MS-C1620 Statistical Inference

Exercise 8

Homework exercise

To be solved at home before the exercise session.

1. a. Show that if in simple linear regression both the explanatory variable x and the response y have been marginally standardized such that $\bar{x} = 0, s_x = 1$ and $\bar{y} = 0, s_y = 1$, then the estimated least squares regression model is simply,

$$\hat{y}_i = \hat{\rho}(x, y)x_i.$$

That is, the regression coefficient of x equals the sample correlation between x and y.

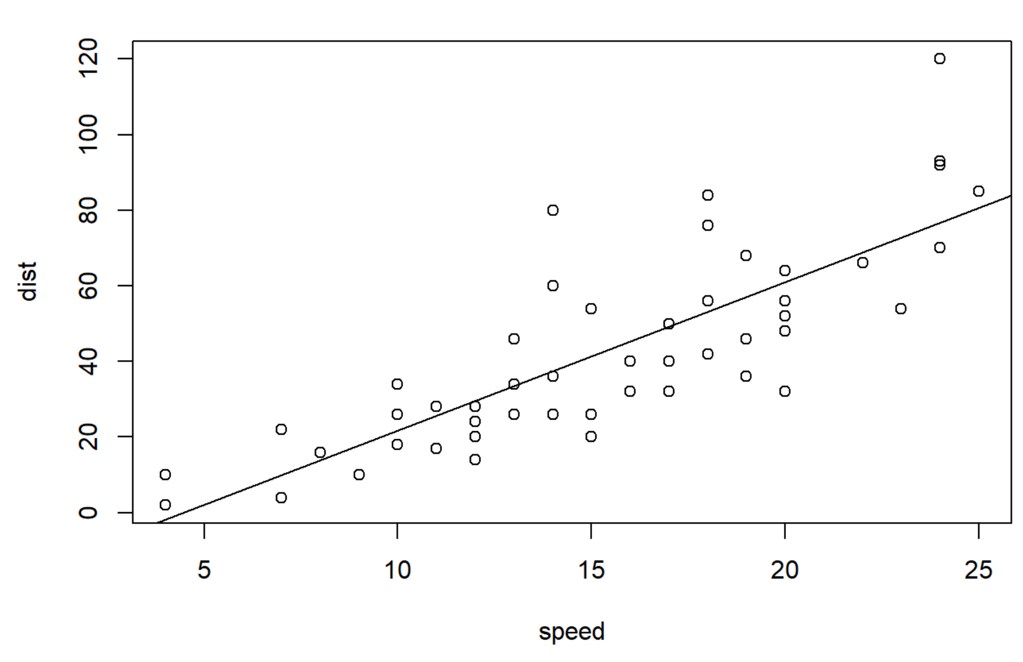
```
# Plug in the means and sample standard deviations to the formula in slide 7.15 to obtain the result.
```

```
b. The `cars` data give the speeds of cars (`speed`, in mph) and the corresponding distances taken to stop (`dist`, in feet). The below shows the model summary of a simple linear regression model fit using `speed` as an explanatory variable and `dist` as a response. Interpret the model results.
```

```
cars_lm <- lm(dist ~ speed, data = cars)
summary(cars_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
plot(dist ~ speed, data = cars)
abline(cars_lm)
```




```
# Regression coefficient ~3.9: an increase of 1 mph in speed increases the expected stopping distance by 3.9 feet (and based on the low p-value, this relationship is not caused by randomness (assuming the model assumptions hold)).
# R-squared 0.6511: the model manages to explain around two thirds of the variation in the response variable, indicating a good fit.
# Based on the scatter plot, the relationship indeed looks to be linear.
```

Class exercise

To be solved at the exercise session.

