AULAS 08 A 14
MÉTODOS QUASE-EXPERIMENTAIS

Ernesto F. L. Amaral

02, 04, 16, 18, 23, 25 e 30 de abril de 2013
Técnicas Avançadas de Avaliação de Políticas Públicas (DCP 098)

Fonte:

Curso "Técnicas Econométricas para Avaliação de Impacto" do "International Policy Centre for Inclusive Growth" (IPC-IG) da "United Nations Development Programme" (UNDP) (http://www.ipc-undp.org/evaluation).

ESTRUTURA DA AULA

– Traduzir a diferença de médias para uma linguagem de regressão linear.

– Aplicar esta linguagem de regressão linear para desenho com dados antes e depois, assim como para desenho com dados de controle e tratamento.

– Erros padrão robustos por heteroscedasticidade.

– Viés causado por seleção, dados em branco (“missing”) e falha no controle de fatores não-observáveis.

– Análise de impacto com regressão linear e não-linear (modelos logísticos e razões de chance).

– Bancos de dados em formato amplo (“wide”) e bancos de dados em painel (“long”).

– Estimando heterogeneidade do impacto.

PARÂMETROS DE INTERESSE

– Ao realizar testes de diferença de médias ou modelos de regressão para avaliação de impacto de políticas públicas, temos os seguintes parâmetros de interesse:

– **ATE:** impacto médio do tratamento na população como um todo (independente de quem foram os grupos de controle e tratamento).

– **ATT ou ATET:** impacto médio sobre o grupo que recebeu a política pública (tratamento).

– **ATU:** representa o quanto o grupo que não está sendo tratado seria afetado caso fosse tratado. Impacto médio sobre o grupo que não recebeu a política (controle). Não é o mesmo que externalidades da política. Mede impacto da expansão do programa para além do grupo já tratado.

– Para que ATU e ATE sejam de interesse, é relevante que amostra de controle represente população de interesse.

DADOS

- Métodos quase-experimentais exigem amostras maiores que métodos experimentais:
  - Somente parte da variação nas variáveis de interesse é utilizada na estimação.
  - Geralmente, o grupo de tratamento tende a ser mais homogêneo, do ponto de vista da investigação, que o grupo de controle.
  - A regra 50%-50% utilizada na amostra de experimentos não é válida nos quase-experimentos.
- Base de dados secundárias:
- Pesquisas para cobrir uma série de outros propósitos.
- Informações relevantes podem estar ausentes.
- Ex.: avaliação de impacto do PBF utilizando a PNAD.

RECAPITULAÇÃO SOBRE TESTE DE HIPÓTESES

DECISÃO SOBRE HIPÓTESES

Hipóteses	$p < \alpha$	$p > \alpha$
Hipótese nula (H_0)	Rejeita	Não rejeita
Hipótese alternativa (H_1)	Aceita	Não aceita

– *p-valor*: probabilidade de não rejeitar a hipótese nula.

– *α*: nível de significância adotado (ex.: 0,10, 0,05, 0,01). É o complementar do nível de confiança (ex.: 90%, 95%, 99%).

– Como Stata calcula *p-valor* bilateral, é só dividir este valor por 2 para obter o *p-valor* unilateral.

HETEROSCEDASTICIDADE

Fonte:
Wooldridge, Jeffrey M. “Introdução à econometria: uma abordagem moderna”. São Paulo: Cengage Learning, 2008. Capítulo 8 (pp.243-271).

HOMOSCEDASTICIDADE

– A hipótese de homoscedasticidade para a regressão múltipla significa que a variância do erro não observável (*u*), condicional nas variáveis explicativas, é constante.

– A homoscedasticidade não se mantém quando a variância dos fatores não-observáveis muda ao longo de diferentes segmentos da população.

– Por exemplo, a heteroscedasticidade está presente se a variância dos fatores não-observados (*u*) que afetam a renda (*y*) aumenta com a idade (*x*).

– A homoscedasticidade é necessária para estimar os testes de t e F, além dos intervalos de confiança.

– A intenção aqui é de: (1) discorrer sobre as consequências da heteroscedasticidade para estimação de MQO; (2) verificar a presença da heteroscedasticidade; (3) discutir soluções para a ocorrência deste problema.

β_j E R^2 NA HETEROSCEDASTICIDADE

– A heteroscedasticidade não provoca viés ou inconsistência nos estimadores MQO de β_j, enquanto a omissão de uma variável importante teria esse efeito.

– O R^2 da população é:

1 – (variância do erro / variância de y)

– Como ambas variâncias no R^2 da população são incondicionais, o R^2 da população não é afetado pela presença de heteroscedasticidade em $Var(u|x_1,..., x_k)$.

– SQR/n estima consistentemente a variância do erro, e SQT/n estima consistentemente a variância de y, seja $Var(u|x_1,..., x_k)$ constante ou não.

– Portanto R^2 e R^2 ajustados são estimadores consistentes do R^2 da população, mantendo ou não a hipótese de homoscedasticidade.

ERROS-PADRÃO NA HETEROSCEDASTICIDADE

– Os estimadores de variâncias [Var(β_j)] são viesados sem a hipótese de homoscedasticidade.

– Como os erros-padrão dos estimadores MQO são baseados diretamente nessas variâncias, eles não mais são válidos para construirmos intervalos de confiança e estatísticas *t*.

– Na presença de heteroscedasticidade, as estatísticas *t* não têm distribuições *t*, as estatísticas *F* não têm distribuição *F*, e a estatística "Multiplicador de Lagrange" (*LM*) não tem distribuição qui-quadrada.

– Portanto, as estatísticas que usamos para testar hipóteses não são válidas na presença de heteroscedasticidade.

– Os estimadores MQO são os melhores estimadores lineares não-viesados na hipótese de homoscedasticidade: isso ocorre quando Var(u|x) for constante.

INFERÊNCIA ROBUSTA

– É possível ajustar erros-padrão, estatísticas *t*, *F* e *LM* de forma a torná-las válidas na presença de heteroscedasticidade de forma desconhecida.

– Isso significa que é possível descrever novas estatísticas que funcionam independentemente do tipo de heteroscedasticidade presente na população.

– Esses métodos são os procedimentos robustos em relação à heteroscedasticidade, já que são válidos mesmo que a variância dos erros não seja constante.

– É possível então estimar variâncias consistentes na presença de heteroscedasticidade.

– A aplicação de métodos robustos em relação à heteroscedasticidade é bastante fácil, pois muitos programas estatísticos e econométricos calculam essas estatísticas como uma opção.

ESTIMANDO VARIÂNCIA COM HETEROSCEDASTICIDADE

– No caso da regressão simples e sem a hipótese de homoscedasticidade, a variância do estimador é:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{SQT_x^2}$$

– Quando $\sigma_i^2 = \sigma^2$ para todo *i*, a fórmula se reduz a: σ^2/SQT_x.

– Quando $\sigma_i^2 \neq \sigma^2$ (heteroscedasticidade), a variância derivada sob homoscedasticidade não é mais válida.

– Como o erro-padrão é baseado diretamente na estimativa da variância, é preciso estimar a equação acima quando a heteroscedasticidade está presente.

– Sendo u_i os resíduos da regressão simples de y sobre x, um estimador válido da variância para a heteroscedasticidade é:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\hat{u}_i^2}{SQT_x^2}$$

EM REGRESSÃO MÚLTIPLA

– No caso de: (1) regressão múltipla; (2) r_{ij} ser o *i*-ésimo resíduo da regressão de x_j sobre todas as outras variáveis independentes; e (3) SQR_j ser a soma dos resíduos quadrados da regressão, temos:

$$Var(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SQR_j^2}$$

– A raiz quadrada desta fórmula é o erro-padrão robusto em relação à heteroscedasticidade de beta estimado.

– Os erros-padrão robustos são atribuídos a White (1980).

– A estatística *t* robusta em relação à heteroscedasticidade é calculada após obter os erros-padrão robustos:

$$t = \frac{estimativa - valor\ hipotético}{erro\ padrão}$$

ERROS-PADRÃO USUAIS E ROBUSTOS

– Geralmente, os erros-padrão robustos são frequentemente maiores do que os erros-padrão usuais.

– Os erros-padrão robustos podem ser estimados mesmo sem que se saiba se a heteroscedasticidade está presente.

– Os novos erros-padrão são válidos (assimptoticamente) haja ou não presença de heteroscedasticidade.

– Com frequência, as diferenças entre os erros-padrão usuais e os robustos são pequenas.

– Erros-padrão usuais podem ser usados se a hipótese de homoscedasticidade se mantiver e erros forem normalmente distribuiídos, já que estatísticas *t* usuais terão distribuições *t*.

– Em amostras pequenas, as estatísticas *t* robustas podem ter distribuições que não sejam próximas da distribuição *t*.

– Em amostras grandes, sempre podemos levar em conta somente os erros-padrão robustos.

ESTATÍSTICAS *F* E *LM*

– É possível obter estatísticas *F* e *LM* robustas em relação à heteroscedasticidade de forma desconhecida.

– A estatística *F* robusta em relação à heteroscedasticidade é chamada de estatística de Wald robusta em relação à heteroscedasticidade.

– O cálculo do teste *F* robusto não tem uma forma simples, mas pode ser computado por alguns programas estatísticos.

MULTIPLICADOR DE LAGRANGE (*LM*) ROBUSTO

– Nem todos programas econométricos calculam estatísticas *F* que sejam robustas em relação à heteroscedasticidade.

– Uma estatística *LM* robusta pode ser obtida manualmente em qualquer programa econométrico:

1. Obtenha os resíduos *u* do modelo restrito.
2. Faça a regressão de cada uma das variáveis independentes excluídas, conforme a hipótese nula, sobre todas as variáveis independentes incluídas, e salve os resíduos (r_1, r_2, ..., r_q).
3. Encontre os produtos de cada r_j por *u* (para todas as observações).
4. Faça a regressão de 1 sobre r_1u, r_2u, ..., r_qu, sem um intercepto.
5. Use a soma dos resíduos quadrados da última regressão para calcular a estatística *LM* robusta (n - SQR), a qual terá distribuição de qui-quadrado.

TESTE DE EXISTÊNCIA DE HETEROSCEDASTICIDADE

– Os erros-padrão robustos em relação à heteroscedasticidade oferecem um método simples para calcular estatísticas *t* que sejam assimptoticamente distribuídas como *t*, haja ou não a presença de heteroscedasticidade.

