

Relatório 3

Luca Klein

2022-12-15

1. Modelos Tobit

1.1 Tobit

As regressões lineares muitas vezes se deparam com problemas de amostragem, sendo censura e truncagem alguns deles. A definição de truncagem é dada pela ausência de observações nas variáveis explicativas e na explicada, ao passo que a censura demonstra a falta de dados somente para variáveis dependentes. Para o tratamento da problemática de censura, temos o modelo de Tobin (1958) conhecido como Tobit ou regressão normal censurada. Quando nos deparamos com a censura a utilização do método Mínimos Quadrados Ordinários (MQO) não chega em resultados com estimadores consistentes, tendo em vista que a amostra não trás valores suficientes para estimação dos verdadeiros valores populacionais. Para evitar esse tipo de erro, o Tobit assume algumas propriedades como a gaussianidade dos resíduos com média zero e variância constante e, pelo lado da variável latente, temos que sua distribuição é normal , com média $X'\beta$ e σ^2 , e seus valores devem ser positivos, caso contrário, ela não existe.

1.11 Aplicação

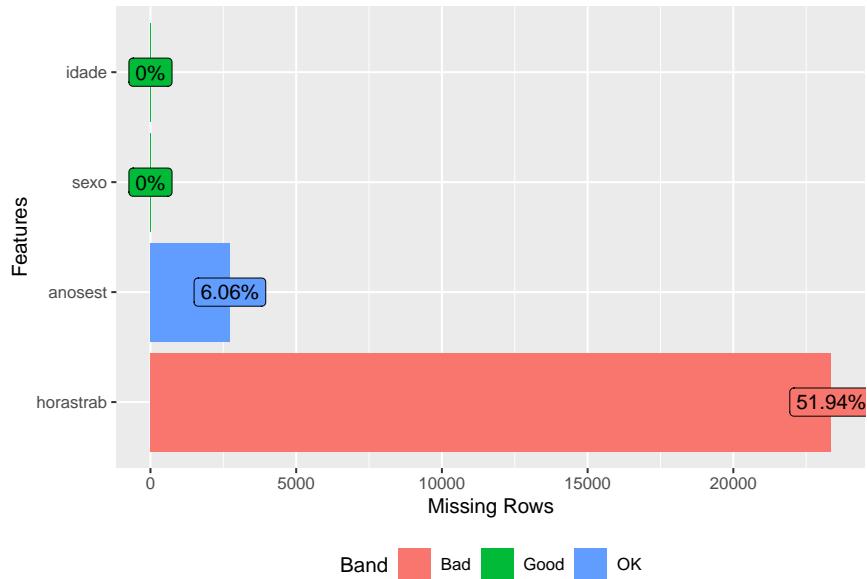
Para estimação dos modelos foi selecionado dois Estados nos períodos do 2T2012 e do 2T2022, a partir do microdados da PNAD. O modelo considerou como variável dependente as horas de trabalho (horastrab) e como explicativas as variáveis anos de estudos (anoest), idade e sexo. Por fim, a escolha das unidades federativas (UFs) foi pautada na média de horas habitualmente trabalhadas por semana no trabalho principal das pessoas de 14 anos ou mais de idade, em que foi selecionado São Paulo (SP) e Piauí (PI), em virtude dessas UFs apresentarem a maior e a menor média de horas de trabalho, respectivamente, segundo a PNAD do 2T2012.

1.12 Análise Descritiva

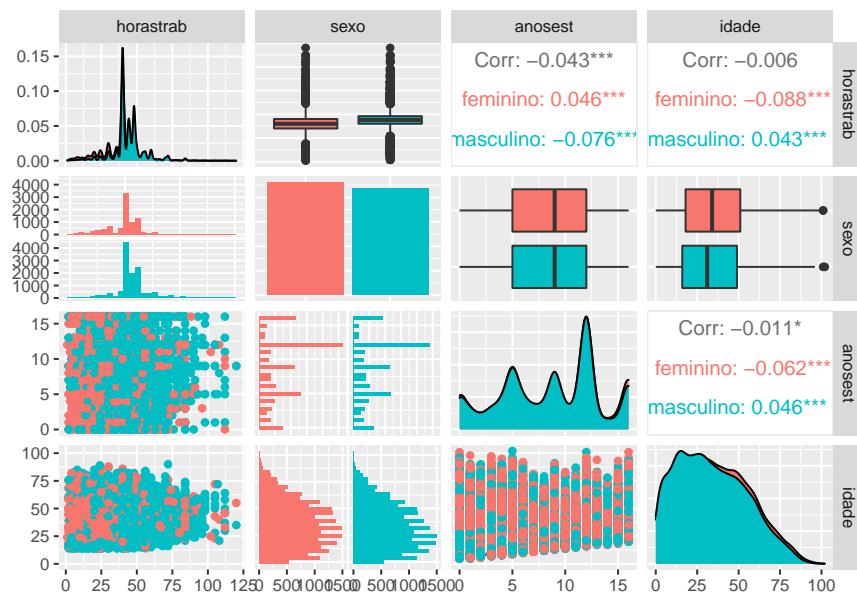
Nesta seção será apresentado as informações descritivas e o comportamento das variáveis para as UFs e períodos.

São Paulo 2T2012

```
##      horastrab          sexo          anoest         idade
##  Min.   : 1.00   Length:44981   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 40.00   Class :character  1st Qu.: 5.000   1st Qu.: 17.00
##  Median : 40.00   Mode  :character  Median : 9.000   Median : 33.00
##  Mean   : 41.77                    Mean   : 8.598   Mean   : 34.58
##  3rd Qu.: 48.00                    3rd Qu.:12.000   3rd Qu.: 50.00
##  Max.   :120.00                   Max.   :16.000   Max.   :102.00
##  NA's    :23362                  NA's    :2726
```

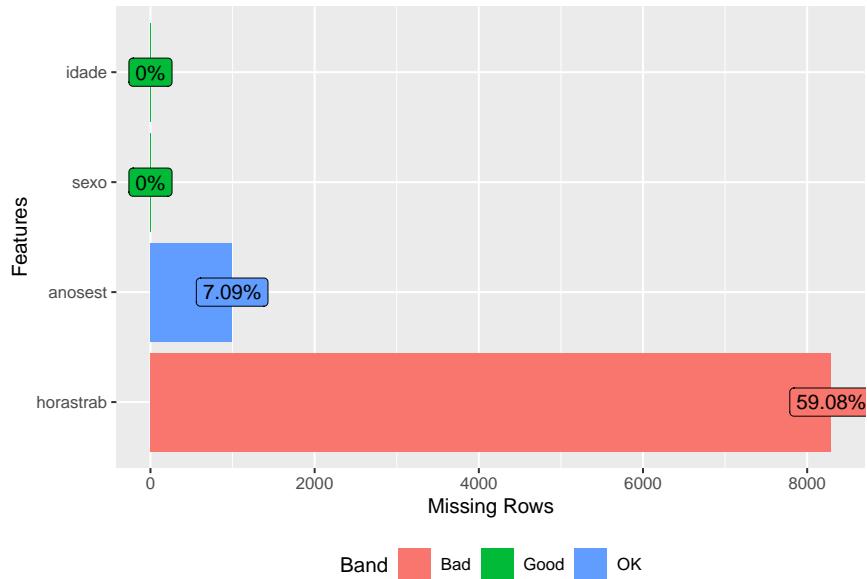


```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

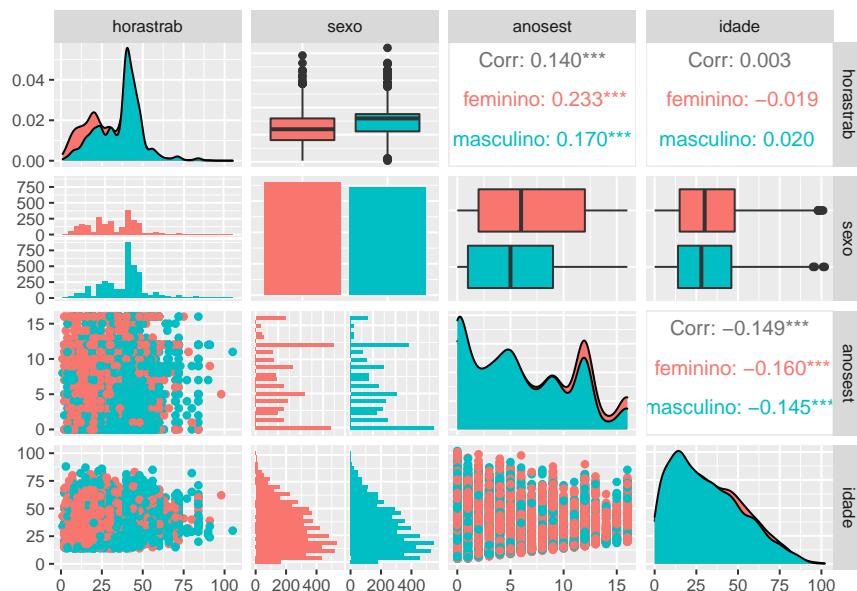


Piauí 2T2012

