Hw4

Dandong Tu

2017/9/20

2.16.1

```r
library(alr4)
```

```
## Loading required package: car

## Loading required package: effects

##
## Attaching package: 'effects'

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

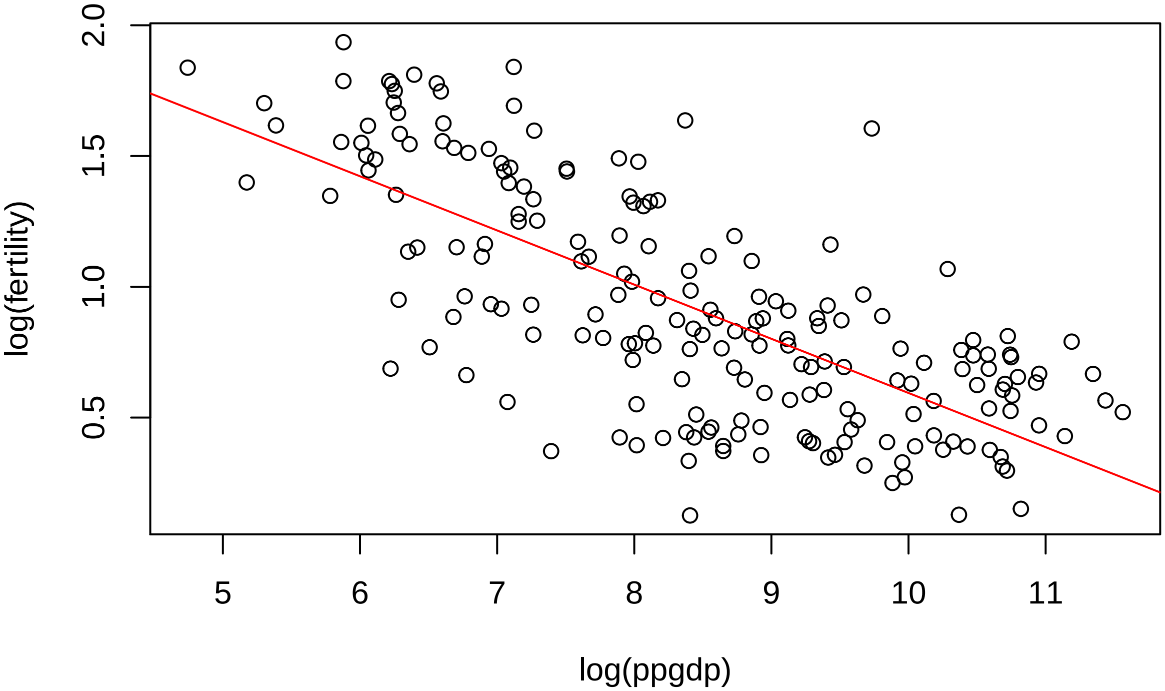
```r
summary(lm(log(UN11$fertility)~log(UN11$ppgdp)))
```

```
##
## Call:
## lm(formula = log(UN11$fertility) ~ log(UN11$ppgdp))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79828 -0.21639  0.02669  0.23424  0.95596
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.66551    0.12057   22.11   <2e-16 ***
## log(UN11$ppgdp) -0.20715    0.01401  -14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 197 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.5236
## F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

According to the summary, the simple linear regression is y=-0.20715x+2.66551

2.16.2

```r
plot(log(fertility)~log(ppgdp), data=UN11)
abline(lm(log(fertility)~log(ppgdp),data=UN11),col="red")
```



2.16.3

Ho: the slope of the regression line is equal to zero H1: the slope of the regression line is negative

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is negative. We give the significance level of 0.01 Based on the result of the data summary from 2.16.1, we observed that the t=b1/se=-14.79 and Pr(>|t|)<2e-16. Since the P-value which is <2.2e-16 is less than the significance level (0.01), we cannot accept the null hypothesis but we do not have much information to reject the NH.

2.16.4

The coefficient of determination R-squared is 0.5236. It means that 52.36% of the response variable variation that is explained by the linear model.

