Correlation, Regression

Veres Dániel Sándor

2019

Settings

Used software: R

Dataset

Using Framingham dataset

```
setwd("C:/pendrivok/oktatos/oktatas_2019tavasz/nemet_stat")
frmgham2_vds <-
read.csv("C:/pendrivok/oktatos/oktatas_2019tavasz/nemet_stat/frmgham2_vds.csv
",
    sep = ";")
fr <- frmgham2_vds[frmgham2_vds$PERIOD == 1, ]
```

Documentation

I think this is one of the most important thing in a statistical analysis…

Important: write it for yourself and others - be able to reproduce and understand!

- title
- name
- date
- source of dataset

Why?

Questions:

- Is there any relation, connection… between variables?
- If there is, how we can describe it?
- How to estimate the value of a variable based on the value of other variable(s)?
- Which variables are the most important?
- ….

Answer: using ***correlation***, ***regression***

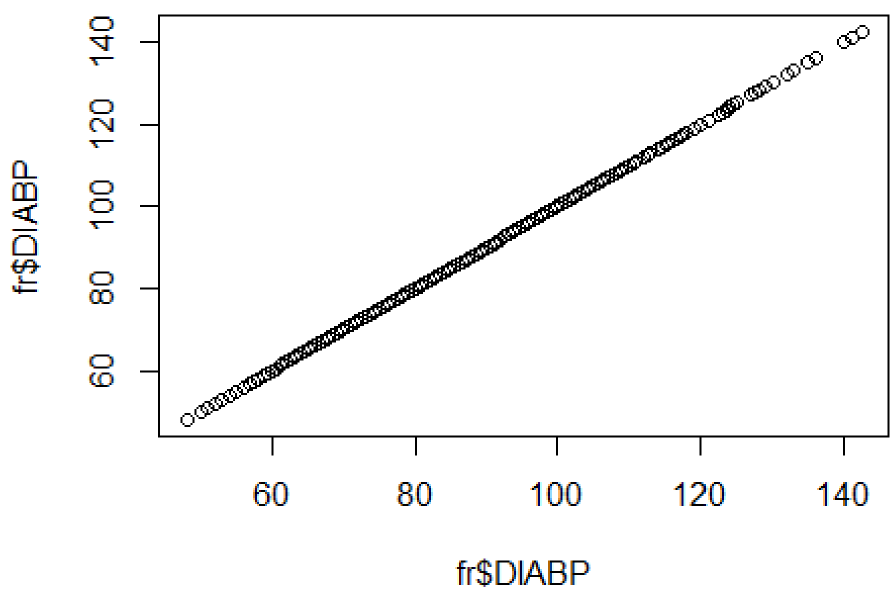
Correlation

symmetric relation of **2**, **random** variable,

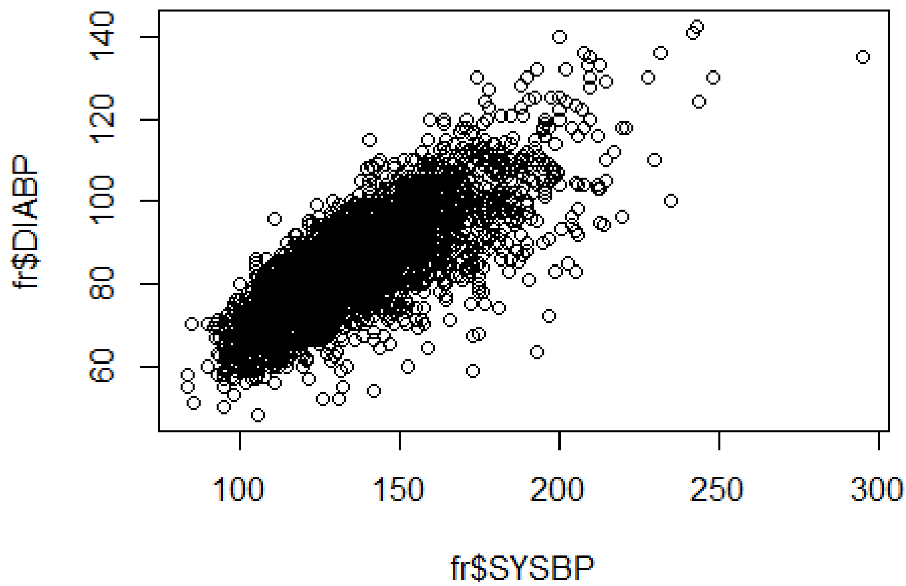
type of *relation*:

