Aula Prática - Dados em Painel

Introdução

Nas duas aulas práticas anteriores, trabalhamos com dados de corte transversal: na primeira, a variação no tratamento vinha de um **experimento aleatório**; na segunda, tentamos eliminar o viés de variável omitida **controlando** por características observáveis dos estudantes. O problema da segunda estratégia é evidente: sempre restam determinantes do resultado que não conseguimos observar e, portanto, não conseguimos incluir na regressão.

Dados em painel oferecem uma saída parcial para esse problema. Quando acompanhamos as **mesmas unidades ao longo do tempo**, podemos eliminar todos os fatores omitidos que são fixos no tempo dentro de cada unidade — mesmo aqueles que não observamos — comparando cada unidade com ela mesma. O objetivo da aula de hoje é implementar essa ideia no R, partindo do modelo mais simples ("antes e depois") até o modelo de efeitos fixos duplos (*two-way fixed effects*).

Leituras de referência

Seções 10.1 a 10.7 de Stock e Watson. Introduction to econometrics. Slides das aulas 7 e 8.

Discussão

Impostos sobre bebidas alcoólicas reduzem mortes no trânsito? Antes de olhar para os dados, pense: por que uma simples regressão da taxa de fatalidade sobre o imposto estadual sobre a cerveja, usando dados de um único ano, provavelmente **não** recuperaria o efeito causal do imposto? Que características dos estados estariam no termo de erro? Quais delas mudam de um ano para o outro e quais são essencialmente fixas?

Descrição da atividade

A ideia para aula prática de hoje será reproduzir a Tabela 10.1 do livro de Stock e Watson reproduzida abaixo.

TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36 (0.05) [0.26, 0.46]	−0.66 (0.29) [−1.23, −0.09]	−0.64 (0.36) [−1.35, 0.07]	−0.45 (0.30) [−1.04, 0.14]	−0.69 (0.35) [−1.38, 0.00]	−0.46 (0.31) [−1.07, 0.15]	−0.93 (0.34) [−1.60, −0.26]
Drinking age 18		0.10		0.03 (0.07) [−0.11, 0.17]	−0.01 (0.08) [−0.17, 0.15]		0.04 (0.10) [−0.16, 0.24]
Drinking age 19				−0.02 (0.05) [−0.12, 0.08]	−0.08 (0.07) [−0.21, 0.06]		−0.07 (0.10) [−0.26, 0.13]
Drinking age 20				0.03 (0.05) [−0.07, 0.13]	−0.10 (0.06) [−0.21, 0.01]		−0.11 (0.13) [−0.36, 0.14]
Drinking age						0.00 (0.02) [−0.05, 0.04]	
Mandatory jail or community service?				0.04 (0.10) [−0.17, 0.25]	0.09 (0.11) [−0.14, 0.31]	0.04 (0.10) [−0.17, 0.25]	0.09 (0.16) [−0.24, 0.42]
Average vehicle miles per driver				0.008 (0.007)	0.017 (0.011)	0.009 (0.007)	0.124 (0.049)
Unemployment rate				−0.063 (0.013)		−0.063 (0.013)	−0.091 (0.021)
Real income per capita (logarithm)				1.82 (0.64)		1.79 (0.64)	1.00 (0.68)
Years	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 & 1988 only
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	yes	yes	yes	yes	yes	yes
***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**							
Time effects = 0			4.22 (0.002)	10.12 (<0.001)	3.48 (0.006)	10.28 (<0.001)	37.49 (<0.001)
Drinking age coefficients = 0				0.35 (0.786)	1.41 (0.253)		0.42 (0.738)
Unemployment rate, income per capita = 0				29.62 (<0.001)		31.96 (<0.001)	25.20 (<0.001)
$\overline{R}^2$	0.091	0.889	0.891	0.926	0.893	0.926	0.899

These regressions were estimated using panel data for 48 U.S. states. Regressions (1) through (6) use data for all years 1982 to 1988, and regression (7) uses data from 1982 and 1988 only. The data set is described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, 95% confidence intervals are given in square brackets under the coefficients, and *p*-values are given in parentheses under the *F*-statistics.

