Ivanildo Batista Marckis Lima

Regressão linear

Recife-PE
22 de novembro de 2021

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Os pesquisadores estão na maioria das vezes interessados em encontrar relações entre uma variável e outras variáveis. Uma forma de encontrar essa relação é por meio da **análise de correlação** que é usada para quantificar ou medir a força ou grau de associação linear entre duas variáveis. A forma gráfica de verificar essa relação é por meio do **gráfico de dispersão**, que permite visualizar como uma variável se comporta em relação a outra. A medida mais comumente usada na análise de correlação é o coeficente de correlação (de *Pearson*, de *Spearman* ou de *Kendall*). Entretanto outra forma de realizar dessa identificação é por meio da **análise de regressão**.

O termo regressão foi usado pela primeira vez por Francis Galton em seu artigo *Family likeness in stature*, onde ele observou que, embora pais altos tivesse filhos altos e pais baixos tivesse filhos baixos, a estatura média das crianças tendia a "regredir"(daí dar o nome de regressão) para a média populacional. A análise de Galton visava saber se havia estabilidade na distribuição das alturas, entretanto a preocupação da análise moderna é descobrir como a altura média dos filhos varia, dada a altura dos pais.

A análise de regressão trata do estudo de uma variável que é dependente de outra ou outras variáveis (explanatórias ou independentes). O objetivo é prever/estimar o valor médio da variável dependente em termos dos valores das variáveis independentes, conforme Gujarati e Porter (2011). A regressão é o modelo matemático que relaciona a variável dependente com as variáveis independentes. Para Angrist e Pischke (2008) a regressão é um dispositivo computacional para estimar diferenças entre um grupo de controle e de tratamento em um experimento. Andrade e Tiryaki (2019) afirma que o modelo de regressão é o mais popular para estudar a relação entre variáveis e isso deve-se pela sua fácil aplicação, baixo custo computacional, interpretação simples, apresenta propriedades que são desejáveis e a maioria dos pacotes estatísticos incluem rotinas prontas para a estimação desse modelo.

A regressão pode ser usada em diversas áreas como **economia** (relação entre despesas de consumo pessoal e renda pessoal, descobrir a resposta da demanda por um produto frente a variação dos preços, relação entre salário e desemprego, etc), **saúde** (relação entre pressão ocular e idade, mortalidade e consumo de drogras lícitas ou ilícitas), **agronomia** (dependência do rendimento de uma plantação em relação à temperatura, à quantidade de chuva e de sol e à aplicação de fertilizantes), entre outras.

Yan e Su (2009) explica que existem três tipos de regressão. A primeira é a **regressão linear simples**, que será o objeto de dsicussão nesse trabalho mais a frente. O segundo tipo de modelo de regressão é a **regressão linear múltipla**, que serve para para modelar a relação entre uma variável dependente y com mais de uma variável independente $(x_1, x_2, \ldots, x_p)$. O modelo de regressão pode ser escrito conforme abaixo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{1.1}$$

o y é a variável dependente, os coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são os regressores, $x_1, x_2, \ldots, x_p$ são as variáveis dependentes e ε é o termo de erro, que assim como no modelo anterior, é normalmente distribuído com $E(\varepsilon) = 0$ e a variância é constante $Var(\varepsilon) = \sigma^2$. Notar que a regressão linear simples é um caso particular da regressão linear múltipla, quando $x_2 = x_3 = \cdots = x_p = 0$.

O terceiro tipo de regressão é a **regressão não linear** a qual assume que a relação entre a variável dependente a as variáveis independentes é não linear nos parâmetros (diferentes das anteriores que se assume a linearidade dos parâmetros). Um exemplo desse tipo de regressão pode ser visto abaixo

$$y = \frac{\alpha}{1 + e^{\beta t}} + \varepsilon \tag{1.2}$$

Onde y é o crescimento de um determinado organismo em função do tempo t, α e β são os parâmetros do modelo e o ε é o erro. Em termos de estimação dos parâmetros, seleção do modelo, diagnóstico, seleção de variáveis, detecção de *outliers* e de observações influentes, os modelos não lineares são mais complicados.

Yan e Su (2009) ainda explica que o modelo de regressão possui três objetivos:

1- Estabelecer uma relação causal entre a variável dependente y e os seus regressores $x_1, x_2, \ldots, x_p$.

2- Prever o valor de y com base no conjunto de valores de $x_1, x_2, \ldots, x_p$. É preciso definir se o modelo pode ser escrito conforme abaixo

$$\text{Variável resposta} = \text{função do regressores} + \text{erro}$$

ou apenas no formato matemático

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon$$

3- Realizar uma seleção criteriosa dos regressores $x_1, x_2, \ldots, x_p$ para identificar quais as variáveis são mais importantes do que outras para explicar o comportamento da variável dependente y, para que a relação causal possa ser determinada de forma mais precisa e eficiente.

Nesse trabalho iremos abordar os conceitos por trás na regressão linear simples no Capítulo 2 , seus pressupostos no Capítulo 3, formas de estimação dos parâmetros do modelo no Capítulo 4, avaliação do modelo no Capítulo 5, inferência dos parâmetros no Capítulo 6 e, por fim, serão realizados exemplos no Capítulo 7 com a linguagem *R*.

2 Modelo de regressão linear simples

A regressão linear simples (também chamada de regressão de primeiro grau) serve para modelar a relação linear entre duas variáveis. Uma deles é a variável dependente e outra é a variável independente.

Ela possui o nome de *linear*, pois o valor da variável dependente é uma função linear da variável independente. Dessas variáveis são gerados os parâmetros da função de regressão, são esses parâmetros ($\beta's$) que definem a linearidade da função, pois eles são elevados a primeira potência. Sendo assim a regressão linear simples tem como notação genérica a fórmula abaixo

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

A função acima possui cinco componentes, que são:

- A variável dependente y é variável que deseja-se saber seu comportamento, também é chamada de **variável alvo** ou **variável resposta**.
- A variável x é a variável independente e será usada para explicar o comportamento da variável resposta, também é chamada de **preditor** ou **variável exploratória**.
- Para a regressão linear simples, o β_0 é o intercepto do modelo ou o valor y, quando o valor de x é igual a zero (variação média de y quando x não varia). Também é conhecido coomo **constante da regressão**.
- Para a regressão linear simples, o β_1 é a inclinação da reta de regressão. Ele é o **coeficiente de regressão** e representa a variação de y em função da variação de uma unidade da variável x.
- ε é o erro ou **resíduo**, diferença entre o valor observado de y e o correspondente ponto da reta de regressão. Assume-se que tem valor esperado igual a zero, $E(\varepsilon) = 0$ e que a sua variância é constante, $Var(\varepsilon) = \sigma^2$. Pode ser interpretado como aquilo que o modelo não consegue explicar.

A Equação 2.1 também pode ser chamada de função de regressão populacional (FRP), entretanto as informações sobre uma população muitas vezes são desconhecidas, por conta disso usa-se amostras populacionais para estimar os parâmetros. Por conta disso os parâmetros são estimativas, então essa função deve ser representada como abaixo

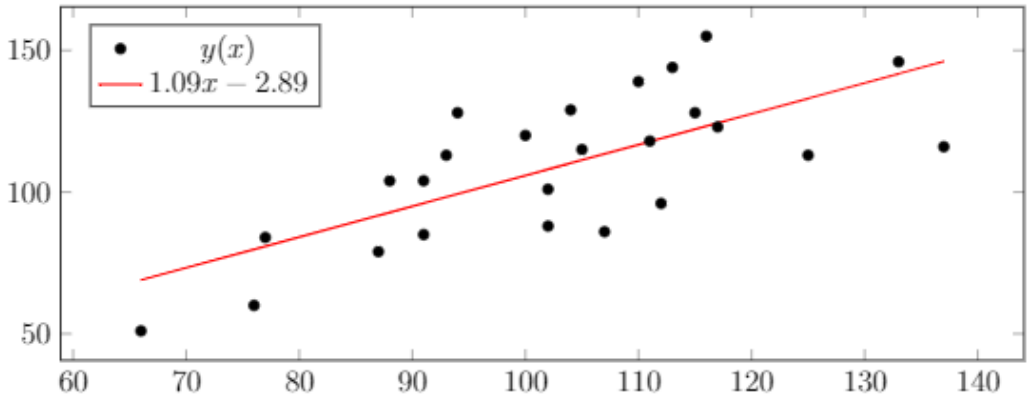
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\varepsilon} \tag{2.2}$$

Onde $\hat{y}$ é o estimador de y, $\hat{\beta}_0$ é o estimador de β_0, $\hat{\beta}_1$ é o estimador de β_1 e $\hat{\varepsilon}$ o erro estimado. Para essa função damos o nome de função de regressão amostral (FRA).

2.1 Exemplo de regressão linear simples

Na Figura 3 temos um exemplo de uma reta de regressão com a função de regressão em sua legenda. Os dados usados são de consumo de cigarros e taxa de mortalidade, sendo obtidos do exemplo da introdução do livro *Linear Regression Analysis* do Yan e Su (2009). Esses mesmos dados estão na Tabela 1.

Figura 1 – Exemplo de reta de regressão



Na figura vemos a reta de regressão na cor vermelha e as observações como pontos na cor preta. A reta de regressão é a reta que tem a menor distância para todas as observações. Na legenda da Figura 3 há a equação da regressão

$$y(x) = -2.89 + 1.09x \tag{2.3}$$

Observa-se Figura 3 que a correlação entre as variáveis é positiva: a medida que o valor do número de fumantes aumenta, o valor da taxa de mortaludade também aumenta. A vantagem da análise de regressão é trazer informações adicionais sobre essa relação : o $\hat{\beta}_0$ dessa equação é o valor de -2.89 e o $\hat{\beta}_1$ o valor de 1.09. A interpretação dada a esse modelo é que a variação média de uma unidade de número de fumantes leva a um aumento (o valor da inclinação é positivo) médio de 1.09 unidade na taxa de mortalidade.

Essa função de regressão estimada permite saber para qual número de fumante a taxa de mortalidade média será igual a zero, $y(x) = 0$

$$0 = -2.89 + 1.09x \tag{2.4}$$

$$2.89 = 1.09x \tag{2.5}$$

$$x = \frac{2.89}{1.09} = 2.65 \approx 3 \tag{2.6}$$

Isolando o valor de x temos que quando o número de fumantes for aproximadamente (arredondando de 2.65 para 3, já que não existem 2.65 pessoas), então a taxa de mortalidade média será aproximadamente zero.

Tabela 1 – Número de fumantes x Taxa de mortalidade

Número de fumantes	77	112	137	113	117	110	94	125	116	**133**
Mortalidade	84	96	116	144	123	139	128	113	155	146
Número de fumantes	102	115	111	105	93	87	88	91	102	100
Mortalidade	101	128	118	115	113	79	104	85	88	120
Número de fumantes	91	76	104	66	107					
Mortalidade	104	60	128	51	86					

Podemos usar a Equação 2.3 para estimar os valores da taxa de mortalidade média para cada número de fumantes observado. Na Tabela 2 temos os valores reais e os valores estimados da taxa de mortalidade.

Tabela 2 – Mortalidade observada x Mortalidade estimada

Observada	84	96	116	144	123	139	128	113	155	146
Estimada	81,04	119,2	146,44	120,3	124,6	117,01	99,6	133,4	123,6	142,1
Observada	101	128	118	115	113	79	104	85	88	120
Estimada	108,3	122,5	118,1	111,6	98,5	91,94	93,0	96,3	108,3	106,11
Observada	104	60	128	51	86					
Estimada	96,3	79,95	110,5	69,1	113,74					

Por ser um modelo simples é normal que os valores não sejam exatos, mas pode-se observar valores estimados bem próximos dos reais, como por exemplo $(84 - 81.04), (123 - 120.3), (146 - 142.1), (118 - 118.1)$. E a diferença entre esses valores reais e os estimados (ou observados) será o nosso erro ε ou **resíduo**.

