Regression

Computer Lab 4

Your Name

2025-03-16

The purpose of today's lab is for you to:

- become familiar with functions in R for correlation and regression analysis
- practice constructing a reasonable statistical model using collected data, as well as critically reviewing the model and its ability to describe reality
- apply your knowledge and analyse a biostatistical dataset using R
- practice presenting assumptions, models, and conclusions from a statistical analysis in writing

Table of contents

Preparation Tasks

You must have worked thoroughly with the essential concepts in the chapter on regression in the course book. Repeat the concepts *regression line, residuals, confidence interval for expected value*, and *prediction interval* if necessary.

You should have completed the following tasks *before* you attend the lab.

Homework 1

Complete exercise 6.10 in the workbook.

1. Introduction - Regression Analysis in R

1 Length and Age of Cod Fish

For 10 cod fish, we have values for the variables Length (cm) and Age (years).

Individual	1	2	3	4	5	6	7	8	9	10
Length (cm)	15	30	35	50	55	60	58	25	12	43
Age (yearr)	1	2	3	4	5	6	5	2	1	4

Task 1.1

Start by entering these values into R. Place the values in a table (dataframe), which you name `codfish`, with the values in two columns: Length and Age. Remember that R is case sensitive and sensitive to the use of lower and upper case letters and special characters such as ä or a. Thus, the dataset should contain 2 columns with 10 values in each column.

```
codfish <- data.frame(Length=c(15,30,35,50,55,60,58,25,12,43),
                      Age=c(1,2,3,4,5,6,5,2,1,4))
codfish
```

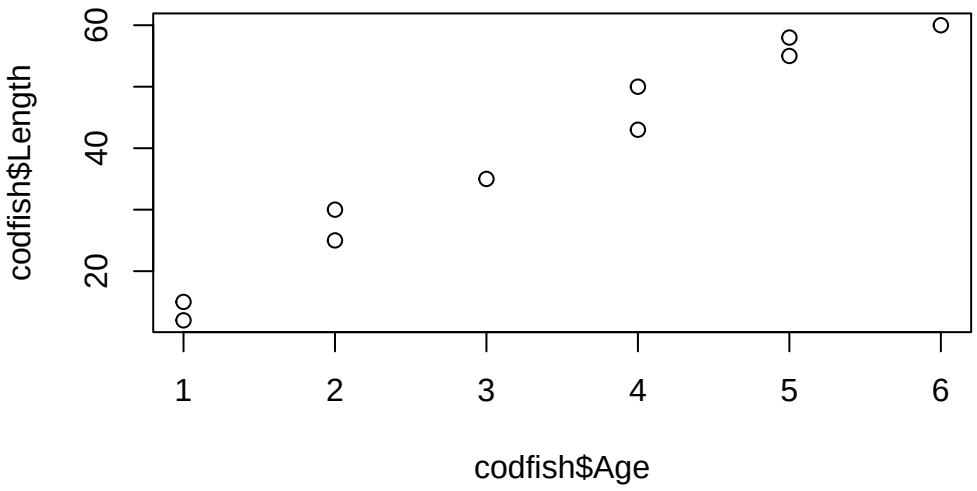
```
   Length Age
1      15   1
2      30   2
3      35   3
4      50   4
5      55   5
6      60   6
7      58   5
```

```
8       25    2
9       12    1
10      43    4
```

Task 1.2

Describe the data. Make a graphical description of the relationship by plotting a scatter plot with Age on the x-axis and Length on the y-axis. Does it appear that there is a linear relationship?

```
plot(codfish$Age, codfish$Length)
```



Answer:

2. Correlation analysis of the relationship between length and age in codfish

We begin by examining the relationship between the variables using the correlation coefficient.

Task 2.1

Calculate the correlation coefficient (Pearson) and test if it is different from zero. Do the results indicate that there is a linear relationship between Length and Age?

