Fecundidade em São Paulo

Maria Cruz e Guilherme Davi

1 Introdução

O objetivo deste trabalho é usar conhecimentos da disciplina de Econometria 1 para analisar a fecundidade da cidade de São Paulo. Fecundidade representa o potencial reprodutivo, normalmente abordado para determinada região. A taxa de fecundidade é o número médio de filhos por mulher em idade reprodutiva (15 a 49 anos).

2 Dados

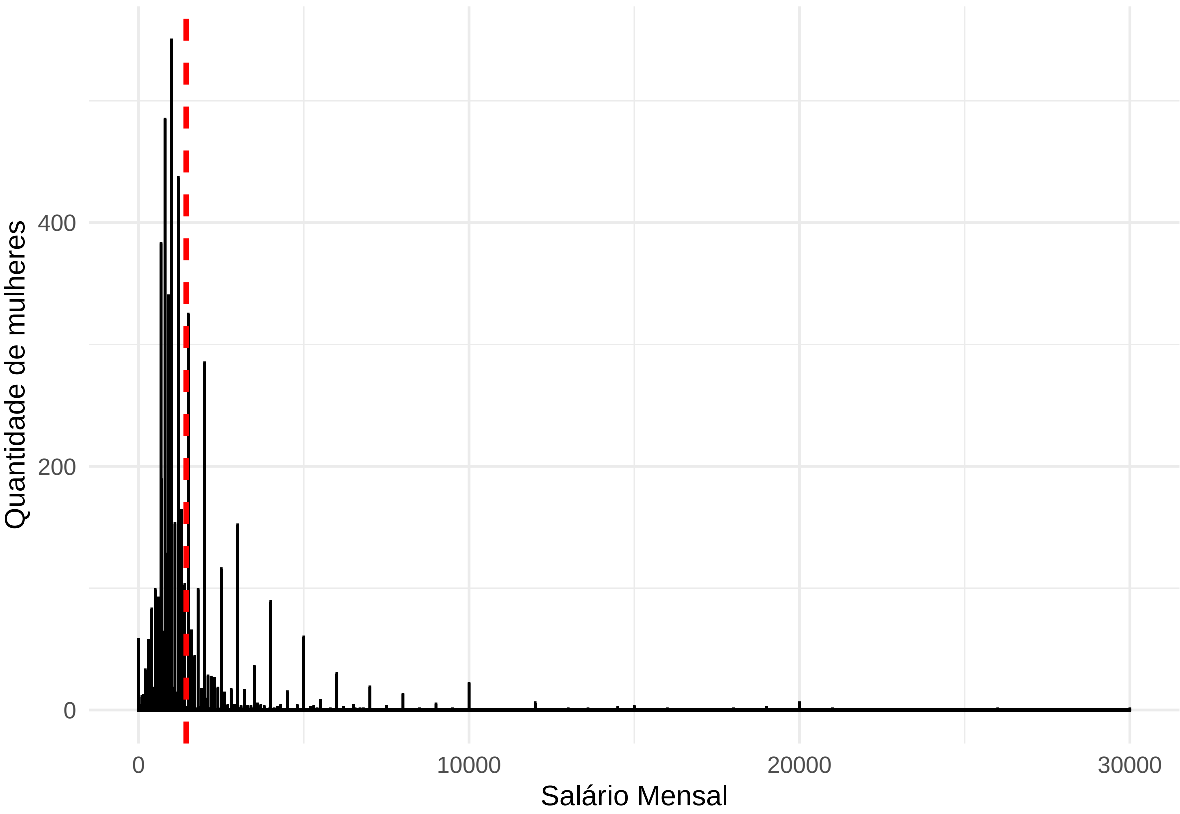
A base possui 8146 observações e 13 variáveis que abordam indicadores pessoais como idade das mulheres, raça, aspectos de inserção ao mercado de trabalho (como atuação no trabalho doméstico, carga horária, salário e educação), e por fim, variáveis de quantidade de filhos por mulher.

A faixa etária da base vai de 7 a 93 anos de idade. Como esse grupo não é considerado nas análises de fecundidade de um determinado local, pois não está inserido na faixa de idade reprodutiva, consideramos dois modelos por recorte de faixa etária: a base completa, e a de idade reprodutiva.

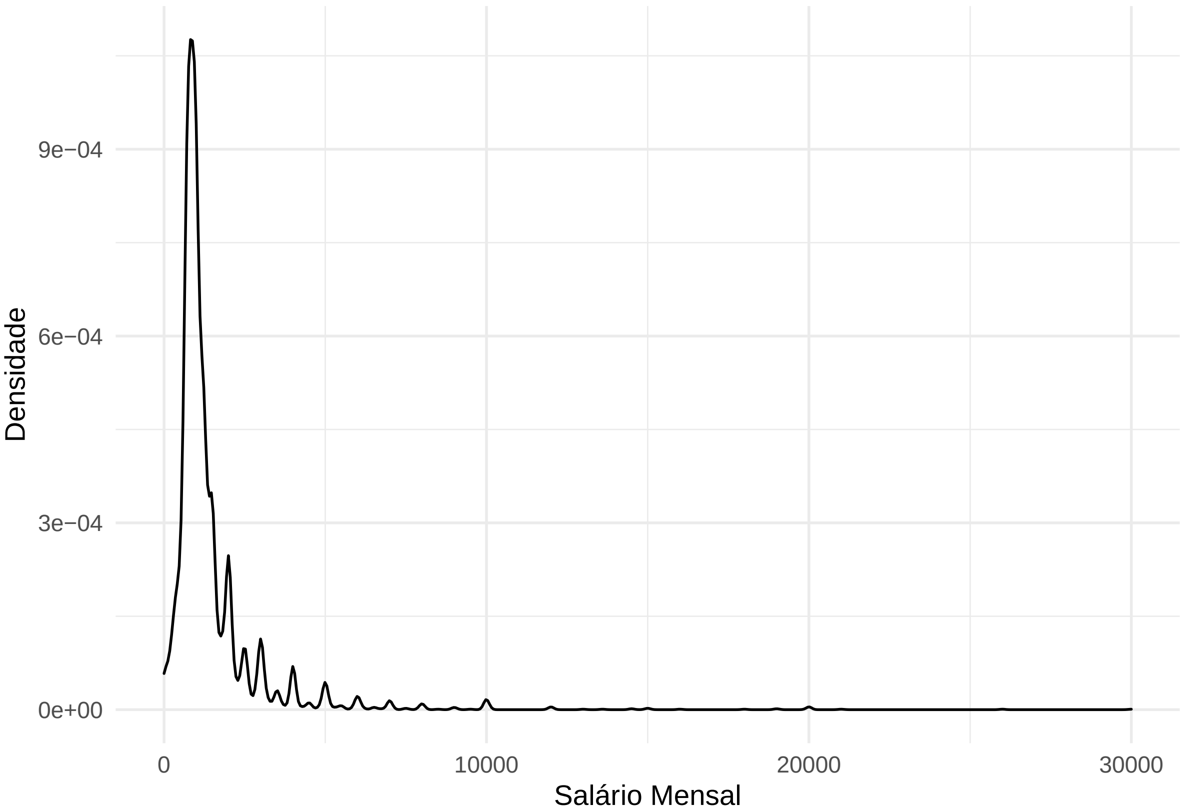
A ideia é que a população de mulheres mais novas que 15 anos e mais velhas que 49 anos não são representativas do potencial de reprodução do local, afinal nessas idades questões comportamentais e intrínsecas a sua saúde reprodutiva afetam a possibilidade de conceber um filho.

