HW9

Dandong Tu

2017/11/1

1

a

```
library(alr4)
```

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

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

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

```
## The following object is masked from 'package:car':
##
##     Prestige
A=factor(BGSall$Sex,c("1","0"))
levels(A)=list("Male"="0","Female"="1")
BGSall$sex=A
scatterplot(HT18~HT9+sex,data = BGSall)
```
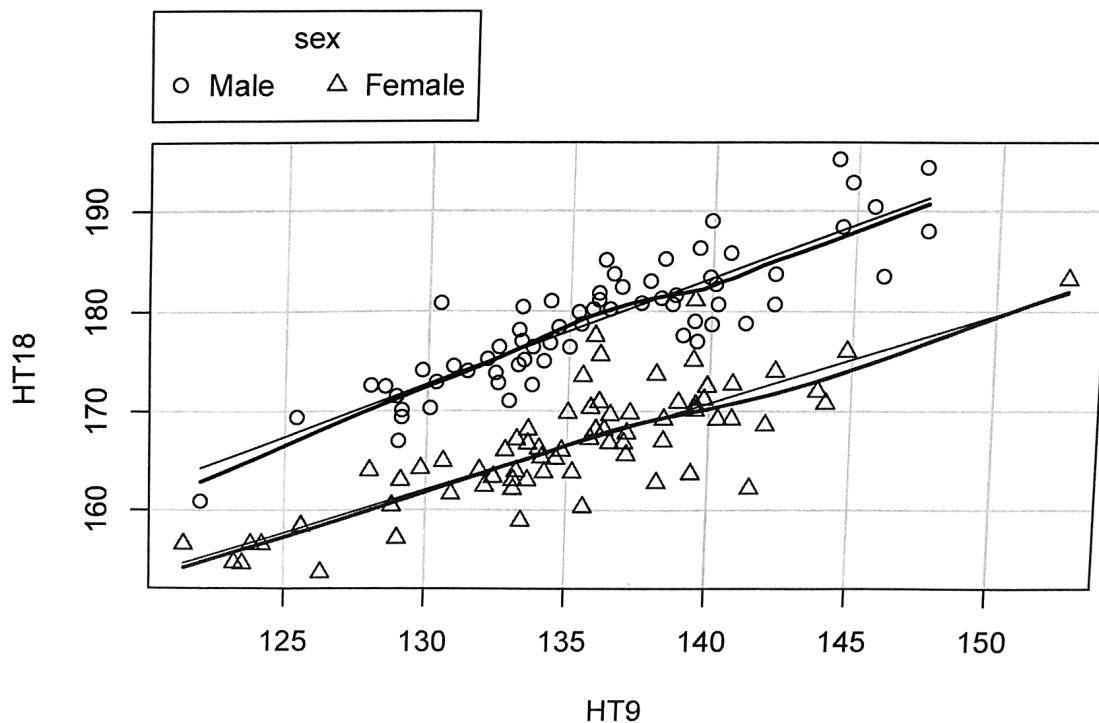
From the graph, two lines are the ols lines fit separately for each **Sex**. The straight line mean function appears appropriate for each group. The slop does not seems same and there is difference in intercepts for two different factor levels.

b

```
m1=lm(HT18~HT9,data = BGSall)
m2=lm(HT18~HT9+sex,data = BGSall)
m3=lm(HT18~HT9+sex+sex:HT9,data = BGSall)
m4=lm(HT18~HT9+sex:HT9,data = BGSall)
anova(m1,m2,m3,m4)

## Analysis of Variance Table
##
## Model 1: HT18 ~ HT9
## Model 2: HT18 ~ HT9 + sex
## Model 3: HT18 ~ HT9 + sex + sex:HT9
## Model 4: HT18 ~ HT9 + sex:HT9
##   Res.Df    RSS Df Sum of Sq        F  Pr(>F)
## 1    134 6190.9
## 2    133 1566.9  1    4624.0 398.2872 < 2e-16 ***
## 3    132 1532.5  1      34.4   2.9638 0.08749 .
## 4    133 1542.2 -1      -9.7   0.8394 0.36125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(m2)

##
## Call:
## lm(formula = HT18 ~ HT9 + sex, data = BGSall)
##
## Residuals:
##      Min      1Q   Median      3Q     Max
## -10.4694  -2.0952  -0.0136  1.7101  10.4467
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   48.51731    7.33385   6.616 8.27e-10 ***
## HT9            0.96006    0.05388  17.819  < 2e-16 ***
## sexFemale    -11.69584    0.59036 -19.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

In part(a) we observed that there is an obvious diffenece between the intercept of the two lines. We can use statistical hypothesis testing to determine the significance. In anova analysis, we are comparing models. From the results, **m1** looks like not needed. **m4** is rejected. And we only considering **m2** and **m3** where **m2** is the model of parallel regression.

c

