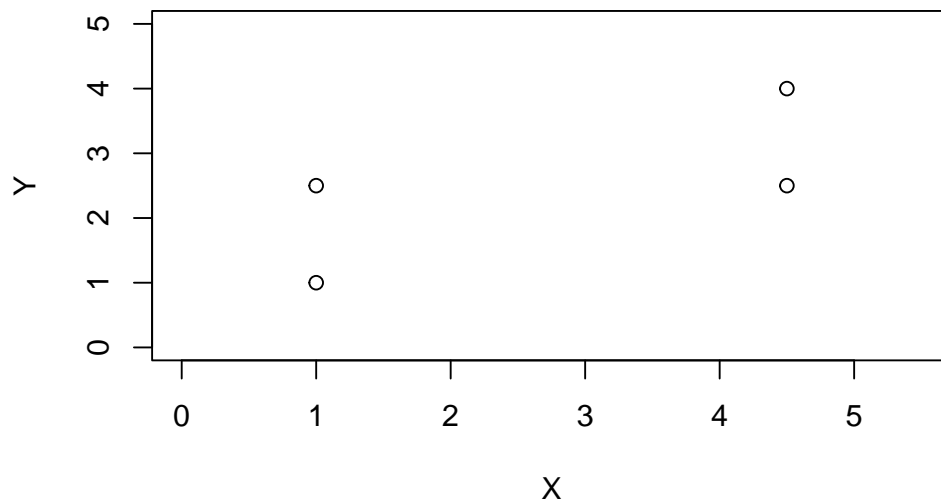
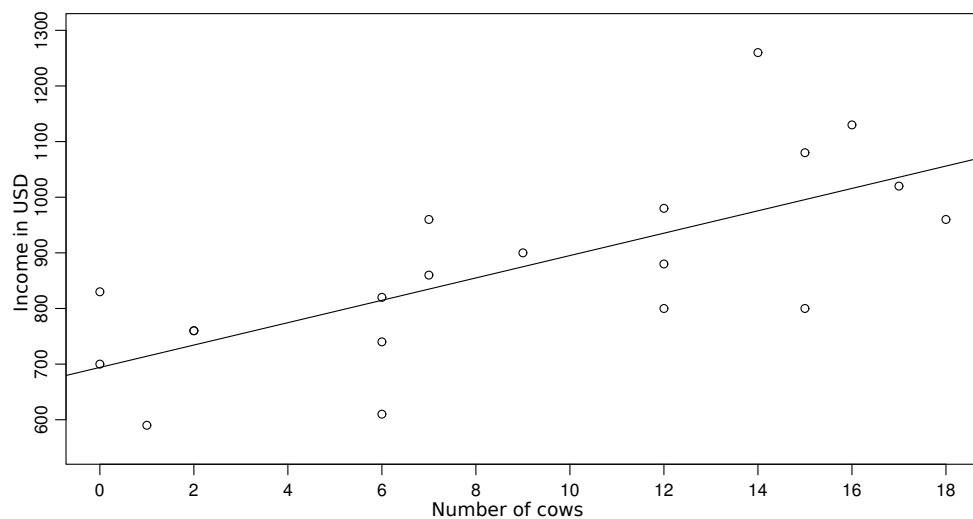


Series 1

1. a) In the plot below, draw the regression line for Y being the dependent and X being the independent variable and vice versa.



- b) In the plot below, we depicted for several farms the yearly income in Dollars versus the number of cows.



- (i) Give the approximate equation for the least squares line.
 - (ii) What is your estimate for the average deviation of the points with respect to the regression line?
 - (iii) Estimate the income of a farm with 15 cows and of a farm with 100 cows? How meaningful are these estimates?
2. In this exercise we prove that the residuals of least squares regression always have mean zero. For ease of notation, we only consider the simple linear model, but this result also holds for multiple linear regression.