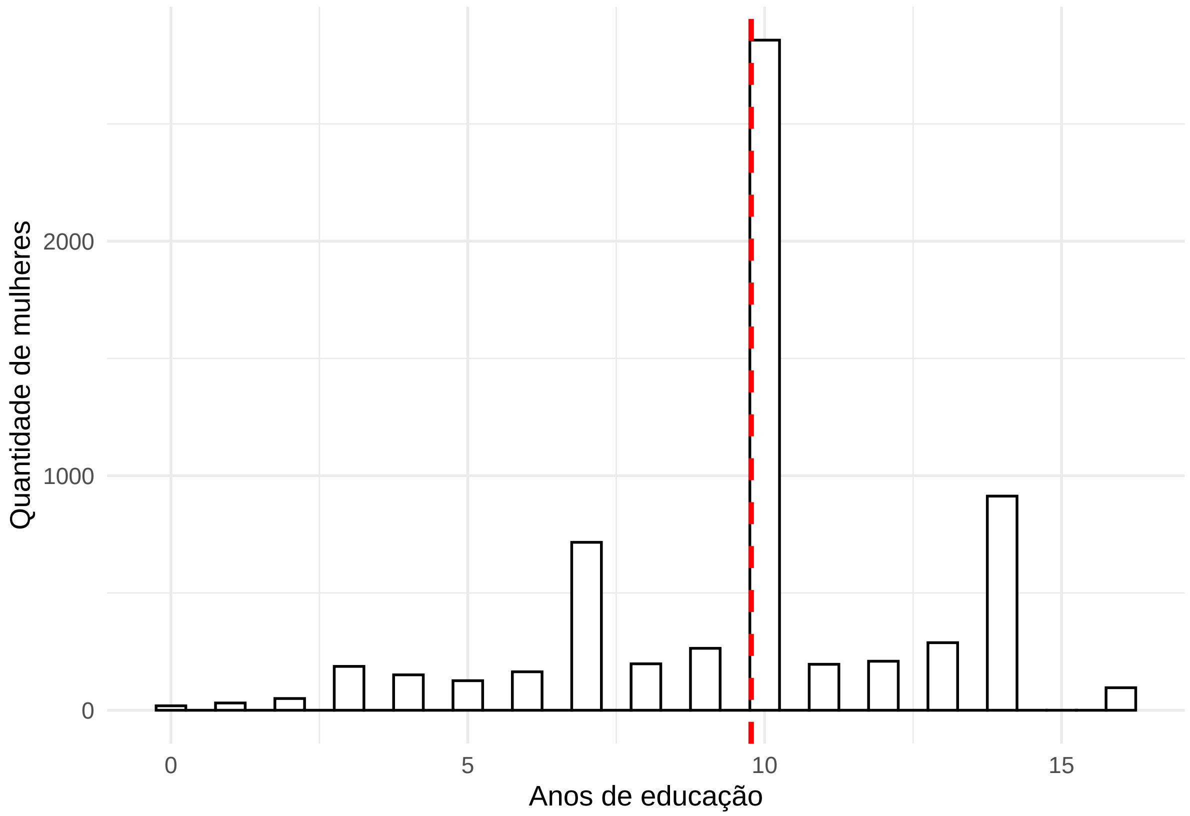
Assim, para a base com amostra de mulheres em idade reprodutiva temos 6465 observações. Dentre essas, temos a distribuição de salário e escolaridade da base:

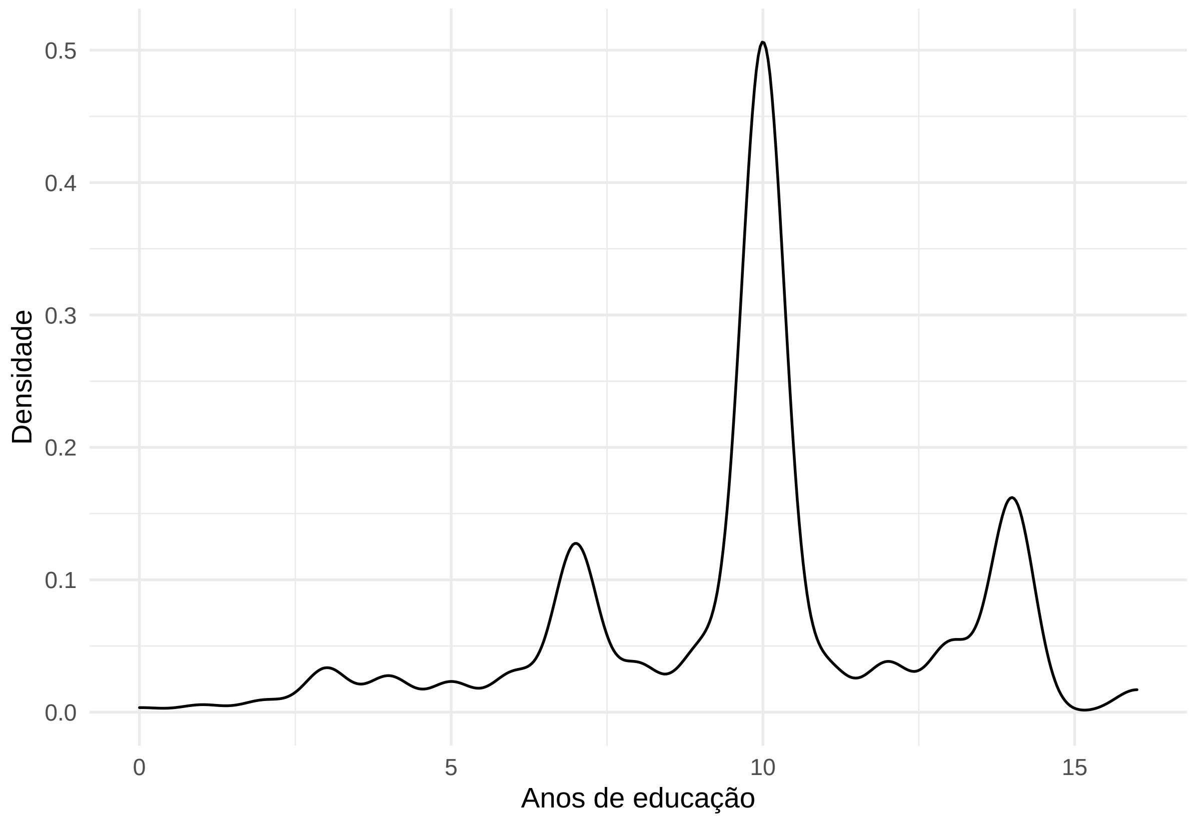
```
## Warning: Removed 374 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 374 rows containing non-finite values (stat_density).
```







Por fim, temos a tabela de indicadores descritivos abaixo:

Análise Descritiva			
	Média	Desvio Padrão	Mediana
Salário mensal	1439,3	1602,4	1000
Anos de educação	9,8	2,9	10
Idade	32,8	9,2	33
Filhos	1,2	1,4	1

3 Literatura

Sabe-se que a fecundidade depende de variáveis de escolha e condição das mulheres. A Pesquisa Nacional de Demografia e Saúde (PNDS) comparou níveis de fecundidade entre 1996 e 2006 para o Brasil e, como evidência, é possível observar que os níveis de fecundidade diminuíram em média para o país. Esse fenômeno esteve presente para mulheres entre 20 e 24 anos. Entretanto, para mulheres mais novas que 20 anos, tais níveis não caíram. Além disso, a pesquisa apresenta que houve uma diminuição na idade do primeiro filho para a região nordeste, que possui maiores níveis de fecundidade para mulheres com menos de 20 anos.

Este fenômeno de queda da fecundidade é denominado transição da fecundidade. As principais mudanças que acarretam esse processo são inserção feminina no mercado de trabalho, maior acesso à educação, maior taxa de urbanização, maior acesso a contraceptivos.

Na literatura de fecundidade, os primeiros estudos são de Becker (1991), em que se propõe a ideia de maximizar o número de crianças dada restrição econômica. Posteriormente Black (2008) questiona se crianças são como bens normais, afinal mais renda aumenta fecundidade. Na realidade brasileira, Ponczek and Souza (2012) discutem qualidade vs. número de crianças no Brasil, ou seja, como a quantidade de crianças pode afetar no investimento de educação, saúde, entre outros. Nesse último questionamento existe um debate de que o primeiro filho tende a ter mais investimento dos pais quando comparados aos outros filhos.

Dessa forma, compreendemos que dada as variáveis disponíveis, poderíamos discutir a relação de fecundidade com renda -no nosso caso, somente salário da mulher- e escolaridade da mulher.

4 Questão 1

4.1 Questão A

Dado o objetivo de realizar uma análise sistemática dos dados de fecundidade utilizando a estimação por métodos de regressão linear, é conveniente produzir uma primeira aproximação ao problema com a estimação de uma regressão simples, sem transformações profundas nas variáveis da base.

Nesse caso, uma especificação de modelo pertinente e interessante (mas não necessariamente correta), é:

$$\text{Quantidade de filhos} = \beta_0 + \beta_1 \text{Salário mensal} + u$$

Assumindo que a criação dos filhos é custosa e que o custo não é distribuído uniformemente entre os responsáveis pela criança na família (talvez por fatores sociais), sendo relativamente maior para as mulheres, poderíamos esperar que salários maiores estejam associados a uma menor quantidade de filhos. Nesse caso, obteríamos um β_1 negativo.

4.2 Questão B

Pelo modelo estimado, temos um intercepto $\hat{\beta}_0$ de 1,29 e um $\hat{\beta}_1$ de -0,00005. Pelo coeficiente do intercepto, temos que, para mulheres com um salário mensal de R$0,00, o número esperado de filhos é de 1,29. Já pelo coeficiente da variável independente, temos que, para cada aumento de R$1,00 no salário mensal das mulheres, estima-se que o número de filhos caia em -0,00005. Para colocar esse número em perspectiva, devemos levar em consideração que, na tentativa de encontrar o efeito causal da renda sobre a fecundidade, estamos muito mais interessados no efeito de centenas ou milhares de reais sobre o número de filhos do que pequenas diferenças unitárias. Isso significa que um aumento de R$1000,00 no salário mensal está associado com uma queda estimada de -0.05 no número de filhos.