– Porém, há razões para saber se realmente há presença de heteroscedasticidade, antes de estimar erros-padrão robustos:

  – As estatísticas *t* usuais são preferíveis se não há heteroscedasticidade.

  – É possível obter um estimador melhor que o MQO quando a forma da heteroscedasticidade é conhecida.

TESTE DE EXISTÊNCIA DE HETEROSCEDASTICIDADE

– Considere um modelo linear:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

– A hipótese nula de que a homoscedasticidade se mantém é:

$$H_0: \text{Var}(u|x_1,x_2,...,x_k) = \sigma^2$$

– Precisamos analisar os dados para saber se a hipótese nula é adequada ou não.

– Se não rejeitamos H_0, concluímos que a heteroscedasticidade não será um problema.

– Como *u* tem esperança condicional zero, $\text{Var}(u|x)=E(u^2|x)$, a hipótese nula será:

$$H_0: E(u^2|x_1,x_2,...,x_k) = E(u^2) = \sigma^2$$

TESTE *F* DE EXISTÊNCIA DE HETEROSCEDASTICIDADE

– Estimamos então esta equação:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + erro$$

– Utilizando o R^2 da equação acima e o número de regressores (k), estimamos a estatística *F*:

$$F = \frac{R^2_{\hat{u}^2}/k}{\left(1 - R^2_{\hat{u}^2}\right)/(n-k-1)}$$

– A estatística *F* tem uma distribuição $F_{k,n-k-1}$ sob a hipótese nula de homoscedasticidade, permitindo o cálculo de sua significância.

TESTE *LM* DE EXISTÊNCIA DE HETEROSCEDASTICIDADE

– A estatística *LM* para a heteroscedasticidade é o tamanho da amostra multiplicado pelo R^2 da equação com u^2 como variável dependente:

$$LM = n * R^2_{\hat{u}^2}$$

– Essa versão *LM* do teste é geralmente chamada teste de Breusch-Pagan da heteroscedasticidade (teste BP).

RESUMINDO O TESTE BP

– Estime o modelo MQO em que *y* é a variável dependente e obtenha os resíduos quadrados (u^2) para cada observação.

– Estime o modelo em que *u* é a variável dependente para obter o R-quadrado.

– Construa a estatística *F* e calcule o p-valor usando a distribuição $F_{k,n-k-1}$.

– Construa a estatística *LM* e calcule o p-valor usando a distribuição de qui-quadrado.

– Se o p-valor ficar abaixo do nível de significância selecionados, então rejeitamos a hipótese nula de homoscedasticidade.

– Se for constatada que não há homoscedasticidade, os erros-padrão robustos em relação à heteroscedasticidade e suas estatísticas de testes poderão ser utilizadas.

– Sabemos ainda que há menos heteroscedasticidade com a variável dependente em forma logarítmica.

TESTE DE WHITE PARA HETEROSCEDASTICIDADE

– A hipótese de homoscedasticidade [$Var(u|x_1,...,x_k)$] pode ser substituída por outra hipótese:
  – O erro quadrado (u^2) é não-correlacionado com:
    – Todas as variáveis independentes (x_j).
    – Os quadrados das variáveis independentes (x_j^2).
    – Todos os produtos cruzados ($x_j x_h$ para $j \neq h$).
– White sugeriu testar formas de heteroscedasticidade que invalidem os erros-padrão e as estatísticas de testes.
– Para um modelo com três variáveis independentes, temos:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + erro$$

–O teste de White para a heteroscedasticidade é a estatística *LM* para testar se todos δ_j na equação sejam zero, exceto δ_0.

TESTE DE WHITE PARA HETEROSCEDASTICIDADE

– O teste de White usa muitos graus de liberdade para modelos com um número moderado de variáveis independentes.

– É possível obter um teste que seja mais facilmente implementado que o teste de White.

– Uma sugestão é usar os valores estimados MQO para verificar a existência de heteroscedasticidade.

– Os valores estimados são apenas funções lineares das variáveis independentes.

– Se eles forem elevados ao quadrado, estamos na prática obtendo uma função particular de todos os quadrados e produtos cruzados das variáveis independentes:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + erro$$

– Podemos usar as estatísticas *F* ou *LM* para a hipótese nula:

$$H_0: \delta_1 = 0, \delta_2 = 0$$

RESUMINDO O TESTE DE WHITE

– Estime o modelo MQO em que *y* é a variável dependente e obtenha os resíduos (*u*) e os valores estimados de *y*.

– Calcule os resíduos quadrados (u^2) e os quadrados dos valores estimados.

– Estime o modelo em que *u* é a variável dependente e *y* e y^2 sejam as variáveis independentes para obter o R^2.

– Construa a estatística *F* e calcule o p-valor usando a distribuição $F_{2,n-3}$.

– Construa a estatística *LM* e calcule o p-valor usando a distribuição de qui-quadrado.

– Se o p-valor ficar abaixo do nível de significância selecionados, então rejeitamos a hipótese nula de homoscedasticidade.

CONSIDERAÇÃO IMPORTANTE

- Se omitirmos um ou mais termos quadráticos em um modelo de regressão ou usarmos o modelo em nível ao invés de usar o log, um teste de heteroscedasticidade pode vir a ser significante, rejeitando a hipótese de homoscedasticidade.
- Isso tem levado alguns pesquisadores a verem estes testes como testes de má especificação do modelo:
  - Porém, há outros testes que podem testar melhor a má especificação de formas funcionais das variáveis.
- Ou seja, é mais apropriado:
  - Primeiro, realizar testes específicos de formas funcionais, já que a má especificação da forma funcional é mais importante que a heteroscedasticidade.
  - Depois de satisfeitos com as formas funcionais das variáveis, estimar o teste para verificar a existência de heteroscedasticidade.

ESTIMAÇÃO DE MÍNIMOS QUADRADOS PONDERADOS

– Se for detectada heteroscedasticidade com o uso de testes estatísticos, é possível estimar erros padrão robustos em relação à heteroscedasticidade após a estimação MQO.

– Porém, antes das estatísticas robustas, é possível modelar e estimar a forma específica da heteroscedasticidade, calculando um estimador mais eficiente que o MQO, além de estatísticas *t* e *F* não enviesadas.

– Isso requer mais trabalho, pois é preciso ser específico sobre a natureza de qualquer heteroscedasticidade.

CONSTANTE MULTIPLICATIVA

– Considere que ***x*** representa todas as variáveis explicativas em:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

– Assuma que *h(**x**)* é alguma função das variáveis explicativas que determina a heteroscedasticidade:

$$Var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$$

– Como variâncias devem ser positivas, *h(**x**)*>0 para todos valores possíveis das variáveis independentes.

– Supomos que a função *h(**x**)* é conhecida. Assim, mesmo que o parâmetro populacional σ^2 seja desconhecido, teremos condições de estimá-lo a partir de uma amostra de dados.

EQUAÇÃO TRANSFORMADA

– Com o objetivo de obter estimadores de β_j que tenham propriedades de eficiência melhores que MQO, estimamos esta equação:

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1\left(x_{i1}/\sqrt{h_i}\right) + \cdots + \beta_k\left(x_{ik}/\sqrt{h_i}\right) + u_i/\sqrt{h_i}$$

– Esta equação é linear em seus parâmetros (RLM.1), a hipótese de amostragem aleatória não se alterou (RLM.2), o termo de erro tem média condicional zero (RLM.3) e não há colinearidade perfeita entre variáveis independentes (RLM.4).

– A equação transformada satisfará as hipóteses do modelo linear clássico, se o modelo original também o fizer, com exceção da hipótese de homoscedasticidade (RLM.5).

MÍNIMOS QUADRADOS GENERALIZADOS (MQG)

– É necessário estimar os parâmetros da nova equação por mínimos quadrados ordinários.

– Os novos betas são estimadores de mínimos quadrados generalizados (MQG).

– Estes estimadores MQG são usados para explicar a heteroscedasticidade nos erros.

– Os erros-padrão, estatísticas *t* e estatísticas *F* podem ser obtidas de regressões que usem as variáveis transformadas.

– Por serem os melhores estimadores lineares não-viesados de beta, os estimadores MQG são mais eficientes que os estimadores MQO.

– A interpretação dos resultados deve ser feita com base na equação original.

– O R^2 indica o quanto da variação do novo y é explicado pelo novo x, o que não é informativo como grau de ajuste.

MÍNIMOS QUADRADOS PONDERADOS (MQP)

– Os estimadores de mínimos quadrados generalizados (MQG) para correção da heteroscedasticidade são chamados de estimadores de mínimos quadrados ponderados (MQP).

– Os novos betas minimizam a soma ponderada dos quadrados dos resíduos.

– A idéia é colocar menos peso nas observações com uma variância de erro mais alta.

– O método MQO atribui pesos iguais a todas as observações, pois isso é melhor quando a variância do erro é idêntica para todas as partições da população.

MÍNIMOS QUADRADOS PONDERADOS (MQP)

– A maioria dos programas econométricos tem um recurso para computar mínimos quadrados ponderados.

– Juntamente com as variáveis dependentes e independentes originais, especificamos a função de ponderação ($1/h_i$).

– Especificamos pesos proporcionais ao inverso da variância.

– Isso nos permite interpretar as estimativas de mínimos quadrados ponderados no modelo original.

– Podemos escrever a equação estimada da maneira habitual.

– As estimativas e os erros-padrão serão diferentes do MQO, mas a maneira como interpretamos essas estimativas, erros-padrão e estatísticas de testes é a mesma.

– Esse procedimento corrige estimativas dos betas (aweight).

– Se considerarmos que a heteroscedasticidade seria um problema para os erros-padrão, deveríamos computar também os erros-padrão robustos (pweight).

MAS NA PRÁTICA...

– Na prática, raramente sabemos como a variância do erro se comporta em relação a uma variável independente.

– Em equações de regressão múltipla, é complicado saber com qual variável independente há heteroscedasticidade nos erros e qual a forma deste problema.

– Existe um caso no qual os pesos necessários para o MQP surgem naturalmente de um modelo econométrico subjacente.

– Isso acontece quando os dados estão em médias de algum grupo ou região, e não em nível individual.

DADOS EM MÉDIAS POR GRUPOS

– Se a equação no nível individual satisfizer a hipótese de homoscedasticidade, então a equação do nível agrupado deverá ter heteroscedasticidade.

– Assim, se para todo grupo *i* e indivíduo *j*:

$$Var(u_{i,j}) = \sigma^2$$

– Então, a variância do termo de erro médio diminui com o tamanho do grupo:

$$Var(\bar{u}_i) = \sigma^2 / m_i$$

– Neste caso, $h_i = 1/m_i$.

