1o. Trabalho Computacional
(TIP7077 – Inteligência Computacional Aplicada)

Programa de Pós-Graduação em Engenharia de Teleinformática (PPGETI)
Departamento de Engenharia de Teleinformática (DETI)
Universidade Federal do Ceará (UFC)
Responsável: Prof. Guilherme de Alencar Barreto

Aluno: Weslley Lioba Caldas

Matricula: 372598

1)

A) Veja que existe uma relação entre pressão e ponto de ebulição. Logo podemos utilizar deste fato para construir um modelo matemático da forma G(f)=p, onde nesse caso f será o ponto de ebulição medido em graus fahrenheit(F°) e p será a pressão medida em Polegadas de mercúrio(Hg°)

Neste caso a pressão será uma variável dependente, visto que para saber seu valor precisaremos do ponto de ebulição, que por usa vez é uma variável independente.

$$G(f) = p$$

B) Inicialmente iremos calcular a estimativa do modelo G(f)=p a partir de uma regressão linear. Feito isso podemos através de histfit, boxplot e R2, verificar se o modelo encontra-se aceitável ou não.

Obtive o seguinte valor para R2 = 0.9944

Como um valor tão alto para R2 fica aparente que o modelo de fato se encontra satisfatório. Porém ainda validaremos sua generalização para os dados através de gráficos como segue abaixo:

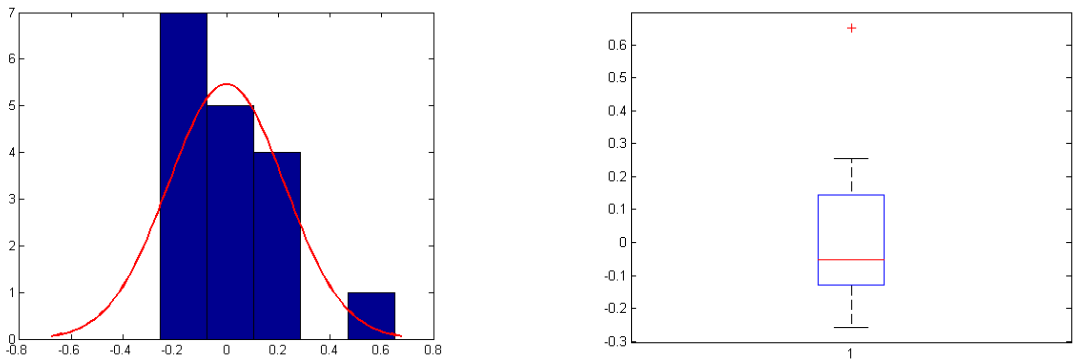

Figura 1:Hisfit(Residuos) a esquerda e Boxplot(Residuos) a direita.

Agora podemos extrair mais informações acerca do nosso modelo. Pelo teorema central do limite se obtivermos uma quantidade muito grande de dados a média da probabilidade será distribuída por uma distribuição normal. Pelo comando histfit é possível ver que nossa aproximação se assemelha a distribuição normal, mas que ainda não o suficiente para dizermos que temos essa grande quantidade de dados. Observe também que o comando boxplot acusa um outlier.

Não tenho conhecimento suficiente da base para de fato validar se este ponto indicado no gráfico é um outlier ou um ponto representativo, o gráfico via boxplot também não me garante que eu tenha dados os suficientes para dizer que tenho um conjunto 100% representativo. Vendo então que o modelo já apresenta um alto coeficiente R2, optarei por não retirar o outlier.

C) Após calcularmos os coeficientes B0 e B1, podemos calcular também a variância do ruído como se segue:

B1h = 0.5229

B0h = -81.0637

Vnoiseh = 0.0542

D) Agora iremos aplicar uma transformação nos dados de forma que o ponto de ebulição medido em graus fahrenheit(F°) passe a ser log(ponto de ebulição). Em seguida, repetiremos o processo já feito anteriormente.

