ECONOMETRIA

Análise de Regressão com Dados Seccionais

Teoria e Exemplos

Esmeralda A. Ramalho

Tópicos do programa

1ª Parte: Análise de Regressão com Dados Seccionais — Continuação

1. Informação Qualitativa: Variáveis Explicativas Binárias ("Dummy")
2. Modelos para Variáveis Dependentes Binárias

2ª Parte: Análise de Regressão com Séries Temporais

3. Análise de Regressão Básica
4. Tópicos Adicionais sobre a Utilização do OLS
5. Autocorrelação e Heteroscedasticidade

3ª Parte: Dados de Painel

6. Introdução aos Métodos para Dados de Painel

Tipos de Dados em Análise

Dados seccionais:

- N indivíduos
- 1 observação por indivíduo

Dados temporais:

- 1 indivíduo
- T observações por indivíduo

Dados de painel

- N indivíduos
- T observações por indivíduo

Bibliografia

Básica

— Wooldridge, J. M. (2016) (W), Introductory Econometrics: a Modern Approach, 6th. ed., Cengage Learning.

Complementar

— Stock, J. H. e Watson, M. W. (2015), Introduction to Econometrics, 3rd ed., Pearson.

Avaliação

Avaliação em EN: envolve 2 provas individuais (PI1 e PI2), cada uma com ponderação de 50%.

Provas escritas individuais: serão realizadas em dois momentos, na semana de paragem das aulas e no final do semestre. Na primeira prova será avaliada a matéria respeitante às aulas que decorridas até à sua data e terá a duração de 1h. Na segunda prova, também com duração de 1 h, será avaliada a matéria respeitante às restantes aulas. Um aluno que obtenha uma nota inferior a 6 numa destas duas provas, será considerado reprovado. Note-se que este sistema requer obrigatoriamente a realização de ambas a PI1 e PI2.

Avaliação em ER: a classificação final será igual à obtida no exame de recurso.

Descrição de informação qualitativa

Até ao momento, as variáveis explicativas tinham uma natureza quantitativa. Este tópico introduz variáveis explicativas qualitativas, que basicamente incorporam no modelo informação qualitativa:

- Preco de imóveis = f(área, nº quartos, qualidade da localização, tipo de imóvel)
- Salario = f(idade, experiencia profissional, tipo de profissão, género, região local trabalho)

Descrição de informação qualitativa

Este tipo de variáveis, designadas de variáveis dummy, têm natureza binária $X_{i1} = \{0,1\}$, sendo o seu efeito parcial sobre a variável de interesse interpretado de forma diferente relativamente às variáveis quantitativas

O modelo de regressão, a estimação e a inferência são os habituais

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i \qquad (i = 1, \cdots, N)$$

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Uma única variável dummy

Uma variável dummy designada de d_i, define-se como

$$d_i = \begin{cases} 1\ se\ atributo\ observado\ no\ individuo\ i \\ 0\ se\ atributo\ não\ observado\ no\ individuo\ i \end{cases}$$

Esta variável permite distinguir **duas categorias**, por exemplo

$$d_i = \begin{cases} 1\ se\ i\ mulher \\ 0\ se\ i\ homem \end{cases}$$

A interpretação do efeito sobre o valor esperado, ceteris paribus, é feita por referência ao atributo omitido: no exemplo, o ser homem.

Uma única variável dummy

Exemplo:

