Aplicações do Teorema de Frisch-Waugh-Lovell

Apoio aos alunos de Econometria e capacitação de futuros docentes

Gustavo de Oliveira Vital

Universidade Federal Fluminense - Faculdade de Economia

18/10/2018

Sumário

Apresentação do Teorema

Dado um modelo Linear, estimado por Mínimos Quadrados Ordinários (MQO):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

Em que y é a variável dependente, sendo representada por um vetor $\mathbf{n} \times \mathbf{1}$; X é a matriz de variáveis explicativas, tal que assumimos que essa seja $\mathbf{n} \times \mathbf{k}$; β é o vetor de coeficientes da regressão, sendo esse um vetor $\mathbf{k} \times \mathbf{1}$ e $\mathbf{u}$ o vetor do termo de erro, $\mathbf{n} \times \mathbf{1}$.
Podemos transformar esse modelo tal que esse seja representado da seguinte forma:

$$\mathbf{y} = \mathbf{X_1}\beta_\mathbf{1} + \mathbf{X_2}\beta_\mathbf{2} + \mathbf{u}$$

Apresentação do Teorema

A proposta de FWL é mostrar que os estimadores de qualquer variável no modelo múltiplo podem ser obtidos se “neutralizarmos ” os efeitos das demais variáveis. Dito de outra forma, se dividirmos a matriz de regressores em 2 grupos de variáveis, basta eliminar os efeitos de um deles para obter o estimador do grupo de interesse do pesquisador.

Apresentação do Teorema

O teorema de FWL apresenta dois resultados:

- As estimativas de mínimos quadrados das regressões (1) e (2) abaixo descritas são numericamente idênticas;
- Os resíduos de mínimos quadrados das regressões (1) e (2) abaixo descritas são numericamente idênticos.

$$\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{u} \tag{1}$$

$$\mathbf{M_1 y} = \mathbf{M_1 X_2}\beta_2 + \textbf{ resíduos} \tag{2}$$

em que $\mathbf{M_1}$ é a matriz que projeta no $Span^{\perp}(\mathbf{X_1})$
No modelo acima, particionamos $\mathbf{X} = \begin{bmatrix} \mathbf{X_1} & \mathbf{X_2} \end{bmatrix}$. Além disso, β_1 terá k_1 componentes e β_2 terá k_2 componentes, em que $k = k_1 + k_2$.

Apresentação do Teorema

Nota 1

A matriz $\mathbf{M_1}$ é dita aniquiladora de $\mathbf{X_1}$ pois projeta ortogonalmente as colunas de $\mathbf{X_1}$ no $Span^{\perp}(\mathbf{X_1})$. Como $\mathbf{X_1}$ é parte de $\mathbf{X}$ (lembre que particionamos $\mathbf{X} = \begin{bmatrix}\mathbf{X_1} & \mathbf{X_2}\end{bmatrix}$), podemos escrever que:

$$\mathbf{M_1 X_1} = (\mathbf{I} - \mathbf{P_1})\mathbf{X_1} = \mathbf{X_1} - \mathbf{P_{X_1} X_1} = \mathbf{X_1} - \mathbf{X_1} = \mathbf{0}$$

A penúltima igualdade decorre do fato de $\mathbf{X_1}$ ser invariante sob a ação de $\mathbf{P_1}$, a matriz de projeção no $Span(\mathbf{X_1})$.

Nota 2

Veja que não incluímos nenhuma caracterização mais objetiva dos resíduos. Isso decorre fundamentalmente de razões geométricas.

