Assignment 4. Statistics 9850/1

March 16, 2011. Due March 30.

1. Degrees of Freedom

Show that for linear regression with p inputs, intercept term included,

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i) = p$$

2. Homicide Data

Homicide data for 1961-73 for the city of Detroit is available in the dataframe Detroit in the rda file, Detroit.rda. You have use the load() or attach() function in R to access this rda-file.

The data frame as $n = 13$ observations on the following $p = 13$ input variables and output variable, HOM.

FTP.1
: Full-time police per 100,000 population

UEMP.2
: Percent unemployed in the population

MAN.3
: Number of manufacturing workers in thousands

LIC.4
: Number of handgun licences per 100,000 population

GR.5
: Number of handgun registrations per 100,000 population

CLEAR.6
: Percent homicides cleared by arrests

WM.7
: Number of white males in the population

NMAN.8
: Number of non-manufacturing workers in thousands

GOV.9
: Number of government workers in thousands

HE.10
: Average hourly earnings

WE.11
: Average weekly earnings

ACC
: Death rate in accidents per 100,000 population

ASR

Number of assaults per 100,000 population

HOM

Number of homicides per 100,000 of population

Use subset selection, ridge and lasso to find the most important variables.

Find the first two principal components and discuss what this suggests.

3. GLM, Box-Cox Analysis and Loess

The file ShipUnload.dat contains data on time to unload (output) for ships with various loads of cargo (Tonnage). Fit a model to describe and summarize the relationship. Try Box-Cox analysis, loess and a GLM with the gamma distribution.

4. Spectroscopy data

A Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle was used to collect data on samples of finely chopped pure meat. 215 samples were measured. For each sample, the fat content was measured along with a 100 channel spectrum of absorbances. Since determining the fat content via analytical chemistry is time consuming we would like to build a model to predict the fat content of new samples using the 100 absorbances which can be measured more easily.

The R workspace, **`MEAT.Rdata`**, on our course website, contains the dataframe, `xmeatspec` corresponding columns V1 to V100 correspond to absorbances across 100 wavelengths and is the fat content.

Divide the data into training and test data. Use about 1/3 of the data for the test sample.

Use the R pls package to fit PCR and PLS models to the training data. Compare the estimated EPE obtained by cross-validation with the prediction error on the test data.

Also fit using LASSO and Ridge regression for the training samples and compare the predictions on the training and test samples.