**Mini project #6**

**Group Member:** Chaoran Li, Wenting Wang

**Contribution of each member:**

Firstly, we discussed the mathematical models and code details together. Then, we divided the project into two part and finished our respective work. Chaoran Li worked on coding and Wenting Wang worked on analyzing. Then, we checked and reviewed our report together. Each member makes contribution to this project as the details shown in table 1.

|  |  |
| --- | --- |
|  | Question1 |
| Chaoran li | 50% |
| Wenting wang | 50% |

Table 1: Member contribution table

**Question 1:**

Build a “reasonably good” linear model for PSA level in prostate.cancer.csv data set.

1) Load and prepare data

Text

Description automatically generated

Text

Description automatically generated

2) Analyze the data with simple linear regression first

a) cancervol:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

b) weight:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

c) age:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

d) benpros:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

e) factor(vesinv)1:

Chart

Description automatically generatedChart

Description automatically generated

Text

Description automatically generated

f) capspen:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

g) gleason:

Chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

Nearly all regressions with log(pas) are better than that with psa, except for capspen. And regressions with log(pas) show better linear trend in our observation. Even for capspen, the diagram shows a good linear trend. Hence, we would use log(pas) afterwards.

3) Multiple linear regression

a) Start with full model

Table

Description automatically generated

b) Drop age

Table

Description automatically generated

c) Drop weight

Table

Description automatically generated

d) Drop capspen (which prefer pas better than log(psa))

Table

Description automatically generated

Reject all H0 here. Compare existing models:

Text

Description automatically generated

We stop here. For it seems that we dropped 3 parameters which are clear not important and the other 4 parameters show good linear trend.

4) Verify by stepwise selection with BIC

a) Forward selection:

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Text

Description automatically generated

A picture containing text

Description automatically generated

2) Backward elimination:

Text

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

A picture containing table

Description automatically generated

3) Stepwise selection:

A picture containing text, table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

All stepwise selections choose the save model as fit4 which is chosen manually by us.

5) Verify the model assumptions

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generated

6) Predict the PSA level

Predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

Table

Description automatically generated

Graphical user interface, text, application

Description automatically generated

10.28357