PSTAT126_Homework5

Yanjie Qi

2019/9/7

1.

```
library(faraway)
data(prostate)
attach(prostate)
```

The log-transformed fiited model given is lpsa = 1.507 + 0.719lcavol; since lpsa and lcavol are all log-transformation, we could know that the estimated coefficient for lcavol means that if lcavol is changed by 1%, then the expected value of lpsa is changed by 100((1+0.01)^0.719-1)%, which is 100(1.01^0.719-1)%.

2.

a).

the meaning of the intercept is that the expected salary of a person, if not considering the sex,will be 24,697; And if the Sex coefficient is 1 that means a female person would earn 21,357 which got from (24,697 - 3340) and if male, then the income would be 24,697.

b).

With the increase of years the college has established, the salary for many original employees are not only considered by their sexuality but also their employment time in the company to show their contribution; therefore, the whole salary pattern of the company changes with both sex and time being considered for employee's salary.

3.

install the package:

```
library("alr4")

## Loading required package: car

## Loading required package: carData
```

```
##
## Attaching package: 'car'

## The following objects are masked from 'package:faraway':
##
##     logit, vif

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

##
## Attaching package: 'alr4'

## The following objects are masked from 'package:faraway':
##
##     cathedral, pipeline, twins

data(cakes)
attach(cakes)
```

set up the terms for the quadratic and the linear model:

```
X1_sq = (X1*X1)
X2_sq = (X2*X2)
X1_X2 = (X1*X2)
mod_3 = lm(Y~X1+X2+X1_sq+X2_sq+X1_X2)
```

a).

Get the summary and anova table:

```
summary(mod_3)

##
## Call:
## lm(formula = Y ~ X1 + X2 + X1_sq + X2_sq + X1_X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## X1_sq       -1.569e-01  3.945e-02  -3.977 0.004079 **
## X2_sq       -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## X1_X2       -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

```r
anova(mod_3)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1  4.3232  4.3232  23.515 0.0012730 **
## X2         1  7.4332  7.4332  40.432 0.0002186 ***
## X1_sq      1  2.1308  2.1308  11.591 0.0092987 **
## X2_sq      1 10.5454 10.5454  57.361 6.462e-05 ***
## X1_X2      1  2.7722  2.7722  15.079 0.0046537 **
## Residuals  8  1.4707  0.1838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the table, the significance levels for the quadratic terms and the interaction are all less than 0.005.

b).

Now get the summary adding block effect:

```r
mod_4 = lm(Y~X1+X2+X1_sq+X2_sq+X1_X2+block)
```

Get the summary and anova table for the model with block effect:

```
summary(mod_4)

##
## Call:
## lm(formula = Y ~ X1 + X2 + X1_sq + X2_sq + X1_X2 + block)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## X1           2.592e+01  4.903e+00   5.287 0.001140 **
## X2           9.918e+00  1.228e+00   8.080 8.56e-05 ***
## X1_sq       -1.569e-01  4.151e-02  -3.779 0.006898 **
## X2_sq       -1.195e-02  1.660e-03  -7.197 0.000178 ***
## X1_X2       -4.163e-02  1.128e-02  -3.690 0.007754 **
## block1       1.143e-01  2.412e-01   0.474 0.650014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129

anova(mod_4)

## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1  4.3232  4.3232 21.2361 0.0024600 **
## X2         1  7.4332  7.4332 36.5132 0.0005198 ***
## X1_sq      1  2.1308  2.1308 10.4670 0.0143465 *
## X2_sq      1 10.5454 10.5454 51.8009 0.0001779 ***
## X1_X2      1  2.7722  2.7722 13.6177 0.0077544 **
```

```
## block       1  0.0457  0.0457  0.2246 0.6500138
## Residuals   7  1.4250  0.2036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the tables, we know the blcok variables is not significant.

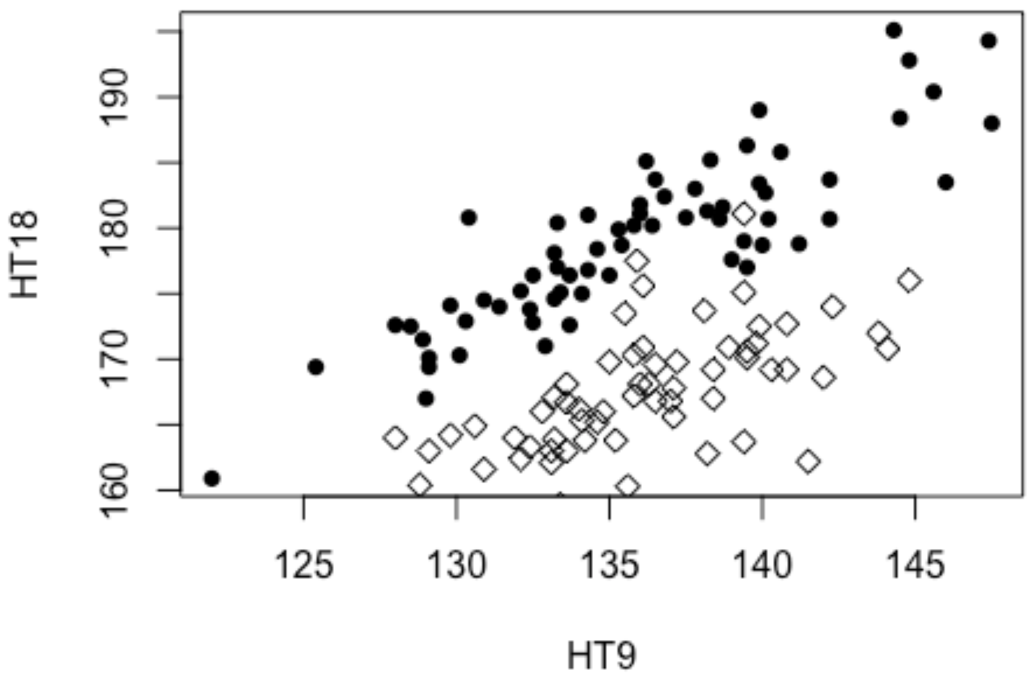
4.

