PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
Departamento de Engenharia Industrial
Rua Marquês de São Vicente, 225
22453-900 – Rio de Janeiro
Brasil

ENG1536 – Inferência Estatística

Laboratório – Guia de Estudo 9

Neste guia, você aprenderá sobre a estimação de regressões lineares em R. Na metodologia básica de regressão linear, postulamos, com fundamento em teoria e/ou observações, que uma variável de interesse (Y) seja uma função linear de uma ou mais variáveis explicativas (X_1, X_2, etc.). Devido a erros de medição que sempre existem e a variáveis desconhecidas, a relação linear entre Y e as variáveis explicativas nunca é perfeitamente observada, mesmo que ela exista. Daí a inclusão de um termo de erro (u) na equação, o qual é modelado como uma variável normal com média nula e variância constante. Como em outros estudos anteriores, trabalharemos sobre dados simulados:

```
x1 <- rnorm(25,2,5); x2 <- rnorm(25,-1,5); u <- rnorm(25,sd=2)
y <- 100 + 1.5*x1 + u
```

Os comandos acima criam 25 observações de quatro variáveis normalmente distribuídas (qual é o valor esperado e o desvio padrão de cada uma delas?). As variáveis `x1` e `x2` farão o papel das duas variáveis explicativas (também chamadas "variáveis independentes" ou "regressores"), que um cientista elegeu para explicar as variações da variável dependente `y`. A quarta variável `u` fará o papel do erro no modelo explicativo. Nesta simulação, como se vê no último comando acima, existe uma verdadeira relação linear entre o valor esperado de `y` e `x1`, mas não há qualquer relação entre o valor esperado de `y` e `x2`. O cientista não sabe disso.

Numa análise exploratória dos dados, a primeira coisa que normalmente fazemos num estudo de regressão é examinar o **diagrama de dispersão** entre a variável Y e cada um dos regressores. Trata-se apenas de um gráfico cartesiano em que representamos por pontos os pares (X_i,Y). Digite:

```
plot(x1,y)  #Os pontos não estão perfeitamente alinhados, mas sugerem uma reta
plot(x2,y)  #Os pontos formam uma nuvem que não sugere relação linear entre y e x2
```

A aglutinação dos pontos num diagrama de dispersão ao redor de uma reta invisível é medida pelo coeficiente de correlação linear amostral, calculado em R pela função `cor`:

```
cor(x1,y)  #O valor deve ser alto, acima de 0,9
cor(x2,y)  #O valor deve ser baixo, entre -0,25 e +0,25
```

O coeficiente correlação linear não é capaz de revelar a **inclinação** da reta que associa linearmente Y e X. Ele apenas indica pelo seu sinal se a inclinação é positiva ou negativa, bem como a força da associação linear. É só através da regressão linear que podemos estimar os verdadeiros coeficientes de inclinação linear do modelo. Em R, isso é feito através da função `lm` (*linear model*), com a qual o você teve um primeiro contato no guia sobre análise de variância. Como a maioria das funções na linguagem R, `lm` é bastante versátil, mas estudaremos aqui apenas como estimar um modelo simples. Digite:

```
lm(y~x1+x2)
```

Na sintaxe desta função, o primeiro argumento especifica o modelo linear de uma maneira que lembra a equação matemática, mas é mais econômica. Primeiro, escreve-se o nome do objeto que contém as observações da variável dependente (neste caso, `y`). Depois, o caractere "~" seguido de uma soma apenas simbólica dos objetos que contêm as observações das variáveis explicativas. Abaixo ilustramos o resultado do comando para uma simulação específica:

```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
   100.4398       1.5079       0.0624
```

O que se vê são os coeficientes estimados da equação de regressão. O primeiro é o "intercepto" estimado, a constante na equação. O verdadeiro valor na simulação é 100 e o valor estimado foi 100,4398. Os dois valores seguintes são os coeficientes de X_1 e X_2 na equação. O

verdadeiro valor do coeficiente de X_1 é 1,5 (veja o comando que criou `y` na simulação) e o valor estimado foi 1,5079. O verdadeiro valor do coeficiente de X_2 é zero (porque X_2 não entra na equação que simula os valores de Y!) e o valor estimado foi 0,0624. O que R mostra no resultado padrão da função `lm` é apenas um subconjunto de todas as informações produzidas pelo processo de estimação do modelo linear. Para se examinar um conjunto mais completo de informações sobre o modelo estimado, usa-se a função `summary`:

```
reg <- lm(y~x1+x2)
summary(reg)
```

