Problemas de Regressão

Problemas diversos sobre Regressão Linear

1. Instale e carregue o pacote UsingR e carregue o conjunto de dados *father.son*. Faça a regressão linear, onde a altura dos filhos seja o resultado e a altura do pai seja preditor. Ache a interseção e a inclinação, plote os dados e sobreponha a linha de regressão.

```
data(father.son)
#
# Podemos calcular os coeficientes da regressão pelas fórmulas:
# Y = b0 + b1x
# b0 = intersecção, b1 = inclinação
# b0 = Y_hat - b1X_hat  e b1 = cor(y,x)*sd(y)/sd(x)

x <- father.son$fheight
y <- father.son$sheight

b1 <- cor(x,y)*sd(y)/sd(x)
b0 <- mean(y) - b1*mean(x)

fit <- lm(y ~ x)
rbind(coef(fit), c(b0, b1))
```

```
##      (Intercept)      x
## [1,]       33.89 0.5141
## [2,]       33.89 0.5141
```

```
cat("\n")
```

```
cat("O resumo dos coeficientes da regressão são: \n")
```

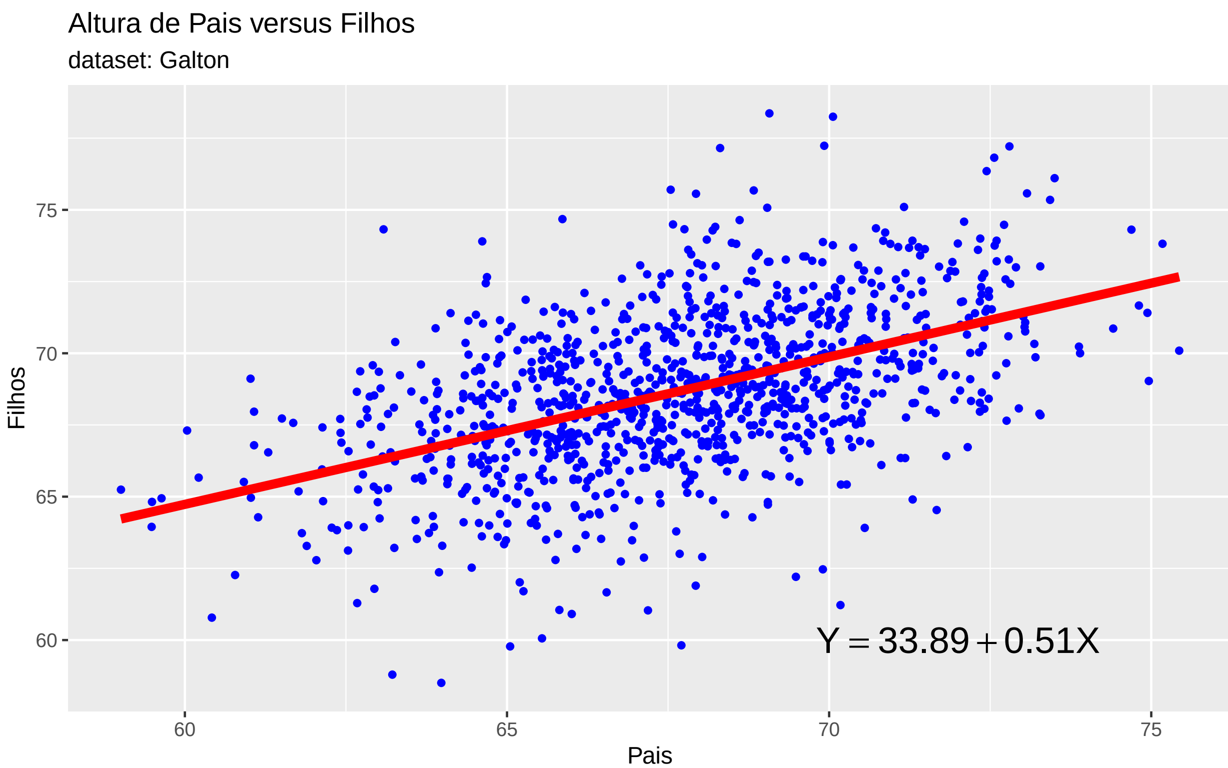
```
## O resumo dos coeficientes da regressão são:
```

```
cat("\n")
```

```
summary(fit)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  33.8866    1.83235   18.49 1.604e-66
## x             0.5141    0.02705   19.01 1.121e-69
```

```
p <- father.son %>%
     ggplot(aes(x = fheight, y = sheight)) +
     geom_point(color = "blue", size = 1.5, shape = 16)
p <- p + labs(title = "Altura de Pais versus Filhos", subtitle = "dataset: Galton", x = "Pais", y = "Fi
p <- p + geom_smooth(method = "lm", formula = y ~ x, color = 'red', lwd = 2, se = FALSE) + annotate("te
p
```



