UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
CURSO DE PÓS-GRADUAÇÃO EM ESTATÍSTICA E MODELAGEM QUANTITATIVA

IDENTIFICAÇÃO DE VARIÁVEIS DETERMINANTES NA SELEÇÃO DE CANDIDATOS, PARA OS CURSOS DE ENGENHARIA, NO PROCESSO SELETIVO DA UNIVERSIDADE FEDERAL DE SANTA MARIA, RS

MONOGRAFIA DE ESPECIALIZAÇÃO

Fabiane Tubino Garcia

Santa Maria, RS, Brasil
2010

IDENTIFICAÇÃO DE VARIÁVEIS DETERMINANTES NA SELEÇÃO DE CANDIDATOS, PARA OS CURSOS DE ENGENHARIA, NO PROCESSO SELETIVO DA UNIVERSIDADE FEDERAL DE SANTA MARIA, RS

por

Fabiane Tubino Garcia

Monografia apresentada ao Curso de Especialização em Estatística e Modelagem Quantitativa, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Especialista em Estatística e Modelagem Quantitativa**

Orientador: Prof. Dr. Luis Felipe Dias Lopes
Co-Orientador: Prof. Dr. Castelar Braz Garcia

Santa Maria, RS, Brasil

2010

UNIVERSIDADE FEDERAL DE SANTA MARIA

CENTRO DE CIÊNCIAS NATURAIS E EXATAS

CURSO DE PÓS-GRADUAÇÃO EM ESTATÍSTICA E MODELAGEM QUANTITATIVA

A Comissão Examinadora, abaixo assinada,
aprova a Monografia de Especialização

IDENTIFICAÇÃO DE VARIÁVEIS DETERMINANTES NA SELEÇÃO DE CANDIDATOS, PARA OS CURSOS DE ENGENHARIA, NO PROCESSO SELETIVO DA UNIVERSIDADE FEDERAL DE SANTA MARIA, RS

elaborada por
Fabiane Tubino Garcia

como requisito parcial para obtenção do grau de
Especialista em Estatística e Modelagem Quantitativa

COMISSÃO EXAMINADORA:

Luis Felipe Dias Lopes, Dr.
(Presidente/Orientador)

Ivanor Müller, Dr. (UFSM)

Gilnei Luiz de Moura, Dr. (UFSM)

Santa Maria, 17 de setembro de 2010.

AGRADECIMENTOS

A Universidade Federal de Santa Maria pela oportunidade de realizar o curso de pós-graduação.

Ao amigo e Prof. Dr. Luis Felipe Dias Lopes, pela orientação, conhecimento, apoio, incentivo e atenção dispensada a esse trabalho.

Ao meu co-orientador, amigo, pai, Prof. Dr. Castelar Braz Garcia pela ajuda, ensinamentos, carinho, paciência e por passar longas horas me auxiliando para o crescimento desta pesquisa.

Aos professores do Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, em especial as professoras Dra. Anaelena Bragança de Moraes, Dra. Luciane Flores Jacobi e ao Prof. Dr. Adriano Mendonça Souza pelos conhecimentos transmitidos, profissionalismo e dedicação.

A coordenadora do Programa de Pós-Graduação em Estatística e Modelagem Quantitativa, Profa. Dra. Roselaine Ruviaro Zanini pela sua atenção, carinho e contribuições no desenvolvimento deste trabalho.

Aos professores Dr. Ivanor Müller e Dr. Gilnei Luiz de Moura, membros da banca examinadora, pelas suas sugestões e colaboração para o aprimoramento deste estudo.

Aos meus colegas do curso pelas horas de estudo que passamos juntos, pelos momentos de descontração, troca de conhecimentos e material.

A minha família, em especial meus pais, Sr. Castelar Braz Garcia, Sra. Sônia Tubino Garcia e meu irmão Leandro Castelar Tubino Garcia por serem o meu porto seguro para todas as horas. Obrigada pelo apoio, compreensão, incentivo, carinho, amor, proteção, pela força para cumprir meus objetivos e principalmente por confiarem em mim.

A todos que de alguma maneira colaboraram para a realização deste trabalho.

RESUMO

Monografia de Especialização
Curso de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

IDENTIFICAÇÃO DE VARIÁVEIS DETERMINANTES NA SELEÇÃO DE CANDIDATOS, PARA OS CURSOS DE ENGENHARIA, NO PROCESSO SELETIVO DA UNIVERSIDADE FEDERAL DE SANTA MARIA, RS

Autor: Fabiane Tubino Garcia
Orientador: Dr. Luis Felipe Dias Lopes
Co-orientador: Dr. Castelar Braz Garcia
Data e Local de Defesa: Santa Maria, 17 de setembro de 2010.

O panorama atual é de valorização pela formação e da educação continuada para o cidadão. Esta valorização deve-se aos avanços tecnológicos em diversas áreas do conhecimento, a globalização, e a necessidade de maior qualificação da mão-de-obra, os quais obrigam a todas as pessoas estarem bem preparadas para atuar e garantir seu espaço em um mercado competitivo. Percebe-se que a demanda por acesso ao ensino superior segue crescendo nas universidades, as quais têm buscado realizar uma seleção mais justa, bem como, de evitar a evasão dos estudantes ingressos. Devido a isto, as IES buscando conhecer o futuro acadêmico têm solicitado a seus candidatos o preenchimento de um questionário sociocultural durante o processo de inscrição ao vestibular. Assim, o objetivo desta investigação foi identificar e testar as variáveis socioculturais capazes de determinar a seleção dos candidatos aos novos cursos de Engenharia, viabilizados pelo programa REUNI, do processo seletivo da UFSM realizado em maio de 2009. Para análise dos dados foi realizado um estudo descritivo e exploratório, utilizando a técnica estatística multivariada de regressão logística múltipla. Obteve-se, uma amostra de 535 candidatos e foram testadas 34 variáveis socioculturais do respectivo banco de dados da COPERVES. Neste estudo a variável dependente corresponde à ocorrência ou não da seleção ao vestibular. Os resultados obtidos no modelo ajustado indicam que somente duas variáveis preditoras foram estatisticamente significativas ($p<0,05$) para estimar a probabilidade de seleção dos candidatos, que são: se já prestou vestibular uma ou mais vezes ($b_1=1,170$) e se possui a renda total mensal familiar até 5 salários mínimos ($b_2=.-1,143$).

Palavras-chave: Processo Seletivo; Engenharias; Variáveis Socioculturais; Seleção; Modelo de Regressão Logística Múltipla.

ABSTRACT

Monografia de Especialização
Curso de Pós-Graduação em Estatística e Modelagem Quantitativa
Universidade Federal de Santa Maria, RS, Brasil

IDENTIFICATION OF VARIABLES IN DETERMINING SELECTION OF CANDIDATES FOR ENGINEERING COURSES IN THE SELECTION PROCESS AT THE UNIVERSITY OF SANTA MARIA, RS

Author: Fabiane Tubino Garcia
Adviser: Dr. Luis Felipe Dias Lopes
Co adviser: Dr. Castelar Braz Garcia
Date and Place of Defense: Santa Maria, September 17, 2010.

The current outlook and upward movement by the formation and continuing education for the citizen. This upward movement is due to technological advances in diverse areas of knowledge, globalization, and the need for greater supply of skilled manpower, which require all persons to be well prepared to work and ensure its place in a competitive market. It is noticed that the demand for access to higher education continues to grow in the universities, which has sought to make a selection more fair and, to prevent evasion of student. Because of this, IES seeking for the future students has asked its candidates to fill out a questionnaire during the sociocultural process of inscription to the entrance exam. Thus, the goal of this investigation was to identify and test the sociocultural variables that determine the selection of candidates for new engineering courses, made possible by the program REUNI of the selection process UFSM held in May 2009. For data analysis was performed an exploratory and descriptive study, using technique multivariate statistics of multiple logistic regression. We obtained a sample of 535 candidates were tested and 34 of their sociocultural variables database COPERVES. In this study the dependent variable corresponds to the occurrence or not of the vestibular selection. The adjusted model results indicate that only two predictor variables were statistically significant ($p < 0.05$) to determine the probability of selection of candidates, which are: the university entry exams is already one or more times ($b_1 = 1.170$) and has total monthly household income up to 5 times the minimum wage ($b_2 =.- 1.143$).

Keywords: Selection Process; Engineering; Sociocultural Variables; Selecion; Mutliple Logistic Regression Model.

LISTA DE TABELAS

TABELA 01 – Variável dependente utilizada no modelo de regressão logística.....28
TABELA 02 – Informações pessoais dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535)..34
TABELA 03 – Informações sobre a formação educacional dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).....34
TABELA 04 – Informações sobre a vida econômica familiar dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).....36
TABELA 05 - Informações sobre hábitos e costumes dos candidatos inscritos e que realizaram o vestibular extraordinário aos cursos de Engenharia, UFSM, 2009 (n=535).....36
TABELA 06 – Resultados da Análise de Regressão Logística Univariada.....37
TABELA 07 – Resultados do Coeficiente de Correlação de Spearman para a multicolineariedade.....38
TABELA 08 – Modelos estimados de regressão logística.....40
TABELA 09 – Tabela de Classificação – Previsão do Modelo.....42
TABELA 10 – Tabela de Classificação – Validação do Modelo.....43

LISTA DE ILUSTRAÇÕES

FIGURA 1 – Forma da relação logística entre variáveis dependente e independente.....14

LISTA DE QUADROS

QUADRO 1 – Efeitos das variáveis independentes na seleção do vestibular.....30

LISTA DE ABREVIATURAS E SIGLAS

AAS – Amostragem Aleatória Simples

COPERVES – Comissão Permanente do Vestibular

EAD – Educação á Distância

ENEM – Exame Nacional do Ensino Médio

G.L. - graus de liberdade

I.C. – Intervalo de Confiança

IES – Instituição de Ensino Superior

IFES – Instituição Federal de Ensino Superior

KDD - *Knowledge Discovery in Databases*

OR - *odds ratio* ou razão de chance

PEIES - Programa de Ingresso ao Ensino Superior

PROUNI – Programa Universidade para todos

Q.I. – quociente de inteligência

REUNI - Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

RS – Rio Grande do Sul

UFSM – Universidade Federal de Santa Maria

V.L. – *Likelihood Value* (valor da verossimilhança)

LISTA DE ANEXOS

ANEXO A - Questionário Sociocultural do Vestibular Extraordinário da UFSM, abril de 2009.....52

SUMÁRIO

1 INTRODUÇÃO.....01

1.1 Contextualização.....01
1.2 Tema da pesquisa.....04
1.3 Justificativa e importância da pesquisa.....04
1.4 Problema de Pesquisa.....05
1.5 Objetivos da pesquisa.....06
1.5.1 Objetivo Geral.....06
1.5.2 Objetivos Específicos.....06
1.6 Delimitação da pesquisa.....06
1.7 Estrutura do trabalho.....07

2 REVISÃO DE LITERATURA.....08

2.1 O Ensino Superior e o Vestibular Tradicional.....08
2.2 A análise de Regressão.....09
2.3. A técnica de Regressão Logística Múltipla Binária.....11
2.3.1 Método da Máxima Verossimilhança.....18
2.3.2 O Teste Wald.....20
2.3.3 O Teste Hosmer e Lemeshow.....21
2.3.4 Pseudo R^2.....22

3 METODOLOGIA.....24

3.1 Delineamento.....24
3.2 Definição do Universo e a Estatística Descritiva.....25
3.3 Dimensionamento do tamanho da amostra.....26
3.4 Instrumento de obtenção dos dados.....27
3.5 Variável Dependente.....28
3.6 Variáveis Independentes ou Covariáveis.....28
3.7 Desenvolvimento do modelo.....31

4 RESULTADOS E DISCUSSÃO.....34

4.1 Análise Descritiva.....34
4.2 Análise de Regressão Logística Univariada.....37
4.3 Análise da Multicolinearidade.....38
4.4 Avaliação do ajuste do modelo.....39
4.5 Modelo de estimação da função Seleção.....41
4.6 Capacidade de Previsão do Modelo.....41
4.7 Validação do Modelo.....43
4.8 Estimação da probabilidade de ocorrência da seleção de um candidato....43

5 CONSIDERAÇÕES FINAIS.....46

5.1 Conclusões.....46

6 REFERÊNCIAS 49

7 ANEXOS.....52

1 INTRODUÇÃO

1.1 Contextualização

Nos últimos anos muito tem se falado e percebido sobre a importância da qualificação de pessoas detentoras de um curso superior, onde estas quando o possuem acabam se tornando um referencial para as demais.

No mercado de trabalho, as empresas buscam por profissionais que estejam cursando ou com a formação superior completa para vagas cuja atividade, às vezes, sequer exigem tamanha qualificação.

Com isso, verifica-se, que a formação superior melhora tanto a condição pessoal, social como a profissional dos indivíduos, ou seja, permite ao ser humano ter uma visão mais contextualizada do mundo globalizado e uma compreensão mais real e sólida de todos os acontecimentos que o cercam (MARTINHAGO, 2005). Isso não quer dizer que pessoas com grande experiência de vida e sem formação superior não possam atingir o mesmo grau de desenvolvimento, mas pessoas com ensino superior compreendem melhor os fenômenos que os cercam.

Segundo Silva e Periçaro (2009) a educação tem sido um dos principais fatores que sustentam o desenvolvimento tecnológico e profissional, atuando de forma direta na qualificação pessoal e social.

Salienta-se que cada estudante é dotado de personalidades divergentes entre si, mas que podem indicar informações valiosas se analisadas em conjunto. Com isso, estas informações podem ser utilizadas como banco de dados para análises futuras, revelando resultados significativos para a tomada de decisão em diversas situações.

A descoberta de conhecimento por meio de um banco de dados, ou prospecção de conhecimento (*Knowledge Discovery in Databases – KDD*), conforme Carvalho (1999) apud Martinhago (2005) é um processo multidisciplinar, que combina técnicas, algoritmos e definições de todas as áreas com a finalidade principal de extrair conhecimento a partir de grandes bases de dados. Esse desiderato desenvolve e valida técnicas, ferramentas e métodos que buscam extrair padrões até então implícitos no banco de dados (SILVA e PERIÇARO, 2009).

Sendo os métodos estatísticos aplicáveis em qualquer área do conhecimento pode-se adequar a educação para este estudo, utilizando dados numéricos para identificar e validar conceitos associados à formulação de hipóteses, por meio de um processo de simplificação, predição e direcionamento.

Dado exposto percebe-se, a existência da relação entre o armazenamento de dados e a sua contribuição na identificação de determinantes relacionados à educação (SILVA e PERIÇARO, 2009).

Para Panizzi (2004) apud Martinhago (2005) os órgãos governamentais não devem apenas se preocupar com o ingresso dos jovens no ensino superior, mas principalmente com a permanência destes nas instituições. Devido a isto, observa-se a importância de conhecer o perfil dos candidatos ao vestibular, de forma a auxiliar na elaboração de projetos que atendam às necessidades dos acadêmicos, e consequentemente forneçam subsídios à permanência destes na IFES (SILVA e PERIÇARO, 2009).

