Regressão simples

Danilo de Paula Santos/Gabriela Wunsch Lopes

17/10/2019

```
# install.packages("compareGroups")
# install.packages("ggplot2")
# install.packages("data.table")


library("compareGroups")
library("ggplot2")
library("haven")
library("ggpubr")
library("dplyr")
library(knitr)
```

Suwit dataset

Estudo da relação entre infecção por ancilóstomo e perda de sangue. Tailândia 1970

```
suwit <- read_sav("Bancos/Suwit.sav")
```

Primeiro passo: verificar o banco

```
compare_suwit <- compareGroups( ~ ., data = suwit)

summary(compare_suwit)
```

```
##
##  --- Descriptives of each row-variable ---
##
## -------------------
## row-variable: Identificação
##
##       N  mean sd       lower    upper
## [ALL] 15 8    4.472136 5.523414 10.47659
##
## -------------------
## row-variable: número de vermes
##
##       N  mean  sd       lower    upper
## [ALL] 15 552.4 513.9007 267.8113 836.9887
##
## -------------------
## row-variable: Perda de sangue por dia
##
##       N  mean     sd       lower    upper
## [ALL] 15 33.45267 24.85249 19.68982 47.21551
```

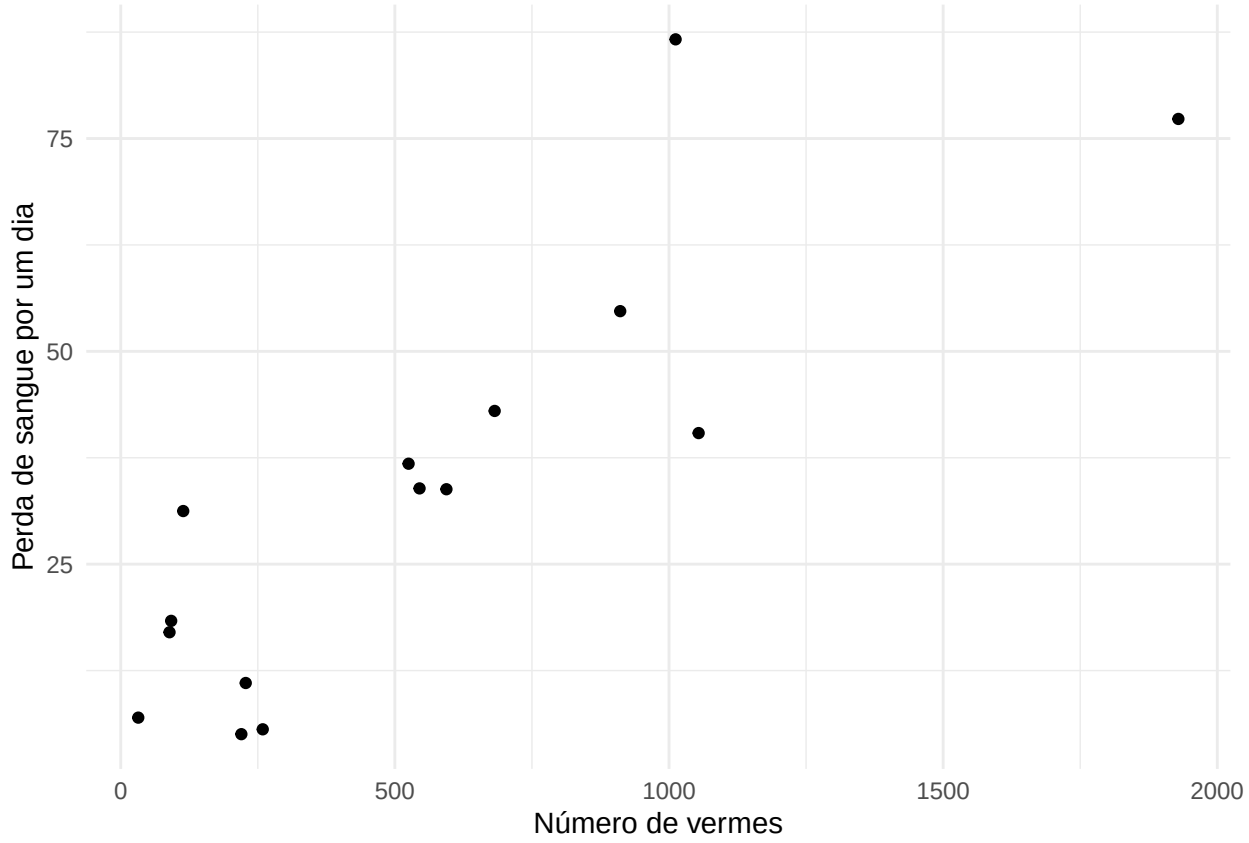
```r
tabela_suwit <- createTable(compare_suwit)

tabela_suwit
```

```
##
## --------Summary descriptives table ---------
##
## ---------------------------------------
##                              [ALL]    N
##                              N=15
## ---------------------------------------
## Identificação           8.00 (4.47) 15
## número de vermes         552 (514)  15
## Perda de sangue por dia 33.5 (24.9) 15
## ---------------------------------------
```

Segundo passo: verificar linearidade com gráfico de dispersão e identificar pontos

1) Gráfico de dispersão

```r
suwit_scatter <- ggplot(suwit, aes(x = vermes, y = perda_sangue))+
  geom_point() +
  scale_x_continuous("Número de vermes") +
  scale_y_continuous("Perda de sangue por um dia") +
  theme_minimal()

suwit_scatter
```



2) Identificando pontos: qual o ID do ponto mais extremo na vertical?

```
suwit %>%
  filter(perda_sangue>75, vermes<1500)
```

```
## # A tibble: 1 x 3
##      id vermes perda_sangue
##   <dbl>  <dbl>        <dbl>
## 1    13   1012         86.7
```

Terceiro passo: estimar modelo

O modelo de regressão linear simples para a perda de sangue da amostra em função da variável preditora "número de vermes" pode ser visualizado como uma linha reta no gráfico, estimada pelo método dos mínimos quadrados, que corresponde à melhor representação de uma linha que percorre os dados tentando minimizar a distância entre a linha e as observações (pontos)

```
lm_suwit <- lm(formula = perda_sangue ~ vermes,
                    data = suwit)

summary(lm_suwit)
```

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

```
## lm(formula = perda_sangue ~ vermes, data = suwit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.846 -10.812   0.750   4.356  34.390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.847327   5.308569   2.043   0.0618 .
## vermes       0.040922   0.007147   5.725 6.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.74 on 13 degrees of freedom
## Multiple R-squared:  0.716,  Adjusted R-squared:  0.6942
## F-statistic: 32.78 on 1 and 13 DF,  p-value: 6.99e-05
```

Além da estimativa de B0 / intercepto (Intercept) e B1/ inclinação (vermes), também observamos o resultado de um teste de significância do preditor B1 (Pr(>|t|)). O coeficiente angular da reta da amostra (b= 0.04092) é significativamente diferente de zero, sendo o p-valor do teste 6.99e-05 *** O coeficiente de determinação (Multiple R-squared) é 0.716, indicando que 71,6% da variabilidade da perda de sangue (o quanto varia em relação ao valor de perda média da amostra) é explicada pelo aumento do número de vermes.

Gerando intervalo de confiança para os parâmetros estimados. O intervalo de confiança estima quanto, em média, é a perda sanguínea de um grupo de pacientes com determinado número de vermes

```
confint(lm_suwit)
```

```
##                   2.5 %      97.5 %
## (Intercept) -0.62113883 22.31579336
## vermes       0.02548089  0.05636321
```

A cada aumento de 1 verme, ocorre um aumento de 0.04 (95%IC 0.026 - 0.056) na perda de sangue.

Estimando intervalo de predição para cada observação do banco. O intervalo de predição prediz a perda de sangue para um paciente que apresenta determinado número de vermes

```
suwit_pred <- predict(lm_suwit, interval = 'predict')
```

```
## Warning in predict.lm(lm_suwit, interval = "predict"): predictions on current data refer to _future_
```

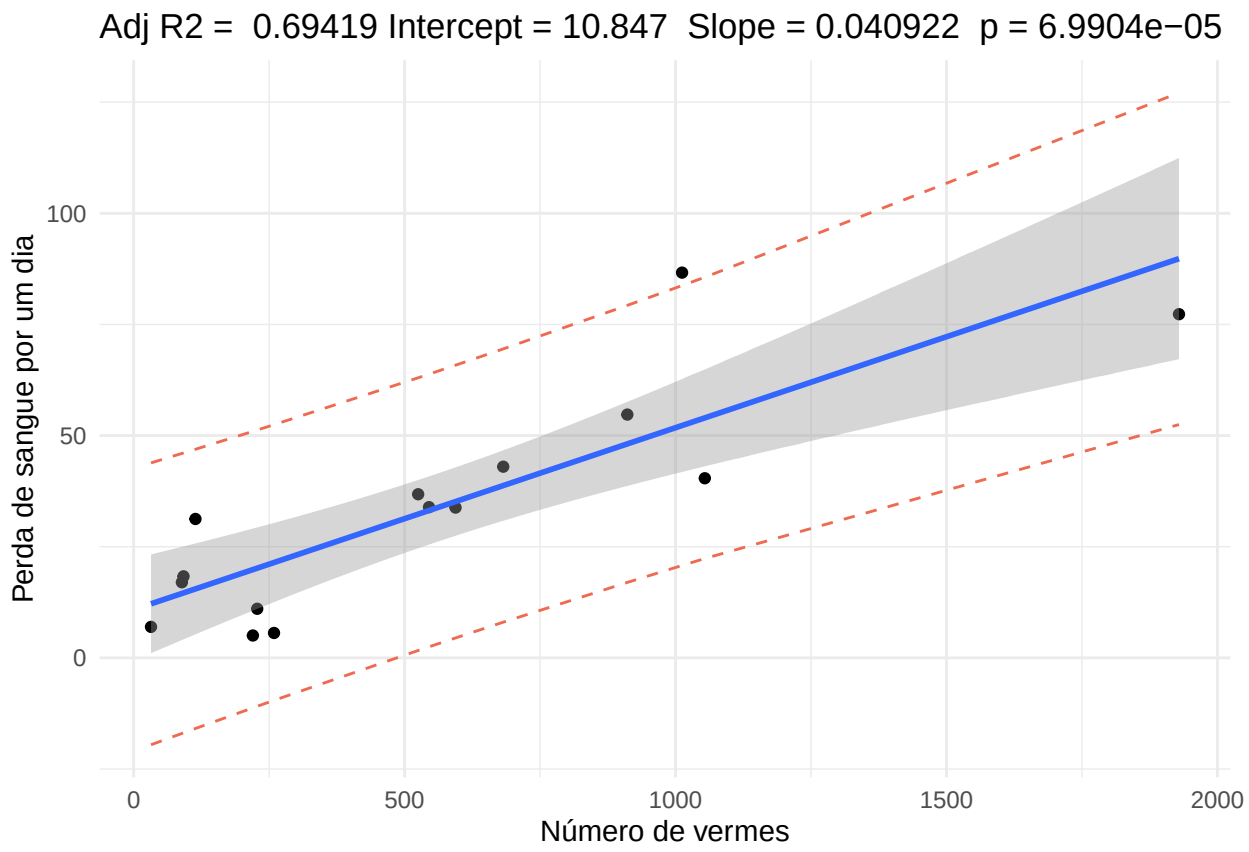
Incluindo o intervalo de predição para cada observação do banco

```
suwit <- cbind(suwit, suwit_pred)
```

Gerando um gráfico contendo intervalo de confiança e intervalo de predição

```
suwit_pred_plot <- ggplot(suwit, aes(x = vermes, y = perda_sangue)) +
                                 geom_point() +
                                 geom_smooth(method = "lm") +
                                 scale_x_continuous("Número de vermes") +
                                 scale_y_continuous("Perda de sangue por um dia") +
```

```
                                  theme_minimal() +
                                  geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") +
                                  geom_line(aes(y = upr), col = "coral2", linetype = "dashed") +
  labs(title = paste("Adj R2 = ",signif(summary(lm_suwit)$adj.r.squared, 5),
                           "Intercept =",signif(lm_suwit$coef[[1]],5 ),
                           " Slope =",signif(lm_suwit$coef[[2]], 5),
                           " p =",signif(summary(lm_suwit)$coef[2,4], 5)))



suwit_pred_plot
```



A faixa cinza corresponde ao intervalo de confiança e as linhas laranja são os limites do intervalo de predição.