$$salario_i = \beta_0 + \delta_0 fem_i + \beta_1 educ_i + \cdots + u_i,$$

onde *fem*=1 para género feminino e *fem*=0 para género masculino

Ceteris paribus

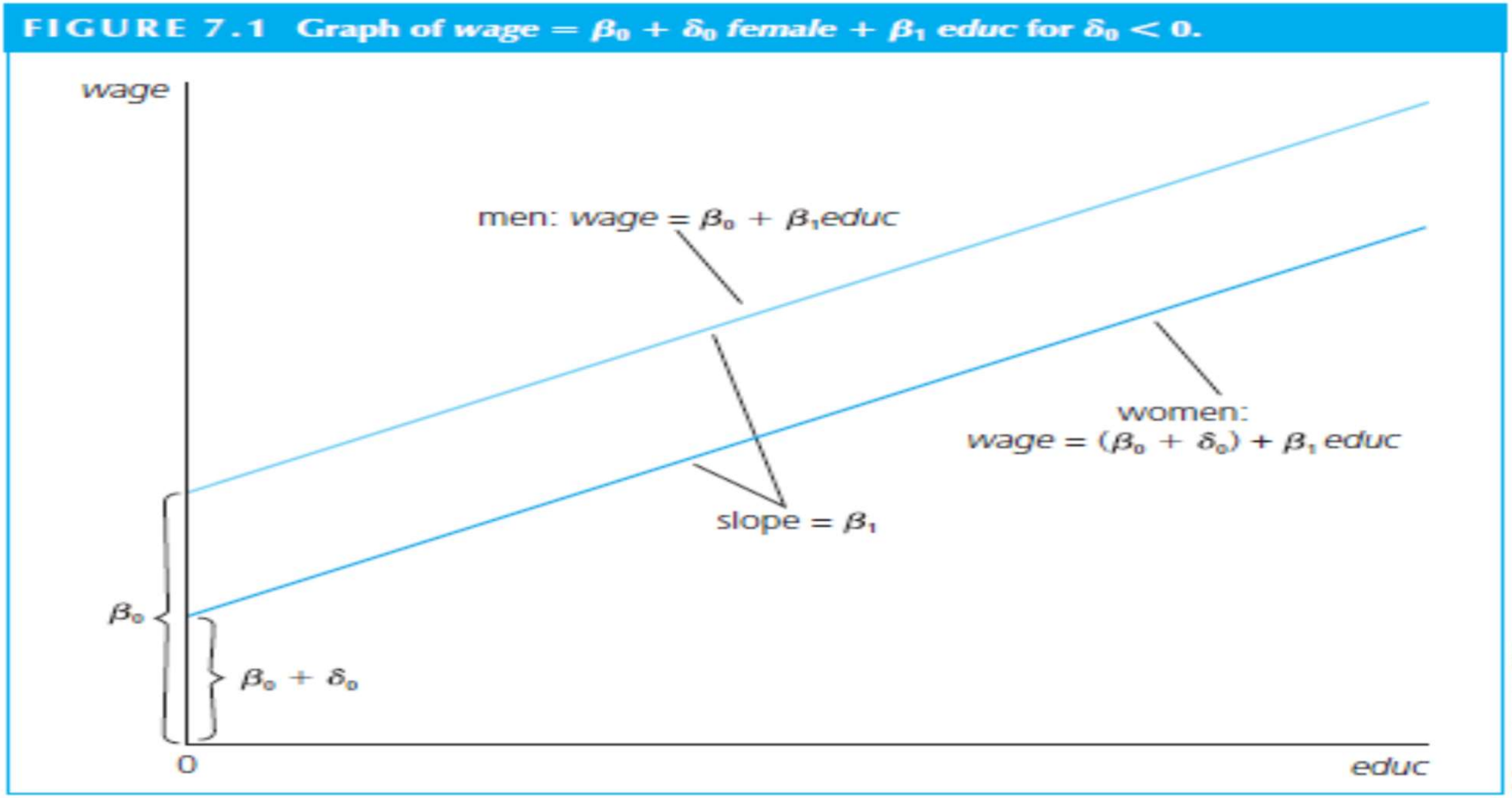
δ_0: diferença média no salário de uma mulher relativamente ao de um homem

A introdução destas *dummy* altera o intercepto do modelo:

	Homens	Mulheres
Intercepto	β_0	$\beta_0 + \delta_0$

Uma única variável dummy

Wooldridge (2016): ilustração



Uma única variável dummy

Exemplo (uma dummy para dois tipos de categorias):

$$salario_i = \beta_0 + \beta_1 fem_i + \beta_2 norte_i + \beta_3 educ_i + u_i,$$

onde *fem*=1 para género feminino e *fem*=0 para género masculino, *norte* =1 para região de residencia Norte e *norte*=0 outra região.

Ceteris paribus

β_1 : diferença no salário de uma mulher relativamente ao de um homem;

β_2: diferença no salário de um habitante do Norte relativamente a um habitante de outras regiões.

Mais uma vez, a introdução destas *dummies* altera o intercepto do modelo:

	Homens	**Mulheres**
Norte	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2$
Outra região	β_0	$\beta_0 + \beta_1$

Uma única variável dummy

Ainda a interpretação:

Os modelos considerados como exemplo são de tipo lin-lin

Pode acontecer ter-se um modelo de tipo log-lin com a dummy:

$$lnsalario_i = \beta_0 + \delta_0 fem_i + \beta_1 educ_i \ldots + u_i,$$

Neste caso, aplicando a interpretação habitual em casos log-lin

- δ_0*100%: diferença no do salário de uma mulher relativamente ao de um homem
- Contudo, esta interpretação é uma aproximação, que pode ser muito distante do efeito real para valores grandes de δ_0. O efeito exacto é dado por

$$[exp(\delta_0) - 1]*100\%:$$

Variáveis dummy para várias categorias

A inclusão no modelo de informação de M categorias requer a introdução de M-1 dummies. Por exemplo, para três categorias (Norte, Centro, Sul):

$$d_1 = \begin{cases} 1\ se\ atributo\ 1\ observado\ (ex{:}\ região\ Norte) \\ 0\ se\ atributo\ não\ observado\ (ex{:}\ outras\ regiões) \end{cases}$$

$$d_2 = \begin{cases} 1\ se\ atributo\ 2\ observado\ (ex{:}\ região\ Centro) \\ 0\ se\ atributo\ não\ observado\ (ex{:}\ outras\ regiões) \end{cases}$$

Mais uma vez, a interpretação do efeito sobre o valor esperado é feita por referência ao atributo omitido: o pertencer à região Sul

Variáveis dummy para várias categorias

Exemplo:

$$Y_i = \beta_0 + \beta_1 fem_i + \beta_2 norte_i + \beta_3 centro_i + \beta_4 educ_i + u_i$$

β_1: diferença em Y de uma mulher relativamente a um homem;

β_2: diferença em Y de um individuo que vive no Norte, relativamente a um individuo que vive no Sul

β_3: diferença em Y de um individuo que vive no Centro, relativamente a um individuo que vive no Sul

Novamente, a introdução destas *dummies* altera o intercepto

	Homens	**Mulheres**
Norte	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2$
Centro	$\beta_0 + \beta_3$	$\beta_0 + \beta_1 + \beta_3$
Sul	β_0	$\beta_0 + \beta_1$

Interacções envolvendo variáveis dummy

Variável de interacção: resulta da multiplicação de uma dummy por outra variável

Exemplo:

$$salario_i = \beta_0 + \delta_0 fem_i + \beta_1 educ_i + \delta_1 fem_i * educ_i + u_i$$

Escrevendo o modelo para cada grupo:

- mulheres (fem=1): $salario_i = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) educ_i + u_i$
- homens (fem=0): $salario_i = \beta_0 + \beta_1 educ_i + u_i$

Logo:

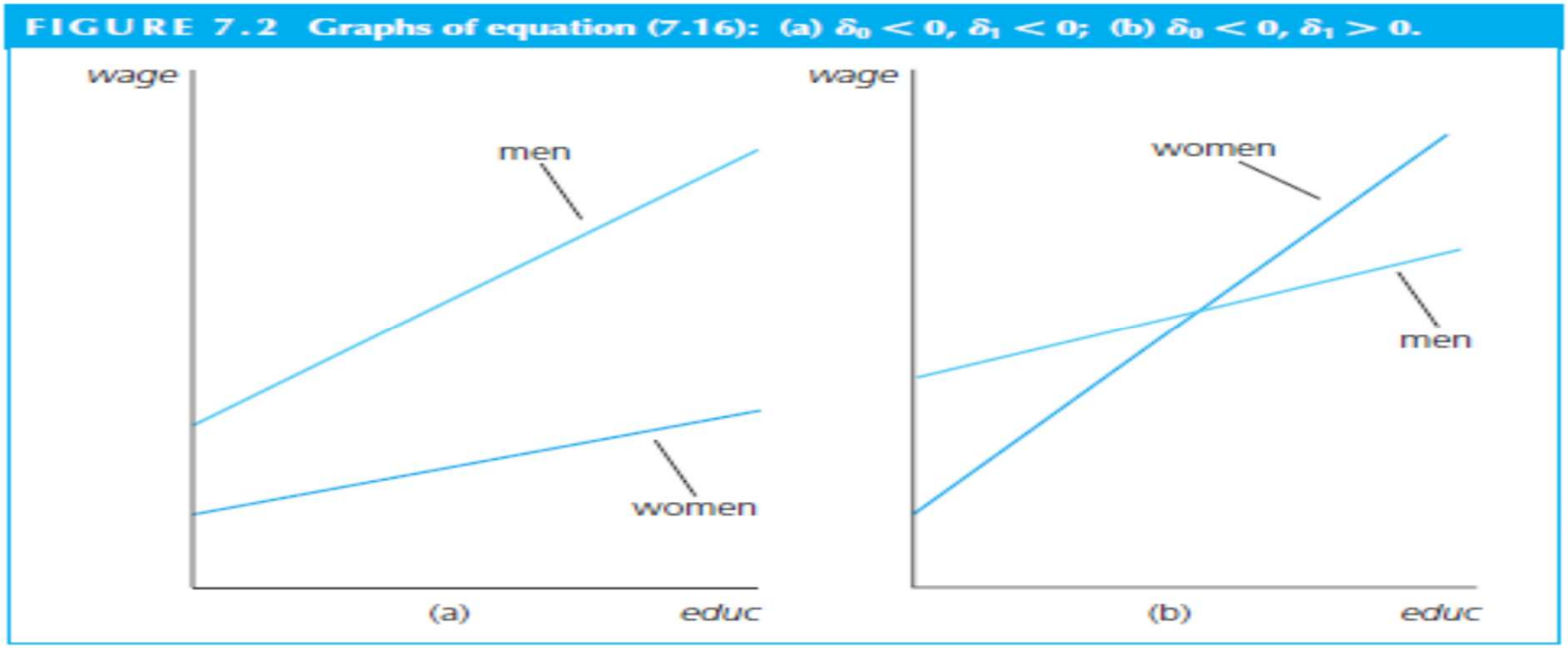
- β_0: termo independente relativo aos homens
- δ_0: diferença no termo independente entre mulheres e homens
- β_1: acréscimo que se verifica no salário dos homens por cada ano de escolaridade adicional
- δ_1: diferença que se verifica no efeito anterior entre mulheres e homens

Interacções envolvendo variáveis dummy

Síntese

	Homens	Mulheres
Efeito de educ	β_1	$\beta_1 + \delta_1$
Intercepto	β_0	$\beta_0 + \delta_0$

Wooldridge (2016): ilustração



Interacções envolvendo variáveis dummy

Teste Chow para quebra de estrutura

Contexto:

- Dois grupos de indivíduos / empresas: G_A, G_B (homens/mulheres, …)
- Suspeita-se que as variáveis explicativas influenciem de forma diferente os dois grupos

Implementação 1:

- Considerar a variável binária

$$D = \begin{cases} 1 & \text{se a observação pertence a } G_A \\ 0 & \text{se a observação não pertence a } G_A \end{cases}$$

- Estimar o modelo original de forma 'duplicada':

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_k X_k + \gamma_0 D + \gamma_1 D X_1 + \cdots + \gamma_k D X_k + v, R^2$$

- Aplicar um teste F para a significância das variáveis envolvendo D:

$H_0: \gamma_0 = \gamma_1 = \cdots = \gamma_k = 0$ (sem quebra de estrutura)

H_1: Não H_0 (com quebra de estrutura)

$$F = \frac{(R^2 - R_*^2)/(k+1)}{(1-R^2)/(N-K-1)} \sim F\big((k+1), N-k-1\big)$$

Interacções envolvendo variáveis dummy

Implementação 2:

- Considerar a variável binária

$$D = \begin{cases} 1 & \text{se a observação pertence a } G_A \\ 0 & \text{se a observação não pertence a } G_A \end{cases}$$

