Regressão Linear e Previsão com R

Agda Costa

13/07/2020

## Etapas

1. Importação dos dados
2. Gráfico de Dispersão Area x Preço
3. Cálculo da correlação Area x Preço
4. Ánálise Exploratória Tempo x Preço
5. Cálculo da correlação Tempo x Preço
6. Análise dos dados
7. Descobrindo os valores maior que certo quantil
8. Ajuste do Modelo de Regressão: Área x Preço
9. Preço estimado de um casa com 70m²
10. Gráfico de dispersão com o modelo de regressão linear
11. Verificando se o modelo é mesmo relevante
12. Comprovando o modelo de Regressão: Área x Preço
13. Replicando as etapas para o modelo Tempo x Preço
14. Suposições
15. Teste do Durbin Watson
16. Teste de Breusch Pagan
17. Teste de Shapiro Wilks
18. Conclusão

**1. Importação dos dados**

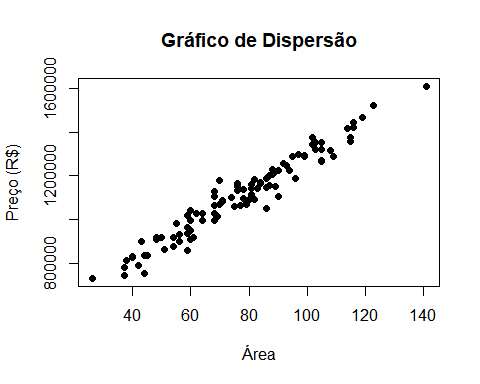
library(readxl)  
dados <- read\_excel("~/RegressaoLinear\_e\_Previsao/dados.xlsx",   
 col\_types = c("numeric", "numeric", "numeric"))  
head(dados)

## # A tibble: 6 x 3  
## preco area tempo  
## <dbl> <dbl> <dbl>  
## 1 1068539 79 437  
## 2 949609 60 596  
## 3 1190573 86 316  
## 4 1465552 119 286  
## 5 1422688 116 141  
## 6 1266837 105 213

**2. Gráfico de Dispersão Area x Preço**

R. A primeira impressão é: quanto maior a área, maior o preço.

plot(dados$area, dados$preco,  
 ylab = "Preço (R$)",   
 xlab = "Área",   
 main = "Gráfico de Dispersão",  
 pch = 19)



**3. Cálculo da correlação Area x Preço**

R. Quando o coeficiente de relação é próximo de 1, entende-se que há uma relação linear forte positiva, que é quando uma váriavel aumenta, a outra também aumenta.

cor(dados$area, dados$preco)

## [1] 0.9711363

R. Como p-valor < 0.05 rejeito a hipotese de que a correlação é zero, então a relação entre área e preço existe, é positiva e estatísticamente significativa.

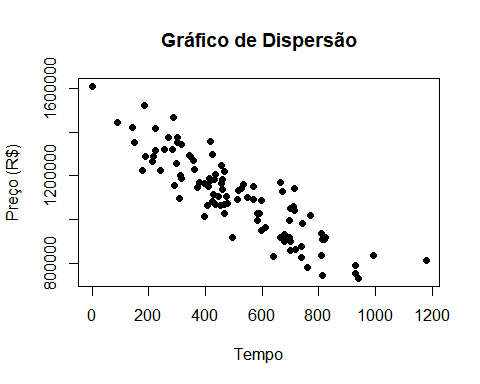
cor.test(dados$area, dados$preco)

##   
## Pearson's product-moment correlation  
##   
## data: dados$area and dados$preco  
## t = 40.305, df = 98, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9573272 0.9805212  
## sample estimates:  
## cor   
## 0.9711363

**4. Ánálise Exploratória Tempo x Preço**

R. A primeira impressão é: quanto maior o tempo, menor o preço.

plot(dados$tempo, dados$preco,  
 ylab = "Preço (R$)",   
 xlab = "Tempo",   
 main = "Gráfico de Dispersão",  
 pch = 19)



