FUNDAÇÃO
GETULIO VARGAS
EPGE
Escola de Pós-Graduação
em Economia

Ensaios Econômicos

Escola de
Pós Graduação
em Economia
da Fundação
Getulio Vargas

Nº 638 ISSN 0104-8910

Diferenciais de Salários por Raça e Gênero: Aplicação dos procedimentos de Oaxaca e Heckman em Pesquisas Amostrais Complexas

Alexandre Pinto de Carvalho, Marcelo Côrtes Neri, Denise Britz Silva

Dezembro de 2006

Os artigos publicados são de inteira responsabilidade de seus autores. As opiniões neles emitidas não exprimem, necessariamente, o ponto de vista da Fundação Getulio Vargas.

Diferenciais de Salários por Raça e Gênero:
Aplicação dos procedimentos de Oaxaca e Heckman em Pesquisas Amostrais Complexas[1]

Alexandre Pinto de Carvalho[2]

Marcelo Côrtes Neri[3]

Denise Britz Silva[4]

RESUMO

Este artigo decompõe o diferencial de salários por cor e sexo dos trabalhadores brasileiros usando os microdados da Pesquisa Nacional por Amostra de Domicílio (PNAD). A metodologia consiste em estimar a equação de salários (Mincer, 1974) com a correção do viés de seleção das informações dos salários (Heckman, 1979). Em seguida, a decomposição do diferencial da média do logaritmo do salário/hora foi obtida pelo procedimento de Oaxaca (1973) apresentada em dois efeitos: características produtivas e discriminação. A análise empírica tem como foco o uso adequado de procedimentos de modelagem estatística em pesquisas, por amostragem complexa, conforme os trabalhos de Skinner e Smith (1989) e Pessoa e Silva (1998).

Os resultados indicam a necessidade de se incorporar o plano amostral e a correção do viés de seleção da informação dos salários, visando melhorar a qualidade das estimativas das equações de salários e avaliar adequadamente as medidas de discriminação. Como exemplo, a estimativa do coeficiente de discriminação, D, entre homens e mulheres de cor branca é 0,37 sem a correção do viés e 0,30 com a correção do viés de seleção das informações dos salários.

---

[1] Este trabalho foi apresentado no XV Encontro de Estudos Populacionais, 18 a 22 setembro de 2006, ABEP.

[2] Mestre pela Escola Nacional de Ciências Estatísticas/IBGE

[3] Centro de Políticas Sociais/FGV e EPGE/FGV

[4] Escola Nacional de Ciências Estatísticas/IBGE

1 Introdução

Nos últimos anos, a estimação da equação de salários[5] tem sido amplamente utilizada em estudos relativos à discriminação no mercado de trabalho brasileiro segundo cor e sexo. Os resultados revelam os efeitos dos determinantes do salário segundo características individuais e do posto de trabalho. O propósito destes trabalhos é avaliar a situação em que indivíduos com atributos produtivos semelhantes, exceto pela cor ou sexo, têm salários tão diferenciados. Este conhecimento auxilia na formulação de políticas públicas em nosso mercado de trabalho. No entanto, alguns dos estudos que utilizam dados provenientes de pesquisas amostrais complexas, como a Pesquisa Nacional por Amostra de Domicílios (PNAD) realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), não incorporam o plano amostral em suas análises. Rodrigues (2003) revela em seu estudo que o fato de não incorporar os pesos e o plano amostral na análise gera estimativas incorretas tanto para os coeficientes como para as suas variâncias.

O objetivo principal desta dissertação é decompor o diferencial da média do logaritmo do salário/hora com ênfase na cor e sexo dos trabalhadores brasileiros (homens de cor branca, mulheres de cor branca, homens de cor preta ou parda e mulheres de cor preta ou parda). A metodologia, baseada no modelo de capital humano, consiste em estimar a equação de salários (Mincer, 1974) com a correção do *viés de seleção das informações* dos salários através do procedimento de Heckman (1979). O procedimento da decomposição do diferencial de salários é baseado no trabalho de Oaxaca (1973) composto por dois efeitos: características produtivas individuais e da discriminação. A análise empírica tem como foco o uso adequado de procedimentos de modelagem estatística em pesquisas por amostragem complexa, conforme os trabalhos de Skinner e Smith (1989) e Pessoa e Silva (1998).

O artigo encontra-se organizado em mais quatro seções, além desta introdução e as considerações finais. Na seção 2, apresenta-se uma análise não-controlada do diferencial da média dos salários entre os homens de cor branca, homens de cor preta ou parda, mulheres de cor branca e mulheres de cor preta ou parda. Avalia-se a inserção e participação no mercado de trabalho e, adicionalmente, caracterizam-se os grupos de cor e sexo segundo estatísticas descritivas da escolaridade, idade, experiência de trabalho e local da residência.

A seção 3 elucida a importância do uso adequado dos dados na inferência estatística com base em pesquisas amostrais complexas como a PNAD 2003. Basicamente, é um resumo dos trabalhos de Skinner, Holt e Smith (1989), Pessoa e Silva (1998), Phillippe (2001) e Rodrigues (2003).

A seção 5 é calculada a decomposição do diferencial da média do logaritmo do salário/hora através da modelagem estatística. O procedimento consiste em estimar a equação de salários para cada grupo e trabalhador com o uso da análise de regressão (*equação de salários* ou *equação minceriana*), através do método de Máxima Pseudo-Verossimilhança (MPV), incorporando o plano amostral. Em seguida, com base no trabalho de Oaxaca (1973), o diferencial é decomposto em dois efeitos: das características individuais e da discriminação.

A seção 6 destaca que para determinadas pessoas seria vantajoso trabalhar se o salário recebido (ou salário potencial) fosse maior que o custo de oportunidade de não trabalhar (ou salário reserva). Desta forma, existe um *viés de seleção das informações do salário*. Neste caso, as estimativas dos coeficientes da equação de salários obtidas a partir das informações dos indivíduos que trabalham na data de referência da pesquisa podem estar viesadas sob a ótica do *viés de seleção das informações*

[5] Modelagem estatística através da Análise de Regressão.

considerando-se que o modelo utilizado na seção 5 não incorpora a informação sobre a avaliação dos indivíduos que não trabalham no que se refere ao custo de oportunidade. Heckman (1979) apresentou este fenômeno e sua solução não apenas para pesquisas amostrais. De modo geral, o viés de seleção de informação pode ser decorrente de duas razões, como destaca Heckman: ou em virtude de seletividade das informações dos indivíduos ou devido ao desenho amostral da pesquisa. Cabe ressaltar que, ao utilizar os dados da PNAD, o efeito que deseja-se incorporar na modelagem visa corrigir *o viés de seletividade da informação dos salários* para os indivíduos que, apesar de estarem devidamente representados na amostra da PNAD 2003, não trabalharam na data de referência da pesquisa supostamente devido a uma avaliação do salário potencial e do custo de oportunidade envolvido nesta escolha.

A seção 7, por sua vez, apresenta as considerações finais e as propostas de trabalhos futuros. Uma extensão natural desta dissertação seria o cálculo da decomposição para os percentis do logaritmo do salário/hora através do uso de regressões quantílicas com a correção de Heckman e a incorporação do plano amostral nas análises.

2 Análise Não Controlada do Diferencial de Salários

2.1 A Utilização de Estatísticas Descritivas

A análise não controlada do diferencial de salários é obtida através de estatísticas descritivas do rendimento do trabalho principal (salário) e indicadores do mercado de trabalho dos grupos de cor e sexo (cor/sexo). Estes resultados são obtidos sem o uso de modelagem estatística.

Considera-se quatro grupos no mercado de trabalho: mulheres de cor branca, mulheres de cor preta ou parda, homens de cor branca e homens de cor preta ou parda. A fonte de dados utilizada é a Pesquisa Nacional por Amostra de Domicílios (PNAD) realizada pelo IBGE em 2003 (maiores detalhes na seção 3).

Em 2003, a População Economicamente Ativa[6] (PEA) no Brasil era em torno de 142 milhões de brasileiros[7] dos quais 79 milhões tinham trabalho, 8 milhões procuravam trabalho e 55 milhões estavam fora da força de trabalho. A população ocupada era composta por aproximadamente 42 milhões de homens e mulheres que declararam sua cor branca e 37 milhões de homens e mulheres de cor preta ou parda.

Na tabela 2.1, verifica-se que, embora a População Economicamente Ativa das mulheres seja superior a dos homens, a sua taxa de participação é inferior. A inserção no mercado de trabalho, representada pela taxa de desocupação, revela que existe um efeito entre as variáveis cor e sexo. Destaca-se a situação das mulheres de cor preta ou parda que apresentam a menor taxa de participação (47,91%) e a maior taxa de desocupação (14,47%).

Tabela 2.2 – Taxas de Participação e Desocupação no Brasil segundo cor e sexo

Indicadores	Homens		Mulheres	
	Branca	Preta ou Parda	Branca	Preta ou Parda
Taxa de Participação	68,87%	68,56%	49,41%	47,91%
Taxa de Desocupação	7,47%	9,14%	11,46%	14,47%

Fonte: Elaboração própria a partir dos microdados da PNAD/IBGE 2003

[6] Pessoas com 10 anos ou mais de idade.
[7] Exceto as pessoas que declararam a cor ignorada ou indígena em 2003.

2.2 Análise das Características Produtivas dos Trabalhadores

As estatísticas descritivas para escolaridade, experiência de trabalho, idade, local da residência (área urbana ou rural), salário efetivamente recebido no mês, jornada de trabalho semanal são utilizadas para caracterizar os grupos de trabalhadores segundo cor e sexo. Em seguida, define-se a média da razão entre o salário e a jornada de trabalho mensal para o cálculo do diferencial não-controlado entre os trabalhadores.

Os resultados da tabela 2.2 revelam que existe uma disparidade entre os salários dos trabalhadores brasileiros segundo cor e sexo. Como exemplo, embora as mulheres de cor branca tenham um nível de escolaridade superior ao dos homens de cor branca o seu salário é inferior, R$657 versus R$1.010. O mesmo resultado é encontrado entre homens e mulheres de cor preta ou parda. Além disso, os salários das pessoas de cor branca são superiores ao de pessoas de cor preta ou parda em nosso mercado de trabalho. Nota-se que a jornada de trabalho dos homens e das mulheres é diferente e, por isso, faz-se necessário calcular o diferencial através de uma medida padronizada denominada salário/hora. O salário/hora é a razão entre o salário recebido no mês multiplicado pela jornada de trabalho mensal (i.e., jornada de trabalho semanal multiplicada 4.2 semanas).

Tabela 2.2 – Estatísticas descritivas das características produtivas dos trabalhadores

Indicadores	Homens		Mulheres	
	Branca	Preta ou Parda	Branca	Preta ou Parda
Salário	1.010,02	478,38	656,76	348,58
Jornada de trabalho semanal	45,40	44,55	38,10	37,31
Escolaridade	7,97	5,65	9,29	7,17
Experiência	23,54	24,25	20,71	22,23
Idade	37,51	35,90	36,00	35,40

Fonte: Elaboração própria a partir dos microdados da PNAD/IBGE 2003.

O diferencial não-controlado é apresentado na tabela 2.3 através da comparação entre homens de cor preta/parda, mulheres de cor branca e mulheres de cor preta/parda com o grupo dos homens de cor branca – definido como grupo base. O propósito desta análise é avaliar se a diferença entre o salário/hora é estatisticamente diferente de zero – o teste é realizado pela estatística t. Como exemplo, os homens de cor branca ganham R$ 3,26 a mais que os homens de cor preta. O menor diferencial é registrado entre homens e mulheres de cor branca (R$1,34). Vale ressaltar que todos os resultados são estatisticamente diferentes de zero.

Tabela 2.3 – Diferença da média do salário/hora[1] segundo cor e sexo

Diferença entre o salário/hora	Estimativa	Estatística t
Homens de cor branca - Homens de cor preta	3,26	26,50
Homens de cor branca - Mulheres de cor branca	1,34	8,52
Homens de cor branca - Mulheres de cor preta ou parda	3,54	28,70

Fonte: Elaboração própria a partir dos microdados da PNAD/IBGE 2003.

