Hw10

Dandong Tu

2017/11/8

1

a

```r
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: effects
```

```
##
## Attaching package: 'effects'
```

```
## The following object is masked from 'package:car':
##
##     Prestige
```

```r
data1=read.table("/Users/dandongtu/Downloads/Robey.txt",header = T)
m1=lm(tfr~region+contraceptors+region:contraceptors,data=data1)
m2=lm(tfr~contraceptors+region+region:contraceptors,data=data1)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: tfr
##                      Df Sum Sq Mean Sq  F value    Pr(>F)
## region                3 44.304  14.768  44.9534 3.576e-13 ***
## contraceptors         1 45.045  45.045 137.1158 8.226e-15 ***
## region:contraceptors  3  0.365   0.122   0.3706    0.7746
## Residuals            42 13.798   0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                      Sum Sq Df  F value    Pr(>F)
## region                1.677  3   1.7018    0.1812
## contraceptors        45.045  1 137.1158 8.226e-15 ***
## region:contraceptors  0.365  3   0.3706    0.7746
## Residuals            13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Type I anova

region

NH:$\beta_1 = 0$

AH:β_1 not equal to 0.

For type I anova, order matter, so that, the *region*, it's testing the main effect for *region*. And the tiny *p-value* we reject the H_0 and conclude that the *region* is statisticly signifant.

contraceptors

NH:tfr ~ region

AH:tfr~region+contraceptors

It's testing the main effect of *contraceptors* after the main effect of *region*. We reject the H_o because of the tiny *p-value* and we could say that *contraceptors* is statisticly signifant, and should include to the model.

region:contraceptors

NH:tfr~region+contraceptors

AH:tfr~region+contraceptors+region:contraceptors

It's testing the interation effect after the main effect of *region* and *contraceptors*. We could not reject the *H_0* due to the large *p-value* and its indicates that interation is statisticly insignifant and we could drop it from the model.

Type II anova (the model already has the effect from other regressions)

For type II anova with test the result from bottom to top.

region:contraceptors

NH:tfr~region+contraceptors

AH:tfr~region+contraceptors+region:contraceptors

The *p-value=0.7746* indicates that the interation is statisticly insignifant, thus we only consider the model with main effect.

contraceptors

NH:$\beta_1 = \beta_2 = 0$

AH:β_1 and β_2 are not equal to 0$

The *p-value=8.226e-15* indicates that the *contraceptors* is statisticly signifant.

region

$NH : \beta_2 = \beta_1 = 0$

$AH : \beta_2$ and β_1 are not equal to 0

The *p-value=3.576e-13* indicates that the *region* is statisticly signifant.

b

```r
Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                       Sum Sq Df  F value    Pr(>F)
## region                 1.677  3   1.7018    0.1812
## contraceptors         45.045  1 137.1158 8.226e-15 ***
```

```
## region:contraceptors  0.365  3   0.3706    0.7746
## Residuals            13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                      Sum Sq Df  F value    Pr(>F)
## contraceptors        45.045  1 137.1158 8.226e-15 ***
## region                1.677  3   1.7018    0.1812
## contraceptors:region  0.365  3   0.3706    0.7746
## Residuals            13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: tfr
##                      Df Sum Sq Mean Sq  F value    Pr(>F)
## region                3 44.304  14.768  44.9534 3.576e-13 ***
## contraceptors         1 45.045  45.045 137.1158 8.226e-15 ***
## region:contraceptors  3  0.365   0.122   0.3706    0.7746
## Residuals            42 13.798   0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: tfr
##                      Df Sum Sq Mean Sq  F value Pr(>F)
## contraceptors         1 87.672  87.672 266.8706 <2e-16 ***
## region                3  1.677   0.559   1.7018 0.1812
## contraceptors:region  3  0.365   0.122   0.3706 0.7746
## Residuals            42 13.798   0.329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, it is easy to see that the type II anova provides the same output for both models while type I anove doesn't.