Regressão simples do impacto do salário mensal sobre a quantidade de filhos

	Quantidade de Filhos
Intercepto	1.290*** (0.02378)
Salário mensal	-0.00005*** (0.00001104)
N	6,314
R-quadrado	0.003358
R-quadrado ajustado	0.003194
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	

Apesar dos coeficientes serem significativos ao nível de 0,1% e do $\hat{\beta}_1$ estimado ter sinal negativo (como esperávamos), vale ressaltar que o modelo estimado é muito limitado. A forma especificada é simples demais e injustificada diante de outras especificações e das hipóteses necessárias para a estimação não-viesada dos coeficientes. Isso será discutido nas questões posteriores.

4.3 Questão C

Sabemos que, em um modelo de regressão simples, o coeficiente $\hat{\beta}_1$ da variável independente do salário mensal é apenas o coeficiente de correlação amostral entre a quantidade de filhos e o salário mensal multiplicada pela razão dos desvios-padrão amostrais de cada variável. A inferência de causalidade, a partir daí, requer diferentes hipóteses sobre como essas variáveis são correlacionadas e interagem com os fatores não-observados no modelo: se a nossa amostra for de fato aleatória e a esperança do erro no modelo for a mesma, dado o salário mensal, o nosso $\hat{\beta}_1$ será também um estimador do efeito causal médio.

Essa última hipótese é muito forte e não deve ser assumida sem uma análise cuidadosa. Se o modelo for especificado incorretamente, o que ocorre quando alguma variável é omitida ou quando o modelo verdadeiro apresenta alguma outra relação não-linear com a variável utilizada no modelo, essa hipótese irá falhar e a correlação obtida será espúria e/ou viesada, não representando uma relação causal.

Isso significa, dentro do contexto de nossa análise, que se houver uma correlação entre o salário mensal e a variável idade das mulheres — contida, por definição, nos fatores não-observáveis do modelo simples estimado — a ausência dessa variável no modelo faz com que o $\hat{\beta}_1$ estimado seja viesado.

Com efeito, é mais do que razoável esperar uma relação entre o salário mensal das mulheres e sua idade por diferentes motivos: mulheres mais velhas podem ter mais experiência profissional (que por sua vez está associada a um salário mensal maior); mulheres com renda maior podem ter maior acesso à saúde, educação e segurança, vivendo, em média, mais anos; essas relações poderiam ainda ser acentuadas em nossa análise pela necessidade de limitar a amostra para mulheres em idade reprodutiva, pois tendências de queda do salário após a aposentadoria não seriam observadas.

4.4 Questão D

Na direção do que foi discutido na última questão, podemos ter noção do viés se assumirmos, um pouco mais realisticamente, que o modelo verdadeiro é dado por:

$$\text{Quantidade de filhos} = \beta_0 + \beta_1 \text{Salário mensal} + \beta_2 \text{Idade} + u$$

Nesse caso simples, o viés causado pela omissão da variável de idade na regressão é dado por $\beta_2 * \hat{\delta}_{1,2}$, no qual β_2 é o parâmetro verdadeiro do modelo enunciado e $\hat{\delta}_{1,2}$ é o coeficiente estimado da regressão do salário mensal na idade. Como esperamos um β_2 positivo (mulheres mais velhas têm, em média, mais filhos) e um $\hat{\delta}_{1,2}$ também positivo (relação positiva entre idade e renda), o sinal do viés seria, também, positivo. Disso, resultaria que o nosso modelo simples estimado subestima a relação negativa entre o salário mensal e a quantidade de filhos.

4.5 Questão E

Para uma regressão simples, gostaríamos de testar

$$H_0 : \beta_0 = 0 H_1 : \beta_0 \neq 0$$

e

$$H_0 : \beta_1 = 0 H_1 : \beta_1 \neq 0$$

Temos:

$$t_0 = \frac{\hat{\beta}_0}{ep(\hat{\beta}_0)} = \frac{1.290007}{0.0238} = 54.201$$

e

$$t_1 = \frac{\hat{\beta}_1}{ep(\hat{\beta}_1)} = \frac{5 * 10^{-5}}{1.104 * 10^{-5}} = -4.530026$$

Como $|t_0| > 2,58$ e $|t_1| > 2,58$, ambos coeficientes são estatisticamente significantes ao nível de 1%, o que já poderia ser visto na tabela.

4.6 Questão F

Para um modelo de regressão simples, as hipóteses clássicas são: - Linearidade nos parâmetros; - Amostragem aleatória; - Variabilidade em x; - Exogeneidade do x ou Média condicional zero; - Homocedasticidade; - Normalidade dos erros (para realização de testes de hipótese).

Para o nosso modelo de regressão simples (no qual x é a variável salário mensal), as três primeiras hipóteses não apresentam problemas: não há motivo para acreditar que o modelo não deveria ser linear nos parâmetros, a amostra da população de interesse é aleatória e apresenta variação na variável salário mensal.

Quanto à hipótese da média condicional zero, como foi discutido anteriormente, temos bons motivos para acreditar que a variável salário mensal é correlacionada com pelo menos um dos fatores não-observados que compõem os erros. Um modelo de regressão linear múltipla, possivelmente com relações não-lineares entre as variáveis (como funções logarítmicas ou quadráticas), seria necessário para especificações de modelos mais realistas. Vimos como a inclusão de uma outra variável, idade, no modelo verdadeiro atesta que a relação negativa entre o salário mensal e a quantidade de filhos está subestimada no coeficiente estimado $\hat{\beta}_1$. Mais adiante, são discutidas ainda outras especificações, mais razoáveis do que a nossa aproximação inicial.

