

# **AULA 10**

## **Problemas Adicionais de Especificação e de Dados**

**Ernesto F. L. Amaral**

**22 de julho de 2011**

**Análise de Regressão Linear e Análise de Dados Categóricos (MQ 2011)**

**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

$$\log(\text{salário}_h) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 \\ + \beta_4 \text{feminino} + \beta_5 \text{feminino} * \text{educ} + u$$

- 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:

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

- Se ele satisfizer RLM.3 (termo de erro tem média condicional zero), nenhuma função não-linear das variáveis independentes deve ser significativa 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:

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

- 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:  $\hat{y}$ .

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

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

- Se teste  $t$  de  $\theta$  é significativo, 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  $\delta_3$  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  $\delta_0$  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  $\beta_1$  e  $\beta_2$ .
- Não obteremos estimadores não-viesados de  $\beta_0$  e  $\beta_3$ .
- 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  $\beta_1$  e  $\beta_2$ .

# HIPÓTESES

- As hipóteses necessárias para que a solução plugada forneça estimadores consistentes de  $\beta_1$  e  $\beta_2$  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  $\beta_1$  e  $\beta_2$  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:

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

- 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:

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

- 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 $\beta_j$ 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  $\beta_0$ , 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  $t$  e  $F$ ) 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  $\beta_0$  e  $\beta_1$  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 ( $x_1$ ):  $\text{Cov}(x_1, e_1) = 0$ .
- 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  $\beta_0$  e  $\beta_1$ .
- Exceto quando  $\beta_1=0$ , 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 econométristas geralmente supõem que o **erro clássico nas variáveis (ECV)** é o erro de medida não-correlacionado com a variável explicativa *não-observada* ( $x_1^*$ ):  $\text{Cov}(x_1^*, e_1) = 0$ .
- 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  $x_3$ ,  $\beta_2$  e  $\beta_3$  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 frequência.
- 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 frequência 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 nos 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 das 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*.