5 Descrição e Análise dos Dados

Nesta etapa da pesquisa serão descritos e analisados os dados coletados. Nesse sentido, o primeiro subtópicos deste capítulo consiste em apresentar uma análise e descrição dos dados coletados com caracterização da amostra (frequência, média, desvio padrão, máximo e mínimo). O subtópicos seguinte irá descrever como ocorreu a preparação da base de dados (variáveis selecionadas e os tratamentos realizados). Em seguida, é aplicada a técnica de regressão múltipla em que serão apresentados os resultados da aplicação avaliando o valor de R2 e a significância do modelo por meio do teste de ANOVA. Por fim, esta seção se encerrará com a exposição dos resultados.

5.1 Descrição dos dados Coletados

A preparação do banco de dados desta pesquisa iniciou-se imediatamente após a coleta de dados. Nesta etapa de preparação dos dados foi realizado alguns ajustes isso porque a base inicial continha 2.049 casos, porém, no decorrer do desenvolvimento do trabalho foi necessário realizar a exclusão de dados faltosos e também dados que apresentaram inconsistência de acordo com a avaliação do pesquisador. Desse modo, o total final de casos foi de 1.300.

Quadro 1 – Identificação das variáveis de estudo

|  |  |
| --- | --- |
| **Siglas** | **Indicador** |
| ID1 | Ano de referência |
| ID4 | Número de Internações |
| ID5 | Dias de Permanência |
| ID6 | Número de Óbitos |
| ID7 | Proporção Internações <= Número de Internações / População Total |
| ID8 | Proporção de dias Permanência <= Dias de Permanência / Número de Internações |
| ID9 | Proporção de Óbitos<= Número de Óbitos/Número Internações |
| ID10 | IDHM |
| ID11 | PIB per Capita série revisada |
| ID12 | População total do município do ano de referência |
| ID13 | População total atendida com esgotamento sanitário |
| ID14 | Quantidade de ligações ativas de esgotos |
| ID15 | Volume de esgotos coletado |
| ID16 | Volume de esgotos tratado |
| ID17 | Índice de coleta de esgoto |
| ID18 | Índice de tratamento de esgoto |

Fonte: Resultados da pesquisa.

Inicialmente as variáveis dependentes foram coletadas em sua forma absoluta, ou seja, quantificavam o número de pessoas internadas, quanto tempo ficavam internadas e o número de óbitos registrados no município. Neste estudo viu-se a necessidade de transformá-las em proporções, criando variáveis derivadas como apresentado no Quadro1. Esta decisão se dá ao fato de abordarmos várias cidades do estado, que possuem peculiaridades diferentes, como diferentes tamanhos de população, que podem vir a influenciar os resultados da pesquisa.

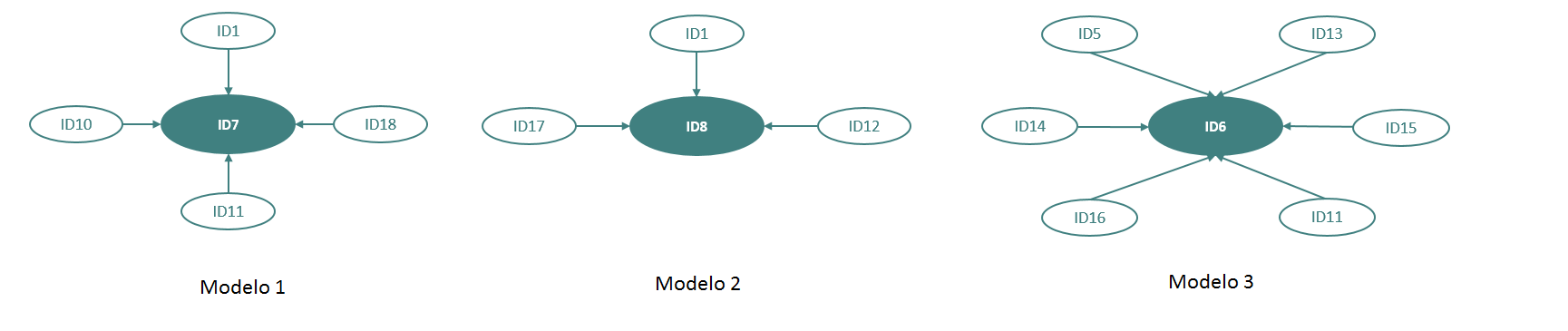
Tabela 1 – Descrição das variáveis do estudo.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variável** | **N** | **Mínimo** | **Máximo** | **Média** | **Desvio Padrão** |
| ID1  ID4 | 1.300  1.300 | 2008  0,00 | 2014  3.230,00 | 2011,33  88,19 | 1,930  195,43 |
| ID5 | 1.300 | 0,00 | 17.231,00 | 347,57 | 1.060,59 |
| ID6 | 1.300 | 0,00 | 31,00 | 1,10 | 2,36 |
| ID7 | 1.300 | 0,00 | 0,03 | 0,00 | 0,00 |
| ID8 | 1.300 | 0,00 | 14,33 | 3,71 | 1,42 |
| ID10 | 1.300 | 0,42 | 0,81 | 0,68 | 0,07 |
| ID11 | 1.300 | 3.279,70 | 144.320,88 | 17.256,06 | 14.760,16 |
| ID12 | 1.300 | 3.432,00 | 2.491.109,00 | 74.436,10 | 196.709,52 |
| ID13 | 1.300 | 33,00 | 2.491.109,00 | 64.123,60 | 195.182,11 |
| ID14 | 1.300 | 0,00 | 574.936,00 | 17.613,82 | 44.861,00 |
| ID15 | 1.300 | 0,07 | 120.242,67 | 2.934,81 | 9.494,70 |
| ID16 | 1.300 | 0,00 | 107.399,70 | 1.363,14 | 7.705,89 |
| ID17 | 1.300 | 0,02 | 100,00 | 67,05 | 22,99 |
| ID18 | 1.300 | 0,00 | 100,00 | 30,11 | 41,72 |