```
##      horastrab          sexo          anoest         idade
##  Min.   : 1.00   Length:14039   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 24.00   Class :character  1st Qu.: 2.00   1st Qu.: 14.00
##  Median : 40.00   Mode  :character  Median : 5.00   Median : 29.00
##  Mean   : 34.17                               Mean   : 6.24   Mean   : 31.89
##  3rd Qu.: 44.00                               3rd Qu.:10.00   3rd Qu.: 47.00
##  Max.   :105.00                               Max.   :16.00   Max.   :102.00
##  NA's    :8294                                NA's    :995
```

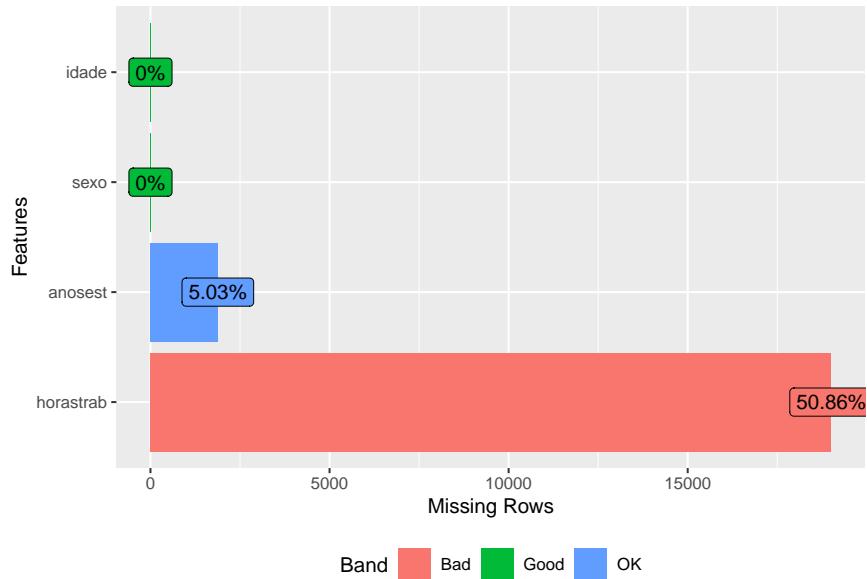


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

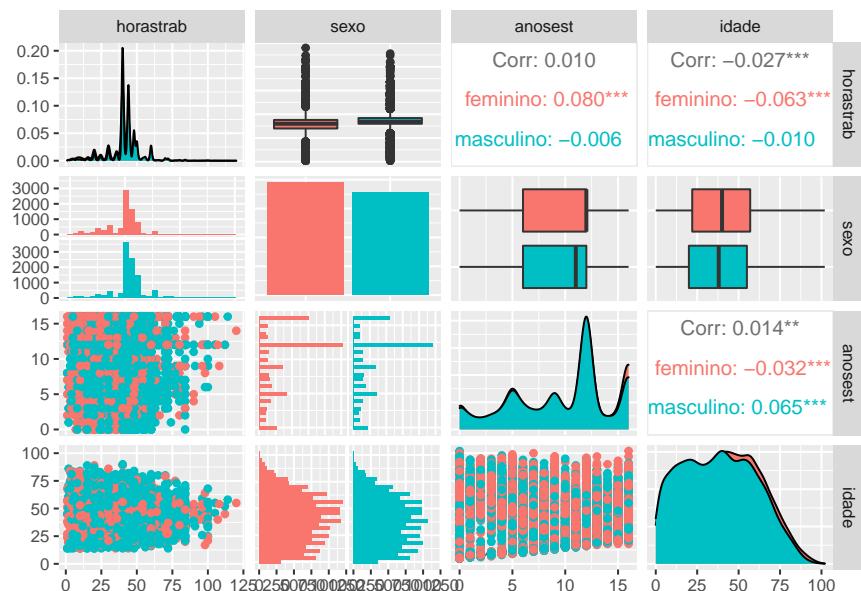


São Paulo 2T2022

```
## horastrab           sexo            anosest          idade
## Min.   : 1.00    Length:37377      Min.   : 0.000  Min.   : 0.00
## 1st Qu.: 40.00   Class :character  1st Qu.: 6.000  1st Qu.: 21.00
## Median : 40.00   Mode  :character  Median :12.000  Median : 39.00
## Mean   : 40.62                    Mean   : 9.612  Mean   : 39.21
## 3rd Qu.: 44.00                    3rd Qu.:12.000  3rd Qu.: 56.00
## Max.   :120.00                    Max.   :16.000  Max.   :102.00
## NA's    :19010                     NA's    :1881
```

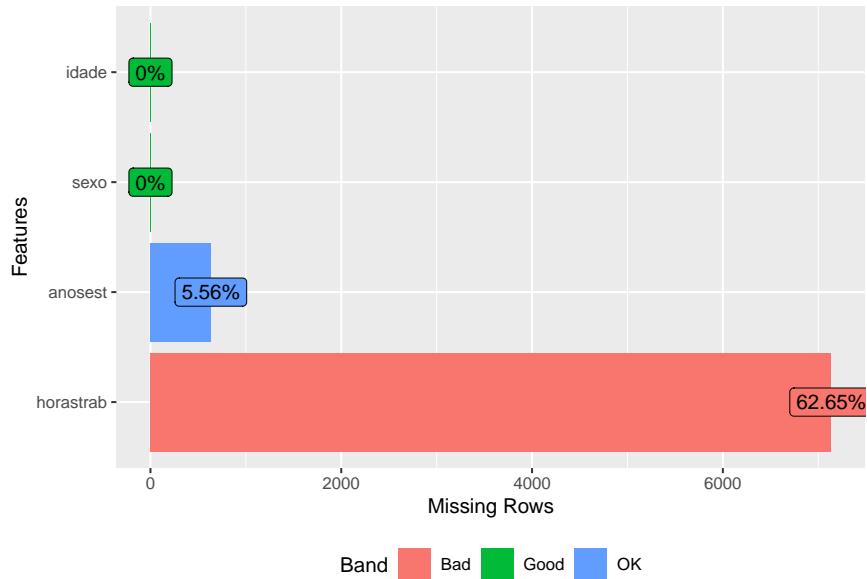


```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

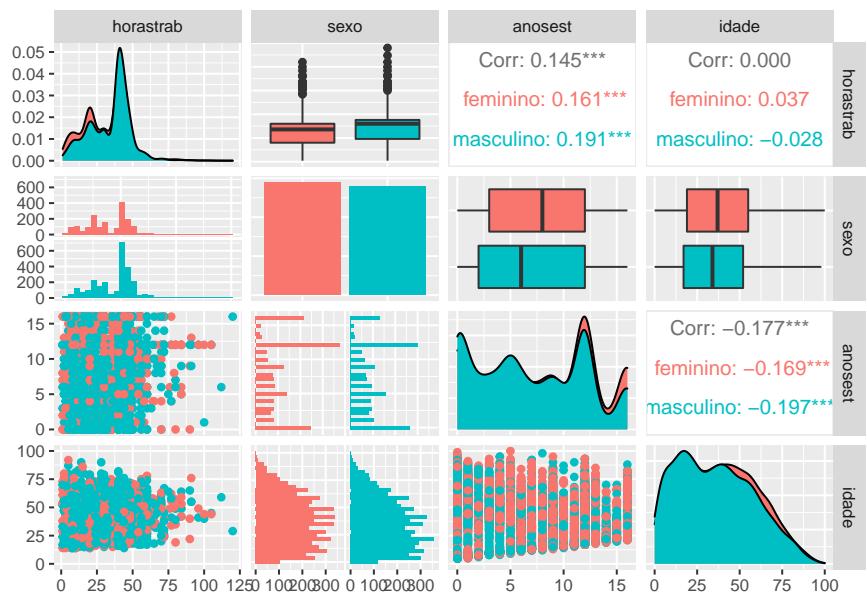


Piauí 2T2022

```
##      horastrab          sexo         anosest        idade
##  Min.   : 1.00   Length:11392   Min.   : 0.000   Min.   : 0.00
##  1st Qu.: 20.00   Class :character  1st Qu.: 3.000   1st Qu.: 18.00
##  Median : 40.00   Mode  :character  Median : 7.000   Median : 36.00
##  Mean   : 33.21                               Mean   : 7.358   Mean   : 36.79
##  3rd Qu.: 44.00                               3rd Qu.:12.000   3rd Qu.: 54.00
##  Max.   :120.00                               Max.   :16.000   Max.   :100.00
##  NA's    :7137                                NA's    :633
```



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Para o primeiro período selecionado, os dados apontam que SP apresentou, na média, estatísticas superiores às do PI para horas de trabalho, anos de estudo e idade. Em termos de correlação cruzada, o sexo feminino e masculino para SP acusaram impactos diferentes da escolaridade para horas de trabalho, sendo o primeiro positivo e o segundo negativo. Enquanto a idade sugeriu uma dinâmica inversa, em que as mulheres impactam negativamente as horas ofertadas e os homens positivamente. Para o PI, as correlações dos gêneros demonstraram influências positivas da escolaridade para horas de trabalho, sendo as mulheres aquelas com o maior impacto. Por outro lado, no que se refere à idade, a relação com as horas de trabalho segue a dinâmica de SP, em que à medida que idade avança, na média, as mulheres tendem a ter um impacto negativo para as horas de trabalho e homens positivo.