Abaixo temos a Tabela 3 com os resíduos do nosso modelo que, no Capítulo 5, serão fruto de análise.

Tabela 3 – Resíduos

2,96	-23,19	-30,44	23,72	-1,64	21,99	28,43	-20,36	31,45	3,92
-7,29	5,54	-0,1	3,44	14,52	-12,94	10,97	-11,3	-20,29	13,89
7,7	-19,95	17,53	-18,05	-27,74					

3 Pressupostos

O modelo de regressão linear baseia-se em várias suposições que determinam o quão bem ele opera. A maioria deles diz respeito às características dos dados populacionais e enfoca os erros de previsão (ε_i). Mas ter acesso às informações de uma população é pouco comum, então devemos avaliar, de forma aproximada ou indireta, as suposições do modelo de regressão linear com informações de uma amostra. Em outras palavras, como não temos informações da população, não podemos calcular ε_i diretamente. A amostra inclui apenas de x_i e de y_i, portanto, devemos usar uma estimativa de ε_i. Esta estimativa, descrita anteriormente como o termo de erro na Equação 2.1, é representado pelos resíduos do modelo, que são calculados como $(y_i - \hat{y}_i)$. Em vez de distinguir os erros de previsão da população e da amostra, no entanto, presumiremos que a amostra fornece uma boa estimativa de Y_i (valores de y_i para a população) com $\hat{y}_i$, de modo que $(y_i - \hat{y}_i) \cong (y_i - Y_i)$.

3.1 Suposições para o modelo

As suposições necessárias para o Modelo de Regressão Linear são:

i O erro tem média zero e variância σ^2 , desconhecida;

ii Os erros são não correlacionados;

iii Os erros têm distribuição normal;

iv A variável regressora explicativa assume valores fixos.

As suposições i e iii, simbolicamente, podem ser representadas por:

$$\varepsilon_i \sim N(0, \sigma^2) \tag{3.1}$$

3.2 Pressupostos da Análise de Regressão

Para obtenção dos resultados, a análise de regressão baseia-se em quatro pressupostos básicos :

3.2.1 Linearidade

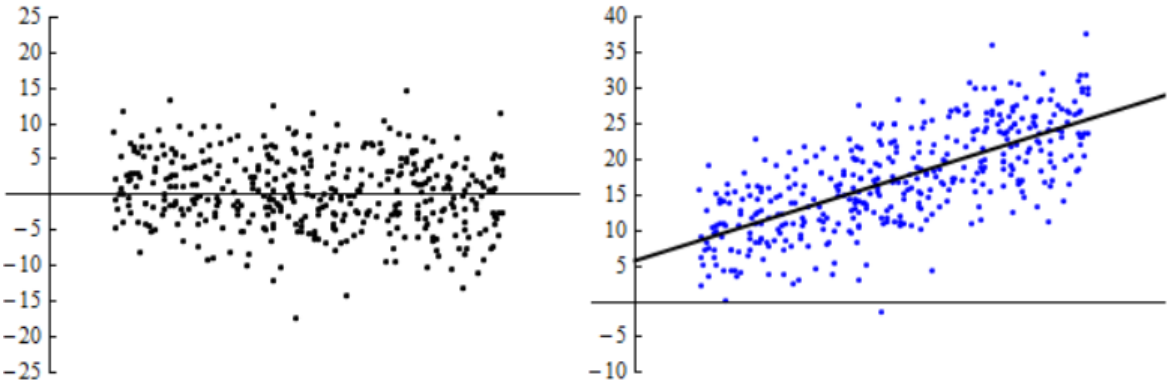
O valor médio de y_i é uma função em linha reta de x_i. Em outro palavras, y_i e x_i têm uma relação linear. Apesar de parecer um pressuposto restritivo matematicamente toda função não-linear pode ser transformada numa função linear através de técnicas

logarítmicas, polinomiais e de relações recíprocas. Não nos cabe neste texto discutir as formulações matemáticas de transformação, porém a sua existência é de fundamental importância uma vez que a análise de regressão não pode ser aplicada se a função não puder ser transformada para a forma linear.

3.2.2 Independência dos Resíduos

Os erros de previsão (ε_i) são estatisticamente independentes um do outro. Na prática, isso muitas vezes implica que as observações são independentes. Uma maneira de (quase) garantir isso é para usar amostragem aleatória simples. A violação do pressuposto da independência dos resíduos implica na existência de forte correlação (autocorrelação) entre os residuais sucessivos. Isto é, e_t não é independente de $e_{t-1} \ldots, e_{t-i+1} \ldots, e_{t+1}, e_{t+2}, \ldots, e_{t+n}$. A falta de independência não afeta o valor dos parâmetros estimados, mas afeta diretamente as variâncias estimadas. A falta de independência dos resíduos implica em R^2 e *estatística* F elevados e *teste* t reduzido se a autocorrelação é positiva e todos os testes com resultados elevados se a autocorrelação for negativa. Na Figura 2, considerando os resíduos no eixo $\mathbf{y}$ e os valores treinados $\hat{y}_i$ no eixo $\mathbf{y}$, vemos no primeiro gráfico que não há um padrão nos dados, mas no segundo gráfico já percebe-se a presença de uma tendência nos resíduos que é um comportamento de um modelo com autocorrelação (sem independência nos resíduos).

Figura 2 – Independência *vs* Dependência

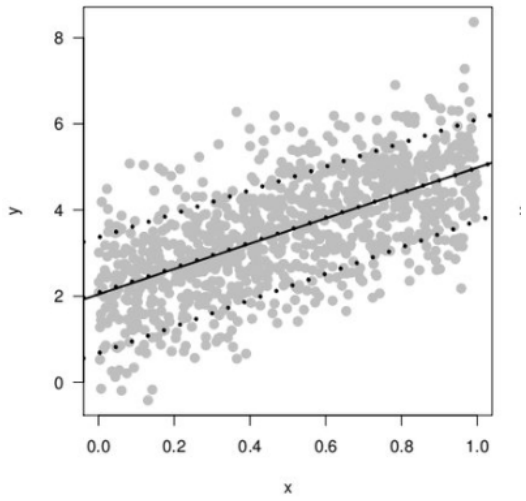


3.2.3 Homocedasticidade (ou variância constante)

Os erros de previsão têm variância equivalente para todos os valores possíveis de x_i. Em outras palavras, a variância dos erros é considerada constante ao longo da distribuição de x_i. Neste ponto, pode ser mais simples, embora impreciso, pensar sobre os valores de y_i e pergunte se sua variabilidade é equivalente em diferentes valores de x_i. Se os resíduos não estão distribuídos ao longo da linha de regressão em torno de todo o intervalo de

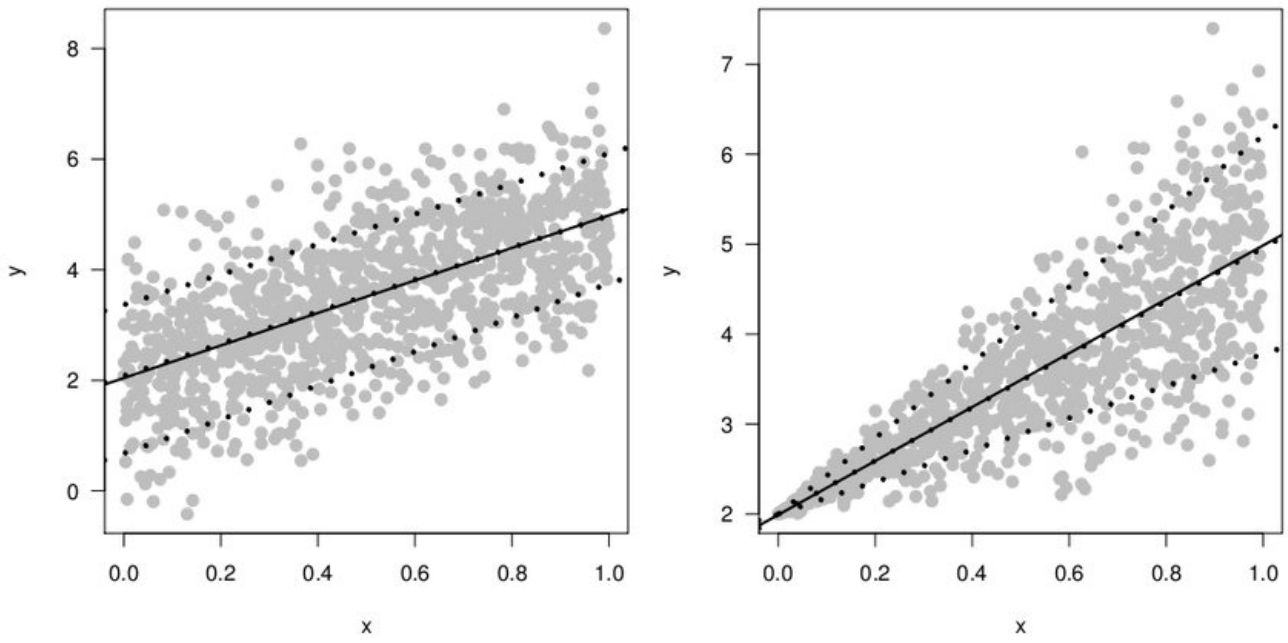
observações, o pressuposto da variância constante, ou homocedasticidade, é violado. A Figura 3 a seguir ilustra o significado da variância constante dos resíduos:

Figura 3 – Variância Constante dos Resíduos



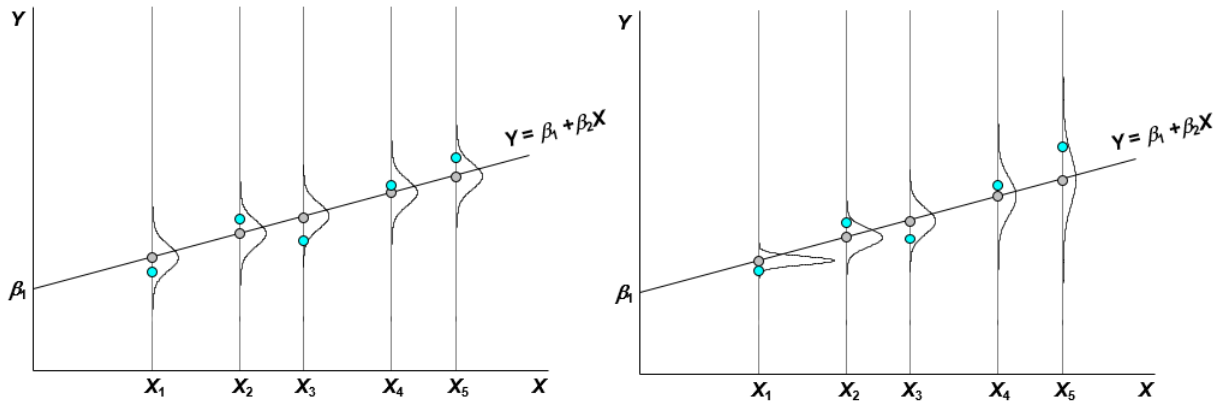
A ocorrência de variâncias não constantes nos resíduos é chamada de **heterocedasticidade**. Sua ocorrência pode estar condicionada a especificações incorretas no modelo de regressão, e sua detecção é possível através do estudo residual dos erros. Na Figura 4 vemos na primeira imagem um exemplo de homocedasticidade onde as variância do modelo é constante. Entretanto na segunda imagem observa-se que a medida que os valores de x vão aumentando a variância também vai aumentando, logo há presença de heterocedasticidade.

Figura 4 – Homocedasticidade *vs* Heterocedasticidade



Outra forma mais intuitiva de entender o problema de encontra-se na Figura 5. Na primeira imagem, para cada observação de x as distribuições são iguais, já na segunda as distribuições são diferentes.

Figura 5 – Homocedasticidade *vs* Heterocedasticidade (densidade)



O teste *Durbin-Watson* pode identificar a presença ou não de heterodasticidade e sua correção esta vinculada à eliminação de algumas variáveis ou a transformação matemática do modelo, trazendo uniformidade dos erros percentuais ao longo da linha de regressão.