Model: Let X be the length and Y be the age of fish. We assume that X and Y belong to a bivariate (2-dimensional) normal distribution with correlation $\rho = \frac{COV(X,Y)}{\sqrt{V(X)\cdot V(Y)}}$. The correlation is estimated by the correlation coefficient, i.e., $\hat{\rho} = r$, where $r = \frac{SS_{XY}}{\sqrt{SS_X \cdot SS_Y}}$.

```
cor(codfish$Age, codfish$Length)
```

```
[1] 0.982841
```

Hypoteses: $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$

Assuming a normal distribution holds, the test statistic $t = r\sqrt{\frac{n-2}{1-r^2}}$ will be approximately normally distributed. Thus, it is possible to test hypotheses using a t-test. This test is performed in R using the `cor.test` routine.

```
cor.test(codfish$Age, codfish$Length)
```

```
    Pearson's product-moment correlation

data:  codfish$Age and codfish$Length
t = 15.071, df = 8, p-value = 3.715e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9266410 0.9960742
sample estimates:
     cor
0.982841
```

Answer:

3. Simple linear regression for the relationship between length and age of codfish

> We will now examine how the relationship between the variables looks by fitting a straight line to the data. The command `lm(y ~ x)` fits a linear model for the dependent variable `y` as a function of one or more explanatory variables `x` (lm is short for linear model). Then, we can obtain various properties of the model and estimates using additional commands.

```
model <- lm(Length ~ Age, data=codfish)
model           # estimates of the parameters intercept and slope
```

```
Call:
lm(formula = Length ~ Age, data = codfish)

Coefficients:
(Intercept)          Age
      5.993        9.790
```

```
summary(model)  # more information, e.g. estimation errors and test if the slope is zero
```

```
Call:
lm(formula = Length ~ Age, data = codfish)

Residuals:
   Min     1Q Median     3Q    Max
-4.733 -1.810 -0.468  2.307  4.847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9929     2.4044   2.492   0.0374 *
Age           9.7900     0.6496  15.071 3.72e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.443 on 8 degrees of freedom
Multiple R-squared:  0.966, Adjusted R-squared:  0.9617
F-statistic: 227.1 on 1 and 8 DF,  p-value: 3.715e-07
```

```
confint(model)  # confidence interval for the estimated parameters
```

```
                2.5 %   97.5 %
(Intercept) 0.4483295 11.53744
Age         8.2920547 11.28802
```

Perform the regression analysis above. Identify the following measures in the output:

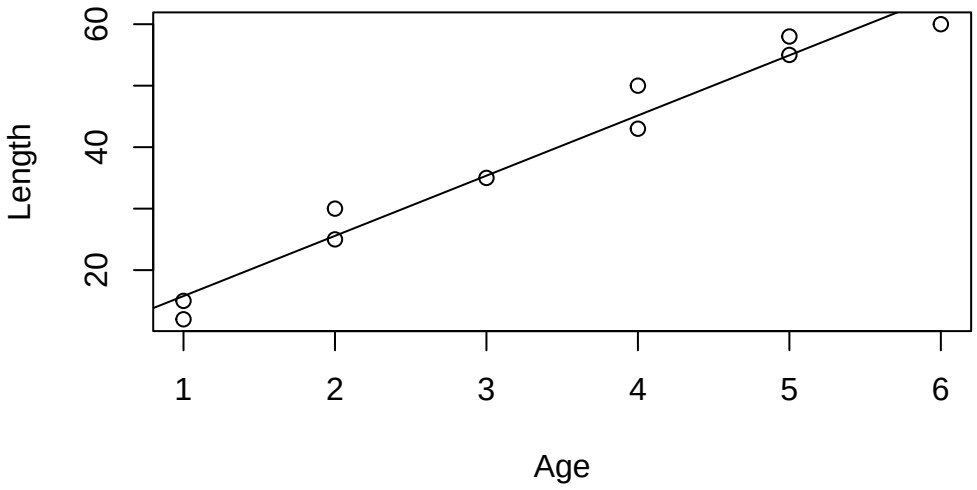
- Coefficient of determination (R^2)
- Standard error of residuals (s_e)
- Estimates of the coefficients β_0 and β_1
- Confidence interval for the slope
- p-value from the t-test of the null hypothesis that the slope is zero