No 2T2022, as variáveis aleatórias contínuas para SP mantiveram-se com média mais elevada em com-

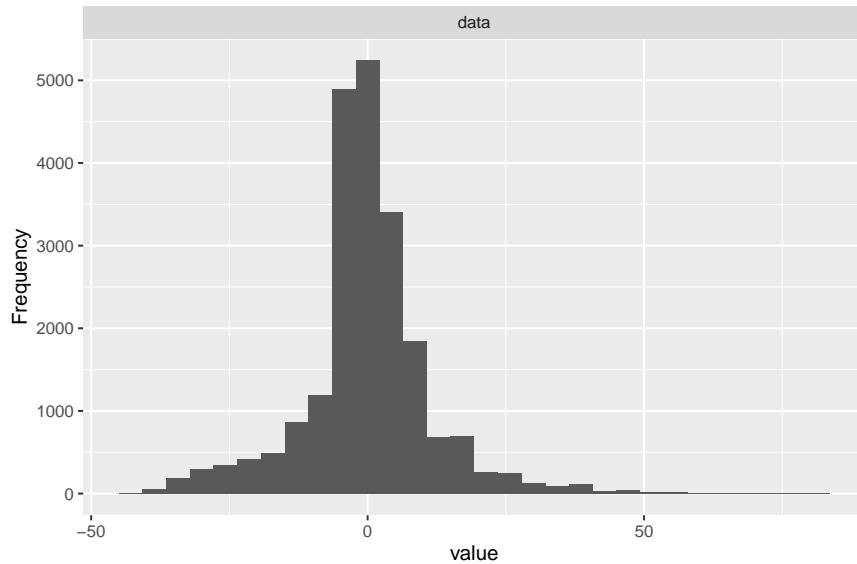
paração as do PI. Em linhas gerais, as horas ofertadas caíram para as duas UFs, ao mesmo tempo que a escolaridade aumentou para ambas, assim como a idade. Pelo lado da correlação cruzada, o impacto médio da escolaridade dos gêneros para horas de trabalho se manteve com dinâmica semelhante ao primeiro período para ambas as UFs. Quanto as influências da idade dos sexos para as horas de trabalho, o período apontou para SP impactos negativos de homens e mulheres, enquanto o PI inverteu a dinâmica do 2T2012, agora as mulheres tendem a ter um impacto médio positivo para horas ofertas e os homens negativos. Por fim, em termos de dados censurados, o PI acusou maior censura que SP em ambos os períodos com maior porcentagem de dados faltantes no segundo recorte de tempo selecionado e SP teve pequena redução na censura frente ao primeiro período.

1.13 Modelo

Os *outputs* dos modelos seguem abaixo, bem como a avaliação de resíduos em termos de média, indicados por [1], e variância:

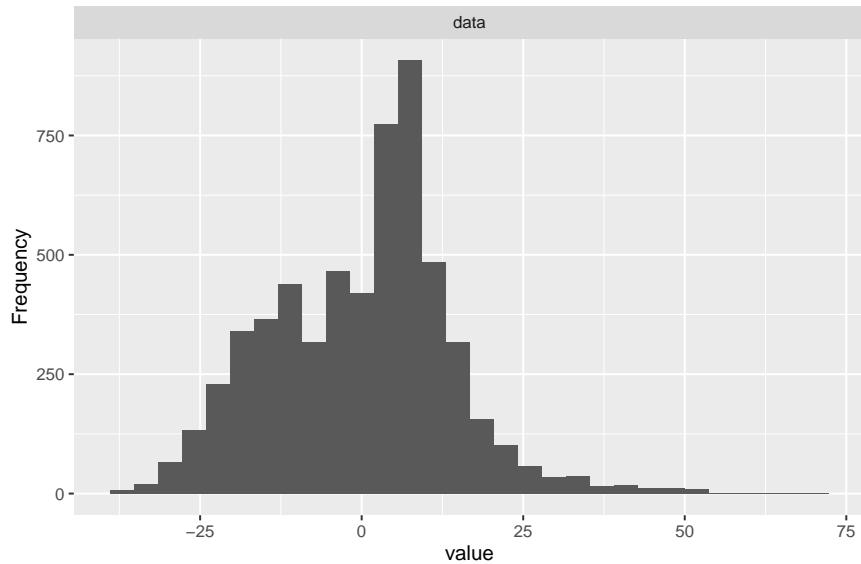
São Paulo 2T2012

```
##
## Call:
## tobit(formula = horastrab ~ sexo + idade + anosest, data = t_sp_1)
##
## Observations: (23362 observations deleted due to missingness)
##           Total    Left-censored     Uncensored Right-censored
##           21619            0            21619            0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 40.164367  0.389355 103.156 < 2e-16 ***
## sexomasculino 5.434287  0.163793  33.178 < 2e-16 ***
## idade      -0.018146  0.006297  -2.882 0.003956 **
## anosest     -0.077653  0.020609  -3.768 0.000165 ***
## Log(scale)    2.472598  0.004809 514.146 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 11.85
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -8.413e+04 on 5 Df
## Wald-statistic: 1151 on 3 Df, p-value: < 2.22e-16
## [1] 2.039166e-15
```



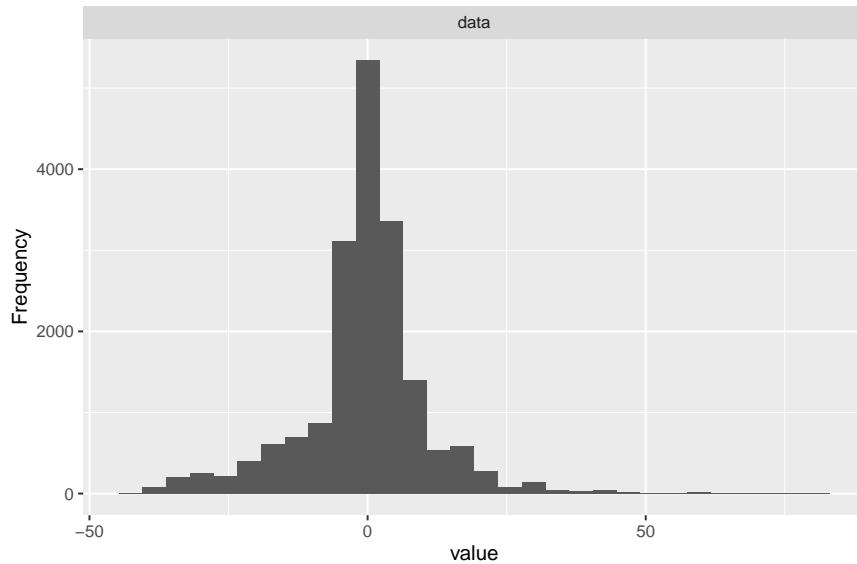
Piauí 2T2012

```
##
## Call:
## tobit(formula = horastrab ~ sexo + anosest, data = t_pi_1)
##
## Observations: (8294 observations deleted due to missingness)
##      Total Left-censored      Uncensored Right-censored
##      5745          0          5745          0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 25.149569   0.451671 55.68 <2e-16 ***
## sexomasculino 7.595745   0.385137 19.72 <2e-16 ***
## anosest     0.578984   0.038078 15.21 <2e-16 ***
## Log(scale)   2.635219   0.009329 282.47 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 13.95
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -2.329e+04 on 4 Df
## Wald-statistic: 511.1 on 2 Df, p-value: < 2.22e-16
## [1] 4.659224e-15
```



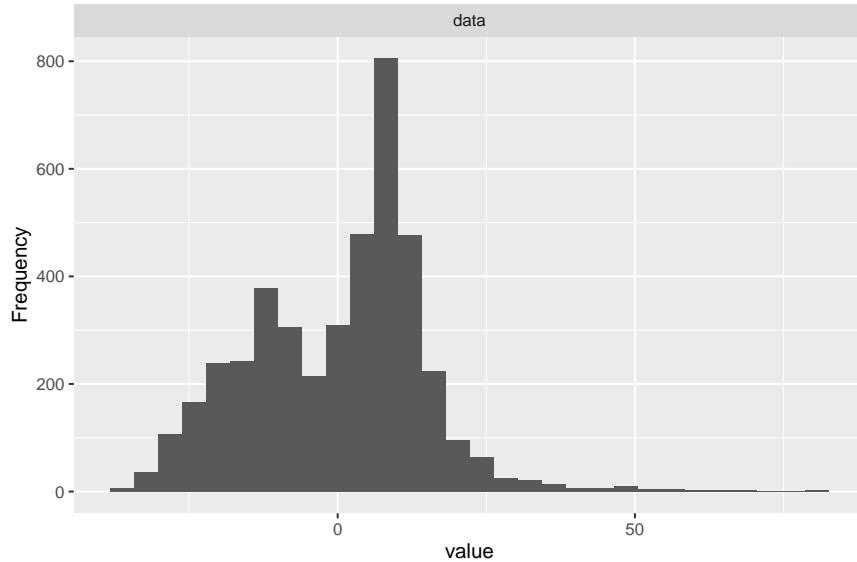
São Paulo 2T2022

```
##
## Call:
## tobit(formula = horastrab ~ sexo + idade + anosest, data = t_sp_2)
##
## Observations: (19010 observations deleted due to missingness)
##      Total Left-censored      Uncensored Right-censored
##      18367          0        18367          0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 38.348884  0.446721 85.845 < 2e-16 ***
## sexo masculino 4.126800  0.171830 24.017 < 2e-16 ***
## idade       -0.022904  0.006311 -3.629 0.000284 ***
## anosest      0.080637  0.023450  3.439 0.000585 ***
## Log(scale)    2.440615  0.005218 467.771 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 11.48
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -7.089e+04 on 5 Df
## Wald-statistic: 590.3 on 3 Df, p-value: < 2.22e-16
##
## [1] -2.497452e-15
```



Piauí 2T2022

```
##
## Call:
## tobit(formula = horastrab ~ sexo + anosest, data = t_pi_2)
##
## Observations: (7137 observations deleted due to missingness)
##      Total Left-censored      Uncensored Right-censored
##        4255          0           4255          0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 24.82987   0.63747  38.95 <2e-16 ***
## sexomasculino 4.82432   0.46631  10.35 <2e-16 ***
## anosest     0.56073   0.04764  11.77 <2e-16 ***
## Log(scale)   2.68153   0.01084 247.37 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 14.61
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -1.745e+04 on 4 Df
## Wald-statistic: 200.4 on 2 Df, p-value: < 2.22e-16
## [1] -2.093599e-15
```



De modo geral, os modelos atendem, em certa medida, os pressupostos básicos dos resíduos com média nula e variância constante. No entanto, ainda que bem comportada em alguns pontos, a variância não apresenta um comportamento homocedástico, algo que pode ser explicado pela grande quantidade de zeros nas horas de trabalho, o método Double Hurdle que será utilizado na próxima seção deve corrigir isso e modelar os zeros a partir da Binomial Negativa.