- Estimar três modelos, um para cada grupo e um para ambos os grupos:

$$G_A\colon Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k + v, com\ SQR_A$$
$$G_B\colon Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \mathrm{e}, com\ SQR_B$$
$$Y = \rho_0 + \rho_1 X_1 + \cdots + \rho_k X_k + \mathrm{w}, com\ SQR$$

Note-se que $\theta_0 + \gamma_0 = \alpha_0, \ldots \theta_\mathrm{k} + \gamma_\mathrm{k} = \alpha_k$

- Aplicar um teste F, mas de versão diferente do anterior, para a significância das variáveis envolvendo D:

$H_0\colon \alpha_0 = \beta_0, \ldots, \alpha_\mathrm{k} = \beta_\mathrm{k}$ (sem quebra de estrutura)

H_1: Não H_0 (com quebra de estrutura)

$$F = \frac{(SQR - SQR_A - SQR_B)/(k+1)}{(SQR_A + SQR_B)/\big(N - 2(k+1)\big)} \sim F\big((k+1), N - 2(k+1)\big)$$

Interacções envolvendo variáveis dummy

Exemplo: Quebra de estrutura - casas de dimensão grande (lote>600) e pequena

$$ln(preco_i) = \beta_0 + \beta_1 ln(area_i) + \beta_1 quartos_i + u_i$$

```
. gen dim=lote>600
. gen larea=log(area)
. gen lpreço=log(preco)
```

Modelo baseado em toda a amostra

```
. reg  lpreço larea quartos
      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(2, 85)        =     54.47
       Model |  4.50364223         2  2.25182112   Prob > F        =    0.0000
    Residual |  3.51396129        85  .041340721   R-squared       =    0.5617
-------------+----------------------------------   Adj R-squared   =    0.5514
       Total |  8.01760352        87  .092156362   Root MSE        =    .20332
------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .8100637   .0987611     8.20   0.000     .6137002    1.006427
     quartos |   .0376464   .0303446     1.24   0.218    -.0226868    .0979795
       _cons |    1.28929   .4666125     2.76   0.007     .3615395    2.217041
------------------------------------------------------------------------------
```

Interacções envolvendo variáveis dummy

Modelo baseado na subamostra de casas grandes

```
. reg  lpreço larea quartos if dim==1

      Source |       SS           df       MS      Number of obs   =        43
-------------+----------------------------------   F(2, 40)        =     50.60
       Model |  2.65666194         2  1.32833097   Prob > F        =    0.0000
    Residual |   1.0500666        40  .026251665   R-squared       =    0.7167
-------------+----------------------------------   Adj R-squared   =    0.7025
       Total |  3.70672854        42  .088255441   Root MSE        =    .16202

------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .7325533   .1151391     6.36   0.000     .4998486     .965258
     quartos |   .0664725   .0327609     2.03   0.049     .0002603    .1326848
       _cons |   1.651792   .5435154     3.04   0.004     .5533069    2.750278
------------------------------------------------------------------------------
```

Interacções envolvendo variáveis dummy

Modelo baseado na subamostra de casas pequenas

```
. reg  lpreço larea quartos if dim==0

      Source |       SS           df       MS      Number of obs   =        45
-------------+----------------------------------   F(2, 42)        =      3.47
       Model |  .316698909         2  .158349455   Prob > F        =    0.0404
    Residual |   1.9189645        42  .045689631   R-squared       =    0.1417
-------------+----------------------------------   Adj R-squared   =    0.1008
       Total |   2.2356634        44  .050810532   Root MSE        =    .21375

------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .4826644   .1841558     2.62   0.012     .1110231    .8543058
     quartos |  -.0150615   .0501592    -0.30   0.765    -.1162869    .0861638
       _cons |   3.079526   .9171463     3.36   0.002      1.22865    4.930402
------------------------------------------------------------------------------
```

$$F = \frac{(3.5140 - 1.0501 - 1.9190)/3}{(1.0501 + 1.9190)/(88 - 6)} = 5.02$$

Rejeita-se Ho de ausência de quebra de estrutura: deverá ser considerado um modelo separado para cada grupo.

Interacções envolvendo variáveis dummy

Modelo com interacções que resume os modelos separados para casas grandes e pequenas

```
. gen dlarea= dim*larea
. gen dquartos= dim*quartos
. reg  lpreço larea quartos dim dlarea dquartos
```

```
      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(5, 82)        =     27.89
       Model |  5.04857242         5  1.00971448   Prob > F        =    0.0000
    Residual |   2.9690311        82  .036207696   R-squared       =    0.6297
-------------+----------------------------------   Adj R-squared   =    0.6071
       Total |  8.01760352        87  .092156362   Root MSE        =    .19028
------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .4826644    .163937     2.94   0.004     .1565416    .8087873
     quartos |  -.0150615   .0446521    -0.34   0.737    -.1038888    .0737658
         dim |  -1.427734   1.036357    -1.38   0.172    -3.489378    .6339111
      dlarea |   .2498889   .2125091     1.18   0.243    -.1728593    .6726371
    dquartos |   .0815341   .0589418     1.38   0.170    -.0357199     .198788
       _cons |   3.079526   .8164513     3.77   0.000     1.455344    4.703708
------------------------------------------------------------------------------
```

$$F = \frac{(0.6297 - 0.5617)/3}{(1 - 0.6297)/(88 - 6)} = 5.02$$

Interacções envolvendo variáveis dummy

Por vezes pode interessar testar a alteração no efeito parcial de uma variável especifica, X_j, de acordo com o G_A e G_B, acrescentando-se apenas essa interação

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_k X_k + \gamma D X_j + v, R^2$$

$H_0: \gamma = 0$ (sem alteração do efeito parcial de X_j entre G_A e G_B)

