Revisão Regressão

Manoel Galdino

2024-03-14

Notação

Vamos começar revisando algumas notações matemáticas que usaremos ao longo do curso.

Somatório

Se eu tenho uma sequências de números $x_1, x_2, ..., x_n$, a soma dessa sequência é dada por:

$$x_1 + x_2 + \ldots + x_n \coloneqq \sum_{i=1}^{n} x_i$$

Ás vezes, quando ficar claro no contexto quais os elementos que estão sendo somados (como nesse caso, que são toda a sequência de x_1 até x_n), dispensaremos os índices do somatório e simplesmente escreveremos $\sum x_i$.

O operador somatório é linear e, portanto, possui algumas propriedades comuns a operadores lineares.

- Para qualquer constante c, $\sum_{i=1}^{n} c \equiv nc$
- Para qualquer constante c, $\sum_{i=1}^{n} cx_i \equiv c\sum_{i=1}^{n} x_i$
- A soma de somatórios é idêntico à somatória das somas, isto é: $\sum_{i=1}^{n}(x_i + y_i) \equiv \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$
- Para quaisquer constantes a e b, $\sum_{i=1}^{n}(ax_i + by_i) \equiv a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$

Por fim, vale destacar relações que não são em geral verdadeiras, isto é, não são propriedades do somatório.

- O somatório de uma razão **não é** a razão do somatório: $\sum_{i=1}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}$
- o somatório de uma variável ao quadrado **não é** igual ao somatório da variável ao quadrado: $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$

Vamos usar somatório para definir a média:

$$\bar{x} \coloneqq \frac{\sum_{i=1}^{n} x_i}{n}$$

Uma propriedade envolvendo a média e o somatório é que somar a diferença de uma variável aleatória para a média é zero.

$$\sum_{i=1}^{n}(x_i - \bar{x}) \equiv 0$$

Uma resultado útil é:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{1}$$

Exercício em sala: prove esse resultado. Dica: expanda o quadrado, aplique as propriedades do somatório, reescreva ora o somatório como uma média, ora a média como somatório, coloque em eivdência e simplifique.

Esperança

O valor esperado de uma variável aleatória é chamado de esperança ou média populacional. Para uma variável aleatória discreta X que pode assumir os valores $x_1, x_2, \ldots, x_n$ cada um com probabilidade $p(x_1), p(x_2), \ldots p(x_n)$ possui esperança definida por:

$$\mathbb{E}[X] \coloneqq p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n = \sum_{i=1}^{2} p(x_i)x_i$$

O operador esperança é linear e, portanto, possui algumas propriedades comuns a operadores lineares.

- Para qualquer constante c, $\mathbb{E}[c] \equiv c$
- Para qualquer constante a, $\mathbb{E}[aX] \equiv a\mathbb{E}[X]$
- Para quaisquer constantes a e b, $\mathbb{E}[aX + b] \equiv a\mathbb{E}[X] + b$
- Para ...
- A soma de somatórios é idêntico à somatória das somas, isto é: $\sum_{i=1}^{n}(x_i + y_i) \equiv \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$
- Para quaisquer constantes a e b, $\sum_{i=1}^{n}(ax_i + by_i) \equiv a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$

Variância

A variância de uma variável aleatória X é dada por:

Definição 1. $Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Uma propriedade útil é a chamada identidade da variância:

$$\begin{align}
\mathrm{Var}(X) &\coloneqq \mathbb{E}[(X - \mathbb{E}[X])^2] \equiv \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{2}\\
& \tag{3}\\
Vamosprovaraidentidade: & \tag{4}\\
\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (X - \mathbb{E}[X])] & \tag{5}\\
&= \mathbb{E}[X^2 - 2 \cdot \mathbb{E}[X] \cdot X + (\mathbb{E}[X])^2] \tag{2. Aplicando a regra do quadrado}\\
&= \mathbb{E}[X^2] - \mathbb{E}[2\mathbb{E}[X]X] + \mathbb{E}[(\mathbb{E}[X])^2] \tag{3. Propriedade da esperança}\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[\mathbb{E}[X]X] + \mathbb{E}[(\mathbb{E}[X])^2] \tag{6}\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X] \cdot \mathbb{E}[X] + (\mathbb{E}[X])^2 \tag{4. $\mathbb{E}[aX] = a\mathbb{E}[X]$}\\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{5. Simplificando}
\end{align}$$

Covariância

A Covariância de duas v.a. X e Y é definida como: $Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$.