Resultados

Coeficientes	SP_1	PI_1	SP_2	PI_2
(Intercept)	40.1644	25.1496	38.3489	24.8299
sexomasculino	5.4343	7.5957	4.1268	4.8243
idade	-0.0181	0.5790	-0.0229	0.5607
anoest	-0.0777	25.1496	0.0806	24.8299

No que tange aos resultados, os coeficientes estimados devem ser interpretados como efeito marginal. Logo, para o primeiro período, temos que o impacto marginal do sexo masculino é maior para PI frente a SP, assim como a escolaridade e a idade que apresentaram efeitos marginais negativos para horas de oferta de trabalho para os paulistas. Quanto ao segundo período, pelo lado do gênero, as UFs apontaram uma redução das influências marginais, sendo a do PI maior do que de SP. Já a idade seguiu a dinâmica do 2T2012, mas com ligeiras diferenças para horas de trabalho na margem. Já a escolaridade, para SP, inverteu-se o impacto na margem e PI seguiu com um efeito similar ao do primeiro recorte de tempo.

1.2 Double Hurdle

A abordagem proposta por Cragg (1971) é uma generalização do modelo de Tobin (1958), sendo esse um método de classificação da amostra em termos de participação de alguma atividade, tendo em vista que sua motivação parte da necessidade de modelar valores nulos presentes nas variáveis, algo que muitas vezes não é tratado em certos tipos de modelos. Nessa linha, para tomar a participação de atividades é cabível a utilização, no primeiro momento, da estimação de um Probit ou Logit e, no segundo momento, a fim de garantir que os participantes assumam somente valores positivos, temos que estimar sua densidade, sendo essa a configuração dos dois estágios do modelo. Finalmente, depois calcular os dois estágios, o método

permite a estimativa das proporções populacionais das categorias segundo as características imputadas nas variáveis independentes.

1.21 Aplicação

Para estimativa do modelo Double Hurdle foi selecionado dois Estados nos períodos do 2T2012 e do 2T2022, a partir dos microdados da PNAD. O modelo considerou como variável dependente as horas de trabalho (horastrab) e como explicativas os anos de estudos (anoest), idade e sexo. Vale ressaltar que para a variável explicada introduziu-se o valor zero nas observações sem valor. Por fim, a escolha das unidades federativas (UFs) foi pautada na média de horas habitualmente trabalhadas por semana no trabalho principal das pessoas de 14 anos ou mais de idade, em que foi selecionado São Paulo (SP) e Piauí (PI), em virtude dessas UFs apresentarem a maior e a menor média de horas de trabalho, respectivamente, também segundo a PNAD do 2T2012.

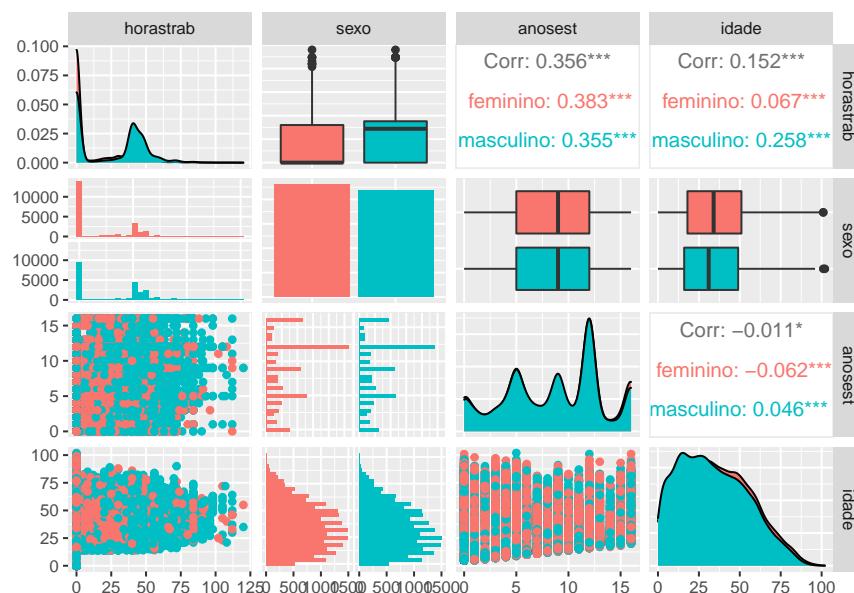
1.22 Análise Descritiva

Nesta seção serão apresentados as informações descritivas e o comportamento das variáveis para as UFs e períodos.

São Paulo 2T2012

```
##      horastrab          sexo         anoest        idade
## Min.   : 0.00  Length:44981   Min.   : 0.000  Min.   : 0.00
## 1st Qu.: 0.00  Class :character  1st Qu.: 5.000  1st Qu.: 17.00
## Median : 0.00  Mode  :character  Median : 9.000  Median : 33.00
## Mean   : 20.08                      Mean   : 8.598  Mean   : 34.58
## 3rd Qu.: 40.00                      3rd Qu.:12.000  3rd Qu.: 50.00
## Max.   :120.00                      Max.   :16.000  Max.   :102.00
## NA's    :2726
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



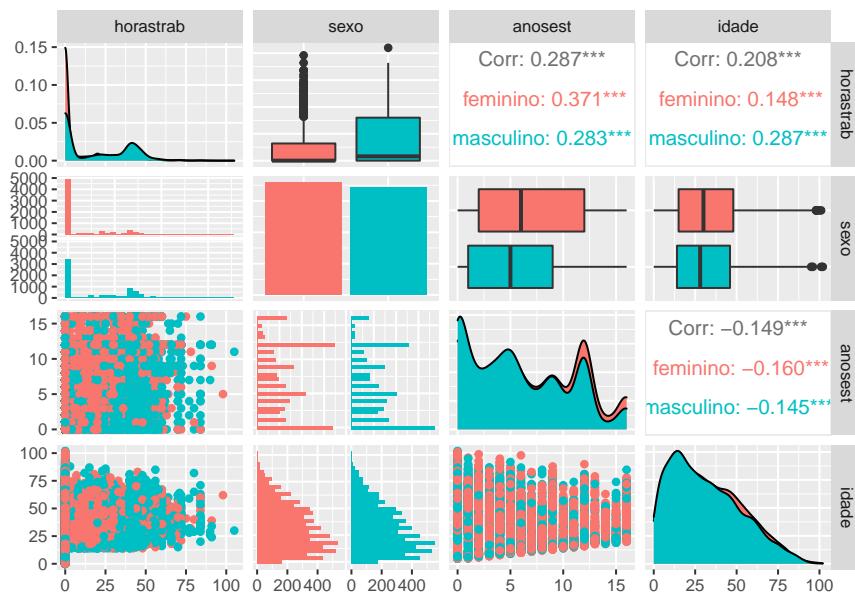
Piauí 2T2012

```

##      horastrab          sexo        anoest        idade
##  Min.   : 0.00  Length:14039   Min.   : 0.00  Min.   : 0.00
##  1st Qu.: 0.00  Class  :character  1st Qu.: 2.00  1st Qu.: 14.00
##  Median : 0.00  Mode   :character  Median : 5.00  Median : 29.00
##  Mean   : 13.98                                     Mean   : 6.24  Mean   : 31.89
##  3rd Qu.: 30.00                                     3rd Qu.:10.00 3rd Qu.: 47.00
##  Max.   :105.00                                     Max.   :16.00  Max.   :102.00
##                                         NA's   :995

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```



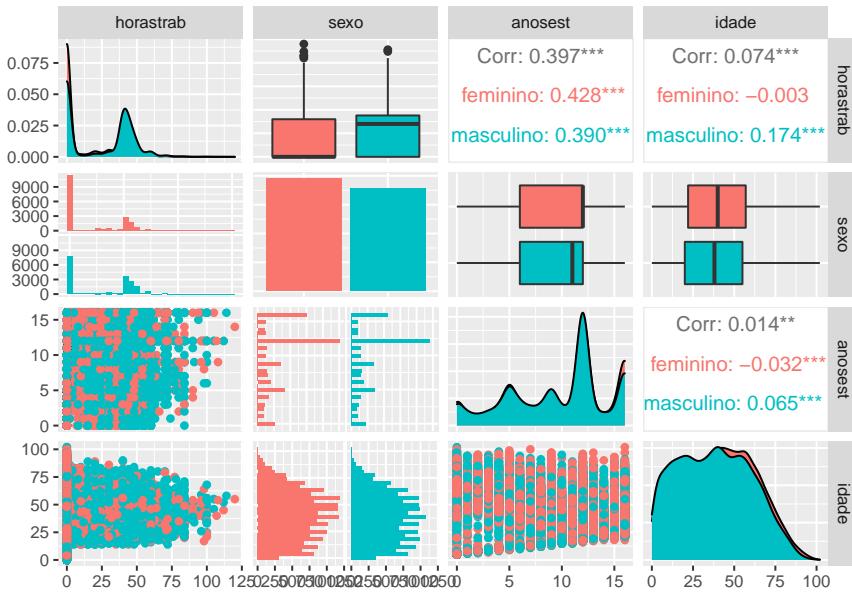
São Paulo 2T2022

```

##      horastrab          sexo        anoest        idade
##  Min.   : 0.00  Length:37377   Min.   : 0.000  Min.   : 0.00
##  1st Qu.: 0.00  Class  :character  1st Qu.: 6.000  1st Qu.: 21.00
##  Median : 0.00  Mode   :character  Median :12.000  Median : 39.00
##  Mean   : 19.96                                     Mean   : 9.612  Mean   : 39.21
##  3rd Qu.: 40.00                                     3rd Qu.:12.000 3rd Qu.: 56.00
##  Max.   :120.00                                     Max.   :16.000  Max.   :102.00
##                                         NA's   :1881

## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

