ECONOMETRIA I - LISTA 2

MODELO DE REGRESSÃO LINEAR SIMPLES

Mateus Cardoso

27/04/2021

As fórmulas

- Variância amostral: $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. A variância é uma medida que mostra o quão distantes da média estão os valores. Para obtê-la, devemos:

1) encontrar a média ($\bar{x}$);

2) subtrair a média de cada valor da amostra ($x_i - \bar{x}$), isso será a variação em torno da média;

3) elevar ao quadrado os resultados dessas subtrações[1];

4) somar os resultados ($\sum(x_i - \bar{x})^2$) e

5) dividir por $n - 1^2$(O 2 é uma nota de rodapé)($\frac{\sum(x_i - \bar{x})^2}{n-1}$).

Note que nos passos 4 e 5 estamos calculando uma média. Esta é a média do quadrado da variação em torno da média amostral.

O problema da variância é que, por estar elevada ao quadrado, possui difícil interpretação. Para resolver isso, convertemos a variância para *desvio-padrão*, que nada mais é do que a raiz quadrada da variância. Com este valor, podemos saber o quão usual é uma variação entre uma observação e a média em relação a variação que normalmente é vista (desvio-padrão).

- Covariância amostral: $Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$. Funciona da mesma forma que a variância, exceto nos passos

[1] Elevamos ao quadrado pois a variância é o segundo momento da distribuição. Os momentos caracterizam as distribuições de probabilidade. Para uma variável aleatória X, o n-ésimo momento em torno de uma constante c é definido como $E[(X - c)^n]$. Sendo $c = E(X)$, temos que o segundo momento central é $E[(X - E(X))^2] = Var(X)$. O terceiro momento ($n = 3$) é chamado assimetria e o quarto momento ($n = 4$) é chamado curtose. Vemos, portanto, que elevar não é uma escolha arbitrária, mas uma consequência dos momentos da distribuição.

[2] Dividimos por $n - 1$ pois estamos fazendo um ajuste de graus de liberdade, eliminando o viés gerado por ter estimado a média.

1) encontrar a média para cada variável ($\bar{x}$ e $\bar{y}$);

2) subtrair a média das duas diferentes variáveis ($(x_i - \bar{x})$ e $(y_i - \bar{y})$) e

3) multiplicar o resultado de uma variável pelo resultado da outra ($(x_i - \bar{x})(y_i - \bar{y})$).

A covariância mede o quanto duas variáveis variam de forma conjunta ou distinta. Se elas tendem a estar acima da média ao mesmo tempo, então multiplicar uma pela outra produzirá um resultado positivo para a maioria das observações, aumentando a covariância. Se elas não tiverem nada a ver uma com a outra, multiplicá-las produzirá ora resultados positivos, ora resultados negativos, que quando somados no passo 4, levarão a covariância para 0.

Agora, podemos ver como a variância e a covariância nos levam para o estimador de Mínimos Quadrados Ordinários.

- Mínimos Quadrados Ordinários (MQO): Sabemos que a formula para o estimador de Mínimos Quadrados Ordinários é $\hat{\beta} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$. Olhando com atenção, vemos que na verdade esta fórmula é $\frac{Cov(x,y)}{Var(x)}$.[3]

De maneira simples, esta fórmula está dizendo: de toda a variação em x, quanto dela varia junto de y?

Tendo estimado o coeficiente angular $\hat{\beta}$, podemos estimar o intercepto por meio de $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

- Soma de quadrados total (SST): $\sum(y_i - \bar{y})^2$. Representa a variação total na variável dependente.

- Soma de quadrados da regressão (SSE): $\sum(\hat{y}_i - \bar{y})^2$. Representa a variação explicada pela regressão.

- Soma de quadrados dos resíduos (SSR): $\sum \hat{u_i}^2$. Representa a variação não explicada pela regressão.

Da mesma forma que $y_i = \hat{y}_i + \hat{u}$, a variação total é igual a variação explicada pela regressão mais a variação não explicada, ou seja $SST = SSE + SSR$.

- R^2: O R-quadrado mede a fração da variação total que é explicada pela regressão. Ou seja, $R^2 = \frac{SSE}{SST}$ ou $R^2 = \frac{SSR}{SST}$(total (1) menos a fração que não é explicada pela regressão).

Exercícios

1) Quais são as hipóteses do modelo de regressão linear simples?

