regression analysis HW2

2019150445/Shin Baek Rok

2020 10 11

1.

a)

```r
#By matrix
x<-c(1,2,3,3,4,5,5)
y<-c(3,7,5,8,11,14,12)
X<-cbind(1,x)
coef<-solve(t(X)%*%X)%*%t(X)%*%y

coef
```

```
##        [,1]
##   0.5319149
## x 2.4468085
```

```r
#By simple code
fitted<-lm(y~x)
fitted
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.5319       2.4468
```

By these coefficients, we can fit the regression equation.

$\hat{y}_i = 0.531949 + 2.4468085x_i, \quad i = 1, ..., 7$

b)

```r
summary(fitted)
```

```
##
## Call:
## lm(formula = y ~ x)
```

```
##
## Residuals:
##        1        2        3        4        5        6        7
##  0.02128  1.57447 -2.87234  0.12766  0.68085  1.23404 -0.76596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5319     1.5881   0.335  0.75127
## x             2.4468     0.4454   5.494  0.00273 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.632 on 5 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8294
## F-statistic: 30.18 on 1 and 5 DF,  p-value: 0.002729
```

```r
t<-(2.4468-2)/0.4454
1-pt(t,7-2)
```

```
## [1] 0.1809193
```

c)

```r
summary(fitted)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##        1        2        3        4        5        6        7
##  0.02128  1.57447 -2.87234  0.12766  0.68085  1.23404 -0.76596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5319     1.5881   0.335  0.75127
## x             2.4468     0.4454   5.494  0.00273 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.632 on 5 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8294
## F-statistic: 30.18 on 1 and 5 DF,  p-value: 0.002729
```

```r
0.5319+(1-pt(0.25,df=5))*1.5881
```

```
## [1] 1.177093
```

```r
0.5319-(1-pt(0.25,df=5))*1.5881
```

```
## [1] -0.1132927
```

Therefore, 95% confidence interval is

$$(-0, 113, 1.177)$$

And thus we cannot reject $H_0 : \beta_0 = 1$ since this 95% CI contains 1.

d)

```r
fitted1<-lm(y~1)
anova(fitted1,fitted)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Res.Df    RSS Df Sum of Sq     F   Pr(>F)
## 1      6 93.714
## 2      5 13.319  1    80.395 30.18 0.002729 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under H_0,

$$SSE(RM) = \sum (y_i - \bar{y})^2 = SST = 93.714$$

$$SSE(FM) = \sum (y_i - \hat{y}_i)^2 = 13.319, \quad (df = 7 - 2 = 5)$$

$$SSE(RM) - SSE(FM) = 93.714 - 13.319 = 80.395, \quad (df = 2 - 1 = 1)$$

Therefore,

$$F = \frac{80.395/1}{13.319/5} = 30.18057$$

```r
1-pf(df1=1,df2=5,q=30.18057) #p-value of F statistic
```

```
## [1] 0.002728809
```

e)

We should find the confidence interval for the mean response $\hat{\mu_0}$ when $x_0 = 4$.

```r
predict(fitted, interval='confidence', newdata=data.frame(x=4),level=0.9)
```

```
##        fit      lwr      upr
## 1 10.31915 8.920531 11.71777
```

2.

```r
mpg<-c(23,21,20,19,22,21,20,19,24,17,19)
engine<-c(3471, 2979, 4195, 4701, 3471, 3960, 4701, 4701, 3311, 4664, 4605)
hp<-c(260,225,275,235,240,195,235,265,230,235,302)
weight<-c(4420,4586,4787,4379,4439,3786,3786,3786,3860,5390,4834)
#load data
```

a)

```r
lm(mpg~engine+hp+weight)
```

```
##
## Call:
## lm(formula = mpg ~ engine + hp + weight)
##
## Coefficients:
## (Intercept)       engine           hp       weight
##   35.180504    -0.002568     0.015389    -0.001843
```

The fitted regression equation is

$$\hat{mpg} = 35.1805 - 0.026 \times engine + 0.0154 \times hp - 0.0018 \times weight$$

b) When weight increases by one unit, while other variables(engine and hp) are fixed, estimated mpg decreases by 0.0018.

c)

