Homework 5

Jiao Qu A20386614, Yuan-An Liu A20375099, Zhenyu Zhang A20287371

1.

```
Auto$origin <- factor(Auto$origin)
regfit.full = regsubsets(mpg~.-name, data = Auto)
summary(regfit.full)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ . - name, data = Auto)
## 8 Variables  (and intercept)
##              Forced in Forced out
## cylinders        FALSE      FALSE
## displacement     FALSE      FALSE
## horsepower       FALSE      FALSE
## weight           FALSE      FALSE
## acceleration     FALSE      FALSE
## year             FALSE      FALSE
## origin2          FALSE      FALSE
## origin3          FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          cylinders displacement horsepower weight acceleration year
## 1  ( 1 ) " "       " "          " "        "*"    " "          " "
## 2  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 3  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 4  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 5  ( 1 ) " "       "*"          " "        "*"    " "          "*"
## 6  ( 1 ) " "       "*"          "*"        "*"    " "          "*"
## 7  ( 1 ) "*"       "*"          "*"        "*"    " "          "*"
## 8  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"
##          origin2 origin3
## 1  ( 1 ) " "     " "
## 2  ( 1 ) " "     " "
## 3  ( 1 ) " "     "*"
## 4  ( 1 ) "*"     "*"
## 5  ( 1 ) "*"     "*"
## 6  ( 1 ) "*"     "*"
## 7  ( 1 ) "*"     "*"
## 8  ( 1 ) "*"     "*"
```

```
regfit.summary = summary(regfit.full)
# a. the best adjusted R2
which.max(regfit.summary$adjr2)
```

```
## [1] 7
```

The best subset: cylinders displacement horsepower weight year origin2 origin3

(a)

```
regfit.summary$adjr2[7]
```

```
## [1] 0.8206916
```

The best adjusted R2 is at 7th: 0.8206916

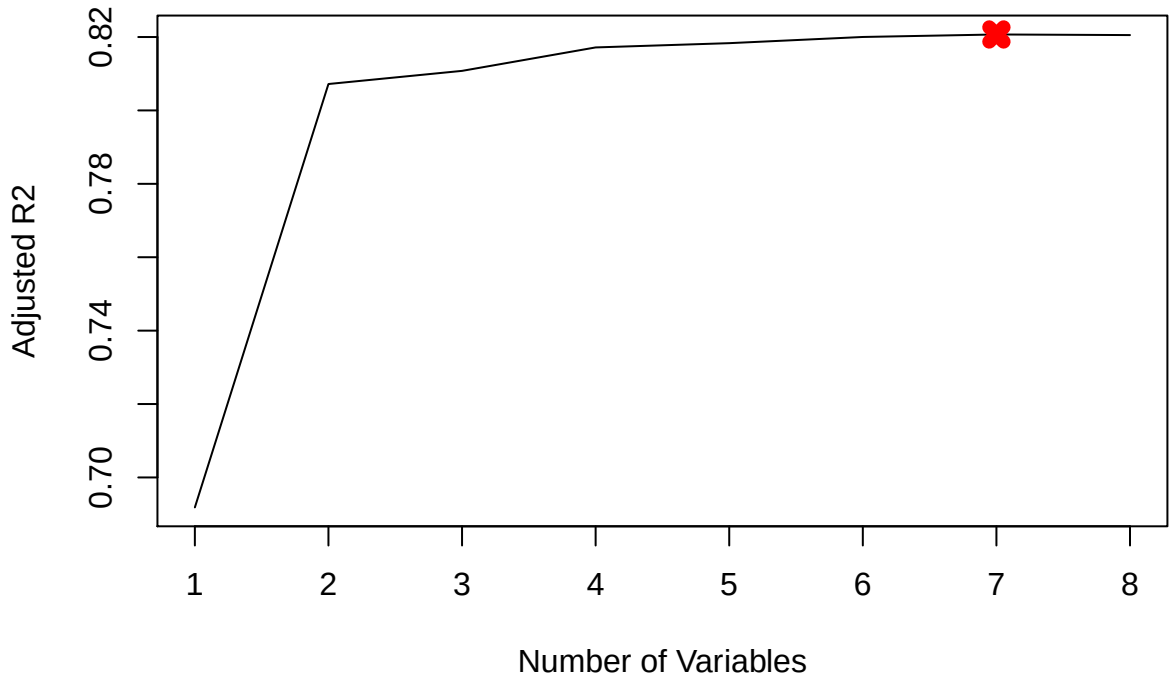
(b)

```
# b. coefficients
coefficients(regfit.full, id = 7)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -16.332312787  -0.502767012   0.023372075  -0.025002677  -0.006459817
##          year       origin2       origin3
##   0.773883341   2.634517472   2.857355960
```

