Hw2

Stone Cai

2024-02-01

Inputting Data

```r
lbw <- read.csv("/Users/stone/Documents/Stat 408/Homework Data Files/Hw2/lowbirthwt.csv")
summary(lbw)
```

```
##     headcirc         length         gestage         birthwt         momage     
##  Min.   :21.00   Min.   :20.00   Min.   :23.00   Min.   : 560   Min.   :14.00  
##  1st Qu.:25.00   1st Qu.:35.00   1st Qu.:27.00   1st Qu.: 880   1st Qu.:23.00  
##  Median :27.00   Median :38.00   Median :29.00   Median :1155   Median :28.00  
##  Mean   :26.45   Mean   :36.82   Mean   :28.89   Mean   :1099   Mean   :27.73  
##  3rd Qu.:28.00   3rd Qu.:39.00   3rd Qu.:31.00   3rd Qu.:1326   3rd Qu.:32.00  
##  Max.   :35.00   Max.   :43.00   Max.   :35.00   Max.   :1490   Max.   :41.00  
##     toxemia    
##  Min.   :0.00  
##  1st Qu.:0.00  
##  Median :0.00  
##  Mean   :0.21  
##  3rd Qu.:0.00  
##  Max.   :1.00
```

```r
mm<-read.csv("/Users/stone/Documents/Stat 408/Homework Data Files/Hw2/musclemass.csv")
summary(mm)
```

```
##   muscle_mass          age       
##  Min.   : 52.00   Min.   :41.00  
##  1st Qu.: 73.00   1st Qu.:50.25  
##  Median : 84.00   Median :60.00  
##  Mean   : 84.97   Mean   :59.98  
##  3rd Qu.: 97.00   3rd Qu.:70.00  
##  Max.   :119.00   Max.   :78.00
```

Problem1

```r
lm_birthwt<-lm(headcirc~birthwt,data=lbw)
summary (lm_birthwt)
```

```
## 
## Call:
## lm(formula = headcirc ~ birthwt, data = lbw)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1622 -0.9399 -0.3071  0.5471 10.0398
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.822e+01  6.447e-01   28.26   <2e-16 ***
## birthwt     7.492e-03  5.699e-04   13.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.531 on 98 degrees of freedom
## Multiple R-squared:  0.6381, Adjusted R-squared:  0.6344
## F-statistic: 172.8 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
anova (lm_birthwt)
```

```
## Analysis of Variance Table
##
## Response: headcirc
##           Df Sum Sq Mean Sq F value    Pr(>F)
## birthwt    1 405.06  405.06  172.82 < 2.2e-16 ***
## Residuals 98 229.69    2.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1a) See above 1b)

$$\hat{\beta_0} = 18.22$$

and

$$\hat{\beta_1} = 0.007492$$

1c) To test hypothesis

$$H_0 : \beta_1 = 0$$

or

$$H_A : \beta_1 \neq 0$$

we find t-statistic given by formula

$$t = \hat{\beta_1}/(SE(\hat{\beta_1}))$$

You can look at R out put from lm_birthwt for estimate of birthwt coefficent and standard error of coefficient

$$t = 0.007492/(5.7 * 10^-4)$$

= 13.15 which gives ~ 0 p-value. Thus, we reject

$$H_0 : \beta_1 = 0$$

in favor of

$$H_A : \beta_1 \neq 0$$

1d) Variance of error can be obtained two ways. It is the square of the Resdiual standard error and it also the Mean Sq output from the anova table for residuals. In either case the estimate is 2.34

1e) SSR = 405.06 SSE= 229.69 SSTO= SSR+ SSE= 634.76. Coefficent of determination is

$$R^2 = SSR/SSTO = 0.6381$$

. This can also be found from above R out put in summary of lm. Coefficent implies that 63.81% of the variability among the observed values of head circumference is explained by the linear relationship between head circumference and birth weight.

Problem 2

```r
lm_momage<-lm(headcirc~momage,data=lbw)
summary (lm_momage)
```

```
##
## Call:
## lm(formula = headcirc ~ momage, data = lbw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1920 -1.7027  0.0908  1.4930  8.8145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.89945    1.20190  20.717   <2e-16 ***
## momage       0.05592    0.04238   1.319     0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.523 on 98 degrees of freedom
## Multiple R-squared:  0.01746,    Adjusted R-squared:  0.007429
## F-statistic: 1.741 on 1 and 98 DF,  p-value: 0.1901
```

