UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
CURSO DE ESPECIALIZAÇÃO EM MÉTODOS ESTATÍSTICOS COMPUTACIONAIS

INDICADORES DE TEMPO DE PASSAGEM ENTRE APRESENTAÇÃO NA SEDE E INÍCIO DA JORNADA EM TREM DA TRIPULAÇÃO DOS TRENS DA MALHA SUDESTE NA REGIÃO DO RIO DE JANEIRO

ADRIANO CANDIDO DA SILVA
RICARDO SAAR RODRIGUES

JUIZ DE FORA
2013

ADRIANO CANDIDO DA SILVA E RICARDO SAAR RODRIGUES

INDICADORES DE TEMPO DE PASSAGEM ENTRE APRESENTAÇÃO NA SEDE E INÍCIO DA JORNADA EM TREM DA TRIPULAÇÃO DOS TRENS DA MALHA SUDESTE NA REGIÃO DO RIO DE JANEIRO

Monografia apresentada como requisito parcial para conclusão do curso de Especialização em Métodos Estatísticos Computacionais do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora, sob orientação do professor Dr. Luiz Cláudio Ribeiro.

JUIZ DE FORA
2013

TERMO DE APROVAÇÃO

ADRIANO CANDIDO DA SILVA E RICARDO SAAR RODRIGUES

INDICADORES DE TEMPO DE PASSAGEM ENTRE APRESENTAÇÃO NA SEDE E INÍCIO DA JORNADA EM TREM DA TRIPULAÇÃO DOS TRENS DA MALHA SUDESTE NA REGIÃO DO RIO DE JANEIRO

Trabalho de Monografia apresentado como parte das exigências de conclusão do Curso de Especialização em Métodos Estatísticos Computacionais, Instituto de Ciências Exatas, Universidade Federal de Juiz de Fora, pela seguinte banca examinadora:

Orientador:

_____

Profº. Dr. Luiz Claudio Ribeiro

Juiz de Fora, 16 de Abril de 2013

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O transporte ferroviário é peça fundamental para a logística de qualquer região produtora de mercadorias de grande volume que necessitam ser transportados por grandes distâncias. Podem-se citar algumas mercadorias como: grãos, *containers*, produtos siderúrgicos acabados (placas, tarugos, lingotes, bobinas, vergalhões, entre outros) e minério de ferro.

Como característica marcante do transporte ferroviário, tem-se, portanto, a movimentação de grandes volumes de mercadorias por grandes distâncias geográficas. A existência de grandes volumes não gera qualquer tipo de empecilho para o transporte, visto que todos os recursos envolvidos (locomotivas, vagões, obras de arte, entre outros) são criados especificamente para esta característica. Já a grande distância geográfica gera problemas para a logística de troca de tripulação, visto que quanto maior essa distância, mais tempo pode-se gastar para essa troca.

A tripulação (também chamada de equipagem na ferrovia) apresenta uma participação significante nos custos variáveis das companhias, o que exige que ela seja utilizada com a maior produtividade possível. Uma etapa do ciclo de trabalho da equipagem muito importante para ser medida e controlada é o tempo gasto pelo maquinista do momento em que é designado a um trem até o momento em que assume de fato a condução dessa composição. Esse tempo é chamado de passagem e pode variar muito em função do ponto onde a equipagem se apresenta e o ponto onde o trem está aguardando essa apresentação. Há pontos determinados para a apresentação desses maquinistas, mas de acordo com o tempo de viagem realizado pelo trem, este recurso pode ser necessário em pontos diferentes da malha ferroviária. Essa necessidade torna-se clara quando analisamos um exemplo: um trem qualquer tem, do ponto A ao ponto B, a viagem planejada em 8 horas para uma velocidade média de 40 km/h (o que deixa clara a distância de 320 km entre os dois pontos). Caso esse trem realize a viagem a uma velocidade média de 30 km/h, ele não conseguirá chegar ao ponto B com a mesma equipagem (para simplificação da análise, será adotado que não é possível extrapolar 8 horas de jornada em trem por maquinista), sendo necessário trocar a tripulação no ponto B’ (situado no km 240). A nova equipagem se apresenta no ponto B (conforme o plano estipulado),

distante em 80 km do ponto onde o trem realmente precisará desse recurso. O tempo necessário para transferir o maquinista do ponto B ao B' é o tempo de passagem.

Quanto maior for esse tempo, mais improdutiva será a utilização da equipagem, o que pode gerar principalmente aumento dos custos. É importante estudar e conhecer o comportamento desse indicador porque ele irá balizar decisões importantes tanto no curto quanto no longo prazo. No curto prazo, a variação desse tempo de passagem pode indicar a necessidade de contratação de mais carros para levar os maquinistas do ponto de apresentação ao ponto de troca (onde o trem estará aguardando a nova tripulação) ou até mesmo mudar o padrão de condução dos trens para enquadrar melhor o tempo de percurso aos pontos de troca existentes. No longo prazo, decisões mais estratégicas como a contratação de mais pessoas (quanto maior a improdutividade desse recurso, mais pessoas serão necessárias para atender o mesmo nível de produção) ou a mudança de pontos de troca dependem em grande parte dos valores esperados para esse indicador.

Para atender tanto a necessidade de curto prazo quanto a de longo prazo, é necessário definir uma forma de estimar os resultados que seja simples ao ponto de não depender de complexos cenários (difíceis de serem montados no curto prazo e passíveis de grandes desvios no longo prazo) e que seja assertiva ao ponto de balizar de forma correta as tomadas de decisões. Com o objetivo de atender a esses dois pressupostos, será utilizada a regressão linear para criar uma forma de estimar esses valores.

Duas informações podem ser consideradas como o pilar de todo o dimensionamento de transporte ferroviário e, em qualquer cenário que se pretenda trabalhar, serão informações de altíssima confiabilidade, que não irão sofrer grandes distorções causadas por simplificações de cálculo. Essas duas informações são: volume diário a ser transportado e tamanho da tripulação (*headcount*). Estas serão as variáveis independentes do modelo de regressão linear, que terá como variável dependente a quantidade de horas de passagem.

2. APRESENTAÇÃO DOS DADOS

Este trabalho irá utilizar dados reais de janeiro de 2010 a agosto de 2012, referentes à utilização de equipagem na região do Rio de Janeiro (a qual é muito importante por apresentar o principal complexo portuário para escoamento da produção de minério de ferro do estado de Minas Gerias). A amostra contém 32 dados para cada variável e foi escolhida por ser toda a base existente com altíssimo nível de confiabilidade. Valores anteriores a esse período podem ter distorções na apuração que iriam enviesar o trabalho.