- monotonic
  - positive
  - negative
  - linear positive
  - ...
- not monotonic
  - parabolic
  - ...
- no relation

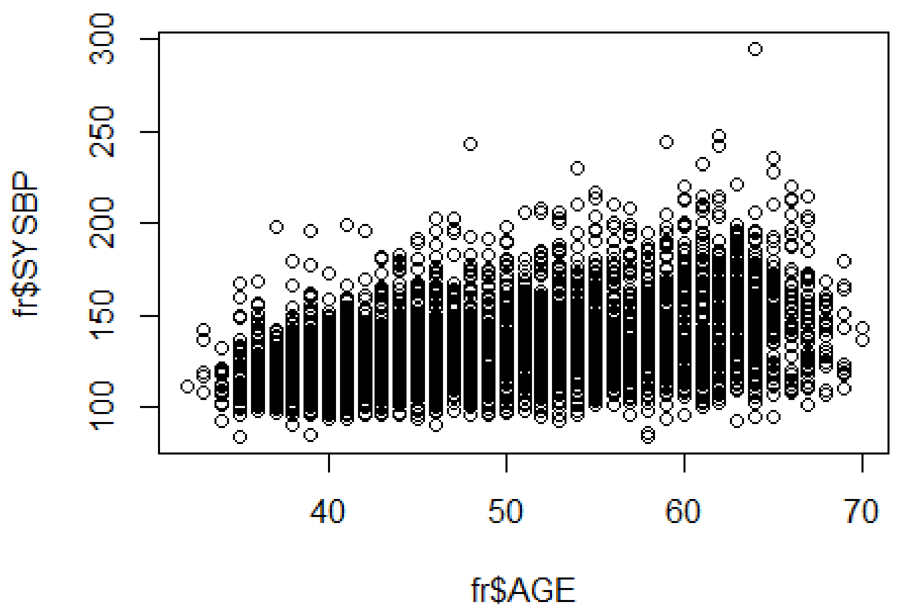
```
plot(fr$DIABP, fr$DIABP)
```



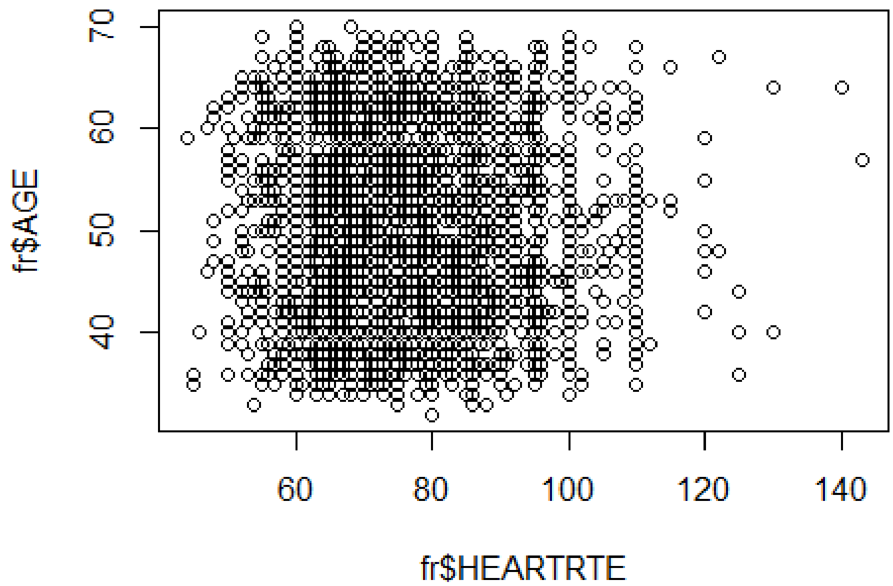

```
plot(fr$SYSBP, fr$DIABP)
```




```
plot(fr$AGE, fr$SYSBP)
```

