Na procura por encontrar um modelo que possa explicar razoavelmente bem a percentagem de alunos aprovados no teste de matemática do 4º ano, agrupamos as variáveis em painel para poder mensurar bem seus efeitos na variável dependente. O estimador mais adequado de acordo com estimativa e com as premissas assumidas deve ser escolhido dentre os possíveis a seguir: POLS; Efeitos Fixos (FE); Efeitos Aleatórios (RE).

Das premissas às vantagens e desvantagens de cada um dos estimadores podemos explicitar:

- POOLED OLS – Exogeneidade estrita; **cov (U_{it},U_{js}|X_{it}, X_{js}) = 0; var (U_{it}|X_{it}) =** $\sigma^2 u$; **full rank.** Utilizar quando os regressores não são correlacionados com a heterogeneidade não observada (**C_i**), pois dessa forma o estimador é consistente.
- FE – Consistente quando há exogeneidade estrita, mas não é eficiente; **full rank.** Utilizar quando os regressores são correlacionados com a heterogeneidade não observada (**C_i**), pois dessa forma o estimador é consistente. Na prática, isso significa fazer uma transformação no modelo e depois aplicação do OLS, resolvendo problema de endogeneidade. Decorre como problema desse método, a estimação das variáveis que não variam ao longo do tempo.

O modelo com efeitos fixos pode consumir muitos graus de liberdade quando há muitas unidades de corte transversal e/ou temporal.

- RE - Exogeneidade estrita; **cov (U_{it},U_{js}|X_{it}, X_{js}) = 0; var (U_{it}|X_{it}) =** $\sigma^2 u$; **full rank.** Utilizar quando os regressores não são correlacionados com a heterogeneidade não observada (**C_i**), pois dessa forma o estimador é consistente. Esse estimador, diferentemente do FE, permite mensurar variáveis que não variam no tempo, ou mensurar de forma mais precisa, aquelas que variam pouco ao longo do tempo.

As matrizes de variância-covariância utilizadas ao longo desse trabalho visam corrigir eventuais problemas de heterocedasticidade e autocorrelação. Nesse sentido, são usadas as matrizes robustas por clusters, permitindo realizar melhores inferências.

Após definir o modelo mais adequado, que relaciona as variáveis, a escolha entre os diferentes estimadores, de efeitos fixos ou aleatórios, será feita mediante teste de Hausman. Esse teste compara as estimativas de efeitos aleatórios com as de efeitos fixos. Diferenças significativas entre as estimativas sugerem a inconsistência do estimador RE. Na ausência de correlação entre regressores e **C_i**, ambos os estimadores serão consistentes, mas RE será mais eficiente.

Primeiramente, descrevemos as variáveis no anexo que são potenciais na entrada do modelo explicativo, com exceção daquelas dummies. Observamos algumas estatísticas desses atributos que facilitam a compreensão dos dados. O primeiro modelo postulado relaciona a variável dependente com o % de alunos elegíveis a almoço grátis, os gastos por aluno e o log dos gastos com estes, considerando a inflação, além das dummies no tempo.

$$\textbf{math4} = \alpha_0 + \alpha_1\,\textbf{lunch} + \alpha_2\,\textbf{exppp} + \alpha_3\,\textbf{lrexpp} + \delta_1\,\textbf{y95} + \delta_2\,\textbf{y96} + \delta_3\,\textbf{y97} + \delta_4\,\textbf{y98}$$

Sob os diferentes estimadores, com uso das variâncias robustas, obtivemos os seguintes coeficientes. É importante ressaltar que alguns desses estimadores não são significativos (ver anexo) e que por isso, sugerem que mudemos a forma funcional do modelo.

