STAT 408 Homework 4

Due by 11:55 pm, Sunday, 11/6/2022

50 points

Please provide detailed calculation and explanation in your solution. Points will be deducted for skimpily written answers. This homework will also require coding in R. On the coding part, the homework solutions should also include detailed description, R code, and output. Write your answers, scan them, and combine to a single pdf file. Name this file as yourname_hw4 and upload to Sakai.

1. (15 points) Use the prostate data with lpsa as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:

    a. Backward elimination (0.05 cutoff)
    b. AIC
    c. Do these model selection methods give you the same result? If not, do you think it is an issue that they are different? Give your insight.

2. (15 points) The aatemp data come from the U.S. Historical Climatology Network. They are the annual mean temperatures (in degrees F) in Ann Arbor, Michigan going back about 150 years. Download this dataset from Sakai and answer following questions.

    a. Fit a linear model of temp~year. Do you think there is a linear trend? (Hint: check plot, parameters, and model goodness of fit)
    b. Observations in successive years may be correlated. Fit a model that estimates this correlation. Does this change your opinion about the trend?
    c. Fit a polynomial model with degree 5. Plot your fitted model on top of the data.
    d. Suppose someone claims that the temperature trend was different before and after 1930. Fit a segmented regression model to check this claim.

3. (15 points) The "longley" dataset includes the following seven social-economic variables from 1947-1962 in the US:

GNP.deflator: GNP implicit price deflator (1954=100)
GNP: Gross National Product.
Unemployed: number of unemployed.
Armed.Forces: number of people in the armed forces.
Population: population $\geqslant$ 14 years of age.
Year: the year (time).
Employed: number of people employed.

Our goal is to explore the relationship between Employed and other variables. Download this dataset from Sakai and answer the following questions.

a. Construct a correlation matrix of six predictors in this dataset. Which predictors do you think are highly correlated? What are the potential reasons for those high correlations?
b. Regress each predictor on others to examine the collinearity. Do you have same conclusion as in (a)?
c. Try to remove some highly correlated predictors. Compare the full model and the smaller model. Do you think the smaller model is better? Give you reason.

4. (15 points) The gala dataset contains 30 Galapagos islands and 7 variables. The relationship between the number of plant species and several geographic variables is of interest.

`Species`
: the number of plant species found on the island

`Endemics`
: the number of endemic species

`Area`
: the area of the island (km^2)

`Elevation`
: the highest elevation of the island (m)

`Nearest`
: the distance from the nearest island (km)

`Scruz`
: the distance from Santa Cruz island (km)

`Adjacent`
: the area of the adjacent island (square km)

The dataset galamiss contains the Galapagos data with missing values left in. Use two datasets to answer following questions.

a. Fit a linear model using gala (the data without missing) with the number of species as the response and the five geographic predictors (without Endemics).
b. In galamiss, which variable(s) includes missing value? How many missing values do we have?
c. Fit the same linear model to galamiss using the deletion strategy for missing values. Compare the fit to that in (a).
d. Use mean value imputation on galamiss and again fit the model. Compare to previous fits.
e. Use a regression-based imputation based on the other four geographic predictors to fill in the missing values in galamiss. Fit the same model and compare to previous fits.