Trabalho aplicativo 1

Análise de regressão simples

Técnicas Estatísticas de Predição – INE5649

Aluno: **Luis Felipe Pelison** - Matrícula**: 14101053**

Professor**: Pedro Alberto Barbetta**

Trabalho aplicativo 1 - Realizar uma análise de regressão, usando R, com a variável dependente **Expectativa de Vida no Município** (**EXPVIDA**, em anos) e com a variável independente **Renda Per Capita do Município** (**RDCP**, em reais) da amostra dada.

Planilha utilizada: **nº 3**

Questões

1.Façam, preliminarmente, um diagrama de dispersão. Os dados sugerem relação entre as variáveis? Por quê? Independente da resposta anterior, você acredita que o ajuste de uma reta possa explicar, em parte, os diferentes valores da expectativa de vida dos municípios? Por quê?

2.Ajustem aos dados uma regressão linear simples. Apresentem a saída computacional. Escrevam a equação de regressão obtida dos dados.

3.Interpretem o coeficiente angular (inclinação) e o R^2.

4.Apresentem um intervalo de 95% de confiança para o coeficiente angular.

5.Digam se a regressão é significante baseando-se no teste F e/ou no teste t. Descrevam claramente em que vocês estão se baseando para justificar as suas afirmativas.

6.Façam uma análise dos resíduos (Diagrama de dispersão: Resíduos x Preditos; façam outros diagnósticos gráficos básicos). Discutam sobre a adequação do modelo baseado nesses gráficos.

7.Apresentem estimativa pontual (mostrando a conta) e intervalo de confiança para o valor médio de expectativa de vida para municípios que tenham renda per capita de R$800,00.

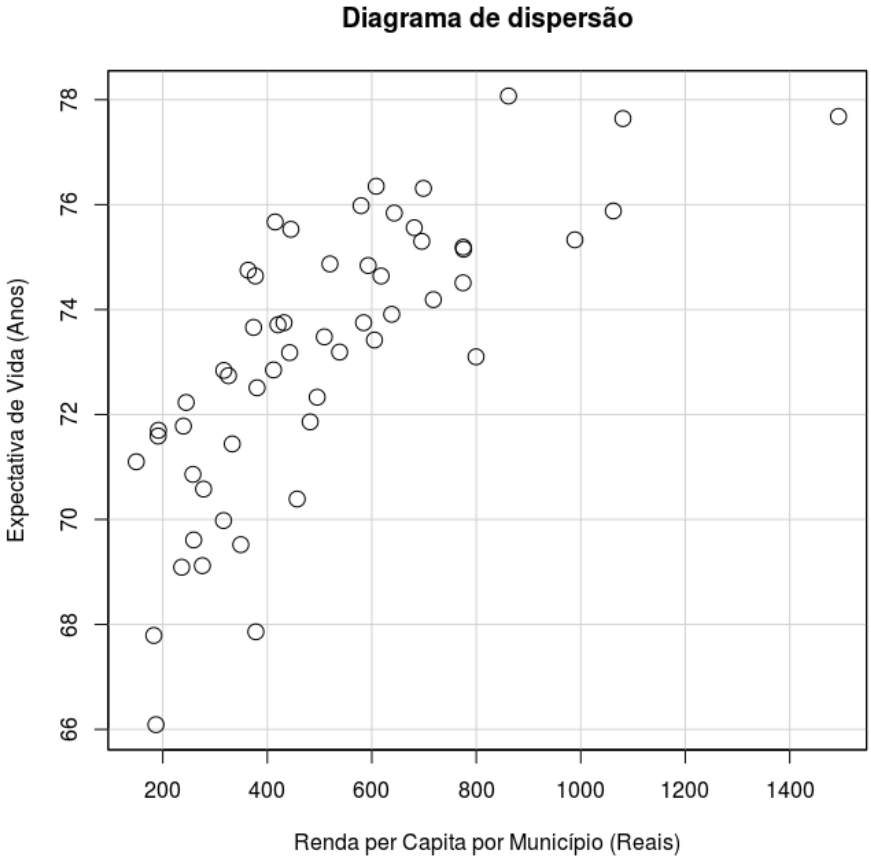
8.Apresentem o intervalo de predição para a expectativa de vida de um município que tenha renda per capita de R$800,00.

