PSTAT 126 Practice 5 (Application)

Saad Mouti

Application

This part comprises questions that involve writing R codes. Please write your codes and answers in the locations indicated and be sure to follow the instructions regarding whether to show codes and output.

A1. Prostate cancer surgery The `faraway::prostate` dataset contains measurements taken from 97 men with prostate cancer due to receive surgery. The first few rows are shown below.

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
```

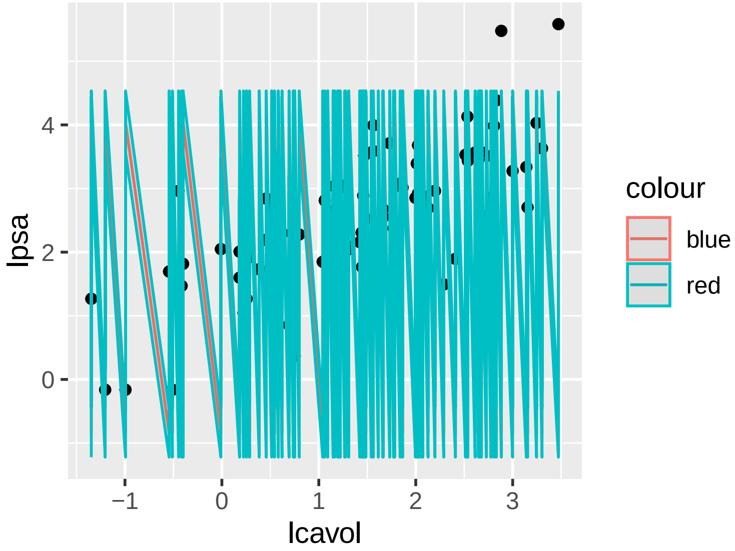
The variable of interest is log-cancer-volume, which reflects the development of prostate cancer in the patient (higher volume indicates later-stage). Prostate-specific antigen is often measured in screening tests to detect prostate cancer. Here you'll explore predicting the cancer development stage based on PSA.

i. Regress log-cancer-volume on age, log-PSA, and log-prostate-weight. Show the model summary only (no codes).

```
##
## Call:
## lm(formula = lcavol ~ age + lpsa + lweight, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07322 -0.54594 -0.01828  0.56280  1.69501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91121    0.81073  -1.124   0.2639
## age          0.02092    0.01147   1.824   0.0713 .
## lpsa         0.76782    0.07526  10.202   <2e-16 ***
## lweight     -0.26773    0.18118  -1.478   0.1429
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7942 on 93 degrees of freedom
## Multiple R-squared:  0.5601, Adjusted R-squared:  0.546
## F-statistic: 39.48 on 3 and 93 DF,  p-value: < 2.2e-16
```

ii. Visualize the estimated relationship between log-cancer-volume and log-PSA for each of the 10th, 50th, and 90th percentiles of ages in the data, with the remaining variable set at its median value. Please include 90% confidence bands; show the graphic only (no codes).

```
##      lcavol     lpsa age lweight       pred     ci.fit    ci.lwr     ci.upr
## 1 -1.347074 -0.16252  56  3.6571 -0.8434950 -0.8434950 -1.221770 -0.4652197
## 2 -1.347074 -0.16252  57  3.6571 -0.8225731 -0.8225731 -1.194848 -0.4502983
## 3 -1.347074 -0.16252  58  3.6571 -0.8016512 -0.8016512 -1.168817 -0.4344851
## 4 -1.347074 -0.16252  59  3.6571 -0.7807293 -0.7807293 -1.143716 -0.4177424
```



iii. Repeat (ii) but fix the variable not shown in the plot at its 90th percentile rather than its median. Does the graphic look much different? Show the graphic only (no codes).

iv. Visualize the estimated relationship between cancer volume and PSA for each of the 10th, 50th, and 90th percentiles of ages in the data and the median value of prostate weight. Please include 90% confidence bands; show the graphic only (no codes).

v. Does it appear that age is associated with much change in cancer volume based on the graphic? Answer in one sentence, and cite a feature of the plot.

Type your answer here (replacing this text)

vi. Suppose a 65-year-old patient arrives and measurements reveal he has a log-PSA level of 0.28 and log-prostate-weight 3.62301. Predict his cancer volume on the log scale with appropriate uncertainty quantification. Show the results, but don't show your codes.

vii. What would the prediction in (v) be if the patient were 20 instead of 65? Show the computed prediction but not your codes.

viii. Is the prediction for the 20-year-old patient more or less uncertain than the prediction for the 65-year-old patient, and why? Explain in 1-2 sentences.

Type your answer here (replacing this text)

ix. Estimate the predictive accuracy of the model you fit in (i). Explain your approach in a sentence or two and report estimated predictive accuracy, but do not show your codes.

x. Does it make sense to use this fitted model to predict whether a patient has cancer in the first place? Why or why not? Explain in 1-3 sentences.

Type your answer here (replacing this text)

Submission instructions

1. Clear your environment and run all codes to check for errors. Resolve any if detected.
2. Input your name in the author information, remove the instructions at the beginning and end of the document, and knit to pdf.
3. Inspect the pdf and fix any display issues.
4. Once the pdf looks good, upload a copy to Gradescope.
5. Save a backup copy of your work in a recoverable location.

Code appendix

```r
# knit options
knitr::opts_chunk$set(echo = F,
                      results = 'markup',
                      fig.width = 4,
                      fig.height = 3,
                      fig.align = 'center',
                      message = F,
                      warning = F)

# packages
library(tidyverse)
library(modelr)
library(faraway)
prostate %>% head()
# fit model and show summary
cancer_model <- lm(lcavol ~ age + lpsa + lweight, data=prostate)
summary(cancer_model)
prostate_10 <- prostate %>%
  filter(quantile(age, 0.1) < age)
prostate_50 <- prostate %>%
  filter(quantile(age, 0.5) < age)
prostate_90 <- prostate %>%
  filter(quantile(age, 0.9) < age)
# construct prediction grid for visualization
pred_df <- data_grid(data = prostate_10, lcavol, lpsa, age, .model = cancer_model)

pred_df <- pred_df %>%
  add_predictions(model = cancer_model)

pred_df <- pred_df %>% cbind(ci = predict(cancer_model, pred_df, interval = 'confidence', level = 0.9))
pred_df %>% head(4)
# construct graphic

ggplot(pred_df, aes(x = lcavol)) + # base layer
geom_point(aes(y = lpsa), data = prostate_10) + # add scatter layer
geom_path(aes(y = ci.fit, color = 'blue')) +
geom_ribbon(aes(ymin = ci.lwr, ymax = ci.upr, color='red'), alpha = 0.1)
# construct prediction grid for visualization

# construct graphic

# construct graphic

# compute prediction

# compute prediction

# estimate predictive accuracy
```