(c)

```
# c. Plot of the adjusted R2 as a function of number of variables
plot(regfit.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R2", pch = 20, type = "l")
points(7, regfit.summary$adjr2[7], pch = 4, col = "red", lwd = 7)
```



2.

```
regfit.fwd=regsubsets(mpg~.-name, data = Auto, method ="forward")
summary(regfit.fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ . - name, data = Auto, method = "forward")
## 8 Variables  (and intercept)
##              Forced in Forced out
## cylinders        FALSE      FALSE
## displacement     FALSE      FALSE
## horsepower       FALSE      FALSE
## weight           FALSE      FALSE
## acceleration     FALSE      FALSE
## year             FALSE      FALSE
## origin2          FALSE      FALSE
## origin3          FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          cylinders displacement horsepower weight acceleration year
## 1  ( 1 ) " "       " "          " "        "*"    " "          " "
## 2  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 3  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 4  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 5  ( 1 ) " "       "*"          " "        "*"    " "          "*"
## 6  ( 1 ) " "       "*"          "*"        "*"    " "          "*"
## 7  ( 1 ) "*"       "*"          "*"        "*"    " "          "*"
## 8  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"
##          origin2 origin3
## 1  ( 1 ) " "     " "
## 2  ( 1 ) " "     " "
## 3  ( 1 ) " "     "*"
## 4  ( 1 ) "*"     "*"
## 5  ( 1 ) "*"     "*"
## 6  ( 1 ) "*"     "*"
## 7  ( 1 ) "*"     "*"
## 8  ( 1 ) "*"     "*"
```

```
regfitFWD.summary = summary(regfit.fwd)
which.max(regfitFWD.summary$adjr2)
```

```
## [1] 7
```