9.Façam um gráfico mostrando a reta de regressão e intervalos de confiança/predição como os descritos nos itens 8 e 9.

10. Você julga razoável usar esse modelo para predizer a expectativa de vida de um município cuja renda per capita é de R$3.000,00 reais? Por quê?

Respostas

1. Plotando o gráfico de dispersão, pelo Rcmdr, obtivemos o seguinte:



Visualizando o gráfico de dispersão, percebe-se que há sim uma relação positiva entre a renda per capita com a expectativa de vida nos municípios. Essa relação pode ser visualizada já que conforme aumenta a renda, tende-se a aumentar também expectativa de vida. Podemos visualizar isso numericamente, calculando o coeficiente de correlação. Pelo Rcmdr, obtivemos que r = 0.7413965. Esse número mostra que há uma correlação

positiva e quase forte entre as variáveis. Portanto, conseguimos realizar uma regressão entre essas variáveis, já que mostrou-se visível uma relação.

Quanto ao ajuste de uma reta em cima dos dados, acredito que sim, em parte, ela pode ajudar. Isso porque os dados tem uma correlação positiva e quase forte. Porém, mostra-se que talvez uma curva (regressão não linear) pode ser melhor, já que os dados não estão totalmente lineares.

2. Calculando a regressão linear simples pelo Rcmdr, obtivemos a seguinte saída computacional:

Output

```
> RegModel.2 <- lm(ESPVIDA~RDPC, data=pred2)

> summary(RegModel.2)

Call:
lm(formula = ESPVIDA ~ RDPC, data = pred2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7088 -1.0026  0.2846  1.1001  3.1898

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.942e+01  5.254e-01 132.126  < 2e-16 ***
RDPC        7.371e-03  9.165e-04   8.043 9.55e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.78 on 53 degrees of freedom
Multiple R-squared:  0.5497, Adjusted R-squared:  0.5412
F-statistic: 64.69 on 1 and 53 DF,  p-value: 9.551e-11
```

Com isso, temos que a equação de regressão é a seguinte:

$$\text{ExpVida} = 69.42 + 0.007371 * \text{RDPC}$$

3. Vamos interpretar os resultados do coeficiente angular (0.007371) e R^2 (0.54). Começando pelo coeficiente angular, vemos que ele é um valor pequeno e positivo. Isso nos diz que conforme a variável expectativa de vida cresce 1 unidade (1 ANO), a variável renda per capita cresce apenas 0.007371 unidades (REAIS), mas positivamente. Assim, uma eleva a outra, por uma taxa bem pequena. São necessários, aproximadamente, 1000 reais de renda para crescer 7.37 anos na expectativa de vida (aproximadamente 135 reais para 1 ano).

Agora, analisando o R^2, podemos ver que realizando a regressão linear, obtivemos um ganho de 54%, comparando com a relação sem regressão, utilizando somente a média da variável expectativa de vida. Assim, utilizar a regressão foi bom para o nosso caso, pois havia uma relação entre as variáveis de fato.

Outra forma de falar a mesma coisa é: 54% da variância da expectativa de vida pode ser explicada por uma regressão linear em termos da renda por capita desse município.

4. Para um intervalo de confiança de 95% para o coeficiente angular (b1), temos a seguinte fórmula:

$$b1 = b1 +- Sb1 * t\ 95\%$$

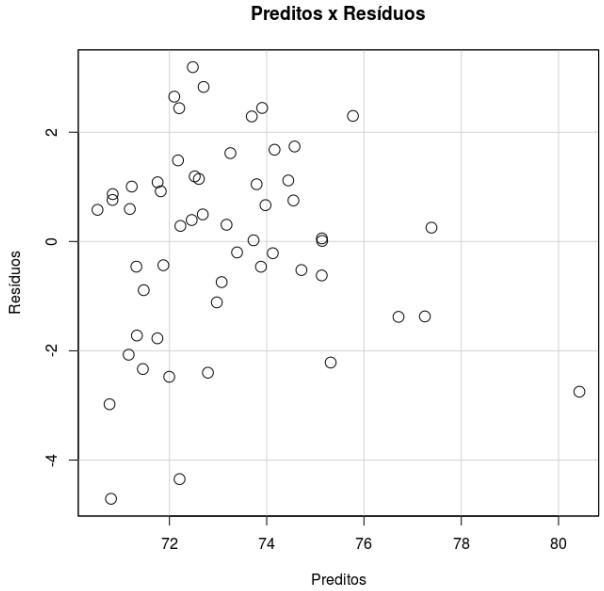
Onde $b1 = 0.007371$, $Sb1 = 0.0009165$ e $t95\%$ (para $G.L = 53$) $\simeq 2$.

