# Statistical inference : part 2, practice session 2

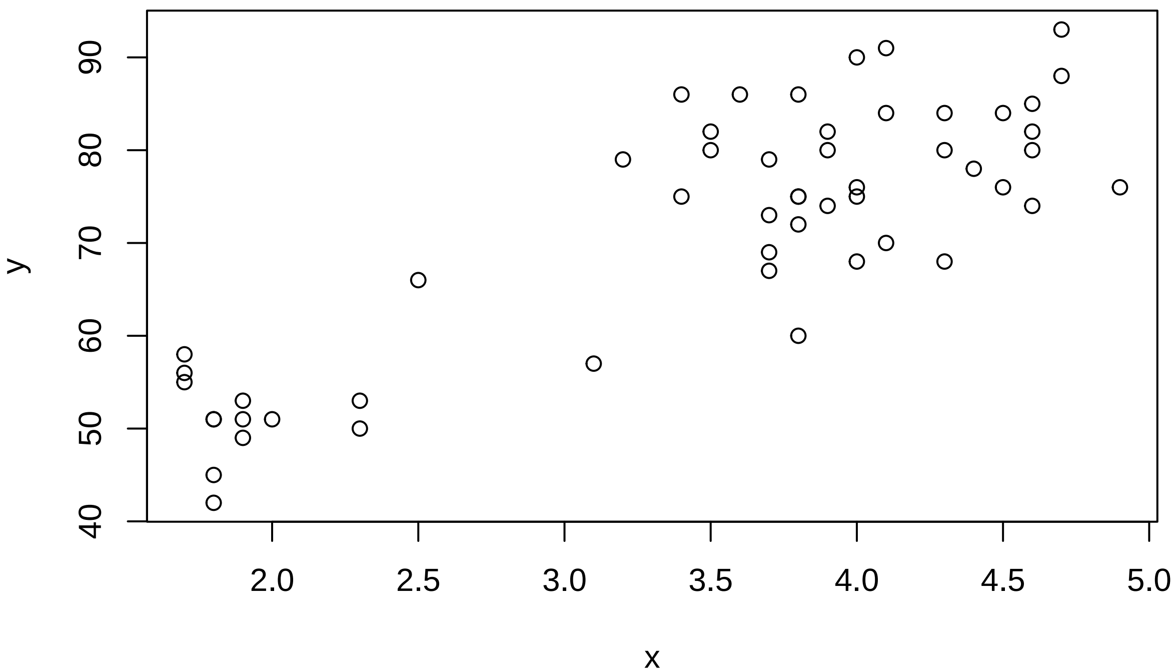
2023-11-30

## Exercicse 6.14

Data importation:

```
geyser = read.csv(file = "geyser.csv", sep = ";")
```

```
plot(y ~ x, data = geyser)
```




(a) Implement the formulas given $\hat{\beta}_1$ and $\hat{\beta}_0$ (you can also use `lm` to fit the model `lm(y ~ x, data = geyser)`)

(b) Under $H_0 : \beta_1 = 0$, we have:

$$t = \frac{\hat{\beta}_1}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$$

where $t(n-2)$ stands for the $t$ distribution with $n-2$ degrees of freedom, and $s^2$ is the unbiaised estimator of the variance of the noise:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Thus, the course (p.132) says to reject $H_0$ if $|t| \geq t_{\alpha/2,n-2}$ where $t_{\alpha/2,n-2}$ is the upper $\alpha/2$ percentage point of the central $t$ distribution with $n-2$ degrees of freedom. This is motivated by the fact that:

$$P_{H_0}(|t| \geq t_{\alpha/2,n-2}) = P_{H_0}(t \geq t_{\alpha/2,n-2}) + P_{H_0}(t \leq -t_{\alpha/2,n-2}) = 2P_{H_0}(t \geq t_{\alpha/2,n-2}) = 2 \times (\alpha/2) = \alpha$$

where the second equality comes from the symmetry of the student distribution. On R $t_{\alpha/2,n-2}$ can be obtained by `qt(alpha/2, n-2, lower.tail = F)`. Answer question by taking $\alpha = 0.05$.

(you can check your results by making `summary(lm(y ~ x, data = geyser))` and looking at the p-value in the summary, you reject the test at risk $\alpha$ if p-value $\leq \alpha$)

(c) Now come back to

$$t = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t(n-2)$$

(here we do not assume $H_0 : \beta_1 = 0$ any more)

Thus the formula can be obtained page 133 of the book: the following $100(1-\alpha)$ confidence interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{s}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Compute the confidence interval with $\alpha = 0.05$.