Prova do Teorema

Pela fórmula geral do estimador de Mínimos Quadrados Ordinários, temos que:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

O estimador de (2) é:

$$(\mathbf{M}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \tag{3}$$

Sejam β_1 e β_2 os dois vetores de coeficientes estimados de (1). Então:

$$\mathbf{y} = \mathbf{P}_x\mathbf{y} + \mathbf{M}_x\mathbf{y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_x\mathbf{y} \tag{4}$$

Premultiplicando o lado esquerdo e o direito da expressão (4) por $\mathbf{X}_2'\mathbf{M}_1$ nós obtemos:

Prova do Teorema

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{y} = \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 \tag{5}$$

O primeiro termo do lado direito de (4) some, pois $\mathbf{M_1}$ aniquila $\mathbf{X_1}$. Podemos observar que o último termo também some:

$$\mathbf{M_X M_1 M_2} = \mathbf{M_X X_2} = \mathbf{0}$$

Agora nós podemos solucionar (5) em relação à $\hat{\beta}_\mathbf{2}$:

$$\hat{\beta}_\mathbf{2} = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X_X})^{-\mathbf{1}}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}$$

Isso é: a expressão (3). Isso prova *parte* do teorema.

Prova do Teorema

Se premultiplicarmos (4) por $\mathbf{M_1}$ ao invés de $\mathbf{X_2'M_1}$ nés iremos obter:

$$\mathbf{M_1 y} = \mathbf{M_1 X_2} \hat{\beta}_{\mathbf{2}} + \mathbf{M_X y} \tag{6}$$

Onde o último termo permanece inalterado, pois $\mathbf{M_1 M_X} = \mathbf{M_X}$. O regressando de (6) é o regressando da regressão (2). Como $\hat{\beta}$ é uma estimação de β (1), pela primeira parte do teorema, o primeiro termo do lado direito de (6) é o vetor de valores estimados da regressão. Ainda, o segundo termo tem que ser o vetor de resíduos da regressão (1). Como $\mathbf{M_X y}$ é , *também*, o vetor de resíduos da regressão (1), isso prova a segunda parte do teorema.

Um Exemplo Prático

Por fim, nesta última sessão do trabalho, o objetivo será apresentar um modelo teórico que tem como objetivo estimar a variação salarial no estado do Rio de Janeiro. A base de dados últilizada sera a PNAD (Pesquisa Nacional por Amostra em Domicílios) de pessoas (PNADPes) do ano de 2015. Ainda que esse não seja o banco de dados atual - a própria PNAD foi descontinuada e hoje é implementada a PNADC (Pesquisa Nacional por Amostra em Domicílios Contínua) - servirá como um exemplo prático da implementação deste teorema. Foi pensado um modelo que leve em consideração o Sexo da pessoa, a Idade e Anos de Estudo.

Um Exemplo Prático

O modelo assumirá a seguinte forma:

$$ln(salario) = \beta_0 + idade\beta_1 + sexo\beta_2 + educ\beta_3 + u$$

Ressalto, ainda, que o objetivo não é melhorar o modelo, seja em relação à acrescimos de variáveis ou mesmo em relação a correção de heterocedasticidade. O presente exemplo visa **somente** demonstrar uma apliação referente ao teorema de FWL.
Os resultados das estimações foram obtidos a partir da linguagem *Python*, especificamente utilizando o modulo *statsmodels*, e a leitura dos dados com o software *R*.

As estimações

Dado o modelo apresentado :

$$ln(salario) = \beta_0 + idade\beta_1 + sexo\beta_2 + educ\beta_3 + u$$

Estimamos ele em sua forma completa após importarmos os modulos necessários:

```
import pandas as pd
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import rc

pnad = pd.read_csv('dadosRJ.csv')
pnad = pnad[['V8005', 'V9532', 'V4803', 'V0302']]
pnad = pnad.rename({'V8005': 'idade', 'V9532': '
    salmes', 'V4803': 'educ', 'V0302': 'sexo'},
    axis='columns')
```

As estimações

Após termos importado os modulos, renomeamos as variáveis e criamos um novo *dataframe* (com o mesmo nome), usando somente as variáveis necessárias.