Assim, b1 pode variar entre 0.005538 e 0.009204, com 95% de confiança.

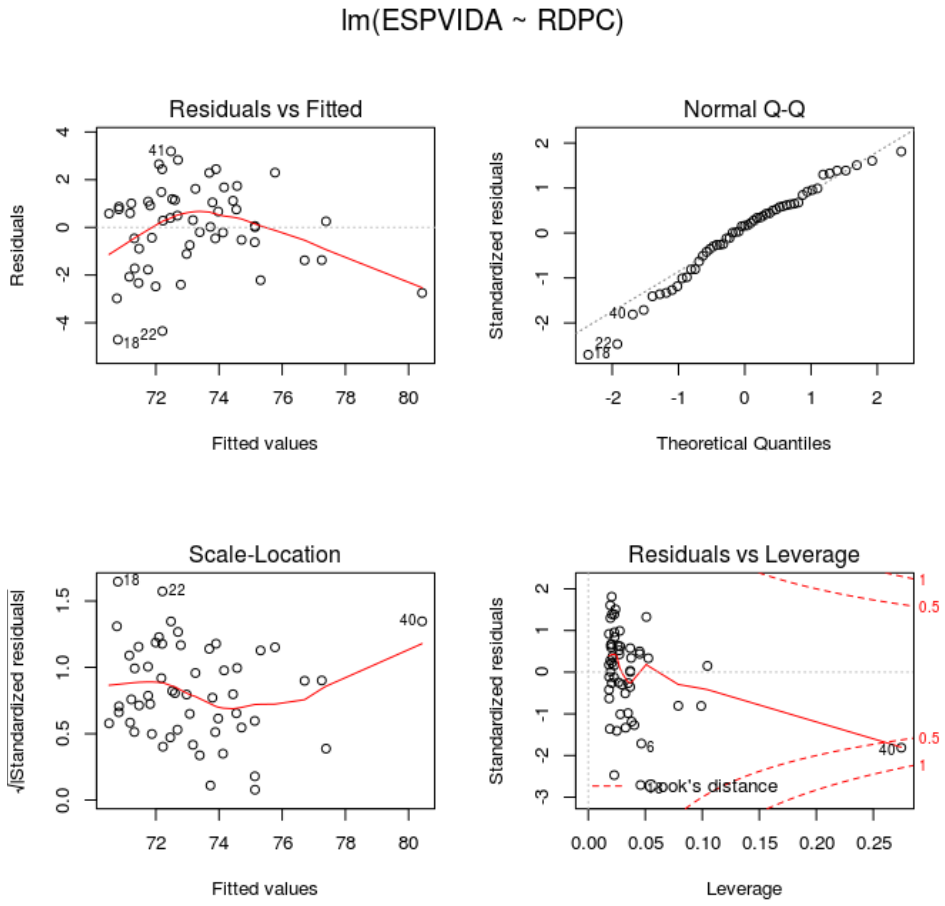
5. Baseando-se nos testes F e t, podemos ver a significância da regressão realizada. Para avaliarmos se a regressão é significante ou não, podemos definir uma hipótese inicial (nula) como sendo Ho => b1 = 0. Isso é, o coeficiente angular b1 é igual a 0, a regressão não existe. Agora, vamos tentar aceitar ou rejeitar essa hipótese com algum teste (F ou t). Vale lembrar que o valor de F é igual o valor de t ao quadrado. Assim, a hipótese será aceita se o valor de t for menor que o módulo do valor de t tabelado para 95% de confiança. Esse valor tabelado, para um grau de liberdade de 53, é aproximadamente 2. Então, calculando nosso t, que é a raíz de F (cujo está presente na figura da questão 2 → 69.64), temos que t = 8,345058418. Logo, como t é maior que o t tabelado, a hipótese é rejeitada. Por tanto, a regressão é significante!

Podemos chegar à mesma conclusão se olharmos para o p-value, cujo também está presente na figura da questão 2 e vale: $9.551e^{-}11$. Pois, se o p-value for menor ou igual a 0.05, o teste é rejeitado. E, como vemos, o p-value é MUITO inferior a 0.05 (100% - 95% = 0.05).