```r
full_model<-lm(mpg~engine+hp+weight)
reduced_model<-lm(mpg~1)
anova(reduced_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ engine + hp + weight
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     10 40.727
## 2      7  5.699  3    35.028 14.341 0.002253 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In anova table, F-statistic is large enough(i.e. p-value for F-statistic is small enough) to reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Thus we can not ignore all variables(There is at least one significant variables).

d)

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = mpg ~ engine + hp + weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54163 -0.06518  0.18154  0.29778  0.89573
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.1805036  3.1118592  11.305 9.48e-06 ***
## engine      -0.0025675  0.0004576  -5.611 0.000806 ***
## hp           0.0153889  0.0112717   1.365 0.214421
## weight      -0.0018431  0.0005841  -3.156 0.016027 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9023 on 7 degrees of freedom
## Multiple R-squared:  0.8601, Adjusted R-squared:  0.8001
## F-statistic: 14.34 on 3 and 7 DF,  p-value: 0.002253
```

In summary of the fitted model, we can find the p-value for individual variables. Since p-value for hp is big, we can say effect of hp variable is not significant on mpg and we may remove the hp variable from the model. Other variables' p-values are small and we can say its' effects are significant on mpg.

e)

```r
summary(lm(mpg~engine+weight))
```

```
##
## Call:
## lm(formula = mpg ~ engine + weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7023 -0.5365  0.1389  0.5184  1.2068
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.9035646  2.9940991  12.325 1.75e-06 ***
## engine      -0.0023726  0.0004576  -5.185 0.000838 ***
## weight      -0.0015555  0.0005734  -2.713 0.026551 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9498 on 8 degrees of freedom
## Multiple R-squared:  0.8228, Adjusted R-squared:  0.7785
## F-statistic: 18.57 on 2 and 8 DF,  p-value: 0.0009858
```

And the fitted regression model is

$$\hat{mpg} = 36.904 - 0.00237 \times engine - 0.00156 \times weight$$

f)

```r
reduced_model2<-lm(mpg~engine+weight)
anova(reduced_model2,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ engine + weight
## Model 2: mpg ~ engine + hp + weight
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1      8 7.2166
## 2      7 5.6990  1    1.5175 1.864 0.2144
```

In anova between reduced model & full model, we cannot reject $H_0 : \beta_2 = 0$ since F-statistic is small & p-value of F-statistic is quiet large. That means the reduced model gives as good a fit as the full model.

3.

g), h)

t-value=Coef/SE, thus (g)=-23.4325/12.74=-1.839286

In the same way, (h)/0.1528=8.32. Thus (h)=8.32*0.1528=1.271296

a), e), i), m)

It is simple linear regression, and there is 20 observations. Thus, (i)=n(#observations)=20 , (e)=(m)=n(#observations)-p(#variables)-1=20-1-1=18, (a)=p(#variables)=1.

b)

(b)=MSR=SSR/p=1848.76/1=1848.76.

c)

In SLR, $F = t_1^2 = 8.32^2 = 69.2224 = (c)$

f)

Since F=MSR/MSE=1848.76/(f)=69.2224, (f)=1848.76/69.2224=26.70754.

d)

Since MSE=SSE/n-p-1=SSE/18=26.70754, SSE=(d)=26.70754*18=480.7357

j)

$R^2 = (j) = SSR/SST = SSR/(SSR + SSE) = 1848.76/(1848.76 + 480.7357) = 0.793631$

k)

$R_a^2 = (k) = 1 - \frac{SSE/18}{SST/19} = 1 - \frac{480.7537/18}{(1848.76+480.7537)/19} = 0.7822$

l)

$\hat{\sigma} = (l) = \sqrt{SSE/18} = \sqrt{480.7537/18} = 5.16803$