3.2.4 Normalidade dos Resíduos

Os erros são uma variável aleatória normalmente distribuída. Também assumimos que os erros têm uma média igual a zero na população, embora isso não seja especialmente importante. Simbolicamente, a suposição de normalidade é freqüentemente apresentada como e na Equação 3.1. A porção de variância da equação (σ^2) tem implicações para a suposição de homocedasticidade. Esta hipótese também apresenta características pouco restritivas uma vez que os resíduos são resultantes de um sem número de fatores menos importantes no que tange a influência no comportamento da variável dependente (senão deveriam ser incluídos na equação de regressão, perdendo sua característica residual). Na média, sua influência pode ser desprezada, uma vez que o erro médio apresenta um comportamento "normalizado".

Estatisticamente se possuímos um número de observações superior a 30 a previsão de dados assume a "normalidade". Isto porque a distribuição amostral dos estimadores pode ser aproximada a curva normal onde n possua amplitude suficiente, o que na maior parte ocorre quando n é igual a 30. O Teorema do Limite Central da estatística permite esta aproximação e torna possível o uso da curva normal na avaliação da dispersão dos dados, inclusive dos resíduos, da amostra em torno do parâmetro central (média). Assim ao calcularmos sua média e variância, a extensão de possíveis erros pode ser avaliada; o que introduz um intervalo de confiança de 30 observações para a variância.

Quando o pressuposto da normalidade dos resíduos é questionado, pode-se realizar dois tipos de análise:

- **Visual** : os resíduos podem ser plotados com vistas a detecção de sua distribuição próxima a normal por meio de um histograma ou de um *QQplot*, com o seu intervalo de variação (o maior menos o menor valor) pode ser medido com vistas a determinação de sua dispersão (se próxima a 6.0 é considerado dentro da distribuição normal).

- **Testes estatísticos**: Podem ser realizados os testes de *Kolmogorov-Smirnov (K-S)*, *K-S* corrigido de *Lilliefors*, *Shapiro-Wilk*, *Anderson-Darling*, *Cramer-von Mises*, teste de assimetria de *D'Agostino*, teste de curtose de *Anscombe-Glynn*, teste omnibus de *D'Agostino-Pearson* e o *Jarque-Bera*. Yazici e Yolacan (2007) compara diversos testes de normalidade.

4 Estimação dos parâmetros

O primeiro passo, na análise de regressão, é obter as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros β_0 e β_1 da regressão. Os valores dessas estimativas serão obtidos a partir de uma amostra de n pares de valores x_i, y_i (com $i = 1, 2, ..., n$), que correspondem a n pontos num gráfico.

Obtemos, então:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

onde $\hat{y}_i$, $\hat{\beta}_0$ e $\hat{\beta}_1$ são, respectivamente estimativas de $E(y_i) = \beta_0 + \beta_1 x_i$, β_0 e β_1. Para cada par de valores x_i, y_i podemos estabelecer o desvio

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$$

4.1 Estimação pelo método de mínimos quadrados ordinários

O método dos mínimos quadrados para o modelo de regressão linear simples consiste em adotar como estimativas dos parâmetros β_0 e β_1 os valores que minimizam a soma dos quadrados dos desvios. A partir da resposta real y_i e da resposta prevista $\hat{y}_i = \beta_0 + \beta_1 xi$ atinge-se o mínimo entre todas as escolhas possíveis de coeficientes de regressão β_0 e β_1, ou seja, temos a função Z

$$Z = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Então queremos minimizar essa função

$$(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} Z$$

ou

$$(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{4.1}$$

Para realizar a estimação, vamos expandir a função Z

$$Z = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(y_i - \beta_0 - \beta_1 x_i)$$

$$Z = \sum_{i=1}^{n} (y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_1 + 2\beta_0 \beta_1 x_i + \beta_0^2 + \beta_1^2 x_i^2)$$

$$Z = \sum_{i=1}^{n} y_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} y_i x_1 + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2$$

$$Z = \sum_{i=1}^{n} y_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} y_i x_i + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2$$

$$Z = \sum_{i=1}^{n} y_i^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} y_i x_i$$

A função Z terá mínimo solucionando as suas derivadas parciais em relação a β_0 e β_1 forem nulas. Assim temos que

$$\frac{\partial Z}{\partial \beta_0} = 2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i = 0 \tag{4.2}$$

$$\frac{\partial Z}{\partial \beta_1} = 2\beta_0 \sum_{i=1}^{n} x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} y_i x_i = 0 \tag{4.3}$$

Por se tratar de uma soma de quadrados de desvios, o ponto extremo será necessariamente um ponto de mínimo da função. Formalmente, pode-se verificar que as condições de segunda ordem para mínimo são satisfeitas. Simplificando Equação 4.2 e Equação 4.3, obtemos as equações normais das condições de primeira ordem: Para a Equação 4.2

$$2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} y_i = 0$$

$$2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i = 2\sum_{i=1}^{n} y_i$$

$$2\sum_{i=1}^{n} y_i = 2n\beta_0 + 2\beta_1 \sum_{i=1}^{n} x_i$$

$$\boxed{\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i}$$

Para a Equação 4.3

$$2\beta_0 \sum_{i=1}^{n} x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} y_i x_i = 0$$

$$2\beta_0 \sum_{i=1}^{n} x_i + 2\beta_1 \sum_{i=1}^{n} x_i^2 = 2\sum_{i=1}^{n} y_i x_i$$

$$\boxed{\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i}$$

Chegamos ao sistema de equações normais

$$\begin{cases} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases} . \tag{4.4}$$

Resolvendo o sistema, obtemos:

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \textbf{ ou } \beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.5}$$

Substituindo em

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Temos

$$\left(\frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \right) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\left(\frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \beta_1 (\sum_{i=1}^{n} x_i)^2}{n} \right) + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\beta_1 \sum_{i=1}^{n} x_i^2 - \frac{\beta_1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} y_i x_i - \left(\frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} \right)$$

Multiplicando ambos os lados por n

$$n\beta_1 \sum_{i=1}^{n} x_i^2 - \beta_1 \left(\sum_{i=1}^{n} x_i \right)^2 = n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

Isolando o β_1

$$\beta_1 \left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] = n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$\beta_1 = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{4.6}$$

Notar que

$$n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i = n \left(\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) \right)$$

e que

$$n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 = n \left(\sum_{i=1}^{n} (x_i - \overline{x})^2 \right)$$

Portanto

$$\beta_1 = \frac{n\left(\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})\right)}{n\left(\sum_{i=1}^{n}(x_i-\overline{x})^2\right)} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \tag{4.7}$$

Dado que encontramos o β_1, podemos encontrar o β_0 . Anteriormente encontramos que

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \tag{4.8}$$

Substituindo o β_1 da Equação 4.6 na Equação 4.5, temos

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \beta_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \left[\frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\right]\frac{\sum_{i=1}^{n} x_i}{n}$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \left[\frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sum_{i=1}^{n}\right] x_i$$

$$n\beta_0 = \frac{n\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \cancel{\sum_{i=1}^{n} y_i (\sum_{i=1}^{n} x_i)^2} - n\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i + \cancel{\sum_{i=1}^{n} y_i (\sum_{i=1}^{n} x_i)^2}}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\cancel{n}\beta_0 = \frac{\cancel{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \cancel{n}\sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\boxed{\beta_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}} \tag{4.9}$$

Pode-se notar também que da Equação 4.8, $\frac{\sum_{i=1}^{n} y_i}{n}$ é média de y $(\overline{y})$ e $\frac{\sum_{i=1}^{n} x_i}{n}$ é a média de x $(\overline{x})$, então β_0 pode escrito como

$$\boxed{\beta_0 = \overline{y} - \beta_1 \overline{x}} \tag{4.10}$$

As variâncias dos estimadores β_0 e β_1 são dadas, respectivamente, por

$$Var(\hat{\beta}_0) = \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

A covariância entre os parâmetros é dada por

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2\left[\frac{\overline{x}}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]$$

4.2 Estimador para a variância

Após encontrar os estimadores para β_0 e $\beta - 1$ é preciso encontrar um estimador para a variância dos erros do modelo. A variância $\sigma^2 = Var(\varepsilon_i) = E(\varepsilon_i^2)$. Podemos usar a variância amostral dos resíduos como um estimador para a variância populacional dos erros.
Ao fazer isso, incorporaremos uma correção de graus de liberdade (número de observações menos o número de parâmetros). Temos, assim, o seguinte estimador da variância dos erros:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.11}$$

Como o modelo de regressão linear simples tem dois parâmetros, então

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.12}$$

Uma relação bastante importante, para futuramente realizar inferência sobre a variância estimada, é

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{4.13}$$

Podemos ver que é a relação entre a variância estimada e uma variância dada, multiplicada pelos graus de liberdade. Vamos citá-la, mas iremos usá-la mais a frente.

4.3 Propriedades dos estimadores

Agora vamos ver algumas propriedades estatísticas do modelo de mínimos quadrados ordinários considerando a suposição que $E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma^2$ e os erros $\varepsilon_i's$ para $i = 1, 2, \ldots$ são independentes. Abaixo seguem os principais teoremas

Teorema 1 *O estimador de mínimos quadrados $\hat{\beta}_0$ é não viesado para β_0.*

Teorema 2 *O estimador de mínimos quadrados $\hat{\beta}_1$ é não viesado para β_1.*

Teorema 3 $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{nSQ_{Total}}$

Teorema 4 *Os estimadores de mínimos quadrados $\hat{\beta}_1$ e $\overline{y}$ não estão correlacionados. Sob a suposição de normalidade de y_i para $i = 1, 2, \ldots, n$, $\hat{\beta}_1$ e $\overline{y}$ são normalmente distribuídos e independentes.*

Teorema 5 $Var(\hat{\beta}_0) = \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{nSQ_{Total}}\right)$

4.4 Exemplo de estimação dos parâmetros

Utilizando o exemplo da Tabela 1 do Capítulo 2 podemos calcular os componentes da fórmula Equação 4.6 e Equação 4.9, que seguem abaixo já calculados.

$$\sum_{i=1}^{n} x_i = 2572 \qquad \sum_{i=1}^{n} y_i = 2725$$

$$\sum_{i=1}^{n} x_i^2 = 271706 \qquad \sum_{i=1}^{n} y_i x_i = 288068$$

$$n = 25$$

Substituindo os valores na Equação 4.6

$$\beta_0 = \frac{(271706 \cdot 2725) - (2572 \cdot 288068)}{25 \cdot 271706 - (2572)^2} = \frac{740398850 - 740910896}{6792650 - 6615184} = -\frac{512046}{177466}$$

$$\beta_0 = -2.88531 \text{ ou } \boxed{\beta_0 \approx -2.89}$$

Temos que o a (ou β_0) é o mesmo encontrado no exemplo do Capítulo 2. Agora para o valor de b, conforme a Equação 4.9.

$$\beta_1 = \frac{25 \cdot 288068 - (2572 \cdot 2725)}{25 \cdot 271706 - (2572)^2} = \frac{7201700 - 7008700}{6792650 - 6615184} = \frac{193000}{177466}$$

$$\beta_1 = 1.0875 \text{ ou } \boxed{\beta_1 \approx 1.09}$$

Estimando os parâmetros manualmente chega-se no mesmo resultado do exemplo, Equação 2.3 $y = -2.89 + 1.09x$. No Capítulo 7 iremos realizar esse procedimento no *software R*.

5 Avaliação do modelo

No processo de seleção de covariáveis, diferentes critérios podem ser usados para comparar os modelos produzidos. Alguns deles são descritos na sequência.

5.1 Coeficiente de Determinação - R^2

Queremos saber o quão bem ajustado o nosso modelo de regressão linear simples está ou o quão bem a linha reta de regressão está ajustada aos dados. Não existe um ajuste perfeito dos dados, raramente isso acontece; mas queremos que pelo menos os resíduos em torno da média sejam os menores possíveis. Para sabe a "qualidade"desse ajuste usamos o coeficiente de determinação R^2 que é uma medida que resume o quão ajustados os dados estão em relação a linha de regressão.