– Portanto, o procedimento mais eficiente será o dos mínimos quadrados ponderados, com pesos correspondentes ao número de indivíduos nos grupos ($1/h_i = m_i$).

– Isso garante que grupos maiores recebam peso maior, o que oferece método eficiente de estimação dos parâmetros no modelo em nível individual quando temos médias.

HETEROSCEDASTICIDADE NO NÍVEL INDIVIDUAL

– Se no caso anterior existisse heteroscedasticidade no nível individual, então a ponderação adequada dependerá da forma da heteroscedasticidade.

– Por isso, vários pesquisadores simplesmente computam erros-padrão e estatísticas de teste robustos na estimação de modelos que usam dados agrupados.

– Uma alternativa é realizar a ponderação pelo tamanho do grupo (aweight), além de estimar as estatísticas robustas em relação à heteroscedasticidade na estimação MQP (pweight).

– Isso assegura que qualquer heteroscedasticidade no nível individual seja representada pela inferência robusta.

MQG FACTÍVEL

– Ao contrário dos exemplos anteriores, a forma exata de heteroscedasticidade não é óbvia na maioria dos casos.

– Em muitos casos podemos modelar a função *h* e utilizar os dados para estimar os parâmetros desconhecidos.

– O uso de h_i-chapéu em lugar de h_i na transformação MQG produz o estimador de mínimos quadrados generalizados factível (MQGF), também chamado de MQG estimado (MQGE).

– Existem várias maneiras de modelar a heteroscedasticidade, mas iremos utilizar um método razoavelmente flexível:

$$Var(u|\boldsymbol{x}) = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k)$$

– É utilizada função exponencial porque modelos lineares não asseguram que os valores previstos sejam positivos, e as variâncias estimadas devem ser positivas para usar o MQP.

ESTIMAÇÃO DO MQG FACTÍVEL

– Para estimar os parâmetros δ_i é preciso transformar a equação anterior em uma forma linear para ser estimada por MQO:

$$log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + e$$

– Na prática (pág. 263):

1. Execute a regressão de *y* sobre x_1, x_2, ..., x_k e obtenha os resíduos de $\hat{u}$.
2. Crie $log(\hat{u}^2)$ elevando ao quadrado os resíduos MQO e depois calculando seu log natural.
3. Execute a regressão na equação acima dos parâmetros δ_i [ou log(u^2) sobre *y*, y^2] e obtenha os valores estimados.
4. Calcule o exponencial dos valores estimados, resultando em: $\hat{h}$.
5. Estime a equação $y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u$, pelo método MQP, usando pesos (aweight) $1/\hat{h}$.

ESTATÍSTICAS *F*

– Ao calcular estatísticas *F*, é importante que os mesmos pesos sejam usados para estimar os modelos com e sem restrições.

– Devemos estimar o modelo sem restrições por MQO com os pesos.

– Usamos os mesmos pesos para estimar o modelo restrito.

– Posteriormente, a estatística *F* pode ser calculada.

– Lembrem-se que o Stata permite utilizar o comando “test” para testar restrições conjuntas após a estimação de um modelo, não sendo necessário calcular manualmente a regressão restrita.

MODELO DE PROBABILIDADE LINEAR REVISITADO

– Quando a variável dependente é binária, o modelo deve conter heteroscedasticidade, a menos que todos parâmetros de inclinação sejam nulos.

– A maneira mais simples de tratar a heteroscedasticidade neste caso é usar a estimação MQO, e calcular os erros-padrão robustos nas estatísticas de testes.

– As estimativas MQO do MPL são simples e geralmente produzem resultados satisfatórios, mas são ineficientes.

– É possível utilizar o MQP para estimar o MPL. No entanto, o método falhará se $\hat{h}$ for negativo (ou zero) em qualquer observação.

ESTIMAÇÃO DO MPL POR MQP

– Estime o modelo por MQO e obtenha os valores estimados de *y*.

– Verifique se todos os valores estimados estão dentro do intervalo unitário:

  – Se assim for, prossiga para o passo seguinte.

  – Caso contrário, alguns ajustes serão necessários para trazer todos os valores estimados para dentro do intervalo unitário:

    – $y_i = 0{,}01$ se $y_i < 0$

    – $y_i = 0{,}99$ se $y_i > 1$

– Construa as variâncias estimadas com esta equação:

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$$

– Estime a equação $y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u$, pelo método MQP, usando pesos (aweight) $1/\hat{h}$.

PROBLEMAS ADICIONAIS DE ESPECIFICAÇÃO E DE DADOS

Fonte:
Wooldridge, Jeffrey M. “Introdução à econometria: uma abordagem moderna”. São Paulo: Cengage Learning, 2008. Capítulo 9 (pp.272-303).

E OS PROBLEMAS NÃO TERMINAM...

– Como vimos anteriormente, a heteroscedasticidade nos erros pode ser vista como uma má especificação do modelo, porém é um problema de menor importância.

– A presença de heteroscedasticidade não causa viés ou inconsistência nos estimadores MQO.

– É ainda possível ajustar intervalos de confiança e estatísticas *t* e *F* para obter inferência válida após a estimação MQO.

– Por fim, os mínimos quadrados ponderados permitem obter estimadores mais eficientes que aqueles do MQO.

– Agora trataremos de um problema mais sério da correlação entre o erro (*u*) e uma ou mais variáveis independentes.

NOVOS PROBLEMAS

– Se *u* for correlacionado com *x*, então *x* é uma **variável explicativa endógena**.

– Quando uma variável omitida é uma função de uma variável explicativa, há **má especificação da forma funcional**.

– A omissão de uma variável importante pode causar correlação entre o erro e variáveis explicativas, o que pode gerar viés e inconsistência em estimadores MQO.

– Tópicos deste capítulo:

  – **Conseqüências** da má especificação da forma funcional e como testar sua existência.

  – Como o uso de **variáveis *proxy*** pode resolver ou aliviar o viés de omissão.

  – Explicação do viés no método MQO que pode aparecer sob certas formas de **erros de medida**.

  – Discussão de **problemas adicionais**: ausência de dados, amostras não-aleatórias e observações extremas.

MÁ ESPECIFICAÇÃO DA FORMA FUNCIONAL

– Um modelo de regressão múltipla sofre de má especificação da forma funcional quando não explica de maneira apropriada a relação entre variáveis explicativas e a dependente.

– Se a renda for explicada pela educação, experiência e experiência ao quadrado, mas omitimos o termo elevado ao quadrado, há má especificação da forma funcional.

– Isso conduz a estimadores viesados das demais variáveis independentes.

– Neste exemplo, a magnitude do viés depende do tamanho do beta de educação e da correlação entre educação, experiência e experiência ao quadrado.

– Usar apenas o estimador viesado de experiência pode ser enganoso, especialmente nos valores extremos de experiência.

OUTRO EXEMPLO

$$log(salárioh) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2$$
$$\beta_4 feminino + \beta_5 feminino * educ + u$$

– Se omitirmos o termo de interação (feminino*educ), estaremos especificando mal a forma funcional.

– Com essa omissão, obteremos estimadores viesados dos outros parâmetros.

– Como retorno de educação depende do sexo, não fica claro que tipo de retorno estaríamos estimando quando omitimos o termo de interação.

NÃO É UM PROBLEMA GRAVE

– A omissão de funções de variáveis independentes não é a única maneira de um modelo sofrer o problema de má especificação da forma funcional.

– Se for necessário utilizar o logaritmo da variável dependente, mas a utilizamos em sua forma original, não obteremos estimadores não-viesados ou consistentes dos efeitos parciais.

– Há testes para detectar esse tipo de problema da forma funcional.

– Esse é um problema secundário, já que temos dados de todas variáveis necessárias para obter uma relação funcional que se ajuste bem aos dados.

– Ou seja, não há omissão de variáveis.

IMPORTÂNCIA DO TESTE *F*

– Uma ferramenta para detectar uma forma funcional mal-especificada é o teste *F* para restrições de exclusões conjuntas.

– Faz sentido adicionar termos quadráticos de variáveis significantes no modelo e executar um teste conjunto de significância.

– Se termos quadráticos adicionados forem significantes, eles podem ser adicionados ao modelo, mas interpretação será mais complicada.

– Além da adição de termos quadráticos, o uso de logaritmos é suficiente para detectar muitas relações não-lineares importantes em ciências sociais aplicadas.

TESTE RESET

– O teste de erro de especificação da regressão (RESET) é útil para detectar a má especificação da forma funcional.

– Suponha este modelo:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

– Se ele satisfizer RLM.3 (termo de erro tem média condicional zero), nenhuma função não-linear das variáveis independentes deve ser significante quando adicionada à equação.

– Se testarmos todas possibilidades de termos quadráticos das variáveis explicativas para testar problemas de forma funcional, teremos a desvantagem de gastar muitos graus de liberdade se houver muitas variáveis independentes.

– Além disso, certos tipos de não-linearidades (logaritmo, por exemplo) não serão detectados por termos quadráticos.

REALIZANDO O TESTE RESET

– O teste RESET adiciona polinômios na equação para detectar má especificação de formas funcionais.

– Para realizar o teste, temos que decidir quantas funções dos valores estimados devem ser incluídas.

– Não há resposta certa para isto, mas os termos quadráticos e cúbicos têm demonstrado utilidade nestas aplicações:

  – Primeiro estimamos a equação original (restrita).

  – Depois, salvamos os valores preditos e geramos seus termos quadráticos e cúbicos.

  – Em seguida, estimamos esta equação (irrestrita) para testar se a equação original têm não-linearidades importantes ausentes:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + erro$$

  – Por fim, geramos a estatística do teste RESET que é a estatística *F* para testar: H_0: $\delta_1 = 0$, $\delta_2 = 0$

LIMITAÇÃO DO TESTE RESET

– Uma desvantagem do teste RESET é que ele não fornece orientação prática de como proceder se modelo for rejeitado.

– A equação irrestrita pode conter termos quadráticos e cúbicos, mas também pode conter logaritmos.

– Modelos com logaritmos das variáveis independentes e dependente são fáceis de serem interpretados e suas variáveis tendem a apresentar distribuição normal.

– O teste RESET é um teste da forma funcional, e não um teste de heteroscedasticidade.

TESTES CONTRA ALTERNATIVAS NÃO-ANINHADAS

– Obter testes para outros tipos de má especificação da forma funcional, nos leva para fora do âmbito dos testes de hipótese clássicos.

