Homework 7

陈旭鹏

2018/4/15

1

a

```
# read the data
setwd('~/Desktop/三春/5线性回归分析/作业/HW7/')
Data<-read.table("hw7.txt")
names(Data) = c("Hours","Cases","Costs","Holiday")
Fit = lm(Hours~Cases+Costs+Holiday, data=Data)
anova(Fit)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Cases      1  136366  136366  6.6417   0.01309 *
## Costs      1    5726    5726  0.2789   0.59987
## Holiday    1 2034514 2034514 99.0905 2.941e-13 ***
## Residuals 48  985530   20532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSTO = sum( anova(Fit)[,2] )
MSE = anova(Fit)[4,3]
SSR = sum( anova(Fit)[1:3,2] )
MSR = SSR / 3
SSE = anova(Fit)[4,2]
```

From the table we have: $SSR(X_1) = 136366$,$SSE(X_1, X_2, X_3) = 985530$

```
Fit2 = lm(Hours~Cases+Holiday, data=Data)
anova(Fit2)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Cases      1  136366  136366   6.7344   0.01244 *
## Holiday    1 2033565 2033565 100.4276 1.875e-13 ***
## Residuals 49  992204   20249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_3|X_1) = 2033565$

$SSR(X_2|X_1,X_3) = SSE(X_1,X_3) - SSE(X_1,X_2,X_3)$ = 992204 - 985530 = 6674

b

$$H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0.$$

From a we have:

$$SSR(X_2|X_1,X_3) = 6674, SSE(X_1,X_2,X_3) = 985530$$
$$F^* = \frac{(6674/1)}{985530/48} = 0.32491$$
$$F(0.95,1,48) = 4.04265$$
$$If\ F^* \leqslant 4.04265\ concludeH_9, otherwise\ concludeH_a$$
$$P - value = 0.5713$$

c

```
Fit2 = lm(Hours~Cases+Costs, data=Data)
anova(Fit2)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Cases      1  136366  136366  2.2125 0.1433
## Costs      1    5726    5726  0.0929 0.7618
## Residuals 49 3020044   61634
```

```
Fit3 = lm(Hours~Costs+Cases, data=Data)
anova(Fit3)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Costs      1   11395   11395  0.1849 0.6691
## Cases      1  130697  130697  2.1206 0.1517
## Residuals 49 3020044   61634
```

So we have $SSR(X_2|X_1) + SSR(X_1) = 136366 + 5726 = 11395 + 130697 = SSR(X_1|X_2) + SSR(X_2)$

Yes, it is always true because:
$SSR(X_2|X_1) + SSR(X_1) = SSE(X_1) - SSE(X_1,X_2) + SSR(X_1) = SSTO - SSE(X_1,X_2)$

$SSR(X_1|X_2) + SSR(X_2) = SSE(X_2) - SSE(X_1,X_2) + SSR(X_2) = SSTO - SSE(X_1,X_2)$

2

From question1, We have $SSR(X_1) = 136366, SSR(X_2) = 5726, SSR = 2176606, SSTO = 3162136$ \ So $R^2_{Y_1} = 0.0431, R^2_{Y_2} = 0.00181, R^2 = 0.6883$ \ From homework6 we have $r_12 = 0.10059216$, so $R^2_{12} =0.0101 $ \

```
Fit4 = lm(Hours~Costs, data=Data)
anova(Fit4)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Costs      1   11395   11395  0.1808 0.6725
## Residuals 50 3150741   63015
```

```
Fit5 = lm(Hours~Cases, data=Data)
anova(Fit5)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Cases      1  136366  136366  2.2534 0.1396
## Residuals 50 3025770   60515
```

$R^2_{Y1|2} = \frac{SSR(X1|X2)}{SSE(X2)}$ = 130697/3150741 = 0.04148\ $R^2_{Y2|1} = \frac{SSR(X2|X1)}{SSE(X1)}$ = 5726/3025770 = 0.001892\

```
Fit6 = lm(Hours~Cases+Holiday, data=Data)
anova(Fit6)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Cases      1  136366  136366   6.7344   0.01244 *
## Holiday    1 2033565 2033565 100.4276 1.875e-13 ***
## Residuals 49  992204   20249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$R^2_{Y2|13} = \frac{SSR(X2|X1,X3)}{SSE(X1,X3)}$