A hipótese da homocedasticidade, apesar de não ser necessária para garantir o não-viés dos estimadores, facilita a estimação de suas variâncias, tendo impacto direto nos testes de hipóteses realizados sobre os parâmetros. Como já foi enunciado, a nossa aproximação inicial com um modelo de regressão simples é mais vulnerável a especificações incorretas: Se, por exemplo, a fecundidade cair em proporção logarítmica com o salário, a estimação do modelo com uma função linear da variável salário mensal tende a ter diferentes variâncias para diferentes valores dessa variável. Nesse caso, podemos utilizar "hetero" como input para o argumento "vcov" da função "feols" do pacote "fixest", de forma a obtermos coeficientes com um estimador robusto à heterocedasticidade. No nosso modelo simples, os resultados com a estimação robusta à heterocedasticidade são os mesmos daqueles utilizando o método não-robusto.

A sexta (e última) hipótese clássica, de normalidade dos erros, também é necessária para realizarmos os testes de hipótese sobre os parâmetros do modelo. Essa hipótese também é forte. Todavia, como ela implica as hipóteses 4 e 5, a perda com essa restrição não é tão grande. Além disso, a grande quantidade de observações justifica que essa hipótese seja feita sem maiores problemas em causa do Teorema do Limite Central.

5 Questão 2

5.1 Questão A

Para avaliar fecundidade na cidade de São Paulo, podemos usar um modelo que permite formas funcionais, portanto, pode-se fazer uso de variáveis binárias, função log entre outros.

No nosso estudo a variável explicada se refere aos filhos, logo, não parece necessário usar logaritmo natural nessa variável.

As variáveis dummy disponibilizadas são: raça (1 para mulheres negras e 0 para brancas, amarelas, pardas); trabalha (1 para mulheres que trabalham e 0 para aquelas que não trabalham); doméstico (1 para aquelas que fazem trabalho doméstico e 0 para aquelas que não fazem trabalho doméstico); educ1 (1 para aquelas que possuem entre 1 a 3 anos de educação e 0 para mais que 3 anos); educ2 (1 para aquelas que possuem entre 4 a 7 anos de educação e 0 para mais que 7 anos); educ3 (1 para aquelas que possuem entre 8 a 11 anos de educação e 0 para mais que 11 anos); educ4 (1 para aquelas que possuem de 12 ou mais anos de educação e 0 para menos que 12 anos); filho (1 se possui pelo menos um filho e 0 não possui nenhum filho).

A variável de comparação para as dummies de educação seria a de mulheres que não possuem nenhum ano de escolaridade, mas como existem poucas observações, criou-se a variável categórica para Fundamental 1 incompleto, Fundamental 2 incompleto, Ensino médio incompleto, Ensino médio completo e Ensino superior.

Estimou-se os três seguintes modelos para cada um de dois grupos: "Baixa escolaridade" (Fundamental 1 e 2 Incompleto, Ensino médio Incompleto) e "Alta escolaridade"

(Ensino médio completo e Superior).

5.1.1 Modelo 1 - Relação de quantidade de filhos e educação

$$\text{Quantidade de filhos} = \beta_0 + \sum_{i=1}^{n=4} \beta_i educ_i + u_1$$

5.1.2 Modelo 2 - Relação de quantidade de filhos e salário por hora

$$\text{Quantidade de filhos} = \beta_0 + \beta_1 log(\text{Salário por hora}) + u_2$$

5.1.3 Modelo 3 - Relação de quantidade de filhos, salário por hora linear e ao quadrado

$$\text{Quantidade de filhos} = \beta_0 + \beta_1 log(\text{Salário por hora}) + \beta_1 log(\text{Salário por hora})^2 + u_3$$

5.2 Questão B

As variáveis criadas foram de salário por hora (somente dividindo salário mensal por total de horas de trabalho mensal), uma variável categórica de educação -Fundamental 1 Incompleto; Fundamental 2 Incompleto; Ensino Médio Incompleto; Ensino Médio Completo e Ensino Superior-, e alteração de variáveis com erros.

Em relação à inclusão, usou-se com base na literatura a motivação de entender os seguintes mecanismos: Maior escolaridade implica que mulheres terão menos filhos - seja por condições biológicas ou seja por escolha? Maior renda tem a mesma relação que escolaridade?

Entendendo as relações destas variáveis que tangem ao tema de inserção das mulheres no mercado de trabalho e renda, podemos observar as heterogeneidades para raça, desigualdade econômica e potencial dedicação ao trabalho doméstico.

Temos algumas limitações quanto às medidas de renda: temos dados sobre salário por hora como indicador de renda, mas sabemos que renda compõe-se de outros rendimentos como de capitais, ademais, o rendimento familiar completo seria mais adequado, entretanto não temos informação sobre a renda de outros responsáveis pela criança.

Outra dificuldade é que a base apresenta 58 mulheres com salário igual a 0, 57 fazem trabalho doméstico e 1 não o faz, sendo que todas têm horas mensais de trabalho. No dicionário provido a variável consta como "Salário mensal do trabalho", então não sabemos exatamente se isso inclui trabalho doméstico ou não, assim ao criar o indicador de salário/hora podemos estar superestimando para mulheres que não fazem trabalho doméstico.

5.3 Questão C

Para o modelo 1, temos que o coeficiente de impacto de grau de escolaridade diminui em $\hat{\beta}_i \ \forall \ i \in (1,4)$ unidades de filhos, quando comparadas a uma mulher com maior grau de escolaridade como fundamental 1 incompleto, as mulheres com mais anos completos de ensino têm menos filhos. Uma surpresa é a diferença do estimador de ensino médio completo para o estimador de ensino superior. Há um aumento desses graus de ensino, que não sabemos se é estatisticamente significativo entre si, afinal comparamos com o estimador de fundamental incompleto 1 e vemos significância de 1%.