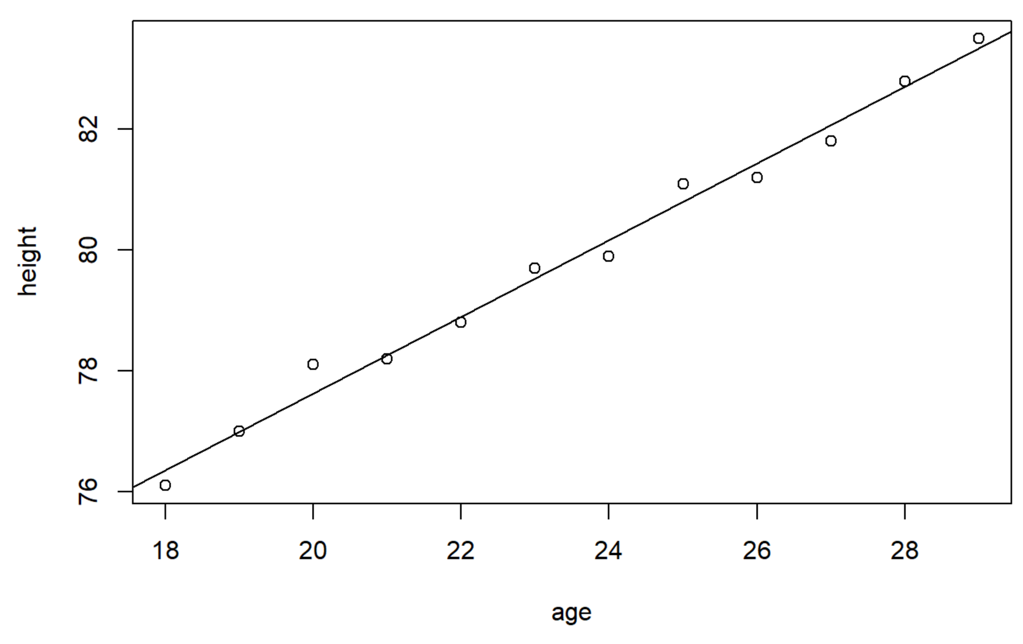
1. The file `data_children.txt` contains data on children's ages (`age` , in months) and heights (`height` , in centimeters). Investigate whether there is a linear relationship between the two variables.
   a. Read the file into R using the command `read.table` .
   b. Draw a scatter plot of `age` and `height` .
   c. Fit a linear model to the data using `height` as a response variable.
   d. Add the fitted regression line to the scatter plot. Does the fit appear good?
   e. Interpret the estimated regression coefficient of `age` and the R^2-value of the model.

```
# a.
# Replace "params$your_path_here_1" with your path to the .txt file in the code below (and remember that "/" is used to navigate sub-folders in R)
children <- read.table(params$your_path_here_1, sep = "\t", header = TRUE)


# b.
plot(children)


# c.
children_lm <- lm(height ~ age, data = children)


# d.
# The line appears to fit the data well.
abline(children_lm)
```




```
# e.
summary(children_lm)
```

```
##
## Call:
## lm(formula = height ~ age, data = children)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27238 -0.24248 -0.02762  0.16014  0.47238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
## age           0.6350     0.0214   29.66 4.43e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.256 on 10 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9876
## F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
```

```
# The regression coefficient is about 0.64, meaning that for an increase of a single month in a child's age, the expected value of her/his height goes up by 0.64cm.

# The coefficient of determination is ~0.99, indicating an excellent fit.
```

