Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política

FLS 5028
Métodos Quantitativos e Técnicas de Pesquisa em Ciência Política

FLP0406
Métodos e Técnicas de Pesquisa em Ciência Política
1º semestre / 2016

Glauco Peres da Silva

LISTA DE EXERCÍCIOS 09 (GABARITO)

Data de entrega: 30/05/2016 (noturno) e 01/06/2016 (vespertino).

Exercício 01 (3 pontos)

Assinale os enunciados abaixo com (V) verdadeiro ou (F) falso. Em caso de falso, indique o erro e o corrija, justificando sua resposta.

(F) I. O modelo de regressão linear é determinista: cada valor de x corresponde a um único valor de y.

Falso. O modelo de regressão linear é probabilístico, e leva em consideração a variabilidade de y para cada valor da x. Em outras palavras, em um modelo probabilístico, para cada x, os valores de y tendem a variar seguindo uma distribuição normal (AGRESTI & FINLAY, 2012:298).

(F) II. A ocorrência de valores *outliers* em uma distribuição afeta a equação da reta de um modelo linear, mas não a correlação entre as variáveis.

Falso. Valores *outliers* afetam tanto a equação da reta quanto a correlação entre as variáveis. Lembrando, a correlação entre duas variáveis pode ser obtida pela seguinte fórmula $r = \left(\frac{S_x}{S_y}\right) b$. O b, por sua vez, é encontrado pela seguinte equação $b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$. Portanto, um valor *outlier* afeta a variância, e esta impacta sobre a estimação do Beta, o qual, por sua vez, influência o valor da correlação.

(F) III. Um modelo de regressão linear pode ser aplicado com sucesso ainda que seja violado o princípio da linearidade, desde que haja aleatorização.

Um modelo de regressão não é robusto contra a violação da condição de linearidade. Em outras palavras, é possível que as variáveis possuam associação, mas que esta não seja linear. Se esse for o caso e um modelo linear for imposto, o pesquisador tenderá a crer que há independência quando não o há, incorrendo no Erro de Tipo II (a hipótese nula é falsa, mas não é rejeitada). Conforme recomendam Agresti e Finlay (2012), um gráfico de

dispersão é necessário para que seja observado se as variáveis possuem um comportamento passível de ser modelado de forma linear.

(F) IV. O coeficiente r^2 informa sobre os ganhos preditivos adquiridos quando utilizamos a média de y para prever os valores das observações de y. Quanto maior o coeficiente r^2, mais seguros ficamos de que o conhecimento de x não nos fornece melhores meios para prever y, em comparação à média de y.

O r^2 nos informa sobre os ganhos preditivos adquiridos quando utilizamos, para prever y, os valores preditos de y(obtidos por meio de uma equação linear), em comparação a uma previsão que parta da média de y. Quanto maior o coeficiente r^2, mais seguros ficamos de que a equação linear nos fornece informações mais precisas quanto aos possíveis valores de y. Um r^2 equivalente a 0,6, por exemplo, nos diz que y varia 60% menos quando utilizamos a informação de x em comparação ao uso da média de y.

(F) V. As estimativas dos parâmetros populacionais α e β devem ser estimativas por ponto. Em outras palavras, em regressões lineares, a construção de intervalos de confiança para os parâmetros mencionados não é possível.

As estimativas dos parâmetros populacionais α e β não precisam ser estimativas por ponto. A construção de intervalos de confiança é possível, e se faz de modo semelhante aos intervalos construídos para médias: $b \pm$ t(ep).

<u>Exercício 02</u> (4 pontos)

Parte da literatura brasileira (AMES, 2003; AMORIM NETO; SANTOS, 2003; PEREIRA & MUELLER, 2002, 2003) argumenta que as emendas individuais orçamentárias – instrumento à disposição dos parlamentares para a alteração do orçamento – são essenciais para a garantia da estabilidade na relação entre o Executivo e o Legislativo. Como o orçamento no Brasil é autorizativo e não impositivo, ou seja, o Executivo é quem decide quais e como os recursos serão gastos, esses autores argumentam que o presidente utilizaria dessas emendas como uma ferramenta para garantir que sua agenda de políticas seja aprovada. Nesse argumento, os parlamentares somente votariam de acordo com o governo mediante a execução de suas emendas individuais. Figueiredo e Limongi (2008), por outro lado, mostraram empiricamente que a associação entre a liberação das emendas individuais orçamentárias e o apoio ao governo não é tão forte como os autores supracitados argumentam.