– Por exemplo, tentar decidir se uma variável independente deveria aparecer em nível:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

– Ou em forma logarítmica:

$$y = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u$$

– Estes são modelos não-aninhados e não podemos usar o teste *F* padrão.

TESTE DE MIZON E RICHARD

– Dois métodos diferentes podem ser usados para modelos não aninhados.

– O primeiro teste foi sugerido por Mizon e Richard (1986).

– Podemos construir um modelo abrangente que contenha cada modelo como um caso especial e, em seguida, testar as restrições que conduziram a cada um dos modelos:

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 log(x_1) + \gamma_4 log(x_2) + u$$

– Podemos primeiro testar H_0: $\gamma_3 = 0$, $\gamma_4 = 0$.

– Podemos também testar H_0: $\gamma_1 = 0$, $\gamma_2 = 0$.

TESTE DE DAVIDSON-MACKINNON

– O segundo é o método de Davidson e MacKinnon (1981), os quais dizem que se esta equação for verdadeira:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

– Então os valores estimados na equação abaixo deveriam ser não significantes na equação acima:

$$y = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u$$

– Para testar a primeira equação, estimamos a segunda equação por MQO e obtemos os valores preditos: $\hat{\hat{y}}$.

– O teste baseia-se na estatística *t* sobre $\hat{\hat{y}}$ na equação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + erro$$

– Se teste *t* de θ é significante, há rejeição da 1ª equação.

– Também podemos fazer o teste inverso.

– Esse teste pode ser usado para testar quaisquer dois modelos não-aninhados com a mesma variável dependente.

PROBLEMAS COM MODELOS NÃO-ANINHADOS

– Não necessariamente um dos modelos será claramente o escolhido, já que ambos os modelos, ou nenhum deles, podem ser rejeitados:
  – Se nenhum for rejeitado, podemos usar o R^2 ajustado para selecionar um deles.
  – Se ambos forem rejeitados, teremos mais trabalho.
  – Se efeitos de importantes variáveis independentes sobre *y* não forem diferentes, não importa qual modelo será usado.
– O teste de Davidson-MacKinnon indica a rejeição de um modelo pela má especificação da forma funcional, mas não necessariamente indica qual o modelo correto.
– É difícil obter testes não-aninhados quando os modelos concorrentes têm variáveis dependentes diferentes.

VARIÁVEIS *PROXY* PARA VARIÁVEIS NÃO-OBSERVADAS

– Um problema mais difícil surge quando um modelo exclui uma variável importante, normalmente devido à não-disponibilidade de dados.

– Se omitirmos uma variável que esteja correlacionada com outra variável independente, os estimadores MQO serão viesados.

– Para resolver o problema de viés de variáveis omitidas de uma equação, podemos obter uma variável *proxy* da variável omitida.

– Uma variável *proxy* é algo que está relacionado com a variável não-observada que gostaríamos de controlar.

– A variável *proxy* não precisa ser a mesma coisa que a variável omitida, mas simplesmente deve ser correlacionada com ela.

EXEMPLIFICAÇÃO DE VARIÁVEIS *PROXY*

– Assumimos que os dados estão disponíveis para y, x_1 e x_2, enquanto a variável x_3^* é não-observada:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

– Temos uma variável *proxy* de x_3^*, que chamamos de x_3, as quais se relacionam desta forma:

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3$$

– O erro v_3 ocorre por x_3^* e x_3 não serem exatamente relacionadas.

– O parâmetro δ_3 mede a relação entre x_3^* e x_3.

– Esperamos que x_3^* e x_3 sejam positivamente relacionadas ($\delta_3>0$).

– Se $\delta_3=0$, então x_3 não é uma *proxy* adequada de x_3^*.

– O intercepto δ_0 pode ser positivo ou negativo, permitindo que x_3^* e x_3 sejam medidas em diferentes escalas.

OPERACIONALIZANDO

– Proposta é simular que x_3^* e x_3 sejam as mesmas, de forma que possamos computar a regressão de y sobre x_1, x_2, x_3.

– O objetivo desta equação é obter boas estimativas dos parâmetros β_1 e β_2.

– Não obteremos estimadores não-viesados de β_0 e β_3.

– Isso é chamado de solução plugada do problema de variáveis omitidas, já que a variável x_3 está “plugada” em x_3^*.

– Se x_3 for verdadeiramente relacionada com x_3^*, essa solução será apropriada.

– Como x_3 e x_3^* não são as mesmas variáveis, devemos determinar quando esse procedimento produzirá estimadores consistentes de β_1 e β_2.

HIPÓTESES

– As hipóteses necessárias para que a solução plugada forneça estimadores consistentes de β_1 e β_2 são:

  – O erro u é não-correlacionado com x_1, x_2 e x_3^*, além de não ser correlacionado com x_3:
    – Ou seja, o valor esperado de u, dadas todas essas variáveis, é zero.

  – O erro v_3 é não-correlacionado com x_1, x_2 e x_3:
    – Supor que v_3 é não-correlacionado com x_1 e x_2 exige que x_3 seja uma boa *proxy* de x_3^*.
    – O valor esperado de x_3^* não depende de x_1 ou de x_2, ou seja, x_3^* tem correlação zero com x_1 e com x_2.

O VIÉS PODE CONTINUAR EXISTINDO

– Se não utilizarmos uma boa *proxy*, os parâmetros β_1 e β_2 continuarão sendo viesados.

– Porém, podemos ter alguma esperança de que esse viés será menor do que se ignorarmos totalmente o problema da variável omitida.

– Variáveis *proxy* também podem aparecer na forma de informação binária para o caso de uma variável dicotômica não-observada.

USO DE VARIÁVEIS DEPENDENTES DEFASADAS

– Quando temos uma idéia de qual fator não-observado devemos controlar, é mais fácil escolher variáveis *proxy*.

– Em alguns casos, suspeitamos que uma ou mais variáveis independentes sejam correlacionadas com uma variável omitida, mas não temos idéia de como obter uma *proxy*.

– Podemos incluir uma variável dependente de um **período anterior** (variável defasada) como variável independente.

– Isso é útil para a **análise de políticas públicas**.

– Uma variável dependente defasada pode ser difícil de ser obtida, mas fornece uma maneira simples de explicar **fatores históricos** que causam diferentes tendências na variável dependente que são difíceis de explicar de outras maneiras.

– Muitos dos mesmos **fatores não-observados** contribuem para os níveis da variável dependente atuais e passados.

– **Efeitos inerciais** também são capturados com defasagens.

IMPORTÂNCIA PARA POLÍTICAS PÚBLICAS

– O uso de uma variável *y* defasada como um método geral para controlar variáveis não-observadas não é uma técnica perfeita.

– Porém, esta prática pode auxiliar na obtenção de uma melhor estimativa dos efeitos de variáveis de políticas de governo (independentes) em diferentes variáveis dependentes.

ERROS DE MEDIDA

– Em alguns casos, não podemos coletar dados da variável que verdadeiramente afetam o comportamento econômico.

– Quando utilizamos uma medida imprecisa de uma variável em um modelo de regressão, nosso modelo conterá um erro de medida.

– O intuito aqui é de estimar as conseqüências do erro de medida para a estimação do MQO e inferir o tamanho do viés.

– O problema do erro de medida tem estrutura estatística similar ao problema da variável omitida e sua substituição pela variável *proxy*.

VARIÁVEL *PROXY* ≠ ERRO DE MEDIDA

– Porém, o problema da variável omitida e do erro de medida são conceitualmente diferentes.

– No caso da variável *proxy*, procuramos uma variável que é associada à variável não-observada:

  – A idade é uma *proxy* de experiência, por exemplo.

  – O efeito parcial da variável omitida não é de interesse central.

– No caso do erro de medida, a variável que não observamos tem significado quantitativo bem definido, mas as medidas sobre elas podem conter erros:

  – A poupança anual registrada é diferente da poupança anual real, por exemplo.

  – A variável independente mal medida é a de maior interesse.

ERRO DE MEDIDA NA VARIÁVEL DEPENDENTE

– Vamos chamar de *y** a variável na população que queremos explicar:

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

– O erro de medida na população é definido como a diferença entre o valor observado e o valor real ($e_0 = y - y^*$).

– O modelo que pode ser estimado é dado por *y*, que é a medida observável de *y** na população:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u + e_0$$

– Na verdade, simplesmente ignoramos o fato de que *y* é uma medida imperfeita de *y** e prosseguimos da maneira habitual.

QUANDO *y* AO INVÉS DE *y** PRODUZ β_j CONSISTENTES?

– Como o modelo original satisfaz as hipóteses de RLM, *u* tem média zero e é não-correlacionado com cada x_j.

– É natural assumir que o erro de medida tem média zero:
  – Se não for assim, teremos um estimador viesado do intercepto β_0, o que não é motivo de preocupação.

– Mais importante é a suposição de que o erro de medida (e_0) é estatisticamente independente das variáveis explicativas (x_j):
  – Se isso for verdade, então os estimadores MQO de *y* em lugar de *y** são não-viesados e consistentes.
  – Além disso, os procedimentos de inferência do método MQO (estatísticas *t*, *F* e *LM*) são válidos.

PONTO PRINCIPAL

– Se e_0 e u forem não-correlacionados, então:

$$\text{Var}(u + e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$$

– Isso significa que o erro de medida na variável dependente resulta em uma variância de erro maior do que quando não ocorre nenhum erro.

– Isso produz variâncias maiores dos estimadores MQO, ou seja, maiores erros-padrão e menores estatísticas *t*.

– A única forma de evitar esse problema é coletar dados melhores.

– O ponto principal é que o erro de medida na variável dependente pode causar vieses no método MQO se ele for sistematicamente relacionado com uma ou mais variáveis explicativas:

  – Se erro de medida for aleatório, o método MQO possuirá boas propriedades e é perfeitamente apropriado.

ERRO DE MEDIDA EM UMA VARIÁVEL EXPLICATIVA

– O erro de medida em uma variável explicativa tem sido considerado um problema mais importante do que o erro de medida em uma variável dependente.

– Um modelo de regressão simples que satisfaz as hipóteses RLM produz estimadores de β_0 e β_1 não-viesados e consistentes:

$$y = \beta_0 + \beta_1 x_1^* + u$$

– O problema é que x_1* não é observado.

– Por exemplo, ao invés da verdadeira renda (x_1*), temos somente a renda declarada (x_1).

– O erro de medida na população é: $e_1 = x_1 - x_1^*$.

– Assumimos que o erro de medida médio na população é zero: E(e_1) = 0.

