**Universidade de Coimbra**

FCTUC

Amostragem e Sondagens

# Projeto Computacional

João Marcelino
João Nuno Freitas
Ricardo Marques

21 de Maio de 2021

# Introdução

Neste projeto pretendemos comparar os estimadores do quociente e da regressão em duas populações convenientemente escolhidas, usando o método de Monte Carlo, onde utilizamos um plano SSR. As populações que vamos utilizar são populações reais que estão incluídas nas bases de dados data(belgianmunicipalities) e data(swissmunicipalities) no *package* "sampling" do R.
Vamos comparar também estes dois estimadores com a média amostral e por fim vamos analisar a probabilidade de cobertura e a margem de erro dos intervalos de confiança produzidos pelos mesmos.

# Simulação

## Resultados teóricos

Na elaboração deste trabalho tivemos em consideração os seguintes resultados.

Num plano SSR, o estimador do total é dado por

$$\hat{t}_y = \frac{N}{n} \sum_{k \in S} y_k \, .$$

A variância deste estimador é dada por

$$\mathrm{Var}(\hat{t}_y) = N^2 \left(1 - \frac{n}{N}\right) \frac{{s_y}^2}{n} \, .$$

Por sua vez, o estimador da média no plano SSR é dado por

$$\hat{\bar{y}} = \frac{\hat{t}_y}{N} \, ,$$

e a sua variância definida por

$$\mathrm{Var}(\hat{\bar{y}}) = \frac{\mathrm{Var}(\hat{t}_y)}{N^2} \, .$$

Quando se verifica uma relação de proporcionalidade entre $y$ e a variável auxiliar $x$, tem-se para $\lambda \in \mathbb{R}$ a seguinte relação

$$y_k = \lambda x_k, k \in U \, .$$

Assim,

$$t_y = \lambda t_x.$$

E para toda a amostra aleatória S, temos que

$$\hat{t}_y = \lambda \hat{t}_x.$$

Quando se verifica esta relação podemos utilizar o estimador do quociente definido por

$$\hat{t}_q = \frac{\hat{t}_y}{\hat{t}_x} t_x.$$

O estimador $\hat{t}_q$ é assim um estimador de tipo rácio que não é cêntrico, contudo, quando temos um tamanho de amostra $n$ grande podemos desprezar o viés do estimador. A variância do estimador $\hat{t}_y$ é dada por

$$\mathrm{Var}(\hat{t}_q) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{{s_y}^2}{n} \left(1 - \frac{2r s_{yx} - r^2 {s_x}^2}{{s_y}^2}\right)$$

Resultados semelhantes se conclui para o estimador do quociente da média, representado por,. Sendo definido por:

$$\hat{\bar{y}}_q = \frac{\hat{\bar{y}}}{\hat{\bar{x}}} \bar{x}$$

Por sua vez a variância deste último estimador é dado por

$$\mathrm{Var}(\hat{\bar{y}}_q) \approx \frac{\mathrm{Var}(\hat{t}_q)}{N^2}\,.$$

Quando se verifica uma relação linear entre a variável de interesse $y$ e a variável auxiliar $x$, da qual supomos conhecer o respetivo total $t_x$, conseguimos estimar $t_y$, utilizando o estimador da regressão definido por:

$$\hat{t_r} = \hat{t_y} - \hat{a}(\hat{t_x} - t_x)$$

Com

$$\hat{a} = \frac{\hat{s_{yx}}}{\hat{s_x^2}}$$

Sabemos que $\hat{t_r}$ é em geral enviesado, contudo para amostras grandes o viés pode ser desprezado. A expressão da variância deste estimador é definida por

$$\mathrm{Var}(\hat{t}_r) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{{s_y}^2}{n} \left(1 - \rho_{xy}^2\right)$$

com

$$\rho_{xy} = \frac{s_{xy}}{s_y s_x}$$

Por outro lado, quando pretendemos estimar a média da população, temos:

$$\hat{\bar{y}}_r = \hat{\bar{y}} - \hat{a}(\hat{\bar{x}} - \bar{x})$$

Cuja variância é dada aproximadamente por

$$\mathrm{Var}(\hat{\bar{y}}_r) \approx \frac{\mathrm{Var}(\hat{t}_r)}{N^2}\,.$$

## População da Belgica (*belgianmunicipalities*)

Temos uma população U de municípios da Bélgica de tamanho $N = 589$ e desta população vamos retirar amostras de tamanho $n$.
A nossa variável de interesse é o total dos impostos em cada um dos municípios que representamos por $y$. Para estimar $\bar{y}$ vamos utilizar uma variável auxiliar $x$ que representa o total da população em cada um dos municípios.
Inicialmente começámos por verificar que tipo de relação existia entre estas duas variáveis. Para isso, construimos o seguinte gráfico de dispersão.

