UNIVERSIDADE DE PASSO FUNDO
FACULDADE DE CIÊNCIAS ECONÔMICAS, ADMINISTRATIVAS E CONTÁBEIS
CENTRO DE PESQUISA E EXTENSÃO DA FEAC
(www.upf.br/cepeac)

Texto para discussão

Texto para discussão Nº 03/2019

APOSTILA DE REGRESSÃO LINEAR MÚLTIPLA COM O USO DO SPSS

Amanda Regina Leite
Rangel Matos
Nadiesca Manica

Passo Fundo - RS - Brasil

PPGAdm

Programa de Pós-Graduação em Administração

Faculdade de Ciências Econômicas, Administrativas e Contábeis - Feac

Regressão Linear Múltipla

Apostila de Regressão Linear Múltipla com o uso do SPSS

Amanda Regina Leite
Rangel Matos
Nadiesca Manica

Universidade de Passo Fundo

SUMÁRIO

INTRODUÇÃO

O método de regressão, a estimação de parâmetros de uma equação, é cada vez mais divulgado e utilizado para resolver problemas nos estudos do campo da Administração e na prática das empresas. A regressão demonstra quantitativamente a força atrás de uma causalidade ou um simples relacionamento que ocorre de X_t para Y_t. Nesse sentido, Y_t é a variável dependente da variável X_t, denominada variável independente.

Quando queremos analisar a relação de dependência entre duas variáveis, em que uma assume o papel de dependente e outra de independente, rodamos uma Regressão Linear Simples (RLS). No entanto, ao definirmos que uma variável Y_t. relaciona-se entre mais de duas variáveis de X_t, em que duas ou mais variáveis assumem o papel de independentes, precisamos calcular uma Regressão Linear Múltipla.

A partir disso, esta apostila tem por objetivo explicar teoricamente e, também, de forma prática como realizar a análise de dados por meio da **Regressão Linear Múltipla** em um software SPSS. O SPSS é um software específico para cálculos estatísticos, muito comum àqueles que realizam análises estatísticas com mais frequência. Para isso, apresentaremos um exemplo prático, o qual será utilizado ao longo de toda a apostila.

INTRODUZINDO A REGRESSÃO MÚLTIPLA

A regressão múltipla envolve três ou mais variáveis, portanto, estimadores. Ou seja, ainda uma única variável dependente, porém duas ou mais variáveis independentes (explanatórias). A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. Isto é, reduzir o coeficiente do intercepto, o qual, em regressão, significa a parte da variável dependente explicada por outras variáveis, que não a considerada no modelo. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar Y, efetuando uma análise de regressão múltipla, por 2 razões:

a) Para reduzir os resíduos estocásticos. Reduzindo-se a variância residual (ERRO PADRÃO DA ESTIMATIVA), aumenta a força dos testes de significância;
b) Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

Em geral, uma variável dependente Y depende de várias variáveis independentes (x1, X2, …, Xk). Na análise de regressão múltipla, procuramos construir um modelo estatístico-matemático para se estudar, objetivamente, a relação entre as variáveis independentes e a variável dependente e, a partir do modelo, conhecer a influência de cada variável independente, como também, predizer a variável dependente em função do conhecimento das variáveis independentes. veja o Quadro 1.

Quadro – 1 Exemplos de aplicação da regressão linear múltipla

Variáveis independentes ($X_1, X_2, ..., X_k$)	Variável dependente Y
X_1 = renda (R$) X_2 = poupança (R$) X_3 = taxa de juros (%)	Y = consumo
X_1 = memória RAM (Gb) X_2 = sistema operacional X_3 =tipo de processador	Y = tempo de resposta do sistema computacional (segundos)
X_1 = área construída do imóvel (m²) X_2 = padrão de qualidade (custo do m²) X_3 = localização	Y = preço de um imóvel novo (R$)
X_1 = valor do modelo novo (R$) X_2 = quilometragem X_3 = idade do veículo X_4 = estado de conservação X_5 = opcionais	Y = valor de revenda de carro seminovo (R$)

Fonte: Barbetta, Reis, Bornia (2010)

2 Definição

Modelos multivariados de pesquisa envolvem análise do relacionamento entre múltiplas variáveis explicativas e, em alguns casos, múltiplas variáveis dependentes. Grande parte das pesquisas delineadas para examinar o efeito exercido por duas ou mais variáveis independentes sobre uma variável dependente utiliza a análise de Regressão Múltipla.

A Regressão Múltipla (RM) é definida por Tabachnick e Fidell (1996) como um conjunto de técnicas estatísticas que possibilita a avaliação do relacionamento de uma variável dependente com diversas variáveis independentes. Essas técnicas são muito úteis nas pesquisas da área de ciências sociais aplicadas, onde grande parte dos estudos envolve variáveis independentes correlacionadas entre si. Para Dunlap e Landis (1998), o uso de preditores redundantes, correlacionados entre si, é uma característica dos estudos da Psicologia que, ao construírem medidas internamente consistentes, incluem múltiplas medidas correlacionadas ao mesmo construto, o que pode levar, em alguns casos, à exclusão de variáveis importantes na explicação da variável em foco. Nestes casos, é mais segura a utilização de técnicas estatísticas como a RM. Embora esta técnica seja sensível à natureza redundante dos preditores, suas limitações já são bastante conhecidas, como, por exemplo, a sua sensibilidade ao erro Tipo II (Dunlap & Landis, 1998). Para facilitar a discussão sobre as aplicações e problemas relacionados ao uso da RM, são apresentadas algumas definições relacionadas à técnica.

O resultado final de uma RM é uma equação da reta que representa a melhor predição de uma variável dependente a partir de diversas variáveis independentes. Esta equação representa um modelo aditivo, no qual as variáveis preditoras somam-se na explicação da variável critério. A equação da regressão linear pode ser representada por: “y = a + bxi + Î”, onde: “y” é a variável dependente, ou critério; “a” é a constante, ou o intercepto entre a reta e o eixo ortogonal; “b” é o parâmetro, coeficiente padronizado de regressão, ou peso; “xi” são as variáveis independentes (preditoras) e “Î” é o erro ou resíduo, que se refere à diferença entre os valores observados e preditos.

Para que o uso desta equação seja eficaz na predição da variável dependente em estudo, o pesquisador deve examinar previamente os pressupostos da RM, bem como identificar as conseqüências da sua violação. Entre os pressupostos citados por Tabachnick e Fidell (1996), estão: (1) a multicolinearidade, (2) a singularidade, (3) a homogeneidade nas variâncias, (4) a normalidade e (5) a linearidade.

Embora seja imprescindível que o pesquisador examine esses pressupostos antes de iniciar suas análises, nota-se que a RM é um modelo eficaz contra a violação de grande parte dos pressupostos. Por exemplo, no caso da inclusão de variáveis multicolineares ou singulares nas análises, o pesquisador estará perdendo graus de liberdade, o que conseqüentemente reduz o poder estatístico de suas conclusões. O pesquisador pode, ainda, estar excluindo de seu modelo de estudo variáveis importantes para a explicação do fenômeno em questão, as quais podem estar correlacionadas com uma variável multicolinear.

A violação do pressuposto de normalidade pode ser atenuada por meio do aumento do tamanho da amostra da população pesquisada. Esse aumento, além de afetar todos os parâmetros da equação, poderá também reduzir os problemas advindos da violação desse pressuposto. Segundo o teorema do limite central, quanto maior a amostra, maior a chance de que as distribuições das médias das variáveis envolvidas estejam normalmente distribuídas, apesar de não terem individualmente o formato normal. Logo, aumentando-se o tamanho da amostra, os efeitos da não-normalidade das variáveis são reduzidos, aumentando a robustez da análise, e tornando menos necessária a transformação dessas variáveis (Tabachnick & Fidell, 1996).

Quando o pressuposto da linearidade é violado, o pesquisador deve estar ciente de que o modelo de regressão linear não é o melhor modelo explicativo para o estudo das variáveis envolvidas, e que outros modelos (e.g. o quadrático) devem ser utilizados. Finalmente, a violação do pressuposto da homogeneidade das variâncias não, necessariamente, invalida a análise, a depender da sua finalidade, mas a enfraquece. A heterogeneidade das variâncias, ou violação da homogeneidade das variâncias, pode ser reduzida por intermédio da transformação de variáveis que não possuem distribuição normal (e.g., assimetria positiva ou negativa).

É necessário ressaltar que a qualidade do modelo de investigação adotado pelo pesquisador pode ser avaliada por meio do valor do coeficiente de determinação, R2, e da distribuição dos resíduos. Tomando como base uma equação de regressão linear (y = a + bxi + Î), diz-se, por exemplo, que um R2 = 0,401 significa que o(s) preditor(es) explica(m) 40% da variância de y. Em outras palavras, o R2 é a quantidade da variância da variável dependente que é explicada conjuntamente pela(s) variável(is) independente(s) e é a estatística mais utilizada para interpretar os resultados da regressão (Tabachnick & Fidell, 1996).

Como observado anteriormente, a regressão permite verificar o quanto cada variável preditora aumenta o poder explicativo da equação de regressão (DR2). Na equação de RM, obtém-se um coeficiente de correlação, o parâmetro b (ou peso padronizado), que representa a magnitude do relacionamento entre cada um dos preditores e o critério, sendo que sua interpretação depende do conhecimento dos erros padrões ele associados (Dunlap & Landis, 1998). O valor de b é influenciado por todas as variáveis preditoras incluídas na equação e está sujeito a mudanças em sua magnitude, dependendo do conjunto de preditores investigados. na próxima seção serão discutidas algumas das principais aplicações deste conjunto de técnicas.

