Multiple Comparisons

David M. Rocke

March 11, 2025

Multiple Comparisons

The coefficients in the linear models for the poison data set are tests of the hypothesis that the given poison/treatment differs from the default (first) level. So we have tests of the difference between Poison II and Poison I and between Poison III and Poison I, but not between Poison III and Poison II. Similarly, we have tests of differences between each of Treatments B, C, and D and Treatment A, but no tests of the three differences within the the three non-default treatment levels. From the F-statistic on each factor, we know differences exist, but not the full list of significant differences.

The Treatment factor has four levels and therefore six distinct pairwise comparisons, and well as possibly other linear hypotheses, and if we conduct lots of tests on the four coefficients in the table, we may have false positives. Of course a test of whether all the rates are equal is obtainable by the F-test in `drop1()`. We computed all the pairwise comparisons, but with no control of false positives from multiple comparisons.

Note that Wald tests of any linear hypothesis can be conducted for any type of regression model for which asymptotically valid covariance matrices can be derived. This includes not just linear regression, but logistic regression, Poisson regression, and many others.

R Package `multcomp`

This package allows post-hoc comparisons among levels of factors, with adjustment to protect against false positives. This can be important in any situation in which comparisons of levels of factors are made when there are more than two levels. We will eventually apply this to the poison data and compare the uncorrected statistics we have already computed with the multiplicity-corrected ones.

A linear hypothesis on a vector of coefficients β of length p with estimates $\hat{\beta}$ is of the form

$$H_0 : L^\top \beta = k,$$

where L is a vector of numbers of length p; often L is a contrast meaning that the sum of the entries is zero and k is also often zero. If $\hat{\beta}$ has estimated covariance matrix $\hat{V}$, then the estimated variance of $L^\top \hat{\beta}$ is $L^\top \hat{V} L$ and an approximate z-statistic for the hypothesis as stated is

$$z = \frac{L^\top \hat{\beta} - k}{\sqrt{L^\top \hat{V} L}}.$$

This can be referred to a t-distribution in linear regression, though not always in other methods.

Comparison with a Control

```
recovery {multcomp}
Recovery time after surgery.

This data frame contains the following variables

blanket
blanket type, a factor at four levels: b0, b1, b2, and b3.

minutes
response variable: recovery time after a surgical procedure.

Details
A company developed specialized heating blankets designed to help the body heat
following a surgical procedure. Four types of blankets were tried on surgical
patients with the aim of comparing the recovery time of patients.
One of the blanket was a standard blanket that had been in use already
in various hospitals.

Source
P. H. Westfall, R. D. Tobias, D. Rom, R. D. Wolfinger, Y. Hochberg (1999).
Multiple Comparisons and Multiple Tests Using the SAS System. Cary,
NC: SAS Institute Inc., page 66.
```

```
> library(multcomp)
> data(recovery)
> recovery.lm <- lm(minutes~blanket,data=recovery)
> summary(recovery.lm)

Call:
lm(formula = minutes ~ blanket, data = recovery)

Residuals:
   Min     1Q Median     3Q    Max
-6.133 -1.800  0.200  2.200  4.867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.59 on 37 degrees of freedom
Multiple R-squared:  0.3797,    Adjusted R-squared:  0.3294
F-statistic:  7.55 on 3 and 37 DF,  p-value: 0.0004619
```

It looks like blanket b2 is better than b0, but we did conduct three hypothesis tests to obtain that finding. The F-test shows that not all the blankets are the same, so it might be reasonable to attribute that only to b2, but we can test that allowing for the multiple comparisons and the correlations between the tests using the Dunnett procedure and also obtain confidence intervals adjusted for multiple comparisons. This is based on the multivariate t distribution of the coefficients and is implemented in the `glht()` command in the R package `multicomp`.

```
> recovery.mc <- glht(recovery.lm,linfct=mcp(blanket="Dunnett"))
> summary(recovery.mc)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

```
> names(recovery.mc)
[1] "model"       "linfct"      "rhs"         "coef"        "vcov"        "df"
[7] "alternative" "type"        "focus"

> recovery.mc$linfct

        (Intercept) blanketb1 blanketb2 blanketb3
b1 - b0           0         1         0         0
b2 - b0           0         0         1         0
b3 - b0           0         0         0         1
attr(,"type")
[1] "Dunnett"

> recovery.mc$rhs
[1] 0 0 0
> recovery.mc$focus
[1] "blanket"
```

Some attributes of an object have extractor functions, including `coef` and `vcov`. All the components can be accessed as attributes of the object. The three linear hypotheses require the linear vectors L and the right-hand sides k.

