STAT393_Assignment2

SHI YUNQI

2020/9/9

```r
library(MASS)
attach(Boston)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

Q1.

```r
# Design matrix
X <- model.matrix(medv ~ lstat + age, data = Boston)
X[1:10,] ## the fist 10 rows of the matrix.
```

```
##    (Intercept) lstat   age
## 1            1  4.98  65.2
## 2            1  9.14  78.9
## 3            1  4.03  61.1
## 4            1  2.94  45.8
## 5            1  5.33  54.2
## 6            1  5.21  58.7
## 7            1 12.43  66.6
## 8            1 19.15  96.1
## 9            1 29.93 100.0
## 10           1 17.10  85.9
```

The fist 10 rows of the matrix are shown above.

Q2.

```r
y = medv
beta_hat <- solve(t(X)%*%X) %*% t(X) %*% y
library(pander)
pander(data.frame(beta_hat))
```

	beta_hat
(Intercept)	33.22
lstat	-1.032
age	0.03454

LSE of $\hat{\beta}_0$ is 33.22, LSE of $\hat{\beta}_1$ is -1.032, LSE of $\hat{\beta}_2$ is 0.03454.

Q3.

```r
y_hat <- X %*% beta_hat
y_hat[1:10,] ## t the first 10 predicted values
```

```
##         1         2         3         4         5         6         7         8
## 30.335350 26.515202 31.174183 31.770610 29.594138 29.873436 22.694801 16.778358
##         9        10
##  5.787382 18.541747
```

The first 10 predicted values are 30.335350, 26.515202, 31.174183, 31.770610, 29.594138, 29.873436, 22.694801, 16.778358, 5.787382, 18.541747.

Q4.

```r
SSE = t(y-y_hat)%*%(y-y_hat)
pander(c(SSE=SSE))
```

SSE
19168

SSE is 19168.

Q5.

```r
n <- length(y)
n
```

```
## [1] 506
```

```r
p <- ncol(X)
p
```

```
## [1] 3
```

```r
RSE <- sqrt(SSE/(n-p))
RSE
```

```
##          [,1]
## [1,] 6.173136
```

Residual standard error is 6.173136.

Q6.

```r
Var_beta_hat=as.numeric(RSE^2)*solve(t(X)%*%X)
Var_beta_hat
```

```
##              (Intercept)        lstat           age
## (Intercept)  0.534137493 -0.0050496041 -0.0057591486
## lstat       -0.005049604  0.0023223465 -0.0003548703
## age         -0.005759149 -0.0003548703  0.0001494620
```

```r
SE = c(sqrt(Var_beta_hat[1,1]),sqrt(Var_beta_hat[2,2]),sqrt(Var_beta_hat[3,3]))
SE_beta_hat0 <-sqrt(Var_beta_hat[1,1])
SE_beta_hat0 ## se(betahat 0)
```

```
## [1] 0.7308471
```

```r
SE_beta_hat1 <- sqrt(Var_beta_hat[2,2])
SE_beta_hat1 ## se(betahat 1)
```

```
## [1] 0.04819073
```

```r
SE_beta_hat2 <- sqrt(Var_beta_hat[3,3])
SE_beta_hat2 ## se(betahat 2)
```

```
## [1] 0.01222547
```

The variance matrix of beta_hat is shown above.

SE($\hat{\beta}_0$) is 0.7308471, SE($\hat{\beta}_1$) is 0.04819073, SE($\hat{\beta}_2$) is 0.01222547.

Q7.

```r
# Get the coefficient matrix
model <- lm(medv ~ lstat + age, data = Boston)
Coef <- summary(model)$coefficients
Coef
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 33.22276053 0.73084711  45.457881 2.943785e-180
## lstat       -1.03206856 0.04819073 -21.416330  8.419554e-73
## age          0.03454434 0.01222547   2.825605  4.906776e-03
```

Interpretation: When the value of lstat and age is zero, then the value of the median value of owner-occupied homes is 33.22276053. For given amount of age, an additional 1 unit on lstat leads to an deccrease in the median value of owner-occupied homes by approximately 1.03206856 units. For given amount of lstat, an additional 1 unit on age leads to an increase in the median value of owner-occupied homes by approximately 0.03454434 units.