2. Referindo-se ainda ao problema. Centre as variáveis de altura dos pais e dos filhos e refit o modelo omitindo a intersecção. Verifique que a inclinação estimada é a mesma do problema 1.

```
xc <- x - mean(x)
yc <- y - mean(y)
refit <- lm(yc ~ xc - 1)
cat("inclinação estimada(centrada) = ", sum(xc * yc) / sum(xc^2), '\n')
```

```
## inclinação estimada(centrada) =  0.5141
```

```
cat("fit sem ajuste ao centro = ", coef(fit), '\n')
```

```
## fit sem ajuste ao centro =  33.89 0.5141
```

```
# cat("fit ajustado ao centro = ", coef(refit))
```

3. Usando os dados do problema. Normalize os dados (father, son) e veja se o slope da reta é a correlação.

```
r <- cor(x,y)
cat("A correlação entre os dados de altura dos pais e filhos é ", r, "\n")
```

```
## A correlação entre os dados de altura dos pais e filhos é  0.5013
```

```
xn <- (x - mean(x))/sd(x)
yn <- (y - mean(y))/sd(y)
fit_n <- lm(yn ~ xn - 1)
coef(fit_n)
```

```
##     xn
## 0.5013
```

4. Volte ao problema de regressão linear do problema 1 (acima). Se a altura do pai for 63 inches, qual seria a altura predita do filho, usando o modelo do problema 1

```
#
predict(fit, newdata = data.frame(x = 63))
```

```
##     1
## 66.27
```

```
b0 = coef(fit)[1]
b1 = coef(fit)[2]
y <- b0 + b1*63
```

5. Considere um dataset onde o desvio padrão da variável resposta é o dobro do variável controle. Sabendo que as variáveis tem uma correlação de 0.3. Se calcularmos um modelo de regressão linear, qual seria a estimativa da inclinação?

```
# A inclinação entre a variável resposta e a controle é dada, pela equação abaixo:

#  b1 <- cor(x,y)*sd(y)/sd(x)

cor_xy <- 0.3
sd_x <- 1
sd_y <- 2*sd_x

b1 <- cor_xy*sd_y/sd_x
b1
```

```
## [1] 0.6
```

A inclinação estimada será de $\beta_1 = 0.6$.

6. Considere o problema anterior. A variável resposta tem uma média de 1 e a variável de controle tem uma média de 0.5. Qual seria o valor da intersecção? A estimativa da intersecção é dada pela equação:

```
# bo = mean(y) - b1*mean(x)
avg_y <- 1
avg_x <- 0.5

b0 <- avg_y - b1*avg_x
b0
```

```
## [1] 0.7
```

7. Verdadeiro ou Falso, se a variável controle tem uma média de 0, a interseção estimada a partir da regressão linear também será uma média da variável resposta ?

Considerando a equação: $\beta_0 = Y - \beta_1 * X$ e aonde Y, X são as médias, e fazendo $X = 0$. A equação anterior se torna $\beta_0 = Y$, verdadeira a proposição.

8. Considere o problema 5 novamente. Qual seria a inclinação estimada, se a variável controle e a variável resposta fossem invertidas.

```
# A inclinação entre a variável resposta e a controle é dada, pela equação abaixo:

#  b1 <- cor(x,y)*sd(y)/sd(x)

cor_xy <- 0.3
sd_y <- 1      # invertendo x e y teremos
sd_x <- 2*sd_y

b1 <- cor_xy*sd_y/sd_x
```

```r
cat("A inclinação invertendo x e y no problema 5 será", b1, "\n")
```

```
## A inclinação invertendo x e y no problema 5 será 0.15
```

Regression to the mean

Ao estudar as estaturas de pais e filhos, Galton observou que filhos de pais com altura baixa em relação à media tendem ser mais altos que seus pais, e filhos de pais com estatura mais alta em relação a média tendem a ser mais baixos que seus pais, ou seja, as alturas dos seres humanos em geral tendem **regredir** à *média*.