Para calcular o R^2 para o modelo de regressão linear simples partimos de que

$$y_i = \hat{y}_i + \hat{\varepsilon}_i \tag{5.1}$$

Subtraindo ambos os lados por $\overline{y}$ temos

$$y_i - \overline{y} = \hat{y}_i - \overline{y} + \hat{\varepsilon}_i \tag{5.2}$$

Definindo $(y_i - \overline{y}) = \mathrm{y}$ e $(\hat{y}_i - \overline{y}) = \hat{\mathrm{y}}$, então

$$\mathrm{y} = \hat{\mathrm{y}} + \hat{\varepsilon}_i \tag{5.3}$$

Elevando ambos os lados da Equação 5.3 e somando na amostra temos

$$\sum_{i=1}^{n} \mathrm{y}^2 = \sum_{i=1}^{n} \hat{\mathrm{y}}^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n} \hat{\mathrm{y}}\hat{\varepsilon}_i \tag{5.4}$$

Como $\sum_{i=1}^{n} \hat{\mathrm{y}}\hat{\varepsilon}_i = 0$

$$\sum_{i=1}^{n} \mathrm{y}^2 = \sum_{i=1}^{n} \hat{\mathrm{y}}^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{5.5}$$

Então,

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{5.6}$$

O componente $\sum_{i=1}^{n}(y_i - \overline{y})^2$, também representado por SQ_{Total} (soma do quadrados totais) é a variabilidade total dos dados (corrigida pela média). O componente $\sum_{i=1}^{n}(\hat{y}_i -$

$\overline{y})^2$ ou SQ_{Reg} (soma dos quadrados da regressão) é variabilidade dos dados explicada pela regressão; e o componente $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ ou SQ_{Res} é a soma do quadrado dos resíduos, variabilidade que o modelo de regressão não consegue explicar. Então a Equação 5.6 pode ser escrita como

$$SQ_{Total} = SQ_{Reg} + SQ_{Res} \tag{5.7}$$

Como dito anteriormente o coeficiente de determinação R^2 corresponde à proporção da variação dos dados explicada pela regressão:

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}} = 1 - \frac{SQ_{Res}}{SQ_{Total}}$$

Características do coeficiente de determinação R^2

- Notar que $SQ_{Total} > 0$ e que o $SQ_{Reg} \geq 0$. Isso implica que o $R^2 \geq 0$. Note também que $SQ_{Reg} \geq SQ_{Total}$, implicando que $R^2 \geq 1$. Ou seja, $\boxed{0 \geq R^2 \geq 1}$.
- O R^2 é o coeficiente de correlação linear entre y_i e $\hat{y}_i$.
- Quanto maior o R^2, maior o poder explicativo do modelo.
- é uma proporção. Ele mede a proporção da variação da resposta em torno de sua média amostral que pode ser explicada usando o modelo de regressão ao invés do modelo simples $y_i = \beta_0 + \varepsilon$.
- O que garante que o R^2 varie entre 0 e 1 é a presença do intercepto no modelo, caso o modelo não possua intercepto deve usar o R^2 não centrado ($\tilde{R^2}$).
- R^2 é não decrescente, ou seja, a inserção de variáveis no modelo nunca o fará decresce (limitação).
- O coeficiente de determinação não é apropriado para comparar modelos com diferentes números de parâmetros, uma vez que R^2 sempre aumenta com a inclusão de novas covariáveis.

5.2 Coeficiente de Determinação Ajustado - R^2_{Aj}

Como dito anteriomente, o coeficiente de determinação possui a limitação de ser não decrescente. Entretanto o **coeficiente de determinação ajustado**,R^2_{ajustado}, não sofre dessa limitação. O R^2_{ajustado} é definido por:

$$1 - \frac{SQ_{Reg}/(n-p)}{SQ_{Total}/(n-1)} \tag{5.8}$$

Estabelecendo uma relação entre R^2_{ajustado} e o R^2

$$R^2_{Aj} = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SQ_{Reg}}{SQ_{Total}}\right)$$

$$R^2_{Aj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right)$$

em que n e p são o número de observações e o número de parâmetros do modelo.

- Diferentemente do que ocorre para R^2, o valor de R^2_{Aj} pode não aumentar mediante inclusão de novas variáveis ao modelo. Deve-se optar por modelos com maiores valores de R^2_{Aj}.
- $R^2_{\text{ajustado}} \leq R^2$.
- O R^2_{ajustado} pode ser menor que zero (negativo).

5.3 Quadrado Médio de Resíduos

Definido por:

$$QM_{Res} = \frac{SQ_{Res}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-p}$$

também pode ser usado para comparação e seleção de modelos de regressão

- Deve-se optar por modelos com menores valores para QM_{Res};
- Pode-se mostrar que minimizar QM_{Res} é equivalente a maximizar R^2_{Aj}, de forma que os dois critérios conduzem à seleção do mesmo conjunto de covariáveis.

5.4 C_p de *Mallows*

O coeficiente C_p de *Mallows* (MALLOWS, 2000) é definido por:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E[\widehat{y}_i - E(y_i)]^2$$

podendo ser estimado por:

$$C_p = \frac{SQ_{Res}}{\sigma^2} + 2p - n$$

em que σ^2 é dado pelo quadrado médio de resídus do modelo que inclui todas as covariáveis.

- Para o modelo completo, com p parâmetros, $C_p = p$.

- Para submodelos, definidos por subconjunto das covariáveis, menores valores para C_p são preferíveis.
- Uma estratégia para selecionar modelos com base nos valores de C_p é plotar C_p versus p e adicionar ao gráfico a reta $C_p = p$.
- Modelos com viés reduzido terão pontos próximos a reta.
- Dentre os modelos com pontos próximos à reta, deve-se optar por aquele com menor C_p (e p, consequentemente).

5.5 Estatística *PRESS*

A Estatística *PRESS* (ALLEN, 1971) permite avaliar a qualidade preditiva dos modelos de regressão, sendo definida por:

$$PRESS = \sum_{i=1}^{n} [y_i - \widehat{y}_i]^2 = \sum_{i=1}^{n} \left(\frac{r_i}{1 - h_{ii}} \right)^2$$

em que $\widehat{y}_i$ é obtido com base no modelo ajustado apenas com as demais $n - 1$ observações $(i = 1, 2, ..., n)$.

- Menores valores da estatística PRESS indicam modelos com maior poder preditivo.

5.6 Critérios de informação

Critério de informação são métricas que mensuram a qualidade de um modelo estatístico visando também a sua simplicidade. Fornece, portanto, uma métrica para comparação e seleção de modelos, em que menores valores do critério escolhido representa uma maior qualidade e simplicidade.

5.6.1 Critério de Informação de Akaike - AIC

O critério de informação de Akaike ou *Akaike Information Criterion* (AKAIKE, 1974), ou simplesmente AIC, é definido por:

$$AIC = -2log(\widehat{\theta}) + 2p$$

em que $log(\widehat{\theta}$ é a log-verossimilhança maximizada do modelo (calculada com base nos emv's dos parâmetros) e p o número de parâmetros.

- O AIC pode ser usado para qualquer modelo ajustado por máxima verossimilhança. No caso de um modelo de regressão linear temos:

$$AIC = -nlog(SQ_{Res}/n) + 2p$$

- O componente $2p$, na expressão do AIC, atua como termo de penalização atribuído à complexidade (número de parâmetros) do modelo.

5.6.2 Critério de Informação Bayesiano - BIC

Um critério alternativo ao AIC é o Critério de Informação Bayesiano ou BIC (SCHWARZ, 1978), definido, para um modelo de regressão linear, por:

$$BIC = -nlog(SQ_{Res}/n) + log(n)p$$

- O BIC penaliza mais fortemente a complexidade do modelo que o AIC ao substituir p por $log(n)$ como fator de penalização.
- Devemos selecionar modelos com menores valores de AIC (ou BIC).

Esse dois critérios são os mais conhecidos, mas existem outros como: *Deviance Information Criterion* (SPIEGELHALTER et al., 2002), *Focused Information Criterion* (CLAESKENS; HJORT, 2003), *Watanabe–Akaike information criterion* (WATANABE, 2013) e o *Hannan–Quinn information criterion* (HANNAN; QUINN, 1979).

6 Inferência dos parâmetros

Nesse capítulo vamos realizar a inferência dos parâmetros. Como saber a distribuição dos parâmetros, como criar intervalos de confiança para os parâmetros e quais os principais testes de hipóteses.

6.1 Distribuição

Considere a Equação 4.1, onde para encontrar os parâmetros β_0 e β_1 é preciso minimizar os erros, logo os parâmetros estão em função dos erros ε_i, por isso a distribuição de probabilidade de β_0 e β_1 dependerá da hipótese adotada sobre a distribuição de probabilidade de ε_i. Em princípio o método de mínimos quadrados ordinários não faz qualquer tipo de suposição sobre a natureza de probabilidade dos erros ε_i, por esse motivo o mais comum é suposição de normalidade. Considerando a hipótese de normalidade da explicitada na Equação 3.1 (ε_i é normal e identicamente distribuído) os estimadores de mínimos quadrados ordinários seguem uma distribuição normal e são identicamentes distribuídos. Dado que $\hat{\beta}_0$ e $\hat{\beta}_1$ estão em função de ε_i, então eles seguem uma distribuição normal com

- Média de $\hat{\beta}_0 = \beta_0$ e Variância de $\hat{\beta}_0 = \sigma^2_{\hat{\beta}_0}$ $\therefore$ $\boxed{\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})}$
- Média de $\hat{\beta}_1 = \beta_1$ e Variância de $\hat{\beta}_1 = \sigma^2_{\hat{\beta}_1}$ $\therefore$ $\boxed{\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})}$

Pelas propriedades da distribuição normal, a variável W que é definida como $W = \dfrac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}}$, segue uam distribuição normal padrão com média zero e variância igual a um ($= 1$), ou $W \sim N(0, 1)$. O mesmo raciocínio se aplica a $\hat{\beta}_1$, com $W = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ e $W \sim N(0, 1)$. Sendo $\sigma_{\hat{\beta}_0}$ e $\sigma_{\hat{\beta}_1}$ os erros padrões de, respectivamente, $\hat{\beta}_0$ e $\hat{\beta}_1$.

Em resumo : Assumindo normalidade dos resíduos, então a distribuição dos parâmetros estimados do modelo de regressão linear simples será um distribuição normal.

6.2 Intervalo de confiança

6.2.1 Para os estimadores

Como visto anteriormente W é uma variável normal padronizada. Podemos empregar a distribuição normal para afirmações probabilísticas sobre β_1 contando que a verdadeira

variância da população, σ^2, seja conhecida. Se σ^2 for conhecida, uma propriedade importante de uma variável normalmente distribuída com média μ e variância σ^2 é que a área sob a curva normal entre $\mu \pm \sigma$ corresponde a cerca de 68%, aquela entre os limites $\mu \pm 2\sigma$ é de cerca de 95% e a que está entre $\mu \pm 3\sigma$ é de cerca de 99.7%. Mas raramente é conhecida e, na prática, é determinada pelo estimador não viesado $\hat{\sigma}^2$. Se substituirmos σ por $\hat{\sigma}$ podemos escrever a equação $W = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ como

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{ep}(\hat{\beta}_1)} = \frac{\text{Estimador - Parâmetro}}{\text{Erro padrão estimado do estimador}} \tag{6.1}$$

onde $\text{ep}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$.