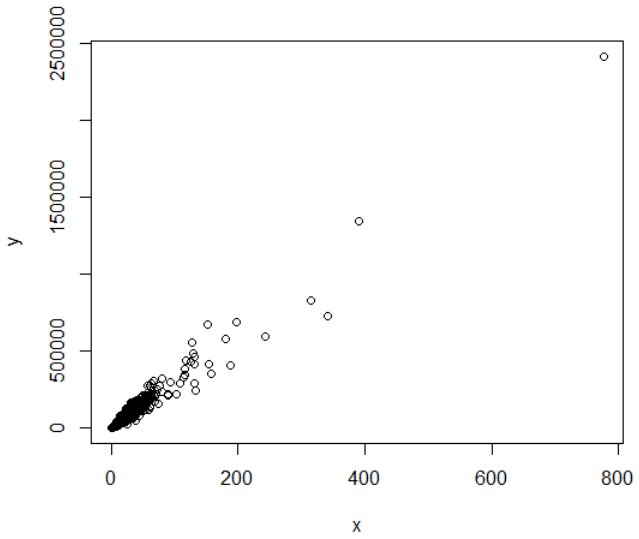



Figura 1: Gráfico de dispersão entre $x$ e $y$

Como podemos observar existe uma relação linear entre as duas variáveis e, portanto é razoável considerar o estimador do quociente e da regressão para estimar $\bar{y}$.
Para a simulação dos dados considerámos 2000 amostras de tamanho 100 retiradas através de um plano SSR. Começámos por gerar os diagramas de extremos de quartis para cada um dos estimadores.

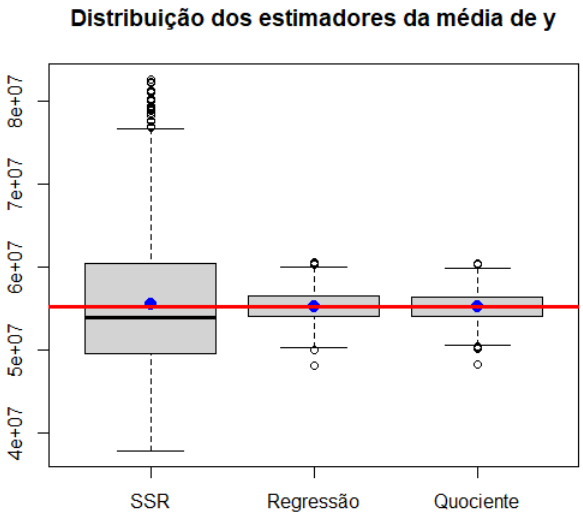



Figura 2: Diagrama de extremos e quartis dos estimadores

A média da variável $y$ está representada a vermelho. Como podemos ver, no caso dos estimadores da regressão e quociente a sua previsão parece coincidir com a média $y$. No caso da média amostral no plano SSR, nota-se que essa coincidência não é tão notória, ainda que exista.
Mais, no estimador do plano SSR existe maior quantidade de outliers, o que leva a uma maior dispersão nos valores da amostra, comparativamente aos estimadores da regressão e do quociente. Assim, podemos concluir que os estimadores da regressão e do quociente são, de facto, fiáveis para estimar a média de $y$.
Analisando o gráfico de extremos e quartis apenas dos dois estimadores em estudo observamos que estes são muito idênticos, sendo, por isso, difícil de concluir qual destes é o melhor.