The best subset: cylinders displacement horsepower weight year origin2 origin3

(a)

```
regfitFWD.summary$adjr2[7]
```

```
## [1] 0.8206916
```

The best adjusted R2 is at 7th: 0.8206916

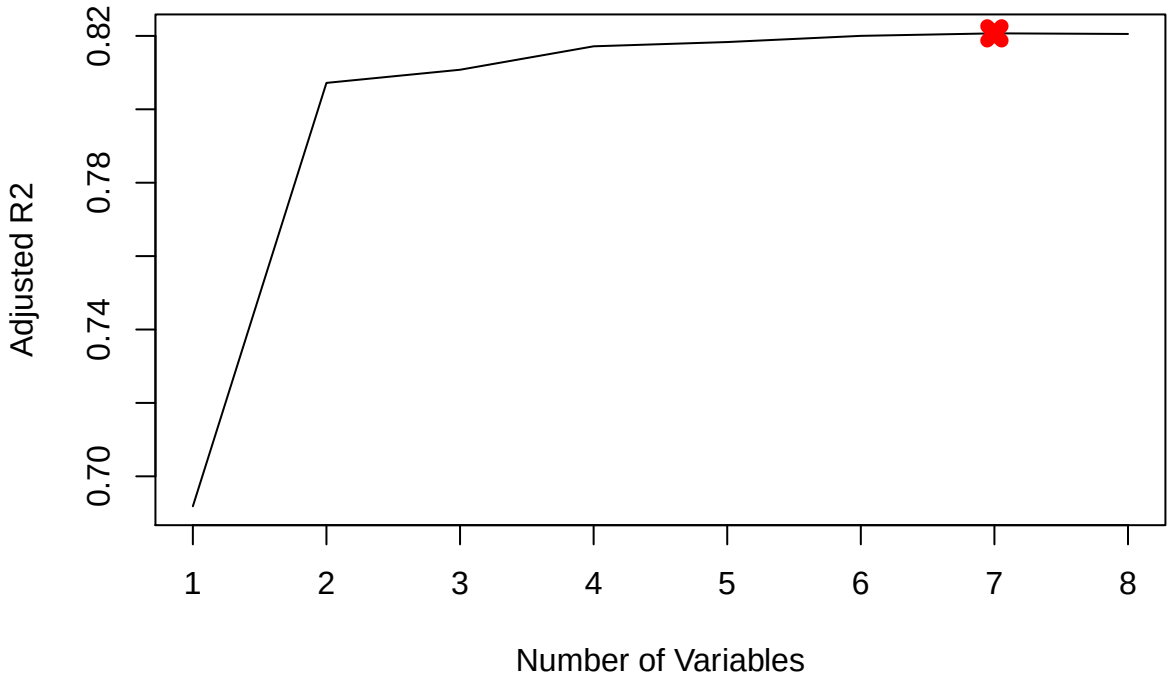
(b)

```
# b. coefficients
coefficients(regfit.fwd, id = 7)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -16.332312787  -0.502767012   0.023372075  -0.025002677  -0.006459817
##          year       origin2       origin3
##   0.773883341   2.634517472   2.857355960
```

(c)

```
# c. Plot of the adjusted R2 as a function of number of variables
plot(regfitFWD.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R2", pch = 20, type = "l")
points(7, regfitFWD.summary$adjr2[7], pch = 4, col = "red", lwd = 7)
```



(d)

It is the same as the best subset.

(e)

It is the same subset. The same K.

3

```
regfit.bwd=regsubsets(mpg~.-name, data = Auto, method ="backward")
summary(regfit.bwd)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ . - name, data = Auto, method = "backward")
## 8 Variables  (and intercept)
```

```
##              Forced in Forced out
## cylinders        FALSE      FALSE
## displacement     FALSE      FALSE
## horsepower       FALSE      FALSE
## weight           FALSE      FALSE
## acceleration     FALSE      FALSE
## year             FALSE      FALSE
## origin2          FALSE      FALSE
## origin3          FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: backward
##          cylinders displacement horsepower weight acceleration year
## 1  ( 1 ) " "       " "          " "        "*"    " "          " "
## 2  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 3  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 4  ( 1 ) " "       " "          " "        "*"    " "          "*"
## 5  ( 1 ) " "       "*"          " "        "*"    " "          "*"
## 6  ( 1 ) " "       "*"          "*"        "*"    " "          "*"
## 7  ( 1 ) "*"       "*"          "*"        "*"    " "          "*"
## 8  ( 1 ) "*"       "*"          "*"        "*"    "*"          "*"
##          origin2 origin3
## 1  ( 1 ) " "     " "
## 2  ( 1 ) " "     " "
## 3  ( 1 ) " "     "*"
## 4  ( 1 ) "*"     "*"
## 5  ( 1 ) "*"     "*"
## 6  ( 1 ) "*"     "*"
## 7  ( 1 ) "*"     "*"
## 8  ( 1 ) "*"     "*"
```

```r
regfitBWD.summary = summary(regfit.bwd)
which.max(regfitBWD.summary$adjr2)
```

```
## [1] 7
```