Variable	beta_POLS	beta_RE	beta_FE
lunch	-.42608832	-.37838136	-.01942969
exppp	.00050263	.00020915	.00024757
lrexpp	5.9713034	5.0333187	2.6411033
y95	11.71157	11.616944	11.817712
y96	13.248996	13.398143	13.694357
y97	10.413762	10.723641	10.847291
y98	23.804401	24.201507	24.248205
_cons	14.829132	21.703464	28.17641

O segundo modelo proposto para explicar a variável dependente se diferencia, pois inclui a dummy de escolas pequenas, além de excluir a variável de gastos por aluno. Isso se deve ao fato de que acreditamos que lrexpp e exppp são bem correlacionadas e por isso, a inserção da variável mais completa (lrexpp), que já contempla inflação, permitiria melhores estimações e obtenção de coeficientes mais significativos, como se pode ver no anexo.

$$\textbf{math4} = \alpha_0 + \alpha_1\,\textbf{lunch} + \alpha_2\,\textbf{lrexpp} + \delta_1\,\textbf{small} + \delta_2\,\textbf{y95} + \delta_3\,\textbf{y96} + \delta_4\,\textbf{y97} + \delta_5\,\textbf{y98}$$

Os coeficientes estimados com os diferentes estimadores são:

Variable	beta_POLS	beta_RE	beta_FE
lunch	-.42651465	-.37870805	-.01950011
lrexpp	7.8705589	5.8173444	3.6326745
small	5.1721261	4.3720591	(omitted)
y95	11.730723	11.624135	11.812404
y96	13.331117	13.429921	13.714798
y97	10.551431	10.77788	10.891048
y98	23.976922	24.27094	24.310102
_cons	.96476883	15.977785	20.916836

Cabe ressaltar ainda que os testes F de significância conjunta nos trazem mais evidência estatística de que as variáveis conjuntamente são estatisticamente significantes.

Posteriormente, realizou-se o teste aos efeitos fixos no tempo no FE para verificar se seria possível retirar as variáveis dummies referentes aos anos. O teste tem como hipótese nula que os coeficientes referentes aos anos são iguais a zero, enquanto que a hipótese alternativa diz que os coeficientes são diferentes de zero. De acordo com o output do teste abaixo, verificou-se que há evidencias de efeitos fixos no tempo, rejeitando-se a hipótese nula, reforçando a utilização das dummies no tempo.

```
. test( y95 y96 y97 y98)

 ( 1)  y95 = 0
 ( 2)  y96 = 0
 ( 3)  y97 = 0
 ( 4)  y98 = 0

       F(  4,  1682) =  555.66
            Prob > F =    0.0000
```

Em seguida, houve a realização do teste de Hausman para a proposta do estimador mais adequado. O output do teste de Hausman está abaixo e verificou-se que há evidências para a rejeição da hipótese nula e, portanto, o estimador mais adequado é o estimador de efeitos fixos. O FE ainda tem a vantagem de ser menos restritivo quanto às hipóteses, como por exemplo, não assumir que não haja correlação entre os regressores e $\mathbf{C_i}$. Apesar da matriz de covariâncias da diferenças entre os estimadores do teste de Hausman não ser definida positiva. No entanto, o STATA consegue calcular a estatística através de artifícios computacionais, e também, não há relatos na literatura se a distribuição assintótica sofre alguma distorção.

```
. hausman beta_FE_NR beta_RE_NR
```