```



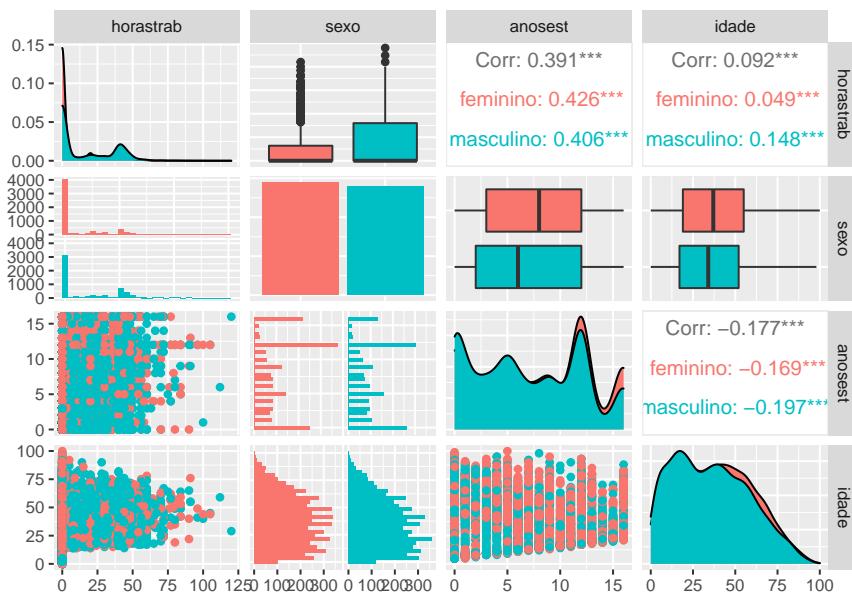
Piauí 2T2022

```
##      horastrab          sexo         anosest        idade
## Min.   : 0.0   Length:11392   Min.   : 0.000   Min.   : 0.00
## 1st Qu.: 0.0   Class :character 1st Qu.: 3.000   1st Qu.: 18.00
## Median : 0.0   Mode  :character Median : 7.000   Median : 36.00
## Mean   : 12.4                    Mean   : 7.358   Mean   : 36.79
## 3rd Qu.: 25.0                    3rd Qu.:12.000   3rd Qu.: 54.00
## Max.   :120.0                    Max.   :16.000   Max.   :100.00
## NA's    :633
```

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.

'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.



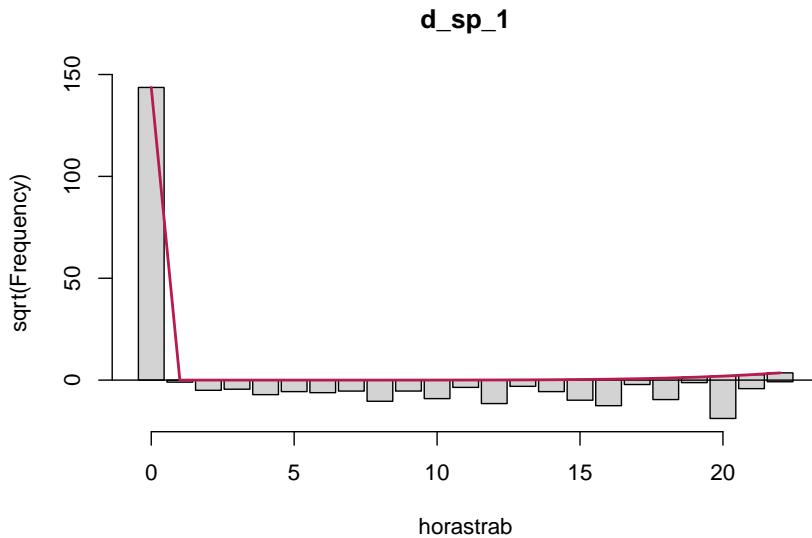
Para o primeiro período, nota-se que a média de horas de trabalho para SP é superior em 6 horas, ao passo que no 2T2022 tal distância se mantém, contudo, há uma redução nas horas para ambas as UFs. Quando observamos os anos de estudo, verificou-se que SP apresenta uma média mais elevada que PI para ambos os períodos e que houve uma evolução para os Estados no segundo período em comparação ao primeiro. O indicador de idade também sugere uma média mais elevada para SP em relação ao PI para as duas épocas em questão. No que tange à correlação entre as variáveis, nota-se, de maneira generalizada, que na média as mulheres são quem mais afetam a oferta de horas de trabalho à medida que a escolaridade cresce. Por outro lado, também de forma generalizada, o impacto da idade para as horas de trabalho é superior para homens frente às mulheres, na média.

1.23 Modelo

Os *outputs* dos modelos seguem abaixo, bem como o *fit* de cada um:

São Paulo 2T2012

```
## 
## Call:
## hurdle(formula = horastrab ~ sexo + idade + anosest, data = h_sp_1)
## 
## Pearson residuals:
##      Min     1Q Median     3Q    Max 
## -3.3400 -0.7620 -0.3011  0.7717  8.5573 
## 
## Count model coefficients (truncated poisson with log link):
##                               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)            3.691e+00  5.091e-03 724.944 < 2e-16 ***
## sexomasculino       1.315e-01  2.163e-03  60.813 < 2e-16 ***
## idade                -4.328e-04 8.203e-05 -5.277 1.32e-07 *** 
## anosest              -1.856e-03 2.688e-04 -6.906 4.97e-12 *** 
## Zero hurdle model coefficients (binomial with logit link):
##                               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -2.6284468  0.0381212 -68.95 <2e-16 ***
## sexomasculino        0.9331313  0.0222229   41.99 <2e-16 *** 
## idade                 0.0121704  0.0005658   21.51 <2e-16 *** 
## anosest               0.2063831  0.0026376   78.25 <2e-16 *** 
## ---                
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Number of iterations in BFGS optimization: 10
## Log-likelihood: -1.239e+05 on 8 Df
```

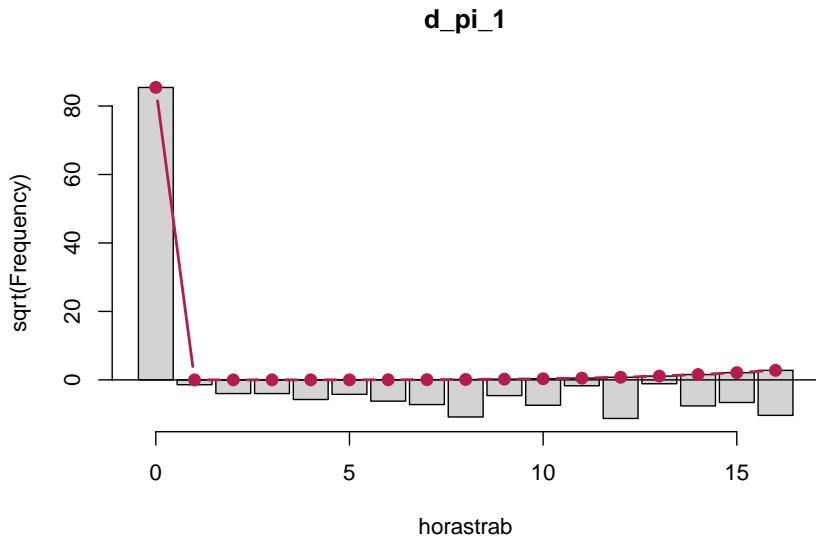


Piauí 2T2012

```

## 
## Call:
## hurdle(formula = horastrab ~ sexo + idade + anosest, data = h_pi_1)
## 
## Pearson residuals:
##      Min     1Q Median     3Q    Max 
## -4.2109 -0.7454 -0.4404  0.7742  7.4980 
## 
## Count model coefficients (truncated poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  3.1534634  0.0099179 317.96   <2e-16 ***
## sexo masculino 0.2311809  0.0048409  47.76   <2e-16 ***  
## idade        0.0022513  0.0001735  12.97   <2e-16 ***  
## anosest      0.0189189  0.0004898  38.62   <2e-16 ***  
## Zero hurdle model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.849616   0.065583  -43.45   <2e-16 ***  
## sexo masculino 1.165492   0.040685   28.65   <2e-16 ***  
## idade        0.027701   0.001043   26.57   <2e-16 ***  
## anosest      0.167629   0.004528   37.02   <2e-16 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Number of iterations in BFGS optimization: 11
## Log-likelihood: -4.089e+04 on 8 Df

