Hands-on Biostatistics

Discovering relationships

18/7/2020

Install necessary packages

```
#install.packages(readr, ggpubr)

#To successfully install ggpubr you need to install Rtolls first from cran.
#If you don't succeed no worries,
#this package is needed only for a graphical display in correlation.
```

Chi-square test for independence

Hypothetical Example: Effectiveness of a Drug Treatment.

```
#read the data
library(readr)
treat.drug <- read_csv("C:/Users/user/Desktop/HOB(2020)/R/chi-square.csv",
    col_types = cols(id = col_number(), improvement = col_character(),
        treatment = col_character()))
```

- Organize the data in a crosstab

```
#Create crosstab
dt<- table(treat.drug$treatment, treat.drug$improvement)
dt #print the crosstab
```

```
##
##                improved not-improved
##   not-treated        26           29
##   treated            35           15
```

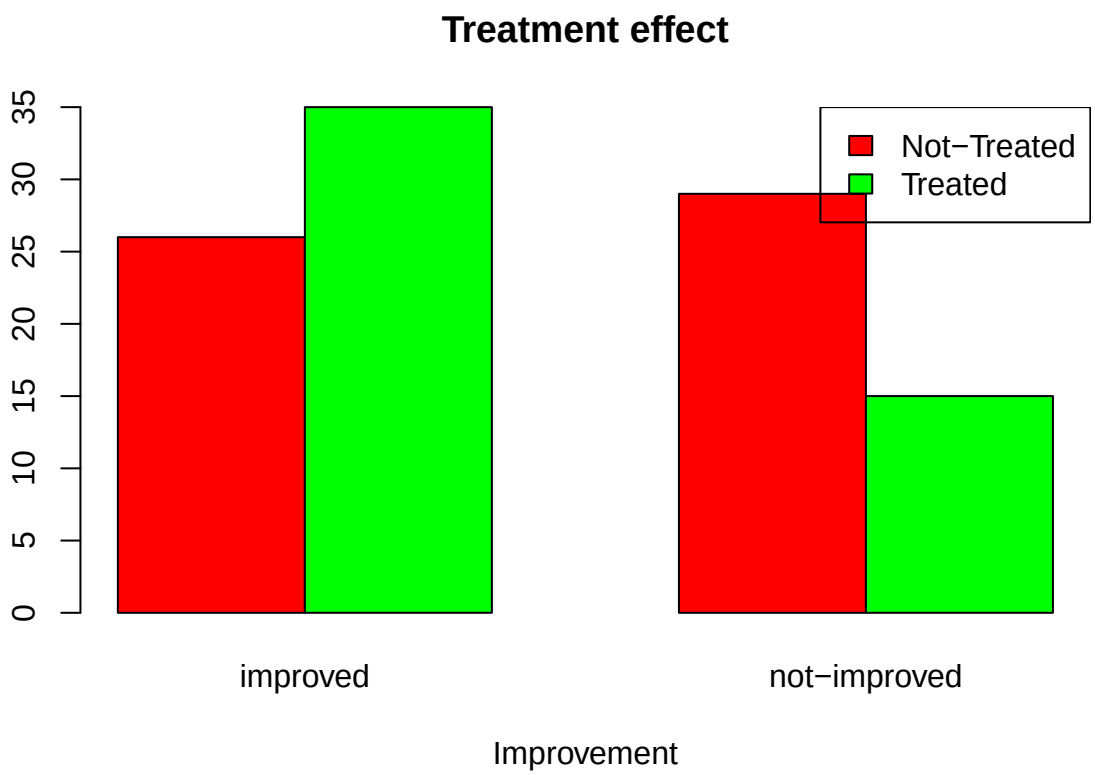
- From the crosstab we see that there are two categorical variables with two levels each. We have treatment, treated and not-treated, and improvement of health: improved and not-improved. We want to see if treatment is related to improvement of health.

