UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

IMPUTAÇÃO MÚLTIPLA UTILIZANDO O SOFTWARE SPSS

NATALIA VAIS RODRIGUES

Porto Alegre
2016

NATALIA VAIS RODRIGUES

IMPUTAÇÃO MÚLTIPLA UTILIZANDO O SOFTWARE SPSS

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de bacharelado em Estatística.

Orientadora:
Professora Dra. Luciana Neves Nunes

Porto Alegre
2016

CIP - Catalogação na Publicação

Rodrigues, Natalia Vais
Imputação Múltipla Utilizando o Software SPSS / Natalia Vais Rodrigues. -- 2016.
44 f.

Orientador: Luciana Neves Nunes.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Instituto de Matemática, Curso de Estatística, Porto Alegre, BR-RS, 2016.

1. Dados Faltantes. 2. Imputação Múltipla. 3. SPSS. 4. Regressão Logística. 5. Regressão de Poisson. I. Nunes, Luciana Neves, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

Instituto de Matemática e Estatística
Departamento

Imputação Múltipla Utilizando o Software SPSS

Natalia Vais Rodrigues

Banca examinadora:

Professora Dra. Luciana Neves Nunes
Universidade Federal do Rio Grande do Sul

Jaimar de Barros Monteiro
Departamento Estadual de Trânsito do RS – Detran/RS

AGRADECIMENTOS

Primeiramente a minha família, que sempre me apoiou e me incentivou a buscar crescimento e que fizeram de tudo para que eu tivesse uma educação de qualidade. Em especial meus pais, Paula e Carlos, meus tios, Magda e Pedro e minha avó América por todo o amor, carinho e força. Mãe você deu a sua vida, para que a minha fosse melhor, obrigada por tudo.

Ao meu namorado, Maycon, por todo companheirismo e incentivo. Divido esta vitória com você.

A minha orientadora, Luciana, que sempre me deu todo o suporte para que este trabalho pudesse ser feito. E me transmitiu todos os seus conhecimentos, para que eu pudesse crescer e concluir esta etapa.

Ao Detran, onde fiz meu estágio obrigatório, por todos os ensinamentos e por ter cedido os dados para que este trabalho fosse possível.

Ao Jaimar, por ter sido meu supervisor de estágio e por ter aceito ser banca deste trabalho.

E por fim, a todos os professores e colegas/amigos que fizeram parte desta jornada.

DEDICATÓRIA

Dedico este trabalho aos meus irmãos, Kauã e Miguel. Que sirva de inspiração para que eles continuem brilhando cada vez mais.

RESUMO

Introdução: Muitos dos métodos estatísticos existentes, como por exemplo modelos de regressão, exigem informações completas para que os resultados sejam produzidos. Porém, sabemos que na prática muitas vezes este pressuposto não é facilmente atendido. Como forma de tratamento aos dados faltantes temos a Imputação Múltipla (IM). Basicamente a técnica da IM consiste em criarmos cópias do banco de dados onde os dados faltantes são substituídos por diferentes valores imputados. O objetivo deste trabalho será aplicar e descrever a técnica de Imputação Múltipla no software SPSS. **Aplicação:** O**s** dados utilizados se referem aos resultados obtidos em prova prática de candidatos a terem sua primeira habilitação para veículos de 4 rodas. Os dados originais não possuem valores faltantes, assim para proceder a IM no SPSS, perdas (5%, 10% e 20%) foram simuladas para uma variável preditora e para variável resposta correspondendo ao modelo de regressão utilizado. Para a qualificação das IM os bancos imputados foram analisados através de regressão logística e regressão de Poisson. Os resultados foram comparados entre os níveis de perda, e com os resultados obtidos para o banco completo. **Resultados:** Foi bastante satisfatório os resultados das regressões para os bancos com IM. As diferenças foram bem pequenas, então se realmente existisse a perda e a IM fosse necessária não teríamos grandes prejuízos nos resultados e suas respectivas interpretações. O software SPSS possui mais de um modelo de IM disponível e sua interface é bem intuitiva e amigável, o que facilita o uso da IM.

Palavras-chave: Dados Faltantes. Imputação Múltipla. SPSS. Regressão Logística. Regressão de Poisson.

ABSTRACT

Introduction: Several of the existing statistical methods, such as regression models, require complete information for the results to be produced. However, we know that in practice this assumption is often not easily accomplished. As a form of treatment to the missing data we have the Multiple Imputation (MI). Basically the MI technique consists on creating copies of the database where the missing data is replaced by different imputed values. The purpose of this paper will be to apply and describe the Multiple Imputation technique in SPSS software. **Application:** The used data refers to the results obtained in a practical test of candidates to have their first driver's license for four wheeled vehicles. The original data does not have missing values, so to proceed to MI in SPSS, losses (5%, 10% and 20%) were simulated for a predictive variable and for response variable corresponding to the regression model used. In order to qualify MI, the imputed banks were analyzed through logistic regression and Poisson regression. The results were compared between the loss levels, and with the results obtained for the complete bank**. Results:** We found the results of the regressions for banks with MI were quite satisfactory. The differences were very small, so if there was indeed a loss and MI was necessary we would not have much damage in the results and their respective interpretations. SPSS software has more than one MI model available and its interface is intuitive and user-friendly, making it easy to use MI.

Keywords: Missing Data. Multiple Imputation. SPSS. Logistic Regression. Poisson regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Muitos dos métodos estatísticos existentes, como por exemplo, modelos de regressão, exigem informações completas para que os resultados sejam aplicáveis. Porém, sabemos que na prática muitas vezes este pressuposto não é facilmente atendido. É bastante comum em pesquisas e estudos a não observância de informações. Para tal existem várias causas justificáveis, como a recusa de quem fornece a informação, a incapacidade de responder, problema no preenchimento ou registro da resposta, perda da informação em algum estágio da pesquisa/estudo, entre outras. Esta perda de informação gera o que chamamos de *dados faltantes*, tradução de *missing data*[1,3].

Uma das formas usadas para lidar com dados faltantes é a exclusão daqueles indivíduos que não possuem seus dados completos para se fazer a análise estatística, mas esta abordagem por muitas vezes restringe a amostra a uma proporção menor da amostra original, e possivelmente o preço a se pagar por isso é a perda de poder e precisão[3]. Para contornar este problema uma alternativa comumente recorrida é a imputação de dados.

Imputação de dados é a técnica na qual se insere valores plausíveis onde temos dados faltantes. Na imputação simples são usadas as estatísticas descritivas mais comuns como média, mediana e valor (ou categoria) mais frequente. Porém inserir um valor constante para os dados faltantes significa que podemos estar subestimando a variabilidade desta variável ou estamos ignorando outros valores admissíveis[1,2,3].

Como alternativa a Imputação Simples, temos a Imputação Múltipla (IM). Na IM os dados faltantes são imputados **m** vezes, gerando **m** bancos de dados distintos e completos. Destes **m** bancos distintos obtemos **m** resultados, de acordo com o objetivo desejado. Esses **m** resultados são compilados de forma que obtenhamos somente um resultado final acerca do qual faremos as inferências desejadas. Uma vantagem da IM é que ao imputarmos **m** vezes os dados faltantes, estamos levando em consideração a variabilidade da imputação[1,2].

Atualmente existem vários softwares que trabalham com a imputação de dados, tais como: SOLAS, STATA, SPSS, SAS e R². O objetivo deste trabalho será aplicar e descrever a técnica de Imputação Múltipla no software SPSS[4]. Para tal aplicação será usado um banco de dados fornecido pelo Detran/RS contendo dados referentes a candidatos que prestaram prova prática para a obtenção da primeira habilitação.

2 METODOLOGIA

2.1 Imputação Múltipla

Basicamente a técnica da IM consiste em criarmos cópias do banco de dados onde os dados faltantes são substituídos por diferentes valores imputados. Ou seja, um número **m** de bancos distintos e completos será gerado, e cada um deles será analisado conforme o objetivo pretendido. Desta forma teremos **m** resultados que serão compilados para que por fim tenhamos um único resultado final para as devidas conclusões. Tipicamente o **m** varia entre 5 e 10, embora ainda seja viável um número maior de imputações[1,2 e 3].

Em 1980, Rubin publicava seu livro onde apresentava a técnica e sua vantagem em relação a imputação única. Pois, como a IM gera valores diferentes para cada **m-ésima** imputação, ganhamos com isso uma variabilidade que não se levaria em conta se imputássemos uma única vez. O problema estava nos recursos computacionais da época, que eram limitados, criando assim uma dificuldade na implementação da técnica. Atualmente, com os avanços computacionais, a IM tornou-se uma abordagem mais usual para tratarmos o problema dos dados falantes[2].

Um passo importante para implementação da IM, é avaliar tanto os mecanismos de não-resposta como o padrão de não-resposta. Pois estes que definem qual método de IM deverá ser utilizado para que as **m** imputações distintas sejam geradas[1,2].

Mecanismos de não – resposta

Rubin apresenta três mecanismos de não-resposta: completamente aleatório, aleatório e não aleatório[1].

Dados Faltantes Completamente Aleatórios

Citado também como MCAR, sigla para *Missing Completely at Random*. São classificados assim dados que a sua ausência não pode ser relacionada com nenhuma outra variável presente no estudo e também não pode ser explicada pela própria variável[1].

Além de possuir uma suposição forte, dificilmente é satisfeito na prática, assim podemos considerá-lo ignorável, ou seja, não existe a necessidade de especificar um modelo para a não resposta[6,8]. Na sua ocorrência, os dados não observados constituem uma subamostra aleatória.

Dados Faltantes Aleatórios

Citado também como MAR, sigla para *Missing at Random*. Quando a ausência pode ser explicada através de outras variáveis disponíveis no banco, e não tem relação com a variável em si. Por exemplo, uma pesquisa onde as mulheres são menos suscetíveis a responderem seu peso, assim temos que os dados faltantes referentes ao peso podem ser explicados pelo sexo do indivíduo, assim se tivermos a variável sexo completa, o peso das mulheres que responderam serão uma amostra aleatória do peso de todas as mulheres[1]. Também pode ser considerado ignorável. Ressalta-se que ignoramos o mecanismo de dados faltantes e não os dados ou as unidades onde temos dados faltantes[6,8].

Dados Faltantes Não Aleatórios

Citado também como NMAR, sigla para *Missing not at Random*. Não possuem relação com as outras variáveis, somente com os valores não observados. Ou seja, quando a própria variável explica a não-resposta. Comumente são valores extremos. Por exemplo, pessoas de baixa ou alta renda omitiram seu salário ao responderem uma pesquisa[1,8,9].

Devido a não aleatoriedade da não-resposta consideramos este mecanismo não ignorável, ou seja, existe a necessidade de especificar um modelo para a mesma.

Padrão de Não-Resposta

Considerando que podemos arranjar os dados de forma retangular ou de matriz, onde as linhas serão cada unidade respondente e as colunas serão as variáveis. Desta forma podemos identificar o padrão de não-resposta como monotônico e não motônico[2,7].

Padrão Monotônico

Quando somente uma das variáveis possui dados faltantes, podendo ser ela resposta ou preditora, temos um padrão univariado. Caso particular do padrão monotônico (Figura 1a). No

caso de dados faltantes em mais de uma variável, consideramos padrão monotônico quando podemos ordenar as variáveis de forma que as colunas x_{j+1} estejam "completas" para todos os casos que não possuem dados faltantes nas colunas x_j (Figura 1b) [2,7].

