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,

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

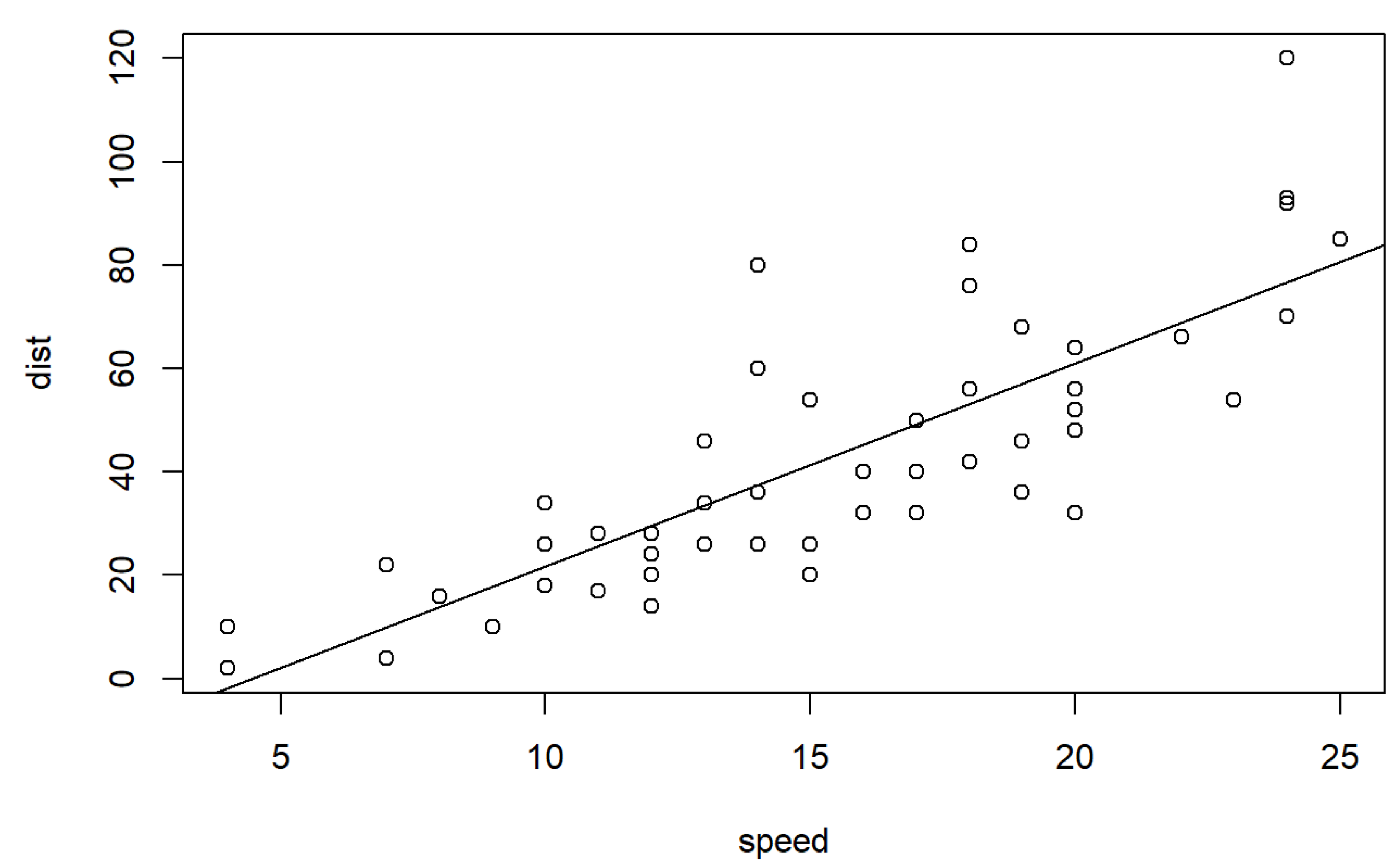
   That is, the regression coefficient of x equals the sample correlation between x and y.

   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)
```



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.

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?

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.