```
# 95% confidence intervals
lolim=Coef[,1] - qt(0.975,n-p)*Coef[,2]
uplim=Coef[,1] + qt(0.975,n-p)*Coef[,2]
pander(data.frame(lolim,uplim))
```

	lolim	uplim
(Intercept)	31.79	34.66
lstat	-1.127	-0.9374
age	0.01053	0.05856

Interpretation: β_0 is the slope of this model. We are 95% confident that the median value of owner-occupied homes is expected to be as low as 31.79 units and as high as 34.66 units in $1000s if the value of lstat and age is zero. With 95% of confidence, for an additional 1 unit of lower status of the population, the increase is as low as -1.127 units and as high as -0.9374 units. For an additional 1 unit of proportion of owner-occupied units built prior to 1940, with 95% of confidence, the increase is as low as 0.01053 units and as high as 0.05856 units.

Q8.

```
## t-test statistic for testing H0: beta_i=0 vs H1: beta_i is not equal to 0. i = 0,1,2.

T <- beta_hat/SE ## t-test statistic
T
```

```
##                    [,1]
## (Intercept)  45.457881
## lstat       -21.416330
## age           2.825605
```

```
p_val = 2 * (1-pt(abs(T),n-p))
p_val
```

```
##                     [,1]
## (Intercept) 0.000000000
## lstat       0.000000000
## age         0.004906776
```

For testing H_0: β_0=0 vs H_1: β_0 is not equal to 0. The t-test statisticis is 45.457881 with t_{503} df, p-value is nearly equal to 0, which is < 0.05, so β_0 is not equal to 0, which means we have enough evidence to reject H_0, the parameter β_0 is statistically significant different from 0, which has influence on the median value of owner-occupied homes in $1000s.

For testing H_0: β_1=0 vs H_1: β_1 is not equal to 0. The t-test statisticis is -21.416330 with t_{503} df, p-value is is nearly equal to 0, which is < 0.05, so β_1 is not equal to 0, which means we have enough evidence to reject H_0, the parameter β_1 is statistically significant different from 0, which has influence on the median value of owner-occupied homes in $1000s.

For testing H_0: β_2=0 vs H_1: β_2 is not equal to 0. The t-test statisticis is 2.825605 with t_{503} df. p-value is $0.004906776 < 0.05$, so β_2 is not equal to 0, which means we have enough evidence to reject H_0, the parameter β_2 is statistically significant different from 0, which has influence on the median value of owner-occupied homes in $1000s.

Q9.

```
y_bar=mean(y)
SST = t(y-y_bar)%*%(y-y_bar)
SSR = t(y_hat-y_bar)%*%(y_hat-y_bar)
pander(c(SST=SST,SSR=SSR, SSE=SSE))
```

SST	SSR	SSE
42716	23548	19168

```
##Check the equation: SST = SSR + SSE
S=round(SSR + SSE,0) ## round to integer
S
```

```
##       [,1]
## [1,] 42716
```

We have SST = 42716, SSR = 23548, SSE = 19168 and then we get S=SSR+SSE, which is equal to 42716 by using R to calculate, i.e. equal to the value of SST (round to integer), so SST = SSR + SSE.

Q10.

```
p=ncol(X)
F=(SSR/(p-1))/(SSE/(n-p))## F test statistic
p_val=pf(F, (p-1), (n-p), lower.tail = FALSE)
pander(c(F=F, p_value = p_val))
```

F	p_value
309	2.982e-88

Since the p-value is very small, which is nearly equal to 0, we reject H_0. We conclude that we have strong evidence that at least one at lstat and age have effect on the median value of owner-occupied homes in $1000s.

Q11.

```
## compute R square
R_square <- SSR/SST
R_square
```

```
##           [,1]
## [1,] 0.5512689
```

```r
## compute adjusted R square
adjusted_R_square <- 1-((SSE/(n-p))/(SST/(n-1)))
adjusted_R_square
```

```
##           [,1]
## [1,] 0.5494847
```

So R square is 0.5512689, adjusted R square is 0.5494847. Interpretation: R squared means 55.13% of the variation in the output variable is explained by the input variables. Adjusted R square calculates R square from only those variables whose addition in the model which are significant is 54.95%.

Q12.