3 Aplicações da Regressão Linear Múltipla

A RM é uma análise estatística muito utilizada em pesquisas em Ciências Sociais para investigar questões referentes: (1) ao grau de relacionamento entre as variáveis, indicando se uma correlação é significativamente diferente de zero; (2) à importância relativa das variáveis preditoras na explicação da variável dependente; (3) à magnitude do aumento da correlação múltipla resultante da adição de uma ou mais variáveis na equação; (4) à maneira pela qual uma variável independente se comporta no contexto de outra(s) variável(is); (5) à natureza do relacionamento entre as variáveis independentes e dependentes, indicando se o relacionamento é linear ou não-linear (e.g., valores quadrados, cúbicos, produtos cruzados entre as variáveis); (6) à comparação entre conjuntos diferentes de variáveis independentes na predição da variável dependente; (7) ao cálculo estimativo dos escores da variável dependente para os membros de uma nova amostra ainda não pesquisada; e (8) à identificação de relacionamentos causais entre variáveis quando aplicada como um caso especial de path analysis ou equação estrutural.

No caso do uso da regressão para a finalidade mencionada no item 5 (verificar se o relacionamento entre as variáveis é linear ou não), a RM pode ser empregada na identificação de variáveis mediadoras e moderadoras. Em função da grande difusão desta estratégia nas pesquisas internacionais (KROMREY; FOSTER-JOHNSON, 1999) e na área das Ciências Sociais, os fenômenos de mediação e moderação serão definidos a seguir, assim como a forma pela qual a regressão pode ser utilizada para identificá-los. Mediação: o conceito de mediação implica suposição de relacionamentos causais entre as variáveis envolvidas. Uma variável mediadora é aquela que, ao estar presente na equação de regressão, diminui a magnitude do relacionamento entre uma variável antecedente e uma variável dependente ou critério. Para melhor ilustrar a definição de uma variável mediadora, podemos analisar o relacionamento entre três variáveis hipotéticas, sendo a variável B a mediadora do relacionamento de A com C (A è B è C). Note-se que a relação entre as variáveis A e C ficará enfraquecida na presença da variável B. No caso de uma variável mediadora pura, o relacionamento entre A e C deixa de existir na presença da variável B.

Segundo Tabachnick e Fidell (1996) e Keppel (1991), a identificação de variáveis mediadoras pode ser feita, por exemplo, com base na observação dos padrões assumidos pelos pesos b das variáveis envolvidas. No caso de uma variável mediadora pura, tem-se um b significativo de A para C, antes da entrada de B na equação. Contudo, uma vez que B é adicionado à equação, o b de B torna-se significativo, enquanto a significância do b de A desaparece. No caso de uma mediação pura, o B captura totalmente a relação entre A e C. Contudo, quando a mediação não é total, pode ainda existir uma relação entre A e C mesmo na presença de B.

4 Definição das variáveis

Para construir um modelo complexo com um conjunto de variáveis independentes, segundo Field (2009), deve se ter muito cuidado ao selecionar as variáveis, pois os valores dos **coeficientes de regressão** (R^2) dependem delas. A escolha das variáveis que serão coletadas parte, em primeiro lugar, da teoria. Após apreender o que outros autores da mesma área do conhecimento estão discutindo sobre o seu tema e definir quais variáveis serão analisadas na sua RLM, é preciso escolher qual delas será a variável dependente e quais serão as independentes.

Desta forma, as variáveis independentes incluídas e a forma com que elas são inseridas na RLM podem ter um grande impacto. Num mundo ideal, as variáveis independentes deveriam ser selecionadas baseadas em pesquisas anteriores. Não se deve de forma alguma selecionar centenas de variáveis independentes ao acaso, juntá-las todos em uma análise de regressão e torcer pelo melhor.

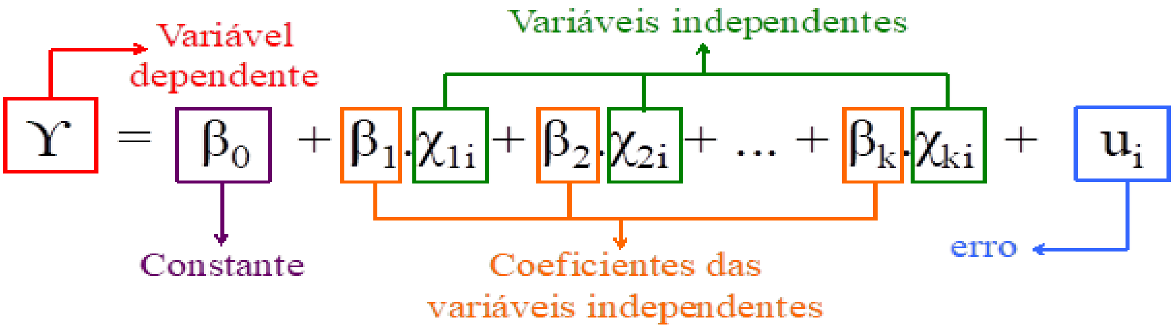
5 O Modelo Matemático

Na análise de regressão, ajustamos um modelo preditivo aos nossos dados e então usamos esse modelo para prever valores da variável dependente (VD) a partir de uma ou mais variáveis independentes (VIs).1 A regressão simples procura prever uma variável de saída a partir de um única variável previsora, enquanto que a regressão múltipla busca prever um resultado a partir de diversas variáveis previsoras. Essa é uma ferramenta bastante útil porque nos permite ir um passo além dos dados que de fato temos. Na Seção seguinte é apresentada a ideia de que podemos prever qualquer valor utilizando a seguinte equação genérica:

$$\text{Saída}_1 = (\text{Modelo}_i\ ) + \text{Erro}_1$$

Isso significa apenas que a saída ou resposta que estamos tentando prever para uma determinada pessoa pode ser prevista por qualquer modelo que ajustarmos aos dados mais algum tipo de erro. Na regressão, o modelo que ajustamos é linear ("modelo linear" significa "modelo baseado em uma linha reta") e você pode imaginar que está tentando resumir um conjunto de dados com uma linha reta. Assim, a palavra "modelo" na equação acima pode ser substituída por algo que defina a linha que ajustamos aos dados. Com qualquer conjunto de dados existem várias linhas que podem ser utilizadas para resumir a tendência geral e, desse modo, precisamos encontrar uma forma de decidir entre as muitas possibilidades. O modelo de regressão linear múltipla é apresentado na Figura 1.

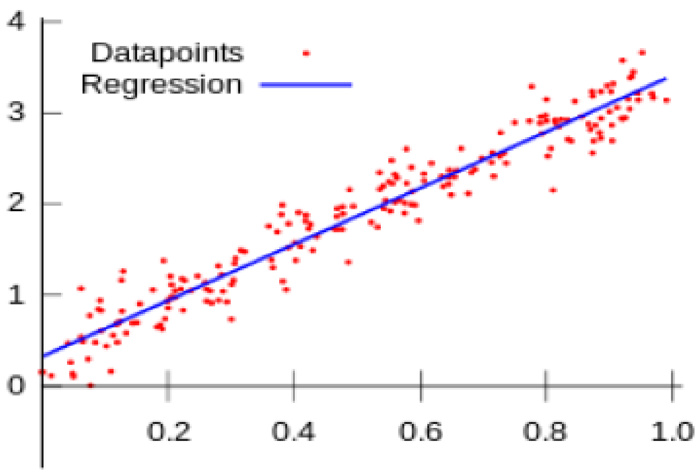
Figura 1 – Modelo de regressão linear múltipla



Fonte: Manosso, Fossati, Berti (2018)

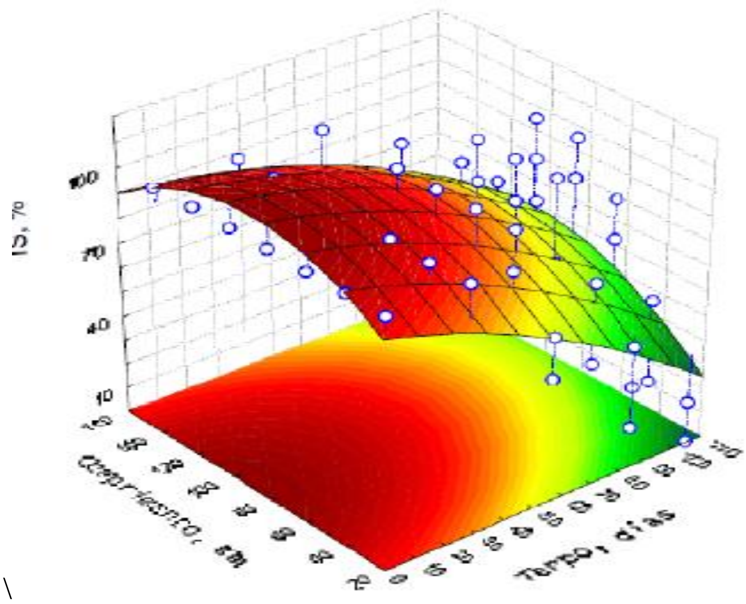
Em que Y representa a variável resposta e x_2 representam as variáveis explicativas (o emissor de tempo e o emissor de dose, respectivamente) e ε representa o erro experimental. Esse é um modelo de regressão linear múltipla com duas variáveis independentes ou explicativas (x_1 e x_2). O termo linear é usado pois a equação é uma função linear de parâmetros desconhecidos β_0, β_1e β_2, denominados coeficientes da regressão. A regressão simples resulta numa reta, conforme ilustrado na Figura 2.

Figura 2 – Reta regressão simples



Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis implica num plano, e um problema de k variáveis implica em um hiperplano, conforme a Figura 3.

Figura 3 – Reta regressão simples



\

Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados Yi e os valores ajustados Yc.

Para que seja possível tirar conclusões precisas, é necessário ajustar o melhor modelo que descreve os dados. Existem muitas maneiras de ajustar uma linha reta aos dados que foram coletados. A mais simples é utilizar seus olhos para imaginar uma linha que pareça resumir bem os dados. Entretanto, esse método é bastante subjetivo e não assegura que o modelo é o melhor que se poderia ter escolhido. Portanto, utiliza-se técnicas matemáticas para determinar a linha que melhor descreve os dados observados. Esse método é denominado método dos mínimos quadrados.

