Lab 9 Solutions

Kaushik Raj V Nadar (208160499)

2024-10-16

Problem 1

Consider the Time Delivery Data (full with 40 observations combined Table 3.2 and Table 11.2 from INTRODUCTION TO LINEAR REGRESSION ANALYSIS by MONTGOMERY et.al.). Split the data in Prediction and Estimation data sets equally. Use the Estimation Data.

Load Required Libraries

```r
library(MPV)  # for time delivery dataset
library(leaps)  # for subset selection
```

Load Data

```r
time_delivery_data <- p15.4
print(time_delivery_data)
```

```
   x1   x2     y set
1   7  560 16.68   e
2   3  220 11.50   e
3   3  340 12.03   p
4   4   80 14.88   e
5   6  150 13.75   p
6   7  330 18.11   p
7   2  110  8.00   p
8   7  210 17.83   e
```

```
9  30 1460 79.24   e
10  5  605 21.50   p
11 16  688 40.33   e
12 10  215 21.00   p
13  4  255 13.50   e
14  6  462 19.75   p
15  9  448 24.00   e
16 10  776 29.00   e
17  6  200 15.35   e
18  7  132 19.00   p
19  3   36  9.50   p
20 17  770 35.10   e
21 10  140 17.90   p
22 26  810 52.32   p
23  9  450 18.75   e
24  8  635 19.83   p
25  4  150 10.75   e
26 22  905 51.00   p
27  7  520 16.80   p
28 15  290 26.16   e
29  5  500 19.90   e
30  6 1000 24.00   p
31  6  225 18.55   p
32 10  775 31.93   p
33  4  212 16.95   e
34  1  144  7.00   e
35  3  126 14.00   p
36 12  655 37.03   p
37 10  420 18.62   e
38  7  150 16.10   e
39  8  360 24.38   p
40 32 1530 64.75   e
```

Split the data into Estimation and Prediction sets

```r
# Using Random Split
set.seed(123)  # for reproducibility
n <- nrow(time_delivery_data)
indices <- sample(1:n, size = n/2)

estimation_data <- time_delivery_data[indices, c("x1", "x2", "y")]
prediction_data <- time_delivery_data[-indices, c("x1", "x2", "y")]
```

```
# Using predefined split
estimation_data <- time_delivery_data[time_delivery_data$set=='e',
                                      c("x1", "x2", "y")]
prediction_data <- time_delivery_data[time_delivery_data$set=='p',
                                      c("x1", "x2", "y")]
```

Part (a)

Based on some subset selection criteria, propose two regression models.

```
# Function to perform subset selection and return the best model
get_best_subset <- function(data, nvmax) {
  regsubsets_out <- regsubsets(y ~ ., data = data, nvmax = nvmax,
                               method = "forward")
  summary_out <- summary(regsubsets_out)
  which.min(summary_out$bic)
}

# Model 1: Based on BIC
best_subset_1 <- get_best_subset(estimation_data, nvmax = 2)
formula_1 <- as.formula(paste("y ~",
paste(names(coef(lm(y ~ ., estimation_data)))[2:best_subset_1],
      collapse = " + ")))
formula_1
```

```
y ~ x1
```

```
# Model 2: Full model
formula_2 <- y ~ x1 + x2
```

Part (b)

For each models, compare the regression coefficients for Prediction and Estimation data sets.

```
# Function to fit model and extract coefficients
fit_and_extract <- function(formula, data) {
  model <- lm(formula, data)
  coef(model)
}
```

```r
# Fit models and extract coefficients
coef_est_1 <- fit_and_extract(formula_1, estimation_data)
coef_pred_1 <- fit_and_extract(formula_1, prediction_data)
coef_est_2 <- fit_and_extract(formula_2, estimation_data)
coef_pred_2 <- fit_and_extract(formula_2, prediction_data)

# Print coefficients
cat("Model 1 coefficients:\n")
print(rbind(Estimation = coef_est_1, Prediction = coef_pred_1))

cat("\nModel 2 coefficients:\n")
print(rbind(Estimation = coef_est_2, Prediction = coef_pred_2))
```

```
Model 1 coefficients:
           (Intercept)       x1
Estimation    3.950143 2.086936
Prediction    6.762536 1.887002

Model 2 coefficients:
           (Intercept)       x1         x2
Estimation    3.507326 1.387630 0.01563308
Prediction    4.423364 1.533742 0.01239923
```

Part (c)

Compute the PRESS statistics and R2 prediction for both models using Prediction and Estimation data sets.

```r
# Function to compute PRESS and R2 prediction
compute_press_r2pred <- function(formula, train_data, test_data) {
  model <- lm(formula, train_data)

  # PRESS statistic
  press <- sum((resid(model) / (1 - hatvalues(model)))^2)

  # R2 prediction
  predictions <- predict(model, newdata = test_data)
  sse <- sum((test_data$y - predictions)^2)
  sst <- sum((test_data$y - mean(train_data$y))^2)
  r2_pred <- 1 - sse / sst
  list(press = press, r2_pred = r2_pred)
}

# Compute PRESS and R2 prediction for both models
results_1 <- compute_press_r2pred(formula_1, estimation_data, prediction_data)
results_2 <- compute_press_r2pred(formula_2, estimation_data, prediction_data)

# Print results
cat("\nModel 1 results:\n")
print(results_1)

cat("\nModel 2 results:\n")
print(results_2)
```

```
Model 1 results:
$press
[1] 781.2205

$r2_pred
[1] 0.8676914


Model 2 results:
$press
[1] 677.0287

$r2_pred
[1] 0.9367248
```