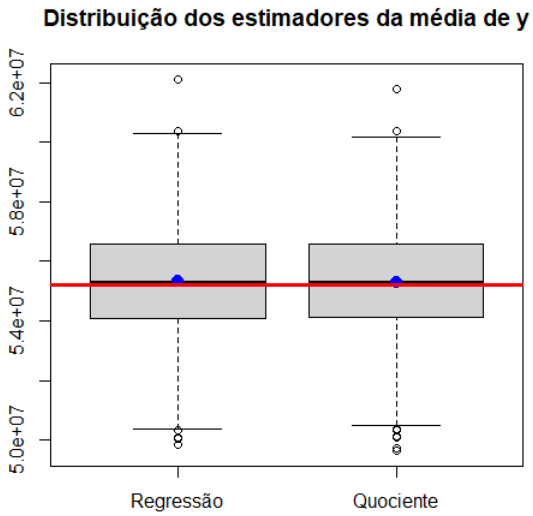



Figura 3: Diagrama de extremos e quartis dos estimadores em estudo

Teoricamente sabemos que um estimador é melhor que o outro quanto menor for o seu EQM. Como estamos perante amostras de dimensão 100, podemos considerar que o viés dos estimadores é próximo de zero, logo basta comparar as variâncias destes estimadores.
Veja-se a seguinte imagem.

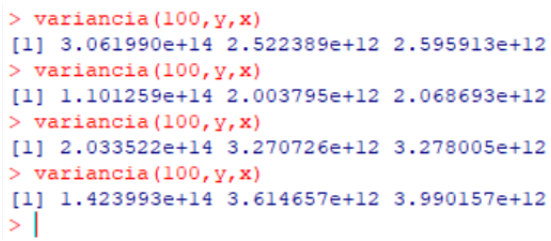



Figura 4: Quatro simulações da variância dos estimadores

Em cada simulação temos 3 valores apresentados: o valor da variância da média amostral num plano SSR, o valor da variância do estimador da regressão e o valor da variância do estimador do quociente.
Como podemos observar, o segundo valor em cada simulação é mais baixo que os outros dois valores, em particular, mais baixo que o terceiro. Logo, concluiu-se que o estimador da regressão é o mais preciso para estimar $\bar{y}$.

## População da Suíça (*swissmunicipalities*)

Temos agora uma população U de municípios da Suíça de tamanho $N = 2896$ e o estudo será análogo ao caso da população da Bélgica.
A nossa variável de interesse é o número de famílias por município que representamos por $u$. No entanto, pretendemos agora estimar $t_u$ e, para isso, vamos utilizar uma variável auxiliar $v$ que representa o total da população em cada um dos municípios.
Novamente, começámos por verificar que tipo de relação existia entre estas duas variáveis. Para isso, construimos o seguinte gráfico de dispersão.

