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 10

O modelo clássico de regressão linear faz suposições importantes sobre o erro aleatório na equação: (i) o erro é normalmente distribuído com média nula e variância constante; (ii) os erros são independentes entre si; e (iii) os erros e os regressores são descorrelatados. Quando um software estima uma equação de regressão, muitos dos resultados calculados dependem da validade daqueles dois pressupostos. Por isso, é importante testarmos a validade deles usando os resíduos da regressão. Neste estudo, você aprenderá alguns recursos básicos para isso. Como em outros estudos, usaremos simulações. A nossa estratégia será simular dados que violem algum dos pressupostos clássicos sobre o erro aleatório, para examinar o efeito da violação sobre os resultados da estimação. Usaremos extensamente o conteúdo do guia anterior, mas nos ateremos a modelos com apenas um regressor.

Para começar, simulemos uma regressão com um erro cuja variância é uma função simples dos valores do regressor:

```
x <- runif(1000,1,20); u <- rnorm(1000,sd=x); y <- 15 + 2*x + u
```

Nesta simulação, os valores do regressor são uniformemente distribuídos entre 1 e 20. Os erros são normais com média nula e desvio padrão igual ao valor do regressor. Construa um diagrama de dispersão entre `x` e `y` e observe que a nuvem de pontos sugere a relação linear entre as duas variáveis, mas ela tem uma forma cônica. Isso indica que não apenas o valor médio de `y` aumenta com `x`, mas também a variabilidade `y` aumenta. Esse tipo de fenômeno, chamado "heterocedasticidade", é comum em dados econômicos/financeiros. Em presença de heterocedasticidade, os estimadores dos coeficientes continuam sendo não-enviesados e consistentes, mas o estimador da variância do erro torna-se enviesado. Isso torna inválidas as

estimativas de variância dos coeficientes estimados da equação. Estime um modelo linear tendo `y` como variável dependente e `x` como regressor, com a função `lm`, e use a função `summary` para examinar os resultados da sua regressão. Você observará que os valores estimados do intercepto e do coeficiente de x na equação são muito próximos dos verdadeiros valores (15 e 2). No entanto, os valores nas colunas seguintes não são corretos. O remédio, nesse caso, é utilizar um método de estimação alternativo, e.g. mínimos quadrados generalizados. Depois de estimar qualquer regressão linear, a primeira providência do analista deveria ser examinar graficamente os resíduos da regressão, que são uma estimativa do verdadeiro erro aleatório. Para isso, use a função `residuals` do R. Experimente:

```
uhat <- residuals(lm(y~x))
plot(uhat)   #Superficialmente, não parece haver problemas
hist(uhat)   #O histograma parece sugerir a forma normal esperada
plot(x,uhat) #MAS AQUI HÁ UM PROBLEMA
```

No último comando acima, você observará que o diagrama de dispersão entre o regressor e os resíduos da regressão mostra que existe uma estrutura nos dados que não foi capturada pela equação de regressão. Isso se revela na forma cônica do diagrama, que não deveria existir numa regressão que tenha sucesso em modelar uma relação entre duas variáveis.

Na próxima simulação, manteremos os valores do regressor, mas criaremos uma série de erros normais (com variância agora constante) que não são independentes. Esse fenômeno é muito comum em séries temporais.

```
u <- rnorm(1); for(i in 2:1000) {u[i] <- 0.95*u[i-1]+rnorm(1)}; y <- 15 + 2*x + u
```

Estime novamente a equação de regressão e examine os resultados. O efeito da autocorrelação dos erros é similar ao efeito da heterocedasticidade: as estimativas pontuais continuam não enviesadas, porém você não pode mais confiar nos intervalos de confiança calculados para elas. Estudando os resíduos outra vez:

```
uhat <- residuals(lm(y~x))
plot(uhat)   #Agora aqui se vê a autocorrelação nos resíduos: ondulações no gráfico
hist(uhat)   #O histograma parece sugerir a forma normal esperada
plot(x,uhat) #Nenhum problema visível no diagrama de dispersão
```

Diante de autocorrelação residual, o analista deveria corrigir o seu modelo, por exemplo incluindo um termo defasado de `y` na equação.

Por último, vamos examinar o efeito da violação do pressuposto de descorrelação entre o erro e o regressor (em econometria, essa situação é chamada de "endogeneidade"). Para isso, criamos um erro cujo valor esperado é uma função do regressor:

```
u <- rnorm(1000,mean=x); y <- 15 + 2*x + u
```

Primeiro, confirme que `x` e `u` são (muito) correlacionados, usando a função `cor`. Estime novamente a equação de regressão e estude os resultados. Você constatará que, agora, o estimador do coeficiente de inclinação da reta superestima o verdadeiro valor: o coeficiente estimado será próximo de 3, quando o verdadeiro valor é 2. Além disso, o seu desvio padrão estimado deverá ser na sua simulação menor do que 0,01. Logo, rejeitaríamos a qualquer nível de significância razoável a hipótese nula de que o verdadeiro valor fosse 2 (que é o valor correto). Em suma, o efeito da violação do pressuposto de descorrelação entre erro e regressor é muito mais sério do que nas violações anteriores: o estimador de mínimos quadrados ordinários nesta situação se torna enviesado. O problema é ainda mais sério porque ele não é visível nos resíduos da regressão, como as violações anteriores. Mas existem testes formais para a presença de endogeneidade que são estudados em cursos avançados de econometria (e.g. teste Hausman). E existem soluções também para o problema, tais como a utilização das chamadas "variáveis instrumentais", outro tópico avançado em econometria.

Lista de exercícios 10

(1) Na primeira simulação que você realizou neste guia, existe um problema no histograma dos resíduos que não é visível a olho nu: excesso de curtose. Estude o verbete "Curtose" na Wikipédia e crie uma função para estimar a curtose num vetor de dados. Verifique que a curtose nos resíduos da primeira simulação é maior do que a curtose de uma amostra normal.

(2) A econometria de séries temporais é particularmente difícil na presença de autocorrelação forte. Crie um vetor `x` no qual o primeiro valor é 100 e cada valor subsequente é igual ao anterior mais um sorteio normal padrão. Crie um vetor `y` da mesma forma. Estime uma regressão linear tendo `y` por variável dependente e `x` como regressor. Critique os resultados da regressão.

(3) Examine os resíduos da terceira simulação e confirme que os gráficos não indicam, à primeira vista, nenhuma violação dos pressupostos do modelo clássico. Calcule também a curtose dos resíduos.

* * *