

## Home Work 6

Yi Chen

3/6/2020

```
data <- read.csv(file = 'mvregex.dat', header = F)
colnames(data) = c("locus", "self", "motiv", "read", "write", "science",
"prog", "prog1", "prog2", "prog3")
data$prog = as.factor(data$prog)
```

### a) Regression of locus on read and write.

#### a) hard-code solution

```
n <- nrow(data)
Z <- cbind(rep(1,n), data$read, data$write)
colnames(Z) <- c("intercept", "read", "write")
y <- as.matrix(data$locus)
colnames(y) <- c("beta.hat")
(b.hat <- solve(t(Z) %*% Z) %*% t(Z) %*% y)

##          beta.hat
## intercept -1.48627258
## read      0.01623081
## write     0.01413381
```

#### b) package solution

```
reg = lm(locus ~ read + write, data = data)
reg$coefficients

## (Intercept)      read      write
## -1.48627258  0.01623081  0.01413381
```

### b) Regression of locus on read, write, and science.

#### hard-code solution

```
Z <- cbind(rep(1,n), data$read, data$write, data$science)
r <- ncol(Z) - 1
colnames(Z) <- c("intercept", "read", "write", "science")
y <- as.matrix(data$locus)
colnames(y) <- c("beta.hat")
b.hat <- solve(t(Z) %*% Z) %*% t(Z) %*% y
t(y - Z %*% b.hat) %*% (y - Z %*% b.hat)/(n-r-1)

##          beta.hat
## beta.hat 0.3756474
```

### package solution

```
reg = lm(locus ~ read + write + science, data = data)
(t(as.matrix(reg$residuals)) %*% as.matrix(reg$residuals)) / (n - r - 1)

##           [,1]
## [1,] 0.3756474
```

### c) F test for overall regression of self on read, write, and science.

#### hard code way

```
n <- nrow(data)
reg <- lm(self~read+ write+ science, data = data)
SSE <- sum((reg$residuals)^2 )
SST <- sum((data$self -mean(data$self))^2)
(F_test <- ((SST-SSE)/3) / (SSE/(n-3-1)))

## [1] 1.268962

pf(q = F_test,df1 = 3,df2 = n-3-1,lower.tail = F)

## [1] 0.2841237
```

#### Using existing package for check

```
reg <- lm(self~read+ write+ science, data = data)
summary(reg)

##
## Call:
## lm(formula = self ~ read + write + science, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54070 -0.46122  0.01644  0.45703  2.04669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.227841    0.178804  -1.274    0.203
## read         0.003088    0.004294   0.719    0.472
## write       -0.003294    0.003923  -0.840    0.401
## science      0.004735    0.004228   1.120    0.263
##
## Residual standard error: 0.705 on 596 degrees of freedom
## Multiple R-squared:  0.006347,    Adjusted R-squared:  0.001345
## F-statistic: 1.269 on 3 and 596 DF,  p-value: 0.2841
```

### d) Regression of locus and self on read, write, and science.

#### hard code way

```
n <- nrow(data)
r <- 3 ## two independent variable
m <- 2 ## three dependent variables
```

```
x <- as.matrix(cbind(rep(1,n),data$read,data$write,data$science))
colnames(x) <- c("intercept","read","write","science")
```

```
y <- as.matrix(data[,c("locus","self")])
(b.hat <- solve(t(x) %*% x) %*% t(x) %*% y)
```

```
##           locus           self
## intercept -1.555277255 -0.227840654
## read      0.013346683  0.003087669
## write     0.012919172 -0.003294353
## science   0.005454143  0.004734544
```

*Using existing package for check*

```
reg = lm (cbind(locus, self) ~ read + write + science, data = data)
reg$coefficients
```

```
##           locus           self
## (Intercept) -1.555277255 -0.227840654
## read      0.013346683  0.003087669
## write     0.012919172 -0.003294353
## science   0.005454143  0.004734544
```