Sabe-se que algumas Instituições Federais de Ensino Superior solicitam a seus candidatos o preenchimento de um questionário sociocultural durante o processo de inscrição ao vestibular. Este procedimento é comum na maioria das IFES as quais visam estruturar um banco de dados que forneça informações precisas e relevantes dos estudantes.

A Universidade Federal de Santa Maria, UFSM, é uma instituição federal de ensino superior que há 50 anos esta voltada ao compromisso com o ensino, pesquisa e extensão, bem como, com a ampliação de experiências visando à formação de cidadãos para atuar no mercado de trabalho nas mais diversas áreas do conhecimento (COPERVES, 2009).

A UFSM está localizada no centro geográfico do Estado do Rio Grande do Sul, no município de Santa Maria, fundada em 18 de março de 1961 e idealizada pelo Prof. Dr. José Mariano da Rocha.

A atual estrutura, determinada pelo estatuto da universidade, estabelece a constituição de oito unidades universitárias: Centro de Ciências Naturais e Exatas, Centro de Ciências Rurais, Centro de Ciências da Saúde, Centro de Educação, Centro de Ciências Sociais e Humanas, Centro de Tecnologia, Centro de Artes e Letras e Centro de Educação Física e Desportos.

A universidade possui, hoje, em pleno desenvolvimento, cursos, programas e projetos nas mais diversas áreas do conhecimento humano. A Instituição mantém 66

cursos de Graduação Presenciais (oferecidos no Vestibular 2009 - 1º semestre/2009), e 28 cursos oferecidos no Vestibular Extraordinário 2009 - 2° semestre/2009; dez cursos de Educação a Distância, 72 de Pós-Graduação Permanente, isto é, 17 de Doutorado, 41 de Mestrado e 14 de Especialização. Além disso, possui um curso de Pós-Doutorado e cinco cursos de Especialização/EAD (1º semestre de 2009).

O contingente educacional desta instituição é de 18.489 alunos (1º semestre de 2009) em cursos permanentes, distribuídos entre os três níveis de ensino, dos quais 13.322 são do ensino de Graduação, 2.261 do ensino de Pós-Graduação e 2.906 do ensino Médio e Tecnológico, destes 200 alunos em estágio. Ainda, a UFSM possui 1.021 alunos matriculados no 2º semestre 2009, Vestibular Extraordinário. O corpo docente é constituído de 1.242 professores do quadro efetivo (Graduação, Pós-Graduação e Ensino Médio e Tecnológico) e 202 professores de contrato temporário; e o quadro de pessoal técnico administrativo totalizada em 2.642 servidores (dezembro de 2008).

Na UFSM, por ano, são realizados dois vestibulares, de inverno que ocorre no mês de julho e de verão no mês de janeiro. Portanto, em abril de 2009 a Instituição realizou o vestibular extraordinário com a finalidade de oferecer novos cursos de graduação viabilizados pelo programa REUNI.

O Vestibular Extraordinário baseado no edital da COPERVES foi constituído de uma única etapa seletivo-classificatória, em dois dias consecutivos, através da aplicação de provas de múltipla escolha e redação, visando ao ingresso de estudantes aos novos cursos de graduação desta Universidade.

A plena execução do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI viabilizou 14 novos cursos de graduação instalados na sede em Santa Maria, disponibilizando 566 vagas para estudantes ingressos em agosto de 2009.

O programa REUNI foi instituído pelo Decreto n° 6.096, de 24 de abril de 2007, é uma das ações que a UFSM tem buscado para ampliar o acesso e a permanência na educação superior. Novos cursos estão sendo criados; novas unidades universitárias construídas, com isto, alternativas de ingresso se expandem e tornam-se mais igualitárias, criando-se um ambiente favorável ao fomento da atividade fim da universidade.

Com o REUNI, o Governo Federal adotou uma série de medidas para retomar o crescimento do ensino superior público, criando condições para que as universidades federais promovam a expansão física, acadêmica e pedagógica da rede federal de educação superior.

As ações do programa contemplam o aumento na oferta de vagas nos cursos de graduação, a ampliação de cursos noturnos, a promoção de inovações pedagógicas e o combate à evasão, entre outras metas que têm o propósito de diminuir as desigualdades sociais no país (REUNI, 2009).

Cabe salientar que a criação de cursos no turno da noite viabilizados pelo programa, cumpre um importante compromisso social na medida em que atende a reivindicação de acesso a universidade pública aos alunos trabalhadores.

O presente estudo foi desenvolvido na Universidade Federal de Santa Maria, utilizando as informações contidas no questionário sociocultural preenchido pelos candidatos inscritos e que concorreram às vagas nos novos cursos de Engenharia do vestibular extraordinário, viabilizado pelo programa REUNI, realizado em maio de 2009.

Busca-se com este trabalho identificar quais são as variáveis determinantes que influenciam na seleção dos candidatos no processo seletivo extraordinário da UFSM, utilizando a técnica estatística multivariada de regressão logística múltipla. Ou seja, procura-se aferir as variáveis de natureza qualitativa e quantitativa que determinam a classificação dos candidatos ao vestibular.

1.2 Tema da pesquisa

O tema a ser estudado nesta pesquisa refere-se em identificar as variáveis socioculturais que influenciam na seleção dos candidatos no processo seletivo da UFSM. Para conhecer estas variáveis será utilizada a análise multivariada, tendo como foco principal a técnica de regressão logística múltipla.

1.3 Justificativa e importância da pesquisa

A importância deste estudo é identificar quais são as variáveis socioculturais que influenciam na classificação dos candidatos os cursos de Engenharia no

vestibular da UFSM. Para isto, será utilizada a técnica multivariada de regressão logística múltipla, que proporcionará obtenção um modelo de seleção.

Cabe salientar que por meio do questionário sociocultural preenchido pelos candidatos no ato da inscrição também será possível identificar quem são estes estudantes, buscando assim conhecer o perfil dos vestibulandos.

Optou-se por estudar os candidatos inscritos e que concorreram as vagas ofertadas aos novos cursos de Engenharia no processo seletivo extraordinário, pelo fato destes cursos apresentarem o maior número de candidatos, com uma demanda média de 12 candidatos/vaga.

É importante salientar que o processo seletivo extraordinário foi viabilizado pelas ações do programa REUNI do qual a UFSM é integrante.

Por estas razões uma investigação desta natureza proporcionará através de um banco de dados e técnicas estatísticas obter informações sobre os ingressos aos cursos de Engenharia do processo seletivo extraordinário realizado pela Instituição.

1.4 Problema de Pesquisa

O panorama atual é de valorização pela formação e da educação continuada para o cidadão. Esta valorização é percebida pela sociedade e no mercado de trabalho e deve-se aos avanços tecnológicos em diversas áreas do conhecimento.

Com isto, a demanda pelos cursos superiores nas instituições de ensino segue crescendo, fazendo com que as IES passem por diversas mudanças na tentativa de realizar uma seleção mais justa.

. Dentre as diversas formas de avaliação realizadas destacam-se o vestibular tradicional; o ENEM; análise de histórico escolar e currículo; avaliação seriada; habilidade específica; aptidão física; sistema de cotas, PEIES e o PROUNI (sendo este último nas universidades privadas).

È importante ressaltar que as IES buscando conhecer o futuro acadêmico têm solicitado a seus candidatos ao vestibular o preenchimento de um questionário socioeconômico e cultural durante o processo de inscrição. E, com base neste cenário, e utilizando as informações socioculturais dos candidatos, se estabelece o seguinte problema de pesquisa: **é possível identificar quais as variáveis socioculturais que influenciam na seleção dos candidatos, aos cursos de**

Engenharia, do processo seletivo extraordinário da Universidade Federal de Santa Maria?

1.5 Objetivos da pesquisa

1.5.1 Objetivo Geral

Identificar e testar as variáveis socioculturais capazes de determinar a seleção dos candidatos aos novos cursos de Engenharia, viabilizados pelo programa REUNI, do processo seletivo extraordinário da UFSM, no ano de 2009, utilizando um modelo de Regressão Logística Múltipla.

1.5.2 Objetivos Específicos

a) Identificar o perfil dos candidatos dos novos cursos de graduação das Engenharia, utilizando as variáveis socioculturais do respectivo banco de dados da COPERVES;

b) Construir um modelo de Regressão Logística Múltipla Binária, que permita predizer se os candidatos inscritos aos cursos de Engenharia serão selecionados ou não no processo seletivo da UFSM;

c) Ajustar um modelo de Regressão Logística com as variáveis socioculturais estatisticamente significativas a seleção dos candidatos;

d) Estimar a precisão do modelo;

e) Validar o modelo de Regressão Logística Múltipla;

f) Estimar a probabilidade de ocorrência da seleção dos candidatos, aos novos cursos de Engenharia viabilizados pelo programa REUNI, no vestibular extraordinário da UFSM.

1.6 Delimitação da pesquisa

É importante salientar que o modelo destina-se especificamente aos novos cursos de graduação nas Engenharia que foram oferecidos no vestibular extraordinário, visto que os agentes envolvidos na análise serão os candidatos inscritos e que realizam as provas para estes cursos.

Também, ressalta-se que dentre os novos cursos de Engenharia, viabilizados pelo programa REUNI, objeto de estudo, estão: Engenharia de Controle e Automação, de Produção, Sanitária e Ambiental, de Computação, Acústica, Química e Ambiental.

Observa-se, que a base deste estudo envolve a Universidade Federal de Santa Maria com sede no município de Santa Maria, portanto, a aplicação deste modelo em outra região, ou outra universidade poderá proporcionar outros resultados, devido a características econômicas, sociais e culturais distintas.

1.7 Estrutura do trabalho

Esta pesquisa será dividida em 5 capítulos. No primeiro capítulo é apresentada a introdução, o tema, a justificativa e importância do estudo, o problema de pesquisa e os objetivos da investigação.

O segundo capítulo trata da revisão de literatura onde é abordada o tema sobre o ensino superior e o vestibular tradicional, bem como a técnica estatística de análise multivariada utilizada no estudo.

No terceiro capítulo é apresentada a metodologia aplicada na construção do modelo probabilístico de seleção ao vestibular, iniciando pelo delineamento, definição do universo a ser investigado, dimensionamento da amostra, identificação da variável dependente e independente, instrumento de obtenção dos dados e desenvolvimento do modelo.

No quarto capítulo, apresentam-se os resultados obtidos com a aplicação da técnica proposta no estudo, bem como a análise descritiva dos dados, a avaliação do ajuste do modelo, a capacidade de previsão, validação e a estimação da probabilidade de ocorrência de seleção.

O quinto capítulo trata das conclusões obtidas na pesquisa, sugestões para trabalhos futuros e algumas considerações finais.

2 REVISÃO DE LITERATURA

Neste capítulo, aborda-se a revisão bibliográfica que deu embasamento ao desenvolvimento desta pesquisa. Os temas abrangem o ensino superior, o vestibular tradicional bem como a técnica estatística de análise multivariada que será utilizada no estudo.

2.1 O Ensino Superior e o Vestibular Tradicional

O panorama atual é de valorização pela formação e da educação continuada para o cidadão. Esta valorização é percebida pela sociedade e no mercado de trabalho e deve-se aos avanços tecnológicos em diversas áreas do conhecimento, a globalização, e a necessidade de maior qualificação da mão-de-obra, os quais obrigam a todas as pessoas estarem bem preparadas para atuar e para mudanças permitindo garantir seu espaço em um mercado competitivo.

O ingresso aos cursos superiores nas instituições de ensino é realizado mediante a seleção de candidatos em um exame classificatório que se tornou obrigatório em 1911 devido ao aumento da procura de estudantes pelas universidades brasileiras, os quais ultrapassavam o número de vagas disponíveis (ALVES, 2008). Com isso, o então Ministro da Justiça e dos Negócios, Rivadavia da Cunha Correa, propôs a lei que exigia o exame de admissão, a difusão dos critérios das provas, a existência de bancas, do calendário e das taxas de inscrição (MARTINHAGO, 2005).

Em 1915, conforme o Decreto n°11.530, as provas passaram a ser chamadas de “vestibular”. Neste período os testes eram constituídos de uma prova escrita e oral. Este tipo de seleção se manteve até meados dos anos 60, quando surgiram as questões de múltipla escolha. Estes testes eram processados em computadores facilitando a correção, dada sua complexidade e o crescente aumento dos candidatos. Porém, o critério de nota mínima aprovava candidatos acima do limite de vagas que a universidade comportava. Sendo assim, os candidatos excedentes aguardavam a expansão de ofertas.

Para solucionar este problema, em 1968, com a implementação da lei n° 5.540, o Governo passa a instituir o sistema classificatório, com corte por nota máxima.

Em 1996 foi aprovada a Lei das Diretrizes e Bases de Educação, e com isso o ingresso ao ensino superior passa a ser feito via processo seletivo a critério de cada instituição (FRANÇA, 2008).

Segundo Alves (2008), a palavra vestibular vem do latim *vestibulum*, que significa entrada. Antigamente usava-se a expressão “exame vestibular” (exame de entrada), com o passar do tempo passou-se a usar apenas “vestibular ou processo seletivo” para designar esse tipo de prova.

Cabe lembrar que existem diferentes formas de avaliação para ingresso em uma IES, no qual se destacam o vestibular tradicional; o ENEM; análise de histórico escolar e currículo; avaliação seriada; habilidade específica; aptidão física; sistema de cotas, PEIES e o PROUNI (sendo este último nas universidades privadas).

Apesar das novas formas de avaliação, em especial o ENEM, o vestibular tradicional continua a ser a principal forma de acesso as universidades federais.

Percebe-se que a demanda por acesso ao ensino superior segue crescendo e devido a isto as IES têm passado por diversas mudanças na tentativa de realizar uma seleção mais justa bem como, de evitar a evasão dos estudantes ingressos.

As IES buscando conhecer o futuro acadêmico têm solicitado a seus candidatos ao vestibular o preenchimento de um questionário socioeconômico e cultural durante o processo de inscrição.

A importância em conhecer o perfil dos candidatos aos cursos de Engenharia do vestibular extraordinário deve-se ao fato de elaborar projetos que atendam às necessidades dos acadêmicos, adaptados ao ensino superior e, consequentemente forneçam subsídios à permanência destes na IFES.

2.2 A análise de Regressão

Um modelo de regressão pode ser definido como uma equação matemática não-determinística que expressa à relação entre variáveis. Nestes modelos, define-se uma variável explicada ou dependente (Y_i), e busca-se verificar a influência de

uma ou mais variáveis explicativas, independentes ou causais (X_i) sobre esta variável explicada (ZANINI, 2007).

Conforme Anderson, Sweeney e Williams (2003), na terminologia de regressão, a variável que esta sendo calculada é chamada de variável dependente, enquanto que a variável ou variáveis que estão sendo usadas para calcular a variável dependente são chamadas de variáveis independentes. Em notação estatística, Y_i denota a variável dependente e X_i denota a variável independente.