Fat_dat dataset

Estudo Fat_dat "Fitting Percentage of Body Fat to Simple Body Measurements"

```
fat_dat <- read_sav("Bancos/fat_dat.sav")
```

Primeiro passo: verificar o banco e corrigir os erros de digitação:

- The body densities for cases 48, 76, and 96, for instance, each seem to have one digit in error as can be seen from the two body fat percentage values.
- Case 42) over 200 pounds in weight who is less than 3 feet tall (the height should presumably be 69.5 inches, not 29.5 inches)!
- The percent body fat estimates are truncated to zero when negative (case 182)

```
# Altura

fat_dat$altura_pol <- ifelse( fat_dat$numero == 42, 69.5, fat_dat$altura_pol)

# Densidades

fat_dat$densidade <- ifelse( fat_dat$numero == 48, 1.0865, fat_dat$densidade)
fat_dat$densidade <- ifelse( fat_dat$numero == 76, 1.0566, fat_dat$densidade)
fat_dat$densidade <- ifelse( fat_dat$numero == 96, 1.0591, fat_dat$densidade)
```

```
compare_fat <- compareGroups( ~ ., data = fat_dat)

summary(compare_fat)
```

```
##
##  --- Descriptives of each row-variable ---
##
## --------------------
## row-variable: numero
##
##       N   mean  sd       lower    upper
## [ALL] 252 126.5 72.89033 117.4569 135.5431
##
## --------------------
## row-variable: fat_Brozek
##
##       N   mean     sd       lower    upper
## [ALL] 252 18.93849 7.750856 17.97689 19.9001
##
## --------------------
## row-variable: fat_Siri
##
##       N   mean     sd      lower    upper
## [ALL] 252 19.15079 8.36874 18.11253 20.18906
##
## --------------------
## row-variable: densidade
##
##       N   mean     sd       lower    upper
## [ALL] 252 1.055455 0.018909 1.053109 1.057801
##
## --------------------
## row-variable: idade
##
##       N   mean     sd       lower    upper
## [ALL] 252 44.88492 12.60204 43.32146 46.44838
##
```

```
## -------------------
## row-variable: Peso em libras
##
##       N   mean     sd       lower    upper
## [ALL] 252 178.9244 29.38916 175.2783 182.5706
##
## -------------------
## row-variable: altura_pol
##
##       N   mean     sd       lower    upper
## [ALL] 252 70.30754 2.609583 69.98378 70.6313
##
## -------------------
## row-variable: kg/m2
##
##       N   mean    sd       lower   upper
## [ALL] 252 25.4369 3.648111 24.9843 25.88951
##
## -------------------
## row-variable: peso da massa magra
##
##       N   mean     sd       lower   upper
## [ALL] 252 143.7139 18.23164 141.452 145.9758
##
## -------------------
## row-variable: Circunferencia do pescoço
##
##       N   mean     sd       lower    upper
## [ALL] 252 37.99206 2.430913 37.69047 38.29365
##
## -------------------
## row-variable: Circunferencia do peito
##
##       N   mean     sd       lower    upper
## [ALL] 252 100.8242 8.430476 99.77829 101.8701
##
## -------------------
## row-variable: Circ do abdomem
##
##       N   mean     sd       lower    upper
## [ALL] 252 92.55595 10.78308 91.21816 93.89375
##
## -------------------
## row-variable: circ do quadril
##
##       N   mean     sd       lower    upper
## [ALL] 252 99.90476 7.164058 99.01596 100.7936
##
## -------------------
## row-variable: circ do coxa
##
##       N   mean     sd       lower    upper
## [ALL] 252 59.40595 5.249952 58.75462 60.05728
##
```

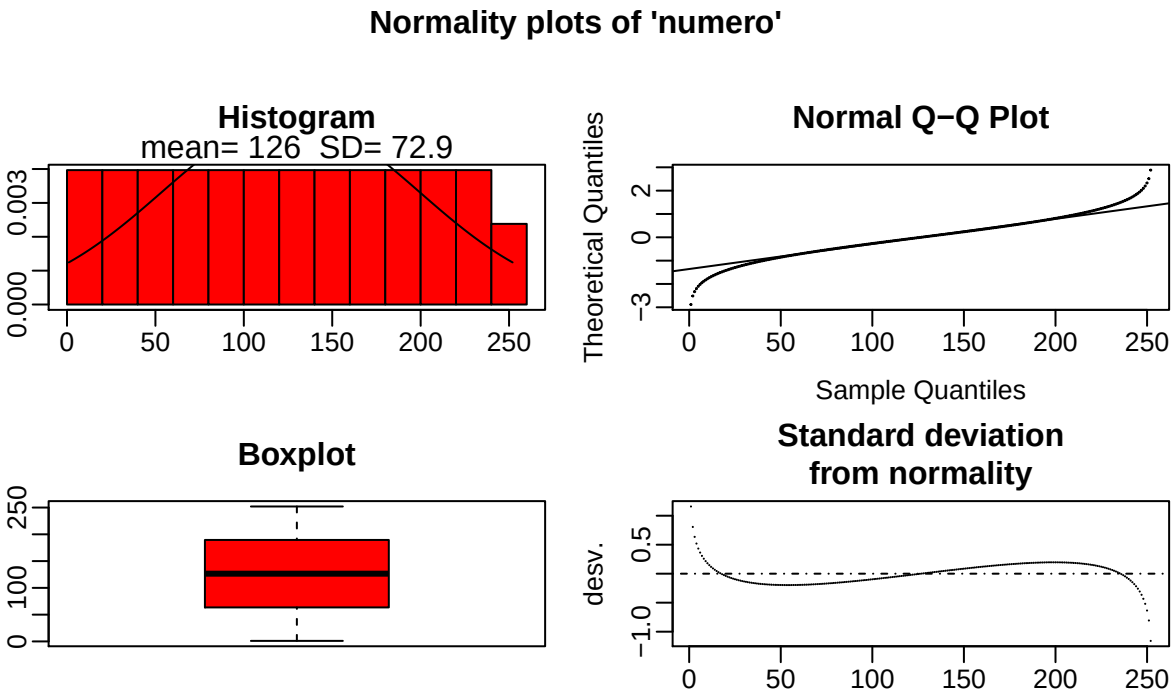
```
## --------------------
## row-variable: circ do joelho
##
##       N   mean     sd       lower    upper
## [ALL] 252 38.59048 2.411805 38.29126 38.8897
##
## --------------------
## row-variable: circ do tornozelo
##
##       N   mean     sd       lower    upper
## [ALL] 252 23.10238 1.694893 22.89211 23.31266
##
## --------------------
## row-variable: circ do biceps
##
##       N   mean     sd       lower    upper
## [ALL] 252 32.27341 3.021274 31.89858 32.64825
##
## --------------------
## row-variable: circ do antebraco
##
##       N   mean     sd       lower    upper
## [ALL] 252 28.66389 2.020691 28.41319 28.91458
##
## --------------------
## row-variable: circ do pulso
##
##       N   mean     sd       lower    upper
## [ALL] 252 18.22976 0.933585 18.11394 18.34559
```

```
tabela_fat <- createTable(compare_fat)

tabela_fat
```

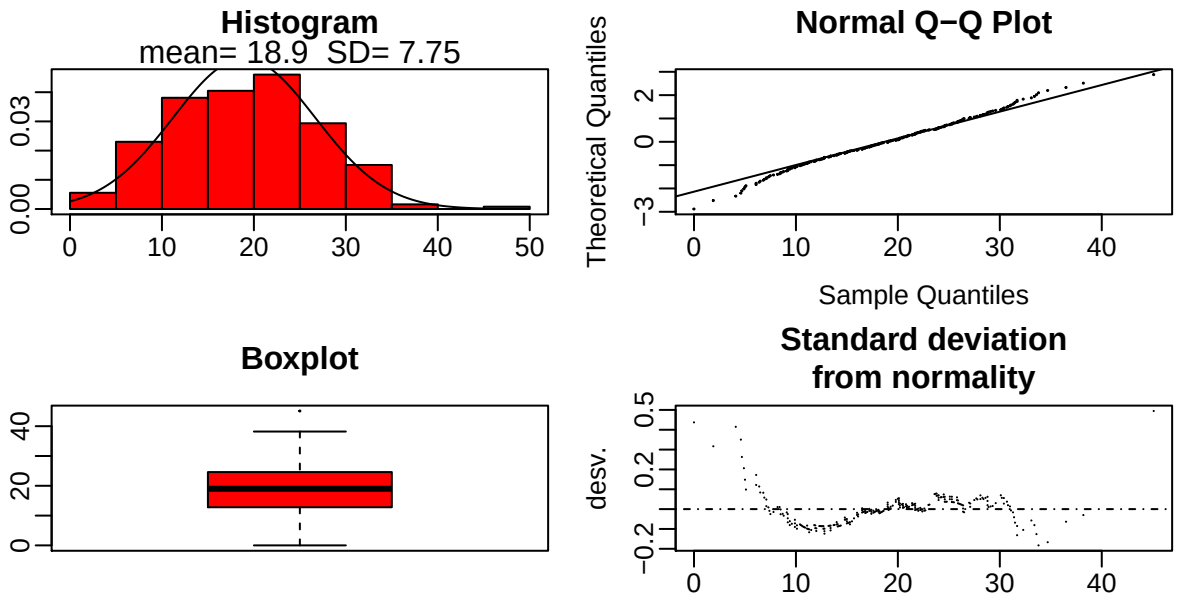
```
##
## --------Summary descriptives table ---------
##
## ------------------------------------------
##                               [ALL]     N
##                               N=252
## ------------------------------------------
## numero                     126 (72.9)  252
## fat_Brozek                 18.9 (7.75) 252
## fat_Siri                   19.2 (8.37) 252
## densidade                  1.06 (0.02) 252
## idade                      44.9 (12.6) 252
## Peso em libras             179 (29.4)  252
## altura_pol                 70.3 (2.61) 252
## kg/m2                      25.4 (3.65) 252
## peso da massa magra        144 (18.2)  252
## Circunferencia do pescoço 38.0 (2.43) 252
## Circunferencia do peito    101 (8.43)  252
## Circ do abdomem            92.6 (10.8) 252
## circ do quadril            99.9 (7.16) 252
```

```
## circ do coxa              59.4 (5.25) 252
## circ do joelho            38.6 (2.41) 252
## circ do tornozelo         23.1 (1.69) 252
## circ do biceps            32.3 (3.02) 252
## circ do antebraco         28.7 (2.02) 252
## circ do pulso             18.2 (0.93) 252
## ----------------------------------------
```