Os resultados revelam, em termos percentuais, o salário do grupo base versus os demais grupos de cor/sexo. Os homens de cor branca ganham, em média, o dobro dos homens e mulheres de cor preta/parda e 28% a mais que as mulheres de cor branca. Os resultados da análise não-controlada do diferencial de salários revelam as disparidades segundo cor e sexo. Em geral, as pessoas de cor preta/parda ganham menos que o grupo base. No entanto, tais comparações não são suficientes para avaliar o quanto deste diferencial é explicado pela discriminação ou pelas características individuais (escolaridade, experiência de trabalho, local de residência e idade). As próximas sub-seções têm por objetivo testar a existência da discriminação através da modelagem estatística. Além disso, decompor este diferencial na parcela referente à discriminação ou às características individuais.

3 Análise de Dados Amostrais Complexos

O esquema de seleção da amostra da PNAD, conforme descreve Silva, Pessoa e Lilá (2002), é estratificado e conglomerado com um, dois ou três estágios de seleção, de acordo com o tipo de estrato. O "primeiro tipo de estrato" é composto pelas seguintes unidades da federação: Acre, Alagoas Amapá, Amazonas, Distrito Federal, Espírito Santo, Goiás, Maranhão, Mato Grosso, Mato Grosso do Sul, Paraíba, Piauí, Rio Grande do Norte, Rondônia, Roraima, Santa Catarina Sergipe e Tocantins. Enquanto que, o "segundo tipo de estrato" é formado pelas unidades da federação: Bahia, Ceará, Minas Gerais, Pará, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul e São Paulo.

O uso de dados de amostrais complexos, conforme destacada na seção anterior, envolve probabilidades distintas de seleção das unidades, conglomeração das unidades e estratificação. Para maiores detalhes ver os trabalhos de Skinner, Holt e Smith (1989) e Pessoa e Silva (1998).

A utilização de métodos adequados para realização de inferência em dados amostrais complexos permite estimar valores de uma variável de interesse e avaliar o grau de precisão das estimativas (através de suas variâncias). As estimativas das variâncias, por sua vez, são influenciadas pelo plano amostral utilizado. Com isso, é importante ressaltar a importância da incorporação do plano nos procedimentos de inferência com base em dados amostrais complexos como a PNAD. Já existem programas estatísticos que suportam tais análises, como exemplo, o SAS, STATA (2003).

A justificativa para a incorporação do desenho amostra nas inferências analíticas, partindo de dados amostrais complexos (Côrrea, 2001), é que os pesos podem ser usados para proteger *contra planos amostrais não-ignoráveis (Pessoa e Silva, 1998), que poderiam introduzir ou causar vícios, e má especificação do modelo.*

Esta seção tem por objetivo apresentar o Efeito do Plano Amostral em pesquisas de dados amostrais complexos como a PNAD. Além disso, apresentar o procedimento para o cálculo dos intervalos de confiança, os testes de hipóteses e o Método de Máxima Pseudo-Verossimilhança (MPV) aplicado na inferência analítica (modelagem).

3.1 Efeito do Plano Amostral

O Efeito do Plano Amostral (EPA[8]) tem por finalidade avaliar o impacto em desconsiderar o esquema de seleção da amostra no cálculo das estimativas. Esta medida foi proposta inicialmente por Kish (1965) e aperfeiçoada por Kish e Frankel (1974), para maiores detalhes ver Pessoa e Silva (1998). Na inferência estatística, para um parâmetro θ, o EPA é obtido pela razão entre a variância do plano amostral complexo (verdadeiro) e a variância da distribuição do estimador $\hat{\theta}$ de θ induzida pelo plano de amostragem aleatória simples (AAS[9]) - $V_{AAS}\left(\hat{\theta}\right)$.

$$\mathrm{EPA}\left(\hat{\theta}\right) = \frac{\mathrm{V_p}\left(\hat{\theta}\right)}{\mathrm{V_{AAS}}\left(\hat{\theta}\right)} \qquad (4.1)$$

onde, $\mathrm{V_p}\left(\hat{\theta}\right)$ = variância da distribuição de $\hat{\theta}$ induzida pelo plano amostral complexo.

Skinner, Holt e Smith (1989, p.24) destacam que esta medida é importante para avaliar a eficiência quando comparamos desenhos alternativos na concepção das pesquisas. Além disso, o uso do EPA apresenta dificuldades no seu cálculo em inferências analíticas (modelagem) e, por isso, definiram o conceito do EPA ampliado (misspecification effect – meff). Esta medida mensura a tendência de um estimador usual (consistente), calculado sob hipótese de observações independentes e identicamente distribuídas (IID), subestimar ou superestimar a variância verdadeira do estimador pontual. O EPA ampliado (também denominado por meff - misspecification effect) é a razão entre a variância do estimador sob o plano amostral ou modelo correto $\mathrm{V_{VERD}}\left(\hat{\theta}\right)$ sobre a esperança do estimador da variância de $\hat{\theta}$ sob a hipótese de observações IID da variância $\mathrm{E_{VERD}}\left(\mathrm{v_0}\right)$.

Dado que $\mathrm{v_0} = \hat{\mathrm{V}}_{\mathrm{IID}}\left(\hat{\theta}\right)$ um estimador da variância de $\hat{\theta}$ para uma Amostra Aleatória Simples sem reposição. Então:

$$\mathrm{meff} = \mathrm{EPA_{ampliado}}\left(\hat{\theta}, \mathrm{v_0}\right) = \frac{\mathrm{V_{VERD}}\left(\hat{\theta}\right)}{\mathrm{E_{VERD}}\left(\mathrm{v_0}\right)} = \frac{\mathrm{V_{VERD}}\left(\hat{\theta}\right)}{\mathrm{E_{VERD}}\left(\hat{\mathrm{V}}_{\mathrm{IID}}\left(\hat{\theta}\right)\right)} \qquad (4.2)$$

O $EPA_{ampliado}\left(\hat{\theta}, v_0\right)$ mensura a tendência de v_0 subestimar ou superestimar $V_{VERD}\left(\hat{\theta}\right)$, variância verdadeira sob o modelo e/ou plano amostral de $\hat{\theta}$. Quanto mais afastado de 1 for o valor de $EPA_{ampliado}\left(\hat{\theta}, v_0\right)$, mais incorreta será considerada a especificação do plano amostral ou do modelo nos procedimentos analíticos.

Desta forma, enquanto a medida proposta por Kish baseia-se nas distribuições induzidas pela aleatorização dos planos amostrais comparados, o $EPA_{ampliado}\left(\hat{\theta}, v_0\right)$ pode ser calculado com respeito a distribuições de aleatorização ou do modelo envolvido.

Em geral, são esperadas as seguintes conseqüências sobre o Efeito do Plano Amostral ao ignorar o plano amostral efetivamente adotado e admitir que o desenho da amostra foi AAS:

1. Ignorar os pesos em v_0 pode inflacionar o meff (ou EPA ampliado);
2. Ignorar conglomerações em v_0 pode inflacionar o meff;
3. Ignorar estratificação em v_0 pode reduzir o meff.

[8] É apresentado nos softwares estatísticos como design effect (deff).
[9] As informações são coletadas de forma independente e são identicamente distribuídas – IID

3.2 Estatística teste

Para uma população finita com um parâmetro de interesse θ e sua estimativa pontual $\hat{\theta}$ o intervalo de confiança com nível de confiança (1- α) a partir da distribuição assintótica de $t_0 = \frac{\hat{\theta} - \theta}{v_0^{0,5}}$, sob a hipótese de que as observações são Independente e Identicamente distribuídas (IID) com distribuição N(0;1), é dado por: $\left[\hat{\theta} - z_{\alpha/2} v_0^{0,5}; \hat{\theta} + z_{\alpha/2} v_0^{0,5}\right]$ onde $z_{\alpha/2} = \int_{\alpha/2}^{+\infty} \varphi(t)dt$ e φ é uma função de densidade da distribuição normal padrão.

Para um plano amostral complexo a estatística de teste é dada por: $t_0 = \frac{\hat{\theta} - \theta}{\left[\hat{V}_{VERD}\left(\hat{\theta}\right)\right]^{1/2}}$ tal que $t_0 \dot{\sim} N[0, EPA(\hat{\theta}, v_0)]$ e $\hat{EPA} = \frac{\hat{V}_{verd}(\hat{\theta})}{v_0}$

Desta forma, ao ignorar os pesos e o efeito de conglomeração do desenho amostral pode-se inflacionar o EPA, ampliando-se os intervalos de confiança para os parâmetros de interesse.

4 Análise Controlada do Diferencial de Salários e Decomposição Segundo o Procedimento de Oaxaca.

4.1 Análise de Regressão e a Teoria do Capital Humano

O capital humano é o conjunto das habilidades do indivíduo ligadas à capacidade produtiva, e incorporadas no conhecimento e qualificação para determinadas atividades Becker (1993). Até 1950 os economistas geralmente assumiam que o poder do salário como dados e não adquiridos. As análises sofisticadas do investimento em educação e treinamento por Adam Smith, Alfred Marshall e Milton Friedman não incluíam em suas discussões a produtividade. Então Theodore W. Schultz, entre outros, iniciaram uma exploração pioneira nas implicações do investimento do capital humano nas questões econômicas.

Esta seção tem por finalidade testar a existência de discriminação dos trabalhadores segundo cor e sexo através da análise controlada do diferencial de salários. Os resultados da análise controlada são obtidos pelo uso de modelagem estatística considerando o plano amostral da PNAD 2003. Adicionalmente, decompõe-se o diferencial de salários em efeitos provenientes das características individuais (escolaridade, experiência ou local de residência) e da discriminação.

A metodologia consiste em estimar a equação de salários (Mincer, 1974) e decompor o diferencial de salários através da metodologia apresentada por Oaxaca (1973). Para isso, incorpora-se o esquema amostral complexo da PNAD 2003 apresentado na seção 3. Desta forma, a sinergia entre as técnicas econométricas e as ferramentas estatísticas, fundamentadas na teoria econômica, resulta em medidas concretas que propiciam um maior conhecimento do nosso mercado de trabalho. A principal ferramenta utilizada é a Análise de Regressão[10] com modelos estimados pelo Método de Máxima Pseudo-Verssomilhança.

A análise de regressão é utilizada no estudo da relação de uma variável resposta com um conjunto de variáveis denominadas explicativas. O seu propósito é

[10] O termo Regressão, por sua vez, amplamente utilizado em estudos do mercado de trabalho foi introduzido Galton (1886) – Gujarati (2000).

estimar ou prever o valor médio populacional da variável dependente em termos dos valores conhecidos[11] das explicativas.

A *teoria do capital humano* aplicada à análise de regressão fornece um arcabouço para avaliar como decisões individuais influenciam nos retornos dos rendimentos. Os trabalhos de Becker (1962), Mincer (1964) e Ramos (1996) descrevem a ligação entre o ciclo da vida de um indivíduo e os investimentos em capital humano. Ben-Porath (1973) formaliza a equação de produção do capital humano baseado nos trabalhos de Becker e Mincer.

4.2 Estudos sobre Determinantes do Diferencial de Salários no Brasil

Diversos estudos procuram obter informações sobre os determinantes do salário dos trabalhadores, a partir de suas características individuais (escolaridade, experiência, cor/raça ou local de moradia) e informações sobre o mercado de trabalho (setor de atividade ou ocupação). O entendimento do processo pelo qual as pessoas desenvolvem suas habilidades na escola e no trabalho são fundamentais para investigar não apenas o porquê das diferenças de salário, mas para fomentar idéias no desenvolvimento econômico e social do país. A literatura existente sobre o tema revela que outros fatores também influenciam o salário como, por exemplo, habilidade inata ou discriminação.

Coelho e Corseuil (2002) apresentam um resumo dos estudos sobre Diferencial de Salários nos últimos trinta anos no Brasil. Os autores encontraram diferentes abordagens no uso da equação de salário. Em alguns artigos o foco da análise é a avaliação das características dos indivíduos na determinação do salário, enquanto outros tem como objetivo mensurar diferenciais entre grupos de trabalhadores de acordo com suas características sóciodemográficas.

Rodrigues (2003) revela a estrutura salarial do Brasil a partir da Pesquisa de Padrão de Vida (PPV), realizada em 1996 pelo IBGE, levando em consideração os efeitos do esquema amostral da pesquisa. A contribuição do trabalho é a avaliação do diferencial de salário não-controlado (razão entre os salários do grupo) e o diferencial de salário controlado quando estimado considerando-se ou não o esquema amostral da pesquisa.

A equação de salários (Mincer, 1974), estimada pela Análise de Regressão, é um importante instrumento para decompor o diferencial de salários entre os grupos de cor. A próxima seção discute a forma funcional da variável resposta e as variáveis explicativas utilizadas.