– Além disso, assumimos que u é não-correlacionado com x_1* e x_1.

SUBSTITUINDO x_1* POR x_1

– Queremos saber as propriedades de MQO se substituirmos x_1* por x_1 e computarmos a regressão de y sobre x_1:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

– As propriedades dependerão das suposições que fizermos sobre o erro de medida (e_1).

– Duas hipóteses opostas têm sido enfatizadas na literatura econométrica.

PRIMEIRA HIPÓTESE

– A primeira hipótese é que e_1 é não-correlacionado com a medida observada (x_1): Cov(x_1,e_1) = 0.

– Dado que $e_1 = x_1 - x_1^*$, se esta hipótese for verdadeira, então e_1 deve ser correlacionado com a variável não-observada x_1^*.

– Como assumimos que tanto u quanto e_1 têm média zero e são não-correlacionados com x_1, então $u - \beta_1 e_1$ tem média zero e é não-correlacionado com x_1.

– Então, a estimação de MQO com x_1 em lugar x_1^* produz um estimador consistente de β_0 e β_1.

– Exceto quando β_1=0, o erro de medida aumenta a variância do erro.

– Isso não afeta nenhuma das propriedades MQO, mas as variâncias dos betas estimados (e os erros-padrão) serão maiores do que se observarmos x_1^* diretamente.

SEGUNDA HIPÓTESE

– A hipótese anterior de que e_1 é não-correlacionada com x_1 é análoga à hipótese da variável *proxy*.

– Porém, os econometristas geralmente supõem que o **erro clássico nas variáveis (ECV)** é o erro de medida não-correlacionado com a variável explicativa *não-observada* (x_1*): Cov(x_1*,e_1) = 0.

– Neste caso, a medida observada é a soma da variável explicativa verdadeira com o erro de medida: $x_1 = x_1^* + e_1$.

– Supomos que u é não-correlacionado com: x_1*, x_1, e_1.

– Se esta hipótese for verdadeira, então e_1 será correlacionado com a variável observada x_1.

– Neste caso, a regressão de MQO de y sobre x_1 produz um estimador viesado e inconsistente.

– Se a variância de x_1* for grande, com relação à variância em e_1, o erro de medida não causará grandes vieses.

E NO CASO DE REGRESSÃO MÚLTIPLA?

– Ao adicionarmos mais duas variáveis explicativas (x_2 e x_3) e a primeira variável é medida com erro (x_1*), supomos que u é não-correlacionado com x_1*, x_2, x_3 e x_1.

– A hipótese crucial refere-se ao erro de medida (e_1).

– Assume-se que e_1 é não-correlacionado com x_2 e x_3.

– Se e_1 for não-correlacionado com x_1, então a regressão MQO de y sobre x_1, x_2 e x_3 produzirá estimadores consistentes.

– Porém, sob a hipótese ECV, o MQO será viesado e inconsistente, pois e_1 é correlacionado com x_1.

– No caso em que x_1* é não-correlacionado com x_2 e x_3, β_2 e β_3 estimados são consistentes.

– Porém, geralmente o erro de medida em uma única variável provoca inconsistência em todos os estimadores.

ERRO DE MEDIDA EM MAIS DE UMA VARIÁVEL

– O erro de medida pode estar presente em mais de uma variável explicativa e também na variável dependente.

– Qualquer erro de medida na variável dependente é usualmente assumido como não-correlacionado com todas as variáveis explicativas, seja ele observado ou não.

– Porém, o viés nos estimadores MQO no caso da hipótese ECV é complicado e não leva a resultados claros, mas a primeira hipótese não é melhor ou pior que a segunda.

– Se e_1 for correlacionado com x_1^* e x_1, MQO é inconsistente.

– Calcular as implicações do erro de medida que não satisfaçam $\text{Cov}(x_1,e_1)=0$ (primeira hipótese) ou $\text{Cov}(x_1^*,e_1)=0$ (segunda hipótese) é difícil de realizar.

– Os estimadores podem ser consistentemente estimados na presença de erros de medida com uso de variáveis instrumentais.

RESUMINDO O ECV

– Sob as hipóteses do erro clássico nas variáveis (ECV), o erro de medida na variável dependente não tem efeito nas propriedades estatísticas do MQO.

– Sob as hipóteses ECV para uma variável independente, o estimador MQO do coeficiente na variável mal medida é viesado em direção a zero.

– O viés nos coeficientes das outras variáveis pode ser para qualquer lado e é difícil de ser determinado.

PROBLEMAS DE AMOSTRAGEM NÃO-ALEATÓRIA

– O problema do erro de medida pode ser visto como um problema de dados, já que não podemos obter dados sobre as variáveis de interesse.

– Sob o modelo clássico de erro nas variáveis (ECV), o termo erro é correlacionado com a variável dependente mal medida.

– Outro problema discutido em capítulos anteriores é a multicolinearidade entre as variáveis explicativas.

– Quando duas variáveis independentes são altamente correlacionadas, pode ser difícil estimar o efeito parcial de cada uma delas.

– Agora serão discutidos os problemas de dados que podem violar a hipótese de amostragem aleatória (RLM.2).

AUSÊNCIA DE DADOS (*MISSING DATA)*

– O problema de ausência de dados pode surgir de várias formas.

– Muitas vezes coletamos uma amostra aleatória e mais tarde descobrimos que estão faltando informações de algumas variáveis importantes para diversas unidades na amostra.

– Quando estão faltando dados de uma observação na variável dependente ou em uma das variáveis independentes, a observação não pode ser usada em uma análise de regressão múltipla padrão.

– Os programas de computador simplesmente ignoram as observações ao calcularem as estimativas.

– Há conseqüências estatísticas provocadas pela ausência de dados?

CORREÇÃO DA AUSÊNCIA DE DADOS

– Se estes dados estiverem faltando aleatoriamente, então o tamanho da amostra aleatória disponível da população será simplesmente reduzido.

– Embora isso torne os estimadores menos precisos, não haverá a produção de nenhum viés e a hipótese de amostragem aleatória (RLM.2) ainda é válida.

– Existem maneiras de usar informações das observações nas quais somente algumas variáveis estão faltando (imputação de dados), mas na prática não se faz isso com freqüência.

– A melhoria nos estimadores normalmente é pequena, embora o método seja complicado.

– O IBGE realizou imputação para os dados do Censo Demográfico de 2000.

– Na maioria dos casos, simplesmente ignoramos as observações que representam falta de informação.

AMOSTRAS NÃO-ALEATÓRIAS NAS INDEPENDENTES

– A ausência de dados é mais problemática quando resulta de uma amostra não-aleatória da população.

– Se há omissão de dados para um conjunto específico da população, a hipótese de amostragem aleatória está sendo violada e devemos nos preocupar com suas conseqüências.

– Certos tipos de amostragens não-aleatórias não causam viés ou inconsistência no MQO, ao escolher a amostra com base nas **variáveis independentes**.

– A seleção da amostra com base nas variáveis independentes é um exemplo de seleção amostral **exógena**.

– O MQO na amostra não-aleatória é não-viesado, porque a regressão é a mesma nos sub-conjuntos da população.

– Desde que haja variação suficiente nas variáveis independentes na subpopulação, essa seleção não será um problema sério, mas resultará em estimadores ineficientes.

AMOSTRAS NÃO-ALEATÓRIAS NA DEPENDENTE

– A seleção de amostra com base na variável dependente (*y*) é um exemplo de seleção amostral **endógena**.

– Se a amostra tiver como base o fato de a variável dependente estar acima ou abaixo de determinado valor, sempre ocorrerá viés no MQO, ao estimarmos o modelo populacional.

– Os parâmetros serão viesados e inconsistentes porque a regressão populacional não é a mesma que o valor predito da variável dependente coletada.

AMOSTRAS NÃO-ALEATÓRIAS POR ESTRATIFICAÇÃO

– Outros desenhos de amostra levam a amostras não-aleatórias da população, em geral intencionalmente.

– Um método comum de coleta de dados é a **amostragem estratificada**, na qual a população é dividida em grupos não sobrepostos (sexo, raça, escolaridade...).

– Alguns grupos podem aparecer com mais freqüência do que a determinada por sua representação populacional.

– Superdimensionar um grupo que seja relativamente pequeno na população é comum na coleta de amostras estratificadas. O mesmo é feito para grupos de baixa renda.

– O MQO é não-viesado e consistente quando a estratificação é feita com relação a uma variável explicativa.

– Se superdimensionarmos um grupo populacional pela variável dependente, o MQO não estimará consistentemente os parâmetros, porque a estratificação é endógena.

OBSERVAÇÕES EXTREMAS OU ATÍPICAS (*OUTLIERS*)

– As estimativas MQO podem ser influenciadas por uma ou diversas observações extremas ou atípicas (*outliers*).

– Uma observação é extrema se sua eliminação da análise de regressão produzir mudança significativa nas estimativas.

– O MQO é suscetível a observações extremas, porque minimiza a soma dos quadrados dos resíduos.

– Ou seja, grandes resíduos recebem muita carga no problema de minimização de mínimos quadrados.

TEORIA E PRÁTICA

- Teoricamente, no problema de observações extremas:
  - Os dados são vistos como provenientes de uma amostra aleatória de determinada população, mas que tem uma distribuição pouco comum com valores extremos.
  - Presume-se que tais observações provêm de uma população diferente.

- Na prática, tais observações podem ocorrer porque:
  - Houve um engano na entrada dos dados, os quais podem ser detectados com análise de estatísticas descritivas.
  - Ao fazer a amostragem de uma pequena população, alguns membros foram muito diferentes dos demais.

O QUE FAZER?

– Pode ser difícil tomar a decisão de manter ou não *outliers*.

– Não é bom definir uma observação extrema por possuir o maior resíduo em uma regressão, porque eliminar essa observação não alterará muito os resultados.

– O ideal é definir um *outlier* pelos gráficos de dispersão das variáveis independentes e dependente observadas.

– Observações extremas podem fornecer informações importantes ao aumentar a variação das variáveis explicativas, o que reduz os erros-padrão.

– A regressão pode ser apresentada **com e sem as observações extremas**, nos casos em que um ou vários pontos dos dados alteram substancialmente os resultados.

– A **transformação logarítmica** também pode ser usada, já que estreita a amplitude dos dados (diminui a força dos *outliers*) e produz estimativas mais fáceis de interpretar.

MÍNIMOS DESVIOS ABSOLUTOS (MDA)

– Ao invés de tentar encontrar observações extremas no dados antes de aplicar o MQO, podemos usar um método de estimação menos sensível aos *outliers*.