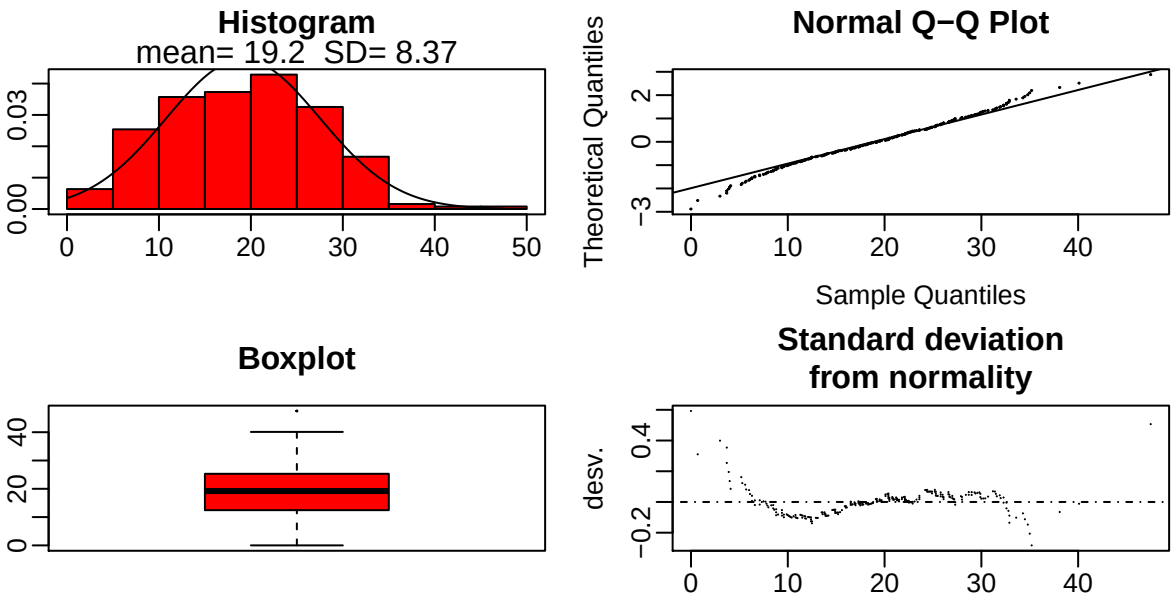
```
plot(compare_fat)
```



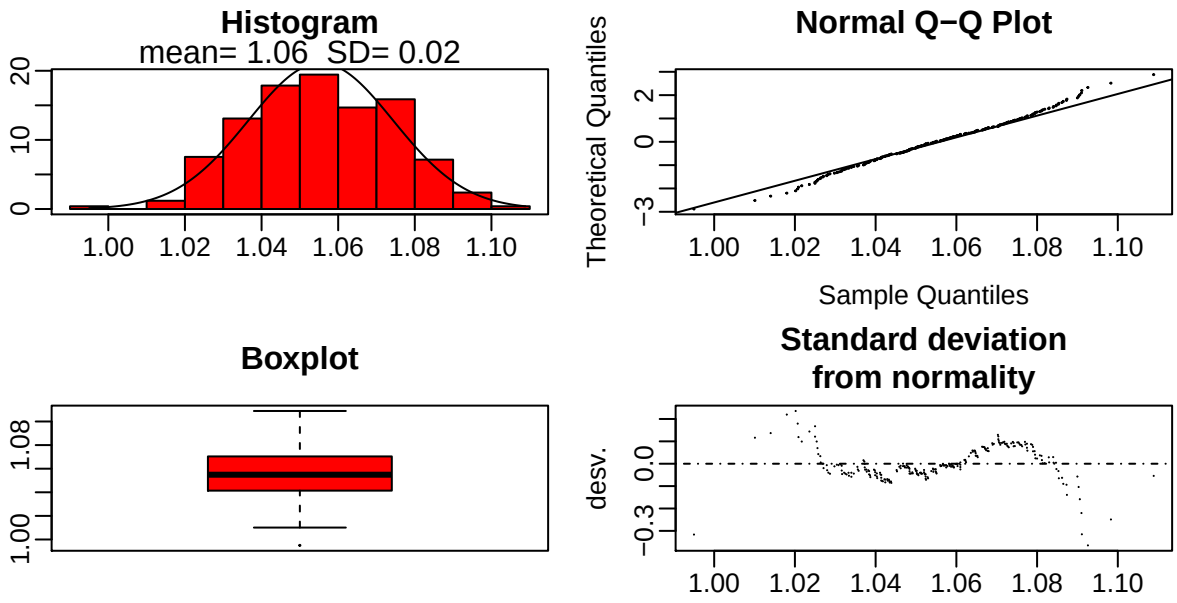
Shapiro−Wilks p−value: <0.001



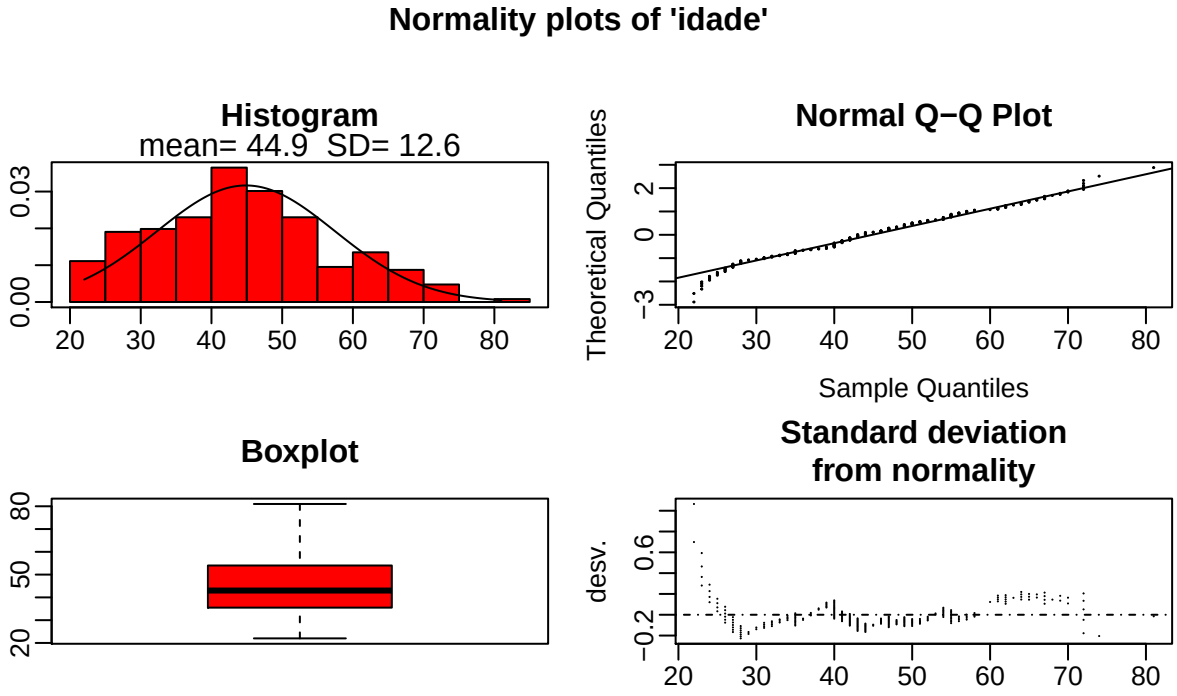
Shapiro−Wilks p−value: 0.275



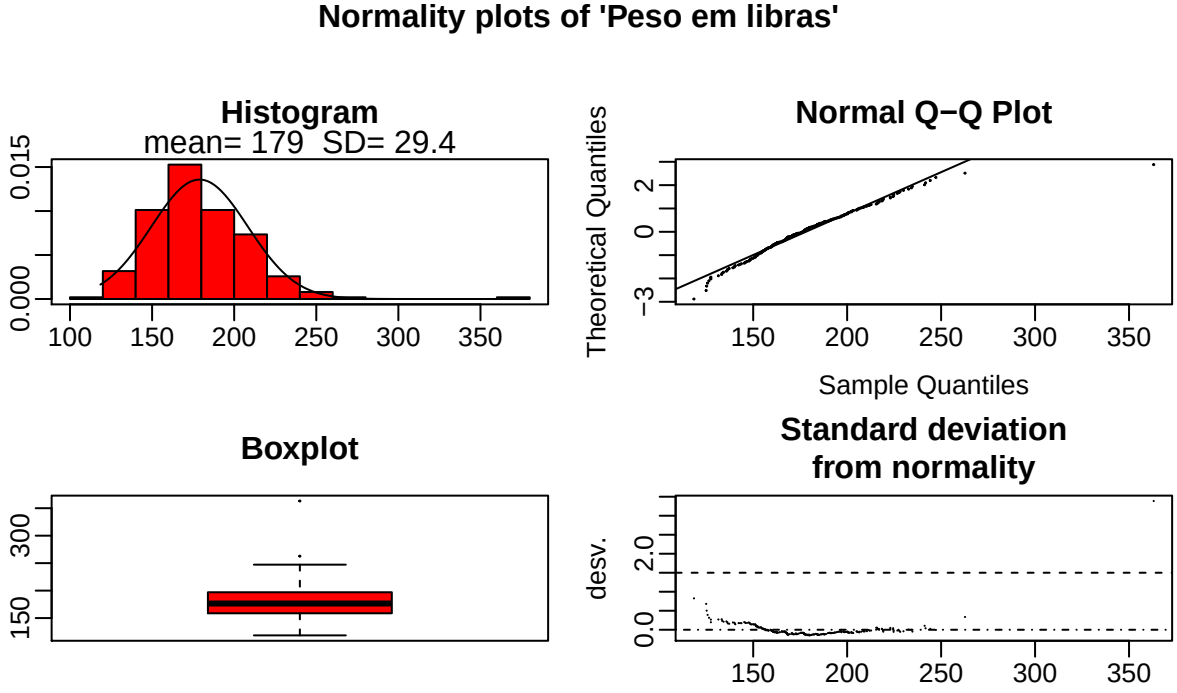
Shapiro−Wilks p−value: 0.165



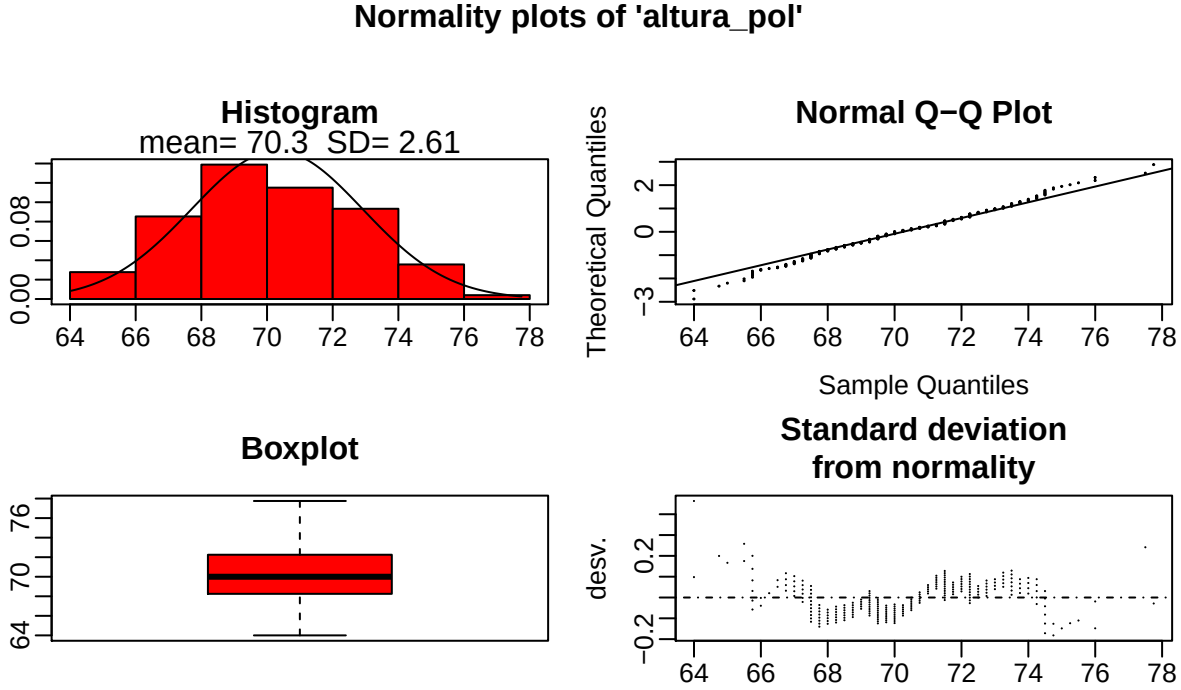
Shapiro−Wilks p−value: 0.522



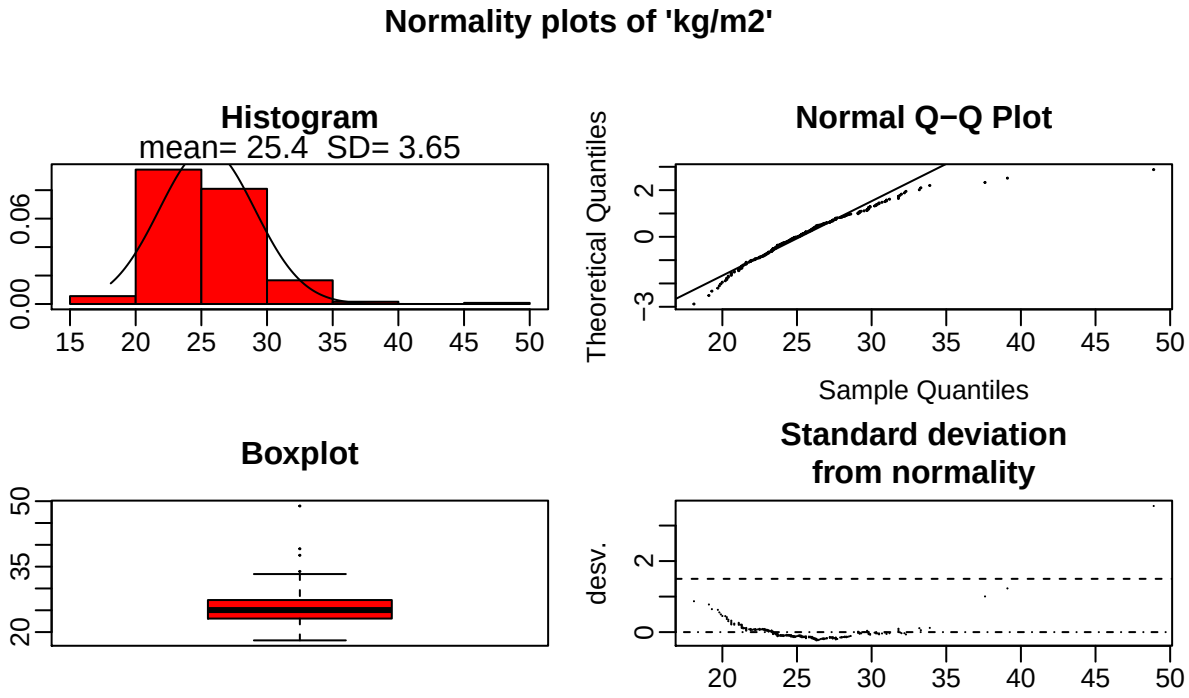
Shapiro−Wilks p−value: 0.001



