



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 01/2019

**Regressão Linear Múltipla
Como simplificar por meio do Excel e SPSS?**

André da Silva Pereira

Thayane Woellner Sviercoski Manosso

Emanuele Canali Fossatti

Sandra Mara Berti

A presente apostila foi elaborada com o objetivo de orientar os leitores em relação ao uso do Excel e do SPSS durante a Regressão Linear Múltipla, a fim de clarear e simplificar os passos deste processo em ambas as ferramentas.

Regressão Linear Múltipla

Como simplificar por meio do
Excel e SPSS?

Thayane Woellner Sviercoski Manosso

Emanuele Canali Fossatti

Sandra Mara Berti

SUMÁRIO

INTRODUÇÃO.....	3
INVESTIGAÇÃO DESCRITIVA	4
INVESTIGAÇÃO CORRELACIONAL.....	5
INVESTIGAÇÃO EXPERIMENTAL	6
<i>A Regressão Linear Múltipla (RLM)</i>	7
1. Definição das variáveis	8
2. Desenho do gráfico de dispersão	8
3. Montagem da equação da RLM	9
4. Rodar a RLM.....	10
5. Substituir os dados na equação da RLM	10
6. Interpretação dos resultados	14
* Variáveis Dummy	14
Exemplo - Exercício Prático.....	16
Utilizando o EXCEL.....	18
Passo 1: Definição das Variáveis.....	18
Passo 2: Desenho do gráfico de dispersão.....	19
Passo 3: Montagem da equação da RLM.....	22
Passo 4: Rodar a RLM.....	22
Passo 5: Substituir os dados na equação da RLM.....	27
Passo 6: Interpretação dos resultados.....	29
* Variáveis Dummy.....	29
Escolhendo o melhor modelo de regressão... ..	38
Utilizando o SPSS.....	44
Passo 1: Definição das Variáveis (Figura 49)	46
Passo 2: Desenho do gráfico de dispersão.....	46
Passo 3: Montagem da equação da RLM.....	50
Passo 4: Rodar a RLM.....	50
Passo 5: Substituir os dados na equação da RLM.....	55
Passo 6: Interpretação dos resultados.....	56
*Variáveis Dummy.....	56
Escolhendo o melhor modelo de regressão... ..	62
REFERÊNCIAS.....	66

INTRODUÇÃO

Estatística é uma palavra que por si só, assusta. Isso ocorre em razão de algumas pessoas simplesmente não gostarem de números, o que não é raro, especialmente em áreas como as Ciências Sociais.

O que acontece, na verdade, é que tantos símbolos e palavras estranhas nos confundem, não permitindo que analisemos os números de forma lógica e simples. Além disso, não há muito material disponível que descomplique a estatística. Pensando nisso, desenvolvemos esta apostila, a qual tem por objetivo explicar de forma prática e bastante ilustrativa como realizar a análise de dados por meio da **Regressão Linear Múltipla** em dois softwares: o EXCEL, velho conhecido e amigo de todos que têm conhecimento básico em informática; e o SPSS, 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.

Mas primeiro é necessário entender **O QUÊ** é, **PORQUÊ** precisamos usar e **COMO** chegamos até a Regressão Linear Múltipla, não é mesmo? Afinal, não esperamos que nossos leitores possuam algum conhecimento prévio de estatística para entender este material.

ESTATÍSTICA E CIÊNCIAS SOCIAIS

Sempre que buscamos entender ou prever um fenômeno, precisamos de dados que nos auxiliem, os quais podem ser coletados e analisados de diversas maneiras. A estatística faz parte da análise de dados **quantitativa**, um método que utiliza a linguagem matemática para descrever as causas de um fenômeno e as relações entre suas variáveis, por exemplo (FONSECA, 2002).

Para que possamos generalizar os resultados que obtemos por meio das análises estatísticas, precisamos que nossa amostra contemple uma representação viável da nossa população. A isso, damos o nome de **inferência estatística**, a qual nos permite tomar decisões generalizadas para toda a população, baseadas em uma amostragem.

As investigações com base em uma amostra, segundo Almeida e Freire (2000), podem ser de três tipos: **descritiva, correlacional ou experimental**. A investigação **descritiva** trata das variáveis separadamente, enquanto as investigações **correlacional** e **experimental** tratam da associação entre uma ou mais variáveis. A seguir, falaremos sobre cada uma delas.

INVESTIGAÇÃO DESCRITIVA

As técnicas que nos proporcionam analisar e interpretar as informações básicas dos dados coletados fazem parte de um conjunto chamado de **Investigação ou Estatística Descritiva**. A descrição dos dados é fornecida pelas **medidas de posição**, também chamadas de tendência central, as quais apresentam a frequência dos dados, conhecidas como **média, mediana e moda**; e **medidas de dispersão**, que nos dizem o quão dispersos ou distantes um do outro estão os valores de um conjunto de dados e são chamadas de **variância e desvio padrão**. Para entendermos melhor e de forma mais fácil, não utilizaremos as fórmulas matemáticas para explicar cada um desses termos, já que os softwares costumam nos fornecer esses valores quando solicitado, e o mais importante é saber interpretá-los.

- **Média:** é o valor médio dos dados e representa onde os dados se concentram. Para obtê-la, somam-se todos os dados de uma variável e divide-se pelo número de dados. Assim, teremos uma média para cada variável.
- **Mediana:** é o valor do meio de um conjunto de dados ordenado, o qual nem sempre é igual a média. Quando o número de dados é ímpar, a mediana é representada pelo número central do conjunto de dados, ou seja, exatamente 50% dos dados estão à sua esquerda (menores) e 50% à sua direita (maiores). Quando o número de dados é par, a mediana é calculada pela média dos dois valores do meio do conjunto.

* Quando existem valores excepcionalmente extremos no conjunto de dados (**outliers**), a mediana pode dar uma ideia melhor de um valor típico do que a média, visto que não é tão distorcida por esses valores.

- **Moda:** é o valor que aparece com maior frequência em um conjunto de dados. Uma variável pode possuir uma única moda (unimodal), duas modas (bimodal), mais de duas modas (multimodal) ou nenhuma moda (amodal).

* A moda é útil quando os valores não são numéricos, e por isso a média e a mediana não podem ser definidas.

- **Variância:** é o valor que mostra o quão distante os valores reais do conjunto de dados estão da média desse conjunto. Quanto menor a variância, mais próximos os valores estão da média; quanto maior a variância, mais distantes os valores estão da média.

- **Desvio padrão:** corresponde ao erro equivalente caso substituíssemos todos os valores reais de um conjunto de dados pela média desse conjunto.

*O desvio padrão é calculado pela raiz quadrada da variância. É mais fácil interpretar os dados utilizando-se o desvio padrão, porque a variância é um valor ao quadrado (2) e, portanto, não pode ser diretamente comparado aos valores reais do conjunto de dados.

INVESTIGAÇÃO CORRELACIONAL

Esse tipo de investigação consiste em descobrir se as variáveis que estão sendo estudadas possuem correlação (dependência) entre si. O índice de correlação é dado pelo coeficiente de correlação de Pearson (r), que pode variar de -1 à 1, conforme mostra a Figura 1.

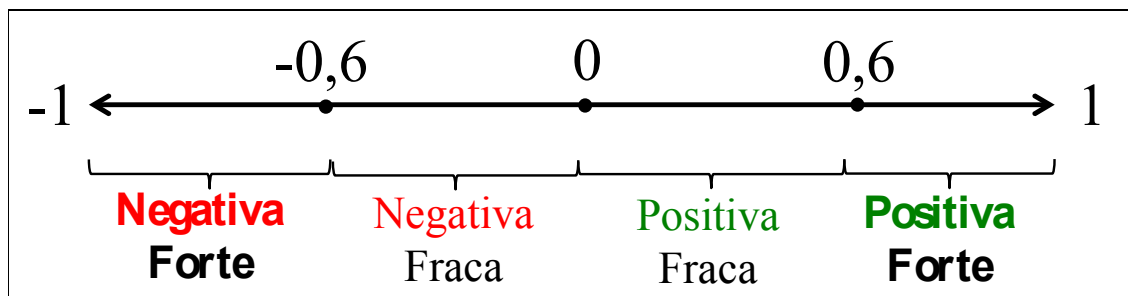


Figura 1. Índice de correlação de Pearson (r)

Valores negativos indicam correlação negativa entre variáveis ($-1 < r < 0$), ou seja, quanto maior o valor de uma das variáveis, menor será o valor da segunda variável. Valores positivos indicam correlação positiva entre variáveis ($0 < r < 1$), indicando que quanto maior

for o valor de uma das variáveis, maior o valor da outra variável. Quando a correlação for zero ($r = 0$), não há nenhuma correlação entre as variáveis. Assim, dizemos que a correlação entre as variáveis é nula. Índices de correlação entre $|0,6|$ ou $|0,7|$ e $|1|$ indicam correlação FORTE entre as variáveis, e valores entre $|0,01|$ e $|0,59|$ ou $|0,69|$ indicam correlação FRACA entre as variáveis (HAIR et al., 2009).

Normalmente, o resultado de uma correlação é apresentado em forma de uma matriz de correlação, o que possibilita que correlacionemos mais de duas variáveis de uma única vez. A correlação entre uma variável e ela mesma sempre será 1,0, de forma que, na matriz de correlação, a diagonal sempre será uma sequência de 1,0, conforme mostra a Figura 2.

Variáveis	1	2	3	4	5	6	7	8	9	10
1	1,000	0,231								
2	0,231	1,000								
3	0,089	0,231	1,000							
4	0,489	0,27	-0,066	1,000						
5	0,412	0,313	0,036	0,564	1,000					
6	0,228	-0,075	0,066	0,168	0,242	1,000				
7	0,326	0,008	0,206	0,336	0,377	0,671	1,000			
8	0,189	0,244	0,409	0,061	0,148	0,146	0,125	1,000		
9	-0,07	0,256	0,483	-0,077	0,045	-0,024	-0,072	0,404	1,000	
10	0,312	0,519	0,211	0,249	0,293	0,154	-0,033	0,397	0,281	1,000

Figura 2. Matriz de Correlação

Note que, neste caso, não há valores em cima da linha diagonal de 1,0. Isso acontece porque o índice de correlação entre a Variável 2 e a Variável 1 é o mesmo que já foi calculado para a Variável 1 x Variável 2, na primeira linha. Algumas vezes, dependendo do software utilizado ou por escolha do pesquisador, todas as casas estarão preenchidas, de forma que os valores se repetem quando as variáveis correlacionadas são as mesmas.

INVESTIGAÇÃO EXPERIMENTAL

Por fim, a investigação experimental procura relações causais e predições entre as variáveis, com o intuito de controlar o fenômeno que se deseja estudar. Inúmeros tipos de procedimentos podem ser utilizados nessa etapa de investigação. Um dos mais difundidos e que possui amplo poder de explicação de previsões, é a **Regressão**.

Quando realizamos a pesquisa teórica sobre o fenômeno que desejamos estudar, encontramos relações anteriores já estudadas e comprovadas por outros autores. A partir disso, temos uma base sólida para acreditar quais variáveis são explicadas e quais são as que explicam determinados fenômenos. À explicada, damos o nome de **variável dependente**, porque o valor assumido por esta depende da variação de outras variáveis. A(s) variável(is) explicativa(s) é(são) chamada(s) de **variável(is) independente(s)**, porque seu(s) valor(es) não se altera(m) quando o valor de outras variáveis muda.

A diferença principal entre a análise de regressão e a correlação, é que na segunda apenas sabemos que há uma associação entre as variáveis, mas não sabemos qual delas é a variável que depende da outra.

A Regressão é sempre **Linear**, porque supõe-se, previamente, que há correlação entre as variáveis que estão sendo analisadas, e isso, graficamente, é representado por uma “linha”. 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)**. Mas quando precisamos analisar a relação entre mais de duas variáveis, em que duas ou mais variáveis assumem o papel de independentes, precisamos calcular uma **Regressão Linear Múltipla**.

A Regressão Linear Múltipla (RLM)

Ressaltamos que a primeira coisa a se fazer, em qualquer estudo científico, é a pesquisa teórica. É a partir desta que poderemos desenvolver nossa ferramenta de coleta de dados, quantos casos serão necessários, quais dados (variáveis) precisamos coletar e qual a relação de causa e efeito entre essas variáveis. Mas esse não é o nosso foco aqui. Supõe-se que, nessa etapa da análise de dados, tudo isso já tenha sido feito. A partir daí, para entendermos a lógica da RLM, seguiremos seis passos para análise de dados:

- 1. Definição das variáveis;**
- 2. Desenho do gráfico de dispersão;**
- 3. Montagem da equação da RLM;**
- 4. Rodar a RLM;**
- 5. Substituir os dados na equação da RLM;**
- 6. Interpretação dos resultados.**

Esses passos serão utilizados ao longo de toda a apostila, para exemplificar o cálculo da RLM tanto no Excel, quanto no SPSS. Mas primeiro, cada um deles será explicado individualmente.

1. Definição das variáveis

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.

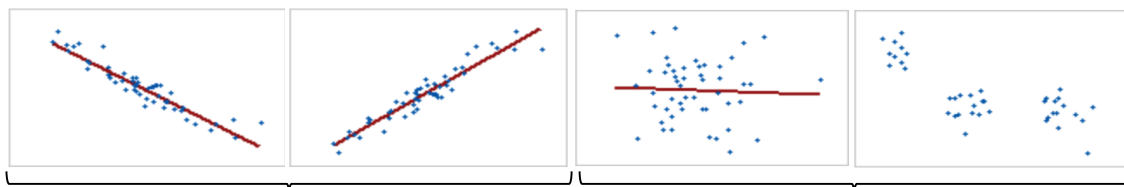
Conforme Field (2009), para construir um modelo complexo com várias variáveis independentes, muito cuidado deve ser tomado ao selecionarmos tais variáveis, porque os valores dos **coeficientes de regressão** (R^2) dependem delas.

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.

2. Desenho do gráfico de dispersão

Após a definição das variáveis, como forma de confirmar o pressuposto de que há correlação entre a variável dependente e cada uma das variáveis independentes, podemos gerar os gráficos de dispersão para cada uma das relações. Por exemplo, se foram escolhidas duas variáveis independentes, A e B, dois gráficos serão gerados, um para a relação entre a variável dependente e A e outro para a relação entre a variável dependente e B.

A interpretação se torna mais fácil quando se gera a linha de tendência sobre os pontos dispersos no gráfico, embora não seja necessário ainda se preocupar com a equação da reta, apenas com a lógica da imagem apresentada. A Figura 3 ilustra a lógica da interpretação de um gráfico de dispersão.



Há correlação entre a variável dependente e a variável independente. A linha de tendência pode aparecer mais ou menos inclinada, o que indica uma correlação forte ou fraca, respectivamente. O resultado apontado sugere que é importante manter a variável independente no modelo.

Não há correlação entre a variável dependente e a variável independente. Provavelmente o r da equação da reta será próximo de zero. O resultado apontado sugere que talvez seja necessário excluir a variável independente do modelo.

Figura 3. Interpretação de gráficos de dispersão

3. Montagem da equação da RLM

Confirmadas as correlações entre as variáveis estudadas, montaremos a equação que descreve a relação de dependência entre essas variáveis. A equação da RLM é a mesma equação da reta ($Y = a + bx$), com a diferença de que há múltiplas variáveis “b” que influenciam na inclinação da reta. A Figura 4 mostra como montar a equação e abaixo de cada termo estão seus respectivos significados.

O diagrama mostra a equação de regressão linear múltipla: $Y = \beta_0 + \beta_1 \cdot \chi_{1i} + \beta_2 \cdot \chi_{2i} + \dots + \beta_k \cdot \chi_{ki} + u_i$. Cada termo é envolto em um retângulo colorido com uma seta apontando para uma explicação:

- Y (retângulo amarelo) → Variável dependente (em vermelho).
- β_0 (retângulo verde) → Constante (em verde).
- $\beta_1, \beta_2, \dots, \beta_k$ (retângulos azuis) → Coeficientes das variáveis independentes (em azul).
- $\chi_{1i}, \chi_{2i}, \dots, \chi_{ki}$ (retângulos vermelhos) → Variáveis independentes (em vermelho).
- u_i (retângulo roxo) → erro (em roxo).

Figura 4. Equação de RLM

Onde:

- Y = valor previsto da **variável dependente** que será obtido por meio do modelo estimado;
- β_0 = representa a **constante** ou **coeficiente linear**; quando todos os χ forem iguais à 0, é o valor de β_0 que corresponde à Y . A constante também nos mostra, no gráfico, qual o valor de Y , quando χ for igual à 0, ou seja, qual o ponto em que a reta inicia no eixo Y do gráfico.
- β_n = é o coeficiente de cada **variável independente** ou **coeficientes angulares**. Esse valor indica quanto a variável dependente (Y) vai variar com a variação de uma unidade de χ , quando todas as outras variáveis forem constantes. No gráfico, representam a inclinação da reta.
- χ_n = é a **descrição** (nome) de cada variável independente;
- u = termo de **erro** ou resíduo, o qual equivale à diferença entre o valor real de Y e o valor previsto de Y . Quando menor o erro, melhor. Se o erro for muito alto, significa que outras variáveis, além das variáveis χ que foram incluídas na equação, afetam Y ;
- i = representa cada uma das variáveis da amostra ($i = 1, 2, 3 \dots n$, em que n é o tamanho da amostra).

4. Rodar a RLM

Depois de montar a equação da RLM fica mais fácil identificar as relações entre as variáveis. O desenvolvimento do cálculo da RLM depende do software que se está utilizando. Por isso, esse passo será melhor explicado no exemplo posterior, que será aplicado tanto no Excel, quanto no SPSS.

De qualquer forma, podemos adiantar que os valores de Y e dos χ sempre serão os nomes das variáveis, constituídos por palavras ou abreviações. Já os β sempre serão valores, e correspondem aos coeficientes das variáveis à que estão associados. O erro (u) também é sempre em formato numérico, e seu valor varia dependendo dos valores das outras variáveis.

5. Substituir os dados na equação da RLM

Após calcular os valores β de cada uma das variáveis, substituiremos esses valores na equação inicial. Mas antes, precisamos saber se as variáveis independentes realmente têm poder preditivo sob a variável dependente.

Para isso, primeiro, vamos falar da **significância**. O nível de significância sempre equivale à 1 menos (-) o intervalo de confiança que se está utilizando, ou seja, se quero ter 90% de confiança nos resultados, meu nível de confiança equivale à $1 - 0,9 = 0,1$; se quero ter 95% de confiança nos resultados, meu nível de confiança equivale à $1 - 0,95 = 0,05$; e se quero ter 99% de confiança nos resultados, meu nível de confiança equivale à $1 - 0,99 = 0,01$. O intervalo de confiança mais utilizado é o de 95%, o que equivale à um nível de significância de 0,05. Mas o que isso significa? Vamos entender por meio da Figura 5.

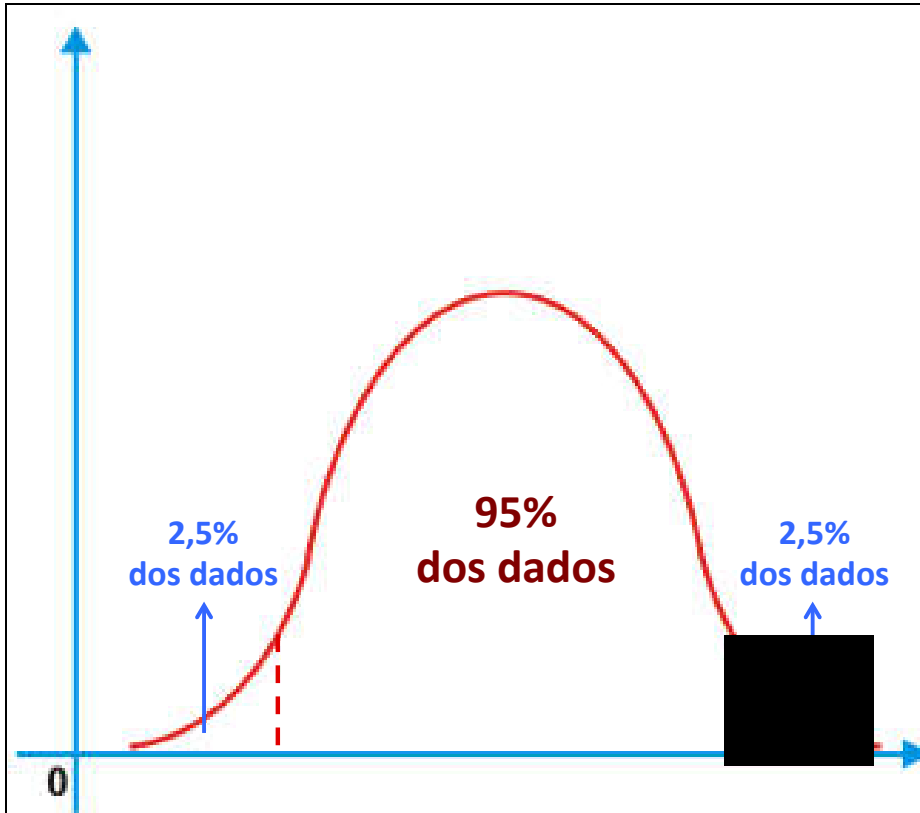


Figura 5. Intervalo de confiança

Quando utilizamos o intervalo de confiança de 95%, consideramos que os resultados válidos estão dentro da área vermelha do gráfico. As “caudas” da curva, que juntas correspondem à 5% dos dados (ou 0,05), abrangem os valores que não fazem parte do intervalo de confiança que desejamos. Por isso, quando dizemos que um valor “é significativo à 95% de confiança” ou que o teste t (valor-P) é menor que 0,05, queremos dizer que ele está na parte vermelha do gráfico, portanto, é válido no modelo. Assim, as variáveis da RLM que apresentarem valores -P ou a significância do teste t menores que 0,05, são mantidas, e as que não apresentarem, são excluídas.

Mas afinal, o que são **teste t**, **valor-P**, **teste z** e **teste f**?

Quando coletamos dados em amostras grandes (mais do que 30, 50 ou 100 casos, dependendo do autor), a média da amostra tende a se comportar da mesma forma que a média de toda a população, visto que ela é representativa e pode ser generalizada. Nesses casos, levamos em conta as significâncias do **TESTE Z**, porque a distribuição é normal, em torno da média, ou seja, valores mais próximos da média tem mais probabilidade de aparecerem do que valores mais distantes da média, como mostra a Figura 6.

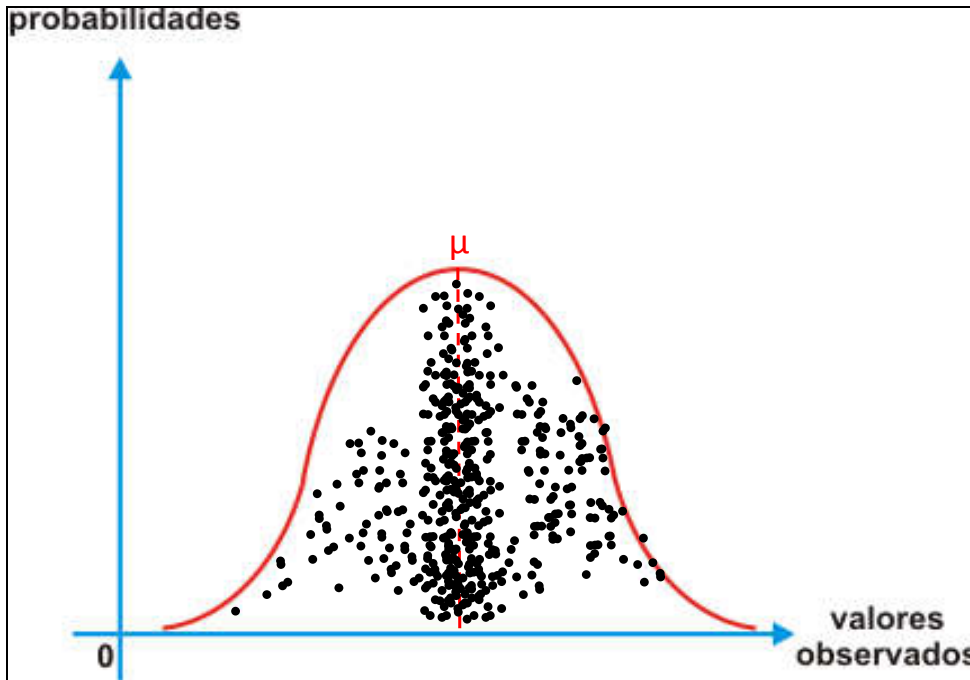


Figura 6. Teste Z

O **Teste T** tem a mesma função do teste z, porém, quando as amostras são pequenas ou quando não se conhece o desvio padrão da população, uma vez que nesse caso não se pode extrapolar o desvio padrão da amostra para a população. O valor-P equivale ao nível de significância do teste t. Alguns softwares fornecem os valores de t e os respectivos valores-P e outros os valores de t e os níveis de significância de cada um, o que quer dizer a mesma coisa.

