Homework

Lucy Lin

2022-11-15

Problem 1

#1a) True, the k+1 variable model requires the k variable model as the base model in order to select the k+1th variable to add.

#1b) True, we remove the k+1th variable from the k+1 variable model (which is the least impactful variable) to get the k variable model.

#1c) False, we are making comparisons among each combination of k predictors to find the one with the lowest SSE. This means that the model with k+1 predictors may have the same predictors in the k predictors model.

#1d) True, both are comparing the models with the lowest SSE.

#1e) False, the backward stepwise keeps the combination of variables that result in the lowest SSE models but this means that a k-variable model may exclude the best (lowest SSE) combination of variables. The best 1 variable model might be excluded in the backward stepwise process.

Problem 2

#2a) True, the one variable model will include RBI therefore so will the two variable model.

#2b) False, this SSE table cannot tell us anything about the two variable models' SSE and we know that the one variable model must have RBI.

#2c) True, the best subset selection for the 1 variable model is the same as the model from #2a, and therefore guarantees RBI is included.

#2d) False, we cannot guarantee that RBI was not removed from the model in the backward stepwise method from this SSE table.

Problem 3

k = 1 predictor: yearOfRegistration

Best k = 2 predictors model: Y ~ 1 + yearOfRegistration + vehicleTypesmall_car

Best k = 5 predictors model: Y ~ 1 + yearOfRegistration + vehicleTypesmall_car + kilometer + notRepairedDamageyes + fuelTypepetrol

```
library(leaps)

data = read.csv("used_car.csv")
fit.best <- regsubsets(price ~ ., data = data, nvmax = 2)
fit1.best <- regsubsets(price ~ ., data = data, nvmax = 5)

summary(fit.best)
```

```
## Subset selection object
## Call: regsubsets.formula(price ~ ., data = data, nvmax = 2)
## 9 Variables  (and intercept)
##                      Forced in Forced out
## vehicleTypecoupe         FALSE      FALSE
## vehicleTypesmall_car     FALSE      FALSE
## vehicleTypesuv           FALSE      FALSE
## yearOfRegistration       FALSE      FALSE
## gearboxmanual            FALSE      FALSE
## kilometer                FALSE      FALSE
## monthOfRegistration      FALSE      FALSE
## fuelTypepetrol           FALSE      FALSE
## notRepairedDamageyes     FALSE      FALSE
## 1 subsets of each size up to 2
## Selection Algorithm: exhaustive
##          vehicleTypecoupe vehicleTypesmall_car vehicleTypesuv
## 1  ( 1 ) " "              " "                  " "
## 2  ( 1 ) " "              "*"                  " "
##          yearOfRegistration gearboxmanual kilometer monthOfRegistration
## 1  ( 1 ) "*"                " "           " "       " "
## 2  ( 1 ) "*"                " "           " "       " "
##          fuelTypepetrol notRepairedDamageyes
## 1  ( 1 ) " "            " "
## 2  ( 1 ) " "            " "
```

```
summary(fit1.best)
```

```
## Subset selection object
## Call: regsubsets.formula(price ~ ., data = data, nvmax = 5)
```

```
## 9 Variables  (and intercept)
##                       Forced in Forced out
## vehicleTypecoupe          FALSE      FALSE
## vehicleTypesmall_car      FALSE      FALSE
## vehicleTypesuv            FALSE      FALSE
## yearOfRegistration        FALSE      FALSE
## gearboxmanual             FALSE      FALSE
## kilometer                 FALSE      FALSE
## monthOfRegistration       FALSE      FALSE
## fuelTypepetrol            FALSE      FALSE
## notRepairedDamageyes      FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          vehicleTypecoupe vehicleTypesmall_car vehicleTypesuv
## 1  ( 1 ) " "               " "                 " "
## 2  ( 1 ) " "               "*"                 " "
## 3  ( 1 ) " "               "*"                 " "
## 4  ( 1 ) " "               "*"                 " "
## 5  ( 1 ) " "               "*"                 " "
##          yearOfRegistration gearboxmanual kilometer monthOfRegistration
## 1  ( 1 ) "*"                " "           " "       " "
## 2  ( 1 ) "*"                " "           " "       " "
## 3  ( 1 ) "*"                " "           "*"       " "
## 4  ( 1 ) "*"                " "           "*"       " "
## 5  ( 1 ) "*"                " "           "*"       " "
##          fuelTypepetrol notRepairedDamageyes
## 1  ( 1 ) " "            " "
## 2  ( 1 ) " "            " "
## 3  ( 1 ) " "            " "
## 4  ( 1 ) " "            "*"
## 5  ( 1 ) "*"            "*"
```

Problem 4

#4a) False, the lowest point on the curve will have a lower error of estimating test error but it still has some bias with overfitting. #4b) True, K2 is more biased and has a lower variance than K1's model so K2 is larger than K1 #4c) True, if K = n+1, there is an instance where there is nothing to exclude from the model #4d) True, there is nothing to validate with K if you remove all the predictors. #4e) True, when it chooses to remove 1 predictor, it is only left with the other predictor (50%) to fit the model on.