4.3 A forma funcional da Equação de Salários

Mincer (1974) integrou a teoria do investimento em capital humano dentro de um contexto empírico, compatível com a teoria econômica. Desde então, o seu trabalho passou a ser amplamente utilizada em estudos do mercado de trabalho que foi denominada como “função salário do capital humano”, ou popularmente conhecida como “equação minceriana”.

O logaritmo da razão entre o salário e a jornada de trabalho (hora) é a variável de interesse para o estudo do diferencial de salários que também é denominada por variável resposta. Seja:

Z_i - salário/hora do i-ésimo indivíduo,

$Y_i = \ln(Z_i)$ - logaritmo de Z_i,

$\underline{X}_i$ - vetor de variável explicativa do i-ésimo indivíduo,

[11] Conhecido ou fixo no sentido de não estocástico.

Considerando um modelo com apenas uma variável explicativa:

$$Y_i = \beta_0 + \beta_k \underline{X}_i + \varepsilon_i , \qquad \varepsilon_i \sim N(0,1) \tag{4.1}$$

A forma funcional da expressão (5.1) é chamada de modelo semilog, pois a variável resposta aparece na forma logarítimica. Desta forma, o coeficiente de inclinação da variável explicativa $\underline{X}_i$ mede a variação proporcional constante em Y para uma dada variação absoluta no valor da variável explicativa, como exemplo:

$$\beta_k = \frac{Variação \quad relativa \quad na \quad \text{var}\, iável \quad resposta}{Variação \quad absoluta \quad no \quad regressor} \tag{4.2}$$

Se multiplicarmos a variação relativa em Y por 100, a expressão (5.2) fornecerá a variação percentual, ou taxa de crescimento/decréscimo em Y para uma variação absoluta em X, o regressor.

Oaxaca (1973) destaca a experiência, escolaridade, posição na ocupação, setor de atividade, grupos de ocupação, migração e estado civil como controles importantes da fonte da discriminação segundo cor e sexo. Como exemplo, o controle definido pela ocupação elimina alguns dos efeitos das barreiras ocupacionais como fonte de discriminação. Como resultado, estaríamos sub-estimando os efeitos da discriminação quando incluímos tais efeitos na análise. Com isso, ele apresentou um outro conjunto de equações de salários que não considera a ocupação e setor de atividade dos trabalhadores denominado por “equação de salários de características individuais” que é adotado na dissertação.

A partir das evidências apresentadas por Oaxaca e o princípio da parcimônia, as variáveis explicativas utilizadas nesta dissertação são: educação, experiência e local de moradia (área urbana ou rural).

A escolaridade dos trabalhadores é definida pelos anos de estudo completos. Tal informação encontra-se definida nos microdados da PNAD 2003. A experiência do trabalhador é obtida pela expressão: experiência = idade em anos completos - anos de estudo completos – 6.

4.4 Análise Controlada do Diferencial de Salários

Apresenta-se a seguir, os resultados da estimação da equação de salários ajustada para a população da PNAD 2003. A distinção entre os modelos 1 e 2 (tabela 4.1) está na utilização das variáveis cor e sexo, além das demais características como escolaridade, experiência e local de moradia. Esta análise compara indivíduos com atributos semelhantes, por exemplo, escolaridade, experiência e local de moradia, exceto a cor e sexo. Os resultados revelam que mulheres de cor branca e homens e mulheres de cor preta e parda ganham menos do que o grupo base (homens brancos).

Para ambos os modelos os sinais dos coeficientes da escolaridade e experiência estão de acordo com os resultados da Teoria do Capital Humano. A interpretação dos coeficientes de escolaridade (linear e quadrático) revela que o salário aumenta conforme a escolaridade a taxas crescentes. No entanto, os sinais dos coeficientes da experiência indicam que o salário aumenta conforme a experiência de trabalho a taxas decrescentes.

Tabela 4.1 – Estimação das Equações de Salários

Características produtivas	Estimativas dos coeficientes	
	Modelo (1)	Modelo (2)
Escolaridade	0.05335	0.04619
	15.07	**13.67**

Escolaridade2	0.00622	0.00640
	30.21	**32.46**
Experiência	0.04951	0.04905
	85.31	**86.65**
Experiência2	-0.00055	-0.00056
	-49.45	**-51.65**
Área Urbana	0.1898968	0.2007406
	27.37	**30.07**
Sexo (1=homens; 0=mulheres)	-	0.29914
	-	**56.03**
Cor (1=cor branca; 0=caso contrário)	-	0.26297
	-	**43.59**
Constante	-0.79963	-1.06887
	-45.32	**-60.37**
R2	**0.40**	**0.44**
P>F	**0.00**	**0.00**

Nota: As informações em negrito correspondem a estatística t

Além do uso de modelagem é possível descrever um indicador para avaliar o grau de discriminação entre os grupos de cor e sexo. A próxima seção descreve o Coeficiente de Discriminação proposto por Becker (1962) e sua generalização descrita por Oaxaca (1973).

4.5 Coeficiente de Discriminação e Procedimentos para o Cálculo da Decomposição de Salários

O Coeficiente de Discriminação apresentado por Becker é definido como a porcentagem do diferencial salarial entre dois tipos de mercados perfeitamente substitutos. Para os casos nos quais os dois fatores não são necessariamente substitutos perfeitos, Becker definiu o coeficiente de discriminação como uma simples diferença entre os salários e a razão dos salários na ausência de discriminação.

Atribui-se como grupo base (L) os homens de cor branca para comparações e L-1 *grupos* definidos pelos homens de cor preta/parda, mulheres de cor branca e mulheres de cor preta/parda.

$$D = \frac{\bar{Z}_L / \bar{Z}_l - \left(\bar{Z}_L / \bar{Z}_l\right)^0}{\left(\bar{Z}_L / \bar{Z}_l\right)^0}, \text{ l =1, 2, 3 e L} \qquad (4.3)$$

onde:

- $\bar{Z}_l$ = média do salário/hora no l-ésimo grupo;
- l = 1(homens de cor preta/parda), 2 (mulheres de cor branca), 3 (mulheres de cor preta/parda) e L (homens de cor branca ou grupo base);
- $\left(\bar{Z}_L / \bar{Z}_l\right)$ = razão entre as médias do salário/hora do grupo base (L) e o l-ésimo grupo;
- $\left(\bar{Z}_L / \bar{Z}_l\right)^0$ = razão entre as médias do salário/hora do grupo base(L) e o l-ésimo grupo na ausência da discriminação;

Oaxaca (1973) apresenta uma generalização da medida de Becker admitindo substitutos perfeitos como caso especial (sendo mais flexível para trabalhos empíricos). Para isso, aplica-se o logaritmo na expressão 4.3.

$$\ln(D) = \ln\left[\frac{\bar{Z}_L / \bar{Z}_I - \left(\bar{Z}_L / \bar{Z}_I\right)^0}{\left(\bar{Z}_L / \bar{Z}_I\right)^0}\right] = \ln(D+1) = \ln\left(\bar{Z}_L / \bar{Z}_I\right) - \ln\left(\bar{Z}_L / \bar{Z}_I\right)^0 \quad (4.4)$$

Como não conhecemos $\left(\bar{Z}_L / \bar{Z}_I\right)^0$, para estimar D faz-se necessário obter valores para $\left(\bar{Z}_L / \bar{Z}_I\right)^0$. A estimativa é realizada considerando ou não a ausência de discriminação. Se não existisse discriminação, a estrutura salarial de um determinado grupo padrão poderia ser aplicada aos demais grupos, ou a estrutura destes grupos poderia ser aplicada ao grupo base.

A estimação pelo Método de Máxima Pseudo-Verossimilhança da equação de salários para determinados grupos de trabalhadores, fornece uma estimativa da estrutura salarial que pode ser aplicada a outros grupos. Este procedimento consiste em estimar separadamente para cada grupo de cor/sexo a equação de Mincer. A definição do modelo é dada pela expressão 5.5.

$$\ln(Z_{li}) = \underline{X}_i' \underline{\beta}^{(l)} + \varepsilon_{li}, \quad \varepsilon_{li} \sim N(0,1) \quad (4.5)$$

Onde:

- $\ln(Z_{li})$ = logaritmo do salário/hora para o i-ésimo indivíduo pertencente ao l-ésimo grupo de cor e sexo;
- $\underline{X}_{li}'$ = matriz das variáveis explicativas (escolaridade, experiência) do i-ésimo indivíduo do l-ésimo grupo de cor e sexo;
- $\underline{\beta}^{(l)}$ = parâmetros a serem estimados do l-ésimo grupo de cor e sexo;
- ε_{li} = erro aleatório

Através da Análise de Regressão, equação de salários, é possível inferir para cada grupo de cor/sexo quais os retornos para cada característica produtiva. Isto é, através dos coeficientes estimados (betas). Com isso seja:

$$G = \frac{\bar{Z}_L - \bar{Z}_I}{\bar{Z}_I} \quad (4.6)$$

Aplicando o logaritmo na expressão (4.6):

$$\ln(G+1) = \ln\left(\bar{Z}_L\right) - \ln\left(\bar{Z}_I\right) = \bar{X}_L' \hat{\beta}_L - \bar{X}_I' \hat{\beta}_I \quad (4.7)$$

Onde $\bar{Z}_L$ e $\bar{Z}_L$ são as médias dos logaritmos dos salários/hora para o grupo-padrão e do l-ésimo grupo de cor/sexo. A partir das propriedades dos estimadores de máximo pseudo-versossimilhança, podemos escrever:

Efeito devido às diferenças das características produtivas

$$\ln\left(\frac{Z_L}{Z_I}\right)^0 = \Delta\bar{X}' \hat{\beta}_I \quad (4.8)$$

Efeitos devido às diferenças da discriminação/retornos

$$\ln(G+1) = -\bar{X}_L' \Delta\hat{\beta} \quad (4.9)$$

4.6 Ajuste das Equações de Salários para os grupos de cor e sexo

Para a realização do procedimento de Oaxaca e obtenção dos efeitos descritos na seção 4.5 é necessário ajustar as equações de salários para os diferentes grupos de cor se sexo. Utilizou-se o programa estatístico STATA 8.0, procedimento svyregress, para o ajuste do modelo de regressão com duas finalidades: ajustar o modelo com o peso e sem a especificação dos estratos (tabela 4.2) e a forma adequada com a especificação dos pesos, estratos e unidades primárias de amostragem (tabela 4.3).

A comparação entre as tabelas 4.2 e 4.3 revela que o fato de não incorporar a estratificação no ajuste dos modelos não altera os valores dos coeficientes, mas as estatísticas de teste (t) são afetadas pela não especificação no ajuste do modelo. Como exemplo, enquanto que a estatística de teste para a variável escolaridade na tabela 4.2 é ***13.18*** na tabela 4.3 é **10.41** e isso significa que o desvio padrão no primeiro caso é menor quando estimamos os coeficientes considerando que tenham sido dados obtidos através de uma amostra aleatória simples. Na tabela 5.3, a interpretação dos coeficientes de escolaridade e experiência de trabalho é realizada em conjunto com os termos quadráticos. Como exemplo, o efeito marginal dos anos destudo das mulheres de cor branca (denominado por *Ef marg MB*) sobre a média do ln(salário/hora) das mulheres é obtido pela expressão:

$$Ef\ marg\ MB = -0.0191 * anos\ de\ estudo + 0.0098 * (anos\ de\ estudo^2) \quad (4.10)$$

Tabela 4.2 - Equações de salários estimadas com o peso e sem a especificação dos estratos. Estimativas dos parâmetros e respectivas estatísticas de teste (negrito)

Variáveis	Cor branca		Cor Preta ou Parda	
	Homens	Mulheres	Homens	Mulheres
Anos de estudo	0,0519	-0,0191	0,0643	0,0261
	13,18	**-3,65**	**19,26**	**5,38**
Anos de estudo2	0,0065	0,0098	0,0046	0,0069
	28,10	**34,96**	**18,74**	**22,87**
Experiência	0,0548	0,0382	0,0524	0,0460
	60,78	**36,00**	**58,18**	**37,25**
Experiência2	-0,0006	-0,0004	-0,0006	-0,0005
	-35,28	**-18,84**	**-40,71**	**-22,48**
Área Urbana	0,2719	0,1972	0,3408	0,3494
	21,02	**9,77**	**32,72**	**18,31**
Constante	-0,8393	-0,5315	-1,0332	-1,1365
	-39,59	**-16,63**	**-60,84**	**-40,20**
R^2	0,44	0,42	0,34	0,32
Prob>F	0,0001	0,0001	0,0001	0,0001
Amostra	44.311	31.115	46.257	28.129
População	22.080.841	15.297.084	19.311.338	11.455.643