– Um desses métodos é o de **mínimos desvios absolutos (MDA)**, o qual minimiza a soma dos desvios absolutos dos resíduos, em lugar da soma dos resíduos quadrados.

– O MDA foi construído para estimar os efeitos da variáveis explicativas sobre a mediana condicional, em vez da média condicional da variável dependente.

– Como a mediana não é afetada por grandes alterações em observações extremas, as estimativas do MDA são resistentes aos *outliers*.

– O MQO atribui mais importância a grandes resíduos, pois cada resíduo é quadrado.

INCONVENIÊNCIAS DO MDA

– Não existem fórmulas para os estimadores, os quais só podem ser encontrados com o uso de métodos iterativos. Mesmo com computadores, esse cálculo é demorado com grandes amostras e com muitas variáveis explicativas.

– As inferências estatísticas são justificadas apenas para amostras grandes, o que dificulta análises de bancos de dados menores.

– Nem sempre estima consistentemente os parâmetros que aparecem na função de média condicional, já que foi construído sobre a mediana condicional.

– Diferença de estimativas de MQO e MDA podem ocorrer por diferenças entre média e mediana (distribuições assimétricas), e não pela existência de *outliers*.

ANÁLISE DE REGRESSÃO LOGÍSTICA

VARIÁVEL DEPENDENTE BINÁRIA

– O modelo de regressão logístico é utilizado quando a variável resposta é qualitativa com dois resultados possíveis.

– Probabilidade de sucesso = p

– Probabilidade de fracasso = 1 – p = q

– Chance = (prob. de sucesso) / (prob. de fracasso)

– Por exemplo, se a probabilidade de sucesso é 0,75, a chance é igual a:

$$p / (1 - p) = p / q = 0,75 / 0,25 = 3$$

RAZÃO DE CHANCES

– Razão de chances para variáveis dependentes binárias é a razão entre a chance de uma linha (ou coluna) de uma tabela 2x2, dividida pela chance da outra linha (ou coluna):

$$\frac{A}{B} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1 * (1-p_2)}{p_2 * (1-p_1)}$$

EXEMPLO DE CÁLCULO DE RAZÃO DE CHANCES

Sexo	Dilma	Serra	Total
Homem	52	39	91
Mulher	43	44	87
Total	**95**	**83**	**178**

- **Chance** de votar na Dilma entre homens:

$$p_1 / (1-p_1) = (52/91) / (39/91) = 0,57 / 0,43 = 1,33$$

- **Chance** de votar na Dilma entre mulheres:

$$p_2 / (1-p_2) = (43/87) / (44/87) = 0,49 / 0,51 = 0,96$$

- **Razão de chances** de votar na Dilma entre homens, em relação às mulheres:

$$[p_1 /(1- p_1)] / [p_2 /(1- p_2)] = 1,33 / 0,96 = 1,39$$

FUNÇÃO DE RESPOSTA
QUANTO VARIÁVEL DEPENDENTE É BINÁRIA

– Vamos considerar o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

– A resposta esperada é dada por:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

– Na regressão logística, Y_i possui uma distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i$$

LOGITO

– O logito (*logit*) equivale ao logaritmo natural (base *e*) da chance:

$$logit(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p)$$

– A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$p = \frac{exp(x)}{1 + exp(x)}$$

RAZÃO DE CHANCES (*ODDS RATIO*)

– Compara a chance de sucesso de um grupo em relação a outro grupo:

$$log(R) = log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$$

$$log(R) = log\left(\frac{p_1}{1-p_1}\right) - log\left(\frac{p_2}{1-p_2}\right)$$

$$log(R) = logit(p_1) - logit(p_2)$$

– Portanto, a diferença entre os logitos de duas probabilidades equivale ao logaritmo da razão de chances.

RAZÃO DE CHANCES (*ODDS RATIO*)

$$\frac{A}{B} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{exp(\beta_0 + \gamma)}{exp(\beta_0)} = \frac{exp(\beta_0) * exp(\gamma)}{exp(\beta_0)} = exp(\gamma)$$

– Razão de chance é dada pela expressão exp(γ): chance de sucesso no grupo A, em relação ao grupo B.

– Se exp(γ) for maior que uma unidade, chance de sucesso em A é maior que em B.
  – Ex.: exp(γ)=1,17, chance de sucesso em A é 1,17 vezes maior do que em B, ou seja, é 17% maior do que em B.

– Se exp(γ) for menor que uma unidade, chance de sucesso em A é menor que em B.
  – Ex.: exp(γ)=0,61, chance de sucesso em A é 0,61 vezes a chance de B, ou seja, é 39% menor do que em B.

DEFINIÇÃO DO VALOR ESPERADO

– Pela definição de valor esperado, obtemos:

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

– Assim, a resposta média, quando a variável resposta é uma variável binária (1 ou 0), representa a probabilidade de $Y = 1$, para o nível da variável independente X_i.

REGRESSÃO LOGÍSTICA
COM UMA VARIÁVEL INDEPENDENTE

– Considerações teóricas e práticas sugerem que quando a variável resposta é binária, a forma da função resposta será frequentemente curvilínea.

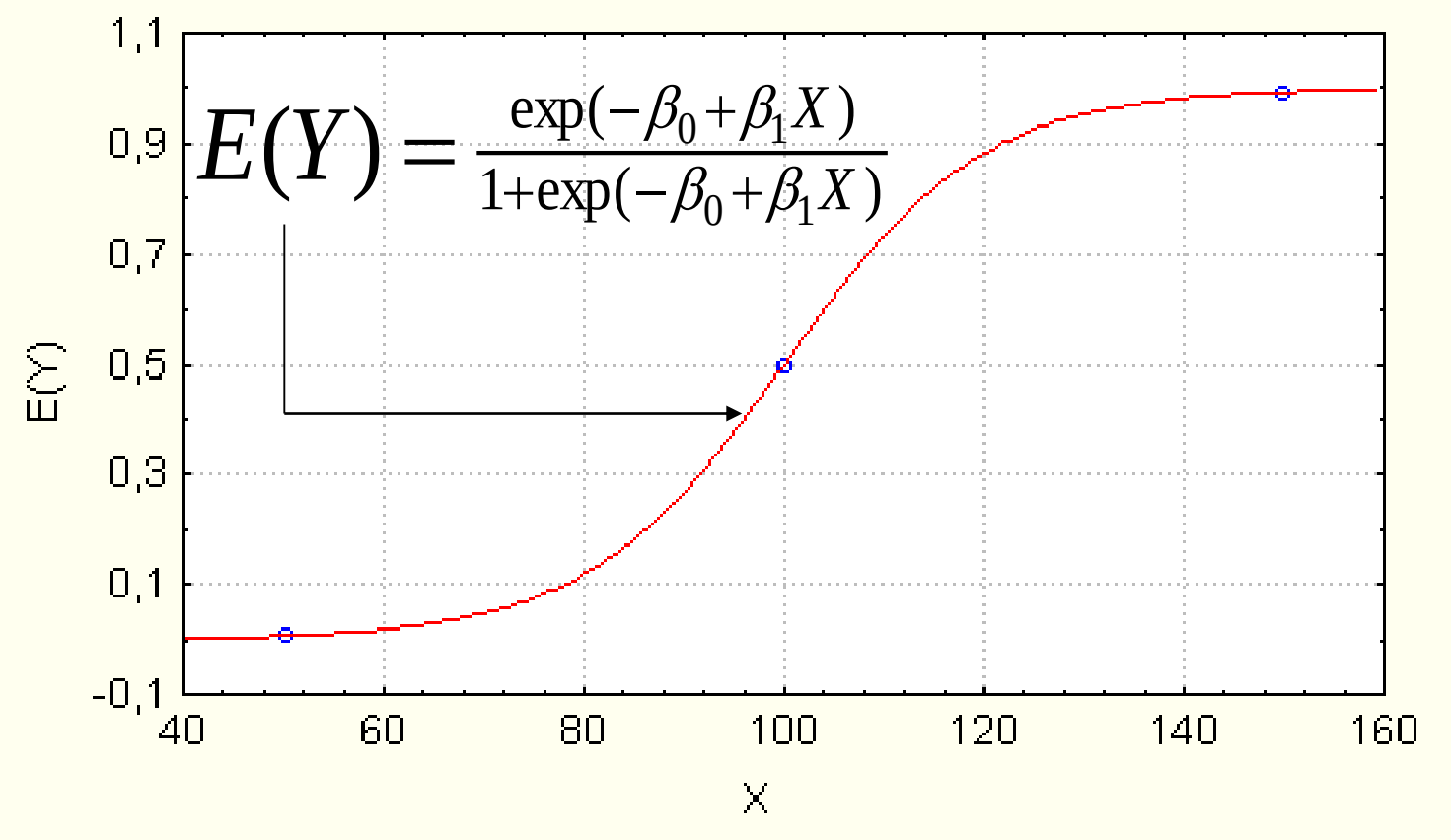
– As funções respostas (valores preditos) das figuras são denominadas funções logísticas, cuja expressão é:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