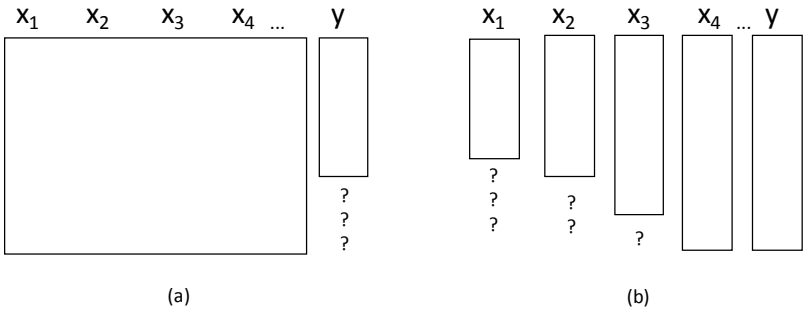


Figura 1 – Padrão de não-resposta Monotônico

Padrão Não Monotônico

Quando temos duas variáveis que nunca são observadas juntas, consideramos um padrão não motônico. Isto ocorre quando os dados provêm de duas amostras, por exemplo, com a mesma variável resposta, mas duas preditoras distintas (Figura 2a).

Outro caso é o padrão arbitrário, quando não conseguimos obter nenhum padrão geral ou estrutura (Figura 2b)[2,7].

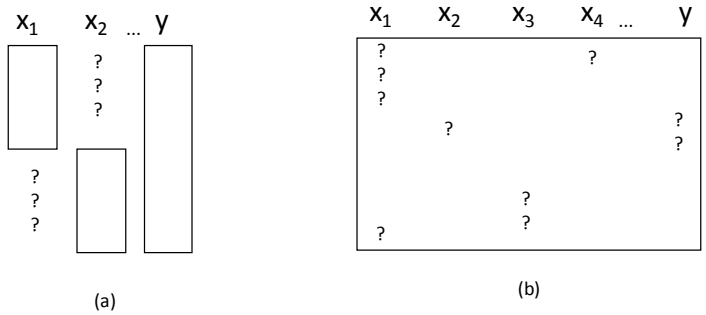


Figura 2 – Padrão de não-resposta Não Monotônico

Métodos de IM

Dos métodos existentes de IM, neste trabalho citaremos os dois métodos usados pelo SPSS para a implementação da IM[4].

Chained Equations

O SPSS utiliza um método de Monte Carlo via cadeias de Markov (MCMC) conhecido como *fully conditional specification* ou *chained equations*[14].

O modelo de imputação é especificado separadamente para cada variável utilizando todas as outras variáveis como preditoras. Em cada etapa do algoritmo, uma imputação é gerada para a variável com dados faltantes, então estes valores imputados são usados para a imputação da próxima variável. O processo continua até que se atinja a convergência. Ou seja, a ideia básica é imputar as variáveis incompletas uma de cada vez, utilizando todas as variáveis e as variáveis já imputadas na etapa anterior[4,5,14].

Regressão linear pode ser usada para variáveis contínuas, ou então, *predictive mean matching*, onde as variáveis imputadas assumem o valor de um conjunto de valores observados mais próximo da base de dados. Para variáveis categóricas, regressão logística é usada[5,14].

Este método pode ser usado tanto em padrões não monotônicos, quanto em padrão monotônico[4].

Pode ser usado para quando temos um padrão arbitrário, sendo monotônico ou não-monotônico.

Método Monotônico

Este é um processo não iterativo, que só pode ser usado para bancos com padrão monotônico. Para cada variável, considerando que as variáveis são usadas nesse método seguindo a ordem da variável com menor número de dados faltantes até a variável com o maior número de dados faltantes, um modelo univariado utilizando é ajustado onde todas as variáveis precedentes a aquela que será imputada são usadas como preditoras, e então os valores preditos são imputados[4,5].

Regras de Rubin

Após as **m** imputações, teremos os **m** bancos completos, esses bancos são analisados de forma individual. As estimativas individuais obtidas das análises podem ser combinadas de forma simples, a modo que tenhamos uma média geral através da seguinte equação: $\bar{Q} = \frac{1}{m}\sum_j^m \widehat{Q_j}$, onde Q_j é cada uma das m estimativas do parâmetro de interesse, sendo j= 1, 2, ..., m[1,6,9].

A variância média dentro das imputações é calculada por: $\bar{U} = \frac{1}{m}\sum_j^m U_j$, onde U_j é a variância dos estimadores Q_j. Já a variância entre as imputações é calculada por: $B = \frac{1}{(m-1)}\sum_j^m (\widehat{Q_j} - \overline{Q})$[1,6,9].

Assim podemos obter a variância total de $\bar{Q}$ através da combinação das variâncias descritas acima: $T = \bar{U} + \left(1 - \frac{1}{m}\right)B$[1,6,9].

A combinação proposta por Rubin, pode ser usado em qualquer dos métodos de IM[1,6].

2.2 Aplicação

O banco de dados utilizado neste trabalho foi fornecido pelo Detran/RS. Possui dados sobre candidatos a terem sua primeira habilitação que prestaram prova prática 4 rodas. Foram utilizados somente os resultados dos candidatos que abriram RENACH entre 1/07/2014 a 30/06/2015.

RENACH significa Registro Nacional de Carteira de Habilitação. No momento em que o indivíduo se apresenta em um Centro de Formação de Condutores (CFC) um processo será aberto com um número de RENACH, que a partir daí conterá todas as informações cadastrais deste indivíduo e os resultados obtidos a cada etapa. Estes dados são armazenados no GID (Gerenciamento de informações do Detran/RS).

Para esse trabalho foram usados os dados somente dos candidatos que pretendiam obter sua primeira habilitação, podendo ser ela na categoria AB ou B. Categoria AB refere-se a carteira com permissão para conduzir motos e carros, e categoria B somente carros.

As informações demográficas dos candidatos presentes no GID e disponibilizadas para este trabalho são, município do CFC de abertura do RENACH, sexo, escolaridade e idade. Já sobre as etapas e seus respectivos resultados referentes ao processo da obtenção da carteira, temos: categoria pretendida (AB ou B), número de provas práticas prestadas, número de reprovações, se obteve aprovação, máximo de pontos obtidos na prova prática (os pontos são ganhos a partir de erros cometidos não podendo ultrapassar 3 pontos para a aprovação), mínimo de pontos obtidos

na prova prática, quantidade (em horas) de aulas práticas (aula em automóvel de 4 rodas, na presença de um instrutor) e quantidade (em horas) de aulas no simulador (aula dentro do CFC, utilizando um simulador de direção).

Como os dados citados acima são necessários para a obtenção da CNH (Carteira Nacional de Habilitação), documento obrigatório e único com valor legal para condução de veículos automotivos, dificilmente teremos dados faltantes ou incorretos no banco de dados. Logo para a aplicação da IM precisamos simular a perda de informação. Para isso será necessário definir cenários em que aleatoriamente 5%, 15% e 20% dos candidatos são selecionados e separadamente as informações de uma variável preditora ou respostados sujeitos selecionados são excluídas de forma a criamos variáveis com *missing* (dados faltantes), para aplicação do módulo de IM no SPSS versão 20. A perda será gerada utilizando a opção *Random sample of cases* do comando *Select cases.*Por exemplo, perda de 5% para a variável sexo pode ser simulada sorteando-se no SPSS 5% dos candidatos e excluindo a informação de sexo destes selecionados.

Após a imputação serão feitos modelos de regressão logística para probabilidade de aprovação e modelos de Poisson para a probabilidade do número de reprovações até o indivíduo obter aprovação. Ambos os modelos serão feitos para cada uma das simulações de perda (5%, 15% e 20%), primeiro para uma variável preditora e depois para a variável resposta. Também serão feitos os dois modelos para o banco completo, a fim de comparações. Assim, no final teremos 14 modelos de regressão para uma breve discussão do uso da IM no SPSS

A variável escolaridade será usada para a simulação da perda para variável preditora, em ambos modelos de regressão, logística e de Poisson. Já para a resposta, aprovação na prova prática terá sua perda simulada quando usarmos o modelo logístico e número de reprovações terá sua perda simulada quando usarmos o modelo de Poisson.

Foi usada escolaridade para simulação de perda, pois segundo NAKANO et al. (2011) o aumento nos anos de escolaridade pode ser associado a uma maior capacidade de atenção e inteligência, consideradas importantes para uma avaliação psicológica do trânsito[10].

O modelo de regressão logística é semelhante ao modelo linear, porém sua variável resposta é binária, assumindo valores 0 ou 1, que significam fracasso e sucesso respectivamente. Permitindo predições para tal variável[11].

Já o modelo de Poisson, é usado em dados de contagem. Tem por objetivo modelar o número de ocorrências de um evento, ou taxa de ocorrência do evento. A variável resposta deve

seguir uma distribuição de Poisson e a média da variável resposta deve ser igual à variância, ou seja, os dados devem possuir a mesma dispersão[12].

Os gráficos para algumas análises descritivas foram gerados pelo software Excel.

2.3 Imputação Múltipla no SPSS

O SPSS possui no seu menu *analyze* a opção *Multiple Imputation*. Este módulo que terá seu uso explicado neste trabalho. Para ilustrar o uso da IM no SPSS, será usado o banco citado acima com simulação de perda de 5% da variável escolaridade. Para o primeiro passo, como abordado anteriormente a necessidade do mesmo, analisa-se o padrão de não-resposta. Para isso, existe a opção *analyze patterns* (Figura 3 a). Na caixa de diálogo que surge (Figura 3 b), vamos usar todas as variáveis disponíveis (mesmo a variável de identificação) como *analyze across table.* Nas opções de saída (*output*) será deixado o default do SPSS, exceto por o mínimo de percentagem de *missings*, pois no caso o default é 10% e no banco usado temos 5% de dados faltantes. Pode-se aumentar o número máximo de variáveis que são mostradas, no caso de haver mais de 25 variáveis no banco.