Fonte: Resultados da pesquisa.

Definiu-se como Modelo 1 aquele que tem como variável dependente a proporção de pessoas internadas. Como Modelo 2 a variável dependente que representa a média, em dias, do tempo de internação e, por fim, o Modelo 3 é composto pela variável dependente número de óbitos. O estudo visa descobrir o impacto das variáveis pré-selecionadas, chamadas de variáveis explicativa, nas respectivas dependentes, conforme mostra a Figura 1

Figura 1 – Modelos propostos para estudo



Na Seção 5.4 será apresentado a teoria referente a aplicação da técnica de Regressão Linear Múltipla.

5.4 Aplicação da Regressão

**5.4.1 Regressão linear múltipla**

Sendo Hair, Anderson Taham & Black (2005) análise de regressão múltipla consiste em uma técnica estatística que pode ser empregada para analisar a relação de causa e efeito entre uma única variável dependente (Y) e diversas variáveis explicativas {x1, x2 …, xn}. De modo, geral a ideia é predizer os valores da variável dependente por meio das variáveis explicativas. Conforme Fávero, Belfiore, Silva & Chan (2009), estima-se o impacto do incremento de cada variável explicativa – que se traduz como peso de cada variável explicativa – sobre a respectiva variação da variável dependente. Além disso, os pesos denotam a contribuição relativa das variáveis explicativas para a previsão geral e facilitam a interpretação sobre a influência de cada variável explicativa em fazer a previsão (Hair, Anderson Taham & Black, 2005).

O modelo de regressão múltipla pode ser expresso por (CHARNET, 2008):

 (1)

Em que: é o fenômeno em estudo (variável dependente);representa o intercepto (constante); são os coeficientes associados a cada variável explicativa (coeficientes angulares); *xki* são as variáveis explicativas (explicativas) com ;  é o termo do erro. O erro , também chamado de resíduo, representa possíveis variáveis que não foram inseridas no modelo, mas que também contribuiriam para a explicação de  (CHARNET, 2008).

**5.4.2 Adequação do modelo**

Após estimar o modelo é necessário verificar se de fato representa o que deseja-se medir, isto é, o modelo expressa a realidade e, consequentemente, proporciona uma melhoria nas decisões (CORRAR *et. al.,* 2007). Portanto, avaliou-se: a qualidade do ajuste por meio dos coeficientes de determinação (*R*²) e determinação ajustado (*R*²*ajustado*) e a multicolinearidade entre as variáveis explicativas por meio da análise dos fatores de inflação da variância (*VIF*).

**5.4.3 Multicolinearidade**

Outro aspecto importante no ajuste de modelos de regressão linear múltipla é a multicolinearidade. Objetiva-se investigar se há multicolinearidade entre as variáveis explicativas, visto que a forte correlação entre elas acarreta vários efeitos negativos no ajuste do modelo de regressão. A multicolinearidade é um problema comum em regressão linear múltipla, indicando que existe uma relação de linearidade entre as variáveis regressoras, prejudicando assim a estimação dos coeficientes de regressão. O problema de multicolinearidade torna a estimativa dos parâmetros imprecisa, por conta de um alto valor do erro padrão, o que não é conveniente estatisticamente (KUTNER et al., 2004; TAMHANE, DUNLOP, 2000).