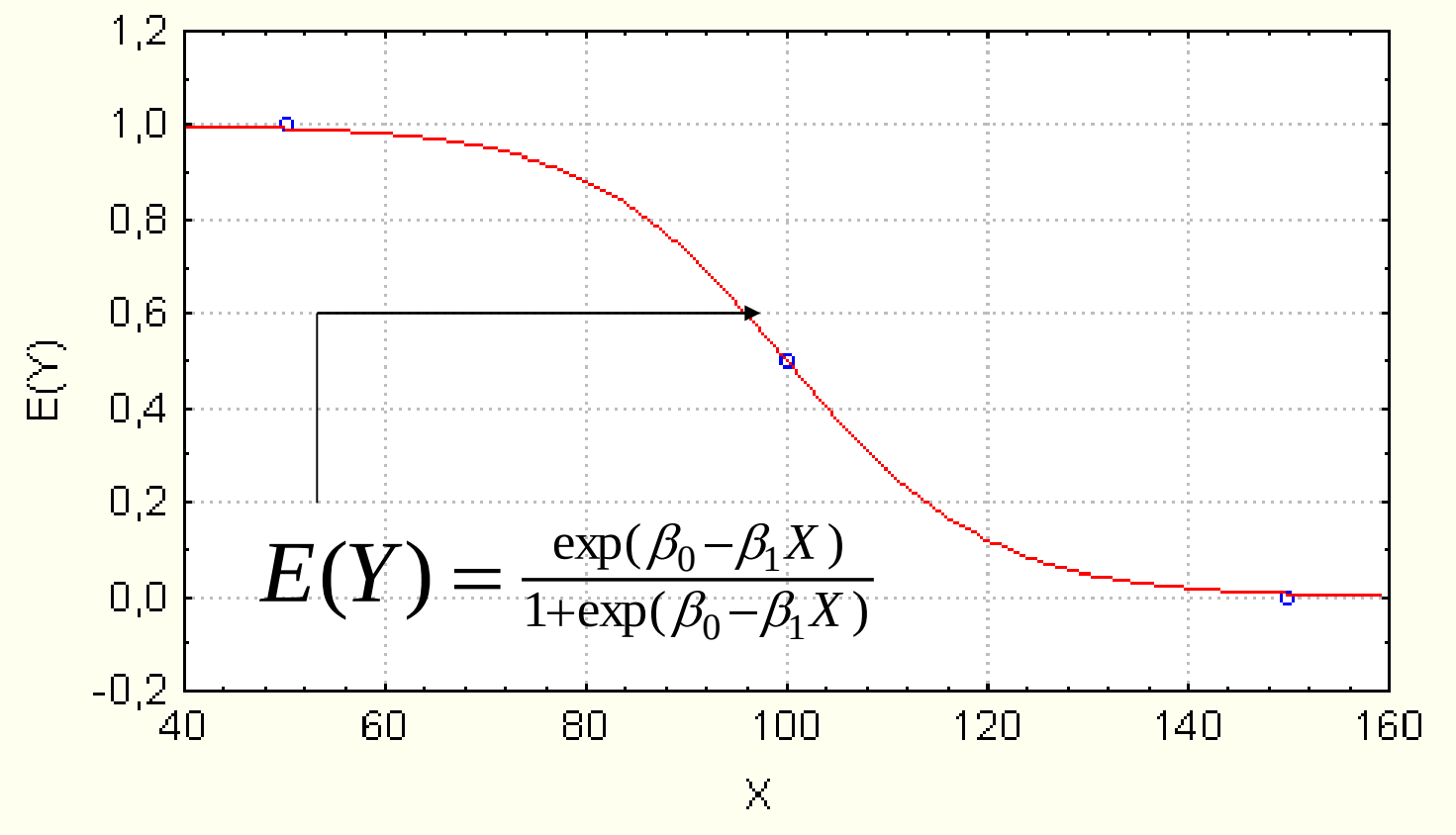
– Forma equivalente:

$$E(Y) = \left[1 + \exp\left(-\beta_0 - \beta_1 X\right)\right]^{-1}$$

VARIÁVEL DEPENDENTE ESTIMADA PELA VARIÁVEL INDEPENDENTE OBSERVADA



VARIÁVEL DEPENDENTE ESTIMADA PELA VARIÁVEL INDEPENDENTE OBSERVADA



REGRESSÃO LOGÍSTICA
COM MAIS DE UMA VARIÁVEL INDEPENDENTE

– Função com uma variável independente:

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)}$$

– Função com uma série de variáveis independentes:

$$E(Y) = \frac{\exp(\boldsymbol{\beta}'\mathbf{X})}{1+\exp(\boldsymbol{\beta}'\mathbf{X})}$$

– Uma forma equivalente é dada por:

$$E(Y) = (1+\exp(-\boldsymbol{\beta}'\mathbf{X}))^{-1}$$

EQUAÇÃO DE REGRESSÃO

– A parte linear da equação da regressão logística é usada para encontrar a probabilidade de estar em uma categoria, baseado na combinação de variáveis independentes.

$$\bar{Y}_i = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

– Os coeficientes de regressão e seus erros padrões são estimados com métodos de máxima verossimilhança.

AJUSTANDO O MODELO

– A função log-verossimilhança estende-se diretamente para o modelo de regressão logística múltipla, dada por:

$$\log_e L(\beta) = \sum_{i=1}^{n} Y_i(\boldsymbol{\beta}'\mathbf{X}_i) - \sum_{i=1}^{n} \log_e(1+\exp(\boldsymbol{\beta}'\mathbf{X}_i))$$

– Métodos numéricos devem ser utilizados para encontrar os valores de $\beta_0, \beta_1,\ldots, \beta_{p-1}$ para maximizar a expressão.

– As estimativas de máxima verossimilhança serão denotadas por $b_0, b_1,\ldots,b_{p-1}$.

– A função resposta logística ajustada e os valores ajustados são dados por:

$$\hat{\pi} = \frac{\exp(\mathbf{b}'\mathbf{X})}{1+\exp(\mathbf{b}'\mathbf{X})} = (1+\exp(-\mathbf{b}'\mathbf{X}))^{-1}$$

$$\hat{\pi}_i = \frac{\exp(\mathbf{b}'\mathbf{X}_i)}{1+\exp(\mathbf{b}'\mathbf{X}_i)} = (1+\exp(-\mathbf{b}'\mathbf{X}_i))^{-1}$$

ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA

– Não existe uma solução analítica para os valores β_0 e β_1 que maximizam a função de verossimilhança.

– Métodos numéricos são necessários para encontrar as estimativas de máxima verossimilhança, b_0 e b_1.

– Encontradas as estimativas b_0 e b_1, substitui-se esses valores para encontrar os valores ajustados.

– O valor ajustado para o *i-ésimo* valor é dado por:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$$

– Se usarmos a transformação *logit*, a função é:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

– A função de resposta ajustada é dada por:

$$\hat{\pi}' = b_0 + b_1 X \quad \text{onde:} \quad \hat{\pi}' = \log_e\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

TESTE DE QUI-QUADRADO DA RAZÃO DE VEROSSIMILHANÇA

– Logaritmo da verossimilhança (*Log-likelihood*):

$$\log - likelihood = \sum_{i=1}^{N} [Y_i \ln(\hat{Y}_i) + (1 - Y_i) \ln(1 - \hat{Y}_i)]$$

– Modelos são comparados com uso dos logaritmos das verossimilhanças dos modelos:

X^2 = – 2 [(*log-likelihood* do modelo restrito)

– (*log likelihood* do modelo irrestrito)]

ou

X^2 = 2 [(*log-likelihood* do modelo irrestrito)

– (*log likelihood* do modelo restrito)]

– Modelos precisam ser aninhados para comparação, ou seja, todas variáveis independentes do menor modelo (restrito) devem estar incluídas no maior modelo (irrestrito).

MAIS TESTE DE QUI-QUADRADO

– O teste de qui-quadrado da razão da verossimilhança é igual ao ajuste do modelo restrito (–2*log. da verossimilhança do modelo anterior) menos o ajuste do modelo irrestrito (–2*log. da versossimilhança do modelo atual).

– O logaritmo da verossimilhança multiplicado por –2 é usado para testar hipóteses entre modelos aninhados, sendo que seu valor não tem um significado específico.

– Esta razão é testada em uma distribuição de qui-quadrado, levando em consideração a diferença entre os graus de liberdade (número de variáveis independentes do modelo irrestrito menos o número de variáveis independentes do modelo restrito).

– Se o teste de qui-quadrado é significante, é afirmado que o modelo irrestrito não pode ter redução de variáveis independentes, dado um nível de significância específico.

TESTE DE QUI-QUADRADO NO STATA (automático)

– Testando hipóteses entre modelos aninhados no STATA, com base no logaritmo da verossimilhança:

```
***Modelo 1
logit y x1 x2
outreg2 using "C:\curso\tabelas\modelo.doc", replace eform word
estimates store m1

***Modelo 2
logit y x1 x2 x3 x4 x5
outreg2 using "C:\curso\tabelas\modelo.doc", append eform word
estimates store m2

***Teste de qui-quadrado para impacto conjunto de x3, x4 e x5
***H0: B3 = B4 = B5 = 0 (impactos de x3, x4 e x5 são iguais a zero)
lrtest m1 m2
```

TESTE DE QUI-QUADRADO NO STATA (manual)

– Se modelos utilizarem peso, teste de qui-quadrado deve ser calculado manualmente:

```
***Modelo 1
logit y x1 x2 [pweight=peso]
***Modelo 2
logit y x1 x2 x3 x4 x5 [pweight=peso]
```

– Multiplicar log. da verossimilhança dos modelos por –2:

```
***Modelo 1: log-likelihood = a
di -2*a
***Modelo 2: log-likelihood = b
di -2*b
```

– Calcular razão dos logs = (–2*log-likelihood do modelo1) menos (–2*log-likelihood do modelo 2):

```
di (-2*a) - (-2*b)
```

– Teste de qui-quadrado entre modelos usa razão dos logs acima (???) e diferença entre graus de liberdade (3) que é o número de independentes inseridas no segundo modelo:

```
tablesq X 3 ???
```

TESTE DE WALD

– Cada coeficiente é avaliado usando o teste de Wald, que é simplesmente um teste de escore z:

$$W_j = \frac{\beta_j}{EP_{\beta_j}}$$

– Os testes dos coeficientes são aproximadamente escores z, os quais são posteriormente elevados ao quadrado, fazendo com que esta estatística tenha distribuição de qui-quadrado.

– Esse teste é usado para avaliar a significância de cada coeficiente (β) no modelo.

– O teste de Wald é conhecido por ser conservador (aumenta o erro II).

ERROS TIPO I E TIPO II

– Ao testar H_0, chegamos a uma conclusão de rejeitá-la ou de deixar de rejeitá-la.

– Tais conclusões pode estar corretas ou erradas.

		Estado verdadeiro da natureza	
		A hipótese nula é verdadeira	A hipótese nula é falsa
Decisão	Decidimos rejeitar a hipótese nula.	**Erro tipo I** (rejeitar uma hipótese nula verdadeira) α	Decisão Correta
	Deixamos de rejeitar a hipótese nula	Decisão Correta	**Erro tipo II** (deixar de rejeitar uma hipótese nula falsa) β

– α: probabilidade de erro tipo I (probabilidade de rejeitar hipótese nula quando ela é verdadeira).

– β: probabilidade de erro tipo II (probabilidade de deixar de rejeitar hipótese nula quando ela é falsa).

PSEUDO R^2

– Há várias medidas de associação que pretendem servir como um R^2 na regressão logística.

– Porém, nenhuma destas medidas é realmente o R^2.

– A interpretação não é a mesma, mas eles podem ser vistos como uma aproximação da variação na variável dependente, devido à variação nas variáveis independentes.