1. Você tem duas escalas ruidosas e um algumas pessoas que vc gostaria de pesar. Você avalia cada pessoa em ambas as escalas. A correlação foi de 0.75. Se você normalizar cada conjunto de pesos, o que você multiplicar o peso na outra escala para obter uma boa estimativa do peso em outra escala?

```r
r <- 0.75 # slope de dados normalizados
# basta multiplicar uma das escalas pela correlação para obter o peso na outra escala
```

2. Considere o problema anterior. Uma pessoa tem um peso 2 desvios padrões acima da média, do grupo na primeira escala. Qtos desvios padrões acima da média seria a estimativa dessa pessoa no segundo grupo?

```r
r <- 0.75
p1 <- 2 # standard deviations above the mean, mean = 0
p2 <- r*p1
p2
```

```
## [1] 1.5
```

O peso no segundo grupo seria dado pela expressa acima, para p2

Statistical linear regression models:

1. Ajuste um modelo de regressão para o conjunto de dados **father.son** com o "the father" como variável controle e a variável **"the son"** como o resultado. Dado um p-value, para o coeficiente angular, faça um teste de hipótese relevante.

```r
data(father.son)


model <- lm(sheight ~ fheight, data = father.son)

summary(model)$coef
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  33.8866    1.83235   18.49 1.604e-66
## fheight       0.5141    0.02705   19.01 1.121e-69
```

Lembrando que o modelo acima pode ser escrito como: $Y = \beta_0 + \beta_1 * X + \epsilon$, onde Y = a variável resposta (the son) e o X = a variável controle (the father).

O teste de hipótese acima, para a variável $x : fheight$ é: $H_0 : \beta_1 = 0$ **-hipótese nula** e a hipótese alternativa: $H_1 \neq 0$, como o p-value é muito menor que o teste estatístico, podemos rejeitar a hipótese nula. Podemos crer que existe uma linearidade entre as duas variáveis do modelo.

2. Usando os dados do exercício 1. Interprete os parâmetros. Recentralize, para intercepção se necessário.

0.5141 = é o coeficiente angular, e significa que a 1 inch de aumento na altura "the father" há um aumento de 0.5141.

33.8866 = é o coeficiente linear, e significa a altura de “the son”, quando a altura do pai, for zero. Como não existe pai com a altura zero, podemos centralizar a variável do eixo x para ter uma melhor interpretabilidade do coeficiente.

```r
model2 <- lm(sheight ~ I(fheight - mean(fheight)), data = father.son)

summary(model2)$coef
```

```
##                            Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 68.6841    0.07421  925.53 0.000e+00
## I(fheight - mean(fheight))   0.5141    0.02705   19.01 1.121e-69
```

A estimativa de `r coef(model2)[1]` é igual a média do valor da variável controle.

3. Usando os dados da questão 1. Faça a previsão da altura do “son” se a altura do pai for de 80 inches. Você recomendaria essa previsão? Pq ou pq não?

```r
p <- 80

predict.lm(model, newdata = data.frame(fheight = p))
```

```
##     1
## 75.01
```

```r
summary(father.son)
```

```
##     fheight         sheight
##  Min.   :59.0   Min.   :58.5
##  1st Qu.:65.8   1st Qu.:66.9
##  Median :67.8   Median :68.6
##  Mean   :67.7   Mean   :68.7
##  3rd Qu.:69.6   3rd Qu.:70.5
##  Max.   :75.4   Max.   :78.4
```

Podemos até usar essa previsão, contudo temos que lembrar que ela está um pouco além da média, e do máximo valor nos dados observados e não temos suficiente informações nessa parte da calda da distribuição dos dados.

4. Carrega o conjunto de dados **mtcars**. Ajuste uma regressão linear com as variáveis **mpg** como a variável de saída, e **horsepower** como variável de controle. Interprete os coeficientes, recentralize se for necessário.

```r
data("mtcars")
fit_mtcars <- lm(mpg ~ hp, data = mtcars)

b1 <- coef(fit_mtcars)[2]
b0 <- coef(fit_mtcars)[1]
```

Podemos ver pelos coeficiente -0.0682 que há uma relação inversa, ou seja cada vez que a cada variação 1 variação em hp, temos um decrescimo de -0.0682.