A presença de multicolinearidade pode ser observada quando o coeficiente de determinação  apresenta um alto valor, mas nenhum dos coeficientes da regressão é estatisticamente significativo. Portanto, uma das formas de detecção é avaliar o Fator de Inflação da Variância (*VIF*). Esse fator mede o grau de associação entre as variáveis explicativas, a partir do coeficiente de determinação do modelo de regressão ajustado apenas entre as variáveis explicativas. O Fator de Inflação da Variância é definido como (BERK, 1977):



Pode-se observar que, quanto maior o , maior é o valor de *VIF*, indicando alta colinearidade.

Valores de *VIFi* maiores que 10 correspondem a um coeficiente de determinação (KUTNER et al., 2004; TAMHANE & DUNLOP, 2000). Partindo dessa ideia/princípio de que coeficiente de determinação acima de 90% (entre variáveis explicativas) caracteriza presença de multicolinearidade, então, justifica-se o porquê de considerar valores de *VIF*i maiores que 10, pois 

**5.4.4 Seleção de variáveis**

Os métodos de seleção têm sido desenvolvidos para identificar se é necessário a inclusão de todas as variáveis regressoras disponíveis ou incluir apenas um subconjunto destas variáveis. No estudo em questão utilizamos o método “passo atrás” (backward), ele consiste na eliminação de uma variável que exerce grande influência sob a soma de quadrados dos resíduos. O ideal é encontrar critérios que balanceiam as duas ideias seguintes, de tendência oposta: (i) por um lado, para melhorar o processo de previsão, maior número possível de variáveis é desejado, pela melhoria de ajustamento decorrente; (ii) por outro lado, devido aos eventuais custos de obtenção dos *x*i’s e às dificuldades de interpretação e de manuseio de um modelo grande, o número de variáveis deve ser tão pequeno quanto possível (DRAPER & SMITH, 1998; CHARNET, 2008).

**5.4.5­­ Método “passo atrás” (*backward*)**

Este procedimento caracteriza-se por incorporar, inicialmente, todas as variáveis auxiliares em um modelo de regressão linear múltipla e percorrer etapas, nas quais uma variável por vez pode vir a ser eliminada. Se em uma dada etapa não houver eliminação de alguma variável, o processo é então interrompido e as variáveis restantes definem o modelo final (CHARNET, 2008).

Como critério de eliminação das variáveis, analisam-se as contribuições individuais das variáveis no modelo. A variável de pior desempenho é eliminada, a não ser que esta atenda a um critério mínimo. No julgamento de uma variável, comparamos o modelo completo com o modelo reduzido, pela retirada de tal variável. Observa-se a soma de quadrados de regressão extra assim como a estatística do teste de sua contribuição, dada por:

** (2.13)**

Sendo  e  calculadas sob o modelo completo e  calculada sob o modelo reduzido. A estatística (2.13) testa a contribuição da variável após a inclusão das demais. A contribuição é significante se o valor da estatística for maior que um quantil especificado da distribuição *F* com 1 e (*n - p*) graus de liberdade, sendo *p* o número de parâmetros do modelo completo. Assim, se o valor da estatística for menor que esse quantil da distribuição *F*, a contribuição não é considerada significante e o modelo reduzido deverá ser preferido. Se observarmos várias variáveis não significantes, apenas uma variável é eliminada em uma etapa (aquela cuja estatística do teste tiver o menor valor). Quando uma variável é eliminada, passamos para a nova etapa cujo modelo completo não contém a variável que foi descartada. Se todas as variáveis são significantes, o processo é concluído, e o modelo completo desta etapa é o modelo final (CHARNET, 2008).

5.1 Análise dos Resultados

Tabela 2 – Estimativa das variáveis em relação aos modelos de regressão proposto e compostos pelas variáveis finais.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variável dependente** | **Período** | **Variáveis Explicativas** | **Modelo estatístico** | **R2** | **R2 ajustado** | **RSE** |
| ID7 - Proporção Internações | 2008 - 2014 | ID1, ID10, ID11 e ID18 | Modelo 1 | 0,406 | 0,404 | 0,003 |
| ID8 - Proporção de dias Permanência | 2008 - 2014 | ID1, ID12 e ID17 | Modelo 2 | 0,880 | 0,880 | 1,379 |
| ID6 - Número de Óbitos | 2008 - 2014 | ID1, ID05, ID13, ID14, ID15 e ID16 | Modelo 3 | 0,728 | 0,726 | 1,360 |
| **Total: 3** | | | | | | |

Fonte: Resultados da pesquisa.

