Universidade de São Paulo

Faculdade de Filosofia, Letras e Ciências Humanas

Departamento de Ciência Política

Métodos Quantitativos de Pesquisa na Ciência Política

FLS 6183 & FLP 0468

2º semestre / 2016

Lista 2 - Erro de Mensuração

Answer Key

Objetivo:

Vários pressupostos do modelo de regressão enfatizam as propriedades que o erro deveria apresentar para conseguirmos utilizar a técnica para produzir as melhores, mais eficientes e não viesadas estimativas de nossos coeficientes. Neste laboratório, vamos utilizar uma simulação para demonstrar como os problemas envolvendo erro de mensuração de nossas variáveis podem levar a viés de nossas estimativas e, principalmente, iremos mostrar o mecanismo de transmissão.

Mesmo que o lab seja voltado para discutir o problema utilizando dados de simulação, como pesquisadores, em nosso dia-a-dia, enfrentamos o problema de viés de mensuração de nossas variáveis. Um caso clássico é o das perguntas em *surveys* sobre a renda dos indivíduos. De um modo geral, os entrevistados não se lembram exatamente de sua renda no mês de referência ou no ano anterior, reportando assim um valor aproximado, inexato. Portanto, um pesquisador que trabalhe com esses dados, por exemplo, tem que estar alerta sobre a possibilidade de se ter viés em sua variável.

Este lab analisa este problema no modelo bivariado com dois exercícios: a) o problema de viés na mensuração da variável independente e, b) o problema de viés na mensuração na variável dependente.

Vamos supor que temos a distribuição populacional da variável dependente (Y) e da variável independente (X). Iremos chamar as medidas populacionais de Yt e Xt, sendo t as "true measures" e as medidas observadas de Y e X. Criamos um banco de dados com N=500, estipulamos que Xt tem uma média de 7 e desvio padrão de 8 e Yt tem média de 10 e desvio padrão de 4. Além disso, estipulamos que a correlação entre Xt e Yt é de 0.7. Desse modo, criamos um banco de dados a partir das propriedades que estabelecemos que devem ter nossas variáveis. É importante ressaltar que Yt e Xt são variáveis aleatórias e com distribuição normal.

Tabela 1. Banco de dados – variáveis populacionais

	média	desvio padrão
dependente (Yt)	10	4
independente (Xt)	7	8

Correlação entre Yt e Xt 0,7.
N = 500 observações

Como queremos entender as consequências de termos erro de mensuração tanto em nossa variável dependente quanto na variável independente, iremos criar duas variáveis aleatórias ey (erro de mensuração de Y) e ex (erro de mensuração de X). Estipulamos que ey apresenta uma distribuição normal com média de 0 e desvio padrão de 3 e ex também apresenta distribuição normal, mas com média de 0 e desvio padrão de 6.

Tabela 2. Banco de dados – erro de mensuração

	média	desvio padrão
Erro de mensuração na variável dependente (ey)	0	3
Erro de mensuração na variável independente (ex)	0	6

Desse modo, a variável observada Y é formada por Yt ("a medida verdadeira") e pelo erro de mensuração ey. E, a variável observada X é composta por Xt ("a medida verdadeira") e ex que é o erro de mensuração dessa variável. Dado que sabemos que há problemas envolvendo erro de mensuração em Y e em X, agora iremos avaliar os efeitos da falta de confiabilidade de nossas medidas nos dois exercícios abaixo.

Exercício 1. Avalie os efeitos da falta de confiabilidade de nossas medidas analisando a média da variável dependente e independente.

O erro de mensuração não provoca viés no valor esperado da variável, uma vez que os erros são aleatórios e, portanto, erro de mensuração em X tem média zero.

$$E(X) = E(X_t + \varepsilon) = E(X_t) + E(\varepsilon)$$

$$E(X) = E(X_t)$$

Em suma, não existe viés na estimativa da média e isso é facilmente comprovável através dos resultados dos testes de hipótese (veja do-file gabarito). No teste de hipótese, partimos da hipótese de que a média de Y é igual a média de Yt (10) e X é igual a média de Xt (7). A hipótese alternativa

que assumimos em cada um dos testes é de que a média da variável com erro de mensuração será maior do que a média populacional pois o erro populacional levaria a um viés positivo em X e em Y. Desse modo, para a variável dependente, assumimos que:

Ho: A média de Y é igual a média de Yt (10).