A Figura 7 ilustra a diferença entre Z e T. Como exemplo, desenhamos as curvas de duas amostras pequenas. Note que quanto menor a amostra, mais distante da média os dados podem estar, e por isso não se pode generalizar a média e o desvio padrão da amostra para a população.

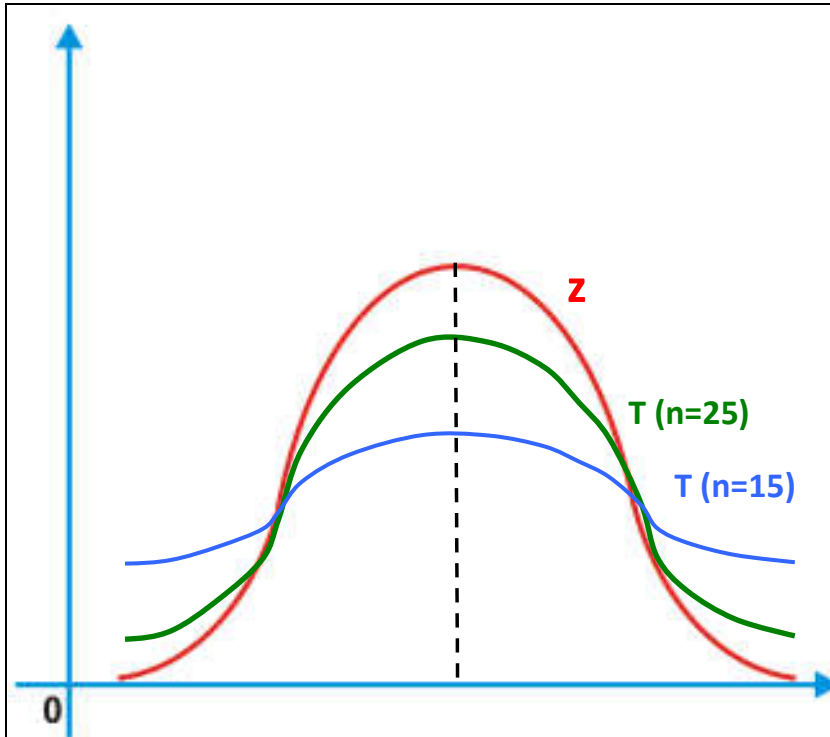


Figura 7. Diferença entre teste z e teste t

Mas então, o que fazer se a amostra for pequena?

Quando não podemos verificar a normalidade dos dados em virtude de nossa amostra ser pequena demais, precisamos testar a normalidade de nossos dados, por meio dos **testes não-paramétricos de hipóteses**. Esse não é o nosso foco aqui, mas precisamos saber que eles existem para que, se necessário, possamos utilizá-los e saibamos pelo que procurar na literatura relacionada a estatística. Retiramos uma explicação básica sobre alguns testes do livro de Bruni (2012), o qual explica detalhadamente cada um deles.

a) **Teste de Kolmogorov-Smirnov:** analisa se os dados da amostra foram extraídos de uma população com uma distribuição peculiar de frequências, como a distribuição normal;

b) **Teste do qui-quadrado:** empregado na análise de frequências, quando uma característica da amostra é analisada;

c) **Teste do qui-quadrado para independência ou associação:** também empregado na análise de frequências, porém quando duas características da amostra são analisadas;

d) **Teste dos sinais:** empregado no estudo de dados emparelhados, quando um mesmo elemento é submetido a duas medidas;

e) **Teste de Wilcoxon:** também analisa dados emparelhados, permitindo, porém, uma consideração das magnitudes encontradas;

f) **Teste de Mann-Whitney:** analisa se dois grupos originam-se de populações com médias diferentes;

g) **Teste da mediana:** analisa se dois grupos originam-se de populações com medianas diferentes;

h) **Teste de Kruskal-Wallis:** analisa se K ($K > 2$) grupos originam-se de populações com médias diferentes.

O **TESTE F**, por fim, testa a equação como um todo, e não fornece os valores de significância de cada variável. É esse valor que nos diz se nossa equação, como um todo, explica a nossa variável dependente.

Resumindo, os valores β que se mantêm na equação da RLM são aqueles que possuem valores de significância menores que 0,05, quando se adota o intervalo de confiança de 95%.

6. Interpretação dos resultados

Finalmente, e talvez o passo mais importante, chega o momento da interpretação de todos esses resultados. Os números possuem diversas informações, mas muitas vezes seus significados estão obscuros. Assim, precisamos traduzi-los em forma de palavras para que outras pessoas vejam o que nós estamos vendo.

Para isso, desenvolveremos um exemplo, seguindo todo o passo a passo que foi apresentado, e interpretaremos os resultados gerados pelos softwares (Excel e SPSS).

* Variáveis Dummy

Mas, e quando as variáveis não são numéricas, mas sim são qualitativas?

Bom, nesse caso, precisaremos transformá-las em variáveis quantitativas, para que possamos realizar as análises estatísticas. Vamos entender a nomenclatura que utilizamos para cada tipo de variável através da Figura 8.

Variáveis	Quantitativa		Qualitativa	
	Discreta	Contínua	Nominal	Ordinal
Descrição	São resultantes de contagens representadas como números inteiros	Podem assumir qualquer valor dentro de um intervalo	Não permitem comparações	Permitem comparações
Exemplos	Nº de filhos	Peso e altura	Nome	Escolaridade e Likert

Figura 8. Diferença entre variáveis quantitativas e qualitativas

Quando transformamos uma variável qualitativa em uma variável quantitativa (numérica), à ela damos o nome de variável *dummy*. Uma variável qualitativa com n

categorias gera $n-1$ variáveis *dummies*. Para atribuir valores numéricos às variáveis qualitativas sempre responderemos perguntas com “Sim” e “Não”. Traduzindo: se a variável qualitativa tem 2 categorias, faremos uma pergunta, o que vai gerar uma *dummy*; se ela possuir 3 categorias, faremos 2 perguntas, gerando duas *dummies*. Uma forma de facilitar esse tipo de transformação é montar colunas diferentes para responder cada pergunta. Excepcionalmente, neste caso, daremos um exemplo a parte para facilitar.

Imaginemos que gostaríamos de analisar o quanto o grau de escolaridade influencia na renda mensal de uma amostra. Para isso, precisamos transformar o grau de escolaridade em uma variável quantitativa. As opções que oferecemos aos respondentes foram: ensino fundamental, ensino médio, ensino superior e pós-graduação. Como temos 4 categorias de escolaridade que se apresentam de forma qualitativa, precisaremos fazer 3 perguntas, criando assim 3 *dummies*. Quando a resposta à nossa pergunta for “Não”, atribuiremos o número 0 ao χ , e quando a resposta for “Sim”, atribuiremos o número 1 ao χ . Escolhemos uma variável (ensino fundamental), a qual não dará origem à nenhuma pergunta e, depois disso, montamos um quadro, conforme apresentado na Figura 9.

1. Tem ensino médio?	2. Tem ensino superior?	3. Tem pós-graduação?	Categoria
0	0	0	Ensino fundamental (E.F.)
1	0	0	Ensino médio (E.M.)
1	1	0	Ensino superior (E.S.)
1	1	1	Pós-Graduação (P.G.)

Figura 9. Elaboração de variáveis dummy

Portanto, o grau de escolaridade da amostra gera 3 variáveis *dummy* na equação da RLM. Vamos montar a equação para entendê-la melhor:

$$\text{RENDA} = \beta_0 + \beta_1 \text{E.M.} + \beta_2 \text{E.S.} + \beta_3 \text{P.G.}$$

Não precisamos de uma *dummy* para o ensino fundamental porque quando todos o χ forem iguais à 0, como mostra o quadro, o ensino fundamental equivale ao valor da constante. Vamos compreender melhor pela substituição da equação para cada categoria:

- i. **Ensino fundamental:** $\text{RENDA} = \beta_0 + \beta_1.0 + \beta_2.0 + \beta_3.0$ (a renda média será o valor da constante, pois todos os outros β são multiplicados por 0);
- ii. **Ensino Médio:** $\text{RENDA} = \beta_0 + \beta_1.1 + \beta_2.0 + \beta_3.0$ (a renda média será o valor da constante + o valor de β_1);

- iii. **Ensino Superior:** $RENDA = \beta_0 + \beta_1.1 + \beta_2.1 + \beta_3.0$ (a renda média será o valor da constante + o valor de β_1 + o valor de β_2);
- iv. **Pós Graduação:** $RENDA = \beta_0 + \beta_1.1 + \beta_2.1 + \beta_3.1$ (a renda média será o valor da constante + o valor de β_1 + o valor de β_2 + o valor de β_3).

Assim, se o grau de escolaridade for um preditor positivo da renda, ou seja, quanto maior o grau de escolaridade, maior a renda, todos os valores de β serão positivos, e à medida que se somam, a renda aumenta. Os exemplos a seguir terão seções específicas para explicar melhor essa transformação de variáveis qualitativas em *dummies* e como interpretá-las.

Exemplo - Exercício Prático

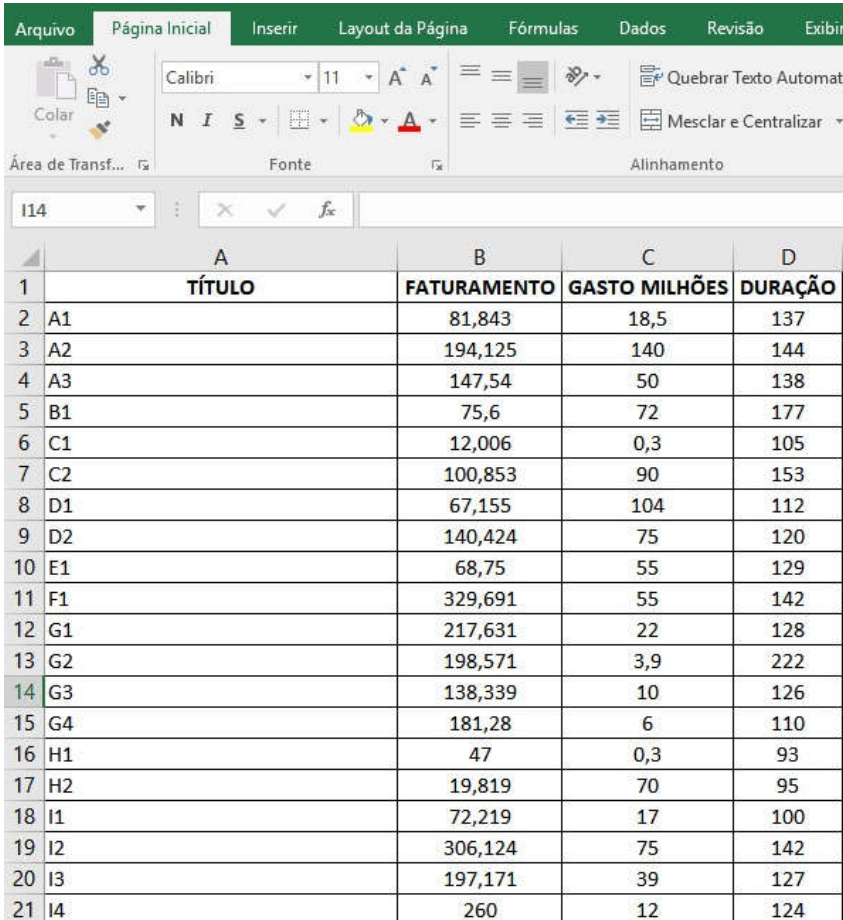
Com o objetivo de ilustrar a RLM, será apresentado um exemplo prático, no qual a regressão será realizada por meio do Excel e do SPSS. No exemplo apresentado na Tabela 1, observa-se um conjunto de dados de uma amostra formada por 36 filmes exibidos nos cinemas. Inicialmente, apresentam-se quatro variáveis: **código** (representando o título do filme), **faturamento** com o filme em milhões, **gasto** com o filme em milhões e **duração** do filme em minutos. Neste exemplo, busca-se identificar a relação existente entre as variáveis independentes (gasto e duração) e a variável dependente (faturamento). Logo depois, seguimos os passos, como apresentado na Introdução desta apostila.

Tabela 1. Amostra de dados – filmes exibidos no cinema

Nº	CÓDIGO	FATURAMENTO (em milhões R\$)	GASTO (em milhões R\$)	DURAÇÃO (em minutos)	Nº	CÓDIGO	FATURAMENTO (em milhões R\$)	GASTO (em milhões R\$)	DURAÇÃO (em minutos)
1	A1	81,843	18,500	137	19	I3	197,171	39,000	127
2	A2	194,125	140,000	144	20	I4	260,000	12,000	124
3	A3	147,540	50,000	138	21	M1	250,147	90,000	98
4	B1	75,600	72,000	177	22	M2	20,100	45,000	117
5	C1	12,006	0,300	105	23	P1	107,930	8,000	154
6	C2	100,853	90,000	153	24	R1	242,374	20,000	115
7	D1	67,155	104,000	112	25	S1	178,091	70,000	170
8	D2	140,424	75,000	120	26	S2	96,067	25,000	197
9	E1	68,750	55,000	129	27	S3	103,001	15,000	111
10	F1	329,691	55,000	142	28	S4	48,068	110,000	121
11	G1	217,631	22,000	128	29	T1	36,900	6,400	108
12	G2	198,571	3,900	222	30	T2	65,000	62,000	114
13	G3	138,339	10,000	126	31	T3	63,540	90,000	126
14	G4	181,280	6,000	110	32	T4	48,265	50,000	128
15	H1	47,000	0,300	93	33	T5	56,876	35,000	132
16	H2	19,819	70,000	95	34	T6	600,743	200,000	195
17	I1	72,219	17,000	100	35	T7	146,261	100,000	144
18	I2	306,124	75,000	142	36	V1	47,474	90,000	102

Utilizando o EXCEL

Inicialmente, iremos realizar cada um dos passos apresentados anteriormente utilizando a ferramenta Excel. Para isso, os dados apresentados na Tabela 1 devem ser transcritos para o Excel, conforme a Figura 10.



The image shows a screenshot of the Microsoft Excel interface. The ribbon at the top includes 'Arquivo', 'Página Inicial', 'Inserir', 'Layout da Página', 'Fórmulas', 'Dados', 'Revisão', and 'Exibir'. The 'Página Inicial' ribbon is active, showing options for 'Colar', 'Fonte' (Font), and 'Alinhamento' (Alignment). The font is set to Calibri, size 11. The formula bar shows 'I14'. Below the ribbon is a table with the following data:

	A	B	C	D
1	TÍTULO	FATURAMENTO	GASTO MILHÕES	DURAÇÃO
2	A1	81,843	18,5	137
3	A2	194,125	140	144
4	A3	147,54	50	138
5	B1	75,6	72	177
6	C1	12,006	0,3	105
7	C2	100,853	90	153
8	D1	67,155	104	112
9	D2	140,424	75	120
10	E1	68,75	55	129
11	F1	329,691	55	142
12	G1	217,631	22	128
13	G2	198,571	3,9	222
14	G3	138,339	10	126
15	G4	181,28	6	110
16	H1	47	0,3	93
17	H2	19,819	70	95
18	I1	72,219	17	100
19	I2	306,124	75	142
20	I3	197,171	39	127
21	I4	260	12	124

Figura 10. Dados no Excel

Passo 1: Definição das Variáveis

No exemplo citado, busca-se saber a influência que as variáveis gasto e duração têm sob o faturamento dos filmes exibidos no cinema. Assim, a variável dependente é o faturamento, enquanto as variáveis independentes são o gasto e a duração.

A fim de facilitar a regressão, sugere-se que as variáveis independentes estejam à direita da variável dependente. O código (título do filme) não é considerado uma variável, pois está sendo apresentado apenas com a intenção de ilustrar quais filmes foram avaliados nesta amostra. A Figura 11 apresenta a definição das variáveis no Excel.

	A	B	C	D
1	TÍTULO	FATURAMENTO	GASTO MILHÕES	DURAÇÃO
2	A1	81,843	18,5	137
3	A2	VARIÁVEL DEPENDENTE	140	144
4	A3	147,54	50	138
5	B1	75,6	72	177
6	C1	12,006	0,3	105
7	C2	100,853	90	153
8	D1	67,155	104	112
9	D2	140,424	75	120
10	E1	68,75	55	129
11	F1	329,691	55	142
12	G1	217,631	22	128
13	G2	198,571	3,9	222
14	G3	138,339	10	126
15	G4	181,28	6	110
16	H1	47	0,3	93
17	H2	19,819	70	95
18	I1	72,219	17	100
19	I2	306,124	75	142
20	I3	197,171	39	127
21	I4	260	12	124

Figura 11. Definição das variáveis no Excel

Passo 2: Desenho do gráfico de dispersão

O segundo passo consiste na criação do gráfico de dispersão. Ressalta-se a importância de gerar o gráfico de dispersão individualmente, relacionando a variável dependente com cada uma das variáveis independentes, visto que, por meio deste gráfico, busca-se verificar se existe relação entre duas variáveis, assim como qual a intensidade desta relação. Assim, serão gerados dois gráficos de dispersão: um que relaciona faturamento (dependente) e gasto (independente) e outro, faturamento (dependente) e duração (independente). Os passos para gerar o primeiro gráfico serão os mesmos utilizados para o segundo gráfico, o que muda é apenas a variável independente.

Para gerar o gráfico de dispersão no Excel, deve-se clicar em “inserir” (1), selecionar as colunas que serão relacionadas (2) e gerar gráfico de dispersão (3), conforme apresentado na Figura 12 e na Figura 13.

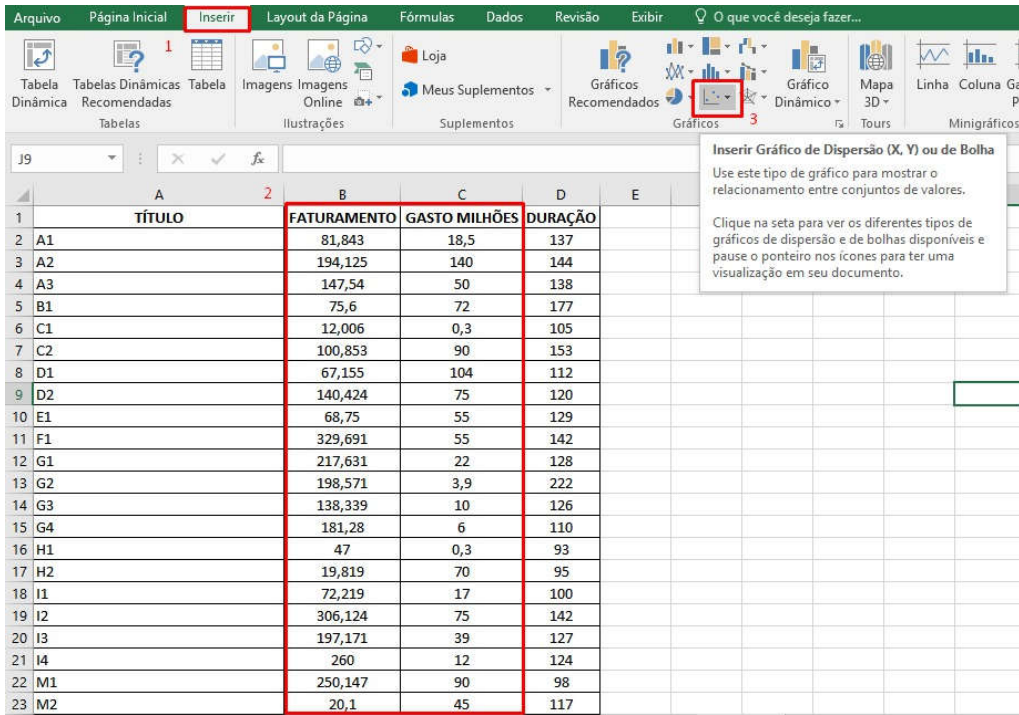


Figura 12. Montagem do gráfico de dispersão de Faturamento X Gasto

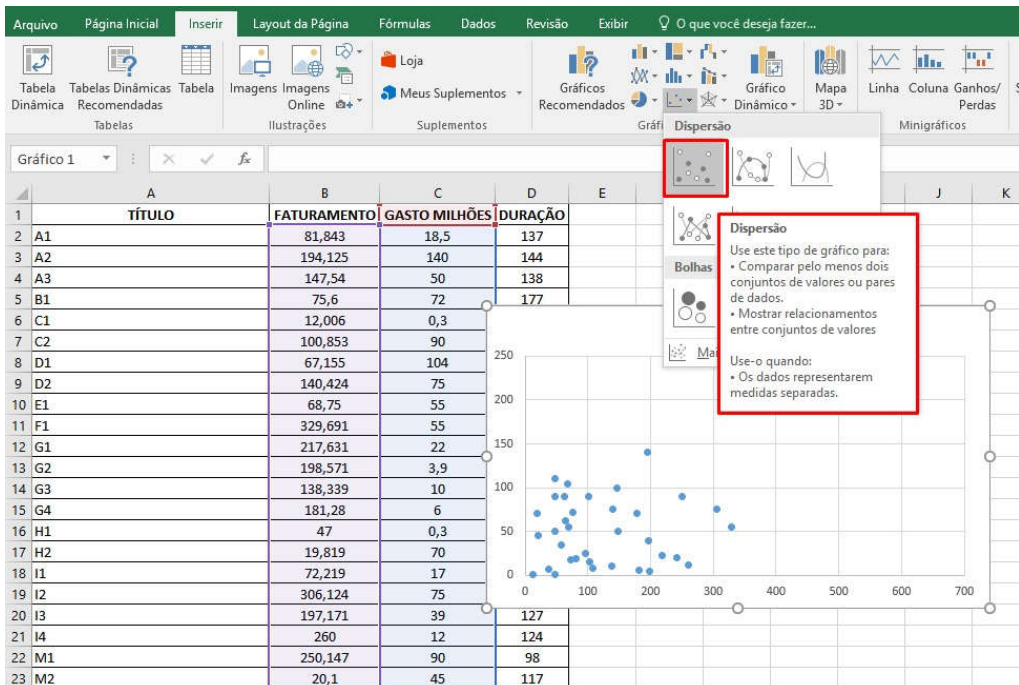


Figura 13. Gráfico de dispersão de Faturamento X Gasto

Para gerar a **linha de tendência** deve-se selecionar o gráfico, clicar no sinal “+” e assinalar o item “linha de tendência”. Com esta linha será possível observar qual a tendência de comportamento entre as variáveis analisadas, conforme a Figura 14.

