Minicurso de Introdução a R

Modelos de Regressão

Caio G. V. Coutinho

PET
ECONOMIA | UFPE

Sumário

Conceitos

Econometria

Econometria significa "**medição econômica**", mas seu escopo é muito mais amplo.

Definição: Econometria

Aplicação da estatística matemática [em especial, a inferência estatística] a dados econômicos para dar suporte empírico aos modelos formulados pela economia matemática e obter resultados numéricos. (Tintner, 1968)

E a arte do econometrista está em encontrar o conjunto de hipóteses suficientemente específicas e realistas que lhe permitam tirar o melhor proveito dos dados de que dispõe (Malinvaud, 1966).

Método

O **método clássico** (não-Bayesiano) da econometria consiste sucintamente em:

1. Exposição da teoria ou hipótese;
2. Especificação do modelo;
3. Obtenção dos dados;
4. Estimação dos parâmetros;
5. Teste de hipótese;
6. Previsão;
7. Uso para fins de controle ou política.

Caso da Propensão Marginal a Consumir (PmgC) de Keynes.

Regressão

Criada por Galton (1886), a regressão é a **principal ferramenta da econometria**.

Definição: Regressão

Estudo da dependência de uma variável (a **dependente**, prevista, endógena, *output*, de interesse) em relação a uma ou mais (as **explicativas**, independentes, regressoras, exógenas, de controle), visando **estimar o valor médio** da população da primeira em termos dos valores conhecidos (em amostragens repetidas) das segundas.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

As variáveis do modelo de regressão são quantitativas (discretas ou contínuas) e estocásticas (aleatórias), i.e., seguem uma distribuição de probabilidade.

É impossível prever com perfeição o resultado de uma variável explicativa sobre a dependente. Há sempre uma variabilidade "intrínseca" de y que não pode ser explicada por x. Dessa forma, a econometria foca em entender o efeito médio.

Correlação e Causalidade

"Uma relação estatística, por mais forte e sugestiva que seja, nunca pode estabelecer uma conexão causal: nossas ideias de causação devem vir de fora da estatística, em última análise, de alguma teoria." Ou seja, **correlação não é causalidade**.

Hipóteses do *core*:

1. A amostra não é viesada, $E\left(\varepsilon|x\right) = 0$;
2. O termo de erro e a variável explicativa são independentes, $E\left(x\varepsilon\right) = 0$;
3. Não há autocorrelação entre os termos de erro, $Cov\left(\varepsilon_i, \varepsilon_j\right) = 0$;
4. O nº de observações é superior ao nº de parâmetros a serem estimados.

Hipóteses auxiliares:

1. Existe uma relação linear entre as variáveis;
2. Os resíduos são homocedásticos.

Acerca da notação empregada, caso a regressão possua apenas duas variáveis, tem-se uma **regressão simples**. Caso mais que duas, uma **regressão múltipla**.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + \cdots + \beta_1 x_{ni} + \varepsilon_i$$

Em que y_i indica a i-ésima observação da variável dependente, x_i indica a i-ésima observação da variável de controle e ε_i representa o erro – a variabilidade da variável de interesse que não é explicada pela(s) variável(is) independente(s).

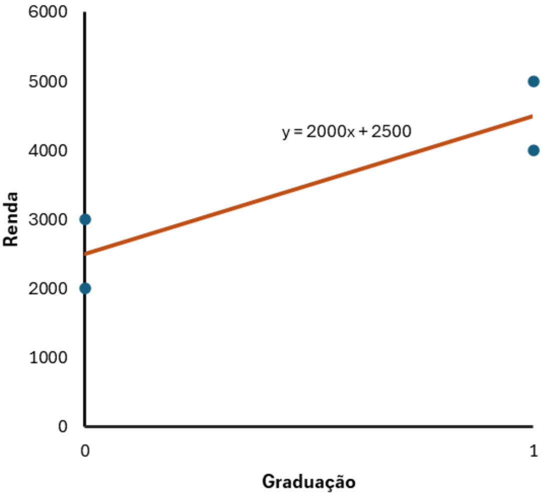
Prática 01

Frequentemente – e em especial na avaliação de choques únicos, como políticas –, as variáveis x_i assumem uma natureza binária. Ou seja, assumem valor 0 ou 1.

Variáveis Binárias

São variáveis *dummy* aquelas que apresentam uma escala nominal, i.e., são qualitativas. E.g., possuir graduação. Diferenciam-se, portanto, das variáveis proporcionais, como renda, altura, etc.

