A Estimação de Parâmetros em Questões Sensíveis com recurso a Informação Auxiliar

Rita Sousa

Doutoramento em Estatística e Gestão de Risco
Especialidade em Estatística

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Maio 9, 2014

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Índice

Variáveis Sensíveis

- Em estudos de pesquisa por inquérito existem muitas situações em que a variável de interesse é sensível.
- A sensibilidade de algumas questões pode dar origem a recusas na resposta ou a falsas respostas dadas de forma intencional.
- Os inquéritos podem assumir diversas configurações, em parte relacionadas com o método de recolha e com o grau de privacidade que é oferecido aos respondentes.
- As estimativas obtidas por inquérito direto em questões sensíveis podem estar sujeitas a erros elevados.

Técnica de Resposta Aleatorizada

- Muitas técnicas têm sido utilizadas para melhorar as respostas através do aumento da privacidade dos inquiridos.
- A Técnica de Resposta Aleatorizada (TRA), introduzida por Warner em 1965, desenvolve uma relação aleatória entre as respostas individuais e as questões.
- Esta técnica providencia confidencialidade aos respondentes e ainda permite aos entrevistadores estimar a característica de interesse num nível mais agregado.

TRA Completa

- Os estudos com utilização da TRA geralmente focam-se na estimação da média ou na taxa de prevalência de uma dada característica sensível na população.
- Na TRA Completa (Warner, 1965; Greenberg et al., 1969) é pedido a todos os respondentes para disponibilizarem uma resposta dissimulada.
- Tendo em conta que nosso principal objetivo é avaliar o desempenho do estimador da média na presença de informação auxiliar, opta-se por utilizar o modelo da TRA aditivo completo para mascarar a variável sensível de interesse.

TRA Aditiva Completa

- Seja Y a variável em estudo, uma variável sensível que não pode ser diretamente observada e seja X uma variável auxiliar não sensível que é fortemente correlacionada com a variável de interesse Y.
- Por exemplo, Y pode estar associada ao consumo de drogas, ao alcoolismo, ao aborto induzido, à orientação sexual, ao rendimento e à fuga aos impostos.
- Considere a TRA em que o respondente dá uma resposta verdadeira para X e uma resposta dissimulada para Y dada por:

$$Z = Y + S,$$

em que S é uma variável que mascara a resposta, que tem média zero e é independente de Y e X.

Plano de Investigação

- O principal objetivo deste trabalho é melhorar a estimação de parâmetros de variáveis sensíveis com recurso a informação auxiliar.
- Nesta investigação propõem-se alguns estimadores para melhorar a estimação da média de uma variável sensível com base no modelo aditivo completo da TRA, recorrendo a informação auxiliar disponível não sensível.
- É desenvolvido um estudo teórico no qual se deduzem as expressões do Erro Médio ($Viés$) e do Erro Quadrático Médio (EQM) para todos os estimadores propostos.

Plano de Investigação

- Para um dado estimador $\hat{\theta}$, $Viés(\hat{\theta}) = E(\hat{\theta}) - \theta$, o que corresponde à diferença entre a média das estimativas e o parâmetro a estimar.
- O $EQM(\hat{\theta}) = Var(\hat{\theta}) + Viés(\hat{\theta})^2$, ou seja, a soma da variância do estimador mais o erro médio ao quadrado; para estimadores centrados o EQM coincide com a variância do estimador.
- Para além do estudo teórico, são desenvolvidas rotinas em R para um extenso estudo de simulação e de aplicação a dados reais, no qual se compara a performance dos principais estimadores em estudo.

Definição

- Considere uma população finita com N unidades $U = (U_1, U_2, ..., U_N)$, a partir da qual se extrai uma amostra de dimensão n pelo método da Amostragem Aleatória Simples Sem Resposição (AASSR).
- Seja Y a variável sensível em estudo que não pode ser diretamente observada e X uma variável auxiliar não sensível que é fortemente correlacionada com Y.
- Considere a TRA aditiva completa para camuflar a variável de interesse com uma resposta dada por $Z = Y + S$. Se a informação auxiliar de X for ignorada, um estimador centrado de μ_Y é a média amostral de Z:

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}. \quad (1)$$

Estimador Proposto da Razão

- Propõe-se uma estimador da razão no qual o estimativa TRA da média para a variável de interesse é melhorada por recurso a informação auxiliar.
- Assim, o estimador proposto da razão (Sousa et al., 2010) estima a média da populaçao da variável sensível Y utilizando informação de uma variável auxiliar não sensível X:

$$\hat{\mu}_R = \bar{z}\left(\frac{\bar{X}}{\bar{x}}\right). \tag{2}$$

Estudo de Simulação

Comparando os valores do Erro Quadrático Médio (EQM) calcula-se a Eficiência Relativa em Percentagem (ERP) do estimador da razão face ao estimador ordinário da média:

$$ERP = \frac{EQM(\hat{\mu}_Y)}{EQM(\hat{\mu}_R)} \times 100.$$

População			Estimação EQM			ERP	
N	ρ_{XY}	n	Empírica	1ª Ordem	2ª Ordem	1ª Ordem	2ª Ordem
1000	0.8783	50	0.0406	0.0392	0.0412	309.31	294.99
		100	0.0183	0.0186	0.0190		302.36
		200	0.0083	0.0083	0.0083		306.18
		500	0.0050	0.0048	0.0048		307.48