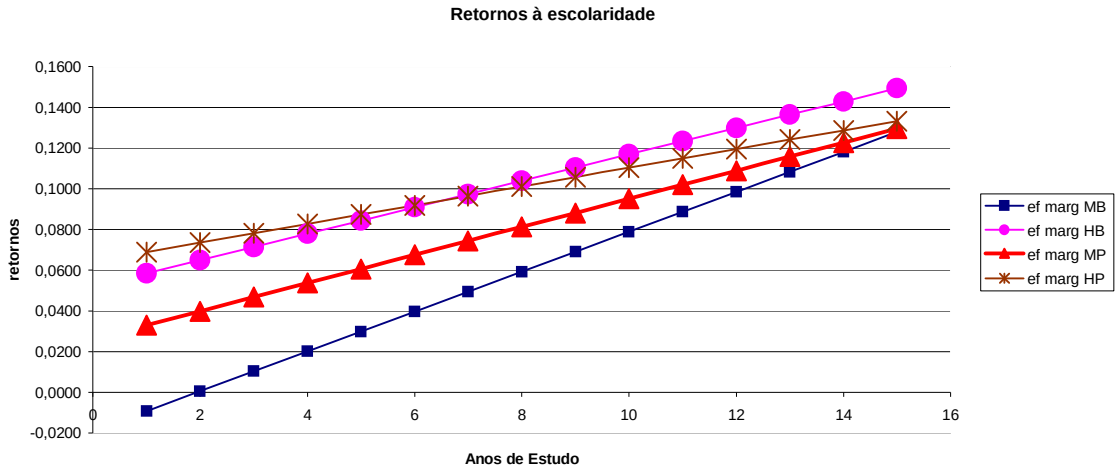
Tabela 4.3 - Equações de Salários para os grupos de cor e sexo
Estimativas dos parâmetros e respectivas estatísticas de teste (negrito)

Variáveis	Cor branca		Cor Preta ou Parda	
	Homens	Mulheres	Homens	Mulheres
Anos de estudo	0,0519	-0,0191	0,0643	0,0261
	10,41	**-3,29**	**15,39**	**4,18**
Anos de estudo2	0,0065	0,0098	0,0046	0,0069
	22,26	**30,98**	**15,61**	**18,95**
Experiência	0,0548	0,0382	0,0524	0,0460
	56,97	**35,64**	**47,73**	**35,07**
Experiência2	-0,0006	-0,0004	-0,0006	-0,0005
	-33,58	**-18,44**	**-32,06**	**-21,65**
Área Urbana	0,2719	0,1972	0,3408	0,3494
	12,57	**7,47**	**17,36**	**10,43**
Constante	-0,8393	-0,5315	-1,0332	-1,1365
	-27,50	**-14,10**	**-35,88**	**-24,73**
R^2	0,44	0,42	0,34	0,32
Prob>F	0,0001	0,0001	0,0001	0,0001
Amostra	44.311	31.115	46.257	28.129
População	22.080.841	15.297.084	19.311.338	11.455.643

Os gráficos 4.1 e 4.2 apresentam os efeitos marginais sobre o ln(salário/hora) para a escolaridade e experiência, respectivamente. Vale ressaltar que, como o efeito marginal é a derivada da equação de salários ele é uma função linear. No gráfico 5.1, as mulheres de cor branca apresentam os menores retornos em todos os anos de estudo ao comparar com outros de grupos de sexo/cor. Além disso, exceto para um ano de estudo, o valor dos retornos é positivo conforme a teoria do capital humano (quanto maior a escolaridade maior o rendimento).

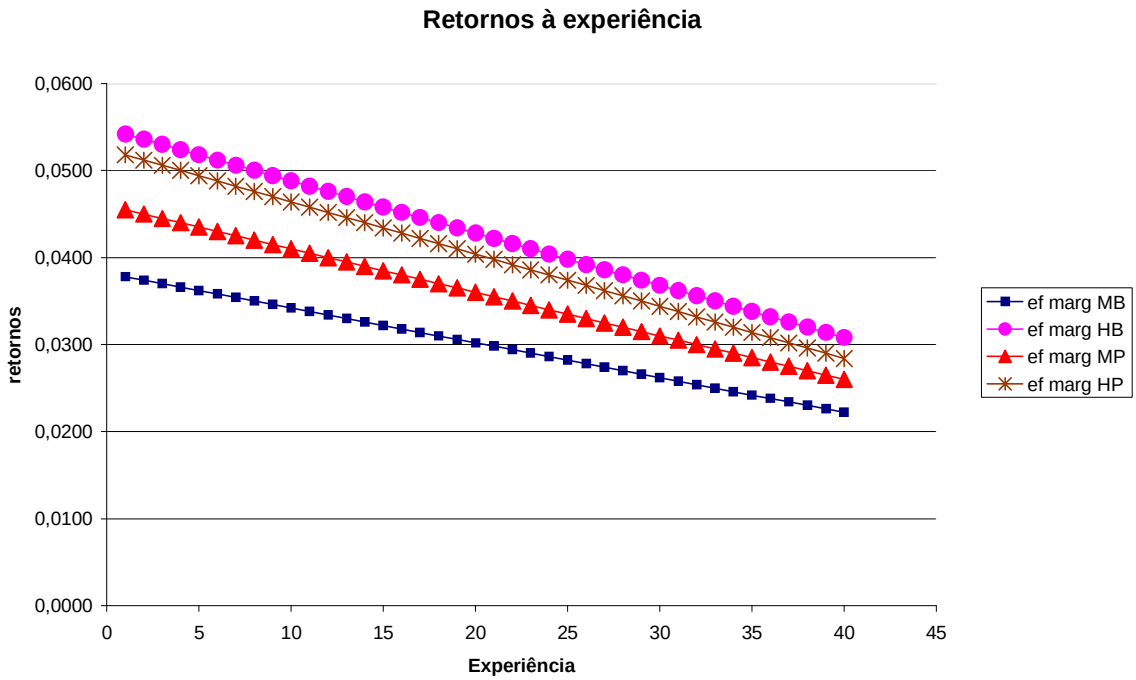
Os homens apresentam os maiores retornos quando comparados às mulheres. Os retornos dos homens de cor branca superam o dos homens de cor preta/parda a partir dos seis anos de estudo.

Gráfico 4.1 – Retornos à escolaridade sobre o ln(salário/hora)



Notas: ef marg MB = efeito marginal das mulheres cor branca, ef marg HB = efeito marginal dos homens de cor branca, ef marg MP = efeito marginal das mulheres de cor preta/parda e ef marg HP = efeito marginal dos homens de cor preta/parda.

Gráfico 4.2 – Retornos à experiência sobre o ln(salário/hora)



4.7 Análise da Equação de Salários com interação

Na seção anterior os resultados foram obtidos de forma independente para grupo de cor e sexo. No entanto, para o cálculo do efeito devido às diferenças dos retornos (discriminação – expressão 4.9) torna-se necessário o cálculo do desvio-padrão incorporando o plano amostral.

A solução para obter a diferença entre os coeficientes dos grupos e o respectivo desvio-padrão com a incorporação do plano amostral é a estimação da equação de salários com interação das variáveis explicativas: escolaridade, experiência, idade e local de residência com o sexo e a cor dos trabalhadores.

Os resultados da tabela 4.4 em negrito representam as estimativas dos coeficientes para o grupo base L (homens de cor branca). As linhas em itálico representam a diferença entre o grupo base (L) e o l-ésimo grupo (homens de cor preta/padra, mulheres de cor branca e mulheres de cor preta/parda).

Como exemplo, o coeficiente para os anos de estudo dos homens de cor branca é **0,0519** (tabela 4.2 e tabela 4.3) enquanto que o para os homens de cor preta/parda é obtido por: **0,0519** + 0,0124 (tabela 5.3) = **0,0643** (tabela 5.2). A amostra considera todos os grupos, aproximadamente 149 mil trabalhadores.

Tabela 4.4 – Ajuste do Modelo de Regressão da Equação dos Salários com a interação sexo*cor e as variáveis explicativas anos de estudo, experiência e local de residência.

Variáveis		Estimativas	Estatística-t
Anos de estudo (Homens de cor branca)		0,0519	**10,41***
	Homens de cor preta/parda	0,0124	**2,13***
	Mulheres de cor branca	-0,0711	**-10,24***
	Mulheres de cor preta/parda	-0,0258	**-3,34***
Anos de estudo^2 (Homens de cor branca)		0,0065	**22,26***
	Homens de cor preta/parda	-0,0020	**-5,21***
	Mulheres de cor branca	0,0033	**8,7***
	Mulheres de cor preta/parda	*0,0004*	***0,78***
Experiência (Homens de cor branca)		0,0548	**56,97***
	Homens de cor preta/parda	*-0,0024*	***-1,67***
	Mulheres de cor branca	-0,0165	**-11,75***
	Mulheres de cor preta/parda	-0,0088	**-5,47***
Experiência ^{2} (Homens de cor branca)		-0,00059	**-33,58***
	Homens de cor preta/parda	-0,00004	**-1,53**
	Mulheres de cor branca	0,00016	**5,82***
	Mulheres de cor preta/parda	*0,00005*	***1,57***
Área Urbana (Homens de cor branca)		0,2719	**12,57***
	Homens de cor preta/parda	0,0689	**2,71***
	Mulheres de cor branca	-0,0747	**-2,75***
	Mulheres de cor preta/parda	0,0775	**2,09***
Constante (Homens de cor branca)		-0,8393	**-27,5***
	Homens de cor preta/parda	-0,1939	**-5,2***
	Mulheres de cor branca	0,3078	**7,28***
	Mulheres de cor preta/parda	-0,2972	**-5,7***
R^2		0,4457	
Prob>F		0,0000	
Amostra		149.812	
População		68.144.906	

Notas: * significativo ao nível de 5% de confiança.

4.8 Resultados da Decomposição do Diferencial de Salários

O objetivo desta seção é detalhar os procedimentos para decomposição da média do logaritmo do salário/hora através da análise controlada (modelagem estatística). Em seguida, investigar o quanto desta diferença é explicada pelas características individuais dos trabalhadores é quanto é proveniente dos retornos dos coeficientes.

Os resultados da tabela 4.5 avaliam se a diferença entre a média do logaritmo do salário/hora do grupo base (homens de cor branca) e o l-ésimo grupo de cor e sexo é estatisticamente significativa.

Tabela 4.5 – Diferença do ln(salário/hora) entre os grupos de cor

Diferenças		Estimativa	Estatística t
ln(salário/hora)	Homens de cor branca - Homens de cor preta	0.59	50.72
	Homens de cor branca - Mulheres de cor branca	0.15	18.83
	Homens de cor branca - Mulheres de cor preta ou parda	0.69	56.74

Fonte: Processado a partir dos microdados da PNAD 2003.

O cálculo da diferença entre as médias das características produtivas (expressão 4.8) foi obtido com a incorporação do plano amostral da PNAD (anexo 3). Esta análise tem por finalidade testar se a diferença entre as características dos grupos de cor e sexo são estatisticamente diferentes de zero.

Verifica-se na tabela 4.6 que os homens têm, em média, uma escolaridade inferior as mulheres de cor branca.