Serão apresentados a seguir os dados que formarão a base deste estudo, permitindo conhecer as características dessas informações.

a) Quantidade de Horas de Passagem

Esta será a variável dependente do modelo de regressão linear. O primeiro ponto a ser analisado é a forma como se dá a distribuição desses dados, logo, será analisado o de teste Kolmogorov-Smirnov (amostra maior que 30 elementos).

TABELA 1 – Teste de normalidade para distribuição dos dados (HR_PASSE)

One-Sample Kolmogorov-Smirnov Test

		HR_PASSE
N		32
Normal Parameters[a,b]	Mean	3113,19
	Std. Deviation	451,352
Most Extreme Differences	Absolute	,103
	Positive	,103
	Negative	-,062
Kolmogorov-Smirnov Z		,583
Asymp. Sig. (2-tailed)		,886

a. Test distribution is Normal.

b. Calculated from data.

FONTE: elaborado pelos próprios autores

Para este teste, tem-se a seguinte formulação das hipóteses:

- H_0: A distribuição da série testada é normal;
- H_1: A distribuição da série testada não tem comportamento normal.

Observa-se para *Sig*. o valor de 0,886, o que deixa claro não existir indícios para rejeitar H_0 para $\alpha = 5\%$ (será este o nível de significância adotado neste trabalho), logo estes dados seguem uma distribuição normal.

Após esta constatação, serão apresentados indicadores para mostrar as principais medidas estatísticas dessa amostra, que são: *amplitude, valor mínimo, valor máximo, média e desvio padrão*. Além desses dados, será apresentado um *box plot* para permitir a visualização de como esses dados estão distribuídos na amostra.

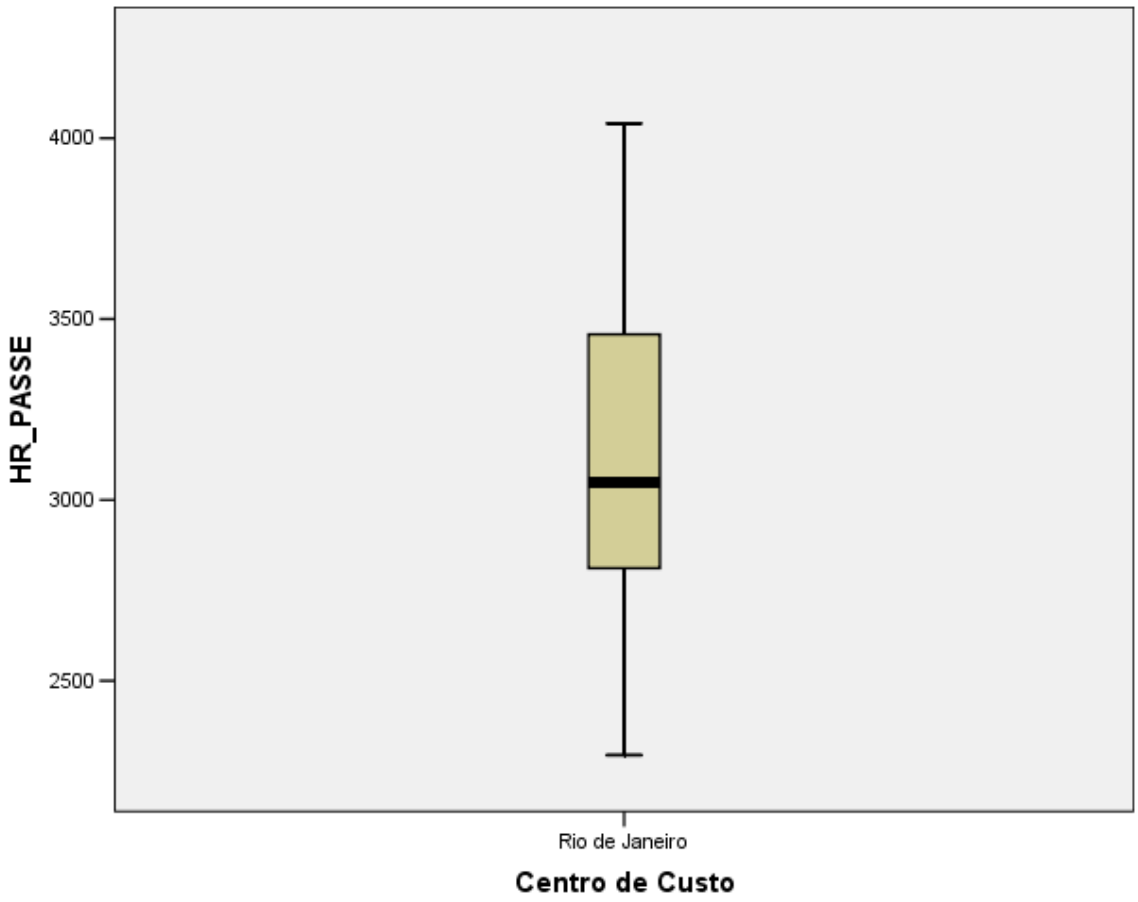
TABELA 2 – Análise descritiva (HR_PASSE)

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
HR_PASSE	32	1747	2294	4041	3113,19	451,352
Valid N (listwise)	32					

FONTE: elaborado pelos próprios autores

GRÁFICO 1 – *Box plot* (HR_PASSE)



FONTE: elaborado pelos próprios autores

Pelo *box plot* acima, não é possível identificar a presença de *outliers*, o que poderia prejudicar o modelo de regressão (GUJARATI, 2006).

b) Volume Diário a ser Transportado

Esta é uma das variáveis elegíveis para compor o modelo de regressão linear como variável independente, visto que é passível de controle para qualquer horizonte de tempo em que se queira empregar o modelo proposto. Analisando a forma como se dá a distribuição desses dados, tem-se novamente uma distribuição normal, visto que *Sig*. é novamente maior que 0,05, conforme apresentado na tabela a seguir:

TABELA 3 – Teste de normalidade para distribuição dos dados (VOL)

One-Sample Kolmogorov-Smirnov Test

		VOL
N		32
Normal Parameters[a,b]	Mean	301884,56
	Std. Deviation	25770,782
Most Extreme Differences	Absolute	,123
	Positive	,076
	Negative	-,123
Kolmogorov-Smirnov Z		,696
Asymp. Sig. (2-tailed)		,718

a. Test distribution is Normal.

b. Calculated from data.

FONTE: elaborado pelos próprios autores

As principais medidas estatísticas e o *box plot* estão apresentados abaixo.