Shapiro−Wilks p−value: <0.001

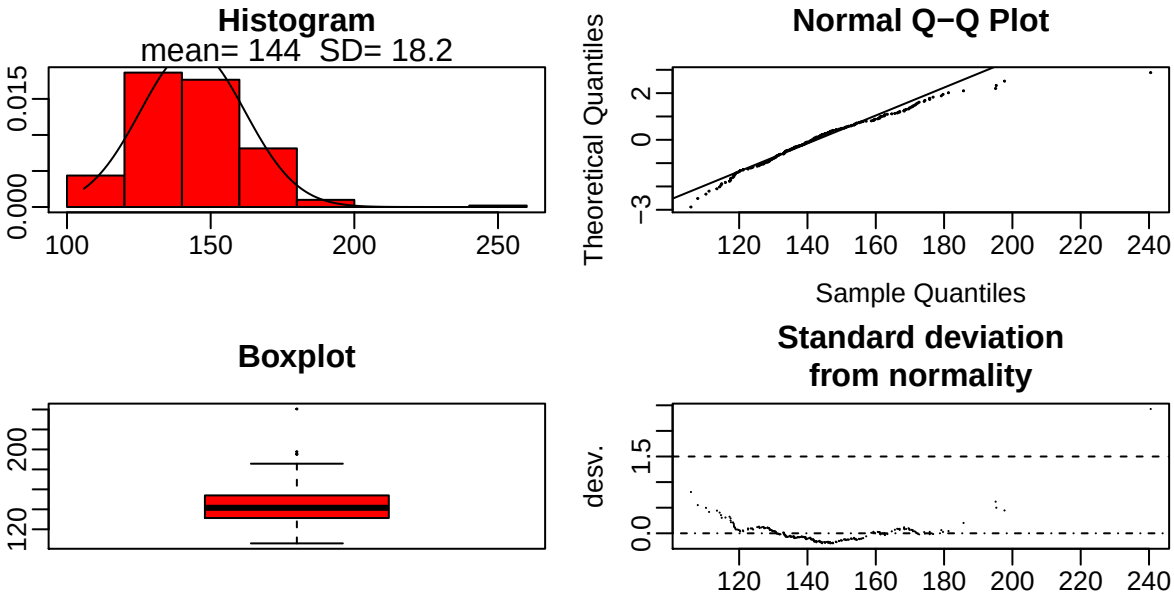


Shapiro−Wilks p−value: 0.237



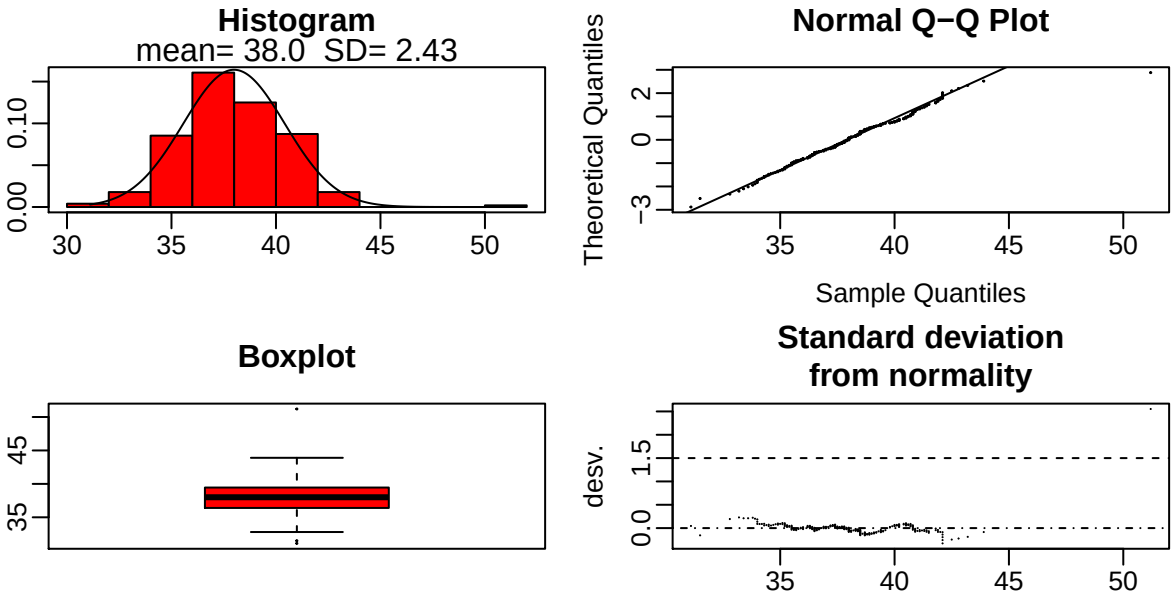
Shapiro−Wilks p−value: <0.001

Normality plots of 'peso da massa magra'



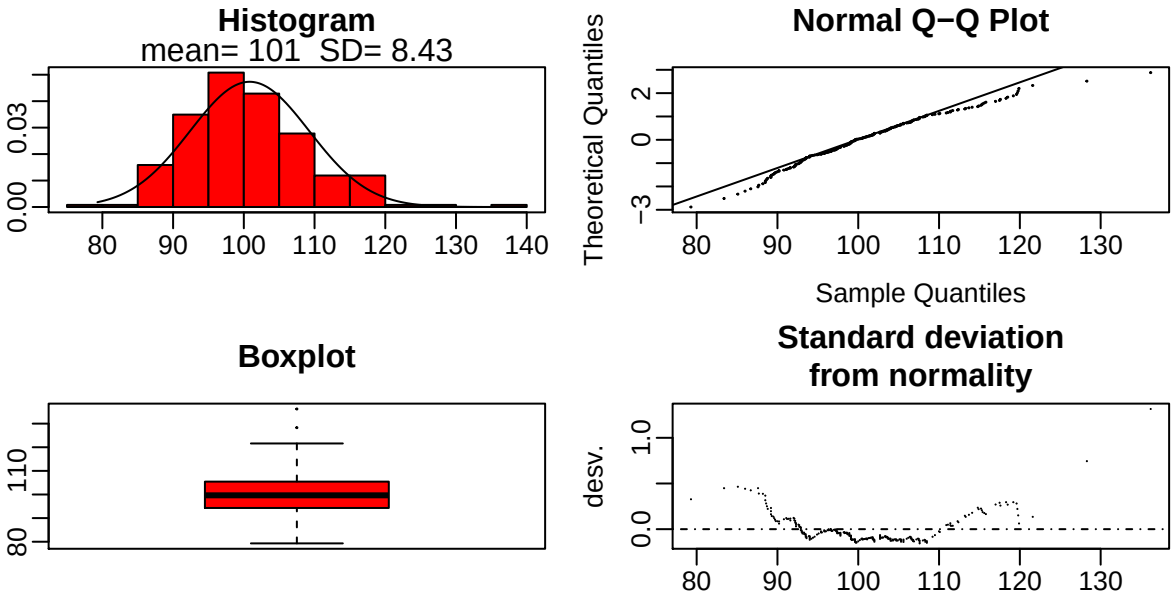
Shapiro−Wilks p−value: <0.001

Normality plots of 'Circunferencia do pescoço'



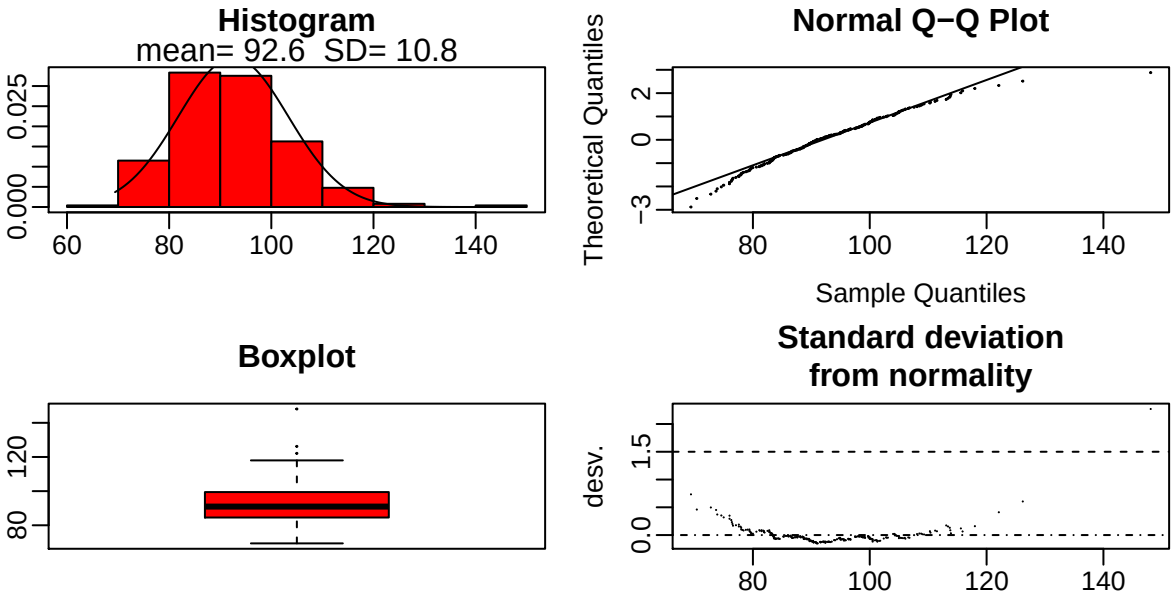
Shapiro−Wilks p−value: <0.001

Normality plots of 'Circunferencia do peito'