**5. Cálculo da correlação Tempo x Preço**

R. Quando coeficiente de relação é próximo de -1, entende-se há uma relação linear forte negativa, que é quando uma váriavel aumenta, a outra diminui. Pelo p-valor < 0.05 rejeitamos a hipótese de que a correlação é zero, então podemos concluir que a relação existe, é negativa e estatísticamente significativa.

cor.test(dados$tempo,dados$preco)

##   
## Pearson's product-moment correlation  
##   
## data: dados$tempo and dados$preco  
## t = -18.16, df = 98, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9163913 -0.8236355  
## sample estimates:  
## cor   
## -0.8780116

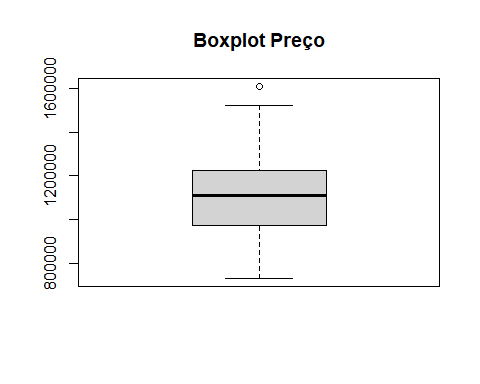
**6. Análise dos dados**

R.Utilizamos o boxplot para visualizar a distribuição dos dados.

Observando o gráfico parece que o preço tem uma distribuição simétrica em torno da mediana.E existe outliers. Analisando melhor as medidas pela função summary, diz que 50% das observações tem um preço variando 979 mil á 1.225 mi.

Usamos a função car para identificar os outliers dentro do boxplot.

boxplot(dados$preco, main = "Boxplot Preço")



summary(dados$preco)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 733343 979788 1111292 1109659 1225775 1607974

library(car)

## Warning: package 'car' was built under R version 4.0.2

## Loading required package: carData

**7.Descobrindo os valores maior que certo quantil**

R. Neste caso, vamos descobrir quais os indices das casas com os preços maiores que o terceiro quantil

which(dados$preco > quantile(dados$preco, 0.75) )

## [1] 4 5 6 10 13 20 21 25 26 27 38 40 50 58 63 66 71 72 74 75 79 84 85 86 89

**8. Ajuste do Modelo de Regressão: Área x Preço**

R. Modelo de regressão linear simples da váriavel Preço em função da area. Isto é preço = bo + b1\*area

mod = lm(dados$preco~dados$area)  
mod

##   
## Call:  
## lm(formula = dados$preco ~ dados$area)  
##   
## Coefficients:  
## (Intercept) dados$area   
## 502347 7851

**9. Preço estimado de um casa com 70m²**

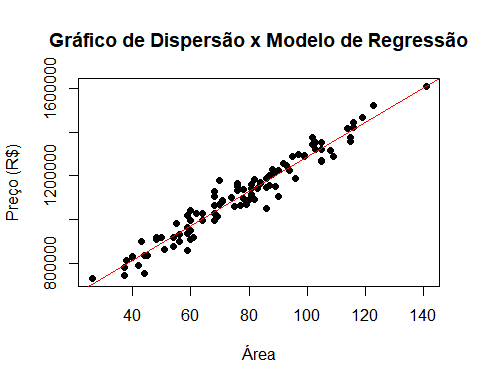
R. Usando os coeficientes gerados pelo função Lm(), temos que o preço estimado para uma casa com 70m² seja de R$1.051.951 Milhões

preço = mod$coefficients[1] + mod$coefficients[2]\*70  
preço

## (Intercept)   
## 1051951

**10. Gráfico de dispersão com o modelo de regressão linear**

plot(dados$area, dados$preco, pch = 19,  
 ylab = "Preço (R$)",  
 xlab = "Área",   
 main = "Gráfico de Dispersão x Modelo de Regressão")   
abline(mod, col = "red")



**11. Verificando se o modelo é mesmo relevante**

R. Já sabemos que a área realmente influi no preço das casas, mas será que é melhor eu considerar mesmo o estimador do preço x area ao invés de simplesmente trabalhar com a média?