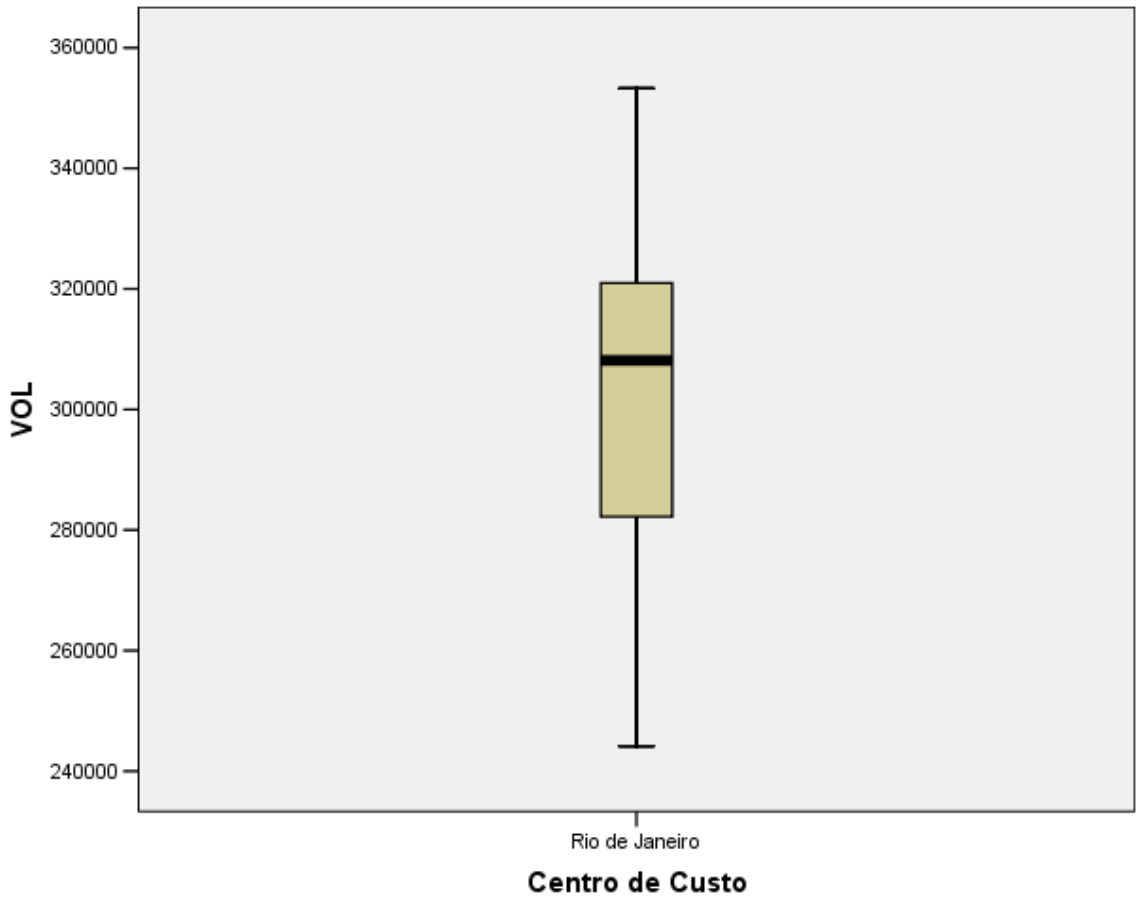
TABELA 4 – Análise Descritiva (VOL)

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
VOL	32	109097	244158	353255	301884,56	25770,782
Valid N (listwise)	32					

FONTE: elaborado pelos próprios autores

GRÁFICO 2 – *Box plot* (VOL)



FONTE: elaborado pelos próprios autores

c) Headcount de Maquinistas

Esta á a última variável a ser analisada e também representa uma variável elegível para compor o modelo de regressão linear como variável independente. Repetindo o teste de Kolmogorov-Smirnov, obtém-se novamente um comportamento normal para a amostra selecionada, como pode-se observar a seguir (*Sig*. maior que 0,05).

TABELA 5 – Teste de normalidade para distribuição dos dados (HCOUNT)

One-Sample Kolmogorov-Smirnov Test

		HCOUNT
N		32
Normal Parameters[a,b]	Mean	265,38
	Std. Deviation	20,144
Most Extreme Differences	Absolute	,144
	Positive	,144
	Negative	-,098
Kolmogorov-Smirnov Z		,817
Asymp. Sig. (2-tailed)		,517

a. Test distribution is Normal.

b. Calculated from data.

FONTE: elaborado pelos próprios autores

As principais medidas estatísticas e o box plot estão apresentados abaixo. Analisando esses últimos elementos, é possível perceber que em nenhuma amostra selecionada há a presença de *outliers*, o que gera uma maior segurança para a construção dos modelos de regressão. Segundo GUJARATI (2006), dependendo da amostra utilizada, é necessário tratar os *outliers* para que não distorçam os resultados encontrados pela regressão. O método mais comum para esse tratamento é a eliminação desses elementos da amostra, porém, como existem apenas 32 valores para cada variável apresentada, essa eliminação poderia causar grande perda para a massa de dados a ser utilizada, pois a exclusão de um único dado representa a redução de aproximadamente 3% no tamanho da amostra.