For type I anova, due to the sequential nature and the fact that factors are tested in a particular order, the first factor are always different compare with type II anova. The second factor are always same as type II anova since the type II anova is testing the factors that already has the effects from other factors.

c

```r
Anova(m1)
```

```
## Anova Table (Type II tests)
```

```
##
## Response: tfr
##                        Sum Sq Df  F value    Pr(>F)    
## region                 1.677  3   1.7018    0.1812    
## contraceptors         45.045  1 137.1158 8.226e-15 ***
## region:contraceptors  0.365  3   0.3706    0.7746    
## Residuals             13.798 42                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the result of type II anova, we observed that the *contraceptors* is the only regressor that is statisticly significant in our model that alreadly contains the *region* and the interation. Therefore, we choose the model that only contains *contraceptors* which is the same result we have in HW8.

The reason behind this is that the process we have done in HW8 is the same process for checking type II anova.

2

Now, we have $\hat{e}_i = (w_i)^{1/2}(y_i - x_i^T\hat{\beta})$

Then:

$$RSS(\beta) = (Y - X\beta)^T W(Y - X\beta) = Y^T WY - Y^T WX\beta - \beta^T X^T WY + \beta^T X^T WX\beta$$
$$\frac{RSS(\beta)}{\beta} = -X^T W^T Y - X^T WY + 2X^T WX\beta$$

where $W^T = W$

When the derivative $\frac{\delta RSS(\beta)}{\delta\beta} = 0$,then

$$-2X^T WY + 2X^T WX\hat{\beta} = 0$$
$$(X^T WX)^{-1}X^T WX\hat{\beta} = (X^T WX)^{-1}X^T WY$$
$$\hat{\beta} = (X^T WX)^{-1}X^T WY$$

So that, if $X^* = W^{1/2}X$ and $Y^* = W^{1/2}Y$ and $W^{1/2}W^{1/2} = W$

then $\hat{\beta} = ((X^*)^T X*)^{-1}(X^*)^T Y^*$ is the weighted least squares coefficient estimateor for β

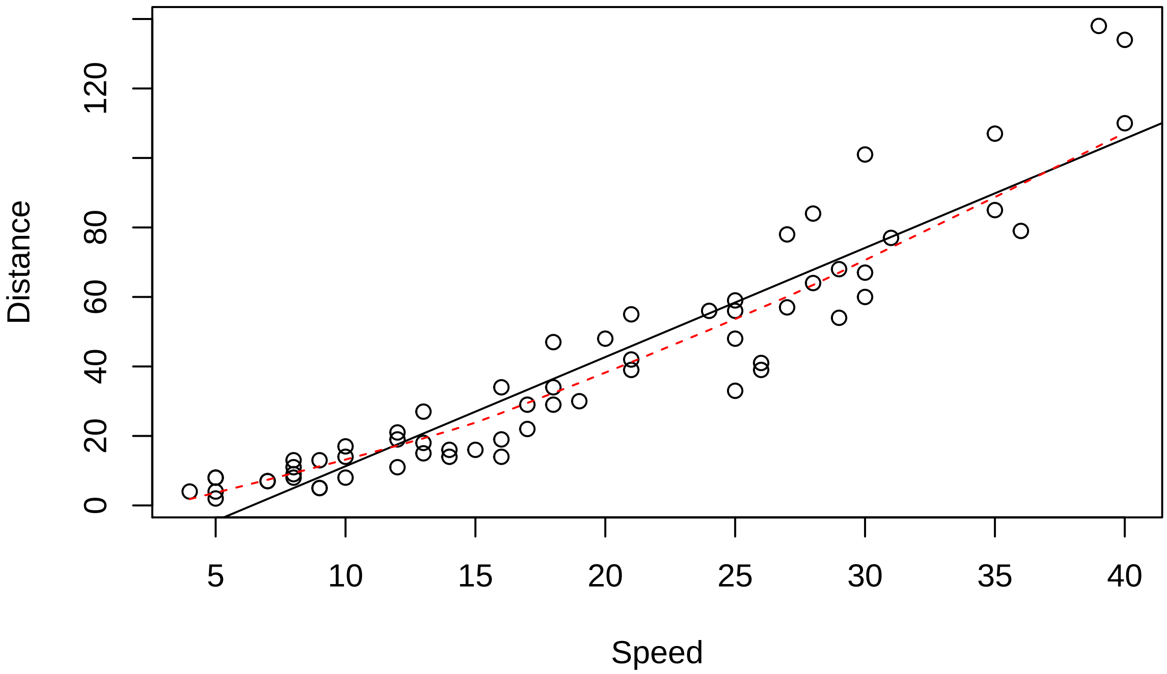
7.6

1