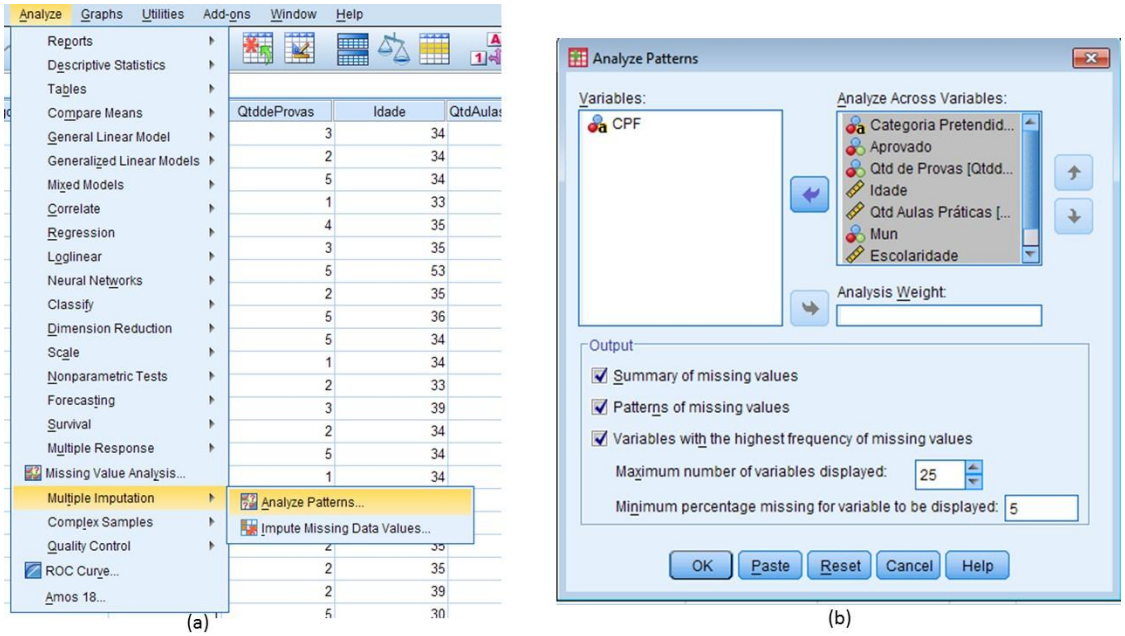


Figura 3 – Módulo *Analyze Patterns* no SPSS

Obteremos como saída um gráfico de pizza com a proporção de dados incompletos para as variáveis, para os casos e para os valores (Figura 4). Uma tabela com análises descritivas das variáveis que apresentaram percentagem de dados faltantes maiores do que o estabelecido (no caso 5%), no exemplo somente escolaridade (Figura 5). Um gráfico de padrões que exibe padrões de dados faltantes para as variáveis de análise (Figura 6). Cada padrão corresponde a um grupo de casos com o mesmo padrão de dados incompletos e completos. Este gráfico ordena as variáveis por dados faltantes, para ajudar na identificação do padrão. No caso, só temos uma variável com dados faltantes, então podemos identificar um padrão monotônico. Um gráfico de barras complementar exibe a porcentagem de casos para cada padrão identificado no gráfico anterior. Isso mostra que quantos dos casos no conjunto de dados têm Padrão 1 ou demais padrões que possam existir (Figura 7).

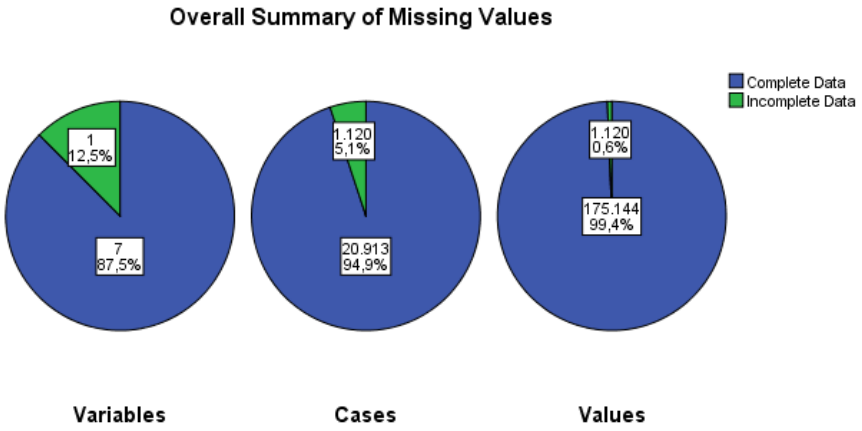


Figura 4 – Gráficos de pizza para diferentes aspectos do banco

Variable Summary[a,b]

	Missing				
	N	Percent	Valid N	Mean	Std. Deviation
Escolaridade	1120	5,1%	20913	2,25	,658

a. Maximum number of variables shown: 25

b. Minimum percentage of missing values for variable to be included: 5,0%

Figura 5 – Tabela de análises descritivas das variáveis com dados faltantes

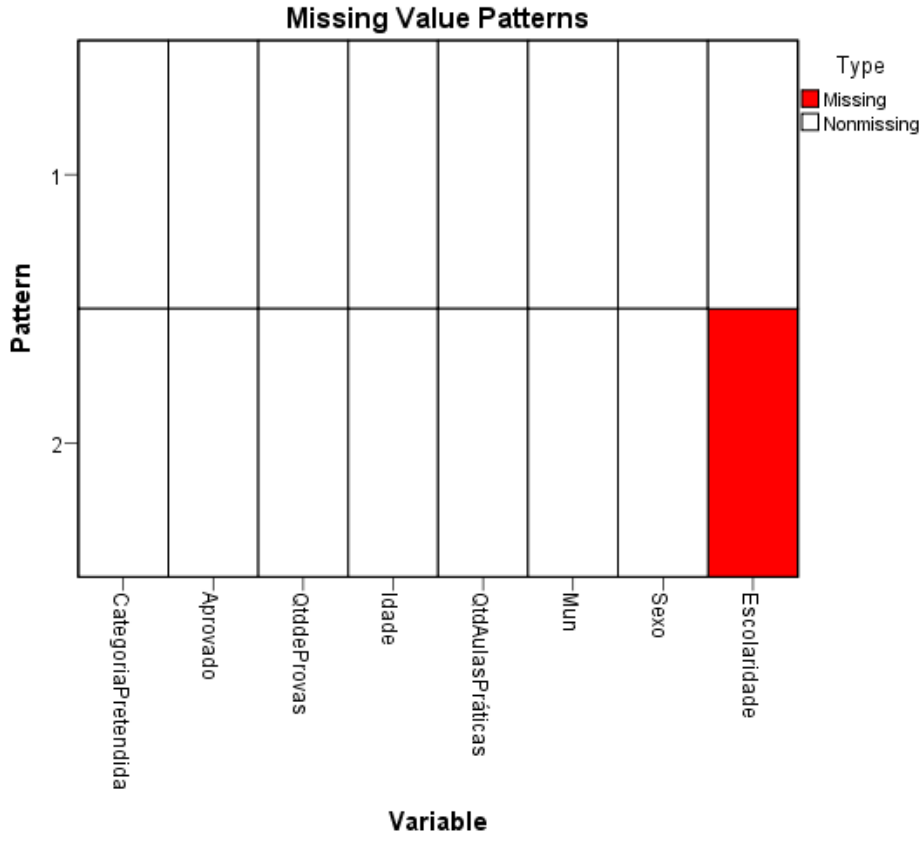


Figura 6 – Gráfico dos padrões de não-resposta

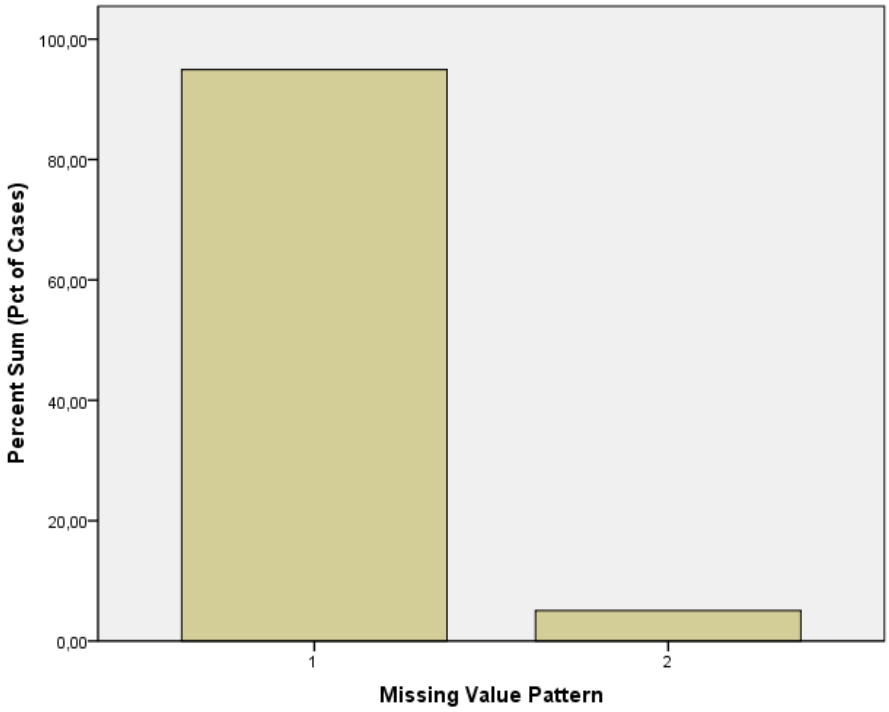


Figura 7 – Gráfico de barras com a percentagem de casos em cada padrão

O segundo passo agora é gerar as **m** imputações. Podemos seguir o mesmo caminho mostrado na figura 3a, mas no caso a opção *impute missing data values* que será utilizada. A caixa de diálogo que surge, tem 4 abas diferentes. Começando pela aba *Variables* (Figura 8). Nesta aba escolhemos as variáveis que entram no modelo usado na IM, no exemplo selecionamos todas as variáveis. Escolhemos o número de imputações, para exemplificar usaremos 10. Um banco novo é criado, onde ficará armazenado o banco original, e as **m** imputações geradas. Por isso, é preciso nomear o mesmo nesta aba.