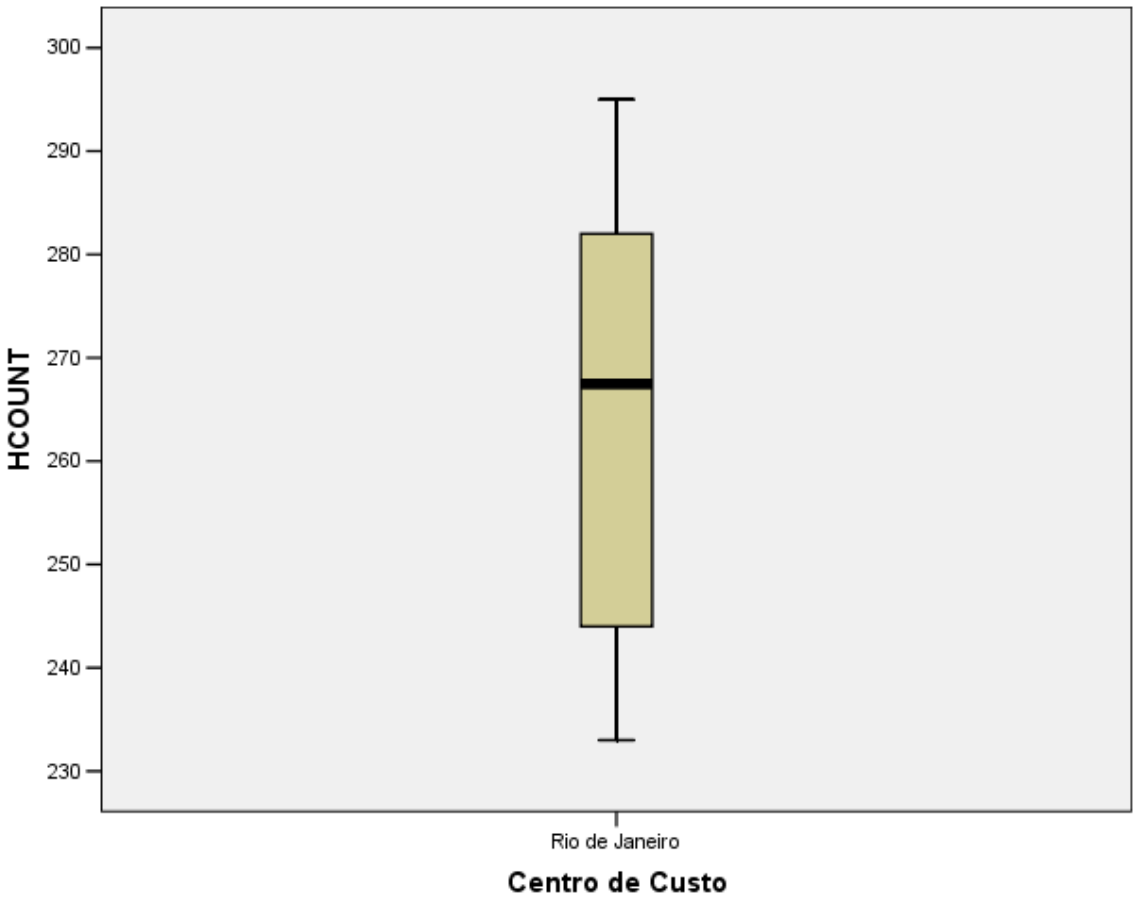
TABELA 6 – Análise Descritiva (HCOUNT)

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
HCOUNT	32	62	233	295	265,38	20,144
Valid N (listwise)	32					

FONTE: elaborado pelos próprios autores

GRÁFICO 3 – *Box Plot* (HCOUNT)



FONTE: elaborado pelos próprios autores

3. METODOLOGIA

3.1. Componentes da Regressão Linear e Medidas de Avaliação

Conforme apresentado por CORRAR, PAULO e DIAS FILHO (2007), a técnica de regressão linear é amplamente empregada na área de negócios principalmente com o propósito de previsão. Essa técnica consiste em determinar uma função matemática para descrever o comportamento de determinado indicador, dado os valores de outros indicadores já conhecidos (pode ser apenas um ou mais de um). O grande objetivo dessa técnica é alcançar valores previstos para o indicador alvo (variável dependente) com maior precisão em relação à simples utilização da média.

Segundo GUJARATI (2006), qualquer modelo de regressão linear pode ser expresso como:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon}$$

Onde,

- Y é a variável dependente;
- X_1, X_2, ..., X_n são as variáveis independentes;
- β_0, β_1, β_2, ..., β_n são os parâmetros da regressão;
- ε é o resíduo ponderado entre todas as observações reais e as estimadas (erro da regressão).

Em relação aos parâmetros, há uma consideração importante a ser feita, pois β_0 é o intercepto (coeficiente linear), isto é, representa o valor da interseção da curva de regressão com o eixo Y, isto é, β_0 é o valor de Y quando todas as variáveis independentes possuem valor igual a zero. Este parâmetro é amplamente conhecido como *constante* do modelo de regressão e pode ou não ser inserido na função matemática. Já o termos β_1, β_2, ..., β_n , que são chamados de coeficientes angulares.

Caso exista apenas β_1 (β_0 pode ou não existir), o modelo será uma regressão linear simples (apenas um coeficiente angular). Caso possua mais de um coeficiente angular, esta será uma regressão linear múltipla.

Os modelos de regressão apresentam, porém, alguns pressupostos que precisam ser seguidos para garantir a qualidade do resultado encontrado. Esses pressupostos são apresentados por CORRAR, PAULO e DIAS FILHO (2007) e GUJARATI (2006), conforme exposto abaixo:

a) A variável Y é aleatória;
b) A esperança matemática dos resíduos é nula, ou seja, a média dos resíduos é nula;
c) A variância de ε (termo de erro) é constante e igual a σ^2 (condição de homoscedasticidade dos resíduos);
d) Os resíduos são independentes entre si;
e) Os resíduos têm distribuição normal.

Observando os pontos já apresentados, fica claro que o objetivo de qualquer modelo de regressão é estimar os parâmetros de modo a alcançar o melhor valor de Y, isto é, minimizar ε.

Para a estimação do modelo, utiliza-se o Método dos Mínimos Quadrados (MMQ), cujo objetivo é obter a menor soma de quadrados dos resíduos (SQR) possível (CORRAR, PAULO e DIAS FILHO, 2007). O objetivo de trabalhar com os quadrados dos resíduos é eliminar a contraposição de sinais, visto que, caso essa eliminação não ocorra, qualquer modelo proposto será, no máximo, tão bom quanto à utilização da média, pois para esse caso, a soma dos resíduos será sempre igual a zero.

Além da análise de SQR, deve-se analisar R *Square* (R^2) como medida de qualidade do modelo proposto. O R^2 é denominado coeficiente de determinação ou poder explicativo da regressão (GUJARATI, 2006) e pode ser obtido elevando ao quadrado o coeficiente de correlação (R), que representa o grau de associação entre as variáveis dependentes e independentes (CORRAR, PAULO e DIAS FILHO, 2007).

O R^2 indica quanto da variação de Y é explicado pelas variações nas variáveis independentes.

Outro ponto de análise é o erro-padrão da estimativa, que representa uma espécie de desvio-padrão em torno da curva de regressão. Quanto menor o erro-padrão da estimativa, melhor o modelo estimado (CORRAR, PAULO e DIAS FILHO, 2007).

Todas as análises de qualidade do modelo, bem como de seleção das variáveis independentes a serem utilizadas serão explicadas ao longo da construção do modelo, no próximo capítulo, cabendo ainda ao capítulo de método o detalhamento dos pressupostos na análise de regressão. É importante ressaltar, porém, que a explicação de como validar os pressupostos também será apresentada no próximo capítulo, no decorrer da construção do modelo proposto.