Entretanto a variável t não possui uma distribuição normal, mas uma distribuição t com $n-2$ graus de liberdade. Portanto, ao invés de usarmos a distribuição normal, vamos usar a distribuição t de *student* para estabelecer um intervalo de confiança para β_1 como a seguir:

$$Pr(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha \tag{6.2}$$

em que o valor t entre as duas desigualdades é o valor t dado pela equação Equação 6.1 e $t_{\alpha/2}$ é o valor da variável t obtido da distribuição t de *student* para um nível de significância $\alpha/2$ e $n-2$ graus de liberdade; muitas vezes é chamado de valor crítico de t em um nível de significância de $\alpha/2$. Substituindo Equação 6.1 na Equação 6.2, obtemos

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\text{ep}(\hat{\beta}_1)} \leq t_{\alpha/2}\right] = 1 - \alpha \tag{6.3}$$

Reorganizando a equação Equação 6.3, temos

$$Pr[\hat{\beta}_1 - t_{\alpha/2}\text{ep}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}\text{ep}(\hat{\beta}_1)] = 1 - \alpha \tag{6.4}$$

A Equação 6.4 oferece um intervalo de confiança de $100(1-\alpha)\%$ para β_1 e que pode ser reescrito de forma simplificada como

$$\hat{\beta}_1 \pm t_{\alpha/2}\text{ep}(\hat{\beta}_1) \tag{6.5}$$

Por analogia, pode-se obter o intervalo de confiança para β_0, conforme abaixo

$$\hat{\beta}_0 \pm t_{\alpha/2}\text{ep}(\hat{\beta}_0) \tag{6.6}$$

A característica importante dos intervalos de confiança dados na Equação 6.5 e na Equação 6.6: nos dois casos a *amplitide do intervalo de confiança é proporcional ao erro padrão do estimador*. Quanto maior o erro padrão, maior a amplitude do intervalo de confiança. Em outras palavras, quanto maior o erro do estimador, maior a incerteza da estimação

do verdadeiro valor do parâmetro desconhecido. O erro padrão é descrito muitas vezes como uma medida de precisão do estimador (da exatidão com que o estimador mede o verdadeiro valor da população).

6.2.2 Para a variância

A Equação 4.13, sob a hipótese de normalidade, segue uma distribuição χ^2 com $(n-2)$ graus de liberdade, onde

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{6.7}$$

portanto podemos usar a distribuição χ^2 para estabelecer um intervalo de confiança para σ^2:

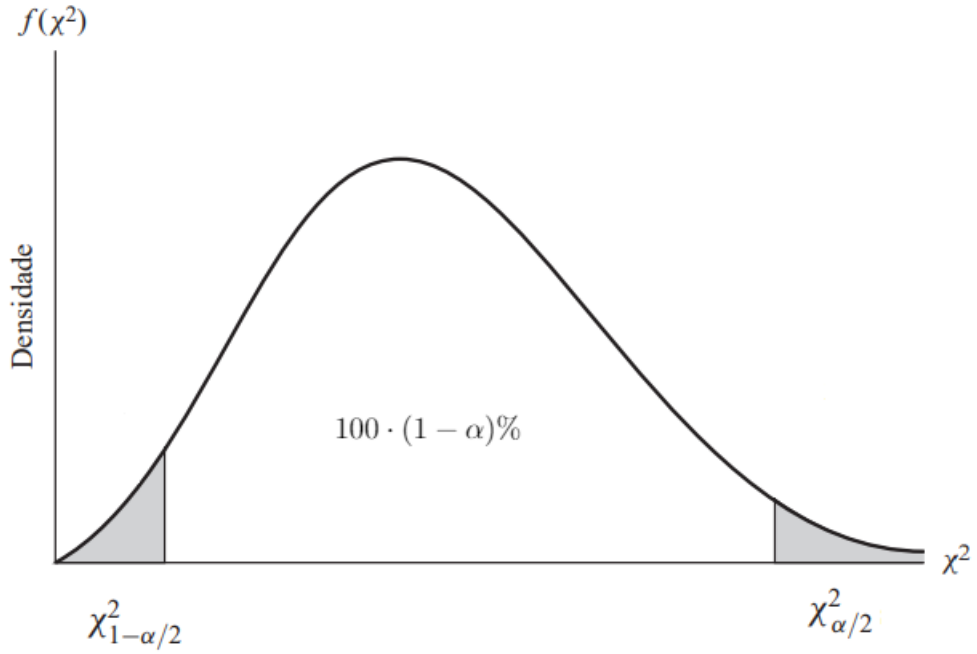
$$Pr(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}) = 1-\alpha \tag{6.8}$$

onde o valor da distribuição chi^2 na desigualdade é dado pela Equação 6.7 e os valores críticos $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$ são obtidos da tabela de qui-quadrado para $n-$ graus de liberdade. Substituindo Equação 6.7 na Equação 6.8 e reorganizando, temos

$$Pr\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] \tag{6.9}$$

o dá o intervalo de confiança $100(1-\alpha)\%$. Na Figura 6 vemos as áreas caudais da distribuição qui-quadrado.

Figura 6 – Áreas caudais da distribuição χ^2



6.3 Teste de hipóteses

Uma hipótese é uma declaração sobre um parâmetro da população. As duas hipóteses complementares em um problema envolvendo um teste de hipóteses são chamadas *hipótese nula* e *hipótese alternativa*, denotadas por H_0 e H_1, respectivamente.

Dado um parâmetro populacional θ, o formato geral da hipótese nula e da hipótese alternativa é $H_0 : \theta \in \Theta$ e $H_1\ \theta \in \Theta^C$, onde Θ é um algum subconjunto do espaço de parâmetros e Θ^C é seu complemento. Por exemplo, suponha que a hipótese nula seja que o verdadeiro valor de θ é θ_0. Assim,

$$H_0 : \theta = \theta_0$$

A hipótese alternativa, considerada aceitável caso H_0 seja rejeitada, pode pode ter formas como

$$H_0 : \theta \neq \theta_0$$

$$H_0 : \theta > \theta_0$$

$$H_0 : \theta < \theta_0$$

a depender das informações do problema.
Um procedimento para testar uma hipótese, ou um teste de hipótese, é uma regra que especifica: (a) para quais valores amostrais a decisão aceita H_0 como verdadeira; e (b) para quais valores amostrais H_0 é rejeitada e H_1 é aceita como verdadeira. O subconjunto do espaço amostral para o qual H_0 será rejeitada é chamado de *região de rejeição*, ou *região crítica*. O complemento da *região de rejeição* é chamado de *região de aceitação*.

Se a hipótese alternativa

- testar desigualdade, o teste será bicaudal (a distribuição terá dois valores críticos);
- testar se o parâmetro é maior que um valor, o teste será unicaudal na direita (um valor crítico à direita);
- estar se o parâmetro é menor que um valor, o teste será unicaudal na esquerda (um valor crítico à esquerda).

6.3.1 *p*-valor

Depois que um teste de hipóteses é realizado, as conclusões devem ser relatadas de algum modo estatisticamente significativo. Um método para relatar os resultados de um teste é expor o nível de significância α utilizado e a decisão de rejeitar ou aceitar H_0. Se α

for pequeno, a decisão de rejeitar H_0 é bastante convincente, mas se α for grande, a decisão de rejeitar H_0 não é muito convincente porque o teste tem uma grande probabilidade de levar, incorretamente, a esta decisão.

Outro meio de relatar os resultados de um teste é expor o chamado p-valor do teste. O p-valor $p(X)$ é uma estatística que satisfaz $0 < p(x) < 1$ para cada ponto amostral x, e corresponde à probabilidade de ocorrer valores da estatística de teste $W(X)$ mais extremos do que o observado para x, sob a hipótese de H_0 ser verdadeira. Ou seja,

$$p(x) = P(W(X) \leq W(x)|\theta \in \Theta)$$

Rejeitaremos H_0 para aqueles níveis de significância α maiores do que o p-valor encontrado.

6.4 Testes estatísticos para o modelo de regressão linear simples

6.4.1 Teste para significância dos parâmetros

Esse teste serve para saber se os parâmetros do modelo de regressão linear simples são estatisticamente diferentes de zero. Como observado na Equação 6.1

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{ep}(\hat{\beta}_1)}$$

t segue uma distribuição t de *student* com $n-2$ graus de liberdade. As nossas hipóteses (nulas e alternativa) para β_1 (também vale para β_0) é que ele é estatisticamente igual a zero.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Observação: nesse caso o nosso teste será bicaudal, pois H_1 não é uma igualdade.

Sob H_0 ($\beta_1 = 0$) temos que

$$t = \frac{\hat{\beta}_1 - 0}{\text{ep}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{ep}(\hat{\beta}_1)}$$

Regra de rejeição: Rejeita-se H_0, para um nível de significância α, se $|t| > t_{\alpha/2,(n-2)}$. Caso queira verificar a tabela t de *student* basta clicar aqui.

6.4.2 Teste de significância da variância

Testar a significância estatística do valor estimado da variância dos erros $\hat{\sigma}^2$. Queremos testar a hipótese nula $H_0 : \hat{\sigma}^2 = \sigma^2$ contra a hipótese alternativa $H_1 : \hat{\sigma}^2 \neq \sigma^2$. Conforme a Equação 6.7

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \tag{6.10}$$

Se, para um determinado nível de significância α, $\chi^2 > \chi^2_{\alpha/2;(n-2)}$ ou $\chi^2 < \chi^2_{(1-\alpha)/2;(n-2)}$, então rejeitamos a hipótese nula e aceitamos a hipótese alternativa.

6.4.3 Teste para significância SQ_{Reg} - Análise de Variância (*ANOVA*)

Conforme a Equação 5.8 no Capítulo 5, a soma dos quadrados totais SQ_{Totais} é decomposta em outros dois componentes: soma dos quadrados explicados ou da regressão SQ_{Reg} e a soma dos quadrados dos resíduos SQ_{Res}. Um estudo desses elementos da SQ_{Totais} é conhecido como **análise de variância** ($ANOVA$) do ponto de vista da regressão.
Associados a esses componentes estão seus graus de liberdade, número de observações menos o número de parâmetros. A SQ_{Totais} tem $n-1$ graus de liberdade, porque perde-se 1 grau de liberdade ao calcular $\overline{y}$. A SQ_{Res} tem $n-2$ graus de liberdade (para o modelo de regressão simples, que possui dois parâmetros - β_0 e β_1). A SQ_{Reg} tem 1 grau de liberdade.

Tabela 4 – Componentes e seus graus de liberdade

SQ_{Totais}	=	SQ_{Reg}	+	SQ_{Res}
$(n-1)$	=	1	+	$(n-2)$

Nota: $SQ_{Reg} = \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2$. Logo se β_1 for estatisticamente não significativo, então SQ_{Reg} será estatisticamente não significativo. Isso implica que x não tem nenhum influência linear sobre y.

Para testar a significância de SQ_{Reg} usamos

$$F = \frac{SQ_{Reg}/(p-1)}{SQ_{Res}/(n-p)} \tag{6.11}$$

Como o modelo de regressão linear simples tem dois parâmetros

$$F = \frac{SQ_{Reg}/(2-1)}{SQ_{Res}/(n-2)} = \frac{SQ_{Reg}}{SQ_{Res}/(n-2)} \tag{6.12}$$

$$F = \frac{SQ_{Reg}}{SQ_{Res}/(n-2)} \sim F_{1,n-2} \tag{6.13}$$

Regra de rejeição: Rejeita-se $H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$ ($H_1 : \beta_n \neq 0$) para um nível de significância α se $F > F_{\alpha;1,(n-2)}$.

Na Tabela 5 abaixo temos uma típica análise de variância (*ANOVA*).

Tabela 5 – ANOVA

Partição da variância	**SQ**	**gl**	**Média SQ**	**F**
SQ_{Reg}	$\sum_{i=1}^{n}(\hat{y}_i)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i)^2$	1	$\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i)^2$	$F = \frac{SQ_{Reg}}{SQ_{Res}/(n-2)}$
SQ_{Res}	$\sum_{i=1}^{n} \hat{\varepsilon}_i^2$	$(n-2)$	$\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2} = \hat{\sigma}^2$	
SQ_{Totais}	$\sum_{i=1}^{n} y_i^2$	$(n-1)$		

6.4.4 Teste para hipótese de normalidade

Como citado no Capítulo 3, existem diversos testes estatísticos para testar a hipótese de normalidade dos resíduos, um dos mais conhecidos é o Teste de Normalidade *Jarque-Bera* (JARQUE; BERA, 1987). O teste *Jarque-Bera* testa se a distribuição dos dados segue uma distribuição normal (H_0) em comparação com uma hipótese alternativa (H_1) em que os dados seguem alguma outra distribuição. A estatística do teste é baseada em dois momentos dos dados, a assimetria e a curtose, e possui uma $\chi^2_{2;1-\alpha}$ distribuição assintótica.

