Home Exercises 6

30.10.2023

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

1. Return to Moodle by **9.00am, Mon 30.10.** (to section "BEFORE").
2. Watch the exercise session video available in Moodle by **10.00am, Mon 30.10.**
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 6.11.**

Problem 1. Read in file "systbp_ldlc.txt". (See Exercise set 1 for how to read it if you don't remember.) Check whether there are any NAs in the data frame by command `anyNA( )`.

(i) Plot `systbp` on x-axis and `ldlc` on y-axis and add their correlation value in the title of the plot.

(ii) Fit a linear model of `ldlc ~ systbp`. Add the regression line to the existing plot. Show the `summary()` of the model fit. Is `systbp` a significant predictor of `ldlc` at significance level 0.05? Is it a useful predictor for any practical purposes, for example, could you use this model in a clinical setting to predict reliably `ldlc` from measured `systbp`? (Hint: Look at the plot and the R-squared value and think whether there is any useful predictive power here.)

(iii) What is the 95%CI of the coefficient of `systbp` in this model?

Problem 2. Let's consider the prostate cancer data set from Exercise set 4. Read it in by `y = read.table("prostate.txt", as.is = T, header = T)` and apply `head(y)`. Our variables are: log of 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), percent of Gleason scores 4 or 5 (pgg45) and log(arithm) of PSA (lpsa).

(i) Visualize the correlation matrix of the 9 variables mentioned above by `corrplot.mixed()` from package `corrplot` (See Lecture example 6.1.4.) You can pick the 9 variables by indexing columns with `2:10` rather than writing the column names explicitly.

(ii) We want to predict `lcavol`. From corrplot we see that `lpsa` is a good candidate predictor as it is highly correlated with `lcavol`. Compute correlation with 95% confidence interval between `lcavol` and `lpsa`. For CI, use `r.con()` function from package `psych`.

(iii) Fit a linear model `lcavol ~ lpsa`. Print out its `summary()`. How much variation in `lcavol` does the model explain?

(iv) Plot values of `lpsa` on x-axis and `lcavol` on y-axis. Add the linear model fit to the same plot using `abline()` function.

(v) Estimate visually from the plot what is an average `lcavol` for an individual whose `lpsa = 2.0`.

(vi) Use the model coefficients from the linear model of (2.ii) to compute the exact linear model prediction for individual with `lpsa = 2.0`.

Problem 3. Continue with the prostate cancer data from Problem 2.

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

(ii) What is the formula by which the model from (3.i) turns the values of `lpsa` and `lcp` into the predicted value of `lcavol`?

(iii) If an individual has values `lpsa=3` and `lcp=-1`, what is the predicted value for `lcavol`.

Problem 4. Continue with the prostate cancer data from Problem 2 & 3.

Split data into two parts based on the `lpsa` values of the individuals. Part I are individuals with `lpsa` at most the median `lpsa` in these data, and Part II are individuals with `lpsa` above the median `lpsa` in these data. Fit separate linear regression models in each Part of the data where you regress `lcavol` on `age` of the individuals. Do you notice differences how age predicts `lcavol` depending on whether `lpsa` values are low or high?