Lab 8: One-Way ANOVA in R

Nov, 2019 STA 206

This handout is based on chapter 16.1 of Julian J. Faraway's book *Practical Regression and Anova using R*, and lecture notes of STA206.

One-Way ANOVA Model

Model equation: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, \cdots, I, \quad j = 1, \cdots, n_i.$

Due to identifiability issues, we need to put some constraint on α_i's for this model to be estimable:

- Set $\alpha_1 = 0$, this corresponds to treatment contrasts, i.e. we set the first level as the baseline.

Coagulation Data

The example data set we will use is a set of 24 blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

Read in the data and check model assumptions

```
> coagulation = read.table('coagulation.txt',header=TRUE)
> summary(coagulation)
      coag       diet
 Min.   :56.00   A:4
 1st Qu.:61.75   B:6
 Median :63.50   C:6
 Mean   :64.00   D:8
 3rd Qu.:67.00
 Max.   :71.00
```

The first step is to plot the data, box-plots are useful:

```
plot(coag~diet, data=coagulation)
```

The output is the left panel in Figure 1.
Before we look at the plot, we note what information a box-plot can provide:

1. **Outliers**: these will be apparent as separated points on the box-plots.

2. **Skewness**: this will be apparent from an asymmetrical form for the boxes.
3. **Unequal variance (heteroskedasticity)**: this will be apparent from clearly unequal box sizes.
4. **Factor Effects**: Whether the means of factor levels appear to be different.

Fitting the model and diagnostics

We fit the model under the restriction $\alpha_1 = 0$, i.e. we set level A as our reference class.

```
> g <- lm(coag~diet, coagulation)
```

By inspection of the model design matrix, we note that this design matrix is the same as if we use dummy variables for diet B,C, and D where diet A is the reference class (as in regression).

```
> head( model.matrix(g) )
    (Intercept) dietB dietC dietD
1             1     0     0     0
2             1     0     0     0
3             1     0     0     0
4             1     0     0     0
5             1     1     0     0
6             1     1     0     0
```

Now we explore the diagnostic plots from our model fit in Figure 1 to assess our model assumptions.

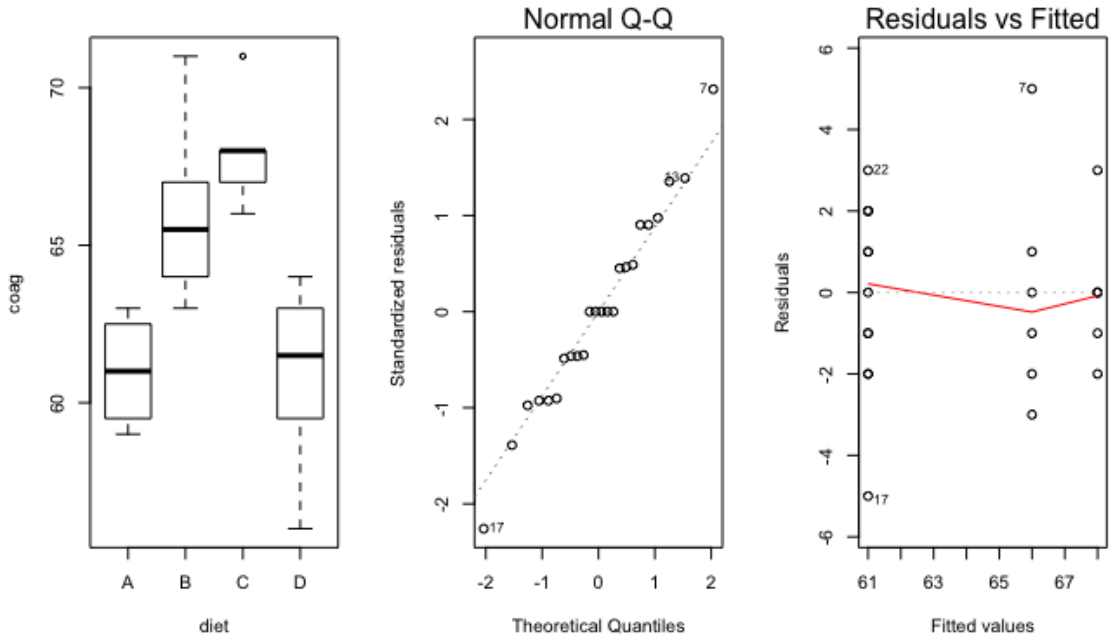

Figure 1

For the first plot in figure 1 (box-plot), we observe:

1. **Outliers**: Group C has 1 distinct outlier that may cause problems.
2. **Skewness**: No group is strongly skewed (left or right).
3. **Unequal variance (heteroskedasticity)**: The groups seem to vary differently but we do have small group sizes.
4. **Factor Effects**: There seems to be strong evidence that the factor level means are different. Also, we can see that the factor level means between Diet A and D does not seem significantly different.

Now consdier the second and third plots of Figure 1. From the QQ-norm plot, we see that the error distribution is approximately normal (slightly heavy-tailed). The residual vs fitted plot shows no sign of unequal error variance.

Similar to regression models, we obtain the summary of this fitted model with the *summary()* function:

```
> summary(g)

Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
   Min     1Q Median     3Q    Max
 -5.00  -1.25   0.00   1.25   5.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared:  0.6706,    Adjusted R-squared:  0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

This provides most of the information we need to perform statistical inference. For example, the estimated group means are

$$\widehat{\mu}_A = 61, \quad \widehat{\mu}_B = \widehat{\mu}_A + 5 = 66, \quad \widehat{\mu}_C = \widehat{\mu}_A + 7 = 68, \quad \widehat{\mu}_D = \widehat{\mu}_A + 0 = 61.$$

We can also construct C.I.'s of the group means; consider a 95% confidence interval for the mean of group B is:

$$C.I._B^{0.95} = \widehat{\mu}_B \pm s(\widehat{\mu}_B) t\left(1 - \frac{\alpha}{2}; n_T - I\right)$$

where $\widehat{\mu}_B = 66$, $s(\widehat{\mu}_B) = \sqrt{MSE/n_B} = 2.366 * \sqrt{1/6} = 0.9660918$. Hence

$$C.I._B^{0.95} = [63.98477, \quad 68.01523].$$

Calculating this in R,

```
> ## C.I. for group B
> muB = g$coef[1]+g$coef[2]
```

```
> sdB = summary(g)$sig*sqrt(1/6)
> muB+qt(0.975,20)*sdB*c(-1,1)
[1] 63.98477 68.01523
```

We can conduct pair-wise comparisons as well. For example, we would like to compare the means of group A and group B.

$$\widehat{D}_{BA} = \widehat{\mu}_B - \widehat{\mu}_A = 66 - 61 = 5,$$

$$s(\widehat{D}_{BA}) = \sqrt{MSE(1/n_A + 1/n_B)} = 2.366 * \sqrt{1/4 + 1/6} = 1.527525.$$

Test statistic $\frac{\widehat{D}_{BA}}{s(\widehat{D}_{BA})} = 3.27$ is a $t(n_T - I)$ distribution under the null hypothesis ($H_0 : D_{BA} = 0$). Therefore, the corresponding p-value is 0.0038 which agrees with what we have for dietB in the summary of the fitted model. And we reject the null hypothesis.

```
> ## pairwise comparison B,A
> d_BA = g$coef[2]
> sd_BA = summary(g)$sig*sqrt(1/4+1/6)
> 2*(1-pt(d_BA/sd_BA,nT-I)) ##p-value
      dietB
0.003802505
```

If we would like to compare the group means between group C and group B, then

$$\widehat{D}_{CB} = \widehat{\mu}_C - \widehat{\mu}_B = \hat{\alpha}_C - \hat{\alpha}_B = 2,$$

$$s(\widehat{D}_{CB}) = \sqrt{MSE(1/n_B + 1/n_C)} = 2.366 * \sqrt{1/6 + 1/6} = 1.36626,$$

$$C.I._{CB}^{0.95} = \widehat{D}_{CB} \pm s(\widehat{D}_{CB}) * t(0.975; n_T - I) = [-0.85, 4.85].$$

Since the 95% C.I. of the difference between the means of group B and C contains 0, we can not reject the null hypothesis $H_0 : D_{CB} = 0$.