Impacto da escolaridade na quantidade de filhos

	Quantidade de Filhos
Intercepto	2.742*** (0.1053)
Fundamental 2 incompleto	-0.9257*** (0.1150)
Ensino médio incompleto	-1.701*** (0.1074)
Ensino médio completo	-2.048*** (0.1263)
Ensino superior	-1.959*** (0.1094)
N	6,314
R-quadrado	0.12497
R-quadrado ajustado	0.12441
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	

Quando estimamos em regressões separadas para cada regressão, temos no segundo modelo os seguintes resultados:

Impacto linear do salário por hora na quantidade de filhos

	Quantidade de Filhos	
	< Ens. Médio	≥ Ens. Médio
Intercepto	1.395*** (0.0623)	0.5177*** (0.0982)
Log Salário por hora	-0.0372 (0.0333)	0.0973** (0.0377)
N	4,486	1,464
R-quadrado	0.00034	0.00556
R-quadrado ajustado	0.00012	0.00488
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

Podemos observar a relação de salário por hora, a variável está na função log com adição de 0,1 para aqueles que tem salário igual a 0, dado que a média da variável é aproximadamente 10.

Para o grupo com menos que ensino médio, vemos que essa relação é insignificante estatisticamente. Enquanto para o grupo com ensino médio ou mais, uma variação percentual do salário hora aumenta 0.0973 a quantidade de filhos. Dado que a média de filhos é de 1,22, esse aumento representa algo expressivo. Portanto, para mulheres mais escolarizadas a escolha de ter um filho a mais é muito expressiva.

Para o terceiro modelo, quando fazemos regressões separadas por grupos, temos a seguinte tabela:

Impacto não-linear do salário por hora na quantidade de filhos

	Quantidade de Filhos	
	< Ens. Médio	≥ Ens. Médio
Intercepto	1.429*** (0.0602)	0.5670*** (0.1353)
Log Salário por hora	-0.1177** (0.0401)	0.0562 (0.0875)
Log(Salário por hora)^2	0.0300* (0.0125)	0.0078 (0.0154)
N	4,486	1,464
R-quadrado	0.00202	0.00576
R-quadrado ajustado	0.00157	0.00440
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

Usando o mesmo tratamento para a variável de salário por hora, na função log, temos que para mulheres menos escolarizadas, a relação linear é negativa, mas a aceleração é positiva. Isso significa que as mulheres com menos que ensino médio têm menos filhos quando seus salários aumentam, entretanto essa relação muda a determinado nível salarial, assim a quantidade de filhos começa a aumentar.

5.4 Questão D

Temos diversos problemas para identificar uma causalidade completa na análise que requer melhor estrutura de dados para obter-se um desenho de modelo. Seria importante ter um painel com dados longitudinais em que poderíamos acompanhar a evolução dessa mulher antes e depois de um filho nos seus níveis educacionais e de renda, para então poder separar os efeitos relacionados e entender a direção dessa causalidade.

Entretanto, dadas as ferramentas presentes, temos a possibilidade de separar e entender o efeito de salário e escolaridade separadamente das outras condições que afetam a escolha de filhos. No caso quando vemos a significância das variáveis nos modelos, podemos compreender a forma com a qual elas impactam a variável dependente.

Assim, não é completamente correto dizer que mais escolaridade é uma causa para menos filhos, pois modelos mais elaborados seriam mais eficientes para determinação, mas no momento, temos que tudo mais constante, mais escolaridade diminui a quantidade de filhos.

5.5 Questão E

Como dito anteriormente, gostaríamos de estimar o efeito de renda, mas não temos dados de recebimentos para além do salário e nem sabemos o rendimento do restante dos responsáveis pela criança. Sabe-se que para pessoas nos quantis mais altos de distribuição de renda, parte relativamente importante é composta por ganhos de capital e não de trabalho. Além disso, pessoas nos quantis mais altos de renda se relacionam com pares da mesma distribuição de renda. Provavelmente existe um viés positivo quando observamos renda (mesmo que usando salário como um indicador possível).

Se temos um modelo original dado por:

$$\text{Qtd filhos} = \beta_0 + \beta_1 log(\text{Salário/hora}) + \beta_2 log(\text{Outra renda mulher}) + \beta_3 log(\text{Renda familiar} + e$$

Se temos o modelo estimado como:

$$\text{Quantidade de filhos} = \beta_0 + \beta_1 log(\text{Salário por hora}) + u_2$$

O estimador de $\hat{\beta_1}$ é dado por:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})y_i}{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2}$$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e)}{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2}$$

$$E(\hat{\beta_1}|X) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})x_{2i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2} + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})x_{3i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x_1})^2}$$

Para que seja não viesado temos que:

$$\sum_{i=1}^{n}(x_{1i} - \bar{x_1})x_{2i} = 0 \quad \vee \beta_2 = 0$$

e

$$\sum_{i=1}^{n}(x_{1i} - \bar{x_1})x_{3i} = 0 \quad \vee \beta_3 = 0$$

Temos por hipótese que esses fatores são positivos, as covariâncias entre as variáveis são positivas e os parâmetros verdadeiros são positivos, logo temos um viés positivo sobre o modelo quando estimamos o impacto da renda.

O argumento segue o mesmo para o modelo com a variável ao quadrado, seria interessante ver se com os dados o estimador dessa variável ao quadrado é ainda mais positivo quando separamos em grupos de escolaridade.

Para escolaridade e primeiro filho, temos que o modelo original seria dado por:

$$\text{Quantidade de filhos} = \beta_0 + \sum_{i=1}^{n=4} \beta_i educ_i + \sum_{j=1}^{n=8} \beta_j educ_i * \text{primeiro filho}_k + e_2$$

A variável de primeiro filho poderia separar em dois grupos: mulheres que tiveram o primeiro filho com menos que 20 anos, e mulheres que tiveram com mais de 20 anos. Assim, esperamos que para mulheres com menor escolaridade e menor grupo de idade do primeiro filho, a quantidade de filhos seja maior que mulheres com mais escolaridade e menor grupo de idade quando tiveram o primeiro filho. Assim, conseguimos separar o efeito de que uma mulher parou de estudar por conta do primeiro filho, ou estudou menos por ter o filho mais nova, afinal, temos um contrafactual, as mulheres que tiveram o filho mais novas e não pararam de estudar.