Para testar, usamos a função summary no modelo criado, e pelo teste F apresentado na última linha, vemos que o modelo é bom comparado com a média e para nos dizer se os coeficientes estimados é diferente de zero, observamos as informações de ‘Coefficients’,e no caso, rejeitamos a hipótese de que são iguais a zeros, vistos que os p-valor < 0.05.

summary(mod)

##   
## Call:  
## lm(formula = dados$preco ~ dados$area)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -126070 -31605 9852 28943 129419   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 502346.8 15715.4 31.96 <2e-16 \*\*\*  
## dados$area 7851.5 194.8 40.30 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44650 on 98 degrees of freedom  
## Multiple R-squared: 0.9431, Adjusted R-squared: 0.9425   
## F-statistic: 1624 on 1 and 98 DF, p-value: < 2.2e-16

**12. Comprovando o modelo de Regressão: Área x Preço**

R. Ainda usando a função summary em cima do modelo, analisamos o R² e o R² ajustado a amostra. Então com esse modelo eu consigo explicar 94% de variabilidade dos meus dados

summary(mod)$r.squared

## [1] 0.9431058

**13. Replicando as etapas para o modelo Tempo x Preço**

R. A gente faz o teste T, para avaliar a importância da variável tempo no modelo e a hipótese que a gente tem é que ela não é importante, que o efeito dela no preço é zero, é nulo. Como a gente faz o teste T e obtém um p-valor menor do que 0,05. Nós concluímos que a variável do tempo, ela é estatisticamente significativa para estimar o preço das casas.

Além disso, da estatística F, temos um p-valor que é menor do que 0,05, concluímos que o modelo que tem o tempo das casas, é melhor do que considerar simplesmente a média do preço das casas e nesse caso, temos um coeficiente de determinação de 0,79.

Nesse caso, ele foi um pouquinho menor do que a gente tinha comparado anteriormente, com o modelo que tinha a área das casas, ou seja, com esse modelo do preço em função do tempo, eu só consigo explicar 77%, aproximadamente de variabilidade dos meus dados.

Então, esse modelo, ele não é muito bom como que a gente quer, e se eu for comparar ele, por exemplo, com o modelo que só tem a área das casas, ele é pior do que aquele outro modelo, por quê? Porque o coeficiente de determinação dele é menor do que o outro.

mod2 = lm(dados$preco~dados$tempo)  
summary(mod2)

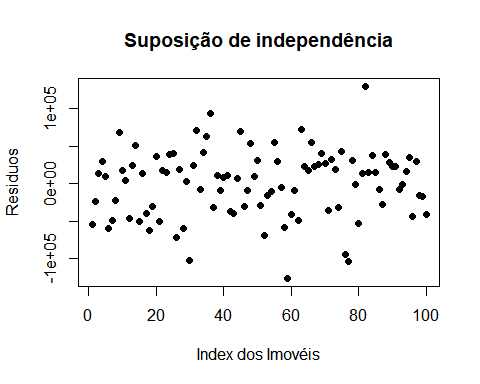
##   
## Call:  
## lm(formula = dados$preco ~ dados$tempo)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -194948 -59657 -6641 53192 201034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1482853.91 22418.73 66.14 <2e-16 \*\*\*  
## dados$tempo -737.03 40.59 -18.16 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 89590 on 98 degrees of freedom  
## Multiple R-squared: 0.7709, Adjusted R-squared: 0.7686   
## F-statistic: 329.8 on 1 and 98 DF, p-value: < 2.2e-16

**14. Suposições**

R. Toda a análise realizada até agora, pelo modelo linear simples foi baseado em algumas suposições sobre o modelo, algumas delas são independencia, homocedasticidade e tem que ter normalidade. Podemos visualizar as suposições por meio de gráficos e fazendo testes para comprovar.