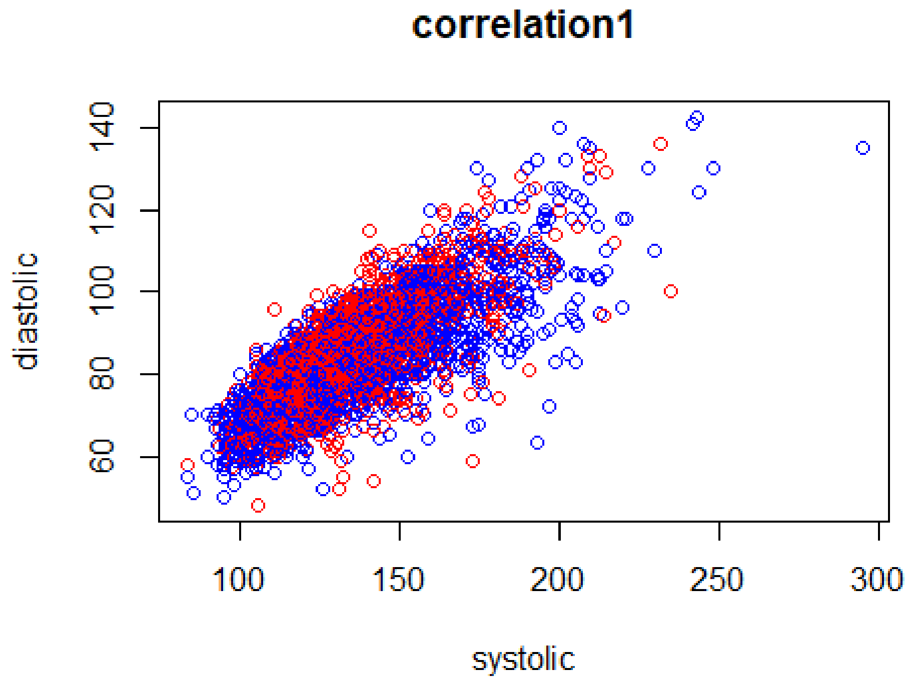



```
plot(fr$HEARTRTE, fr$AGE)
```



correlation: if **monotonic relation** (positive or negative….)

```
plot(fr$SYSBP, fr$DIABP, xlab = "systolic", ylab = "diastolic",
    main = "correlation1", col = c("blue", "red")[fr$SEX])
```



“Strength” of the correlation

Measures, usually:

- Pearson r: if linear
- Spearman rho: monotonic (not necessary linear) - Pearson for ranks
- (Kendall tau: monotonic (not necessary linear) - same “weights” for observations)
- ...

Values: between -1 and 1

Calculation: Measures the distance from the “middle” (check google for formula… :)

```
cor(fr$SYSBP, fr$DIABP, method = "pearson")

[1] 0,7842

cor(fr$SYSBP, fr$DIABP, method = "spearman")

[1] 0,7762
```

Hypothesis test: Assumptions: independent H0: R = 0 Statistics: $t = \sqrt{r^2 * \frac{n-2}{1-r^2}}$

```
cor.test(fr$SYSBP, fr$DIABP, use = "complete.obs", method = "pearson")


    Pearson's product-moment correlation
```

```
data:  fr$SYSBP and fr$DIABP
t = 84, df = 4400, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0,7726 0,7953
sample estimates:
   cor
0,7842
```

Meaning:…

RELEVANT?!

Regression

Function relation (NOT symmetric) between dependent (outcome, result, Y) variable and independent (explanatory, predictor, X) variable(s).

Y depends on X!!! - knowledge, not statistical

Questions:

- is there a (given kind of) relation? (statistical relation, not causality)
- what is the value of Y if X is:…? (estimation)
- what is the value of X if Y is:…?
- what is the best function that describe the relation?

Now we will talk about linear regression

Assumptions:

- X and Y are at least interval scale
- at least Y is a random variable (if we could not decide which is Y and which is X, both X and Y have to be random variable) - [remember: correlation: both are random]
- Y = f(X) +ϵ, where E(ϵ) = 0 and normally distributed (… or)

For 2 variables correlation - regression questions are “transformable”.