Notem que $Cov(X,X) = Var(X)$.

A covariância é positiva quando ambos X e Y tendem a ter valores acima (ou abaixo) de sua média simultaneamente, enquanto ela é negativa quando uma v.a. tende a ter valores acima da sua média e a outra abaixo.

3. Identidade da Covariância

$Cov(X,Y) = \mathbb{E}[X * Y] - \mathbb{E}[X] * \mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X]) * (Y - \mathbb{E}[Y])]$ Exercício para o leitor. Prove que isso é verdade.

4, Covariância é simétrica

$Cov(X,Y) = Cov(Y,X)$

5. Variância não é linear $Var(a * X + b) = a^2 * Var(x)$
6. Covariância não é linear

$Cov(a * X + b, Y) = a * Cov(Y,X)$

Regressão

Normalmente, nós assumimos que existe um modelo populacional, dado pela equação de regressão populacional:

$$y = \beta_0 + \beta_1 x + u$$

Nós chamamos o y de variável dependente, resposta, explicada, predita etc. Chamamos o x de preditor, variável independente, variável explicativa etc.

Os parâmetros β_0 e β_1 não são dados e, portanto, não podem ser observados. No máximo, podem ser estimados, a partir de dados e suposições críveis. Portanto, a **credibilidade das suposições** será um tema constante no curso. Como veremos, a moderna prática de inferência causal gira em torno de estratgégias de identificação críveis, isto é, que adotam suposições críveis.

E o termo de erro u resume todas as demais variáveis que impactam y e que não estão explicitamente consideradas no modelo na forma funcional especificada.

Suposições

1. Sem perda de generalidade (portanto, uma suposição simplificadora sem maiores consequências) é que o valor esperado de μ é zero na população. Formalmente:

$$\mathbb{E}[u] = 0 \tag{7}$$

2. Independência na média

Vamos assumir que o termo de erro u é independente do preditor para cada valor de x. Formalmente,

$$\mathbb{E}[u|x] = \mathbb{E}[u] \tag{8}$$

Combinando equação 6 com 7, chegamos a equação 8:

$$\mathbb{E}[u|x] = 0 \tag{9}$$

Essa suposição é chamada de suposição de média condicional zero e é crítica em modelos de regressão. Isso porque ela implica que:

$$\mathbb{E}[y|x] = \beta_0 + \beta_1 x \tag{10}$$

A equação 9 mostra que a função de regressão populacional é linear em X, o que Angrist e Pischke chamam de Função de Esperança Condicional (CEF, na sigla em inglês).

Quando a suposição 8 é satisfeita, poderemos interpretar β_1 como um prâmetro causal.

OLS

Lembremos que:

$$\mathbb{E}[u|x] = 0 \implies \mathbb{E}[ux] = 0 \tag{11}$$

Podemos provar isso usando a Lei da Esperanças Iteradas:

$$\begin{aligned} \mathbb{E}[ux] = \mathbb{E}[\mathbb{E}[ux]] = & \qquad (12) \\ \mathbb{E}[x\mathbb{E}[u]] = & \qquad \text{(1. usando o fato de que } x \text{ é fixo)} \\ \mathbb{E}[x \cdot 0] = \mathbb{E}[0] = 0 & \qquad (13) \end{aligned}$$

Portanto, basta usar as condições nas equações 6 e 8 para obter estimativas para os parâmetros da regressão.

