}{\sigma}\right)$$

Então, o logaritmo da função de verossimilhança para a regressão censurada é:

$$\log L = \sum_{y_i > 0} -\frac{1}{2}\left[\log(2\pi) + \log\sigma^2 + \frac{\left(y_i - \beta' X_i\right)^2}{\sigma^2}\right] + \sum_{y_i = 0} \log\left[1 - \Phi\left(\frac{\beta' X_i}{\sigma}\right)\right]$$

Efeito Marginal não condicional:

$$E\left(y_i \middle| X_i\right) = \beta' X_i \cdot \Phi\left(\frac{\beta' X_i}{\sigma}\right) + \sigma\phi\left(\frac{-\beta' X_i}{\sigma}\right)$$

$$\phi\left(\frac{-\beta' X}{\sigma}\right) = \phi(\alpha) \rightarrow \phi'(\alpha) = -\alpha\phi$$

$$\frac{\partial E\left(y_i\right)}{\partial X_i} = \beta\Phi + \beta' X_i \phi \frac{\beta}{\sigma} + \sigma\left(\frac{\beta' X_i}{\sigma}\right)\phi \frac{-\beta}{\sigma}$$

$$\frac{\partial E\left(y_i\right)}{\partial X_i} = \beta\left[\Phi + \phi \frac{\beta' X_i}{\sigma} - \frac{\beta' X_i}{\sigma}\phi\right] = \beta\Phi$$

Modelos com seletividade amostral ou incidentalmente truncado.

Ex: Modelo de oferta de trabalho de mulheres.

Considere duas equações:

1. equação de salários.

A diferença entre o salário de mercado e o salário reserva é função de características, tais como: educação, idade, número de crianças, local de habitação, etc.

2. equação de horas.

O número desejado de horas de trabalho ofertado depende do salário, número de crianças na família, estado civil, etc.

O problema ocorre quando consideramos a segunda equação que descreve o número desejado de horas de trabalho, mas este número só é observado se o indivíduo trabalha, isto é, se o salário de mercado é maior do que o salário reserva. Então a variável horas de trabalho é incidentalmente truncada.

Em notação matemática temos:

Equação que determina a seletividade amostral:

$$Z_i^* = \gamma' W_i + u_i$$

$$Z_i = 1 \quad se \quad Z_i^* > 0$$

$$Z_i = 0 \quad em \ caso \ contrario$$

Equação de interesse:

$$Y_i = \beta' X_i + \varepsilon_i$$

Então, Y_i é observado somente se Z_i^* é maior que zero. Suponha que ε_i e u_i tem uma distribuição normal bivariada com média 0 e correlação ρ. Então,

$$E(Y_i | Y_i \text{ é observado}) = E(Y_i | Z_i^* > 0) = E(Y_i | u_i > -\gamma' W_i)$$

$$= \beta' X_i + E(\varepsilon_i | u_i > -\gamma' W_i) = \beta' X_i + \rho\sigma_\varepsilon \frac{\phi(\alpha_u)}{\Phi(\alpha_u)} = \beta' X_i + \beta_\lambda \lambda_i(\alpha_u)$$

onde

$$\alpha_u = \frac{-\gamma' W_i}{\sigma_u}$$

Então,

$$Y_i \Big| Z_i^* > 0 = E(Y_i \Big| Z_i^* > 0) + v_i = \beta' X_i + \beta_\lambda \lambda_i(\alpha_u) + v_i$$

Fazer a regressão por mínimos quadrados usando apenas os dados observados – por exemplo, MQO da regressão de horas em função de seus determinantes usando apenas dados de mulheres no mercado de trabalho – produz estimativas inconsistentes de β. Este é um caso de omissão de variáveis. A regressão por mínimos quadrados de Y em função de X e λ produziria estimativas consistentes.

Os parâmetros do modelo de seleção podem ser estimados por máxima verossimilhança, mas Heckman (1979) propôs um método mais simples:

1) estimar a primeira equação (modelo próbite – 1 se participa do mercado de trabalho e 0 se não) por máxima verossimilhança e obter as estimativas dos parâmetros γ. Para cada observação calcula-se

$$\hat{\lambda}_i = \frac{\phi(\hat{\gamma}' W_i)}{\Phi(\hat{\gamma}' W_i)}$$

2) Rodar a regressão por MQO de Y em função de X e λ estimado para obter estimativas consistentes dos parâmetros β.

Efeito marginal condicional

$$E\left(Y_i \middle| Z_i^* > 0 \right) = \boldsymbol{\beta}'\mathbf{x}_i + \beta_\lambda \lambda(\alpha_u)$$

$$\frac{\partial E\left(Y_i \middle| Z_i^* > 0 \right)}{\partial X_i} = \beta - \frac{\gamma}{\sigma_u} \rho \sigma_\varepsilon \delta$$

$$\delta = \lambda(\lambda - \alpha)$$

Efeito marginal não condicional

$$E(Y_i) = \left[\beta' x_i + \beta_\lambda \lambda \right] \Phi$$

$$\frac{\partial E(Y_i)}{\partial x}$$

Modelo Próbite Sequencial

crianças

Estuda

(P_1)

outros

trabalho remunerado

(P_2)

outros

trabalho não remunerado

(P_3)

não trabalha

(P_4)

$$P_1 = F(x\beta_1)$$

$$P_2 = [1 - F(x\beta_1)]F(x\beta_2)$$

$$P_3 = [1 - F(x\beta_1)][1 - F(x\beta_2)]F(x\beta_3)$$

$$P_3 = [1 - F(x\beta_1)][1 - F(x\beta_2)][1 - F(x\beta_3)]$$

Segundo Maddala esses modelos são fáceis de se analisar porque a função de verossimilhança para estes modelos pode ser maximizada, maximizando funções de verossimilhança de modelos dicótomos repetitivamente.