AULA 17 - Variáveis binárias

Econometria I

Área quantitativa - IE/UFRJ

Variáveis binárias

- A variável binária (ou dummy) é um simples exemplo de variável aleatória, o qual é chamada de função indicadora de um conjunto A:

$$I_A(x) = 1, \text{ se } x \in A \quad \text{ou} \quad I_A(x) = 0, \text{ se } x \notin A$$

- São utilizadas para diversos fins:
  - Reconhecer diferenças no intercepto e na inclinação
  - Sazonalidade
  - Variáveis qualitativas: diferenças em idade, sexo, escolaridade,...
  - Como variável dependente: Logit e Probit

Variáveis independentes binárias

- Em modelos com corte transversal é uma variável que na maior parte dos modelos está presente.
- Para modelos de séries de tempo é comum aparecer trocas no intercepto e inclinação, por causa de trocas de política econômica, trocas de regulamentações ou legislações, crises econômicas, atentados, etc.
- Também em geral temos a presença de fatores sazonais nas variáveis econômicas.
- A dummy também pode ser utilizada para eliminar outlier, colando 1 para o período do outlier e 0 caso contrário. Ou ainda poderíamos excluir esta observação.

Variáveis independentes binárias

Example

Modelo corte transversal para identificar o que determina diferenças salariais.
Regressão do salário sendo explicado pela experiência em anos e variáveis dummies qualitativas para o nível de educação e responsabilidade administrativa, que são as seguintes:
DumEF=1 ensino fundamental e 0 caso contrário;
DumEM=1 ensino médio e 0 caso contrário;
A dummy para graduação em diante é 0;
DumADM=1 para responsabilidade administrativa e 0 caso contrário.
O modelo a ser estimado é o seguinte:

$$sal = \alpha + \beta_1 exp + \beta_2 dumEF + \beta_3 dumEM + \beta_4 dumADM + u$$

Variáveis independentes binárias

Example

Resultado R:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	11009.46	386.52	28.484	$<$ 2e-16 ***
exp	552.47	30.78	17.948	$<$ 2e-16 ***
dumEF	-2974.95	415.30	-7.163	9.7e-09 ***
dumEM	125.04	391.00	0.320	0.751
dumADM	6863.29	316.62	21.676	$<$ 2e-16 ***

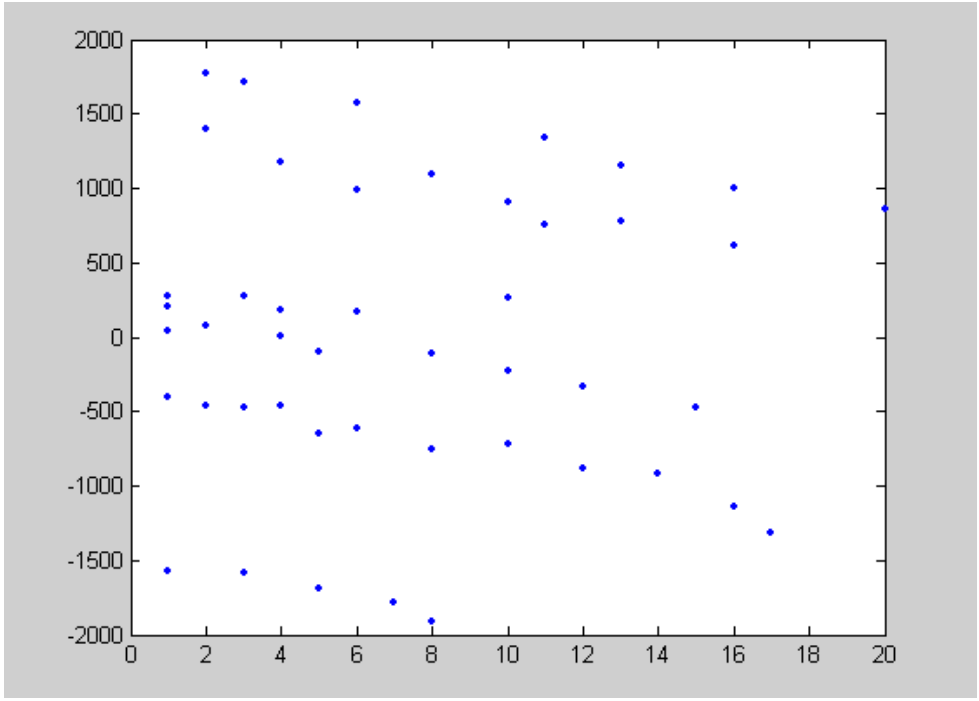
Residual standard error: 1036 on 41 degrees of freedom
Multiple R-squared: 0.9561, Adjusted R-squared: 0.9519
F-statistic: 223.5 on 4 and 41 DF, p-value: $<$ 2.2e-16

- O modelo tem um bom poder de explicação (95%) sendo que a única variável que não foi significativa foi a dummy para o ensino médio.
- Fazendo o teste de White é detectado o problema de heterocedasticidade, que pode estar relacionado com a variância dos dados ou pela omissão de alguma variável.