Visualize the data

```
# Plot the data

barplot(dt,
main = "Treatment effect",
xlab = "Improvement", beside = TRUE,
col = c("red","green")
)
legend("topright",
c("Not-Treated","Treated"),
```

```
fill = c("red","green")
)
```



Conduct the chi-square test for association/independence

```
# Chi-sq test
chisq <- chisq.test(dt, correct=FALSE)
chisq
```

```
##
##  Pearson's Chi-squared test
##
## data:  dt
## X-squared = 5.5569, df = 1, p-value = 0.01841
```

We have a chi-squared value of 5.55. Since we get a p-Value less than the significance level of 0.05, we can reject the null hypothesis and conclude that the two variables are in fact dependent.

Check assumptions

```
# Expected counts
round(chisq$expected,2)
```

```
##
##               improved not-improved
##   not-treated    31.95        23.05
```

```
##   treated        29.05        20.95
```

Documentation

A chi-square test for association was conducted between treatment and health improvement. All expected cell frequencies were greater than five. There was a statistically significant association between treatment and health improvement, ??2(1) = 5.56, p = 0.018.

Fisher's exact test

The function fisher.test is used to perform Fisher's exact test when the sample size is small. We will run the Fisher's exact test in the same data as in chi-square to see if there is a difference.

```
fisher <- fisher.test(dt)
fisher
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dt
## p-value = 0.02889
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1582959 0.9212947
## sample estimates:
## odds ratio
##  0.3878601
```

From the output we see that the p-value is less than the significance level of 0.05, we can reject the null hypothesis. In our context, rejecting the null hypothesis for the Fisher's exact test of independence means that there is a significant relationship between the two categorical variables (treatment and improvement).

Simple linear regression

Linear regression it is used when we want to predict the value of a dependent/outcome variable based on the value of another independent/predictor variable. In our example, we will use linear regression to understand whether height (in inches) can be predicted based on age.

Read the data

```
library(readxl)
ageandheight <- read_excel("ageandheight.xls", sheet = "Hoja2") #Upload the data

print(ageandheight)
```

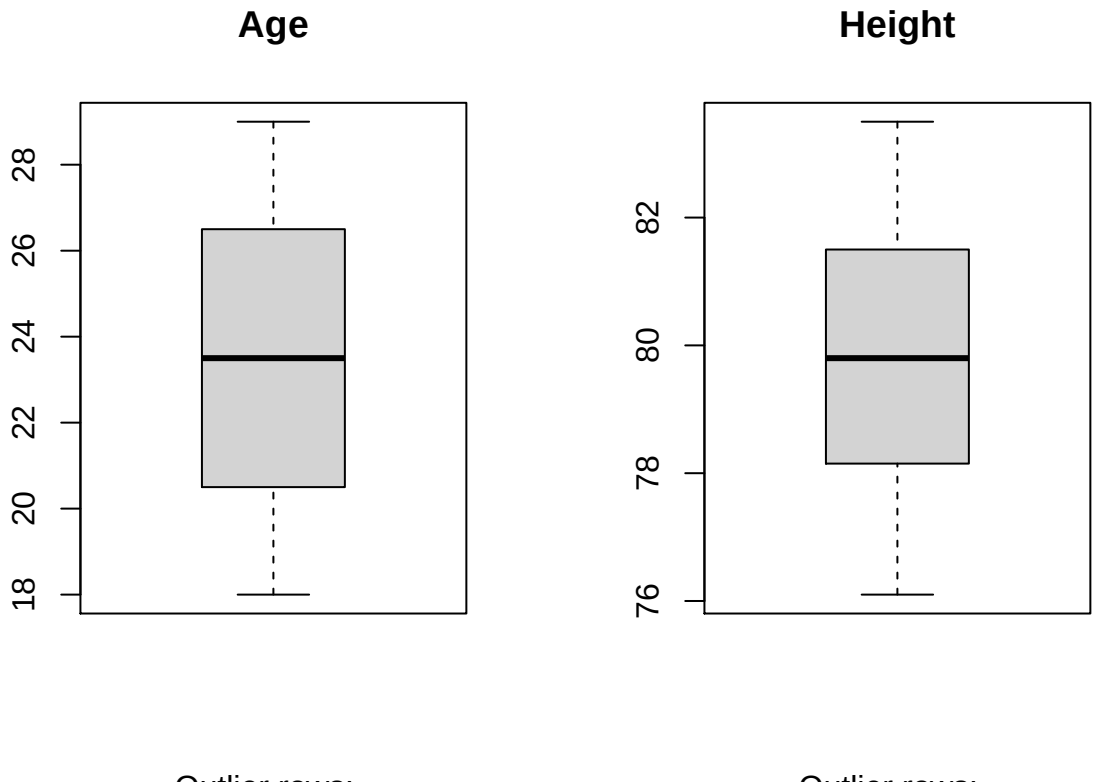
```
## # A tibble: 12 x 2
##      age height
##    <dbl>  <dbl>
##  1    18   76.1
##  2    19   77
##  3    20   78.1
##  4    21   78.2
##  5    22   78.8
##  6    23   79.7
##  7    24   79.9
##  8    25   81.1
```

```
##  9    26   81.2
## 10    27   81.8
## 11    28   82.8
## 12    29   83.5
```