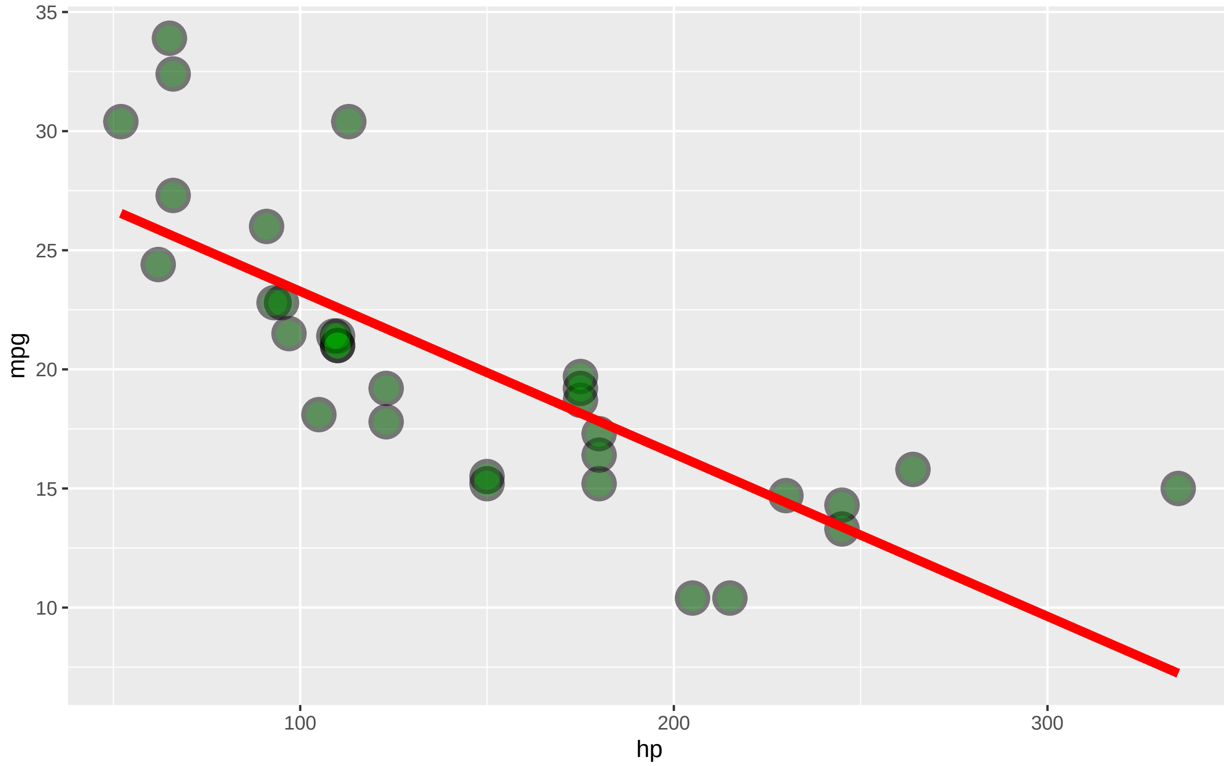
```r
# centralizando o modelo

fit_mtcars2 <- lm(mpg ~ I(hp - mean(hp)), data = mtcars)
summary(fit_mtcars2)$coef
```

```
##                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)      20.09062    0.68288  29.420 1.102e-23
## I(hp - mean(hp)) -0.06823    0.01012  -6.742 1.788e-07
```

Agora podemos interpretar `r coef(fit_mtcars2)[1]` como sendo o valor para a média do mpg.

5. Em relação a questão 4, plote a reta ajustada ao diagrama de dispersão das variáveis usadas para criar o modelo.

```
g <- ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point(size = 7, colour = "black", alpha = 0.5)
g = g + geom_point(size = 5, colour = "green", alpha = 0.2)
g = g + geom_smooth(method = "lm", colour = "red", se = FALSE, lwd = 2)
g
```



6. Utilizando o modelo da questão 4. Teste a hipótese de relacionamento não linear entre horsepower e milhas por galão.

```
summary(fit_mtcars2)$coef
```

```
##                   Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       20.09062    0.68288  29.420 1.102e-23
## I(hp - mean(hp)) -0.06823    0.01012  -6.742 1.788e-07
```

Pelo valor do teste estatístico de β_1 que é a inclinação da reta de ajuste e explica uma relação negativa entre hp e mpg, podemos rejeitar a hipótese nula de não relação, porque o teste é significativo e o p-value é quase zero. O que significa que é bem significativo a relação de linearidade entre as variáveis do modelo o que é reforçado pela rejeição da hipótese nula e aceitação da hipótese alternativa.

