Correlation and Linear Regressions

sn

2024-05-22

Linear Regression

A simple linear regression describes the association between an independent variable and a dependent one

- How to find such "good" straight line?

We estimate b0 and b1 by minimizing the sum of the squared differences between the observed and the predicted values of the outcome.

- *Goodness of Fit:* R^2 = 1 - [sum(residuals)^2 / sum(yo-ye)^2]

Represents the proportion of the variance of the dependent variable that it is explained by the independent variable in the regression model.

- summary(lm(y~x))

Regression and correlation

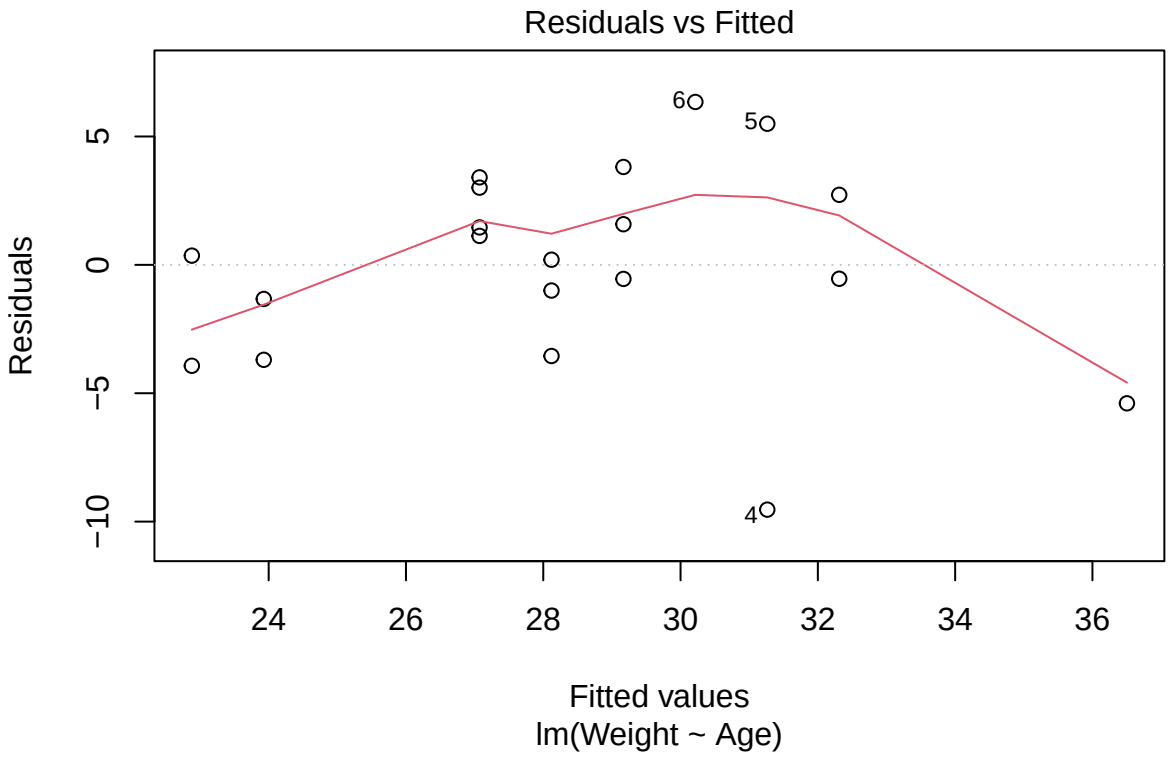
Formulating a linear model and exploring it

```
model_mice <- lm(Weight~Age, data = WT)
```