```r
head(stopping)
```

```
##   Speed Distance
## 1     4        4
## 2     5        2
## 3     5        4
## 4     5        8
## 5     5        8
## 6     7        7
```

```r
plot(Distance~Speed,data=stopping)
abline(lm(Distance~Speed,data=stopping))
lines(lowess(stopping$Distance~stopping$Speed),lty=2,col="red")
```



The solid line is for simple linear regression, and the red dashed line is a quadratic fit line. It is easy to see thatfor small and large values of *Speed*,the linear fit is not very good at predicting *Distance*

```r
m3=lm(Distance~Speed+I(Speed^2),data=stopping)
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: Distance
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Speed        1  59639   59639 605.198 < 2.2e-16 ***
## I(Speed^2)   1   2496    2496  25.329 4.835e-06 ***
## Residuals   59   5814      99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(m3)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + I(Speed^2), data = stopping)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5192  -5.4527  -0.5519   3.8442  27.9373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.58036    5.10266   0.310    0.758
## Speed        0.41607    0.55641   0.748    0.458
```

```
## I(Speed^2)   0.06556    0.01303   5.033 4.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.927 on 59 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9115
## F-statistic: 315.3 on 2 and 59 DF,  p-value: < 2.2e-16
```

From the result, we observed that the quadratic regression model is better fit.

2

```r
speed=(stopping$Speed)^2
ncvTest(m3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 22.97013    Df = 1     p = 1.645386e-06
```

```r
ncvTest(m3,~stopping$Speed)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ stopping$Speed
## Chisquare = 23.39216    Df = 1     p = 1.321162e-06
```

```r
ncvTest(m3,~stopping$Speed+speed)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ stopping$Speed + speed
## Chisquare = 23.46559    Df = 2     p = 8.026245e-06
```

For each test, we have the lower *p-vale*s, so that we are able to reject the hypothesis that the variance is constant.

3

```r
m4 = lm(Distance ~ Speed+speed, data=stopping, weights=1/Speed)
summary(m4)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + speed, data = stopping, weights = 1/Speed)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0037 -1.4120 -0.1054  1.2586  5.0984
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.32590    3.09898   0.428    0.670
## Speed        0.44801    0.42065   1.065    0.291
## speed        0.06479    0.01122   5.777 3.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.011 on 59 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9204
## F-statistic: 353.8 on 2 and 59 DF,  p-value: < 2.2e-16
```

From the result, we see that the coeffient is larger than the previous one while the standard errors are smaller than the previous one.

4

```r
(a=hccm(m3,typle="hc3"))
```

```
##             (Intercept)       Speed    I(Speed^2)
## (Intercept) 18.45413036 -2.65217062  0.0690953828
## Speed       -2.65217062  0.39729977 -0.0106319865
## I(Speed^2)   0.06909538 -0.01063199  0.0002974929
```

The table shows the estimated covariance matrix.

```r
sqrt(diag(hccm(m3, type="hc3")))
```

```
## (Intercept)       Speed  I(Speed^2)
##  4.29582709  0.63031720  0.01724798
```

```r
summary(m3)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + I(Speed^2), data = stopping)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5192  -5.4527  -0.5519   3.8442  27.9373
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.58036    5.10266   0.310    0.758
## Speed        0.41607    0.55641   0.748    0.458
## I(Speed^2)   0.06556    0.01303   5.033 4.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.927 on 59 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9115
## F-statistic: 315.3 on 2 and 59 DF,  p-value: < 2.2e-16
```

```r
summary(m4)
```

```
##
## Call:
## lm(formula = Distance ~ Speed + speed, data = stopping, weights = 1/Speed)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0037 -1.4120 -0.1054  1.2586  5.0984
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.32590    3.09898   0.428    0.670
## Speed        0.44801    0.42065   1.065    0.291
## speed        0.06479    0.01122   5.777 3.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.011 on 59 degrees of freedom
## Multiple R-squared:  0.923,  Adjusted R-squared:  0.9204
## F-statistic: 353.8 on 2 and 59 DF,  p-value: < 2.2e-16
```

The summary from *m3* which contain the standard errors value of *5.10*,*0.556* and *0.013* relatively.

The summary from *m4* which contain the standard errors value of *3.09*,*0.42* and *0.01122*relatively.

With the result, we have the estimated standard error of *4.29*,*0.63* and *0.017*, thus the new values are in between of value of *m3* and *m4*