\*Legenda: ID1 (Ano Referência); ID5 (Dias de Permanecia); ID6 (Número de Óbitos); ID7 (Proporção Internações <= Número de Internações / População Total); ID8 (Proporção de dias Permanência <= Dias de Permanência / Número de Internações); ID10 (IDHM – Índice de Desenvolvimento Humano); ID11 (PIB per capita); ID12 (População total do município do ano de referência); ID13 (População total atendida com esgotamento sanitário); ID14 (Quantidade de ligações ativas de esgotos); ID15 (Volume de esgotos coletados); ID16 (Volume de esgotos tratados); ID17 (Índice de coleta de esgoto); ID18 (Índice de tratamento de esgoto);

Com relação aos modelos, apresentado na Tabela 2, obteve-se o valor de de cerca 0,406, 0,880 e 0,728 para Modelo 1, Modelo 2 e Modelo 3, respectivamente. No entanto, é recomendado avaliar o R2 ajustado que penaliza a inclusão de regressores poucos explicativos.

Entre os modelos é possível notar que as variáveis explicativas são diferentes, como apesentado anteriormente na Figura 1. O estudo comtempla o período de 2008 a 2014.

Tabela 2 – Significância dos modelos propostos - ANOVA

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| **Modelos** | | **Soma dos quadrados** | **df** | **Média dos quadrados** | **F** | **Sig.** |
| **Modelo 1** | Regressão | ,008 | 4 | ,002 | 221,712 | ,000e |
| Resíduos | ,012 | 1296 | ,000 |
| Total | ,020d | 1300 |  |  |  |
| **Modelo 2** | Regressão | 18108,360 | 3 | 6032,739 | 3170,182 | ,000e |
| Resíduos | 2457,998 | 1297 | 1,903 |
| Total | 20566,359d | 1300 |  |  |  |
| **Modelo 3** | Regressão | 6386,052 | 6 | 1064,342 | 575,789 | ,000e |
| Resíduos | 2391,948 | 1294 | 1,848 |
| Total | 8778,000d | 1300 |  |  |  |

Na Tabela 2 (ANOVA) são apresentados os resultados referentes a significância dos modelos propostos. Por meio dos resultados é possível afirmar que os três modelos propostos possuem significância estatística, ou seja, para todos os modelos pelo menos uma das variáveis explicativas que o compõem pode explicar o comportamento da variável dependente (Y). Com relação ao nível de significância (sig.) nota-se que o p-valor é inferior a 0,05.

Tabela 3 – Significância dos parâmetros da regressão múltipla.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coeficientesa,b** | | | | | | | | | |
| **Modelo** | | **Coeficientes não padronizado** | | **Coeficentes padronizado** | | **t** | | **Sig.** | |
| **B** | **Erro padronizado** | | **Beta** | |  | |  | |
| **Modelo 1** | **ID1** | 5,488E-06 | ,000 | | 2,812 | | 12,259 | | ,000 | |
| **ID10** | -,012 | ,001 | | -2,127 | | -8,821 | | ,000 | |
| **ID11** | -1,899E-08 | ,000 | | -,110 | | -2,952 | | ,003 | |
| **ID18** | -3,998E-06 | ,000 | | -,052 | | -1,975 | | ,049 | |
| **Modelo 2** | **ID17** | -,005 | ,002 | | -,086 | | -2,904 | | ,004 | |
| **ID12** | 1,784E-06 | ,000 | | ,094 | | 9,146 | | ,000 | |
| **ID1** | ,002 | ,000 | | ,982 | | 33,034 | | ,000 | |
| **Modelo 3** | **ID5** | ,002 | ,000 | | ,772 | | 21,279 | | ,000 | |
| **ID13** | -1,236E-05 | ,000 | | -,977 | | -5,623 | | ,000 | |
| **ID14** | 4,292E-05 | ,000 | | ,796 | | 4,966 | | ,000 | |
| **ID15** | 9,971E-05 | ,000 | | ,381 | | 3,889 | | ,000 | |
| **ID16** | -5,400E-05 | ,000 | | -,163 | | -3,364 | | ,001 | |
| **ID1** | ,000 | ,000 | | ,112 | | 5,738 | | ,000 | |