Vejam que $\mathbb{E}[ux] = 0$, então podemos provar que $\mathbb{C}\rtimes\gtrapprox[u, x] = 0$.

$$\mathbb{E}[u] = \mathbb{E}[y - \beta_0 - \beta_1 x] = 0\mathbb{E}[u|X] = \mathbb{E}[y - \beta_0 - \beta_1 x] = 0 \tag{14}$$

$$\mathbb{E}[xu] = \mathbb{E}[x(y - \beta_0 - \beta_1 x)] = 0 \tag{15}$$

Essas são as duas condições que determinam os valores dos parâmetros β_0 e $beta_1$ na população. Como em geral temos apenas uma amostra para estimar os parâmetros, podemos estimá-los pelos chamados "plug-in estimators", isto é, a contraparte amostral da fórmula populacional. No caso da esperança, precisamos calcular as médias:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{16}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)) = 0 \tag{17}$$

Simplificando as equações 15 e 16, temos:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \tag{18}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 x_i = \tag{19}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \tag{20}$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \tag{21}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{22}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i)) = 0 \tag{1. substituindo a equação de $\hat{\beta}_0$ em 16}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_1\bar{x}x_i - \hat{\beta}_1 x_i x_i) = 0 \tag{23}$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i^2 - \bar{x}x_i) = \tag{24}$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i) = \tag{25}$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}\bar{x}) = \tag{26}$$

$$\tag{27}$$

Lemrbando o resultado 1 de somatório, temos:

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \tag{28}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \tag{1. usando resultado 2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{29}$$

E o resíduo (para distringuir do erro) é a diferença entre a previsão do modelo, dada por $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1$ e o valor observado y_i.

$$\hat{u} = y_i - \hat{y_i} = y_i - (\hat{\beta}_0 + \hat{\beta}_1)$$

É possível mostrar, usando cálculo, que o estimador de OLS é exatamente dado por essa fórmula. Basta escrever a fórmula do quadrado dos resíduos e minimizá-la.

Propriedades do modelo de regressão

Por contrução, a soma dos resíduos é sempre zero.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \frac{1}{n}\sum_{i=1}^{n}\hat{u} = 0 \tag{30}$$

Também por contrução, a covariância amostral entre os resíduos e os preditores é sempre zero.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)) = 0 \frac{1}{n}\sum_{i=1}^{n}(x_i\hat{u}) = 0 \tag{31}$$

Por fim, como o $\hat{y}_i$ é uma função linear de x_i, a covariância amostral entre $\hat{y}_i$ e o resíduo $\hat{u}_i$ também é zero.

Propriedades do estimador

Com a suposição de independência média do erro em relação a x, podemos provar que o estimador de OLD é não-viesado, isto é:

$$\mathbb{E}[\hat{\beta}_0] = \beta_0$$

E também:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

As suposições chave nessa demonstração são: modelo linear, amostra aleatória iid e independência na média do erro em relação aos preditores.

Teorema da anatomia da regressão

Uma última propriedade da regressão que quero discutir aqui (e que é útil pela conexão com o viés de variável omitida) é o que Angrist and Psichke chamaram de anatomia da regressão.

Vamos supor que queremos estimar o efeito causal do número de filhas mulheres sobre a ideologia dos pais. A hipótese é que mais filhas mullheres tornam os pais mais liberais (no sentido americano). Se o número de filhas mulheres for aleatório, então a nossa equação de regressão $Y_i = \beta_0 + \beta_1 x_i + u_i$ captura, por meio de β_1 o efeito causal médio de número de filhas mulheres sobre a ideologia dos pais. Isso porque, se for realmente aleatório, então $\mathbb{E}[u|x] = 0$. Porém, x provavelmente não é aleatório, já que muitas pessoas tentam ter um certo número de filhas. Para fins pedagógicos, vamos supor que é condicionalmente aleatório, isto é, se condicionarmos na idade e renda dos pais. Nossa equação de regressão fica então:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 R_i + \beta 3 I_i + u_i$$

em que R_i é a renda da família i e I_i é a idade média da família i. O teorema da anatomia da regressão detalha como interpretar β_1.