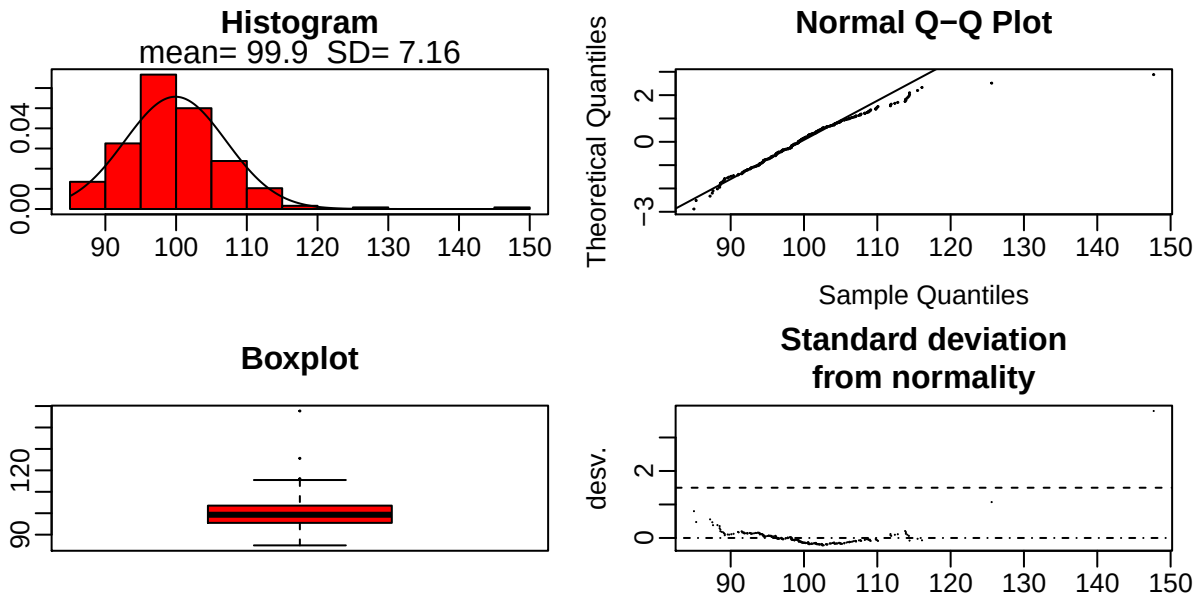


Shapiro−Wilks p−value: <0.001

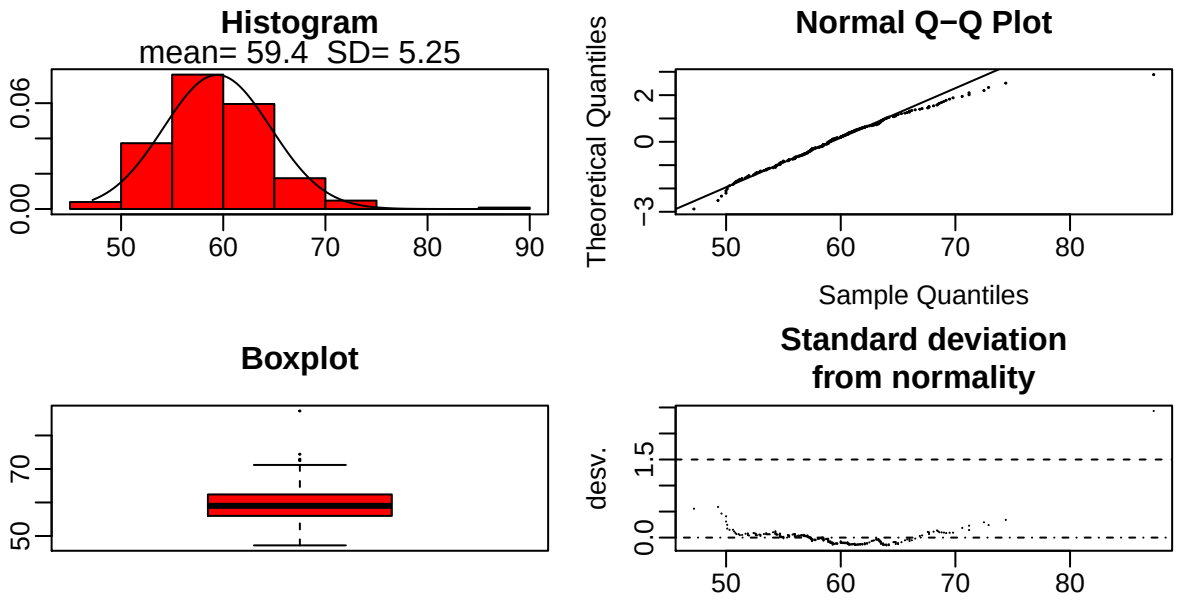
Normality plots of 'Circ do abdomem'