De acordo com o dicionario da PNAD de 2015, salário (*proxy*) é representado pelo código: 'V9532'; sexo pelo código: 'V0302'; idade pelo código: 'V8005'; e escolaridade(*proxy*) pelo código: 'V4803'. Retiramos os *missings* das variáveis e o valor '17' do código referente à educação, por ser não determinado. Além disso, aplicaremos o *log* na variável salmes, para trabalharmos com a variação. Ainda, por sexo ser uma variável *dummy*, assumiremos mulher $= 1$; homem $= 0$.

As estimações

```
lnsalmes = np.log(pnad['salmes'])
pnad['lnsalmes'] = lnsalmes  # len = 25858
pnad = pnad[(pnad.educ < 17)]  # len = 25829
pnad = pnad.dropna()  # len = 11471
pnad['sexo'] = pnad['sexo'].map({4: 1, 2: 0})
```

Como, por *defaut* o modulo *statsmodels* não inclui constante no modelo, temos que adicionar uma coluna de "uns":

```
pnad = sm.add_constant(pnad)
```

Feito isso, só nos resta estimar o modelo: coluna de "uns":

```
y = pnad['lnsalmes']
X = pnad[['const', 'idade', 'sexo', 'educ']]

modelo = sm.OLS(y, X)
ajuste = modelo.fit()

print(ajuste.summary().as_latex())
```

As estimações

Após estimarmos o modelo, sob a forma $ln(salario) = \beta_0 + idade\beta_1 + sexo\beta_2 + educ\beta_3 + u$, obtivemos o seguinte resultado:

Dep. Variable:	lnsalmes	**R-squared:**	0.062
Model:	OLS	**Adj. R-squared:**	0.061
Method:	Least Squares	**F-statistic:**	250.9
Date:	Sat, 06 Oct 2018	**Prob (F-statistic):**	9.75e-158
Time:	02:10:11	**Log-Likelihood:**	-28027.
No. Observations:	11476	**AIC:**	5.606e+04
Df Residuals:	11472	**BIC:**	5.609e+04
Df Model:	3		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	4.8823	0.123	39.660	0.000	4.641	5.124
idade	0.0251	0.002	12.476	0.000	0.021	0.029
sexo	-0.3950	0.053	-7.494	0.000	-0.498	-0.292
educ	0.1755	0.007	25.793	0.000	0.162	0.189

Omnibus:	13384.632	**Durbin-Watson:**	1.369
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	936746.793
Skew:	6.430	**Prob(JB):**	0.00
Kurtosis:	45.351	**Cond. No.**	209.

As estimações

A partir de agora, faremos o mesmo procedimento regredindo as variáveis explicativas e a variável dependente contra *idade*, e extraindo de cada uma os resíduos das regressões:

```
M1y = sm.OLS(pnad['lnsalmes'], pnad[['const', '
    idade']]).fit().resid
M1X2 = sm.OLS(pnad['educ'], pnad[['const', 'idade
    ']]).fit().resid
M1X3 = sm.OLS(pnad['sexo'], pnad[['const', 'idade
    ']]).fit().resid
```

Adiciono os resultados no *dataframe* pnad e faço outra estimação - dessa vez referente ao modelo com as matrizes aniquiladoras de *idade*:

```
pnad['M1y'] = M1y
pnad['M1X2'] = M1X2
pnad['M1X3'] = M1X3
modelo2 = sm.OLS(pnad['M1y'], pnad[['const', '
    M1X2', 'M1X3']])
ajuste2 = modelo2.fit()
```

As estimações

Aplicando o seguinte comando iremos obteremos o *output*.

```
print(ajuste2.summary())
```

Dep. Variable:	M1y	**R-squared:**	0.057
Model:	OLS	**Adj. R-squared:**	0.056
Method:	Least Squares	**F-statistic:**	343.8
Date:	Sat, 06 Oct 2018	**Prob (F-statistic):**	1.01e-145
Time:	02:00:10	**Log-Likelihood:**	-28027.
No. Observations:	11476	**AIC:**	5.606e+04
Df Residuals:	11473	**BIC:**	5.608e+04
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	5.496e-15	0.026	2.12e-13	1.000	-0.051	0.051
M1X2	0.1755	0.007	25.794	0.000	0.162	0.189
M1X3	-0.3950	0.053	-7.494	0.000	-0.498	-0.292

Omnibus:	13384.632	**Durbin-Watson:**	1.369
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	936746.793
Skew:	6.430	**Prob(JB):**	0.00
Kurtosis:	45.351	**Cond. No.**	7.80

De fato, isso nos leva a crer que $\hat{\beta}_2$ e $\hat{\beta}_3$ são iguais em âmbos os modelos.