Tabela 4.6 – Diferença entre as médias das características produtivas e coeficientes estimados para o Grupo Base (L) e o l-ésimo grupo de cor/sexo

Características	(Grupo padrão) - (l-ésimo grupo de cor/sexo)					
	Entre as médias			Entre os coeficientes estimados		
	Mulher de cor branca	Mulher de cor preta ou parda	Homem de cor preta ou parda	Mulher de cor branca	Mulher de cor preta ou parda	Homem de cor preta ou parda
*Anos de estudo	-1,32	0,80	2,32	-0,0711	-0,02578	0,0124
	-36,17	**15,32**	**44,55**	**-10,2**	**-3,3**	**2,1**
Anos de estudo^2	-21,83	12,77	33,41	0,0033	0,00035	-0,0020
	-36,92	**15,87**	**44,21**	**8,7**	**0,8**	**-5,2**
Experiência	2,84	1,31	-0,71	-0,0165	-0,00883	-0,0024
	23,89	**9,92**	**-5,35**	**-11,8**	**-5,5**	**-1,7**
Experiência^2	159,82	85,74	-40,09	0,0002	0,00005	-0,000039
	23,66	**11,11**	**-4,91**	**5,8**	**1,57**	**-1,5**
Área Urbana	-0,07	-0,03	0,06	-0,0747	0,07745	0,0689
	-21,54	**-5,46**	**9,77**	**-2,8**	**2,1**	**2,7**
Constante				0,3078	-0,29722	-0,1939
				7,28	**-5,70**	**-5,20**

Com base nos resultados das tabelas 4.5 e 4.6 é possível a construção da decomposição do diferencial da média do ln(salário/hora) de forma adequada. A finalidade da tabela 4.7 é avaliar o quanto das características produtivas explicam o diferencial e o que é proveniente da discriminação. Entre as pessoas de cor branca, homens e mulheres, as características produtivas explicam –107,1% e o efeito da discriminação explica 207% da diferença entre a média do ln(salário/hora). Para esta análise vale ressaltar que o efeito da discriminação (0,31) é maior que a diferença dos logaritmos dos salários (0,15), pois as mulheres têm mulheres atributos pessoais

(escolaridade media superior) quando comparados aos homens, e se fossem remuneradas igualmente seus rendimentos seriam maiores.

A diferença da média do ln(salário/hora) entre os trabalhadores do grupo e as pessoas de cor preta e parda revela que as mulheres sofrem uma maior discriminação. Enquanto que o efeito da discriminação explica 84% do diferencial entre as mulheres de cor branca e o grupo base, entre os homens é responsável por 47%.

Tabela 4.7 - Efeitos da Discriminação Estimados pelas Características Pessoais

	Efeitos	Mulher branca		Homem preto		Mulher preta	
		(1)a	(2)b	(1)a	(2)b	(1)a	(2)b
	Diferencial de salários = (3)	**0,152**	**100,0%**	**0,593**	**100,0%**	**0,688**	**100,0%**
	Anos de estudo	0,0254	16,7%	0,1489	25,1%	0,0209	3,04%
	Anos de estudo2	-0,2143	-140,9%	0,1526	25,7%	0,0878	12,76%
	Experiência	0,1085	71,4%	-0,0371	-6,3%	0,0602	8,74%
	Experiência2	-0,0686	-45,1%	0,0253	4,3%	-0,0467	-6,79%
	Área Urbana	-0,0138	-9,1%	0,0212	3,6%	-0,0108	-1,58%
	Somatório do efeitos = (4)	**-0,1629**	**-107,1%**	**0,3110**	**52,4%**	**0,1113**	**16,2%**
Efeito Discriminação	**Estimativa de ln(D+1) = (5)**	**0,3150**	**207,1%**	**0,2825**	**47,6%**	**0,577**	**83,8%**
	Estimativa de D = (6)	0,37		0,33		0,78	

Notas: (a) é igual ao produto entre o coeficiente da variável explicativa e a diferença das características médias da tabela 5.5

(b) é igual a coluna (a) expressa como percentual da diferencial de salários.

(3) Representa o diferencial de salários entre o grupo padrão e o l-ésimo grupo de cor/sexo

(4) Somatório dos efeitos sobre o diferencial de salários

(5) É igual a (3) - (4)

(6) É igual a exponencial do item (5)

A estimativa do coeficiente de discriminação (D), por sua vez, sintetiza os resultados da análise controlada da discriminação presente no mercado de trabalho brasileiro em 2003. A conclusão é que a discriminação no mercado de trabalho é mais evidente contra as mulheres de cor preta. Desta forma, considerando como grupo base os homens de cor branca, o maior valor registrado de D foi de 0,78 entre as mulheres de cor preta e parda e o grupo base. Em seguida, 0,37 para as mulheres de cor branca e 0,33 para os homens de cor preta e parda.

Esta seção desenvolveu os procedimentos teóricos e operacionais para o cálculo da decomposição da média do ln(salário/hora) em pesquisas como a PNAD 2003. Foi identificado que a melhor maneira para estimar as equações de salários é através de um modelo de interação que já fornece as estimativas das diferenças dos coeficientes com a incorporação do plano amostral e facilita na construção do quadro de decomposição. A próxima seção tem por finalidade avaliar corrigir os possíveis vieses de seleção da informação do salário em pesquisas como a PNAD 2003 que registra os salários daqueles que trabalham.

5 Decomposição do Diferencial de Salário com Correção do Viés de Seleção do Salário na Análise de Regressão

5.1 O viés de seleção da informação dos salários

Esta seção tem por finalidade apresentar os procedimentos necessários para corrigir o viés de seleção das informações dos salários reportados pelos indivíduos. Com isso, avaliar os impactos nas equações de salários e nos resultados da decomposição de salários. Com relação à variável resposta, ln(salário/hora), vale ressaltar que na PNAD, como em outras pesquisas, as informações coletadas são fornecidas pelas pessoas que tinham trabalho na época da pesquisa. Isto é, os salários observados na PNAD 2003 estão relacionados com a decisão de um indivíduo trabalhar ou não.

De fato, para determinadas pessoas seria vantajoso trabalhar se o salário recebido (ou salário potencial) fosse maior que o custo de oportunidade (ou salário reserva). Desta forma, existe um viés de seleção das informações do salário. Neste caso, as estimativas dos coeficientes da equação de salários obtidas a partir das informações dos indivíduos que trabalham na data de referência da pesquisa podem estar viesadas sob a ótica do **viés de seleção das informações,** considerando-se que o modelo utilizada na seção 4 não incorpora a informação sobre a avaliação dos indivíduos que não trabalham no que se refere ao custo de oportunidade.

Heckman (1979) apresentou o fenômeno e sua solução não apenas para pesquisas amostrais. De modo geral, o viés de seleção de informação pode ser decorrente de duas razões, como destaca Heckman: ou em virtude de seletividade das informações dos indivíduos ou devido ao desenho amostral da pesquisa. Cabe ressaltar que, ao utilizar os dados da PNAD, o efeito que deseja-se incorporar na modelagem visa corrigir o viés de seletividade da informação dos salários para os indivíduos que, apesar de estarem devidamente representados na amostra da PNAD 2003, não trabalhavam na data de referência da pesquisa supostamente devido a uma avaliação do salário potencial e do custo de oportunidade envolvido nesta escolha.

5.2 A Correção do Viés de Seletividade da Informação dos Salários através do Procedimento de Heckman em Pesquisas Amostrais Complexas

Conforme apresentado na seção anterior, o problema de estimar a equação de salários é que não observamos o salário para toda a amostra, mas apenas para aqueles que trabalham. A metodologia apresentada é um resumo dos trabalhos de Heckman (1979) e Kassouf (1994) e sua motivação tem como base a incorporar a complexidade do desenho amostral (seção 3) na estimação dos modelos.

A equação de participação avalia a probabilidade do indivíduo trabalhar segundo algumas varáveis explicativas[12]. O modelo utilizado é o probit com a incorporação do plano amostral. A variável dependente assume o valor “1” se o indivíduo tem rendimento (ocupado) e “0” caso contrário (descocupados ou inativos).

Com base nos trabalhos de Heckman (1979) e Kassouf (1994) as variáveis explicativas selecionadas foram:

- Escolaridade em anos de estudo;
- Escolaridade ao quadrado;
- Experiência
- Experiência ao quadrado
- Chefe - condição no domicílio chefe;

[12] Ressalta-se que para identificação é necessário que algumas destas variáveis não estejam incluídas na equação de salários.

- Filho – condição no domicílio filho;
- Crianças no domicílio entre 0 e 5 anos – assume o valor 1 se há crianças e o caso contrário.

Os resultados da equação de salários – modelo probit – estão descritos na tabela 5.1. Verifica-se que conforme aumenta a experiência de trabalho maior é a probabilidade de um indivíduo participar no mercado de trabalho. O sinal negativo do termo quadrático da experiência indica que a probabilidade de participar no mercado de trabalho cresce a taxas decrescentes. O efeito do termo linear da escolaridade em todos os grupos de cor reflete que a probabilidade aumenta conforme os anos de estudo. No entanto, o termo quadrático indica que para os homens a probabilidade cresce a taxas decrescentes e as mulheres a taxas crescentes. Ser chefe do domicílio é uma variável importante para determinar a probabilidade do indivíduo particpar do mercado de trabalho. Com isso, independente do sexo e a da cor a probabilidade de participar no mercado de trabalho de chefes de família é positiva em todos os grupos de cor e sexo. Para os indivíduos que declararam ter uma relação de filho com o chefe do domicílio o comportamento é o contrário. Isto é, o sinal negativo expressa que estes indivíduos estão menos propensos a participar do mercado de trabalho.

Além das características pessoais e responsabilidades no domicílio foi incluída uma informação que tem impacto sobre todas as pessoas que vivem no mesmo domicílio: *crianças de 0 a 5 anos de idade*. Enquanto os homens estão mais propensos a trabalhar, as mulheres têm um comportamento contrário. Heckman (1979) e Kassouf (1994) justificam este comportamento pela regra de decisão entre o salário potencial e o salário reserva descrito na seção anterior.

Tabela 5.1 – Equação de Participação no Mercado de Trabalho Brasileiro em 2003

Estimadores	Cor branca		Cor Preta ou Parda	
	Homens	Mulheres	Homens	Mulheres
Anos de estudo	0,1484	0,0752	0,1027	0,0560
	30,05	**16,49**	**19,38**	**11,36**
Anos de estudo2	-0,0043	0,0017	-0,0001	0,0044
	-12,14	**5,51**	**-0,31**	**11,26**
Experiência	0,0944	0,0745	0,1087	0,0911
	63,52	**50,16**	**74,94**	**65,62**
Experiência2	-0,0016	-0,0014	-0,0017	-0,0015
	-61,00	**-48,44**	**-61,37**	**-56,98**
Área Urbana	-0,2875	0,2482	-0,1858	0,2398
	-13,75	**9,57**	**-8,47**	**10,28**
Chefe do domicílio	0,5883973	0,4216303	0,5646465	0,4643732
	25,89	**23,2**	**25,7**	27,02
Filho	-0,26	-0,12	-0,23	-0,05
	-14,37	**-6,56**	**-13,83**	**-2,81**
Criança de 0 a 5 anos	0,02	-0,18	0,04	-0,21
	1,53	**-12,79**	**3,01**	**-14,27**
Constante	-1,1674	-1,7858	-1,3662	-2,0279
	-43,90	**-54,57**	**-47,77**	**-62,18**
Amostra	87.601	96.528	98.593	99.112

A partir dos coeficientes da equação de participação (modelo probit) – tabela 5.1 – é calculado a variável lambda (ou razão inversa de Mills) e utilizado como variável explicativa para estimação da equação de salários que é apresentado na próxima seção.

5.2.1 Ajuste das Equações de Salário com Correção do Viés de Seletividade da Informação.

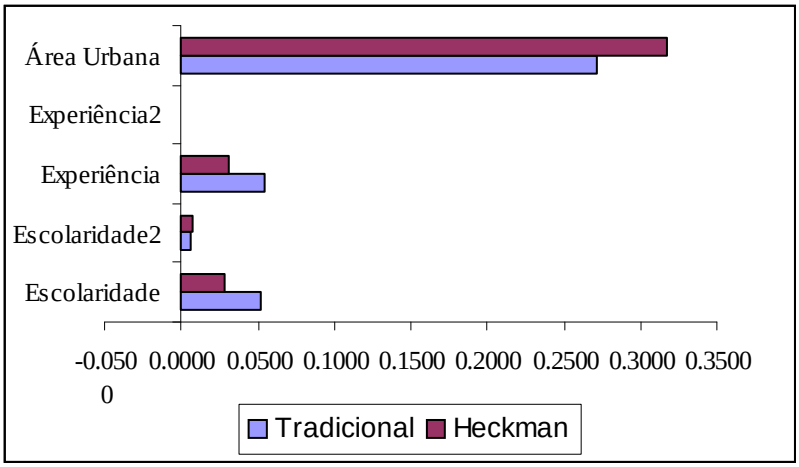
Os resultados da tabela 5.2 apresentam as equações de salários com a correção de Heckman. Vale ressaltar que o coeficiente da variável lambda é estatisticamente significativo, indicando que a inclusão é necessária para correção do viés de seleção da informação do salário.