:::

Task 3.2

To plot the estimated regression line in the figure you created earlier, you can use the command

```
plot(Length ~ Age, data=codfish)
abline(model)
```



4. Assumption Checking

We will now check three of the assumptions in the analysis, namely (1) residuals are independent, (2) normally distributed, and (3) have equal variance. The assumption of independence is assessed by understanding how the data was collected. In this case, we don't have much information, except that the measurements were taken on different individuals caught in a day. We find no reason to assume dependence, and we examine the remaining assumptions by studying the residuals. The residuals in the fitted model are computed with the command"

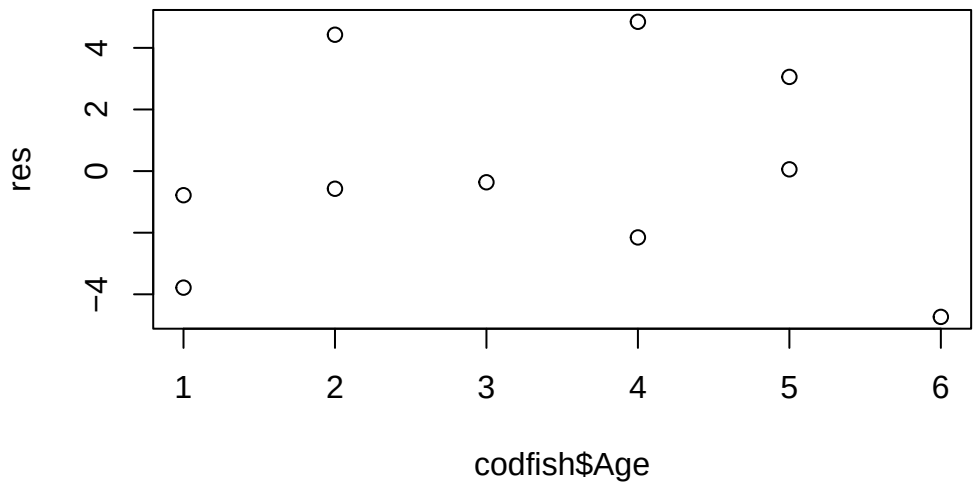
```
res <- residuals(model)
res
```

```
         1          2          3          4          5          6          7
-0.7829181  4.4270463 -0.3629893  4.8469751  0.0569395 -4.7330961  3.0569395
         8          9         10
-0.5729537 -3.7829181 -2.1530249
```

Task 3.3

Examine whether the residuals have equal variances (i.e., constant spread around the line) by plotting the residuals against the x-variable Age. Does the assumption seem to hold?

```
plot(codfish$Age, res)
```



Answer:

Task 3.4

Examine if the residuals are normally distributed by creating a Q-Q plot using the function `qqnorm`.
Does the assumption of normality seem to be met?

```
# wrtie your R code here
```

Answer:

5. Predictions and confidence intervals

A regression model can be used to forecast or predict the expected value or the individual value of y given $x = x_0$. In the formula collection, we have the following formulas for confidence intervals (refer to the formula collection for explanations of all symbols).

Confidence interval of the expected value:

$$I_{\mu(x_0)} : b_0 + b_1 x_0 \pm t_{1-\alpha/2,n-2} \sqrt{s_e^2 (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x})}$$

Confidence interval of the next value (also known as prediction interval):