Observação: Na 1ª edição do livro é a tabela 8.2.

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.

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

2. Este painel é balanceado?

Um painel é **balanceado** quando todos os estados possuem o mesmo número de observações ao longo do tempo. Caso contrário, o painel é **não balanceado**, o que pode ocorrer por dados faltantes para algum estado em determinado ano. Como vemos abaixo, há 48 estados na base de dados e, para cada estado, o número mínimo e o número máximo de observações são iguais a 7 (os anos de 1982 a 1988). Como $48 \times 7 = 336$, que é exatamente o número de linhas da base, o painel é **balanceado**.[1]

```
# A tibble: 1 × 3
  n_states obs_min obs_max
     <int>   <int>   <int>
1       48       7       7
```

Atenção: painel balanceado não significa que não há valores faltantes em *nenhuma* variável. Como veremos adiante, as variáveis de punição (`jaild` e `comserd`) têm um valor faltante, o que fará com que algumas especificações sejam estimadas com 335, e não 336, observações.

3. Apresente um conjunto de estatísticas descritivas para fatalidades e imposto sobre bebidas.

Como a base de dados possui um conjunto muito grande de informações, vamos selecionar apenas as variáveis que representam os impostos sobre bebidas e a taxa de fatalidade de trânsito para termos uma melhor visualização dos resultados. Olhando o dicionário de variáveis vemos que as variáveis de interesse possuem os seguintes nomes na base de dados: *beertax* e *mrall*.[2]

Name	select(Fatalities, mrall,...
Number of rows	336
Number of columns	2
_____	
Column type frequency:	
numeric	2
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
mrall	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
beertax	0	1	0.51	0.48	0.04	0.21	0.35	0.65	2.72	

Os resultados acima mostram que a taxa média de fatalidade é de aproximadamente 2 mortes por 10.000 residentes (a variável `mrall` está medida por residente, daí a média de 0,0002) e que o imposto médio sobre a cerveja é de US$ 0,51 por caixa. Note que há bastante variabilidade nos dados, principalmente em relação ao imposto sobre a bebida: o desvio-padrão (0,48) é quase do mesmo tamanho da média.

Como vemos acima, a variável `mrall` está registrada na base de dados como a taxa de mortes em acidente de trânsito por residente. Assim, para facilitar a replicação dos resultados abaixo, vamos criar uma variável chamada `fatality_rate` que será medida como taxa de mortes por 10.000 residentes.

4. Modelo em Diferenças T=2 ou "Antes e Depois"

[1] Leia a documentação da função count para entender como foi feita a análise. Seria possível também fazer essa conferência olhando quantos anos há na base de dados e qual seria o número máximo e mínimo de estados por ano (tente após ler a documentação!).

[2] Consulte a documentação do pacote skim para entender detalhes sobre como o pacote funciona.

Como vimos na aula expositiva, uma forma de considerar o efeito fixo de unidade no modelo é fazer uma regressão em diferença usando T = 2, ou seja, fazendo uma comparação "antes e depois". Isto permite controlar por todos os fatores fixos no tempo, mas que variam entre os estados.

Antes de fazer a regressão de MQO, precisamos salvar em um objeto uma base de dados que só contenha informações do ano inicial e do ano final. Para isso, usamos a função `filter` do pacote `dplyr` que faz parte do pacote `tidyverse`. Abaixo, apresentamos duas formas equivalentes de fazer esse filtro condicional. Para maiores detalhes, ver a documentação dos pacotes.

Agora vamos criar as variáveis em primeira diferença. Inicialmente, vamos trabalhar apenas com a taxa de fatalidade e a taxa de imposto sobre a cerveja.

