Aula Prática - Inferência Causal Baseada em Regressões

Introdução

Antes de prosseguir com a aula prática sobre modelos de regressão, iremos assistir a dois vídeos. O primeiro vídeo apresenta de modo intuitivo a ideia de *ceteris paribus* em um contexto de regressão. Já o segundo vídeo discute em maior detalhe o problema do viés.

Ceteris Paribus

https://www.youtube.com/watch?v=iPBV3BlV7jk&list=PL-uRhZ_p-BM5ovNRg-G6hDib27OCvcyW8&index=2

Viés de Seleção = Viés de Variável Omitida em Regressão

https://www.youtube.com/watch?v=6YrIDhaUQOE&t=61s

Dale, S., Krueger, A. Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. The Quarterly Journal of Economics, Volume 117, Issue 4, November 2002, Pages 1491–1527.

Descrição da atividade

A ideia para a aula prática de hoje será reproduzir a Tabela 7.1 do livro de Stock e Watson, apresentada abaixo.

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28 (0.52) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for income assistance (X_4)				−0.790 (0.068)	0.048 (0.059)
Intercept	698.9 (10.4)	686.0 (8.7)	700.2 (5.6)	698.0 (6.9)	700.4 (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

Observação: na 1ª edição do livro, é a Tabela 5.2.

1. Fazer download da base de dados e do dicionário de variáveis. Importar dados para o R.

Antes de iniciar a atividade, vamos chamar os pacotes que serão utilizados. Nesta resolução, são utilizados os seguintes pacotes do R.[1]

Agora, vamos importar os dados e salvá-los no objeto chamado `caschool`.

2. Estimar os modelos de regressão apresentados nas colunas 1 a 5 da tabela.

Abaixo estimamos as 5 especificações da tabela utilizando o comando `lm`. Em todos os casos, salvamos o resultado em um objeto para que possamos aplicar outras funções a ele para a obtenção de outros resultados de interesse.

Uma vez que o modelo estimado foi salvo, podemos ver os resultados principais com as opções padrão utilizando a função `summary`. Vejamos o caso da primeira especificação:

```

Call:
lm(formula = testscr ~ str, data = caschool)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.727 -14.251   0.483  12.822  48.540 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
str          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,   Adjusted R-squared:  0.04897 
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Vocês notam alguma diferença em relação à tabela? Os desvios-padrão dos coeficientes são diferentes! Isso porque a tabela já reporta desvios-padrão robustos, ou seja, que já contemplam a possibilidade de termos um modelo heterocedástico. Para ver o resultado com desvio-padrão robusto, utilizaremos a função `coeftest` e o pacote `sandwich` para obter o resultado:[2]

```

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 698.93295   10.36436 67.4362 < 2.2e-16 ***
str          -2.27981    0.51949 -4.3886 1.447e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] Como não quero reportar as mensagens de erro e avisos que o R reproduz após a execução do comando `library`, inseri opções específicas para esse bloco de código. Para mais detalhes, veja a página de referência sobre opções de execução.

[2] Alternativamente, é possível utilizar o pacote `estimatr` para obter o resultado de modo mais direto. Veja a documentação do pacote aqui.

Podemos proceder de forma análoga para obter os resultados das demais especificações. Não faremos isso nesta etapa, pois iremos trabalhar no item seguinte para já montar uma tabela de resultados.

3. Utilizando o pacote `gt` e seu estagiário de IA, prepare a tabela para reportar os resultados. Não se preocupe em reproduzir exatamente a formatação, desde que apresente as informações principais (coeficientes, desvios-padrão robustos). Não há necessidade de mostrar na tabela o intervalo de confiança, como é feito apenas para a variável *student-teacher ratio* no livro.

Abaixo, seguem o código e a Tabela 1 com os resultados gerados a partir dos modelos estimados.[3]

Tabela 1

Tabela. Desempenho em Testes e Razão Aluno-Professor

	Variável dependente: testscr				
	(1)	(2)	(3)	(4)	(5)
Razão aluno-professor (str)	−2.280	−1.101	−0.998	−1.308	−1.014
	(0.519)	(0.433)	(0.270)	(0.339)	(0.269)
% de aprendizes de inglês (el_pct)		−0.650	−0.122	−0.488	−0.130
		(0.031)	(0.033)	(0.030)	(0.036)
% de alunos com refeição gratuita (meal_pct)			−0.547		−0.529
			(0.024)		(0.038)
% de famílias no CalWORKS (calw_pct)				−0.790	−0.048
				(0.068)	(0.059)
Constante	698.933	686.032	700.150	697.999	700.392
	(10.364)	(8.728)	(5.568)	(6.920)	(5.537)
Observações	420	420	420	420	420
R^2	0.051	0.426	0.775	0.629	0.775

Nota: erros-padrão robustos à heterocedasticidade entre parênteses.