Visualize the data

Boxplots to check for outliers

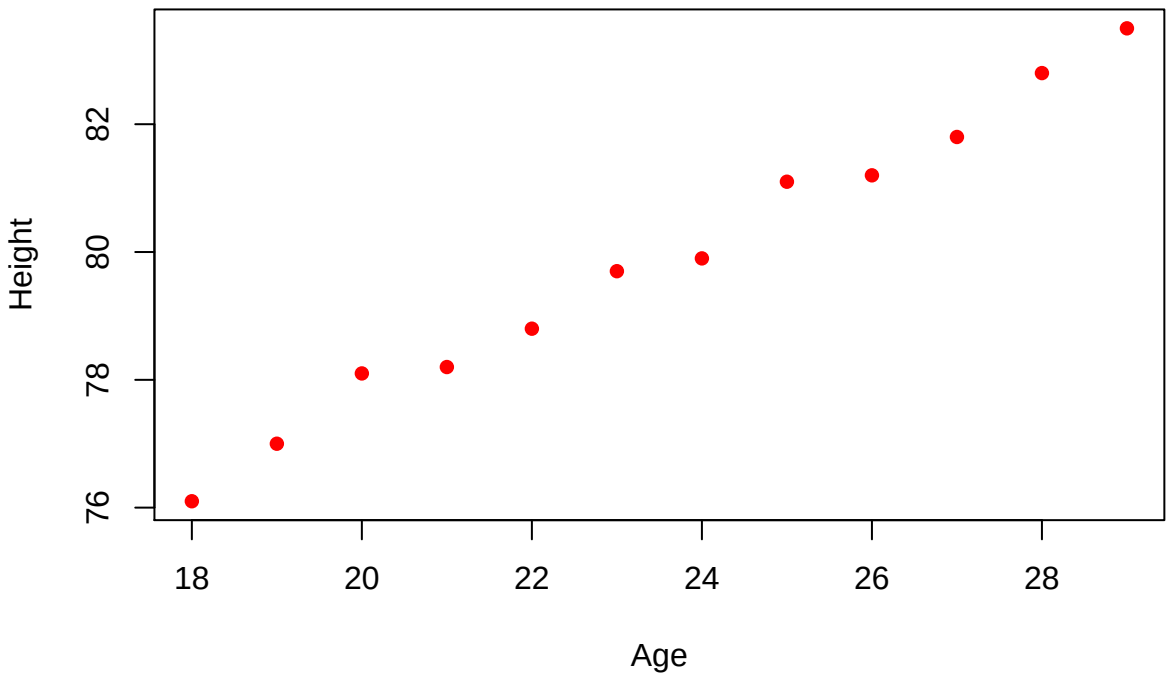
```
par(mfrow=c(1, 2))  # divide graph area in 2 columns
boxplot(ageandheight$age, main="Age", sub=paste("Outlier rows: ", boxplot.stats(ageandheight$age)$out))
boxplot(ageandheight$height, main="Height", sub=paste("Outlier rows: ", boxplot.stats(ageandheight$heigh
```



- There are no outliers in the data.