	——— Coefficients ———			
	(b) beta_FE_NR	(B) beta_RE_NR	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
lunch	-.0195001	-.3787081	.3592079	.0291607
lrexpp	3.632675	5.817344	-2.18467	.9219224
y95	11.8124	11.62414	.1882694	.1594307
y96	13.7148	13.42992	.2848771	.1894086
y97	10.89105	10.77788	.1131677	.220322
y98	24.3101	24.27094	.0391617	.2264399

```
                b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      146.72
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

Anexo

Variable		Mean	Std. Dev.	Min	Max	Observations
math4	overall	63.57726	20.19047	2.9	100	N = 7150
	between		16.08074	11.75	98.94	n = 1683
	within		12.37335	13.71059	122.3439	T-bar = 4.24837
lunch	overall	36.7446	25.10881	0	100	N = 7150
	between		25.10869	.306	98.61333	n = 1683
	within		4.329699	3.244596	66.3826	T-bar = 4.24837
exppp	overall	3972.765	843.9141	1521	13581	N = 7150
	between		669.6237	2301.8	9139	n = 1683
	within		528.6379	-442.2347	9972.765	T-bar = 4.24837
cpi	overall	1.571522	.0506793	1.482	1.63	N = 7150
	between		.0175765	1.525	1.601333	n = 1683
	within		.0481114	1.481189	1.656189	T-bar = 4.24837
rexppp	overall	4048.838	814.8421	1539.028	14708.17	N = 7150
	between		673.4252	2373.791	9110.974	n = 1683
	within		472.1131	-229.745	10823.15	T-bar = 4.24837
lrexpp	overall	8.287368	.1924372	7.338906	9.596158	N = 7150
	between		.1543265	7.763933	9.043225	n = 1683
	within		.1165782	7.564195	9.070365	T-bar = 4.24837

As variáveis da base de dados se referem a:

distid	district identifier
schid	school identifier
lunch	% eligible for free lunch
exppp	expenditure per pupil
math4	4th grade math test
year	1992=school yr 1991-2
cpi	consumer price index
rexppp	(exppp/cpi)*1.695: 1997 $
lrexpp	log(rexpp)
y94	=1 if year == 1994
y95	=1 if year == 1995
y96	=1 if year == 1996
y97	=1 if year == 1997
y98	=1 if year == 1998
small	=1 if the school has less than 100 students enrolled

Pooled OLS Modelo 1

Linear regression

Number of obs	= 7150
F(7, 1682)	= 597.51
Prob > F	= 0.0000
R-squared	= 0.4012
Root MSE	= 15.631

(Std. Err. adjusted for 1683 clusters in schid)

math4	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lunch	-.4260883	.0133935	-31.81	0.000	-.452358	-.3998187
exppp	.0005026	.0016997	0.30	0.767	-.0028311	.0038363
lrexpp	5.971303	7.082661	0.84	0.399	-7.920454	19.86306
y95	11.71157	.5788769	20.23	0.000	10.57617	12.84696
y96	13.249	.6200841	21.37	0.000	12.03278	14.46521
y97	10.41376	.6953555	14.98	0.000	9.049909	11.77762
y98	23.8044	.7608678	31.29	0.000	22.31205	25.29675
_cons	14.82913	52.24064	0.28	0.777	-87.63437	117.2926

Efeitos Aleatórios Modelo 1

Random-effects GLS regression
Group variable: schid

Number of obs	= 7150
Number of groups	= 1683

R-sq: within = 0.3443
between = 0.4267
overall = 0.3996

Obs per group: min	= 3
avg	= 4.2
max	= 5

corr(u_i, X) = 0 (assumed)

Wald chi2(7)	= 4085.62
Prob > chi2	= 0.0000

math4	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lunch	-.3783814	.0111848	-33.83	0.000	-.4003032	-.3564595
exppp	.0002091	.0011138	0.19	0.851	-.0019739	.0023922
lrexpp	5.033319	4.718923	1.07	0.286	-4.215601	14.28224
y95	11.61694	.5435046	21.37	0.000	10.55169	12.68219
y96	13.39814	.5384819	24.88	0.000	12.34274	14.45355
y97	10.72364	.5806352	18.47	0.000	9.585617	11.86167
y98	24.20151	.6112494	39.59	0.000	23.00348	25.39953
_cons	21.70346	34.88973	0.62	0.534	-46.67914	90.08607
sigma_u	10.727036					
sigma_e	11.333623					
rho	.47252444	(fraction of variance due to u_i)				

Efeitos Fixos Modelo 1

```
Fixed-effects (within) regression               Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

R-sq:  within  = 0.3592                         Obs per group: min =         3
       between = 0.0224                                        avg =       4.2
       overall = 0.1470                                        max =         5

                                                F(7,1682)          =    431.86
corr(u_i, Xb)  = 0.0057                         Prob > F           =    0.0000