Tabela 5.2 – Equação de Salários com Correção de Heckman

Estimadores	Cor branca		Cor Preta ou Parda	
	Homens	Mulheres	Homens	Mulheres
Anos de estudo	0,0280	-0,0257	0,0467	0,0217
	5,42	**-3,98**	**10,88**	**3,44**
	0,0051697	0,0064646	0,0042961	0,0063076
Anos de estudo2	0,0072	0,0098	0,0046	0,0067
	24,06	**30,82**	**15,66**	**18,10**
	0,0003012	0,000318	0,0002966	0,0003717
Experiência	0,0311	0,0324	0,0255	0,0401
	24,51	**11,71**	**14,54**	**17,94**
	0,0012673	0,0027657	0,0017507	0,0022357
Experiência2	-0,0002	-0,0003	-0,0002	-0,0005
	-9,96	**-6,28**	**-8,62**	**-11,94**
	0,0000214	0,0000517	0,0000275	0,000038
Área Urbana	0,3169	0,1750	0,3711	0,3312
	14,42	**6,1**	**18,89**	**9,93**
	0,0219826	0,0286843	0,0196445	0,0333521
Lambda	-0,3456	-0,1068	-0,3569	-0,0908
	-24,38	**-2,32**	**-19,39**	**-3,10**
	0,0141757	0,0460311	0,0184066	0,0293129
Constante	-0,3256	-0,3040	-0,4560	-0,9317
	-9,12	**-2,87**	**-11,78**	**-12,92**

O sinal negativo da variável *lambda* indica que os fatores não mensurados, na equação de salários, por um aumentam a probabilidade de participação diminuem os retornos do salário. Como exemplo, nos gráficos 6.1 e 6.2, para os homens de cor branca, o coeficiente da escolaridade pelo método tradicional era 0.052 (seção 5) e com a correção de Heckman passa a ser 0.028 (tabela 6.2). Quando realizamos a mesma análise para a variável experiência de trabalho, para todos os grupos de cor e sexo, os coeficientes também sofrem reduções. A comparação com o método tradicional (tabela 5.4) revela que tanto os coeficientes, quanto as variâncias, se alteram com a utilização deste procedimento. Mesmo assim todas os coeficientes continuam significativos a 95% de confiança.

Gráfico 5.1 – Comparação entre os coeficientes da Equação de Salários pelo Método Tradicional e com a Correção de Heckman.



De posse das estimativas da equação de salários corrigidas do viés de seleção da informação do salário o próximo passo é decompor o diferencial entre os grupos de trabalhadores na próxima seção.

5.3 Decomposição dos Diferenciais de Salários segundo Oaxaca

Os resultados da tabela 5.3 são referentes às diferenças entre as médias das características produtivas e a diferença entre os coeficientes. Vale destacar que, por limitações operacionais o STATA não incorpora o procedimento de interação no Heckman incluindo a variável lambda. Desta forma, a hipótese adotada é que os dados são "paired data" entre os regressores, ou seja, não há um correspondente homem negro (ou l-ésimo grupo de cor e sexo) para cada homem branco (grupo base). Seja:

$$Var\left(\underline{\beta}^{(L)} - \underline{\beta}^{(l)}\right) = Var\left(\underline{\beta}^{(L)}\right) - Var\left(\underline{\beta}^{(l)}\right) - 2Cov\left(\underline{\beta}^{(L)} - \underline{\beta}^{(l)}\right) \quad (5.1)$$

Onde:

$\underline{\beta}^{(L)}$ = coeficiente do grupo base

$\underline{\beta}^{(l)}$ = coeficiente do l-ésimo grupo de cor e sexo.

Assumindo que:

$Cov\left(\underline{\beta}^{(L)} - \underline{\beta}^{(l)}\right)$

Tabela 5.3 – Diferenças entre as médias dos regressores e coeficientes estimados com a correção de Heckman.

	(Grupo padrão) - (l-ésimo grupo de cor/sexo)					
	Entre as médias			Entre os coeficientes estimados		
	Mulher de cor branca	Mulher de cor preta ou parda	Homem de cor preta ou parda	Mulher de cor branca	Mulher de cor preta ou parda	Homem de cor preta ou parda
Anos de estudo	-1,32	0,80	2,32	-0,0537	0,0187	-0,0064
	-36,17	**15,32**	**44,55**	**-4,6193**	**1,9769**	**-0,5544**
				0,0116	0,0095	0,0115
Anos de estudo^2	-21,83	33,41	33,41	0,0026	-0,0026	-0,0005
	-36,92	**15,87**	**44,21**	**4,1266**	**-4,3515**	**-0,7658**
				0,0006	0,0006	0,0007
Experiência	2,84	-0,71	-0,71	0,0013	-0,0056	0,0090
	23,89	**9,92**	**-5,35**	**0,3283**	**-1,8592**	**2,5816**
				0,0040	0,0030	0,0035
Experiência^2	159,82	-40,09	-40,09	-0,0001	0,0000	-0,0002
	23,66	**11,11**	**-4,91**	**-1,5198**	**-0,4806**	**-4,0455**

				0,0001	0,0000	0,0001
Área Urbana	-0,07	0,06	0,06	-0,1420	0,0542	0,0143
	-21,54	**-5,46**	**9,77**	**-2,8024**	**1,3020**	**0,2583**
				0,0507	0,0416	0,0553
Lambda	-0,36	-0,42	-0,42	0,2388	-0,0114	0,2547
				3,9667	**-0,3489**	**5,8578**
				0,0602	0,0326	0,0435
Constante		-	-	0,0216	-0,60614	-0,1305

Na tabela 5.4, o diferencial observado entre homens de cor branca e mulheres de cor preta/parda o efeito da discriminação explica 97% contra apenas 3% das características pessoais. Entre homens e mulheres de cor branca o efeito da discriminação explica 175% do diferencial de salários.

Tabela 5.4 – Efeitos da Discriminação Estimados pelas Características Pessoas **–** Correção de Heckman

	Efeitos	Mulher branca		Homem preto		Mulher preta	
		(1)a	(2)b	(1)a	(2)b	(1)a	(2)b
	Diferencial de salários = (3)	**0,152**	**100,0%**	**0,593**	**100,0%**	**0,688**	**100,0%**
	Anos de estudo	0,0341	22,4%	0,1083	18,3%	0,00040551	0,05891%
	Anos de estudo2	-0,2139	-140,7%	0,1551	26,1%	-0,00001751	-0,00254%
	Experiência	0,0919	60,4%	-0,0180	-3,0%	-0,00022505	-0,03269%
	Experiência2	-0,0519	-34,1%	0,0095	1,6%	0,00000001	0,00000%
	Área Urbana	-0,0123	-8,1%	0,0231	3,9%	0,01795235	2,60798%
	Lambda	0,0382	25,1%	0,1488	25,1%	0,00103260	0,15001%
	Somatório do efeitos = (4)	**-0,1139**	**-74,9%**	**0,4268**	**71,9%**	**0,0191**	**2,8%**
Efeito Discriminação	**Estimativa de ln(D+1) = (5)**	**0,27**	**174,9%**	**0,17**	**28,1%**	**0,67**	**97,2%**
Coeficiente de Discriminação	**Estimativa de D = (6)**	0,30		0,18		0,95	

Notas: (a) é igual ao produto entre o coeficiente da variável explicativa e a diferença das características médias da tabela 5.5

(b) é igual o ajuste de (a) expresso como percentual da diferencial de salários.

(3) Representa o diferencial de salários entre o grupo padrão e o l-ésimo grupo de cor/sexo

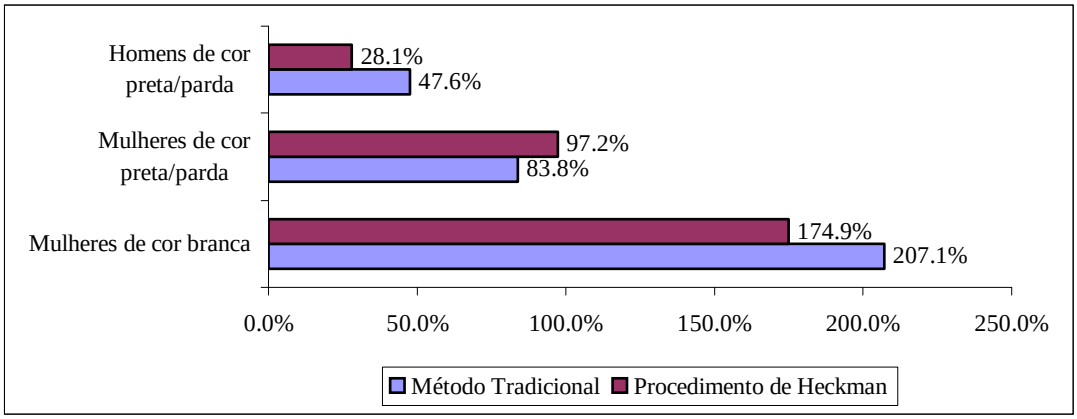
(4) Somatório dos efeitos sobre o diferencial de salários

(5) É igual a (3) - (4)

(6) É igual a exponencial do item (5)

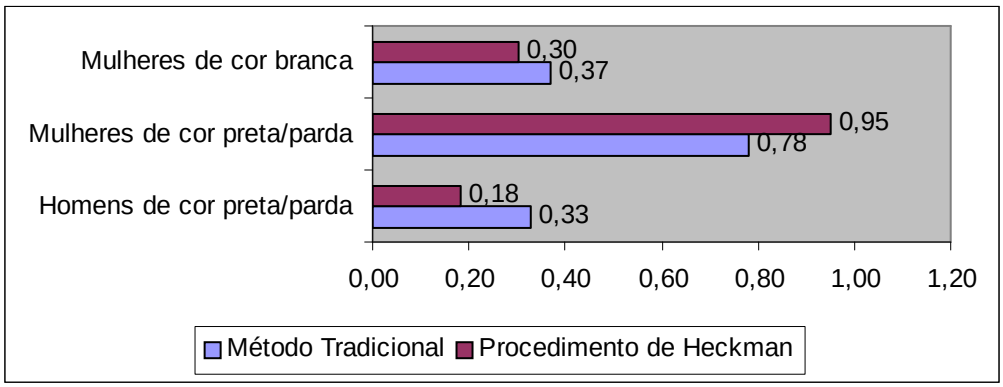
O gráfico 5.2 revela que se a decomposição fosse realizada pelo método tradicional a conclusão seria que 84% do diferencial entre homens de cor branca e mulheres de cor preta é provenientes da discriminação. No entanto, com a correção de Heckman este indicador é de 97%, ou seja, estaríamos subestimando o efeito da discriminação. Destaca-se o efeito da discriminação entre homens e mulheres de cor branca que cai 7 pontos percentuais com a correção de Heckman.

Gráfico 5.2 - Efeito da Discriminação – Método Tradicional vs Heckman



A estimativa do coeficiente de discriminação, pelo método tradicional, estaria subestimada pelo método tradicional (gráfico 6.3), exceto na comparação entre os homens de cor branca e as mulheres de cor preta cujo coeficiente aumentou de 0.78 (método tradicional) para 0.95 (correção de Heckman).

Gráfico 5.3 – Estimativa do Coeficiente de Discriminação – Método Tradicional vs Procedimento de Heckman



6 **Considerações Finais**

Em geral, as análises da equação de salários que utilizam os microdados da PNAD não incorporam a complexidade da pesquisa. Além disso, os resultados são utilizados para o estudo da decomposição do diferencial de salários. Contudo, esse procedimento pode não ser o mais adequado na avaliação da discriminação no mercado de trabalho brasileiro. Este artigo conclui que o procedimento de Heckman corrige o viés de seleção das informações dos salários sendo adequado para obter estimativas consistentes dos parâmetros da equação. Como exemplo, a estimativa do coeficiente de discriminação a partir da equação de salários sem correção, 0.37 para 0.30 com correção entre homens e mulheres de cor branca. A análise da equação de participação revela que quanto maior a escolaridade mais propensos os indivíduos estão para trabalhar. A discriminação é responsável por 97% do diferencial de salários entre homens de cor branca e as mulheres de cor preta ou parda.