- A diferença entre a 1ª e a 2ª ordem de aproximação não é significativa, mesmo para amostras de dimensão reduzida.
- O estimador da razão, apesar de enviesado, apresenta melhor ERP quando comparado com o estimador TRA da média ($\hat{\mu}_Y$).
- O ganho pode ser bastante significativo se a correlação entre a variável em estudo Y e a variável auxiliar X for elevada.

Estimadores Transformados da Razão

- Estudou-se também o estimador transformado da razão (Sousa et al., 2010):

$$\hat{\mu}_{TR} = \bar{z}\left(\frac{c\bar{X}+d}{c\bar{x}+d}\right), \tag{3}$$

em que c e d são parâmetros independentes da unidade de medida, que podem corresponder a medidas, tais como o coeficiente de simetria $skewness$ e o coeficiente de achatamento $kurtosis$ em X.

- Para amostras de pequena dimensão este estimador dá origem a uma redução no $Viés$ quando o parâmetro aditivo d corresponde à $kurtosis$.
- Para situações de forte correlação entre a variável de interesse Y e a variável auxiliar X o estimador transformado da razão com a medida $skewness$ no parâmetro d permite reduzir o EQM relativamente ao estimador da razão.
- Os estimadores transformados da razão produzem ganhos mínimos face ao estimador proposto da razão.

Estimador Proposto da Regressão

- Considerando uma relação linear entre Y e X, propõe-se o seguinte estimador da regressão (Gupta et al., 2012) para a média populacional de Y:

$$\hat{\mu}_{Reg} = \bar{z} + \hat{\beta}_{zx} \left(\bar{X} - \bar{x} \right), \tag{4}$$

em que $\hat{\beta}_{zx} = \dfrac{S_{zx}}{S_x^2}$ é o coeficiente amostral de regressão entre Z e X; $Z = Y + S$ é a resposta mascarada de Y.

Propriedades

Pode-se verificar que:

1. $EQM(\hat{\mu}_{Reg}) < EQM(\hat{\mu}_Y)$ se $\rho_{yx}^2 > 0$;
2. $EQM(\hat{\mu}_{Reg}) < EQM(\hat{\mu}_R)$ se $(C_x - C_z\rho_{zx})^2 > 0$.

- Estas condições são sempre verdadeiras, concluindo-se que, segundo a aproximação de 1ª ordem, o estimador da regressão é mais eficiente do que o estimador ordinal da média $(\hat{\mu}_Y)$ e do que o estimador da razão $(\hat{\mu}_R)$.

Estimador Acumulado da Razão e Regressão

- Ray e Singh (1981), Perri (2004), Kadilar e Cingi (2004, 2006) desenvolveram estimadores acumulados da razão e da regressão que combinam os dois estimadores.
- De forma similar, considera-se o estimador híbrido como generalização do estimador acumulado da razão e regressão, cujos coeficientes minimizam o valor do EQM. Este estimador (Gupta et al., 2012) é definido por:

$$\hat{\mu}_{GRR} = \left[k_1 \bar{z} + k_2 \left(\bar{X} - \bar{x}\right)\right] \left(\frac{\bar{X}}{\bar{x}}\right), \tag{5}$$

em que k_1 e k_2 são constantes.

Propriedades

Pode-se verificar que:

1. $EQM(\hat{\mu}_{GRR})_{min} < EQM(\hat{\mu}_Y)$ se $\left(\frac{1-f}{n}\right)\{S_y^2 + S_s^2\} > 0$, que é sempre verdadeira;
2. $EQM(\hat{\mu}_{GRR})_{min} < EQM(\hat{\mu}_R)$ se $\left(\frac{1-f}{n}\right) C_x^2 < 1$, que tem elevada probabilidade de ser verdadeira;
3. $EQM(\hat{\mu}_{GRR})_{min} < EQM(\hat{\mu}_{Reg})$ se $\left(\frac{1-f}{n}\right) C_z^2 \left(1 - \rho_{zx}^2\right) > 0$, que é sempre verdadeira;

Em que $f = n/N$ corresponde à fração de amostragem.

Exemplo Numérico

- Apresenta-se de seguida um exemplo de aplicação a dados reais relativos ao Inquérito às Tecnologias da Informação e Comunicação nas Empresas (IUTICE).
- Seja Y o montante anual de compras realizadas pelas empresas em 2009 e X a variável auxiliar do valor do volume de negócios (VVN) disponibilizada, para cada empresa, pelos dados administrativos da Informação Empresarial Simplificada (IES).
- Considera-se o universo de 5336 respondentes do IUTCE em 2010 como sendo a nossa população em estudo, da qual se extraem diferentes amostras.
- Para simular a TRA considera-se S como sendo uma variável gerada aleatoriamente com média zero e desvio padrão igual a 10% do desvio padrão de X.