6 Interpretação da regressão

Na regressão simples:

b = aumento em Y, decorrente de um aumento unitário em X.

Na regressão múltipla:

b_i = aumento em Y se X_i for aumentado de 1 unidade, mantendo-se constantes todas as demais variáveis X_i.

7 *Regressão e análise da variância (ANOVA)*

Há 3 casos principais de aplicação da regressão múltipla:

a) Regressão "padrão": é a regressão somente sobre valores numéricos.
b) Análise da variância (ANOVA): equivale somente à regressão sobre variáveis mudas.
c) Análise da covariância (ANOCOVA): é a regressão sobre variáveis mudas e variáveis numéricas.

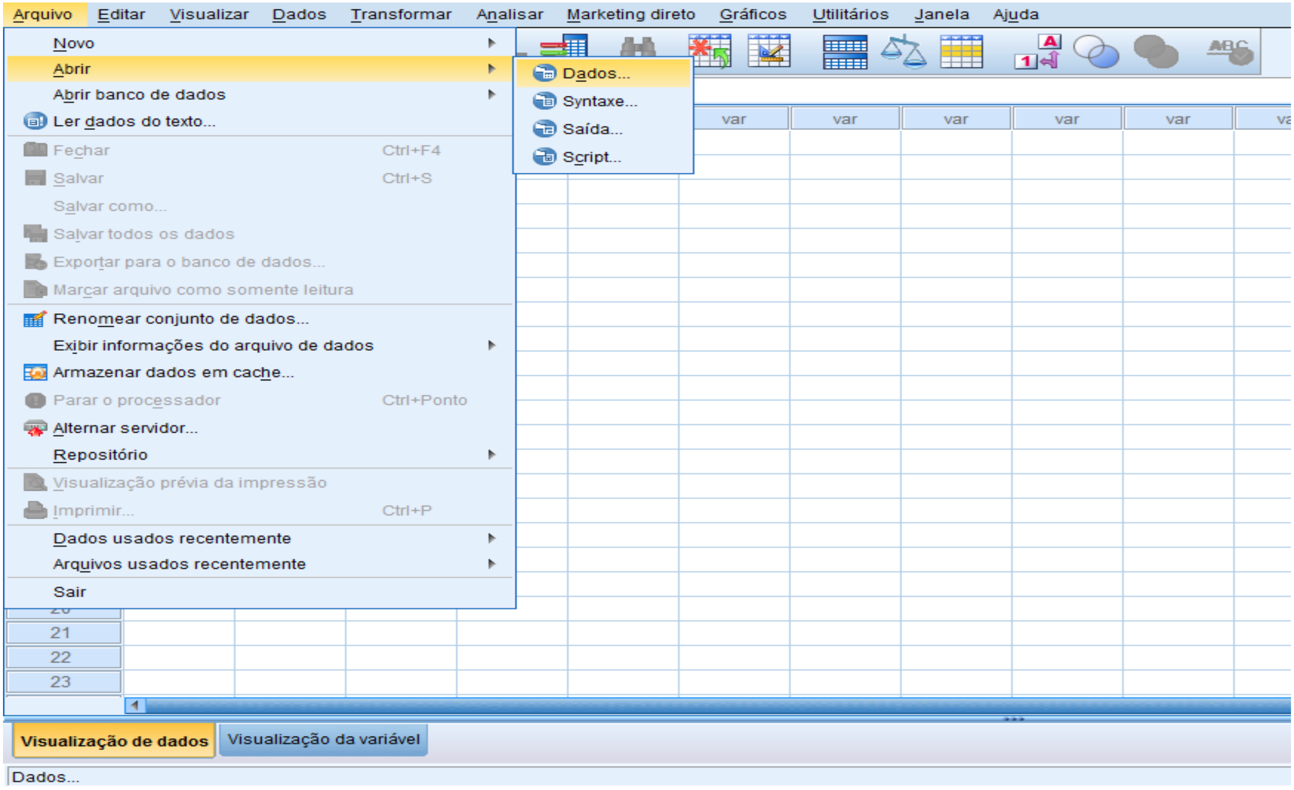
Em resumo, a regressão padrão é o instrumento mais poderoso quando a variável independente, X, é numérica. Já a análise da variância é adequada quando a variável independente é um conjunto de categorias não-ordenadas.

Executando a RLM no SPSS

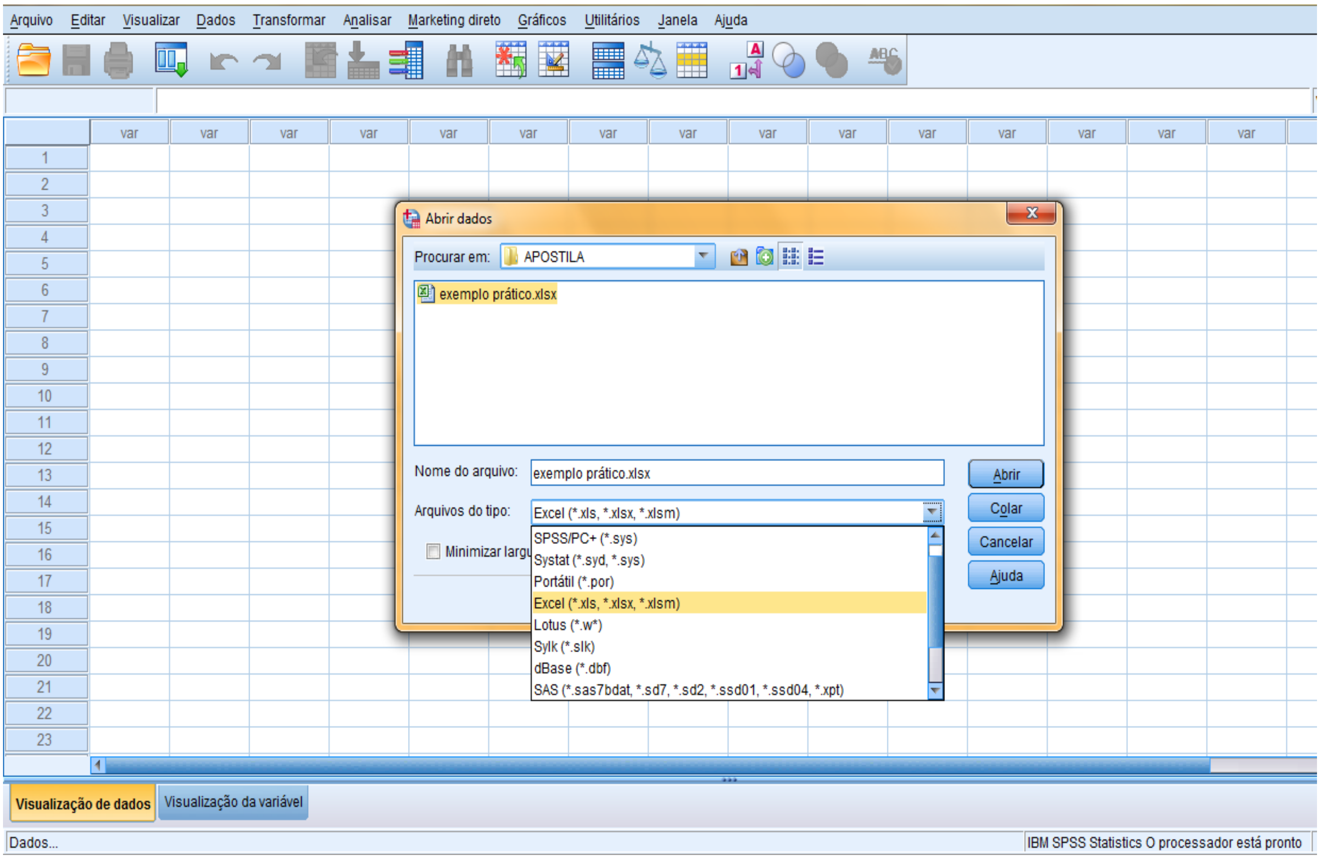
O SPSS é um software criado pela IBM para análises estatísticas nas Ciências Sociais. Por isso, vamos utiliza-lo para resolver um exercício prático sobre análise linear múltipla.

O primeiro passo é digitar uma base de dados no SPSS, ou então, se você já tem uma base de dados digitada em outro programa, por exemplo, o Excel, é necessário somente importa-la conforme a figura X.