	(y)	(x)
i	Renda	Graduação
1	2000	0
2	3000	0
3	4000	1
4	5000	1



Prática 01

Práticas

Dados

Será utilizada a tradicional base de dados auto.dta. As variáveis dessa base:

01. make: Modelo do carro;

02. price: Preço do carro;

03. mpg: Milhas por galão (uma medida de eficiência de combustível);

04. rep78: Avaliação de reparo de 1978 (com valores de 1 a 5);

05. headroom: Espaço acima da cabeça (em polegadas);

06. trunk: Capacidade do porta-malas (em pés cúbicos);

07. weight: Massa do carro (em libras);

08. length: Comprimento do carro (em polegadas);

09. turn: Diâmetro de viragem (em pés);

10. turn: displacement: Deslocamento do motor (em polegadas cúbicas);

11. gear_ratio: Relação de transmissão final;

12. foreign: Indica se o carro é estrangeiro ou doméstico.

Pacotes

```r
# Pacotes necessários
install.packages("readstata13")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("broom")
install.packages("car")
install.packages("MASS")
install.packages("repr")

# Bibliotecas necessárias
library(repr)
library(readstata13)
library(dplyr)
library(ggplot2)
library(broom)
library(car)
library(MASS)
library(scales)
library(lmtest)
library(quantreg)
```

Prática 01: Regressão Simples e Múltipla

```
# Importando os dados
df <- read.dta13("seu_caminho/auto.dta")
head(df)

# Gráfico de dispersão
theme_set(theme_minimal())
options(repr.plot.width = 16, repr.plot.height = 9, repr.plot.res = 700)
ggplot(df, aes(x = weight, y = price)) +
  geom_point(shape = 21, fill = "white", color = "#222631", size = 3,
             stroke = 0.5) +
  labs(title = "Prática 01: Gráfico de Dispersão",
       x = "Massa",
       y = "Preço") +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black")
  )
```

```r
# Um modelo de regressão simples
model_lm <- lm(price ~ weight, data = df)
summary(model_lm)

# Gráfico de dispersão com a reta de regressão
ggplot(df, aes(x = weight, y = price)) +
  geom_point(shape = 21, fill = "white", color = "#222631", size = 3,
      stroke = 0.5) +
  geom_smooth(method = "lm", color = "#aa3f3b", se = FALSE) +
  scale_x_continuous(labels = comma_format(big.mark = ".")) +
  scale_y_continuous(labels = comma_format(big.mark = ".")) +
  labs(title = "Prática 01: Regressão Simples",
       x = "Massa",
       y = "Preço") +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black")
  )
```

```r
# Prevendo o preço do automóvel de Massa (4000)
predict(model_lm, newdata = data.frame(weight = 4000))

# Encontrando previsões para o modelo
predictions <- predict(model_lm)
head(predictions)

# Plotando os valores observados e ajustados
df$predicted <- predictions
theme_set(theme_minimal())
options(repr.plot.width = 16, repr.plot.height = 9, repr.plot.res = 700)
ggplot(df, aes(x = seq_along(price))) +
  geom_line(aes(y = price), color = "#16215b") +
  geom_line(aes(y = predicted), color = "red", linetype = "dashed") +
  labs(title = "Prática 01: Previsão",
     x = "Massa",
     y = "Preço") +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black"))
```

```r
# Um modelo de regressão múltipla (Massa e comprimento)
model <- lm(price ~ weight + length, data = df)
summary(model)

# Um modelo de regressão múltipla (outras variáveis)
excluir <- c("foreign", "make", "predicted")
df1 <- df[, !(names(df) %in% excluir)]
model <- lm(price ~ ., data = df1, na.action = na.omit)
summary(model)
```

Voltar

Prática 01: Variáveis *Dummy*

```r
# Variáveis dummy para foreign
df <- df %>% mutate(dum_Foreign = as.factor(foreign))
df1 <- cbind(df1, model.matrix(~dum_Foreign - 1, data = df))
model <- lm(price ~ ., data = df1, na.action = na.omit)
summary(model)
```

Voltar

Prática 02: Modelos não-lineares

```r
# Regressão log-linear
model_ln <- lm(log(price) ~ weight + length, data = df)
summary(model_ln)

# Gráfico de dispersão com a reta de regressão
ggplot(df, aes(x = weight, y = log(price))) +
  geom_point(shape = 21, fill = "white", color = "#222631", size = 3,
      stroke = 0.5) +
  geom_smooth(method = "lm", color = "#aa3f3b", se = FALSE) +
  scale_x_continuous(labels = comma_format(big.mark = ".")) +
  scale_y_continuous(labels = comma_format(big.mark = ".")) +
  labs(title = "Prática 02: Modelos não-lineares",
       x = "Massa",
       y = "Preço") +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black"))
```

