ADA HW7

Liangquan Zhou lz2377

Thursday, October 23, 2014

1. Consider the Duodenal Ulcer data given in Problem 25, Chapter 5.

First input the data.

```
controls = c(.11, .11, .11, .19, .21, .22, .24, .25, .31)
gallstone = c(.18, .27, .36, .37, .39, .47, .37, .57)
ulcer = c(.29, .30, .40, .45, .47, .52, .57, 1.10)
Y = as.numeric(c(controls, gallstone, ulcer))
name = c(rep("controls", length(controls)), rep("gallstone", length(gallstone)),
  rep("ulcer", length(ulcer)))
dat = data.frame(cbind2(Y, name),stringsAsFactors=F)
names(dat) = c("Y", "name")
dat$Y = as.numeric(dat$Y)
dat$name = as.factor(dat$name)
```

a). Using an appropriate ANOVA model, determine whether there is a significant difference among the group means. Use both an F test and simultaneous confidence interval procedures.

Use F test:

```
anov1 = aov(Y ~ name, data = dat)
summary(anov1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## name         2  0.433  0.2164    7.95 0.0025 **
## Residuals   22  0.599  0.0272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result we can see that $p < 0.05$, then we should reject $H_0 : \mu_{Controls} = \mu_{Gallstone} = \mu_{Ulcer}$.

Use simultaneous confidence interval procedures:

Here we use `Bonferroni` method:

```
pairwise.t.test(dat$Y, dat$name, p.adjust.method = "bonf", alternative = c("two.sided"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dat$Y and dat$name
##
```