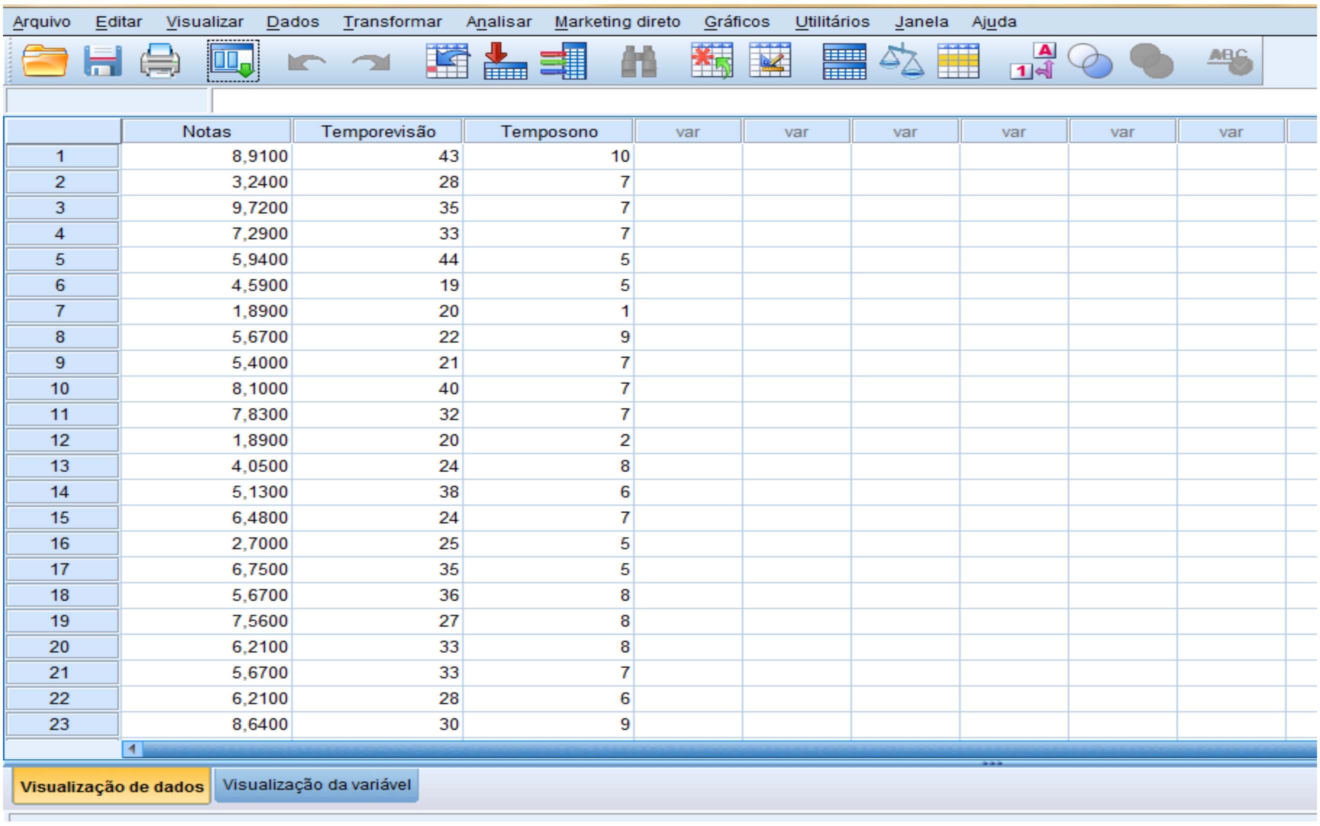
Primeiro clicar em arquivo > Abrir > Dados.



Na caixa de diálogo, em “Arquivos do tipo”, selecione a extensão em que o arquivo original está (SPSS, Excel, SAS ou texto) > Abrir.



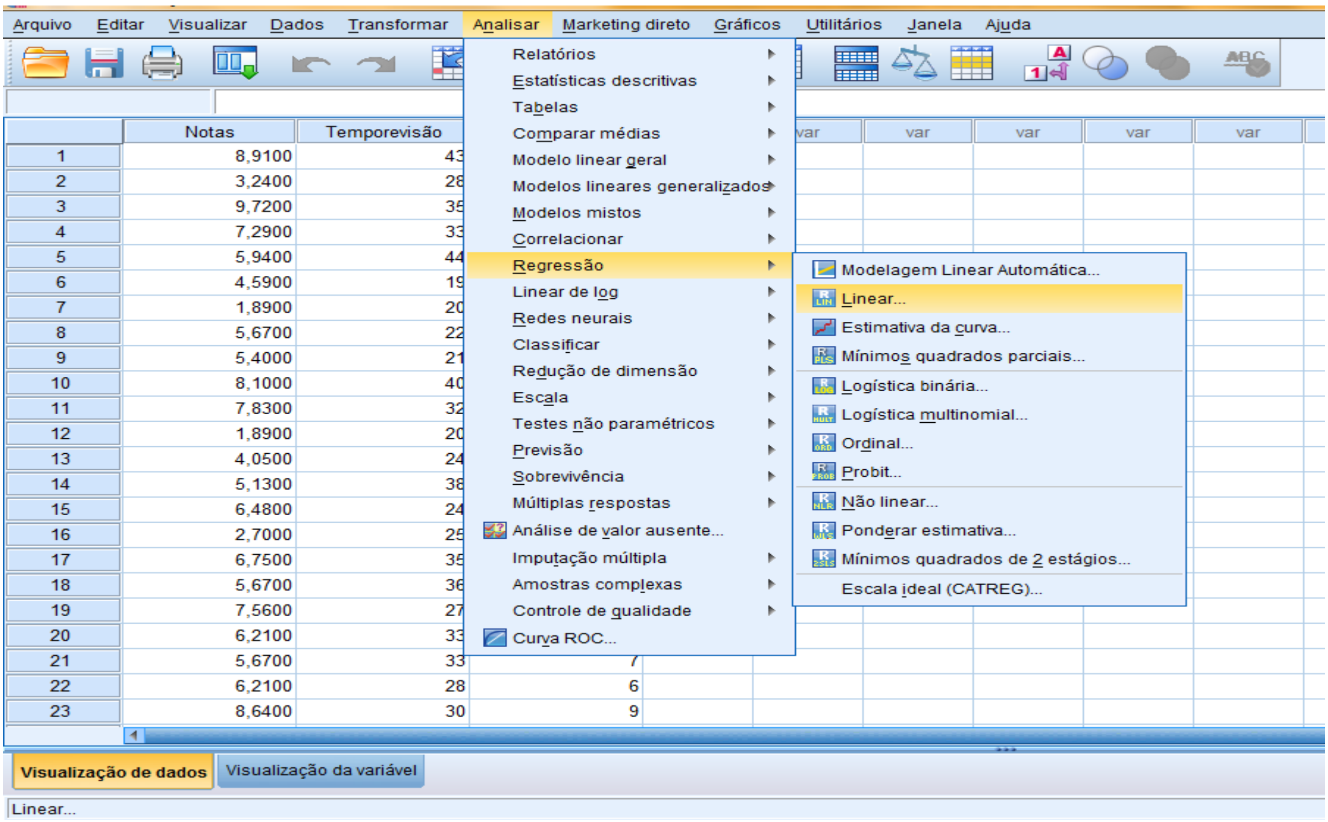
Pronto, o arquivo está apto para trabalhar.



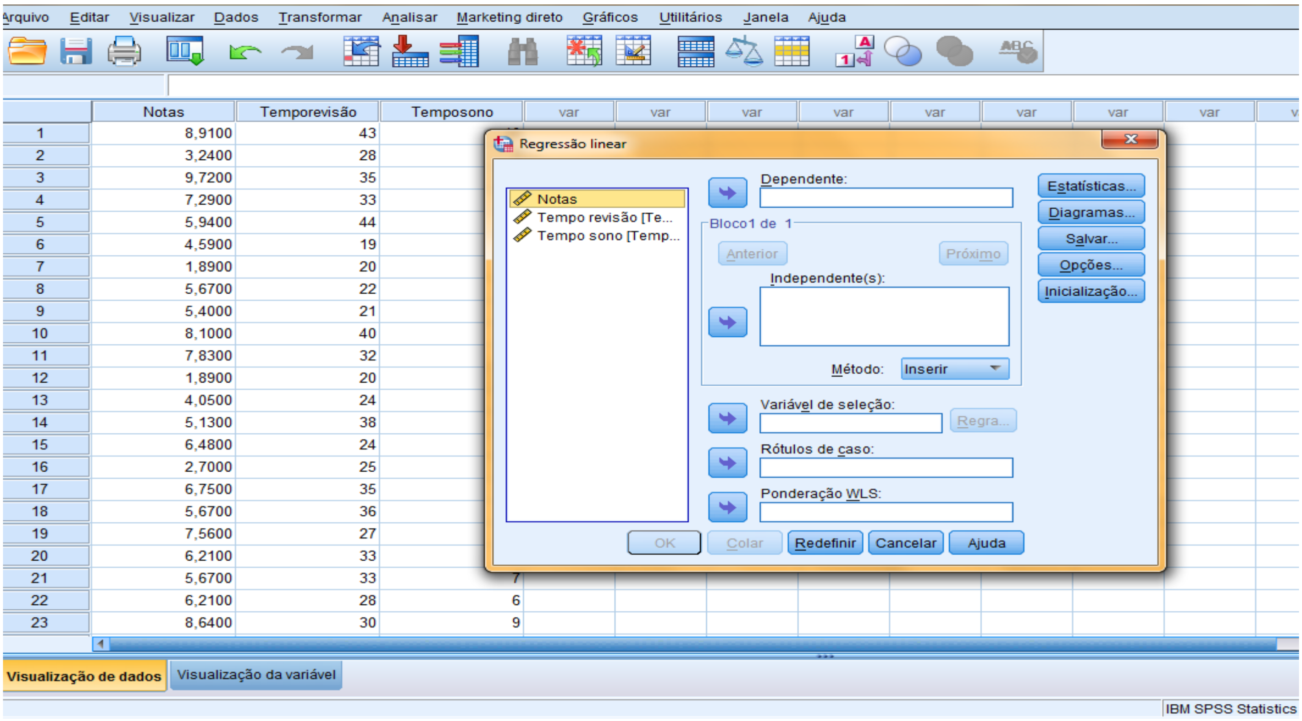
	Notas	Temporevisão	Temposono
1	8,9100	43	10
2	3,2400	28	7
3	9,7200	35	7
4	7,2900	33	7
5	5,9400	44	5
6	4,5900	19	5
7	1,8900	20	1
8	5,6700	22	9
9	5,4000	21	7
10	8,1000	40	7
11	7,8300	32	7
12	1,8900	20	2
13	4,0500	24	8
14	5,1300	38	6
15	6,4800	24	7
16	2,7000	25	5
17	6,7500	35	5
18	5,6700	36	8
19	7,5600	27	8
20	6,2100	33	8
21	5,6700	33	7
22	6,2100	28	6
23	8,6400	30	9

O segundo passo para fazer a regressão linear múltipla é:

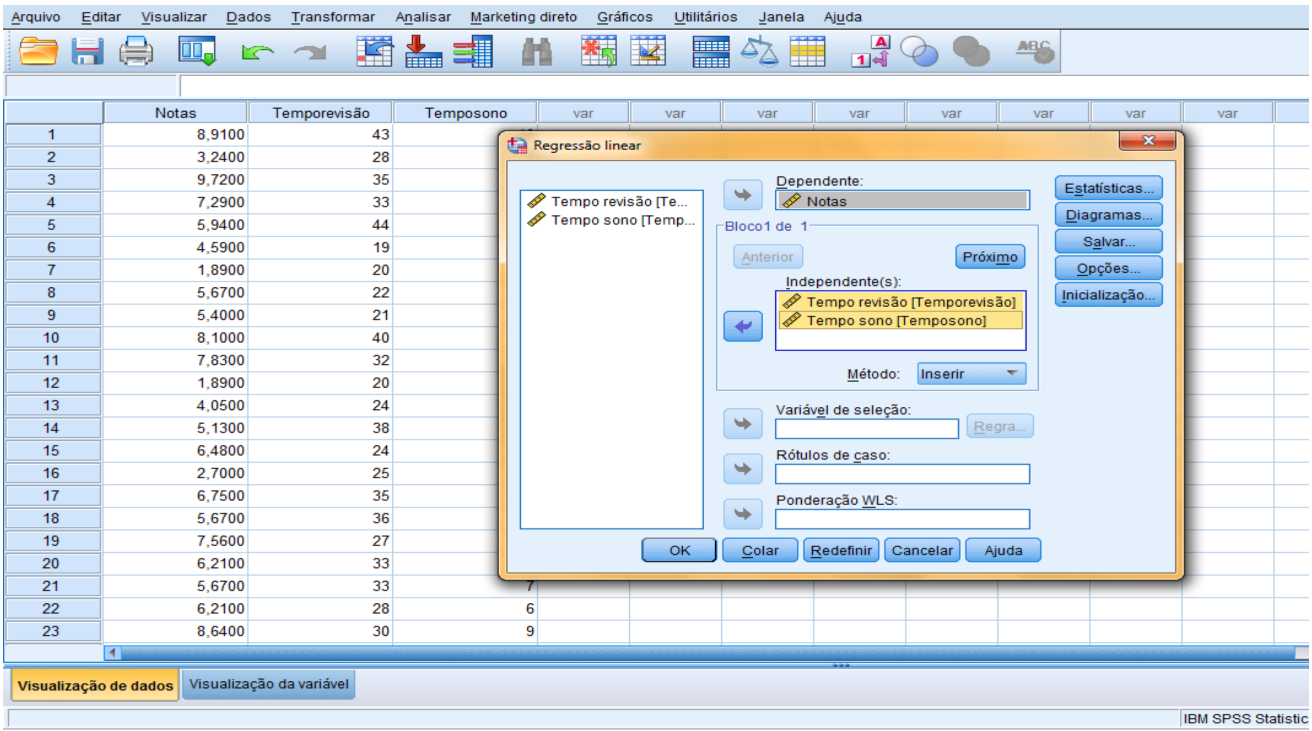
Clicar em ANALISAR > REGRESSÃO > LINEAR



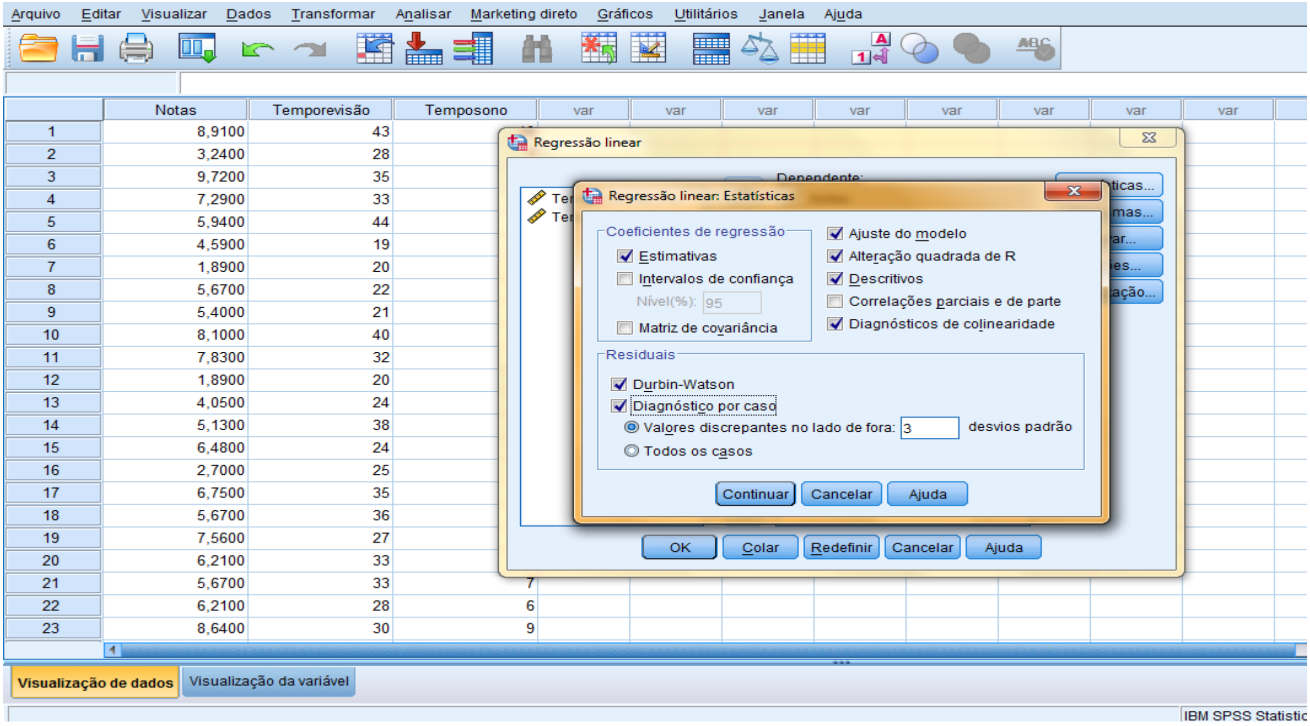
Abrirá uma janela com espaço para colocar quais são as variáveis independentes e qual é a variável dependente:



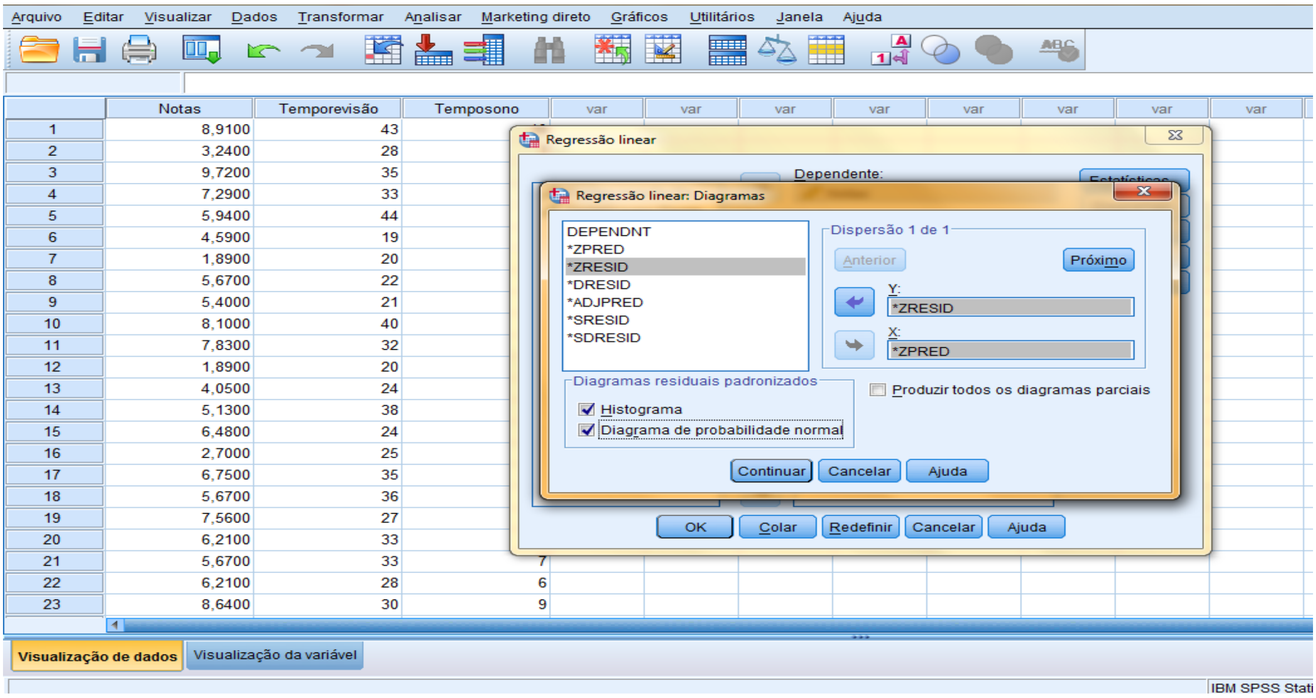
Aqui se coloca a variável dependente e as variáveis independentes em seus respectivos lugares, conforme mostra a figura abaixo.