Características da População:

$N = 5336, \rho_{XY} = 0.9632$
$\mu_X = 22.99, \mu_Y = 30.19$ (em milhares de €)
$\sigma_X = 172.09, \sigma_Y = 138.653$ e $\beta_{YX} = 0.7763$

Resultados

Tabela: EQM e ERP para o estimador da razão e para os estimadores da regressão relativamente ao estimador ordinário da média

População				Estimação *EQM*		
N	ρ_{XY}	n	Estimador	Empírica	Teórica	ERP
5336	0.9636	100	$\hat{\mu}_R$	11.5741	16.4778	1162.46
			$\hat{\mu}_{Reg}$	0.8601	16.4153	1166.88
			$\hat{\mu}_{GRR}$	11.6905	15.3461	1248.18
		1000	$\hat{\mu}_R$	1.4224	1.3645	1162.46
			$\hat{\mu}_{Reg}$	1.4265	1.3594	1166.88
			$\hat{\mu}_{GRR}$	1.4287	1.3253	1196.91

- A ERP no estimador generalizado da razão e da regressão é maior do que nos restantes estimadores, particularmente quando a dimensão da amostra é reduzida.
- Ambos os estimadores apresentam melhor desempenho do que o estimador ordinário da média, no entanto o ganho de eficiência é superior no estimador da regressão.
- Ganho adicionais, ainda que modestos, são possíveis de obter com a versão generalizada do estimador da razão e regressão.

Estimador do Tipo Exponencial

- O primeiro estimador proposto corresponde a uma generalização do estimador acumulado da diferença e do tipo exponencial, seguindo Grover (2010) e Shabbir e Gupta (2007), que se define pela seguinte expressão:

$$\hat{\mu}_{DE} = \left[w_1 \bar{z} + w_2 \left(\bar{X} - \bar{x}\right)\right] exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right), \qquad (6)$$

onde w_1 e w_2 são pesos adequados que otimizam o estimador.

Exemplo Numérico

Tabela: EQM e ERP para o estimador da razão, para os estimadores da regressão e para o estimador exponencial relativamente ao estimador ordinário da média

População				Estimação *EQM*		
N	ρ_{XY}	n	Estimador	Empírica	Teórica	ERP
5336	0.9636	500	$\hat{\mu}_R$	2.7259	3.0367	1164.94
			$\hat{\mu}_{Reg}$	3.0170	3.0252	1169.39
			$\hat{\mu}_{GRR}$	2.7509	3.0069	1176.50
			$\hat{\mu}_{DE}$	2.8631	2.9173	1212.61
		1000	$\hat{\mu}_R$	1.3092	1.3614	1164.94
			$\hat{\mu}_{Reg}$	1.3592	1.3562	1169.39
			$\hat{\mu}_{GRR}$	1.3175	1.3526	1172.47
			$\hat{\mu}_{DE}$	1.3381	1.3345	1188.38

- O estimador generalizado acumulado da diferença e do tipo exponencial apresenta valores mais elevados de ERP.
- O 1º estimador proposto do tipo exponencial é mais eficiente do que os existentes na literatura.

Estimador Exponencial Otimizado

- Propõe-se agora um estimador exponencial otimizado que resulta de uma versão modificada do estimador acumulado da diferença e do tipo exponencial e que se define pelo seguinte expressão:

$$\hat{\mu}_{EO} = [d_1 \bar{z} + d_2] \, exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right), \tag{7}$$

em que d_1 e d_2 são constantes.

Exemplo Numérico

- Apresenta-se de seguida um exemplo de aplicação a dados reais relativos ao Inquérito Mensal de Conjuntura (IMC).
- Considera-se como população alvo as empresas do setor da indústria respondentes ao IMC em 2010.
- Seja Y o valor dos salários mensais registados em 2010 e X o número de pessoas ao serviço, conhecido para todas as empresas, disponibilizado pelos dados administrativos da Informação Empresarial Simplificada (IES).
- Consideram-se os dados de 26980 salários recolhidos nesse ano, dos quais se extraem diferentes amostras.
- Para simular a TRA considera-se S como sendo uma variável gerada aleatoriamente com média zero e desvio padrão igual a 10% do desvio padrão de X.

Características da População:

$N = 26980, \rho_{XY} = 0.8599$
$\mu_X = 113.91, \mu_Y = 167.18$ (em milhares de €)
$\sigma_X = 215.8, \sigma_Y = 501.4$ e $\sigma_{YX} = 93040$

Resultados

Tabela: EQM para o estimador proposto do tipo exponencial ($\hat{\mu}_{DE}$) e para o estimador exponencial otimizado ($\hat{\mu}_{EO}$)

População				**Estimação EQM**	
N	ρ_{XY}	n	Estimador	Empírica	Teórica
26980	0.8599	5000	$\hat{\mu}_{DE}$	10.88	10.75
			$\hat{\mu}_{EO}$	1.09	1.07
		10000	$\hat{\mu}_{DE}$	4.19	4.15
			$\hat{\mu}_{EO}$	0.41	0.41

- O estimador proposto acumulado da diferença e do tipo exponencial pode trazer ganhos significativos.
- O estimador otimizado é mais eficiente do que o estimador acumulado da diferença e do tipo exponencial proposto anteriormente, que por sua vez é melhor os estimadores definidos para a média de uma populaçao finita.