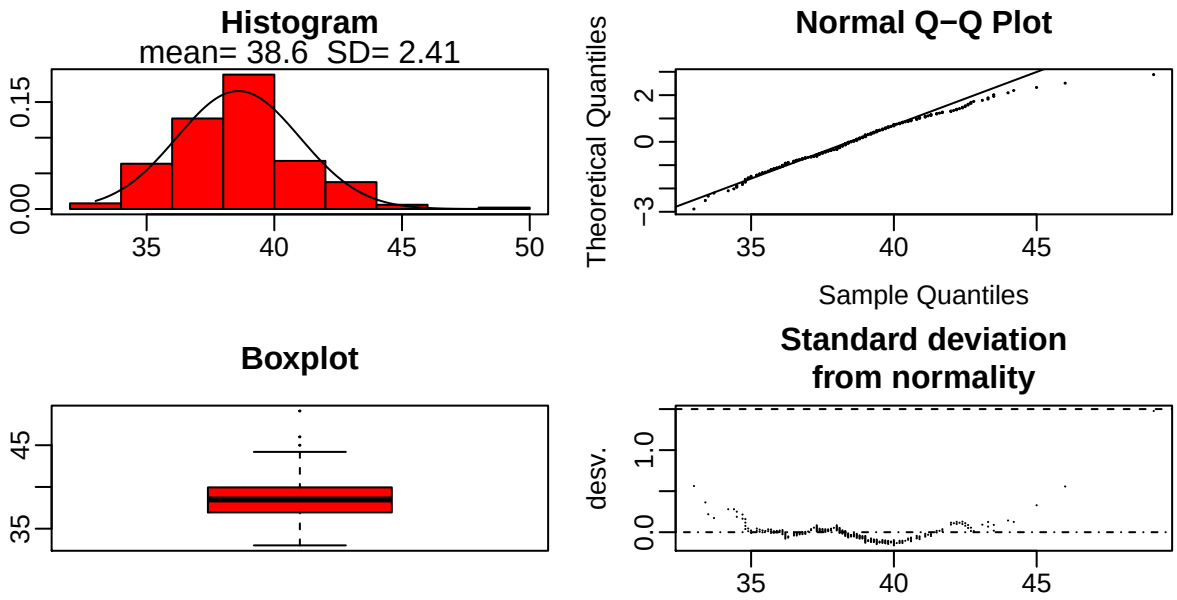
Shapiro−Wilks p−value: <0.001



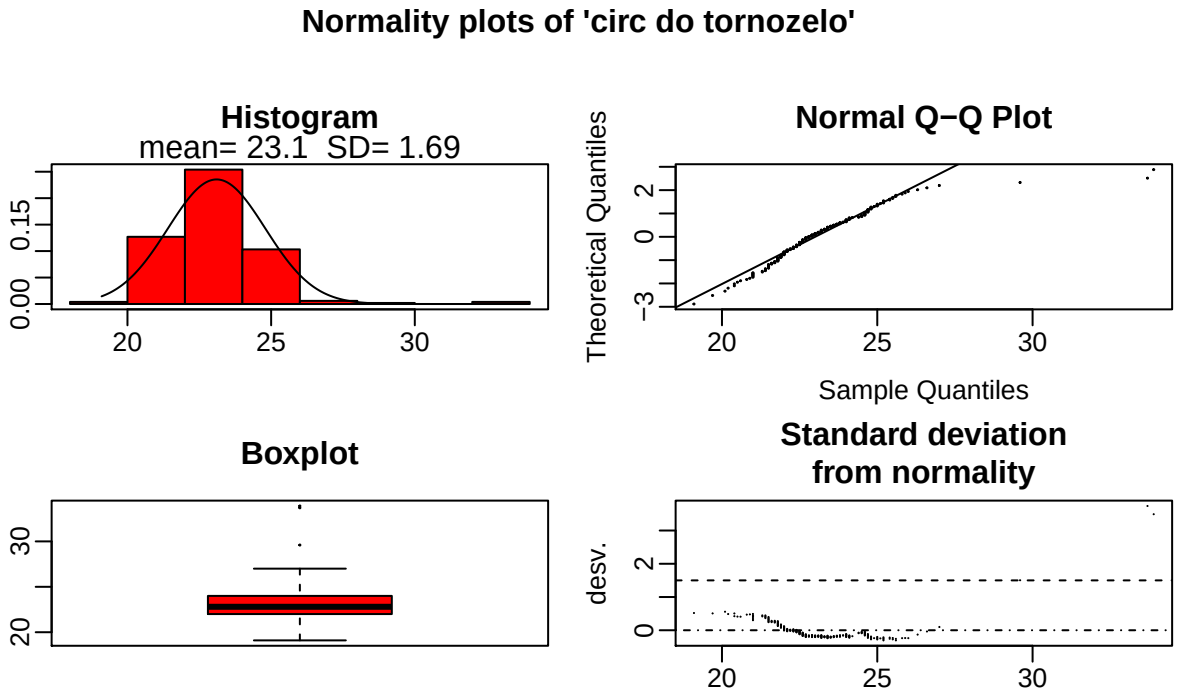
Shapiro−Wilks p−value: <0.001



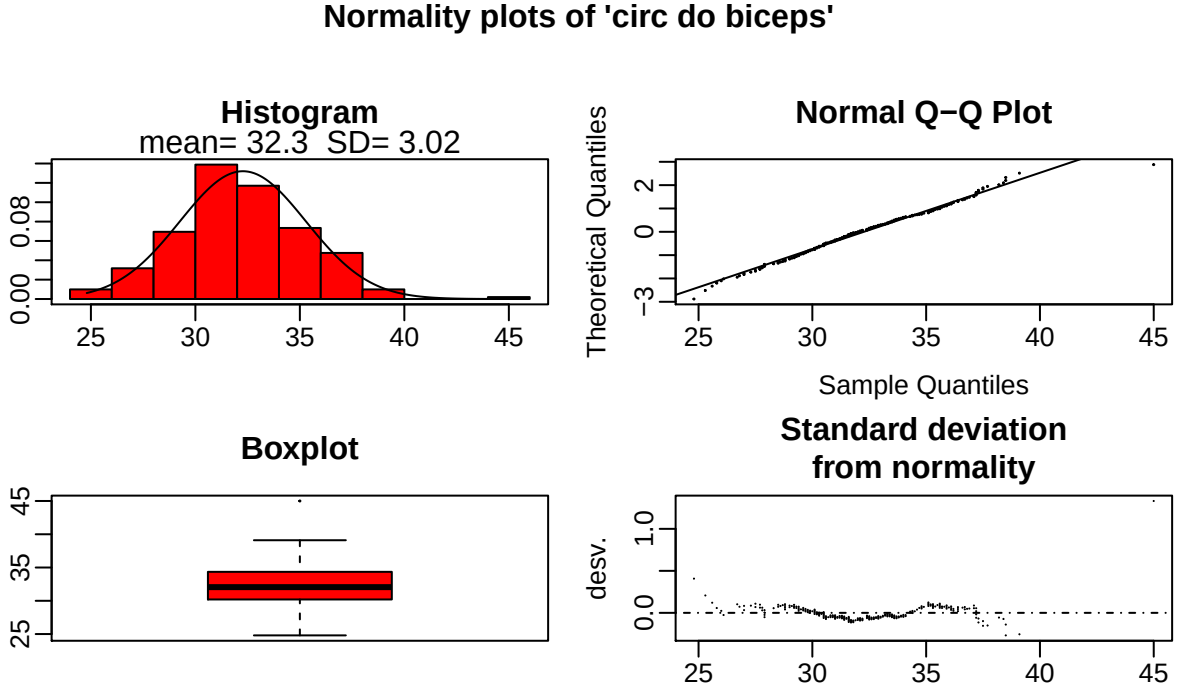
Shapiro−Wilks p−value: <0.001



Shapiro−Wilks p−value: 0.003

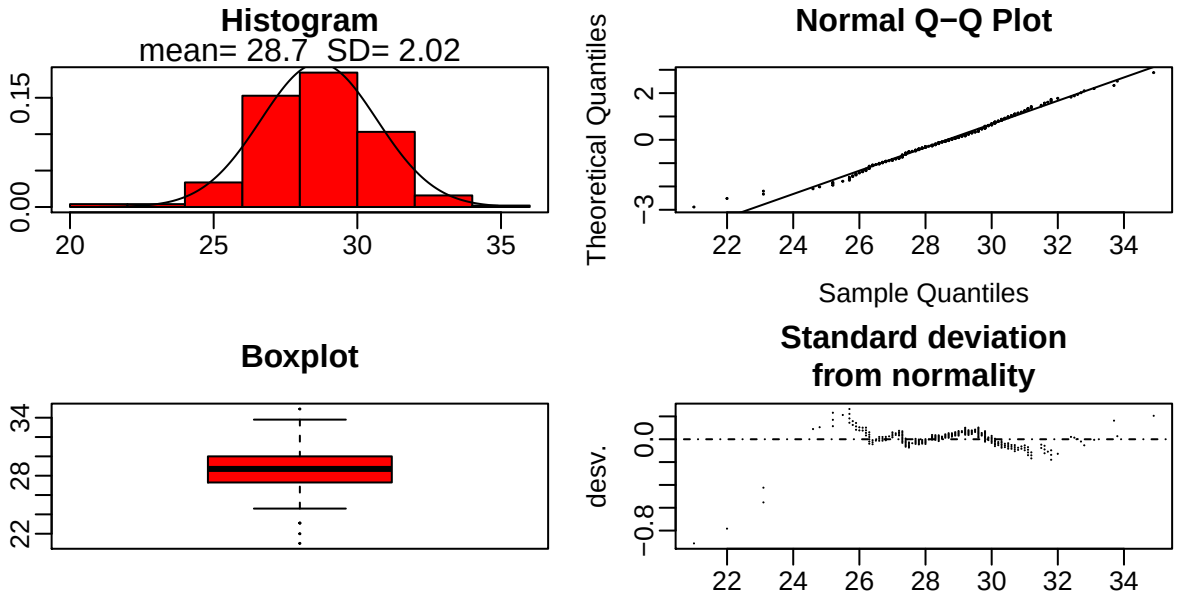


Shapiro−Wilks p−value: <0.001

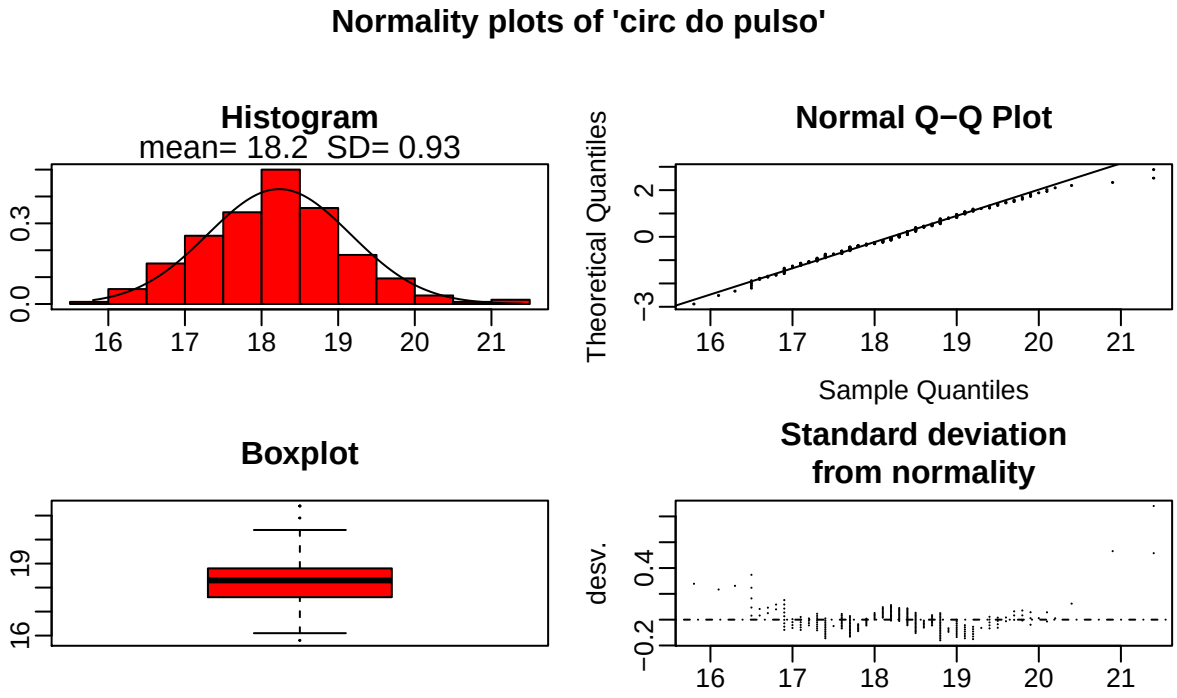


Shapiro−Wilks p−value: 0.046

Normality plots of 'circ do antebraco'



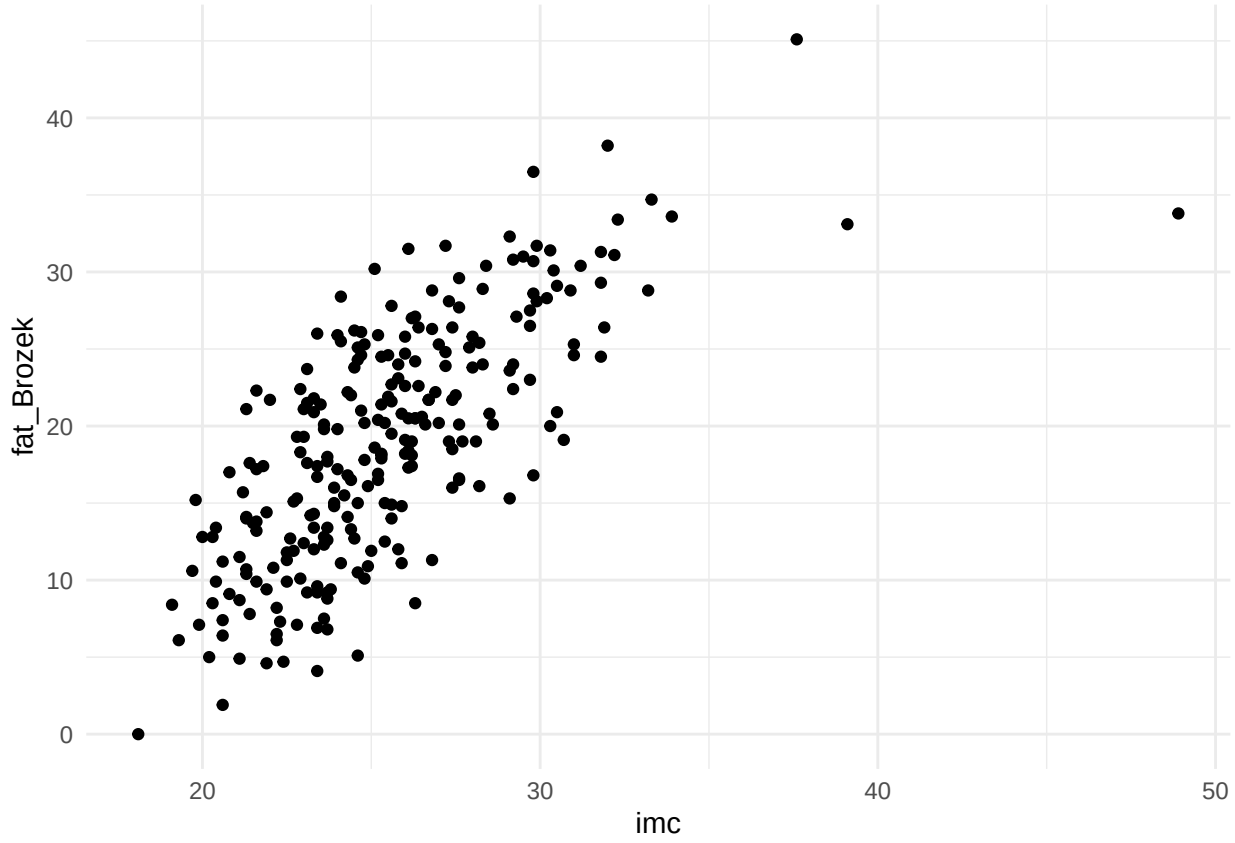
Shapiro−Wilks p−value: 0.048



Segundo passo: verificar linearidade com gráfico de dispersão. Identificar pontos

1) Gráfico de dispersão

```
disp_fat_Brozek<- ggplot(fat_dat, aes(x = imc, y = fat_Brozek))+
  geom_point() +
  theme_minimal()

disp_fat_Brozek
```



2) Identificando pontos outliers: outlier na horizontal

```
fat_dat %>%
  filter(imc>40)
```

```
## # A tibble: 1 x 19
##   numero fat_Brozek fat_Siri densidade idade peso_lbs altura_pol   imc
##    <dbl>      <dbl>    <dbl>     <dbl> <dbl>    <dbl>      <dbl> <dbl>
## 1     39       33.8     35.2      1.02    46     363.       72.2  48.9
## # ... with 11 more variables: FFW <dbl>, pescoco <dbl>, peito <dbl>,
## #   abdomen <dbl>, quadril <dbl>, coxa <dbl>, joelho <dbl>,
## #   tornozelo <dbl>, biceps <dbl>, antebraco <dbl>, pulso <dbl>
```

Terceiro passo: estimar modelo

```
lm_imc_brozek <- lm(formula = fat_Brozek ~ imc,
                    data = fat_dat)

summary(lm_imc_brozek)
```

```
##
## Call:
## lm(formula = fat_Brozek ~ imc, data = fat_dat)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4292  -3.4478   0.2113   3.8663  11.7826
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.40508    2.36723   -8.62 7.78e-16 ***
## imc           1.54671    0.09212   16.79  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.324 on 250 degrees of freedom
## Multiple R-squared:   0.53,  Adjusted R-squared:  0.5281
## F-statistic: 281.9 on 1 and 250 DF,  p-value: < 2.2e-16
```

Estimando intervalo de confiança

```
confint(lm_imc_brozek)
```

```
##                   2.5 %    97.5 %
## (Intercept) -25.067331 -15.74283
## imc           1.365275   1.72815
```

A cada aumento de 1 unidade no IMC ocorre aumento de 1.55% na gordura corporal pelo índice Brozek. O intervalo de confiança é de 1.36 a 1.73.