3.2. Pressupostos na Análise de Regressão

Serão apresentados os principais pressupostos requeridos para a análise de regressão, sendo que é importante destacar que cada um possui a sua importância e todos devem ser respeitados para a aplicação apropriada desse procedimento estatístico. Vale ressaltar que o descumprimento de alguns desses pressupostos não inutiliza todo o trabalho desenvolvido, pois há formas de se trabalhar com os dados (transformações) a fim de buscar o enquadramento do resultado aos pressupostos necessários.

3.2.1. Normalidade dos Resíduos

Conforme apresentado por GUJARATI (2006), o conjunto dos resíduos produzidos em todo o intervalo das observações deve apresentar distribuição normal (normalidade dos resíduos).

> A condição de normalidade dos resíduos não é necessária para obtenção dos estimadores pelo métodos dos mínimos quadrados, mas sim para a definição de intervalos de confiança e testes de significância. (CORRAR, PAULO e DIAS FILHO, 2007, p. 152).

A omissão de variáveis explicativas importantes, a presença de *outliers* ou a definição incorreta da forma funcional da equação podem ser possíveis causas para a falta de normalidade dos resíduos.

Para avaliar o atendimento desse pressuposto, deve-se utilizar os testes estatísticos para avaliação de distribuição normal já comumente utilizados, como Kolmogorov-Smirnov (para amostras maiores que 30 elementos) e Shapiro-Wilk (para amostras menores que 30 elementos) (CORRAR, PAULO e DIAS FILHO, 2007).

3.2.2. Homocedasticidade

Este pressuposto preocupa-se com a variância residual. Para atende-lo, é necessário que os resíduos apresentem comportamento aleatório, sem padrão.

> O conjunto de resíduos referentes a cada observação de X deve ter variância constante ou homogênea em toda a extensão das variáveis independentes; isto é, a dispersão de Y em relação às observações deve manter consistência ou ser constante em todas as dimensões desta variável. Tal característica se define como homocedasticidade, ou seja, dispersão homogênea das ocorrências de Y em relação a cada observação de X. (CORRAR, PAULO e DIAS FILHO, 2007, p. 152).

Os principais causadores de quebra desse pressuposto são a seleção de apenas uma seção dos dados (concentração da amostra em um dado intervalo de tempo, de modo a examinar apenas uma parte da realidade da população), erro de especificação das variáveis ou da função matemática (CORRAR, PAULO e DIAS FILHO, 2007).

3.2.3. Ausência de Autocorrelação Serial

O modelo de regressão linear pressupõe que não exista correlação entre os resíduos, isto é, um valor encontrado para a variável Y não impacta no valor encontrado da próxima variável Y.

> A variável Y só sofre influência da própria variável X considerada e não dos efeitos defasados de X1 sobre X2 e desta sobre Y. Dito de outro modo, os resíduos são independentes entre si e só se observa o efeito de X sobre Y, ou seja, não existe autocorrelação residual. (CORRAR, PAULO e DIAS FILHO, 2007, p. 152).

Um dos métodos utilizados para diagnosticar a ausência de autocorrelação serial é o teste estatístico de Durbin-Watson, que será empregado e explicado no próximo capítulo.

3.2.4. Linearidade dos Coeficientes

Este pressuposto indica que a relação entre as variáveis escolhidas para compor o modelo deve ser representada matematicamente por uma função de primeiro grau (CORRAR, PAULO e DIAS FILHO, 2007).

3.2.5. Multicolinearidade

A multicolinearidade é o fenômeno no qual se analisa a correlação existente entre as diversas variáveis independentes, sendo que este fenômeno não é uma questão de natureza (existência ou não de multicolinearidade), mas sim de grau, pois sempre existirá correlação entre as variáveis independentes, mas quanto menor este grau, menor serão as dificuldades de interpretação dos resultados (GUJARATI, 2006).

A utilização de variáveis independentes altamente correlacionadas geram erros-padrão maiores, menor eficiência dos estimadores, estimativas mais imprecisas, estimadores sensíveis a pequenas variações nos dados (gerando dificuldade na separação dos efeitos de cada uma das variáveis). (CORRAR, PAULO e DIAS FILHO, 2007).

4. ANÁLISE E RESULTADOS

Para a realização da análise proposta nesse trabalho, será utilizado o software SPSS para gerar o modelo de regressão linear e todos os demais pontos a serem analisados. O nível de significância (α) adotado será de 5%. A variável dependente será tratada por HR_PASSE e as variáveis dependentes serão tratadas por:

- VOL → Volume diário transportado
- HCOUNT → *Headcount* de maquinistas

A construção do modelo será realizada em duas partes. Primeiro será utilizada apenas uma variável independente e depois será inserida a segunda variável. O modelo que apresentar o melhor desempenho será o escolhido para determinar o comportamento da quantidade de horas de passe (HRPASSE).

4.1. Modelo de Regressão Linear Simples

Dando início ao trabalho de criação do modelo, a primeira etapa consiste em analisar a matriz de correlação das variáveis. O resultado dessa matriz pode ser encontrado na tabela abaixo.

TABELA 7 – Matriz de correlação

Correlations

		HR_PASSE	VOL	HCOUNT
HR_PASSE	Pearson Correlation	1	,749**	,788**
	Sig. (2-tailed)		,000	,000
	N	32	32	32
VOL	Pearson Correlation	,749**	1	,562**
	Sig. (2-tailed)	,000		,001
	N	32	32	32
HCOUNT	Pearson Correlation	,788**	,562**	1
	Sig. (2-tailed)	,000	,001	
	N	32	32	32

**. Correlation is significant at the 0.01 level (2-tailed).

FONTE: elaborado pelos próprios autores

Nessa matriz observa-se que todas as variáveis independentes selecionadas possuem grande correlação com HR_PASSE (*Sig*. menor que α), sendo que as variáveis HCOUNT e VOL são, nessa ordem, as mais correlacionadas com HR_PASSE.

Como a maior correlação se dá com a variável HCOUNT (R = 0,788), esta será selecionada para a construção da equação de regressão usando a apenas uma variável independente. Abaixo estão apresentados os resultados dessa regressão.

TABELA 8 – Sumário do modelo (regressão linear simples)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,788[a]	,621	,608	282,446

a. Predictors: (Constant), HCOUNT

FONTE: elaborado pelos próprios autores

O coeficiente de correlação (R) representa apenas o grau de associação entre a variável dependente HCOUNT e a variável independente HR_PASSE, que é de 0,788 (esse foi o parâmetro utilizado para a escolha desta variável entre as duas existentes).

O coeficiente de determinação (R^2) indica que 62,1% da variação da variável dependente HR_PASSE é explicada pelas variações ocorridas na variável independente HCOUNT.