The best subset: cylinders displacement horsepower weight year origin2 origin3

(a)

```r
regfitBWD.summary$adjr2[7]
```

```
## [1] 0.8206916
```

The best adjusted R2 is at 7th: 0.8206916

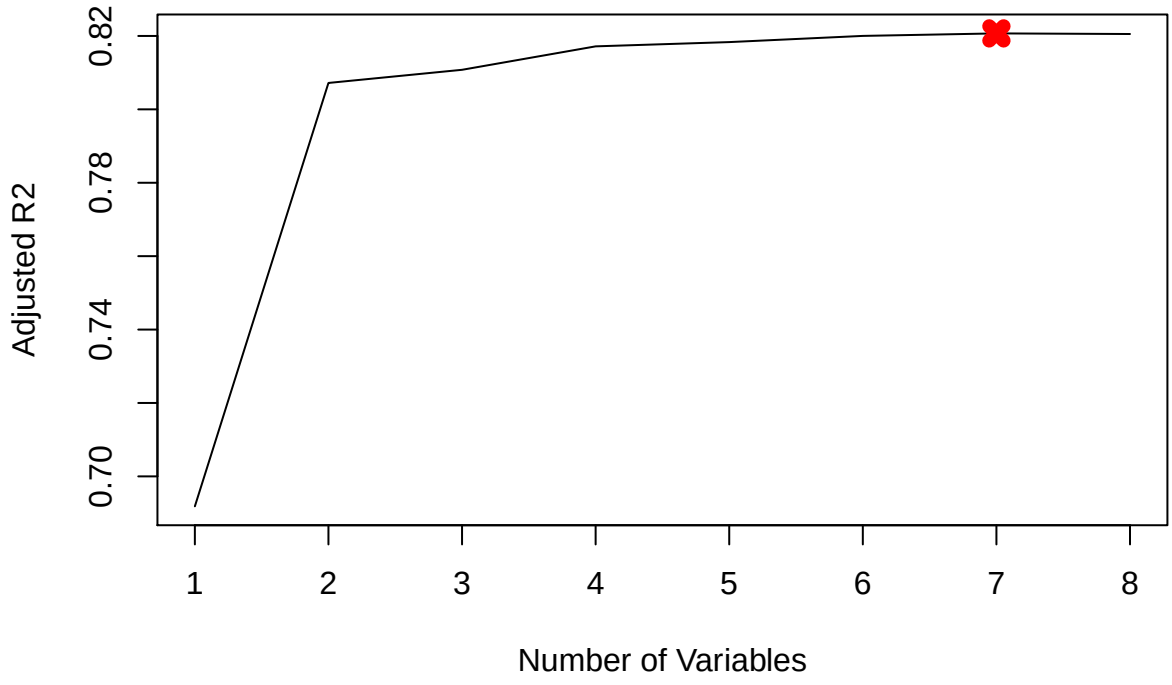
(b)

```r
# b. coefficients
coefficients(regfit.bwd, id = 7)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -16.332312787  -0.502767012   0.023372075  -0.025002677  -0.006459817
##          year       origin2       origin3
##   0.773883341   2.634517472   2.857355960
```

(c)

```
# c. Plot of the adjusted R2 as a function of number of variables
plot(regfitBWD.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R2", pch = 20, type = "l")
points(7, regfitBWD.summary$adjr2[7], pch = 4, col = "red", lwd = 7)
```



(d)

It is the same as the best subset.

(e)

It is the same subset. The same K.

4

(a)

```
set.seed(1)
train=sample(c(TRUE,FALSE),nrow(Auto),replace=TRUE)
test=(!train)
Auto$origin=factor(Auto$origin)
regfit.best=regsubsets(mpg~.-name,data=Auto[train,],nvmax=7)
test.mat=model.matrix(mpg~.-name,data=Auto[test,])
val.errors=rep(NA,7)
 for(i in 1:7){
     coefi=coef(regfit.best,id=i)
```