```
contrMat(n, type = c("Dunnett", "Tukey", "Sequen", "AVE",
                     "Changepoint", "Williams", "Marcus",
                      "McDermott", "UmbrellaWilliams", "GrandMean"),
                    base = 1)
Arguments
n                    a (possibly named) vector of sample sizes for each group.
type                 type of contrast.
base                 an integer specifying which group is considered the baseline
                     group for Dunnett contrasts.
```

This lists the types of pre-specified contrasts. Any set of linear hypotheses can also be specified just as a matrix `linfct` and right-hand side vector `rhs`, as we did by hand with the poison data. A base level can be given for Dunnett comparisons, which for general hypotheses is the `focus` attribute.

```
recovery.lm
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .

recovery.mc
Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182
```

Note that the t-scores are the same, but the p-values are adjusted for multiple comparisons so that the chance that one or more is significant at level α in the null case is less than or equal to α. The only hypothesis to survive the multiplicity correction is that b2 is better than the standard, b0.

```
> summary(recovery.mc,test  = adjusted(type="bonferroni"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330 0.574796
b2 - b0 == 0  -7.4667     1.6038  -4.656 0.000122 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884 0.202439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- bonferroni method)
```

This method allows changing the multiplicity adjustment method. In this case, we replace the Dunnett method, which accounts for correlations in the tests, with the Bonferroni method, which does not. Note that `adjusted(type = "none")` gives the original tests with no multiplicity adjustment.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.8000     0.5792  25.552  < 2e-16 ***          lm
blanketb1    -2.1333     1.6038  -1.330   0.1916
blanketb2    -7.4667     1.6038  -4.656 4.07e-05 ***
blanketb3    -1.6667     0.8848  -1.884   0.0675 .

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330    0.456           Dunnett
b2 - b0 == 0  -7.4667     1.6038  -4.656   <0.001 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884    0.182

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
b1 - b0 == 0  -2.1333     1.6038  -1.330 0.574796           Bonferroni
b2 - b0 == 0  -7.4667     1.6038  -4.656 0.000122 ***
b3 - b0 == 0  -1.6667     0.8848  -1.884 0.202439
```

Both Dunnett and Bonferroni protect the familywise error rate, but Dunnett has smaller p-values because it uses the correlations of the tests.

```
> confint(recovery.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Quantile = 2.489
95% family-wise confidence level


Linear Hypotheses:
              Estimate lwr      upr
b1 - b0 == 0  -2.1333  -6.1251   1.8584
b2 - b0 == 0  -7.4667 -11.4584  -3.4749
b3 - b0 == 0  -1.6667  -3.8688   0.5355
```

There is at least a 95% chance that all the true values of the contrasts lie in their stated intervals. We can use the argument `calpha=qt(.975,37)` (2.026192) to get standard, uncorrected confidence intervals if desired.

```
> confint(recovery.mc,calpha=qt(0.975,37))

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Dunnett Contrasts


Fit: lm(formula = minutes ~ blanket, data = recovery)

Quantile = 2.0262
95% confidence level


Linear Hypotheses:
              Estimate lwr      upr
b1 - b0 == 0  -2.1333  -5.3829   1.1162
b2 - b0 == 0  -7.4667 -10.7162  -4.2171
b3 - b0 == 0  -1.6667  -3.4594   0.1261
```

```
> confint(recovery.lm)
                 2.5 %     97.5 %
(Intercept)  13.626389 15.9736107
blanketb1    -5.382914  1.1162474
blanketb2   -10.716247 -4.2170859
blanketb3    -3.459387  0.1260532

Linear Hypotheses:                                  with standard t quantile
              Estimate lwr      upr
b1 - b0 == 0  -2.1333  -5.3829   1.1162
b2 - b0 == 0  -7.4667 -10.7162  -4.2171
b3 - b0 == 0  -1.6667  -3.4594   0.1261

Linear Hypotheses:                                  with adjusted t quantile
              Estimate lwr      upr
b1 - b0 == 0  -2.1333  -6.1251   1.8584
b2 - b0 == 0  -7.4667 -11.4584  -3.4749
b3 - b0 == 0  -1.6667  -3.8688   0.5355
```

The confidence intervals from `lm` are individually valid, but if we consider them to be independent the chance that at least one does not contain the true value is $1 - (0.95)^3 = 0.14$. We could use Bonferroni confidence intervals at 98.3% confidence, but the Dunnett ones will be narrower because they use the correlations of the variables.