Amostragem Aleatória Estratificada

- Alguns estudos têm sido desenvolvidos também para melhorar a estimação de parâmetros com diferentes esquemas amostrais. Nesse contexto destacam-se autores como Kadilar e Cingi (2005), Shabbir e Gupta (2005, 2006), Singh e Vishwakarma (2008) e Koyuncu e Kadilar (2008, 2009).
- Nesta tese sugerem-se o estimador combinado da razão (Sousa et al., 2010) e o estimador combinado da regressão (Gupta et al., 2012) para a média da população de uma variável sensível com recurso a informação auxiliar não sensível. Assim, utiliza-se a TRA no contexto da amostragem aleatória estratificada (AAE).

Estimador Combinado da Razão

- Considere uma amostra aleatória estratificada s selecionada da população em estudo U, que se divide em L estratos de dimensão N_h tal que $\sum_{h=1}^{L} N_h = N (h = 1, ..., L)$.
  Propõe-se o seguinte estimador combinado da razão:

$$\hat{\mu}_{Rst} = \bar{z}_{st} \left(\frac{\bar{X}}{\bar{x}_{st}} \right), \tag{8}$$

em que $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ e $\bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h$ são as médias amostrais estratificadas de X e Z, respetivamente; o peso de cada estrato h define-se por $W_h = N_h/N$.

Estimador Combinado da Regressão

- Assumindo uma relação linear entre Y e X, propõe-se o seguinte estimador combinado da regressão para a média populacional de Y:

$$\hat{\mu}_{Regst} = \bar{z}_{st} + \hat{\beta}_c \left(\bar{X} - \bar{x}_{st} \right), \tag{9}$$

em que $\hat{\beta}_c = \dfrac{\sum_{h=1}^{L} W_h^2 \gamma_h s_{zxh}}{\sum_{h=1}^{L} W_h^2 \gamma_h s_{xh}^2}$ é o coeficiente amostral de regressão entre Z e X e $\gamma_h = \left(\dfrac{1}{n_h} - \dfrac{1}{N_h} \right)$.

Exemplo Numérico

- Apresenta-se de seguida um exemplo de aplicação a dados reais relativos ao Inquérito às Tecnologias da Informação e Comunicação nas Empresas (IUTICE).
- Seja Y o montante anual das compras realizadas pelas empresas em 2009 e X a variável auxiliar do valor do volume de negócios (VVN) disponibilizada, para cada empresa, pelos dados administrativos da Informação Empresarial Simplificada (IES).
- Considera-se o universo de 1698 pequenas e médias empresas que responderam ao IUTCE em 2010 como sendo a nossa população em estudo, da qual se extraem diferentes amostras.
- Para simular a TRA considera-se S como sendo uma variável gerada aleatoriamente com média zero e desvio padrão igual a 10% do desvio padrão de X.
- As amostras foram selecionadas por alocação proporcional à dimensão dos estratos na população.

Características da População:

$N = 1698, \rho_{XY} = 0.9368$
$\mu_X = 25.31, \mu_Y = 17.97$ (em milhões de €)
$\sigma_X = 25.31, \sigma_Y = 22.39$ e $\beta_{YX} = 0.8284$
$h = 3$ estratos (VVN:$< 10, 10 - 30, \geq 30$)

Instituto Nacional de Estatística
Statistics Portugal

Resultados

Tabela: EQM teórico e ERP para os estimadores da razão e da regressão relativamente ao estimador ordinário da média e ERP para a AAS relativamente à AAE $(Deff)$

População			n	Estimador	EQM Teórico		ERP	$Deff$
N	N_h	ρ_{XY}			AAS	AAE		
1698	$N_1 = 979$ $N_2 = 362$ $N_3 = 357$	0.9368	250	$\hat{\mu}_{Yst}$	2.0403	0.6948	100.00	293.66
				$\hat{\mu}_{Rst}$	0.2909	0.2397	289.91	851.35
				$\hat{\mu}_{Regst}$	0.2499	0.1863	373.00	1095.35
			500	$\hat{\mu}_{Yst}$	0.8440	0.2903	100.00	290.71
				$\hat{\mu}_{Rst}$	0.1204	0.0992	292.66	850.81
				$\hat{\mu}_{Regst}$	0.1034	0.0785	369.91	1075.37

- A vantagem de utilizar a TRA com recurso a informação auxiliar permanece válida no contexto da amostragem aleatória estratificada.
- O ganho é mais evidente na AAS porque a estratificação já reduz significativamente os valores do EQM.