Ha: A média de Y é maior do que a média de Yt.

O resultado do teste de hipótese indica que o intervalo de confiança da diferença entre Y e Yt é [9,12; 10,02], isto é, o valor 10 (que é a média populacional - Yt) está dentro do intervalo. Portanto, a conclusão é de que não temos evidências para rejeitar a hipótese de que as médias são iguais. Do mesmo modo, para a variável independente, assumimos que:

Ho: A média de X é igual a média de Xt (7).

Ha: A média de Y é maior do que a média de Xt.

O resultado do teste de hipótese indica que o intervalo de confiança da diferença entre X e Xt é [5,48; 7,21], isto é, o valor 7 (que é a média populacional - Xt) está dentro do intervalo. Portanto, a conclusão é de que não temos evidências para rejeitar a hipótese de que as médias são iguais.

Exercício 2. Avalie os efeitos da falta de confiabilidade de nossas medidas analisando a variância das variáveis.

Uma vez que temos erro de mensuração aleatório, as variâncias serão viesadas, pois:

$$V(X) = V(X_t + \varepsilon) = V(X_t) + V(\varepsilon)$$

Nesse sentido, a variância da variável observada será maior que o valor verdadeiro. A existência de viés pode ser identificada através de teste que aponta que a variância amostral é estatisticamente superior à populacional simulada (veja do-file gabarito). Testamos se as variâncias de Xt e X, bem como Yt e Y, são iguais. O comando *sdtest* testa se as variáveis têm o mesmo desvio padrão. Utilizamos as seguintes hipóteses:

Ho: O desvio padrão de Y é igual ao desvio padrão de Yt (8).

Ha: O desvio padrão de Y é diferente do que o desvio padrão de Yt.

O resultado do teste de hipótese indica que temos evidências para rejeitar a hipótese de que os desvio padrão são iguais, pois o p-value é < 0.001. Do mesmo modo, para a variável independente, assumimos que:

Ho: O desvio padrão de Y é igual ao desvio padrão de Yt (4).

Ha: O desvio padrão de Y é diferente do que o desvio padrão de Yt.

O resultado do teste de hipótese indica que temos evidências para rejeitar a hipótese de que os desvio padrão são iguais, pois o p-value é < 0.001.

Exercício 3. Supondo que Yt mensura perfeitamente a variável Y, avalie os efeitos da falta de confiabilidade de X analisando a covariância entre as variáveis.

Dado que:

$$COV(X, Y_t) = COV(X_t + \varepsilon, Y_t) = COV(X_t, Y_t) + COV(\varepsilon, Y_t) = COV(X_t, Y)$$

Portanto, as covariâncias não são viesadas pelo erro de mensuração aleatório.

Exercício 4. Supondo que Yt mensura perfeitamente a variável Y, porém X apresenta um problema de erro de mensuração aleatório, avalie a confiabilidade de X analisando a correlação entre as variáveis.

O erro de mensuração aleatório produz um viés para baixo na correlação entre X e Yt comparado a correlação de Xt com Yt. Nesse caso, há um viés de atenuação.

$$r_{XY_t} = \frac{\sigma_{XY_t}}{\sigma_X + \sigma_{Y_t}}$$

$$r_{X_tY_t} = \frac{\sigma_{X_tY_t}}{\sigma_{X_t} + \sigma_{Y_t}} = \frac{\sigma_{XY_t}}{\sigma_X + \sigma_{Y_t}}$$

Quando X e Yt covariam positivamente, $CORR(X, Y_t) \leq CORR(X_t, Y_t)$. Portanto, o erro de mensuração aleatório produz um viés para baixo na correlação bivariada.