7. Em relação ainda à questão 04. Prediga, mpg para um valor de hp = 111.

```
summary(mtcars$hp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    52.0    96.5   123.0   146.7   180.0   335.0
```

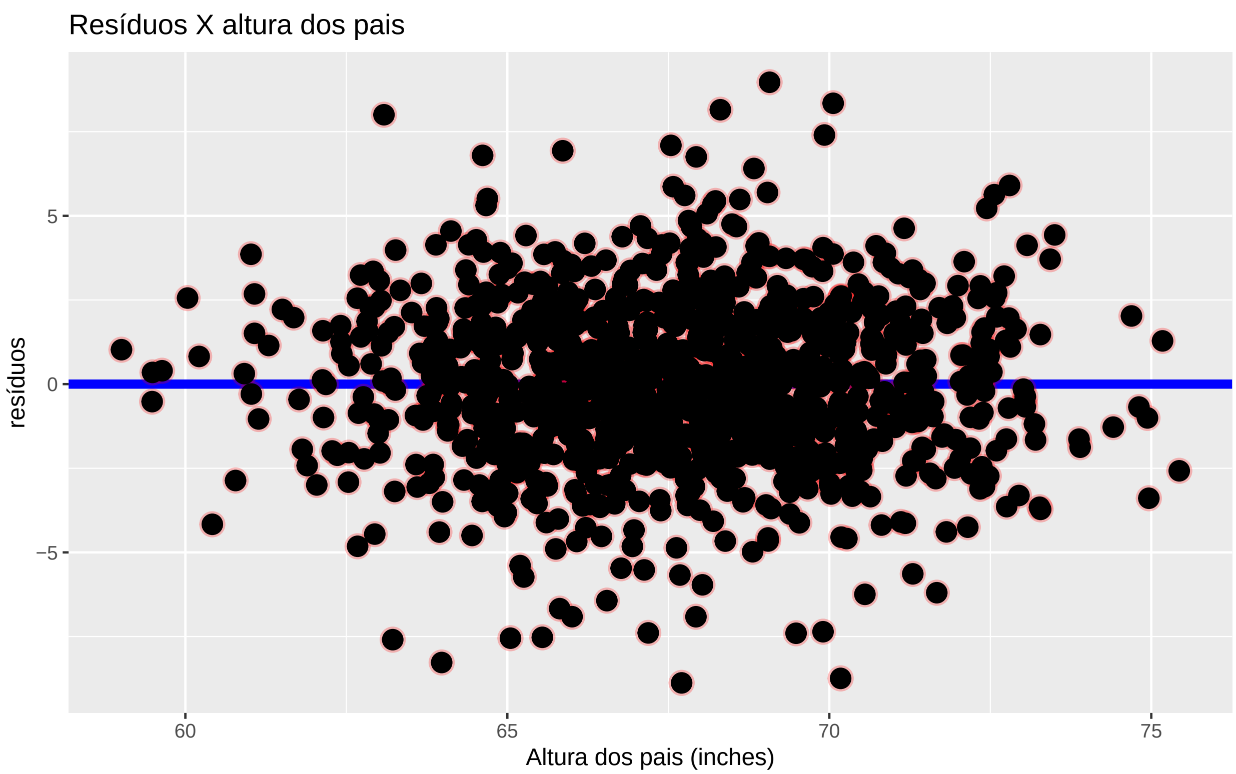
```r
predict.lm(fit_mtcars, newdata = data.frame(hp = 111))
```

```
##     1
## 22.53
```

Resíduos:

1. Ajuste um modelo de regression linear para o conjunto de dados **father.son** com "the father" como variável explicativa e a variável "the son" como a variável resposta. Plote a altura do "father" versus os resíduos (eixo vertical).

```r
data("father.son")
x <- father.son$fheight
y <- father.son$sheight
fit <- lm(y ~ x, data = father.son)
father.son$y_hat <- predict(fit)
father.son$e <- resid(fit)

g <- father.son %>%
     ggplot(aes(x = fheight, y = e))
g = g + geom_hline(lwd = 2, color = "blue", yintercept = 0)
g = g + geom_point(color = "red", size = 5, alpha = 0.25)
g = g + geom_point(color = "black", size = 4)
g = g + labs(title = "Resíduos X altura dos pais", x = "Altura dos pais (inches)", y = "resíduos")
g
```