Scatterplot to see if there is a trend for linear relationship

```
plot(ageandheight$height~ageandheight$age, xlab="Age", ylab="Height", pch = 16, col = "red")
```



Perform linear regression

```
lmHeight = lm(height~age, data = ageandheight) #Create the linear regression

summary(lmHeight) #Review the results
```

```
##
## Call:
## lm(formula = height ~ age, data = ageandheight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27238 -0.24248 -0.02762  0.16014  0.47238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
## age           0.6350     0.0214   29.66 4.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.256 on 10 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9876
## F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

- Interpretation of the coefficients

Documentation

A linear regression established age could significantly predict height, F(1,10) =880, p < 0.0001 and age accounted for 98.8 % (R squre=0.988) of the explained variability in height. The regression equation was: predicted height = 64.92 + (0.635 * age).

Note:

In the output, you can see the values of the intercept (alfa value) and the slope (beta value) for the age. If there is a child that is 20.5 months old, alfa is 64.92 and b is 0.635, the model predicts (on average) that its height in centimeters is around 64.92 + (0.635 * 20.5) = 77.93 cm.

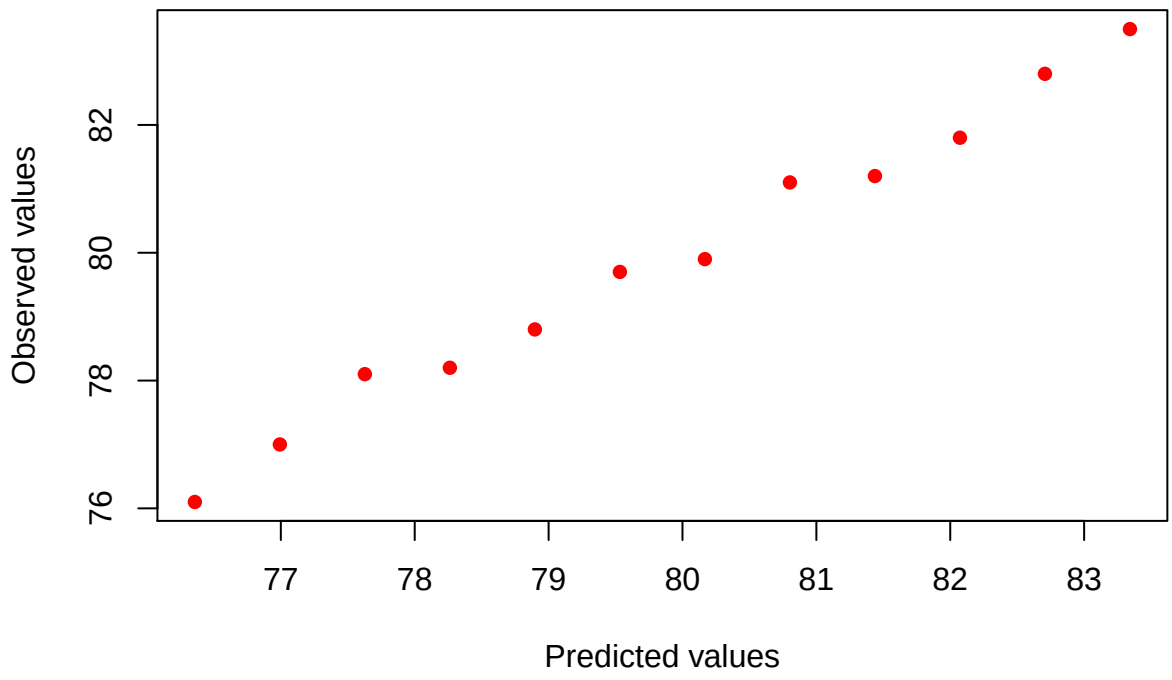
Extract fitted/predicted values