A estatística do teste *Jarque-Bera* é dada pela equação abaixo:

$$S_{JB} = T\left[\frac{\alpha_1^2}{6} + \frac{(\alpha_2 - 3)^2}{24}\right] \tag{6.14}$$

onde α_1 é a medida padrão de assimetria de uma distribuição ou o **coeficiente de assimetria** e α_2 o coeficiente de curtose. Ambos podem ser estimados por

$$\hat{\alpha_1} = \frac{\sum_{i=1}^{n} \varepsilon_i^3}{n(\hat{\sigma}^2)^{3/2}} = \frac{\hat{\mu}^3}{n(\hat{\sigma}^2)^{3/2}} \qquad \hat{\alpha_2} = \frac{\sum_{i=1}^{n} \varepsilon_i^4}{n(\hat{\sigma}^2)^2} = \frac{\hat{\mu}^3}{n(\hat{\sigma}^2)^2}$$

$\frac{\sum_{i=1}^{n} \varepsilon_i^3}{n}$ e $\frac{\sum_{i=1}^{n} \varepsilon_i^4}{n}$ são os estimadores consistentes para, respectivamente, os terceiro e quarto momentos amostrais. Assim podemos reescrever a Equação 6.14 como

$$S_{JB} = T\left[\frac{\hat{\alpha}_1^2}{6} + \frac{(\hat{\alpha}_2 - 3)^2}{24}\right] \tag{6.15}$$

Regra de rejeição : Se $S_{JB} > \chi^2_{2;1-\alpha}$, devemos rejeitar H_0. Uma vantagem desse teste é que ele pode ser desmembrado em outros dois testes: Um para assimetria e outro para curtose com distribuição $\chi^2_{1;1-\alpha}$ para cada um.

O teste *Jarque-Bera* é bastante conhecido, porém outros também podem ser aplicados para testar normalidade como o *Shapiro-Wilk* (SHAPIRO; WILK, 1965) e o *Anderson-Darling* (ANDERSON; DARLING, 1954), que possuem a mesma hipótese nula do *Jarque-Bera* e utilizados nesse trabalho

6.4.5 Teste para autocorrelação

Para detectar autocorrelação ou a hipótese de dependência entre os resíduos, o teste mais comumente usado é o *Durbin-Watson* (DURBIN; WATSON, 1992). As hipóteses do teste são $H_0 : \rho = 0$ (não há autocorrelação nos resíduos) contra a $H_1 : \rho \neq 0$ (há autocorrelação residual).

A estatística do teste é dada por

$$d = \frac{\sum_{i=2}^{n}(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=2}^{n}\hat{\varepsilon}_i^2} \tag{6.16}$$

a estatística d varia entre 0 e 4 e interpretamos o seus resultados como:

- se o valor do teste estiver próximo de 4 ($d \approx 4$), então há evidência para autocorrelação negativa;
- se estiver próximo de 2 ($d \approx 2$), evidência para ausência de autocorrelação;
- e se estiver perto de 0 ($d \approx 0$), há evidência para autocorrelação positiva.

Existem outros testes para identificação de autocorrelação como, o teste de *Wallis* (WALLIS, 1972), teste de *Breush-Godfrey* (GODFREY, 1978), *Ljung-Box* (LJUNG; BOX, 1978), etc.

6.4.6 Teste para identificar heterocedasticidade

Principais formas de identificar heterocedasticidade:

- **Forma gráfica**: gerando uma gráfico de dispersão com os resíduos elevados ao quadrado e a variável explicativa x.
- **Testes estatísticos**: Testes *Goldfeld-Quandt*, *Breusch-Pagan* ou de *White*.

Nesse trabalho usaremos os testes *Breusch-Pagan* e *Goldfeld-Quandt*, cuja a hipótese nula é que os resíduos são homocedásticos (variância constante). Se o p-valor do teste for maior que o nível de significância (por padrão é 5%), então aceitamos a hipótese nula, caso contrário os resíduos são heterocedásticos.

6.4.6.1 Teste *Breusch-Pagan*

Falando um pouco sobre o teste *Breusch-Pagan* a ideia desse teste é gerar uma regressão liner entre os resíduos quadrados (ε^2) do modelo e a variável explicativa (x), para testar se os resíduos tem relação com o regressor. Temos então

$$\hat{\varepsilon_i^2} = \delta_0 + \delta_1 x_i + u_i$$

onde u_i são os resíduos desse modelo. É verificado se δ_1 é estatisticamente significativo (se $\delta_1 \neq 0$). Se $\delta_1 \neq 0$, então a variância dos resíduos dependem da variável explicativa; mas se $\delta_1 = 0$, então teremos do modelo apenas δ_0 e u_i ($E(u_i) = 0$), então a variância é contante, pois δ_0 é constante.

7 Exemplos

Nesse capítulo serão realizados as etapas dos capítulos anteriores para gerar um modelo de regressão linear simples na linguagem *R*. Iremos seguir todas as etapas dos capítulos anteriores, desde a estimação do modelo até a realização dos testes estatísticos para avaliar o modelo.

7.1 Base de dados

A base de dados que usaremos será a *mtcars*. Os dados foram extraídos da revista *Motor Trend US* de 1974 e abrangem o consumo de combustível e 10 aspectos do *design* e desempenho de automóveis para 32 automóveis (modelos de 1973-74).

Na Tabela 6 temos o nome das variáveis na base de dados e as suas respectivas descrições.

Tabela 6 – Variáveis do *dataset mtcars*

Nome da coluna	**Descrição da variável**
mpg	Milhas/galão (EUA)
cyl	Número de cilindros
disp	Deslocamento (*cu.in.*)
hp	Potência bruta
drat	Relação do eixo traseiro
wt	Peso (1000 libras)
qsec	Tempo de 1/4 de milha
vs	Motor (0 = em forma de V, 1 = reto)
am	Transmissão (0 = automática, 1 = manual)
gear	Número de marchas para frente

Para gerar o modelo vamos usar como variável dependente ou alvo a coluna *mpg* (milha por galão), que é uma medida de consumo de combustível do veículo, e como variável explicativa usaremos a variável *hp* (*horse power* ou cavalos de força), uma medida de potência do veículo.

O objetivo é analisar a variação média do consumo do veículo a medida que a potência do mesmo varia.

7.2 Comandos *R*

Realizando a instalação dos pacotes que serão usados nesse capítulo. No pacote *lmtest* encontram-se os teste para avaliarmos o modelo. No pacote *tseries* o teste *Jarque-Bera* para analisar a hipótese de normalidade dos resíduos. E o pacote *ggplot2* para gerar gráficos, se preciso. No Código 7.1 realizamos a instalação dos pacotes.

Código 7.1 – Instalação dos pacotes

```
#instalando os pacotes
library(lmtest)
library(tseries)
library(nortest)
library(olsrr)
```

Primeiras linhas da base de dados

Código 7.2 – Modelo de regressão

```
#primeiras linhas da base de dados
head(mtcars)
```

Figura 7 – Primeiras linhas da base de dados

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Últimas linhas da base de dados

Código 7.3 – Modelo de regressão

```
#linhas finais da base de dados
tail(mtcars)
```

Figura 8 – Últimas linhas da base de dados

```
##                mpg cyl  disp  hp drat    wt qsec vs am gear carb
## Porsche 914-2 26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
## Lotus Europa  30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
## Ford Pantera L 15.8  8 351.0 264 4.22 3.170 14.5  0  1    5    4
## Ferrari Dino  19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
## Maserati Bora 15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
## Volvo 142E    21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
```

No Código 7.4 temos o sumário das colunas que selecionamos.

Código 7.4 – Sumário das variáveis

```
#linhas finais da base de dados
summary(data.frame(mtcars$mpg,mtcars$hp))
```

Figura 9 – Sumário das colunas

```
##    mtcars.mpg      mtcars.hp
##  Min.   :10.40   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.: 96.5
##  Median :19.20   Median :123.0
##  Mean   :20.09   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:180.0
##  Max.   :33.90   Max.   :335.0
```

No Código 7.5 vemos a correlação entre as variáveis *mpg* e *hp*.

Código 7.5 – Tabela de correlação

```
cor(data.frame(mtcars$mpg,mtcars$hp))
```

Figura 10 – Correlação

```
##             mtcars.mpg  mtcars.hp
## mtcars.mpg   1.0000000 -0.7761684
## mtcars.hp   -0.7761684  1.0000000
```

No Código 7.6 temos o teste para saber se a correlação entre as variáveis é estatisticamente significativa.

Código 7.6 – Sumário das variáveis

```
#linhas finais da base de dados
summary(data.frame(mtcars$mpg,mtcars$hp))
```

Figura 11 – Teste de correlação

```
##
##  Pearson's product-moment correlation
##
## data:  mtcars$mpg and mtcars$hp
## t = -6.7424, df = 30, p-value = 1.788e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8852686 -0.5860994
## sample estimates:
##        cor
## -0.7761684
```

Para estimar um modelo de regressão linear simples no R basta usar o Código 7.6.

Código 7.7 – Modelo de regressão

```
#gerando o modelo
modelo = lm(mpg~hp, data=mtcars)
modelo
```

Figura 12 – Modelo de regressão linear simples

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Coefficients:
## (Intercept)           hp
##    30.09886     -0.06823
```

Temos os coeficientes β_0 e β_1 cujo os valores são, respectivamente, 30.09886 e -0.06823. A interpretação do modelo é: **Variação de uma unidade de cavalo de potência do veículo diminui, em média,** 0.06823 **unidade o consumo de combustível do veículo, em Milhas por galão.**

Após gerar o modelo podemos gerar o sumário com as informações do mesmo. O comando *summary( )* no Código 7.3 gera a saída na Figura 13.

Código 7.8 – Sumário do modelo

```
#resumo do modelo
summary(modelo)
```

Figura 13 – Sumário do modelo

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

7.2.1 Interpretanto o sumário do modelo

A primeira informação impressa pelo resumo de regressão linear após a fórmula é a estatística de resumo residual. As estatísticas de resumo residual fornecem informações sobre a simetria da distribuição residual. A mediana deve ser próxima a zero com a média esperada dos resíduos sendo zero, sendo a distribuição simétrica ambos serão iguais. O terceiro quartil ($3Q$) e o primeiro quartil ($1Q$) devem ser próximos um do outro em magnitude, assim como o máximo e o mínimo.

Figura 14 – Estatísticas dos resíduos

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
```

Na Figura 14 mostra que a mediana está próxima de zero, mas os quartis e o máximo e o mínimo não estão próximos. Mais a frente serão realizados os teste de normalidade que verificarão a normalidade, entretanto os resíduos não aparentam seguir uma distribuição normal.

Agora temos os coeficientes do modelo que já foram interpretados anteriomente. Nessa parte do sumário podemos ver se os parâmetros são ou não estatisticamente significativos (estatisticamente diferentes de zero). Como vista na subseção 6.4.1, testa-se a hipótese deles serem iguais a zero, se o valor do p-valor for menor que o nível de significância $\alpha = 0.05(5\%)$ (padrão), então rejeitamos a hipótese nula.

Figura 15 – Coeficientes do modelo

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A Figura 15 mostra o valor dos coeficientes (*Estimated*), o erro padrão (*Std. Error*), o valor da estatística t sob a hipótese nula (t *value* - razão entre o valor do parâmetro e o erro padrão) e os p-valores ($\Pr(>|t|)$). Todos os p-valores ficaram abaixo do nível de significância de 5%), então rejeitamos H_0 e os nossos coeficientes são estatisticamente significativos e diferentes de zero.

