Home Exercises 7

Your Name

6.11.2023

Write your name at the beginning of the file as "author:".

1. Return to Moodle by **9.00am, Mon 6.11.** (to section "BEFORE").
2. Watch the exercise session video available in Moodle by **10.00am, Mon 6.11.**
3. If you observe during the exercise session that your answers need some correction, return a corrected version to Moodle (to section "AFTER") by **9.00 am, Mon 13.11.**

Problem 1.

Read in the data from "prostate.txt" using command

```
pr = read.table("prostate.txt", as.is = TRUE, header = TRUE)
```

when the file is in the same directory as your .Rmd file.

These data are from *Stamey et al. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, Journal of Urology 141(5), 1076-1083.* They studied the level of prostate specific antigen (PSA) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy. The variables include the log(arithm) of PSA (lpsa), log cancer volume (lcavol), log prostate weight (lweight), age, log of benign prostatic hyperplasia amount (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45).

(i) Fit a linear model for `lcavol` that uses `lpsa` as the only predictor. Show `summary()` of the fitted model. How much variance does it explain?

(ii) Plot the four diagnostic plots of the linear model from part (i) using plot( ) command on the `lm`-object and a 2x2 plotting area. (See Lecture 7.) Explain what you should be looking at each of the four plots and whether you detect any problems with these diagnostic plots?

(iii) If an individual has value `lpsa = 3`, what is the predicted value for `lcavol` and what is its 95% prediction interval? (Use `predict( )` from Lecture 7).

Problem 2.

Continue with prostate data set from Problem 1.

(i) Fit a model for `lcavol` that uses variables `lpsa` and `lcp` as predictors. How much variance does it explain?

(ii) Make a histogram of `lcp`. Consider 5 individuals that all have `lpsa = 3` and they have different values for `lcp`, namely, -1, 0, 1 , 2 and 3, respectively. Make a data.frame that corresponds to such 5 individuals and has 5 rows and two columns (columns named `lpsa` and `lcp`). Apply `predict( )` function to the linear model fitted in part (i) to get the predicted values for `lcavol` with 95% prediction intervals for these 5 individuals.

(iii) Based on part (ii), if an individual has `lpsa = 3` and `lcavol = 4` would you consider that he rather has `lcp = -1` or `lcp = 3`?

Problem 3.

Let's continue with prostate cancer data from Problems 1 & 2.

(i) Fit linear model `lcavol ~ svi`. Use `summary( )` on the `lm`-object. What is the coefficient for `svi` in this model? What is its P-value? How much variation in `lcavol` the model explains?

(ii) Fit linear model `lcavol ~ lpsa + svi`. What is the coefficient for `svi` in this model? How much variation in `lcavol` the model explains? What has happened to P-value of `svi` compared to model in part (i)? What is your conclusion about predictive power of `svi` vs. `lpsa`?

(iii) Fit linear model `lcavol ~ lpsa + lcp + svi`. What is the coefficient for `svi` in this model? How much variation in `lcavol` the model explains? What has happened to P-value of `svi` compared to model in part (ii)? What is your conclusion about the predictive power of `svi` vs. `lpsa` and `lcp`?

Problem 4.

Let's study the data on social factors from lecture 7. Read it in using `y = read.csv("UN98.csv", as.is = TRUE, header = TRUE, sep =",")` as in lecture material and rename the columns using command

```
colnames(y) = c("country","region","tfr","contr","eduM","eduF","lifeM",
                "lifeF","infMor","GDP","econM","econF","illiM","illiF")
```

(Note: By default, the code block above is not evaluated by `Knit` because it has `eval = FALSE` in its initialization. You can either copy only the command to your own solution or set `eval = TRUE` in this code block after you have first read in the data set.)

Let's study the life expectancies in males (`lifeM`) and females (`lifeF`) as functions of total fertility rate (`tfr`) and infant mortality `infMor`.

(i) Plot histograms of `lifeM` and `lifeF` as well as a scatter plot where `lifeF` is on the x-axis and `lifeM` is on the y-axis. Which sex is typically having higher life expectancy? (Hint: You can add line y=x by `abline(0,1)` to make it easier to visually compare which value is larger.)

(ii) Fit linear models `lm.m` for `lifeM ~ tfr + infMor` and `lm.f` for `lifeF ~ tfr + infMor`. Is there a difference how `tfr` and `infMor` predicts life expectancy in males vs females? (Compare coefficients and total variance explained by the model.)

(iii) Add a column `lifeD` to data frame `y` as the difference between life expectancies of males and females by command `y$lifeD = y$lifeM - y$lifeF`. Are `infMor` and `tfr` important predictors of `lifeD` in linear regression and if so what kind of an effect they have on it?

(iv) In (iii) you saw how `lifeD` changes as function of `tfr`. Plot `tfr` on x-axis and `lifeD` on y-axis and determine from the plot for which kind of `tfr` values the difference in life expectancy between the sexes is the largest.