**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) Firstly, we need to prepare and explore the data. Because the vesinv is quantitative variable, we need to convert it to factor with 1 dummy variable.

Text

Description automatically generated

Text

Description automatically generated

2) Then, we analyze the data with simple linear regression for both psa and the log transformation log(psa).

a) cancervol: From the two plots below, we can find that, as univariate regression, log(psa) performs much better positive linear trend.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

b) weight: From the two plots below, we can find that, as univariate regression, both psa and log(psa) do not perform very good linear trend, but log(psa) is better.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

c) age: From the two plots below, we can find that, as univariate regression, log(psa) performs much better positive linear trend.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

d) benpros: From the two plots below, we can find that, as univariate regression, psa shows negative correlation, but log(psa) shows positive linear relationship.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

e) factor(vesinv)1: Because vesinv is a qualitative variable, there exist only two possible values i.e. 0 or 1. And there are more people have Seminal vesicle invasion, who are the people with dummy variable 0.

Chart

Description automatically generatedChart

Description automatically generated

Text

Description automatically generated

f) capspen: From the two plots below, we can find that, as univariate regression, both psa and log(psa) show linear relationship, but psa shows better positive linear relationship.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

g) gleason: Although there only exist three value for gleason, we still can find that, as univariate regression, log(psa) performs much better positive linear trend.

Chart

Description automatically generatedChart, scatter chart

Description automatically generated

Text

Description automatically generated

Above all, nearly all regressions with log(pas) are better than that with psa. Hence, we would use log(pas) afterwards.

3) Next, we explore multiple linear regression

a) Start with full model:

Table

Description automatically generated

From the Coefficients, we find the P-value of age is 0.81186, which fail to reject the Null Hypothesis . Thus, we can try to drop the most impossible predictor ‘age’.

b) Drop age

Table

Description automatically generated

‘Weight’ has a P-value of 0.45643 far more large than 0.05. Hence, we can drop the predictor ‘weight’.

c) Drop weight

Table

Description automatically generated

‘Capspen’ has a P-value of 0.413237 far more large than 0.05. Hence, we can drop the predictor ‘capspen’. Besides, ‘capspen’ is the only predictor which prefers psa better than log(psa). If we drop it, the final model would have better linear trend.

d) Drop capspen

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 our result by model selection with BIC

We would use all three kind of model selections and check the results with the model we got above.

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 three model selections choose the model as fit4 which is chosen manually by us. For there exist a clear boundary in this question and we can clearly find out where to stop.

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

Assumptions:

a) Errors have mean zero and constant variance.

Proved by residual plot. The horizontal line nearly split all data points in half. The points are scattered around zero and have less pattern. Hence, errors have nearly mean zero and constant variance.

b) Errors are normally distributed.

Proved by QQ plot. The QQ plot nearly fits the QQ line which means that the errors are nearly normal distributed.

c) Errors are independent.

Proved by time series plot. In all, the time series plot shows a positive trend while index increasing. But this trend is quite tiny. We can roughly announced that errors are independent.

6) Predict the PSA level

From the QQ plot of residuals above, we see that residuals hold normalization very well. Thus, we can conclude that our model assumptions hold and we can choose model fit4 as our ‘reasonably good’ model.

PSA = exp(-0.65013 + 0.06488\*(cancervol) + 0.09136\*(benpros) +

0.68421\*(factor(vesinv)1) + 0.33376\*(gleason))

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

The mean of cancerval = 6.998682

The mean of benpros = 2.534725

The mode of vesinv = 0

The mean of gleason = 6.876289

Thus, the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors is 10.28357.