Series 11

1. a) Consider the linear regression model

$$y_i = \mu + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, ..., n. \tag{1}$$

Define $\beta := (\beta_1, ..., \beta_p)^T$ and the generalized residuals as

$$r_i(\beta) := y_i - \mu - \sum_{j=1}^{p} \beta_j \cdot x_{ij}, \qquad i = 1, ..., n. \tag{2}$$

Show that taking $\mu = \bar{y} - \sum_{j=1}^{p} \beta_j \bar{x}_{.j}$ the generalized residuals can be written as

$$r_i(\beta) = y_i - \bar{y} - \sum_{j=1}^{p} \beta_j (x_{ij} - \bar{x}_{.j}), \qquad i = 1, ..., n, \tag{3}$$

where $\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$. Note that in equations (2) and (3), $\beta_1, ..., \beta_p$ are the *same*. Hence by centering the response and predictor variables it is always possible to get rid of the intercept μ in equation (1) to estimate $\beta_1, ..., \beta_p$. Moreover, having equal generalized residuals implies having the same value of the Residual Sum of Squares, and therefore the Least Squares estimation of $\beta_1, ..., \beta_p$ are the same in both the model with an intercept and in the model without it.

b) (Optional) Show that the ridge-regression solution defined as

$$\tilde{\boldsymbol{\beta}}^*(s) = \underset{\|\boldsymbol{\beta}\|^2 \le s}{\arg\min} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2$$

is given by

$$\hat{\boldsymbol{\beta}}^*(\lambda) = (X^\intercal X + \lambda \mathbb{I})^{-1} X^\intercal \mathbf{Y}.$$

where λ is a suitably chosen Lagrange-multiplicator. Therefore the ridge estimator is still linearly depending on the response $\mathbf{Y}$. Note that for λ large enough the ridge solution exists even if $X^\intercal X$ does not have full rank or if it is computationally close to singular. Therefore ridge regression is practicable also if $n \ll p$.
Hint: Use the method of Lagrange multipliers with one-sided inequality constraint from convex optimization.

In sub-tasks **c)** and **d**) we will use bold lower case letters to denote vectors and upper case letters to denote matrices.

c) Let $n \ge p$. The *ridge traces* $\hat{\boldsymbol{\beta}}^*(\lambda)$ can be determined computationally easily by using the *singular value decomposition* of the data matrix $X = UDV^\intercal$, where $U(n \times p)$ and $V(p \times p)$ are orthogonal and D is diagonal. Use the result of **b)** to show that:

$$\hat{\boldsymbol{\beta}}^*(\lambda) = V(D^2 + \lambda \mathbb{I})^{-1} D U^\intercal \mathbf{y}.$$

d) Show that the ridge regression fit is just a linear combination of shrinked response-components y_i with respect to the orthogonal basis defined by U. More explicitly show that:

$$\hat{\mathbf{y}}_{ridge}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u_j} \mathbf{u_j}^\intercal \mathbf{y},$$

where d_j, $j = 1, ..., p$, are the diagonal elements of D and $\mathbf{u_i}$, $i = 1, ..., p$, are the columns of U. In fact one can show that the directions defined by $\mathbf{u_j}$ are the so called *principal components* of the dataset X. The smaller the corresponding d_j-value, the smaller the data variance in direction $\mathbf{u_j}$. For directions with small data variance, the gradient estimation for the minimization problem is difficult, therefore ridge regression shrinks the corresponding coefficients the most.

e) Ridge regression can also be motivated by Bayesian theory. We assume that

$$\mathbf{Y}|\beta \sim \mathcal{N}(X\beta, \sigma^2 I) \text{ and } \beta \sim \mathcal{N}(\mathbf{0}, \tau I).$$

Show that the ridge estimator $\hat{\beta}^*(\lambda)$ is the maximum a posteriori estimator (MAP) and deduce that it corresponds to the mean of the posterior distribution. What is the relationship between λ, τ and σ^2?

2. In this task we revisit the ozone dataset that you have already encountered several times. You may get the data as indicated in Series 8. The aim here is to get acquainted with penalized regresssion methods, i.e. ridge regression, the lasso and elastic net, in R .

a) Load the data, apply a log transformation on the response `upo3` and remove the outlier (observation number 92) as done in Series 8.

b) Generate a R -formula and the according design matrix for a cubic penalized regression model that accounts for all 3-way interactions.
R-Hints:
Use the `wrapFormula` function of the `sfsmisc` package to set up formula and `model.matrix` to get the design matrix.

```
require(sfsmisc)
ff <- wrapFormula(logupo3 ~ ., data=?, wrapString="poly(*,degree=?)")
ff <- update(ff, logupo3 ~ ?)
mm <- model.matrix(?, data=?)
```

c) Fit a cubic penalized regression model that accounts for all 3-way interactions to the data. Use ridge and lasso regression for the regularization problem. Plot the ridge and lasso traces. How do they differ?
R-Hints:
To perform penalized regression via ridging and lasso use the `glmnet` function in the package of the same name.

```
require(glmnet)
ridge <- glmnet(mm, ?, alpha=?)
lasso <- glmnet(mm, ?, alpha=?)
plot(?, xvar="lambda")
```

d) Select an optimal tuning parameter λ with an elastic net penalty $\alpha = 0.5$ via 10-fold cross validation. Find an optimal λ according to the "1-std error rule" from a plot that shows the mean squared error as a function of $\log(\lambda)$.
R-Hints:
To perform cross validation for the elastic net use the `cv.glmnet` function

```
set.seed(1)
cv.eln <- cv.glmnet(?,?,alpha=?, nfolds=?)
plot(cv.eln)
```

e) Compare your results from **d**) with findings from Series 8. Which model is more suited for prediction: the `gam` (Series 8) or the elastic net model?

Preliminary discussion: Friday, May 27.

Deadline: Friday, June 03.

Question hour: Monday, 04.07.2016 and Friday, 05.08.2016, both 15:00 - 16:00, HG G 26.1 .

Exam Consultation: Monday, 26.09.2016, 12 p.m. - 1 p.m., HG G26.1 .