TABELA 9 – ANOVA (regressão linear simples)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3922009	1	3922008,821	49,163	,000[a]
	Residual	2393271	30	79775,702		
	Total	6315280	31			

a. Predictors: (Constant), HCOUNT

b. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

A soma total dos resíduos quadrados apresenta o valor de 6.315.280, isto é, este é o resíduo quadrado que ocorreria se fosse utilizada apenas a média da variável dependente HR_PASSE para predição. Utilizando a variável independente HCOUNT, esse resíduo cai para 2.392.271, mostrando que o modelo gerado até agora já é melhor do que a utilização da média.

Como o *Sig*. do Teste F - ANOVA apresenta o valor de 0,000 (menor que $\alpha = 0,05$), rejeita-se a hipótese de que R^2 é igual a zero. Desse modo, pode-se dizer que a variável estatística exerce influência sobre a variável dependente, ou seja, o modelo é significativo.

TABELA 10 – Coeficientes (regressão linear simples)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1572,599	670,151		-2,347	,026
	HCOUNT	17,657	2,518	,788	7,012	,000

a. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

Como foi utilizada uma constante na geração desse modelo, tem-se que o valor previsto para cada observação é o valor do intercepto (constante) -1.572,599 mais o coeficiente de regressão (HCOUNT) 17,657 multiplicado pelo valor da variável independente. Dessa forma, a equação de regressão pode ser escrita da seguinte forma:

HR_PASSE = -1.572,599 + 17,657 HCOUNT

Através dessa equação, pode-se dizer que, a cada 1 ponto percentual de aumento na quantidade de maquinistas, a quantidade de horas de passe sofre, em média, um aumento de 17,657 pontos percentuais.

Outra observação importante é que o Teste t mostrou que o *Sig*. do intercepto é maior que α, o que pode significar que o mesmo não deveria ser utilizado para fins preditivos. Em termos práticos, porém, não é necessário testar o termo constante (CORRAR, PAULO e DIAS FILHO, 2007). Já o coeficiente de regressão da variável independente difere significativamente de zero (*Sig*. menor que α).

4.2. Modelo de regressão linear múltipla

A próxima etapa é a inserção da nova variável no modelo. Para a realização dessa etapa, deve-se avaliar a correlação parcial controlada pela variável que já entrou na regressão (CORRAR, PAULO e DIAS FILHO, 2007), que nesse caso é HCOUNT. Como há apenas uma variável a ser escolhida, essa avaliação não se faz necessária, partindo-se logo para a rotina de estimação do modelo com mais de

uma variável independente. A tabela abaixo apresenta os resultados dsse novo modelo utilizando duas variáveis independentes.

TABELA 11 – Sumário do modelo (regressão linear múltipla)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,870[a]	,758	,741	229,692

a. Predictors: (Constant), VOL, HCOUNT

FONTE: elaborado pelos próprios autores

O coeficiente de determinação (R^2) com a inclusão da variável VOL aumentou 13,7% (poder explicativo adicional de VOL). O conjunto de variáveis independentes explica, assim, 75,8% da variação na variável dependente HR_PASSE.

O coeficiente de determinação ajustado (R^2 ajustado) é uma medida modificada do coeficiente de determinação que considera o número de variáveis independentes incluídas no modelo e o tamanho da amostra. Quando o objetivo é a comparação entre equações, é uma medida mais útil que o R^2 (CORRAR, PAULO e DIAS FILHO, 2007). O primeiro modelo apresentou R^2 ajustado de 0,608, contra 0,741 do modelo atual, demonstrando que o modelo de regressão múltipla é superior em relação modelo de regressão simples neste caso.

O erro padrão da estimativa (*Std. Error of the Estimate*) também é considerada uma medida de precisão das previsões (CORRAR, PAULO e DIAS FILHO, 2007) e, sua diminuição de 282,466 para 229,692 corrobora com a afirmação de maior ajustamento do modelo de regressão múltipla.

TABELA 12 – ANOVA (regressão linear múltipla)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4785284	2	2392642,072	45,351	,000[a]
	Residual	1529996	29	52758,474		
	Total	6315280	31			

a. Predictors: (Constant), VOL, HCOUNT

b. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

Os resíduos quadrados deixados pelo modelo com duas variáveis (1.529.996) são menores que os da regressão simples (2.392.271). O modelo estimado com duas variáveis independentes é, portanto, mais preciso que a equação com uma única variável.

O Teste F – ANOVA apresenta o *Sig*. menor que α, rejeitando-se, portanto, a hipótese de que o R^2 é igual a zero. Dessa forma, tem-se que, pelo menos uma das variáveis independentes exerce influência sobre a HR_PASSE, logo, o modelo é significativo como um todo.

TABELA 13 – Coeficientes (regressão linear múltipla)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2442,960	585,920		-4,169	,000
	HCOUNT	12,034	2,475	,537	4,862	,000
	VOL	,008	,002	,447	4,045	,000

a. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

A equação de regressão do novo modelo pode ser expressa da seguinte forma:

HR_PASSE = -2.442,960 + 12,034 HCOUNT + 0,008 VOL

A variação de 1 ponto percentual em HCOUNT provoca uma acréscimo de 12,034 pontos percentuais, em média, na variável dependente e, a variação de 1 ponto percentual em VOL provoca um acréscimo de 0,008 pontos percentuais, em média, em HR_PASSE.

As variáveis independentes estão expressas em unidades diferentes, o que torna difícil a comparação do peso de cada coeficiente no modelo de regressão. A padronização dos coeficientes é a ferramenta estatística que permite comparar esses valores e nada mais é do que a divisão do coeficiente pelo seu desvio padrão (CORRAR, PAULO e DIAS FILHO, 2007). Desse modo, percebe-se que os pesos

dos coeficientes são muito próximos (HCOUNT é apenas 20% maior que VOL, aproximadamente).

Pelo Teste t, percebe-se que a probabilidade dos coeficientes de HCOUNT e VOL serem estatisticamente nulos tende a zero (*Sig*. menor que α). Neste modelo de regressão múltipla, essa afirmação é verdadeira também para a constante do modelo.

Considerando a significância estatística dos estimadores, estes podem ser usados para predizer a quantidade de horas de passagem, dado o *headcount* de maquinistas e o volume diário transportado. É importante apenas lembrar que há 24,2% de variação de HR_PASSE que não estão sendo explicadas por esse modelo, sendo este um percentual aceitável para a proposta desse estudo.

Finalizada a escolha do modelo, dá-se início a última parte do trabalho, que consiste na avaliação dos pressupostos que garantem a integridade dos testes de ajustamento e de significância do modelo.