Com relação na Tabela 3 é possível notar que todas as variáveis explicativas apresentam sig. t < 0,05, ou seja, os modelos (Modelo 1, Modelo 2 e Modelo 3) apresentam apenas variáveis com parâmetros de significantes a 5%. Nota-se também a inexistência dos valores de B0 em cada um dos modelos, causado pelo melhor ajuste do na ausência do mesmo. Desse modo, a partir desses modelos são possíveis elaborar previsões de (Yi) em função de possíveis valores das variáveis explicativas que compõem cada um dos modelos.

A seguir serão apresentadas as composições das equações para estimação das variáveis dependentes de cada modelo proposto.

Equações:

### 3.5.1 - Modelo 1

**ID7**(Prop\_Internados) *=* (5,488E-06 \* **ID1**) **+**(-0,012 \* **ID10**) **+** (- 1,899E-08 \* **ID11**) + (- 3,998E-06 \* **ID18**)*+ є*

Analisando o Modelo 1, temos a composição da nossa variável dependente, proporção de internados, de acordo com suas variáveis explicativas. Notamos que 1 unidade da variável ID1 resulta no acréscimo de 0,00005488 na proporção de pessoas internada ao manter as demais variáveis explicativas constantes. Uma possível justificativa é que com o passar dos anos a população passou a ter mais acesso a hospitais, aumentando o número de pessoas com condições de serem internadas. Já as demais variáveis explicativas causaram decréscimo na proporção de pessoas internadas, ou seja, o aumento de 1 unidade de IDH (ID10) diminui em 0,012 a proporção de pessoas internadas. A mesma lógica é aplicada para as demais variáveis.

### 3.5.2 - Modelo 2

**ID8** (Dias\_Permanência) *=* (- 0,005 \* **ID17**)+ (1,784E-06 \* **ID12**)+(0,002 \* **ID1**)*+ є*

Seguindo a mesma lógica descrita no Modelo 1, percebemos que Modelo 2 tanto o ID1 quanto o ID12, que representam respectivamente o ano de análise e a população total do município, resultam no acréscimo nos dias de permanência dos pacientes. Já a variável ID17, que representa o índice de coleta de esgoto, causa um decréscimo nos dias de permanência internados, ou seja, o investimento na coleta de esgoto impacta diretamente na economia em gastos hospitalares resultantes das internações.

### 3.5.3 - Modelo 3

**ID6** (Proporção de Óbitos) *=* (0,002 \* **ID5**) + (- 1,236E-05 \* **ID13**) + (4,292E-05 \* **ID14**) + (9,971E-05 \* **ID15**) + (- 5,400E-05 \* **ID16**) + (0,000 \* **ID1**) *+ є*

O Modelo 3, que busca entender os impactos no número de óbitos, vemos que quanto maior o tempo de internação (ID5) maior a nossa dependente. Esta mesma interpretação é feita com a ID14 e ID15 que representam, respectivamente, quantidade de ligações ativas de esgoto e volume de esgoto coletado. Já a ID13 e a ID16 diminuem o número de óbito, resultado este que ilustra o impacto dos investimentos em tratamento de esgoto na saúde da população.

6 Referências

BERK, K. N. Tolerance and condition in regression computations. Journal of the American Statistical Association, 72 (360), 863-866, 1977.

CHARNET, R. Freire, C. D. L., Charnet, E. M., & Bonvino, H. Análise de modelos de regressão linear: com aplicações. 2. ed. Campinas, São Paulo: Editora da Unicamp, 2008.

CORRAR, S. L.; PAULO, E.; DIAS FILHO, J.M. (Coord.). Análise multivariada para cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007, 539p.

DRAPER, N.R.; SMITH, H. Applied regression analysis.3. ed. New York: John Wiley e Sons, 1998. 706p.

FÁVERO, Luiz Paulo et al. **Análise de Dados:**Modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

HAIR, J. F.; ANDERSON, R.E.; TATHAM, R. L.; BLACK, W. C.; BABIN, B. J. Análise multivariada de dados. 5. Ed. Porto Alegre: Bookman, 2005. KUTNER, M. H. et al. Applied linear models. 5thed. New York: McGraw Hill Irwin, 2004.

TAMHANE, A.; DUNLOP, D. **Statistics and data analysis: From Elementary to Intermediate**. Prentice Hall, 2000.