```
> ## pairwise comparison C,B
> d_CB = g$coef[3]-g$coef[2]
> sd_CB = summary(g)$sig*sqrt(1/6+1/6)
> c(d_CB-qt(0.975,20)*sd_CB,d_CB+qt(0.975,20)*sd_CB)
     dietC      dietC
-0.8499686  4.8499686
```

Multiple Comparison

We consider pairwise comparisons first. A simple $(1-\alpha)$-C.I. for $\mu_i - \mu_j = \alpha_i - \alpha_j$ is

$$(\hat{\alpha}_i - \hat{\alpha}_j) \pm t\left(1 - \frac{\alpha}{2}; n_T - I\right)\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

A test for $H_0 : \mu_i = \mu_j$ amounts to seeing whether zero lies in this interval or not. This is fine for just one test but suppose we do a lot of pairwise tests at $\alpha = 5\%$, the family-wise type-I error rate will be much bigger than 5%.

Now we return to our real data. We've found that there is a significant difference among the diets. But which diets can be said to be different and which diets are not distinguishable? The estimated difference between diet B and diet C is 2. First we do the regular t-distribution calculation:

```
> ## Regular CI
> qt(1-.05/2,20)
[1] 2.085963
> qt(1-.05/2,20)*summary(g)$sig*sqrt(1/6+1/6)
[1] 2.849969
> c(2-2.85,2+2.85)# 95% regular CI of difference between B and C
[1] -0.85  4.85
```

If we perform the Tukey's procedure:

```
> ## Tukey's simultaneous CI
> qtukey(0.95,4,20)
[1] 3.958293
> (3.96/sqrt(2))*summary(g)$sig*sqrt(1/6+1/6)
[1] 3.825723
> c(2-3.83,2+3.83)# 95% Tukey's CI of difference between B and C
[1] -1.83  5.83
```

The $1-\alpha$ confidence intervals constructed by Tukey's method is computed as

$$\hat{u_i} - \hat{u_j} \pm se(\hat{u_i} - \hat{u_j}) * \frac{1}{\sqrt{2}} q_{1-\alpha,I,n-I}$$

where $q_{1-\alpha,I,n-I}$ is the $1-\alpha$ quantile of the studentized range distribution with parameter I ,n-I.
Another convenient way of obtaining Tukey's intervals is as following:

```
> TukeyHSD(aov(coag~diet, coagulation))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
    diff         lwr       upr     p adj
B-A    5   0.7245544  9.275446 0.0183283
C-A    7   2.7245544 11.275446 0.0009577
D-A    0  -4.0560438  4.056044 1.0000000
C-B    2  -1.8240748  5.824075 0.4766005
D-B   -5  -8.5770944 -1.422906 0.0044114
D-C   -7 -10.5770944 -3.422906 0.0001268
```

So at family-wise significance level 0.05, we reject the following null hypothese: $\mu_A = \mu_B, \mu_A = \mu_C, \mu_B = \mu_D, \mu_C = \mu_D$, as zero is not contained in the respective intervals. Suppose two comparisons were pre-specified (including the comparison between group B and group C), then using the Bonferroni correction we can get the C.I. of the difference between group B and group C by:

```
> ## Bonferroni CI
> qt(1-0.05/(2*2),20)*summary(g)$sig*sqrt(1/6+1/6)
[1] 3.310607
> c(2-3.31,2+3.31)# 95% Bonferroni CI of difference between B and C
# when doing 2 comparisons
[1] -1.31  5.31
```

This is narrower than Tukey's interval.

Contrasts

A contrast among the effects $\alpha_1, \cdots, \alpha_I$ is a linear combination $\sum_i c_i \alpha_i$ which satisfies $\sum_i c_i = 0$. For example comparison of the mean for group B and group C is a contrast with $c_1 = 0, c_2 = 1, c_3 = -1, c_4 = 0$.