Para resolver o exercício, utilize o banco de dados XXX e seu codebook, disponíveis no Moodle.

Mínimo de 3 e máximo de 10 linhas por item (a, b, c, d).

a-) Um primeiro passo para começar a investigar a relação entre duas variáveis é através da representação gráfica. Pensando o problema acima, responda:

I. Quais variáveis você escolheria para testar?Dado o enunciado do exercício, qual é a dependente e qual é a independente? De que tipo elas são?

A variável dependente é "Taxa_disciplina" e a independente "Taxa_execução", elas são quantitativas contínuas (proporções). Segundo o codebook, sabemos que a primeira nos diz a disciplina dos parlamentares com relação ao governo e a segunda a média de recursos proveniente de emendas individuais que o parlamentar teve liberada ou executada pelo governo.

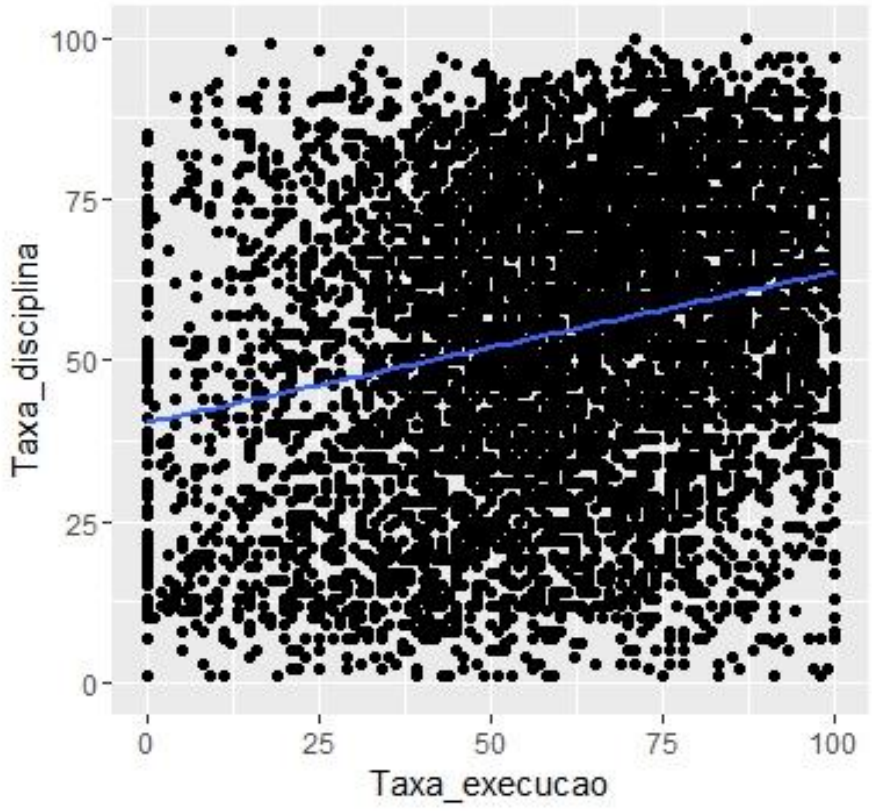
II. Assumindo a perspectiva de Ames (2003), Santos (2003) e Pereira e Mueller (2002; 2003), o que você esperaria encontrar? Justifique.

Assumindo que AMES, 2003; AMORIM NETO; SANTOS, 2003; PEREIRA & MUELLER, 2002, 2003 estão certos, deveríamos esperar uma correlação positiva entre as duas variáveis, uma vez que o aumento da taxa de execução junto com a disciplina é uma evidência do argumento segundo o qual os parlamentares obedeceriam o governo (taxa de disciplina) em troca da aprovação de suas propostas (taxa de execução).

III. Qual o melhor gráfico para ver essa relação? Execute-o e interprete-o.

Uma vez que estamos lidando com duas variáveis quantitativas, o melhor gráfico para visualizar a relação é o de dispersão. Como a *taxa de execução* é a variável explicativa, ela deve ficar no eixo X, por sua vez, a variável dependente *taxa de disciplina* deve ficar no eixo Y.