5.6 Questão F

Para os testes, gostaríamos de testar os estimadores de escolaridade de ensino médio completo para o estimador de ensino superior, além de olhar para os estimadores com as heterogeneidades dispostas.

Primeiro, temos o teste de escolaridade:

$$H_0 : \gamma = \beta_3 - \beta_4 = 0, H_1 : \gamma = \beta_3 - \beta_4 \neq 0$$

Assim, estimamos o modelo:

$$\text{Quantidade de filhos} = \beta_0 + \sum_{i=1}^{n=4} \beta_i educ_i + u$$

Entretanto, nosso grupo de comparação é ensino médio completo, portanto os estimadores são:

$$t_1 = \frac{\hat{\gamma}}{ep(\hat{\gamma})}$$

$$t_1 = \frac{0.0896}{0.0758}$$

$$t_1 = 1.182058$$

Como $|t_1| < 1,96$, a diferença não é estatisticamente significante - o que já era observado na tabela de resultados. Portanto, o estimador mais negativo para ensino médio completo que cresce não é um problema para o nosso argumento.

Agora, no modelo 1 de escolaridade, incluímos as variáveis dummy para ver se existem diferenças para raça (mulheres negras e não negras), mulheres que fazem trabalho doméstico, e riqueza (salário por hora acima da mediana).

Impacto da escolaridade na quantidade de filhos

	Quantidade de Filhos
Intercepto	2.157*** (0.1099)
Fundamental 2 incompleto	-0.9078*** (0.1142)
Ensino médio incompleto	-1.667*** (0.1071)
Ensino médio completo	-2.026*** (0.1277)
Ensino superior	-1.926*** (0.1126)
Trabalho Doméstico	0.5793*** (0.0389)
Raça	0.0948 (0.0677)
Riqueza	0.1158** (0.0361)
N	6,314
R-quadrado	0.14923
R-quadrado ajustado	0.14829
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	

Podemos ver que mulheres que fazem trabalho doméstico (ou seja tem a variável igual a 1), possuem mais filhos para todas os períodos de escolaridade pois o intercepto delas comparado a mulheres que não fazem trabalho doméstico é maior.

Além disso, vemos que raça não é estatisticamente significante, ou seja mulheres negras não parecem ter mais ou menos filhos que outras mulheres.

E quando olhamos para mulheres que possuem salário por hora maior que a mediana, essas têm mais filhos, o seu intercepto é maior, ou seja, para todas as escolaridades, quando a mulher tem maior salário, possui mais filhos.

Os testes são os seguintes, primeiro para trabalho doméstico:

$$t_2 = \frac{\hat{\beta}_5}{ep(\hat{\beta}_5)}$$

$$t_2 = \frac{0.5793}{0.0389}$$

$$t_2 = 14.89203$$

Para raça:

$$t_3 = \frac{\hat{\beta}_6}{ep(\hat{\beta}_6)}$$

$$t_3 = \frac{0.0948}{0.0677}$$

$$t_3 = 1.400295$$

Para riqueza:

$$t_4 = \frac{\hat{\beta}_7}{ep(\hat{\beta}_7)}$$

$$t_4 = \frac{0.1158}{0.0361}$$

$$t_4 = 3.207756$$

Portanto, trabalho doméstico e riqueza são estatisticamente significantes, mas raça não.

Em seguida, temos a tabela com os resultados para modelo 2:

Impacto linear do salário por hora na quantidade de filhos

	Quantidade de Filhos	
	<Ens. Médio	≥ Ens. Médio
Intercepto	0.7410*** (0.0841)	0.1964. (0.1131)
Log Salário por hora	-0.0483 (0.0406)	0.1149** (0.0405)
Trabalho doméstico	0.7344*** (0.0551)	0.4727*** (0.0553)
Raça	0.1101 (0.0805)	-0.0088 (0.1609)
Riqueza	0.0594 (0.0512)	-0.1007 (0.1148)
N	4,486	1,464
R-quadrado	0.03104	0.04130
R-quadrado ajustado	0.03018	0.03867
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

Nesse caso, os testes t seguem os mesmos, mas como podemos confirmar pelo tabela de resultados, vemos que somente trabalho doméstico é significante - e é significante para os dois grupos de escolaridade-, e o efeito tem o mesmo sinal, quando a mulher faz trabalho doméstico, ela tem mais filhos, entretanto, mulheres mais escolarizadas que fazem trabalho doméstico têm menos filhos que mulheres menos escolarizadas que fazem trabalho doméstico.

E por fim temos o modelo 3:

Impacto não-linear do salário por hora na quantidade de filhos

	Quantidade de Filhos	
	<Ens. Médio	$\geq$ Ens. Médio
Intercepto	0.7661*** (0.0825)	0.2276. (0.1339)
Log Salário por hora	-0.1068* (0.0425)	0.0820 (0.0924)
Log(Salário por hora)^2	0.0260. (0.0133)	0.0058 (0.0153)
Trabalho Doméstico	0.7311*** (0.0551)	0.4732*** (0.0553)
Raça	0.1134 (0.0804)	-0.0073 (0.1609)
Riqueza	0.0318 (0.0551)	-0.0884 (0.1195)
N	4,486	1,464
R-quadrado	0.03224	0.04140
R-quadrado ajustado	0.03116	0.03811
Significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1		

Podemos ver que os padrões são os mesmos que o modelo 2, ou seja, a existência da variável ao quadrado não muda a relação para trabalho doméstico, raça e riqueza.

5.7 Questão G

Temos que as hipóteses clássicas são: - Linearidade nos parâmetros; - Amostragem aleatória; - Inexistência de multicolinearidade perfeita (ou existência de variabilidade nas variáveis dependentes); - Exogeneidade estrita; - Homocedasticidade e inexistência de autocorrelação; - Normalidade dos erros.