4. Utilizando a tabela e interpretando os resultados apresentados, responda às seguintes perguntas:

   Utilizando os resultados da Tabela 1, seguem sugestões de respostas breveds para as perguntas realizadas em cada item. **Para uma análise mais aprofundada de todos os resultados, sugiro fortemente a leitura do capítulo do livro em que os resultados são discutidos detalhadamente.**

   a. Qual especificação do modelo você julga ser a mais apropriada?

      Sugestão de resposta: A especificação (3) poderia ser escolhida como especificação mais apropriada, uma vez que, além da variável de interesse, controla por características dos alunos que poderiam gerar viés de variável omitida caso não fossem incluídas no modelo. As especificações (4) e (5) apresentam formas alternativas de se controlar pelas condições socioeconômicas dos alunos. Note que os resultados não mudam muito entre as especificações (4) e (5).

   b. Utilizando a especificação escolhida, qual o impacto de reduzir o tamanho médio das turmas em 1 aluno? Interprete o valor do coeficiente e também discuta o resultado em termos de inferência estatística.

      Sugestão de resposta: O resultado reportado na coluna (3) indica que, ao reduzir a razão aluno-professor em um aluno por professor, a nota média no teste aumenta aproximadamente 1 ponto, mantendo-se constantes as características dos alunos. Note que o resultado é robusto, ou seja, o coeficiente é relativamente similar nas demais especificações com controles.

[3] Para essa tabela, eu pedi auxílio ao Claude. Veja o prompt inicial no anexo. Depois, revisei o código e fiz alguns ajustes para corrigir erros e detalhes de formatação.

c. Existe relação causal entre tamanho das turmas e notas? Caso sua resposta seja sim, explique qual foi a hipótese de identificação que você utilizou. Se não, diga qual hipótese o modelo não atende e por que ela não seria atendida.

   Sugestão de resposta: Caso responda sim, o aluno deve justificar por que o modelo não sofre de viés de variável omitida, ou seja, que não há nenhuma variável não incluída no modelo que seja correlacionada com a razão aluno-professor e com as notas médias do teste.

   Caso responda não, o aluno deve apresentar outros mecanismos omitidos que estejam correlacionados com a razão aluno-professor e com as notas médias do teste.

d. Imagine que você seja prefeito de uma cidade no Brasil. Seu secretário de educação lhe apresenta os resultados desse estudo utilizando os dados da Califórnia e sugere que vocês tentem reduzir o tamanho médio das turmas no município com o intuito de melhorar o desempenho escolar. O que você acharia dessa proposta? Justifique, apresentando argumentos de validade interna ou externa.

   Sugestão de resposta: Eu pediria ao secretário para me apresentar por quais motivos ele acreditava que os resultados do estudo feito para Califórnia poderiam ser extrapolados para a população do município, dadas as grandes diferenças socioeconômicas, culturais e institucionais entre o estado americano e o nosso país. Caso a explicação não fosse convincente, eu chegaria à conclusão de que o argumento teria um problema de validade externa e, portanto, não se aplicaria à população residente no município.

Anexo - Prompt

Abaixo, segue o prompt enviado ao Claude para elaboração da tabela em formato de publicação.

Prompt:

Estou trabalhando em um arquivo quarto com o header YAML abaixo. O objetivo é montar uma tabela usando o pacote gt do R. Esta tabela deve reportar o resultados das 5 especificações abaixo em formato para publicação em periódico científico. Note que meu arquivo exporta tanto para o formato html quanto para pdf. As cinco especificações também seguem abaixo. Os pacotes a serem utilizados já foram importados no arquivo. Os desvios-padrão a serem reportados devem ser desvios-padrão robusto a heteroscedásticidade. Além disso, não há necessidade de incluir * para indicar significância estatística.

#Especificação 1 espec_1 <- lm(testscr ~ str, data = caschool)

#Especificação 2 espec_2 <- lm(testscr ~ str + el_pct, data = caschool)

#Especificação 3 espec_3 <- lm(testscr ~ str + el_pct + meal_pct, data = caschool)

#Especificação 4 espec_4 <- lm(testscr ~ str + el_pct + calw_pct, data = caschool)

#Especificação 5 espec_5 <- lm(testscr ~ str + el_pct + meal_pct+ calw_pct, data = caschool)

Header YAML:

```yaml
---
title: "Aula Prática - Inferência Causal Baseada em Regressões"
reference-location: margin
format:
  html:
    echo: true
    warning: false
    error: false
  pdf:
    echo: false
    warning: false
    error: false
    babel-lang: pt
---
```