2.16.5

y1=-0.20715x1+2.66551 when ppgdp=1000 the x1=log(1000)=3

```
y1=-0.20715*3+2.66551
t=abs(qt(0.025,198)) #95%predictive interval
SXX=sum((log(UN11$ppgdp)-mean(log(UN11$ppgdp)))^2)
sepred=0.12057*(1+1/200+(sum(3-mean(log(UN11$ppgdp)))^2/(SXX))^(1/2))
#since sepred=sigma(1+1/n+sum(x-x-bar)^2/(SXX))^(1/2)
aa=y1-t*sepred
bb=y1+t*sepred
aa
```

```
## [1] 1.745839
```

```
bb
```

```
## [1] 2.342281
```

```
exp(aa)
```

```
## [1] 5.730705
```

```
exp(bb)
```

```
## [1] 10.40495
```

```
#Thus, a 95% predictive interval is given by(5.730705,10.40495) for fertility
```

2.16.6

1

```
UN11[which.max(UN11$fertility),]
```

```
##       region  group fertility ppgdp lifeExpF pctUrban
## Niger Africa africa     6.925 357.7    55.77       17
```

The locality with the highest value of fertility is Niger in Africa

2

```
UN11[which.min(UN11$fertility),]
```

```
##                        region group fertility  ppgdp lifeExpF pctUrban
## Bosnia and Herzegovina Europe other     1.134 4477.7     78.4       49
```

The locality with the lowerest value of fertility is Bosnia and Herzegovina in Europe

3

```
m1=lm(log(UN11$ppgdp)~log(UN11$fertility))
residual=resid(m1)
head(sort(residual))
```

```
##       134       118       123        14       196       126
## -2.812794 -2.442884 -2.323145 -2.320328 -2.283591 -2.084928
```

```
head(sort(residual,decreasing=TRUE))
```

```
##       58      148      135       88      105       20
## 3.028584 2.416462 2.258575 2.216974 2.104354 2.091633
```

```
UN11[134,]
```

```
##             region group fertility ppgdp lifeExpF pctUrban
## North Korea   Asia other     1.988   504    72.12       60
```

```
UN11[118,]
```

```
##         region group fertility  ppgdp lifeExpF pctUrban
## Moldova Europe other      1.45 1625.8    73.48       48
```

```
UN11[58,]
```

```
##                   region  group fertility   ppgdp lifeExpF pctUrban
## Equatorial Guinea Africa africa      4.98 16852.4    52.91       40
```

```
UN11[148,]
```

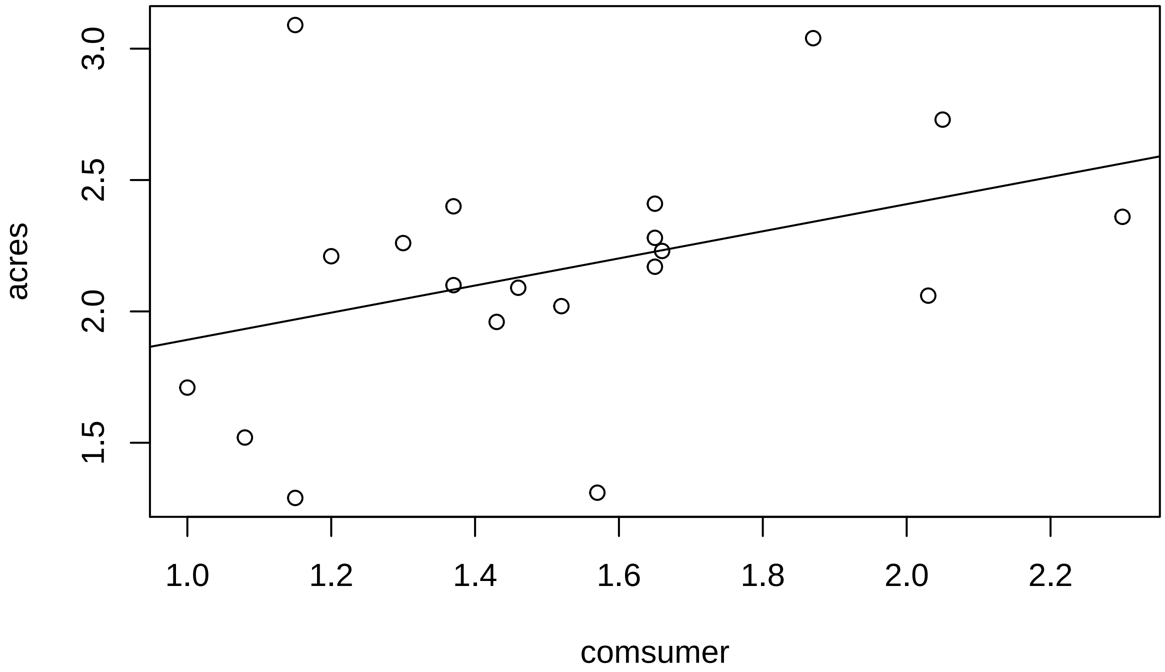
```
##       region group fertility   ppgdp lifeExpF pctUrban
## Qatar   Asia other     2.204 72397.9    78.24       96
```