Estimando intervalo de predição

```
brozek_pred <- predict(lm_imc_brozek, interval = "predict")
```

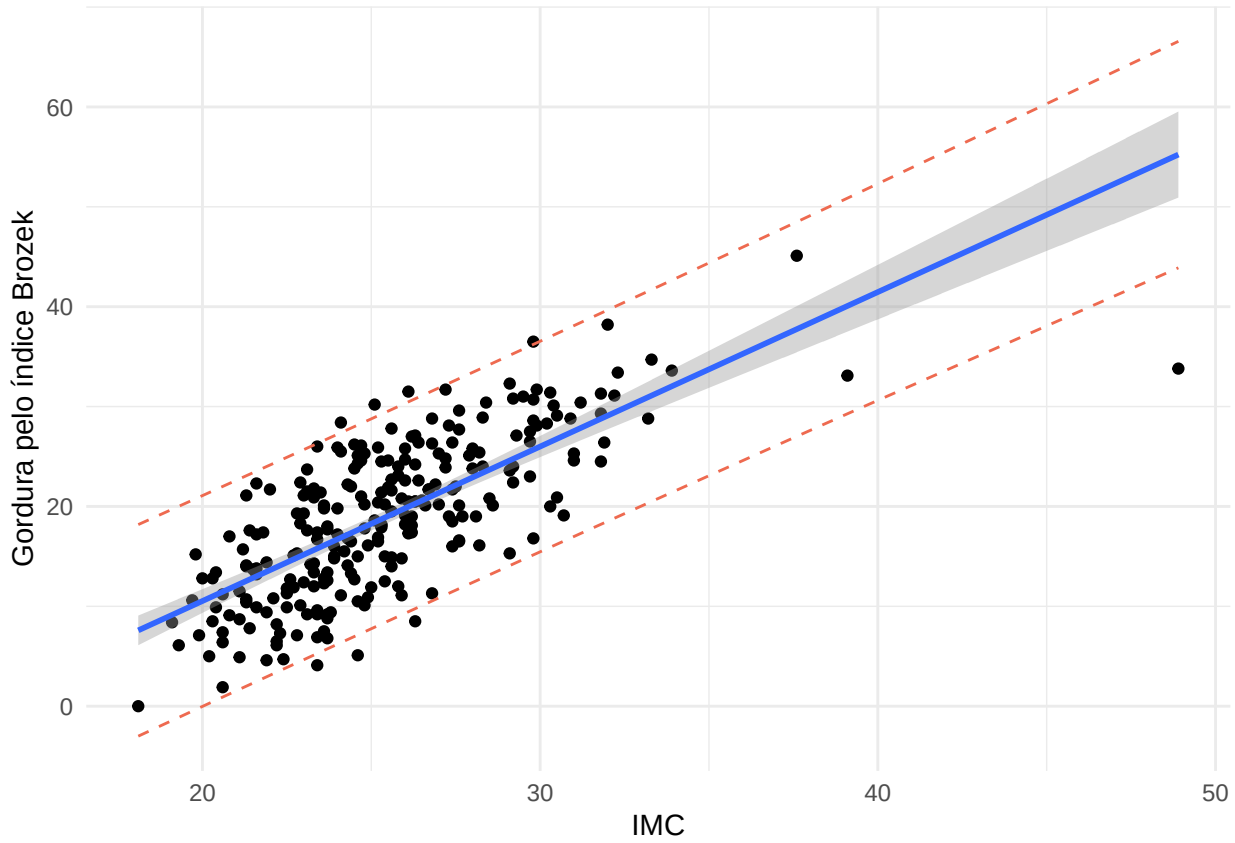
```
## Warning in predict.lm(lm_imc_brozek, interval = "predict"): predictions on current data refer to _fu
```

Incluindo o intervalo de predição para cada observação do banco

```
dat_brozek <-cbind(fat_dat, brozek_pred)
```

Gerando um gráfico com intervalo de confiança e intervalo de predição

```
plot_fat_Brozek <- ggplot(dat_brozek, aes(x = imc, y = fat_Brozek))+
  geom_point() +
  geom_smooth (method = "lm",)+
  scale_x_continuous("IMC") +
  scale_y_continuous("Gordura pelo índice Brozek") +
  theme_minimal() +
  geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") +   geom_line(aes(y = upr), col = "coral

plot_fat_Brozek
```



Gerando um modelo para predizer a gordura segundo Siri conforme o IMC

```
lm_imc_siri <- lm(formula = fat_Siri  ~ imc,
                  data = fat_dat)

summary(lm_imc_siri)
```

```
##
## Call:
## lm(formula = fat_Siri ~ imc, data = fat_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1070  -3.7418   0.2101   4.2070  12.8114
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.29940    2.55796  -9.109   <2e-16 ***
## imc           1.66884    0.09955  16.764   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.753 on 250 degrees of freedom
## Multiple R-squared:  0.5292, Adjusted R-squared:  0.5273
## F-statistic:   281 on 1 and 250 DF,  p-value: < 2.2e-16
```

Gerando intervalo de confiança para os parâmetros estimados

```
confint(lm_imc_siri)
```

```
##                   2.5 %     97.5 %
## (Intercept) -28.337296 -18.261512
## imc           1.472787   1.864899
```

Para o aumento de 1 ponto no IMC ocorre aumento de 1.7% (95%IC 1.5-1.86) na gordura corporal. O modelo explica 52% da variabilidade da gordura corporal.

Gerando um intervalo de predição para cada observação do banco na variável fat_siri

```
siri_pred <- predict(lm_imc_siri, interval = "predict")
```

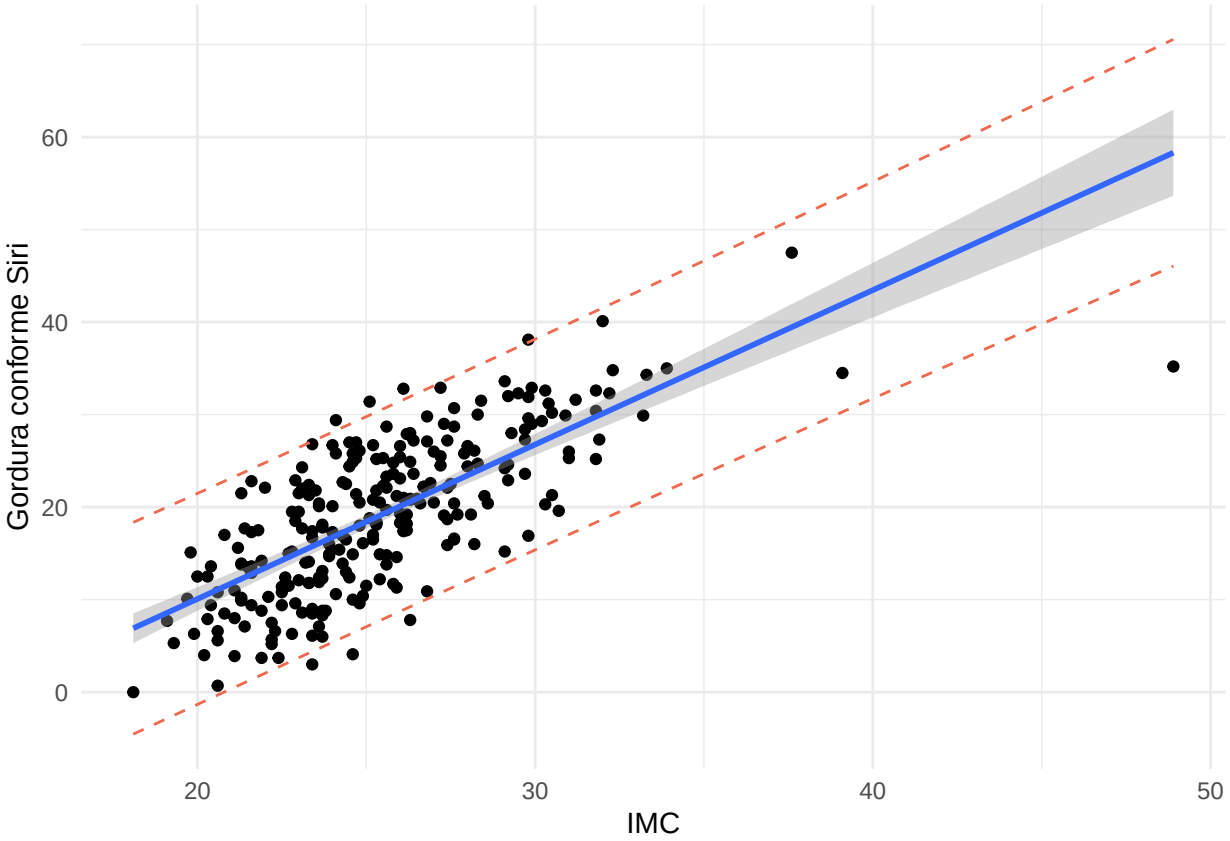
```
## Warning in predict.lm(lm_imc_siri, interval = "predict"): predictions on current data refer to _futu
```

Incluindo o intervalo de predição para cada observação do banco

```
dat_siri <-cbind(fat_dat, siri_pred)
```

Gerando gráfico com intervalo de confiança e intervalo de predição

```
plot_fat_siri <- ggplot(dat_siri, aes(x = imc, y = fat_Siri))+
  geom_point() +
  geom_smooth (method = "lm",)+
  scale_x_continuous("IMC")+
  scale_y_continuous("Gordura conforme Siri")+
  theme_minimal() +
  geom_line(aes(y = lwr), col = "coral2", linetype = "dashed") +   geom_line(aes(y = upr), col = "coral

plot_fat_siri
```



Anscombe dataset

Famoso banco de dados criado por Francis Anscombe com propriedades numéricas idênticas (estatísticas resumo e linhas de regressão), mas diferentes formas funcionais

Esse banco está disponível no R, basta chamá-lo

```
anscombe_dataset <- data("anscombe")
```

Estatísticas descritivas

```
compare_anscombe <- compareGroups(data = anscombe, ~ x1+
                                    x2+
                                    x3+
                                    x4)

createTable(compare_anscombe)
```

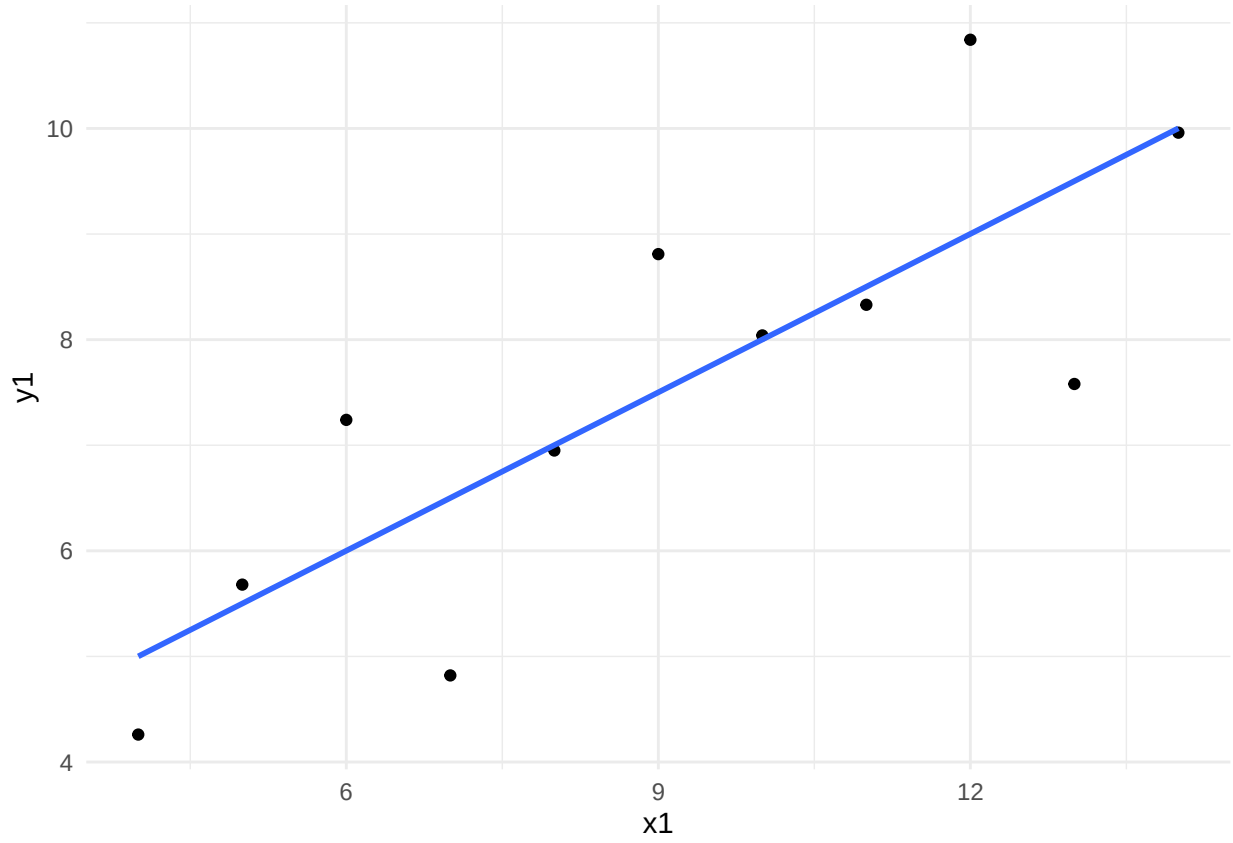
```
##
## --------Summary descriptives table ---------
##
## ------------------
##       [ALL]    N
##       N=11
```

```
## -----------------
## x1 9.00 (3.32) 11
## x2 9.00 (3.32) 11
## x3 9.00 (3.32) 11
## x4 9.00 (3.32) 11
## -----------------
```