2. Com relação a questao 1. Estime, diretamente a variância residual e compare com a estimativa de saída da função lm.

```r
n <- nrow(father.son)
sum(resid(fit)^2)/(n - 2)
```

```
## [1] 5.937
```

```r
summary(fit)$sigma^2
```

```
## [1] 5.937
```

A variação residual é o que resta após modelo ser explicado pela variável resposta.

3. Com relação a questão 1. Calcule o R^2 para este modelo. Sabemos das aulas anteriores que, a correlação é derivada assim :

$$\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$$

   Então R^2 é literalmente r ao quadrado.

```r
# No summary(fit), tem a informação de Adjusted - R squared,
# que é o ajuste para o número de coeficientes que se tem no modelo
# como esse modelo só tem 2 variáveis, x, y e o número de dados é grande
# não terá muita diferença na resposta final, mais para uma amostra de
# dados menor esse termo pesará.

r <- cor(x,y)^2
R_squared <- r
R_squared
```

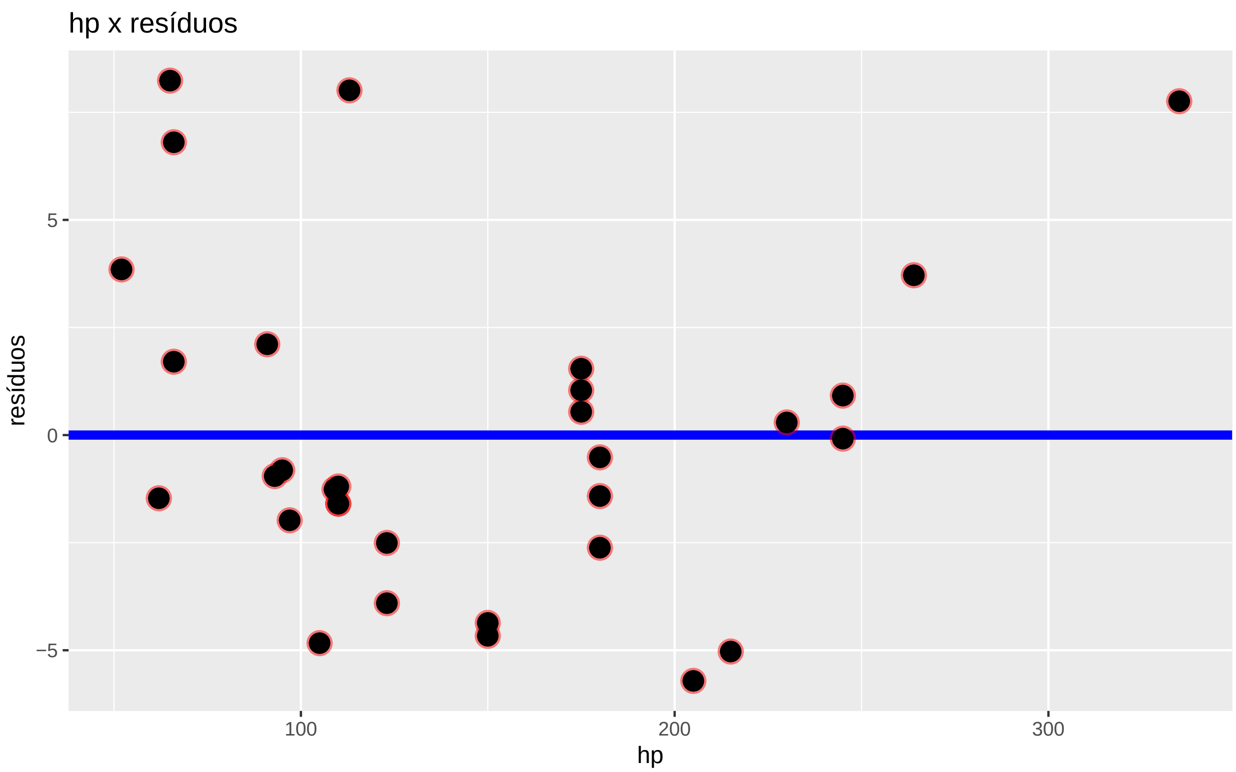
```
## [1] 0.2513
```

```r
summary(fit)$r.squared
```

```
## [1] 0.2513
```

Importante lembrar que nesse modelo apenas 25% da variável resposta é explicada pela linearidade com a variável explicativa.