H_1: Não H_0

$$t \sim t(N - k - 1)$$

Pode ainda interessar testar a alteração no efeito parcial de uma variável especifica, X_j, de acordo com 3 ou mais grupos. Por exemplo em G_A G_B e G_C

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_k X_k + \gamma_1 D_A X_j + \gamma_2 D_B X_j + v, R^2$$

$H_0: \gamma_1 = \gamma_2 = 0$ (sem alteração do efeito parcial de X_j)

H_1: Não H_0

$$F = \frac{(R^2 - R_*^2)/m}{(1 - R^2)/(N - K - 1)} \sim F(m, N - k - 1)$$

Modelos para Variáveis Dependentes Binárias

Variável dependente binária define-se como

$$y_i = \begin{cases} 1 \; se \; atributo \; observado \; no \; individuo \; i \\ 0 \; se \; atributo \; não \; observado \; no \; individuo \; i \end{cases}$$

Exemplos:

- Um adulto detém ($y_i = 1$) ou não ($y_i = 0$) licenciatura
- Um agregado obtém ($y_i = 1$) ou não ($y_i = 0$) crédito bancário
- Uma empresa abre falência ($y_i = 1$) ou não ($y_i = 1$)

Modelos e respectivo método de estimação

- probabilístico linear → método dos mínimos quadrados
- logit e probit → método da máxima verosimilhança

Modelo probabilístico linear

Especificação

Considera-se o modelo linear típico:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i \qquad (i = 1, \cdots, N)$$

$$E(\mathrm{Y}|\mathrm{X}) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_k X_k$$

Contudo, neste caso específico, como $Y_i \in \{0,1\}$

$$E(Y|X) = Pr(Y = 0|X)0 + Pr(Y = 1|X)1 = Pr(Y = 1|X)$$

implicando que

- $Pr(Y = 1|X) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_k X_k$
- Os efeitos parciais são interpretados em termos da probabilidade de sucesso: $\Delta X_j = 1 \Longrightarrow \Delta E(Y|X) = \Delta Pr(Y = 1|X) = \beta_j$

Modelo probabilístico linear

Interpretação e estimação

Interpretação dos parâmetros da regressão

- Interpretação habitual do modelo de regressão linear não faz sentido:
  - $\Delta X_j = 1 \Rightarrow \Delta y = \beta_j$
  - Δy apenas pode ser igual a 0, 1 ou -1 mas β_j pode assumir qualquer valor
- Interpretação possível:

$$\Delta X_j = 1 \Rightarrow \Delta P\left(y = 1|X\right) = \beta_j$$

Estimação

- Estimação pelo método dos mínimos quadrados
- Variável dependente estimada:

$$\widehat{P\left(y = 1|X\right)} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

Previsão de Y: $\hat{y} = 1 \; se \; Pr(\widehat{y = 1}|x) \geq 0.5$

Modelo probabilístico linear

Limitações do MLP

Problema #1

- Por definição, $0 \leq P(y = 1|X) \leq 1$, mas nada garante que $0 \leq \widehat{P(y = 1|X)} \leq 1$
- Problema impossível de resolver no âmbito do modelo probabilístico linear

Problema #2

- O modelo linear implica a existência de efeitos parciais constantes, o que não é razoável assumir quando a variável dependente é limitada

Modelo probabilístico linear

Problema #3

- O modelo probabilístico linear é sempre heteroscedástico pois $Var\,(u|X) = X\beta\,(1 - X\beta)$, o que implica que:
  - Os estimadores dos mínimos quadrados, embora consistentes, não são eficientes
  - A inferência é inválida se realizada com base em fórmulas que assumem homoscedasticidade para o cálculo da variância dos estimadores
- Problema fácil de resolver usando os métodos habituais:
  - Método dos mínimos quadrados robustos: calcular a variância com base em fórmulas robustas à heteroscedasticidade $\longrightarrow$ inferência válida mas estimadores não eficientes
  - Método dos mínimos quadrados ponderados $\longrightarrow$ inferência válida e estimadores eficientes:

$$\frac{y}{\sqrt{X\beta\,(1 - X\beta)}} = \beta_0 \frac{1}{\sqrt{X\beta\,(1 - X\beta)}} + \beta_1 \frac{X_1}{\sqrt{X\beta\,(1 - X\beta)}} + \ldots + u^*$$

Modelo probabilístico linear

Exemplo: Loan application

- Variável dependente: *aprove* (=1 se obteve crédito à habitação)
- Variáveis explicativas: *hrat* (peso da prestação no rendimento mensal em %), *married* (=1 se casado), *dep* (número de dependentes do agregado familiar)

```
. reg approve married dep hrat
      Source |       SS           df       MS      Number of obs   =     1,986
-------------+----------------------------------   F(3, 1982)      =      6.86
       Model |  2.19826059         3   .73275353   Prob > F        =    0.0001
    Residual |  211.823894     1,982  .106873812   R-squared       =    0.0103
-------------+----------------------------------   Adj R-squared   =    0.0088
       Total |  214.022155     1,985  .107819725   Root MSE        =    .32692
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0526323   .0165946     3.17   0.002     .0200876     .085177
         dep |  -.0152046   .0071213    -2.14   0.033    -.0291707   -.0012386
        hrat |  -.0029698   .0010333    -2.87   0.004    -.0049962   -.0009434
       _cons |   .9278138     .02892    32.08   0.000     .8710969    .9845306
------------------------------------------------------------------------------
```