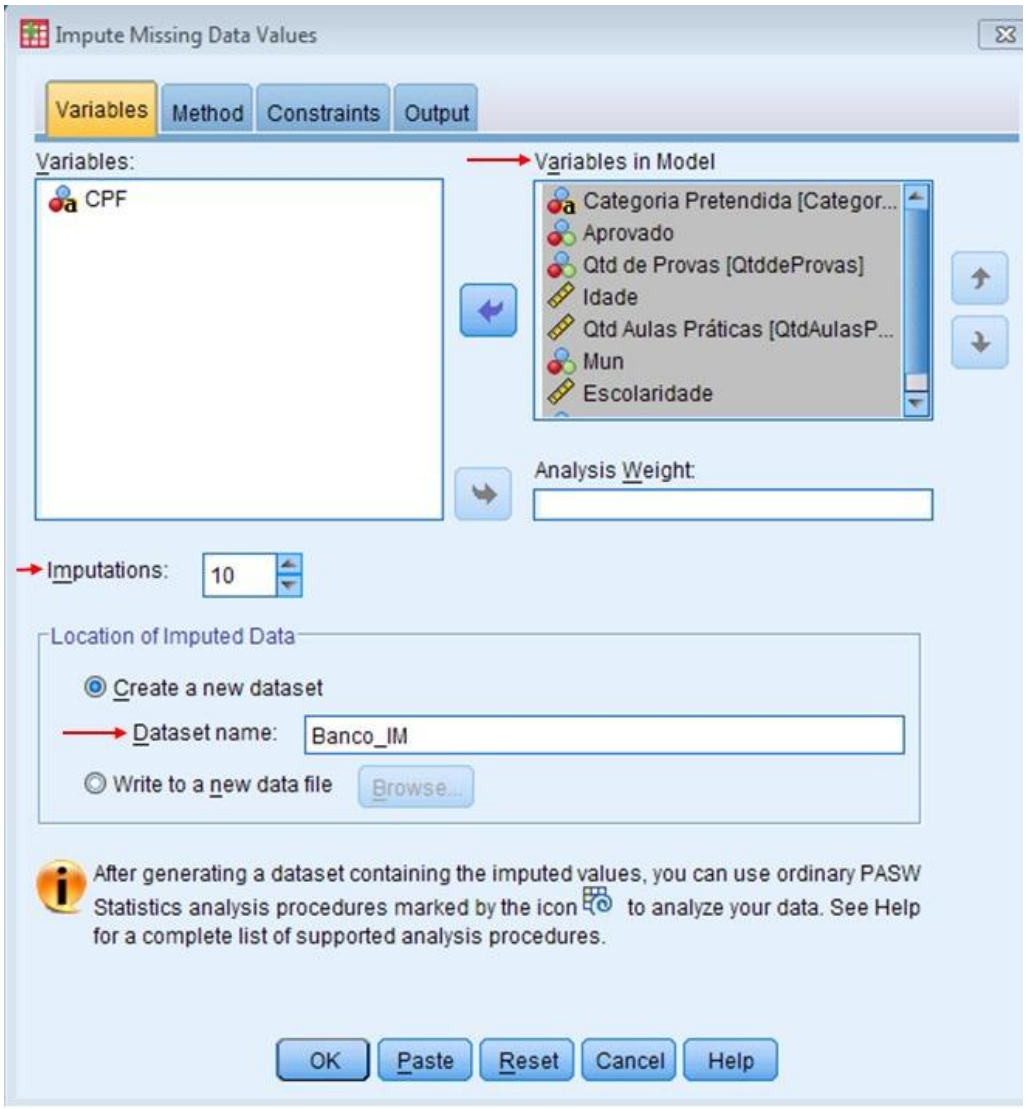


Figura 8 – Aba *Variables* da caixa *Impute Missing Data Values*

Na próxima aba *Method* (Figura 9) o modelo a ser utilizado na IM deve ser definido. O SPSS tem como opção de modelos para a IM os dois modelos citados anteriormente: *fully conditional specification* e o *monotone*. Existe também a possibilidade de deixar que o SPSS escolha o modelo a ser usado, utilizando a opção *automatic*, assim quando o padrão for monotônico o modelo *monotone* será utilizado, caso contrário o modelo FHS será escolhido. Pode-se definir também o modelo usado para variáveis quantitativas: regressão linear ou *Predictive Mean Matching (PMM)*. No exemplo, será usada a opção de escolha automática do modelo.

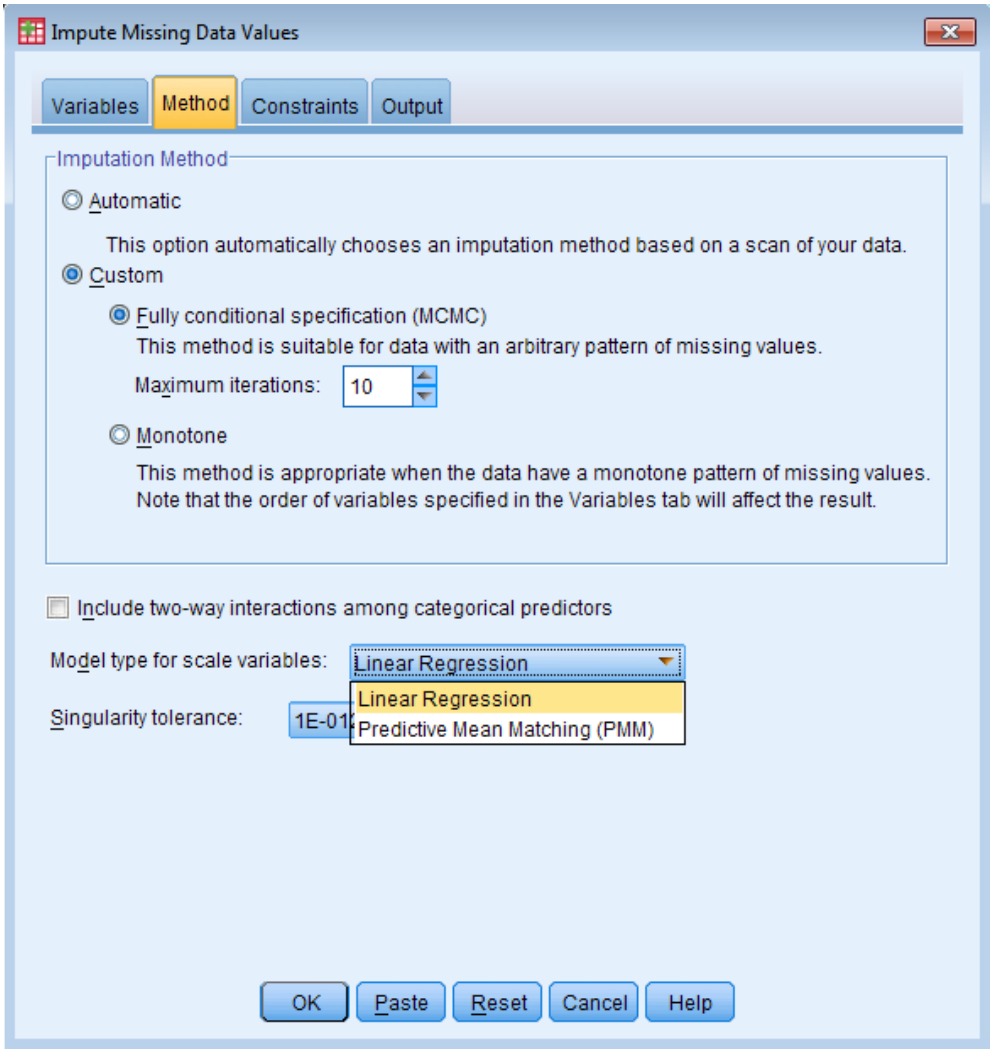


Figura 9 - Aba *Method* da caixa *Impute Missing Data Values*

Na próxima aba *Constraints* (Figura 10) podemos escanear os dados, e obter dados descritivos das variáveis, tais como: percentagem de *missings*, valores mínimo e máximo observados. Também onde definimos quais variáveis são usadas como preditoras das imputações, e quais devem ser imputadas, ou ambas as opções (Figura 11). Pode-se também optar pela exclusão de variáveis que possuírem mais de um determinado limite de dados faltantes definido pelo próprio usuário. No exemplo, escolaridade será definida para imputar e ser usada como preditora, e as demais serão definidas somente como preditoras.

Quando um dos métodos de IM é escolhido e definido, e quando é usada a regressão linear como tipo de modelo para variáveis escalares, precisa se definir (Figura 10) os limites da variável a ser imputada em *Define Constraints*, assim como se determina se será utilizado o valor inteiro ou com casas decimais.

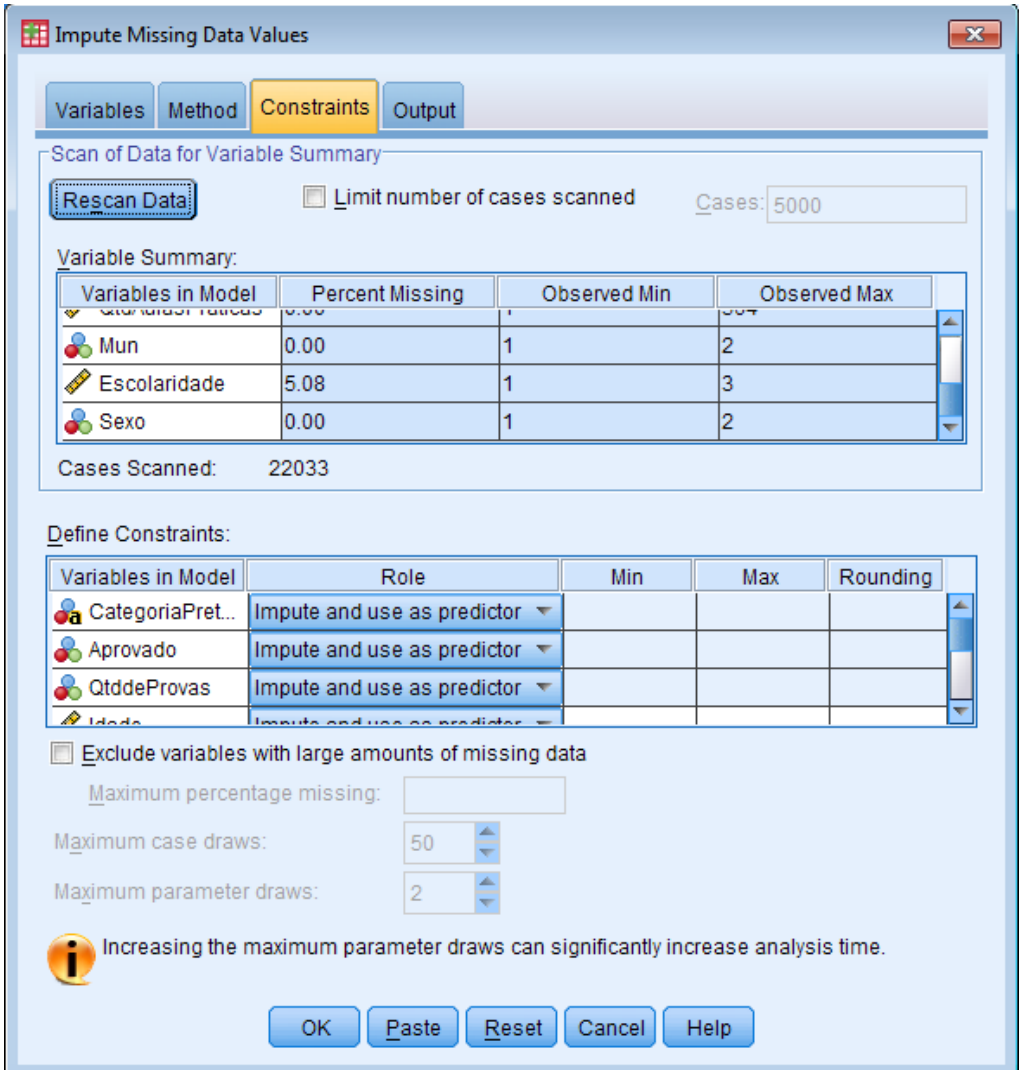


Figura 10 - Aba *Constraints* da caixa *Impute Missing Data Values*

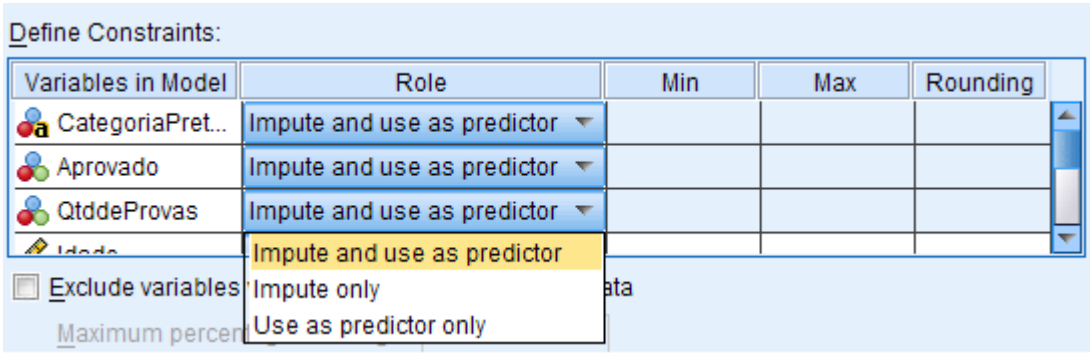


Figura 11 – Opção *Define Constraints*

Na última aba *Output* (Figura 12) definimos o que queremos visualizar na saída. O modelo de imputação já vem automaticamente selecionado, já as medidas descritivas das variáveis imputadas não. Pode ser escolhida esta opção agora, ou depois de rodar a imputação e o banco estiver completo pode ser pedida uma análise descritiva de todo o banco. Existe a possibilidade de um histórico de iteração ser exibido.