```
##           controls gallstone
## gallstone 0.111    -
## ulcer     0.002    0.311
##
## P value adjustment method: bonferroni
```

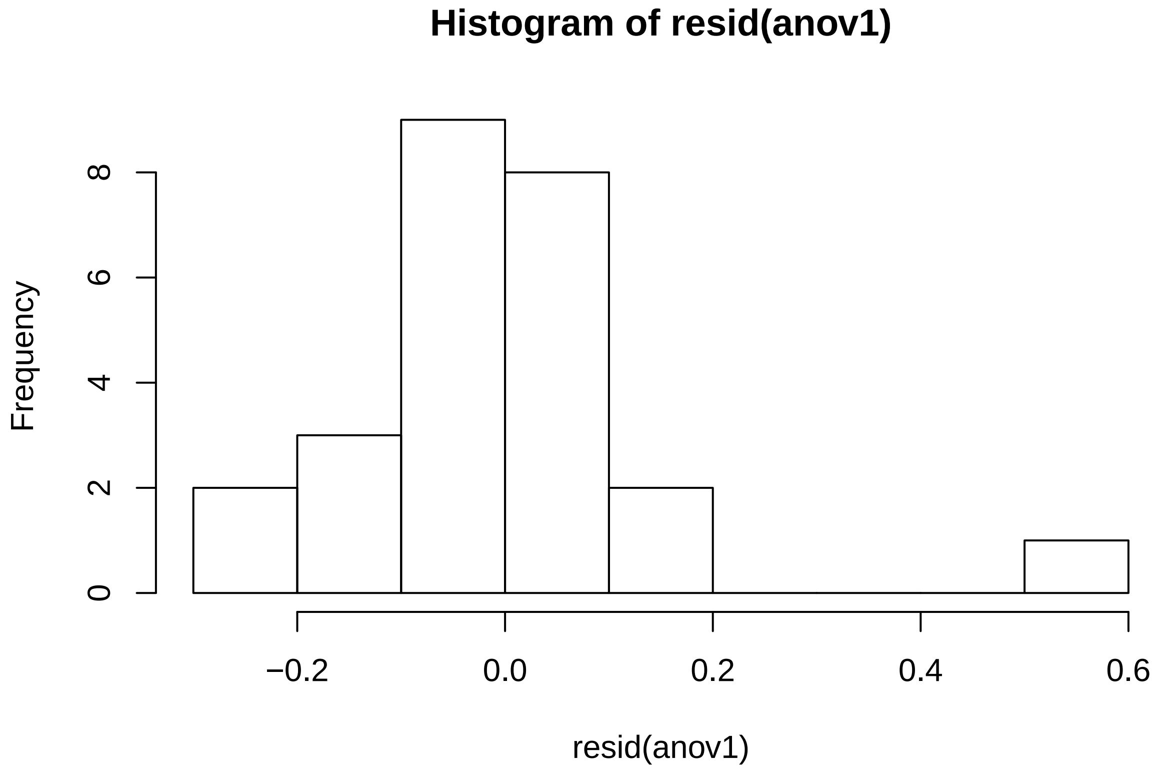
Then we can see that the group `Controls` has a different mean with the group `Ulcer`.

b). Assess the assumptions of the ANOVA model.

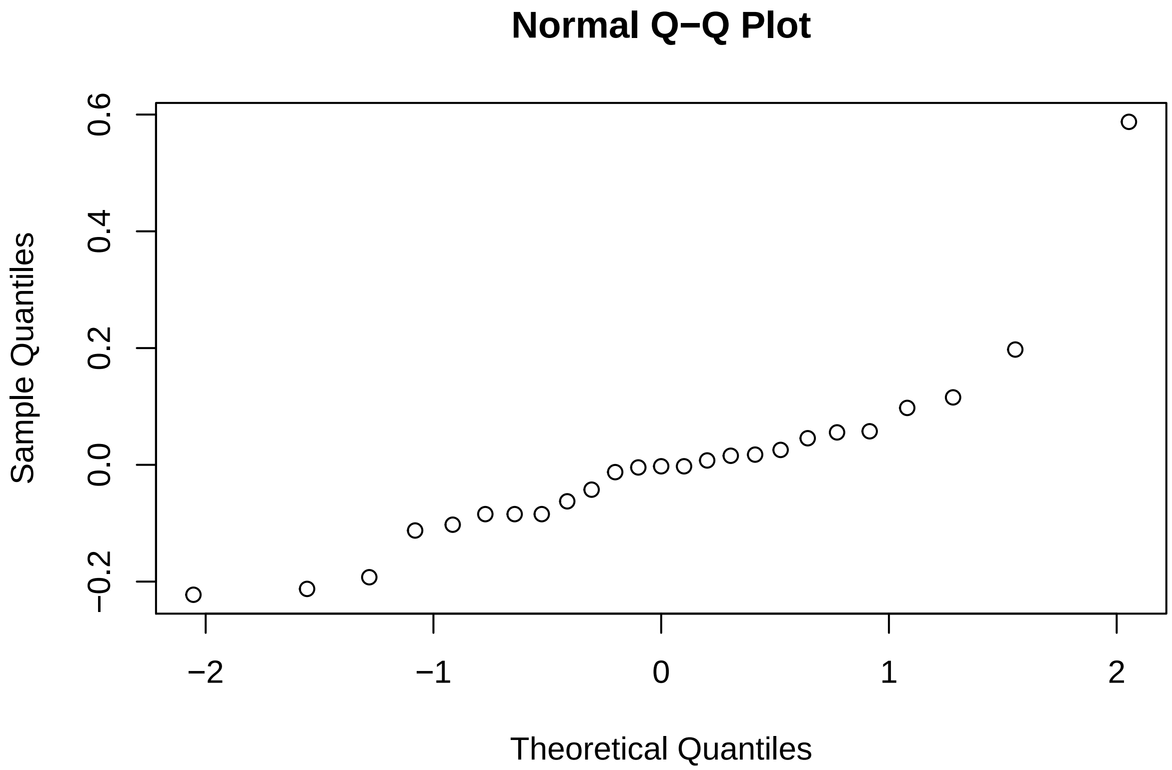
1). Non-normality.

First plot the residual's bar plot and qq plot for explanatory analysis:

```
hist(resid(anov1), 10)
```



```
qqnorm(resid(anov1))
```



It's hard to tell from the bar plot and qq plot for the residuals. For more information we can use Shapiro-Wilk normality test:

```
shapiro.test(resid(anov1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(anov1)
## W = 0.8155, p-value = 0.0004176
```

p value is smaller than 0.05, so the assumption of normality is invalid

2). Unequal Variances.

Use Bartlett's test. Sine Bartlett's test is highly depend on normality assumption, and from the above we see that the normality assumption is valid, we can use Bartlett's test.

```
library(car)
leveneTest(y = dat$Y, g = dat$name)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2     1.3   0.29
##       22
```

p value is greater than 0.05, so the homogeneity of variances is valid

c). Compare the results to those obtained using a non-parametric procedure.

Use Kruskal-Wallis test:

```
kruskal.test(x = dat$Y, g = dat$name)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dat$Y and dat$name
## Kruskal-Wallis chi-squared = 13.87, df = 2, p-value = 0.0009744
```

Since p value is smaller 0.05, we reject H_0 and think there is a significant difference among the group means.

2. Consider the IQ scores data of Display 13.24, problem 19, Chapter 13.

Frist input the data.