Modelo probabilístico linear

$$\widehat{Pr(approve} = 1|X\ ) = 0.928 + 0.053married - 0.015dep - 0.003\ hrat$$

Ceteris paribus:

- Um individuo casado, tem uma probabilidade de aprovação de credito superior em 0.053 relativamente a um individuo não casado
- Por cada dependente adicional no agregado, a probabilidade de ter o credito aprovado decai 0.015
- Quando o peso da prestação mensal aumenta em 1 pp., a probabilidade de ter o crédito aprovado reduz-se em 0.003

Modelo probabilístico linear

- Versão robusta à heteroscedasticidade

```
. reg approve married dep hrat, robust

Linear regression                               Number of obs     =      1,986
                                                F(3, 1982)        =       5.53
                                                Prob > F          =     0.0009
                                                R-squared         =     0.0103
                                                Root MSE          =     .32692
------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0526323   .0174539     3.02   0.003     .0184025    .0868621
         dep |  -.0152046   .0074779    -2.03   0.042      -.02987   -.0005393
        hrat |  -.0029698   .0012641    -2.35   0.019     -.005449   -.0004906
       _cons |   .9278138   .0336977    27.53   0.000     .8617271    .9939004
------------------------------------------------------------------------------
```

Os regressores são individualmente e conjuntamente significativos a 5% de significância

Modelo probabilístico linear

- Modelos separados para casados e não casados

```
reg approve dep hrat if married==1, robust
Linear regression                               Number of obs     =      1,308
                                                F(2, 1305)        =       2.26
                                                Prob > F          =     0.1043
                                                R-squared         =     0.0049
                                                Root MSE          =     .31068
------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dep |  -.0082572    .007336    -1.13   0.261    -.0226488    .0061345
        hrat |  -.0027316    .001595    -1.71   0.087    -.0058607    .0003975
       _cons |   .9672602   .0393522    24.58   0.000     .8900598    1.044461
------------------------------------------------------------------------------
```

Modelo probabilístico linear

```
. reg approve dep hrat if married==0, robust

Linear regression                               Number of obs     =        678
                                                F(2, 675)         =       3.87
                                                Prob > F          =     0.0214
                                                R-squared         =     0.0177
                                                Root MSE          =     .35512
------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dep |  -.0635972    .028055    -2.27   0.024    -.1186827   -.0085116
        hrat |  -.0037526    .002132    -1.76   0.079    -.0079388    .0004336
       _cons |     .95826   .0542735    17.66   0.000     .8516948    1.064825
------------------------------------------------------------------------------
```

Modelo probabilístico linear

- Modelo que permite intercepto e efeitos parciais diferenciados (base do teste de Chow)

```
. gen depmarried=dep*married
. gen hratmarried=hrat*married
. reg approve married dep hrat depmarried hratmarried, robust

Linear regression                               Number of obs     =      1,986
                                                F(5, 1980)        =       3.57
                                                Prob > F          =     0.0032
                                                R-squared         =     0.0137
                                                Root MSE          =     .32651
------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0090002   .0670164     0.13   0.893    -.1224297    .1404302
         dep |  -.0635972   .0280352    -2.27   0.023    -.1185788   -.0086155
        hrat |  -.0037526   .0021305    -1.76   0.078    -.0079309    .0004258
  depmarried |     .05534   .0289798     1.91   0.056    -.0014942    .1121742
 hratmarried |   .0010209   .0026618     0.38   0.701    -.0041993    .0062411
       _cons |     .95826   .0542353    17.67   0.000     .8518958    1.064624
------------------------------------------------------------------------------
```

Modelo probabilístico linear

```
. reg approve dep hrat, robust

Linear regression                               Number of obs     =      1,986
                                                F(2, 1983)        =       3.80
                                                Prob > F          =     0.0226
                                                R-squared         =     0.0052
                                                Root MSE          =     .32766

------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dep |  -.0070851    .006939    -1.02   0.307    -.0206936    .0065234
        hrat |  -.0031418   .0012585    -2.50   0.013    -.0056099   -.0006736
       _cons |   .9604815    .031376    30.61   0.000     .8989481    1.022015
------------------------------------------------------------------------------

. display ((0.0137-0.0052)/3)/((1-0.0137)/(1986-6))
5.6879246
```

Modelo probabilístico linear

- Modelo que permite que efeito parcial de *dep* seja diferenciado

```
 reg approve dep hrat depmarried, robust

Linear regression                               Number of obs     =      1,986
                                                F(3, 1982)        =       4.48
                                                Prob > F          =     0.0039
                                                R-squared         =     0.0118
                                                Root MSE          =     .32667

------------------------------------------------------------------------------
             |               Robust
     approve |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dep |  -.0724662   .0270075    -2.68   0.007    -.1254323   -.0195001
        hrat |   -.003258   .0012708    -2.56   0.010    -.0057502   -.0007658
  depmarried |   .0709102   .0270861     2.62   0.009       .01779    .1240304
       _cons |   .9644555   .0317553    30.37   0.000     .9021782    1.026733
------------------------------------------------------------------------------
```

- Efeito de *dep* para não casados: -0,072
- Efeito de *dep* para casados: -0,072+0,071=-0,001
- A 5% de significância, pode-se dizer que o efeito parcial difere nos dois grupos em análise (p-value 0.009)