4. Carrega o dataset **mtcars**. Ajuste uma regressão linear com as variáveis mpg como resposta hp como variável explicativa. Plote hp x resíduos.



5. Com os dados do modelo da questão anterior, estime diretamente a variância residual e compare com a estimativa da saída da função lm.

```
n <- nrow(mtcars)
sum(resid(model)^2)/(n - 2)
```

```
## [1] 14.92
```

```
summary(model)$sigma^2
```

```
## [1] 14.92
```

A variância residual é o que o modelo não consegue explicar. 6. A partir do modelo de ajuste linear da questão 4, derive o R^2.

```
summary(model)$r.squared
```

```
## [1] 0.6024
```

60% da variação mpg é explicada pela relação linear com hp.

Estatística inferencial para modelos de regressão Linear.

1. Teste se o coeficiente angular para o dataset "father.son" é diferente de zero (*father* como variável independente e o *son* como variável dependente.)

Solução:

```
fit <- lm(sheight ~ fheight, data = father.son)
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = sheight ~ fheight, data = father.son)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.877 -1.514 -0.008  1.629  8.968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.887      1.832    18.5   <2e-16 ***
## fheight        0.514      0.027    19.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.44 on 1076 degrees of freedom
## Multiple R-squared:  0.251,  Adjusted R-squared:  0.251
## F-statistic:  361 on 1 and 1076 DF,  p-value: <2e-16
```

Podemos ver na tabela acima, o coeficiente angular $\beta_1 = 0.514$ tem um resultado t value bem diferente de zero e com p-value igual a zero, o coeficiente angular é significante. Isto é um teste de hipótese para β_1. Um teste para a relação de linearidade entre os coeficiente linear e angular da reta ajustada.

$$H_0 : \beta_1 = 0 H_a : \beta_1 \neq 0$$

2. Usando o modelo do problema 1 (anterior).Forme o intervalo de confiança para coeficiente angular. Usando o extrator de funções, teremos:

```
confint(fit, parm = 2)
```

```
##         2.5 % 97.5 %
## fheight 0.461 0.5672
```

O intervalo de confiança nos mostra que há um grau de incerteza na estimação dos coeficientes do modelo.

3. Usando o modelo da questão 1, forme um intervalo de confiança para intercepção (coeficiente linear da reta), (centralize a variável independente para ficar mais fácil a interpretação da intersecção)

Para resolver esse problema, vamos primeiro centralizar a variável x do modelo, e então refazemos o modelo, com o centro na média, isso não mudará a inclinação da reta, mas irá dar uma interpretação a intercecção β_0.

```
fit <- lm(sheight ~ I(fheight - mean(fheight)), data = father.son)

confint(fit, parm = 1)
```

```
##             2.5 % 97.5 %
## (Intercept) 68.54  68.83
```

A idéia de centrar a média da variável independente, agora o β_0 significa o valor predito para a média da variável independente. Traduzindo para o problema estudado, é a altura do filho predita, para um valor médio de altura do pai.

4. Referenciando ainda à questão 1, e usando a informação obtida na questão anterior (centralizando a média), forme um intervalo para a altura esperada do filho a altura média do pai.

```
avg <- mean(father.son$fheight)
fit <- lm(sheight ~ fheight, data = father.son)
predict(fit, newdata = data.frame(fheight = avg), interval = "confidence")
```

```
##     fit   lwr   upr
## 1 68.68 68.54 68.83
```

5. Usando os dados da questão. Forme um intervalo de predição para a altura do filho à média da altura do pai.

```
fit <- lm(sheight ~ I(fheight - mean(fheight)), data = father.son)
predict(fit, newdata = data.frame(fheight = avg), interval = "prediction")
```

```
##     fit  lwr   upr
## 1 68.68 63.9 73.47
```