```
#Extract fitted values

lmHeight$fitted.values
```

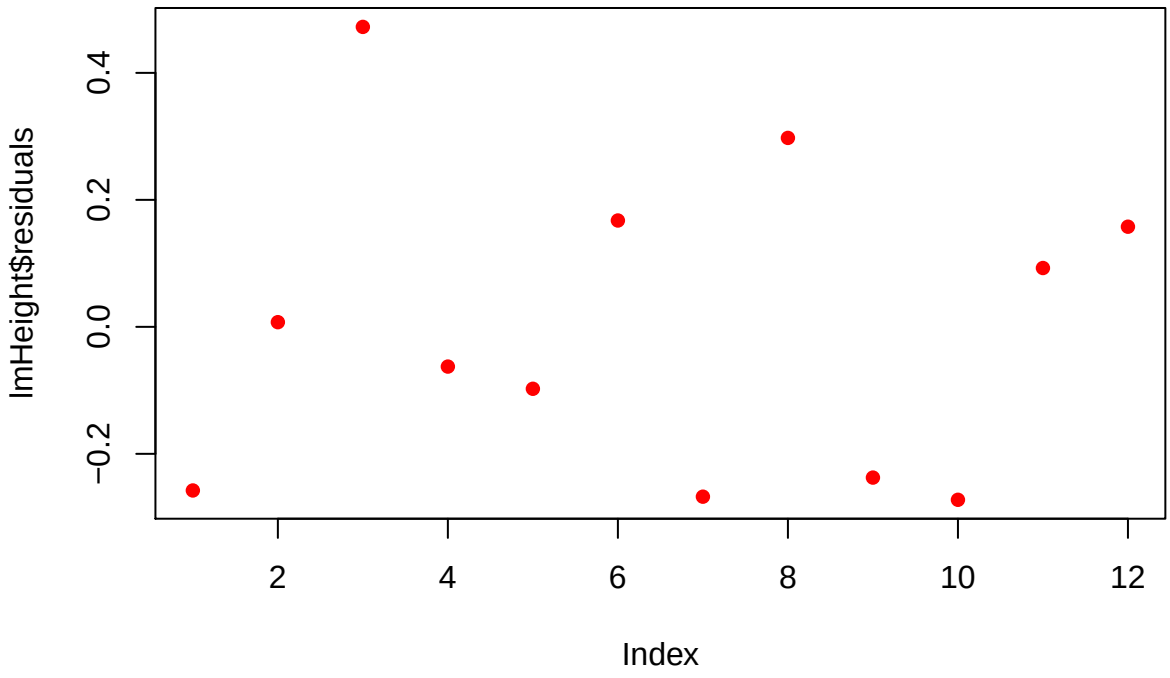
```
##        1        2        3        4        5        6        7        8
## 76.35769 76.99266 77.62762 78.26259 78.89755 79.53252 80.16748 80.80245
##        9       10       11       12
## 81.43741 82.07238 82.70734 83.34231
```