$SSR(X_2|X_1, X_3) = SSE(X_1, X_3) - SSE(X_1, X_2, X_3)$ = 992204 - 985530 = 6674

$SSE(X_1, X_3)$ = 992204, so $R^2_{Y2|13}$ = 6674/992204 = 0.006726

3

a

```
Fit = lm(Hours~Cases, data=Data)
summary(Fit)
```

```
##
## Call:
## lm(formula = Hours ~ Cases, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -356.18 -164.64  -56.07  111.23  619.01
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.080e+03  1.917e+02  21.283   <2e-16 ***
## Cases       9.355e-04  6.232e-04   1.501     0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246 on 50 degrees of freedom
## Multiple R-squared:  0.04312,    Adjusted R-squared:  0.02399
## F-statistic: 2.253 on 1 and 50 DF,  p-value: 0.1396
```

So regression function is $\hat{Y} = 4080 + 0.0009355X_1$ \

b

The regression function in 6.10a is $Y = 0.0007871X_1 - 13.17X_2 + 623.6X3 + 4150$

The coefficient β_1 is bigger than coefficient in 6.10a.

c

No, from question 1, $SSR(X_1) = 136366, SSR(X_1|X_2) = 130697$. It's not substantial

d

The correlation of X_1, X_2 is highest in all predictors, so the $SSR(X_1) and SSR(X_1|X_2)$ don't have substantial difference.

4

a

To run a polynomial regression model on one or more predictor variables, it is advisable to first center the variables by subtracting the corresponding mean of each, in order to reduce the intercorrelation among the variables.

```
x1 <- Data$Cases - mean(Data$Cases)
x3 <- Data$Holiday - mean(Data$Holiday)
x1sq <- x1^2
x3sq <- x3^2
x1x3 <- x1 * x3
Grocery <- cbind( Data, x1, x3, x1sq, x3sq, x1x3 )
Poly <- lm(  Hours ~ x1 + x3 + x1sq + x3sq + x1x3, data=Grocery )
summary(Poly)
```

```
##
## Call:
## lm(formula = Hours ~ x1 + x3 + x1sq + x3sq + x1x3, data = Grocery)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -288.253 -102.112   -7.251   72.363  294.646
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.367e+03  2.607e+01 167.500  < 2e-16 ***
## x1           8.610e-04  5.514e-04   1.561    0.125
## x3           6.237e+02  6.571e+01   9.491 1.68e-12 ***
## x1sq        -1.154e-09  5.481e-09  -0.211    0.834
## x3sq                NA         NA      NA       NA
## x1x3        -8.870e-05  8.760e-04  -0.101    0.920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 145.2 on 47 degrees of freedom
## Multiple R-squared:  0.6865, Adjusted R-squared:  0.6599
## F-statistic: 25.74 on 4 and 47 DF,  p-value: 2.476e-11
```

So the model is $\hat{Y} = 4367 + 8.61 \times 10^{-4} X_1 + 623.7 X_3 - 1.154 \times 10^{-9} X_1^2 - 8.87 \times 10^{-5} X_1 X_3$

b

```
anova(Poly)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x1         1  136366  136366  6.4663  0.01435 *
## x3         1 2033565 2033565 96.4287 5.74e-13 ***
## x1sq       1     815     815  0.0386  0.84500
## x1x3       1     216     216  0.0103  0.91978
## Residuals 47  991173   21089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Fit7 <-lm(  Hours ~ x1 + x3, data=Grocery )
anova(Fit7)
```

```
## Analysis of Variance Table
##
## Response: Hours
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x1         1  136366  136366   6.7344   0.01244 *
## x3         1 2033565 2033565 100.4276 1.875e-13 ***
## Residuals 49  992204   20249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(0.95,3,46)
```

```
## [1] 2.806845
```

$$
\begin{gathered}
H_0 : \beta_3, \beta_4, \beta_5 = 0, H_a : \text{not all } \beta_k in H_0 = 0 \\
F^* = \frac{SSR(X_1^2, X_3^2, X_1X_3|X_1, X_3)/3}{SSE(X_1^2, X_3^2, X_1X_3, X_1, X_3)/(n-6)} \\
= \frac{(SSE(X_1, X_3) - SSE(X_1^2, X_3^2, X_1X_3, X_1, X_3))/3}{991173/46} \\
= \frac{(992204 - 991173)/3}{991173/46} \\
= 0.01594945 \\
F(0.95, 3, 46) = 2.806845, So\ F^* < F(0.95, 3, 46), \text{Do not reject H\_0.}
\end{gathered}
$$

```r
pf(0.01594945,3,46)
```

```
## [1] 0.002785933
```

p-value = 0.002785933