6th February, 2020

STK-2100

Mandatory assignment 1 of 2

Submission deadline

Thursday 20th February 2020, 14:30 at Canvas.

Instructions

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LaTeX). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. 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, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (e-mail: studieinfo@math.uio.no) well before the deadline.

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

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

GOOD LUCK!

Problem 1. Consider the subsample of 326 US women, from a study of Luke et al. (1997) on the relationship between percentage body fat content `pbfm` and body-mass index `bmi`, that you can download from:
https://www.imbi.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/res_bodyfat.zip
The aim of the study was to find how well `bmi`, easily computed as the ratio between weight and squared hight (so measured in kg/m^2), can be used to predict `pbfm`, whose measurement involves instead a complex bioelectrical impedance analysis (Royston & Sauerbrei, 2008).

(a) Plot the data and fit a simple linear Gaussian model. On this model, perform a graphical diagnostic analysis and comment on its results, focusing, in particular, to what the Ascombe plot reveals.

(b) From the graphical results of the previous point, would you suggest to use a logarithmic or a quadratic transformation of the explanatory variable? Why? Fit a model with the explanatory variable logarithmically transformed and a model in which both the linear and the quadratic effect of the explanatory variable are included: report the results for both models and comment on the results.

(c) A polynomial of higher order may be beneficial here: Use cross-validation to select the best (in terms of mean square error of the resulting model) degree of the polynomial (try up to a polynomial of order 10) and report the related model.

(d) Repeat the analysis of point (c) using a procedure based on an information criterion. Do you select the same model?

Problem 2. From a large population-based case–control study on Oral cancer conducted in the US (Day et al., 1993), the data related to the African American population (194 cases, `ccstatus` $= 1$, and 203 controls, `ccstatus` $= 0$) have been selected. You can download them from:
https://www.imbi.uni-freiburg.de/imbi/Royston-Sauerbrei-book/Multivariable_Model-building/downloads/datasets/oral_ca.zip
The aim of the study was to evaluate the risk of Oral cancer based on the variables `drinks` (number of 1oz ethanol-equivalent drinks consumed per week), `sex`, `age` and `cigs` (number of cigarettes smoked per day).

We are first interested in the effect of smoking alone:

(a) Dichotomize the variable `cigs` in two categories, smokers and not smokers, and create a table with the observed frequencies for cases and controls. In addition, provide the estimated probabilities (including their standard errors) of experiencing an oral cancer (i.e., being a case) for the two sub-populations.

(b) Test the hypothesis that the two probabilities are equal and comment on the result.

(c) Fit a linear logistic model using the dichotomized variable as explanatory variable and comment on the result: does being a smoker increase or decrease the risk of experiencing oral cancer? How much, in terms of log-odds?

(d) Repeat the analysis of point (c) by considering, now, the number of cigarettes as a continuous variable. Comment on the result: What does the regression coefficient for this variable mean now? Why does the value of the intercept change with respect to point (c), although they are both related to the odds for non-smokers?

Consider now the other three variables (`drinks`, `sex` and `age`) as well:

(e) Fit a linear logistic model including all the explanatory variables and report the result. What is the increase in terms of log-odds for an increasing number of cigarettes per day smoked estimated by this model? Why did it change from the one obtained in model fitted in point (c)?

(f) What about `age`? Do older people risk more? Is that statistically significant? Fit a model excluding `age` from the explanatory variables and compare the resulting model with the one obtained at point (e). Which one would you keep?

(g) Use a polynomial of degree 2 to model the effect of `drinks`. Does it improve the model?

(h) What about doing that for `cigs` instead of `drinks`? What is the effect on the model?

(i) For both `drinks` and `cigs`, is there any advantage in adding a cubic term?

(j) What does it mean to include a second order effect in the model? Does it tell us anything about the importance of the risk factor? If yes, what does it tell us? If no, what does it tell us instead?

(k) Split the data in a training and a test set and fit the models of the previous points. Report their training and test error and comment on the results. Based on these results, which one is the best model for prediction?

Bibliography

Day, G. L., Blot, W. J., Austin, D. F., Bernstein, L., Greenberg, R. S., Preston-Martin, S., Schoenberg, J. B., Winn, D. M., McLaughlin, J. K. & Fraumeni Jr, J. F. (1993). Racial differences in risk of oral and pharyngeal cancer: alcohol, tobacco, and other determinants. *JNCI: Journal of the National Cancer Institute* **85**, 465–473.

Luke, A., Durazo-Arvizu, R., Rotimi, C., Prewitt, T. E., Forrester, T., Wilks, R., Ogunbiyi, O. J., Schoeller, D. A., McGee, D. & Cooper, R. S. (1997). Relation between body mass index and body fat in black population samples from nigeria, jamaica, and the united states. *American Journal of Epidemiology* **145**, 620–628.

Royston, P. & Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. Wiley, Chichester.