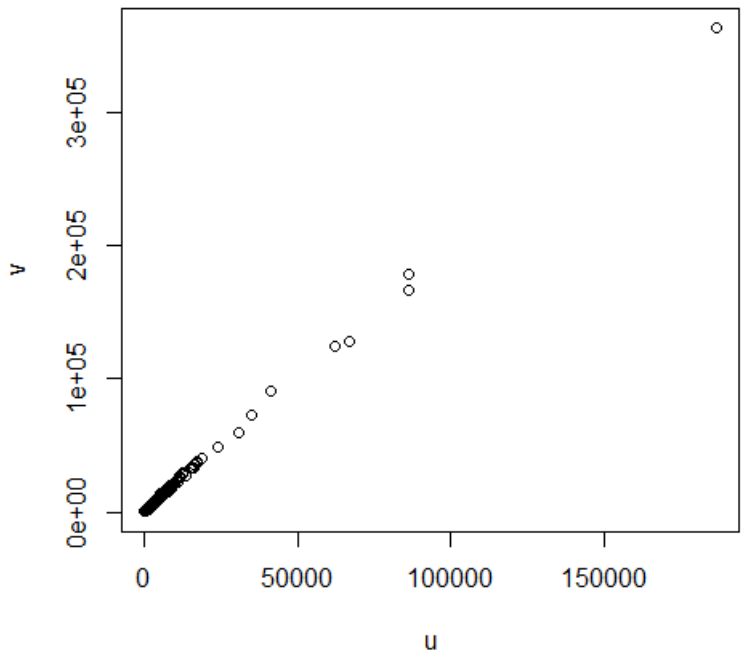
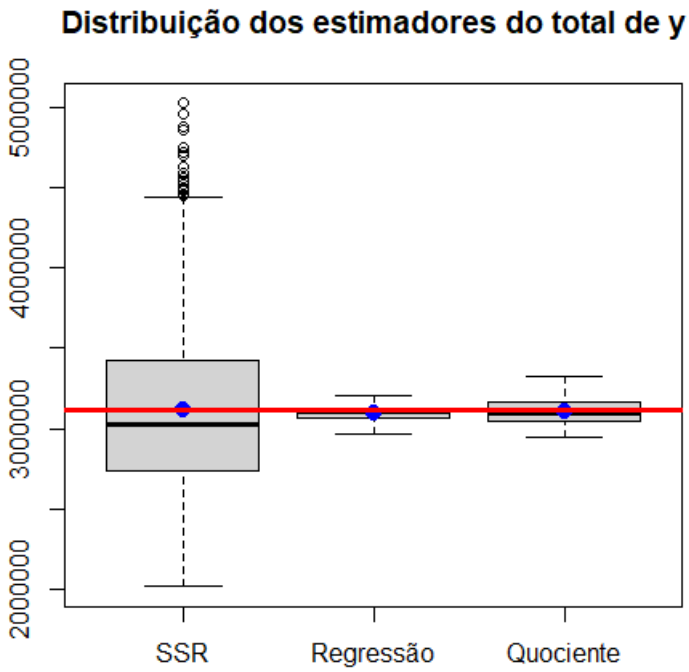
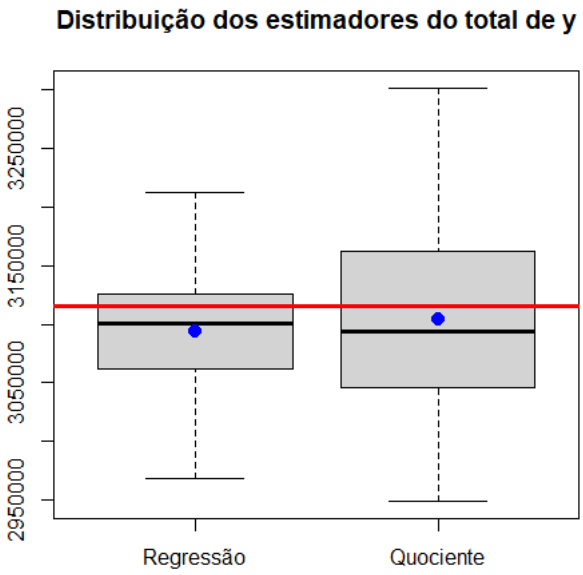



Figura 5: Gráfico de dispersão entre $u$ e $v$

Como podemos observar existe claramente uma relação linear entre as duas variáveis e, portanto, é razoável considerar o estimador do quociente e da regressão para estimar $\bar{y}$.
Para a simulação dos dados considerámos 2000 amostras de tamanho 600 retiradas através de um plano SSR. Começámos por gerar os diagramas de extremos de quartis para cada um dos estimadores.




O total da variável $u$ está representada a vermelho. Como podemos ver, no caso dos estimadores da regressão e quociente a sua previsão parece coincidir com o total de $u$, no entanto a amplitude interquartil no SSR é maior. Mais, no estimador do plano SSR existe maior quantidade de outliers, o que leva a uma maior dispersão nos valores da amostra, comparativamente aos estimadores da regressão e do quociente. Assim, podemos concluir que os estimadores da regressão e do quociente são, de facto, fiáveis para estimar o total de $y$.
Analisando o gráfico de extremos e quartis apenas dos dois estimadores em estudo observamos que há menor dispersão no estimador da regressão.




Teoricamente sabemos que um estimador é melhor que o outro quanto menor for o seu EQM. Como estamos perante amostras de dimensão 600, podemos considerar que o viés dos estimadores é próximo de zero, logo basta comparar as variâncias destes estimadores.
Veja-se a seguinte imagem.

```
> variancia(600,u,v)
[1] 238483108070    941087631   2675270046
> variancia(600,u,v)
[1] 51864421486   221164344   263511904
> variancia(600,u,v)
[1] 170129272326   1178133228   2443297719
> variancia(600,u,v)
[1] 286446223401   2256124015   4834661978
>
```

Em cada simulação temos 3 valores apresentados: o valor da variância da média amostral num plano SSR, o valor da variância do estimador da regressão e o valor da variância do estimador do quociente.
Como podemos observar, o segundo valor em cada simulação é mais baixo que os outros dois valores, em particular, mais baixo que o terceiro. Logo, concluiu-se que o estimador da regressão é o mais preciso para estimar $t_y$.