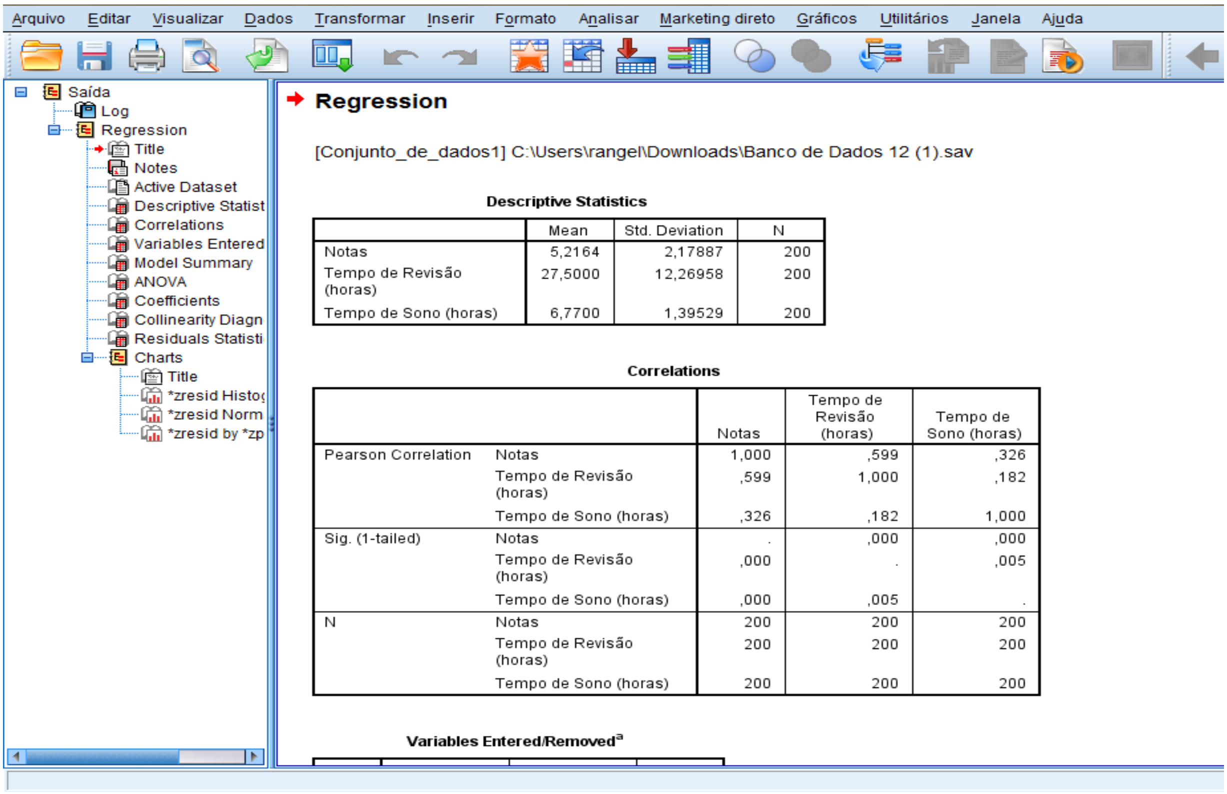
O próximo passo é clicar em estatísticas, onde abrirá uma nova janela, e nesta janela é importante marcar as opções de interesse do pesquisador. Neste exemplo serão marcadas as seguintes opções: estimativa > ajuste do módulo > alteração quadrada de R > descritivos > diagnósticos de colinearidade > durbin-watson > diagnóstico por caso > valores discrepantes e depois é só clicar em continuar, conforme mostra a figura a seguir.



Depois disso, é interessante assinalar a opção diagrama, assim será possível fazer um diagrama dos valores previstos padronizados (clicar em ZPRED e flecha que passa para a caixa X), e depois os resíduos padronizados (clicar em ZRESID e frecha que passa para a caixa Y). Por fim clicar em histograma > diagrama de probabilidade normal, pois é importante para verificar a normalidade dos resíduos. Após clicar em continuar e depois OK. Segue figura:



Após fazer isso, abrirá uma janela com várias informações conforme mostra a figura abaixo:



Descriptive Statistics

	Mean	Std. Deviation	N
Notas	5,2164	2,17887	200
Tempo de Revisão (horas)	27,5000	12,26958	200
Tempo de Sono (horas)	6,7700	1,39529	200

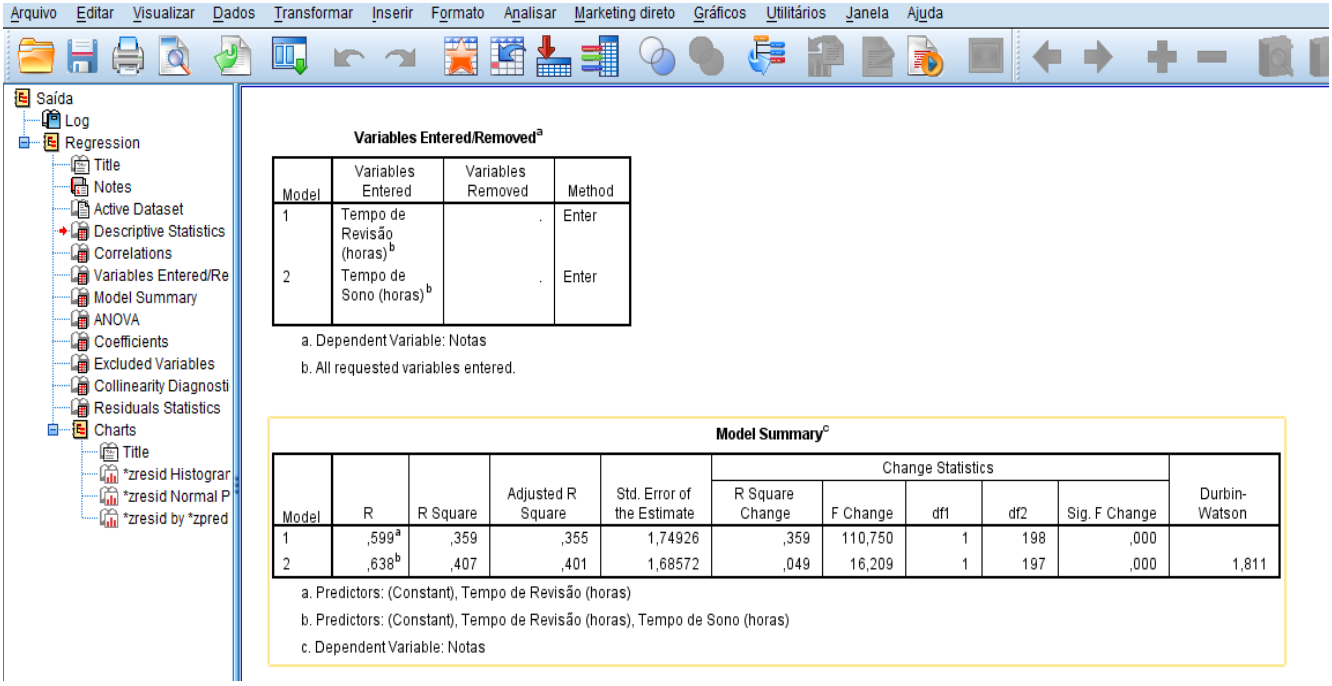
Correlations

		Notas	Tempo de Revisão (horas)	Tempo de Sono (horas)
Pearson Correlation	Notas	1,000	,599	,326
	Tempo de Revisão (horas)	,599	1,000	,182
	Tempo de Sono (horas)	,326	,182	1,000
Sig. (1-tailed)	Notas	.	,000	,000
	Tempo de Revisão (horas)	,000	.	,005
	Tempo de Sono (horas)	,000	,005	.
N	Notas	200	200	200
	Tempo de Revisão (horas)	200	200	200
	Tempo de Sono (horas)	200	200	200

Na primeira tabela é possível identificar a média, o desvio padrão das variáveis, o tempo médio de horas revisando o conteúdo e o tempo de sono.

A segunda tabela é a de correlações, isto é, mostra a correlação existente entre as variáveis. Nota-se que a variável nota tem uma variação moderada com o tempo de revisão (0,599) e uma correlação média com o tempo de sono (0,326), isso se torna um indicio de que essas variáveis vão ser importantes para o modelo. Além disso, ao observar a tabela da correlação é possível identificar se existe ausência da multicolinearidade, ou seja, isso significa que não pode ocorrer uma alta correlação entre as variáveis independentes e, conforme mostra à tabela de correlação isso não ocorre, pois, a correlação média está em 0,18.

Na próxima figura é possível identificar mais duas tabelas.



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Tempo de Revisão (horas)[b]	.	Enter
2	Tempo de Sono (horas)[b]	.	Enter

a. Dependent Variable: Notas

b. All requested variables entered.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	,599[a]	,359	,355	1,74926	,359	110,750	1	198	,000	
2	,638[b]	,407	,401	1,68572	,049	16,209	1	197	,000	1,811

a. Predictors: (Constant), Tempo de Revisão (horas)