```
summary(m2)$coef
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  48.5173098 7.33385131   6.61553 8.265885e-10
## HT9           0.9600564 0.05387964  17.81854 4.778019e-37
## sexFemale   -11.6958406 0.59035893 -19.81141 1.659509e-41
```

```
confint(m2)
```

```
##                    2.5 %      97.5 %
## (Intercept)   34.0112360   63.023384
## HT9            0.8534845    1.066628
## sexFemale    -12.8635477  -10.528134
```

From the summary and by using **confint**, we obtained a 95% confidence interval of [-12.85,10.53] for the difference between males and females.

2

Please see last page.

3

a

Since the NH and AH both contain interations, the meaning of NH is the difference in intercept

b

```
NH=lm(lifeExpF~log(ppgdp)+group:log(ppgdp),data = UN11)
AH=lm(lifeExpF~group+log(ppgdp)+group:log(ppgdp),data = UN11)
summary(NH)
```

```
##
## Call:
## lm(formula = lifeExpF ~ log(ppgdp) + group:log(ppgdp), data = UN11)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -18.6121  -2.5029   0.3037  2.4489  15.3486
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              43.8040     2.6231  16.699  < 2e-16 ***
## log(ppgdp)                3.7245     0.2677  13.912  < 2e-16 ***
## log(ppgdp):groupother    -0.0698     0.1153  -0.605    0.546
## log(ppgdp):groupafrica   -1.4303     0.1726  -8.285 1.87e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.18 on 195 degrees of freedom
## Multiple R-squared:  0.7422, Adjusted R-squared:  0.7382
## F-statistic: 187.1 on 3 and 195 DF,  p-value: < 2.2e-16
summary(AH)

##
## Call:
## lm(formula = lifeExpF ~ group + log(ppgdp) + group:log(ppgdp),
##     data = UN11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.634  -2.089   0.301   2.255  14.489
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             59.2137    15.2203   3.890 0.000138 ***
## groupother             -11.1731    15.5948  -0.716 0.474572
## groupafrica            -22.9848    15.7838  -1.456 0.146954
## log(ppgdp)               2.2425     1.4664   1.529 0.127844
## groupother:log(ppgdp)    0.9294     1.5177   0.612 0.540986
## groupafrica:log(ppgdp)   1.0950     1.5785   0.694 0.488703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.129 on 193 degrees of freedom
## Multiple R-squared:  0.7498, Adjusted R-squared:  0.7433
## F-statistic: 115.7 on 5 and 193 DF,  p-value: < 2.2e-16
```

Comparing summaries,in model **AH** we have large **p-value** for every regression.And in model**NH**,all **p-values** are small except the **p-value** for log(ppgdp):groupother. So that we are unable to reject H0. The **0.7498** of R^2 in dicates that it is a good explain for this model. While in model **NH**, we found that the **ppgdp** is significant and the **group** is also significant by the **p-values**. And the value of R^2, which is **0.7422** also indicates that it is a good fit with the overall model being significant.

c

$H_0 : \beta_{02} - \beta_{03} = 14 and \beta_{12} + \beta_{13} = .2$

H_A: at least one equality doesn't hold.

Full model : $E(lifeExpF|Group) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x + \beta_{12}U_2 x + \beta_{13}U_3 x$

```
L = matrix(c(0, 1, -1, 0, 0, 0,
             0, 0, 0, 0, 1, 1 ),
           byrow=TRUE, nrow=2)
c.vec=c(14,.2)
m33=lm(lifeExpF~group+log(ppgdp)+group:log(ppgdp),data = UN11)
m33