Criando gráfico de dispersão para os 4 bancos dentro de Anscombe

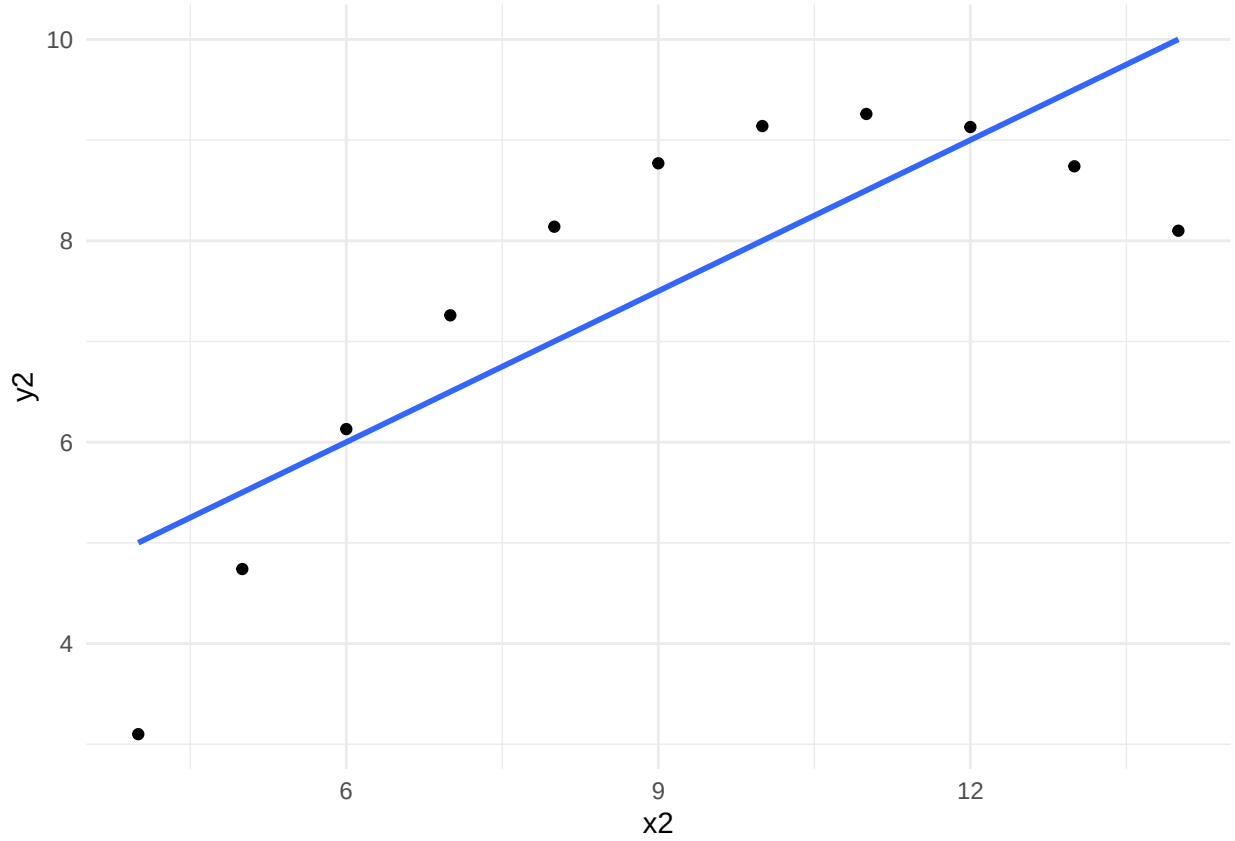
Banco 1

```
x1_scatter <- ggplot(anscombe, aes(x = x1, y = y1))+
  geom_point() +
  theme_minimal()+
  geom_smooth(method = "lm", se = FALSE)

x1_scatter
```



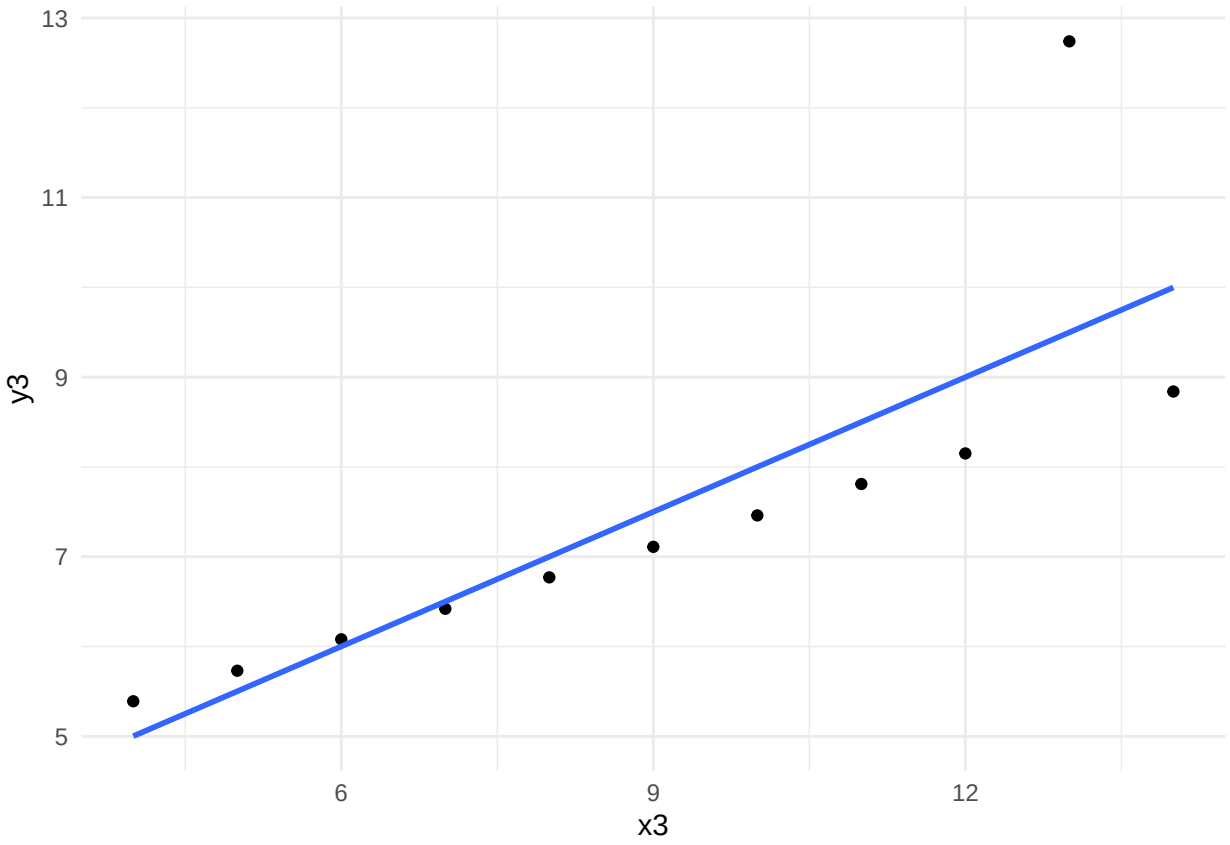
Banco 2

```
x2_scatter <- ggplot(anscombe, aes(x = x2, y = y2))+
  geom_point() +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE)

x2_scatter
```



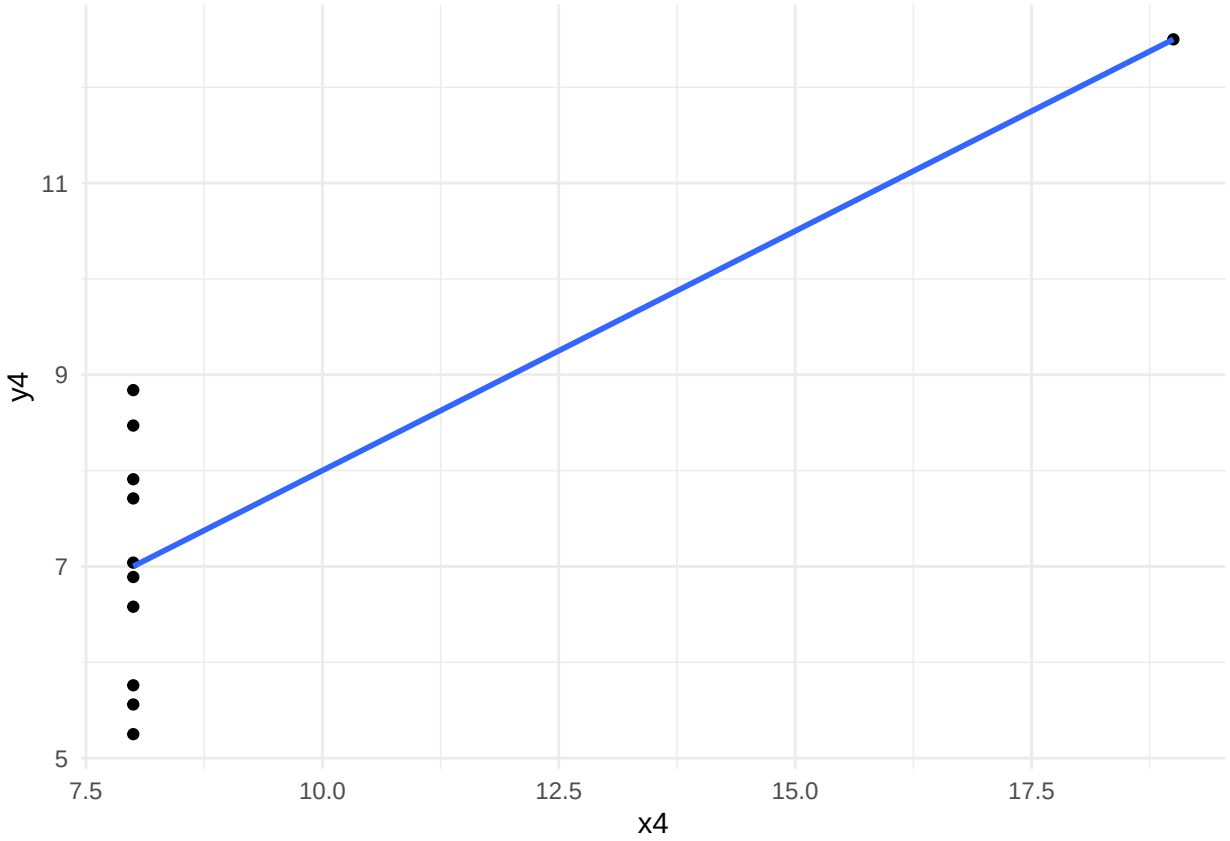
Banco 3

```
x3_scatter <- ggplot(anscombe, aes(x = x3, y = y3))+
  geom_point() +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE)

x3_scatter
```



Banco 4

```
x4_scatter <- ggplot(anscombe, aes(x = x4, y = y4))+
  geom_point() +
  theme_minimal() +
  geom_smooth(method = "lm", se = FALSE)

x4_scatter
```


11
9
7
5
y4
7.5
10.0
12.5
15.0
17.5
x4