Segundo Barrow (2007), as principais características de uma análise de regressão são:

- a regressão permite investigar as relações entre duas ou mais variáveis;
- aponta-se uma direção de causalidade, da(s) variável(is) explicativa(s) para a variável dependente;
- mede-se a influência de cada variável explicativa sobre a variável dependente;
- pode-se determinar a significância de cada variável explicativa.

O tipo mais simples de análise de regressão, envolvendo uma variável independente e uma dependente na qual a relação entre as variáveis é aproximada por uma linha reta, é chamado de regressão linear simples. A análise de regressão envolvendo duas ou mais variáveis independentes é chamada de análise de regressão múltipla (ANDERSON, SWEENEY e WILLIAMS, 2003).

Em uma regressão linear simples a variável de saída (Y_i) é prevista a partir da equação:

$$Y_i = b_0 + b_1X_i + \xi_i, \quad (2.1)$$

No caso de um modelo de regressão linear múltipla, pode-se ajustar o modelo da forma:

$$Y_i = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + \xi_i, \quad (2.2)$$

onde:

Y_i = representa a variável dependente;

b_0 = intercepto da linha;

b_i = são os coeficientes de regressão;

X_i = são as variáveis independentes;
ξ_i = erro aleatório.

2.3 A técnica de Regressão Logística Múltipla Binária

Segundo Hosmer e Lemeshow (1989) a regressão logística é uma técnica semelhante à regressão linear, sendo utilizada quando a variável dependente é categórica e, em geral dicotômica ou binária.

A utilização da técnica é adequada em muitas situações porque permite que se analise o efeito de uma ou mais variáveis independentes (discretas ou contínuas) sobre uma variável dependente dicotômica, representando a presença (1) ou ausência (0) de uma característica (HOSMER; LEMESHOW, 1989).

Zanini (2007) destaca que a regressão logística é aplicável ou preferida quando se tem uma variável dependente categórica dicotômica, ou seja, uma variável nominal ou não métrica que possui apenas dois grupos ou classificações, como resultados possíveis, como, por exemplo, alto ou baixo, sim ou não etc.

Conforme Corrar, Paulo e Dias Filho (2009), a técnica de regressão logística foi desenvolvida por volta de 1960 em resposta ao desafio de realizar predições ou explicar a ocorrência de determinados fenômenos em que a variável dependente fosse de natureza binária. Um dos estudos pioneiros que mais contribuíram para o seu avanço foi o famoso *Framingham Heart Study*, realizado com a Universidade de Boston. O objetivo principal do estudo foi identificar os fatores que contribuíam para a ocorrência de doenças cardiovasculares.

Esta técnica vem sendo utilizada em diversas áreas do conhecimento, e seu objetivo é identificar quais são as variáveis independentes que influenciam no resultado da variável dependente e utilizá-las em uma equação para estimar a probabilidade das variáveis independentes explicarem o desfecho (HOSMER; LEMESHOW, 1989).

Ainda segundo os autores Hosmer e Lemeshow (1989), a regressão logística tornou-se, portanto, um método padrão de análise de regressão para variáveis medidas de forma dicotômica. Desta forma, a diferença principal da regressão logística comparada ao modelo linear clássico é que a distribuição da variável resposta segue uma distribuição binomial, e não uma distribuição normal.

Segundo Corrar, Paulo e Dias Filho (2009), uma das razões para o uso da regressão logística para realizar predições é o fato da técnica apresentar um número reduzido de suposições, assim, o pesquisador consegue contornar certas restrições encontradas em outros modelos de análise multivariada de dados.

Para utilizar a técnica de regressão logística é necessário observar os seguintes requisitos (PAULO; CORRAR; DIAS FILHO, 2009):

- incluir todas as variáveis preditoras no modelo para que ele obtenha maior estabilidade;
- o valor esperado do erro deve ser zero;
- inexistência de autocorrelação entre os erros;
- inexistência de correlação entre os erros e as variáveis independentes e;
- a ausência de multicolinearidade perfeita entre as variáveis independentes.

Hair, Anderson, Tatham e Black (2005) e Field (2009) destacam algumas características da regressão logística:

- não é necessário supor a normalidade multivariada;
- é uma técnica mais genérica e mais robusta, pois sua aplicação é apropriada numa grande variedade de situações;
- é uma técnica similar a regressão linear múltipla, mas com variável de saída categórica dicotômica;
- em vez de prever o valor da variável “y” a partir de um previsor X_i, se prevê a probabilidade de “y” ocorrer;
- a equação da regressão logística expressa uma regressão linear simples ou múltipla em termos logarítmicos e dessa forma resolve o problema da violação da hipótese de linearidade; e
- os coeficientes dos parâmetros são estimados utilizando a estimação de máxima verossimilhança que seleciona os coeficientes que tornam os valores observados mais prováveis de terem ocorrido.

Para Corrar, Paulo e Dias Filho (2009), a regressão logística busca encontrar uma função logística formada por meio de ponderações das variáveis (atributos), cuja resposta permita estabelecer a probabilidade de ocorrência de determinado evento e a importância das variáveis para esta ocorrência.

A este respeito, Hair, Anderson, Tatham e Black (2005), afirmam que a regressão logística se assemelha em muitos aspectos a regressão linear, mas se difere basicamente sentido de prever a probabilidade de um evento ocorrer.

Segundo Penha (2002) na regressão logística as variáveis independentes podem ser tanto fatores quanto covariáveis, enquanto que as variáveis dependentes podem ser apresentadas em duas ou mais categorias. Entende-se que as covariáveis são representadas por dados contínuos, enquanto os fatores são dados categóricos.

Penha (2002) também destaca que existem três procedimentos distintos para manipular os dados, e são denominados Regressão Logística Binária, Ordinal e Nominal. A escolha do método depende do número de categorias e das características da variável resposta.

A variável binária é aquela que aceita apenas dois níveis de resposta, como adimplente ou inadimplente. Já a variável ordinária segue uma ordenação natural das coisas, como pequeno, médio e grande e a variável nominal pode ter mais de três níveis e não considera nenhuma ordenação, como por exemplo, a previsão do tempo: ensolarado, nublado e chuvoso (PENHA, 2002).

Corrar, Paulo e Dias Filho (2009, p.284) definem que “A regressão logística é uma técnica de análise multivariada que permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias”.

Hair, Anderson, Tatham e Black (2005), também salientam que uma das vantagens em utilizar a regressão logística é saber apenas se um evento aconteceu para que então se possa usar um valor binário como variável dependente. A partir desse valor dicotômico, o procedimento prevê sua estimativa da probabilidade de que o evento ocorrerá ou não

Para Brito e Assaf Neto (2005, p.8) a regressão logística:

> “É uma técnica de análise multivariada, apropriada para as situações nas quais a variável dependente é categórica e assume um entre dois resultados possíveis (binária), tais como: normal ou anormal, cliente ou não cliente e solvente ou insolvente”.

Em um modelo de regressão logística binária, a variável resposta poderá assumir dois valores: 0 (zero) indicando a ausência de um determinado atributo e 1 (um) indicando a presença.

Uma das semelhanças da regressão logística com a regressão múltipla é no que diz respeito ao formato e aos dados nominais e categóricos poderem ser incluídos como variáveis independentes por meio de alguma forma de codificação

binária. No entanto, a técnica difere da regressão múltipla no sentido de prever diretamente a probabilidade de um evento acontecer. Ou seja, os valores de probabilidade podem ser qualquer valor entre zero e um, mas o valor previsto deve ser limitado, de modo a recair no intervalo de zero a um (HAIR; ANDERSON; TATHAM; BLACK, 2005).

Para definir uma relação delimitada por zero e um, a regressão logística usa uma relação assumida entre as variáveis independente e dependente (Figura 1), e, que lembra uma curva em forma de “S” (HAIR; ANDERSON; TATHAM; BLACK, 2005).

Os modelos lineares de regressão não podem acomodar tal relação entre as variáveis, já que ela é inerentemente não-linear. Por isso a regressão logística foi desenvolvida para lidar especificamente com essas questões. Destaca-se que a regressão logística deriva seu nome justamente dessa transformação logística utilizada com a variável dependente (HAIR; ANDERSON; TATHAM; BLACK, 2005).

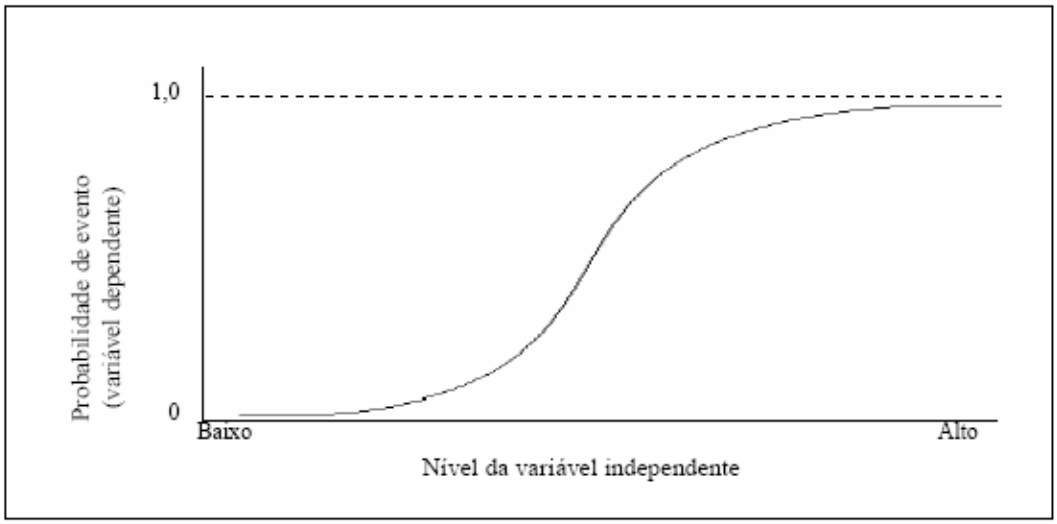

Figura 1 – Forma da relação logística entre variáveis dependente e independente.
Fonte: Hair, Anderson, Tatham e Black, 2005, p.232.

Para Hair, Anderson, Tatham e Black (2005), quando a variável independente aumenta, a probabilidade aumenta para cima da curva, mas em seguida a inclinação começa a diminuir, de forma que, em qualquer nível da variável independente, a probabilidade irá tender a um, mas jamais excederá a esse valor.

Corrar, Paulo e Dias Filho (2009) acrescentam que o uso do modelo linear poderia conduzir a predições de valores menores que zero e maiores que um, e com

isso torna-se necessário converter as observações em razão de chances (*odds ratio*) e submetê-las a uma transformação logarítmica. Com isso, o modelo passa a evidenciar mudanças nas inter-relações dos *logs* da variável dependente.

Verifica-se que a técnica de regressão logística gera um modelo matemático, cuja resposta permite estabelecer a probabilidade de uma observação pertencer a um grupo previamente determinado, em razão do comportamento de um conjunto de variáveis independentes (BRITO; ASSAF NETO, 2005). Para construção deste modelo efetua-se uma transformação logística na variável dependente, sendo esse processo constituído de duas etapas. A primeira consiste em convertê-la numa razão de chance e a segunda, em transformá-la numa variável de base logarítmica (CORRAR; PAULO; DIAS FILHO, 2009).

Primeiramente converte-se a probabilidade associada a cada observação em razão de chance (*odds ratio*), que representa a probabilidade de sucesso (p) comparada com a de fracasso (1 – p):

$$\text{Razão de chance ou } \textit{Odds ratio} = \frac{p}{(1-p)} . \qquad (2.3)$$

Segundo Kahn e Sempos (1989) apud Mezzomo (2009), a razão de chances ou *odds ratio* (OR) é a razão da chance do desfecho (Y) a seleção, em relação à chance do desfecho (Y) a não-seleção. Esta razão permite conhecer qual a chance de um evento ocorrer, em relação a ele não ocorrer sob as mesmas condições (PENHA, 2002).

Logo, obtêm-se o logaritmo natural da razão de chance:

$$\ln \left(\frac{p}{1-p} \right) = b_0 + b_1 X_1 \qquad (2.4)$$

Percebe-se que do lado esquerdo da equação tem-se o logaritmo natural da razão de chance e no lado direito, as variáveis independentes e os coeficientes estimados que expressam mudanças no *log* da razão de chance.

Na equação acima, se X for uma variável categórica binária, poderá assumir o valor 0 ou 1, substituindo-se:

$$\ln (odds)_{x=0} = b_0 + b_1 . 0 = b_0 ; \quad (2.5a)$$

$$\ln (odds)_{x=1} = b_0 + b_1 . 1 = b_0 + b_1 . \quad (2.5b)$$

Destaca-se que o *odds ratio* (OR) corresponde ao aumento de uma variável na variável independente (X) e tem-se (MEZZOMO, 2009):

$$OR = e^{b1} ; \quad (2.6)$$

sendo e= 2,718, que é o número neperiano, sendo a base do logaritmo utilizado no modelo de regressão múltipla.

Em geral, para n variáveis independentes têm-se:

$$\ln odds = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n . \quad (2.7)$$

Conforme Corrar, Paulo e Dias Filho (2009) é importante considerar que a regressão logística calcula mudanças nas inter-relações dos *logs* da variável dependente e não na própria variável, como acontece com a linear.

Logo que o modelo logístico tenha sido ajustado a um conjunto de dados pode-se obter a razão de chance estimada. Para isto, eleva-se a constante matemática (e) ao expoente composto dos coeficientes estimados:

$$\left(\frac{p}{1 - p} \right) = e^{(b0 + b1\ X1 + b2\ X2\ .....\ + bn\ Xn)} . \quad (2.8)$$

Se a razão de chance estiver estimada, alcança-se o objetivo final, ou seja, é possível identificar a probabilidade associada à ocorrência de um determinado evento.

Simplificando, o modelo de regressão logística pode ser escrito como expresso pela equação:

$$P = \frac{1}{1 + e^{-(b0 + b1\ X1 + b2\ X2\ .....\ + bn\ Xn)}} = \frac{1}{1 + e^{-(\ln odds)}} . \quad (2.9)$$

Com essa transformação logística, a variável dependente passa a ser linear em relação à variável independente, assim como os parâmetros (coeficientes).

Logo identificada à equação, estima-se os coeficientes com a utilização do método da máxima verossimilhança.

Na regressão logística, os coeficientes medem o efeito de alterações nas variáveis independentes sobre o logaritmo natural da razão de probabilidades, chamado de *logit* (BRITO; ASSAF NETO, 2005) e, em conseqüência, os modelos são denominados modelos *logit* (GUJARATI, 2004).

Segundo Selau (2008) os coeficientes estimados (b_0, b_1, ... ,b_n) são medidas das variações na proporção das probabilidades, chamada de razão de desigualdade. São expressos em logaritmos, necessitando serem transformados para facilitar a interpretação.

Ao utilizar a técnica de regressão logística, o interesse pode estar na identificação do efeito de um fator de risco específico ou em determinar quais são os vários fatores associados com a variável resposta (SELAU, 2008). Para Hosmer e Lemeshow (1989), a função logística vem sendo bastante aplicada não apenas pela simplicidade de suas propriedades teóricas, mas principalmente, devido a sua simples interpretação como o logaritmo da razão de chances (*odds ratio*).

