

TMA4267 - Linear Statistical Models

Exercise 7 - V2014

May 17, 2014

Last changes: Problem 3a, 0.4 changed to 40 for penalty parameter to yield correct solution.

Problem 1: Gasoline data

Consider the multiple linear regression model, with response vector \mathbf{Y} of dimension $(n \times 1)$ and p covariates in \mathbf{X} (dimension $(n \times p)$).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether Y , the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

x_1 :	TankTemp	tank temperature ($^{\circ}\text{F}$)
x_2 :	GasTemp	gasoline temperature ($^{\circ}\text{F}$)
x_3 :	TankPres	vapor pressure in tank (psi)
x_4 :	GasPres	vapor pressure of gasoline (psi)

The data set is called `sniffer.dat`.

Since many of the methods we have studied in Part 7 (Model selection, shrinkage and dimension reduction) will perform differently with standardized variables, for simplicity we start by standardizing the covariates (make the mean 0 and the variance 1).

- a) Standardize the covariates, and then make scatter plots of the response and the covariates. Does this point to a MLR model?

Calculate the estimated covariance matrix of the standardized covariates. Do you see a potential problem here?

<http://www.math.ntnu.no/~bakke/TMA4267/2015V/sniffer.dat>

```
ds <- read.table("http://www.math.ntnu.no/~mottela/  
TMA4267/V2014/sniffer.dat", header=TRUE)  
x <- apply(ds[, -5], 2, scale)  
y <- ds[, 5]  
cov(x)
```

- b) Fit a MLR, with all four covariates. Find confidence intervals for the regression coefficients. Explain the fitted model. Assess the model fit.

```

full <- lm(y~x)
summary(full)
confint(full)
plot(full$fitted,rstudent(full))
qqnorm(rstudent(full))
qqline(rstudent(full))

```

- c) Based on the fitted full model in b), suggest a reduced model. Find confidence intervals (CI) for the regression coefficients. Compare the CIs to the CIs found in b). Test, using an F-test, if the reduced model is preferable to the full model. Explain your findings.

```

red <- lm(y~x[,c(2,4)])
summary(red)
confint(red)
anova(full,red)

```

- c) We now turn to best subset selection. Perform best subset selection using R_{adj}^2 and Mallows C_p to choose the best model. How does this model relate to the model in b)?

Remark: in our textbook “An Introduction to Statistical Learning, with Applications to R” (ISLR), chapter 6, AIC is seen to be closely related to Mallows C_p . Different implementations of Mallows C_p exists. ISLR suggests to choose the model with the smallest AIC, and due to the connection to Mallows C_p then also to choose the model with the smallest Mallows C_p . We will do so here. With other implementations of Mallows C_p the model with Mallows C_p closest to the number of parameters estimated is chosen.

```

library(leaps)
bests <- regsubsets(x,y)
sumbests <- summary(bests)
which.max(sumbests$adjr2)
which.min(sumbests$cp)

```

- d) Then turn to shrinkage methods. Fit ridge regression and lasso regression to the data. Use crossvalidation to choose the optimal value for the penalty parameter.

Compare the ridge coefficients with the coefficients in the full model. Compare the lasso model with the models in b) and c).

```

library(glmnet)
fit.ridge=glmnet(x,y,alpha=0)
plot(fit.ridge,xvar="lambda",label=TRUE)
cv.ridge=cv.glmnet(x,y,alpha=0)
plot(cv.ridge)
coef(cv.ridge)

fit.lasso=glmnet(x,y)
plot(fit.lasso,xvar="lambda",label=TRUE)
cv.lasso=cv.glmnet(x,y)
plot(cv.lasso)
coef(cv.lasso)

```

- e) Finally, turn to Principal component regression. How many PCs will you use? Compare the final estimated coefficients with the estimated coefficients of the full model.

```
res <- prcomp(x,scale=FALSE) #already scaled
biplot(res)
plot(res)
summary(res)
res$rotation
z <- res$x[,1:2]
fit <- lm(y~z)
summary(fit)
# for interpretation, back to original data
betas <- res$rotation[,1:2]%*matrix(fit$coeff[2:3],ncol=1)
betas
# compared to the findings for the full model
cbind(full$coeff[-1],betas)
```

- f) Compare the estimated coefficients for the models in b)-e). Do you see any trend in the results? How would you compare the results from ridge regression and PCR?

Problem 2: Best subsets and stepwise methods

Answer exercise 1 in chapter 6.8 (page 259) of An Introduction to Statistical Learning with Applications in R (ISLR).

Problem 3: Ridge regression

This problem is taken, with permission from Wessel van Wieringen, from a course in High-dimensional data analysis at Vrije University of Amsterdam.

- a) Find the ridge regression solution for the data below for a general value of λ and for the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ (only apply the ridge penalty to the slope parameter, not to the intercept). Show that when λ is chosen as 40, the ridge solution fit is $\hat{Y} = 40 + 1.75X$.

Data: $\mathbf{X}^T = (X_1, X_2, \dots, X_8)^T = (-2, -1, -1, -1, 0, 1, 2, 2)^T$, and $\mathbf{Y}^T = (Y_1, Y_2, \dots, Y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

- b) The coefficients β of a linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, are estimated by $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. The associated fitted values then given by $\hat{\mathbf{Y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{H} \mathbf{Y}$, where $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. The matrix \mathbf{H} is a projection matrix and satisfies $\mathbf{H} = \mathbf{H}^2$. Hence, linear regression projects the response \mathbf{Y} onto the vector space spanned by the columns of \mathbf{X} . Consequently, the residuals $\hat{\varepsilon}$ and $\hat{\mathbf{Y}}$ are orthogonal.

Next, consider the ridge estimator of the regression coefficients: $\hat{\beta}(\lambda) = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^T \mathbf{Y}$. Let $\hat{\mathbf{Y}}(\lambda) = \mathbf{X} \hat{\beta}(\lambda)$ be the vector of associated fitted values.

Show that the matrix $\mathbf{Q} = \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_{pp})^{-1} \mathbf{X}^T$, associated with ridge regression, is not a projection matrix (for any $\lambda > 0$). Hint: a projection matrix is idempotent.

- c) Show that the ‘ridge fit’ $\hat{\mathbf{Y}}(\lambda)$ is not orthogonal to the associated ‘ridge residuals’ $\hat{\varepsilon}(\lambda)$ (for any $\lambda > 0$).

Problem 4: The cork data and PCA

We have worked with the data on bark deposits on cork trees in the lectures. The data set is available from

```
corkds <-  
as.matrix(read.table("http://www.math.ntnu.no/~mettela/  
TMA4267/Data/corkMKB.txt"))  
dimnames(corkds)[[2]] <- c("N", "E", "S", "W")
```

The four column represent the weight of bark deposits in the directions North, East, South and West, thus the column names set above.

- a) Perform principal component analysis (PCA) on the cork data.
Do this both using the `prcomp` function and by programming this yourself using eigen value analysis or singular value decomposition.
Should the PCA be carried out on the covariance or the correlation matrix?
- b) Give an interpretation of the first two principal components. How many principal components would you use?
- c) If we assume that the cork data is a random sample from a multivariate normal population of cork trees. How may you interpret and graphically present the first two principal components?