```



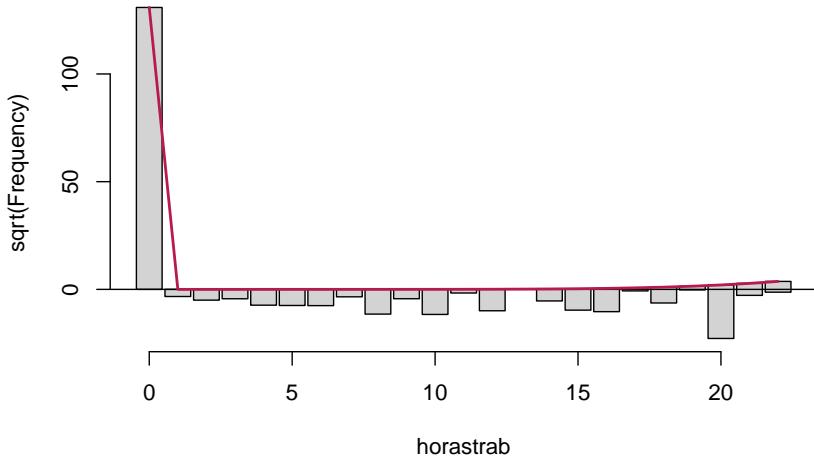
São Paulo 2T2022

```

## 
## Call:
## hurdle(formula = horastrab ~ sexo + idade + anosest, data = h_sp_2)
## 
## Pearson residuals:
##      Min     1Q Median     3Q    Max 
## -2.4124 -0.7684 -0.2861  0.7899  6.2094 
## 
## Count model coefficients (truncated poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 3.647e+00 6.119e-03 595.987 < 2e-16 ***
## sexo masculino 1.022e-01 2.363e-03  43.240 < 2e-16 *** 
## idade       -5.628e-04 8.611e-05 -6.535 6.35e-11 *** 
## anosest      1.994e-03 3.214e-04   6.202 5.57e-10 *** 
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.4916917 0.0432749 -57.578 <2e-16 *** 
## sexo masculino 0.8465540 0.0244407  34.637 <2e-16 *** 
## idade        0.0004596 0.0005911   0.778  0.437  
## anosest      0.2211700 0.0029236   75.650 <2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1 
## 
## Number of iterations in BFGS optimization: 11 
## Log-likelihood: -1.049e+05 on 8 Df

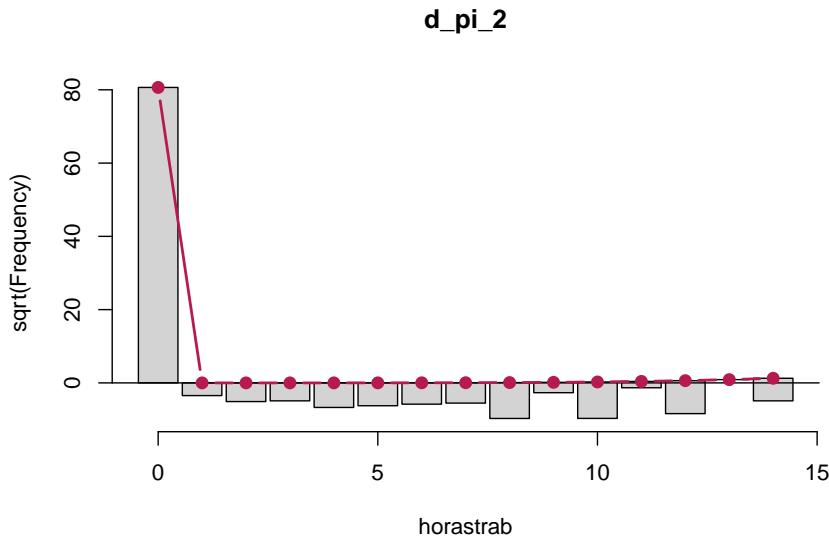
```

d_sp_2



Piauí 2T2022

```
##  
## Call:  
## hurdle(formula = horastrab ~ sexo + idade + anosest, data = h_pi_2)  
##  
## Pearson residuals:  
##      Min       1Q   Median       3Q      Max  
## -3.2368 -0.6797 -0.3846  0.6573 13.5208  
##  
## Count model coefficients (truncated poisson with log link):  
##                 Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  3.1421292  0.0129384 242.852  <2e-16 ***  
## sexomasculino 0.1497733  0.0056010  26.741  <2e-16 ***  
## idade        0.0020132  0.0002050   9.819  <2e-16 ***  
## anosest      0.0189612  0.0006034  31.426  <2e-16 ***  
## Zero hurdle model coefficients (binomial with logit link):  
##                 Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -3.260596   0.083140 -39.22  <2e-16 ***  
## sexomasculino 1.037904   0.046720  22.21  <2e-16 ***  
## idade        0.016745   0.001166  14.36  <2e-16 ***  
## anosest      0.212293   0.005050  42.04  <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Number of iterations in BFGS optimization: 16  
## Log-likelihood: -3.245e+04 on 8 Df
```



Resultados

Odds	SP_1	PI_1	SP_2	PI_2
(Intercept)	40.0708	23.4170	38.3436	23.1531
sexomasculino	1.1406	1.2601	1.1076	1.1616
idade	0.9996	1.0023	0.9994	1.0020
anosest	0.9981	1.0191	1.0020	1.0191

Primeiramente, os modelos Double Hurdle apresentam *outputs* no formato de regressão de Poisson para atividades diferentes de zero e no formato Binomial Negativa para atividade zero. Nesse sentido, a interpretação dos resultados é dada pela exponencialização dos coeficientes estimados, sendo eles uma razão de chances. Logo, para valores acima de 1, temos uma maior probabilidade do evento horas de trabalho acontecer e tomamos o sexo masculino como referência.

Para valores positivos no 2T2012, temos que o sexo masculino tem maiores chances de ofertar horas de trabalho frente as mulheres e que para o PI essa probabilidade é maior do que para SP. Na medida que a idade avança e com maior escolaridade, as mulheres tendem a ofertar mais horas que homens em SP, ao mesmo tempo que para o PI tal dinâmica se inverte, são os homens que trabalham mais.

Quando se trata do segundo período, note-se uma redução na chance de oferta de horas para homens em SP e PI, sendo que a probabilidade para o PI se manteve acima da paulista. Para idade, a dinâmica foi semelhante, ambas as UFs acusaram uma ligeira queda nas chances e o PI segue com probabilidade maiores que SP. No que se refere aos anos de estudo, para SP houve uma inversão na probabilidade, nesse período a chance dos homens apresentarem mais horas ofertadas superou as chances das mulheres, enquanto o PI manteve a maior probabilidade para homens e seguiu acima de SP.

Ademais, em termos validação do modelo, utilizou-se o *rootogram* que, basicamente, demonstra a qualidade de previsão do modelo para cada ponto da variável, em que podemos pensar no *rootogram* como a relação entre as contagens ajustadas - se uma barra não atingir a linha zero, o modelo prediz uma determinada caixa de contagem e, se a barra exceder a linha zero, ela sob prevê. Nesse sentido, a alta frequência de zeros em todos os modelos indicaram uma sob previsão, ao passo que para as demais contagens os valores foram sub previsto, porém, com erros mais comportados.

Odds	SP_1	PI_1	SP_2	PI_2
(Intercept)	0.0722	0.0579	0.0828	0.0384
sexomasculino	2.5425	3.2075	2.3316	2.8233
idade	1.0122	1.0281	1.0005	1.0169
anoest	1.2292	1.1825	1.2475	1.2365

Pelo lado dos resultados nulos, o primeiro período acusou de forma generalizada que os homens tem maiores chances de não ofertarem horas de trabalho que as mulheres. Tal probabilidade é maior no primeiro período, contudo, o PI apresenta uma redução significativa para o 2T2022. No mais, o PI supera SP nos dois períodos para idade, com ligeira redução de chances no 2T2022. Enquanto SP supera o PI na escolaridade, com pequeno avanço nas chances para ambos os Estados.

2. Modelos de correção de seleção amostral

2.1 Modelo de Heckman

Os modelos estimados via MQO muitas vezes sofrem de viés amostral e/ou da omissão de regressores relevantes, algo que viola os pressupostos de Gauss-Markov e, com isso, temos que os estimadores da regressão não são BLUE (Best Linear Unbiased Estimator), dado a inconsistência deles. Desta maneira, o modelo em dois estágios de Heckman é uma opção para contornar tais problemas, uma vez que incluímos uma equação adicional que liga as covariadas de interesse com a equação principal. Sendo assim, a ideia central do estimador Heckit é que os fatores não observáveis irão afetar tanto a variável de interesse quanto as probabilidades de seleção amostral, tendo em vista que tais informações devem estar contidas no resíduo da equação de seleção. Logo, a estimação consiste, em seu primeiro estágio, a modelagem via Probit para encontrar o valor esperado dos resíduos de seleção da parte não observável e, no segundo estágio, deve ser feita uma estimação via MQO com os resíduos estimados no primeiro estágio, para que assim seja possível encontrar estimadores consistentes.

2.11 Aplicação

Para tal método utilizou-se os microdados da PNAD para os mesmos períodos das aplicações acima e para duas UFs. A motivação por trás do modelo surge para tentar estimar a renda daqueles que não trabalham, vulgo salário de reserva, uma vez que a renda é representada na amostra por aqueles que trabalham e os demais são subrepresentados. Para tal, o primeiro estágio foi estimado a partir de um vetor binário que explicita quem trabalha por 1 e quem não trabalha por 0, onde assumiu-se quem trabalha aquele cuja renda era diferente de zero, e como covariadas incluiu-se os anos de estudo (anoest), idade e número de filhos (nfilhotot). Para o segundo estágio, estimou-se o logaritmo da renda contra os anos de estudo e idade. Para efeitos de comparação, foi estimado o mesmo modelo do segundo estágio por meio dos Mínimos Quadrados Ordinários. Por fim, a escolha das UFs foi pautada no rendimento médio real do trabalho principal, efetivamente recebido no mês, para o maior e menor valor recebido no 2T2012, sendo assim os modelos são referentes a São Paulo (SP) e Maranhão (MA)

2.12 Análise Descritiva

Nesta seção será apresentado as informações descritivas e o comportamento das variáveis para as UFs e períodos.