Tais procedimentos não alteram a forma de leitura e interpretação do sinal do coeficiente. Um coeficiente positivo revela que aquela variável aumenta a probabilidade de ocorrência do evento, enquanto que um valor negativo diminui a probabilidade prevista.

Mesmo a regressão logística sendo uma técnica robusta, existe o pressuposto importante da alta correlação entre as variáveis independentes, já que o modelo é sensível a colinearidade entre as variáveis (HAIR; ANDERSON; TATHAM; BLACK, 2005). A utilização de variáveis altamente correlacionadas para a estimação do modelo pode ocasionar estimativas extremamente exageradas dos coeficientes de regressão (HOSMER; LEMESHOW, 1989).

Corrar, Paulo e Dias Filho (2009) destacam alguns fatores que contribuem para o êxito na utilização da técnica:

- comparada a outras técnicas de dependência, a regressão logística acolhe com mais facilidade variáveis categóricas;

- mostra-se mais adequada à solução de problemas que envolvem estimação de probabilidades, pois trabalha com uma escala de resultados que vai de zero a um;

- requer um número menor de suposições iniciais, se comparada com outras técnicas utilizadas para discriminar grupos;

- admite variáveis independentes métricas e não-métricas, simultaneamente;

- facilita a construção de modelos destinados à previsão de um evento em diversas áreas de conhecimento;

- tendo em vista que o referido modelo é mais flexível quanto às suposições iniciais, tende a ser mais útil e a apresentar resultados mais confiáveis;

- os resultados da análise podem ser interpretados com relativa facilidade, já que a lógica do modelo se assemelha em muito à de outras técnicas bem conhecidas, como a regressão linear;

- apresenta facilidade computacional, tendo sido incluída em vários pacotes estatísticos amplamente difundidos em todo o mundo.

Sendo assim, se o pesquisador tem um problema que envolva uma variável dependente dicotômica não é necessário apelar para métodos elaborados para suprir as limitações da regressão múltipla, nem precisa forçar-se a usar uma outra técnica, principalmente se suas suposições estatísticas não são satisfeitas. A regressão logística aborda satisfatoriamente esses problemas e oferece um método de analise desenvolvido especialmente para lidar com esse tipo de situação da forma mais eficiente possível (HAIR; ANDERSON; TATHAM; BLACK, 2005).

2.3.1 Método da Máxima Verossimilhança

A regressão logística diferencia da regressão múltipla no método de estimar os coeficientes. Ao invés de minimizar os desvios quadráticos (método dos mínimos quadrados), a regressão logística maximiza a “verossimilhança” de que um evento ocorra (HAIR; ANDERSON; TATHAM; BLACK, 2005).

Assim sendo, a estimação do modelo é realizada pelo método da máxima verossimilhança, devido à natureza não linear da transformação logística. Este método é usado para encontrar as melhores estimativas “mais prováveis” para os coeficientes e a variável dependente é transformada em uma variável de base logarítmica. Observa-se, que o valor de verossimilhança é utilizado no lugar da

soma dos quadrados dos resíduos, usado na regressão múltipla, quando se calcula a medida de ajuste geral do modelo (HAIR; ANDERSON; TATHAM; BLACK, 2005).

O valor de verossimilhança é dado pela expressão -2LL, – 2VL ou -2log verossimilhança, que é o logaritmo natural do *Likelihood Value* multiplicado por -2, seguindo-se uma distribuição Qui-quadrado:

$$X^2 = 2\ [VL\ (novo) - VL\ (básico)]\ . \qquad (2.10)$$
$$(gl = K_{novo} - K_{básico})$$

sendo:
X^2 = distribuição de Qui-Quadrado;
VL (novo) = valor de verossimilhança incluindo a constante e os coeficientes dos previsores;
VL (básico) = valor de verossimilhança somente com a constante;
gl = graus de liberdade;
K_{novo} = número de parâmetros estimado (somente a constante);
$K_{básico}$ = número parâmetros estimados (constante e o número de previsores).

Observa-se que o valor da verossimilhança também pode ser comparado entre equações, onde a diferença representa a mudança no ajuste preditivo de uma equação para outra. Programas estatísticos têm testes automáticos para a significância dessas diferenças. O teste qui-quadrado para a redução no valor do logaritmo da verossimilhança fornece uma medida que possibilita melhora devido à introdução das variáveis independentes (HAIR; ANDERSON; TATHAM; BLACK, 2005).

A distribuição qui-quadrado utilizada tem graus de liberdade igual ao número de parâmetros no novo modelo menos o número de parâmetros no modelo básico. O número de parâmetros, K no modelo básico é sempre igual a 1, pois a constante é o único parâmetro a ser estimado, e qualquer modelo subseqüente terá um número de graus de liberdade igual ao número de previsores mais 1, ou seja, o número de previsores mais o parâmetro representando a constante (FIELD, 2009).

Corrar, Paulo e Dias Filho (2009) ressaltam que o *Log Likelihood Value* é uma das principais medidas de avaliação geral da regressão logística, que busca

aferir a capacidade do modelo estimar a probabilidade associada à ocorrência de determinado evento.

Para Field (2009, p.224):

> A verossimilhança-log é, portanto baseada na soma das probabilidades associadas com a saída real e a prevista. A estatística de verossimilhança-log é análoga à soma dos resíduos ao quadrado na regressão múltipla, no sentido de que ela é um indicador de quanta informação não explicada ainda existe após o modelo ter sido ajustado. Como conseqüência, tem-se que valores altos da estatística de verossimilhança-log indicam uma aderência pobre do modelo, porque quanto maior for esse valor, mais observações não explicadas existirão.

Um modelo bem ajustado terá um valor pequeno para -2LL sendo que o valor mínimo é 0 (zero). Um modelo com ajuste perfeito terá como resposta um valor de verossimilhança igual a 1 (um) e, portanto, -2LL será igual a 0 (zero).

2.3.2 O Teste Wald

Para Corrar, Paulo e Dias Filho (2009), a finalidade da estatística Wald é testar o grau de significância de cada coeficiente da equação logística, inclusive a constante, ou seja, verifica se cada parâmetro estimado é significativamente diferente de zero.

Hair, Anderson, Tatham e Black (2005), sugerem o uso da estatística de Wald para testar a significância dos coeficientes. Este teste fornece a significância estatística para cada coeficiente estimado, de modo que o teste de hipóteses pode ocorrer com acontece na regressão múltipla.

Hosmer e Lemeshow (1989) destacam que o teste Wald é obtido pela comparação da estimativa da máxima verossimilhança do parâmetro de inclinação b_i em relação à estimativa do seu erro padrão. A razão resultante, sob a hipótese que $b_i = 0$, segue uma distribuição normal padrão. O teste é calculado pela equação:

$$\text{Wald} = \frac{b_i}{SE\,(b_i)}, \qquad (2.11)$$

sendo:

b_i = estimativa do coeficiente de uma variável independente incluída no modelo;

$SE\,(b_i)$ = erro padrão do coeficiente.

As hipóteses para o teste são:

$$\begin{cases} H_o: b_i = 0 \\ H_1: b_i \neq 0 \end{cases}$$

Comparando-se o teste Wald calculado com um valor de $Z_{\alpha/2}$ tabelado (bilateral) a um nível de significância, torna-se possível concluir pela aceitação ou rejeição da hipótese nula (H_o).

O teste Wald assim como o teste da razão de verossimilhança precisam da estimativa de máxima verossimilhança dos parâmetros β_i´s (MEZZOMO, 2009).

Para Hosmer e Lemeshow (1989), o método para testar a significância dos coeficientes de uma variável na regressão logística é similar ao utilizado na regressão linear, mas usa uma função de verossimilhança para uma variável dicotômica.

Hauck e Donner (1977) apud Hosmer e Lemeshow (1989) destacam que a estatística Wald frequentemente rejeita coeficientes que seriam significativos, e devido a isto, os autores recomendam que o teste da razão de verossimilhança deva ser usado.

2.3.3 O Teste Hosmer e Lemeshow

Outro mecanismo que pode auxiliar a identificar a capacidade preditiva do modelo logístico é o teste de Hosmer e Lemeshow, que se trata de um teste Qui-quadrado capaz de prever as possíveis diferenças significativas existentes entre as classificações realizadas pelo modelo e a realidade observada (CORRAR; PAULO; DIAS FILHO, 2009).

As hipóteses para o teste são:

$$\begin{cases} H_o: Y = \hat{Y} \\ H_1: Y \neq \hat{Y} \end{cases}$$

Considerando Y como o valor real da variável observada e $\hat{Y}$ como o valor previsto, o teste é feito com o intuito de medir a proximidade de ambos. Portanto, busca-se a um nível de significância, aceitar a hipótese nula (H_o) de que não existem

diferenças significativas entre as classificações preditas pelo modelo e as observadas. Caso a hipótese nula seja rejeitada, ocorre a aceitação da hipótese alternativa (H_1) e isto revela que o modelo não representa a realidade de forma satisfatória, ou seja, o modelo não é capaz de produzir estimativas e classificações muito confiáveis.

Segundo Zanini (2007), quanto menor é o valor da diferença entre Y e Ŷ, mais os valores previstos se aproximam dos reais e, melhor o desempenho preditivo do modelo.

2.3.4 Pseudo R²

Os chamados Pseudos – R-Quadrado são indicadores que cumprem um papel semelhante ao coeficiente de determinação da regressão linear, ou seja, medem o poder de explicação ou ajuste geral do modelo. Dentre eles, segundo Corrar, Paulo e Dias Filho (2009) estão:

- R² de Cox-Snell - trata-se de um mecanismo que pode ser utilizado para comparar o desempenho de modelos concorrentes. Baseia-se na verossimilhança-log do modelo (VL (novo)) e a verossimilhança-log do modelo original (VL (básico)) e o tamanho da amostra (n):

$$R^2_{CS} = 1 - e^{[-2/n\ (VL\ (novo) - VL\ (básico))]}. \qquad (2.12)$$

sendo:
R^2_{CS} = indicador de Cox-Snell;
n = tamanho da amostra;
VL (novo) = valor de verossimilhança incluindo a constante e os coeficientes dos previsores;
VL (básico) = valor de verossimilhança somente com a constante.

Cabe lembrar que este indicador baseia-se no *Likelihood Value* e situa-se numa escala que começa em 0 (zero), mas não chega a 1 (um) em seu limite superior.

- R^2 de Nagelkerke – este coeficiente foi proposto por Nagelkerke em 1991 com a finalidade de ajustar o índice de Cox-Snell para que ele pudesse chegar ao referido limite máximo, em uma escala que vai de zero a um. Sua finalidade é a mesma do coeficiente mencionado anteriormente, sendo seu valor obtido pela expressão:

$$R^2_N = \frac{R^2_{CS}}{1 - e^{[2\,(VL(básico)\,/\,n]}} \quad . \qquad (2.13)$$

sendo:

R^2_N = indicador de Nagelkerke;

R^2_{CS} = indicador de Cox-Snell;

e = base dos logaritmos neperianos (2,718);

VL (básico) = valor de verossimilhança somente com a constante;

n = tamanho da amostra.

Cabe salientar que não existe consenso quanto a superioridade deste ou de outro coeficiente enquanto medida de adequação do modelo e como não são conflitantes entre si, pode-se utilizá-las em conjunto (CORRAR; PAULO; DIAS FILHO, 2009). Também, sugere-se que os Pseudos R^2 sejam utilizados apenas como uma medida aproximada do poder preditivo de cada modelo (FIELD, 2009)

2 METODOLOGIA

Neste capítulo serão abordadas todas as etapas realizadas para a construção do modelo proposto desta pesquisa empírica.

3.1 Delineamento

Para Inácio Filho (2004), metodologia consiste em um conjunto de procedimentos e técnicas utilizadas no processo de investigação, incluindo os aspectos relacionados à como fazer a pesquisa.

Segundo Lakatos e Marconi (2008), metodologia é um conjunto de atividades sistemáticas e racionais que com segurança e economia, permite alcançar o objetivo traçando o caminho a ser seguido, detectando erros e auxiliando as decisões do pesquisador.

A pesquisa desenvolvida neste estudo classifica-se como quantitativa, descritiva e exploratória.

Conforme Hair, Babin, Money e Samouel (2005), a pesquisa quantitativa é aquela que utiliza números para representar as propriedades em estudo analisando-os por meio de técnicas estatísticas.

Para Gil (2008), a pesquisa descritiva tem como objetivo a descrição das características de determinada população ou fenômeno ou o estabelecimento de relações entre as variáveis dependentes e independentes.

Segundo Sampieri (2006), estudos exploratórios são realizados quando se deseja examinar um tema ou problema de pesquisa pouco estudado. São desenvolvidos com o objetivo de proporcionar visão geral, de tipo aproximativo, acerca de determinado fato, buscando desenvolver, esclarecer e modificar conceitos e idéias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores (GIL, 2008).

Nesta investigação, optou-se, por aplicar a análise multivariada, que pode ser definida como o conjunto de métodos que permitem a análise simultânea dos dados recolhidos para um ou mais conjuntos de indivíduos (populações ou amostras) caracterizados por mais de duas variáveis correlacionadas entre si (CORRAR, PAULO e DIAS FILHO, 2009).

A técnica aplicada foi a de Regressão Logística Múltipla que, trata-se de uma ferramenta bastante utilizada em investigações acadêmicas e que tem a finalidade de fornecer informações significativas sobre o efeito destes fatores associados à ocorrência de um determinado evento.

3.2 Definição do Universo e a Estatística Descritiva

Para cumprir com o objetivo proposto nesta investigação foram coletadas as informações socioculturais do respectivo banco de dados da COPERVES, referentes aos candidatos inscritos que realizaram o processo seletivo.

Define-se Universo ou População como o conjunto de todas as unidades elementares de interesse (BOLFARINE e BUSSAB, 2005).

Fonseca e Martins (2008) destacam que o conceito de Universo é intuitivo, trata-se do conjunto de indivíduos ou objetos que apresentam em comum determinadas características definidas para o estudo.

O universo desta pesquisa esta associado aos candidatos do processo seletivo extraordinário de admissão para o ensino superior aos novos cursos de Engenharia da UFSM, totalizando em 2708 vestibulandos.

Primeiramente, foi dimensionado o tamanho da amostra acerca da população em estudo e logo foi utilizada a técnica de estatística descritiva para identificar o perfil destes candidatos.

A estatística descritiva é uma ferramenta utilizada para descrever e resumir um conjunto de dados, de modo que eles possam ser facilmente descritos e interpretados.

Crespo (2010) destaca que a coleta, a organização e a descrição dos dados estão a cargo da estatística descritiva.

Para Bruni (2007), a principal função da estatística descritiva consiste em resumir dados e informações investigadas, expondo-os da maneira mais prática e simples possível. Dessa maneira reduz-se o conjunto de dados, tornando-o mais maleável, constituindo tabelas, gráficos ou sumarizando os seus valores através de medidas descritivas (LOPES, 2008).