Erro padrão residual : fornece o desvio padrão dos resíduos e nos informa sobre quão grande é o erro de predição na amostra ou nos dados de treinamento . Na Figura 16 temos o valor dessa medida : 3.863.

Figura 16 – Erro padrão residual

```
## Residual standard error: 3.863 on 30 degrees of freedom
```

Coeficiente de determinação - R^2 e R^2_{ajustado}: Como já explicado na seção 5.1 e na seção 5.2, o R^2 mede o quão bem ajustado está o nosso modelo e o R^2_{ajustado} mede apenas a variação das variáveis que de fato são relevantes para o modelo. A Figura 17 vemos que o valor do R^2 foi de 0.6024 e do R^2_{ajustado} foi um pouco menor, 0.5892.

Figura 17 – Coeficiente de determinação

```
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
```

Estatística F: A estatística F gerada pelo sumário é para testar se pelo menos um dos coeficientes é estatisticamente igual a zero. O p-valor na Figura 18 ficou abaixo de 5%, logo rejeita-se H_0 de que todos os parâmetros são conjuntamente estatisticamente não significativos. (**Observação**: como estamos usando uma regressão linear simples era de se esperar o resultado encontrado, mas essa estatística é melhor aplicada quando utiliza-se uma regressão linear múltipla).

Figura 18 – Estatística F

```
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

7.2.2 Análise da variância - *ANOVA*

Tabela *ANOVA* do modelo com a soma dos quadrados explicados da regressão (SQ_{Reg}).

Código 7.9 – ANOVA

```
anova(modelo)
```

Figura 19 – Análise de variância - ANOVA

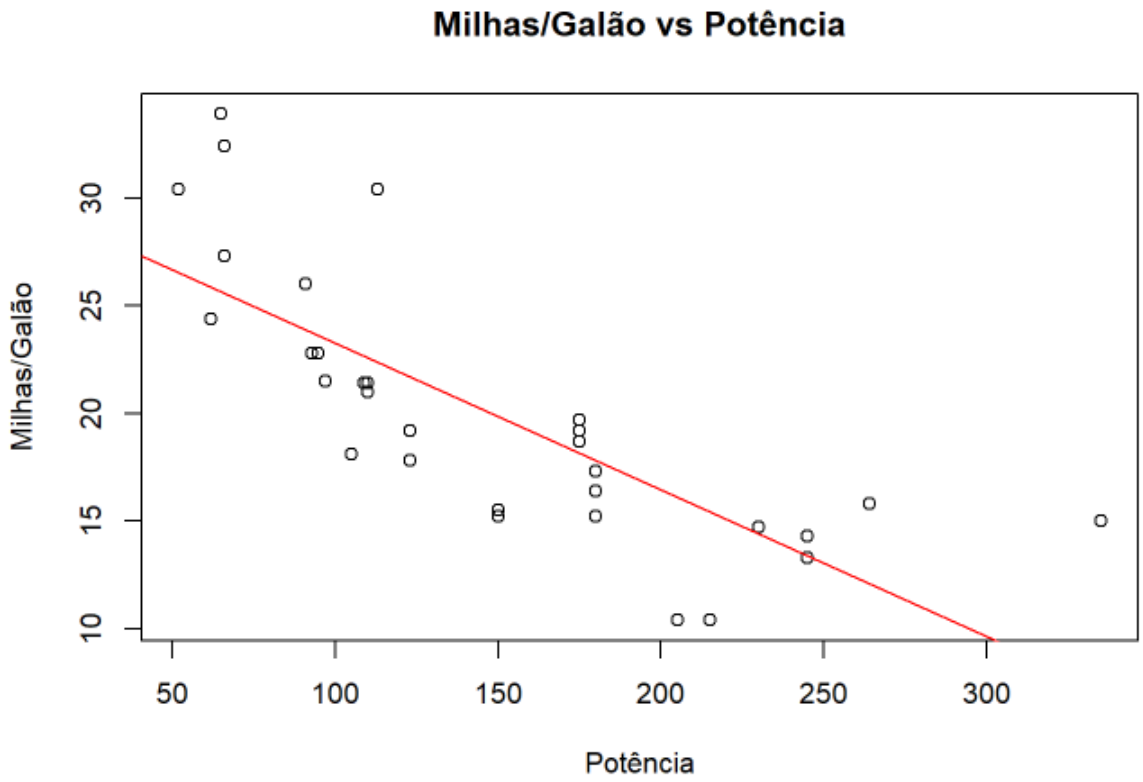
```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## hp         1 678.37  678.37   45.46 1.788e-07 ***
## Residuals 30 447.67   14.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na Figura 20 temos o gráfico de consumo *vs* potência com a reta de regressão do nosso modelo.

Código 7.10 – Consumo *vs* Potência

```
plot(mtcars$mpg~mtcars$hp, main='Milhas/Galao_vs_Potencia',
xlab = "Potencia", ylab = "Milhas/Galao")
abline(lm(mpg~hp, data=mtcars), col="red")
```

Figura 20 – Gráfico - Consumo *vs* Potência



A linha reta vermelha do gráfico é dada pela função

$$\text{Consumo (Milha/galão)} = 30.09886 - 0.06823 \text{ Potência (Cavalos de força)} \tag{7.1}$$

Informações do modelo podem ser extraídas, conforme o Código 7.11.

Código 7.11 – Informações do modelo

```
#coeficientes
modelo$coefficients
#residuos
modelo$residuals
#valores treinados do modelo
modelo$fitted.values
#graus de liberdade dos residuos
modelo$df.residual
#formula do modelo no R
modelo$call
```

Os intervalos de confiança dos parâmetros, para diferentes valores de α (10%, 5% e 1%), podem ser obtidos conforme códigos abaixo.

Código 7.12 – IC com α a 10%

```
confint(modelo, level = 0.90)
```

Figura 21 – Intervalo de confiança - $\alpha = 10\%$

```
##                      5 %        95 %
## (Intercept) 27.32567042 32.87205066
## hp          -0.08540338 -0.05105318
```

Código 7.13 – IC com α a 5%

```
confint(modelo)
```

Figura 22 – Intervalo de confiança - $\alpha = 5\%$

```
##                    2.5 %     97.5 %
## (Intercept) 26.76194879 33.4357723
## hp          -0.08889465 -0.0475619
```

Código 7.14 – IC com α a 1%

```
confint(modelo, level = 0.99)
```

Figura 23 – Intervalo de confiança - $\alpha = 1\%$

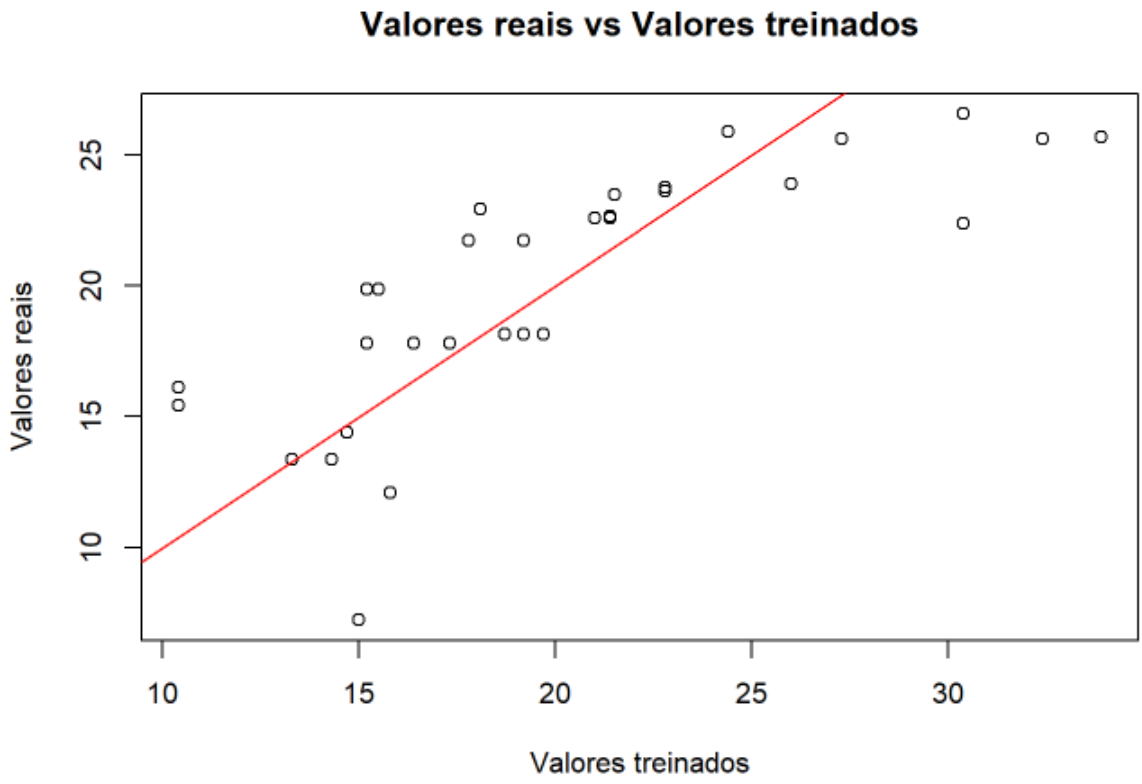
```
##                    0.5 %      99.5 %
## (Intercept) 25.60558503 34.59213605
## hp          -0.09605632 -0.04040024
```

Na Figura 24 gerada pelo Código 7.15, vemos o ajuste entre os dados reais e os valores gerados pelo modelo.

Código 7.15 – Valores reais vs valores treinados

```
plot(mtcars$mpg, modelo$fitted.values,
main="Valores␣reais␣vs␣Valores␣treinados",
     xlab = "Valores␣treinados",ylab = "Valores␣reais")
abline(lm(mtcars$mpg~modelo$fitted.values),col="red")
```

Figura 24 – Valores reais vs Valores treinados



7.2.3 Análise dos resíduos

Aqui nessa subseção iremos avalisar o modelo por meio dos seus resíduos e aplicar testes para verificar as principais hipóteses e pressupostos do modelo de regressão linear simples. Primeiramente iremos

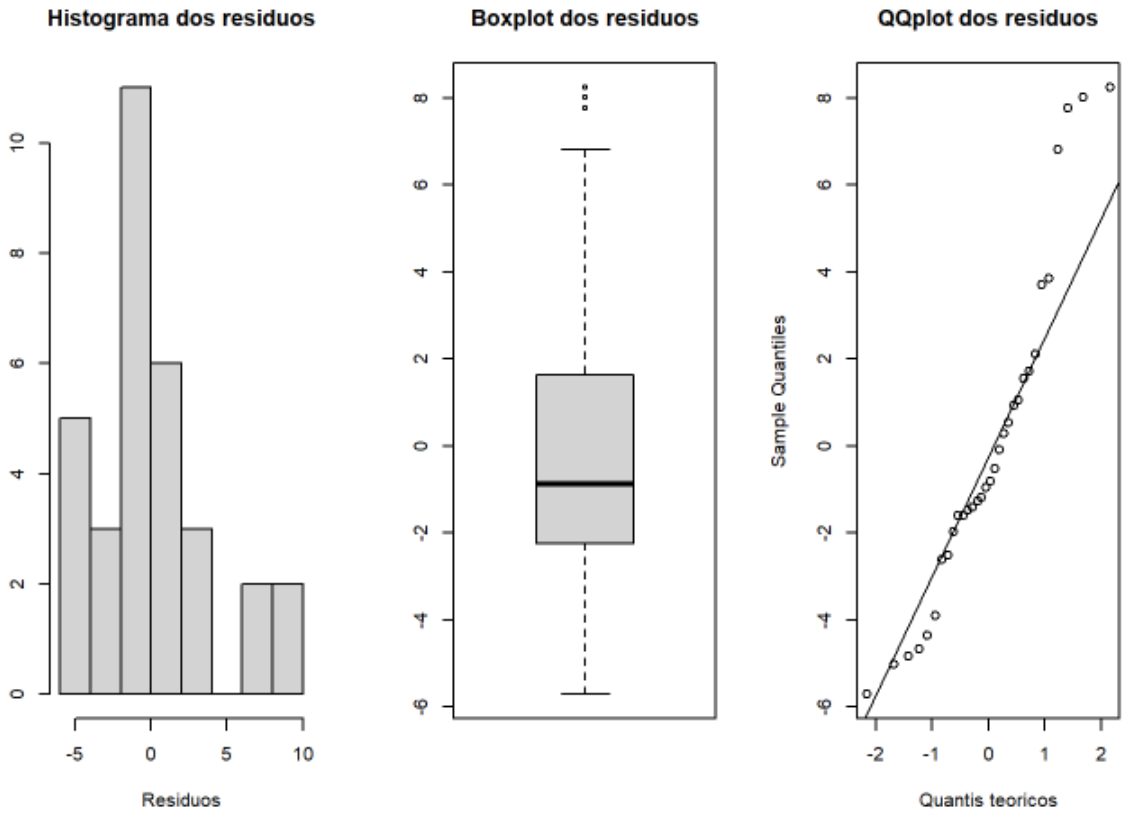
- Criar gráficos sobre os resíduos como *boxplot*, *qqplot*, histograma, etc;
- Realizar testes estatísticos para autocorrelação, heterocedasticidade e normalidade dos resíduos.