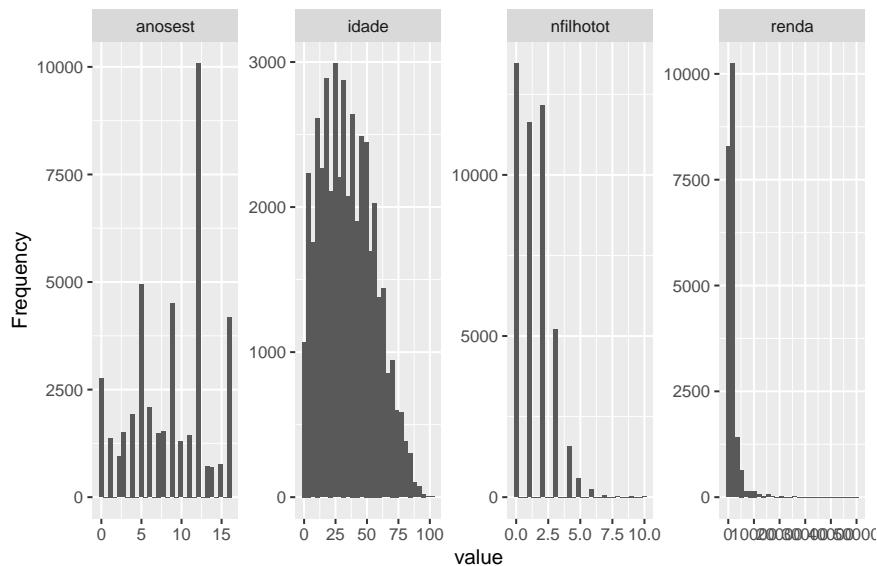
São Paulo 2T2012

```
##      idade      anoest      nfilhotot      renda
##  Min.   : 0.00  Min.   : 0.000  Min.   : 0.000  Min.   : 1
```

```

##   1st Qu.: 17.00    1st Qu.: 5.000    1st Qu.: 0.000    1st Qu.: 700
##   Median : 33.00    Median : 9.000    Median : 1.000    Median : 1000
##   Mean    : 34.58    Mean   : 8.598    Mean   : 1.408    Mean   : 1658
##   3rd Qu.: 50.00    3rd Qu.:12.000    3rd Qu.: 2.000    3rd Qu.: 1665
##   Max.    :102.00    Max.   :16.000    Max.   :10.000    Max.   :50000
##   NA's     :2726          NA's     :23731
##
##      trab
##   Min.    :0.0000
##   1st Qu.:0.0000
##   Median :0.0000
##   Mean   :0.4724
##   3rd Qu.:1.0000
##   Max.   :1.0000
##   NA's

```

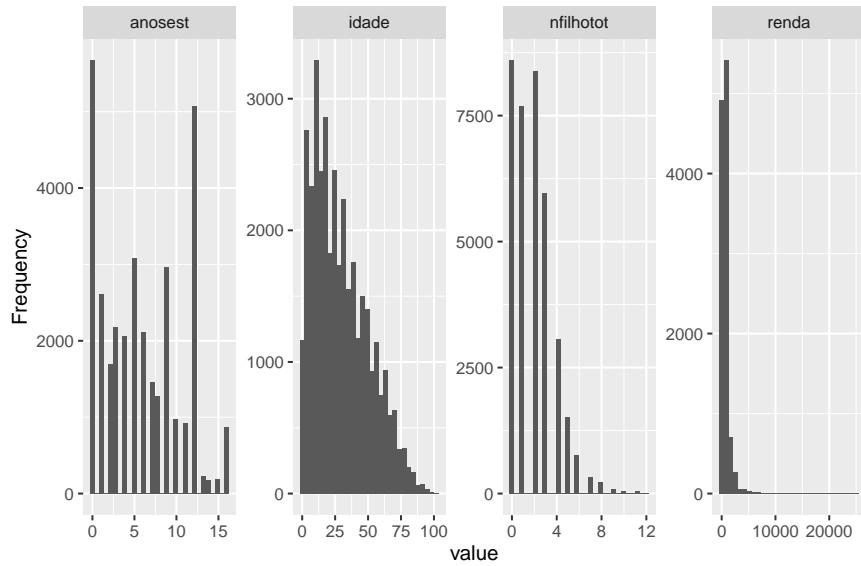


Maranhão 2T2012

```

##      idade      anosest      nfilhotot      renda
##   Min.    : 0.00    Min.   : 0.00    Min.   : 0.000    Min.   : 8.0
##   1st Qu.: 12.00   1st Qu.: 2.00   1st Qu.: 1.000    1st Qu.: 200.0
##   Median : 25.00   Median : 5.00   Median : 2.000    Median : 600.0
##   Mean   : 29.22   Mean   : 5.93   Mean   : 1.985    Mean   : 698.9
##   3rd Qu.: 43.00   3rd Qu.:10.00   3rd Qu.: 3.000    3rd Qu.: 800.0
##   Max.   :102.00   Max.   :16.00   Max.   :12.000    Max.   :25000.0
##   NA's    :3190          NA's     :25228
##
##      trab
##   Min.    :0.0000
##   1st Qu.:0.0000
##   Median :0.0000
##   Mean   :0.3136
##   3rd Qu.:1.0000
##   Max.   :1.0000
##   NA's

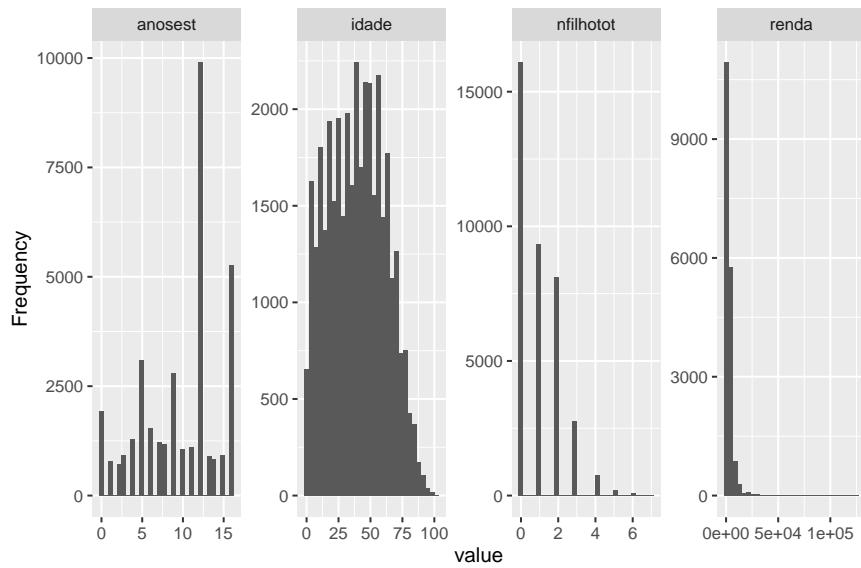
```



São Paulo 2T2022

```

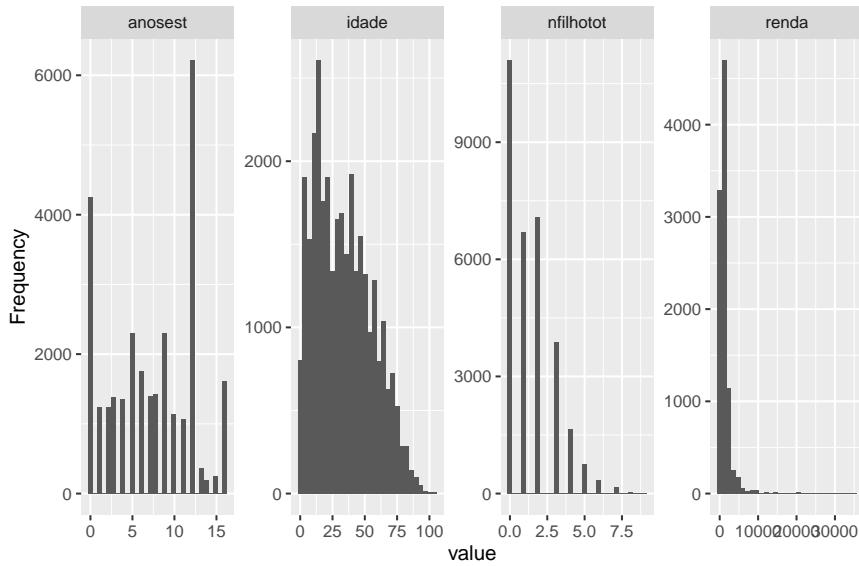
##      idade          anosest        nfilhotot       renda
##  Min.   : 0.00   Min.   :0.0000   Min.   :0.000   Min.   : 30
##  1st Qu.:21.00   1st Qu.: 6.000   1st Qu.:0.000   1st Qu.:1221
##  Median :39.00   Median :12.000   Median :1.000   Median :1900
##  Mean   :39.21   Mean   : 9.612   Mean   :1.034   Mean   :2999
##  3rd Qu.:56.00   3rd Qu.:12.000   3rd Qu.:2.000   3rd Qu.:3000
##  Max.   :102.00   Max.   :16.000   Max.   : 7.000   Max.   :125000
##           NA's   :1881                 NA's   :19216
##      trab
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4859
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```



Maranhão 2T2022

```

##      idade          anosest        nfilhotot       renda
##  Min.   : 0.00   Min.   :0.000   Min.   :0.000   Min.   : 10
##  1st Qu.:15.00   1st Qu.:3.000   1st Qu.:0.000   1st Qu.: 500
##  Median :31.00   Median :7.000   Median :1.000   Median :1212
##  Mean   :33.36   Mean   :7.189   Mean   :1.473   Mean   :1370
##  3rd Qu.:49.00   3rd Qu.:12.000  3rd Qu.:2.000   3rd Qu.:1500
##  Max.   :104.00  Max.   :16.000  Max.   :9.000   Max.   :35000
##           NA's   :2277           NA's   :22005
##      trab
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3077
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```



Com relação ao primeiro período, temos que as variáveis aleatórias para SP, na média, apresentaram números maiores do que o MA, com exceção ao número de filhos cujo valor médio é idêntico. Pelo lado categórico, aqueles que não trabalham representaram 52,7% e 68,6% das amostras de SP e do MA, respectivamente. Para o segundo período analisado, temos que a dinâmica passada se manteve, ou seja, as variáveis aleatórias na média apresentaram valores maiores para SP em comparação ao MA, à exceção do número de filhos. No tocante à parte dicotômica, a proporção dos que não trabalham em SP caiu para 51,4% da amostra e no MA aumentou para 69,2%.