# IC e probabilidade de cobertura

Dado um intervalo de confiança, $I_n$ para um parâmetro $\theta$, entende-se por *probabilidade de cobertura* a $P(\theta \in I_n)$. Se temos um intervalo de confiança com nível de confiança $\alpha$, então espera-se que as probabilidades de cobertura sejam aproximadamente $\alpha$.

## População da Belgica (*belgianmunicipalities*)

Inicialmente começamos por construir um intervalo de confiança para cada um dos estimadores da média de $y$. Para uma amostra de tamanho 100, fizemos o estudo para um intervalo de confiança com nível de confiança 95% e obtivemos os seguintes resultados.

```
> IC(100,y,x,1.96)
         [,1]     [,2]     [,3]
[1,] 23998928 53860750 53045855
[2,] 70582628 56785715 58876872
>
```

Figura 6: IC para os estimadores SSR, regressão e quociente

Na figura anterior temos uma matriz 2 por 3 em que as colunas representam os extremos dos intervalos de confiança. Podemos ver que para o mesmo nível de confiança as amplitudes dos IC para os estimadores da regressão e do quociente são muito inferiores quando comparados ao do estimador SSR. Note-se também que a amplitude do IC obtido para o estimador da regressão é inferior ao do estimador do quociente.
De seguida fomos calcular as probabilidades de cobertura.

```
> PC(100,2000,y,x,1.96)
[1] 0.9910 0.9090 0.9145
> PC(100,2000,y,x,1.96)
[1] 0.9945 0.9080 0.9150
> PC(100,2000,y,x,1.96)
[1] 0.9940 0.9010 0.9145
> PC(100,2000,y,x,1.96)
[1] 0.9945 0.9170 0.9245
> PC(100,2000,y,x,1.96)
[1] 0.9905 0.9005 0.9080
```

Figura 7: Probabilidades de cobertura dos estimadores

Como se pode ver na figura 7 foram feitas 5 simulações e como a nossa amostra é de tamanho 100 podemos ver que as probabilidades de cobertura não coincidem com o nível de confiança, ficando proximas de 90%.
No entanto, sabemos que se aumentarmos o tamanho da nossa amostra verificamos que as probabilidades de cobertura também aumentam. Quando se tem $n \approx N$ estas probabilidades aproximam-se do nível de confiança 95%.
Para $n = 500$ e $n = 550$ obtivemos os seguintes resultados.

```
> IC(500,y,x,1.96)
         [,1]     [,2]     [,3]
[1,] 51229788 54610025 54591155
[2,] 58652318 55252278 55730179
>
```

Figura 8: IC para os estimadores SSR, regressão e quociente

```
> PC(500,2000,y,x,1.96)
[1] 0.946 0.942 0.947
> PC(550,2000,y,x,1.96)
[1] 0.9370 0.9435 0.9520
>
```

Figura 9: Probabilidades de cobertura dos estimadores

Como podemos ver na figura 9, foram feitas duas simulações, a primeira para uma amostra de tamanho 500 e a segunda para uma amostra de tamanho 550. Em ambas as amostras podemos agora ver que as probabilidades de cobertura já se encontram mais próximas dos 95%, como era de esperar.

No entanto, como o melhor estimador para estimar $\hat{t}_y$ seria o da regressão, seria de esperar que o valor das probabilidades de cobertura deste estimador fossem as melhores, tanto na amostra de tamanho 100, como para estas duas últimas. O que não acontece. Os valores destas probalidades são sempre melhores no estimador do quociente.
Isto pode dever-se ao facto de estarmos a lidar com populações reais, podendo, por isso, haver valores muito discrepantes e os resultados não irem de encontro ao esperado teoricamente. Ou pode ser devido ao facto de se estar a utilizar um quantil de uma lei normal, quando o estimadores não são bem aproximados por uma lei normal.
Desta forma, decidimos construir os gráficos QQ Plot de cada um dos estimadores.