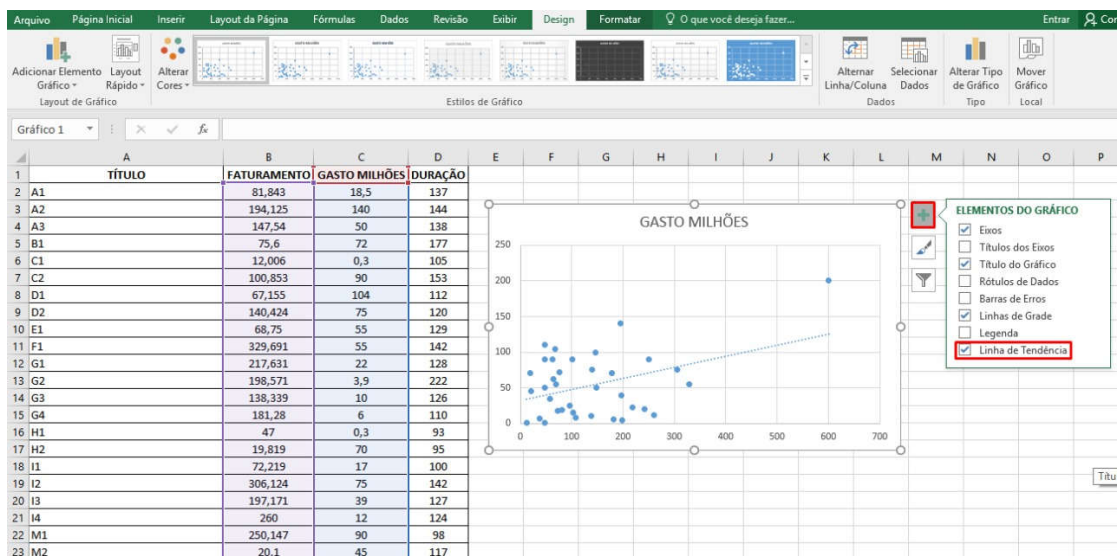


Figura 14. Linha de tendência da relação entre Faturamento X Gasto

Agora, faremos o mesmo com a variável duração, conforme Figura 15:

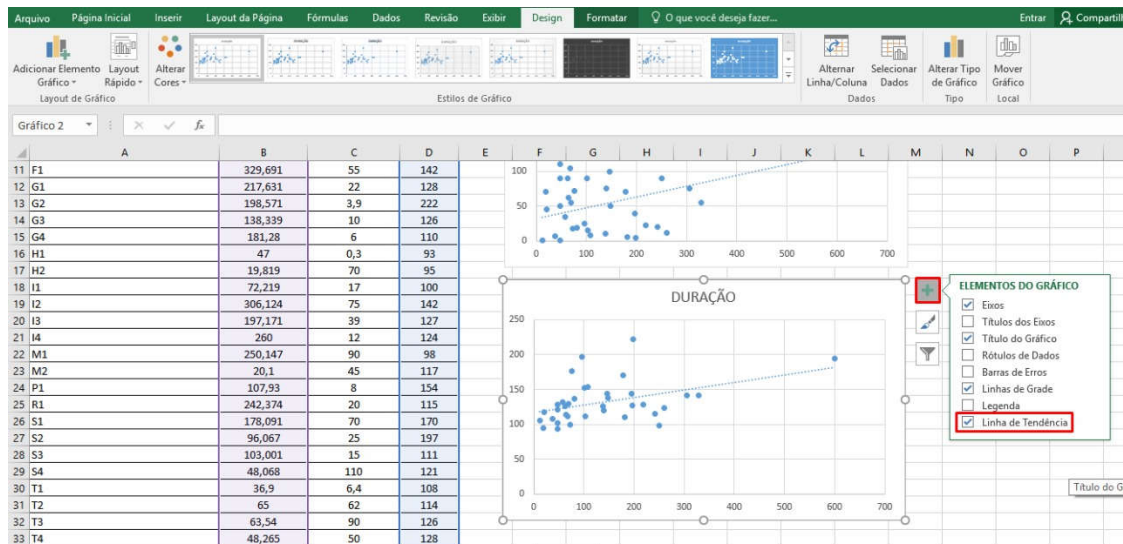
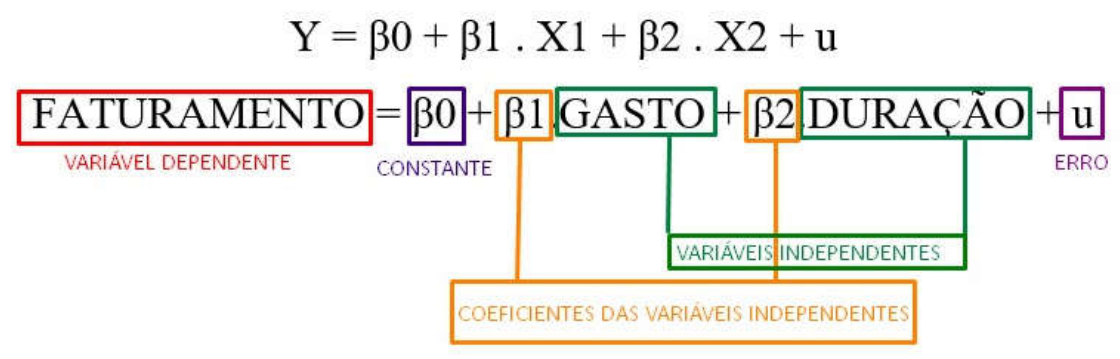


Figura 15. Linha de tendência da relação entre Faturamento X Duração

Observa-se que, por meio do gráfico de dispersão, é possível verificar como os dois conjuntos de dados comparáveis concordam entre si. Quanto mais os conjuntos de dados concordarem, mais os pontos dispersos tendem a se concentrar ao redor (próximo) da linha.

Passo 3: Montagem da equação da RLM

A Figura 16 apresenta a montagem da equação da RLM, conforme o exemplo prático apresentado.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$


O diagrama mostra a equação $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$ com os seguintes componentes rotulados:

- FATURAMENTO**: VARIÁVEL DEPENDENTE
- β_0** : CONSTANTE
- β_1** : COEFICIENTES DAS VARIÁVEIS INDEPENDENTES
- GASTO**: VARIÁVEL INDEPENDENTE
- β_2** : COEFICIENTES DAS VARIÁVEIS INDEPENDENTES
- DURAÇÃO**: VARIÁVEL INDEPENDENTE
- u**: ERRO

Figura 16. Equação da RLM - Exemplo Prático

Passo 4: Rodar a RLM

O **quarto passo** consiste em gerar a RLM no Excel. Para isso, é necessário utilizar a ferramenta “Análise de Dados”, conforme apresentado na Figura 17.

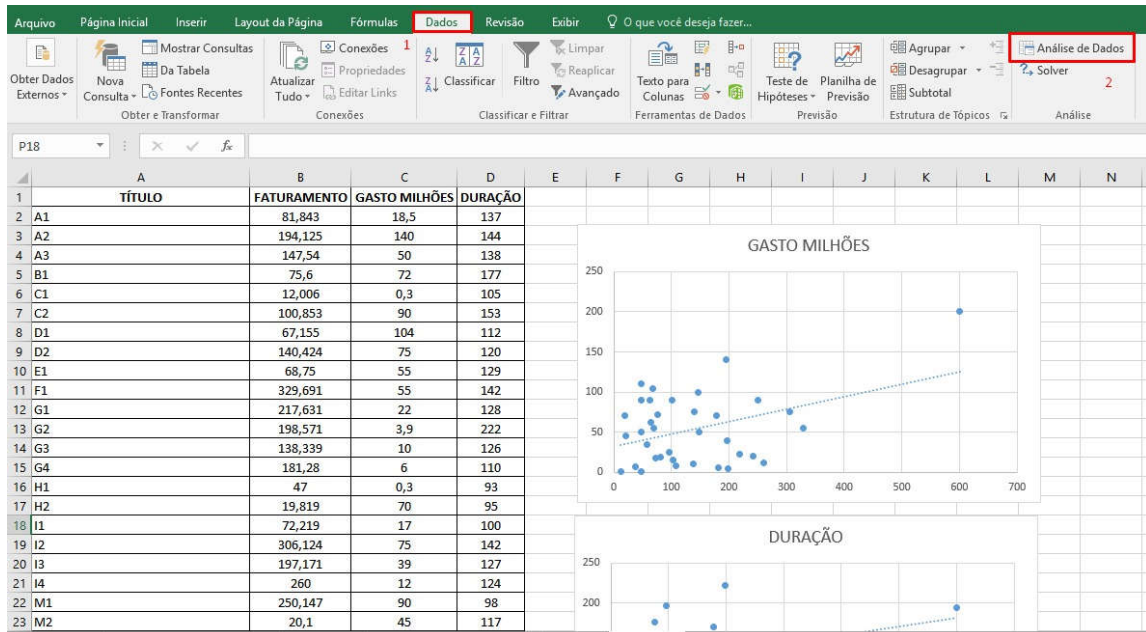


Figura 17. Rodando a RLM

Caso essa ferramenta não esteja habilitada em seu Excel, é possível habilitá-la seguindo os seguintes passos: Arquivo > Opções > Suplementos > Suplementos do Excel > Ir > Ferramentas de Análise > Selecionar Análise de Dados > OK. Para gerar a RLM é

necessário seguir os seguintes passos: Dados > Análise de Dados > Regressão > OK, conforme Figura 18.

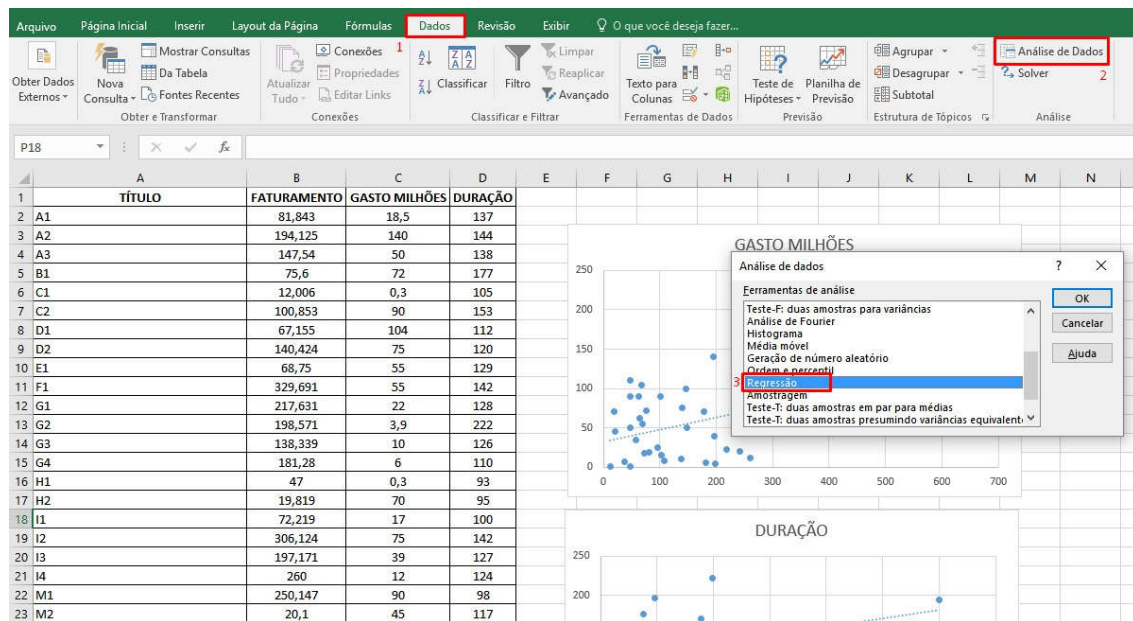


Figura 18. Rodando a RLM no Excel

No “intervalo Y de entrada” deve-se selecionar **toda a coluna da variável dependente**; no “intervalo X de entrada” deve-se selecionar **todas as colunas das variáveis independentes**. É importante selecionar o campo “**rótulos**” para que os nomes de cada coluna estejam visíveis posteriormente. Do mesmo modo, é importante selecionar o campo “**nível de confiança**”, visto que o pesquisador pode alterar o intervalo de confiança, se desejar, conforme Figura 19. O Excel, automaticamente, gera o intervalo padrão de confiança de 95% e assim, se o pesquisador desejar, pode inserir um novo intervalo para comparação. Quanto menores os níveis de confiança, mais estreitos serão os intervalos para conter um determinado parâmetro. Por outro lado, quanto maiores forem os níveis de confiança, maior amplitude terão os intervalos para conter este parâmetro.

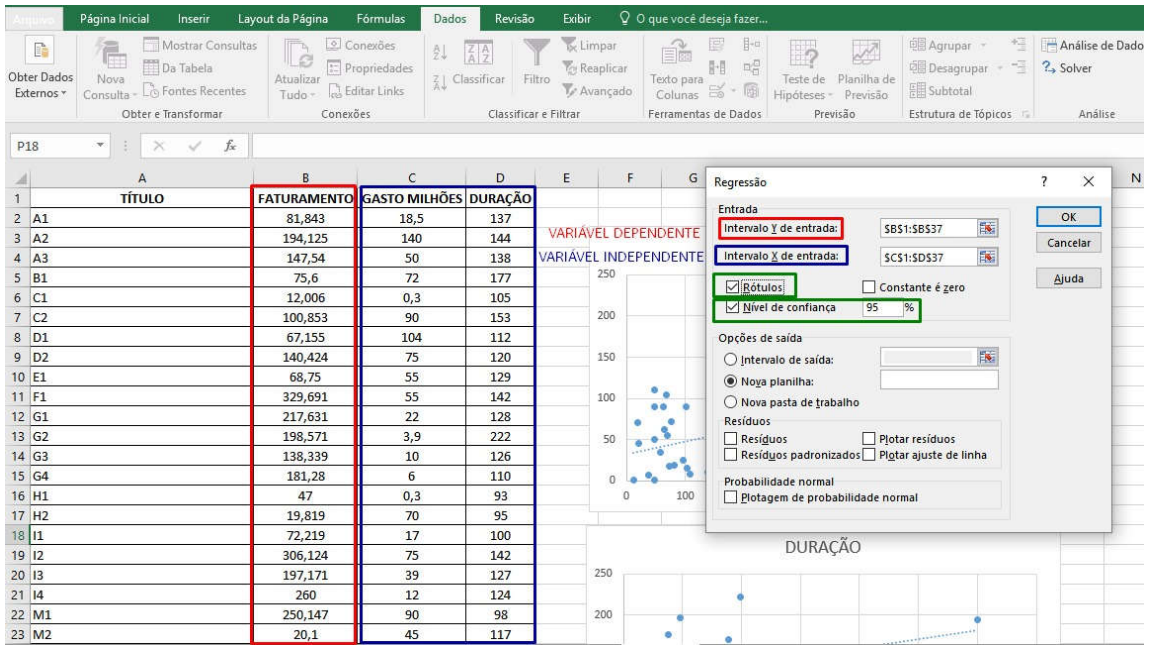


Figura 19. Selecionando as variáveis da regressão

Após, o Excel irá gerar uma nova aba na planilha que está sendo utilizada, conforme a Figura 20.

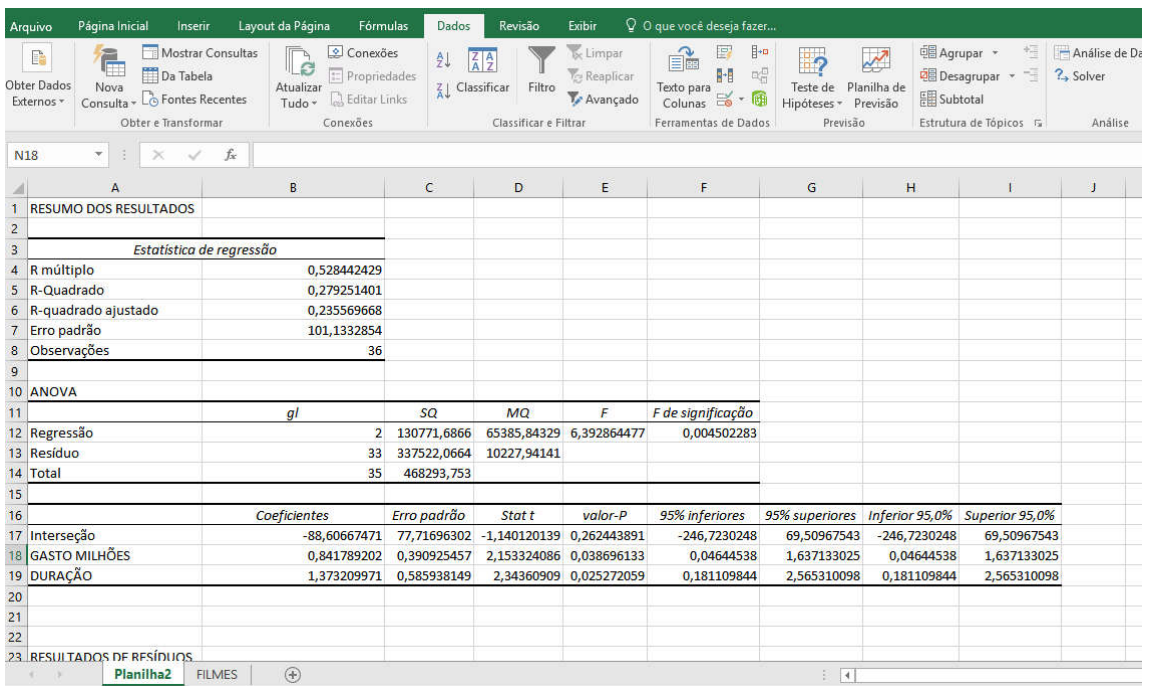


Figura 20. Planilha de resultados da RLM

Abaixo, cada um dos dados apresentados na Figura 20 serão apresentados separadamente, por meio da Figura 21, Figura 22 e Figura 23.

- **ESTATÍSTICA DE REGRESSÃO:**

<i>Estatística de regressão</i>	
R múltiplo	0,528442429
R-Quadrado	0,279251401
R-quadrado ajustado	0,235569668
Erro padrão	101,1332854
Observações	36

Figura 21. Estatística de regressão

O **R múltiplo** é utilizado quando existem vários previsores. /ele representa a correlação entre os valores de Y observados e previstos pelo modelo de regressão múltipla. Valores grandes de R múltiplo (mais próximos de 1) representam alta correlação entre os valores previstos e observados da variável dependente.

O **R²** explica se a relação entre as variáveis é forte ou fraca. Quanto mais perto de 1 for o resultado, mais forte será a relação. Ressalta-se que o R² relaciona todas as variáveis com a variável dependente. Se o R² for igual a 1, o que dificilmente ocorrerá, não haverá resíduos para cada uma das observações da amostra em estudo e a variabilidade da variável Y estará totalmente explicada pelo vetor de variáveis X consideradas no modelo de regressão.

Quando há o intuito de comparar o coeficiente de ajuste (R²) entre dois modelos ou entre um mesmo modelo com tamanhos de amostras diferentes, faz-se necessário o uso do **R² ajustado**, o qual é uma medida do R² da regressão estimada pelo método de mínimos quadrados ordinários ajustada pelo número de graus de liberdade, uma vez que a estimativa amostral de R² tende a superestimar o parâmetro populacional.

O **termo de erro ou resíduo** equivale à diferença entre o valor real de Y e o valor previsto de Y, visto que por meio de “u” é possível capturar o efeito das demais variáveis não incluídas no modelo de regressão utilizado.

O número de **observações** equivale ao número de casos analisados. Neste caso, foram 36 filmes.

- **ANOVA**

ANOVA					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	2	130771,6866	65385,84329	6,392864477	0,004502283
Resíduo	33	337522,0664	10227,94141		
Total	35	468293,753			

Figura 22. Resultados ANOVA

Os **graus de liberdade (gl ou df)** representam a quantidade de informação, fornecida pelos dados, que você pode "gastar" para estimar os valores de parâmetros populacionais desconhecidos, e calcular a variabilidade dessas estimativas. Esse valor é determinado pelo número de observações em sua amostra e o número de parâmetros em seu modelo. Aumentar seu tamanho amostral fornece mais informações sobre a população e, desta forma, aumenta os graus de liberdade em seus dados. Adicionar parâmetros ao seu modelo (aumentando o número de termos em uma equação de regressão, por exemplo) "gasta" informações dos seus dados, e reduz os graus de liberdade disponíveis para estimar a variabilidade das estimativas de parâmetro. Na RLM deve-se estimar um parâmetro para cada termo que você escolha incluir no modelo, e cada um consome um grau de liberdade. Portanto, incluir termos em excesso em um modelo de RLM reduz os graus de liberdade disponíveis para estimar a variabilidade dos parâmetros, e pode torná-lo menos confiável.

O **SQ** é a soma dos quadrados dos desvios totais, que representa a dispersão da variação aleatória de y em relação a sua média y total. O resultado final da SQ é obtido por meio da SQ da Regressão, que é a variação dos valores de y em torno de sua média (explicada pela regressão) + a SQ dos Resíduos, que é diferença entre os valores de y determinados e y' estimados (variação residual não explicada pela regressão).

O **Quadrado Médio (MQ)** é obtido pela divisão da Soma de Quadrados por seus respectivos graus de liberdade.

O **F/ F de significação** relaciona-se ao teste F, o qual avalia se o modelo proposto é útil para explicar a variável dependente, ou seja, busca identificar se pelo menos uma das variáveis independentes está relacionada à variável dependente. Assim, o F de significação deve ser $< 0,005$ para que o modelo seja considerado útil, visto que avalia a significância estatística geral do modelo estimado.

• **RESULTADO DA REGRESSÃO**

	<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>	<i>Inferior 95,0%</i>	<i>Superior 95,0%</i>
Interseção	-88,60667471	77,71696302	-1,140120139	0,262443891	-246,7230248	69,50967543	-246,7230248	69,50967543
GASTO MILHÕES	0,841789202	0,390925457	2,153324086	0,038696133	0,04644538	1,637133025	0,04644538	1,637133025
DURAÇÃO	1,373209971	0,585938149	2,34360909	0,025272059	0,181109844	2,565310098	0,181109844	2,565310098

Figura 23. Resultado da regressão

Os **coeficientes** são os números pelos quais as variáveis da equação serão multiplicadas. Esse valor representa o quanto a variável dependente irá variar quando a respectiva variável independente variar 1 unidade.

O **erro padrão** é o mesmo desvio padrão de uma estimativa. O erro padrão do coeficiente mede o grau de precisão com que o modelo estima o valor desconhecido do coeficiente. Quanto menor o erro padrão, mais precisa é a estimativa.

Dividir o **coeficiente** pelo **erro padrão** calcula o valor-t, ou **stat t**.

O **valor-P** representa o teste de significância individual. Este dado fornece a significância estatística de cada parâmetro a ser considerado no modelo de regressão. Assim, busca saber quais variáveis estão relacionadas com a variável dependente. Para que exista significância, o valor-P deve ser $< 0,05$ (à 95%).

Os **95% inferiores e superiores** equivalem-se ao nível de confiança da regressão. As duas primeiras colunas são geradas automaticamente pelo Excel e as duas últimas são geradas de acordo com a escolha do pesquisador, caso queira comparar o modelo com outro grau de confiança que não seja 95%. Como, neste caso, não foi escolhido um nível de confiança diferente de 95%, o Excel replica as duas primeiras colunas.