2.13 Modelo

Os *outputs* dos modelos seguem abaixo, bem como a comparação com os respectivos modelos estimados somente via MQO:

São Paulo 2T2012

```
## -----
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 42255 observations (21005 censored and 21250 observed)
## 10 free parameters (df = 42246)
## Probit selection equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.290637   0.024890 -51.854 <2e-16 ***
## idade        0.006656   0.000372  17.892 <2e-16 ***
## anosest      0.119372   0.001480  80.669 <2e-16 ***
## nfilhotot    0.013014   0.005626   2.313  0.0207 *
## Outcome equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.712504   0.372750 -7.277 3.47e-13 ***
## idade        0.032387   0.001516  21.356 < 2e-16 ***
## anosest      0.469638   0.018226  25.767 < 2e-16 ***
## Multiple R-Squared:0.2828,   Adjusted R-Squared:0.2827
## Error terms:
##           Estimate Std. Error t value Pr(>|t|)
## invMillsRatio 5.2624     0.1989   26.45 <2e-16 ***
```

```

## sigma          4.0409      NA      NA      NA
## rho           1.3023      NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## -----
## 
## 
## =====
##                               Dependent variable:
## -----
##                               log(renda)
## OLS                      selection
## (1)                      (2)
## -----
## idade          0.016***    0.032*** 
##                  (0.0004)   (0.002)
## 
## anoest         0.097***    0.470*** 
##                  (0.001)    (0.018)
## 
## Constant       5.402***   -2.713*** 
##                  (0.022)   (0.373)
## 
## -----
## Observations     21,250        42,255
## R2              0.242
## Adjusted R2     0.242
## rho             1.302
## Inverse Mills Ratio      5.262*** (0.199)
## Residual Std. Error     0.699 (df = 21247)
## F Statistic       3,387.015*** (df = 2; 21247)
## =====
## Note:           *p<0.1; **p<0.05; ***p<0.01

```

Maranhão 2T2012

```

## -----
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 33563 observations (22038 censored and 11525 observed)
## 10 free parameters (df = 33554)
## Probit selection equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.6538273  0.0262370 -63.034 < 2e-16 ***
## idade        0.0184188  0.0004298  42.859 < 2e-16 ***
## anoest        0.0971741  0.0016945  57.347 < 2e-16 ***
## nfilhotot    0.0142424  0.0046757   3.046  0.00232 **
## Outcome equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.99931   1.00787  -2.976  0.00292 **
## idade        0.06071   0.00664   9.142 < 2e-16 ***
## anoest        0.38016   0.03495  10.878 < 2e-16 ***
## Multiple R-Squared:0.2963,   Adjusted R-Squared:0.2962
## Error terms:
##           Estimate Std. Error t value Pr(>|t|)

```

```

## invMillsRatio 4.0709      0.5237    7.773 7.89e-15 ***
## sigma          3.4326      NA        NA        NA
## rho            1.1860      NA        NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## -----
##
## =====
##                               Dependent variable:
## -----
##                               log(renda)
## OLS                      selection
## (1)                      (2)
## -----
## idade                     0.013***      0.061*** 
##                           (0.001)       (0.007)
## 
## anoest                    0.121***      0.380*** 
##                           (0.002)       (0.035)
## 
## Constant                  4.615***      -2.999*** 
##                           (0.033)       (1.008)
## 
## -----
## Observations              11,525        33,563
## R2                        0.283
## Adjusted R2                0.283
## rho                       1.186
## Inverse Mills Ratio       4.071*** (0.524)
## Residual Std. Error       0.880 (df = 11522)
## F Statistic                2,277.988*** (df = 2; 11522)
## -----
## Note:                      *p<0.1; **p<0.05; ***p<0.01

```

São Paulo 2T2022

```

## -----
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 35496 observations (17335 censored and 18161 observed)
## 10 free parameters (df = 35487)
## Probit selection equation:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.1735638  0.0279215 -42.031 <2e-16 ***
## idade       -0.0005497  0.0003975  -1.383   0.167
## anoest       0.1273095  0.0016230   78.440 <2e-16 ***
## nfilhotot   -0.0085086  0.0071975  -1.182   0.237
## Outcome equation:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.334359      NaN      NaN      NaN
## idade       0.009519  0.001092   8.717 <2e-16 ***
## anoest       0.519086      NaN      NaN      NaN
## Multiple R-Squared:0.2561,   Adjusted R-Squared:0.256
## Error terms:

```

```

##           Estimate Std. Error t value Pr(>|t|)
## invMillsRatio   5.363      NaN      NaN      NaN
## sigma          4.082      NA      NA      NA
## rho            1.314      NA      NA      NA
## -----
## 
## =====
##             Dependent variable:
## -----
##                   log(renda)
##                   OLS      selection
##                   (1)      (2)
## -----
## idade           0.014***    0.010***  

##                 (0.0004)    (0.001)
## 
## anoest          0.102***    0.519  

##                 (0.001)
## 
## Constant        5.873***   -2.334  

##                 (0.027)
## 
## -----
## Observations     18,161      35,496
## R2              0.214
## Adjusted R2     0.214
## rho             1.314
## Inverse Mills Ratio      5.363
## Residual Std. Error   0.730 (df = 18158)
## F Statistic       2,467.851*** (df = 2; 18158)
## =====
## Note:           *p<0.1; **p<0.05; ***p<0.01

```

Maranhão 2T2022

```

## -----
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 29510 observations (19728 censored and 9782 observed)
## 10 free parameters (df = 29501)
## Probit selection equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.7349981  0.0301859 -57.477 <2e-16 ***
## idade        0.0132317  0.0004657  28.414 <2e-16 ***
## anoest        0.1069280  0.0017753  60.230 <2e-16 ***
## nfilhotot   -0.0071180  0.0060643  -1.174   0.241
## Outcome equation:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.519237   1.085444  -3.242  0.00119 **
## idade        0.053037   0.005266  10.071 < 2e-16 ***
## anoest        0.427899   0.040441  10.581 < 2e-16 ***
## Multiple R-Squared:0.2518,   Adjusted R-Squared:0.2516
## Error terms:
##           Estimate Std. Error t value Pr(>|t|)

```

```

## invMillsRatio 4.3168      0.5238     8.241    <2e-16 ***
## sigma          3.6373       NA        NA        NA
## rho            1.1868       NA        NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## -----
##
## -----
##                               Dependent variable:
## -----
##                               log(renda)
##           OLS             selection
##           (1)              (2)
## -----
## idade            0.016***      0.053*** 
##                   (0.001)        (0.005)
## 
## anoest          0.113***      0.428*** 
##                   (0.002)        (0.040)
## 
## Constant        5.039***      -3.519*** 
##                   (0.040)        (1.085)
## 
## -----
## Observations    9,782          29,510
## R2              0.240
## Adjusted R2     0.240
## rho              1.187
## Inverse Mills Ratio   4.317*** (0.524)
## Residual Std. Error  0.890 (df = 9779)
## F Statistic     1,547.828*** (df = 2; 9779)
## -----
## Note:           *p<0.1; **p<0.05; ***p<0.01

```

Resultados

```

## -----
##                               Dependent variable:
## -----
##                               log(renda)
##           SP 1          MA 1          SP 2          MA 2
##           (1)          (2)          (3)          (4)
## -----
## idade            0.032***      0.061***     0.010***    0.053*** 
##                   (0.002)        (0.007)        (0.001)        (0.005)
## 
## anoest          0.470***      0.380***     0.519        0.428*** 
##                   (0.018)        (0.035)          (0.040)
## 
## Constant        -2.713***     -2.999***    -2.334        -3.519*** 
##                   (0.373)        (1.008)          (1.085)
## 
## -----

```

```

## Observations      42,255        33,563        35,496        29,510
## rho              1.302         1.186         1.314         1.187
## Inverse Mills Ratio 5.262*** (0.199) 4.071*** (0.524) 5.363   4.317*** (0.524)
## =====
## Note:                                     *p<0.1; **p<0.05; ***p<0.01

```

Os valores acima correspondem as estimações do 2º estágio e, primeiramente, cabe comentar sobre o nível de significância da inversa de Mills. O fator de ajuste de Mills, basicamente, é calculado a partir da estimação do 1º estágio quando existe correlação entre os dados observáveis e não observáveis, nesse sentido a inversa de Mills entra como fator de correção da heterocedasticidade no 2º estágio e produz estimadores consistentes e eficientes. Portanto, quando aceitamos a hipótese nula do teste T-student para a inversa de Mills entende-se que o modelo não sofre de viés amostral. No que tange aos resultados, o modelo para SP_2 acusou que não há viés de seleção, sendo assim ele perde o efeito de comparação com os demais modelos acima. O modelo estimado via MQO seria uma opção para o exercício, porém os valores dos parâmetros parecem divergir consideravelmente, isso em relação ao modelo do primeiro período, para tirar qualquer *insight*, além de que nenhuma validação do modelo foi feita, por conta disso, o trabalho descartou o uso dele.

A interpretação dos resultados é feita por meio de semi-elasticidades, tendo em vista que são modelos log-nível. Logo, para o primeiro período, temos que SP tem um impacto de médio de 3,2% sobre a renda quando aumentamos a idade em uma unidade, ao passo que MA o impacto tende a ser maior. Pelo lado da escolaridade, os efeitos são maiores para ambas as UFs, sendo que SP apresenta maiores influências na média. Quando tratamos do segundo período, observou-se a evolução do impacto médio dos regressores sobre a renda no MA. Por fim, é interessante notar a diferença nos parâmetros estimados nas duas abordagens, MQO e Heckman, onde de fato havia viés de seleção na amostra das UFs, uma vez que a razão da inversa de Mills se deu significativa para ampla maioria dos modelos analisados.

Referências

- Cameron, A. C. and P. K. Trivedi (2005), Microeometrics: Methods and Applications, Cambridge University Press, New York, Chapters 14-16
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39(5), 829–844
- TOBIN, James. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, p. 24-36, 1958.