Informação Auxiliar Retrospetiva

- O estudo de determinadas variáveis de interesse nem sempre é possível por observação direta das mesmas, independentemente do seu grau de sensibilidade. No entanto, é frequente essa informação estar disponível para um período de tempo precedente.
- Os dados da população são muitas vezes conhecidos para um período de tempo em que se realiza um recenseamento ou em que se disponibilizam dados administrativos.
- A utilização de informação auxiliar pode melhorar significativamente os resultados da estimação, nomeadamente no que diz respeito ao parâmetro da média.

Estimadores Estudados

- Se estimarmos a média de uma dada variável de interesse Y_t, no período de referência t, ignorando a informação retrospetiva do período $(t-m)$, o estimador usual é a média amostral: $\hat{\mu}_{Y_t} = \bar{y}_t$.
- Por outro lado, se considerarmos a informação auxiliar disponível no período $(t-m)$, o estimador da razão define-se por:

$$\hat{\mu}_{R_t} = \bar{y}_t \left(\frac{\bar{Y}_{(t-m)}}{\bar{y}_{(t-m)}} \right), \tag{10}$$

com $\bar{Y}_{(t-m)}$ e $\bar{y}_{(t-m)}$ médias populacional e amostral de Y no período $(t-m)$.

- Admitindo a existência de uma correlação linear significativa, entre a variável de interesse Y_t e a variável auxiliar $Y_{(t-m)}$, o estimador da regressão define-se por:

$$\hat{\mu}_{Reg_t} = \bar{y}_t + \hat{\beta}_{y_{(t-m)}y_t} \left(\bar{Y}_{(t-m)} - \bar{y}_{(t-m)} \right), \tag{11}$$

em que $\hat{\beta}_{y_{(t-m)}y_t}$ é o coeficiente de regressão estimado entre Y_t e $Y_{(t-m)}$.

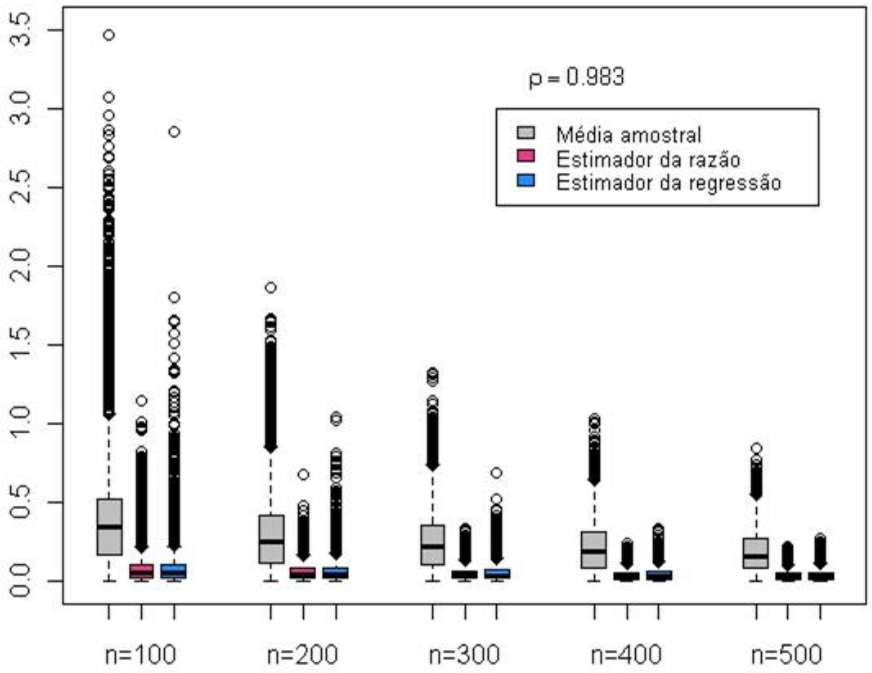
Aplicação Prática

- Os estimadores da razão e da regressão, com recurso a informação auxiliar retrospetiva, foram testados com dados reais do Inquérito à Utilização das Tecnologias da Informação e da Comunicação nas Empresas (IUTICE).
- Os estimadores foram testados considerando, como população, os dados de 2080 (N) empresas comuns às amostras do IUTICE em 2008 e 2009.

  Variável de interesse Y: valor das exportações
  Período de referência t: ano de 2009
  Período da variável auxiliar, $(t-m)$: ano de 2008 $(m=1)$
  Dimensão das amostras: $n = 100, 200, 300, 400$ e 500
  Coeficiente de correlação: $\rho_{y_{(2008)}y_{(2009)}} = 0.9830$
- Para cada valor de n foram selecionadas aleatoriamente 5000 amostras que permitiram obter um valor empírico do $Viés$ e do EQM.

Resultados

Figura: $Viés$ Absoluto Relativo



- Os estimadores com informação auxiliar retrospetiva apresentam menor enviesamento no $Viés$ Absoluto Relativo:

$$\left| \frac{Viés(\hat{\mu}_Y)}{\bar{Y}} \right|.$$

- Os estimadores da razão e da regressão, que utilizam informação auxiliar, apresentam valores de $Viés$ muito próximos sobretudo para amostras de maior dimensão.