Para as hipóteses 1 a 3 não temos problemas aparentes, afinal definimos os modelos como lineares nos parâmetros, temos uma amostra aleatória e não fazemos uso de variáveis que possam gerar multicolinearidade (como variáveis dummy com a mesma informação).

Para a hipótese 4, sobre exogeneidade precisamos garantir que tudo que está nos erros não é correlacionado com as variáveis explicativas, portanto fazemos algumas análises do que poderia interferir no modelo. Como discutido anteriormente, provavelmente existe um viés que subestima o efeito de renda na escolha de filhos, dado que só captamos o efeito de trabalho da mulher. Além disso, podemos ter um efeito de que a mudança na escolaridade é diferente para as idades em que as mulheres têm seu primeiro filho.

Portanto, existem sim, diversos fatores associados às variáveis explicativas que não estão inclusos no modelo, pois não são observados na amostra, logo não podemos concluir que não existe viés, entretanto como argumentamos, a direção do efeito não parece ser viesada, e sim a sua magnitude. Se nossas hipóteses são de que o efeito de ter menos escolaridade deve impactar em ter mais filhos quanto menor a idade da mãe com o primeiro ou que quanto maior a renda familiar mais filhos, nossos resultados possuem o mesmo sentido, porém são viesados em sua magnitude.

Para a hipótese 5, fazemos a correção para heterocedasticidade e autocorrelação dos erros no próprio estimador do software R, utilizamos o parâmetro “vcov” na função “feols” do pacote “fixest”.

Para a hipótese 6, assumimos que os erros são tratados com distribuição normal para realizar os testes no software, e não parece haver evidência de que isso seria um problema na análise.

5.8 Questão H

Para o primeiro modelo temos para cada nível de escolaridade:

Escolaridade	Previsão de filhos	IC
Fundamental 1 incompleto	2.7421603	[2.535699, 2.9486212]
Fundamental 2 incompleto	1.8164452	[-1.151160, -0.7002703]
Ensino médio incompleto	1.0413024	[-1.911369, -1.4903468]
Ensino médio completo	0.6937799	[-2.296024, -1.8007366]
Ensino superior	0.7833462	[-2.173271,-1.7443572]

Aqui a média de quantidade de filhos diminui com maiores etapas concluídas de escolaridade, como visto anteriormente, além de que os intervalos de confiança são consistentes com as significâncias apresentadas nas tabelas de resultados.

Para o segundo modelo temos:

Salário para níveis de escolaridade	Previsão de filhos	IC
Salário ($<$ Ens. Médio)	1.318831	[-0.1025323, 0.02822815]
Salário ($\geq$ Ens. Médio)	0.811509	[0.02335572, 0.1712812]

Temos que para mulheres com mais salário e menos escolaridade, a previsão é de uma quantidade maior adicionada, mas como não é estatisticamente significante, o argumento não é válido.

Para o terceiro modelo temos:

Salário para níveis de escolaridade	Previsão de filhos	IC
Salário ($<$ Ens. Médio)	2.970334	[-0.196304898,-0.03912853]
Salário ao quadrado ($<$ Ens. Médio)		[0.005446964, 0.05457166]
Salário ($\geq$ Ens. Médio)	3.988776	[-0.11546158, 0.22787181]
Salário ao quadrado ($\geq$ Ens. Médio)		[-0.02245694,0.03796151]

Nesse modelo, mulheres com maior escolaridade e maior salário têm mais filhos que a média do grupo de mulheres com menor escolaridade.

6 Questão 3

6.1 Questão A

A base de dados ideal deveria incluir além das informações que possuímos, a idade da mãe quando teve o primeiro filho, o rendimento total familiar, a escolaridade do pai da criança, e seria interessante obter dado um questionário, questões sobre escolhas para ter filhos. A PND possui questões para diferentes faixas etárias sobre como ter ou não filhos afetou sua carreira profissional, formação acadêmica, relações pessoais, entre outros.

Além dessas variáveis, seria interessante saber de que bairro ou setor censitário essas mulheres moram - será que indices de urbanização afetam a escolha de quantidade de filhos?-, além de alguma medida temporal para observar essas mulheres ao longo do tempo.

$$\text{Qtd filhos} = \beta_0+\sum_{i=1}^{n=4}\beta_i educm_{i,t}+\beta_5 log(renda_t)+\beta_6 log(renda_t^2)+\beta_7 idade+\beta_8 educh+\beta_9 urbano+\epsilon$$

Em que educm é o nível de educação da mãe pelas classificações que usamos, renda renda é o rendimento familiar, idade é uma dummy se a mãe teve um filho antes dos 20 anos, educh é uma dummy se o marido possui ensino médio completo ou mais, urbamo é uma variável dummy que medimos se o local é mais urbanizado, para a cidade de SP podemos pensar na existência de estação de metrô, trem e ônibus próximas.

6.2 Questão B

Espera-se que para $\beta_i \quad i \ \in (1,4)$ a cada nível maior de educação o estimador se torne mais negativo quando comparado ao menor nível de educação. Para β_5 espera-se que seja negativo para uma base completa, para β_6 espera-se que seja positivo. Para β_7 espera-se que seja positivo, e para β_8 espera-se que seja negativo e para β_9 também espera-se que seja negativo.

6.3 Questão C

Na medida em que o modelo ideal e a base de dados ideal elaboradas nesse exercício hipotético já visavam a conformidade com as hipóteses necessárias para a estimação precisa dos efeitos causais dos fatores relevantes, ao considerar as hipóteses clássicas de regressão linear, a única hipótese que ainda acreditamos ter uma dificuldade maior de aceitação, mesmo nesse cenário fantástico, é a hipótese de exogeneidade estrita. Entretanto, uma vez que tentamos abarcar, de forma explícita, as principais motivações de escolha de filhos citadas na literatura sobre fecundidade, não há grandes motivos para acreditar que a hipótese geraria, no final das contas, algum problema.