Modelos logit e probit e estimação por MV

- A variável dependente define-se como:

$$Y_i = \begin{cases} 1 & \text{com probabilidade } Pr(Y_i = 1|x_i) = G(x_i'\beta) \\ 0 & \text{com probabilidade } 1 - G(x_i'\beta) \end{cases}$$

onde $G(x_i'\beta)$ é uma função não linear, $0 \leq G(x_i'\beta) \leq 1$

- $Y_i|X \sim B[1, G(x_i'\beta)]$:
  - $E(Y|X) = G(x_i'\beta)$
  - $V(Y|X) = G(x_i'\beta)\big(1 - G(x_i'\beta)\big)$
  - $\mathrm{f}(Y_i|X) = G(x_i'\beta)^{y_i}[1 - G(x_i'\beta)]^{1-y_i}$

Modelos logit e probit

Especificação

- A variável dependente define-se como:

$$Y_i = \begin{cases} 1 & \text{com probabilidade } Pr(Y_i = 1|x_i) = G(x_i'\beta) \\ 0 & \text{com probabilidade } 1 - G(x_i'\beta) \end{cases}$$

onde $0 \leq G(x_i'\beta) \leq 1$, com

- Probit: $G(x_i'\beta) = \Phi(x_i'\beta) = \int_{-\infty}^{x\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(x_i'\beta\right)^2}{2}} dx\beta$
- Logit: $G(x_i'\beta) = \Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$

Modelos logit e probit

Efeitos parciais

Os coeficientes $\boldsymbol{\beta}$ não têm interpretação directa, apenas o seu sinal e significância tem significado. Note-se também que os **coeficientes $\boldsymbol{\beta}$** não são comparáveis entre modelos (ex: $\beta_{logit} \cong 1{,}6\beta_{probit}$)

Os efeitos parciais, ceteris paribus, são dados por

- $\Delta X_j = 1 \Longrightarrow \Delta E(Y|X) = \Delta Pr(Y = 1|X) = \beta_j \frac{\partial G(x_i'\beta)}{\partial x_i'\beta} = \beta_j g(x_i'\beta)$
  - Probit: $g(x_i'\beta) = \phi(x_i'\beta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i'\beta)^2}{2}}$
  - Logit: $g(x_i'\beta) = \lambda(x_i'\beta) = \Lambda(x_i'\beta)[1 - \Lambda(x_i'\beta)] = \frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})^2}$

Modelos logit e probit

- Se X_j for discreta ou a variação ΔX_j for grande, quando se passa de $X_j = c$ para $X_j = c + 1$

  $\Delta Pr(Y = 1|X) = G\left(\beta_0 + \cdots \beta_j(c + 1) + \cdots\right) - G\left(\beta_0 + \cdots \beta_j c + \cdots\right)$

- Avaliação:

  - média dos regressores: efeito parcial avaliado na média
  - para todos os indivíduos e depois feita a média: efeito parcial médio

Modelos logit e probit e estimação por MV

O logit e o probit são modelos não lineares, estimando-se pelo Método da Máxima Verosimilhança, baseado na função

$$L = \prod_{i=1}^{N} G(x_i'\beta)^{y_i}[1 - G(x_i'\beta)]^{1-y_i}$$

de onde

$$LL = \sum_{i=1}^{N} \{y_i \ln[G(x_i'\beta)] + (1 - y_i)\ln[1 - G(x_i'\beta)]\}$$

Os estimadores destes modelos têm as propriedades usuais dos estimadores MV:

- consistentes: $plim(\hat{\beta}) = \beta$
- assimptoticamente normais
- eficientes

Principais testes

No âmbito da máxima verosimilhança existem três tipos de testes principais:

1. Razão de Verosimilhanças (LR)
2. Wald
3. Score / LM

Os três testes apenas são válidos assimptoticamente e são todos equivalentes nessa situação

Modelos logit e probit : inferência no âmbito do método da MV

Inferência (principais testes)

1. Razão de verosimilhanças (LR): significância conjunta

H_0: $\beta_{g+1} = \cdots = \beta_k = 0$ $[P(Y|x) = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g) \rightarrow LL_*]$

H_1: Não H_0 $[P(Y|x) = G(\beta_0 + \beta_1 x_1 + \cdots \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k) \rightarrow LL]$

- Estatística de teste:

$$LR = 2(LL - LL_*) \sim \chi^2_{k-g}$$

  - Requer a estimação de dois modelos: o restrito e o não restrito

- Caso particular:

H_0: $\beta_1 = \cdots = \beta_k = 0$ $[P(Y|x) = F(\beta_0) \rightarrow LL_*]$

H_1: Não H_0 $[P(Y|x) = G(\beta_0 + \beta_1 x_1 + \cdots \beta_g x_k) \rightarrow LL]$

$$LR = 2(LL - LL_0) \sim \chi^2_k$$

Modelos logit e probit : inferência no âmbito do método da MV

2. Teste de Wald: significância individual

$$H_0: \beta_j = 0$$

- Estatística de teste:

$$W = \frac{\hat{\beta}_j^{\,2}}{\hat{\sigma}_{\hat{\beta}_j}^{\,2}} \sim \chi_1^2$$

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim N(0,1)$$

  - Requer apenas a estimação do modelo não restrito

Modelos logit e probit

- Para escolher o melhor modelo é comum calcular a percentagem de previsões correctas produzida por cada modelo:

	$Y_i = 1$	$Y_i = 0$	Total
$\hat{Y}_i = 1$	n_{11}		
$\hat{Y}_i = 0$		n_{00}	
Total	n_1	n_0	n

  - % previsões correctas: $(n_{11} + n_{00})/n$
  - % 1's correctamente previstos: n_{11}/n_1
  - % 0's correctamente previstos: n_{00}/n_0