Agora vamos estimar o modelo de MQO em primeira diferença:

```
Call:
lm(formula = d_fatality_rate ~ d_beer_tax, data = Fatalities_1982_1988)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22715 -0.09619  0.09212  0.22290  0.67745

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07204    0.06064  -1.188   0.2410
d_beer_tax  -1.04097    0.41723  -2.495   0.0162 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.394 on 46 degrees of freedom
  (48 observations deleted due to missingness)
Multiple R-squared:  0.1192,    Adjusted R-squared:    0.1
F-statistic: 6.225 on 1 and 46 DF,  p-value: 0.01625
```

O coeficiente estimado é −1,04: um aumento de US$ 1,00 no imposto por caixa de cerveja está associado a uma redução de 1,04 morte por 10.000 residentes. Note que 48 observações foram descartadas (`48 observations deleted due to missingness`): são as observações de 1982, para as quais não existe uma defasagem dentro do estado.

Questões:

A. É necessário controlar por alguma variável constante no tempo para evitar viés de seleção no modelo acima? Se sim, por que? Se não, por que?

Sugestão de resposta: Não — e, mais do que desnecessário, seria impossível. Escrevendo o modelo como $FatalityRate_{it} = \beta_1 BeerTax_{it} + \alpha_i + u_{it}$, em que α_i captura *tudo* o que é fixo no tempo em cada estado (cultura em relação ao álcool, densidade populacional, geografia e clima, traçado das estradas, composição religiosa etc.), ao tomar a diferença entre 1988 e 1982 obtemos

$$\Delta FatalityRate_i = \beta_1 \Delta BeerTax_i + \Delta u_i,$$

já que $\alpha_i - \alpha_i = 0$. Ou seja, a própria diferenciação elimina todos esses fatores, observados ou **não observados**. Se tentássemos incluir uma variável constante no tempo na regressão em diferenças, sua

diferença seria igual a zero para todos os estados e o coeficiente não seria identificado (multicolinearidade perfeita).

O ponto importante é que isso **não** resolve tudo: o modelo continua vulnerável a fatores que **variam no tempo dentro do estado** e que estejam correlacionados com a variação do imposto — por exemplo, o ciclo econômico entre 1982 e 1988, mudanças nas leis de trânsito ou campanhas de conscientização. A hipótese de identificação aqui é exatamente que Δu_i não seja correlacionado com $\Delta BeerTax_i$.

B. Podemos fazer inferência estatística a partir dos resultados reportados acima? Se sim, por que? Se não, por que?

Sugestão de resposta: Podemos, mas não com os erros-padrão reportados pelo `summary`. Com $T = 2$, após a diferenciação resta **uma única observação por estado**, de modo que a correlação serial dentro do estado deixa de ser um problema (não há o que correlacionar) — e é por isso que, neste caso específico, não precisamos de erros-padrão clusterizados. O problema que permanece é a **heteroscedasticidade**: estados muito diferentes em tamanho e população dificilmente terão a mesma variância do erro. Assim, o correto é reportar erros-padrão robustos:

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072037   0.065355 -1.1022 0.276091
d_beer_tax  -1.040973   0.355006 -2.9323 0.005229 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note que o erro-padrão robusto (0,355) é **menor** que o convencional (0,417) neste caso, de modo que a conclusão não muda: o coeficiente é estatisticamente significativo a 5%. Mas essa é uma coincidência da amostra — o mais comum é que a correção aumente o erro-padrão.

5. Modelo com efeito fixo de unidade

O pacote atualmente mais utilizado para dados em painel com efeitos fixos é o pacote fixest. A função que iremos utilizar para estimar os modelos é a função `feols` (*Fixed-effects OLS Estimation*). Antes de prosseguir, veja a documentação do pacote para entender como montar a especificação.

```
OLS estimation, Dep. Var.: fatality_rate
Observations: 336
Fixed-effects: state: 48
Standard-errors: Clustered (state)
         Estimate Std. Error  t value Pr(>|t|)
beertax -0.655874   0.291856 -2.24725 0.029358 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.17547     Adj. R2: 0.889129
                Within R2: 0.040745
```

Questões

A. Acima estimamos o modelo utilizando efeito fixo com a função `feols`. Como poderíamos estimar o mesmo modelo mas utilizando a função `lm`? Há dois caminhos. Quais são? Caso não consiga escrever o código, apenas descreva quais especificações utilizaria.