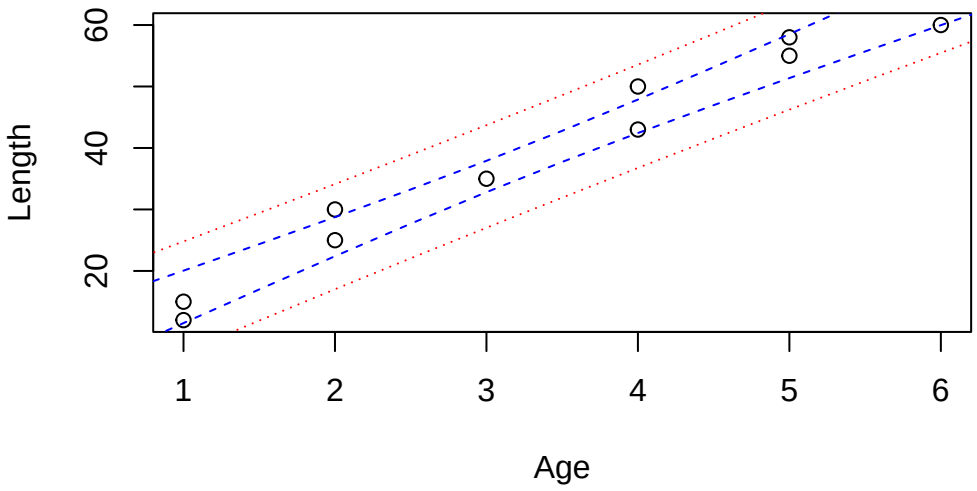
$$I_{y(x_0)} : b_0 + b_1 x_0 \pm t_{1-\alpha/2,n-2} \sqrt{s_e^2 (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x})}$$

Task 3.5

Draw predictions of length given age with confidence intervals for expected values and for individual values (prediction intervals). First, we use the model to make predictions for different values of age.
To plot the intervals in the figure, we draw lines with age on the x-axis and corresponding interval limits on the y-axis. To distinguish between the intervals, we draw the confidence interval with dashed blue lines and the prediction interval with dotted red lines.
Why is the prediction interval (the blue one) wider than the confidence interval for the expected value (the red one)?

```
# Do predictions for a sequence of ages with steps of half years:
# 0.5, 1.0, 1.5, …, 7.0, 7.5:
x0 <- data.frame(Age=seq(0.5,7.5,0.5))
mu0conf <- predict(model, x0, interval="confidence")  # confidence interval
y0pred <- predict(model, x0, interval="prediction")  # prediction interval


plot(Length ~ Age, data=codfish)
lines(x0$Age, mu0conf[,"lwr"], col="blue", lty=2) # lower limit
lines(x0$Age, mu0conf[,"upr"], col="blue", lty=2) # upper limit
lines(x0$Age, y0pred[,"lwr"], col="red", lty=3)
lines(x0$Age, y0pred[,"upr"], col="red", lty=3)
```



Answer:

> Task3.6
>
> What length is a seven year old codfish expected to have? Provide with confidence interval!.

```
# write your R code here
```

Answer:

6. Regression Analysis of the Relationship between Concentration and Absorbance

In the lab "Protein Determination According to the Bradford Method" in the cell biology course, absorbance was investigated for samples with different dilutions of Bovine Serum Albumin (BSA) standard. Samples with 0-10 µg of protein were diluted to 100 µl with water, and two samples were prepared per concentration.

Data for one lab group is available in the file `data/lab4_filer/Labbdata.RData`.

Model: According to the Lambert-Beer's law, absorbance (A) can be described as a linear function of concentration (c): $A = k \cdot c$, where the constant k depends on the substance's molar absorption coefficient at a specific wavelength and the cuvette's length. During measurements, one naturally expects some random variation. A reasonable model is that the

absorbance at measurement number i, A_i, is linearly described by the concentration c_i plus a random error:

$$A_i = \beta_0 + \beta_1 \cdot c_i + \epsilon_i$$

where ϵ_i is independent and normal distributed random errors with expected value 0 and standard deviation σ. Here β_0 is the true absorbance in the solution and β_1 corresponds to k in the equation.

> Task 4.1
>
> Examine based on data from the lab whether the linear regression model above is reasonable to fit.

```
load("data/lab4_filer/Labbdata.RData")
model2 <- lm(absorbans ~ koncentration, data=Labbdata)

# write your R code here
```

Answer:

> Task 4.2
>
> How much does the absorbance increase when the concentration increase one unit? Provide a 95 % confidence interval for this quantity.