```
data("BGSall")
attach(BGSall)
```

a).

Draw the scatterplot of HT18 versus HT9:

```
with(subset(BGSall, Sex == "0"), plot(HT9,HT18,pch = 16))
with(subset(BGSall, Sex == "1"), points(HT9,HT18,pch = 5))
```



From the plot, it seems that the increase in female height from age 9 to age 18 is less as compared with boys. For instance, a female with age 9 and height around 135 has height of 165 at age 18, but a male with age 9 and height around 135 has height in the range 175-180 at age 18.

b).

To obtain the appropriate test, we have to use anova function to get the p-value of reduced model, which is without the interaction:

```
m1 <- lm(HT18 ~ HT9 + factor(Sex), data = BGSall)
m2 <- lm(HT18 ~ HT9*factor(Sex), data = BGSall)
anova(m1,m2)

## Analysis of Variance Table
##
```

```
## Model 1: HT18 ~ HT9 + factor(Sex)
## Model 2: HT18 ~ HT9 * factor(Sex)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    133 1566.9
## 2    132 1532.5  1    34.409 2.9638 0.08749 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

since the p-value is 0.08749, larger than the alpha 0.05, we could accept the null hypothesis that the coefficient for interaction is 0 at 5% confidence level. Therefore, there is sufficient evidence to conclude that the parallel regression model is better

c).

set up the model with HT9 as one quantitative predictor and factor(Sex) as one binary predictor:

```
summary(m1)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + factor(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 ***
## factor(Sex)1 -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
```

From table, we can see that The difference in the intercepts of the two groups (Sex = 0 and Sex = 1) is 11.69584. That is the estimated difference between males and females is 11.69584 with standard error of 0.59036.Degree of freedom of residual is 133.

t-value for the test:

```
abs(qt(0.05/2, 133))

## [1] 1.977961
```

Therefore, 95% confidence interval between males and females: (11.69584 - 1.98 * 0.59036, 11.69584 + 1.98 * 0.59036) which is (10.52693, 12.86475)

5.

```
library("faraway")
data("infmort")
names(infmort)

## [1] "region"    "income"    "mortality" "oil"

head(infmort)

##                            region income mortality            oil
## Australia                    Asia   3426      26.7 no oil exports
## Austria                    Europe   3350      23.7 no oil exports
## Belgium                    Europe   3346      17.0 no oil exports
## Canada                   Americas   4751      16.8 no oil exports
## Denmark                    Europe   5029      13.5 no oil exports
## Finland                    Europe   3312      10.1 no oil exports
```

a).

Hypothesis Test: H0:beta1 = beta2 = beta12 = 0 versus H1: at lease one of betai not zero.

set up the model:

```
fit <- lm(log(mortality)~region+log(income)+log(income)*region, data =
infmort)
summary(fit)

##
## Call:
## lm(formula = log(mortality) ~ region + log(income) + log(income) *
```

```
##     region, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46809 -0.26530 -0.02148  0.27478  3.14219
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.9385     0.6362   7.763 1.06e-11 ***
## regionEurope                   2.0882     1.8422   1.134   0.2599
## regionAsia                     1.2634     0.8561   1.476   0.1434
## regionAmericas                 1.5661     1.1856   1.321   0.1898
## log(income)                   -0.0112     0.1235  -0.091   0.9280
## regionEurope:log(income)      -0.5205     0.2516  -2.069   0.0413 *
## regionAsia:log(income)        -0.3798     0.1580  -2.404   0.0182 *
## regionAmericas:log(income)    -0.3978     0.1979  -2.010   0.0473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5971 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6464, Adjusted R-squared:  0.6198
## F-statistic: 24.29 on 7 and 93 DF,  p-value: < 2.2e-16
```

From Summary, we could know that F is 24.29 on 7 and 93 DF and p-value is 2.2*10^(-16). Therefore, we reject the null hypothesis since p-value is smaller the alpha=0.05, and we can conclude that the fitted model is significant.

b).

In the model above, beat2 = beta12 = 0 implies that region has no impact on the relationship between income and mortality, which means log(mortality) is independent of region and interaction between region and log(income).

c).

From b), the new model could be: E(log(mortality)|income, region) = $\beta 0 + \beta 1$ log(income)

```r
n.mod <- lm(log(mortality)~log(income), data = infmort)
o.mod <- lm(log(mortality)~log(income)+ region +log(income)*region,
data = infmort)
anova(n.mod, o.mod)
```

```
## Analysis of Variance Table
##
## Model 1: log(mortality) ~ log(income)
## Model 2: log(mortality) ~ log(income) + region + log(income) *
region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     99 46.685
## 2     93 33.152  6    13.533 6.3274 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From Anova, we know the p-value is 1.31e-05, which is much smaller than alpha = 0.05. Therefore, we reject the null hypothesis that beta12 and beta2 are 0. There is sufficient evidence to conclude that region and interaction between region and log(income) are significant variable in determining the log(mortality) for given income and region.