(For comparing two raters, two measuring devices (with no gold standard)…. - do NOT use correlation, regression - next lecture…)

Linear regression:

Linear function: $Y = \beta_0 + \beta_1 * X + \epsilon$

With the fitted line we **estimate** the y value ($\hat{y}$) for a given x

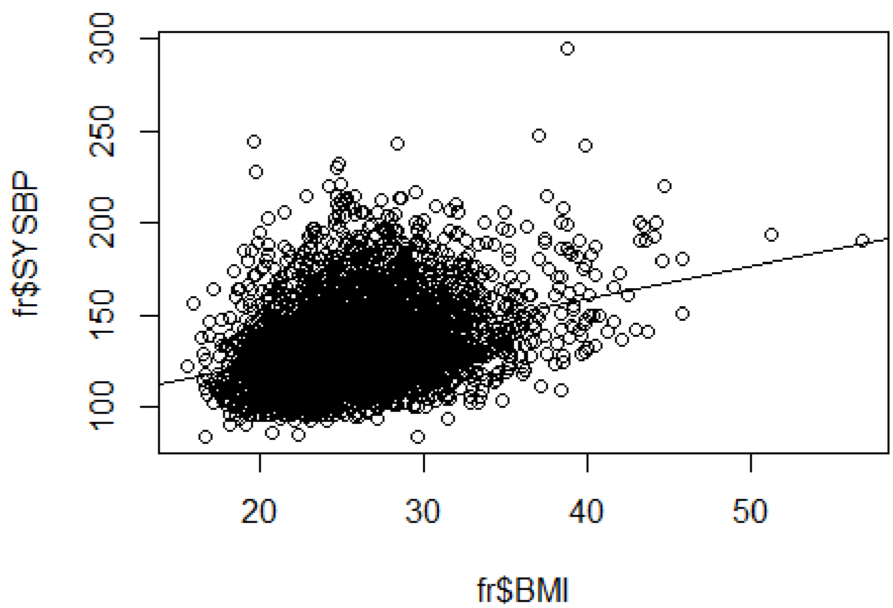
$\hat{y} = b_0 + b_1 * x$

Meaning of intercept and slope: ….. learned before!, but reveal it

For estimation of the linear we (usually) use the OLS (Ordinary Least Square method)

Residuals: points-line **vertical** differences (difference of measured and estimated values)

```
plot(fr$BMI, fr$SYSBP)
abline(lm(fr$SYSBP ~ fr$BMI))
```




```
lm(fr$SYSBP ~ fr$BMI)


Call:
lm(formula = fr$SYSBP ~ fr$BMI)

Coefficients:
(Intercept)       fr$BMI
      86,67         1,79
```

Hypothesis tests

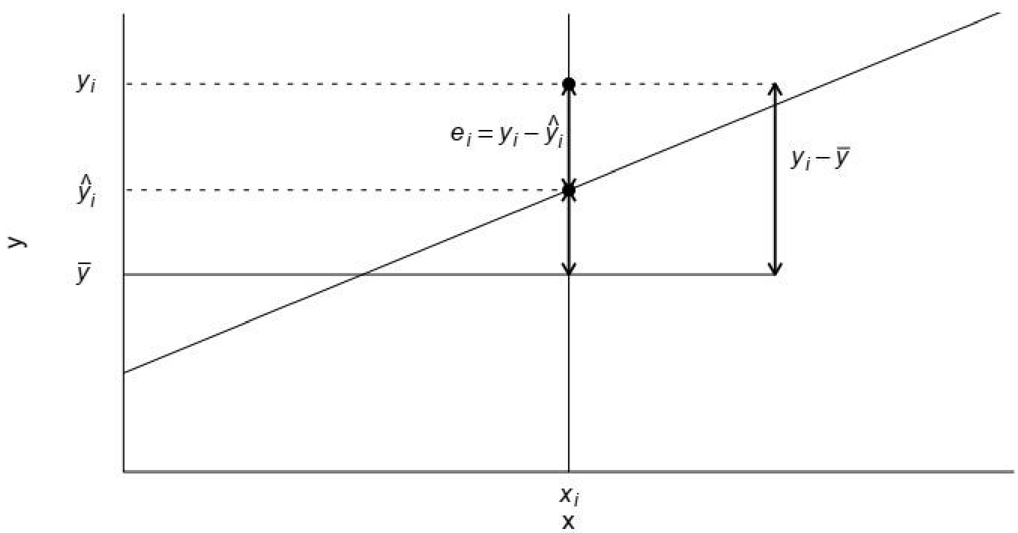
H0: theoretical slope = 0 Statistics: t = b1/SE(b1)

```
summary(lm(fr$SYSBP ~ fr$BMI))


Call:
lm(formula = fr$SYSBP ~ fr$BMI)

Residuals:
```

```
   Min      1Q Median      3Q     Max 
-56,21 -14,78  -3,83  10,42 138,90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  86,6668     2,0286    42,7   <2e-16 ***
fr$BMI        1,7885     0,0775    23,1   <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 21,1 on 4413 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0,108, Adjusted R-squared:  0,107 
F-statistic:  532 on 1 and 4413 DF,  p-value: <2e-16
```