Variáveis independentes binárias

Example

Plotamos os resíduos da equação contra a variável explicativa experiência:



Observamos três níveis de resíduos. Possivelmente as variáveis dummies que têm sido definidas não são adequadas para explicar efeitos de educação e a de quem está em uma posição administrativa.

Variáveis independentes binárias

Example

Então a idéia é fazer uma dummy da educação multiplicada pela variável administrativa. O modelo tem o seguinte formato:

$$\begin{aligned} sal =& \alpha + \beta_1 exp + \beta_2 dumEF + \beta_3 dumEM + \beta_4 dumADM \\ &+ \beta_5 dumEF * dumADM + \beta_6 dumEM * dumADM + u \end{aligned}$$

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	11187.433	85.618	130.667	$<$ 2e-16 ***
exp	503.141	6.028	83.471	$<$ 2e-16 ***
dumEF	-1709.622	114.063	-14.988	$<$ 2e-16 ***
dumADM	7018.027	111.091	63.174	$<$ 2e-16 ***
dumEM	-383.335	105.654	-3.628	0.000817 ***
dumEF:dumADM	-3071.315	161.706	-18.993	$<$ 2e-16 ***
dumADM:dumEM	1862.503	142.037	13.113	7.08e-16 ***

Residual standard error: 188.2 on 39 degrees of freedom
Multiple R-squared: 0.9986, Adjusted R-squared: 0.9984

F-statistic: 4717 on 6 and 39 DF, p-value: $<$ 2.2e-16

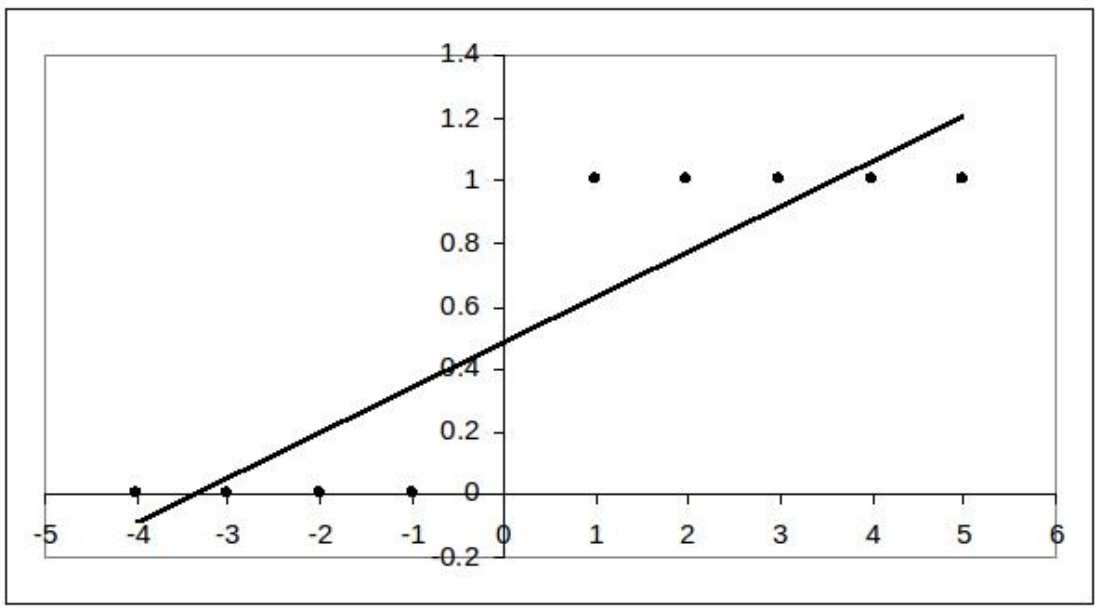
Agora não temos mais o problema de heterocedasticidade e este problema estava relacionado com a omissão de variável explicativa.

