Homework 6 Solution

Solution

Question 1. Accuracy of software effort estimates

6.4(a) (5 pts) In Step 1 of the stepwise regression, how many different one_variable models are fit to the data?

Answer: There are eight different one-variable models that can be fit to the data.

6.4(b) (5 pts) In Step 1, the variable x_1 is selected as the 'best' one-variable predictor. How is this determined?

Answer: $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ is larger (absolute) than any of the other t-values. Or equivalently, p-value is smaller than any of the other p-values.

6.4(c) (5 pts)In Step 2 of the stepwise regression, how many different two-variable models (where x_1 is one of the variables) are fit to the data?

Answer: There would be an additional seven two-variable combinations that could be fitted to the model with x_1 being one of the variables.

6.4(d) (10 pts) The only two variables selected for entry into the stepwise regression model were x_1 and x_8. The stepwise regression yielded the following prediction equation:

$$\hat{y} = .12 - .28x_1 + 0.27x_8$$

Give a practical interpretation of the β estimates multiplied by x_1 and x_8.

Answer:

$\hat{\beta}_1 = -0.28$. The difference in the mean relative error in estimating effort (y) between developer and project leader is estimated to be -0.28.

$\hat{\beta}_2 = 0.27$. The difference in the mean relative error in estimating effort (y) between previous accuracy of more that 20% and previous accuracy less than 20% is estimated to be 0.27.

6.4(e) (5 pts) Why should a researcher by wary of using the model, pard d, as the final model for predicting effect (y).

Answer: One should be wary of using the model in part d as the final model because it contains only indicator variables and only first-order terms.

Question 2. Clerical staff work hours (Data set: CLERICAL)

6.8(a) (10 pts) Conduct a stepwise regression analysis of the data using R.

```
### Import data and conduct a stepwise regression

clerical = read.csv("STAT 3113 Data Sets/CLERICAL.csv", header = TRUE, sep = ",", dec = ".")
clerical = clerical[, c(-1,-2)]

library(leaps)
```

```
step_model = regsubsets(Y~., data = clerical, nvmax = 7, method = "seqrep")
summary(step_model)
```

```
## Subset selection object
## Call: regsubsets.formula(Y ~ ., data = clerical, nvmax = 7, method = "seqrep")
## 7 Variables  (and intercept)
##    Forced in Forced out
## X1     FALSE      FALSE
## X2     FALSE      FALSE
## X3     FALSE      FALSE
## X4     FALSE      FALSE
## X5     FALSE      FALSE
## X6     FALSE      FALSE
## X7     FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: 'sequential replacement'
##          X1  X2  X3  X4  X5  X6  X7
## 1  ( 1 ) " " " " " " " " "*" " " " "
## 2  ( 1 ) " " "*" " " " " "*" " " " "
## 3  ( 1 ) " " "*" " " "*" "*" " " " "
## 4  ( 1 ) " " "*" " " "*" "*" "*" " "
## 5  ( 1 ) " " "*" "*" "*" "*" "*" " "
## 6  ( 1 ) "*" "*" "*" "*" "*" "*" " "
## 7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
fit_complete = lm(Y~., data = clerical)
ols_step_both_p(fit_complete, details = FALSE, pent = 0.05, prem = 0.05)
```

```
##
##                               Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                      Added/                   Adj.
## Step    Variable    Removed     R-Square    R-Square     C(p)        AIC        RMSE
## ----------------------------------------------------------------------------------------
##    1       X5       addition       0.345       0.332    18.7750    415.8625    12.7007
##    2       X2       addition       0.436       0.413    11.4660    410.0556    11.9018
##    3       X4       addition       0.481       0.448     8.9480    407.7976    11.5428
## ----------------------------------------------------------------------------------------
```

(b) (15 pts) Interpret the β estimates in the resulting stepwise model.

```
fit = lm(Y~X5 + X2 + X4, data = clerical)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X5 + X2 + X4, data = clerical)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.259  -9.075  -1.938   6.882  29.303
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.725640   6.910199  11.248 4.69e-15 ***
## X5           0.058268   0.009714   5.998 2.52e-07 ***
## X2           0.136264   0.045413   3.001  0.00426 **
## X4          -0.034689   0.017140  -2.024  0.04857 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.54 on 48 degrees of freedom
## Multiple R-squared:  0.4806, Adjusted R-squared:  0.4481
## F-statistic:  14.8 on 3 and 48 DF,  p-value: 5.91e-07
```

Answer:

$\hat{\beta}_0 = 77.73$. This is simply the y-intercept 0 which is not in the observed ranges for x_5, x_2, and x_4.