Registra-se como uma extensão natural deste trabalho a análise da decomposição do salário/hora para os diferentes percentis da distribuição do rendimento de trabalho. Esta idéia já foi realizada por Soares (2000) como uma extensão da metodologia de Oaxaca, porém sem incorporar o plano amostral na modelagem. O procedimento consiste em estimar equações de salários por percentis (regressões quantílicas) da

distribuição do salário/hora com a correção do viés de seleção da informação do salário pelo procedimento de Heckman. O que se pretende é avaliar os efeitos da discriminação entre os 10% mais pobres, os 10% mais ricos e comparar os resultados com a média da população.

Referências Bibliográficas

Carvalho, A. P. Decomposição do Diferencial de Salários no Brasil em 2003: Uma aplicação dos procedimentos de Oaxaca e Heckman em Pesquisas Amostrais Complexas, Tese de Mestrado, Escola Nacional de Ciências Estatísticas, 2005

HECKMAN, J.J. Sample Selections Bias as a Specification Error Econometrica, vol. 47, nº 1, 1979.

KASSOUF, A. L. The Wage Rate Estimation Using The Heckman Procedure Revista de Econometria Rio de Janeiro, v.14, nº1, p.89-107, 1994.

MINCER, J. Schooling, Experience, and Earnings. New York 1974

OAXACA, R. Male-Female Wage Differentials in Urban Labor Market. In International Economic Review, v. 14, n.3, p. 693-709. 1973

PESSOA, D.G.C.; SILVA, P.L.N. Análise de dados amostrais complexos. São Paulo: ABE, 1998. (13º Simpósio Nacional de Probabilidade e Estatística – Caxambu) 187p.

RODRIGUES, S.C. Análise da Estrutura Salarial Revelada pela PPV Incorporando peso e plano amostral. Rio de Janeiro, 2003. (Dissertação de Mestrado em Estudos Populacionais e Pesquisas Sociais).

SKINNER, C.J.; HOLT, D e SMITH, T.M.F. Analysis of Complex Surveys. Chichester: John Wiley & Sons, 1989.

Anexo 1

O algoritmo a seguir foi processado pelo SAS e tem por objetivo identificar estratos com apenas uma unidade primário de amostragem e novas construções.

```
DATA DOMICS;
 INFILE 'C:\TEMP_MDADOS\PNAD2003\DOM2003.TXT' LRECL=182 MISSOVER;

   INPUT

   @00005  NDOM     11.    /* IDENTIFICAÇAO DO DOMICILIO       */

            @00005  UF       2.    /* UNIDADE DA FEDERAÇÃO          */
```

```
                @00166  UPA       3.    /* DELIMITAÇÃO DO MUNICÍPIO        */
        @00169  STRAT     7.    /* IDENTIFICAÇÃO DE AUTO E NÃO AUTO */
        @00176  PSU       7.    /* UNIDADE PRIMÁRIA DE AMOSTRAGEM   */

        @00005  V0102     8.    /* NÚMERO DE CONTROLE              */
        @00013  V0103     3.    /* NÚMERO DE SÉRIE                 */
        @00016  V0104     2.    /* TIPO DE ENTREVISTA              */
        @00018  V0105     2.    /* TOTAL DE MORADORES              */
        @00020  V0106     2.    /* TOTAL MORADORES 10 ANOS OU +    */

;

*FILTRO PARA AS ENTREVISTAS REALIZADAS;
IF V0104=1;

Select;
when (strat=110006)  nstrat=119999;
when (strat=110007)  nstrat=119999;
when (strat=110009)  nstrat=119999;
when (strat=130005)  nstrat=130004;
when (strat=130007)  nstrat=130006;
when (strat=140004)  nstrat=140003;
when (strat=150011)  nstrat=159999;
when (strat=150014)  nstrat=159999;
when (strat=150018)  nstrat=159999;
when (strat=160003)  nstrat=169999;
when (strat=160005)  nstrat=169999;
when (strat=170004)  nstrat=179999;

when (strat=170005)  nstrat=179999;
when (strat=170007)  nstrat=179999;
when (strat=170012)  nstrat=179999;
when (strat=170014)  nstrat=179999;
when (strat=210003)  nstrat=219999;
when (strat=210005)  nstrat=219999;
when (strat=220002)  nstrat=220001;
when (strat=230015)  nstrat=239999;
when (strat=230016)  nstrat=239999;
when (strat=230017)  nstrat=239999;
when (strat=230019)  nstrat=239999;
```

```
when (strat=230020)  nstrat=239999;
when (strat=230022)  nstrat=239999;
when (strat=230029)  nstrat=239999;
when (strat=230032)  nstrat=239999;
when (strat=230035)  nstrat=239999;
when (strat=230037)  nstrat=239999;
when (strat=230039)  nstrat=239999;
when (strat=230041)  nstrat=239999;
when (strat=240004)  nstrat=249999;
when (strat=240005)  nstrat=249999;
when (strat=240006)  nstrat=249999;
when (strat=240012)  nstrat=249999;
when (strat=240014)  nstrat=249999;
when (strat=240016)  nstrat=249999;
when (strat=250003)  nstrat=259999;
when (strat=250004)  nstrat=259999;
when (strat=250007)  nstrat=259999;
when (strat=250010)  nstrat=259999;
when (strat=250014)  nstrat=259999;
when (strat=260016)  nstrat=269999;
when (strat=260017)  nstrat=269999;
when (strat=260018)  nstrat=269999;
when (strat=260019)  nstrat=269999;
when (strat=260020)  nstrat=269999;
when (strat=260021)  nstrat=269999;
when (strat=260031)  nstrat=269999;
when (strat=260040)  nstrat=269999;
when (strat=270003)  nstrat=279999;
when (strat=270008)  nstrat=279999;
when (strat=280004)  nstrat=289999;
when (strat=280005)  nstrat=289999;
when (strat=280006)  nstrat=289999;
when (strat=280008)  nstrat=289999;
when (strat=280010)  nstrat=289999;
when (strat=280012)  nstrat=289999;
when (strat=290019)  nstrat=299999;
when (strat=290020)  nstrat=299999;
when (strat=290021)  nstrat=299999;
when (strat=290023)  nstrat=299999;
when (strat=290025)  nstrat=299999;
when (strat=290040)  nstrat=299999;
```

```
when (strat=310026)  nstrat=319999;
when (strat=310027)  nstrat=319999;
when (strat=310028)  nstrat=319999;
when (strat=310029)  nstrat=319999;
when (strat=310031)  nstrat=319999;
when (strat=310035)  nstrat=319999;
when (strat=310036)  nstrat=319999;
when (strat=310037)  nstrat=319999;
when (strat=310038)  nstrat=319999;
when (strat=310039)  nstrat=319999;
when (strat=310047)  nstrat=319999;
when (strat=310050)  nstrat=319999;
when (strat=310054)  nstrat=319999;
when (strat=310057)  nstrat=319999;
when (strat=310059)  nstrat=319999;
when (strat=310073)  nstrat=319999;
when (strat=310076)  nstrat=319999;
when (strat=310082)  nstrat=319999;
when (strat=310090)  nstrat=319999;
when (strat=310095)  nstrat=319999;
when (strat=320008)  nstrat=329999;
when (strat=320010)  nstrat=329999;
when (strat=320015)  nstrat=329999;
when (strat=330030)  nstrat=339999;
when (strat=330031)  nstrat=339999;
when (strat=330032)  nstrat=339999;
when (strat=330033)  nstrat=339999;
when (strat=330034)  nstrat=339999;
when (strat=330035)  nstrat=339999;
when (strat=330036)  nstrat=339999;
when (strat=330037)  nstrat=339999;
when (strat=330038)  nstrat=339999;
when (strat=330039)  nstrat=339999;
when (strat=330041)  nstrat=339999;
when (strat=330042)  nstrat=339999;
when (strat=330043)  nstrat=339999;
when (strat=330044)  nstrat=339999;
when (strat=330048)  nstrat=339999;
when (strat=330053)  nstrat=339999;
when (strat=350052)  nstrat=359999;
when (strat=350054)  nstrat=359999;
```

```
when (strat=350055)  nstrat=359999;
when (strat=350056)  nstrat=359999;
when (strat=350058)  nstrat=359999;
when (strat=350060)  nstrat=359999;
when (strat=350061)  nstrat=359999;
when (strat=350062)  nstrat=359999;
when (strat=350063)  nstrat=359999;
when (strat=350064)  nstrat=359999;
when (strat=350067)  nstrat=359999;
when (strat=350069)  nstrat=359999;
when (strat=350070)  nstrat=359999;
when (strat=350071)  nstrat=359999;
when (strat=350072)  nstrat=359999;
when (strat=350073)  nstrat=359999;
when (strat=350075)  nstrat=359999;
when (strat=350076)  nstrat=359999;
when (strat=350077)  nstrat=359999;
when (strat=350084)  nstrat=359999;
when (strat=350086)  nstrat=359999;
when (strat=350097)  nstrat=359999;
when (strat=350101)  nstrat=359999;
when (strat=350105)  nstrat=359999;
when (strat=350112)  nstrat=359999;
when (strat=350118)  nstrat=359999;
when (strat=350123)  nstrat=359999;
when (strat=350126)  nstrat=359999;
when (strat=410021)  nstrat=419999;
when (strat=410022)  nstrat=419999;
when (strat=410024)  nstrat=419999;
when (strat=410025)  nstrat=419999;
when (strat=410027)  nstrat=419999;
when (strat=410029)  nstrat=419999;
when (strat=410031)  nstrat=419999;
when (strat=410033)  nstrat=419999;
when (strat=410035)  nstrat=419999;
when (strat=410040)  nstrat=419999;
when (strat=410052)  nstrat=419999;
when (strat=410057)  nstrat=419999;
when (strat=410059)  nstrat=419999;
when (strat=420009)  nstrat=429999;
when (strat=420010)  nstrat=429999;
```

```
when (strat=420012)  nstrat=429999;
when (strat=420013)  nstrat=429999;
when (strat=420014)  nstrat=429999;
when (strat=420015)  nstrat=429999;
when (strat=420019)  nstrat=429999;
when (strat=420026)  nstrat=429999;
when (strat=420028)  nstrat=429999;
when (strat=430034)  nstrat=439999;
when (strat=430035)  nstrat=439999;
when (strat=430037)  nstrat=439999;
when (strat=430038)  nstrat=439999;
when (strat=430039)  nstrat=439999;
when (strat=430042)  nstrat=439999;
when (strat=430043)  nstrat=439999;
when (strat=430044)  nstrat=439999;
when (strat=430045)  nstrat=439999;
when (strat=430049)  nstrat=439999;
when (strat=430050)  nstrat=439999;
when (strat=430051)  nstrat=439999;
when (strat=430052)  nstrat=439999;
when (strat=430056)  nstrat=439999;
when (strat=430066)  nstrat=439999;
when (strat=430076)  nstrat=439999;
when (strat=500007)  nstrat=509999;
when (strat=500008)  nstrat=509999;
when (strat=500012)  nstrat=509999;
when (strat=500016)  nstrat=509999;
when (strat=500018)  nstrat=509999;
when (strat=510007)  nstrat=519999;
when (strat=510011)  nstrat=519999;
when (strat=520015)  nstrat=529999;
when (strat=520016)  nstrat=529999;
when (strat=520017)  nstrat=529999;
when (strat=520020)  nstrat=529999;
when (strat=520021)  nstrat=529999;
when (strat=520026)  nstrat=529999;
when (strat=520035)  nstrat=529999;
```

```
when (strat=170008)  nstrat=179999;
when (strat=170009)  nstrat=179999;

when (strat=430031)  nstrat=439999;
when (strat=430054)  nstrat=439999;

when (strat=500019)  nstrat=509999;
when (strat=500020)  nstrat=509999;

when (strat=520022)  nstrat=529999;
when (strat=520023)  nstrat=529999;
otherwise            nstrat=strat;
end;

RUN;
```

Apêndice 1 - Análise de Dados Amostrais Complexos

A.1 A Inferência em Dados Amostrais Complexos

O uso de dados de amostrais complexos, conforme destacado no capítulo anterior, envolve probabilidades distintas de seleção das unidades, conglomeração das unidades e estratificação. Para maiores detalhes ver os trabalhos de Skinner, Holt e Smith (1989) e Pessoa e Silva (1998).

A utilização de métodos adequados para realização de inferência em dados amostrais complexos permite estimar valores de uma variável de interesse e avaliar o grau de precisão das estimativas (através de suas variâncias). As estimativas das variâncias, por sua vez, são influenciadas pelo plano amostral utilizado. Com isso, é importante ressaltar a importância da incorporação do plano nos procedimentos de inferência com base em dados amostrais complexos como a PNAD. Já existem programas estatísticos que suportam tais análises, como exemplo, o SAS, STATA (2003).