Conclusão

De fato, podemos concluir que os coeficientes são iguais, sendo essa uma maneira alternativa de estimação de MQO. Além disso, podemos chegar a uma outra conclusão. Os resíduos dos dois modelos são os mesmos. Como "*idade*" está incorporada ao segundo modelo por meio da matriz aniquiladora essa variável ainda se apresenta por meio dessa matriz. Com mais algumas linhas de códigos podemos conferir ao gráfico de dispersão dos resíduos dos dois modelos. Assim, não só os coeficientes $\hat{\beta}_2$ e $\hat{\beta}_3$ são iguais, mas os resíduos $\hat{u}_1$ e $\hat{u}_2$ também - sendo o primeiro referente ao modelo com *idade* e o segundo ao modelo com a matriz aniquiladora de *idade*.

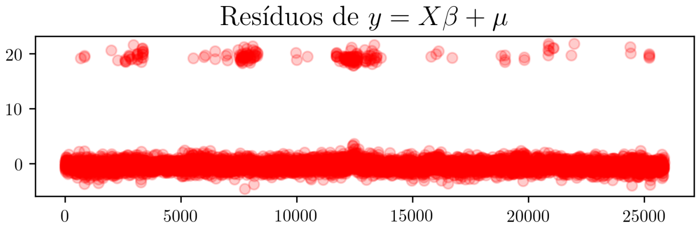
Visualizando o gráfico de resíduos

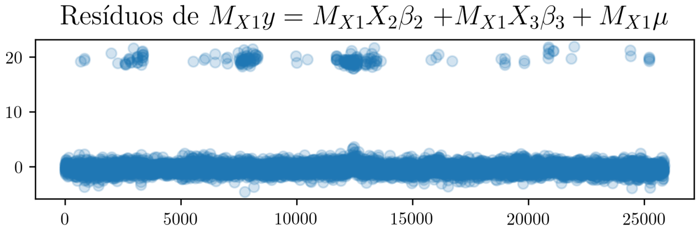
```
rc('font', **{'family': 'sans-serif', 'sans-serif': ['Helvetica']})
rc('text', usetex=True)

plt.subplot(2, 1, 1)
plt.plot(ajuste.resid, 'o', alpha=0.2, color='red')
plt.title(r"Resíduos de $y =X\beta + \mu$ ", fontsize=16)
plt.subplot(2, 1, 2)
plt.plot(ajuste2.resid, 'o', alpha=0.2)
plt.title(r"Resíduos de $M_{X1}y =M_{X1}X_{2}\beta_{2}$ "r"$ + M_{X1}X_{3}\beta_{3} + M_{X1}\mu$", fontsize=16)
plt.tight_layout()
plt.show()
```

Que nos dá o seguinte resultado:

Gráficos de Resíduos





Outras Aplicações

- Sazonalidade
- Tendência
- Sub e superspecificação

Bibliografia

Davidson, R., MacKinnon, J. G., et al. (2004).
Econometric theory and methods, volume 5.
Oxford University Press New York.

Frisch, R. and Waugh, F. V. (1933).
Partial time regressions as compared with individual trends.
Econometrica: Journal of the Econometric Society, pages 387–401.

Lovell, M. (2007).
A simple proof of the fwl (frisch-waugh-lovell) theorem.
Wesleyan Economics Working Papers 2005-012, Wesleyan University, Department of Economics.

Seabold, S. and Perktold, J. (2010).
Statsmodels: Econometric and statistical modeling with python.
In *9th Python in Science Conference*.