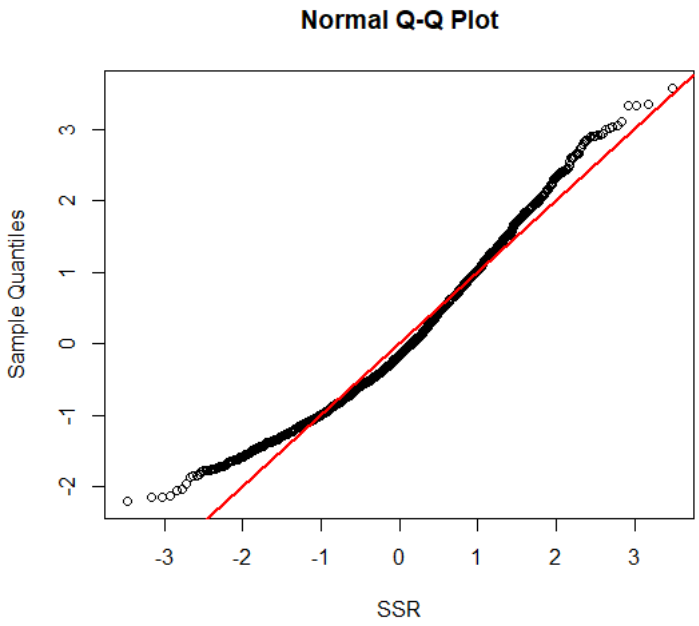



Figura 10: QQ-Plot do estimador SSR

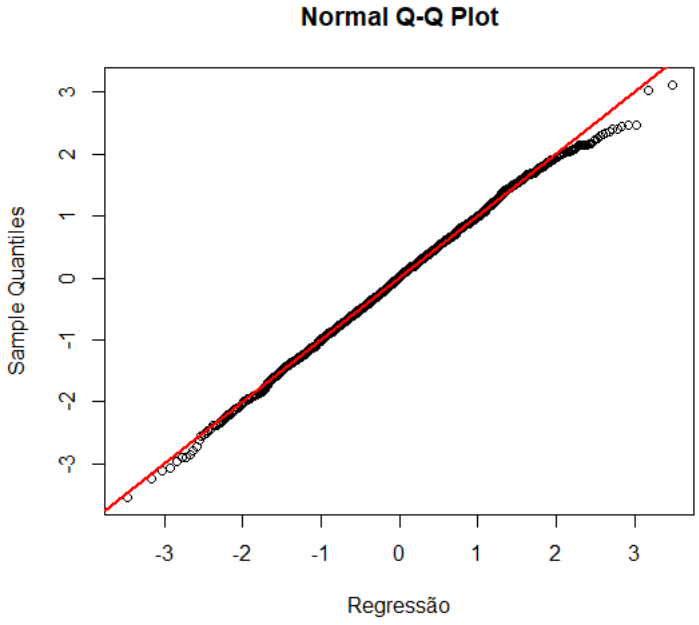



Figura 11: QQ-Plot do estimador da regressão

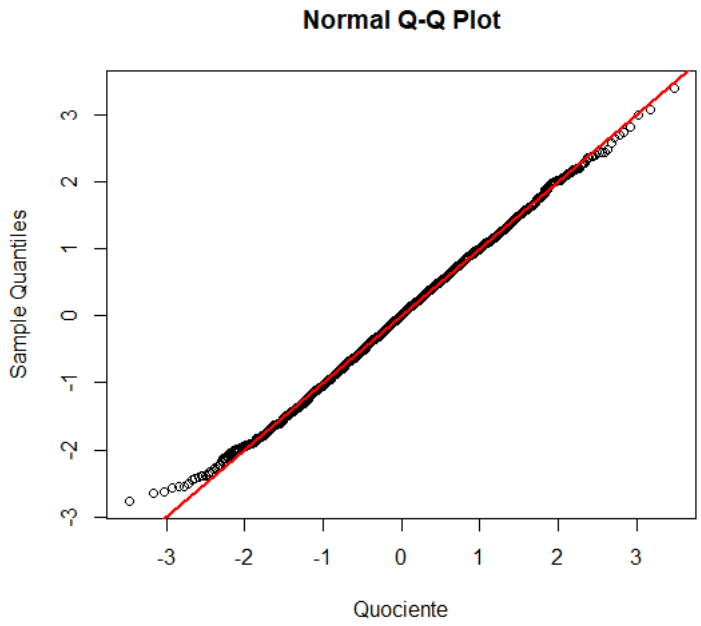



Figura 12: QQ-Plot do estimador do quociente

Observamos então que os estimadores da regressão e do quociente são melhor aproximados por uma lei normal centrada e reduzida. No entanto, isto não justifica o porquê de as probabilidades de cobertura do quociente serem melhores do que as do estimador da regressão. Portanto, este resultado deve-se ao facto de se estar a trabalhar com valores reais.

## População da Suíça (*swissmunicipalities*)

Novamente começamos por construir um intervalo de confiança para cada um dos estimadores do total $\hat{t_u}$. Para uma amostra de tamanho 600, fizemos o estudo para um intervalo de confiança com nível de confiança 95% e obtivemos os seguintes resultados.

```
> IC(600,u,v,1.96)
        [,1]    [,2]    [,3]
[1,] 2323581 3077042 2986229
[2,] 2811790 3124815 3153520
>
```

Figura 13: IC para os estimadores SSR, regressão e quociente

Na figura anterior temos uma matriz 2 por 3 em que as colunas representam os extremos dos intervalos de confiança. Podemos ver que para o mesmo nível de confiança as amplitudes dos IC para os estimadores da regressão e do quociente são muito inferiores quando comparados ao do estimador SSR. Note-se também que a amplitude do IC obtido para o estimador da regressão é inferior ao do estimador do quociente.
De seguida fomos calcular as probabilidades de cobertura para $n = 600$ e obtivemos os seguintes resultados:

```
> PC(600,2000,u,v,1.96)
[1] 0.8735 0.7890 0.7635
> PC(600,2000,u,v,1.96)
[1] 0.8770 0.7835 0.7645
> PC(600,2000,u,v,1.96)
[1] 0.8840 0.7935 0.7725
> PC(600,2000,u,v,1.96)
[1] 0.8720 0.7850 0.7675
> PC(600,2000,u,v,1.96)
[1] 0.8855 0.7925 0.7650
```

Figura 14: Probabilidades de cobertura dos estimadores

Como se pode ver na figura 14 foram feitas 5 simulações e como a nossa amostra é de tamanho 600 podemos ver que as probabilidades de cobertura não coincidem com o nível de confiança, ficando proximas de 90% para o estimador do SSR e de 80% para os estimadores da regressão e do quociente.

No entanto, sabemos que se aumentarmos o tamanho da nossa amostra verificamos que as probabilidades de cobertura também aumentam. Para $n = 1500$ e $n = 2000$ obtivemos os seguintes resultados.

```
> IC(1500,u,v,1.96)
        [,1]    [,2]    [,3]
[1,] 2319545 3042501 2916147
[2,] 3409753 3168127 3122245
>
```

Figura 15: IC para os estimadores SSR, regressão e quociente

```
> PC(1500,2000,u,v,1.96)
[1] 0.9160 0.8955 0.8700
> PC(2000,2000,u,v,1.96)
[1] 0.924 0.920 0.928
```

Figura 16: Probabilidades de cobertura dos estimadores

Como podemos ver na figura 16, foram feitas duas simulações, a primeira para uma amostra de tamanho 1500 e a segunda para uma amostra de tamanho 2000. Em ambas as amostras podemos agora ver que as probabilidades de cobertura já se encontram mais próximas dos 95%, como era de esperar.

Desta forma, decidimos construir os gráficos QQ Plot de cada um dos estimadores.