$\hat{\beta}_1 = 0.05827$. We estimate the mean number of hours worked per day by the clerical staff to increase by 0.058 hours for every one additional check cashed, holding the other variables constant.

$\hat{\beta}_2 = 0.1363$. We estimate the number of hours worked per day by the clerical staff to increase by 0.136 hours for every one additional money order or gift certificate sold, holding the other variables constant.

$\hat{\beta}_3 = -0.035$. We estimate the number of hours worked per day by the clerical staff to decrease by 0.035 hours for every one additional change order transaction processed, holding the other variables constant.

(c) (5 pts) What are the dangers associated with drawing inferences from the stepwise model?

Answer: This model was selected as the best model after many different models were fit to the data. There are two main dangers associated with using and interpreting this model:

- The overall probability with making at least one Type I error is extremely high.
- This model only includes linear components. Before using this model, the analyst should consider both quadratic and interaction relationship in the model.

Question 3. Cooling Method for Gas Turbines (Data set: GASTURBINE)

6.10(a) (5 pts) Use stepwise regression (with stepwise selection) to find the 'best' predictors of heat rate.

```
gasturbine = read.csv("STAT 3113 Data Sets/GASTURBINE.csv")
fit.complete = lm(HEATRATE ~., data = gasturbine)
stepwise.turbine = ols_step_both_p(fit.complete, details=FALSE, pent=0.05, prem=0.05)
stepwise.turbine$model
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Coefficients:
## (Intercept)          RPM   INLET.TEMP     EXH.TEMP
##  14359.7168       0.1051      -9.2226      12.4260
```

Answer: The best predictors are: RPM, INLET-TEMP, and EXH-TEMP.

6.10(b) (5 pts) Use stepwise regression (with backward elimination) to find the 'best' predictors of heat rate.

```
backward.turbine = ols_step_backward_p(fit.complete, details=FALSE, prem=0.05)
backward.turbine$model
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Coefficients:
## (Intercept)          RPM   INLET.TEMP     EXH.TEMP
##  14359.7168       0.1051      -9.2226      12.4260
```

Answer: The best predictors are: RPM, INLET-TEMP, and EXH-TEMP.

6.10(c) (15 pts) Use all-possible-regression-selection to find the 'best' predictors of heat rate.

```
library(leaps)

bestsubset = regsubsets(HEATRATE~., data = gasturbine, nvmax=9)
bestsubset.summary = summary(bestsubset)

bestsubset.summary$outmat
```