Passo 5: Substituir os dados na equação da RLM

Conforme apresentado anteriormente, este passo consiste em substituir os valores encontrados por meio da RLM na equação original, conforme Figura 24 e Figura 25. Abaixo apresenta-se a substituição de valores na equação da RLM.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$
$$\text{FAT} = - 88,6 + 0,84 \cdot \text{GAS} + 1,37 \cdot \text{DUR} + 101,13$$

Estatística de regressão							
R múltiplo		0,528442429					
R-Quadrado		0,279251401					
R-quadrado ajustado		0,235569668					
Erro padrão	u	101,1332854					
Observações		36					

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	2	130771,6866	65385,84329	6,392864477	0,004502283
Resíduo	33	337522,0664	10227,94141		
Total	35	468293,753			

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
Interseção	B0 -88,60667471	77,71696302	-1,140120139	0,262443891	-246,7230248	69,50967543	-246,7230248	69,50967543
GASTO MILHÕES	B1 0,841789202	0,390925457	2,153324086	0,038696133	0,04644538	1,637133025	0,04644538	1,637133025
DURAÇÃO	B2 1,373209971	0,585938149	2,34360909	0,025272059	0,181109844	2,565310098	0,181109844	2,565310098

Figura 24. Substituição de valores na RLM

É importante também analisar a significância de cada variável independente em relação à variável dependente por meio do “valor-P”, conforme segue:

	Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
Interseção	-88,60667471	77,71696302	-1,140120139	0,262443891	-246,7230248	69,50967543	-246,7230248	69,50967543
GASTO MILHÕES	0,841789202	0,390925457	2,153324086	0,038696133	0,04644538	1,637133025	0,04644538	1,637133025
DURAÇÃO	1,373209971	0,585938149	2,34360909	0,025272059	0,181109844	2,565310098	0,181109844	2,565310098

Figura 25. Valor-p (teste t)

O exemplo abaixo, exposto na Figura 26, busca clarear a significância de cada uma das variáveis.

$$Y = - 88,6 + 0,84 \cdot \text{GASTO} + 1,37 \cdot \text{DURAÇÃO} + 101,13$$

0,26

0,03

0,02

SIGNIFICÂNCIA

Figura 26. Significância das variáveis

Passo 6: Interpretação dos resultados

Analisando os resultados, observa-se que ao relacionar as variáveis independentes com a variável dependente no modelo proposto, a relação entre elas pode ser considerada fraca, uma vez que o resultado do **R² foi 0,27**. Além disso, por meio do valor do **erro padrão (101,1)**, é possível que existam variáveis independentes que não estão sendo consideradas no modelo proposto. Ou seja, a diferença entre o valor real de Y e o valor previsto de Y pode ser considerada significativamente alta. Por outro lado, o teste F aponta que o modelo proposto é útil para explicar a variável dependente, visto que o **F de significação foi de 0,004**, mantendo-se abaixo de 0,05. Ainda, por meio do “**valor – P**” observa-se que as variáveis gasto e duração são significativamente relacionadas com a variável dependente, visto que ambas, individualmente, **apresentaram valor < 0,05**, em um intervalo de **95% de confiança**. Por fim, constata-se que a cada aumento da variável gasto, o **valor de Y aumentará 0,84**. Do mesmo modo, para cada aumento na duração do filme, **o valor de Y aumentará 1,37**.

Conforme citado anteriormente, observa-se que outras variáveis podem não estar sendo consideradas neste modelo, visto que o valor de “u” é alto. Em outras palavras, é possível que **outras variáveis além do gasto e do tempo de duração do filme influenciem o faturamento de um filme**.

**** Variáveis Dummy***

A partir da interpretação dos resultados apresentada, acrescentamos aqui uma variável independente relativa ao período de lançamento dos filmes, separando os filmes lançados antes de 1990 dos lançados após 1990. Assim, antes de realizar novamente a RLM, é necessário transformar as variáveis qualitativas em variáveis *dummy*, atribuindo a elas valores numéricos. Neste exemplo, faz-se a pergunta: o filme foi lançado após de 1990? A resposta SIM equivale a 1, enquanto a resposta NÃO equivale à 0. Deste modo, neste modelo, os filmes lançados antes de 1990 estão identificados pelo número 0, enquanto os filmes lançados após 1990 estão identificados com o número 1, conforme Figura 27.

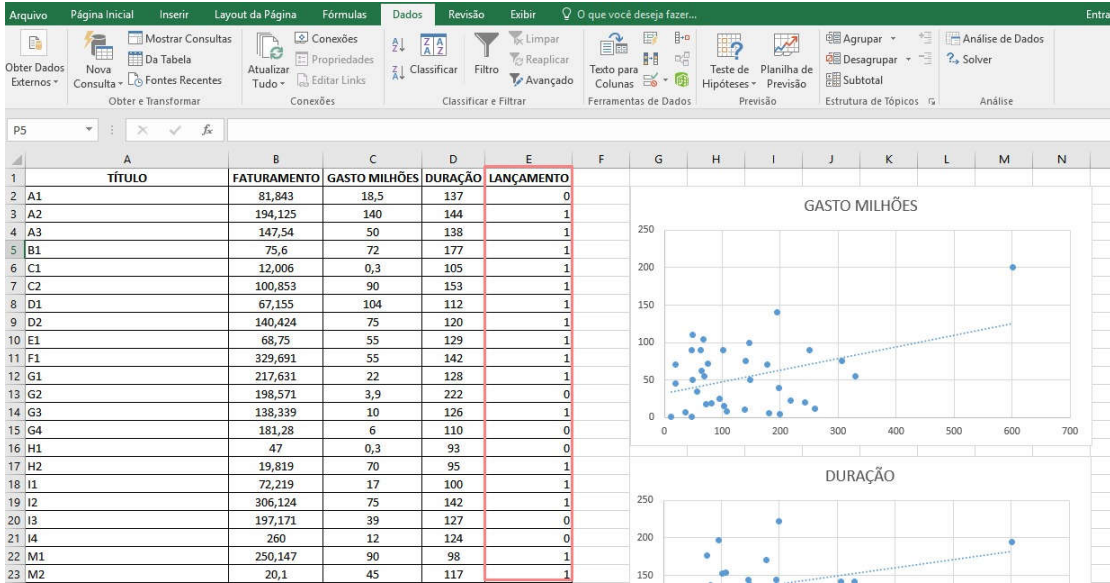


Figura 27. Inclusão da variável lançamento

Após substituir os anos por 0 e 1, é necessário seguir os mesmos passos apresentados anteriormente. O primeiro deles, como já sabemos quais são as variáveis independentes e dependentes, é desenhar o gráfico de dispersão, conforme Figura 28.

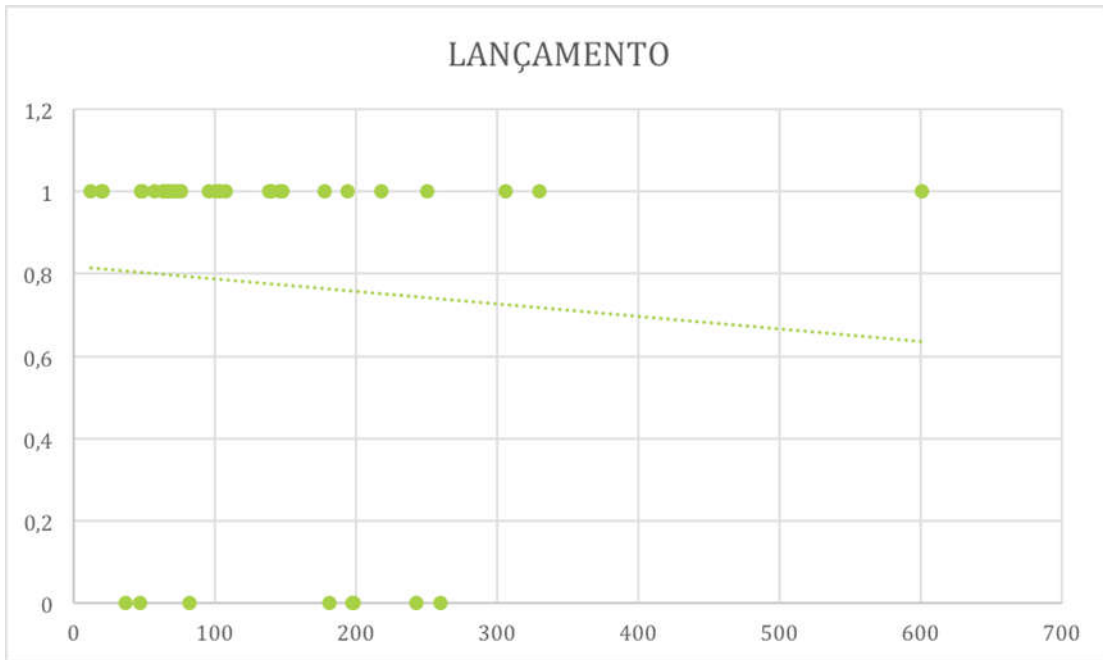


Figura 28. Gráfico de dispersão Faturamento X Lançamento

Após, é necessário gerar a **equação básica** de RLM. Observa-se que, com a inclusão de uma nova variável (lançamento), acrescentou-se também na equação mais uma variável.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + u$$

O próximo passo consiste em realizar a RLM no Excel. O “intervalo Y de entrada” permanece o mesmo, porém no “intervalo X de entrada” é necessário selecionar todas as variáveis independentes, inclusive a *dummy* (lançamento), conforme a Figura 29.

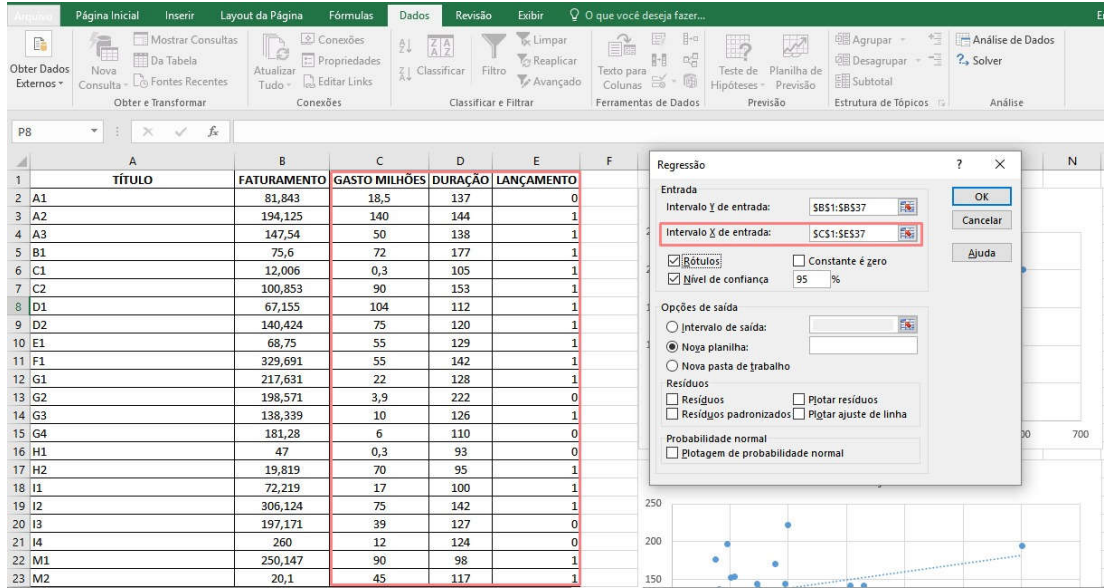


Figura 29. RLM com variável dummy

Os resultados da RLM são apresentados na Figura 30.

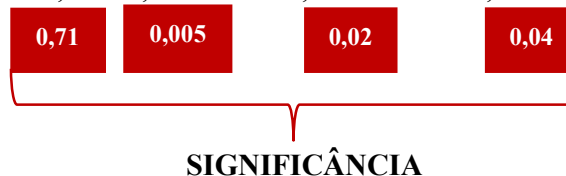
	A	B	C	D	E	F	G	H	I
1	RESUMO DOS RESULTADOS								
2									
3	<i>Estatística de regressão</i>								
4	R múltiplo	0,606485796							
5	R-Quadrado	0,36782502							
6	R-quadrado ajustado	0,308558616							
7	Erro padrão	96,18400234							
8	Observações	36							
9									
10	ANOVA								
11		<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>			
12	Regressão	3	172250,1592	57416,71973	6,206298902	0,001910885			
13	Resíduo	32	296043,5938	9251,362306					
14	Total	35	468293,753						
15									
16		<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>	<i>Inferior 95,0%</i>	<i>Superior 95,0%</i>
17	Interseção	-29,44842078	79,0177091	-0,372681278	0,711842811	-190,4022272	131,5053856	-190,4022272	131,5053856
18	GASTO MILHÕES	1,285654149	0,426817829	3,012184734	0,005035474	0,416254681	2,155053616	0,416254681	2,155053616
19	DURAÇÃO	1,297604692	0,55840613	2,32376513	0,026647909	0,160168628	2,435040757	0,160168628	2,435040757
20	LANÇAMENTO	-93,83568537	44,3158922	-2,117427422	0,042088364	-184,1042038	-3,567166901	-184,1042038	-3,567166901

Figura 30. Resultados da regressão com variável dummy

Com base nos resultados apontados, é necessário novamente substituir a equação de RLM:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + u$$

$$\text{FAT} = -29,44 + 1,28 \cdot \text{GAS} + 1,29 \cdot \text{DUR} - 93,83 \cdot \text{LANÇ} + 96,18$$



Novamente os resultados serão interpretados. Observa-se que, ao relacionar as variáveis independentes com a variável dependente, a relação entre elas continua a ser considerada fraca, uma vez que o resultado do **R² foi 0,36**. Porém, observa-se que com a inclusão de mais uma variável independente a força da relação entre as variáveis aumentou. Os dados novamente apontam que ainda é possível existir outras variáveis independentes que não estão sendo consideradas no modelo, visto que o valor do **erro padrão** ainda é considerado alto (**96,1**). Por meio do **Teste F** observa-se que o modelo proposto é útil para explicar a variável dependente, visto que o **F de significação foi de 0,001**, mantendo-se abaixo de 0,05. Por meio do “valor – P” observa-se que as variáveis **gasto, duração e lançamento** são individualmente significativamente relacionadas com a variável dependente, visto que todas apresentaram valor **< 0,05**, em um intervalo de **95%** de confiança. Por fim, ressalta-se que a cada aumento da variável **gasto, o valor de Y aumentará 1,28**; para cada aumento da duração do filme, **o valor de Y aumentará 1,29** e para cada aumento do ano de lançamento **o valor de Y diminuirá 93,83**. Neste caso o valor diminuirá devido ao resultado negativo apontado na regressão.

Conforme citado anteriormente, **observa-se ainda que outras variáveis podem não estar sendo consideradas neste modelo**. Assim, será acrescentada a variável independente “**faixa etária**”, a qual pode ser: livre, maior que 14 anos e maior que 16 anos. Nas figuras seguintes, cada uma das abas da planilha foi renomeada de acordo com a regressão que está sendo realizada, a fim de nortear o leitor desta apostila. Assim, RLM 1 corresponde à primeira RLM, sem variáveis *dummy*; RLM 2 corresponde à RLM realizada com a variável *dummy* de lançamento e RLM 3 corresponde à RLM que está sendo realizada neste momento.

Conforme apresentado anteriormente, novamente deparamo-nos com uma variável qualitativa, visto que a faixa etária, neste caso, não corresponde à um valor quantitativo. Seguindo o exemplo anterior, poderíamos atribuir os seguintes valores para cada categoria (Tabela 2):

Tabela 2. Categorias Dummy

Categoria da Variável	Valor atribuído
Livre	0
Maior que 14	1
Maior que 16	2

Porém, é válido ressaltar que, diferente do exemplo anterior onde havia apenas duas categorias (lançados antes de 1990 ou lançados após 1990), **esta variável *dummy* apresenta**

três categorias. Assim, segundo Belfiore (2015), quando houver mais de duas categorias em uma variável *dummy*, é possível seguir dois caminhos:

1. Se a variável independente for constante, pode-se utilizar os valores 1, 2, 3, 4, 5, e assim por diante. Porém, observa-se que geralmente a variável independente não é constante. Portanto:
2. Criam-se duas variáveis *dummy*. Por exemplo, opta-se por fazer duas perguntas: o filme é para maiores de 14 anos? e; O filme é para maiores de 16 anos? A resposta **SIM equivale ao número 1** e a resposta **NÃO equivale ao número 0**. Deste modo, quando ambas as perguntas receberem a resposta não, o filme terá uma faixa etária livre (Tabela 3).

Tabela 3. Categorias dummy com mais de duas variáveis

Categoria da Variável	É para maiores de 14?	É para maiores de 16?
Livre	0	0
Maior que 14	1	0
Maior que 16	0	1

Assim, seguindo o procedimento sugerido, as novas colunas inseridas na amostra serão “Faixa etária + 14” e “Faixa Etária + 16”, conforme Figura 31.

	A	B	C	D	E	F	G	H
	TÍTULO	FATURAMENTO	GASTO MILHÕES	DURAÇÃO	LANÇAMENTO	FAIXA ETÁRIA GERAL	É FAIXA + 14	É FAIXA ETÁRIA + 16
2	A1	81,843	18,5	137	0	1	1	0
3	A2	194,125	140	144	1	1	1	0
4	A3	147,54	50	138	1	0	0	0
5	B1	75,6	72	177	1	1	1	0
6	C1	12,006	0,3	105	1	0	0	0
7	C2	100,853	90	153	1	1	1	0
8	D1	67,155	104	112	1	1	1	0
9	D2	140,424	75	120	1	2	0	1
10	E1	68,75	55	129	1	0	0	0
11	F1	329,691	55	142	1	1	1	0
12	G1	217,631	22	128	1	0	0	0
13	G2	198,571	3,9	222	0	0	0	0
14	G3	138,339	10	126	1	1	1	0
15	G4	181,28	6	110	0	0	0	0
16	H1	47	0,3	93	0	2	0	1
17	H2	19,819	70	95	1	0	0	0
18	I1	72,219	17	100	1	2	0	1
19	I2	306,124	75	142	1	1	1	0
20	I3	197,171	39	127	0	0	0	0
21	I4	260	12	124	0	2	0	1
22	M1	250,147	90	98	1	0	0	0
23	M2	20,1	45	117	1	1	1	0

Figura 31. Inclusão das colunas dummy

Novamente os passos apresentados anteriormente serão retomados. O primeiro deles consiste no **gráfico de dispersão**. Neste caso, o gráfico não será gerado, pois não podemos

inferir que há uma relação linear entre mais de duas variáveis. Assim, conforme citado anteriormente, é necessário **gerar a equação básica de RLM**. Observa-se que foram acrescentados o $\beta_4 \cdot X_4$ e o $\beta_5 \cdot X_5$ na equação, os quais correspondem respectivamente às variáveis independentes maior de 14 anos e maior de 16 anos, relacionadas à **faixa etária** do filme.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + u$$

O próximo passo consiste em realizar a RLM no Excel, conforme apresentado anteriormente. O “intervalo Y de entrada” permanece o mesmo, porém no “intervalo X de entrada” é necessário selecionar todas as variáveis independentes, inclusive as *dummy* (lançamento e faixa etária). A Figura 32 demonstra como gerar a RLM e a Figura 33 apresenta os resultados gerados por meio da respectiva regressão.

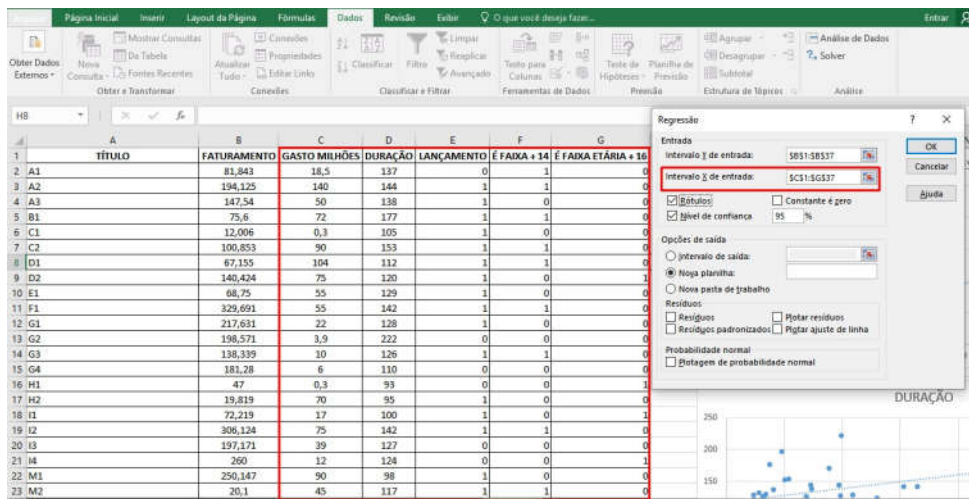


Figura 32. Gerando a RLM com mais de uma variável dummy

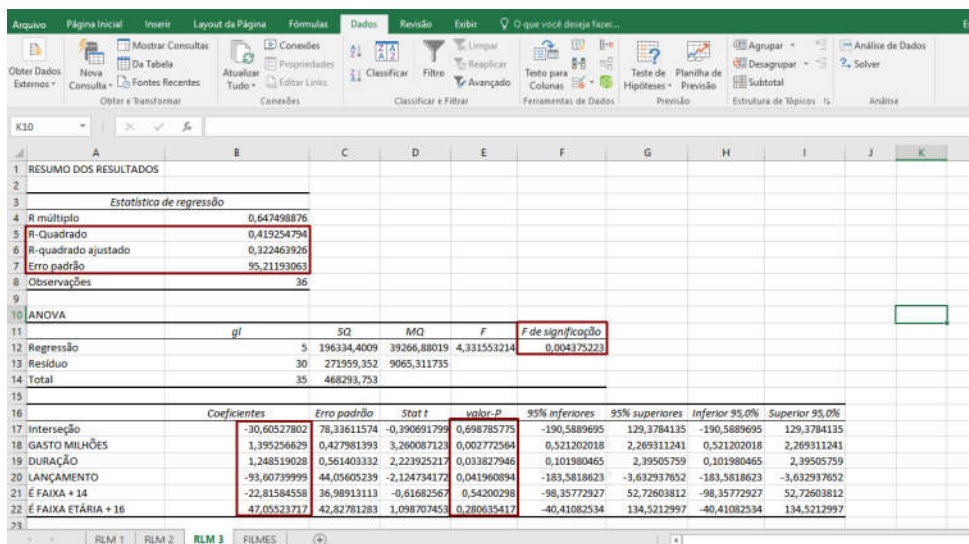
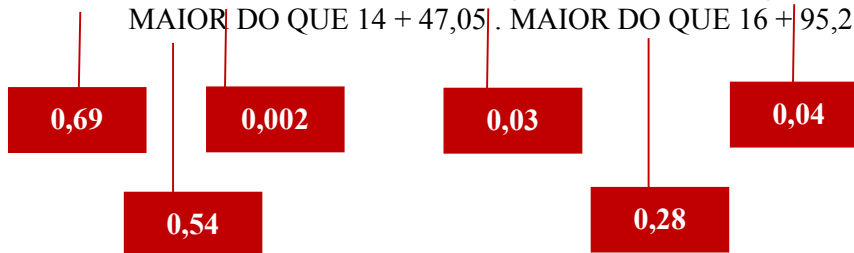


Figura 33. Resultado da RLM com mais de uma variável dummy

Com base nos resultados apontados, é necessário novamente substituir a equação:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + u$$

$$Y = -30,60 + 1,39 \cdot \text{GASTO} + 1,24 \cdot \text{DURAÇÃO} - 93,60 \cdot \text{LANÇAMENTO} - 22,8 \cdot \text{MAIOR DO QUE 14} + 47,05 \cdot \text{MAIOR DO QUE 16} + 95,2$$



Novamente os resultados serão interpretados. Ao relacionar as variáveis independentes com a variável dependente, observa-se que a relação entre elas pode ser considerada fraca, uma vez que o resultado do **R²** foi **0,41**. Porém, observa-se que com a inclusão de mais uma variável, a força da relação entre as variáveis aumentou. Ainda é possível que existam outras variáveis independentes que não estão sendo consideradas no modelo proposto, visto que o valor do **erro padrão** ainda é considerado alto (**95,2**). Porém, observa-se que com a inclusão de outras variáveis o valor do erro padrão diminuiu. Por meio do Teste F observa-se que o modelo é útil para explicar a variável dependente, visto que o **F de significação** foi de **0,004**, mantendo-se ainda abaixo de 0,05. Por meio do “valor – P” observa-se que apenas as variáveis **gasto, duração e lançamento** são **significativamente relacionadas com a variável dependente**, visto que estas apresentaram valor < 0,05 em um intervalo de **95%** de confiança. As variáveis relacionadas à faixa etária não apresentaram alta significância. Por fim, ressalta-se que a cada aumento da variável gasto, **o valor de Y aumentará 1,39**; para cada aumento

da duração do filme, o valor de Y aumentará 1,24; para cada aumento do ano de lançamento, o valor de Y diminuirá 93,60; para cada faixa etária maior de 14 anos, o valor de Y diminuirá 22,8 e para cada faixa etária maior de 16 anos, o valor de Y diminuirá 47,05. Para saber a relação com a variável **faixa etária livre**, é necessário substituir β_4 e β_5 por 0.