```
library(Sleuth2)
dat = ex1319
```

a). (Do problem 19) Does the difference in mean socres for those with high and low SES biological parents depend on whether the adoptive parents were high or low SES? If not, how much is the mean IQ score affected by the SES of adoptive parents, and how much is it affected by the SES of the biological parents? Is one of these effects larger than the other?

Use two-way anova and see the p value for the interaction term:

```
anov2 = aov(IQ ~ Adoptive * Biologic, data = dat)
summary(anov2)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## Adoptive           1   1478    1478    8.46 0.00637 **
## Biologic           1   2291    2291   13.11 0.00094 ***
## Adoptive:Biologic  1      2       2    0.01 0.91744
## Residuals         34   5941     175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value of the interaction term is greater than 0.05, which implies that the interaction term is not significant, so `Adoptive` and `Biologic` are not depend on each other.

Split the data into `biologic.high` and `biologic.low` two subsets, and apply anova on these two datasets to see whether the `Adoptive` is significant on IQ score.

```
biologic.high = subset(dat, Biologic == "High")
summary(aov(IQ ~ Adoptive, data = biologic.high))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Adoptive     1    651     651    4.43  0.051 .
## Residuals   16   2348     147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
biologic.low = subset(dat, Biologic == "Low")
summary(aov(IQ ~ Adoptive, data = biologic.low))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Adoptive     1    627     627    3.14  0.093 .
## Residuals   18   3593     200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above we can see that two test both has a p value greater than 0.05, so we think in both `biologic.high` and `biologic.low` subsets, the SES of adotive parents is not significant on `IQ`. Thus we think that the difference in mean scores for those with high and low SES biological parents does NOT depend on whether the adoptive paremts were high or low SES.

To analyze how much is the mean IQ score affected by each of them, we use ancova:

```
ancova = aov(IQ ~ Adoptive + Biologic, data = dat)
summary(ancova)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Adoptive     1   1478    1478     8.7 0.00564 **
## Biologic     1   2291    2291    13.5 0.00079 ***
## Residuals   35   5943     170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

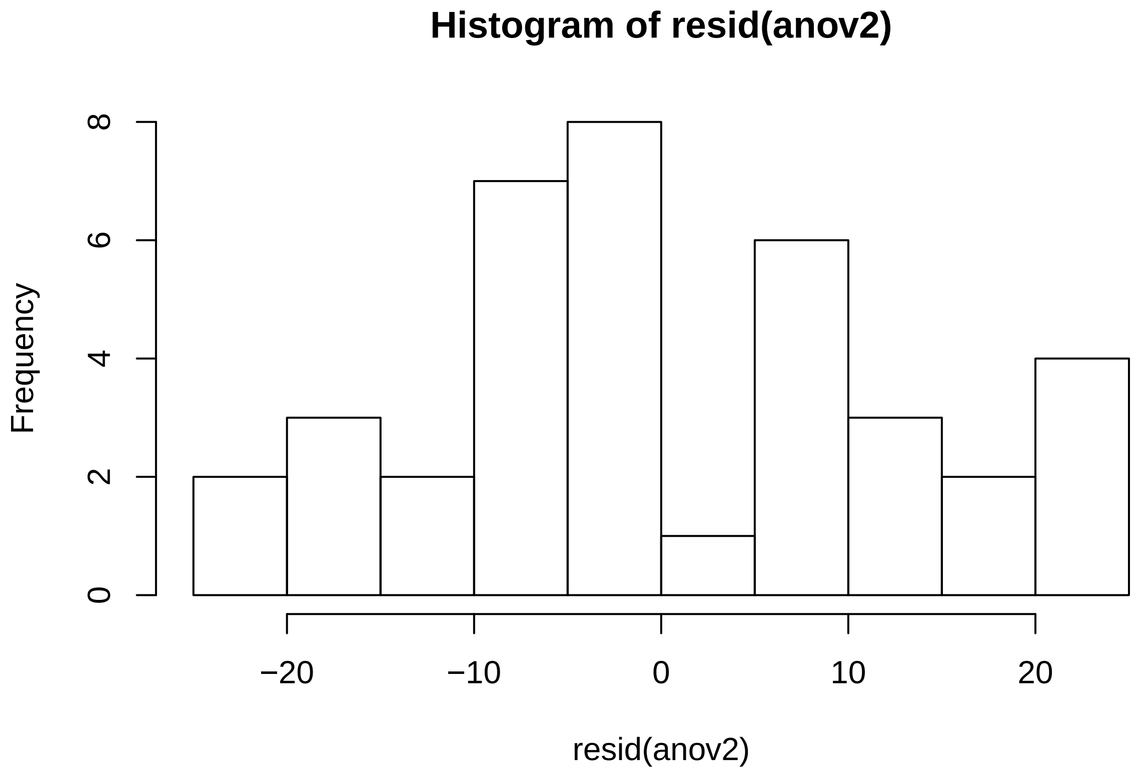
p values we can see that both `Adoptive` and `Biologic` are significant. Since the p value for `Biologic` is smaller than p value for `Adoptive`, we think `Biologic` affected on mean scores more than `Adoptive`.