Sugestão de resposta: Os dois caminhos são (i) incluir **dummies de estado** na regressão e (ii) aplicar a transformação *within* (subtrair de cada variável a sua média dentro do estado) e rodar o MQO nas variáveis centradas. Os dois produzem exatamente o mesmo coeficiente de `beertax` que o `feols`:

```
   beertax
-0.6558736
```

```
beertax_dm
-0.6558736
```

Duas observações práticas. Primeiro, o caminho 1 fica inviável quando o número de unidades é muito grande (imagine um painel de milhões de CPFs): o `feols` resolve isso porque "absorve" os efeitos fixos sem nunca construir as dummies. Segundo, no caminho 2 os graus de liberdade reportados pelo `lm` estão **errados**, pois o `lm` não sabe que gastamos n parâmetros ao estimar as médias por estado — por isso os erros-padrão saem subestimados. Em nenhum dos dois casos o `lm` entrega erros-padrão clusterizados automaticamente.

B. Por que utilizamos a opção `vcov` na estimação acima com `feols`? Se não utilizarmos essa opção, qual problema teremos em nossos resultados?

Sugestão de resposta: Porque as observações de um mesmo estado **não** são independentes entre si. Os fatores omitidos que compõem u_{it} (condições das estradas, fiscalização, hábitos locais) mudam lentamente, de modo que u_{it} é **serialmente correlacionado**: $\text{corr}(u_{it}, u_{i,t+1}) \neq 0$. Os erros-padrão usuais de MQO — tanto os que supõem homocedasticidade quanto os robustos à heterocedasticidade — assumem ausência de correlação serial e, por isso, tipicamente **subestimam** a incerteza amostral. Na prática, isso gera estatísticas t infladas, intervalos de confiança estreitos demais e rejeição da hipótese nula com frequência maior do que o nível de significância escolhido. Compare:

```
beertax
0.18785
```

```
  beertax
0.2032797
```

```
  beertax
0.2918556
```

O erro-padrão clusterizado (0,29) é mais de 50% maior que o convencional (0,19). Usar o erro-padrão errado aqui não mudaria a conclusão de significância, mas mudaria bastante o intervalo de confiança reportado.

C. A partir do que vimos sobre a função `feols`, você consegue estimar um modelo que possui somente efeito fixo de tempo?

Sugestão de resposta: Basta trocar o que aparece depois da barra vertical `|`, que é onde declaramos os efeitos fixos a serem absorvidos:

```
OLS estimation, Dep. Var.: fatality_rate
Observations: 336
```

```
Fixed-effects: year: 7
Standard-errors: Clustered (year)
        Estimate Std. Error t value   Pr(>|t|)
beertax 0.366336   0.048247 7.59295 0.00027154 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.540533     Adj. R2: 0.079412
                 Within R2: 0.094538
```

Note o resultado: sem o efeito fixo de estado, o coeficiente do imposto volta a ser **positivo** (0,37). Controlar apenas pelo que é comum a todos os estados em cada ano não resolve o viés — os estados com impostos mais altos são sistematicamente diferentes dos demais.

D. E efeito fixo duplo (*two-way fixed effect*)?

Sugestão de resposta: Declaramos os dois efeitos fixos separados por `+`:

```
OLS estimation, Dep. Var.: fatality_rate
Observations: 336
Fixed-effects: state: 48,  year: 7
Standard-errors: Clustered (state)
        Estimate Std. Error  t value Pr(>|t|)
beertax -0.63998   0.357078 -1.79227 0.079528 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.171819     Adj. R2: 0.891425
                 Within R2: 0.036065
```

Esta é a especificação da coluna (3) da Tabela 10.1: o efeito fixo de estado elimina tudo o que é fixo no tempo dentro de cada estado e o efeito fixo de tempo elimina os choques comuns a todos os estados em um mesmo ano (como a recessão do início dos anos 1980 ou mudanças na legislação federal).

