IND522
Advanced Statistical Modelling

Fall 2025

Professor Orhan FEYZİOĞLU

Assignment ♯6: Due December 25

1) The **peanuts** data set contains measurements of the alfatoxin (**X**) and the corresponding percentage of noncontaminated peanuts in the batch (**Y**). Do a scatterplot of these data. What is a good model for these data? Use cross-validation to choose the best model.

2) Use Monte Carlo simulation to compare the performance of the bootstrap and the jackknife methods for estimating the standard error and bias of the sample second central moment. For every Monte Carlo trial, generate 100 standard normal random variables and calculate the bootstrap and jackknife estimates of the standard error and bias. Show the distribution of the bootstrap estimates (of bias and standard error) and the jackknife estimates (of bias and standard error) in a histogram or a box plot. Make some comparisons of the two methods.

3) Using the **law** data set, find the jackknife replicates of the median. How many different values are there? What is the jackknife estimate of the standard error of the median? Use the bootstrap method to get an estimate of the standard error of the median. Compare two estimates of the standard error of the median.

4) Generate data according to $y = 4x^3 + 6x^2\text{–}1 + \epsilon$, where ϵ represents some noise with constant variance. Fit a first-degree model to it and plot the residuals versus the observed predictor values x_i (residual dependence plot). Construct also box plots and histograms of the residuals. Do they show that the model is not adequate? Repeat for $d = 2, 3$.

5) Fit the **wais** data using probit and the complementary log log link. Compare the results with the logistic regression using the techniques to assess model fit.

6) Convert the number of satellites in the horseshoe crab data to a binomial response variable, keeping the carapace width as the predictor. Set $Y = 1$ if a female crab has one or more satellites, and $Y = 1$ if she has no satellites. Fit various models and assess the results.

7) The **cpunish** data contains $n = 17$ observations. The response variable is the number of times that capital punishment is implemented in a state for the year 1997. The explanatory variables are median per capita income (dollars), the percent of the population living in poverty, the percent of African-American citizens in the population, the log(rate) of violent crimes, a variable indicating whether a state is in the South, and the proportion of the population with a college degree. Use the **glmfit** function to fit a model using the Poisson link function. Note that **glmfit** automatically puts a column of ones for the constant term. Get the deviance (**dev**) and the statistics (**stats**). What are the estimated coefficients and their standard errors? Find the 95% Wald confidence intervals for each estimated coefficient.

8) Apply the penalized approaches —ridge regression, lasso, and elastic net— to the **airfoil** data. Compare your results with subset selection.