3.3 Dimensionamento da amostra

Fonseca e Martins (2008, p.177), definem que “amostra é um subconjunto da população”. Corresponde a parcelas do todo e costumam ser extraídas e analisadas quando o estudo envolve populações finitas com tamanhos consideráveis ou populações infinitas, que apresentam elementos que não podem ser contados (BRUNI, 2007).

Segundo Bruni (2007, p.169), “a amostra consiste em uma maneira de não estudar o conjunto como um todo, mas uma parte dele, sem que ocorra a perda das características essenciais da população”.

Portanto, para construção do modelo é necessário obter uma amostra significativa e representativa da população em estudo. A melhor maneira de se obter uma amostra representativa é empregar um procedimento aleatório para a seleção dos indivíduos (CALLEGARI-JAQUES, 2003).

Dessa forma a amostra será classificada como probabilística, visto basear-se em algum instrumento aleatório que lhes dá uma chance conhecida de serem selecionados, assim minimizando a tendenciosidade de seleção. As estimativas baseadas em uma amostra probabilística podem ser generalizadas para a população-alvo com um nível específico de segurança (HAIR, BABIN, MONEY e SAMOUEL, 2005).

O dimensionamento da amostra aleatória probabilística adotado foi obtido para uma população finita foi obtido utilizando um erro de amostragem (precisão) de 5%, com um nível de confiança de 99% e com a proporção de elementos favoráveis e desfavoráveis de 50% para cada um deles.

A fórmula utilizada para o cálculo do tamanho da amostra é definida por:

$$n = \frac{Z^2 . p . q . N}{e^2 (N - 1) + Z^2 . p . q} \quad , \qquad (3.1a)$$

onde:

n = tamanho da amostra;

N = tamanho do universo = 2.708;

z = valor obtido na curva normal com 99% de probabilidade = 2,58;

p = proporção de elementos favoráveis = 0,5;
q = proporção de elementos desfavoráveis = 0,5;
e = erro de amostragem = 0,05;

Sendo assim:

$$n = \frac{2{,}58^2 . 0{,}5 . 0{,}5 . 2708}{0{,}05^2 \ (2707) \ \ 2{,}58^2 . 0{,}5 . 0{,}5} \cong 535 \text{ candidatos} \quad . \qquad (3.1b)$$

O critério de amostragem aplicado foi o da Amostragem Aleatória Simples (AAS), no qual todos os indivíduos da população têm igual probabilidade de serem selecionados. De acordo com Bolfarine e Bussab (2005, p.16), na AAS, “cada unidade elementar é sorteada com igual probabilidade, individualmente, sem estratificação, e com um único estágio e seleção aleatória”. Para realizar os sorteios foram utilizadas as “tábuas de números aleatórios”, que consistem em tabelas que apresentam sequências dos dígitos de 0 a 9 distribuídos aleatoriamente (FONSECA e MARTINS, 2008)

3.4 Instrumento de obtenção dos dados

O instrumento de coleta de dados utilizado para a obtenção das variáveis foi um questionário sociocultural de múltipla escolha formulado e aplicado pela COPERVES no ato da inscrição no vestibular extraordinário, no ano de 2009 (ANEXO A).

Sendo assim, o banco de dados utilizado corresponde às informações socioculturais dos candidatos inscritos e que realizaram as provas do processo seletivo. Das 39 questões que compunham o questionário foram selecionadas 34, que, segundo o ponto de vista do pesquisador (QUADRO 1), podem afetar na determinação da seleção dos candidatos aos novos cursos de Engenharia para ingresso na IFES em estudo. Estas questões, tratam-se de informações pessoais, de formação educacional, sobre a vida econômica familiar e, sobre hábitos e costumes dos candidatos.

3.5 Variável Dependente

Para estimar o modelo utilizou-se como variável dependente binária (Y) a seleção dos candidatos, ou seja, o resultado da análise possibilitou associação às categorias selecionado ou não selecionado no processo seletivo (Tabela 1).

TABELA 1 - Variável dependente utilizada no modelo de regressão logística.

Variável Dependente	Codificação
Selecionado	1
Não Selecionado	0

Fonte: Elaboração Própria

3.6 Variáveis Independentes ou Covariáveis

No estudo foram consideradas inicialmente as seguintes covariáveis para construir o modelo proposto de Regressão Logística Múltipla, representadas por:

X_1: Sexo Masculino
X_2: Estado Civil Solteiro
X_3: Língua Estrangeira Inglês
X_4: Idade até 25 anos
X_5: Natural de Santa Maria
X_6: Concluiu ou concluirá o atual Ensino Médio
X_7: Concluiu ou concluirá seu Ensino Médio em Escola Pública
X_8: Realiza ou realizará o Ensino Médio no turno diurno
X_9: Concluiu ou concluirá o Ensino Médio nos últimos três anos
X_{10}: Freqüentou ou freqüenta curso pré-vestibular
X_{11}: O principal fator para ter sucesso no vestibular é muito estudo pessoal
X_{12}: Já prestou vestibular uma ou mais vezes
X_{13}: Já iniciou algum curso superior
X_{14}: O principal motivo que o levou a ter interesse em ingressar num curso superior é a formação profissional
X_{15}: Pais aprovam a escolha profissional

X_{16}: Pai possui Ensino Fundamental

X_{17}: Pai possui Ensino Médio

X_{18}: Pai possui Ensino Superior e Pós-Graduação

X_{19}: Pai sem escolaridade

X_{20}: Mãe possui Ensino Fundamental

X_{21}: Mãe possui Ensino Médio

X_{22}: Mãe possui Ensino Superior e Pós-Graduação

X_{23}: Mãe sem escolaridade

X_{24}: Não participa na vida econômica da família (não trabalha e os gastos são financiados pela família)

X_{25}: Possui renda total mensal familiar até 05 salários mínimos

X_{26}: Possui renda total mensal familiar acima de 05 salários mínimos até 09 salários mínimos

X_{27}: Possui renda total mensal familiar acima de 09 salários mínimos

X_{28}: Costuma ler livros, jornais e revistas

X_{29}: O tipo de leitura que mais ocupa seu tempo é a informativa

X_{30}: Costuma assistir televisão

X_{31}: O tempo que utiliza para assistir televisão é de uma ou mais horas

X_{32}: Tem acesso a computador

X_{33}: Tem acesso a internet em casa

X_{34}: Pratica esporte

Cabe salientar que nas 34 covariáveis utilizadas no modelo estatístico, por se tratarem de variáveis qualitativas originalmente não mensuráveis, foram utilizadas variáveis *dummies*, denominadas de variáveis binárias que podem tomar um de dois valores, em geral 0 ou 1, isto é, servem para descrever qualquer evento que tenha apenas dois resultados possíveis (HILL, 2003).

Neste modelo as variáveis explicativas podem assumir dois valores: 0 (zero) indicando a ausência de um determinado atributo e 1 (um) indicando a presença.

No Quadro 1, explica-se como as covariáveis selecionadas podem afetar na determinação da seleção dos candidatos para ingresso na IFES em estudo, ou seja, verificar se há alguma relação com a variável dicotômica a ser explicada. Sendo assim, busca-se investigar se as variáveis obtidas a partir do banco de dados

explicam a situação de seleção ou não seleção dos candidatos ao processo seletivo dos cursos de Engenharia da UFSM.

Sigla	Variáveis Explicativas	Relação com a seleção no vestibular
Informações pessoais dos candidatos		
X_1	Sexo	Existência de diferenças na preparação para o vestibular atribuído a homens e mulheres.
X_2	Estado Civil	Candidatos casados ou morando com companheiro apresentam menos tempo para estudar do que os que são solteiros.
Informações pessoais dos candidatos		
X_4	Idade	Tendência do aluno com menos idade de depender de seus pais ou familiares e com isso apresentam mais tempo para se dedicar aos estudos.
X_5	Naturalidade	Acredita-se que os candidatos naturais de Santa Maria têm mais chances de ingressar na UFSM, pois na prova existem questões regionais
Informações sobre formação educacional dos candidatos e familiares		
X_3	Língua Estrangeira	Os candidatos que escolhem a língua inglesa têm maior influência pelo fato deste idioma ser estudado desde o ensino fundamental.
X_6,X_7,X_8,X_9	Ensino Médio	O tipo de ensino médio, turno, escola e ano de conclusão podem ter influencia sobre a seleção do candidato.
X_{10},X_{11} e X_{12}	Sucesso no Vestibular	Acredita-se que quanto mais estudo pessoal e a experiência no processo melhor o desempenho do candidato
X_{13},X_{14} e X_{15}	Sobre Formação Superior	O que motiva um candidato a estudar pode surtir efeito no processo seletivo
X_{16},X_{17},X_{18}, X_{19},X_{20},X_{21}, X_{22} e X_{23}	Nível de Escolaridade dos pais	Pais com nível de escolaridade mais elevado tendem a oferecer aos filhos um nível de escolaridade superior que facilita no ingresso a IFES
Informações sobre a vida econômica familiar		
X_{24}	Participação na vida econômica familiar	O candidato que não trabalha e tem seus gastos financiados pela família possui maior tempo para dedicar-se aos estudos
X_{25}, X_{26} e X_{27}	Renda Familiar	Quanto maior a renda da família maiores condições os candidatos terão em freqüentar bons cursinhos pré-vestibulares e bons colégios de ensino, com isto suas possibilidades aumentam em ingressar em uma IFES
Informações sobre hábitos e costumes dos candidatos		
X_{28} e X_{29}	Leitura	O hábito da leitura facilita a escrita que é importante nas questões dissertativas e redação do processo seletivo.
X_{30} e X_{31}	Assistir televisão	É importante o candidato se manter atualizado de todos os acontecimentos.
X_{32} e X_{33}	Acesso a computador e internet	O acesso a computador e a internet facilita na aquisição de conhecimentos e no estudo.

Sigla	Variáveis Explicativas	Relação com a seleção no vestibular
Informações sobre hábitos e costumes dos candidatos		
X_{34}	Praticar esportes	A prática esportiva torna a vida mais saudável e com isso ocorre mais disposição para estudar.

QUADRO 1. Efeitos das variáveis independentes na seleção do vestibular.
Fonte: Adaptado de Camargos et al. (2008) apud Ribeiro (2008).

3.7 Desenvolvimento do modelo

Após serem identificadas e codificadas as variáveis dependentes e independentes partiu-se para o desenvolvimento do modelo de regressão.

Primeiramente foi realizada a seleção das variáveis candidatas ao modelo de regressão logística múltipla por meio de uma análise de regressão logística univariada. Este procedimento foi realizado para verificar a existência de associação de cada covariável com a variável dependente.

As associações entre cada variável independente com a variável dependente pode ser verificada pelo teste de independência do qui-quadrado ou pela análise de regressão logística univariada, no qual fornecem a mesma informação.

O critério utilizado para verificar a significância de associação entre cada covariável com a variável dependente foi de um $p \leq 0,25$ pois, segundo Hosmer e Lemeshow (1989, p.86), “toda covariável que tiver um p-valor menor ou igual a 0,25 deve ser considerada como uma candidata para o modelo múltiplo junto com todas as variáveis de importância conhecidas”.

Hosmer e Lemeshow (1989, p.86) também salientam que o uso do 0,25 como um critério para a seleção de variáveis é baseado no trabalho de Bendel e Afifi (1977) sobre regressão linear e no trabalho de Mickey e Greenland (1989) sobre regressão logística. Estes autores mostram que o uso do tradicional nível (tais como 0,05) frequentemente não identifica as variáveis conhecidas como importantes.

Com isso, a covariável que apresentou um p-valor menor ou igual a 0,25 no teste univariado, verificado pela significância do teste de Wald, foi incluída como possível candidata a fazer parte do modelo múltiplo.

A regressão logística univariada também possibilitou a estimação das razões de chance (*odds ratio*) e os respectivos intervalos de confiança com 95%.

O próximo passo foi verificar a ausência de colinearidade ou multicolinearidade entre as variáveis independentes, com a finalidade de identificar

a correlação entre as mesmas. Para tanto, utilizou-se um p-valor significativo ao nível de 5%.

A análise da multicolineariedade verifica se existe correlação entre duas ou mais variáveis explicativas (X_i), levando a dificultar a separação dos efeitos de cada uma delas sozinha sobre a variável explicada (Y_i) (CORRAR; PAULO e DIAS FILHO, 2009).

Corrar; Paulo e Dias Filho (2009, p.156) também ressaltam que:

> Do ponto de vista técnico, a multicolinearidade tende a distorcer os coeficientes angulares estimados para as variáveis que a apresentam, prejudicando a habilidade preditiva do modelo e a compreensão do real efeito da variável independente sobre o comportamento da variável dependente.

O coeficiente utilizado para a análise da multicolineariedade entre as variáveis explicativas (X_i) foi o de Correlação de Spearman.

Segundo HAIR, BABIN, MONEY e SAMOUEL (2005), o coeficiente de Correlação de Spearman é considerado uma estatística mais conservadora e é utilizado em escalas nominais ou ordinais (não-métricas).

Na sequência, foi realizada a análise de regressão logística múltipla, na qual, utilizou-se simultaneamente no modelo, as covariáveis independentes significativas da análise de regressão logística univariada e não colineares. Para isto utilizou-se o método *enter,* que consiste no método da entrada forçada, ou seja, em que todas as covariáveis são colocadas no mesmo modelo de regressão em um único bloco e as estimativas dos parâmetros são calculadas para cada bloco (FIELD, 2009).

Na etapa seguinte foram eliminadas do modelo múltiplo, uma a uma, as covariáveis não significativas ($p > 0,05$), sempre verificando o efeito da saída de cada uma nos coeficientes (b_i) que ficavam no modelo. A cada eliminação de uma covariável (com maior valor de p), o modelo era processado para que os coeficientes fossem reajustados para um novo conjunto de covariáveis. Para decidir pela continuidade da covariável no modelo, optou-se pela verificação da significância do teste Wald com um $p \leq 0,05$.

Para avaliação e ajuste do modelo de estimação final permaneceram as covariáveis significativas ($p \leq 5\%$) verificadas pelo Teste Wald e também, comparou-se entre modelos alternativos os resultados dos testes de Pseudo R^2 de Cox-Snell e

Nagelkerke, o cálculo do -2LL (razão de verossimilhança) e a significância do teste de Hosmer e Lemeshow.

Logo após foi realizada a capacidade de previsão do modelo no qual foi adotado o ponto de corte no valor de 0,5 e sendo examinada pela estruturação de uma tabela de classificação, que verifica os erros e acertos para o modelo estimado.

A próxima etapa foi verificada pelo processo de validação do modelo estimado onde foi utilizada uma nova amostra de validação (n=535). É importante ressaltar que a amostra de validação partiu da amostra original, constituindo uma abordagem de validade interna.

Finalizando, com a função logit, foram estimadas algumas probabilidades de ocorrência do evento (seleção no vestibular) utilizando os valores dos coeficientes das covariáveis que foram inseridas no modelo final.

Os procedimentos estatísticos foram realizados por meio do auxílio dos programas computacionais Excel e SPSS *(Statistical Package for Windows*).