Resultados

Tabela: EQM e respetiva ERP

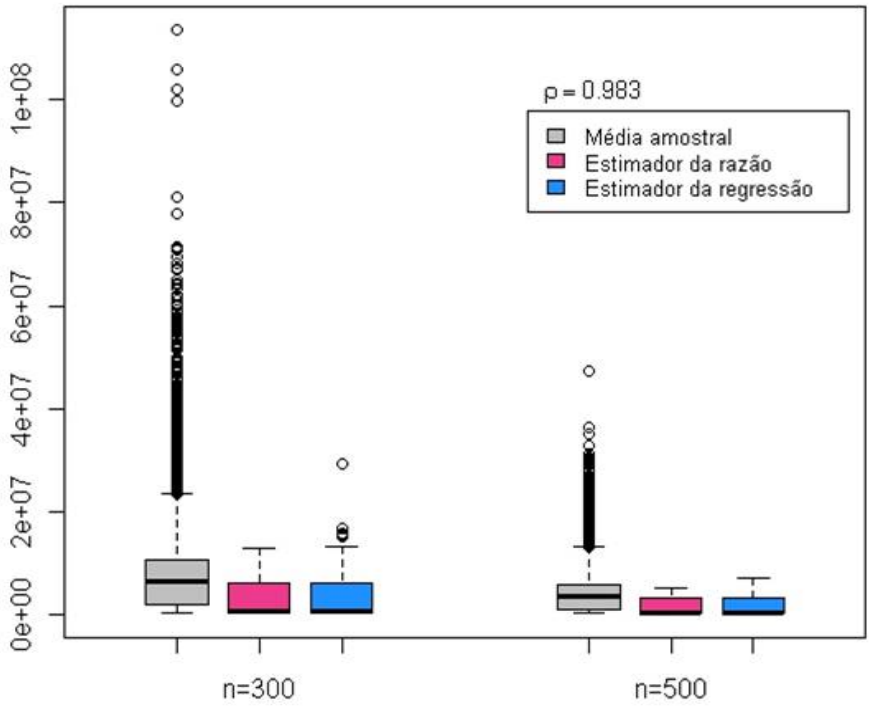
População			Estimação *EQM*		
N	ρ	n	**Empírica**	**Teórica**	***ERP***
		100	21038830,91	20394272,26	
			1072762,19	743315,21	
			1707420,94	**685741,26**	
		200	9463804,74	9682129,26	
			429391,56	352887,02	
			589879,88	**325553,93**	
		300	6218862,58	6111414,92	2743.69
2080	**0,9830**		261136,45	222744,29	**2974.05**
			318169,80	**205491,49**	
		400	4428255,68	4326057,75	
			173577,60	157672,92	
			195887,40	**145460,27**	
		500	3275266,72	3254843,45	
			135603,54	118630,10	
			147706,07	**109441,53**	

- Comparando os valores do Erro Quadrático Médio (EQM), os estimadores da razão e da regressão apresentam vantagens claras face à usual média amostral que não utiliza informação auxiliar.

Média Amostral
Estimador da razão
Estimador da regressão

Resultados

Figura: Erro Quadrático Médio



- Pode-se observar que a utilização de informação auxiliar (retrospetiva) aumenta a eficiência dos estimadores, diminuindo significativamente o enviesamento para valores elevados na distribuição empírica do EQM.

Exemplo Numérico

- Apresenta-se um exemplo numérico com o objetivo de estabelecer uma análise comparativa de desempenho dos principais estimadores propostos.
- Considere-se um conjunto de dados reais relativos aos dados económicos recolhidos mensalmente pelo Inquérito Mensal de Conjuntura (IMC).
- Seja Y o montante anual de compras efetuadas pelas empresas em 2009 e X a variável auxiliar relativa a esse montante registado em 2008, disponível por dados administrativos para o universo das empresas.
- Considera-se como população em estudo as 608 empresas comuns, respondentes em 2008 e 2009.
- Para a TRA define-se S como uma variável aleatória de média zero e desvio padrão igual a 10% do desvio padrão de X.

Características da População:

$N = 608, \rho_{XY} = 0.9447$
$\mu_X = 21357.69, \mu_Y = 17828.2$ (em milhares de €)
$\sigma_X = 65874.83, \sigma_Y = 57489.53$ e $\sigma_{XY} = 3577597688$

Resultados

Compara-se o estimador ordinário da média $(\hat{\mu}_Y)$ com os principais estimadores propostos em estudo:

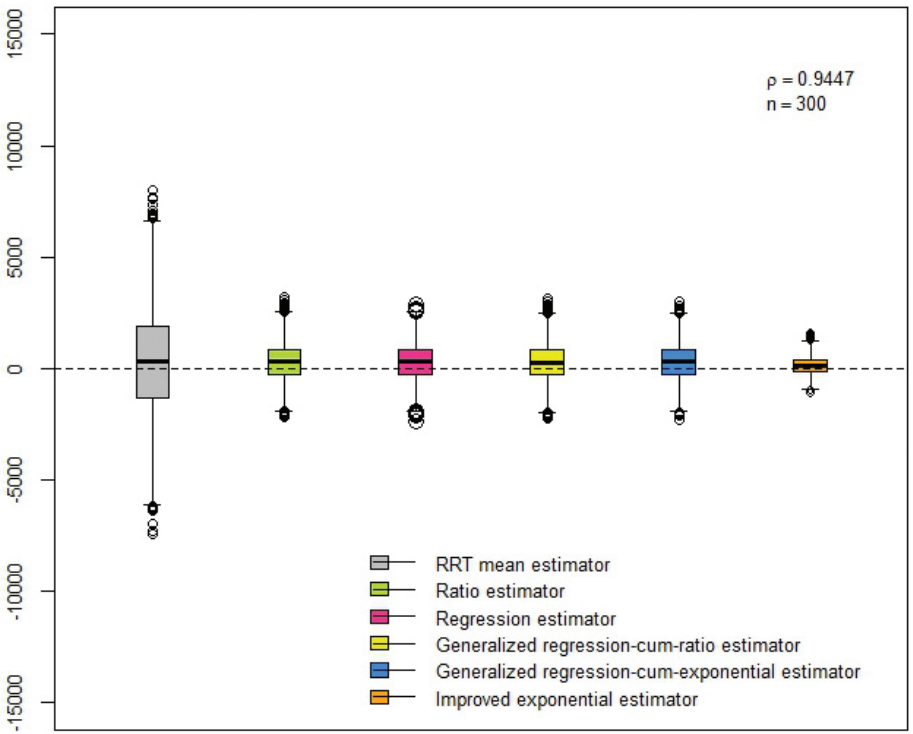
- Estimador da Razão $(\hat{\mu}_R)$ (Sousa et al., 2010);
- Estimador da Regressão $(\hat{\mu}_{Reg})$ (Gupta et al., 2012);
- Estimador Acumulado Razão-Regressão $(\hat{\mu}_{GRR})$ (Gupta et al., 2012);
- Estimador Acumulado da Diferença-Exponencial $(\hat{\mu}_{DE})$ (Koyuncu et al., 2013);
- Estimador Exponencial Otimizado $(\hat{\mu}_{EO})$ (Gupta et al., 2013).

Tabela: *Viés* Absoluto Relativo para os estimadores em comparação

População			*Viés* Absoluto Relativo			
N	ρ_{XY}	Estimador	$n = 50$	$n = 100$	$n = 200$	$n = 300$
608	0.9447	$\hat{\mu}_R$	0.0022	0.0010	0.0004	0.0002
		$\hat{\mu}_{Reg}$	0.0080	0.0036	0.0015	0.0007
		$\hat{\mu}_{GRR}$	0.0225	0.0104	0.0042	0.0021
		$\hat{\mu}_{DE}$	0.0215	0.0093	0.0036	0.0018
		$\hat{\mu}_{EO}$	0.0042	0.0022	0.0009	0.0005

Resultados

Figura: Distribuição Empírica do $Viés$



- Pelo gráfico que se segue podemos ver que todos so estimadores têm uma distribuição de $Viés$ em torno de zero mas é o estimador exponencial otimizado que apresenta menor dispersão.
- Apesar do estimador TRA ordinal da média ser centrado, apresenta valores empíricos de $Viés$ mais elevados face aos estimadores que utilizam informação auxiliar.

Resultados

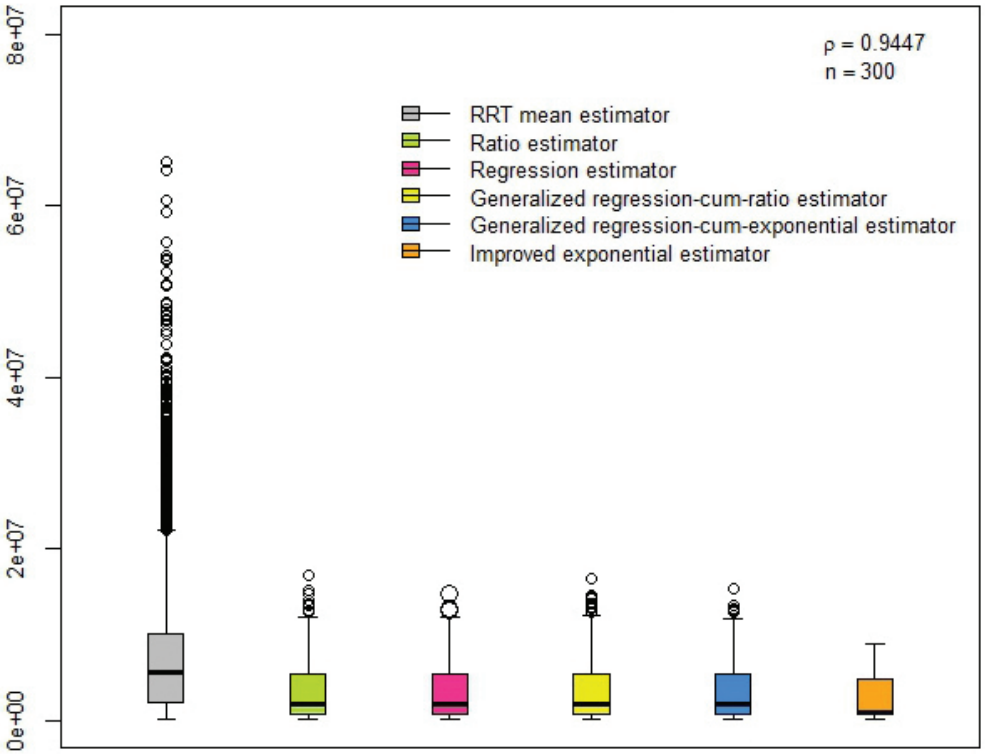
Tabela: EQM e PRE para todos os estimadores em comparação relativamente ao estimador ordinário da média