Com base no modelo proposto, ainda há um **próximo passo** a ser seguido, o qual consiste em verificar a existência ou não de **multicolinearidade** entre as variáveis independentes, ou seja, verificar se existe alta relação entre as variáveis independentes. Embora as variáveis independentes tenham relação com a variável dependente, quando existe alta correlação entre as variáveis independentes o R^2 tende a diminuir e assim é necessário retirá-las da equação. Para verificar a multicolinearidade do modelo proposto é necessário realizar a **Matriz de Correlação**, disponível em Dados > Análise de Dados > Correlação, conforme apresentado na Figura 34.

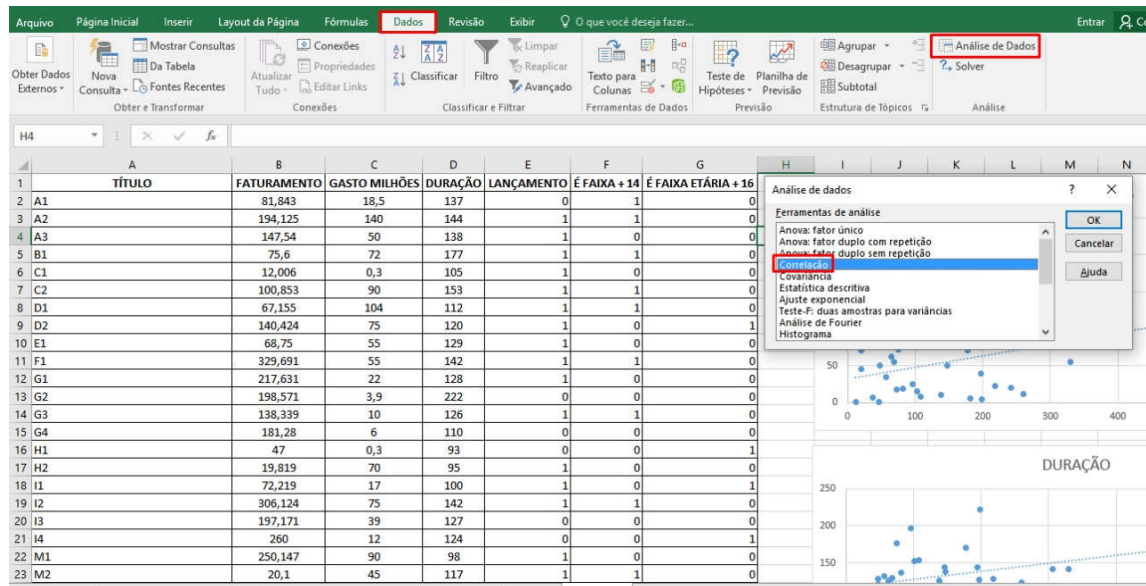


Figura 34. Gerando a matriz de correlação

O “intervalo Y de entrada” corresponde à variável dependente, enquanto o “intervalo X de entrada” corresponde à todas as variáveis independentes, conforme Figura 35.

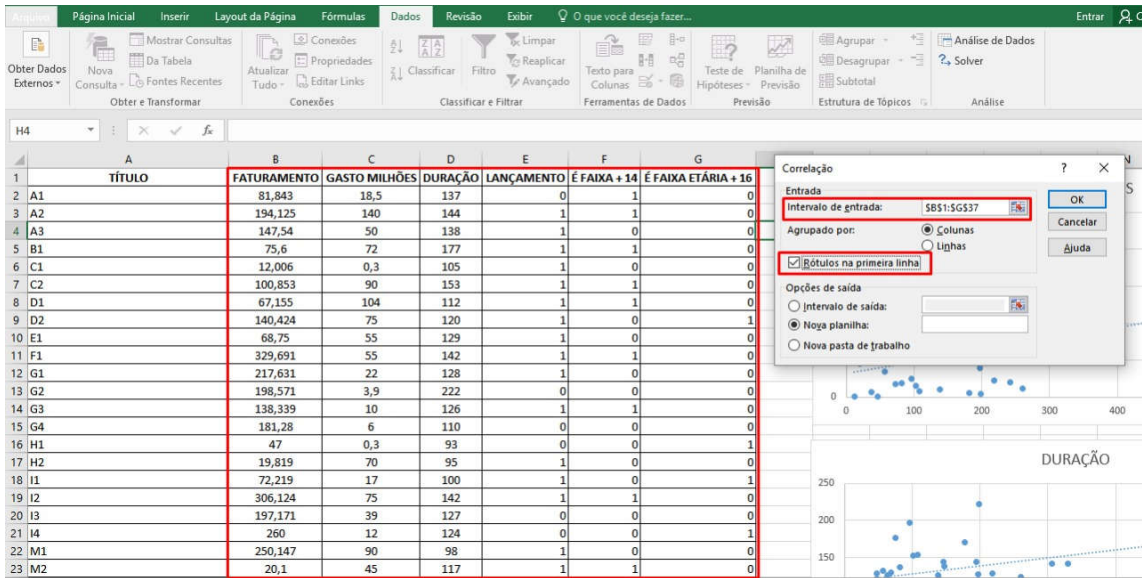


Figura 35. Intervalos de entrada para matriz de correlação

A partir da geração da correlação, o Excel irá gerar uma nova planilha, conforme a Figura 36, que apresentará a Matriz de Correlação.

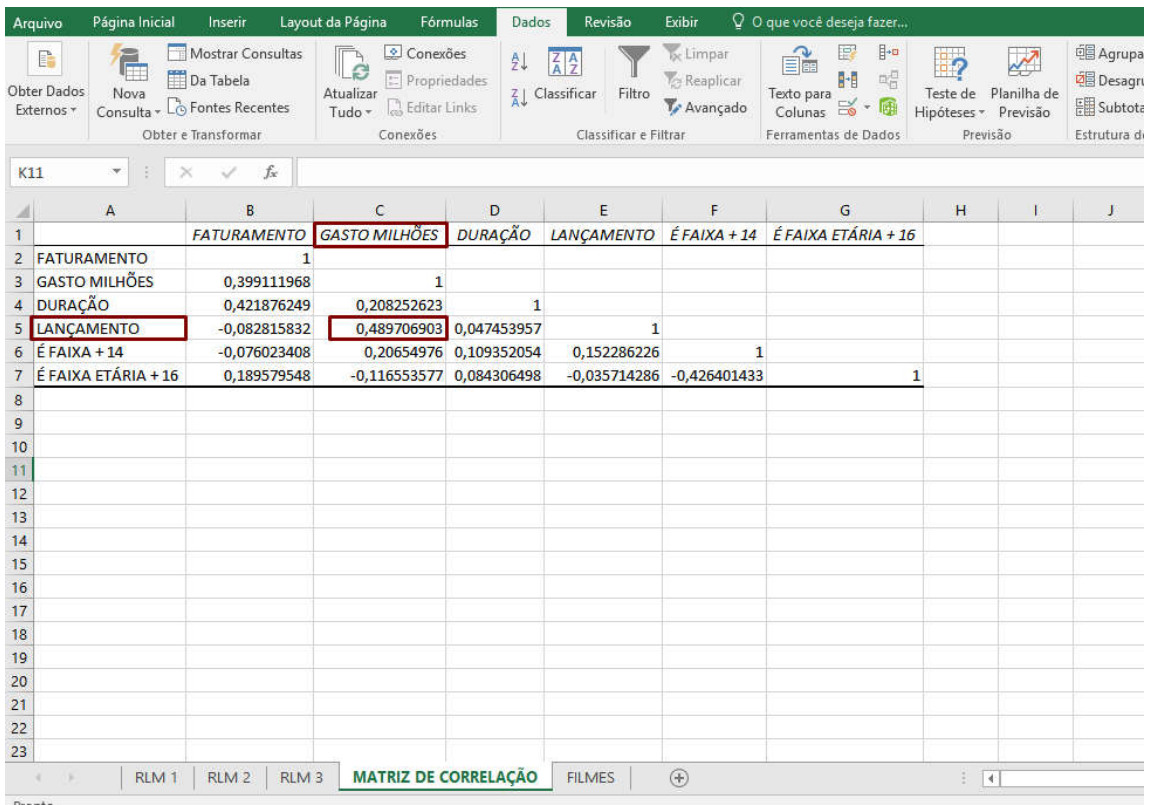


Figura 36. Matriz de correlação

Caso exista uma correlação entre as variáveis independentes maior do que 0,6 é necessário excluí-las do modelo. No exemplo que está sendo utilizado, houve um valor

próximo à 0,6 entre lançamento e gasto, sugerindo uma possível multicolinearidade. Assim, iremos testar alguns modelos para análise, verificando qual deles apresenta o maior R^2 ajustado, já que este é utilizado quando há o intuito de comparar o coeficiente de ajuste (R^2) entre dois modelos ou entre um mesmo modelo com tamanhos de amostras diferentes.

Escolhendo o melhor modelo de regressão...

Modelo 1 – Considerando todas as variáveis

O **Modelo 1** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas. Observa-se que este modelo já foi realizado no item anterior (RLM3) e apresentou **R^2 ajustado de 0,32**. Ou seja, as cinco variáveis independentes **explicam juntas 32%** da variável dependente. A Figura 37 apresenta os resultados desta RLM, a qual considera todas as variáveis.

RESUMO DOS RESULTADOS									
Estatística de regressão									
R múltiplo	0,647498876								
R-Quadrado	0,419254794								
R-quadrado ajustado	0,322463926								
Erro padrão	95,21193063								
Observações	36								
ANOVA									
	gl	SQ	MQ	F	F de significação				
Regressão	5	196334,4009	39266,88019	4,331553214	0,004375223				
Resíduo	30	271959,352	9065,311735						
Total	35	468293,753							
	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%	
Interseção	-30,60527802	78,33611574	-0,390691799	0,698785775	-190,5889695	129,3784135	-190,5889695	129,3784135	
GASTO MILHÕES	1,395256629	0,427981393	3,260087123	0,002772564	0,521202018	2,269311241	0,521202018	2,269311241	
DURAÇÃO	1,248519028	0,561403332	2,223925217	0,033827946	0,101980465	2,39505759	0,101980465	2,39505759	
LANÇAMENTO	-93,60739999	44,05605239	-2,124734172	0,041960894	-183,5818623	-3,632937652	-183,5818623	-3,632937652	
É FAIXA + 14	-22,81584558	36,98913113	-0,61682567	0,54200298	-98,35772927	52,72603812	-98,35772927	52,72603812	
É FAIXA ETÁRIA + 16	47,05523717	42,82781283	1,098707453	0,280635417	-40,41082534	134,5212997	-40,41082534	134,5212997	

Figura 37 Resultados da RLM do modelo 1

Modelo 2 – Considerando todas as variáveis exceto faixa etária

O **Modelo 2** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, exceto a faixa etária, visto que esta não apresentou alta significância anteriormente (valor-P). Observa-se que este modelo já foi realizado (RLM2) e

apresentou **R² ajustado de 0,3**. Ou seja, as variáveis independentes **explicam juntas 30%** da variável dependente. A Figura 38 apresenta os resultados desta RLM, a qual considera todas as variáveis, exceto a faixa etária.

Estadística de regressão									
R múltiplo	0,606485796								
R-Quadrado	0,36782502								
R-quadrado ajustado	0,308558616								
Erro padrão	96,18400234								
Observações	36								

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	3	172250,1592	57416,71973	6,206298902	0,001910885
Resíduo	32	296043,5938	9251,362306		
Total	35	468293,753			

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
Interseção	-29,44842078	79,0177091	-0,372681278	0,711842811	-190,4022272	131,5053856	-190,4022272	131,5053856
GASTO MILHÕES	1,285654149	0,426817829	3,012184734	0,005035474	0,416254681	2,155053616	0,416254681	2,155053616
DURAÇÃO	1,297604692	0,55840613	2,32376513	0,026647909	0,160168628	2,435040757	0,160168628	2,435040757
LANÇAMENTO	-93,83568537	44,3158922	-2,117427422	0,042088364	-184,1042038	-3,567166901	-184,1042038	-3,567166901

Figura 38. Resultados da RLM do modelo 2

Modelo 3 – Considerando todas as variáveis, exceto gasto

O **Modelo 3** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, **exceto** gasto. Este modelo será aplicado, uma vez que se observou certa **multicolinearidade** entre as variáveis **gasto e lançamento**. Assim iremos retirá-las individualmente de cada modelo, a fim de analisar as possíveis diferenças no R² ajustado. A Figura 39 apresenta a RLM sendo gerada e a Figura 40 apresenta os resultados desta RLM, a qual considera todas as variáveis, exceto gasto.

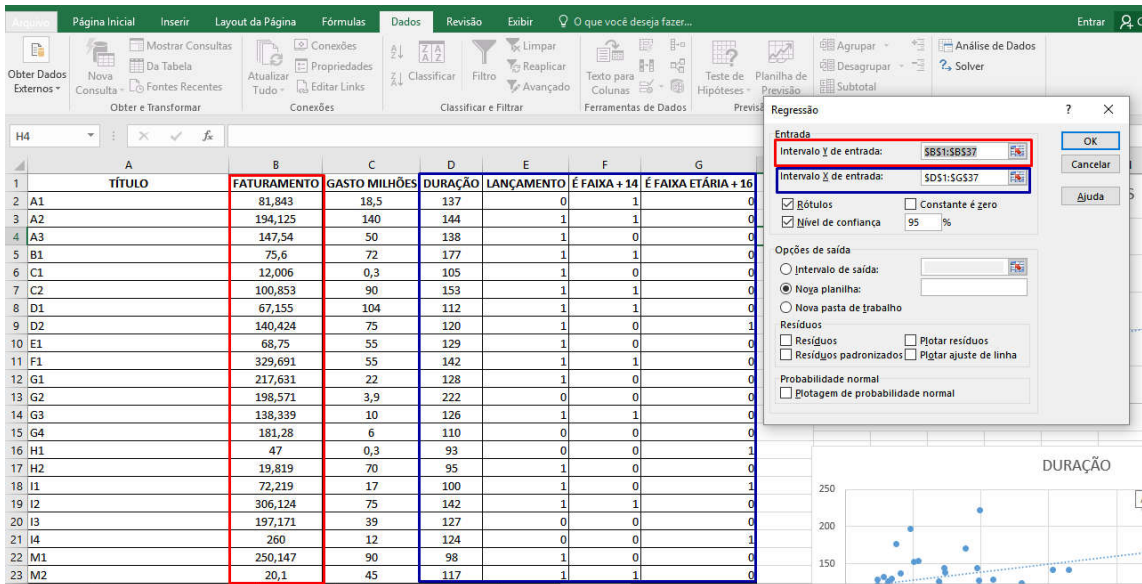


Figura 39. Gerando a RLM do modelo 3

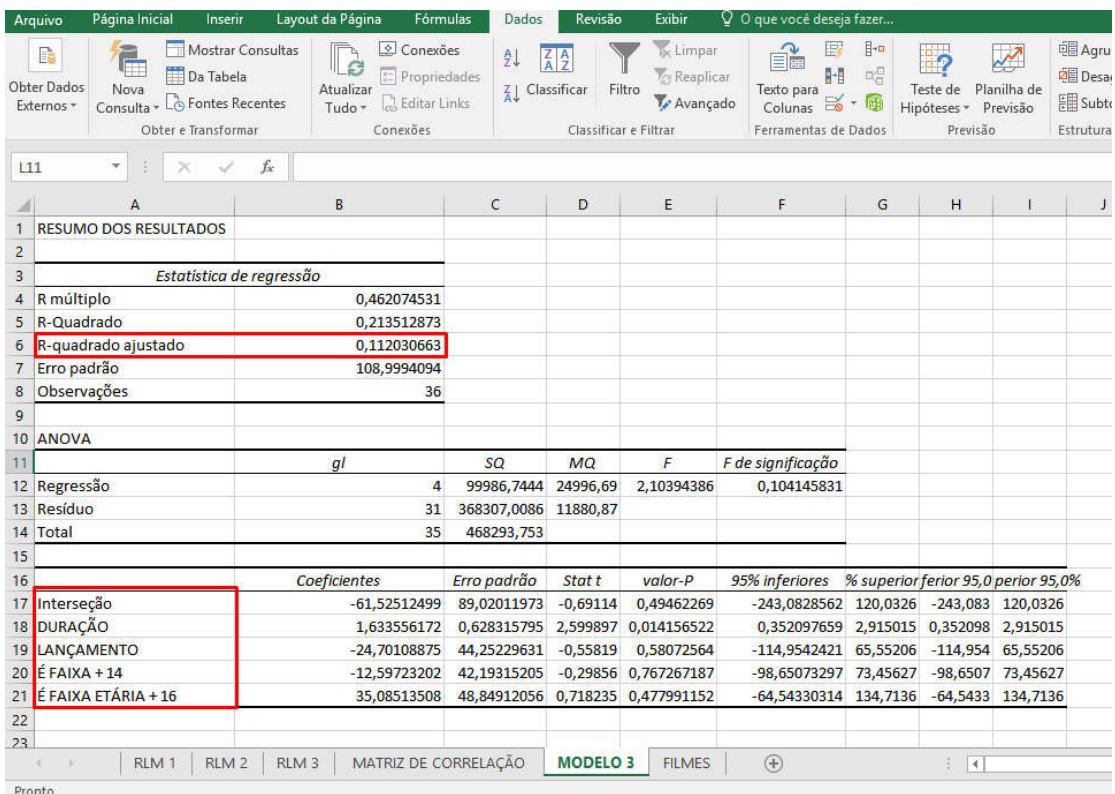


Figura 40. Resultados da RLM do modelo 3

Observa-se que neste modelo o R^2 ajustado foi de 0,11. Ou seja, as variáveis independentes explicam juntas 11% da variável dependente.

Modelo 4 – Considerando todas as variáveis, exceto lançamento

O **Modelo 4** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, **exceto** lançamento. Este modelo será aplicado, uma vez que se observou certa **multicolinearidade** entre as variáveis **gasto e lançamento**. Assim iremos retirá-las individualmente de cada modelo, a fim de analisar as possíveis diferenças no R^2 ajustado. Para realizar a regressão, é necessário **modificar o local das colunas** para que todas as variáveis independentes permaneçam lado a lado, conforme Figura 41. A Figura 42 demonstra a RLM do Modelo 4 sendo gerada e a Figura 43 apresenta os resultados desta RLM, considerando todas as variáveis, exceto lançamento.

	A	B	C	D	E	F	G	H	I
1	TÍTULO	FATURAMENTO	GASTO MILHÕES	DURAÇÃO		É FAIXA + 14	É FAIXA ETÁRIA + 16	LANÇAMENTO	
2	A1	81,843	18,5	137		1	0	0	
3	A2	194,125	140	144		1	0	1	
4	A3	147,54	50	138		0	0	1	
5	B1	75,6	72	177		1	0	1	
6	C1	12,006	0,3	105		0	0	1	
7	C2	100,853	90	153		1	0	1	
8	D1	67,155	104	112		1	0	1	
9	D2	140,424	75	120		0	1	1	
10	E1	68,75	55	129		0	0	1	
11	F1	329,691	55	142		1	0	1	
12	G1	217,631	22	128		0	0	1	
13	G2	198,571	3,9	222		0	0	0	
14	G3	138,339	10	126		1	0	1	
15	G4	181,28	6	110		0	0	0	
16	H1	47	0,3	93		0	1	0	
17	H2	19,819	70	95		0	0	1	
18	I1	72,219	17	100		0	1	1	
19	I2	306,124	75	142		1	0	1	
20	I3	197,171	39	127		0	0	0	
21	I4	260	12	124		0	1	0	
22	M1	250,147	90	98		0	0	1	
23	M2	20,1	45	117		1	0	1	

Figura 41. Reorganização das colunas para RLM do modelo 4

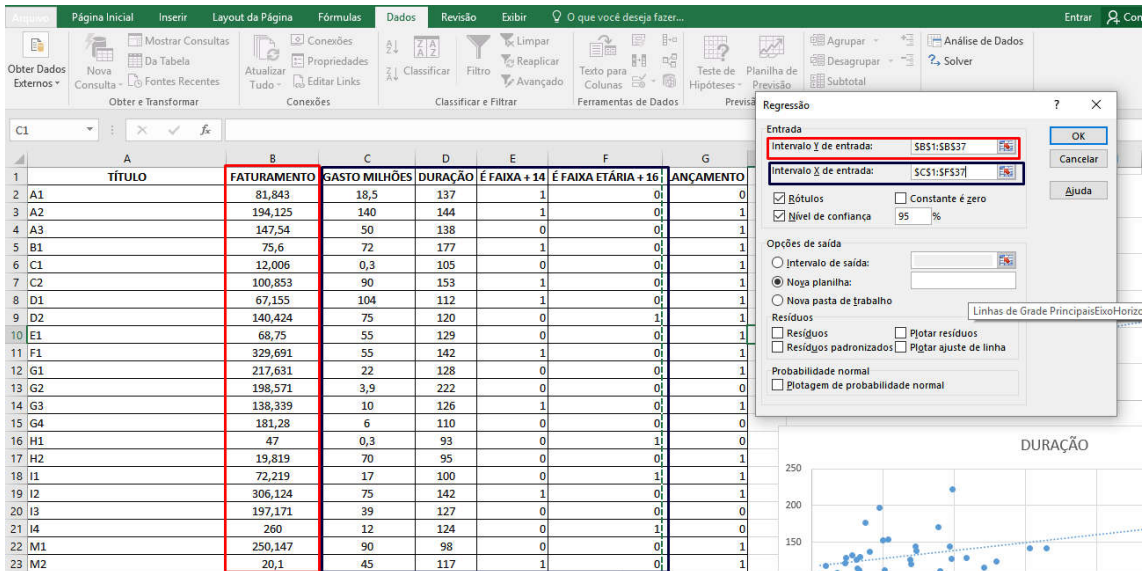


Figura 42. Gerando a RLM do modelo 4

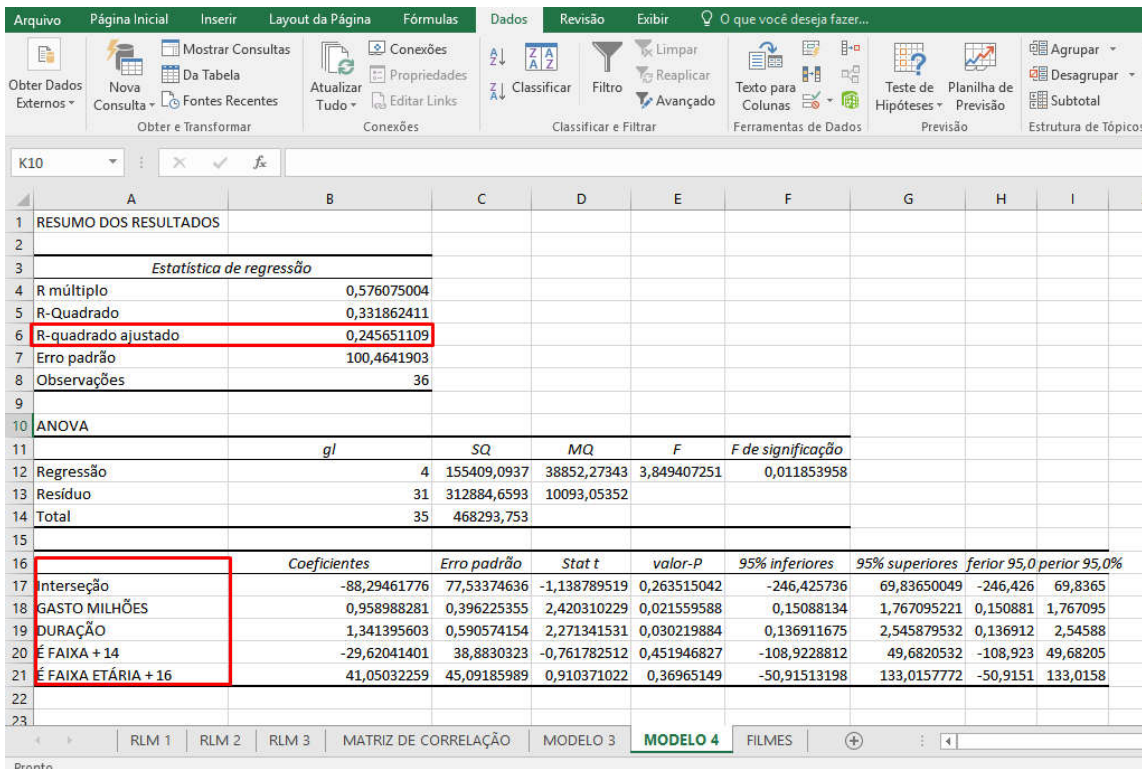


Figura 43. Resultados da RLM do modelo 4

Observa-se que neste modelo o R^2 ajustado foi de 0,24. Ou seja, as variáveis independentes explicam juntas 24% da variável dependente.