```r
anova (lm_momage)
```

```
## Analysis of Variance Table
##
## Response: headcirc
##           Df Sum Sq Mean Sq F value Pr(>F)
## momage     1  11.08  11.080   1.741 0.1901
## Residuals 98 623.67   6.364
```

2a) Data loaded in above. b). $\hat{\beta}_0 = 24.899$ and $\hat{\beta}_1 = 0.05592$ c).To test hypothesis

$$H_0 : \beta_1 = 0$$

or

$$H_A : \beta_1 \neq 0$$

we find t-statistic given by formula

$$t = \hat{\beta}_1/(SE(\hat{\beta_1}))$$

You can look at R out put from lm_birthwt for estimate of birthwt coefficent and standard error of coefficient

$$t = 0.05592/(0.04238)$$

= 1.32 which gives ~ 0.19 p-value. Thus, we FAIL TO REJECT

$$H_0 : \beta_1 = 0$$

in favor of

$$H_A : \beta_1 \neq 0$$

d) Estimator of variance is given by Residual standard error squared= 6.364

e) SSR = 11.08 SSE= 623.67 SSTO= SSR+ SSE= 634.76. Coefficent of determination is

$$R^2 = SSR/SSTO = 0.0175$$

. This can also be found from above R out put in summary of lm. Coefficent implies that 1.75% of the variability among the observed values of head circumference is explained by the linear relationship between head circumference and momage.

f) Based on first 2 questions we have evidence from both t-test and R squared values that there is strong relationship between headcirc and birthwt. Based on t-test we determined that the coefficient of bithwt in a simple regression model was significant whereas momage was not.

-Also

$$R^2$$

was much better for simple regression model with birthwt.

Problem 3)

```r
lm_mm<-lm(muscle_mass~age,data=mm)
summary (lm_mm)
```

```
##
## Call:
## lm(formula = muscle_mass ~ age, data = mm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

```r
anova (lm_mm)
```

```
## Analysis of Variance Table
##
## Response: muscle_mass
```

```
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## age         1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qt(0.025,58)
```

```
## [1] -2.001717
```

```r
confint(lm_mm)
```

```
##                  2.5 %     97.5 %
## (Intercept) 145.312572 167.380556
## age          -1.370545  -1.009446
```

a) Created simple LM with muscle_mass as response and age as predictor.Note: Age of people sampled ranges from 41-78 (is relevant for our interpretation).

Assume

$$\alpha = 0.05$$

and conduct one sided t-test with

$$H_0 : \beta_1 = 0$$

or

$$H_A : \beta_1 < 0$$

.

Test statistic and

$$SE(\hat{\beta}_1)$$

are given by R output.

$$t^\star = -13.19.$$

P-value from out put is 2*10^-16, but for a one sided test we halve the p-value. It is still ~ 0 so we reject null in favor of

$$H_A$$

. Alternatively, if we use decision rule approach for one-sided t-test we reject

$$H_0$$

if

$$t^\star < t(0.025, 58) = -2.001$$

. Since

$$t^\star = -13.19 < -2.001$$

we reject null.

b) From above context no. We can only make interpretation of the data for the range of those who were sampled. Only women from ages 41-78 were sampled, so it would not make sense to say

$$\hat{\beta}_0$$

provides relevant information even if p-value is significant. Would a new born ever have 156lbs of muscle mass contextually?

C) Contextually, coefficient intercept for age represents change in muscle mass per year increase in age. The difference in expected muscle mass for women who differ by a year is the age coefficent. Interval is given by formula

$$\hat{\beta}_1 \pm t\star(1-\alpha/2, n-2) * SE(\hat{\beta}_1)$$

. All values necessary to calculate CI are in R output, but can also be directly calculated using conf-int.

From R output this is (-1.371,-1.009). Age is not necessary to make CI since the CI is dependent on estimated slope of coefficient, its standard error, and the parameters of its t-statistic. None of these are reliant on specific age.

Problem 4) ## same output as from problem 3

```
qf(0.9,1,58)
```

```
## [1] 2.794089
```

a). Anova table set up in R output from above.

b).

$$H_0 : \beta_1 = 0$$

or

$$H_A : \beta_1 \neq 0$$

F-distribution follows degree freedom (p-1, n-2) which in this case gives (1,58).

So if

$$F^\star > F(0.9, 1, 58)$$

then reject null hypothesis. F-star value is obtained from output above. 174.06> 2.80.

Thus we reject

$$H_0$$

and conclude there is a linear relationship between muscle mass and age.

c). Variance that remains "unexplained" is

$$1 - R^2$$

. From out put

$$1 - R^2$$

is 0.2499 so 24.99% of variance of model remains unexplained. This is relatively small.

d)

$$R^2 = SSR/SSTO$$

which from Anova table is 11627.5 / 15501.93 = 0.75. This is confirmed by linear model r-squared output for same model.

$$r = -\sqrt{R^2} = -0.866$$

Remember we want to take the negative square root to reflect that slope of our age coefficient is also negative.