6. Carregue o dataset **mtcars**. Ajuste um modelo de regressão a variáveis mpg (dependente) e hp como independente. Teste se hp é ou não estatisticamente diferente de zero. Interprete o resultado. Vamos definir o modelo e usar o R para calcular o ajuste do modelo primeiramente.

```
data("mtcars")
model <- lm(mpg ~ hp, data = mtcars)
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.712 -2.112 -0.885  1.582  8.236
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.0989     1.6339   18.42  < 2e-16 ***
## hp           -0.0682     0.0101   -6.74  1.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.86 on 30 degrees of freedom
## Multiple R-squared:  0.602,  Adjusted R-squared:  0.589
## F-statistic: 45.5 on 1 and 30 DF,  p-value: 1.79e-07
```

- Solução :Da tabela acima, que contém as principais estatísticas relacionadas ao modelo adotado, que é $mpg = 30.10 - 0.01 * hp$, temos que t-value é diferente de zero, e p-value tem um valor próximo de zero, o que corrobora que $\beta_1 \neq 0$ e podemos rejeitar a hipótese nula e aceitar a hipótese alternativa de que o coeficiente angular estatisticamente diferente de zero e que **mpg** tem uma relação linear com hp, o sinal negativo de **hp**, nos informa que a medida que a medida que aumenta-se a potência dos motores nos carros, diminui-se a autonomia dos veículos.

7. Com o resultado para os coeficientes do modelo da questão anterior, forme um intervalo de confiança para o coeficiente angular.

- Solução: O intervalo de confiança para o coeficiente angular é interessante porque nos mostra a incerteza associada a determinação do ajuste dos coeficiente ao modelo adotado, e para o resultado obtido ainda temos a confirmação da significância estatística para β_1, podemos ver que o zero não está incluso no intervalo.

```
model <- lm(mpg ~ hp , data = mtcars)

confint(model, parm = 2)
```

```
##        2.5 %   97.5 %
## hp -0.08889 -0.04756
```

8. Usando os dados da questão 6. Forme um **IC** para a intercecção da reta de ajuste(centre a variável hp primeiro).

- Solução: Primeiro temos que refitar o modelo fazendo a centralizando de hp, para dar significado ao β_0, que será o consumo considerando a potência média dos carros no modelo adotado.

```
model <- lm(mpg ~ I(hp - mean(hp)), data = mtcars)
confint(model, parm = 1)
```

```
##             2.5 % 97.5 %
## (Intercept)  18.7  21.49
```

9. Usando o dataset do problema 6. Forme um intervalo de predição para o valor esperado de mpg condicionado ao valor médio de hp.

- Solução: Vamos primeiramente calcular o valor médio de hp, para o conjunto de dados e usar esse valor, para predizer o valor mpg, dado o valor médio de potência dos carros.

```
avg_hp <- mean(mtcars$hp)
avg_hp
```

```
## [1] 146.7
```

```
model <- lm(mpg ~ hp, data = mtcars)
predict(model, newdata = data.frame(hp = avg_hp), interval = "confidence")
```

```
##     fit  lwr   upr
## 1 20.09 18.7 21.49
```

10. Forme um intervalo de predição para o valor esperado de mpg para o valor médio de hp.

```
model <- lm(mpg ~ I(hp - mean(hp)), data = mtcars)
predict(model, newdata = data.frame(hp = avg_hp), interval = "prediction")
```

```
##     fit   lwr  upr
## 1 20.09 12.08 28.1
```

11. Cria um gráfico, com a linha de regressão e os valores esperados e os intervalos de predição.

```
x <- mtcars$hp
y <- mtcars$mpg
fit <- lm(y ~ x)
newx = data.frame(x = seq(min(x), max(x), length = 100))
p1 = data.frame(predict(fit, newdata = newx,interval = ("confidence")))
p2 = data.frame(predict(fit, newdata = newx,interval = ("prediction")))

p1$interval = "confidence"
p2$interval = "prediction"
p1$x = newx$x
p2$x = newx$x
dat = rbind(p1, p2)
names(dat)[1] = "y"


 g = ggplot(dat, aes(x = x, y = y))
 g = g + geom_ribbon(aes(ymin = lwr, ymax = upr, fill =   interval), alpha = 0.2)
 g = g + geom_line(color =  "blue", lwd = 1.5)
 g = g + geom_point(data = data.frame(x = x, y = y),
 aes(x = x, y = y), size = 4, color = "red")
 g = g + labs(title = "Regression: hp x mpg",
```

```
subtitle = "dataset mtcars", x = "hp", y = "mpg" )
g
```