6. Reproduzir a tabela 10.1 do livro utilizando os pacotes `fixest` e `gt`.

Para reproduzir os demais resultados da tabela (especificações 4 a 7), precisamos primeiro preparar as variáveis de controle. Assim, precisamos primeiro criar as variáveis dummies da idade mínima para beber e punição para que possamos estimar os modelos com esses controles. Além disso, a variável milhas média por motorista é medida em milhares e a renda real per capita entra no modelo em logaritmo.

Agora podemos estimar as sete especificações da tabela. Note que a única diferença entre as colunas está (i) nos controles incluídos, (ii) nos efeitos fixos absorvidos e (iii) na amostra utilizada.

A tabela do livro também reporta o teste F de exclusão conjunta dos **efeitos de tempo**. Como o `feols` *absorve* os efeitos fixos declarados depois da barra `|`, eles não aparecem entre os coeficientes e não podem ser testados diretamente. A solução é estimar, apenas para esse teste, versões auxiliares em que o ano entra como conjunto de dummies (`i(year)`), mantendo somente o estado como efeito fixo absorvido.

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

[3] Para essa tabela, eu pedi auxílio ao Claude. Veja o prompt inicial no anexo. Depois, revisei o código e fiz alguns ajustes de formatação e de conteúdo, em especial na construção dos testes F.

Tabela 1: Efeito das leis sobre embriaguez ao volante nas mortes de trânsito

	Variável dependente: taxa de fatalidade (mortes por 10.000)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Imposto sobre a cerveja	0.365	−0.656	−0.640	−0.445	−0.690	−0.456	−0.926
	(0.053)	(0.292)	(0.357)	(0.297)	(0.352)	(0.307)	(0.343)
	[0.26, 0.47]	[−1.24, −0.07]	[−1.36, 0.08]	[−1.04, 0.15]	[−1.40, 0.02]	[−1.07, 0.16]	[−1.62, −0.24]
Idade mínima 18 anos				0.028	−0.010		0.037
				(0.070)	(0.083)		(0.102)
Idade mínima 19 anos				−0.018	−0.076		−0.065
				(0.050)	(0.068)		(0.099)
Idade mínima 20 anos				0.032	−0.100		−0.113
				(0.051)	(0.056)		(0.125)
Idade mínima (em anos)						−0.002	
						(0.022)	
Prisão ou serviço comunitário obrigatório				0.038	0.085	0.039	0.089
				(0.103)	(0.112)	(0.103)	(0.164)
Milhas médias por motorista (mil)				0.008	0.017	0.009	0.124
				(0.007)	(0.011)	(0.007)	(0.049)
Taxa de desemprego				−0.063		−0.063	−0.091
				(0.013)		(0.013)	(0.021)
Renda real per capita (log)				1.816		1.786	0.996
				(0.636)		(0.643)	(0.678)
Anos	1982–88	1982–88	1982–88	1982–88	1982–88	1982–88	1982 e 1988
Efeitos fixos de estado	não	sim	sim	sim	sim	sim	sim
Efeitos fixos de tempo	não	não	sim	sim	sim	sim	sim
Erros-padrão clusterizados	não	sim	sim	sim	sim	sim	sim
Efeitos de tempo = 0			4.22	10.12	3.48	10.28	37.43
			(<0.001)	(<0.001)	(0.002)	(<0.001)	(<0.001)
Coeficientes de idade mínima = 0				0.35	1.41	0.01	0.42
				(0.786)	(0.241)	(0.920)	(0.739)
Taxa de desemprego, renda per capita = 0				29.62		31.96	25.16
				(<0.001)		(<0.001)	(<0.001)
Observações	336	336	336	335	335	335	94
R^2 ajustado	0.091	0.889	0.891	0.926	0.893	0.926	0.900

Notas: erros-padrão entre parênteses e intervalos de confiança de 95% entre colchetes. Os erros-padrão são clusterizados por estado em todas as especificações, exceto na (1), em que são apenas robustos à heterocedasticidade. Os p-valores dos testes F aparecem entre parênteses abaixo das estatísticas. As especificações (1) a (6) usam todos os anos de 1982 a 1988; a especificação (7) usa apenas 1982 e 1988.

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

   Utilizando os resultados da Tabela 1, seguem sugestões de respostas breves 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. Por que o coeficiente do imposto sobre a cerveja muda de sinal entre as colunas (1) e (2)?