H0: X and Y are independent Statistics: $F = \frac{SS_R}{SS_H/(n-2)}$

Variance divided 2 part: Total variance of Y = Variance of X regression + Variance of random error




```
anova(lm(fr$SYSBP ~ fr$BMI))

Analysis of Variance Table

Response: fr$SYSBP
            Df  Sum Sq Mean Sq F value Pr(>F)    
fr$BMI       1  237560  237560     532 <2e-16 ***
Residuals 4413 1969349     446                   
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

In case of 1 X the two Hypothesis are the same.

Meaning of R^2

How much of variability of Y explaned by X

$$R^2 = \frac{SS_R}{SS_T} = 1 - SS_H SS_T$$

Confidence intervals

(If we repeat the measurement, the 95% of the fitted parameters will be in this range)

For the parameters

```
confint(lm(fr$SYSBP ~ fr$BMI))

             2,5 % 97,5 %
(Intercept) 82,690  90,64
fr$BMI       1,637   1,94
```

For an estimated Y

```
reg_mod <- lm(SYSBP ~ BMI, data = fr)
x <- data.frame(BMI = 20)
predict(reg_mod, newdata = x, int = "confidence")

    fit   lwr   upr
1 122,4 121,4 123,5
```

For more x

```
reg_mod <- lm(SYSBP ~ BMI, data = fr)
x <- data.frame(BMI = 20:25)
predict(reg_mod, newdata = x, int = "confidence")