Prática 03: Regressão Robusta

```r
# Teste de Heterocedasticidade
df %>%
  mutate(residuos = model_lm$residuals) %>%
  ggplot(data = ., aes(y = residuos, x = weight)) +
  geom_point() +
  geom_abline(slope = 0) +
  theme_classic()
bptest(model_lm)

# Regressão robusta
model_rlm <- rlm(price ~ weight, data = df)
summary(model_rlm)
```

Prática 03: Regressão Robusta

```r
# Gráfico de dispersão com a reta de regressão
ggplot(df, aes(x = weight, y = price)) +
  geom_point(shape = 21, fill = "white", color = "#222631", size = 3,
      stroke = 0.5) +
  geom_smooth(method = "rlm", color = "#aa3f3b", se = FALSE) +
  scale_x_continuous(labels = comma_format(big.mark = ".")) +
  scale_y_continuous(labels = comma_format(big.mark = ".")) +
  labs(title = "Prática 03: Regressão Robusta",
       x = "Massa",
       y = "Preço") +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black"))
```

```r
# Comparando visualmente os modelos linear e robusto
ggplot(df, aes(x = weight, y = price)) +
  geom_point(shape = 21, fill = "white", color = "#222631", size = 3,
      stroke = 0.5) +
  geom_smooth(aes(color = "Modelo Robusto"), method = "rlm", se = FALSE) +
  geom_smooth(aes(color = "Modelo Linear"), method = "lm", se = FALSE) +
  scale_x_continuous(labels = scales::comma_format(big.mark = ".")) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".")) +
  labs(title = "Prática 03: OLS vs. RLM",
       x = "Massa",
       y = "Preço") +
  scale_color_manual(name = "", values = c("Modelo Linear" = "#152a6d", "
      Modelo Robusto" = "#aa3f3b"),
                     labels = c("Modelo Linear", "Modelo Robusto")) +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.line = element_line(color = "black"),
    legend.position = "bottom")
```

Prática 04: Regressão Quantílica

```r
# Regressão Quantílica
model_quant <- rq(price ~ weight, data = df, tau = 0.5)
model_quant

# Automatizando os gráficos com uma função
rq_smooth <- function(method = "rq", se = FALSE, tau, color) {
  if (method == "rq") {
    return(geom_smooth(aes(color = color), method = method, se = se,
                       method.args = list(tau = tau)))
  } else {
    return(geom_smooth(aes(color = color), method = method, se = se))
  }
}
```

```r
# Plotando o gráfico de dispersão com diferentes retas
ggplot(df, aes(x = weight, y = price)) +
  geom_point(shape = 21, fill = "white", color = "#222631",
             size = 3, stroke = 0.5) +
  rq_smooth(method = "rq", se = FALSE, tau = 0.5,
            color = "Modelo Quantílico Q(50)") +
  rq_smooth(method = "rq", se = FALSE, tau = 0.1,
            color = "Modelo Quantílico Q(10)") +
  rq_smooth(method = "rq", se = FALSE, tau = 0.9,
            color = "Modelo Quantílico Q(90)") +
  geom_smooth(aes(color = "Modelo Linear"), method = "lm", se = FALSE) +
  scale_x_continuous(labels = scales::comma_format(big.mark = ".")) +
  scale_y_continuous(labels = scales::comma_format(big.mark = ".")) +
  labs(title = "Prática 04: Regressão Quantílica",
       x = "Massa",
       y = "Preço") +
```

```
scale_color_manual(name = "",
                   values = c("Modelo Linear" = "#152a6d",
                     "Modelo Quantílico Q(50)" = "#aa3f3b",
                     "Modelo Quantílico Q(10)" = "#255f00",
                     "Modelo Quantílico Q(90)" = "#330c4b"),
                   labels = c("Modelo Linear", "Modelo Quantílico Q(50)"
                       ,
                            "Modelo Quantílico Q(10)",
                            "Modelo Quantílico Q(90)")) +
theme(
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  axis.line = element_line(color = "black"),
  legend.position = "bottom"
)
```

Apêndice

Exogeneidade Estrita

A exogeneidade estrita é o primeiro pressuposto basilar da Econometria.

Exogeneidade Estrita (Hipótese 01)

O somatório dos desvios em relação à média é 0 (zero).

Esse resultado é demonstrado por:

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

Dado que $\sum_{i=1}^{n} d_i = 0$,

$$\frac{1}{n}\sum_{i=1}^{n} d_i = 0\frac{1}{n} \Rightarrow \bar{d} = 0$$

Ou seja, a média dos desvios em relação à média é também igual a 0. Portanto, para N (o equivalente populacional), independentemente da variável:

$$E\left(\varepsilon|x\right) = 0 \tag{1}$$

Dessa forma, demonstra-se a **hipótese 1** (H_1): **a expectância dos erros em relação à média populacional é igual a 0**. Caso contrário, terá sido feita uma seleção amostral enviesada que não permitirá uma análise correta dos dados.