4 RESULTADOS

4.1. Análise Descritiva

Na análise descritiva foram elaboradas tabelas de frequência para traçar o perfil dos candidatos dos novos cursos de Engenharia da UFSM (n = 535). Para esta análise foram utilizadas as informações socioculturais dos candidatos inscritos e que realizaram o vestibular extraordinário, realizado em maio de 2009.

TABELA 2 - Informações pessoais dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).

Variável	Freqüência (%)
Sexo	
Masculino	425 (79,40)
Feminino	110 (20,60)
Estado Civil	
Solteiro	514 (96,10)
Outros	21 (3,90)
Idade	
≤ 25 anos	481 (89,90)
> 25 anos	54 (10,10)
Naturalidade	
Santa Maria	330 (61,70)
Outros	205 (38,30)

Fonte: Elaboração Própria

Na Tabela 2, pode-se verificar que a maioria dos candidatos inscritos e que realizam o vestibular extraordinário aos novos cursos de Engenharia são do sexo masculino (79,40%); solteiros (96,10%); com idade inferior ou igual a 25 anos (89,90%); e naturais do município de Santa Maria (61,70%).

TABELA 3 - Informações sobre a formação educacional dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).

Variável	Freqüência (%)
Tipo de Ensino médio concluído	
Atual Ensino Médio	475 (88,80)
Outros	60 (11,20)
Onde cursou o Ensino Médio	
Maior parte em Escola Pública	322 (60,20)
Outros	213 (39,80)
Turno em que realizou o Ensino Médio	
Diurno	487 (91,00)
Noturno	48 (9,00)
Ano em que concluiu o Ensino Médio	
Nos últimos três anos	353 (66,00)
Outros	182 (34,00)

Variável	Freqüência (%)
Freqüenta ou freqüentou Pré-Vestibular	
Sim	305 (57,00)
Não	230 (43,00)
Principal fator para ter sucesso no vestibular	
Muito estudo pessoal	397 (74,20)
Outros	138 (25,8)
Quantas vezes já fez vestibular	
Uma ou mais vezes	468 (87,50)
Nenhuma	67 (12,50)
Iniciou algum curso superior	
Sim	166 (31,00)
Não	369 (69,00)
Principal motivo que o levou a ter interesse em ingressar em um curso superior	
Formação Profissional	469 (87,70)
Outros	66 (12,30)
Posição dos pais diante da escolha profissional	
Aprovam	465 (86,90)
Não aprovam	70 (13,10)
Escolaridade do Pai	
Sem escolaridade	17 (3,18)
Com Ensino Fundamental	126 (23,55)
Com Ensino Médio	228 (42,62)
Com Ensino Superior e Pós-Graduação	164 (30,65)
Escolaridade da Mãe	
Sem escolaridade	4 (0,75)
Com Ensino Fundamental	104 (19,44)
Com Ensino Médio	194 (36,26)
Com Ensino Superior e Pós-Graduação	233 (43,55)

Fonte: Elaboração Própria

Observa-se que dos 535 candidatos da amostra, 88,8% concluíram ou concluirão o atual Ensino Médio; 60,2% cursaram o Ensino Médio a maior parte em escola pública; 91% realizaram seus estudos de Ensino Médio no turno diurno; e 66% concluíram o Ensino Médio nos últimos três anos. Também, verifica-se que 57% frequentou ou frequenta Pré-Vestibular; que 74,2% dos candidatos acreditam que o principal fator para ter sucesso no vestibular é muito estudo pessoal; 87,5% já prestaram vestibular uma ou mais vezes; 69% nunca iniciaram um curso superior; 87,7% afirmam que a formação profissional é o principal motivo que os levou a ter interesse em ingressar em um curso superior e 86,9% dos pais aprovam a escolha profissional dos seus filhos.

Com relação à escolaridade dos familiares dos candidatos; constata-se que a maioria dos pais (42,62%) possui ensino médio sendo que 3,18% não possuem escolaridade. Os resultados apurados quanto à escolaridade das mães mostraram

que 43,55% possuem ensino superior e pós-graduação e que 0,75% não possuem escolaridade.

TABELA 4 - Informações sobre a vida econômica familiar dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).

Variável	Freqüência (%)
Participa na vida econômica da família	
Sim	123 (23,00)
Não	412 (77,00)
Renda total mensal da família	
≤ 5 S.M*	280 (52,34)
Acima de 5 S.M. até 9 S.M*	150 (28,04)
> de 9 S.M*	105 (19,62)

Fonte: Elaboração Própria
*S.M = salários mínimos.

Na Tabela 4, percebe-se que 77% dos candidatos não participam na vida econômica familiar, ou seja, não trabalham e seus gastos são financiados pela família. Quanto a renda total mensal familiar, verifica-se que a 52,34% apresentam uma renda inferior e igual a 5 salários mínimos, o que teoricamente pode indicar um baixo poder aquisitivo dos mesmos, restando 47,66% com renda total mensal familiar acima de 5 salários mínimos.

TABELA 5 - Informações sobre hábitos e costumes dos candidatos inscritos e que realizaram o vestibular extraordinário aos novos cursos de Engenharia, UFSM, 2009 (n = 535).

Variável	Freqüência (%)
Costuma ler livros, jornais e revistas	
Sim	508 (95,00)
Não	27 (5,00)
Tipo de leitura que mais ocupa o tempo	
Informativa	325 (60,70)
Outras	210 (39,30)
Costuma assistir televisão	
Sim	521 (97,40)
Não	14 (2,60)
Tempo que utiliza para assistir à televisão	
Uma hora ou mais	365 (68,20)
Outros	170 (31,80)
Tem acesso a computador	
Sim	531 (99,30)
Não	4 (0,7)
Tem acesso a internet em casa	
Sim	393 (73,50)
Não	142 (26,50)
Pratica esporte	
Sim	476 (89,00)
Não	59 (11,00)

Fonte: Elaboração Própria

Com relação aos hábitos e costumes dos candidatos, verifica-se que 95% costumam ler livros, jornais e revistas, sendo que 60,7% ocupam seu tempo com leitura informativa. Também, observa-se que dos 535 candidatos analisados 97,4% costumam assistir televisão; 68,2% utilizam uma ou mais horas de seu tempo assistindo televisão; 99,3% têm acesso a computador; 73,5% têm internet em casa e 89% praticam algum tipo de esporte.

4.2. Análise de Regressão Logística Univariada

Na Tabela 6, são apresentados os resultados da análise de regressão logística univariada, utilizando as 34 covariáveis selecionadas referentes as informações socioculturais dos candidatos aos novos cursos de Engenharia do vestibular extraordinário da UFSM.

TABELA 6 – Resultado da análise de Regressão Logística Univariada.

Variáveis Independentes	p-valor*	OR	IC 95%
Língua Estrangeira Inglês (X_2)	0,002	2,291	1,351 – 3,884
Concluiu ou concluirá o atual Ensino Médio (X_6)	0,074	2,959	0,900 – 9,730
Cursou o ensino médio a maior parte em Escola Pública (X_7)	0,052	0,601	0,360 – 1,005
Freqüenta ou freqüentou pré-vestibular (X_{10})	0,207	1,409	0,828 – 2,397
Já prestou vestibular uma ou mais vezes (X_{12})	0,044	3,380	1,031 – 11,081
Pai com ensino fundamental (X_{16})	0,015	0,367	0,163 – 0,825
Pai com ensino superior e pós-graduação (X_{18})	0,010	1,984	1,175 – 3,351
Mãe com ensino fundamental (X_{20})	0,018	0,321	0,125 – 0,820
Mãe com ensino superior e pós-graduação (X_{22})	0,010	1,969	1,174 – 3,304
Não participa na vida econômica da família (X_{24})	0,051	2,075	0,997 – 4,320
Renda total mensal familiar ≤ 5 S.M (X_{25})	0,000	0,314	0,179 – 0,551
Renda total mensal familiar acima de 5 S.M até 9 S.M. (X_{26})	0,131	1,515	0,883 – 2,599
Renda total mensal familiar > 9 S.M. (X_{27})	0,001	2,614	1,493 – 4,575
Costuma ler livros, jornais e revistas (X_{28})	0,187	3,882	0,518 – 29,091

Variáveis Independentes	p-valor*	OR	IC 95%
Tem acesso a internet em casa (X_{33})	0,160	1,581	0,835 – 2,993

Fonte: Adaptado de Mezzomo (2009).
* p ≤ 0,25; OR = *odds ratio* bruto ,Categoria de referência=1; IC 95% = Intervalo de Confiança de 95%.

Percebe-se que das 34 covariáveis analisadas somente 15 apresentaram significância, seguindo o critério de Hosmer e Lemeshow (1989) de um p ≤ 0,25. Dentre as possíveis covariáveis candidatas a fazer parte do modelo múltiplo estão: língua estrangeira Inglês, concluiu ou concluirá o atual Ensino Médio, cursou o ensino médio a maior parte em Escola Pública, freqüenta ou freqüentou pré-vestibular, já prestou vestibular uma ou mais vezes, pai com ensino fundamental, pai com ensino superior e pós-graduação, mãe com ensino fundamental, mãe com ensino superior e pós-graduação, não participa na vida econômica familiar, renda total mensal familiar ≤ 5 S.M, renda total mensal familiar acima de 5 S.M. até 9 S.M., renda total mensal familiar > 9 S.M., costuma ler livros, jornais e revistas e tem acesso a internet em casa. Destaca-se que o valor da significância da covariável foi identificado pelo valor do p global do teste Wald.

Também foram calculadas as razões de chance (OR = *odds ratio* bruto) das covariáveis, na qual, observa-se que as que apresentaram uma maior chance na seleção ao vestibular são as variáveis “Costuma ler livros, jornais e revistas” (OR = 3,882) e “Já prestou vestibular uma ou mais vezes” (OR = 3,380).

4.3 Análise da Multicolinearidade

O coeficiente utilizado para verificar a existência de multicolinearidade entre as covariáveis independentes foi o de Correlação de Spearman e seus resultados seguem apresentados na Tabela 7.

Para esta análise foram utilizadas as 15 covariáveis que apresentaram significância na analise de regressão univariada.

TABELA 7 - Resultados do Coeficiente de Correlação de Spearman para a multicolinearidade.

Variáveis	X_6	X_{10}	X_{12}	X_{16}	X_{20}	X_{22}	X_{24}	X_{25}	X_{26}	X_{27}
X_6	1									
X_{10}	0,146	1								
X_{12}	0,116	0,196	1							
X_{16}	-0,150	-0,058	-0,027	1						
X_{20}	-0,136	-0,072	0,020	0,425	1					

Variáveis	X_6	X_{10}	X_{12}	X_{16}	X_{20}	X_{22}	X_{24}	X_{25}	X_{26}	X_{27}
X_{22}	0,147	0,067	-0,044	-0,249	-0,386	1				
X_{24}	0,284	0,145	0,035	-0,175	-0,243	0,142	1			
X_{25}	-0,114	-0,163	-0,044	0,262	0,301	-0,286	-0,157	1		
X_{26}	0,077	0,038	-0,015	-0,156	-0,180	0,159	0,084	-0,637	1	
X_{27}	0,065	0,197	0,053	-0,168	-0,181	0,179	0,127	-0,496	-0,301	1
X_{28}	-0,001	0,024	-0,036	-0,019	0,022	0,058	-0,045	0,002	0,011	-0,020

Fonte: Elaboração Própria.
Variáveis não significativas a um nível de 5%.
X_6=Concluiu o atual Ensino Médio X_{10}=Freqüentou ou freqüenta pré-vestibular; X_{12}= Já prestou vestibular uma ou mais vezes; X_{16}=Pai com ensino fundamental; X_{20}=Mãe com ensino fundamental; X_{22}=Mãe com ensino superior e pós-graduação; X_{24}=Não participa da vida econômica familiar; X_{25}=Renda total mensal familiar ate 5 S.M.; X_{26}=Renda total mensal familiar acima de 5 S.M até 9 S.M.; X_{27}=Renda total mensal familiar acima de 9 S.M e X_{28}=Costuma ler livros, jornais e revistas.

Os resultados mostraram que 11 covariáveis não estão correlacionadas entre si (TABELA 7), que são: Concluiu o atual Ensino Médio; Freqüentou ou freqüenta pré-vestibular; Já prestou vestibular uma ou mais vezes; Pai com ensino fundamental; Mãe com ensino fundamental; Mãe com ensino superior e pós-graduação; Não participa da vida econômica familiar; Renda total mensal familiar ate 5 S.M.; Renda total mensal familiar acima de 5 S.M até 9 S.M.; Renda total mensal familiar acima de 9 S.M e Costuma ler livros, jornais e revistas. Com isto, estas são as covariáveis selecionadas para serem testadas no modelo múltiplo.

As variáveis independentes que apresentaram correlação a um nível de significância de 5% são: Língua Estrangeira inglês; Cursou o ensino médio a maior parte em Escola Pública; Pai com ensino superior e pós-graduação; e Tem acesso a internet em casa. Devido a isto, estas covariáveis foram retiradas e não farão parte do modelo.

4.4 Avaliação do ajuste do modelo

Na etapa de avaliação do ajuste do modelo de regressão logística múltipla estimado foram realizados testes estatísticos com as covariáveis significativas para validar a sua aplicação.

Com isso, foram analisadas as 11 covariáveis não colineares, com a variável dependente e apenas 2 foram significativas ao nível de 5%, verificado pela significância do teste Wald.

Sendo assim, foram obtidos 3 modelos alternativos de estimação apresentados na Tabela 8. No modelo I foi inserida a constante e a variável “já prestou vestibular uma ou mais vezes” (X_{12}); o modelo II foi verificado pela presença

da constante com a variável “renda total mensal familiar até 5 S.M” (X_{25}) e o modelo III foi estimado com a constante e as duas covariáveis citadas anteriormente.

TABELA 8 - Modelos estimados de regressão logística.

Modelos/Variáveis	b_0	b_1	b_2	-2LL	Cox-Snell	Nagelkerke
I. X_{12}	-3,060	1,218				
Sig. Teste Wald	0,000*	0,044*		397,98	0,010	0,020
OR	0,047	3,380				
IC 95%		1,031-11,081				
II. X_{25}	-1,462	-1,159				
Sig. Teste Wald	0,000*	0,000*		385,57	0,033	0,063
OR	0,232	0,314				
IC 95%		O,179-0,551				
III. X_{12} e X_{25}	**-2,541**	**1,170**	**-1,143**			
Sig. Teste Wald	0,000*	0,050*	0,000*	**380,00**	**0,042**	**0,080**
OR	0,079	3,222	0,319			
IC 95%		0,974-10,658	0,181-0,560			

Fonte: Elaboração Própria.
* Estatisticamente significativo ao nível de 5%. OR: *Odds ratio* ajustado para as outras covariáveis da tabela por meio da regressão logística múltipla OR ajustado = 1:categoria de referência; IC95%: intervalo de confiança de 95%; - 2LL= razão de verossimilhança.
X_{12}= Já prestou vestibular uma ou mais vezes; X_{25}=Possui renda total mensal familiar até 5 S.M.