Answer:

> Task 4.3
>
> What is the expected absorbance for a sample with concentration 50 (mg/l). Provide a 95 % confidence interval for this quantity. To do this, create a data frame where the value for the concentration is specified.

```
x50 <- data.frame(koncentration=c(50))
mu50conf <- predict(model2, x50,interval="confidence")

# write your R code here
```

Answer:

> Task 4.4
>
> Vi har ett prov med koncentration 50 (mg/l). Ange ett 95 %-igt prediktionsintervall för absorbansen i just detta prov.

```
# write your R code here
```

Answer:

> Task 4.5
>
> Huvudsyftet med mätningarna var att erhålla en standardkurva för hur absorbansen påverkas av koncentrationen. Anta att vi på ett prov med okänd koncentration c_0 uppmätte en absorbans på 0.43. Vilken koncentration kan det svara mot.
> Skatta koncentrationen utifrån kännedom av absorbansen. En skattning av c_0 kan vi få fram genom att lösa ut x ur sambandet $0.43 = \beta_0 + \beta_1 \cdot x$ så här (om den anpassade modellen sparats i variabeln `model2`). Vad blev den skattade koncentrationen?

```
beta0 <- model2$coefficients[1]
beta1 <- model2$coefficients[2]
c0 <- (0.43 - beta0) / beta1
c0
```

```
(Intercept)
   46.57378
```

Answer:

Some answers:
2.1. Yes! $r = 0.9828, t = 15.071$; p-värde=0.000 - you can reject the hypothesis that there is no linear association
3.1. $R^2 = 0.966$; $s_e = 3.443$; $\hat{\beta}_0 = b_0 = 5.993$; $\hat{\beta}_1 = b_1 = 9.790$; $I_{\beta_1} = (8.29, 11.29)$; $p = 3.72 \cdot 10^{-7}$
3.6. Expected length at age 7 is 74.52; Confidence interval (68.43, 80.61)
4.1. Normal distributed?: difficult to tell; Constant variance?: not easy to judge (too few values)
4.2. 0.0008 interval: (0.00063, 0.0011)
4.3. confidence interval: (0.425, 0.441)
4.4. prediction interval: (0.407, 0.459)
4.5. c_0 is estimated to be 46.6 mg/l

Summary of R functions

```
cor(x, y)                       #Correlation coefficient
cor.test(x, y)                  #Test for correlation coefficient
lm(y ~ x)                       #Regression of y as a function of x
lm(y ~ x, data)                 #Data = dataframe containing x and y
summary(model)                  #Estimates, p-values, etcs
confint(model)                  #Confidence interval for parameters
predict(model, x0)              #Prediction of an expected value when x=x0
predict(model, x0, interval="confidence")      #… with a confidence interval
```

```r
predict(model, x0, interval="prediction")      #… with a prediction interval
residuals(model)                #Residuals
```

5. Statistic analysis

This is a continuation of the data material from the study of the activation program from laboration 3.

Some researchers argue that cholesterol levels increase with age, while others believe that factors such as smoking and sedentary lifestyle are more significant. In the variables `Aalder` and `Balder`, the ages of the men examined at the start of the study are recorded.

Do our data at the start of the study indicate that age affects cholesterol levels?

Suggested workflow

Since these are measurements before the activation campaign and you want large samples, it is advisable to merge data from groups A and B. Below is an example of how to do this.

```r
load("C:/Rfolder/masb11_students/data/lab3_filer/blodprov.Rdata")
df <- data.frame(
  alder = c(kolesterol$Aalder,kolesterol$Balder),
  kolesterol = c(kolesterol$Afore,kolesterol$Bfore)
)
```

Plot cholesterol content against age and examine with a statistical method whether cholesterol levels, apart from random factors, can be described by a function of age.

In the course, we have gone through simple linear regression and correlation analysis to investigate relationships between continuous variables. Consider which of the methods is most suitable for this particular dataset? Feel free to discuss with the lab/exercise supervisor before proceeding.

Instruktioner för rapportering i del 6

Please specify (if appropriate):

- what assumptions you make about data,
- which models you are specifying,
- which hypothesis you use
- Justify the results of the analysis and what interpretations and conclusions you make.
- To pass task 5, avoid referring to the R-code to describe what you have done.