b. Assess the validity of all assumptions.

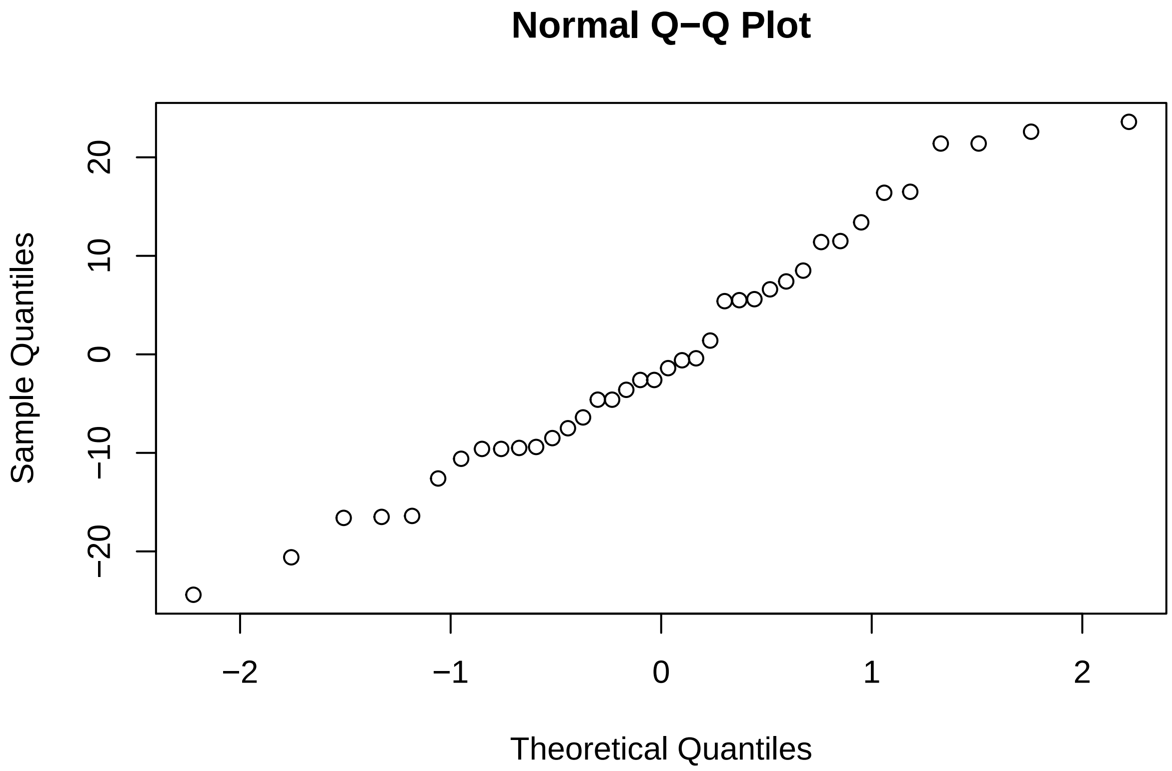
1). Non-normality.

First plot the residual's bar plot and qq plot for explanatory analysis:

```
hist(resid(anov2), 10)
```



```r
qqnorm(resid(anov2))
```



It’s hard to tell from the bar plot and qq plot for the residuals. For more information we can use Shapiro-Wilk normality test:

```
shapiro.test(resid(anov2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(anov2)
## W = 0.9706, p-value = 0.4071
```

p value is greater than 0.05, so the assumption of normality is valid

2). Non-Parallel Regression lines.

Use two-anova:

```
summary(anov2)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## Adoptive           1   1478    1478    8.46 0.00637 **
## Biologic           1   2291    2291   13.11 0.00094 ***
## Adoptive:Biologic  1      2       2    0.01 0.91744
## Residuals         34   5941     175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that interaction term is not significant. So the parallelism is valid.