    fit   lwr   upr
1 122,4 121,4 123,5
2 124,2 123,3 125,2
3 126,0 125,2 126,9
4 127,8 127,0 128,6
5 129,6 128,9 130,3
6 131,4 130,7 132,0
```

Prediction interval

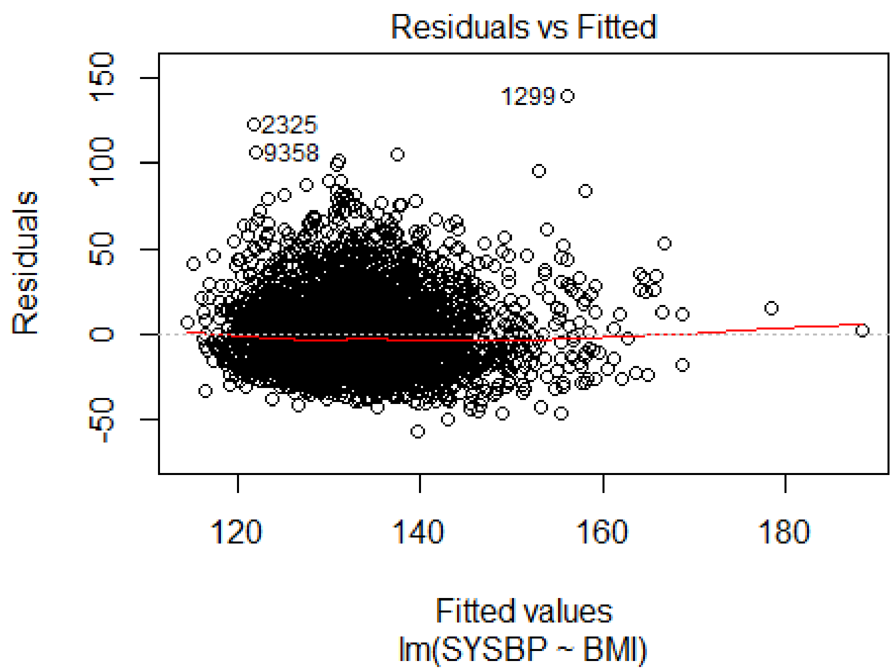
An interval that contains a **new observation** with 95% probability

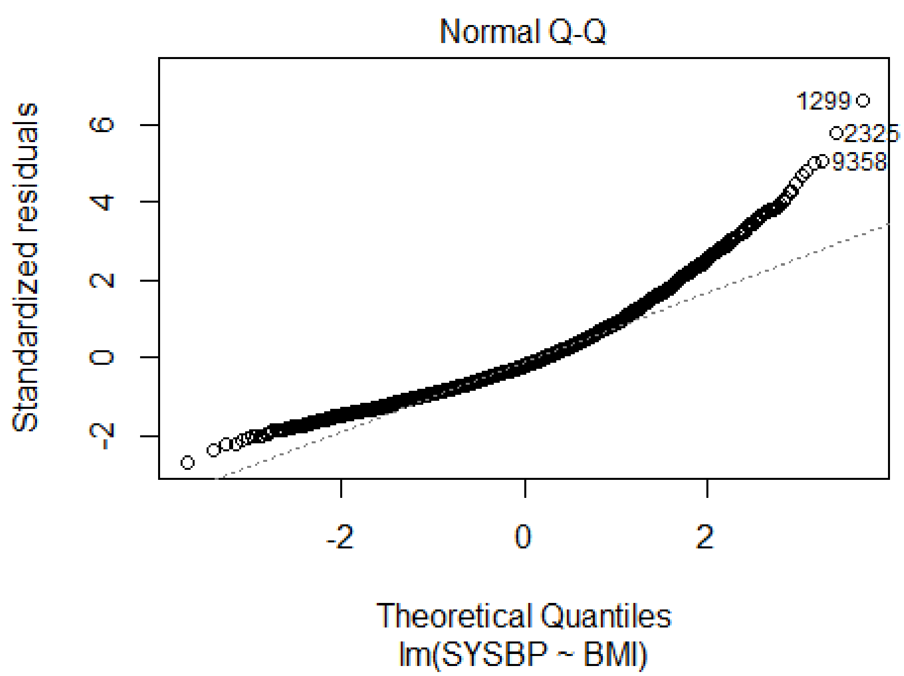
```
reg_mod <- lm(SYSBP ~ BMI, data = fr)
x <- data.frame(BMI = 20)
predict(reg_mod, newdata = x, int = "prediction")