Modelo 5 – Considerando apenas a variável duração

O **Modelo 5** considera a relação entre a variável dependente e a variável independente de duração, somente. Assim, podemos descobrir, por meio de uma RLS, o quanto essa variável independente, sozinha, explica o faturamento dos filmes. A Figura 44 demonstra a RLS do Modelo 5 sendo gerada e a Figura 45 apresenta os resultados desta RLS, considerando apenas a variável duração.

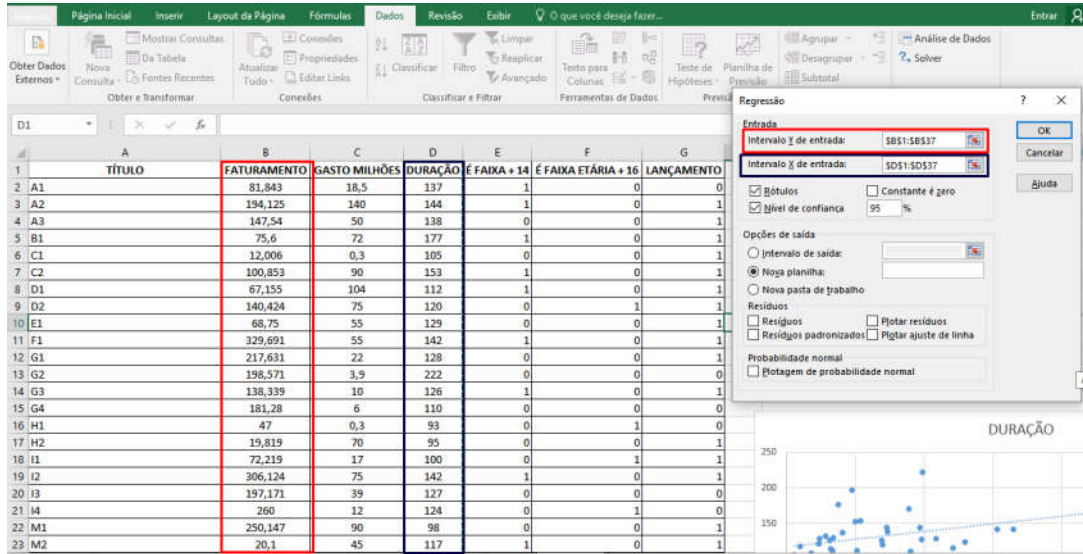


Figura 44. Gerando a RLM do modelo 5

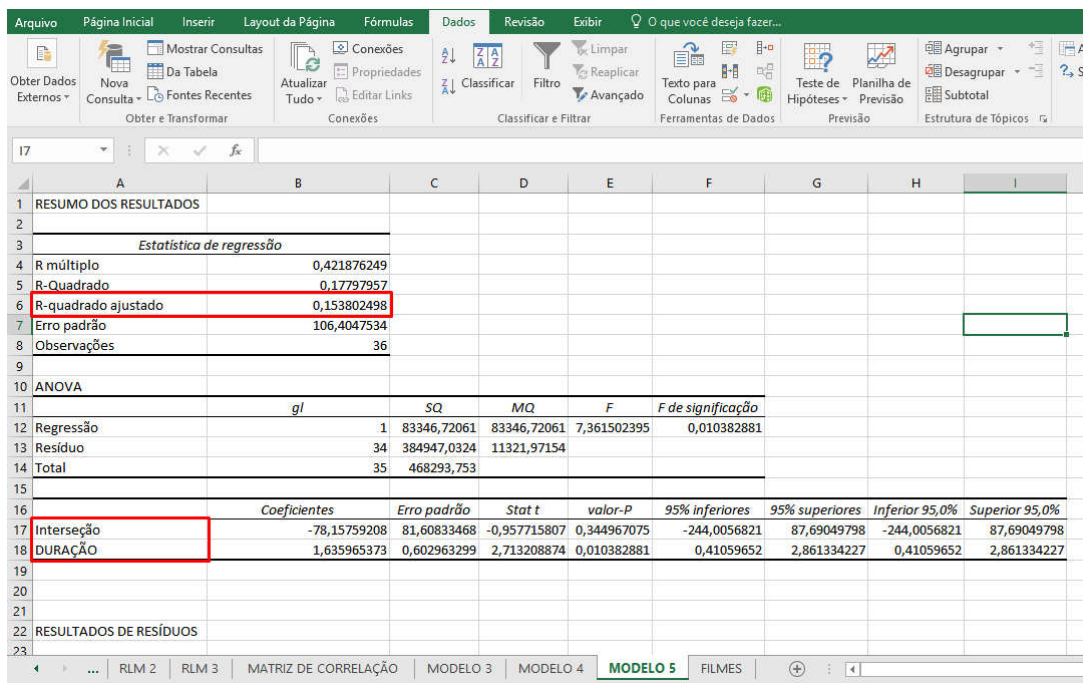


Figura 45. Resultados da RLM do modelo 5

Observa-se que neste modelo o **R² ajustado foi de 0,15**. Ou seja, a variável de duração **explica sozinha 15%** da variável dependente. Esse é um R² alto, quando comparado aos outros modelos gerados. No Modelo 1, em que todas as variáveis foram inseridas, o valor do R² foi de 0,32, ou seja, a duração é responsável por explicar metade do R², sozinha, e portanto, é uma variável muito importante.

Outros modelos também poderiam ser testados, e a escolha dependerá do objetivo de cada pesquisador.

Utilizando o SPSS

O SPSS é um software criado pela IBM para análises estatísticas nas Ciências Sociais. Não é nosso objetivo central ensinar a usá-lo, mas algumas dicas são necessárias para darmos continuidade às nossas análises.

Se você já tem uma base de dados digitada em outro programa, ou mesmo no próprio SPSS, pode importá-la clicando 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, conforme mostram as Figuras 46 e 47.

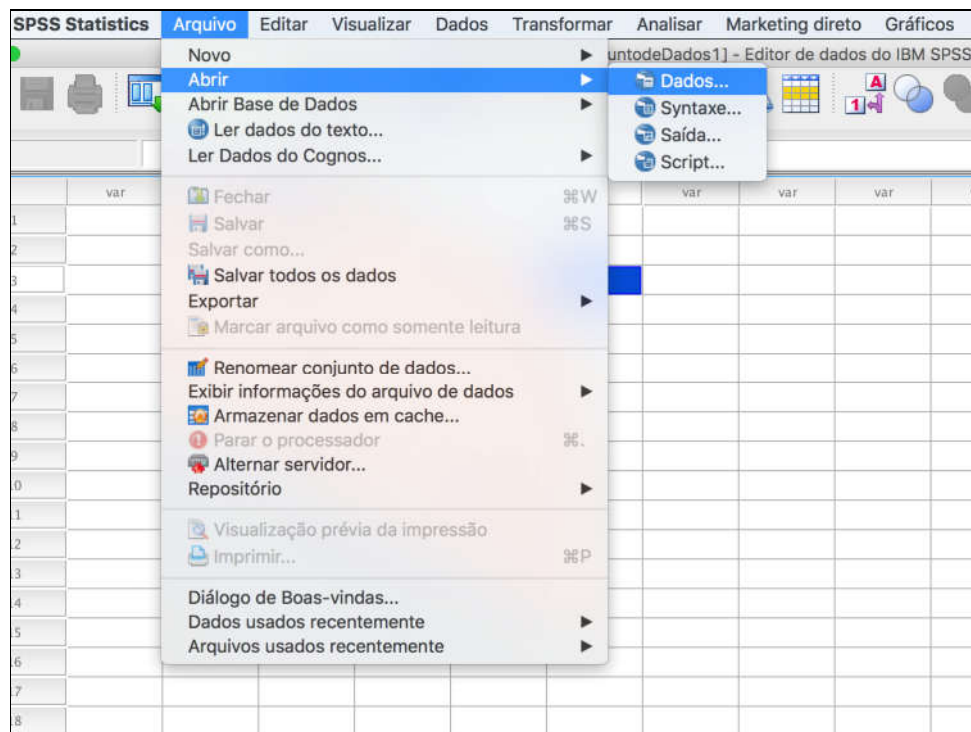


Figura 46. Abrir ou importar arquivo no SPSS

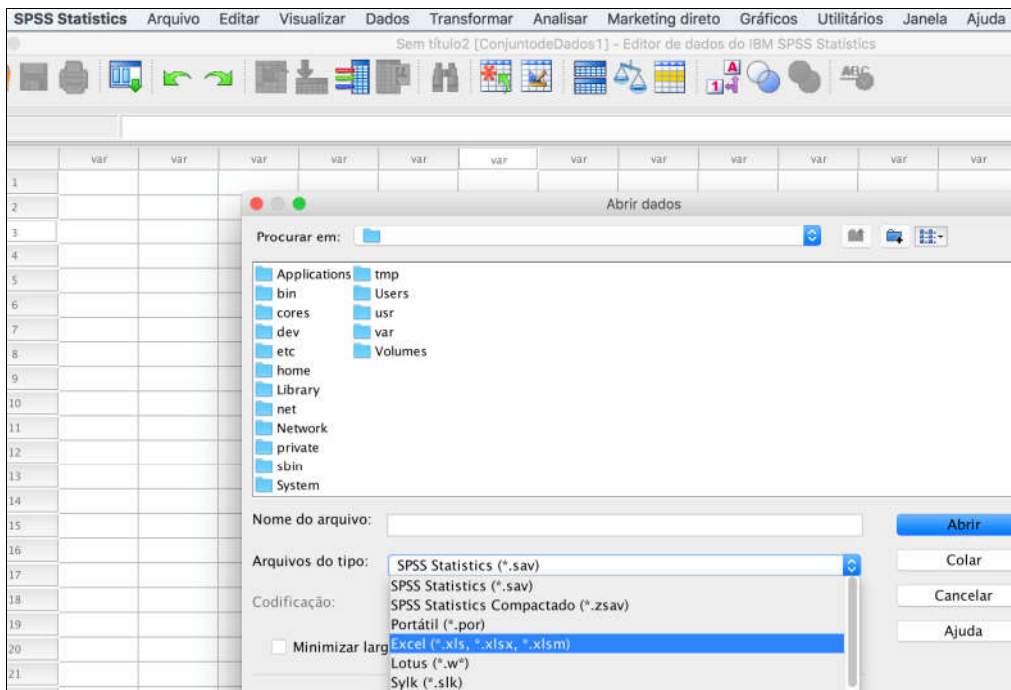


Figura 47. Selecionando a extensão do arquivo

No nosso caso, como já realizamos a mesma análise no Excel, apenas importaremos os dados. Mas você também pode digitá-los direto no SPSS, caso desejar (Figura 48).

	TÍTULO	FATURAMENTO	GASTOMILHOES	DURAÇÃO
1	A1	81,84	18,50	137,00
2	A2	194,13	140,00	144,00
3	A3	147,54	50,00	138,00
4	B1	75,60	72,00	177,00
5	C1	12,01	,30	105,00
6	C2	100,85	90,00	153,00
7	D1	67,16	104,00	112,00
8	D2	140,42	75,00	120,00
9	E1	68,75	55,00	129,00
10	F1	329,69	55,00	142,00
11	G1	217,63	22,00	128,00
12	G2	198,57	3,90	222,00
13	G3	138,34	10,00	126,00
14	G4	181,28	6,00	110,00
15	H1	47,00	,30	93,00
16	H2	19,82	70,00	95,00
17	I1	72,22	17,00	100,00
18	I2	306,12	75,00	142,00
19	I3	197,17	39,00	127,00
20	I4	260,00	12,00	124,00
21	M1	250,15	90,00	98,00
22	M2	20,10	45,00	117,00
23	P1	107,93	8,00	154,00

Figura 48. Digitando os dados no SPSS

Passo 1: Definição das Variáveis (Figura 49)

	TÍTULO	FATURAMENTO	GASTOMILHOES	DURAÇÃO
1	A1	81,84	18,50	137,00
2	A2	194,13	140,00	144,00
3	A3			
4	B1			
5	C1			
6	C2	100,85	90,00	153,00
7	D1	67,16	104,00	112,00
8	D2	140,42	75,00	120,00
9	E1	68,75	55,00	129,00
10	F1	329,69	55,00	142,00
11	G1	217,63	22,00	128,00
12	G2	198,57	3,90	222,00
13	G3	138,34	10,00	126,00
14	G4	181,28	6,00	110,00
15	H1	47,00	,30	93,00
16	H2	19,82	70,00	95,00
17	I1	72,22	17,00	100,00
18	I2	306,12	75,00	142,00
19	I3	197,17	39,00	127,00
20	I4	260,00	12,00	124,00
21	M1	250,15	90,00	98,00
22	M2	20,10	45,00	117,00
23	P1	107,93	8,00	154,00

Figura 49. Definição das variáveis no SPSS

Passo 2: Desenho do gráfico de dispersão

Para gerar o gráfico de dispersão, é necessário seguir o passo a passo descrito a seguir (Figura 50): Gráficos > Construtor de Gráfico > Ok.

	TÍTULO	FATURAMENTO	GASTOMILHOES	DURAÇÃO	LAN
1	A1	81,84	18,50	137,00	
2	A2	194,13	140,00	144,00	
3	A3	147,54	50,00	138,00	1,00
4	B1	75,60	72,00	177,00	1,00
5	C1	12,01	,30	105,00	1,00
6	C2	100,85	90,00	153,00	1,00
7	D1	67,16	104,00	112,00	1,00
8	D2	140,42	75,00	120,00	1,00
9	E1	68,75	55,00	129,00	1,00
10	F1	329,69	55,00	142,00	1,00
11	G1	217,63	22,00	128,00	1,00
12	G2	198,57	3,90	222,00	,00
13	G3	138,34	10,00	126,00	1,00
14	G4	181,28	6,00	110,00	,00
15	H1	47,00	,30	93,00	,00
16	H2	19,82	70,00	95,00	1,00
17	I1	72,22	17,00	100,00	1,00
18	I2	306,12	75,00	142,00	1,00
19	I3	197,17	39,00	127,00	,00
20	I4	260,00	12,00	124,00	,00
21	M1	250,15	90,00	98,00	1,00
22	M2	20,10	45,00	117,00	1,00
23	P1	107,93	8,00	154,00	1,00

Figura 50. Construtor de gráfico no SPSS

Para escolher o tipo de gráfico é necessário seguir o passo a passo descrito:

- Galeria > Dispersão/Ponto > Primeira opção (Gráfico Disperso Simples).

Depois, precisamos definir os eixos do gráfico. Para isso é necessário arrastar com o mouse a variável *faturamento* para o eixo Y e a variável *gastoemmilhoes* para o eixo X > Ok (Figura 51). O gráfico será gerado na tela de saída de dados (Figura 52).

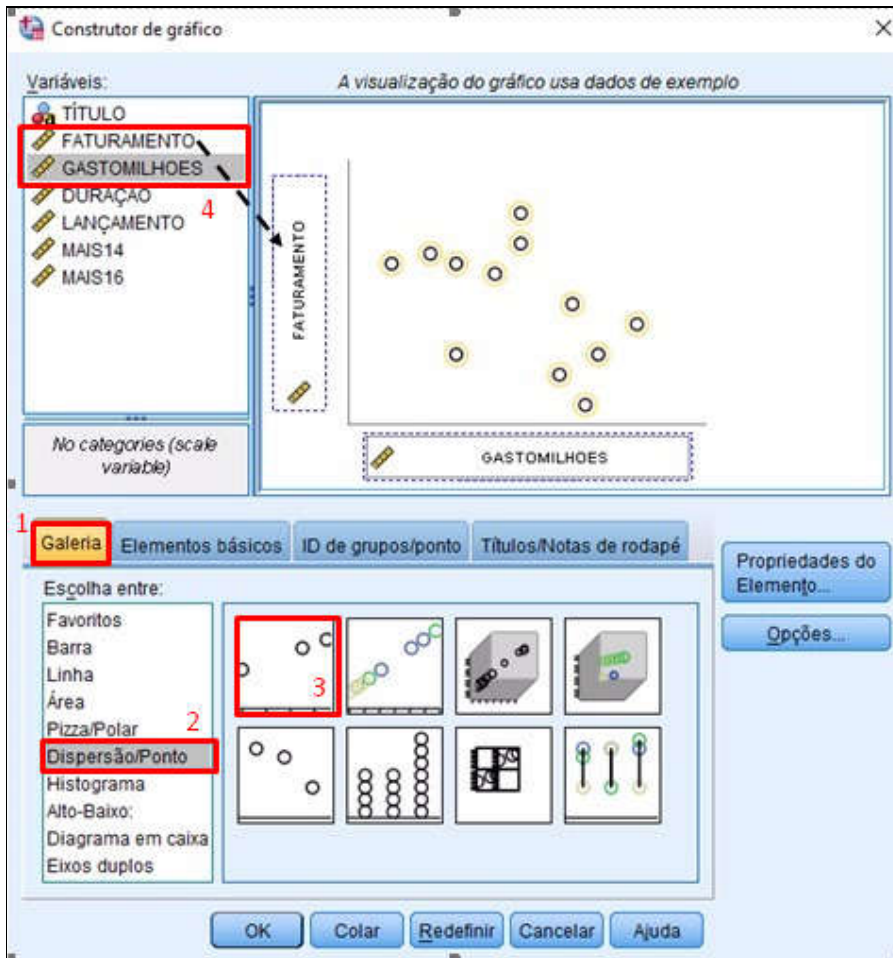


Figura 51. Gerando gráfico no SPSS

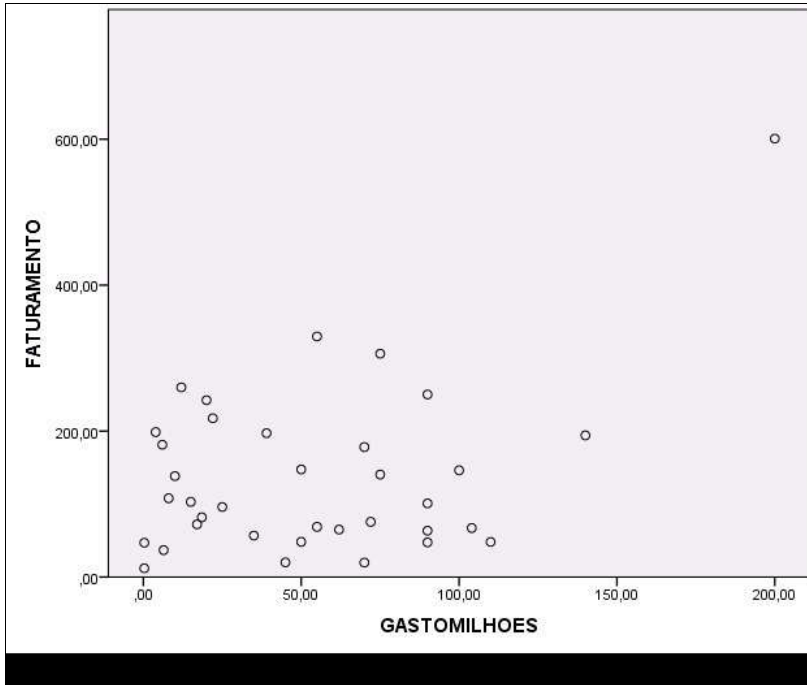


Figura 52. Gráfico de Dispersão Faturamento X GastoMilhões

Observa-se que não é possível interpretar a relação entre as duas variáveis apenas com os pontos dispersos no gráfico. Por isso, gera-se a linha de tendência. Para gerar a linha de tendência é necessário clicar duas vezes sob o gráfico e clicar no ícone “adicionar linha de ajuste no total” (Figura 53).

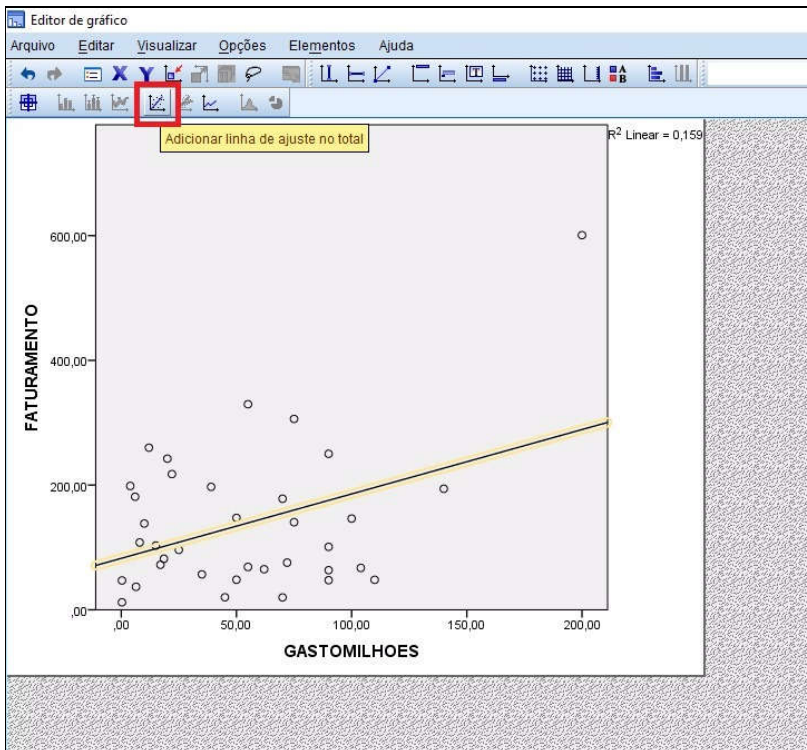


Figura 53. Gerando linha de tendência

É necessário selecionar o tipo de linha que se quer visualizar. Neste caso, o método de ajuste é linear, visto que busca-se descobrir qual relação entre a variável dependente e independente. É necessário desmarcar a opção “anexar rótulo à linha” e clicar em “aplicar”, conforme a Figura 54.

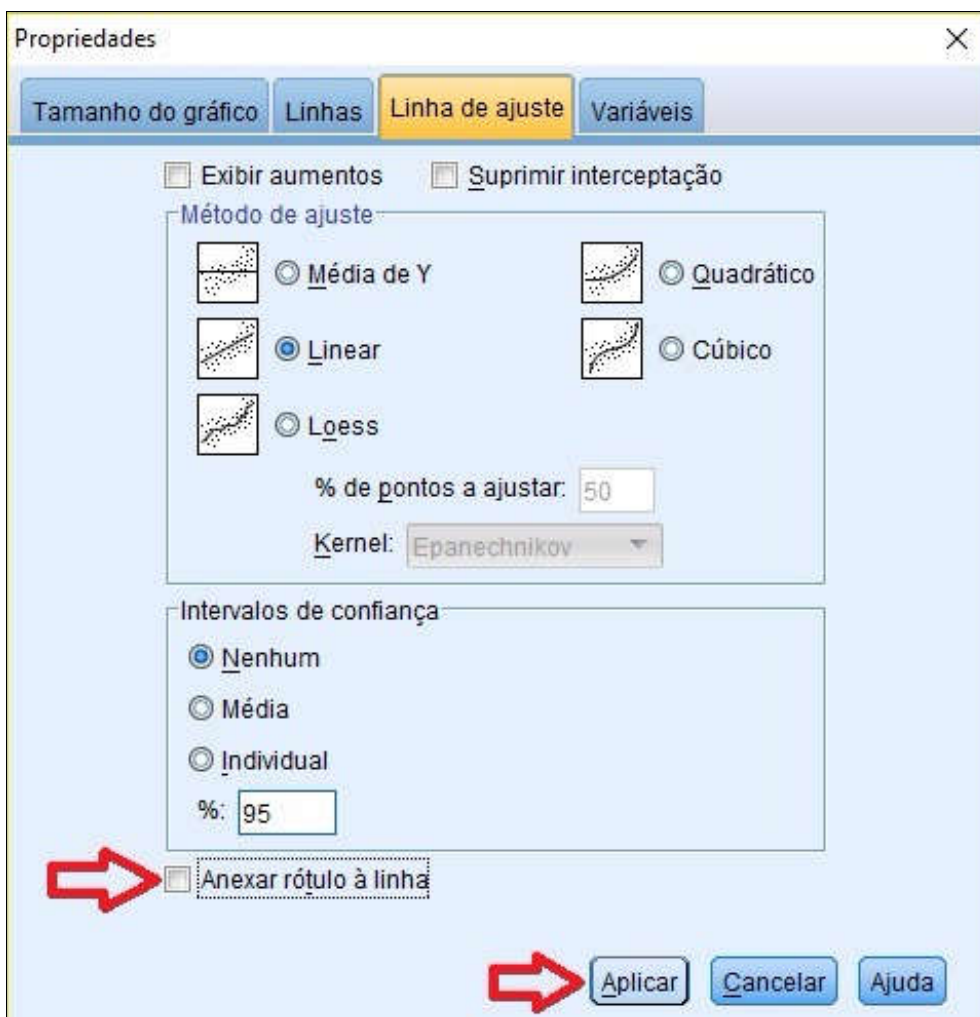


Figura 54. Linha de ajuste

Para gerar o gráfico com a variável duração, é necessário criar um novo gráfico a partir dos passos citados, porém deve-se arrastar com o mouse a variável *duração* para o eixo X (Figura 55).