Observa-se que o modelo estimado de número III originou o modelo proposto de regressão logística, no qual foram inseridas as variáveis X_{12} (já prestou vestibular uma ou mais vezes) e X_{25} (renda total mensal familiar até 5 S.M).

Na comparação dos resultados dos testes de avaliação do ajuste do modelo entre as três simulações de modelos alternativos, verifica-se que o modelo III apresentou *log* da verossimilhança (-2LL) no valor de 380,00. Esta medida mede o grau de aderência do modelo e indica que quanto menor o seu valor melhor, devido ao fato de medir a razão entre os valores de saída previstos e os observados.

Com relação às variáveis independentes, percebe-se que em todos os modelos alternativos os coeficientes da constante e das variáveis independentes (X_{12} e X_{25}) apresentaram significância estatística a um nível de 0,05 a partir da interpretação do teste Wald, que tem a finalidade de verificar se os coeficientes de cada variável independente são significativamente diferentes de zero.

O coeficiente de Cox-Snell e Nagelkerke indicam o poder de previsão aproximado do modelo. As duas medidas podem ser utilizadas juntas para medir a adequação do modelo logístico e, constata-se que os melhores valores para estas medidas apresentam-se no modelo III.

Já o teste de Hosmer e Lemeshow (p=0,999) calculado no modelo III indica uma boa aderência entre os valores observados e previstos. Isto pode ser verificado também pelo percentual de acerto do modelo que alcança 87,5% do total.

4.5 Modelo de estimação da função Seleção

O modelo final de regressão logística foi formado por duas das 34 variáveis explicativas inseridas no estudo.

As variáveis inseridas no modelo foram X_{12} (já prestou vestibular uma ou mais vezes) e X_{25} (renda total mensal familiar até 5 S.M.). Com isto, a função *logit* para estimar o modelo de seleção dos candidatos e que obteve o melhor ajuste aos dados é dada por:

$$P = \frac{1}{1 + e^{-(-2{,}541 + 1{,}170\, X_{12} - 1{,}143\, X_{25})}} \quad . \qquad (4.1)$$

sendo:

P = probabilidade de seleção do candidato ao vestibular nos cursos de Engenharia;

X_{12} = já prestou vestibular uma ou mais vezes;

X_{25} = renda total mensal familiar até 5 salários mínimos.;

e = base dos logaritmos neperianos (2,718).

Após a verificação do modelo, percebe-se que os coeficientes das variáveis explicativas obtiveram o sinal esperado na função *logit*. Ou seja, a variável X_{12} com o coeficiente positivo indica que se o candidato já prestou vestibular uma ou mais vezes maior será a sua probabilidade de seleção enquanto que o coeficiente negativo da variável X_{25} indica que se ele possuir uma renda total mensal familiar até 5 S.M menores serão as chances do estudante ser selecionado.

Isso implica que a probabilidade de seleção ao vestibular esta relacionada ao candidato que já prestou vestibular uma ou mais vezes e ao fato dele ter uma renda total mensal familiar ate 5 S.M.

4.6 Capacidade de Previsão do Modelo

O objetivo da regressão logística é descrever a relação entre uma variável resposta – dependente – e uma ou mais variáveis explicativas – independentes

(SILVA e PERIÇARO, 2009), sendo a sua principal característica o fato da variável resposta estabelecida ser dicotômica, ou seja, poder assumir um entre dois resultados propostos. Nesta pesquisa, foi atribuído o valor zero para a probabilidade do candidato não ser selecionado no vestibular e valor um para a probabilidade do candidato ser selecionado.

O ponto de corte adotado no modelo foi o valor de 0,5, “valor padronizado para a técnica de regressão logística”, segundo Araújo e Carmona (2007). Conforme Hair Jr. et al. (2005), este valor representa a probabilidade de ocorrer o evento pelo critério de aleatoriedade ou chances iguais. Sendo assim, os candidatos que tiveram uma probabilidade estimada de seleção inferior a 0,5, foram classificados como não selecionados e aqueles que obtiveram a probabilidade superior a 0,5 foram classificados como selecionados.

A capacidade de previsão do modelo foi examinada pela estruturação de uma tabela de classificação, que verifica os erros e acertos para o modelo estimado, portanto, mostra os valores previstos e observados dos candidatos selecionados e não selecionados (Tabela 9). Esta tabela informa a capacidade de previsão do modelo quando a constante (b_0) e os coeficientes das covariáveis (X_{12} = já prestou vestibular uma ou mais vezes e X_{25} = renda total mensal familiar até 5 salários mínimos) são incluídas no mesmo (FIELD, 2009).

TABELA 9 - Tabela de Classificação – Previsão do Modelo

Observado	**Estimado**		**Classificações Corretas**
	Não Selecionado	**Selecionado**	
Não Selecionado	468	0	100,0%
Selecionado	67	0	0,0%
Total	535		87,5%

Fonte: Adaptado de Brito e Assaf Neto (2005) apud Ribeiro (2008).
Nota: O valor do ponto de corte é 0,5.

Percebe-se que o percentual de acerto do modelo representa 87,5% no total, tendo sido estimados corretamente 468 dos 535 candidatos da amostra de análise.

4.7 Validação do Modelo

Para Hair, Anderson, Tatham e Black (2005), no processo de validação do modelo de regressão logística, é importante utilizar o método de criação de amostras de análise e de validação.

O principal meio para validar um modelo estimado é pelo uso da amostra de validação, bem como a avaliação de sua precisão preditiva (HAIR, ANDERSON, TATHAM e BLACK, 2005). Sendo assim, a validade é estabelecida se o modelo de regressão logística classifica observações, em nível aceitável, que não foram usadas no processo de estimação.

Neste caso a amostra de validação partiu da amostra original e, então essa abordagem estabelece validade interna.

Obteve-se a seleção de uma nova amostra de 535 candidatos e foram inseridos o valor da constante bem como os parâmetros dos coeficientes das duas variáveis significativas do modelo estimado de regressão logística múltipla, que são: já prestou vestibular uma ou mais vezes (X_{12}) e possui renda mensal total até 5 salários mínimos (X_{25}).

A Tabela 10 apresenta os resultados obtidos a partir da amostra de validação.

TABELA 10 - Tabela de Classificação – Validação do Modelo

Observado	**Estimado**		**Classificações Corretas**
	Não Selecionado	**Selecionado**	
Não Selecionado	463	0	100,0%
Selecionado	72	0	0,0%
Total	535		86%

Fonte: Adaptado de Brito e Assaf Neto (2005) apud Ribeiro (2009)

Verifica-se que o percentual de acerto acumulado foi de 86%, onde foram classificados corretamente 463 candidatos da amostra de validação, enquanto que na amostra de análise o resultado foi de 87,5% de classificação correta.

4.8 Estimação da probabilidade de ocorrência da seleção de um candidato

A seguir, foram realizadas quatro situações hipotéticas utilizando o modelo múltiplo encontrado para estimar a probabilidade da ocorrência da seleção dos candidatos ao vestibular nos cursos de engenharia da UFSM.

Situação 1: Qual a probabilidade de um candidato que prestou vestibular com as características de já ter realizado o processo seletivo uma ou mais vezes e que possui a renda total mensal familiar até 5 salários mínimos, ingressar na universidade?

ln *odds* = - 2,541 + 1,170 . (1) – 1,143 . (1) = 2,514

$$P = \frac{1}{1 + e^{-(\ln odds)}} = \frac{1}{1 + e^{-(2,514)}} = 0,0749 = 7,49\%$$

Situação 2: Qual a probabilidade de um candidato que prestou vestibular ingressar a universidade sendo que já prestou o processo seletivo uma ou mais vezes e que não possui renda total mensal familiar até 5 salários mínimos?

ln *odds* = - 2,541 + 1,170 . (1) – 1,143 . (0) = - 1,371

$$P = \frac{1}{1 + e^{-(\ln odds)}} = \frac{1}{1 + e^{-(-1,371)}} = 0,2024 = 20,24\%$$

Situação 3: Qual a probabilidade de um candidato que prestou vestibular ingressar na universidade sendo que nunca realizou o processo seletivo e que possui renda total mensal familiar até 5 salários mínimos?

ln *odds* = - 2,541 + 1,170 . (0) – 1,143 . (1) = - 3,684

$$P = \frac{1}{1 + e^{-(\ln odds)}} = \frac{1}{1 + e^{-(-3,684)}} = 0,0245 = 2,45\%$$

Situação 4: Qual a probabilidade de um candidato que prestou vestibular ingressar na Universidade sendo que nunca realizou o processo seletivo e que não possui renda total mensal familiar até 5 salários mínimos?

ln *odds* = - 2,541 + 1,170 . (0) – 1,143 . (0) = - 2,541

$$P = \frac{1}{1 + e^{-(\ln odds)}} = \frac{1}{1 + e^{-(-2,541)}} = 0,0730 = 7,30\%$$

Observa-se pelas quatro situações que a situação 2 foi a que apresentou uma maior probabilidade de uma possível seleção (20,24%) e no qual esta inserida a variável “já realizou o vestibular uma ou mais vezes”. As outras situações apresentaram uma probabilidade baixa para estimar a seleção dos mesmos.

5 CONSIDERAÇÕES FINAIS

5.1Conclusões

Este trabalho teve como objetivo principal identificar as variáveis capazes de determinar a seleção dos candidatos em um processo seletivo de uma Instituição Federal do Ensino Superior, com a utilização da técnica estatística de Regressão Logística Múltipla. Para tanto, escolheu-se como objeto de estudo a Universidade Federal de Santa Maria e, obteve-se, uma amostra de 535 candidatos inscritos e que concorreram aos novos cursos de Engenharia, viabilizado pelo programa REUNI, no processo seletivo extraordinário, realizado em maio de 2009.

Os resultados obtidos neste estudo indicam que dentre as 34 variáveis socioculturais testadas somente duas foram estatisticamente significativas para determinar o modelo de seleção dos candidatos no processo seletivo, que são: se o candidato já prestou vestibular uma ou mais vezes e se possui a renda total mensal familiar até 5 salários mínimos.

A partir destas duas variáveis foi construído o modelo de seleção utilizando a técnica de regressão logística múltipla binária, no qual estimou corretamente 87,5% dos candidatos da amostra total de análise, resultado que pode ser considerado satisfatório em termos de estimação da probabilidade.

Ao analisar as relações das variáveis estatisticamente significativas no modelo com as hipóteses de seleção no processo seletivo, os sinais dos coeficientes de regressão têm um papel fundamental na interpretação dos resultados. Nesta pesquisa a variável “Já prestou vestibular uma ou mais vezes” apresentou sinal positivo (b_1 = 1,170) enquanto que outra “Possui renda total mensal familiar até 5 salários mínimos” mostrou sinal negativo (b_2 = -1,143). Isto indica que um candidato que já prestou vestibular uma ou mais vezes possui uma probabilidade maior de ser selecionado, enquanto que se ele possui uma renda total mensal familiar até 5 salários mínimos, esta probabilidade tende a diminuir, o que pode, teoricamente, indicar que se o candidato possui um baixo poder aquisitivo sua probabilidade de seleção no vestibular tende a diminuir.

Este estudo, embora não seja considerado definitivo em termos de estimação de probabilidade da seleção ao vestibular em Instituições Federais do Ensino

Superior, apresentou resultados e conclusões que estavam de acordo com a realidade, o que justifica sua aplicação.

Quanto ao perfil dos candidatos ao vestibular dos novos cursos de Engenharia da UFSM (n = 535), observou-se que a maioria é do sexo masculino (79,40%); solteiros (96,10%); com idade inferior ou igual a 25 anos (89,90%); e naturais do município de Santa Maria (61,70%). Também, percebeu-se que 88,8% concluíram o atual Ensino Médio; 60,2% cursaram o Ensino Médio a maior parte em escola pública; 91% realizaram seus estudos de Ensino Médio no turno diurno; e 66% concluíram o Ensino Médio nos últimos três anos.

Verificou-se que 57% frequentou Pré-Vestibular; que 74,2% dos candidatos acreditam que o principal fator para ter sucesso no vestibular é muito estudo pessoal; 87,5% já prestaram vestibular uma ou mais vezes; 69% nunca iniciaram um curso superior; 87,7% afirmam que a formação profissional é o principal motivo que os levou a ter interesse em ingressar em um curso superior e 86,9% dos pais aprovam a escolha profissional dos seus filhos.

Com relação à escolaridade dos familiares dos candidatos; constata-se que a maioria dos pais (42,62%) possui ensino médio e quanto às mães, 43,55% possuem ensino superior e pós-graduação.

Percebeu-se que 77% dos não trabalham e seus gastos são financiados pela família, que 52,34% apresentam uma renda total mensal familiar igual a 5 salários mínimos, o que teoricamente pode indicar um baixo poder aquisitivo dos mesmos.

Com relação aos hábitos e costumes dos candidatos, apurou-se que 95% costumam ler livros, jornais e revistas, sendo que 60,7% ocupam seu tempo com leitura informativa. Também, observou-se que 97,4% costumam assistir televisão; 68,2% utilizam uma ou mais horas de seu tempo assistindo televisão; 99,3% têm acesso a computador; 73,5% têm internet em casa e 89% praticam algum tipo de esporte.

É importante destacar que o questionário sociocultural de onde foram extraídas as variáveis para construção do modelo estatístico, não foi um instrumento elaborado para este fim e, portanto, a preparação de um instrumento com esta finalidade poderia incluir outras variáveis que talvez fossem significativas no modelo sob o ponto de vista do pesquisador.

Salienta-se, que o modelo de seleção recomendado neste trabalho não deve ser considerado como o único fator determinante para a seleção no vestibular aos

novos cursos de Engenharia da Instituição, visto que existem outros aspectos não quantitativos que podem influenciar nos resultados, como por exemplo, o fator psicológico, tais como, emocional, *stress*, alteração no comportamento, instabilidade financeira da família, ausência de estrutura familiar, Q.I, etc.

Recomenda-se, como sugestão para futuras pesquisas a aplicação do modelo de estimação em outras Instituições Federais do Ensino Superior da mesma região ou não, para que assim possa ser comparada e verificada a similaridade ou a disparidade dos resultados obtidos, e que seja obtida uma amostra de análise estratificada por curso ou por turno, investigando a hipótese da existência de classes sociais distintas para cada curso e buscando avaliar sua importância na análise de determinação de seleção de um candidato.

Também, sugere-se que seja elaborado um questionário voltado somente para os candidatos que concorrem aos novos cursos de Engenharia para que assim, possam-se obter mais informações sobre os futuros acadêmicos; e a partir da amostra analisada buscar selecionar somente os candidatos aprovados para verificar o perfil dos mesmos e assim comparar com o perfil traçado inicialmente do total da amostra.

Por fim, espera-se que este estudo possa servir como referência a trabalhos que utilizem a técnica de regressão logística múltipla binária, bem como auxiliar a Instituição a partir da identificação do perfil dos candidatos, na elaboração de projetos, adaptados ao ensino superior, que atendam as necessidades e, forneçam subsídios à permanência destes na IFES.

6 REFERÊNCIAS

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística aplicada à Administração e Economia**. São Paulo: Pioneira Thomson Learning, 2003.