    fit   lwr   upr
1 122,4 81,01 163,9
```

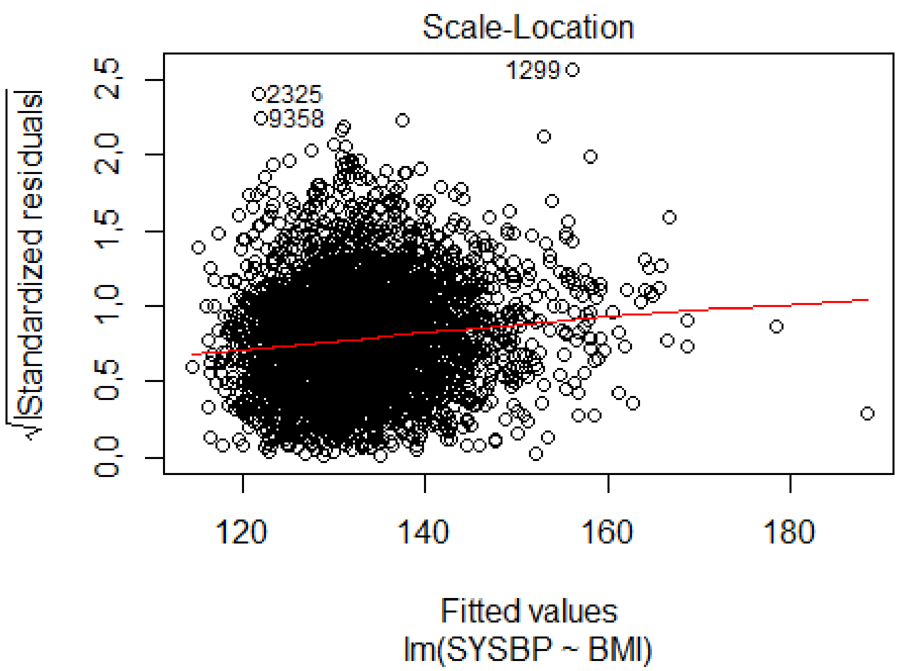
Diagnostic

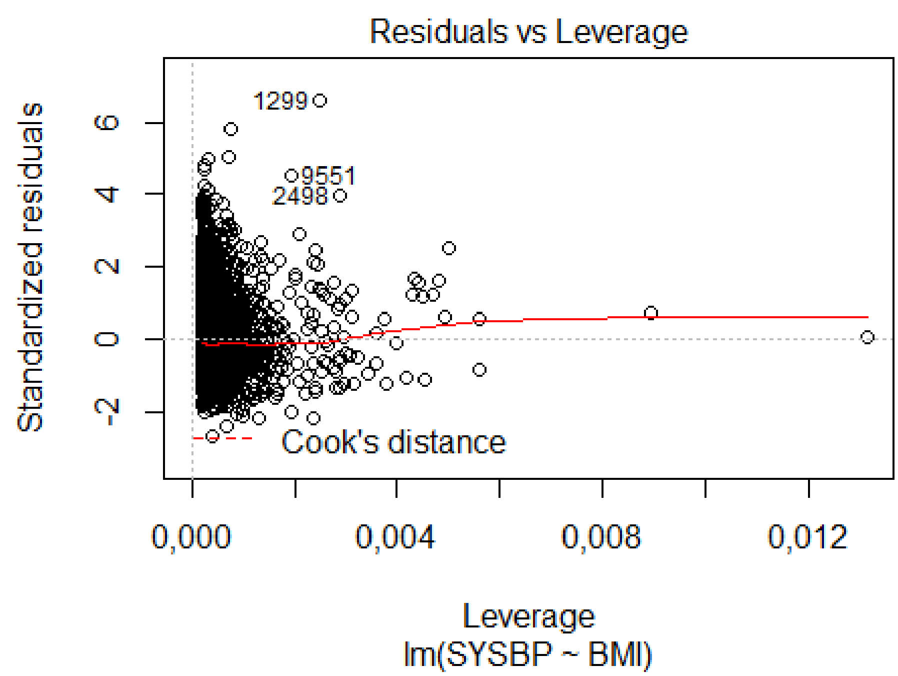
We have assumptions…

```
plot(reg_mod)
```


Residuals vs Fitted
Residuals
150
100
50
0
-50
1299
2325
9358
120
140
160
180
Fitted values
lm(SYSBP ~ BMI)


Normal Q-Q
Standardized residuals
6
4
2
0
-2
1299
2325
9358
-2
0
2
Theoretical Quantiles
lm(SYSBP ~ BMI)





Multiple linear regression

Always write up the equation!

With 2 X variable

```
regmod2 <- lm(formula = SYSBP ~ BMI + TOTCHOL, data = fr, na.action =
na.exclude)
summary(regmod2)


Call:
lm(formula = SYSBP ~ BMI + TOTCHOL, data = fr, na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
-50,73 -14,29  -3,66  10,16 138,96

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 69,94431    2,47293    28,3   <2e-16 ***
BMI          1,68506    0,07746    21,8   <2e-16 ***
TOTCHOL      0,08176    0,00711    11,5   <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 20,8 on 4361 degrees of freedom
  (70 observations deleted due to missingness)
Multiple R-squared:  0,134, Adjusted R-squared:  0,134
F-statistic:  339 on 2 and 4361 DF,  p-value: <2e-16
```

Without X variable?

```
regmod_null <- lm(formula = SYSBP ~ 1, data = fr, na.action = na.exclude)
summary(regmod_null)  # intercept is the mean!


Call:
lm(formula = SYSBP ~ 1, data = fr, na.action = na.exclude)

Residuals:
   Min     1Q Median     3Q    Max
-49,41 -15,41  -3,91  11,09 162,09

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  132,908      0,337     395   <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 22,4 on 4433 degrees of freedom
```

Which model is the best? Which variable is good to put in?