4.2.1. Análise do Pressuposto de Multicolinearidade

É importante ressaltar que o problema da multicolinearidade normalmente tem relação com regressões que apresentam R^2 altos e coeficientes não significativos. Nesse estudo, R^2 pode ser considerado de médio a alto, porém os coeficientes são significativos, logo não se espera essa multicolinearidade. Na tabela abaixo são apresentadas as estatísticas *Tolerance* e VIF (*Variance Inflaction Factor*), que são mediadas recíprocas (mesma interpretação).

> O cálculo da medida *Tolerance* é feito estimando cada variável independente como se dependente fosse e regredindo-a em relação às demais, e obtendo-se, assim, o valor (1 – R2) de tal regressão; portanto, quando *Tolerance* (ou VIF) são próximos da unidade, é indicativo de não-detecção de multicolinearidade, pois o Coeficiente de Determinação terá sido próximo de zero. (CORRAR, PAULO e DIAS FILHO, 2007, p. 187).

TABELA 14 – Estatística de colinearidade (regressão linear múltipla)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-2442,960	585,920		-4,169	,000		
	HCOUNT	12,034	2,475	,537	4,862	,000	,685	1,461
	VOL	,008	,002	,447	4,045	,000	,685	1,461

a. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

A análise de VIF é dada pela literatura (GUJARATI, 2000; HAIR, 2005) da seguinte forma:

- Até 1 – sem multicolinearidade;
- De 1 até 10 – com multicolinearidade aceitável;
- Acima de 10 – com multicolinearidade problemática.

Pelas análises serem recíprocas, o índice de *Tolerance* será o inverso:

- Até 1 – sem multicolinearidade;
- De 1 até 0,10 – com multicolinearidade aceitável;
- Abaixo de0, 10 – com multicolinearidade problemática.

No presente trabalho, portanto, não se detectam problemas de multicolinearidade (dados os testes utilizados).

4.2.2. Análise do Pressuposto de Autocorrelação Serial nos Resíduos

Para essa análise será utilizado o tese de DURBIN-WATSON que baseia-se em cálculo de medida conhecida como Estatística DW, tabelada para valores críticos (segundo o nível de confiança escolhido) (CORRAR, PAULO e DIAS FILHO, 2007).

A tabela abaixo apresenta o resultado desta estatística, que possui a seguinte formulação de hipótese:

- H_0: Não existe correlação serial dos resíduos;
- H_1: Existe correlação serial dos resíduos.

TABELA 15 – Teste de Durbin-Watson (regressão linear múltipla)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,870[a]	,758	,741	229,692	1,738

a. Predictors: (Constant), VOL, HCOUNT

b. Dependent Variable: HR_PASSE

FONTE: elaborado pelos próprios autores

Para esta análise, é preciso considerar o tamanho da amostra (n) e o número de variáveis independentes (p) para estabelecer o valor crítico inferior (d_L) e o valor crítico superior (d_U). A análise se baseia na seguinte regra para comparação apresentada por CORRAR, PAULO e DIAS FILHO (2007) para os valores encontrados na estatística de DURBIN-WATSON (1,738).

- Entre 0 e d_L – Autocorrelação positiva;
- Entre d_L e d_U – Não conclusivo;
- Entre d_U e 4-d_U – Ausência de autocorrelação;
- Entre 4-d_U e 4-d_L – Não conclusivo;
- Entre d_L e 4 – Autocorrelação negativa.

Para uma amostra de 32 elementos e 2 variáveis independentes, tem-se os valores de d_L = 1,309 e d_U =1,574, conforme tabela apresentada por GUJARATI (2006). Logo, 4-d_U = 2,426 e 4-d_L =2,691.

É possível concluir que a Estatística DW apresentada se encontra entre d_U e 4-d_U, caracterizando ausência de autocorrelação serial, atendendo ao pressuposto da regressão. O resultado encontrado corrobora com a regra amplamente adotada de que valores de Estatística DW próximos a 2 atendem ao pressuposto.

4.2.3. Análise do Pressuposto de Normalidade dos Resíduos

A avaliação desse pressuposto é realizada através do teste Kolmogorov-Smirnov, que examina se a série apresenta distribuição próxima à distribuição normal.

Para isto, são formuladas as seguintes hipóteses:

- H_0: A distribuição da série testada é normal;
- H_1: A distribuição da série testada não tem comportamento normal.

Para a estatística K-S, portanto, o resultado é esperado é *Sig*. maior que α (não se pode rejeitar a hipótese nula). Deve-se ressaltar que este teste deve ser realizado nos resíduos padronizados, aqui chamados de ZRE_1.

Conforme observa-se na tabela abaixo, não há indícios para rejeitar a hipótese de normalidade da distribuição dos resíduos padronizados (*Sig*. maior que 0,05), o que confirma o atendimento desse terceiro pressuposto.

TABELA 16 – Teste de normalidade de resíduos (regressão linear múltipla)

One-Sample Kolmogorov-Smirnov Test

		Standardized Residual
N		32
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	,96720415
Most Extreme Differences	Absolute	,083
	Positive	,083
	Negative	-,074
Kolmogorov-Smirnov Z		,469
Asymp. Sig. (2-tailed)		,980

a. Test distribution is Normal.

b. Calculated from data.

FONTE: elaborado pelos próprios autores

Caso a amostra fosse menor do que 30 elementos, esse teste deveria ser substituído pelo teste de normalidade de Shapiro-Wilk (CORRAR, PAULO e DIAS FILHO, 2007).

4.2.4. Análise do Pressuposto de Homocedasticidade dos Resíduos

Para testar esse último caso será empregado o Teste de Pesarán-Pesarán, desenvolvido para examinar se a variância dos resíduos mantém-se constante ao longo de toda a amostra (homocedasticidade).

Como apresentado por CORRAR, PAULO e DIAS FILHO (2007), esse teste implica em regredir o quadrado dos resíduos padronizados (aqui chamados de

ZRE_2) como função do quadrado dos valores estimados padronizados (aqui chamados de ZPR_2). De forma simplificada, pode-se dizer que é uma regressão simples com o quadrado dos resíduos padronizados como variável dependente e o quadrado dos valores estimados padronizados como variável independente.

As hipóteses a serem testadas são:

- H_0: Os resíduos são homocedásticos;
- H_1: Os resíduos são heterocedásticos.

O objetivo é avaliar a significância estatística do coeficiente ZPR_2. Se o coeficiente for estatisticamente significante, indica a presença de heterocedasticidade, pois os resíduos são influenciados pela variável dependente, não tendo um comportamento aleatório em relação às variáveis independentes (não atendimento ao pressuposto em questão).

