**Universidade de São Paulo**

**Faculdade de Filosofia, Letras e Ciências Humanas**

**Departamento de Ciência Política**

**Métodos Quantitativos e Técnicas em Ciência Política**

1º semestre / 2021

**Lab #11. Regressão Linear Simples e Correlação**

Suponha que, teoricamente, acredite-se que exista uma relação entre as variáveis *Y* e *X*. Essa relação pode ser medida quantitativamente a partir de uma equação. Se utilizarmos uma reta de regressão linear simples, a relação entre as variáveis seria representada pela equação:

$y=a+bx$ (1)

$\hat{y}=\hat{a}+\hat{b}x$ (2)

1. Interprete cada um dos termos das equações acima.

De acordo com este modelo de regressão linear simples, o parâmetro *b* pode ser representado pela seguinte relação:

$b=\frac{\sum\_{}^{}\left(y-\overbar{y}\right)\left(x-\overbar{x}\right)}{\sum\_{}^{}\left(x-\overbar{x}\right)^{2}}$ (3)

Esta equação está associada matematicamente aos conceitos de variância e de covariância.

1. Utilizando estes conceitos, reinterprete o termo *b*.

De acordo com a equação (1), para cada valor determinado de *x*, associaremos um único valor *y*. Isto implica dizer que para todas as unidades observadas, o mesmo par *(x, y)* seria encontrado. Esta situação geraria um modelo determinista, o que não estaria de acordo com a interpretação geral dos fenômenos sociais. Para transforma-la em uma relação probabilística, reescrevemos a equação original da seguinte maneira:

$E(y)=a+bx$ (4)

1. Interprete os termos desta nova equação. O que muda em relação à resposta (a)?

Quando se acredita que haja uma relação teórica entre *Y* e *X*, não se espera que a função matemática que indique esta relação seja precisa nas estimativas. Mas se matematicamente a relação é tal como apresentada em (2), inferimos a média, mas não para o valor de cada observação particular. Isto implicaria em representar a equação (1) novamente, introduzindo um novo termo. A equação ficaria assim:

$y=a+bx+e$ (5)

1. O que significa a introdução deste termo *e*? Qual sua interpretação?

Esta é uma equação da relação *populacional*. Na prática, não podemos observar *e*.

1. Como, então, entendemos este termo *e*? Ele é representado por qual outro conceito empírico?

Este conceito empírico será aquele utilizado para que possamos aferir a adequação do nosso modelo aos dados.

1. Como ele é obtido empiricamente? Quais são os limites em seu uso? Explique.

Por fim, suponha que uma pesquisadora calcule o coeficiente de correlação *ρ* para as variáveis *Y* e *X* ao invés de utilizar o modelo de regressão linear simples.

1. Qual a relação entre este índice de correlação e o termo *b*? Há alguma situação em que estes valores se aproximarão? Qual?
2. Por que *ρ* e *b* são diferentes? Explique.