– Para comparação de grau de ajuste entre modelos é mais apropriado fazer o teste de qui-quadrado da razão da verossimilhança.

MODELO LOGÍSTICO MULTINOMIAL

– É possível estimar uma regressão logística em que a variável dependente tem mais de duas categorias.

– Ou seja, o modelo logístico pode ser estendido quando a variável resposta qualitativa tem mais do que duas categorias.

– Por exemplo, posicionamento ideológico: esquerda, centro, direita.

– São geradas k – 1 equações, sendo k o número de categorias.

– As equações geram probabilidades para predizer se uma categoria está acima/abaixo da categoria de referência.

EXEMPLO DE MODELO LOGÍSTICO

IMPACTO DO BOLSA FAMÍLIA SOBRE ABANDONO ESCOLAR

–Banco de dados de Avaliação de Impacto do Programa Bolsa Família (AIBF) de 2005 do Ministério do Desenvolvimento Social e Combate à Fome (MDS).

–Modelos logísticos foram estimados para três grupos de domicílios, segundo limites máximos da renda domiciliar per capita:

1) R$50,00: população com piores condições sócio-econômicas.

2) R$100,00: limite oficial de renda definido para elegibilidade ao PBF.

3) R$200,00: garante representatividade amostral em todos grupos.

VARIÁVEL DEPENDENTE

- Variável dependente indica se a criança abandonou a escola entre 2004 e 2005:
  - No ano passado, frequentava escola ou creche?
  - Frequenta escola ou creche atualmente?

- Foi realizada análise multivariada, controlando as estimativas por características do domicílio, mãe e criança.

VARIÁVEIS INDEPENDENTES DE DOMICÍLIO

- Número de membros da família.
- Presença de idosos.
- Presença de rede geral de água.
- Iluminação elétrica.
- Serviço de coleta de lixo.
- Domicílio em zona urbana ou rural.
- Região de residência (Sul/Sudeste; Norte/Centro-Oeste; Nordeste).

VARIÁVEIS INDEPENDENTES DA MÃE

–Indicação se mãe é chefe do domicílio.

–Cor/raça.

–Anos de escolaridade.

–Idade.

–Residia há menos de 10 anos no município.

–Participação em organizações sociais.

–Horas de trabalho por semana.

–Tempo gasto em cuidados com a casa por dia.

DEMAIS VARIÁVEIS INDEPENDENTES

Variáveis independentes da criança:

– Idade da criança.

– Indicação se criança trabalha.

– Mãe reside no domicílio.

Beneficiário do Programa Bolsa Família:

– Indicação se criança reside em domicílio que recebe o benefício.

DESCRIÇÃO DA AMOSTRA

– Distribuição percentual de crianças por grupos de renda domiciliar per capita e recebimento do benefício.

Programa Bolsa Família	Limite de renda domiciliar per capita		
	R$50,00	R$100,00	R$200,00
Sim	68,39%	64,71%	59,75%
Não	31,61%	35,29%	40,25%
Nº casos (n)	3.312	6.761	9.232

Fonte: AIBF/MDS (2005).

DISTRIBUIÇÃO DA VARIÁVEL DEPENDENTE

– Percentual de crianças que abandonaram a escola entre 2004 e 2005 por grupo de renda e recebimento do benefício.

Programa Bolsa Família	**Limite de renda domiciliar per capita**		
	R$50,00	**R$100,00**	**R$200,00**
Sim	1,10%	1,42%	1,30%
Não	2,39%	1,97%	1,80%
Diferença	1,28%***	0,55%***	0,50***

***Significativo ao nível de confiança de 99%.
Fonte: AIBF/MDS (2005).

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis de domicílio			
Nº de membros da família	1,122	1,124***	1,108***
Idosos no domicílio	1,454	1,678	1,331
Rede de água	1,066	0,767	0,694*
Iluminação elétrica	1,270	1,106	1,293
Coleta de lixo	0,994	0,756	0,621**
Rural	ref.	ref.	ref.
Urbano	1,729	1,910*	2,309***
Sul/Sudeste	ref.	ref.	ref.
Norte/Centro-Oeste	2,536**	1,889**	1,630**
Nordeste	3,035**	2,248***	2,064***

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da mãe			
Mãe é chefe do domicílio	1,974***	1,445*	1,508**
Preta/Parda	ref.	ref.	ref.
Branca	2,248**	2,029***	1,465**
0 anos de estudo	ref.	ref.	ref.
1-4 anos de estudo	1,267	1,195	1,135
5-8 anos de estudo	0,701	0,898	0,902
9+ anos de estudo	0,251*	0,440*	0,481*
0-24 anos	1,507	4,757***	4,534***
25-34 anos	ref.	ref.	ref.
35-49 anos	1,170	1,111	1,109
50+ anos	0,053***	0,532	0,645

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da mãe			
<10 anos no município	1,325	1,411	1,838***
Participa org. social	0,731	0,643*	0,565***
0 hora/semana trabalho	ref.	ref.	ref.
1-20 horas/semana trabalho	0,257*	0,920	1,177
21-39 horas/semana trabalho	0,736	0,744	0,907
40+ horas/semana trabalho	0,904	1,790**	1,529*
0-2 hora/dia trab. casa	ref.	ref.	ref.
3-4 hora/dia trab. casa	2,975	1,089	0,854
5-6 hora/dia trab. casa	2,399	1,241	1,050
7+ hora/dia trab. casa	2,084	1,563	1,443

RAZÕES DE CHANCES DA CRIANÇA TER ABANDONADO A ESCOLA ENTRE 2004 E 2005 (cont.)

Variáveis independentes	R$50,00	R$100,00	R$200,00
Variáveis da criança			
Idade	1,174**	1,226***	1,194***
Criança trabalha	1,417	1,177	1,465
Mãe reside no domicílio	0,218***	0,455**	0,610*
Beneficiário do Programa Bolsa Família	**0,428*** **	**0,662** **	**0,666** **
Número de casos (crianças)	3.312	6.761	9.232

*Significativo ao nível de 90%; **Significativo ao nível de 95%; ***Significativo ao nível de 99%.
Fonte: AIBF/MDS (2005).

USO DE PESOS PARA CONSIDERAR PLANO AMOSTRAL

DIFERENTES PESOS

Indivíduo	Número de observações coletadas na amostra	Peso para expandir para o tamanho da população (N)	Peso para manter o tamanho da amostra (n)
João	1	4	0,8
Maria	1	6	1,2
Total	2	10	2

EXEMPLO:
Peso amostral do João =
Peso de frequência do João * (Peso amostral total / Peso de frequência total)

PESO DE FREQUÊNCIA NO STATA

- **FWEIGHT**:
  - Expande os resultados da amostra para o tamanho populacional.
  - Utilizado em tabelas para gerar frequências.
  - O uso desse peso é importante na amostra do Censo Demográfico e na Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE) para expandir a amostra para o tamanho da população do país, por exemplo.
  - Somente pode ser usado em tabelas de frequência quando o peso é uma variável discreta (não decimal).

```
tab x [fweight = peso]
```

PESO AMOSTRAL PARA PROGRAMADORES NO STATA

– **IWEIGHT**:
  – Não tem uma explicação estatística formal.
  – Esse peso é utilizado por programadores que precisam implementar técnicas analíticas próprias.
  – Pode ser utilizado em tabelas de frequência, mesmo que o peso seja decimal.

```
tab x [iweight = peso]
```

PESO AMOSTRAL ANALÍTICO NO STATA

- **AWEIGHT**:
  - Inversamente proporcional à variância da observação.
  - Número de observações na regressão é escalonado para permanecer o mesmo que o número no banco.
  - Utilizado para estimar uma regressão linear quando os dados são médias observadas, tais como:

group	x	y	n
1	3.5	26.0	2
2	5.0	20.0	3

  - Ao invés de:

group	x	y
1	3	22
1	4	30
2	8	25
2	2	19
2	5	16

UM POUCO MAIS SOBRE O AWEIGHT

– De uma forma geral, não é correto utilizar o **AWEIGHT** como um peso amostral, porque as fórmulas utilizadas por esse comando assumem que pesos maiores se referem a observações medidas de forma mais acurada.

– Uma observação em uma amostra não é medida de forma mais cuidadosa que nenhuma outra observação, já que todas fazem parte do mesmo plano amostral.

– Usar o **AWEIGHT** para especificar pesos amostrais fará com que o Stata estime valores incorretos de variância e de erros padrões para os coeficientes, assim como valores incorretos de "p" para os testes de hipótese.

```
regress y x1 x2 [aweight = peso]
```

PESO AMOSTRAL NAS REGRESSÕES DO STATA

- **PWEIGHT**:
  - Ideal para ser usado nas regressões do Stata.
  - Usa o peso amostral como o número de observações na população que cada observação representa.
  - São estimadas proporções, médias e parâmetros da regressão corretamente.
  - Há o uso de uma técnica de estimação robusta da variância que automaticamente ajusta para as características do plano amostral, de tal forma que variâncias, erros padrões e intervalos de confiança são calculados de forma mais precisa.
  - É o inverso da probabilidade da observação ser incluída no banco, devido ao desenho amostral.

```
regress y x1 x2 [pweight = peso]
```

OUTRAS OBSERVAÇÕES SOBRE PESOS NO STATA

PESOS EM TABELAS DE FREQUÊNCIA		
Tipo do peso	**Expandir para o tamanho da população (N)**	**Manter o tamanho da amostra (n)**
Discreto	**fweight**	**aweight**
Decimal	**iweight**	

PESOS EM MODELOS DE REGRESSÃO devem manter o tamanho da amostra (n)	
Erro padrão robusto	**R^2 ajustado, SQT, SQE, SQR**
pweight **reg y x, robust**	**aweight** **outreg2**

PLANO AMOSTRAL COMPLEXO

– Estatísticas descritivas e modelos de regressão devem levar em consideração a estrutura de planos amostrais complexos.

– PNAD tem amostra complexa (Silva, Pessoa, Lila, 2002):

  – Considerar variáveis de estrato de município autorrepresentativo e não autorrepresentativo (v4617) e de unidade primária de amostragem (v4618), do banco de domicílios.

  – Agregar variáveis acima ao banco de pessoas, o qual possui peso da pessoa (v4729).

  – Lidar com problema de alguns estratos terem somente uma unidade primária de amostragem. Pode-se especificar média deste estrato como sendo a média geral, ao invés da média do próprio estrato.

```
svyset [pweight=v4729], strata(v4617) psu(v4618) singleunit(centered)
```

  – Tabelas e regressões devem ser precedidas de “svy:”.