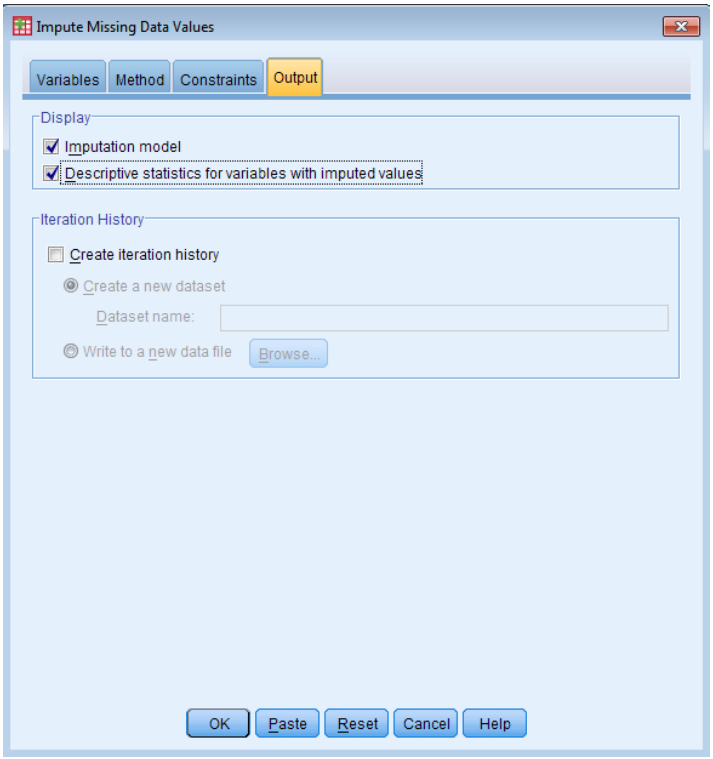
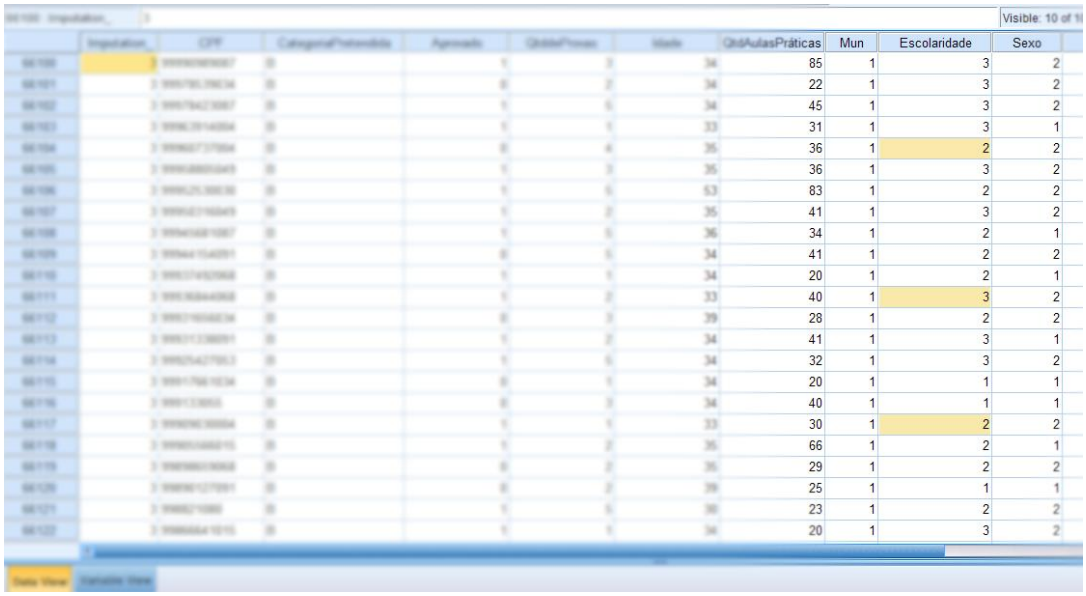


Figura 12 - Aba *Output* da caixa *Impute Missing Data Values*

Após os passos descritos acima o banco com dados imputados será criado, com uma nova variável: Imputation. Esta variável é muito importante, pois ela que nos diz quais linhas correspondem a qual banco imputado. Por exemplo, para a imputação m = 1 esta variável assume valor 1 na primeira coluna em todas as linhas correspondentes ao banco imputado m = 1. Na variável, ou variáveis imputadas, os dados imputados são destacados em laranja/amarelo (Figura 13).



	Imputation_	CPF	CategoriaPretendida	Aprovado	QtdMatTrimes	Idade	QtdAulasPráticas	Mun	Escolaridade	Sexo
98100	1	[illegible]	0	1	3	34	85	1	3	2
98101	1	[illegible]	0	0	2	34	22	1	3	2
98102	1	[illegible]	0	1	5	34	45	1	3	2
98103	1	[illegible]	0	1	1	33	31	1	3	1
98104	1	[illegible]	0	0	4	35	36	1	2	2
98105	1	[illegible]	0	1	3	35	36	1	3	2
98106	1	[illegible]	0	1	5	53	83	1	2	2
98107	1	[illegible]	0	1	2	35	41	1	3	2
98108	1	[illegible]	0	1	5	36	34	1	2	1
98109	1	[illegible]	0	0	5	34	41	1	2	2
98110	1	[illegible]	0	1	1	34	20	1	2	1
98111	1	[illegible]	0	1	2	33	40	1	3	2
98112	1	[illegible]	0	0	3	39	28	1	2	2
98113	1	[illegible]	0	1	2	34	41	1	3	1
98114	1	[illegible]	0	1	5	34	32	1	3	2
98115	1	[illegible]	0	0	1	34	20	1	1	1
98116	1	[illegible]	0	0	3	34	40	1	1	1
98117	1	[illegible]	0	1	1	33	30	1	2	2
98118	1	[illegible]	0	1	2	35	66	1	2	1
98119	1	[illegible]	0	0	2	35	29	1	2	2
98120	1	[illegible]	0	0	2	39	25	1	1	1
98121	1	[illegible]	0	1	5	30	23	1	2	2
98122	1	[illegible]	0	1	1	34	20	1	3	2

Figura 13 - Banco de Dados criado após a imputação

Agora para os fins desejados, as análises podem ser feitas da forma usual. Quando uma opção de análise aceita a IM, um ícone com uma espiral aparece (Figura 14). O *software* SPSS possui várias opções de análises estatísticas para bancos com IM, como as regressões para modelos lineares generalizados abordadas neste trabalho, Regressão de Poisson e Regressão Logística (Figura 14). Além de muitas outras opções: análises descritivas, análises das frequências, comparações de médias pelo teste T, ANOVA, modelo linear misto, correlação, regressão de Cox, entre outras. (Figura 15).

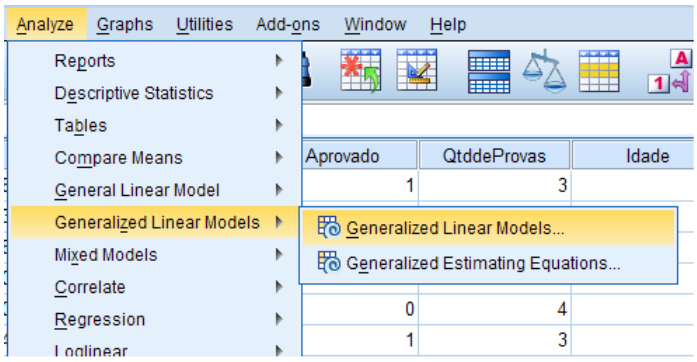


Figura 14 - Menu *analyze* do SPSS após a IM

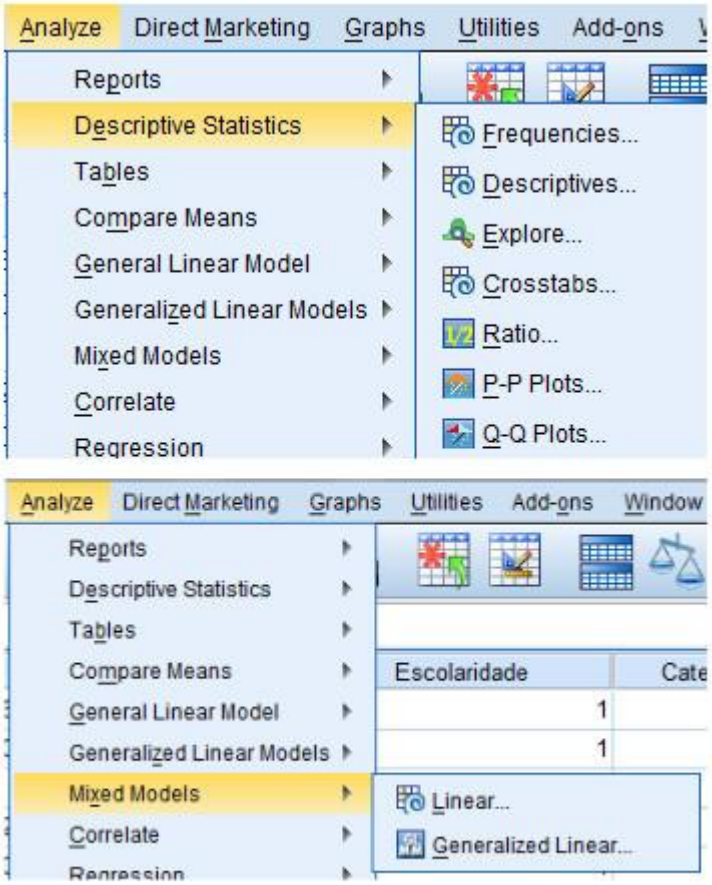


Figura 15 – Exemplos do menu *analyze* do SPSS após a IM

3 RESULTADOS

Para as análises de regressão, não utilizamos as variáveis de aulas no simulador, pois todos os candidatos que fizeram esse tipo de aula, apresentam o mesmo número de horas. Também não foi utilizada a variável pontos obtidos pelo candidato, pois esta variável tem uma relação direta com a aprovação, no caso todos os candidatos aprovados possuem até 3 pontos, assim como os reprovados possuem 4 pontos ou mais. Foram excluídos da análise os candidatos que possuíam escolaridade não informada (0,2%), para que só tivéssemos perda proveniente da simulação. Assim, foram analisados 21.999 candidatos.

Os candidatos apresentaram uma média de 33,09 horas de aulas práticas realizadas, com um desvio padrão de 19,51 apresentando uma amplitude grande, pois o mínimo de horas de aulas práticas apresentado no banco foi 2 horas e o máximo 564 horas. Estes valores mais extremos são casos a parte. Candidatos que possuem menos de 20 horas, o mínimo de horas de aulas práticas exigidas para a obtenção da carteira de motorista, realizaram parte do processo em outro estado, ao qual não temos acesso aos dados. Os valores mais altos são correspondentes ao acúmulo de horas de aulas que são guardados no cadastro do candidato. Porém esses casos representam uma parcela muito pequena (0,06%) do número total de candidatos (Tabela 1).