##
## Call:
## lm(formula = lifeExpF ~ group + log(ppgdp) + group:log(ppgdp),
```

```
##       data = UN11)
##
## Coefficients:
##          (Intercept)                 groupother                groupafrica
##              59.2137                   -11.1731                   -22.9848
##          log(ppgdp)    groupother:log(ppgdp)   groupafrica:log(ppgdp)
##               2.2425                     0.9294                     1.0950
m333 = linearHypothesis(m33, hypothesis.matrix = L, rhs = c.vec)
m333

## Linear hypothesis test
##
## Hypothesis:
## groupother - groupafrica = 14
## groupother:log(ppgdp)  + groupafrica:log(ppgdp) = 0.2
##
## Model 1: restricted model
## Model 2: lifeExpF ~ group + log(ppgdp) + group:log(ppgdp)
##
##    Res.Df     RSS Df Sum of Sq       F Pr(>F)
## 1     195 5092.6
## 2     193 5077.7  2    14.906 0.2833 0.7536
```

From the **linearHypothesis** result, we have the **p-value=0.754**, so that we do not have enough evidence to reject the H_0. Thus, we do not know whether the equalities are hold or not.

a.

$HX = X$

$HX = H[X_1 | X_2] = [HX_1 | HX_2] = [X_1 | X_2]$

so $HX_1 = X_1$

$HH_R = HX_1(X_1^T X_1)^{-1} X_1^T = X_1(X_1^T X_1)^{-1} X_1^T = H_R$

$H_R X_1 = X_1(X_1^T X_1)^{-1} X_1^T X_1 = X_1(X_1^T X_1)^{-1}(X_1^T X_1) = X_1$

So that, $HX_1 = H_R X_1 = X_1$

b. $(H - H_R)(H - H_R) = HH - HH_R - H_R H + H_R H_R$

$$= H - HR - HR + H_R$$

$$= H - H_R$$

$(H - H_R)^T = H^T - H_R^T = H - H_R$

So, $H - H_R$ is symmetric and idempotent

1. $RSS_R = Y^T(I - H_R)Y \qquad RSS = Y^T(I - H)Y$

$RSS_R - RSS_F = SS_{reg} = Y^T(H - H_R)Y$

Full : $Y = X_1\beta_1 + X_2\beta_2$

Null : $Y = X_1\beta_1$ Since $\beta_2 = 0$

$E(Y) = X\beta \qquad E(Y - X\beta) = 0 \quad \Big| \quad \dfrac{SS_{reg}}{\sigma^2} = (Y - X\beta)^T(H - H_R)(Y - X\beta)$ where $X\beta = X_1\beta_1$

$(H - H_R)X_1\beta_1$

$= HX_1\beta_1 - HRX_1\beta_1 = X_1\beta_1 - X_1\beta_1$

$= 0$ from partial

Therefore $Y \sim N(0, V)$, by applying Theorems 2

$$\frac{1}{\sigma^2} SS_{reg} \sim \chi_q^2$$

d. $\quad \sigma^2 = \dfrac{RSS}{n-p-1} = Y^T \left(\dfrac{I-H}{n-p-1} \right) Y$

$\left(\dfrac{H-H_R}{\sigma^2} \right) \left(\sigma^2(I-H) \right) \left[\dfrac{I-H}{n-p-1} \right] = (H-H_R)(I-H) \dfrac{1}{n-p-1}$

$\qquad\qquad\qquad\qquad\qquad = (H-H-H_R+H_R) \dfrac{1}{n-p-1}$

$\qquad\qquad\qquad\qquad\qquad = 0$

Therefore, SS_{reg} and $\hat{\sigma}^2$ are independent

e. $\quad Cov\ [(I-H)Y, (H-H_R)Y] = \sigma^2 (I-H)(H-H_R) = 0$, thus $(I-H)Y$ and $(H-H_R)Y$ are independent

$df_R - df_F = rank\ (H-H_R) = trace\ (H-H_R) = p' - (p'-q) = q$

$\qquad\qquad\qquad$ then RSS and SS_{reg} are independent

$$F = \dfrac{\dfrac{SS_{reg}}{q}}{\dfrac{RSS}{n-p'}} = \dfrac{\dfrac{Y^T(H-H_R)Y}{q}}{\dfrac{Y^T(I-H)Y}{n-p'}} \underset{under\ H_0}{\sim} \dfrac{\dfrac{\chi^2_q}{q}}{\dfrac{\chi^2_{n-p'}}{n-p'}} \sim F_{q,n-p'}$$