Primeiro vamos plotar os residuos de todas as minhas 100 observações do meu modelo. E gráficamente percebemos que eles não seguem um padrão, ou seja, não tem nenhuma tendência, e é isso que a gente espera. Podemos também identificar, pela função identify, os Index das casas com os piores residuos, isto que, que tiveram uma diferença maior do valor estimado.

plot(mod$residuals,  
 ylab = "Residuos",  
 xlab = "Index dos Imovéis",   
 main = "Suposição de independência",  
 pch = 19)  
  
identify(mod$residuals, n=2)



## integer(0)

**15. Teste do Durbin Watson**

R. Agora nós vamos fazer o teste do Durbin Watson para o modelo do preço, em função da casa. Ele faz uma estatística que teste precisamente, se eles são independentes. Aí, o p-valor que a gente compara é com 0.05, nesse caso, o p-valor não foi menor do que 0.05. Logo, a gente não tem evidência suficiente para rejeitar a hipótese nula.

Se eu não consigo rejeitar é porque eu não tenho evidências para dizer que elas não são independentes. Então, eu aceito que são independentes.

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.2

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

dwtest(mod)

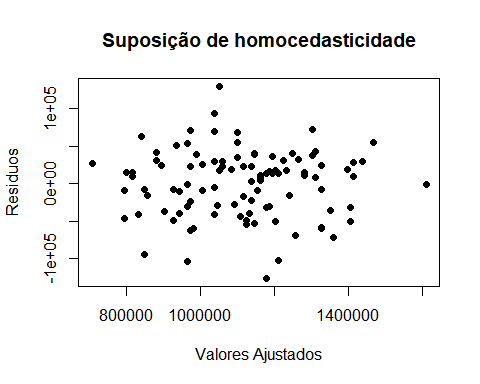
##   
## Durbin-Watson test  
##   
## data: mod  
## DW = 1.7309, p-value = 0.08659  
## alternative hypothesis: true autocorrelation is greater than 0

**16. Teste de Breusch Pagan**

1. A outra suposição que o modelo linear simples faz é que os erros são homocedásticos, ou seja, que as variâncias deles são iguais. Para isso, nós podemos utilizar, por exemplo, o gráfico dos valores ajustados contra os resíduos. Pelo gráfico não há alguma relação entre eles e os valores ajustados, mas novamente, isso é uma forma de verificar de forma visual, a gente precisa realizar um teste, só para confirmar.

Minha Hipótese nula é que eles tem a mesma variância.Nesse caso o p-valor não foi menor do que 0.05, então eu não tenho evidências para rejeitar a hipótese de que os erros são homocedásticos , portanto tudo ok com essa suposição do modelo, eles são homocedásticos , pelo menos não tenho evidências para dizer o contrário.

plot(mod$fitted.values, mod$residuals,   
 xlab = "Valores Ajustados",  
 ylab = "Residuos",  
 pch = 19,  
 main = "Suposição de homocedasticidade"  
 )



library(lmtest)  
bptest(mod)

##   
## studentized Breusch-Pagan test  
##   
## data: mod  
## BP = 0.12005, df = 1, p-value = 0.729

**17. Teste de Shapiro Wilks**

R. O teste que nós podemos utilizar para verificar a normalidade dos dados é o teste de Shapiro Wilks. A hipótese nula é que os dados tem distribuição normal. Então eu quero testar que os erros, em particular o resíduo, que é a estimativa do erro, tem distribuição normal com média zero.

como o p-valor do que 0.05, eu não tenho evidência para rejeitar que os erros provêm de uma distribuição normal.Então, eu aceito que eles têm distribuição normal por enquanto, porque eu não tenho evidência para dizer o contrário.

shapiro.test(mod$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: mod$residuals  
## W = 0.98356, p-value = 0.2492

**18. Conclusão**

A gente testou todas as suposições básicas do modelo de regressão linear simples e até aqui, nós já temos um modelo corretamente estimado e já permite obter estimativas, previsões futuras para novas casas, e determinar qual seria o preço delas.