```
> ## Scheffe's CI
> sqrt((4-1)*qf(0.95,3,20))
[1] 3.048799
> sqrt((4-1)*qf(0.95,3,20))*summary(g)$sig*sqrt(1/6+1/6)
```

```
[1] 4.165452
> c(2-4.17,2+4.17)# 95% Scheffe's CI of difference between B and C
[1] -2.17  6.17
```

In general the Scheffe's C.I. for a contrast $L = \sum_i c_i \alpha_i$ is computed as:

$$\hat{L} = \sum_i c_i \hat{\alpha}_i, \quad s(\hat{L}) = \sqrt{MSE(\sum_i \frac{c_i^2}{n_i})},$$
$$C.I._L^s = \hat{L} \pm S * s(\hat{L}), \quad \text{where } S^2 = (I-1)F(1-\alpha; I-1; n_T - I).$$

Scheffe's procedure can deal with finite or infinitely many contrasts.

If we look at the multipliers (or S, above, for each interval) for all pair-wise comparisons in each procedure for our data set:

```
> ## multipliers for all pairwise comparisons
> sqrt(1/2)*qtukey(0.95,4,20)## Tukey
[1] 2.798936
> qt(1-0.05/12,20)## Bonferroni
[1] 2.927119
> sqrt((4-1)*qf(0.95,3,20))## Scheffe
[1] 3.048799
```

We can see that the Tukey's multiplier is the smallest, hence resulting in narrower C.I. which is preferred for the family of all pair-wise comparisons. And Scheffe's multiplier is the biggest. **Note**: Tukey's procedure only deals with pair-wise comparisons.

In practice which procedure to use depends on: whether it is applicable and whether it results in a smaller multiplier.

Nonparametric rank F test

Nonparametric rank F test is usually used for non-normal data, such as discrete data since it does not require the normality assumption. The procedure goes as follows:

- Get the ranks R of the data.
- Perform ANOVA on the ranks:

$$MSTR(R) = \frac{\sum_{i=1}^{I} n_i(\bar{R}_{i.} - \bar{R}_{..})^2}{I-1},$$

$$MSE(R) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(R_{ij} - \bar{R}_{i.})^2}{n_T - I}.$$

- Under the null hypothesis of equal mean: $F^* = \frac{MSTR(R)}{MSE(R)}$ follows approximately $F(I-1; n_T - I)$ distribution, where $MSTR(R)$ and $MSE(R)$ are based on one-way anova of the R_{ij}'s.

We may perform it here to the coagulation data here since it is discrete. First we need to get the ranks of the data.

```
> coag.rank = rank(coagulation[,1],ties.method="average")
> rbind(coag, coag.rank)[1:2,1:5]
          [,1] [,2] [,3] [,4] [,5]
coag      62.0 60.0 63.0 59.0 63.0
coag.rank  7.5  4.5 10.5  2.5 10.5
```

Then we perform ANOVA on the ranks (as our response).

```
> fit.F = lm(coag.rank~diet)
> summary(fit.F)

Call:
lm(formula = coag.rank ~ diet)

Residuals:
   Min     1Q Median     3Q    Max
-6.000 -2.562  0.375  2.500  7.250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.250      1.925   3.246 0.004043 **
dietB         10.000      2.485   4.024 0.000666 ***
dietC         14.000      2.485   5.633 1.63e-05 ***
dietD          0.750      2.358   0.318 0.753715
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.85 on 20 degrees of freedom
Multiple R-squared:  0.7398,    Adjusted R-squared:  0.7008
F-statistic: 18.95 on 3 and 20 DF,  p-value: 4.599e-06
```

Since the p-value=4.6e-6 which is very small, we can reject the null hypothesis of equal means.