O Gráfico fica:



Como podemos ver, os pontos estão muito dispersos e espalhados nos quatro quadrantes do gráfico, impedindo a visualização de qualquer padrão de relacionamento entre as duas variáveis.

b-) A **covariância** é "um modo estatístico para resumir um padrão de associação geral (ou falta dele) entre duas variáveis contínuas" (Kellstendt&Whitten: 2015, p.183). Seu cálculo se dá pela formula:

$$COV_{xy} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Resultados positivos descrevem uma relação positiva e resultados negativos uma relação negativa.O que o resultado nos diz sobre a relação entre as variáveis? Ele é condizente com o esperado a partir do gráfico?

A covariância entre as duas variáveis é 0,0113, um valor positivo, mas muito pequeno. Ou seja, possuímos mais uma evidência de que as variáveis não estão relacionadas, ou seja, para não aceitar a posição defendida por AMES, 2003; AMORIM NETO; SANTOS, 2003; PEREIRA& MUELLER, 2002, 2003.

c-) Para saber a força dessa relação precisamos dar mais um passo na nossa análise e olhar o **coeficiente de correlação (r)**. Defina esse coeficiente e sua utilidade na análise, em seguida, calcule-o e interprete-o com base na teoria.

A fórmula para o cálculo do coeficiente de Pearson é:

$$r = \frac{COV_{xy}}{\sqrt{var_x var_y}}$$

Em que var_x e var_y representam a variância de X e a variância de Y, respectivamente.

Agresi e Finlay (2012), colocam que a correlação é uma versão padronizada da reta que descreve a relação entre X e Y, de modo que "a correlação é o valor que inclinação assumiria para unidades tais que as variáveis tenham desvios padrão iguais" (p.312). Ela é importante na análise uma vez que é capaz de nos dizer a força da associação entre as variáveis.

O coeficiente de correlação varia entre -1 e 1, de modo que valores absolutos altos indicam relações fortes. No caso em análise, r é igual a 0,23, o que nos diz que a relação entre as duas variáveis é muito fraca, uma vez que o coeficiente de correlação está bem mais próximo de zero do que de um. Pensando na teoria, novamente temos uma evidência de

que os parlamentares não seguem o governo em busca de recursos, via emendas individuais.

d-) Por fim, execute o teste de hipóteses para essa correlação. Adote um nível confiança de 95% (significância de 05%). O que podemos concluir?

O teste em questão é o demonstrado pro Agresti e Finlay (2012) na página 314. O primeiro passo é formular as hipóteses a serem testadas. No caso, verificaremos se a correlação entre as variáveis (ρ) é estatisticamente diferente de zero:

$$H_0: \rho = 0$$

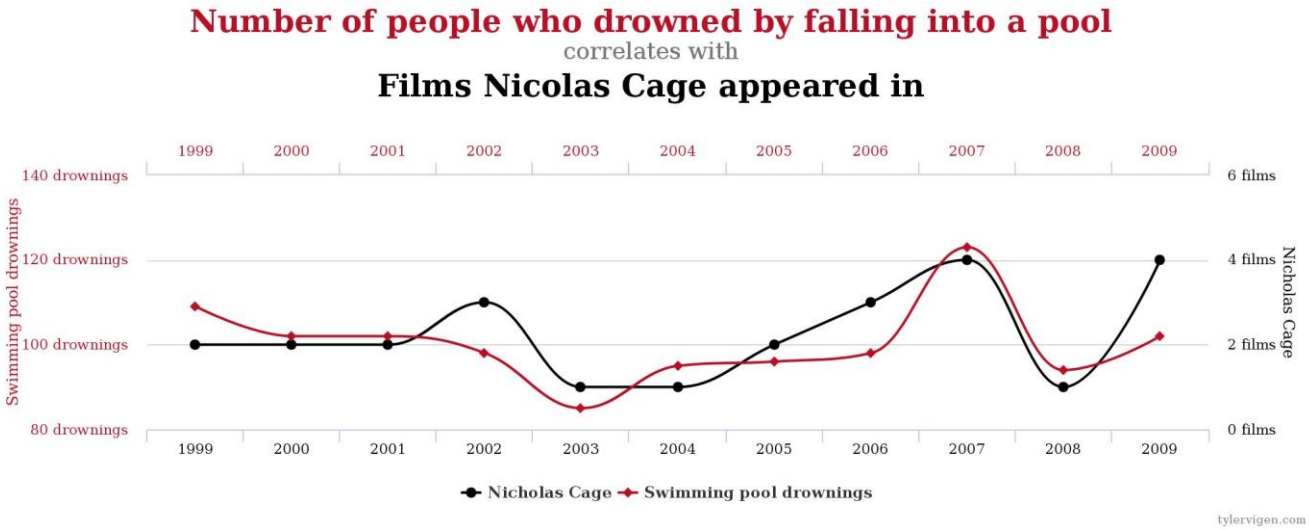
$$H_a: \rho \neq 0$$

O cálculo da estatística t se dá pela fórmula:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n-2)}} = 19{,}70$$

Como podemos ver, essa é uma estatística t extremamente grande. Dessa forma, olhando a tabela de distribuição t podemos concluir, com nível de significância superior a 99%, que o coeficiente de correlação é diferente de zero.

Exercício 03 (3 pontos)



O Gráfico acima apresenta o relacionamento entre as variáveis "número de pessoas afogadas ao cair em uma piscina" e "filmes em que Nicolas Cage apareceu". O índice de correlação é de 0.66, indicando um relacionamento positivo moderado entre as variáveis. Um colega, também pesquisador, diz que quer empreender uma pesquisa acadêmica sobre esse relacionamento, uma vez que conseguiu boas evidências estatísticas de causalidade entre as variáveis.

Diante dessa situação, podemos perceber que o pesquisador está duplamente equivocado. Visando o aprimoramento de seu colega como pesquisador, quais críticas você faria? **Mínimo de 10 linhas.**

Nessa questão, esperava-se que o aluno se atentasse para duas questões. A primeira diz respeito à relevância do que o pesquisador pretende estudar, bem como o erro de sua motivação. Uma pesquisa precisa ser baseada em uma teoria, empreender um estudo apenas a partir de uma correlação não faz sentido, o que nesse exemplo é claro pelo absurdo da relação, mas pode acontecer em situações de identificação mais difícil. O segundo, estreitamente ligado ao anterior, versa sobre a afirmação de que correlação não quer dizer causalidade, ou seja, duas variáveis fortemente correlacionadas não necessariamente possuem uma relação de X causando Y. Esse ponto pode ser defendido de várias formas, uma delas é o fato da correlação não estipular o sentido da relação, nesse caso, a não ser que tenhamos uma justificativa teórica que prove o contrário, Y pode causar X. Ademais, como será estudado adiante, outras precauções precisam ser tomadas para afirmar com mais segurança uma relação causal.

<u>Exercício 04: Pós-Graduação</u> (5 pontos)

Agresti e Finlay (2012) elaboram alguns passos a serem seguidos quando o objetivo é a construção de um modelo baseado em uma regressão linear. No *Moodle* está disponível o banco de dados que será utilizado neste exercício.

a-) O banco de dados apresenta a taxa de homicídio e o IDH para 58 distritos da cidade de São Paulo, amostrados aleatoriamente. Assuma como variável dependente o número de homicídios e, como variável independente, o IDH (a variável foi convertida de modo a variar entre 1 e 10). Estime a equação linear de previsão e interprete cada um dos coeficientes. Dica: utilize as fórmulas e substitua os valores.

$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = -6{,}838$

$a = \bar{y} - b\bar{x} = 11{,}091 - (-6{,}838 \times 8{,}493) = 69{,}167$

$\hat{y} = -69{,}167 - 6{,}838x$

Por se tratar de uma amostra, não conhecemos os verdadeiros parâmetros alpha e beta, já que estes se referem à população. Utilizaremos os dados amostrais para estima-los e, por esse motivo, estes foram indicados na equação como *a* e *b* (em contraposição a α e β). De modo similiar, o y recebe um acento circunflexo, para assinalar que se trata de um valor estimado, mais especificamente, à média estimada de y para determinado valor x.