      Sugestão de resposta: Porque a coluna (1) compara estados **entre si**, enquanto a coluna (2) compara cada estado **com ele mesmo** ao longo do tempo. Na coluna (1), o coeficiente positivo (0,36) reflete viés de variável omitida: estados que historicamente cobram impostos mais altos sobre a cerveja são também estados com características fixas associadas a mais mortes no trânsito (por exemplo, mais quilômetros rodados por habitante, estradas rurais, menor densidade). Ao incluir o efeito fixo de estado, toda essa variação entre estados é descartada e resta apenas a

variação temporal dentro de cada estado — e aí o sinal se inverte para −0,66. A mudança de sinal é, em si, a melhor evidência de que a regressão da coluna (1) estava viesada.

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

Sugestão de resposta: A especificação (4) é uma boa escolha como especificação principal. Ela inclui os dois efeitos fixos (controlando tanto pelo que é fixo em cada estado quanto pelos choques comuns a todos os estados em cada ano) e ainda controla pelas demais políticas de trânsito (idade mínima para beber, punição obrigatória) e pelas condições econômicas (desemprego e renda per capita), que variam no tempo dentro do estado e afetam tanto a quantidade dirigida quanto a arrecadação de impostos. As colunas (5) e (6) servem como checagens de robustez: (5) mostra o que acontece ao omitir os controles econômicos e (6) mostra que medir a idade mínima de forma contínua não altera as conclusões. Note que os testes F indicam que as variáveis de idade mínima **não** são conjuntamente significativas (F = 0,35, p = 0,786), enquanto desemprego e renda **são** (F = 29,62, p < 0,001) — ou seja, os controles econômicos são os que realmente importam.

c. Utilizando a especificação escolhida, qual o impacto de aumentar o imposto sobre a cerveja em US$ 1,00 por caixa? Interprete o valor do coeficiente e também discuta o resultado em termos de inferência estatística.

Sugestão de resposta: O coeficiente de −0,45 na coluna (4) indica que um aumento de US$ 1,00 no imposto por caixa de cerveja está associado a uma redução de 0,45 morte por 10.000 residentes, mantidos constantes os demais controles e os efeitos fixos de estado e de tempo. Como a média da taxa de fatalidade é de cerca de 2 mortes por 10.000, isso corresponde a uma queda de aproximadamente 22% — um efeito grande em termos práticos. Estatisticamente, porém, o resultado é frágil: o erro-padrão clusterizado é 0,30 e o intervalo de confiança de 95% ([−1,04; 0,15]) **inclui o zero**. Ou seja, não podemos rejeitar a hipótese de que o efeito é nulo ao nível de 5%. Vale notar como a precisão varia entre especificações: apenas nas colunas (2) e (7) o intervalo exclui o zero.

d. Existe relação causal entre imposto sobre a cerveja e mortes no trânsito? 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: Para interpretar $\hat{\beta}_1$ como efeito causal, precisamos da hipótese $E(u_{it} \mid BeerTax_{i1}, ..., BeerTax_{iT}, \alpha_i, \lambda_t) = 0$. Isso exige, entre outras coisas: (i) que não haja determinantes omitidos das mortes no trânsito que **variem dentro do estado ao longo do tempo** e sejam correlacionados com mudanças no imposto (por exemplo, o estado aprovar simultaneamente o aumento do imposto e uma campanha de fiscalização); (ii) que não haja **causalidade reversa**, isto é, que um ano com muitas mortes no trânsito não leve o estado a aumentar o imposto no ano seguinte; e (iii) que não haja **efeitos defasados** omitidos — se o imposto de hoje afeta as mortes de amanhã, o modelo estático subestima o efeito total e seria preciso um modelo TWFE dinâmico, com defasagens do imposto.