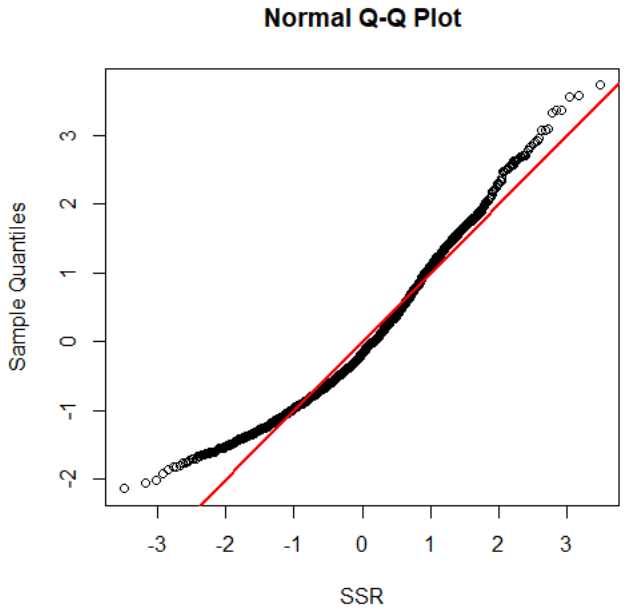



Figura 17: QQ-plot do estimador SSR

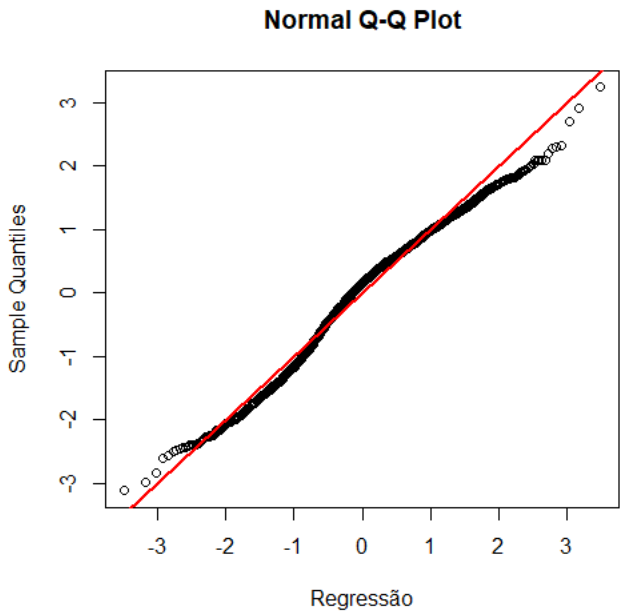



Figura 18: QQ-plot do estimador da regressão

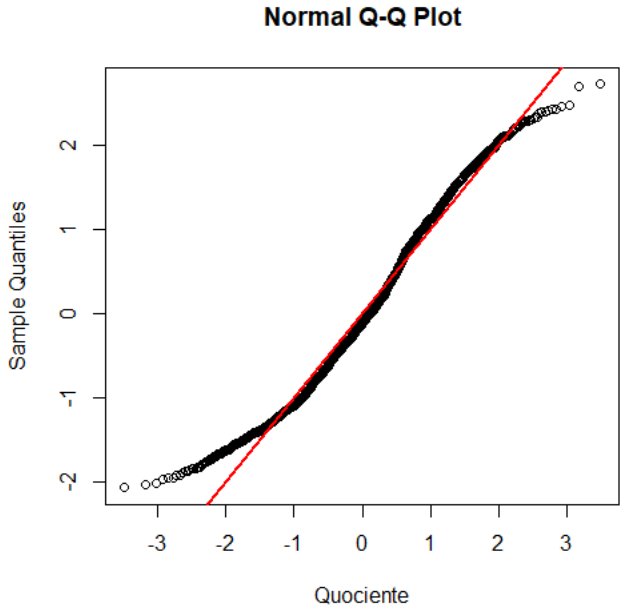



Figura 19: QQ-plot do estimador do quociente

Observamos então que os estimadores da regressão e do quociente são melhor aproximados por uma lei normal centrada e reduzida, no entanto isto não é tão evidente como na população anterior.

# 1 Conclusão

Neste projeto trabalhamos com valores reais, o que nos levou, por vezes, a resultados algo inesperados. Estes acontecem pelas más aproximações das leis de cada estimador.
Face aos resultados obtidos, concluímos que a eficiência de cada estimador depende bastante das características da população e da relação entre as variáveis de interesse com a auxiliar.

# Referências


[1] Carlos Tenreiro *Notas do curso de Amostragem e Sondagens.*

[2] Probabilidades de cobertura,
`https://www.rpubs.com/Loha911/probabilidade-de-cobertura-de-intervalo-de-confianca`