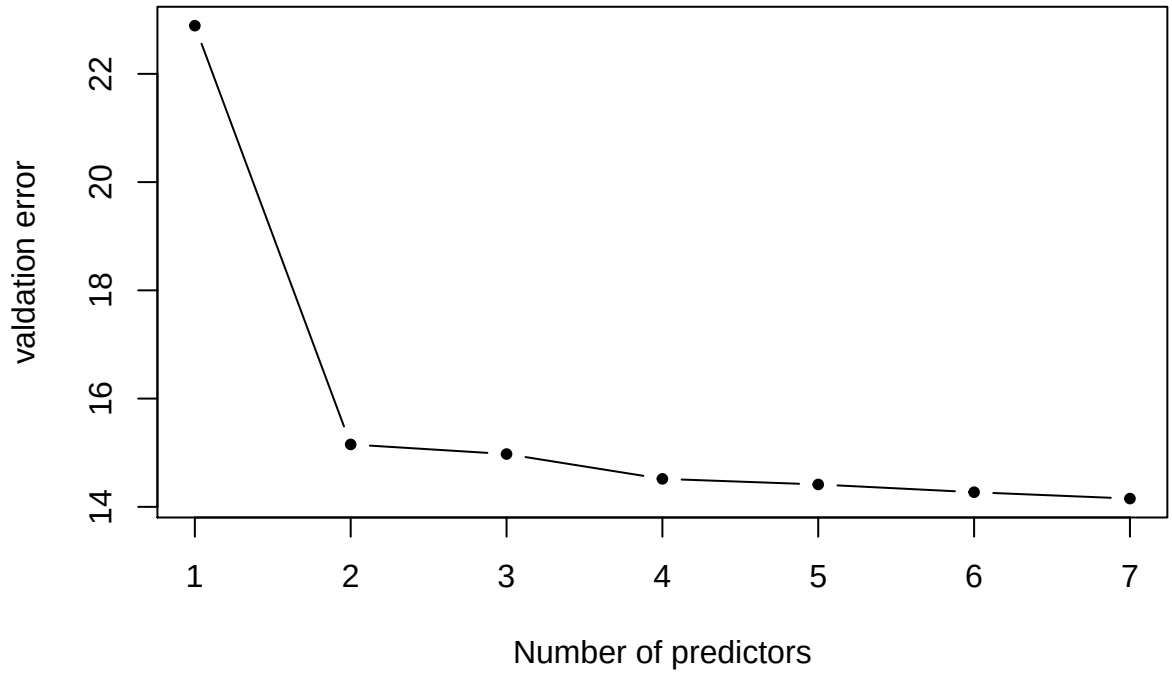
```
    pred=test.mat[,names(coefi)]%*%coefi
    val.errors[i]=mean((Auto$mpg[test]-pred)^2) }
val.errors
```

```
## [1] 22.89000 15.15314 14.97520 14.51691 14.41350 14.27013 14.15268
```

The best validation error is 14.15268.

(b)

```
# Plot the validation error as a function of k, the number of features.
plot(val.errors,xlab = "Number of predictors",ylab = "valdation error",pch=20,type = "b")
```



(c)

```
# Show the coefficients.
which.min(val.errors)
```

```
## [1] 7
```

```
coef(regfit.best,7)
```

```
##  (Intercept)    cylinders displacement       weight acceleration
## -12.85459278  -0.30559961   0.01519892  -0.00734830   0.13138150
##         year      origin2      origin3
##   0.70905703   1.11584946   2.18205585
```

(d)

Is this result different than the one you generated in question 1 for best subsets? This result is same as the one that generated in Q1.

(e)

```
regfit.best=regsubsets(mpg~.-name,data=Auto,nvmax=7)
coef(regfit.best,7)
```

```
##   (Intercept)     cylinders  displacement    horsepower        weight
## -16.332312787  -0.502767012   0.023372075  -0.025002677  -0.006459817
##          year       origin2       origin3
##   0.773883341   2.634517472   2.857355960
```

5

(a)