Recorde-se que $\hat{y} = 1 \; se \; Pr(\widehat{y = 1}|x) \geq 0.5$

Modelos logit e probit

Exemplo: Loan application

```
. logit approve married dep hrat
Logistic regression                             Number of obs     =      1,986
                                                LR chi2(3)        =      20.14
                                                Prob > chi2       =     0.0002
Log likelihood = -729.88535                     Pseudo R2         =     0.0136
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .4848471   .1534378     3.16   0.002     .1841145    .7855796
         dep |  -.1410653   .0642852    -2.19   0.028     -.267062   -.0150686
       hrat |  -.0267698   .0094215    -2.84   0.004    -.0452355    -.008304
       _cons |   2.453316   .2693916     9.11   0.000     1.925318    2.981314
------------------------------------------------------------------------------
. probit approve married dep hrat
Probit regression                               Number of obs     =      1,986
                                                LR chi2(3)        =      19.52
                                                Prob > chi2       =     0.0002
Log likelihood = -730.19351                     Pseudo R2         =     0.0132
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |    .255545   .0812719     3.14   0.002     .0962551     .414835
         dep |  -.0747856   .0345529    -2.16   0.030    -.1425081   -.0070632
       hrat |  -.0134307    .004899    -2.74   0.006    -.0230325   -.0038289
       _cons |   1.397252   .1395427    10.01   0.000     1.123753    1.670751
------------------------------------------------------------------------------
```

Modelos logit e probit

Teste LR para a significância conjunta

```
. logit approve
Logistic regression                             Number of obs     =      1,986
                                                LR chi2(0)        =      -0.00
                                                Prob > chi2       =          .
Log likelihood = -739.95364                     Pseudo R2         =    -0.0000
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    1.965621   .0683551    28.76   0.000     1.831647    2.099594
------------------------------------------------------------------------------
```

- $LR = 2(-729.885 + 739.954) = 20.14$, a 5% $\chi^2_3 = 7.814$

```
. probit approve
Probit regression                               Number of obs     =      1,986
                                                LR chi2(0)        =      -0.00
                                                Prob > chi2       =          .
Log likelihood = -739.95364                     Pseudo R2         =    -0.0000
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    1.160808   .0362193    32.05   0.000     1.089819    1.231796
------------------------------------------------------------------------------
```

- $LR = 2(-730.194 + 739.954) = 19.52$, a 5% $\chi^2_3 = 7.814$

Modelos logit e probit

Teste LR para a significância conjunta de *dep* e *hrat*

```
. logit approve married
Logistic regression                             Number of obs     =      1,986
                                                LR chi2(1)        =       7.08
                                                Prob > chi2       =     0.0078
Log likelihood = -736.41424                     Pseudo R2         =     0.0048
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .3743725   .1394271     2.69   0.007     .1011004    .6476446
       _cons |   1.731135   .1074245    16.11   0.000     1.520587    1.941683
------------------------------------------------------------------------------

. probit approve married
Probit regression                               Number of obs     =      1,986
                                                LR chi2(1)        =       7.08
                                                Prob > chi2       =     0.0078
Log likelihood = -736.41424                     Pseudo R2         =     0.0048
------------------------------------------------------------------------------
     approve |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .1996706   .0747431     2.67   0.008     .0531768    .3461643
       _cons |   1.034537   .0587709    17.60   0.000     .9193486    1.149726
------------------------------------------------------------------------------
```

Modelos logit e probit

Logit

- $LR = 2(-729.885 + 736.414) = 13.30$, a 5% $\chi^2_2 = 5.99$

Probit

- $LR = 2(-730.194 + 736.414) = 12.47$, a 5% $\chi^2_2 = 5.99$

Modelos logit e probit

Efeito parcial de *dep* avaliado na média dos regressores

```
. sum married dep hrat
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     married |      1,986    .6586103    .4742953          0          1
         dep |      1,986    .7708963    1.104321          0          8
        hrat |      1,986    24.78853    7.111969          1         72
```

Logit

```
. display 2.453316+.4848471*0.6586103-0.1410653*0.7708963-0.0267698*24.78853
2.0003106

. display (-0.1410653)*exp(2.0003106)/((1+exp(2.0003106))^2)
-.01480745
```

Probit

```
. display 1.397252+0.255545*0.6586103-0.0747856*0.7708963-0.0134307*24.78853
1.1749773

. display (-0.0747856)*normalden(1.1749773)
-.01496031
```

Modelos logit e probit

Efeito parcial médio de *dep*

Logit

```
. gen XBL=2.453316+0.4848471*married-0.1410653*dep-0.0267698*hrat
. gen gL=exp(XBL)/((1+exp(XBL))^2)
```

Probit

```
. gen XBP=1.397252+0.255545*married-0.0747856*dep-0.0134307*hrat
. gen gP=normalden(XBP)
```

```
. sum gL gP

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          gL |      1,986    .1066199    .0241528   .0501185    .249342
          gP |      1,986    .2008731     .036222   .1063764   .3905759
```

Logit

```
. display (-0.1410653)*.1066199
-.01504037
```

Probit

```
. display (-0.0747856)*.2008731
-.01502242
```

Modelos logit e probit

Contagem de acertos na previsão

```
. quietly logit approve married dep hrat
. estat classification

Logistic model for approve

              -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |      1742            244  |       1986
     -     |         0              0  |          0
-----------+--------------------------+-----------
   Total   |      1742            244  |       1986

Classified + if predicted Pr(D) >= .5
(…)
-------------------------------------------------
Correctly classified                       87.71%
-------------------------------------------------

.
```