```r
model <- lm(medv ~ lstat + age, data = Boston)
model.matrix(model)[1:10,]
```

```
##    (Intercept) lstat   age
## 1            1  4.98  65.2
## 2            1  9.14  78.9
## 3            1  4.03  61.1
## 4            1  2.94  45.8
## 5            1  5.33  54.2
## 6            1  5.21  58.7
## 7            1 12.43  66.6
## 8            1 19.15  96.1
## 9            1 29.93 100.0
## 10           1 17.10  85.9
```

```r
summary(model)$coef
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 33.22276053 0.73084711  45.457881 2.943785e-180
## lstat       -1.03206856 0.04819073 -21.416330  8.419554e-73
## age          0.03454434 0.01222547   2.825605  4.906776e-03
```

```r
summary(model)$sigma
```

```
## [1] 6.173136
```

```r
summary(model)$r.squared
```

```
## [1] 0.5512689
```

```r
summary(model)$adj.r.squared
```

```
## [1] 0.5494847
```

```r
summary(model)$fstatistic
```

```
##    value    numdf    dendf
## 308.9693   2.0000 503.0000
```

```r
confint(model)
```

```
##                   2.5 %      97.5 %
## (Intercept) 31.78687150 34.65864956
## lstat       -1.12674848 -0.93738865
## age          0.01052507  0.05856361
```

```r
vcov(model)
```

```
##              (Intercept)         lstat           age
## (Intercept)  0.534137493 -0.0050496041 -0.0057591486
## lstat       -0.005049604  0.0023223465 -0.0003548703
## age         -0.005759149 -0.0003548703  0.0001494620
```

From above results, the code reproduce the before calculation.

Q13.

```r
new = data.frame(lstat=c(mean(lstat)), age=c(mean(age)))
#95% confidence interval for y
predict(model, newdata=new,interval = "confidence" )
```

```
##        fit      lwr      upr
## 1 22.53281 21.99364 23.07198
```

```r
#95% prediction interval for y
predict(model, newdata=new,interval = "prediction" )
```

```
##        fit      lwr      upr
## 1 22.53281 10.39252 34.67309
```

95% confidence interval for y is (21.99364, 23.07198) with fit value = 22.53281, 95% prediction interval for y is (10.39252,34.67309) with fit value = 22.53281.

Q14.

```r
model1 = lm(medv ~ 1 )
model2 = lm(medv ~ lstat)
model3 = lm(medv ~ lstat + age )
anova(model1, model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ 1
## Model 2: medv ~ lstat
```

```
## Model 3: medv ~ lstat + age
##   Res.Df   RSS Df Sum of Sq       F    Pr(>F)
## 1    505 42716
## 2    504 19472  1   23243.9 609.955 < 2.2e-16 ***
## 3    503 19168  1     304.3   7.984  0.004907 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model1 vs model2: H_0: model1 is true, H_1: model2 is true. test statistic: F = 609.955 with F(1,504) df. P-value $< 2.2e\text{-}16 < 0.05$, reject H_0, indicating given that lstat in the model has effect on medv, so we choose model2.

model2 vs model3: H_0: model2 is true, H_1: model3 is true. test statistic: F = 7.984 with F(1,503) df. P-value $=0.004907 < 0.05$, reject H_0, indicating given that lstat and age in the model has effect on medv, so we choose model3.

Q15.

```
pander(AIC(model1,model2,model3))
```

	df	AIC
model1	2	3684
model2	3	3289
model3	4	3283

```
pander(BIC(model1,model2,model3))
```

	df	BIC
model1	2	3693
model2	3	3302
model3	4	3300

Both AIC and BIC choose the model 3 as the best model among the 3 models.

Q16.

```
##R square for the 3 models
summary(model1)$r.squared
```

```
## [1] 0
```

```
summary(model2)$r.squared
```

```
## [1] 0.5441463
```

```
summary(model3)$r.squared
```

```
## [1] 0.5512689
```

R square for model1 is 0, for model2 is 0.5441463, for model3 is 0.5512689.

```r
## adjust R square for the 3 models
summary(model1)$adj.r.squared
```

```
## [1] 0
```

```r
summary(model2)$adj.r.squared
```

```
## [1] 0.5432418
```

```r
summary(model3)$adj.r.squared
```

```
## [1] 0.5494847
```

adjust R square for model1 is 0, for model2 is 0.5432418, for model3 is 0.5494847.

There are large increase of R square from 0 to 0.5441463 for the addition of variable lstat. The increase is around 0.01 for the addition of variable age. The adjusted R squared can be used for model comparison. Since the model 3 has the largest adj_R_sq = 0.5494847, the model 3 is the best model among the 3 models.