Os candidatos apresentaram uma média de 2,44 provas realizadas com desvio padrão de 1,77. O máximo observado no banco de dados foi de 16 provas realizadas. Contudo a maioria dos candidatos (78,7%) realizaram até 3 provas. O número de reprovações tem média igual a 1,82 e desvio padrão de 1,93 (Tabela 1).

Tabela 1 - Estatísticas Descritivas para as variáveis quantitativas

	Mínimo	Máximo	Média	Desvio Padrão
Quantidade de Aulas	2	564	33,09	19,51
Quantidade de Provas	1	16	2,44	1,77
Nº de reprovações	0	16	1,82	1,93

Dentre os candidatos analisados, 96,6% realizaram o processo em Porto Alegre e o restante fora de Porto Alegre. Podemos perceber que há mais mulheres tentando a primeira habilitação com uma proporção de 55% dos candidatos (Gráfico 1). Também vemos que a maioria (62%) obteve a aprovação (Gráfico 2).

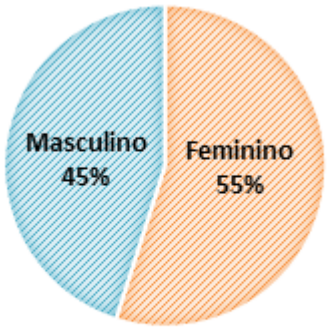


Gráfico 1 - Proporção de sexo entre os candidatos

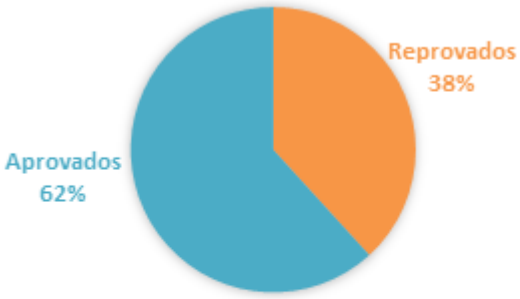


Gráfico 2 - Proporção de aprovação entre os candidatos

Entre os candidatos, 50,1% possui ensino médio e 37,7% tem ensino superior (Gráfico 3). Podemos ver apenas uma pequena proporção dos candidatos que almejam categoria AB (6,1%), a maioria (93,9%) se concentra na categoria B na primeira habilitação (Gráfico 4).

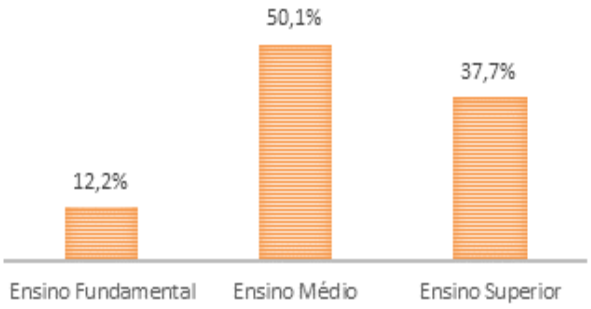


Gráfico 3 - Distribuição do nível escolar entre os candidatos

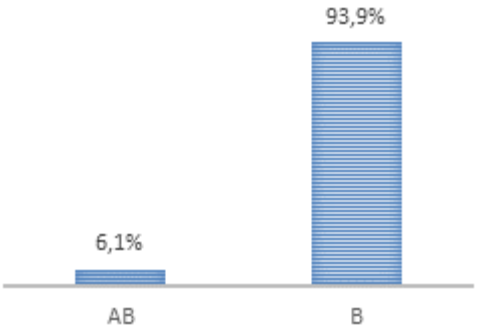


Gráfico 4 - Distribuição da categoria pretendida entre os candidatos

A maioria dos candidatos (64,3%) tem menos de 30 anos de idade. E a faixa de idade em que mais temos candidatos tirando sua primeira habilitação é entre os 21 e 24 anos de idade, concentrando 24,7% dos candidatos (Gráfico 5).

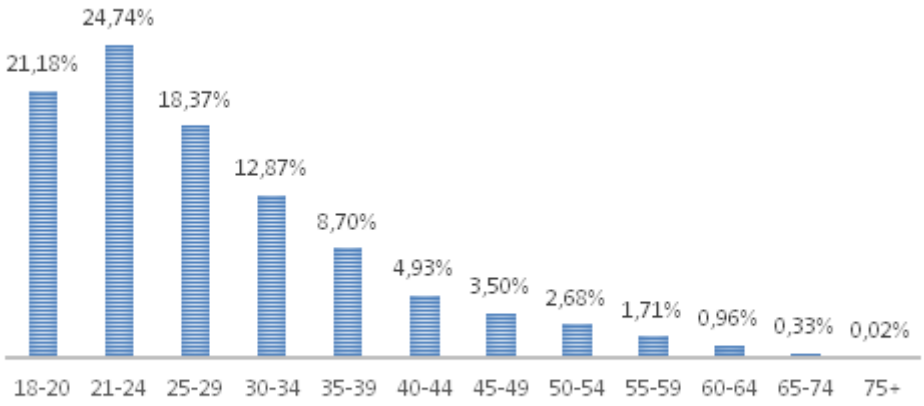


Gráfico 5 - Distribuição da idade entre os candidatos

Regressão Logística com banco de dados completo

Primeiramente se realizou a análise de regressão logística com o banco de dados completo. Nessa regressão foi usada a "aprovação" como variável resposta, utilizando não aprovado como categoria de referência. Vamos citar não aprovação como reprovação, para simplificar as conclusões. Todas as variáveis escolhidas para o modelo se mostraram significativas. Vamos utilizar a sigla RC para Razão de Chances.

A chance de ser reprovado para quem abriu RENACH em Porto Alegre é 1,715 vezes a chance de quem abriu RENACH fora de Porto Alegre. Candidatos do sexo masculino possuem 2,35 vezes a chance de serem reprovados do que candidatos do sexo feminino (Tabela 2).

Ensino fundamental em relação ao ensino superior não se mostrou significativo a um nível de significância de 5%. A chance de ser reprovado para candidatos com ensino médio é 0,838 vezes a chance de quem possui ensino superior. Ou seja, a chance de reprovar é menor para quem possui ensino médio (Tabela 2).

Quem pretende obter a primeira habilitação para categoria AB tem a chance maior de ser reprovado do que quem quer categoria B. Pois a chance de ser reprovado para quem quer AB é 1,188 a chance de quem quer B (Tabela 2).

A chance de reprovar tende a diminuir conforme a idade do candidato aumenta (RC=0,938). Em relação ao número de provas prestadas, essa a chance de reprovação tende a diminuir conforme aumenta o número de provas (RC=0,832). Mas, a chance de reprovar tende a aumentar conforme o número de aulas cresce (RC=1,009; Tabela 2).

Tabela 2 - Regressão logística binária para aprovação do candidato

	RC	IC (95%)	*p* -valor
Município			
Porto Alegre	1,715	(1,462 - 2,01)	0,000
Fora de Porto Alegre	1		
Sexo			
Masculino	2,350	(2,202 - 2,508)	0,000
Feminino	1		
Escolaridade			
Fundamental	0,906	(0,815 - 1,007)	0,068
Médio	0,838	(0,786 - 0,894)	0,000
Superior	1		
Categoria			
AB	1,188	(1,038 - 1,36)	0,012
B	1		
Idade	0,938	(0,935 - 0,942)	0,000
Quantidade de aulas	1,009	(1,007 - 1,011)	0,000
Quantidade de provas	0,832	(0,817 - 0,847)	0,000

Regressão Logística para Banco com IM

Separadamente, foram simuladas perdas de 5%, 10% e 20% tanto na variável resposta “aprovação” quanto na variável preditora “escolaridade”. Para a IM, usamos as variáveis: município, sexo, escolaridade, categoria, idade, quantidade de aulas e quantidade de provas. Ou seja, as mesmas que são usadas na regressão logística. Foi usado um **m** igual a 10, e embora nossos dados apresentam um padrão monotônico o método de IM escolhido foi o *fully*

conditional specification, utilizando *predictive mean matching*. A escolha do método foi porque ao usarmos o método monotônico alguns valores deixavam de ser imputados, pois o modelo não ficava bem ajustado.

Outra consideração a ser feita a respeito da IM no aplicativo é que o SPSS não apresenta a opção *pooled* (combinação dos resultados de cada uma das **m** imputações, conforme regra de Rubin) para todas as estatísticas. No caso temos somente o Beta ($\hat{\beta}$) estimado, seu desvio padrão e seu intervalo de confiança, opções *default* do SPSS. Então a exp($\hat{\beta}$), que á a Razão de Chances, precisou ser calculada posteriormente para as devidas comparações e conclusões.

Uma questão a ser observada é o fato de que para a regressão logística após a IM são processadas **m** regressões, correspondentes a cada **m** banco imputado, assim como combinar estes **m** resultados. Neste trabalho foi utilizado **m=10** imputações. Para cada iteração que o algoritmo executa, o tamanho do passo é reduzido por um fator de 0,5 até o logaritmo da verossimilhança aumente ou o número de *Maximum Step-Halving* ser atingido[4]. Este último deve ser aumentado para que as regressões possam ser feitas. No caso, aumentou se de 5 (*default* do SPSS) para 25. Quanto maior o **m** definido para a IM, maior o *Maximum Step-Halving* deverá ser.

Tabela 3 - Regressão Logística binária para aprovação em bancos com IM na variável preditora escolaridade

	5% de perda			10% de perda			20% de perda		
	RC	IC (95%)	*p* -valor	RC	IC (95%)	*p* -valor	RC	IC (95%)	*p* -valor
Município									
Porto Alegre	1,716	(1,463 - 2,011)	0,000	1,718	(1,465 - 2,013)	0,000	1,718	(1,465 - 2,015)	0,000
Fora de Porto Alegre	1			1			1		
Sexo									
Masculino	2,349	(2,201 - 2,506)	0,000	2,356	(2,207 - 2,514)	0,000	2,354	(2,205 - 2,511)	0,000
Feminino	1			1			1		
Escolaridade									
Fundamental	0,911	(0,817 - 1,015)	0,091	0,890	(0,795 - 0,996)	0,043	0,897	(0,800 - 1,005)	0,062
Médio	0,841	(0,785 - 0,900)	0,000	0,834	(0,777 - 0,894)	0,000	0,830	(0,776 - 0,888)	0,000
Superior	1		.	1			1		
Categoria									
AB	1,189	(1,038 - 1,360)	0,012	1,191	(1,040 - 1,363)	0,011	1,192	(1,041 - 1,364)	0,011
B	1			1			1		
Idade	0,938	(0,934 - 0,941)	0,000	0,938	(0,935 - 0,941)	0,000	0,938	(0,934 - 0,941)	0,000
Quantidade de aulas	1,009	(1,007 - 1,011)	0,000	1,009	(1,007 - 1,011)	0,000	1,009	(1,007 - 1,011)	0,000
Quantidade de provas	0,832	(0,817 - 0,847)	0,000	0,832	(0,817 - 0,847)	0,000	0,832	(0,817 - 0,847)	0,000

Podemos ver que as diferenças entre as razões de chance para os bancos com IM (Tabela 3) e as razões de chance para o banco completo (Tabela 2) aparecem somente na terceira casa decimal, sendo que o mesmo acontece para os intervalos de confiança. A variável escolaridade, nossa variável imputada, apresentou uma diferença maior nas estimativas comparada à análise do banco de dados completo. Destaca-se também que escolaridade apresentou diferença na significância para a categoria nível fundamental em relação ao nível superior no caso da IM para perda de 10%, assim se continuarmos seguindo o nível de significância de 5% rejeitaríamos a hipótese nula diferentemente do que acontece quando analisamos o banco completo. Porém vale ressaltar que se usássemos um nível de significância menos "rígido" de 10% teríamos a mesma conclusão para todos os casos (Tabela 3).