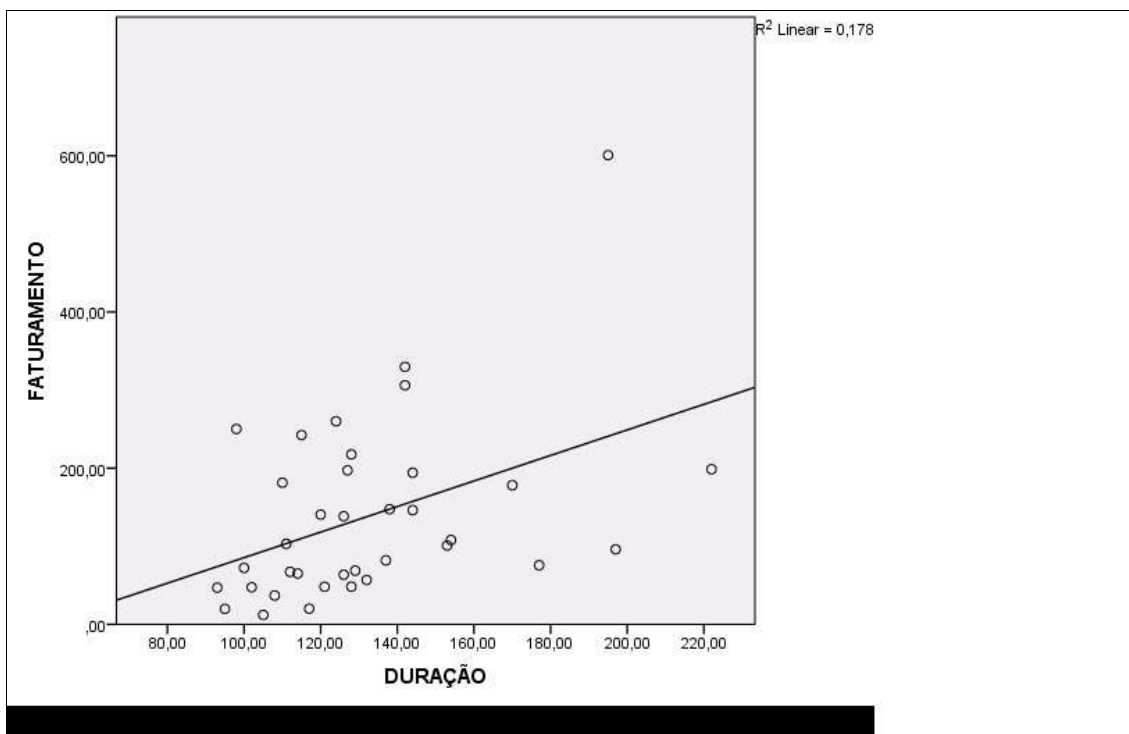


Figura 55. Gráfico de Dispersão Faturamento X Duração

Passo 3: Montagem da equação da RLM

A Figura 56 apresenta a montagem da equação da RLM, conforme o exemplo prático apresentado. A equação é a mesma que montamos no Excel.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$

FATURAMENTO = β_0 + β_1 GASTO + β_2 DURAÇÃO + u

VARIÁVEL DEPENDENTE CONSTANTE VARIÁVEIS INDEPENDENTES ERRO

COEFICIENTES DAS VARIÁVEIS INDEPENDENTES

Figura 56. Equação da RLM

Passo 4: Rodar a RLM

Visto o gráfico e a equação, podemos iniciar a Regressão Múltipla no SPSS. Na tela principal é necessário selecionar Analisar > Regressão > Linear, conforme Figura 57.

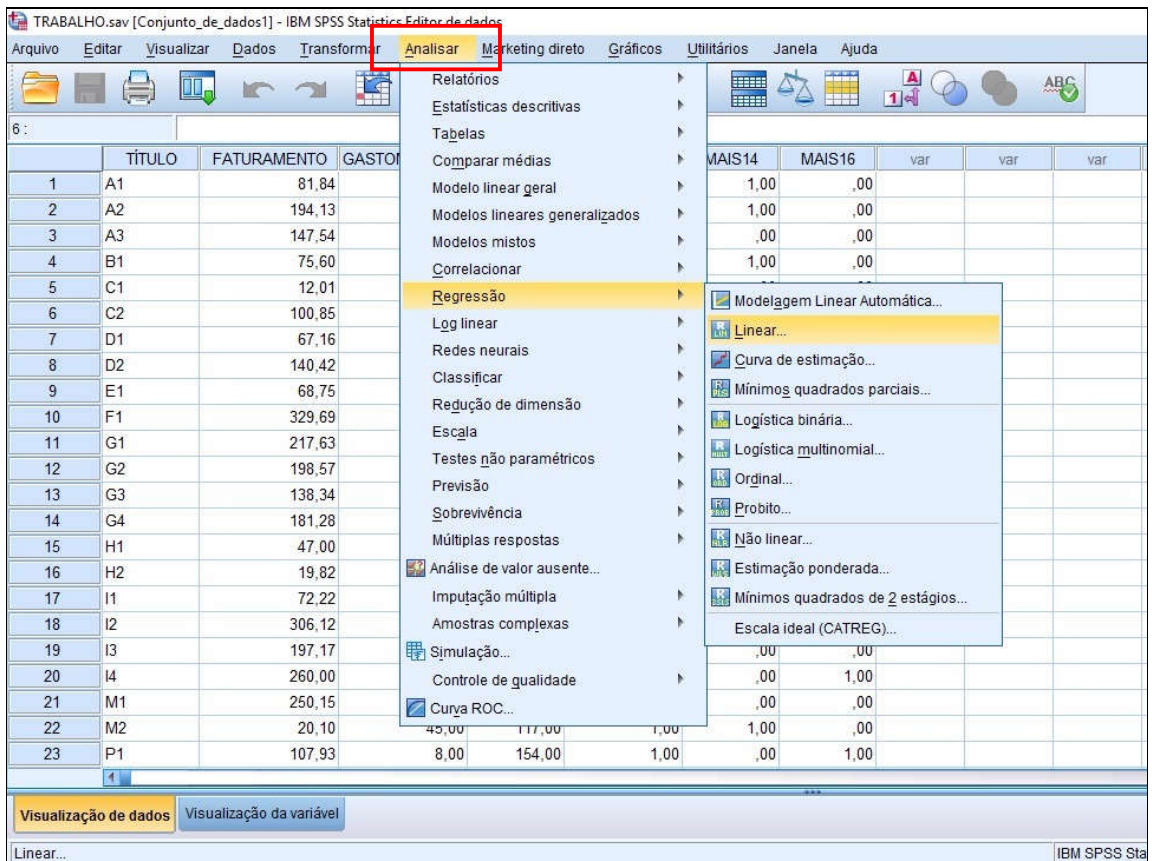


Figura 57. Regressão no SPSS

Na tela que abrirá em seguida, é necessário clicar em FATURAMENTO e na seta azul ao lado da caixa Dependente (1). Em seguida, clicar em GASTOMILHOES e na seta azul da caixa Independente (2). Após, faça o mesmo com a variável DURAÇÃO. Depois, selecione Estatísticas (3). A Figura 58 ilustra este passo a passo.

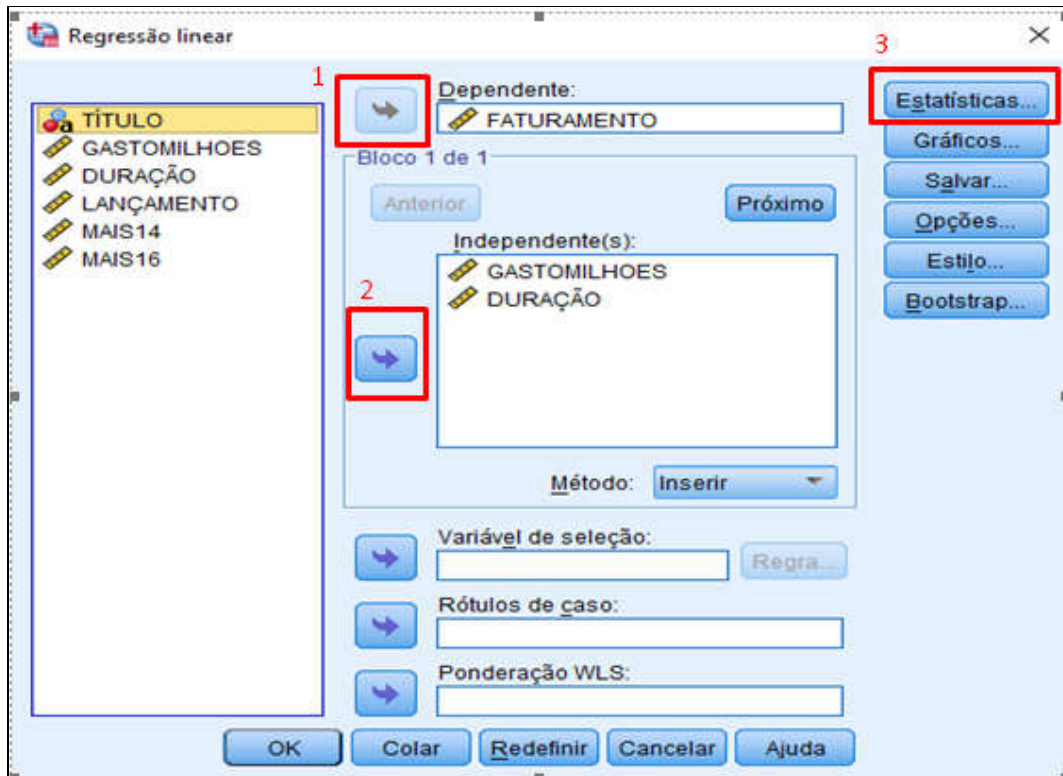


Figura 58. Entrada da variável dependente e independente

Assim, uma nova aba abrirá na qual você pode selecionar os *flags* que desejar. Marcamos aqui todos os *flags* (Figura 59). Nem todos eles são essenciais na análise de RLM e, portanto, não serão todos interpretados nos passos seguintes.

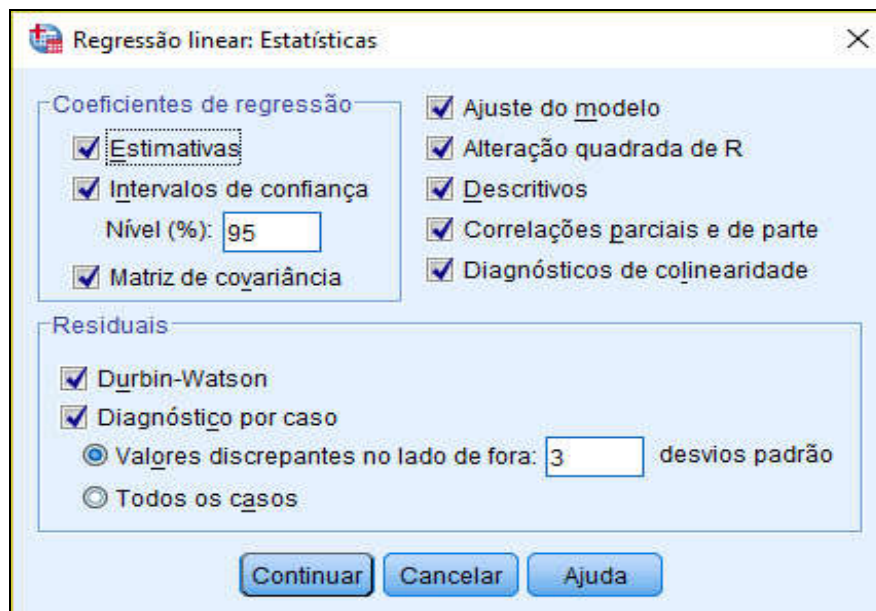


Figura 59. Opções de “Estatísticas”

Os resultados da RLM são obtidos na caixa de saída dos resultados, como nas figuras abaixo. A Figura 61 de estatísticas descritivas nos mostra a média e desvio padrão de cada variável do conjunto de dados, bem como o número de casos que foram avaliados (36 filmes). Por exemplo, sabemos que a média de faturamento destes filmes foi de 137,9716 milhões. Essa tabela não é necessária para interpretar o modelo de regressão, mas é útil como resumo dos dados.

Estatísticas descritivas			
	Média	Desvio Padrão	N
FATURAMENTO	137,9716	115,67118	36
GASTOMILHOES	53,6500	44,70890	36
DURAÇÃO	132,1111	29,82882	36

Figura 60. Estatísticas descritivas

Na matriz de correlações, pode-se observar a correlação de *Pearson* entre cada par de variáveis. O item 1 da Figura 62 representa as correlações entre o faturamento e as demais variáveis. Nota-se que sempre a correlação entre uma variável e ela mesma será 1,000. O item 2 diz respeito à significância das correlações entre as variáveis, ou seja, a correlação entre faturamento e gasto em milhões foi de 0,008, representando uma correlação significativa a 95% de confiança, já que $0,008 < 0,05$.

Correlações				
		FATURAMENTO	GASTOMILHOES	DURAÇÃO
Correlação de Pearson	FATURAMENTO	1,000	,399	,422
	GASTOMILHOES	1	1,000	,208
	DURAÇÃO	,422	,208	1,000
Sig. (1 extremidade)	FATURAMENTO	.	,008	,005
	GASTOMILHOES	2	,008	,111
	DURAÇÃO	,005	,111	.
N	FATURAMENTO	36	36	36
	GASTOMILHOES	36	36	36
	DURAÇÃO	36	36	36

Figura 61. Correlações

A matriz de correlações é extremamente útil para fornecer uma ideia aproximada do relacionamento entre as variáveis independentes e a variável dependente, bem como para o

primeiro exame da multicolinearidade. Se não existir multicolinearidade nos dados, não deve existir valores de correlação substanciais ($R > 0,90$) entre os previsores (FIELD, 2009). Com base na matriz de correlação, o SPSS indica quais variáveis devem ser inseridas no modelo (Figura 63). Neste caso, as duas variáveis independentes em teste possuem relação com a variável dependente.

Variáveis Inseridas/Removidas^a			
Modelo	Variáveis inseridas	Variáveis removidas	Método
1	DURAÇÃO, GASTOMILHO ES ^b	.	Inserir

a. Variável Dependente: FATURAMENTO
b. Todas as variáveis solicitadas inseridas.

Figura 62. Variáveis Inseridas/Removidas

O resumo do modelo descreve se o modelo é eficaz. Para interpretar o resumo do modelo, observa-se os dados fornecidos no R , R^2 , R^2 ajustado, no Erro Padrão da Estimativa e no *Durbin-Watson*, conforme Figura 64. Na figura 64 podemos observar, na coluna denominada R , o valor do coeficiente de correlação múltipla entre os previsores e a saída. As próximas colunas fornecem o valor de R^2 , o R^2 ajustado e o Erro Padrão de Estimativa que são foram melhor explicados na página 25. A alteração de R^2 nos diz se essa mudança de valor é significativa. Neste exemplo, a alteração não se mostrou significativa, pois é maior do que 0,05.

A estatística de *Durbin-Watson*, é encontrada na última coluna da tabela. Essa estatística nos informa se a hipótese de independência dos erros é satisfeita. Como uma regra conservadora, o autor sugere que valores menores do que 1 ou maiores do que 3 devem, definitivamente, ser motivos de preocupação (FIELD, 2009). Quanto mais próximo de 2 o valor estiver, melhor.

Resumo do modelo ^b										
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F	
1	,528 ^a	,279	,236	101,13329	,279	6,393	2	33	,005	1,661

a. Preditores: (Constante), DURAÇÃO, GASTOMILHOES
b. Variável Dependente: FATURAMENTO

Figura 63. Resumo do modelo

Na Figura 65 podemos verificar os valores das análises de variância (ANOVA). Ressalta-se que os significados de cada coluna podem ser revistos nas páginas 25 e 26 desta apostila.

ANOVA ^a						
Modelo		Soma dos Quadrados	df	Quadrado Médio	Z	Sig.
1	Regressão	130771,687	2	65385,843	6,393	,005 ^b
	Resíduo	337522,066	33	10227,941		
	Total	468293,753	35			

a. Variável Dependente: FATURAMENTO
b. Preditores: (Constante), DURAÇÃO, GASTOMILHOES

Figura 64. Análise de variância

A Figura 66 nos apresenta o resultado final da RLM. A constante (b₀), como explicado anteriormente, representa o valor da variável dependente caso todas as variáveis independentes fossem 0. Neste caso, B₀ é igual a -88,607. A significância de T (valor-p) indica se as variáveis são significativas.

Coeficientes ^a													
Modelo		Coeficientes não padronizados		Coeficientes padronizados		95,0% Intervalo de Confiança para B		Correlações			Estatísticas de colinearidade		
		B	Erro Padrão	Beta	t	Sig.	Limite inferior	Limite superior	Ordem zero	Parcial	Parte	Tolerância	VIF
1	(Constante)	-88,607	77,717		-1,140	,262	-246,723	69,510					
	GASTOMILHOES	,842	,391	,325	2,153	,039	,046	1,637	,399	,351	,318	,957	1,045
	DURAÇÃO	1,373	,586	,354	2,344	,025	,181	2,565	,422	,378	,346	,957	1,045

Figura 65. Coeficientes

Passo 5: Substituir os dados na equação da RLM

Conforme já apresentado no Excel, este passo consiste em substituir os valores encontrados por meio da RLM na equação original, conforme Figura 64 e Figura 66. Abaixo apresenta-se a substituição de valores na equação da RLM.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$

$$FAT = - 88,6 + 0,84 \cdot GAS + 1,37 \cdot DUR + 101,13$$

Figura 66: Substituição de valores na equação da RLM

É importante também analisar a significância de cada variável independente em relação à variável dependente por meio do “Sig.”, conforme exemplo abaixo, exposto na Figura 68. Este dado busca clarear a significância de cada uma das variáveis.

$$Y = - 88,6 + 0,84 \cdot GASTO + 1,37 \cdot DURAÇÃO + 101,13$$

0,26 0,03 0,02



SIGNIFICÂNCIA

Figura 67. Significância das variáveis

Passo 6: Interpretação dos resultados

Conforme já apresentado na página 29 desta apostila, a interpretação dos resultados é mesma fornecida no modelo de Excel, visto que os mesmos dados foram utilizados, apenas com ferramentas diferentes.

****Variáveis Dummy***

Conforme citado anteriormente, observa-se que outras variáveis podem não estar sendo consideradas neste modelo, visto que o valor do erro padrão da estimativa é alto (Figura 64).

Assim, acrescentaremos a variável LANÇAMENTO, conforme foi realizado no modelo do Excel (Figura 69).

*TRABALHO.sav [Conjunto_de_dados1] - IBM SPSS Statistics Editor de dados

Arquivo Editar Visualizar Dados Transformar Analisar Marketing direto Gráficos

11:

	TÍTULO	FATURAMENTO	GASTOMILHOES	DURAÇÃO	LANÇAMENTO
1	A1	81,84	18,50	137,00	,00
2	A2	194,13	140,00	144,00	1,00
3	A3	147,54	50,00	138,00	1,00
4	B1	75,60	72,00	177,00	1,00
5	C1	12,01	,30	105,00	1,00
6	C2	100,85	90,00	153,00	1,00
7	D1	67,16	104,00	112,00	1,00
8	D2	140,42	75,00	120,00	1,00
9	E1	68,75	55,00	129,00	1,00
10	F1	329,69	55,00	142,00	1,00
11	G1	217,63	22,00	128,00	1,00
12	G2	198,57	3,90	222,00	,00
13	G3	138,34	10,00	126,00	1,00
14	G4	181,28	6,00	110,00	,00
15	H1	47,00	,30	93,00	,00
16	H2	19,82	70,00	95,00	1,00
17	I1	72,22	17,00	100,00	1,00
18	I2	306,12	75,00	142,00	1,00
19	I3	197,17	39,00	127,00	,00
20	I4	260,00	12,00	124,00	,00
21	M1	250,15	90,00	98,00	1,00
22	M2	20,10	45,00	117,00	1,00
23	P1	107,93	8,00	154,00	1,00

Visualização de dados Visualização da variável

Figura 68. Inclusão da variável lançamento

Antes de realizar novamente a RLM, é necessário transformar as variáveis qualitativas em variáveis *dummy*, atribuindo a elas valores numéricos, conforme apresentado anteriormente. Após, é necessário gerar a **equação básica** de RLM.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + u$$

O próximo passo consiste em realizar a **RLM** no SPSS. A variável dependente permanece a mesma, porém é necessário selecionar todas as variáveis independentes, inclusive a *dummy* (lançamento). A Figura 70 ilustra este passo a passo.

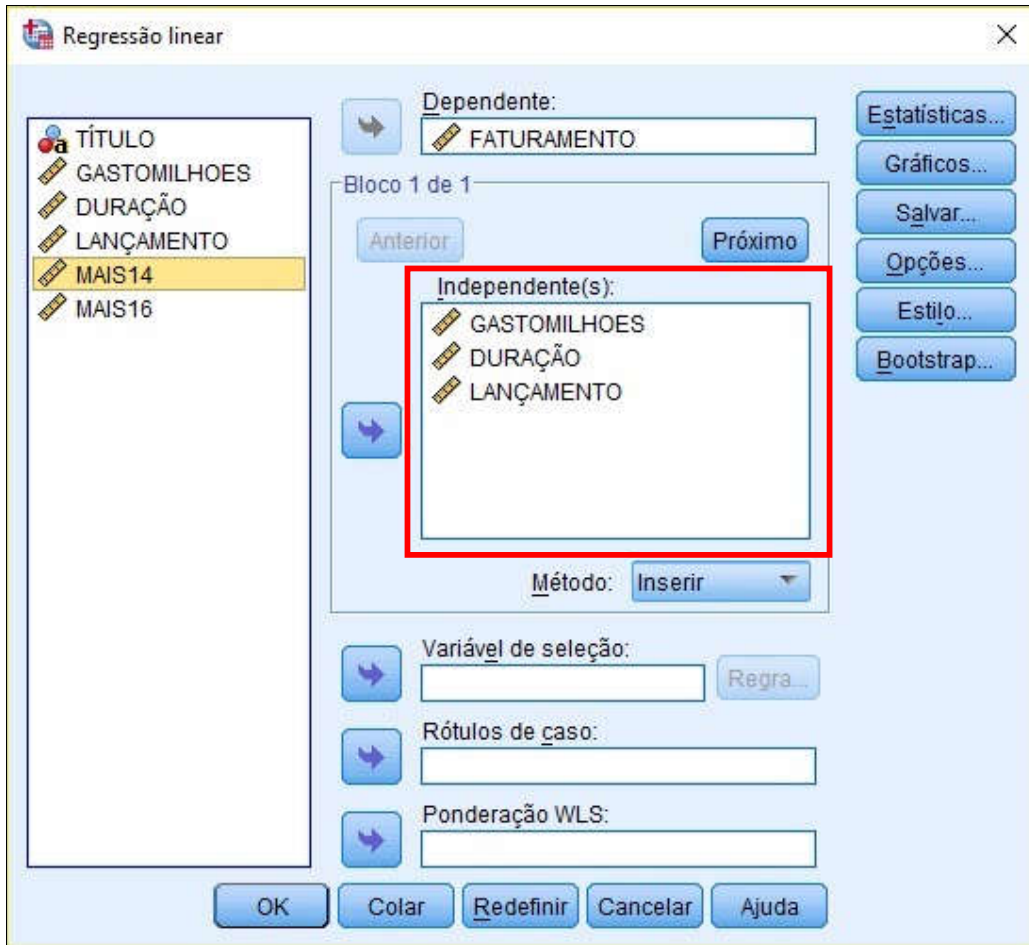


Figura 69. Inclusão da variável lançamento

O resumo do modelo, apresentado na Figura 71, indica que o R^2 ajustado aumentou com relação ao modelo anterior, sem a variável *dummy lançamento*. A Figura 72 mostra os resultados que serão substituídos na equação.