6. O gráfico de Resíduos x Preditos pode ser observado abaixo. Ele foi gerado pelo Rcmdr.



Outros gráficos gerados:



Desses gráficos podemos tirar algumas conclusões. Primeiro, no gráfico de resíduos x preditos, vemos que há uma formação dos dados que nos indica que o modelo linear não é muito bom. O certo seria mesmo uma abordagem Não-Linear, como previso lá no item 1. O modelo linear não é bom pois os dados não estão totalmente balanceados e em torno do 0, mas sim em uma tendência no formato de parábola.

Pelo gráfico de comparação de quantis, podemos ver que os dados não seguem a reta totalmente. No início e no fim dela, os dados “fogem” do padrão. Isso também indica a não-linearidade da relação entre expectativa de vida e renda per capita

7. Considerando 800,00 reais de renda per capita, a estimativa pontual da expectativa de vida tem o seguinte formato:

ExpVida média estimada = 69.42 + 0.007371 * 800

ExpVida média estimada = 75.3168 anos

Sendo o intervalo de confiança para 95% igual a

ExpVida média estimada +- S exp * t95%

com

$$S_{exp} = S_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

e

$$s_e = \sqrt{QME} = \sqrt{\frac{SQE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Calculando isso, temos Sexp = 0,3584 e t95% ≃ 2

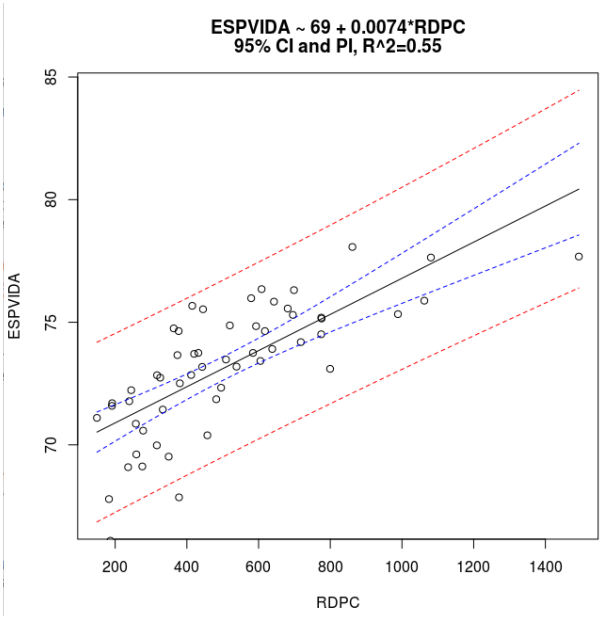
Assim, o intervalo de confiança vai de 74,6000 até 76,0336 anos

8. O intervalo de predição para a expectativa de vida de um município cuja renda per capita é 800,00 pode ser obtido pelo software Rcmdr. A resposta do software é a seguinte:

```
> predict_CI_PI(RegModel.1, data.frame('RDPC'=800), level=0.95)
       fit   lwr.CI   upr.CI   lwr.PI  upr.PI
1 75.31833 74.60001 76.03664 71.67645 78.9602
```

Isso nos mostra que o intervalo é de 71.67645 até 78.9602, onde o valor esperado é o mesmo da questão 7: 75.31. (vale notar que na questão 7 os valores foram calculados manualmente, portanto houveram aproximações e arredondamentos)

9. Podemos visualizar melhor esses intervalos, tanto de predição quanto de confiança para a regressão linear no gráfico abaixo



Onde as linhas tracejadas vermelhas representam os limites superior e inferior do intervalo de predição e as azuis o intervalo de confiança de 95%.

10. Devido a relação entre as variáveis ser não-linear, não temos uma boa estimativa para valores altos da renda. Assim, se tentarmos predizer o valor da expectativa de vida para uma renda de 3.000 reais, iremos ter um grande resíduo. Podemos ver no software que essa predição dá um valor de 91.53513 anos, o que pode ser exageradamente alto, já que não é certo que um munício de 3000 reais de renda, irá ter uma expectativa de vida tão alto. O esperado seria uma expectativa de vida mais baixa. Esse valor alto foi devido a linearidade da regressão, que não serve para 100% dos dados. É apenas uma aproximação linear.