[3] E os $n-1$? Estes são anulados na divisão $\frac{Cov(x,y)}{Var(x)}$. Como $\frac{Cov(x,y)}{Var(x)} = \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1} * \frac{n-1}{\sum(x_i-\bar{x})^2}$. Podemos ver que os $n-1$ se anulam, restando portanto $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$.

2) Explique o que é o coeficiente de determinação (R^2).

3) (Exercício slides Aula 2 e 3) Com base na amostra de 10 valores, calcular as estimativas dos parâmetros do modelo: $Y_i = \alpha + \beta X_i + u_i$.

X	Y
0	3
1	2
1	3
2	5
3	4
3	4
4	7
5	6
5	7
6	9

a) Determine as estimativas dos parâmetros.

b) Calcular o coeficiente de determinação (R^2).

c) Calcular a variância do erro σ^2.

d) (Questão extra) Verifique se $\bar{\hat{y}} = \bar{y}$. Explique o porquê.

4) Explique a diferença entre resíduo e erro.

5) (Wooldridge, Cap. 2) Seja *kids* o número de filhos de uma mulher, e *educ* os anos de educação da mulher. Um modelo simples que relaciona a fertilidade a anos de educação é

$$kids = \beta_0 + \beta_1 educ + u \quad ,$$

em que u é um erro não observável.

a)Que tipos de fatores estão contidos em u? É provável que eles estejam correlacionados com o nível de educação?

b) Uma análise de regressão simples mostrará o efeito *ceteris paribus* da educação sobre a fertilidade? Explique.

6) (Gujarati, Cap.3) A tabela abaixo mostra os dados de renda familiar semanal (X) e despesas familiares de consumo semanal (Y).

X (US$)	Y (US$)
80	70
100	65
120	90
140	95
160	110
180	115
200	120
220	140
240	155
260	150

Calcule:

a) As estimativas dos parâmetros β_0 e β_1.

b) O coeficiente de determinação.

c) A elasticidade-renda do consumo (Modelo log-log), interpretando os coeficientes.

Exercícios no R

1) (Cunningham, Cap. 2.14) Utilizando o comando rnorm(), simule:

- x: 10.000 observações aleatórias que seguem a distribuição normal, com desvio-padrão igual a 9;
- u: 10.000 observações aleatórias que seguem a distribuição normal, com desvio-padrão igual a 36;
- y: resultado da verdadeira relação entre as variáveis $y = 3 + 2x + u$;
- $\hat{y}$: o resultado da regressão linear de y em x;
- $\hat{u}$: os resíduos da regressão.

Verifique:

a) Se $\sum \hat{u}_i = 0$.

b) Se $\bar{y} = \bar{\hat{y}}$.

2) Use os dados do arquivo COUNTYMURDERS para responder a essas questões. Utilize somente os dados de 1996.

(i) Quantos condados tiveram zero assassinatos em 1996? Quantos condados tiveram pelo menos uma execução? Qual é o maior número de execuções?

(ii) Estime a equação abaixo, em que murders corresponde ao número de assassinatos

$$murders = \beta_0 + \beta_1 execs + u$$

por MQO e relate os resultados da forma usual, incluindo o tamanho da amostra e o R-quadrado.

(iii) Interprete o coeficiente de inclinação registrado no item (ii). A equação estimada sugere um efeito dissuasor da pena capital?

(iv) Qual é o menor número de assassinatos que pode ser previsto pela equação? Qual é o resíduo de um condado com zero execuções e zero assassinatos?

(v) Explique por que uma análise de regressão simples não é adequada para determinar se a pena capital tem um efeito dissuasor sobre os assassinatos.

Referências

HUNTINGTON-KLEIN, Nick. **The Effect:** An Introduction to Research Design and Causality. Disponível em: https://www.nickchk.com/book/The_Effect_DRAFT_DO_NOT_REHOST.pdf. Acesso em 28/04/2021.

CUNNINGHAM, Scott. **Causal Inference:** The Mixtape. Disponível em: https://mixtape.scunning.com/index.html. Acesso em 03/05/2021.

WOOLDRIDGE, Jeffrey. **Introdução à Econometria:** Uma Abordagem Moderna. 3. ed. São Paulo: Cengage, 2019.

GUJARATI, Damodar. **Econometria básica**: 4. ed. Rio de Janeiro: Elsevier, 2006.