Independência entre o erro e a variável

O segundo pressuposto basilar da Econometria é demonstrado através da Lei das Expectativas Iteradas (LEI).

LEI

A média de uma variável é igual à média das médias.

$$E\left(y\right) = E\left[E\left(y|x\right)\right]$$

Substituindo a variável qualquer pelo erro (desvio em relação à média populacional):

$$E\left(\varepsilon\right) = E\left[E\left(\varepsilon|x\right)\right]$$

Sob a Hipótese 1:

$$E(\varepsilon) = E[0] = 0$$

Ainda, tendo em vista que $E(xy|x) = xE(y|x)$, é possível afirmar que a **covariância entre o erro e uma variável qualquer é também igual a 0**:

$$cov(x, \varepsilon) = E(x\varepsilon) - E(x)E(\varepsilon) = E(x\varepsilon)$$

$$E(x\varepsilon) = E[E(x\varepsilon|x)] = E[xE(\varepsilon|x)] = E(0) = 0$$

$$cov(x, \varepsilon) = E(x\varepsilon) = 0 \tag{2}$$

Estimação de Parâmetros

Os parâmetros podem ser estimados por 3 métodos centrais: método de máxima verossimilhança (ML), método de momentos (MM) e método de mínimos quadrados ordinários (OLS). Além disso, via somatório e via matriz.

Expresso o erro por $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$, pela simplicidade da regressão simples, a estimação através do método de **mínimos quadrados ordinários** com a **minimização dos desvios quadráticos**:

$$\min_{\{\beta_0,\beta_1\}} \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

$$\frac{\partial \varepsilon_i}{\partial \beta_0} = 2\left(-1\right)\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \Rightarrow \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i\right) - \sum_{i=1}^{n}\left(\beta_0\right) - \sum_{i=1}^{n}\left(\beta_1 x_i\right) = 0$$

Ao serem multiplicados todos os fatores por $1/n$ e sabendo que o somatório de uma constante é n multiplicado por ela, ou seja, $\sum_{i=1}^{n}\left(\beta_0\right) = n\beta_0$:

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

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

Para o parâmetro β_1:

$$\frac{\partial \varepsilon_i}{\partial \beta_1} = 2\left(-1\right)\left(x_i\right)\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0 \Rightarrow \sum_{i=1}^{n}\left[y_i - \left(\bar{y} - \beta_1 \bar{x}\right) - \beta_1 x_i\right]\left(x_i\right) = 0$$

$$\sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) - \left(x_i - \bar{x}\right)\beta_1\right]\left(x_i\right) = 0 \Rightarrow \sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right)\left(x_i\right) - \left(x_i\right)\left(x_i - \bar{x}\right)\beta_1\right] = 0$$

$$\sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right)\left(x_i\right)\right] - \beta_1\sum_{i=1}^{n}\left(x_i\right)\left(x_i - \bar{x}\right) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i\right)}{\sum_{i=1}^{n}\left(x_i\right)\left(x_i - \bar{x}\right)}$$

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

R-Quadrado

R-quadrado (R^2) demonstra a capacidade explicativa do modelo a partir da variabilidade dos valores estimados. Indica o quão bem ajustada é a reta em relação aos dados. Dado que a média do grupo do elemento pode ser expresso por $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i$$

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$

Subtraída média da variável dependente e aplicada a potência:

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + \hat{\varepsilon}_i \Rightarrow (y_i - \bar{y})^2 = (\hat{y}_i - \bar{y} + \hat{\varepsilon}_i)^2$$

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 - 2(\hat{y}_i - \bar{y})\hat{\varepsilon}_i + \hat{\varepsilon}_i^2$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 - \sum 2 (\hat{y}_i - \bar{y}) \hat{\varepsilon}_i + \sum \hat{\varepsilon}_i^2$$

Sob a hipótese 1, tem-se que $\sum 2 (\hat{y}_i - \bar{y}) \hat{\varepsilon}_i = 0$. Portanto:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2$$

Dessa forma, dividido todos os termos por $\sum (y_i - \bar{y})^2$:

$$\frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum \hat{\varepsilon}_i^2}{\sum (y_i - \bar{y})^2}$$

Em que $R^2 = \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2$ demonstra o quão bem o modelo explica a variabilidade encontrada nos dados. Assim, R^2 também pode ser expresso como:

$$R^2 = 1 - \frac{\sum \hat{\varepsilon}_i^2}{\sum(y_i - \bar{y})^2}$$