Este é o resultado com os mesmos dados simulados de antes:

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6817 -0.7386  0.1175  1.9446  4.9919

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.43975    0.55733 180.215  < 2e-16 ***
x1            1.50795    0.10412  14.483 9.93e-13 ***
x2            0.06240    0.08647   0.722    0.478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.333 on 22 degrees of freedom
Multiple R-squared:  0.9051,    Adjusted R-squared:  0.8965
F-statistic: 104.9 on 2 and 22 DF,  p-value: 5.635e-12
```

Na coluna da tabela principal intitulada `Estimate` você encontra as mesmas estimativas de antes. As colunas seguintes contêm informações que permitem testar a significância estatística das estimativas. O p-valor (bilateral) encontrado na última coluna é extremamente pequeno para o intercepto estimado e o coeficiente de X_1 estimado. Isso nos leva a rejeitar a hipótese de que os seus verdadeiros valores sejam zero (e sabemos que não são). O p-valor associado ao coeficiente X_2 é muito alto, 0,478. Ele não nos permite rejeitar a hipótese de que o verdadeiro

valor do coeficiente seja zero (e sabemos que é zero). Outras estatísticas relevantes mostradas abaixo da tabela principal são: (i) o R^2 da regressão, igual a 0,9051, indicando grande aderência dos dados ao modelo proposto; e (ii) O p-valor da estatística F associada ao modelo, muito pequeno igual a $5{,}635 \times 10^{-12}$, indicando rejeição segura da hipótese de que o modelo como um todo não explica as variações de Y.

Os teste de hipótese feitos acima dependem de um pressuposto do modelo clássico de regressão linear: os erros são normalmente distribuídos e independentes. No caso da simulação feita aqui, eles de fato são. Em aplicações reais, no entanto, o analista de dados precisa examinar os resíduos da regressão para testar a sua normalidade. Existem vários testes importantes a serem feitos depois da estimação de uma equação de regressão para que possamos tirar conclusões razoavelmente seguras sobre ela. Estudaremos alguns desses testes no próximo guia.

Lista de exercícios 9

(1) Calcule manualmente, usando comandos em R, a estatística t igual a 0,722 associada ao coeficiente de X_2 na tabela completa mostrada no guia, bem como o p-valor associado a ela, igual a 0,478. Faça o mesmo cálculo para a mesma tabela na sua simulação. DICA: o p-valor mostrado na tabela é de um teste bilateral.

(2) Modifique a simulação original do texto para que a variável X_2 entre na função linear que determina os valores de Y, com um coeficiente de inclinação igual a -3. Examine o diagrama de dispersão entre Y e X_2 e estime o coeficiente de correlação linear entre elas. Depois, estime novamente o modelo e examine a linha correspondente a X_2 na tabela dos resultados.

(3) Crie uma nova simulação com apenas um regressor X (25 observações) que é uma variável aleatória normal padrão e uma variável dependente cujos valores são exatamente X^2 (sem erro). Estime o coeficiente de correlação linear entre Y e X e examine o diagrama de dispersão. Depois estude o sumário de um modelo linear estimado entre Y e X. Você descobrirá que o coeficiente estimado de X não é estatisticamente significativo. Você é capaz de explicar isso?

(4) Existe na instalação padrão do R uma moldura de dados chamada `faithful` que contém dois dados sobre 272 erupções registradas do geiser Old Faithful do Parque Yellowstone, nos EUA: a duração da erupção (Y) e o tempo de espera antes do início da erupção (X). Estude o diagrama de dispersão entre X e Y. Parece haver uma relação linear entre as duas variáveis? Você observa algo curioso no diagrama? Estime o coeficiente de correlação linear. Depois, estime um modelo linear para explicar a duração da erupção pelo tempo de espera até o início da erupção e comente a significância estatística do regressor.

* * *