Therefore, two localities with the largest negative residuals is north Korea and Moldova with residual value -2.812794 and -2.442884 relatively. And two localities with the largest positive residuals are Equatorial Guinea and Qatar with residual value of 3.028584 and 2.416462.

Problem 2

a

```
data1=read.table("/Users/dandongtu/Downloads/Sahlins.txt",header = TRUE)
a1=data1$consumer
a2=data1$acres
plot(x=a1,y=a2,xlab="comsumer",ylab = "acres")
abline(lm(a2~a1)) #to see if it seems a linear relation
```



From the figure, we observed a weak positive linear relation between acres/gardener ~ consumer/gardener. Meanwhile, there are several points look like unsual such as 4th and 17th observation with large positive residual and 3rd and 12th observation with large netative residual.

b

```
summary(lm(a2~a1))
```

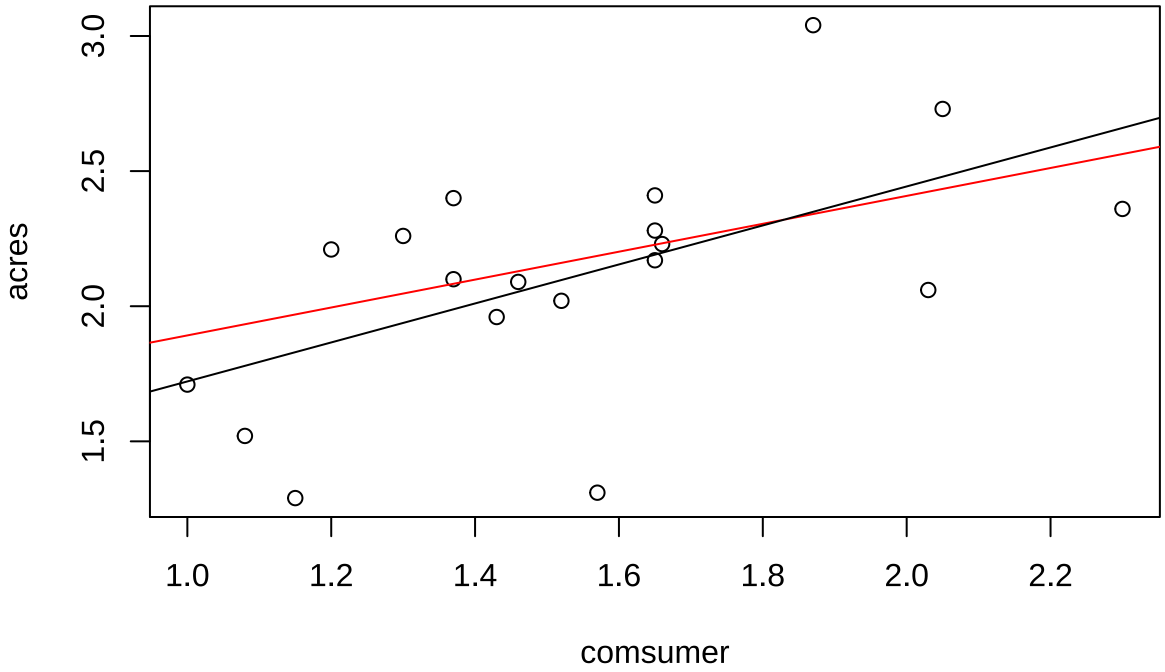
```
##
## Call:
## lm(formula = a2 ~ a1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8763 -0.1873 -0.0211  0.2135  1.1206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3756     0.4684   2.937  0.00881 **
## a1            0.5163     0.3002   1.720  0.10263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 18 degrees of freedom
## Multiple R-squared:  0.1411, Adjusted R-squared:  0.0934
## F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

```
sigma1=0.4543^2
sigma1
```

```
## [1] 0.2063885
```

From the summary result, we observed betahat01=1.3756 betahat11=0.5163 and variance which is sigma1=0.2063885.