```
plot(ageandheight$height~lmHeight$fitted.values, xlab="Predicted values", ylab="Observed values", pch =
```

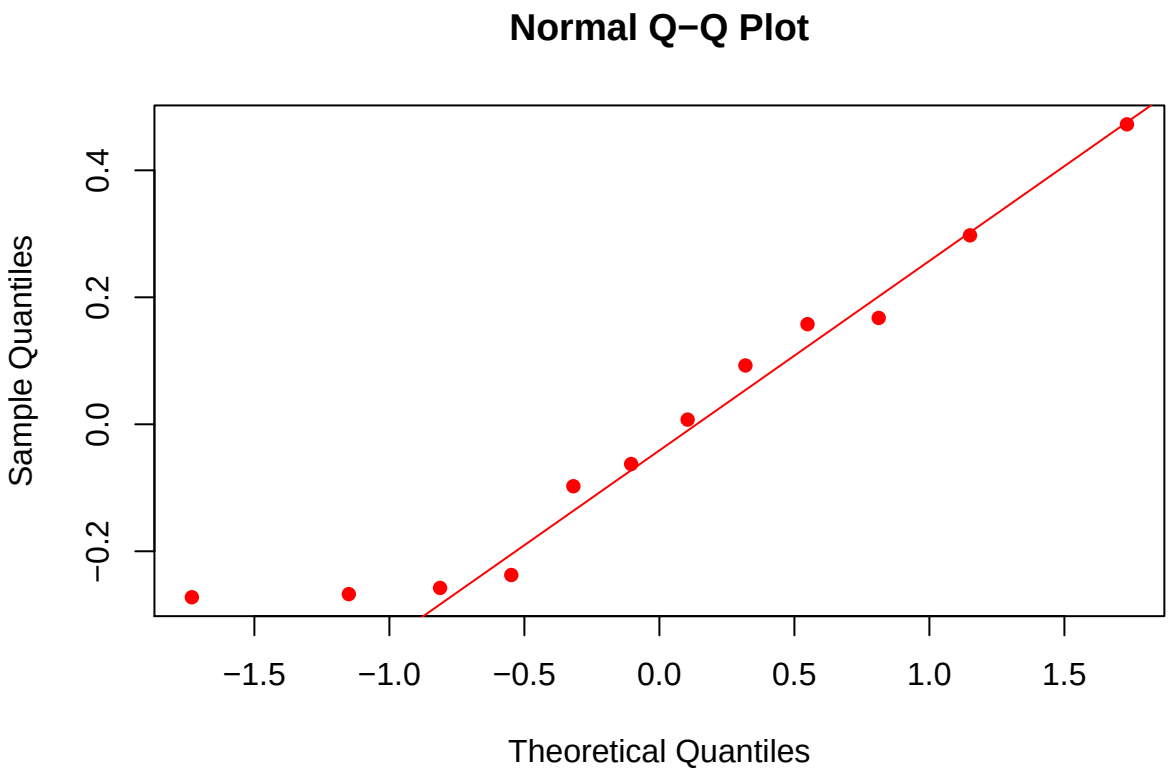


Model fit diagnostic

```
plot(lmHeight$residuals, pch = 16, col = "red")
```




```
qqnorm(lmHeight$residuals, pch = 16, col = "red")
qqline(lmHeight$residuals, pch = 16, col = "red")
```



You don't see any clear patterns on your residuals, which is good!

Correlation

The Pearson product-moment correlation coefficient (Pearson's correlation, for short) is used to measure the strength and direction of association between two continuous variables.

In our example, we could use a Pearson's correlation to understand whether there is an association between age and height.

Test for normality

```
shapiro.test(ageandheight$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ageandheight$age
## W = 0.9669, p-value = 0.8757
```

```
shapiro.test(ageandheight$height)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ageandheight$height
## W = 0.97576, p-value = 0.9609
```

COmpute the Pearson correlation coefficient