                                    (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lunch |  -.0194297   .0392258    -0.50   0.620    -.0963662    .0575068
       exppp |   .0002476   .0013149     0.19   0.851    -.0023315    .0028266
     lrexpp |   2.641103   5.673599     0.47   0.642    -8.486954    13.76916
         y95 |   11.81771   .5406662    21.86   0.000     10.75726    12.87816
         y96 |   13.69436   .5741153    23.85   0.000      12.5683    14.82041
         y97 |   10.84729   .6348292    17.09   0.000     9.602152    12.09243
         y98 |    24.2482   .6715035    36.11   0.000     22.93113    25.56528
       _cons |   28.17641   42.10287     0.67   0.503    -54.40311    110.7559
-------------+----------------------------------------------------------------
     sigma_u |  15.903173
     sigma_e |  11.333623
         rho |  .66317821   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Pooled OLS Modelo 2

```
Linear regression                               Number of obs =      7150
                                                F(  7,  1682) =    599.60
                                                Prob > F      =    0.0000
                                                R-squared     =    0.4017
                                                Root MSE      =    15.624

                                    (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lunch |  -.4265147   .0133945   -31.84   0.000    -.4527862   -.4002431
     lrexpp |   7.870559   1.512585     5.20   0.000     4.903813    10.83731
       small |   5.172126    2.52384     2.05   0.041     .2219283    10.12232
         y95 |   11.73072   .5751393    20.40   0.000     10.60266    12.85879
         y96 |   13.33112   .6041528    22.07   0.000     12.14615    14.51609
         y97 |   10.55143   .6314118    16.71   0.000     9.312995    11.78987
         y98 |   23.97692   .6362909    37.68   0.000     22.72892    25.22493
       _cons |   .9647688   12.21426     0.08   0.937    -22.99198    24.92152
------------------------------------------------------------------------------
```

Efeitos Aleatórios Modelo 2

```
Random-effects GLS regression                   Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

R-sq:  within  = 0.3443                         Obs per group: min =         3
       between = 0.4272                                        avg =       4.2
       overall = 0.4001                                        max =         5

                                                Wald chi2(7)       =   3919.72
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

                                (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lunch |  -.3787081   .0135219   -28.01   0.000    -.4052106   -.3522055
      lrexpp |   5.817344   1.352155     4.30   0.000     3.16717    8.467519
       small |   4.372059   2.690631     1.62   0.104    -.9014805   9.645599
         y95 |   11.62414   .5164185    22.51   0.000     10.61197    12.6363
         y96 |   13.42992   .5419898    24.78   0.000     12.36764    14.4922
         y97 |   10.77788   .5696456    18.92   0.000     9.661395    11.89436
         y98 |   24.27094   .5694026    42.63   0.000     23.15493    25.38695
       _cons |   15.97779   10.93991     1.46   0.144    -5.464052    37.41962
-------------+----------------------------------------------------------------
     sigma_u |  10.720915
     sigma_e |  11.332629
         rho |  .47228364   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Efeitos Fixos Modelo 2

```
Fixed-effects (within) regression               Number of obs      =      7150
Group variable: schid                           Number of groups   =      1683

R-sq:  within  = 0.3592                         Obs per group: min =         3
       between = 0.0227                                        avg =       4.2
       overall = 0.1471                                        max =         5

                                                F(6,1682)          =    504.09
corr(u_i, Xb)  = 0.0059                         Prob > F           =    0.0000

                                (Std. Err. adjusted for 1683 clusters in schid)
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lunch |  -.0195001   .0392341    -0.50   0.619    -.096453    .0574527
      lrexpp |   3.632675   1.733773     2.10   0.036     .2320939   7.033255
       small |          0  (omitted)
         y95 |    11.8124   .5384138    21.94   0.000     10.75637    12.86844
         y96 |    13.7148    .569822    24.07   0.000     12.59716    14.83243
         y97 |   10.89105   .6091401    17.88   0.000     9.696295     12.0858
         y98 |    24.3101   .6053166    40.16   0.000     23.12285    25.49736
       _cons |   20.91684   14.05233     1.49   0.137    -6.645066    48.47874
-------------+----------------------------------------------------------------
     sigma_u |   15.90154
     sigma_e |  11.332629
         rho |  .66317151   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```