População				Estimação EQM		
N	ρ_{XY}	n	Estimador	Empírica	Teórica	EQM
608	0.9447	300	$\hat{\mu}_Y$	5728956.33	5656539.42	100.00
			$\hat{\mu}_R$	791784.85	676854.07	835.71
			$\hat{\mu}_{Reg}$	752752.90	676074.28	836.67
			$\hat{\mu}_{GRR}$	765577.73	674615.91	838.48
			$\hat{\mu}_{DE}$	740274.33	670651.05	843.44
			$\hat{\mu}_{EO}$	175396.47	149770.32	3776.81

- Os estimadores de regressão apresentam ganhos de desempenho face ao estimador da razão.
- Tal como esperado, o estimador exponencial otimizado é o que apresenta maior eficiência, dada a significativa redução verificada no EQM.

Resultados

Figura: Distribuição Empírica do EQM



- De acordo com o gráfico apresentado, o uso de informação auxiliar reduz significativamente a amplitude do EQM, particularmente no estimador exponencial otimizado que apresenta resultados muito próximos de zero.

Notas Finais

- No contexto da pesquisa por inquérito o uso de informação auxiliar pode ser essencial para melhorar a precisão das estimativas, principalmente quando se trata de variáveis sensíveis.
- As novas metodologias propostas foram comparadas entre si e com a estimativa TRA ordinal da média que não utiliza informação auxiliar.
- Concluiu-se que a estimação da média de uma variável sensível pode ser significativamente melhorada pelo uso de uma variável correlacionada auxiliar não sensível.
- Na presença de uma forte correlação linear, entre as variáveis de interesse e auxiliar, o estimador da regressão apresenta melhor desempenho do que o estimador da razão.
- Alguns estimadores do tipo exponencial revelaram-se mais eficientes do que os estimadores da razão e da regressão.

Trabalho Futuro

Nesta área existe ainda muita matéria a explorar e o meu plano futuro de trabalhos passa por:

- Estudar outras combinações de estimadores;
- Fazer aplicações com diferentes desenhos amostrais;
- Testar outras técnicas que proporcionem confidencialidade aos respondentes quando estes têm de responder a questões sensíveis;
- Planear e aplicar um inquérito com questões sensíveis para avaliar o desempenho dos estimadores propostos numa aplicação real da TRA.

Referências Bibliográficas

- COCHRAN, W.G. 1997. *Sampling Techniques*, 3^{rd} Ed., New York, Wiley Eastern Ltd.
- EDWARDS, A. L. 1957. *The social desirability variable in personality assessment and research*, New York: Dryden, Praeger.
- EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.
- GUPTA, S. & SHABBIR, J. 2004. Sensitivity estimation for personal interview survey questions. *Statistica*, 64, 643-653.
- GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.
- KADILAR, C. & CINGI, H. 2004. Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, 151, 893-902.
- KADILAR, C. & CINGI, H. 2006. Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19(1), 75-79.

Referências Bibliográficas

- KOYUNCU, N., GUPTA, S. & SOUSA, R. 2013. Exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary.
- KOYUNCU, N. & KADILAR, C. 2010. On the family of estimators of population mean in stratified random sampling. *Pakistan Journal of Statistics*, 26(2), 427-443.
- MUKHOPADHYAY, P. 1998. *Theory and Methods of Survey Sampling*, New Delhi, Prentice-Hall of India.
- SÄRDNAL, C.-E., SWENSSON, B. & WRETMAN, J. 1997. *Model assisted survey sampling*, Springer series in statistics.
- SOUSA, R., SHABBIR, J., REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.
- SUKHATME, P.V. & SUKHATME, B.V. 1984. *Sampling theory of surveys with applications*, 3^{rd} Ed., Ames, Iowa, Iowa State University Press.
- WARNER, S. L. 1965. Randomized response: a survey technique for elimination evasive answer bias.

FIM

Obrigada pela vossa atenção...