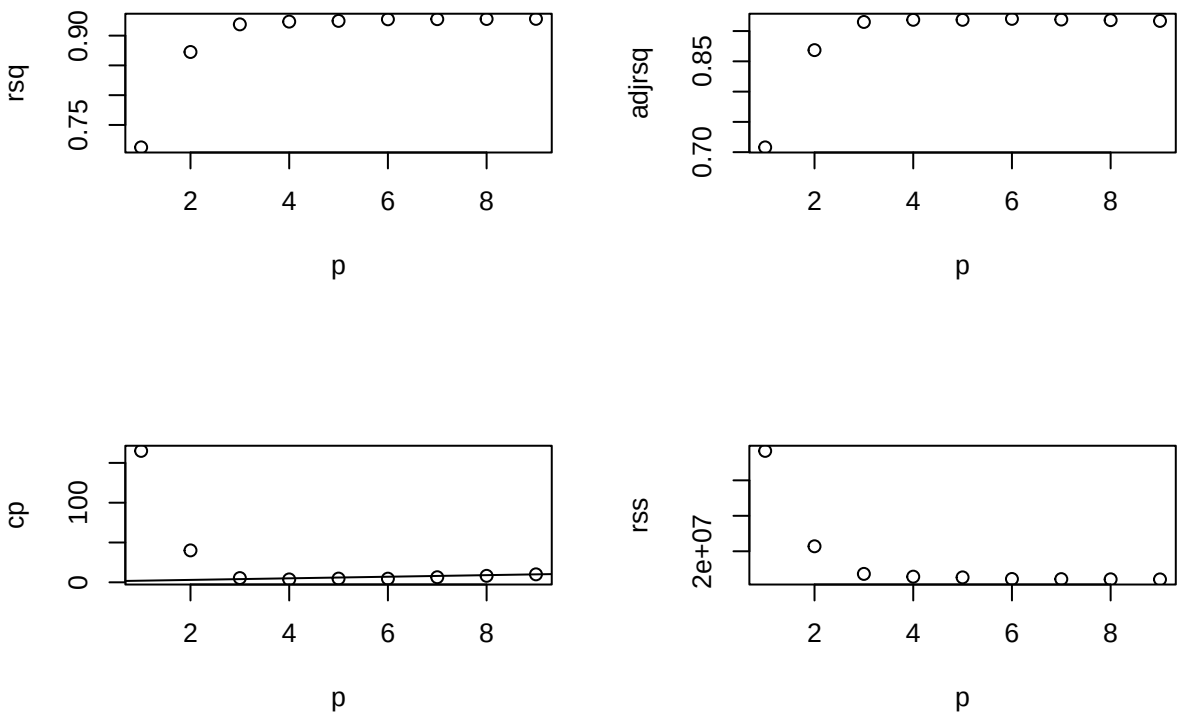
```
##          ENGINEAeroderiv ENGINETraditional SHAFTS RPM CPRATIO INLET.TEMP
## 1  ( 1 ) " "             " "               " "    "*" " "     " "
## 2  ( 1 ) " "             " "               " "    "*" " "     "*"
## 3  ( 1 ) " "             " "               " "    "*" " "     "*"
## 4  ( 1 ) " "             " "               " "    "*" " "     "*"
## 5  ( 1 ) " "             " "               " "    "*" " "     "*"
## 6  ( 1 ) "*"             "*"               " "    "*" " "     "*"
## 7  ( 1 ) "*"             "*"               "*"    "*" " "     "*"
## 8  ( 1 ) "*"             "*"               "*"    "*" "*"     "*"
## 9  ( 1 ) "*"             "*"               "*"    "*" "*"     "*"
##          EXH.TEMP AIRFLOW POWER
## 1  ( 1 ) " "      " "     " "
## 2  ( 1 ) " "      " "     " "
## 3  ( 1 ) "*"      " "     " "
## 4  ( 1 ) "*"      "*"     " "
## 5  ( 1 ) "*"      "*"     "*"
## 6  ( 1 ) "*"      "*"     " "
## 7  ( 1 ) "*"      "*"     " "
## 8  ( 1 ) "*"      "*"     " "
## 9  ( 1 ) "*"      "*"     "*"
```

```
rsq = bestsubset.summary$rsq
adjrsq = bestsubset.summary$adjr2
cp = bestsubset.summary$cp
rss = bestsubset.summary$rss

p=seq(1, 9, 1)

par(mfrow=c(2,2))
plot(p, rsq)
plot(p, adjrsq)
plot(p, cp)
```

```
abline(1,1)
plot(p, rss)
```




```
par(mfrow=c(1,1))
```

Answer: From the output below, the best models would be three or four variable model.

- Three variable model contains variables:RPM, INLET.TEMP, EXH.TEMP
- Four variable model contains variables: RPM, INLET.TEMP, EXH.TEMP, AIRFLOW

6.10(d) (5 pts) Compare the results, parts a-c. Which independent variables consistently are selected as the 'best' predictors?

Answer: Based on the results from a-c above, the three variables (RPM, INLET.TEMP, and EXH.TEMP) are consistently selected as the best predictors. AIRFLOW is marginal, but should be considered.

6.10(e) (10 pts) Explain how you would use the results, parts a-c, to develop a model for heat rate?

Answer: We would use RPM, INLET.TEMP, EXH.TEMP, and possibly AIRFLOW as independent variables for predicting HEATRATE in a multiple regression model, then checking to see if including interaction and higher-order terms would improve the model.