```
data2=data1[-4,]
plot(x=data2$consumers,y=data2$acres,xlab="comsumer",ylab = "acres")
abline(lm(a2~a1),col="red")
abline(lm(data2$acres~data2$consumers))
```



```
summary(lm(data2$acres~data2$consumers))
```

```
##
## Call:
## lm(formula = data2$acres ~ data2$consumers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82291 -0.16808  0.03215  0.23505  0.69061
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.0000     0.3969   2.519   0.0221 *
## data2$consumers   0.7216     0.2514   2.870   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 17 degrees of freedom
## Multiple R-squared:  0.3264, Adjusted R-squared:  0.2868
## F-statistic: 8.238 on 1 and 17 DF,  p-value: 0.01061
```

In figure(red line/with 4th observation; black line/without), we oberserd a stronger linear relation by comparing regression line with and without 4th observation. From the summary(without fourth household),we observed betahat02=1 , betaha12=0.7216 which indicates a stronger liner relation between acres/gardener ~ consumer/gardener.It this case, linear regression doing a better job in summarizing the data.

c

with fourth household

From the summary in part.a, we observed that the standard errors of intercept is 0.4684 and standard errors for slope is 0.3002.

```r
t1=abs(qt(0.025,20))
l1=1.3756-t1*(0.4684)
h1=1.3756+t1*(0.4684)
l1
```

```
## [1] 0.3985347
```

```r
h1
```

```
## [1] 2.352665
```

The interval for betahat01 is 0.3985347<betahat0<2.352665

```r
l2=0.5163-t1*(0.3002)
h2=0.5163+t1*(0.3002)
l2
```

```
## [1] -0.1099062
```

```r
h2
```

```
## [1] 1.142506
```

The interval for betahat1is -0.1099062<betahat1<1.142506 H0:betahat1>0 H1:betahat1<=0 From the resulet, under95% confidence interval, with p value of 0.1026, provding some evidence against NH, so that we may not able to say the population slope is (always)greater than zero.

H0:betahat0>0 H1:betahat0<=0 For betahat0, under 95% two-sided confidence interval, with p value of 0.00881, suggesting no evidence against the NH.

without fourth household.

The summary shows that the betahat0=1 betahat1=0.7216.

```r
t2=abs(qt(0.025,19))
l3=1-t2*(0.3969)
h3=1+t2*(0.3969)
l3
```

```
## [1] 0.1692788
```

```r
h3
```

```
## [1] 1.830721
```

The interval for betahat01 is 0.1692788<betahat0<1.830721

```r
l4=0.7216-t2*(0.2514)
h4=0.7216+t2*(0.2514)
l4
```

```
## [1] 0.1954138
```

```r
h4
```

```
## [1] 1.247786
```

The interval for betahat1is 0.1954138<betahat1<1.247786

H0:betahat1>0 H1:betahat1<=0 From the resulet, under95% confidence interval, with p value of 0.0106, suggesting that no evidence against the NH

H0:betahat0>0 H1:betahat0<=0 For betahat0, under 95% two-sided confidence interval, with p value of 0.0221 which is less than the 0.05,so that it is suggesting no evidence against the NH.

d

```
y_star=coef(lm(data1))[1]+coef(lm(data1))[2]*1.5
y_star
```

```
## (Intercept)
##    1.342066
```

```
lm1=lm(acres~consumers,data1)
new_data=data.frame(consumers=1.5)
predict(lm1,new_data,interval="prediction")
```

```
##        fit     lwr     upr
## 1 2.150125 1.17196 3.12829
```

```
predict(lm1,new_data,interval="confidence")
```

```
##        fit      lwr      upr
## 1 2.150125 1.936202 2.364047
```

From the result, we obsered that the acres/gardener ratio is 2.15 with lwr of 1.17196 to upr of 3.12829. Fro second part with 2.15 and 1.936202 to 2.364047 relatively. The answer will change if instead we asking to determine the mean since the equation for caluating is changed.