A justificativa para a incorporação do desenho amostra nas inferências analíticas, partindo de dados amostrais complexos (Côrrea, 2001), é que os pesos podem ser usados para proteger *contra planos amostrais não-ignoráveis (Pessoa e Silva, 1998), que poderiam introduzir ou causar vícios, e má especificação do modelo.*

Este capítulo tem por objetivo apresentar o Efeito do Plano Amostral em pesquisas de dados amostrais complexos como a PNAD. Além disso, apresentar o procedimento para o cálculo dos intervalos de confiança, os testes de hipóteses e o Método de Máxima Pseudo-Verossimilhança (MPV) aplicado na inferência analítica (modelagem).

A.2 Efeito do Plano Amostral

O Efeito do Plano Amostral (EPA[13]) tem por finalidade avaliar o impacto em desconsiderar o esquema de seleção da amostra no cálculo das estimativas. Esta medida foi proposta inicialmente por Kish (1965) e aperfeiçoada por Kish e Frankel (1974), para maiores detalhes ver Pessoa e Silva (1998). Na inferência estatística, para um parâmetro θ, o EPA é obtido pela razão entre a variância do plano amostral complexo (verdadeiro) e a variância da distribuição do estimador $\hat{\theta}$ de θ induzida pelo plano de amostragem aleatória simples (AAS[14]) - $V_{AAS}(\hat{\theta})$.

$$EPA(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_{AAS}(\hat{\theta})} \qquad (4.1)$$

onde, $V_p(\hat{\theta})$ -= variância da distribuição de $\hat{\theta}$ induzida pelo plano amostral complexo.

Skinner, Holt e Smith (1989, p.24) destacam que esta medida é importante para avaliar a eficiência quando comparamos desenhos alternativos na concepção das pesquisas. Além disso, o uso do EPA apresenta dificuldades no seu cálculo em inferências analíticas (modelagem) e, por isso, definiram o conceito do EPA ampliado

[13] É apresentado nos softwares estatísticos como design effect (deff).
[14] As informações são coletadas de forma independente e são identicamente distribuídas – IID

(misspecification effect – meff). Esta medida mensura a tendência de um estimador usual (consistente), calculado sob hipótese de observações independentes e identicamente distribuídas (IID), subestimar ou superestimar a variância verdadeira do estimador pontual. O EPA ampliado (também denominado por meff - misspecification effect) é a razão entre a variância do estimador sob o plano amostral ou modelo correto $V_{VERD}(\hat{\theta})$ sobre a esperança do estimador da variância de $\hat{\theta}$ sob a hipótese de observações IID da variância $E_{VERD}(v_0)$.

Dado que $v_0 = \hat{V}_{IID}(\hat{\theta})$ um estimador da variância de $\hat{\theta}$ para uma Amostra Aleatória Simples sem reposição. Então:

$$\text{meff} = EPA_{ampliado}(\hat{\theta}, v_0) = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(v_0)} = \frac{V_{VERD}(\hat{\theta})}{E_{VERD}(\hat{V}_{IID}(\hat{\theta}))} \quad (4.2)$$

O $EPA_{ampliado}(\hat{\theta}, v_0)$ mensura a tendência de v_0 subestimar ou superestimar $V_{VERD}(\hat{\theta})$, variância verdadeira sob o modelo e/ou plano amostral de $\hat{\theta}$. Quanto mais afastado de 1 for o valor de $EPA_{ampliado}(\hat{\theta}, v_0)$, mais incorreta será considerada a especificação do plano amostral ou do modelo nos procedimentos analíticos.

Desta forma, enquanto a medida proposta por Kish baseia-se nas distribuições induzidas pela aleatorização dos planos amostrais comparados, o $EPA_{ampliado}(\hat{\theta}, v_0)$ pode ser calculado com respeito a distribuições de aleatorização ou do modelo envolvido.

Em geral, são esperadas as seguintes conseqüências sobre o Efeito do Plano Amostral ao ignorar o plano amostral efetivamente adotado e admitir que o desenho da amostra foi AAS:

4. Ignorar os pesos em v_0 pode inflacionar o meff (ou EPA ampliado);
5. Ignorar conglomerações em v_0 pode inflacionar o meff;
6. Ignorar estratificação em v_0 pode reduzir o meff.

A.3 Estatística teste

Para uma população finita com um parâmetro de interesse θ e sua estimativa pontual $\hat{\theta}$ o intervalo de confiança com nível de confiança (1- α) a partir da distribuição assintótica de $t_0 = \frac{\hat{\theta} - \theta}{v_0^{0,5}}$, sob a hipótese de que as observações são Independente e Identicamente distribuídas (IID) com distribuição N(0;1), é dado por: $\left[\hat{\theta} - z_{\alpha/2} v_0^{0,5}; \hat{\theta} + z_{\alpha/2} v_0^{0,5}\right]$ onde $z_{\alpha/2} = \int_{\alpha/2}^{+\infty} \varphi(t)dt$ e φ é uma função de densidade da distribuição normal padrão.

Para um plano amostral complexo a estatística de teste é dada por: $t_0 = \frac{\hat{\theta} \quad \theta}{\left[\hat{V}_{VERD}(\hat{\theta})\right]^{1/2}}$ tal que $t_0 \dot{\sim} N[0, EPA(\hat{\theta}, v_0)]$ e $\hat{EPA} = \frac{\hat{V}_{verd}(\hat{\theta})}{v_0}$

Desta forma, ao ignorar os pesos e o efeito de conglomeração do desenho amostral pode-se inflacionar o EPA, ampliando-se os intervalos de confiança para os parâmetros de interesse.

A.4 Método de Máxima Pseudo Verossimilhança

Esta seção é um resumo dos trabalhos apresentados por Phillippe (2001) e Rodrigues (2003). A incorporação do plano amostral na inferência analítica consolidada por Skinner, Holt e Smith (1989), propõem um método para a estimação dos parâmetros do modelo denominado Método de Pseudo Máxima Verossimilhança e será detalhado adiante.

Seja i o índice do elemento de uma população, o qual contém as informações dos estratos, das Unidades Primárias de Amostragem (UPA) no estrato e dos elementos dentro das UPA's. Dada as variáveis $(Y_i, \underline{X}_i)$ provenientes da população U, onde Y_i é a variável de interesse (ou resposta) e $\underline{X}_i$ é um vetor de características (variáveis explicativas) associadas a cada *i*, tal que $i \in U$. Assumindo que $Y_1,...,Y_N$ são IID com função de densidade $f(y_i, \underline{\beta})$*, onde* $\underline{\beta}$ é um vetor de parâmetros desconhecidos de interesse. Com isso, a verossimilhança e o logaritmo da verossimilhança populacionais são descritos por:

$$L(\underline{\beta}; y_i \underline{X}_i) = \prod_{i \in U} f(y_i, \underline{\beta}) \tag{4.6}$$

$$l(\underline{\beta}; y_i \underline{X}_i) = \sum_{i \in U} \log\left[f(y_i, \underline{\beta})\right] \tag{4.7}$$

Sendo $l(\underline{\beta}; y_i \underline{X}_i)$ o logaritmo da função de verossimilhança associado ao modelo, onde $\underline{\beta}$ é o vetor de parâmetros com dimensão p x 1, $\underline{X}_i$ é um vetor de dimensão *1 x p,* definido para todo $i \in U$. Então para uma população finita, os parâmetros $\underline{\beta}$ são gerados através da solução de um sistemas de equações definido por:

$$G(\underline{\beta}) = \sum_{i \in U} u_i(\underline{\beta}; y_i, \underline{X}_i) \tag{4.8}$$

onde $u_i = \frac{\partial l(\underline{\beta}; y_i, \underline{X}_i)}{\partial \underline{\beta}}$ é o vetor *p x 1* dos escores do elemento *i, i* $\in U$

A solução deste sistema, quando $G(\underline{\beta}) = 0$, para todo $i \in U$ é o estimador de máxima verossimilhança para $\underline{\beta}$ no caso de um censo.

Como $G(\underline{\beta}) = \sum_{i \in U} u_i(\underline{\beta}, y_i, \underline{x}_i)$ é a soma dos escores, que por sua vez é um vetor de totais, e para estimá-lo $(i \in s)$ - amostra – pode-se utilizar um estimador linear

ponderado da forma $\hat{G}(\underline{\beta}) = \sum_{i \in s} w_i u_i(\underline{\beta}; y_i, \underline{x}_i)$, onde w_i são os pesos[15] propriamente definidos. O estimador de Máxima Pseudo-Verossimilhança de $\underline{\hat{\beta}}$ é a solução da equação:

$$\hat{G}(\underline{\beta}) = \sum_{i \in s} w_i u_i(\underline{\beta}; y_i, \underline{X}_i) = 0 \qquad (4.9)$$

Binder (1983), Pessoa e Silva (1998) apresentam a matriz de primeira ordem da expansão da série de Taylor para o estimador de pseudo máxima-verossimilhança. Com isto, encontra-se um estimador para a variância assintótica, sob o plano amostral, de $\underline{\hat{\beta}}$, que é dado por:

$$\hat{V}(\underline{\hat{\beta}}) = \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}} \right]^{-1} \hat{V}\left[\sum_{i \in s} w_i u_i(\underline{\hat{\beta}}; y_i; \underline{x}_i) \right] \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}} \right]^{-1} \qquad (4.10)$$

onde $\hat{V}\left[\sum_{i \in s} w_i u_i(\underline{\hat{\beta}}; y_i; \underline{x}_i) \right]$ é um estimador consistente para a variância do estimador do total populacional dos escores.

Binder (1983) mostrou que a distribuição assintótica de $\underline{\hat{\beta}}$ é normal multivariada, fornecendo uma base para inferência sobre $\underline{\beta}$ sob amostras grandes, tal que:

$$\hat{V}(\underline{\hat{\beta}})^{-1/2}(\underline{\hat{\beta}} - \underline{\beta}) \sim NM(0; I) \qquad (4.11)$$

Os pesos w_i devem ser tais que satisfaçam as condições de que os estimadores sejam assintoticamente normais e não viciados, além de possuírem estimadores de variância consistentes. Estas condições são satisfeitas quando a probabilidade de inclusão na amostra da *i-ésima* unidade amostral da população, *i=1,2..,N*, é maior do que zero ($\pi_i > 0$) e, simultaneamente, que a probabilidade de inclusão conjunta da *i-ésima* e *j-ésima* unidades amostrais da população, i≠j, seja maior do que zero ($\pi_{ij} > 0$). Os estimadores de máxima pseudo verossimilhança não são únicos, pois existem diversas maneiras de definir os pesos w_i.

Usualmente utilizamos os pesos do estimador de Horwitz-Thompson para totais, dado por:

$$w_i = \frac{1}{\pi_i}, \forall i \in s \qquad (4.12)$$

[15] Peso é igual ao inverso da probabilidade de seleção do indivíduo em uma amostra.

Substituindo (4.12) em (4.9) encontramos o estimador pontual $\underline{\hat{\beta}}_{\pi}$ do parâmetro $\underline{\beta}$ no modelo, e substitindo (4) e $\underline{\hat{\beta}}_{\pi}$ em (4.10) temos que:

$$\hat{V}(\underline{\hat{\beta}}_{\pi}) = \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}_{\pi}} \right]^{-1} \hat{V}\left[\sum_{i \in S} \frac{1}{\pi_i} u_i(\underline{\hat{\beta}}_{\pi}; y_i; \underline{x}_i) \right] \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}_{\pi}} \right]^{-1} \tag{4.13}$$

$$\hat{V}(\underline{\hat{\beta}}) = \left\{ \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}_{\pi}} \right]^{-1} \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} [u_i(\underline{\hat{\beta}}_{\pi}, y_i, \underline{x}_i)][u_i(\underline{\hat{\beta}}_{\pi}, y_i, \underline{x}_i)]' \left[\left. \frac{\partial \hat{G}(\underline{\beta})}{\partial \underline{\beta}} \right|_{\underline{\beta}=\underline{\hat{\beta}}_{\pi}} \right]^{-1} \right\} \tag{4.14}$$

A equação (4.14) fornece o estimador da variância do estimador de pseudo máxima verossimilhança $\underline{\hat{\beta}}_{\pi}$ de $\underline{\beta}_{\pi}$.