Resumo do modelo ^b											
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson	
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F		
1	,606 ^a	,368	,309	96,18400	,368	6,206	3	32	,002	1,761	

a. Preditores: (Constante), DURAÇÃO, LANÇAMENTO, GASTOMILHOES
b. Variável Dependente: FATURAMENTO

Figura 70. Resumo do modelo

Coeficientes ^a													
Modelo	Coeficientes não padronizados		Coeficientes padronizados	t	Sig.	95,0% Intervalo de Confiança para B		Correlações			Estatísticas de colinearidade		
	B	Erro Padrão	Beta			Limite inferior	Limite superior	Ordem zero	Parcial	Parte	Tolerância	VIF	
1	(Constante)	-29,448	79,018										
	LANÇAMENTO	-93,836	44,316	-.342	-.2117	,042	-184,104	-3,567	-.083	-.351	-.298	,757	1,321
	GASTOMILHOES	1,286	,427	,497	3,012	,005	,416	2,155	,399	,470	,423	,726	1,378
	DURAÇÃO	1,298	,558	,335	2,324	,027	,160	2,435	,422	,380	,327	,953	1,050

a. Variável Dependente: FATURAMENTO

Figura 71. Coeficientes

Com base nos resultados apontados, é necessário novamente substituir a equação de RLM (Figura 73):

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + u$$

$$\text{FAT} = -29,44 + 1,28 \cdot \text{GAS} + 1,29 \cdot \text{DUR} - 93,83 \cdot \text{LANC} + 96,18$$

0,71

0,005

0,02

0,04

SIGNIFICÂNCIA

Figura 72. Equação do modelo 1

A interpretação dos resultados é idêntica a do Excel. Conforme citado anteriormente, **observa-se ainda que outras variáveis podem não estar sendo consideradas neste modelo**. Assim, será acrescentada a variável independente “**faixa etária**”, a qual pode ser: livre, maior que 14 anos e maior que 16 anos. Observa-se que esta nova variável, também qualitativa, apresenta três categorias. Portanto, é necessário transformá-la em valores numéricos, conforme apresentado anteriormente. Assim, seguindo o mesmo procedimento sugerido no Excel, gera-se um novo modelo de regressão. O passo a passo para realização desta RLM é apresentado na Figuras 74.

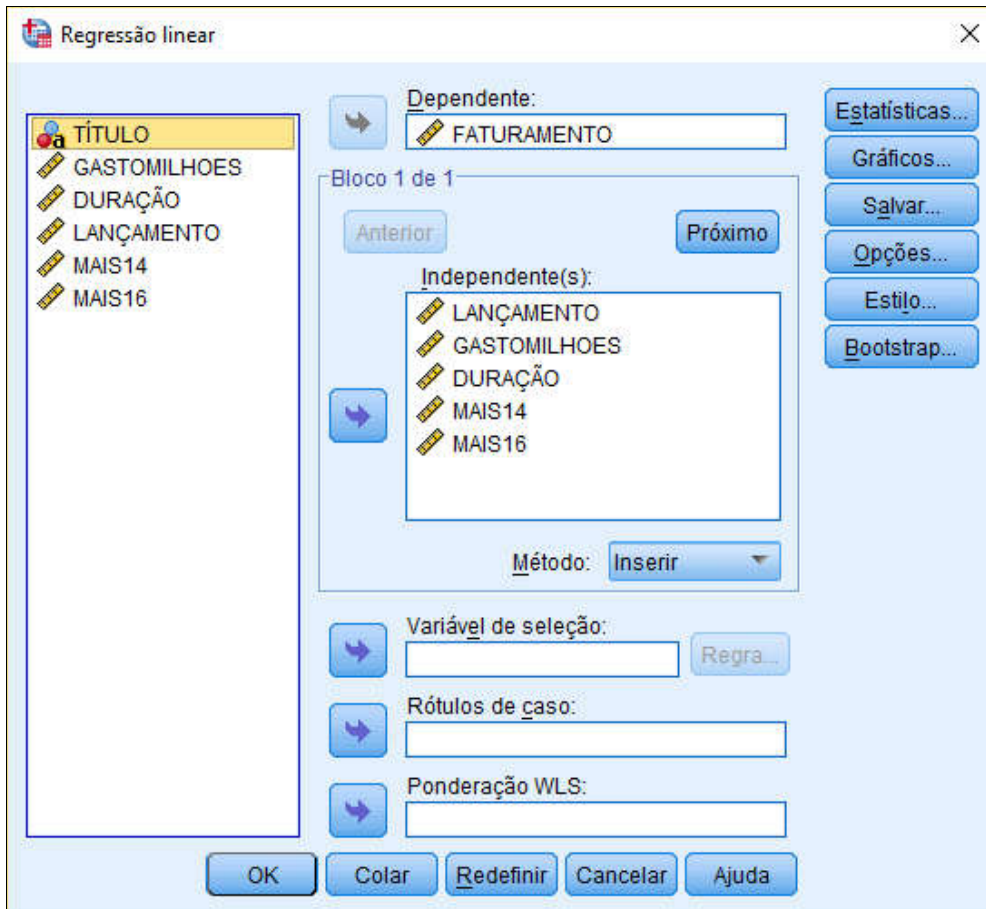


Figura 73. Inserção das variáveis independentes dummy

Observa-se que a variável dependente permanece a mesma, porém na variável independente é necessário selecionar todas as variáveis independentes, inclusive as *dummy* (lançamento e faixa etária). As saídas geradas podem ser vistas nas Figuras 75 e 76.

Resumo do modelo ^b										
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F	
1	,647 ^a	,419	,322	95,21193	,419	4,332	5	30	,004	1,677

a. Preditores: (Constante), MAIS16, LANÇAMENTO, DURAÇÃO, MAIS14, GASTOMILHOES
b. Variável Dependente: FATURAMENTO

Figura 74. Resumo do modelo

Coeficientes ^a															
Modelo		Coeficientes não padronizados		Coeficientes padronizados		t	Sig.	95,0% Intervalo de Confiança para B		Correlações			Estatísticas de colinearidade		
		B	Erro Padrão	Beta				Limite inferior	Limite superior	Ordem zero	Parcial	Parte	Tolerância	VIF	
1	(Constante)	-30,605	78,336			-,391	,699	-190,589	129,378						
	LANÇAMENTO	-93,607	44,056	-,341		-2,125	,042	-183,582	-3,633	-,083	-,362	-,296	,751	1,332	
	GASTOMILHOES	1,395	,428	,539		3,260	,003	,521	2,269	,399	,511	,454	,707	1,414	
	DURAÇÃO	1,249	,561	,322		2,224	,034	,102	2,395	,422	,376	,309	,924	1,083	
	MAIS14	-22,816	36,989	-,098		-,617	,542	-98,358	52,726	-,076	-,112	-,086	,774	1,291	
	MAIS16	47,055	42,828	,172		1,099	,281	-40,411	134,521	,190	,197	,153	,794	1,259	

a. Variável Dependente: FATURAMENTO

Figura 75. Coeficientes

Substituindo os valores encontrados, teremos a equação apresentada na Figura 77.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + u$$

$$Y = -30,60 + 1,39 \cdot \text{GASTO} + 1,24 \cdot \text{DURAÇÃO} - 93,60 \cdot \text{LANÇAMENTO} - 22,8 \cdot \text{MAIOR DO QUE 14} + 47,05 \cdot \text{MAIOR DO QUE 16} + 95,2$$

Figura 76. Equação

A interpretação dos resultados segue o mesmo padrão do Excel. Assim, com base no modelo proposto, ainda há um **próximo passo** a ser seguido, o qual consiste em verificar a existência ou não de **multicolinearidade** entre as variáveis independentes. Para verificar a multicolinearidade do modelo proposto é necessário realizar a **Matriz de Correlação**, na saída gerada pelo SPSS é a tabela com o nome correlações, a qual será apresentada na Figura 78.

		Correlações					
		FATURAMENTO	LANÇAMENTO	GASTOMILHOES	DURAÇÃO	MAIS14	MAIS16
Correlação de Pearson	FATURAMENTO	1,000	-,083	,399	,422	-,076	,190
	LANÇAMENTO	-,083	1,000	,490	,047	,152	-,036
	GASTOMILHOES	,399	,490	1,000	,208	,207	-,117
	DURAÇÃO	,422	,047	,208	1,000	,109	,084
	MAIS14	-,076	,152	,207	,109	1,000	-,426
	MAIS16	,190	-,036	-,117	,084	-,426	1,000
Sig. (1 extremidade)	FATURAMENTO	.	,316	,008	,005	,330	,134
	LANÇAMENTO	,316	.	,001	,392	,188	,418
	GASTOMILHOES	,008	,001	.	,111	,113	,249
	DURAÇÃO	,005	,392	,111	.	,263	,312
	MAIS14	,330	,188	,113	,263	.	,005
	MAIS16	,134	,418	,249	,312	,005	.
N	FATURAMENTO	36	36	36	36	36	36
	LANÇAMENTO	36	36	36	36	36	36
	GASTOMILHOES	36	36	36	36	36	36
	DURAÇÃO	36	36	36	36	36	36
	MAIS14	36	36	36	36	36	36
	MAIS16	36	36	36	36	36	36

Figura 77. Correlações

Escolhendo o melhor modelo de regressão...

Modelo 1 – Considerando todas as variáveis

O **Modelo 1** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas. Observa-se que este modelo já foi realizado anteriormente e apresentou **R² ajustado de 0,32**. Ou seja, as cinco variáveis independentes **explicam juntas 32%** da variável dependente. A Figura 78 apresenta os resultados desta RLM, a qual considera todas as variáveis.

Resumo do modelo ^b										
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F	
1	,647 ^a	,419	,322	95,21193	,419	4,332	5	30	,004	1,677

a. Preditores: (Constante), MAIS16, LANÇAMENTO, DURAÇÃO, MAIS14, GASTOMILHOES
b. Variável Dependente: FATURAMENTO

Figura 78. Resumo do Modelo 1

Modelo 2 – Considerando todas as variáveis exceto faixa etária

O **Modelo 2** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, exceto a faixa etária, visto que esta não apresentou alta significância anteriormente (valor-P). Observa-se que este modelo já foi realizado e apresentou **R² ajustado de 0,3**. Ou seja, as variáveis independentes **explicam juntas 30%** da variável dependente. A Figura 79 apresenta os resultados desta RLM, a qual considera todas as variáveis, exceto a faixa etária.

Resumo do modelo ^b										
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F	
1	,606 ^a	,368	,309	96,18400	,368	6,206	3	32	,002	1,761

a. Preditores: (Constante), DURAÇÃO, LANÇAMENTO, GASTOMILHOES
b. Variável Dependente: FATURAMENTO

Figura 79. Resumo do Modelo 2

Modelo 3 – Considerando todas as variáveis, exceto gasto

O **Modelo 3** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, **exceto** gasto. Este modelo será aplicado, uma vez que se observou certa **multicolinearidade** entre as variáveis **gasto e lançamento**. A Figura 80 apresenta a RLM sendo gerada e a Figura 81 apresenta os resultados desta RLM, a qual considera todas as variáveis, exceto gasto.

Observa-se que neste modelo o **R² ajustado foi de 0,11**. Ou seja, as variáveis independentes **explicam juntas 11%** da variável dependente.

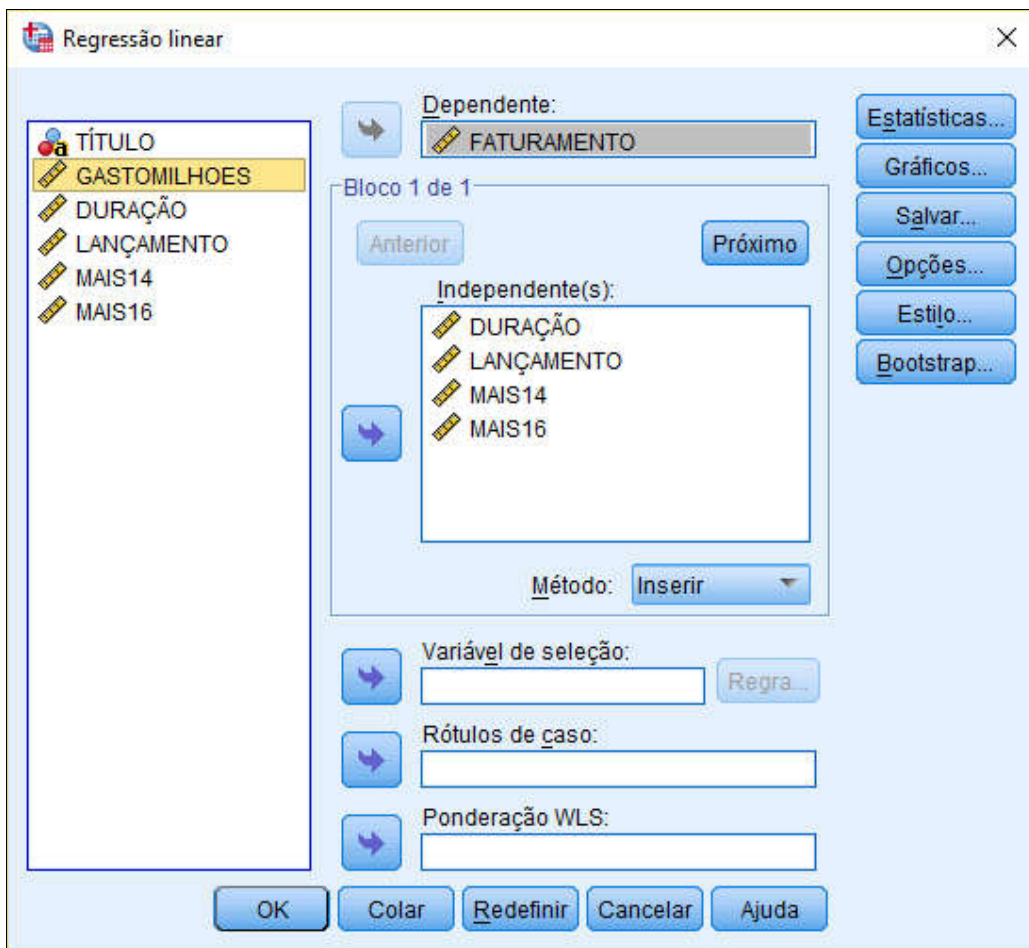


Figura 80. Modelo 3

Resumo do modelo ^b											
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson	
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F		
1	,462 ^a	,214	,112	108,99941	,214	2,104	4	31	,104	1,892	

a. Preditores: (Constante), MAIS16, LANÇAMENTO, DURAÇÃO, MAIS14
b. Variável Dependente: FATURAMENTO

Figura 81. Resumo do Modelo 3

Modelo 4 – Considerando todas as variáveis, exceto lançamento

O **Modelo 4** considera a relação entre a variável dependente e todas as variáveis independentes apresentadas, **exceto** lançamento. Este modelo será aplicado, uma vez que se observou certa **multicolinearidade** entre as variáveis **gasto e lançamento**. A Figura 82 demonstra a RLM do Modelo 4 sendo gerada e a Figura 83 apresenta os resultados desta RLM, considerando todas as variáveis, exceto lançamento. Observa-se que neste modelo o **R²**

ajustado foi de 0,24. Ou seja, as variáveis independentes **explicam juntas 24%** da variável dependente.

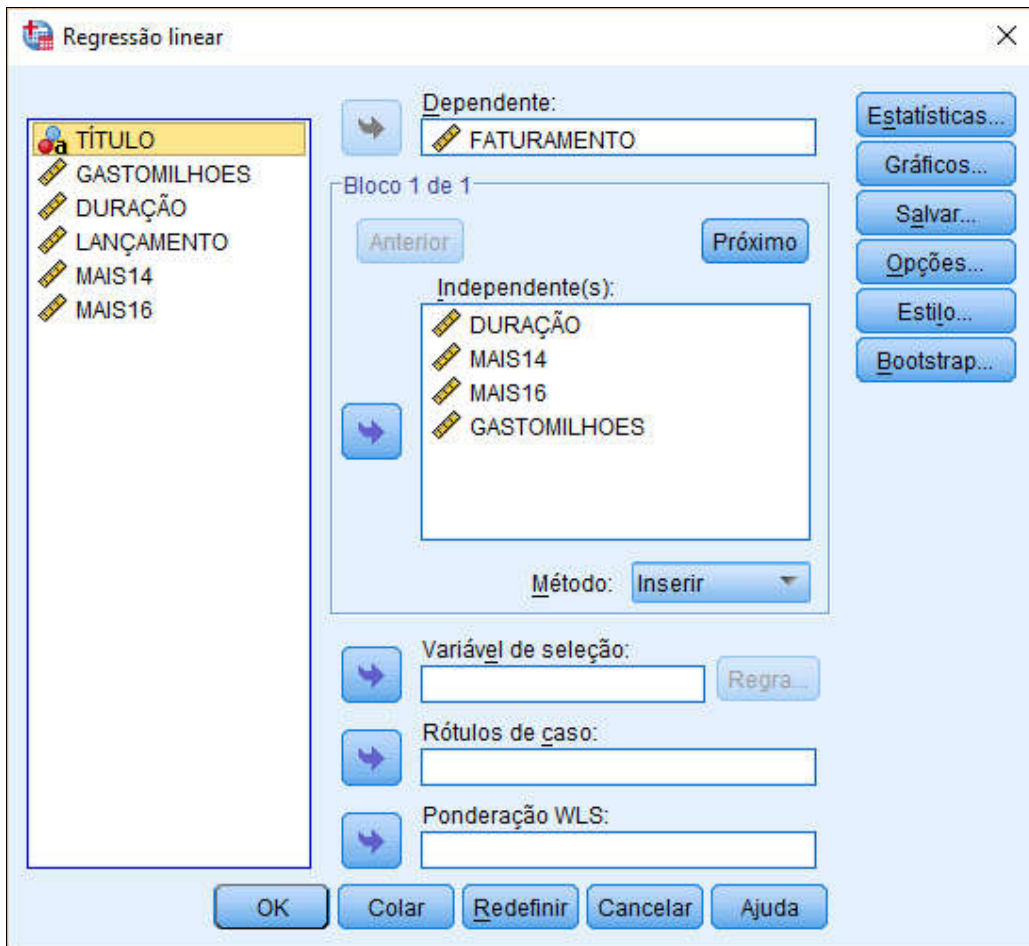


Figura 82. Modelo 4

Resumo do modelo ^b											
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson	
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F		
1	,576 ^a	,332	,246	100,46419	,332	3,849	4	31	,012	1,571	

a. Preditores: (Constante), GASTOMILHOES, MAIS16, DURAÇÃO, MAIS14
b. Variável Dependente: FATURAMENTO

Figura 83. Resumo do Modelo 4

Modelo 5 – Considerando apenas a variável duração

O **Modelo 5** considera a relação entre a variável dependente e a variável independente de duração, somente. A Figura 84 demonstra a RLM do Modelo 5 sendo gerada e a Figura 85 apresenta os resultados desta RLM, considerando apenas a variável duração. Observa-se que neste modelo o **R² ajustado foi de 0,15**. Ou seja, a variável de duração **explica sozinha 15%** da variável dependente.

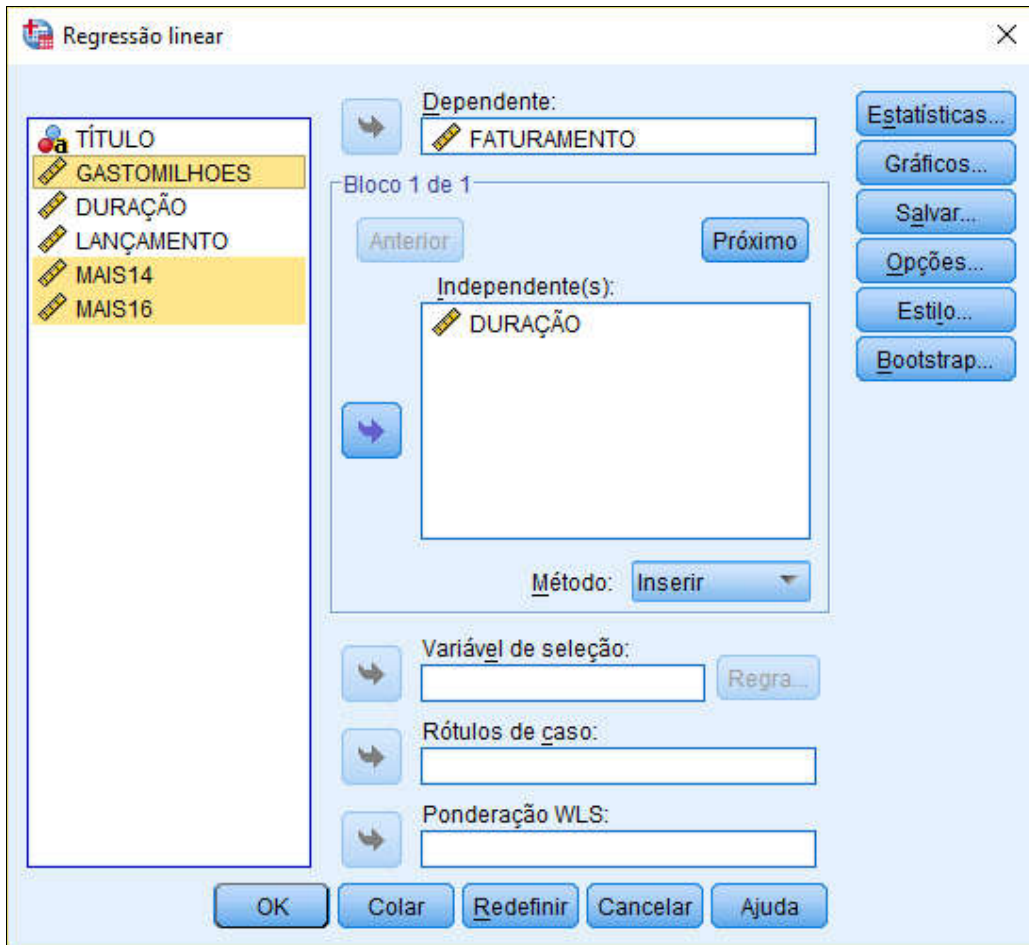


Figura 84. Modelo 5

Resumo do modelo ^b											
Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson	
					Alteração de R quadrado	Alteração F	df1	df2	Sig. Alteração F		
1	,422 ^a	,178	,154	106,40475	,178	7,362	1	34	,010	1,865	

a. Preditores: (Constante), DURAÇÃO
b. Variável Dependente: FATURAMENTO

Figura 85. Resumo do modelo 5

REFERÊNCIAS

ALMEIDA, L. S.; FREIRE, T. **Metodologia da investigação em psicologia e educação**. 2000.

BELFIORE, P. **Estatística aplicada a administração, contabilidade e economia com Excel e SPSS**. Rio de Janeiro: Elsevier, 2015.

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

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

FONSECA, J. J. S. **Metodologia da Pesquisa Científica**. 2002.

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