B1h = 106.4413

B0h = -540.4209

Vnoiseh = 0.0630

R2 = 0.9935

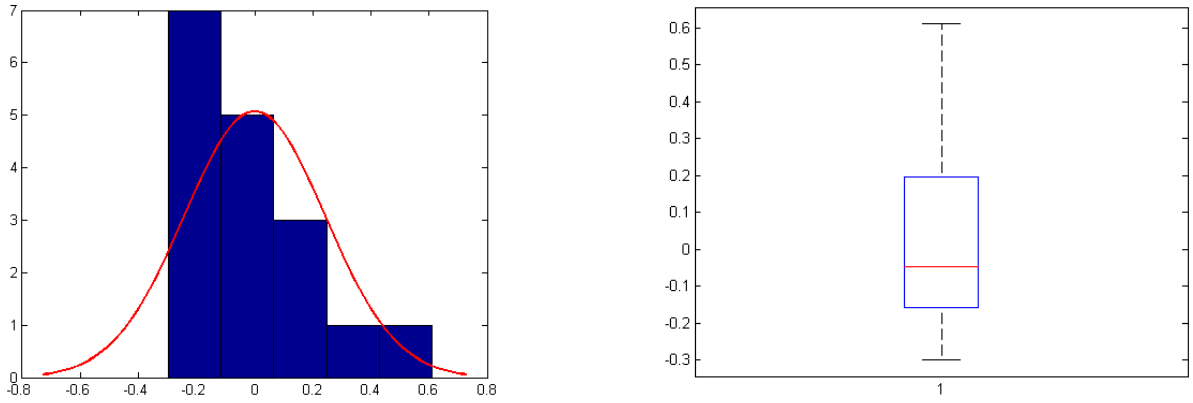

Figura 2:Hisfit(Residuos) a esquerda e Boxplot(Residuos) a direita..

Obtivemos resultados semelhantes aos vistos anteriormente, porém note que o outlier sumiu ao plotarmos o comando boxplot(residuos). O que pode indicar que este é um modelo representativo melhor.

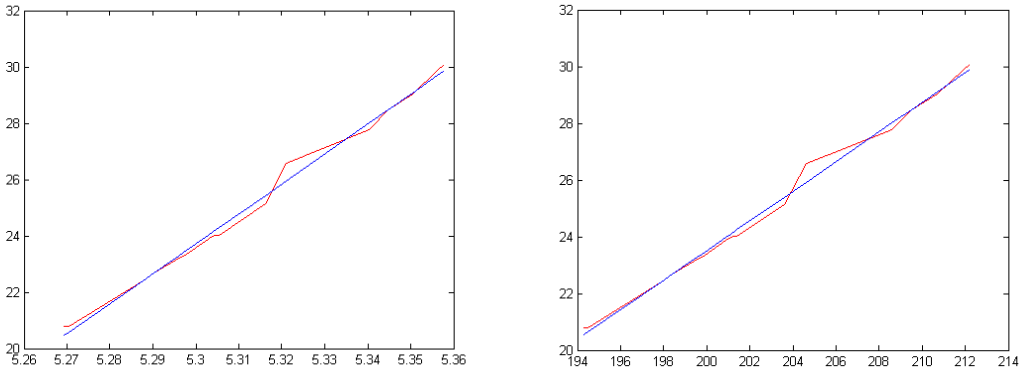

Figura 2: A esquerda temos a projeção de yh obtido da transformação do logaritmo. A esquerda projeção de yh sobre os dados convencionais.

De fato analisando os dois gráficos, não é possível ver muita diferença, mas observe que com a transformação dos dados a partir da função logarítmica, a discrepância que forçou um outlier no nosso primeiro estudo diminui, os dados que antes estavam entre 194 a 214 estão agora de 5,26 a 5,36.

2)

A) Inicialmente iremos normalizar os dados entre 0 e 1, faremos isso para verificar qual dos coeficientes apresenta maior peso no modelo. A seguir calcularemos utilizando regressão múltipla os coeficientes pedidos.

c1	0.056176	c2	0.099775
c3	0.46925	c4	0.044049
c5	0.81	c6	3.6335e-23
c7	0.15622	c8	0.10326
c9	0.9497	c10	0.43285
c11	0.28966	c12	0.024102
c13	0.0027195	(Intercept)	0.15965