Exercício 5. Supondo que Yt mensura perfeitamente a variável Y, porém X apresenta um problema de erro de mensuração aleatório, avalie o efeito de tal problema sobre os resultados da regressão bivariada entre Yt e Xt e da regressão entre Yt e X. Compare os resultados.

O erro de mensuração aleatório na variável independente produz um problema de viés para baixo no coeficiente de inclinação da regressão bivariada. Isso pode ser verificado estimando duas regressões em que a primeira utiliza a variável populacional "verdadeira" como variável explicativa e a segunda utiliza a variável com erro de mensuração como sua variável explicativa.

$$\beta_{Y_t X} = \frac{\sigma_{XY_t}}{\sigma^2_X}$$

$$\beta_{Y_t X_t} = \frac{\sigma_{X_t Y_t}}{\sigma^2_{X_t}}, \text{ então quando } X \text{ e } Y_t \text{ covariam positivamente, } \beta_{Y_t X} \leq \beta_{Y_t X_t}$$

A seguir, apresentamos os resultados dos dois modelos de regressão. No modelo 1, analisamos o impacto da variável "verdadeira" (Xt) sobre a variável Yt, enquanto que no modelo 2, estudamos o impacto de X em Yt. O coeficiente do modelo 2 é menor do que o coeficiente do modelo com as variáveis "verdadeiras".

	Modelo 1 b/se	**Modelo 2** b/se
Xt	0.352*** (0.02)	
X		0.240*** (0.02)
constant	7.441*** (0.16)	8.212*** (0.18)
r2	0.504	0.335
Obs	500.00	500.00

* p<0.05, ** p<0.01, *** p<0.001

Exercício 6. Agora, supondo que Xt mensura perfeitamente a variável X, porém Y apresenta um problema de erro de mensuração aleatório, avalie o efeito do erro de mensuração aleatório sobre os resultados da regressão bivariada entre Yt e Xt e da regressão entre Y e Xt. Compare os resultados.

O erro de mensuração aleatório na variável dependente não produz um viés no coeficiente de inclinação da reta, porém causa um aumento no erro padrão, consequentemente, aumento no intervalo de confiança da estimativa. Isso pode ser verificado estimando duas regressões em que a

primeira utiliza a variável populacional "verdadeira" (Xt) como variável dependente e a segunda utiliza a variável com erro de mensuração como sua variável dependente.

$$\beta_{YX_t} = \frac{\sigma_{X_t Y}}{\sigma^2_{X_t}}$$

$$\beta_{Y_t X_t} = \frac{\sigma_{X_t Y_t}}{\sigma^2_{X_t}}, \text{ então } \beta_{YX_t} = \beta_{Y_t X_t}$$

A seguir, apresentamos os resultados dos dois modelos de regressão. No modelo 1, analisamos o impacto da variável "verdadeira" (Xt) sobre a variável dependente "verdadeira" (Yt), enquanto que no modelo 2, estudamos o impacto de Xt em Y. Podemos notar que os coeficientes são muito próximos (0,352 vs. 0,354).

	Modelo 1 b/se	**Modelo 2** b/se
Xt	0.352***	0.354***
	(0.02)	(0.02)
constant	7.441***	7.268***
	(0.16)	(0.24)
r2	0.504	0.321
Obs	500.00	500.00

* p<0.05, ** p<0.01, *** p<0.001

Exercício 7. Discuta sobre os resultados encontrados no exercício 5 e 6 pensando especialmente sobre os pressupostos do modelo de regressão.

Comparando os resultados, podemos notar que quando há erro de mensuração nas variáveis, o problema é ainda mais grave no caso da variável independente. A estimativa do impacto da variável independente sobre a dependente é viesada quando temos problema de mensuração em X, o mesmo não ocorre quando se trata de erro de mensuração em Y. No primeiro caso, nossas estimativas são viesadas para baixo, enquanto que no segundo caso temos estimativas não viesadas, mas com intervalos de confiança maiores.