

STK-2100

Mandatory assignment 1 of 2

Submission deadline

Thursday 7th of March 2024, 14:30 at Canvas.

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts.

You can choose between scanning handwritten notes or using a typesetting software for mathematics (e.g. LaTeX). Scanned pages must be clearly legible. The assignment must be submitted as a single PDF file. Remember to include any relevant programming code and resulting plots and figures, in the PDF-file.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, you may be asked to give an oral account.

Application for postponed delivery

If you experience technical problems, or need a postponement of the deadline, you must contact the student administration at the Department of Mathematics no later than the same day as the deadline: studieinfo@math.uio.no Note that teaching staff can not grant extensions.

All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

[Read the department's guidelines for mandatory assignments for more information.](#)

GOOD LUCK!

Problem 1. Which of the following models

- (a) $y = \beta_0/(1 + \beta_1 x) + \beta_2 x^{1/2} + \epsilon;$
- (b) $y = \beta_0 + \beta_1/x + \beta_2 x^2 + \epsilon;$
- (c) $y = 1/(\beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon);$
- (d) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \epsilon;$
- (e) $y = \beta_0 + \beta_1 x^{\beta_2} + \epsilon;$

can be written as linear regression models:

- (i) as they are;
- (ii) when a single parameter is held fixed;
- (iii) after a suitable transformation (i.e., use the same transformation on both y and $f(x)$)?

Give the form of the design matrix X , the vector of the regression parameters, and: in the case of (ii), which parameters should be held fixed; in the case of (iii), a transformation that can be used.

Problem 2. Consider the data from [Hastie & Tibshirani \(1990\)](#) about 43 children with diabetes (the R code to download the data, as provided by [Royston & Sauerbrei, 2004](#), is available on the course's web page). The goal is to relate the response **cpep**, the logarithm of C-peptide concentration at diagnosis, to two covariates, **age** (age at diagnosis) and **base** (a measure of acidity).

- (a) Fit a univariate linear regression model for each of the two covariates, contrast their (marginal) effects and comment on the results. If you had to choose to only use one covariate, would you have had a strong preference for one of the two?
- (b) Let us now focus on age. By plotting its values versus those of the response, it does not seem to have a linear effect. Try some polynomial transformations to capture this non-linearity. Do you see any improvement?
- (c) Considering the nature of the covariate **age**, which other transformation do you think should be considered? How are the results obtained by using this transformation when compared to those of point (b)?

- (d) Consider both covariates in a single model. Contrast the estimates of the regression coefficients and the R^2 to those obtained at point (a): what can we say about the two covariates? Do you think we have problems with collinearity here?

Problem 3. Consider the data collected by [Stamey et al. \(1989\)](#) to study potential predictors of prostate-specific antigen (PSA) in patients with prostate cancer (the R code to download the data, as provided by [Royston & Sauerbrei, 2004](#), is available on the course's web page). The response is the logarithm of the PSA concentration (`lpsa`), while the collected covariates are: `age` (age in years); `svi` (seminal vessel invasion, yes/no); `pgg45` (percentage Gleason score 4 or 5); `cavol` (cancer volume in ml); `weight` (prostate weight in grams); `bph` (amount of benign prostatic hyperplasia in grams); and `cp` (amount of capsular penetration in grams).

- (a) Try out the following regression methods: backward elimination or forward selection, lasso, ridge regression, and PCR or PLS). Provide the final model for each method and discuss the results.
- (b) If our goal is predicting the (logarithm of the) concentration of PSA for a new patient, which model would you suggest to use? Justify your answer, providing numerical evidence.
- (c) Does the model chosen at point (b) involve all of the available features? Why, in this case, is it better like this?

Bibliography

- HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. CRC Press.
- ROYSTON, P. & SAUERBREI, W. (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine* **23**, 723–748.
- STAMEY, T. A., KABALIN, J. N., MCNEAL, J. E., JOHNSTONE, I. M., FREIHA, F., REDWINE, E. A. & YANG, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. Radical prostatectomy treated patients. *The Journal of Urology* **141**, 1076–1083.