B) R2: 0.749, R2 Ajustado 0.735

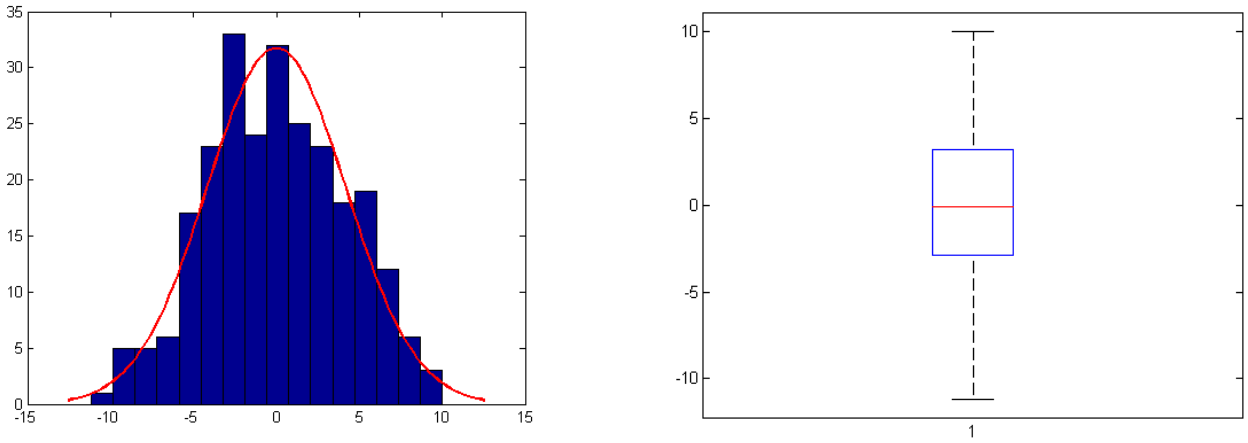

Figura 3: Histfit e boxplot respectivamente.

O gráfico histfit se assemelha a uma normal, indicando que temos uma quantidade interessante dos dados, o comando boxplot no entanto não indicou nenhum outlier. Apesar de nosso coeficiente R estar baixo, isso não indica que o modelo está necessariamente com outlier ou errado. Em alguns casos é da própria natureza dos dados serem dificilmente separados. Nestes casos testar um modelo não linear seria o mais indicado.

C) Inicialmente temos que o sexto termo apresenta um valor maior que os outros isso nos leva a acreditar que ele é o membro mais importante, em contra partida o coeficiente 13 seria o de menor relevância.

c1	0.056176	c2	0.099775
c3	0.46925	c4	0.044049
c5	0.81	**c6**	**3.6335e-23**
c7	0.15622	c8	0.10326
c9	0.9497	c10	0.43285
c11	0.28966	c12	0.024102
c13	0.0027195	(Intercept)	0.15965

Como a quantidade de dados é pequena é possível refazer o modelo excluindo uma ou mais das variáveis e comprar os resultados, abaixo foram feitos 13 modelos onde cada um exclui o coeficiente i.

TermoExcluido = 1

Rsquared =

Ordinary: 0.7452

Adjusted: 0.7324

TermoExcluido = 2

Rsquared =

Ordinary: 0.7462

Adjusted: 0.7334

TermoExcluido = 3

Rsquared =

Ordinary: 0.7485

Adjusted: 0.7359

TermoExcluido = 4

Rsquared =

Ordinary: 0.7447

Adjusted: 0.7319

TermoExcluido = 5

Rsquared =

Ordinary: 0.7490

Adjusted: 0.7364

TermoExcluido = 6

Rsquared =

Ordinary: 0.6204

Adjusted: 0.6014

TermoExcluido = 7

Rsquared =

Ordinary: 0.7469

Adjusted: 0.7342

TermoExcluido = 8

Rsquared =

Ordinary: 0.7462

Adjusted: 0.7335

TermoExcluido = 9

Rsquared =

Ordinary: 0.7490

Adjusted: 0.7364

TermoExcluido = 10

Rsquared =

Ordinary: 0.7484

Adjusted: 0.7358

TermoExcluido = 11

Rsquared =

Ordinary: 0.7479

Adjusted: 0.7352

TermoExcluido = 12

Rsquared =

Ordinary: 0.7436

Adjusted: 0.7307

TermoExcluido = 13

Rsquared =

Ordinary: 0.7394

Adjusted: 0.7263

De fato minha suspeita era verídica o coeficiente 6 é o que quando retirado apresenta maior queda no coeficiente R e o 13 o que apresenta menor impacto. Com isso podemos concluir que o componente 6 seguindo as métricas dadas é o mais importante ao modelo.