De modo ,geral, com k variáveis preditoras, temos:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + ... + \beta_k x_{Ki} + e_i$$

Vamos definir uma regressão auxiliar, em que regredimos x_{1i} em função de todos os outros preditores:

$$x_{1i} = \gamma_0 + \gamma_{k-1}x_{k-1i} + \gamma_{k+1}x_{k+1i} + ... + \beta_K x_{Ki} + f_i$$

Defina o resíduo $\tilde{x} = x_{1i} + \hat{x}_{1i}$. Então, o parâmetro β_1 da equação principal de regressão pode ser reescrito como:

$$\beta_1 = \frac{Cov(y_i, \tilde{x})}{Var(\tilde{x})}$$

Vamos ver o exemplo que o Scott dá no livro dele:

```
library(tidyverse)
library(haven)
library(ggplot2)
library(stargazer)

read_data <- function(df) {
  full_path <- paste0("https://github.com/scunning1975/mixtape/raw/master/",
                      df)
  haven::read_dta(full_path)
}

auto <-
  read_data("auto.dta") %>%
  mutate(length = length - mean(length))

lm1 <- lm(price ~ length, auto)
lm2 <- lm(price ~ length + weight + headroom + mpg, auto)
lm_aux <- lm(length ~ weight + headroom + mpg, auto)
auto <-
  auto %>%
  mutate(length_resid = residuals(lm_aux))

lm2_alt <- lm(price ~ length_resid, auto)

coef_lm1 <- lm1$coefficients
coef_lm2_alt <- lm2_alt$coefficients
resid_lm2 <- lm2$residuals

y_single <- tibble(price = coef_lm2_alt[1] + coef_lm1[2]*auto$length_resid,
                   length_resid = auto$length_resid)

y_multi <- tibble(price = coef_lm2_alt[1] + coef_lm2_alt[2]*auto$length_resid,
                  length_resid = auto$length_resid)
```

```
stargazer(lm1, lm2, type = "latex",
          title = "Resultados das Regressões",
          model.names = FALSE,
          intercept.bottom = FALSE,
          column.labels = c("Modelo 1", "Modelo 2"),
          covariate.labels = c("Intercepto", "Comprimento", "Peso", "Altura Interna", "MPG"),
          omit.stat = c("LL", "ser", "f"),
          digits = 2)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: seg, mar 18, 2024 - 17:42:51

Table 1: Resultados das Regressões

	Dependent variable:	
	price	
	Modelo 1	Modelo 2
	(1)	(2)
Intercepto	6,165.26***	−3,581.38
	(311.40)	(4,538.81)
Comprimento	57.20***	−94.50**
	(14.08)	(40.40)
Peso		4.34***
		(1.16)
Altura Interna		−490.97
		(388.49)
MPG		−87.96
		(83.59)
Observations	74	74
R^2	0.19	0.37
Adjusted R^2	0.18	0.34
Note:	*p<0.1; **p<0.05; ***p<0.01	

```r
stargazer(lm2_alt, type = "latex",
          title = "Resultado das Regressõe Resóiduo",
          omit.stat = c("LL", "ser", "f"),
          digits = 2)
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: seg, mar 18, 2024 - 17:42:51

```r
auto %>%
  ggplot(aes(x=length_resid, y = price)) +
  geom_point() +
  geom_smooth(data = y_multi, color = "blue") +
  geom_smooth(data = y_single, color = "red") +
  annotate("text", x = Inf, y = Inf, label = "bivariada, inclinação 57.2",
           hjust = 1, vjust = 1, color = "red", size = 4, angle = 0) + # Anotação
  annotate("text", x = 0, y = Inf, label = "multivariada, inclinação -94.5",
           hjust = 1, vjust = 1, color = "blue", size = 4, angle = 0)
```

Table 2: Resultado das Regressõe Resóiduo

	Dependent variable:
	price
length_resid	−94.50*
	(48.64)
Constant	6,165.26***
	(336.54)
Observations	74
R^2	0.05
Adjusted R^2	0.04
Note:	*p<0.1; **p<0.05; ***p<0.01