As variáveis contínuas e discretas (idade, quantidade de aulas e quantidade de provas) não apresentaram nenhuma diferença da razão de chances para os bancos com IM em comparação com a razão de chances para o banco completo. Apenas a variável idade apresentou para o seu IC uma pequena diferença na terceira casa decimal (Tabela 3).

Fraction Missing Information (FMI) é uma medida para o risco de não respostas, e serve para medir o nível de incerteza sobre os valores a serem imputados para os casos de falta de dados[15].

Nota-se que conforme o número de dados faltantes cresce, o FMI também cresce, exceto para a variável escolaridade na categoria médio, que para a IM para perda de 10% apresentou um valor mais alto que para a perda de 20%, que por sua vez se mostrou menor que para a perda de 5%. A escolaridade, variável com dados imputados foi a que mostrou uma maior incerteza (FMI) (Tabela 4).

Tabela 4 - Nível de Incerteza (FMI) para Regressão Logística em dados com IM na variável escolaridade

	5%	10%	20%
Município			
Porto Alegre	0,000	0,000	0,001
Fora de Porto Alegre			
Sexo			
Masculino	0,003	0,005	0,008
Feminino			
Escolaridade			
Fundamental	0,047	0,122	0,150
Médio	0,082	0,156	0,073
Superior			
Categoria			
AB	0,000	0,001	0,001
B			
Idade	0,004	0,012	0,014
Quantidade de aulas	0,001	0,003	0,002
Quantidade de provas	0,000	0,000	0,001

Tabela 5 - Regressão Logística binária para aprovação em bancos com IM na variável resposta aprovação

	5% de perda			10% de perda			20% de perda		
	RC	IC (95%)	*p* -valor	RC	IC (95%)	*p* -valor	RC	IC (95%)	*p* -valor
Município									
Porto Alegre	1,719	(1,450 - 2,040)	0,000	1,768	(1,487 - 2,104)	0,000	1,804	(1,517 - 2,146)	0,000
Fora de Porto Alegre	1			1			1		
Sexo									
Masculino	2,375	(2,218 - 2,539)	0,000	2,370	(2,214 - 2,539)	0,000	2,363	(2,205 - 2,531)	0,000
Feminino	1			1			1		
Escolaridade									
Fundamental	0,914	(0,820 - 1,019)	0,106	0,896	(0,804 - 0,999)	0,048	0,908	(0,799 - 1,031)	0,135
Médio	0,843	(0,788 - 0,901)	0,000	0,833	(0,777 - 0,892)	0,000	0,827	(0,769 - 0,887)	0,000
Superior	1		.	1			1		
Categoria									
AB	1,178	(1,025 - 1,353)	0,021	1,147	(0,996 - 1,321)	0,057	1,091	(0,951 - 1,251)	0,213
B	1			1			1		
Idade	0,937	(0,934 - 0,940)	0,000	0,937	(0,933 - 0,940)	0,000	0,937	(0,932 - 0,940)	0,000
Quantidade de aulas	1,009	(1,007 - 1,011)	0,000	1,009	(1,007 - 1,011)	0,000	1,009	(1,007 - 1,011)	0,000
Quantidade de provas	0,832	(0,817 - 0,847)	0,000	0,830	(0,815 - 0,846)	0,000	0,832	(0,817 - 0,847)	0,000

De modo geral quando a IM é para a variável resposta, as diferenças entre as estimativas para os bancos com imputação e as estimativas para o banco completo acabam sendo um pouco maiores, se comparadas às diferenças entre as estimativas para os bancos com imputação na

variável preditora escolaridade e as estimativas para o banco completo. Assim como os IC, com 95% de confiança, na maioria dos casos apresentou maior amplitude, ou seja, menos precisão (Tabela 5).

As variáveis discretas continuam apresentando menor diferença do que as variáveis categóricas. Mesmo com 20% de perda, ainda conseguimos obter através da IM estimativas coerentes e satisfatórias para a regressão logística (Tabela 5).

Tabela 6 - Nível de Incerteza (FMI) para Regressão Logística em dados com IM na variável aprovação

	5%	10%	20%
Município			
Porto Alegre	0,091	0,084	0,211
Fora de Porto Alegre			
Sexo			
Masculino	0,066	0,097	0,106
Feminino			
Escolaridade			
Fundamental	0,051	0,051	0,304
Médio	0,062	0,123	0,179
Superior			
Categoria			
AB	0,057	0,098	0,055
B			
Idade	0,117	0,136	0,177
Quantidade de aulas	0,043	0,089	0,092
Quantidade de provas	0,081	0,078	0,078

Houve um aumento no FMI quando imputamos a variável resposta, em relação ao FMI quando imputamos a variável preditora. Entre 5% e 10% as diferenças não são muito grandes, mas já de 10% para 20% o FMI tem um maior acréscimo para quase todas as variáveis, exceto para categoria pretendida em que decai (Tabela 6).

Regressão de Poisson

Para a regressão de Poisson foi usada a variável "número de reprovações" como variável resposta. Para este modelo não se usou como variável preditora o número de provas, pois a mesma tem relação direta com o número de reprovações. Todas as variáveis escolhidas para o modelo se mostraram significativas.

A partir do ajuste do modelo com o banco de dados completo, foi possível se observar que para quem abriu RENACH em Porto Alegre a média do número de reprovações é 1,34 vezes a média de quem abriu RENACH fora de Porto Alegre. Para candidatos do sexo masculino a média do número de reprovações é 0,697 vezes a média de quem é do sexo feminino. Ou seja, em média, homens tem um número menor de reprovações do que as mulheres.

Para quem almeja categoria AB a média do número de reprovações é 0,851 vezes a média do que quem almeja somente categoria B. Ou seja, quem pretende categoria AB reprova, em média, menos vezes que quem pretende categoria B. A média do número de reprovações para quem possui escolaridade nível fundamental é 1,149 vezes a média do que quem possui escolaridade nível superior. E para quem possui ensino médio é 1,089 vezes maior.

O número de reprovações será 1,3% maior para cada ano a mais na idade do candidato. Assim como, 0,5% maior para cada hora de aula a mais feita pelo candidato (Tabela 7).

Tabela 7 - Regressão de Poisson para nº de reprovações do candidato

	Exp($\hat{\beta}$)	IC (95%)	*p* -valor
Município			
Porto Alegre	1,34	(1,261 - 1,423)	0,000
Fora de Porto Alegre	1		
Sexo			
Masculino	0,697	(0,682 - 0,713)	0,000
Feminino	1		
Escolaridade			
Fundamental	1,149	(1,111 - 1,188)	0,000
Médio	1,089	(1,065 - 1,113)	0,000
Superior	1		
Categoria			
AB	0,851	(0,81 - 0,894)	0,000
B	1		
Idade	1,013	(1,012 - 1,014)	0,000
Quantidade de aulas	1,005	(1,005 - 1,006)	0,000

Regressão de Poisson para Banco com IM

Em momentos distintos, foram simuladas perdas de 5%, 10% e 20% tanto na variável resposta "número de reprovações" quanto na variável preditora "escolaridade". Para a IM, usamos as variáveis: município, sexo, escolaridade, categoria, idade e quantidade de aulas. As mesmas que são usadas na regressão de Poisson. Foi escolhido um **m** igual a 10. Como a variável resposta número de reprovações é uma variável quantitativa discreta, ao tentarmos usar o método monotônico, ou mesmo o método o *fully conditional specification*, utilizando regressão linear, os dados imputados apresentaram um erro onde o software diferenciava dois tipos de zeros, que apresentavam a mesma formatação e o mesmo valor, mas não eram tratados como valores iguais pelo SPSS. Para contornar o problema, foi usado então o método *predictive mean matching*. Outra forma também para a resolução do problema seria a criação de uma variável onde se multiplica o número de reprovações por 100 e depois se divide por 100 (o mesmo que multiplicar por um). Assim os zeros passam a ser tratados todos de forma igual.

Tabela 8 - Regressão de Poisson para bancos com IM na variável preditora escolaridade

	5% de perda			10% de perda			20% de perda		
	Exp($\hat{\beta}$)	IC (95%)	*p* -valor	Exp($\hat{\beta}$)	IC (95%)	*p* -valor	Exp($\hat{\beta}$)	IC (95%)	*p* -valor
Município									
Porto Alegre	1,338	(1,259 - 1,421)	0,000	1,338	(1,258 - 1,420)	0,000	1,339	(1,259 - 1,423)	0,000
Fora de Porto Alegre	1			1			1		
Sexo									
Masculino	0,698	(0,682 - 0,713)	0,000	0,698	(0,682 - 0,713)	0,000	0,698	(0,682 - 0,713)	0,000
Feminino	1			1			1		
Escolaridade									
Fundamental	1,145	(1,105 - 1,185)	0,000	1,140	(1,095 - 1,186)	0,000	1,138	(1,095 - 1,182)	0,000
Médio	1,091	(1,067 - 1,116)	0,000	1,089	(1,062 - 1,116)	0,000	1,081	(1,053 - 1,110)	0,000
Superior	1		.	1			1		
Categoria									
AB	0,851	(0,809 - 0,894)	0,000	0,851	(0,810 - 0,894)	0,000	0,852	(0,811 - 0,894)	0,000
B	1			1			1		
Idade	1,013	(1,012 - 1,014)	0,000	1,013	(1,012 - 1,014)	0,000	1,013	(1,012 - 1,014)	0,000
Quantidade de aulas	1,005	(1,005 - 1,006)	0,000	1,005	(1,005 - 1,006)	0,000	1,005	(1,005 - 1,006)	0,000

Como para os modelos de regressão logística, os modelos de Poisson mostraram diferenças entre as estimativas da exp(β) dos bancos com imputação em relação ao banco completo somente na terceira casa decimal. Pode se notar que mesmo com o aumento da perda, a diferença continua sendo bem pequena, ou quase nenhuma. Escolaridade foi a variável que apresentou maior

diferença entre as estimativas para os bancos com IM e as estimativas para o banco completo (Tabela 8).

De novo, as variáveis quantitativas (idade e quantidade de aulas) não alteraram em nada suas estimativas para os bancos com IM em comparação ao banco completo.

Tabela 9 - Nível de Incerteza (FMI) para Regressão de Poisson com IM na variável preditora escolaridade

	5%	10%	20%
Município			
Porto Alegre	0,000	0,001	0,001
Fora de Porto Alegre			
Sexo			
Masculino	0,004	0,021	0,020
Feminino			
Escolaridade			
Fundamental	0,081	0,306	0,242
Médio	0,059	0,226	0,327
Superior			
Categoria			
AB	0,000	0,002	0,001
B			
Idade	0,010	0,043	0,032
Quantidade de aulas	0,001	0,006	0,005

Nota-se que existe um aumento no FMI conforme a perda cresce de 5% para 10%. Já de 10% para 20% na maioria dos casos vemos uma diminuição no FMI. A variável com valores imputados, escolaridade, mostrou o maior nível de incerteza (Tabela 9).