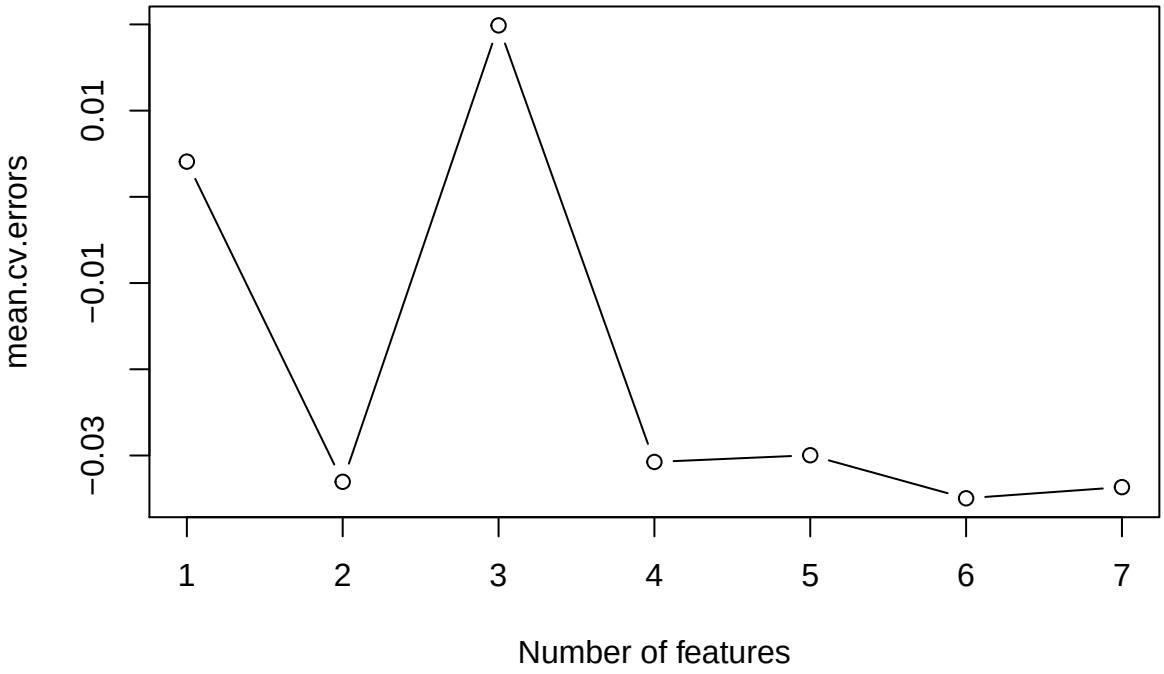
```
k=3
set.seed(1)
Auto$origin=factor(Auto$origin)
folds=sample(1:k,nrow(Auto),replace = TRUE)
cv.errors=matrix(NA,k,7,dimnames = list(NULL, paste(1:7)))
predict.regsubsets = function(object, newdata, id, ...) {
    form = as.formula(object$call[[2]])
    mat = model.matrix(form, newdata)
    coefi = coef(object, id = id)
    xvars=names(coefi)
    mat[, names(coefi)] %*% coefi
}
for(j in 1:k){
  best.fit=regsubsets(mpg~.-name,data = Auto[folds!=j,],nvmax = 7)
  for (i in 1:7) {
    pred=predict(best.fit,Auto[folds==j,],id=i)
    cv.errors[j,i]=mean((Auto$mpg[folds==j]-pred))
  }
}
mean.cv.errors=apply(cv.errors,2,mean)
par(mfrow=c(1,1))
mean.cv.errors
```

```
##           1           2           3           4           5
##  0.004087927 -0.033051909  0.019884940 -0.030749359 -0.029970830
##           6           7
## -0.034964546 -0.033669173
```

The best validation error is -0.034964546.

(b)

```
# Plot the validation error as a function of k, the number of features.
par(mfrow=c(1,1))
plot(mean.cv.errors,xlab="Number of features",type='b')
```



(c)

```
# Show the coefficients.
coef(best.fit,6)
```

```
##   (Intercept)  displacement    horsepower        weight          year
## -11.523520546   0.012343854  -0.016853494  -0.006685956   0.701758918
##       origin2       origin3
##   1.311818705   2.263924981
```

(d)

Is this result different than the one you generated in question 1 for best subsets? This result is different from the one that generated in Q1. The result shows the best model is the one that only contains 6 variables.

(e)

```
reg.best=regsubsets(mpg~.-name,data=Auto,nvmax=7)
coef(reg.best,6)
```

```
##   (Intercept)  displacement    horsepower        weight          year
## -17.503569715   0.015548663  -0.023044411  -0.006565293   0.774863460
##       origin2       origin3
```

```
##   2.595820822   2.772209410
```