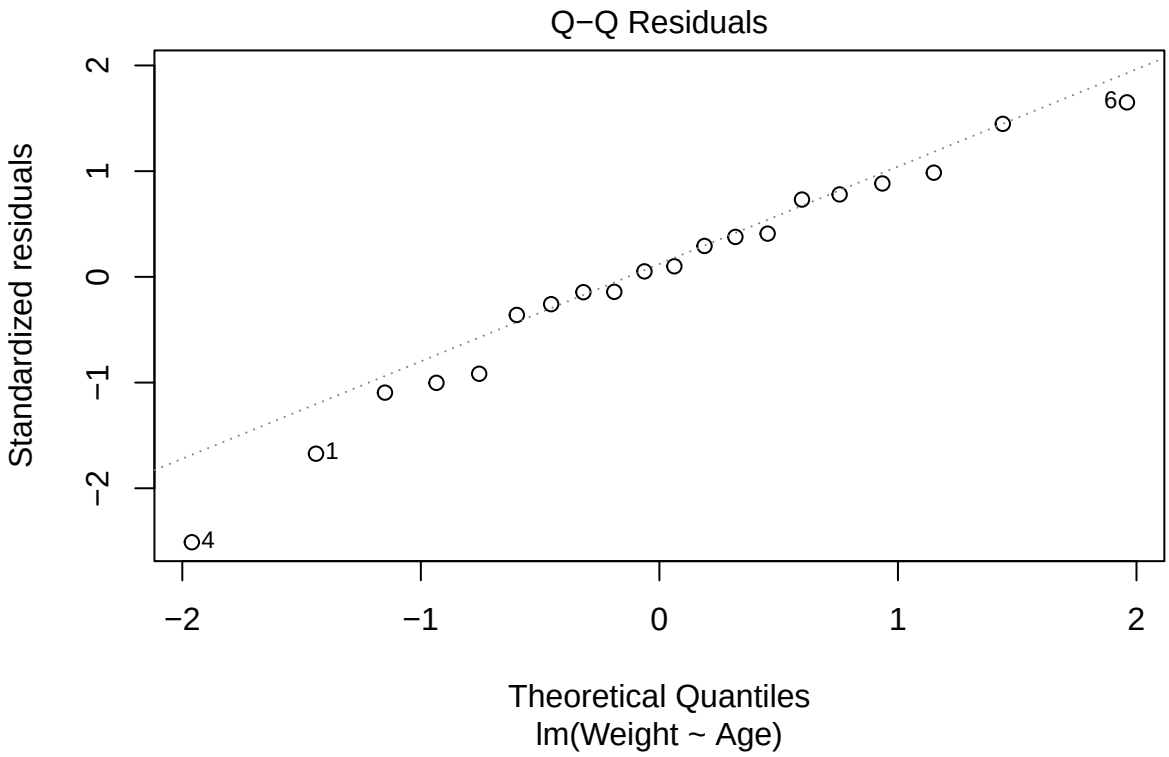
Linear regression assumptions:

- Independence of observations;
- Linear relationship between variables;
- Homoscedacity of residuals;
- Normal distribution of residuals.

```
plot(model_mice, 1) # To test homoscedacity
```




```
plot(model_mice, 2) # To test the normality of residuals hist(resid(model_mice)) # To test the normalit
```




```
# linear model result
summary(model_mice)
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = WT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5335 -1.8841  0.2795  2.7984  6.3443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.594      3.241   5.120 7.16e-05 ***
## Age            1.048      0.277   3.782  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.974 on 18 degrees of freedom
## Multiple R-squared:  0.4428, Adjusted R-squared:  0.4119
## F-statistic: 14.31 on 1 and 18 DF,  p-value: 0.001364
```

```
# predictions
new <- data.frame(Age = c(3, 20))
predict.lm(object = model_mice, newdata = new)
```

```
##        1        2
## 19.73744 37.55047
```

```r
#  Test the correlation between variables
cor.test(WT$Age, WT$Weight, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  WT$Age and WT$Weight
## t = 3.7823, df = 18, p-value = 0.001364
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3159815 0.8559241
## sample estimates:
##       cor
## 0.6654531
```