## e) Regression of locus, self, and motiv on read, write, and science.

*hard code way*

```
n <- nrow(data)
p <- 3 ## three independent variable
q <- 3 ## three dependent variables
x <- as.matrix(cbind(rep(1,n),data$read,data$write,data$science))
colnames(x) <- c("intercept","slope_read","slope_write","slope_sciences")
```

```
y <- as.matrix(data[,c("locus","self","motiv")])
b.hat <- solve(t(x) %*% x) %*% t(x) %*% y
(SS.res <- t(y - x %*% b.hat) %*% (y - x %*% b.hat))
```

```
##           locus           self           motiv
## locus 223.88586  42.43956  48.19605
## self  42.43956 296.25868  98.45040
## motiv 48.19605  98.45040 374.54226
```

*Using existing package for check*

```
reg = lm (cbind(locus, self,motiv) ~ read + write + science, data = data)
t(residuals(reg)) %*% residuals(reg)
```

```
##           locus           self           motiv
## locus 223.88586  42.43956  48.19605
## self  42.43956 296.25868  98.45040
## motiv 48.19605  98.45040 374.54226
```

#### f) Regression of locus, self, and motiv on prog.

```
r <- 1
m <- 3
q <- 0
## using likelihood ratio test
### proposed model
x <- as.matrix(cbind(rep(1,n),data$prog))
colnames(x) <- c("intercept","slope_prog")
y <- as.matrix(data[,c("locus","self","motiv")])
b.hat <- solve(t(x) %*% x) %*% t(x) %*% y
E <- t(y - x %*% b.hat) %*% (y - x %*% b.hat)
### compared model
x1 <- rep(1,n)
b.hat1 <- solve(t(x1) %*% x1) %*% t(x1) %*% y
H <- t(y - x1 %*% b.hat1) %*% (y - x1 %*% b.hat1)

# Chi-sq approx
-(n-r-1-(m-r+q+1)/2)*log(det(E)/det(H))

## [1] 77.43948

# p-value
1-pchisq(-(n-r-1-(m-r+q+1)/2)*log(det(E)/det(H)), m*(r-q))

## [1] 1.110223e-16
```

The overall performance of the proposed model is significantly different from the reduced model by rejecting the null hypothesis. This means that the overall model fit good.

#### g)

```
n <- nrow(data)
p <- 2
q <- 3
m <- 0
## using likelihood ratio test
### proposed model
x <- as.matrix(cbind(rep(1,n),data$prog1,data$prog2))
colnames(x) <- c("intercept","slope_prog1","slope_prog2")
y <- as.matrix(data[,c("locus","self","motiv")])
b.hat <- solve(t(x) %*% x) %*% t(x) %*% y
E <- t(y - x %*% b.hat) %*% (y - x %*% b.hat)
### compared model
x1 <- rep(1,n)
b.hat1 <- solve(t(x1) %*% x1) %*% t(x1) %*% y
H <- t(y - x1 %*% b.hat1) %*% (y - x1 %*% b.hat1)

# Chi-sq approx
- (n - p - 1 - (q - p + m + 1)/2)*log(det(E)/det(H))
```

```
## [1] 79.17591

# p-value
1-pchisq(-(n - p - 1 - (q - p + m + 1)/2)*log(det(E)/det(H)), q*(p-m))

## [1] 5.329071e-15
```

When we add the dependent variable after dummy coding, the p-value of likelihood ratio test become bigger. This indicates that it may be risky to take the ordinal variable as continuous in hypothesis testing. However, in this case, we make the same conclusion that the overall model fit is good since it rejects the null hypothesis that there is no significant difference between proposed model and reduced model.

Further, with manova test we can see each prog variables are significant itself.

```
reg <- lm(cbind(cbind(locus,self,motiv)) ~ prog1 + prog2,data )
summary(manova(reg), test="Wilks")

##              Df    Wilks approx F num Df den Df    Pr(>F)
## prog1         1 0.90209   21.5275      3    595 3.006e-13 ***
## prog2         1 0.96767    6.6271      3    595 0.0002081 ***
## Residuals 597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```