Uma resposta "sim" precisa argumentar por que essas condições são plausíveis no contexto; uma resposta "não" precisa apontar um mecanismo concreto que as viole. O ponto honesto é que o item (ii) é preocupante aqui: impostos sobre bebidas costumam ser alterados justamente em resposta à percepção pública sobre acidentes.

e. Imagine que você seja prefeito de uma cidade no Brasil. Seu secretário de saúde lhe apresenta esses resultados e propõe aumentar o imposto sobre a cerveja no município para reduzir mortes no trânsito. O que você acharia dessa proposta? Justifique, apresentando argumentos de validade interna ou externa.

Sugestão de resposta: Há objeções nos dois planos. Em **validade interna**, os intervalos de confiança das especificações mais completas incluem o zero: o estudo não estabelece com segurança que o efeito é diferente de zero, então basear uma política nele é apostar em um resultado estatisticamente frágil. Em **validade externa**, os dados são de estados americanos entre 1982 e 1988, com padrões de motorização, malha viária, fiscalização e consumo de álcool muito distintos dos brasileiros de hoje. Além disso, há um problema de escala: a política estudada é estadual, enquanto a proposta é municipal — em um município, o consumidor pode simplesmente comprar cerveja na cidade vizinha, o que enfraquece o efeito do imposto sobre o consumo local (o mesmo motivo pelo qual o efeito estimado para um estado não se traduz diretamente para uma cidade). Eu pediria ao secretário evidência sobre o efeito de impostos sobre álcool no contexto brasileiro e uma avaliação de instrumentos alternativos (fiscalização de alcoolemia, por exemplo) antes de decidir.

Anexo - Prompt

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

Prompt:

Estou trabalhando em um arquivo quarto que exporta tanto para html quanto para pdf. O objetivo é reproduzir a Tabela 10.1 do livro de Stock e Watson usando o pacote `gt` do R, em formato para publicação em periódico científico (sem cores, sem asteriscos de significância).

Os sete modelos já foram estimados com a função `feols` do pacote `fixest` e estão salvos nos objetos `espec_1` a `espec_7`. A variável dependente é `fatality_rate` (mortes no trânsito por 10.000 habitantes) e a variável de interesse é `beertax` (imposto sobre uma caixa de cerveja). Os erros-padrão são clusterizados por estado, exceto na especificação (1).

A tabela deve ter uma coluna por especificação e, no corpo, as seguintes linhas para cada variável explicativa: a estimativa, o erro-padrão entre parênteses na linha de baixo e, apenas para `beertax`, o intervalo de confiança de 95% entre colchetes em uma terceira linha.

Abaixo do bloco de coeficientes, incluir linhas informando: os anos utilizados na estimação, se há efeito fixo de estado, se há efeito fixo de tempo e se os erros-padrão são clusterizados.

Em seguida, um bloco com testes F de exclusão conjunta, reportando a estatística F e, na linha abaixo, o p-valor entre parênteses, para três grupos: (i) efeitos de tempo, (ii) coeficientes de idade mínima para beber, (iii) taxa de desemprego e renda per capita. Atenção: como o `feols` absorve os efeitos fixos declarados depois da barra vertical, eles não aparecem entre os coeficientes — me explique como contornar isso para conseguir testar a exclusão dos efeitos de tempo.

Por fim, duas linhas com o número de observações e o R^2 ajustado.

Os rótulos das variáveis devem estar em português:

beertax = Imposto sobre a cerveja d_18 = Idade mínima 18 anos d_19 = Idade mínima 19 anos d_20 = Idade mínima 20 anos mlda = Idade mínima (em anos) jail_com = Prisão ou serviço comunitário obrigatório vmilespd = Milhas médias por motorista (mil) unrate = Taxa de desemprego log_perinc = Renda real per capita (log)

Recursos adicionais para aprofundar e praticar (opcional)

- Documentação do pacote `fixest` — em especial a vinheta Fast fixed-effects estimation.
- R para Ciência de Dados (2ª edição)
- Introduction to Creating gt Tables

Observação: o pacote `fixest` também possui a função `etable`, que gera tabelas de regressão prontas com pouquíssimo código. Vale a pena compará-la com o resultado construído "na mão" com o `gt`: a primeira é muito mais rápida, a segunda dá controle total sobre o que aparece na tabela.