Stat 6021: Pairwise Comparisons from R Output

Read this after going over the tutorial for this module.

1 Setting Up Our Example

In the tutorial, we looked at a data set that contains ratings of various wines produced in California. We focused on the response variable $y = \mathit{Quality}$ (average quality rating), $x_1 = \mathit{Flavor}$ (average flavor rating), and *Region* indicating which of three regions (North / Central / Napa) in California the wine is produced in. We ended up creating two indicator variables to represent the three regions, with Napa as the reference class, i.e.,

$$
\begin{aligned}
I_1 &= \begin{cases} 1 & \text{if North} \\ 0 & \text{otherwise;} \end{cases} \\
I_2 &= \begin{cases} 1 & \text{if Central} \\ 0 & \text{otherwise;} \end{cases}
\end{aligned}
$$

The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \epsilon.$$

So the regression functions are:

North region: $\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) = (\beta_0 + \beta_2) + \beta_1 x_1$
Central region: $\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3) + \beta_1 x_1$
Napa region: $\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 x_1$

We note the following:

- The difference in mean quality rating between the North and Napa regions is denoted by β_2.
- The difference in mean quality rating between the Central and Napa regions is denoted by β_3.
- The difference in mean quality rating between the North and Central regions is denoted by $\beta_2 - \beta_3$.

2 R Output

After fitting the model, we have the following output from R for the coefficients

```
  (Intercept)        Flavor   RegionNorth RegionCentral
        8.318         1.116        -1.223        -2.757
```

and the variance-covariance matrix for the coefficients

```
              (Intercept) Flavor RegionNorth RegionCentral
(Intercept)         1.020 -0.170      -0.277        -0.277
Flavor             -0.170  0.030       0.037         0.037
RegionNorth        -0.277  0.037       0.160         0.113
RegionCentral      -0.277  0.037       0.113         0.202
```

The entries in the diagonals of this matrix give the estimated variance of the corresponding coefficient, for example, the estimated variance of β_0 is 1.020, and so its standard error is $\sqrt{1.020}$.

The entries in the off-diagonals of this matrix give the estimated covariance of the corresponding coefficients, for example, the estimated covariance between β_0 and β_1 is -0.170.

3 Pairwise Comparisons using Bonferroni Procedure

Using the Bonferroni procedure, compute the 95% family confidence intervals for the difference in mean quality rating between wines in the

1. North and Napa regions;
2. Central and Napa regions;
3. North and Central regions.

As we are making three pairwise comparisons here, the multiplier for the confidence interval is $\Delta = t_{1-0.05/(2\times g),n-p} = t_{1-0.05/6,34} = 2.518$ since $g = 3, n = 38$ and $p = 4$, where g denotes the number of pairwise comparisons we are making.

For the difference in mean quality rating between wines in the North and Napa regions, we use the confidence interval for β_2, i.e.,

$$\begin{array}{rcl} \hat{\beta}_2 & \pm & \Delta se(\hat{\beta}_2) \\ = -1.223 & \pm & 2.518 \times \sqrt{0.160} \\ = (-2.230 & , & -0.216) \end{array}$$

Likewise, for the difference in mean quality rating between wines in the North and Napa regions, we use the confidence interval for β_3, which results in (-3.889, -1.625).

For the difference in mean quality rating between wines in the North and Central regions, we use the confidence interval for $\beta_2 - \beta_3$. We need to find the estimated variance of $\beta_2 - \beta_3$ first, i.e.,

$$\begin{aligned} s^2\{\hat{\beta}_2 - \hat{\beta}_3\} &= s^2\{\hat{\beta}_2\} + s^2\{\hat{\beta}_3\} - 2s\{\hat{\beta}_2, \hat{\beta}_3\} \\ &= 0.160 + 0.202 - 2(0.113) \\ &= 0.136. \end{aligned}$$

using the general result (1) shown below. Therefore, the CI for $\beta_2 - \beta_3$ is

$$\begin{aligned} (\hat{\beta}_2 - \hat{\beta}_3) &\pm \Delta se\{\hat{\beta}_2 - \hat{\beta}_3\} \\ = (-1.223 + 2.757) &\pm 2.518\sqrt{0.136} \\ = (1.192 \quad &, \quad 1.876). \end{aligned}$$

Note: A general result for the variance of any linear combination of random variables X and Y is

$$s^2\{aX + bY\} = a^2 s^2\{X\} + b^2 s^2\{Y\} + 2ab \times s\{X, Y\} \tag{1}$$