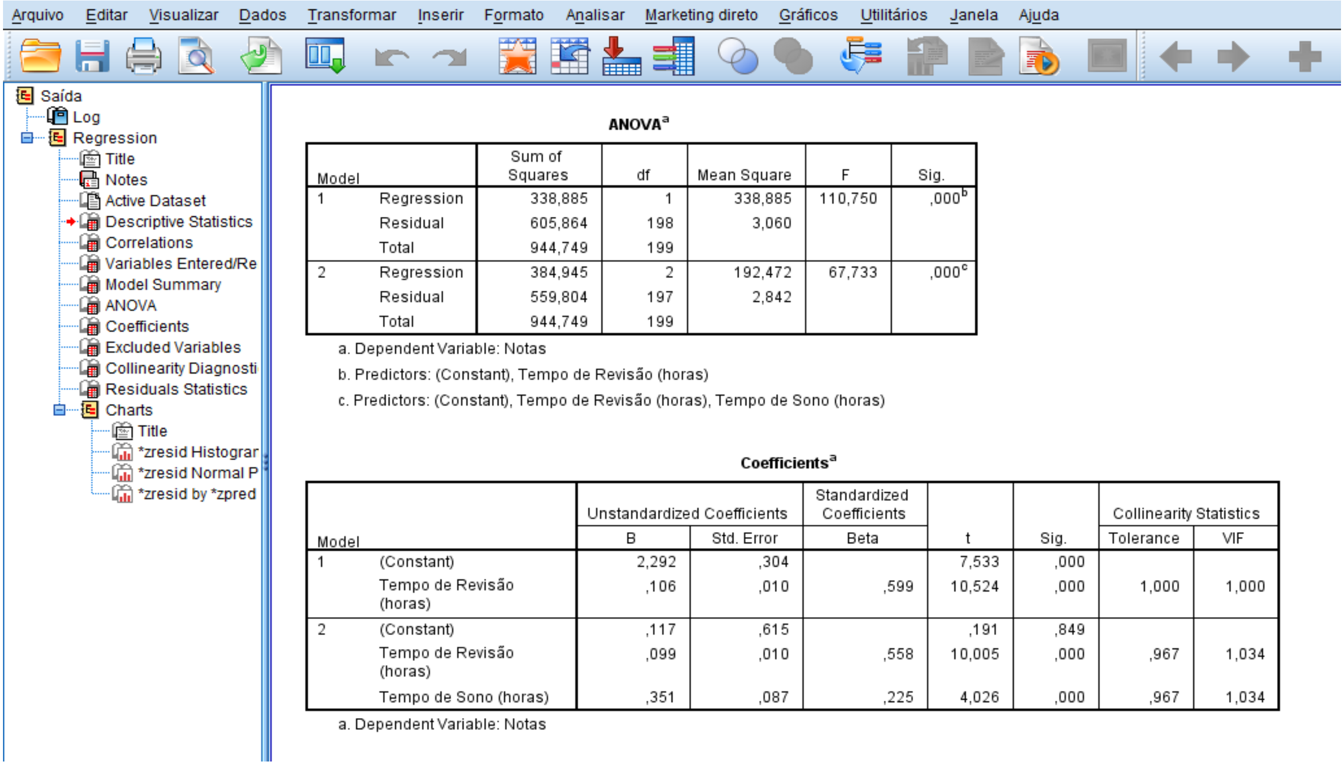
b. Predictors: (Constant), Tempo de Revisão (horas), Tempo de Sono (horas)

c. Dependent Variable: Notas

A primeira tabela faz um resumo dos modelos, mostra que no modelo um foi inserida a variável tempo de revisão e no modelo dois a variável tempo de sono.

A segunda tabela é a tabela Model Summary, sendo muito importante para conseguir comparar os modelos entre si. Na primeira linha ela mostra o modelo 1, que contém apenas o tempo de revisão como variável independente, onde o R (correlação entre a variável dependente e independente). Tenho o R^2 (R square, que explica a porcentagem da variação na variável dependente que é explicada pela variável independente, ou seja, o modelo 1 explica 35.9% das notas ajustadas).

Vejamos agora a próxima figura que apresenta mais duas tabelas.



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	338,885	1	338,885	110,750	,000[b]
	Residual	605,864	198	3,060		
	Total	944,749	199			
2	Regression	384,945	2	192,472	67,733	,000[c]
	Residual	559,804	197	2,842		
	Total	944,749	199			

a. Dependent Variable: Notas

b. Predictors: (Constant), Tempo de Revisão (horas)

c. Predictors: (Constant), Tempo de Revisão (horas), Tempo de Sono (horas)

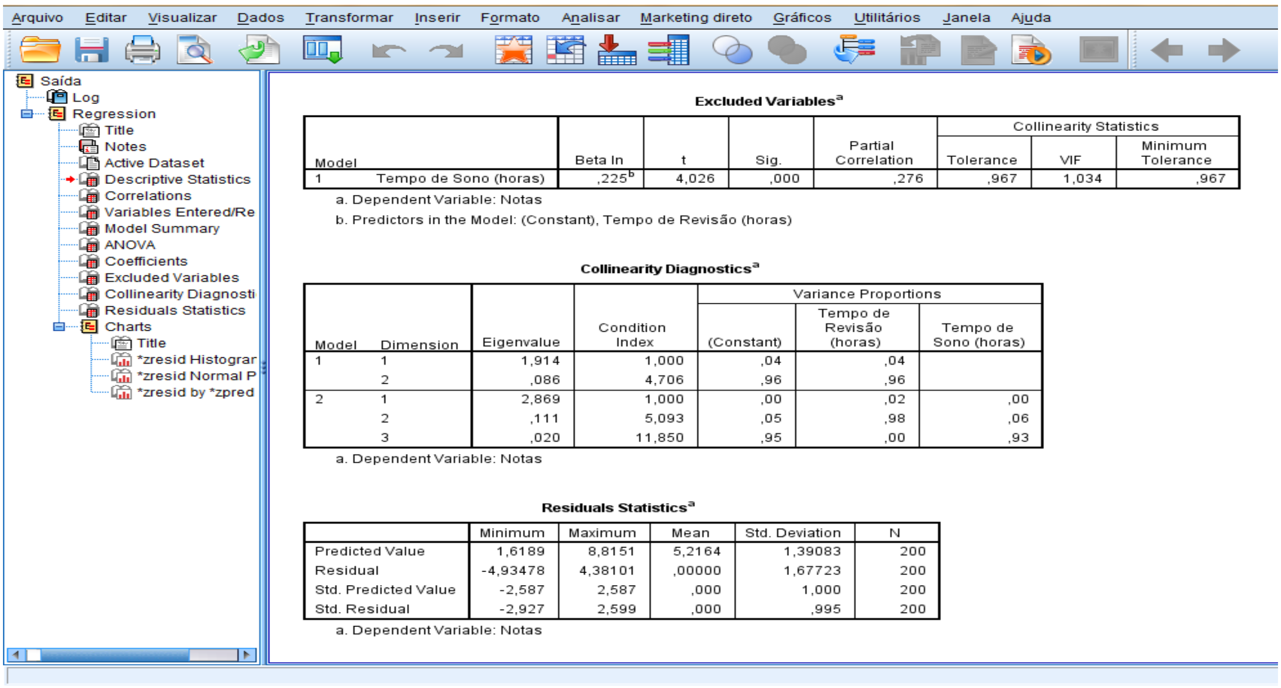
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2,292	,304		7,533	,000		
	Tempo de Revisão (horas)	,106	,010	,599	10,524	,000	1,000	1,000
2	(Constant)	,117	,615		,191	,849		
	Tempo de Revisão (horas)	,099	,010	,558	10,005	,000	,967	1,034
	Tempo de Sono (horas)	,351	,087	,225	4,026	,000	,967	1,034

a. Dependent Variable: Notas

A primeira tabela apresenta a ANOVA, onde é possível analisar, por exemplo, o teste F e o teste p (Sig). A segunda tabela é a Coefficients que mostra os coeficientes necessários para interpretar se realmente as variáveis são importantes para o modelo estudado. O teste p (Sig) é menor que 0,05, ou seja, os valores destas variáveis são significativamente diferentes de zero.

A próxima figura apresenta mais três tabelas, sendo que a mais importante é a tabela dos resíduos.



Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Tempo de Sono (horas)	,225[b]	4,026	,000	,276	,967	1,034	,967

a. Dependent Variable: Notas

b. Predictors in the Model: (Constant), Tempo de Revisão (horas)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Tempo de Revisão (horas)	Tempo de Sono (horas)
1	1	1,914	1,000	,04	,04	
	2	,086	4,706	,96	,96	
2	1	2,869	1,000	,00	,02	,00
	2	,111	5,093	,05	,98	,06
	3	,020	11,850	,95	,00	,93

a. Dependent Variable: Notas

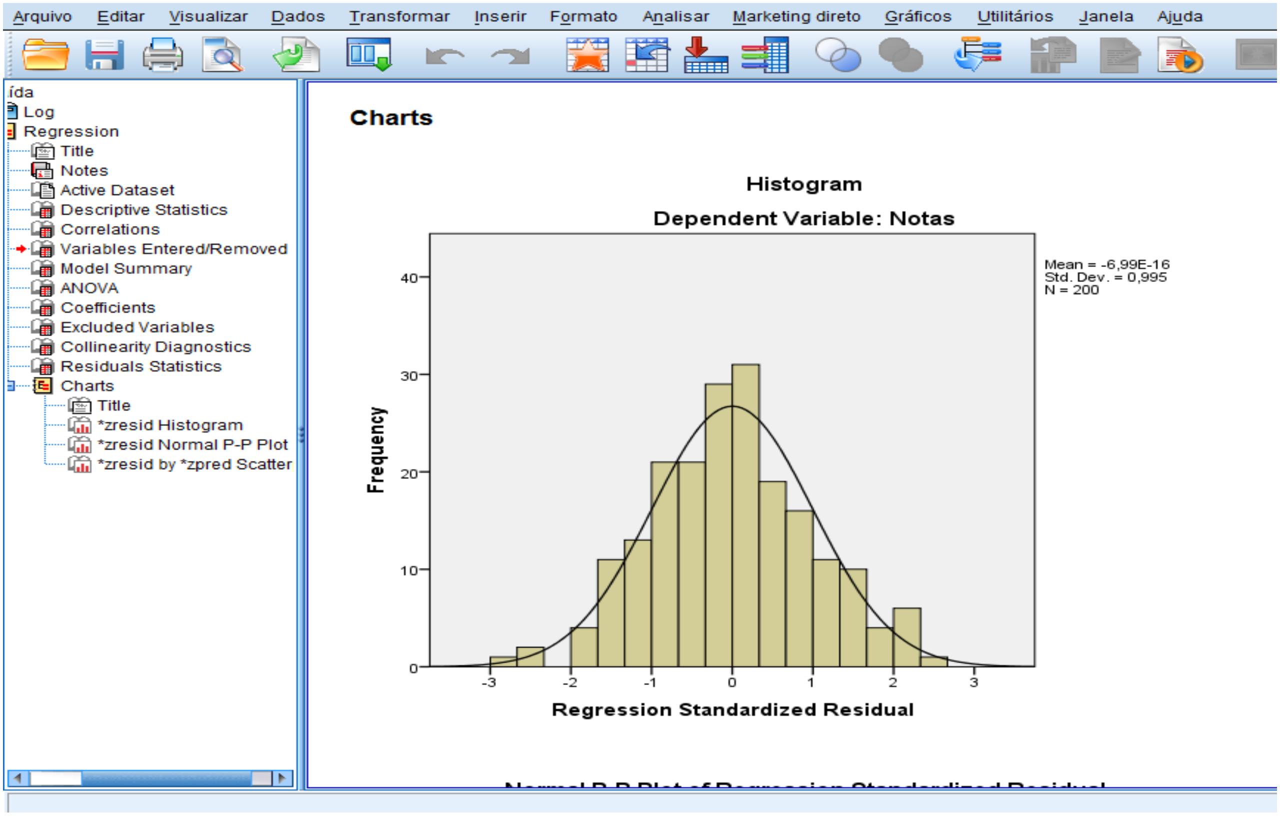
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,6189	8,8151	5,2164	1,39083	200
Residual	-4,93478	4,38101	,00000	1,67723	200
Std. Predicted Value	-2,587	2,587	,000	1,000	200
Std. Residual	-2,927	2,599	,000	,995	200