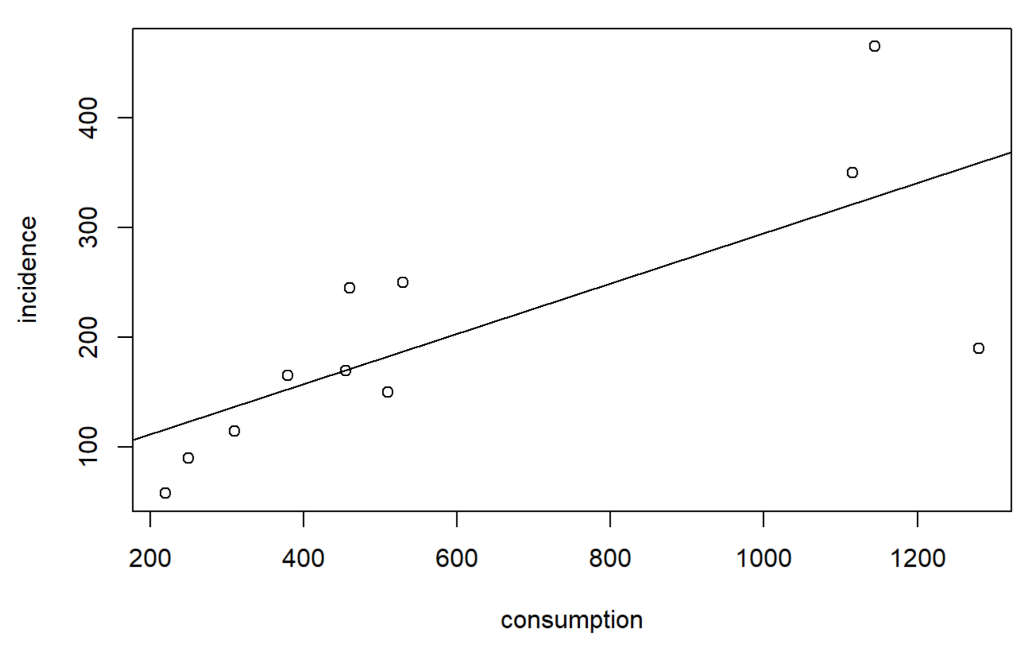
2. The file `data_tobacco.txt` contains data on cigarette consumption and lung cancer incidences from 11 different countries. The variable `consumption` describes the yearly consumption of cigarettes per capita in 1930 and the variable `incidence` tells the lung cancer incidence rates per 100 000 people in 1950. *(Recall exercise 7.2)*
   a. Read the file into R using the command `read.table` .
   b. Draw a scatter plot of `consumption` and `incidence` .
   c. Fit a linear model to the data using `incidence` as a response variable.
   d. Add the fitted regression line to the scatter plot. Does the fit appear good?
   e. Interpret the estimated regression coefficient and p-value of `consumption` .
   f. Interpret the R^2-value of the model.
   g. Drop USA from the data, redo the previous analysis and compare the results to those obtained with the full data. What happened?

```
# a.
# Replace "params$your_path_here_2" with your path to the .txt file in the code below (and remember that "/" is used to navigate sub-folders in R)
tobacco <- read.table(params$your_path_here_2, sep = "\t", header = TRUE)


# b.
plot(incidence ~ consumption, data = tobacco)
# text(tobacco$consumption, tobacco$incidence, labels = tobacco$country, cex= 0.7, pos=3)


# c.
tobacco_lm <- lm(incidence ~ consumption, data = tobacco)


# d.
# The line appears to miss most of the points in favour of trying to reach closer to the outlier in the lower right corner.
abline(tobacco_lm)
```




```
# e.
summary(tobacco_lm)
```

```
##
## Call:
## lm(formula = incidence ~ consumption, data = tobacco)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -169.016  -32.813    0.004   45.804  136.914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.74886   48.95871   1.343  0.21217
## consumption  0.22912    0.06921   3.310  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.13 on 9 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.4989
## F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

```
# The regression coefficient is about 0.23, meaning that for an increase of a single smoked cigarette in year per capita, the expected value of the incidence of lung cancer (20 years aftewards) goes up by 0.23 units.
# The p-value related to the coefficient is small (below the standard 0.05), meaning that this effect is most likely real, and not just caused by randomness (assmuing the model assumptions hold).

# f.
# The R-squared is somewhat high (~0.55), meaning that the model succeeds in explaining a bit over half of the variation of the response variable (however, compare this to the R-squared without the outlier in part g.)



# g.
tobacco <- tobacco[-7, ]
# Running the previous code to remove USA and then redoing the steps yields:
# Examination of the scatter plot shows that the line fits the data much better now
# Coefficient of "consumption" ~ 0.36 (the outlier had "contaminated" the value for the full data)
# p-value is much smaller (removing the outlier helped the model see clearer that the perceived effect is not just randomness)
# R-squared increased to ~0.89, a much better fit.

# NOTE: Again, in practice we should have some clear reason for removing U.S. from the data. We'll investigate this closer next time.
```

3. **(Optional)** Investigate how much a single outlier can affect the results of a linear model: Create a small data set that has a perfect linear relationship between its two variables (such a model has the explanatory variable p-value equal to 0 and the coefficient of determination equal to 1). Then, add a single outlying data point and see how much you can change the p-value and the coefficient of determination by varying the outlier's value.