Na equação, *a* se refere ao intercepto e *b* à inclinação da reta. O intercepto revela o valor estimado de y para quando x é igual a zero. Desse modo, quando o valor do IDH é zero, esperaríamos uma taxa de homicídio equivalente a 69,16. O *b* se refere à alteração prevista em y para o aumento de uma unidade em x. Assim, o aumento de uma unidade no IDH gera uma redução de 6,838 na taxa de homicídio.

b-) Construa um intervalo de confiança de 99% para o valor estimado de Beta e expresse as hipóteses nula a alternativa. Podemos rejeitar a hipótese nula?Dica: utilize as fórmulas e substitua os valores.

H_0: b = 0

H_a: b ≠ 0

$b \pm$ t(ep)

ep = $\frac{s}{\sqrt{\sum(x-\bar{x})^2}}$, onde s = $\sqrt{\frac{SQE}{n-1}}$

SQE = $\sum(y-\hat{y})^2$ = 3237,39

s = $\sqrt{\frac{SQE}{n-1}} = \sqrt{\frac{3237,39}{57}} = \sqrt{56,79}$ = 7,53

ep = $\frac{s}{\sqrt{\sum(x-\bar{x})^2}}$ = ep = $\frac{7,53}{\sqrt{21,57}}$ = 1,622

Grau de liberdade = n – 2 = 56. De acordo com a tabela, t é aproximadamente 2,7.

$b \pm$ t(ep) = b + 2,7(1,622) = -2,456

$b \pm$ t(ep) = b - 2,7(1,622) = -11,219

b está entre -11,219 e -2,456

Podemos rejeitar a hipótese nula com 99% de confiança. Em outras palavras, temos 99% de confiança de que *b* não é igual a zero.

c-) Calcule a soma dos quadrados totais (SQT) e a soma dos quadrados do erro (SQE). Há diferença? O que podemos concluir? Dica: utilize as fórmulas.

SQE = $\sum(y-\hat{y})^2$ = 3237,39

SQT = $\sum(y-\bar{y})^2$ = 4246,24

Como podemos notar, a soma dos quadrados do erro (SQE) é inferior à soma dos quadrados totais (SQT). Isso nos informa que, ao tentarmos prever os valores de y utilizando a equação linear de previsão, erramos menos do que o faríamos caso utilizássemos apenas a média amostral.

d-) Calcule o r^2. O que o seu valor nos informa sobre o poder preditivo do modelo e sobre a relação entre a taxa de homicídios e o IDH dos distritos da cidade de São Paulo? O valor

encontrado é surpreendente? Qual é a relação entre valor obtido e o fato de se tratar de uma análise bivariada?

$r^2 = \frac{SQT - SQE}{SQT} = \frac{4246{,}24 - 3237{,}39}{4246{,}24} = 0{,}237$

O r^2 nos informa sobre os ganhos preditivos adquiridos quando utilizamos, para prever y, os valores preditos de y (obtidos por meio de uma equação linear), em comparação a uma previsão que parta da média de y. Um r^2 equivalente a 0,237 nos diz que y varia 23,7% menos quando utilizamos a informação de x em comparação ao uso da média de y. Em outras palavras, podemos dizer que o IDH explica 23,7% da variabilidade da taxa de homicídio.

Um r^2 equivalente a 0,237 pode ser considerado pequeno. Contudo, a equação de previsão utilizou apenas uma variável independente, dificultando o poder de previsão do modelo. Considerado individualmente, o IDH pode ser considerado um importante preditor para as taxas de homicídio na cidade de São Paulo, embora outras variáveis devam ser incorporadas no modelo.

e-) Explique o que são variabilidade condicional e variabilidade marginal. Como a variabilidade condicional se relaciona ao modelo de regressão linear? Por que a variabilidade condicional tende a ser inferior à variabilidade marginal quando a relação entre as variáveis atende ao princípio da linearidade? Dica: faça referência ao teorema do limite central.

A variabilidade condicional se refere à variação de y para determinado valor de x. Em um modelo de regressão linear, cada valor de x possui uma distribuição condicional de y diferente. O modelo de regressão prevê a média do valor de y para cada x, desse modo, como se trata de uma média, o teorema do limite central permite que as distribuições condicionais sejam consideradas normais. A variabilidade marginal, por sua vez, se refere à variabilidade das observações de toda a distribuição.

Quando as variáveis possuem um relacionamento linear, menor é o erro de previsão quando se utiliza uma equação linear. Quanto menor o erro, menor a variabilidade de y para um mesmo valor de x, ou seja, menor a variação condicional.

Boa lista!