```
cor.test(ageandheight$age,ageandheight$height, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  ageandheight$age and ageandheight$height
## t = 29.665, df = 10, p-value = 4.428e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9793465 0.9984716
## sample estimates:
##       cor
## 0.9943661
```

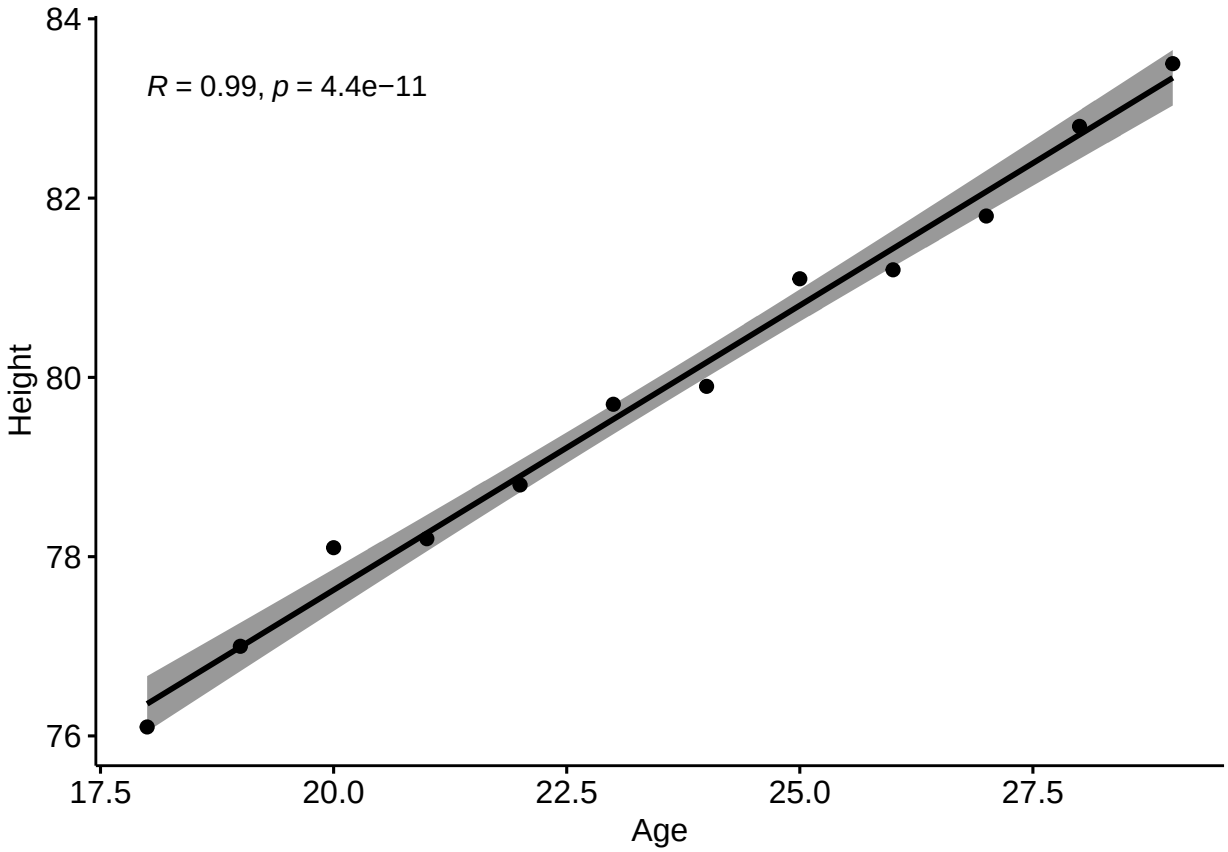
Visualize your results with 95% CI included

```
library("ggpubr")
```

```
## Loading required package: ggplot2
```

```
ggscatter(ageandheight, x = "age", y = "height",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Age", ylab = "Height")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Documentation

There was a strong positive correlation between age and height, r = 0.99. An increase in age led to an increase in height, r(10) = 0.99, p < 0.0001.

References

https://www.r-bloggers.com/chi-squared-test/

https://cran.r-project.org/web/packages

https://towardsdatascience.com/fishers-exact-test-in-r-independence-test-for-a-small-sample-56965db48e87

https://www.datacamp.com/community/tutorials/linear-regression

http://www.sthda.com/english/wiki/correlation-test-between-two-variables

https://statistics.laerd.com/