Advertência quanto ao uso de variáveis dummies

- Se uma variável qualitativa tem m categorias, introduza apenas $(m-1)$ variáveis binárias (Ex: três tipos de escolaridade, introduza apenas duas variáveis binárias. Caso contrário terá a colinearidade perfeita com o intercepto.
- A categoria para a qual nenhuma variável binária é atribuída é conhecida como categoria-base, de controle, de comparação, de referência ou categoria omitida. Todas as comparações são feitas em relação à categoria de referência. (Ex: nível graduação)
- Os coeficientes ligados às variáveis binárias são conhecidos como coeficientes diferenciais de intercepto, porque informam quanto a categoria que recebe o valor de 1 difere do coeficiente do intercepto da categoria de referência.
- Se uma variável qualitativa apresentar mais de uma categoria, como em nosso exemplo, a escolha da categoria de referência ficará estritamente a critério do pesquisador.

Modelos com variável discreta dependente

- Vimos até então modelos em que as variáveis explicativas (x) são dummies, agora veremos modelos em que variáveis explicadas (y) são dummies.
- Por exemplo $y = 1$ se a empresa entra em default e $y = 0$ caso contrário.
- A ideia de fazer uma estimativa por MQO, tentando aproximar uma linha de regressão linear, temos que os valores previstos podem ficar fora do intervalo $[0, 1]$ e assim os erros de previsão serão grandes.



Modelos com variável dependente de escolha discreta

Existe uma variedade de comportamentos econômicos onde as **aproximações contínuas não são boas**.

Por exemplo:

- **Indivíduos**: escolha de emprego; número de filhos; se compra uma casa; se é inadimplente; se participa da escola; se opta casar.
- **Empresas**: se construir uma planta, e se sim, qual localização; qual produto produzir; se paga a dívida; adquirem-se outras empresas.

Estes modelos são conhecidos como **Resposta Qualitativa**.

Modelos de escolha discreta

- Nos modelos de Resposta Qualitativa a **variável dependente** é um indicador de uma **escolha discreta**
- Em geral, os métodos de **regressão convencionais** são **inadequados** e a maioria das estimativas é feita por **máxima verossimilhança**.
- Existem vários modelos. Os mais conhecidos são os modelos com variáveis dependentes **binárias e ordinais**, por exemplo:
  - **Participação da força de trabalho**: atribuimos "não" com 0 e "sim" com 1. Esses as decisões são escolhas qualitativas. A codificação $0/1$ é uma mera conveniência.
  - **Opiniões de um certo tipo de legislação**: Seja 0 representado "contrário fortemente", 1 "contrário", 2 "neutro", 3 "apoia" e '4 "apoia fortemente". Esses números são rankings, e os valores escolhidos não são quantitativos, mas apenas uma ordenação.

Modelos de escolha discreta

- Precisamos de modelos de probabilidade que aproximem o verdadeiro processo gerador dos dados.
- Analisaremos os modelos dentro de uma estrutura geral dos modelos de probabilidade:

$$\text{Prob(evento } j \text{ ocorrer)} = \text{Prob}(Y = j) = \text{F[efeitos relevantes, parâmetros]}$$

- O estudo da escolha qualitativa concentra-se na especificação, estimativa e uso de modelos para as probabilidades de eventos, onde na maioria dos casos, o "evento" é um escolha do indivíduo entre um conjunto de alternativas.

Modelos para escolha binária

Modelos para explicar uma variável dependente binária (0/1) surgem em dois contextos:

- Especificar uma relação entre a variável de interesse e um conjunto de variáveis explicativas.
  Exemplo: a relação entre o comportamento do voto e a renda.
- Modelo no qual a natureza dos dados observados dita o tratamento especial de um modelo de escolha binária.
  Exemplo: demanda por ingressos de um evento esportivo, mas podemos estar interessados em saber apenas se a capacidade foi preenchida (demanda maior ou igual a capacidade $Y = 1$) ou não ($Y = 0$).

Modelos para escolha binária

- Para modelar a escolha binária considere o exemplo em que o agente trabalha ou procura trabalho $(Y = 1)$ ou não $(Y = 0)$, no período em que a pesquisa é feita.

- O conjunto de fatores, como idade, estado civil, educação e histórico de trabalho, do vetor $\mathbf{x}$ explicam a decisão, tal que:

$$Prob(Y = 1|\mathbf{x}) = F(\mathbf{x}, \beta)$$
$$Prob(Y = 0|\mathbf{x}) = 1 - F(\mathbf{x}, \beta)$$

  sendo $F(\cdot)$ uma função.