TABELA 17 – ANOVA para Análise de Homocedasticidade (Regressão Linear Múltipla)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,139	1	,139	,080	,779[a]
	Residual	51,918	30	1,731		
	Total	52,058	31			

a. Predictors: (Constant), ZPR_2

b. Dependent Variable: ZRE_2

FONTE: elaborado pelos próprios autores

Conforme apresentado acima, não há indícios para rejeitar a hipótese de homecedasticidade, visto que *Sig*. é maior que 0,05. Em outras palavras pode-se dizer que a variância dos resíduos da variável dependente é constante para todas as observações referentes a cada conjunto de valores das variáveis independentes (CORRAR, PAULO e DIAS FILHO, 2007).

Analisando todos os resultados encontrados, pode-se dizer que a equação abaixo representa de forma satisfatória o comportamento do tempo de passagem a partir de variações no *headcount* de maquinistas e volume transportado.

HR_PASSE = -2.442,960 + 12,034 HCOUNT + 0,008 VOL

5. CONCLUSÃO

Toda a argumentação trabalhada nos capítulos anteriores garantiu ferramentas para análise do resultado desse trabalho, quanto aos objetivos propostos. Dois modelos de regressão linear (um simples e um múltiplo) foram testados para explicar o comportamento da quantidade de horas de passagem entre a apresentação na sede e o início da jornada em trem da tripulação dos trens da malha sudeste na região do Rio de Janeiro.

O modelo de regressão linear múltiplo obteve o melhor desempenho, visto que apresentou o valor de apenas 1.529.996 resíduos quadrados, frente aos 2.392.271 deixados pelo modelo de regressão linear simples (utilizando a variável dependente HCOUNT) e 6.315.280 deixados pelo cálculo a partir da média. Quando se analisa o erro padrão da estimativa, obtêm-se a redução de 282,466 para 229,692 quando se troca a regressão simples pela múltipla, resultado que corrobora com a afirmação de maior ajustamento deste último modelo.

Partindo para o poder explicativo dessas funções matemáticas, percebe-se que quando se utiliza duas variáveis independentes, 75,8% das variações no tempo de passagem estão sendo explicadas pela função gerada. Em contrapartida, quando se utiliza apenas uma varável independente, apenas 62,1% dessas variações estão sendo explicadas.

Quanto ao atendimento aos pressupostos, estes foram verificados apenas para a regressão múltipla, visto que foi o modelo com maior ajustamento. Todos os pontos analisados obtiveram resultados satisfatórios em relação ao seu atendimento, sendo que foram testados os seguintes pressupostos:

a) Multicolinearidade;
b) Ausência de autocorrelação serial nos resíduos;
c) Normalidade dos resíduos;
d) Homocedasticidade dos resíduos.

O pressuposto de linearidade dos coeficientes também foi atendido, visto que a função matemática gerada foi:

HR_PASSE = -2.442,960 + 12,034 HCOUNT + 0,008 VOL

Tem-se, portanto, que a equação anterior pode ser utilizada para explicar o comportamento da variável HR_PASSE (com R^2 igual a 75,8%) tanto para a tomada de decisões no curto prazo quanto no longo prazo, sendo necessário apenas obter os valores de HCOUNT (*headcount* de maquinistas) e VOL (volume diário transportado).

6. REFERÊNCIAS

CORRAR, L. J. ; PAULO, E. e DIAS FILHO, J. M.. **Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia**. São Paulo: Editora Atlas, 2007

GUJARATI, D. N.. **Econometria Básica. Tradução Maria José Cyhlar Monteiro**. Rio de Janeiro: Editora Elsevier, 2006

SOARES, T.M.. **Curso de Análise de Regressão**. Material de aula do curso de Especialização em Métodos Estatísticos Computacionais do Instituto de Ciências Exatas da Universidade Federal de Juiz de Fora.

7. ANEXOS

- Base de dados utilizada no estudo.

CC	HCOUNT	VOL	HR_PASSE
Rio de Janeiro	233	244158	2500
Rio de Janeiro	233	270585	2625
Rio de Janeiro	238	261330	2294
Rio de Janeiro	241	283639	2964
Rio de Janeiro	243	290303	2839
Rio de Janeiro	243	292752	2742
Rio de Janeiro	243	300839	2807
Rio de Janeiro	244	283216	2733
Rio de Janeiro	244	313227	2526
Rio de Janeiro	246	319997	2900
Rio de Janeiro	253	292013	2817
Rio de Janeiro	255	272708	2839
Rio de Janeiro	258	274638	2815
Rio de Janeiro	265	312141	2967
Rio de Janeiro	267	281232	2618
Rio de Janeiro	266	287626	3158
Rio de Janeiro	276	319446	3161
Rio de Janeiro	272	306292	3475
Rio de Janeiro	268	334858	3301
Rio de Janeiro	275	324524	3463
Rio de Janeiro	272	339211	4027
Rio de Janeiro	272	353255	3485
Rio de Janeiro	279	309964	3361
Rio de Janeiro	281	278762	3073
Rio de Janeiro	283	265666	3022
Rio de Janeiro	285	321967	3211
Rio de Janeiro	293	311231	3131
Rio de Janeiro	295	311034	3547
Rio de Janeiro	290	319325	3452
Rio de Janeiro	295	330154	3950
Rio de Janeiro	295	323408	3775
Rio de Janeiro	289	330800	4041

- *Syntax* do SPSS

```
CORRELATIONS
  /VARIABLES=HR_PASSE VOL HCOUNT
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT HR_PASSE
  /METHOD=ENTER HCOUNT  .


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT HR_PASSE
  /METHOD=ENTER HCOUNT VOL  .



-------> Criação dos valores previstos padronizados e resíduos padronizados

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT HR_PASSE
  /METHOD=ENTER HCOUNT VOL
  /SAVE ZPRED ZRESID .

COMPUTE ZPR_2 = ZPR_1 * ZPR_1 .
EXECUTE .


COMPUTE ZRE_2 = ZRE_1 * ZRE_1 .
EXECUTE .



-------> Relização dos testes para os pressupostos


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT HR_PASSE
  /METHOD=ENTER HCOUNT VOL  .
```

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT HR_PASSE
  /METHOD=ENTER HCOUNT VOL
  /RESIDUALS DURBIN .


NPAR TESTS
  /K-S(NORMAL)= ZRE_1
  /MISSING ANALYSIS.


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT ZRE_2
  /METHOD=ENTER ZPR_2  .
```