STAT 210
Applied Statistics and Data Analysis:
Homework 9

Due on Nov. 20/2022

Question 1

Consider the `prostate` data set in the `faraway` package. Consider `lpsa` as the response variable and exclude the variables `svi` and `gleason` from the analysis.

(a) Do an exploratory analysis of the data. Do a matrix of plots. Which variables seem to have a linear relationship with the response? Compute and plot the correlation coefficients for the regressors. Comment on what you obtain.

(b) Fit a model for `lpsa` with all the other variables as predictors. Calculate the variance inflation factors and eliminate variables with vif greater than two.

(c) Starting with the variables selected in (b), do a variable selection procedure using backward elimination with a p to remove equal to 0.15. Do also variable selection using the BIC criterion. Compare the models that you get. Do residual analysis for both of them. Comment on your results.

(d) Which model would you select and why?

(e) Suppose a new patient with the following values arrives:

Table 1: Variables for a new patient

lcavol	lweight	age	Ibph	lcp	pgg45
1.44692	3.623	65	0.30	-0.799	15.0

Predict the `lpsa` for this patient along with appropriate 98% prediction and confidence intervals.

Question 2

For this question use the data set `Birthweight.csv`. We will consider only the variables `birthwt`, `mppwt` and `smoker`. They represent the weight of the baby at birth, the weight of the mother before pregnancy and whether the mother smokes, with 1 indicating that the mother is a smoker.

(i) Subset the data corresponding to the variables mentioned above. Plot `birthwt` against `mppwt` and color the dots according to the value of `smoker`. Add a regression line for `birthwt` against `mppwt`. Comment. Print the summary table for the regression and interpret the results.

(ii) We want to add `smoker` as a categorical regressor to the previous model. Fit a complete model including interaction and work your way to a minimal adequate model. Write down the equation for you final model and interpret the coefficients.

(iii) Draw a scatter plot of `birthwt` against `mppwt` and color the dots according to the value of `smoker`. Add the regression lines for your model. Predict the `birthwt` value for a `mppwt` value of 120 and both values for `smoker`. Add prediction intervals at the 98% level.

(iv) State clearly the assumptions on which the regression model is based. Using graphs and hypothesis tests, do a diagnostic analysis for the model you fitted and verify whether these assumptions are satisfied.