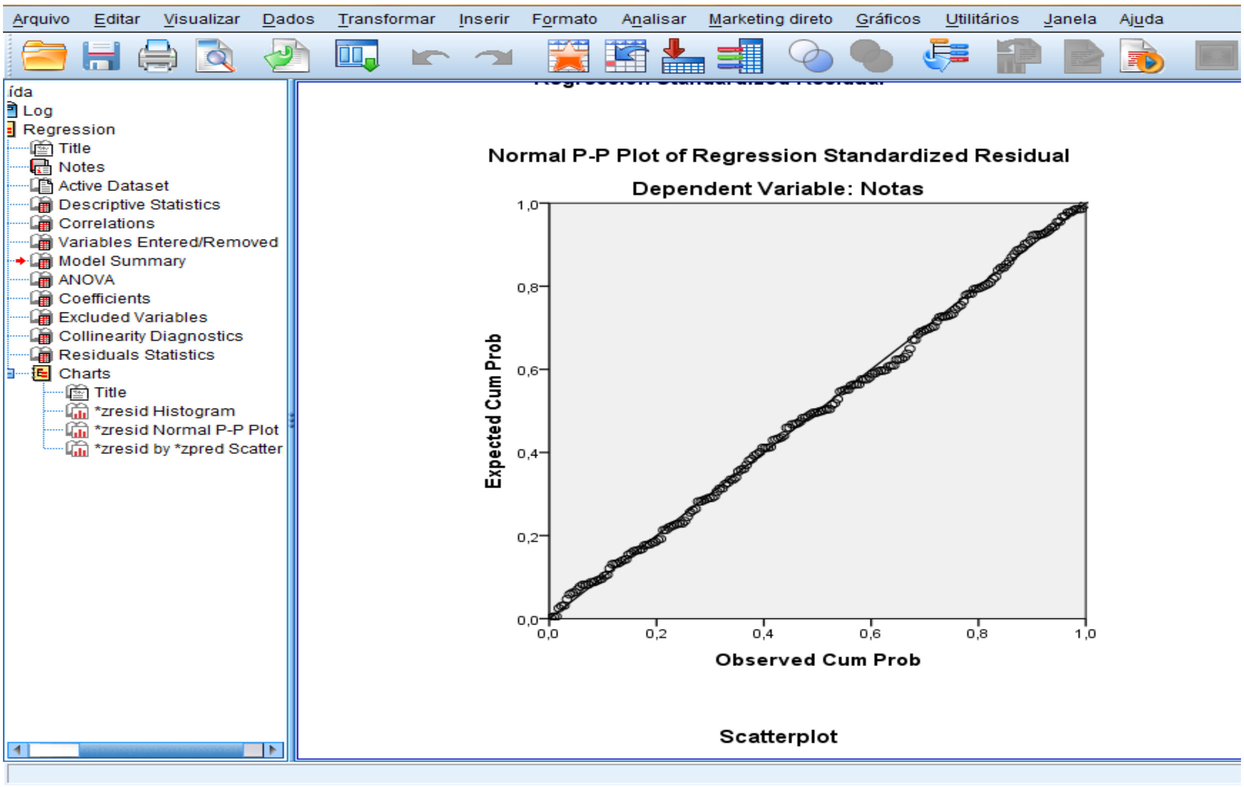
a. Dependent Variable: Notas

Na tabela dos resíduos pode-se observar se existe outiliers que são valores que estão fora das faixas -3 e +3, nesse exemplo observa-se que tanto os valores previstos, quanto os valores residuais ficaram dentro do limite.

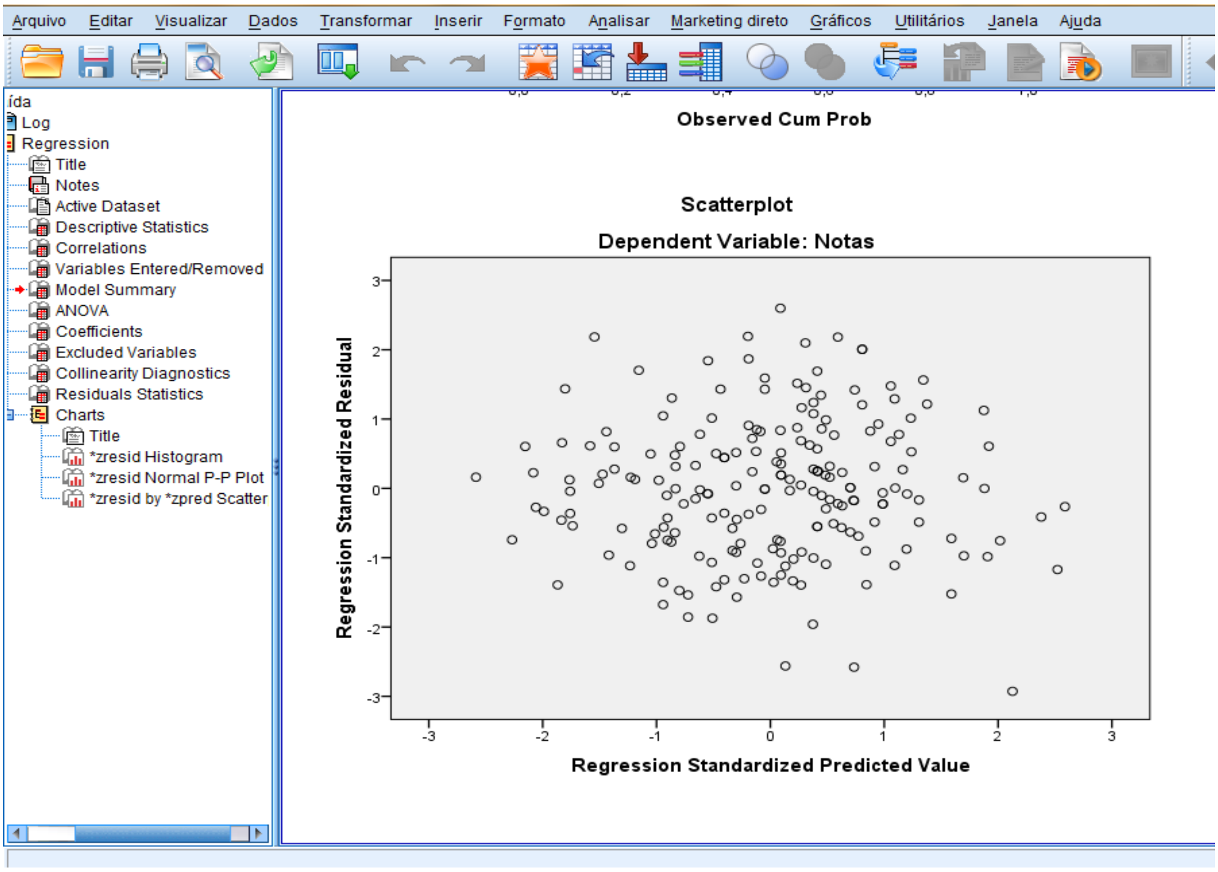
A próxima figura apresenta um histograma para os resíduos, sendo que é importante ter uma distribuição normal. No nosso exemplo, os resíduos não têm uma distribuição perfeitamente normal, mas está quase normal.



A próxima figura apresenta outro gráfico que permite analisar a qualidade dos resíduos, sendo que cada bolinha representa um resíduo e quanto mais perto da linha mais os resíduos se aproximam da distribuição normal.



A última figura nos mostra o gráfico da homocedasticidade. Caso exista homocedasticidade os pontos estarão dispersos de forma aleatória.



REFERÊNCIAS

BARBETTA, P. A.; REIS, M. M.; BORNIA, A.C. Estatística para cursos de engenharia e informática. 3 ed. São Paulo: Atlas, 2010.

BRUNI, A. L. **SPSS-Guia prático para pesquisadores**. São Paulo: Atlas, 2012.

DUNLAP, William P.; LANDIS, Ronald S. Interpretations of multiple regression borrowed from factor analysis and canonical correlation. **The Journal of General Psychology**, v. 125, n. 4, p. 397-407, 1998.

FIELD, A. **Descobrindo a estatística usando o SPSS**. 2 ed. Porto Alegre: Bookman, 2009.

HAIR, J. F.; BLACK, B.; BABIN, B.; ANDERSON, R. E.; TATHAM, R. L. Análise multivariada de dados. 6. ed. Porto Alegre: Bookman, 2009.

TABACHNICK, B.; FIDELL, L. S. **Using multivariate statistics**. 3 ed. New York: Harper Collins, 1996.

KEPPEL, G**. Design and analysis**: A researcher’s handbook. 3 ed. Englewood Cliffs: Prentice-Hall, 1991.

KROMREY, Jeffrey D.; FOSTER-JOHNSON, Lynn. Statistically differentiating between interaction and nonlinearity in multiple regression analysis: A Monte Carlo investigation of a recommended strategy. **Educational and Psychological Measurement**, v. 59, n. 3, p. 392-413, 1999.