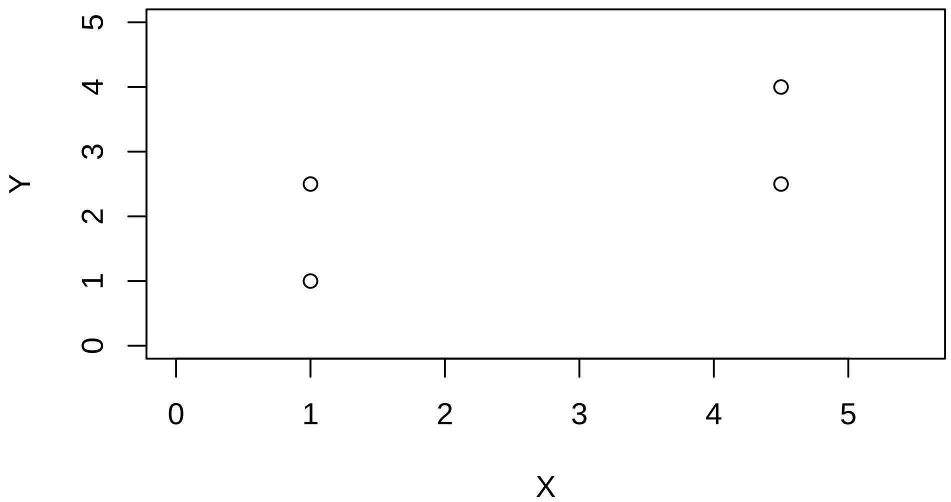
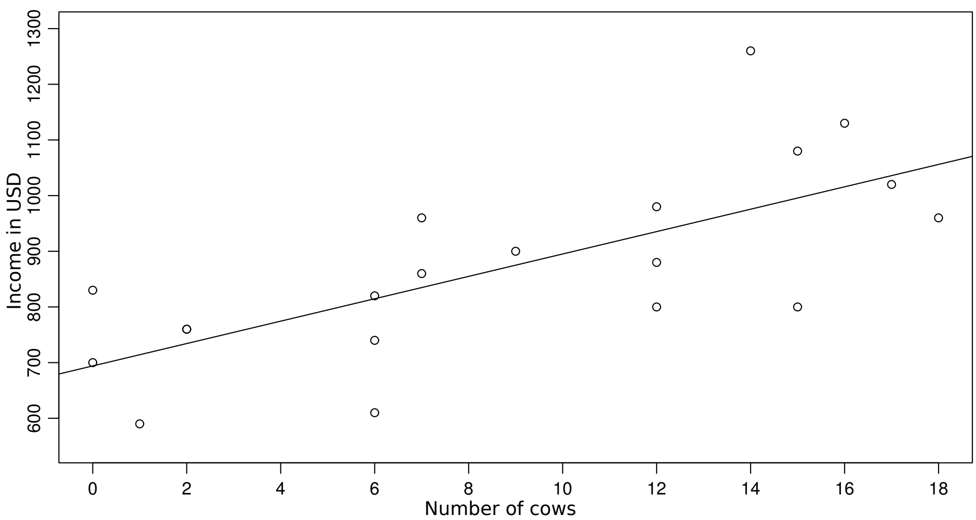
Series 1

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?

c) Let $(x_1, y_1), \cdots, (x_n, y_n)$ be n given data points. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. For any line $b_0 + b_1 x$, let $e_i = y_i - (b_0 + b_1 x_i)$ be the vertical distance of y_i to the line. Show that $\sum_{i=1}^n e_i = 0$ for any line that passes through the point of averages $(\bar{x}, \bar{y})$.