- Se definirmos uma regressão linear $F(\mathbf{x}, \beta) = \mathbf{x}'\beta$ não podemos garantir que $\mathbf{x}'\beta$ esteja no intervalo $[0, 1]$.

Modelos Probit e Logit

- Em princípio, qualquer distribuição de probabilidade contínua adequada definida sobre a linha real será suficiente.
- Um primeiro candidato é a distribuição normal sendo utilizada em muitas aplicações - chamamos modelo **Probit**

$$Prob(Y = 1|\mathbf{x}) = \int_{-\infty}^{x'\beta} \phi(t)dt = \Phi(\mathbf{x}'\beta)$$

em que $\phi(t)$ é a densidade da normal padrão

- Outra distribuição bastante usada é a logística, sendo denomidada modelo **Logit**

$$Prob(Y = 1|\mathbf{x}) = \frac{e^{\mathbf{x}'\beta}}{1+e^{\mathbf{x}'\beta}} = \Lambda(\mathbf{x}'\beta)$$

em que $\Lambda(\mathbf{x}'\beta)$ é a função distribuição cumulativa logística

- A distribuição logística é semelhante à normal, exceto nas caudas, que são mais pesadas.

Interpretação dos modelos de resposta binária

- A interpretação não é tão simples como o modelo OLS (em que podemos achar β como uma derivada parcial).
- Isto é por causa da não linearidade.
- O modelo de probabilidade é uma regressão:

$$E[y|\mathbf{x}] = 0[1 - F(\mathbf{x}', \beta)] + 1[F(\mathbf{x}', \beta)] = F(\mathbf{x}', \beta)$$

- Para aplicar os modelos probit e logit, é importante saber como interpretar os β_j no caso de variáveis explicativas contínuas e discretas.
- Se $\mathbf{x}_j$ é contínua

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = f(\mathbf{x}'\beta)\beta$$

em que $f(\cdot)$ é a função densidade

Interpretação dos modelos de resposta binária

- Para a distribuição normal:

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \phi(\mathbf{x}'\beta)\beta$$

- Para a distribuição logística:

$$\frac{\partial E[y|\mathbf{x}]}{\partial \mathbf{x}} = \frac{e^{\mathbf{x}'\beta}}{(1+e^{\mathbf{x}'\beta})^2}\beta$$

- O efeito marginal para uma variável binária independente, digamos d

$$\begin{aligned}\text{Efeito marginal} =& \text{Prob}[Y = 1|\overline{\mathbf{x}}_{(d)}, d = 1] \\ &- \text{Prob}[Y = 1|\overline{\mathbf{x}}_{(d)}, d = 0]\end{aligned}$$

em que $\overline{\mathbf{x}}_{(d)}$ denota a média de todos as outras variáveis do modelo.

Estimação dos modelos de resposta binária

- A estimação é baseada no método de máxima verossimilhança.
- Ambos modelos Probit e Logit podem ser estimados pelo log da função de verossimilhança

$$\ln L = \sum_{y_i=0} [1 - \ln F(\mathbf{x}_i'\beta)] + \sum_{y_i=1} \ln F(\mathbf{x}_i'\beta)$$

- Assumimos que $F(\cdot)$ é duas vezes diferenciável.
- Usamos a condição de primeira ordem para maximização de L $\frac{\partial \ln L}{\partial \beta}$.
- Para testar hipóteses sobre os coeficientes no caso de uma restrição podemos usar o teste t. Para mais de uma restrição podemos usar o teste Wald, LR, LM.

Variáveis dependentes ordinais

- Constituem uma variação dos modelos logit/probit e são geralmente aplicados quando a variável dependente é uma medida discreta e ordenada - não simplesmente binária, mas uma escala ordinal em vez de uma escala de intervalo.
- Exemplos de variáveis de escolha multinomial são ordenadas:
  - Classificações de títulos
  - Resultados dos testes de preferências
  - Pesquisas de opinião
  - A designação de pessoal militar para classificações de trabalho por nível de habilidade
  - Resultados de votação em determinados programas
  - O nível de cobertura de seguro tomado por um consumidor: nenhum, parte ou total
  - Emprego: desempregado, a tempo parcial ou a tempo inteiro
- Os modelos probit e logit ordenados passaram a ser bastante utilizados como uma estrutura para analisar tais respostas.