- a) Given data $(x_1, y_1), \dots, (x_n, y_n)$, let

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{b_0, b_1} \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

Show that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

- b) Use part a) to show that the least squares regression line $\hat{\beta}_0 + \hat{\beta}_1 x$ passes through the point of averages (\bar{x}, \bar{y}) .
- c) For any line $y = b_0 + b_1 x$, let $e_i = y_i - (b_0 + b_1 x_i)$ be the vertical distance from y_i to the line. Show that $\sum e_i = 0$ for any line that passes through the point of averages (\bar{x}, \bar{y}) .
3. The following table contains some functions which can be linearized by a suitable transformation. Complete the table by inserting the needed transformations of x and y , and the resulting linear forms.

| Function | Transformation | Linear form |
|---|------------------------------|--------------------------------------|
| $y = \alpha x^\beta$ | $y' = \log(y), x' = \log(x)$ | $y' = \log(\alpha) + \beta \cdot x'$ |
| $y = \alpha e^{\beta \cdot x}$ | | |
| $y = \alpha + \beta \cdot \log(x)$ | | |
| $y = x/(\alpha \cdot x - \beta)$ | | |
| $y = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta \cdot x}}$ | | |
| $y = \alpha e^{\beta/x}$ | | |
| $y = 1/(\alpha + \beta e^{-x})$ | | |

4. The behaviour of the least squares estimator can be investigated by a small simulation study. Here are the R-commands for linear regression:

```
> ## simple linear regression
> set.seed(21)                # initializes the random number generator
> x <- seq(1,40,1)            # generates equidistant x-values
> y <- 2*x+1+5*rmnorm(length(x)) # y-values=linear function(x) + error
> reg <- lm(y~x)              # fit of the linear regression
> summary(reg)                # output of selected results
> plot(x,y)                   # scatter plot
> abline(reg)                  # draws regression line
```

- a) Write a sequence of R-commands which randomly generates 100 times a vector of y -values according to the above model with the given x -values and generates a vector of slopes of the regression lines.

R-hint: `help(for)`.

- b) Draw a histogram of the 100 estimated slopes and add the normal density of the theoretically true distribution of the slopes to the histogram.

R-hints: Because of part d), you should use `par(mfrow=c(1,2))`. The histogram must be drawn with parameter `freq=FALSE`, so that the y -axis is suitably scaled for drawing the density. The density can be added by something like

```
lines(seq(1.7,2.3,by=0.01),dnorm(seq(1.7,2.3,by=0.01),mean=?,sd=?)),
```

where you have to find the correct values for `mean` and `sd`. To compute the inverse of a matrix use `solve()`.

- c) Compute the mean and empirical standard deviation of the estimated slopes.
- d) Repeat the simulation with a skew, non-normal error distribution. That is, replace the second line by

```
y <- 2*x+1+5*(1-rchisq(length(x), df=1))/sqrt(2)
# You may try hist(5*(1-rchisq(40, df=1))/sqrt(2))
# to explore the error distribution
```

Repeat part b) using the new slopes. Add the same normal density, which is only an asymptotic approximation to the true distribution in this setup. Why does the normal approximation fit well? Note that the random variable $(1 - \chi_1^2)/\sqrt{2}$ has expectation 0 and variance 1.

5. The dataset `airline` contains the monthly number of flight passengers in the USA in the years 1949-1960.
- Plot the data against time and verbally describe what you observe.
 - Compute the logarithms of the data and plot them against time. Comment on the differences.
 - Define a linear model of the form

$$\log(y_t) = \sum_{j=1}^p \theta_j f_j(t) + \epsilon_t$$

by choosing $f_1(t) = t$ (linear trend in time) and by defining f_2, \dots, f_{13} as indicator functions of the months, e.g.

$$f_2(t) = \begin{cases} 1 & \text{if } t \text{ corresponds to a January month} \\ 0 & \text{otherwise.} \end{cases}$$

Remark: It is not necessary to specify an intercept parameter. Why? (optional)

- Fit the model specified in c) and plot the fitted values and residuals against time. Do you think that the model assumptions hold?

R-hints: read data:

```
airline <- scan("http://stat.ethz.ch/Teaching/Datasets/airline.dat")
```

If this is not possible at home, save the data locally and read it by:

```
airline <- scan("filename")
```

Use `plot(airline, ...)` for a) and consider the parameter `type="l"` of `plot` for the plots. Regression fit by

```
reg <- lm(log(airline) ~ f1+...+fp-1)
```

`fj` must be a vector of length 144 containing the values of $f_j(t)$. The inclusion of `-1` in the `lm`-command prevents the fit of an intercept parameter (which would be done by default otherwise). `f2` can be generated by `rep(c(1,rep(0,11)),12)`. The construction of the other terms `fj` is similar. The commands `fitted(reg)` and `resid(reg)` extract fitted values and residuals from an `lm`-object.

Preliminary discussion: Friday, March 4.

Deadline: Friday, March 11.