2. Let $y = \beta_0 + \beta_1 \log(x) + \epsilon$, where log is the natural logarithm. Given the following R output from such a model, answer the following questions.

```
Call:
lm(formula = y ~ log(x))

Residuals:
    Min      1Q  Median      3Q     Max
-3.5433 -0.5629 -0.0072  0.6564  3.0682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1783     0.1734   12.56   <2e-16 ***
log(x)        1.8232     0.1044   17.46   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9624 on 298 degrees of freedom
Multiple R-squared:  0.5057,       Adjusted R-squared:  0.504
F-statistic: 304.9 on 1 and 298 DF,  p-value: < 2.2e-16
```

a) Compute the predicted value of y for $x = 4$.

b) If we compare two observations i and j where $x_i = 2x_j$, then the fitted value $\hat{y}_i$ compared to $\hat{y}_j$ is increased by a value ____. Please fill in the blank.

Let $\log(y) = \beta_0 + \beta_1 x + \epsilon$. Given the following R output from such a model, answer the following questions.

```
Call:
lm(formula = log(y) ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5474 -0.5645 -0.0144  0.6577  3.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.12022    0.13581   8.248 5.21e-15 ***
x            0.95966    0.02272  42.244  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9619 on 298 degrees of freedom
Multiple R-squared:  0.8569,       Adjusted R-squared:  0.8564
F-statistic:  1785 on 1 and 298 DF,  p-value: < 2.2e-16
```

c) Compute the predicted value of y for $x = 3$.

d) If we compare two observations i and j where $x_i = x_j + 1$, then the fitted value $\hat{y}_i$ compared to $\hat{y}_j$ is multiplied by a value ____. Please fill in the blank.

3. 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   <- rnorm(40, 20, 3)               # generates x-values
> y   <- 1 + 2 * x + 5 * rnorm(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)                        # draw regression line
> plot(reg, which = 1)               # draw Tukey-Anscombe plot
```

a) Write a sequence of R-commands which randomly generates 100 times a vector of y-values according to the given model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with $\beta_0 = 1, \beta_1 = 2$ and $\epsilon_i \sim N(0, 5^2)$ with the given x-values, and generates a corresponding vector of estimated slopes $\hat{\beta}_1$ of the regression lines.
Hint:

- Look at the help file of the function `for`, i.e. `?for`.

b) For the first three generated y-vectors, plot y against x, and add the fitted regression line and construct the corresponding Tukey-Anscombe plot.

c) Compute the empirical mean and standard deviation of the estimated slopes.

d) Compute the theoretical variance of $\hat{\beta}_1$.
Hint: To compute the inverse of a matrix use `solve()`.

e) Draw a histogram of the 100 estimated slopes and add the normal density of the theoretical distribution of $\hat{\beta}_1$ to the histogram. What do you observe? Does it fit well?
Hints: 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
`lines(seq(1.3, 2.6, by = 0.01), dnorm(seq(1.3, 2.6, by = 0.01), mean= ?, sd = ?))`,
where you have to find the correct values for the arguments `mean` and `sd`.

4. We now repeat the simulation from exercise 3 with different error distributions that violate some of the assumptions. In each case, repeat part a) - e) and answer the following questions for all the tasks: Which (if any) assumptions are violated? What properties of the distribution of $\hat{\beta}_1$ are affected by this? Which part of the R output do you still trust?

a) Replace the fourth line of the R code in the previous exercise by

```
y   <- 1 + 2 * x + 5 * (1 - rchisq(length(x), df = 1)) / sqrt(2)
```

Hints: To get an idea of the error distribution, you may look at the following histogram and values

```
errors <- 5 * (1 - rchisq(40, df = 1)) / sqrt(2)
hist(errors)
mean(errors)
var(errors)
```

b) Replace the fourth line of the R code in the previous exercise by

```
y   <- 1 + 2 * x + 5 * rnorm(length(x), mean = x^2 / 5 - 1, sd = 1)
```

c) Replace the fourth line of the R code in the previous exercise by

```
require(MASS)
Sigma <- matrix(0.7,40,40)
diag(Sigma) <- 1
y   <- 1 + 2 * x + 5 * mvrnorm(n = 1, mu = rep(0, length(x)), Sigma = Sigma)
```

d) Replace the fourth line of the R code in the previous exercise by

```
y <- 1 + 2 * x + 5 * rnorm(length(x), mean = 0, sd = (x-15)^2 / 30)
```

Preliminary discussion: Friday, March 01.

Deadline: Friday, March 08.