O Código 7.16 gera os três gráficos da Figura 25 onde há o histograma, o **boxplot** e o *qqplot* dos resíduos.

Código 7.16 – Histograma, *Boxplot*, *QQplot*

```
par(mfrow=c(1,3))
hist(residuals(modelo),
     main="Histograma dos residuos",
     xlab = "Residuos", ylab = "");
boxplot(residuals(modelo),
        main="Boxplot dos residuos");
qqnorm(residuals(modelo),
       main="QQplot dos residuos", xlab = "Quantis teoricos");
qqline(residuals(modelo))
```

Figura 25 – Histograma, *Boxplot*, *QQplot* dos resíduos



O histograma mostra a distribuição dos resíduos, que deveria ter um formato de sino, típico de uma distribuição normal; mas não é o que o gráfico mostra. O *boxplor* deveria ter a linha preta (que representa a mediana) centrada no valor zero, mas está deslocado um pouco abaixo e o limite superior do gráfico indica presenta de *outliers* nos resíduos. Por fim, o *qqplot* deveria mostrar todos os quantis teóricos bem próximos da linha reta do gráfico para indicar normalidade dos resíduos, mas há muitos pontos distantes da reta. Esse resultados, a princípio, indicam que os resíduos não seguem uma distribuição normal.

7.2.3.1 Testes de normalidade

Iremos testar a hipótese de normalidade com os testes mencionados na subseção 6.4.4. O primeiro, que foi explicado, é o *Jarque-Bera* e o resultado para os resíduos do nosso modelo pode ser vista a seguir.

Código 7.17 – Teste *Jarque-Bera*

```
jarque.bera.test(residuals(modelo))
```

Figura 26 – Resultado do teste *Jarque-Bera*

```
##
##  Jarque Bera Test
##
## data:  residuals(modelo)
## X-squared = 2.9836, df = 2, p-value = 0.225
```

O teste *Jarque-Bera* apresentou um p-valor maior que 5% indicando que há normalidade nos resíduos, mas esse teste funciona melhor quando o número de observações é grande, o que não é o nosso caso. Por esse motivo iremos utilizar os testes *Shapiro-Wilk* e o *Anderson-Darling*. O Código 7.18 e gera os resultado da Figura 27.

Código 7.18 – Testes *Shapiro-Wilk* e *Anderson-Darling*

```
shapiro.test(residuals(modelo))
ad.test(residuals(modelo))
```

Figura 27 – Resultados dos testes *Shapiro-Wilk* e *Anderson-Darling*

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo)
## W = 0.92337, p-value = 0.02568
```

```
##
##  Anderson-Darling normality test
##
## data:  residuals(modelo)
## A = 0.79822, p-value = 0.03447
```

Os p-valores desses dois testes ficaram abaixo do nível de significância de 5% (0.02568 e 0.03447), logo rejeitamos a hipótese nula de normalidade nos resíduos.

7.2.3.2 Testes de heterocedasticidade

Vamos testar se os resíduos possuem ou não variância normal. O Código 7.19 gera os resultados da Figura 28.

O teste *Goldfeld-Quandt* separa os resíduos em duas partes (por padrão a proporção entre essas duas parte é igual - *fraction* $= 1/2$), em seguida cria duas regressões entre os resíduos quadrados e variável explicativa e testa se a variância dessas regressões são iguais. Se forem, homocedasticidade, caso contrário, heterocedasticidade.

Código 7.19 – Testes *Goldfeld-Quandt* e *Breusch-Pagan*

```
#teste Goldfeld-Quandt
gqtest(modelo, fraction=1/2, order.by = ~hp, data=mtcars)
#teste Breusch-Pagan
bptest(modelo, varformula = ~hp, data=mtcars, studentize = F)
```

Figura 28 – Resultados dos testes *Goldfeld-Quandt* e *Breusch-Pagan*

```
##
##  Goldfeld-Quandt test
##
## data:  modelo
## GQ = 0.37566, df1 = 6, df2 = 6, p-value = 0.8707
## alternative hypothesis: variance increases from segment 1 to 2


##
##  Breusch-Pagan test
##
## data:  modelo
## BP = 0.047689, df = 1, p-value = 0.8271
```

Para ambos os p-valores ficaram bem acima de 5%, logo aceitamos a hipótese nula de homocedasticidade dos resíduos. Seguindo a ideia do teste *Breusch-Pagan* poderíamos observar a significância estatística do coeficiente da regressão dos resíduos quadrados com a variável explicativa *hp*.

Código 7.20 – Testes de heterocedasticidade

```
residuos2 <- residuals(modelo)^2 #residuos quadrados
summary(lm(residuos2~mtcars$hp)) #regressao
```

Figura 29 – Sumário

```
##
## Call:
## lm(formula = residuos2 ~ mtcars$hp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.096 -12.386  -9.838   5.700  54.766
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.32963    8.49399   1.452    0.157
## mtcars$hp    0.01132    0.05261   0.215    0.831
##
## Residual standard error: 20.08 on 30 degrees of freedom
## Multiple R-squared:  0.001541,   Adjusted R-squared:  -0.03174
## F-statistic: 0.04629 on 1 and 30 DF,  p-value: 0.8311
```

Conforme a Figura 29 o coeficiente da variável explicativa é estatisticamente não significativo, logo não há relação entre os resíduos quadrados e o a potência dos veículos (hp).

7.2.3.3 Teste para autocorrelação

Conforme a subseção 6.4.5 se a estatística do teste *Durbin-Watson* estiver próxima de 2, então há evidência para ausência de autocorrelação; caso contrário há evidência para a presença de autocorrelação.

Código 7.21 – Teste de autocorrelação

```
dwtest(modelo$model)
```

Figura 30 – Resutado do teste *Durbin-Watson*

```
##
##  Durbin-Watson test
##
## data:  modelo$model
## DW = 1.1338, p-value = 0.00411
## alternative hypothesis: true autocorrelation is greater than 0
```

Conforme a imagem Figura 30 a estatística não está próxima do valor 2, então concluímos que há presença de autocorrelação nos resíduos.

7.2.4 Avaliação do modelo

Agora vamos visualizar alguma métricas que citamos no Capítulo 5, entretanto vale lembrar que elas são usadas para realizar comparação entre modelos. Aqui só vamos deixar os códigos em *R*, visto que não há outros modelos para comparar com o nosso.

7.2.4.1 Critério de informação

Critérios de informação de *Akaike* e de *Schwarz*, respectivamente, na Figura 31.

Código 7.22 – Critério de informação

```
AIC(modelo) #Akaike
BIC(modelo) #Bayesian/Schwarz
```

Figura 31 – Critérios *Akaike* e de *Schwarz*

```
## [1] 181.2386        ## [1] 185.6358
```

7.2.4.2 *MSE* - Média do quadrados dos erros

Código 7.23 – MSE

```
mean(residuos2) #residuos2 foi calculado anteriormente
```

Temos como resultado o valor de 13.98982. Vale salientar que existem outras métricas que podem ser calculados usando o pacote *Metrics* (clique **aqui**)

7.2.4.3 Estatística *PRESS*

Código 7.24 – Estatística *PRESS*

```
PRESS(modelo)
```

Figura 32 – Resultado da statística *PRESS*

```
## $stat
## [1] 552.1057
##
## $residuals
##  [1] -1.66099755 -1.66099755 -1.00491036 -1.24411972  0.56173216 -5.05305906
##  [7]  1.01622592 -1.59722856 -0.85980580 -2.59797203 -4.04890010 -1.47510039
## [13] -0.53870740 -2.72362438 -6.04185009 -5.36952627  0.31878954  7.36326825
## [19]  4.24261493  8.92344023 -2.08100452 -4.50576319 -4.81546469 -0.09189952
## [25]  1.08080858  1.84423060  2.22689185  8.33636574  4.24723309  1.59988499
## [31] 10.69415011 -1.31592681
##
## $P.square
## [1] 0.5096958
```

8 Conclusão

Esse trabalho faz parte da atividade 5 da disciplina **Uso de Editores de Texto na Elaboração de Documentos Científicos** do professor Dr. Antonio Samuel, sobre o modelo de **regressão linear simples**. O trabalho foi criado no editor de textos científicos *Overleaf* e foram usadas a documentação da classe *abntex2* e do pacote *abntex2cite*. Todos os capítulos contemplam os tópicos exigidos na atividade.

Referências

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ALLEN, D. M. **The prediction sum of squares as a criterion for selecting predictor variables**. [S.l.]: University of Kentucky, 1971.

ANDERSON, T. W.; DARLING, D. A. A test of goodness of fit. **Journal of the American statistical association**, Taylor & Francis, v. 49, n. 268, p. 765–769, 1954.

ANDRADE, C. S. M.; TIRYAKI, G. F. **Econometria na prática**. [S.l.]: Alta Books Editora, 2019.

ANGRIST, J. D.; PISCHKE, J.-S. **Mostly harmless econometrics**. [S.l.]: Princeton university press, 2008.

CLAESKENS, G.; HJORT, N. L. The focused information criterion. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 464, p. 900–916, 2003.

DURBIN, J.; WATSON, G. S. Testing for serial correlation in least squares regression. i. In: **Breakthroughs in Statistics**. [S.l.]: Springer, 1992. p. 237–259.

GODFREY, L. G. Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 1303–1310, 1978.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica-5**. [S.l.]: Amgh Editora, 2011.

HANNAN, E. J.; QUINN, B. G. The determination of the order of an autoregression. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 41, n. 2, p. 190–195, 1979.

JARQUE, C. M.; BERA, A. K. A test for normality of observations and regression residuals. **International Statistical Review/Revue Internationale de Statistique**, JSTOR, p. 163–172, 1987.

LJUNG, G.; BOX, G. **On a measure of lack of fit in time series models. Biometrika, 65**. [S.l.], 1978.

MALLOWS, C. L. Some comments on cp. **Technometrics**, Taylor & Francis, v. 42, n. 1, p. 87–94, 2000.

SCHWARZ, G. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. **Journal of the royal statistical society: Series b (statistical methodology)**, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002.

WALLIS, K. F. Testing for fourth order autocorrelation in quarterly regression equations. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 617–636, 1972.

WATANABE, S. A widely applicable bayesian information criterion. **Journal of Machine Learning Research**, v. 14, n. Mar, p. 867–897, 2013.

YAN, X.; SU, X. **Linear regression analysis: theory and computing**. [S.l.]: World Scientific, 2009.

YAZICI, B.; YOLACAN, S. A comparison of various tests of normality. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 77, n. 2, p. 175–183, 2007.