ALVES, S. B. A origem do vestibular no Brasil. 2008. Disponível em: http://www.vestibular.brasilescola.com/especial/a-origem-vestibular-no-brasil.htm. Acesso em 20 jun. de 2010.

ARAÚJO, E.; CARMONA, C. U. Desenvolvimento de Modelos *Credit Scoring* com abordagem de Regressão Logística para a gestão da inadimplência de uma Instituição de Microcrédito. **Contabilidade Vista & Revista**, Vol. 18, N. 3, 2007.

BARROW, M. **Estatística para economia, contabilidade e administração**. São Paulo: Ática, 2007.

BOLFARINE, H.; BUSSAB, W. **Elementos de Amostragem**. São Paulo: Editora Blucher, 2005.

BRITO, G. A . S.; ASSAF NETO, A. **Modelo de Classificação de Risco de Crédito de Grandes Empresas**. In: SBFIN, 2005.

BRUNI, A. L. **Estatística aplicada à gestão empresarial**. São Paulo: Atlas, 2007.

CALLEGARI-JAQUES, S. **Bioestatística: princípios e aplicações**. Porto Alegre: Artmed, 2003.

COPERVES. Comissão Permanente do Vestibular. 2009. Disponível em http://coperves.proj.ufsm.br/.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada**: **para os cursos de administração, ciências contábeis e economia**. 1. ed. São Paulo: Atlas, 2009.

CRESPO, A. A. **Estatística Fácil**. 19. ed. São Paulo: Saraiva, 2009.

FIELD, A. **Descobrindo a Estatística usando SPSS**. 2. ed. Porto Alegre: Artmed, 2009.

FONSECA, J.; MARTINS, G. **Curso de Estatística**. 6. ed. São Paulo: Atlas, 2008.

FRANÇA, A. A história do vestibular. 2008. Disponível em: http://www.passeiweb.com/saiba_mais/voce_sabia/vestibular. Acesso em 20 jun. de 2010.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GUJARATI, D. N. **Econometria** 4. ed. México: McGraw-Hill Interamericana, 2004.

HAIR Jr., J.F.; BABIN, B.; MONEY, A. H.; SAMOUEL, P. **Fundamentos de métodos de pesquisa em Administração**. Porto Alegre: Bookman, 2005.

HAIR Jr., J.F. ANDERSON, R.E. ; TATHAM, R.L. ; BLACK, W.C. **Análise Multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005.

HILL, R. C. **Econometria**. 2. ed. São Paulo: Saraiva, 2003.

HOSMER, D.; LEMESHOW, S. **Applied Logistic Regression**. New York: John Wiley & Sons, 1989.

INÁCIO FILHO, G. **A monografia na Universidade**. 7. ed. Campinas: Papirus, 2004.

LAKATOS, E. M.; MARCONI, M. **Fundamentos de Metodologia Científica**. 6. ed. São Paulo: Atlas, 2008.

LOPES, L.F.; et al. **Caderno Didático: estatística geral**. 3. ed. Santa Maria: UFSM, CCNE, 2008.

MARTINHAGO, S. Descoberta de conhecimento sobre o processo seletivo da UFPR. **Dissertação de Mestrado**, Curitiba: 2005.

MEZZOMO, M. **Estudo da Mortalidade Infantil – um estudo de regressão logística múltipla**. Monografia (Especialização em Estatística e Modelagem Quantitativa). Centro de Ciências Naturais e Exatas – Universidade Federal de Santa Maria. Santa Maria, RS, Brasil, 2009).

PENHA, R. N. **Um estudo sobre regressão logística binária.** Monografia (Graduação em Engenharia da Produção). Departamento de Produção – Universidade Federal de Itajubá. Itajubá, MG, Brasil, 2002.

REUNI. Reestruturação e Expansão das Universidades Federais. 2009. Ministério da Educação. Disponível em: WWW.reuni.mec.gov.br.

RIBEIRO, C. F. Proposta de construção de um modelo econométrico para estimar a probabilidade de risco de inadimplência: uma verificação empírica na Universidade Católica de Pelotas. **Dissertação de Mestrado em Ciências Contábeis**. Universidade do Vale do Rio dos Sinos. São Leopoldo, 2008.

SAMPIERI, R. H. **Metodologia de pesquisa**. 3. ed. São Paulo: McGraw-Hill, 2006.

SELAU, L. P. Construção de modelos de previsão de risco de crédito. **Dissertação de Mestrado em Engenharia da Produção**. Escola de Engenharia - Universidade Federal do Rio Grande do Sul. Porto Alegre, 2008

SILVA, T. C.; PERIÇARO, G. A. Classificação dos candidatos ao vestibular da FECILCAM via técnicas estatísticas multivariadas. **Anais do XXXII CNMAC – Congresso Nacional de Matemática Aplicada e Computacional**. v. 2. Cuiabá, MT, 2009.

UFSM. Universidade Federal de Santa Maria. Disponível em www.ufsm.br.

ZANINI, A. Regressão Logística e redes neurais artificiais: um problema de estrutura de preferência do consumidor e classificação de perfis de consumo. **Dissertação de Mestrado em Economia Aplicada**. FEA/ Universidade Federal de Juiz de Fora, 2007.

7 ANEXOS

ANEXO A. Questionário Sociocultural do Vestibular Extraordinário da UFSM, abril de 2009.

1) Onde você reside atualmente? _____

2) Em que município você terminou ou está cursando a última série do Ensino Médio? _____

3) Que tipo de Ensino Médio você concluiu ou concluirá?

1 - Antigo Colegial (científico ou clássico)
2 - Atual Ensino Médio
3 - Profissionalizante (área agrícola, industrial, de saúde, comercial, etc.)
4 - Supletivo ou Madureza
5 - Outro

4) Como você fez ou está fazendo seus estudos de Ensino Médio ou equivalente?

1 - Todo em escola pública
2 - Todo em escola particular
3 - Maior parte em escola pública
4 - Maior parte em escola particular
5 - Supletivo ou Madureza
6 - Outro

5) Em que turno você realizou ou realizará a maior parte do Ensino Médio?
1 - Diurno
2 - Noturno

6) Em que ano você concluiu ou concluirá o Ensino Médio ou equivalente?

7) Você frequentou ou frequenta curso pré-vestibular?

1 - Não
2 - Sim, por menos de um semestre
3 - Sim, por um semestre
4 - Sim, por mais de um semestre

8) Na sua opinião, qual o principal fator para ter sucesso no vestibular?

1 - Muito estudo pessoal
2 - Um bom colégio de Ensino Fundamental e de Ensino Médio
3 - Um bom curso pré-vestibular
4 - Sorte
5 - Outro fator

9) Quantas vezes você já fez vestibular?

1 - Nenhuma
2 - Uma vez
3 - Duas vezes
4 - Três vezes ou mais

10) Você já iniciou algum curso superior?

1 - Sim, pretendo frequentar dois cursos superiores concomitantemente
2 - Sim, estou frequentando e pretendo desistir se aprovado neste Concurso Vestibular
3 - Sim, mas abandonei
4 - Sim, já concluí
5 - Sim, já concluí e estou cursando outro curso
6 - Sim, já concluí um e abandonei outro
7 - Não

11) Qual o principal motivo que o levou a ter interesse em ingressar num curso superior?

1 - Formação profissional
2 - Aquisição de cultura geral
3 - Aquisição de conhecimentos que permitem compreender melhor o mundo
4 - Formação acadêmica para aprofundamento da atividade prática que já desempenho
5 - Aquisição de conhecimento teórico voltado para a pesquisa

12) Qual o principal motivo que o levou a optar pelo curso em que está se inscrevendo?

1 - Pequena concorrência às vagas
2 - Bom nível de exigência do curso
3 - Possibilidade de ganho financeiro considerável
4 - Facilidade de mercado de trabalho
5 - Prestígio social da profissão
6 - Atendimento às minhas aptidões e interesses
7 - Influência de familiares
8 - Aperfeiçoamento de meu exercício profissional atual
9 - Compatibilidade de horário com minha atual profissão
10 - Outro motivo

13) Qual a posição de seus pais diante de sua escolha profissional?

1 - Aprovam
2 - Não aprovam
3 - São indiferentes
4 - Não conhecem a minha escolha
5 - Outra

14) Qual a escolaridade de seu pai?

1 - Ensino Fundamental ou equivalente incompleto
2 - Ensino Fundamental ou equivalente completo
3 - Ensino Fundamental completo
4 - Ensino Médio ou incompleto
9 - Ensino Médio ou completo
5 - Ensino Superior
6 - Sem escolaridade
7 - Pós-Graduação
8 - Outra situação

15) Qual a escolaridade de sua mãe?

1 - Ensino Fundamental ou equivalente incompleto
2 - Ensino Fundamental ou equivalente completo
3 - Ensino Fundamental completo
4 - Ensino Médio ou incompleto
9 - Ensino Médio ou completo
5 - Ensino Superior
6 - Sem escolaridade
7 - Pós-Graduação
8 - Outra situação

16) Qual a ocupação principal de seu pai?

1 - Proprietário de grande empresa (acima de 100 empregados)
2 - Proprietário de média empresa (de 10 a 100 empregados)
3 - Proprietário de microempresa
4 - Administrador de grande empresa
5 - Administrador de média empresa
6 - Administrador em serviço público, banco, repartição
7 - Profissional liberal
8 - Professor do ensino superior
9 - Professor do ensino fundamental e/ou médio
10 - Profissional ou técnico de nível médio
11 - Militar (federal ou estadual)
12 - Trabalhador na área de indústria
13 - Trabalhador na área de comércio
14 - Trabalhador na área de agropecuária
15 - Trabalhador na área de prestação de serviço
16 - Artesão, cabeleireiro, manicure e outras profissões afins
17 - Funcionário público
18 - Atleta ou artista
19 - Dona de casa
20 - Aposentado
21 - Outras
22 - Não trabalha

17) Qual a ocupação principal de sua mãe?

1 - Proprietária de grande empresa (acima de 100 empregados)
2 - Proprietária de média empresa (de 10 a 100 empregados)
3 - Proprietária de microempresa
4 - Administradora de grande empresa
5 - Administradora de média empresa
6 - Administradora em serviço público, banco, repartição
7 - Profissional liberal
8 - Professora do ensino superior
9 - Professora do ensino fundamental e/ou médio
10 - Profissional ou técnica de nível médio
11 - Militar (federal ou estadual)
12 - Trabalhadora na área de indústria
13 - Trabalhadora na área de comércio
14 - Trabalhadora na área de agropecuária
15 - Trabalhadora na área de prestação de serviço
16 - Artesã, cabeleireira, manicure e outras profissões afins
17 - Funcionária pública
18 - Atleta ou artista
19 - Dona de casa
20 - Aposentada
21 - Outras

18) Qual a sua participação na vida econômica da família?

1 - Não trabalho e meus gastos são financiados pela família ou por outras pessoas
2 - Trabalho, mas recebo ajuda financeira da família ou de outras pessoas
3 - Trabalho e sou responsável pelo meu sustento
4 - Trabalho e sou responsável pelo meu sustento e auxilio a família ou outras pessoas
5 - Trabalho e sou o principal responsável pelo sustento da família

19) Qual a renda total mensal, em salários mínimos, da família? (Se for solteiro, inclua rendimentos seus, de seus pais, de seus irmãos e de outras pessoas que contribuam para a renda da família; se for casado, os rendimentos, do cônjuge, filhos e de outras pessoas que contribuam para a renda familiar.)

1 - Até 01 salário mínimo
2 - De 01 até menos de 02 salários mínimos
3 - De 02 até menos de 03 salários mínimos
4 - De 03 até menos de 04 salários mínimos
5 - De 04 até menos de 05 salários mínimos
6 - De 05 até menos de 07 salários mínimos
7 - De 07 até menos de 09 salários mínimos
8 - De 09 até menos de 11 salários mínimos
9 - De 11 até menos de 13 salários mínimos
10 - De 13 até menos de 15 salários mínimos
11 - De 15 até menos de 17 salários mínimos

12 - 17 ou mais Salários Mínimos

20) Como você pretende se manter durante o curso?

1 - Com recursos familiares
2 - Com recursos do meu trabalho
3 - Com auxílio de bolsa de estudos
4 - Com recursos ainda não definidos

21) Qual a situação do imóvel em que você reside?

1 - Próprio
2 - Alugado
3 - Financiado
1 - Próprio

22) Que fator mais o influenciou na escolha desta Universidade?

1 - Tem bom nível de ensino
2 - É a única que oferece o curso pretendido
3 - É gratuita
4 - Há menor concorrência às vagas
5 - Fica geograficamente próxima à minha residência
6 - Outro fator

23) De que atividade você mais gosta de participar?

1 - Social
2 - Artística-cultural
3 – Esportiva
4 - Religiosa
5 - Política-partidária

24) Além das tarefas escolares, você costuma ler livros, jornais e revistas?

1 - Quase todos os dias
2 - Uma vez por semana
3 - Uma vez por mês
4 – Raramente
5 - Não leio

25) Qual o tipo de leitura que mais ocupa seu tempo?

1 – Informativa
2 - Recreativa (esporte, moda, sexo, outra)
3 - Literária
4 - Técnica
5 - Não leio

26) Em média, que tempo você utiliza para assistir à televisão?

1 - Mais de duas horas por dia
2 - Entre uma e duas horas por dia
3 - Em torno de uma hora por dia
4 - Menos de uma hora por dia
5 - Só eventualmente
6 - Não assisto à televisão

27) A que tipo de programa você mais assiste na televisão?

1 - Novela
2 - Noticiário
3 - Filme
4 - Programa cultural (reportagem, entrevista)
5 – Esporte
6 - Outro
7 - Não assisto à televisão

28) Que tipo de música você prefere?

1 - Romântica
2 - Popular Brasileira (MPB)
3 - Sertaneja
4 - Tradicionalista
5 - Clássica
6 - Rock
7 - Outra

29) Que tipo de programa de rádio você prefere?

1 - Noticiário
2 - Esportivo
3 - Informativo-cultural
4 - Musical
5 – Variedades
6 - Outro
7 - Não escuto rádio

30) Onde você tem acesso a microcomputador?

1 - Em casa
2 - No trabalho
3 - Em casa e no trabalho
4 - No shopping
5 - Na casa de amigos
6 - Não tenho acesso

31) Onde você tem acesso à Internet (Rede mundial de computadores)?

1 - Em casa
2 - No trabalho
3 - Em casa e no trabalho
4 - No shopping
5 - Na casa de amigos
6 - Não tenho acesso à Internet

32) Qual dos esportes relacionados a seguir você costuma praticar?

1 - Futebol
2 - Natação
3 - Basquete
4 - Vôlei
5 - Artes marciais
6 – Ginástica
7 - Outro
8 - Não pratico esportes

33) Língua estrangeira escolhida:

1 - Inglês
2 - Espanhol

34) Sexo:

1 - Masculino
2 - Feminino

35) Estado Civil:

1 - Solteiro
2 – Casado
3 – Outros

36) Idade: _____

37) Nome do Pai:_____

38) Nome da Mãe: _____

39) Cor/Raça: _____