Tabela 10 - Regressão de Poisson para bancos com IM na variável resposta nº de reprovações

	5% de perda			10% de perda			20% de perda		
	Exp($\hat{\beta}$)	IC (95%)	*p* -valor	Exp($\hat{\beta}$)	IC (95%)	*p* -valor	Exp($\hat{\beta}$)	IC (95%)	*p* -valor
Município									
Porto Alegre	1,340	(1,256 - 1,429)	0,000	1,347	(1,261 - 1,440)	0,000	1,290	(1,192 - 1,397)	0,000
Fora de Porto Alegre	1			1			1		
Sexo									
Masculino	0,698	(0,682 - 0,713)	0,000	0,696	(0,680 - 0,711)	0,000	0,700	(0,683 - 0,717)	0,000
Feminino	1			1			1		
Escolaridade									
Fundamental	1,140	(1,101 - 1,179)	0,000	1,129	(1,090 - 1,168)	0,000	1,122	(1,076 - 1,168)	0,000
Médio	1,091	(1,067 - 1,116)	0,000	1,090	(1,065 - 1,114)	0,000	1,090	(1,065 - 1,115)	0,000
Superior	1		.	1			1		
Categoria									
AB	0,840	(0,799 - 0,884)	0,000	0,839	(0,797 - 0,881)	0,000	0,851	(0,808 - 0,896)	0,000
B	1			1			1		
Idade	1,013	(1,012 - 1,014)	0,000	1,013	(1,012 - 1,014)	0,000	1,013	(1,012 - 1,014)	0,000
Quantidade de aulas	1,005	(1,005 - 1,006)	0,000	1,005	(1,005 - 1,006)	0,000	1,005	(1,005 - 1,006)	0,000

As estimativas da regressão de Poisson para os bancos de dados com IM na variável resposta número de reprovações foram as que mais diferiram das estimativas do banco completo. Exceto pelas estimativas das variáveis quantitativas (idade e quantidade de aulas) que apresentam quase nenhuma ou nenhuma diferença (Tabela 10).

Em relação ao nível de incerteza, vemos um acréscimo do FMI conforme a perda aumenta. Destaque para a variável Município que apresentou os valores mais altos de FMI (Tabela 11).

Tabela 11 - Nível de Incerteza (FMI) para Regressão de Poisson com IM na variável resposta nº de reprovações

	5%	10%	20%
Município			
Porto Alegre	0,057	0,126	0,374
Fora de Porto Alegre			
Sexo			
Masculino	0,026	0,037	0,188
Feminino			
Escolaridade			
Fundamental	0,046	0,067	0,321
Médio	0,048	0,059	0,104
Superior			
Categoria			
AB	0,056	0,016	0,122
B			
Idade	0,046	0,095	0,124
Quantidade de aulas	0,031	0,029	0,089

4 DISCUSSÃO

O SPSS possui uma interface amigável e intuitiva, que permite o uso fácil da IM. Possui mais de uma opção de modelagem para a IM, além de possuir a opção de deixar o SPSS escolher automaticamente o método de imputação, que se mostrou eficiente em identificar o padrão monotônico que os dados apresentavam. Porém não tem implementado alguns outros modelos de imputação múltipla comuns, como por exemplo, *Bayesian Linear Regression*[2,4]. Outro fator interessante é possuir várias opções de análises finais para bancos com IM.

A IM no SPSS apresentou alguns problemas, como o caso das diferenças dos zeros quando foi preciso definir os limites da variável a ser imputada. Como mostra a Figura 16, alguns dos zeros imputados se juntaram aos zeros já existentes no banco, porém outros 77 zeros imputados foram considerados diferentes. Isso acontece quando deixamos como automática a escolha do método de IM ou quando escolhemos a opção *linear regression* para a IM, tanto para o método monotônico quando para o *fully conditional specification*. Se não definirmos os limites da variável a ser imputada nos métodos citados usando a opção regressão linear, valores fora dos limites, que são os possíveis, podem vir a ser imputados. Outra questão é a necessidade de aumentar o *Maximum Step-Halving* para as regressões dos bancos após IM.

1	Valid	0	77	,4	,4	,4
		0	6280	28,5	28,5	28,9
		1	5332	24,2	24,2	53,1
		2	4006	18,2	18,2	71,3

Figura 16 – *Output* da tabela de frequências da variável nº de reprovações após IM

Quando avaliamos os valores estimados para os dados com valores imputados e comparamos com as estimativas obtidas com os dados completos, podemos notar que a análise de regressão logística apresentou uma menor diferença do que a regressão de Poisson. Assim como para o nível de incerteza, que foi maior para a regressão de Poisson.

Considerando que as diferenças vistas entre as estimativas das regressões para os bancos completos e as estimativas das regressões para os bancos com IM foram quase nulas ou bem pequenas, podemos considerar satisfatório o uso da IM. Então se realmente existisse a perda e a IM fosse necessária não teríamos grandes prejuízos nos resultados e suas respectivas

interpretações. Contudo, é preciso ressaltar que o tamanho de amostra grande, 21.999 candidatos, implica em maior robustez dos modelos. Ou seja, existe a possibilidade de que as diferenças entre as estimativas para o banco completo e as estimativas para os bancos com IM sejam tão pequenas ou quase nulas devido ao tamanho amostral.

Considerando também o fato de que a perda simulada foi produzida em somente uma variável de cada vez, gerando assim um modelo mais simples de imputação, devido ao foco maior do trabalho ser ilustrar a IM no SPSS, a falta de maior complexidade no padrão da perda pode também justificar as diferenças observadas.

O tempo de processamento para a IM na variável escolaridade foi em torno de 1 minuto, não mostrando grande diferença entre os níveis de perda. Para a IM na variável aprovação o tempo de processamento foi em torno de 48 segundos, de novo não mostrando grande diferença entre os níveis de perda. A IM para a variável número de reprovações, única variável quantitativa imputada, apresentou o menor tempo de processamento, ficando em torno de 15 segundos. O tempo de processamento das regressões, tanto de Poisson quanto a Logística, para os bancos imputados não foi maior do que 35 segundos. Ou seja, hoje em dia, mesmo com um banco de dados relativamente grande e o uso de m=10 imputações, o tempo computacional dispendido é pequeno e vem ao encontro da literatura que refere que o avanço computacional favorece o uso da imputação múltipla como uma boa alternativa para o tratamento de dados faltantes[1].

As escolhas dos níveis de perda (5%, 10% e 20%) foram baseadas na recomendação da escolha entre IM ou Imputação Única[2]. Porém, pode ocorrer em casos reais perdas maiores que 20%. Assim fica como ideia para um trabalho futuro, testar a IM no SPSS com outras percentagens de dados faltantes, maiores do que as que foram aqui usadas. Assim como com tamanhos amostrais menores, dado que na prática é bastante recorrente a existência de bancos de dados com poucas observações, devido a vários fatores como, falta de recursos, dificuldade na coleta, entre outros.

Uma limitação desse estudo foi não ter explorado a falta de dados em mais variáveis ao mesmo tempo. Por exemplo, ter bancos de dados com padrão não monotônico que computacionalmente são mais difíceis de lidar do que quando temos um padrão univariado ou um padrão monotônico[13].

Também se deixa como possibilidade de um trabalho futuro, a investigação mais a fundo da causa do problema dos zeros para a imputação em uma variável quantitativa. E as possíveis soluções deste problema.

Por fim, o objetivo do trabalho era mostrar o uso da IM no SPSS e a partir de um banco de dados reais isso foi possível, utilizando duas técnicas estatísticas (Regressão Logística e Regressão de Poisson) que são muito usadas em várias áreas de conhecimento. Entretanto cabe ressaltar que o fato do *software* aqui usado ser pago, pode ser um limitador para uma parte dos usuários de Estatística. Mas, de qualquer forma, espera-se que esse trabalho possa servir de apoio para quem quer usar a técnica da Imputação Múltipla.

Um agradecimento especial ao Departamento Estadual de Trânsito do RS (Detran/RS) pela cedência dos dados para que realizássemos a técnica em dados que representassem a realidade, onde podemos identificar possíveis problemas e procurar suas soluções, ilustrando assim de forma mais verídica o uso da IM no SPSS.

REFERÊNCIAS BIBLIOGRÁFICAS

1. NUNES, Luciana N.; KLÜCK, Mariza M.; FACHEL, Jandyra M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Rev. Bras. Epidemiol., São Paulo, v. 13, n. 4, p. 596-606, Dezembro. 2010.
2. **NUNES, Luciana N.** Métodos de imputação de dados aplicados na área da saúde. 2007. 120 f. Tese (Doutorado em Epidemiologia) – Universidade Federal do Rio Grande do Sul, Porto Alegre.
3. **STERNE, Jonathan A C et al.** Multiple Imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338:b2393, Junho. 2009.
4. IBM Corp. Released 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.
5. **HORTON, Nicholas J.; KLEINMAN,** Ken P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, Alexandria, v. 61, n. 1, p. 79-90, Fevereiro. 2007.
6. **BARACHO, Stella M. L. N.** Tratamentos de dados ausentes em estudos longitudinais. 2003. 24 f. Tese (Mestrado em Estatística) – Universidade Federal de Minas Gerais, Belo Horizonte.
7. **SCHAFER, Joseph L.; GRAHAM,** John W. Missing Data: Our View of the State of the Art. Psychological Methods, v. 7, n. 2, p. 147-177, Junho 2002.
8. **SCHEFFER, Judi;** Dealing with Missing Data. Research Letters in the Information and Mathematical Sciences, Nova Zelândia, v. 3, p. 153-160, Abril. 2002.
9. **YUAN, Yang C.;** Multiple Imputation for Missing Data: Concepts and New Development. SAS Software Technical Papers. 2013.
10. NAKANO, Tatiana C.; SAMPAIO, Maria H. L.; SILVA, Adriana B. Atenção e inteligência em candidatos à primeira carteira nacional de habilitação, Boletim de Psicologia, v. 61, n. 134, p. 63-78, Junho. 2011.
11. Site: Portal Action. Disponível em: http://www.portalaction.com.br/analise-de-regressao/regressao-logistica. Acessado em: 18/10/2016

12. TADANO, Yara S.; UGAYA, Cássia M. L.; FRANCO, A. T. Método de Regressão de Poisson: Metodologia para avaliação do impacto da poluição atmosférica na saúde populacional, Ambiente e Saúde, Campinas, v. 12, n. 2, p. 405-414, Julho. 2009.
13. **DONG, Yiran; PENG**, Chao-Ying J. Principled missing data methods for researchers, SpringerPlus, , v. 2, n. 1, p. 222, Maio. 2013.
14. Site: Applied Missing Data .Disponível em: http://www.appliedmissingdata.com/spss-multiple-imputation.pdf Acessado em: 31/08/2016
15. **WAGNER**, James. The fraction of missing information as a tool for monitoring the quality of survey data, Public Opinion Quarterly, Oxford, v. 74, n. 2, p. 223-243, Março. 2010.