(you can check you results with `confint(lm(y~x, data = geyser))`)

(d)

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

(you can also check that $r^2 = \left(\frac{\text{cov}(x,y)}{\sigma_x \sigma_y}\right)^2$) where $\frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$ is the linear correlation coefficient (you can use `cor(geyser$x, geyser$y)` to get this coefficient)

(you can also look at `Multiple R-squared` value when making `summary(lm(y ~ x, data = geyser))`)

## Exercise 7.53

Import the data:

```
gas = read.csv("gas.csv",sep=";")
head(gas)
```

```
##    y x1 x2   x3   x4
## 1 29 33 53 3.32 3.42
## 2 24 31 36 3.10 3.26
## 3 26 33 51 3.18 3.18
## 4 22 37 51 3.39 3.08
## 5 27 36 54 3.20 3.41
## 6 21 35 35 3.03 3.03
```

```
X = cbind(1,as.matrix(gas[,-1])) # add a first column of 1 for the intercept
head(X)
```

```
##        x1 x2   x3   x4
## [1,] 1 33 53 3.32 3.42
## [2,] 1 31 36 3.10 3.26
## [3,] 1 33 51 3.18 3.18
## [4,] 1 37 51 3.39 3.08
## [5,] 1 36 54 3.20 3.41
## [6,] 1 35 35 3.03 3.03
```

```
y = gas[,1]
```

(a)

Use the formula

$$\hat{\beta} = (X'X)^{-1}X'y$$

With R, the function `t` stands for transposition, `solve` for matrix inversion, and the operator `%*%` stand for matrix multiplication.

And $s^2$ is given by

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

You can compare you results with the results of `lm`:

```
reg = lm(y ~ ., data = gas)
reg$coefficients # hat beta
```

```
## (Intercept)          x1          x2          x3          x4 
##  1.01501756 -0.02860886  0.21581693 -4.32005167  8.97488928
```

```
reg_summary = summary(reg)
reg_summary$sigma # s
```

```
## [1] 2.729995
```

(b) We know that $\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ and $\sigma^2$ can be estimated by $s^2$ thus

$$\widehat{\text{cov}(\hat{\beta})} = s^2(X'X)^{-1}$$

Thus do the computation of $\widehat{\text{cov}(\hat{\beta})}$ using previous results:

You can check the value of $(X'X)^{-1}$

```
reg_summary$cov.unscaled
```

```
##              (Intercept)            x1            x2          x3          x4
## (Intercept)  0.464850012  0.0019446557 -0.0085652766 -0.15591023  0.14388182
## x1           0.001944656  0.0011014057 -0.0002584188 -0.02187407  0.01051298
## x2          -0.008565277 -0.0002584188  0.0006152966  0.01394149 -0.01677282
## x3          -0.155910227 -0.0218740653  0.0139414927  1.09058795 -0.96667144
## x4           0.143881820  0.0105129832 -0.0167728207 -0.96667144  1.03147930
```

or directly obtain $s^2(X'X)^{-1}$ by making:

```
vcov(reg)
```

```
##             (Intercept)           x1           x2         x3          x4
## (Intercept)  3.46446861  0.014493274 -0.063835928 -1.1619793  1.07233308
## x1           0.01449327  0.008208638 -0.001925963 -0.1630247  0.07835194
## x2          -0.06383593 -0.001925963  0.004585728  0.1039042 -0.12500572
## x3          -1.16197929 -0.163024654  0.103904189  8.1280146 -7.20448052
## x4           1.07233308  0.078351940 -0.125005720 -7.2044805  7.68748532
```

The diagonal elements gives the estimated variance of each parameter.

(c)

Find $R^2$ and $R_a^2$, you can use the formulas of exercise 7.29:

$$R^2 = 1 - SSE/\sum_i(y_i - \bar{y})^2$$

and

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{\sum_i (y_i - \bar{y})^2/(n-1)}$$

You can check you results:

```
reg_summary
```

```
## 
## Call:
## lm(formula = y ~ ., data = gas)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -5.586 -1.221 -0.118  1.320  5.106 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  1.01502    1.86131   0.545  0.59001   
## x1          -0.02861    0.09060  -0.316  0.75461   
## x2           0.21582    0.06772   3.187  0.00362 **
## x3          -4.32005    2.85097  -1.515  0.14132   
## x4           8.97489    2.77263   3.237  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.73 on 27 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9151 
## F-statistic: 84.54 on 4 and 27 DF,  p-value: 7.249e-15
```

where `Multiple R-squared` gives the $R^2$ and `Adjusted R-squared` gives $R_a^2$