All Pairs Comparisons

```
immer {MASS}
Yields from a Barley Field Trial
Description
The immer data frame has 30 rows and 4 columns. Five varieties of barley were
grown in six locations in each of 1931 and 1932.

This data frame contains the following columns:

Loc
The location.

Var
The variety of barley ("manchuria", "svansota", "velvet", "trebi" and "peatland").

Y1
Yield in 1931.

Y2
Yield in 1932.
```

```
> library(MASS)
> data(immer)
> immer1 <- data.frame(immer,Yield = (immer$Y1+immer$Y2))
> summary(immer.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  204.403     12.156  16.815 2.88e-13 ***
VarP          16.300     12.156   1.341 0.194983
VarS          -6.517     12.156  -0.536 0.597810
VarT          47.617     12.156   3.917 0.000854 ***
VarV           9.583     12.156   0.788 0.439728
LocD         -52.120     13.316  -3.914 0.000860 ***
LocGR        -56.680     13.316  -4.256 0.000386 ***
LocM          -7.180     13.316  -0.539 0.595705
LocUF        -32.020     13.316  -2.405 0.025996 *
LocW          54.280     13.316   4.076 0.000589 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.05 on 20 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.7924
F-statistic:  13.3 on 9 and 20 DF,  p-value: 1.216e-06
```

```
> drop1(immer.lm,test="F")
Single term deletions

Model:
Yield ~ Var + Loc
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>               8866 190.66
Var     4     10620 19486 206.29  5.9891  0.002453 **
Loc     5     42442 51308 233.33 19.1480 5.212e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> tapply(immer1$Yield,immer1$Var,mean)
       M        P        S        T        V
188.7833 205.0833 182.2667 236.4000 198.3667
> sort(tapply(immer1$Yield,immer1$Var,mean))
       S        M        V        P        T
182.2667 188.7833 198.3667 205.0833 236.4000
```

Both variety and location are significant, but it is not clear which pairs of varieties are shown to differ. Variety T, however, has by far the highest yield.

```
       S        M        V        P        T
182.2667 188.7833 198.3667 205.0833 236.4000

> immer.mc <- glht(immer.lm,linfct=mcp(Var = "Tukey"))
> summary(immer.mc)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Yield ~ Var + Loc, data = immer1)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
P - M == 0   16.300     12.156   1.341  0.67008
S - M == 0   -6.517     12.156  -0.536  0.98242
T - M == 0   47.617     12.156   3.917  0.00675 **
V - M == 0    9.583     12.156   0.788  0.93102
S - P == 0  -22.817     12.156  -1.877  0.36064
T - P == 0   31.317     12.156   2.576  0.11336   T is not provably better than P
V - P == 0   -6.717     12.156  -0.553  0.98035
T - S == 0   54.133     12.156   4.453  0.00201 **
V - S == 0   16.100     12.156   1.324  0.67981
V - T == 0  -38.033     12.156  -3.129  0.03773 *
```

```
> summary(immer.mc,test=adjusted(type="none"))

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = Yield ~ Var + Loc, data = immer1)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
P - M == 0   16.300     12.156   1.341 0.194983
S - M == 0   -6.517     12.156  -0.536 0.597810
T - M == 0   47.617     12.156   3.917 0.000854 ***
V - M == 0    9.583     12.156   0.788 0.439728
S - P == 0  -22.817     12.156  -1.877 0.075185 .
T - P == 0   31.317     12.156   2.576 0.018029 *  T is better than P
V - P == 0   -6.717     12.156  -0.553 0.586700
T - S == 0   54.133     12.156   4.453 0.000244 ***
V - S == 0   16.100     12.156   1.324 0.200289
V - T == 0  -38.033     12.156  -3.129 0.005288 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- none method)
```

```
> confint(immer.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts

Fit: lm(formula = Yield ~ Var + Loc, data = immer1)

Quantile = 2.9932
95% family-wise confidence level


Linear Hypotheses:
            Estimate lwr      upr
P - M == 0   16.3000 -20.0850  52.6850
S - M == 0   -6.5167 -42.9016  29.8683
T - M == 0   47.6167  11.2317  84.0016
V - M == 0    9.5833 -26.8016  45.9683
S - P == 0  -22.8167 -59.2016  13.5683
T - P == 0   31.3167  -5.0683  67.7016
V - P == 0   -6.7167 -43.1016  29.6683
T - S == 0   54.1333  17.7484  90.5183
V - S == 0   16.1000 -20.2850  52.4850
V - T == 0  -38.0333 -74.4183  -1.6484
```

The confidence intervals and tests that result from uncorrected `lm` and other regression models are often called *Least Significant Difference* $=$ *LSD* tests and intervals. When there are many levels of a factor, this can result in false positives. One possible intermediate choice is to use the LSD tests and intervals, but only if the anova test for the factor is significant. This method is sometimes called the *Protected LSD*. This protects against the case where all the levels have equal effect, but not against partial equalities. However, for some applications this may be enough.

Poison Example

```
> poison$Poison <- factor(poison$Poison)
> poison$Treatment <- factor(poison$Treatment)
> poisont.lm <- lm(invsurv~Poison+Treatment,data=poison)
> summary(poisont.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.961      2.906  15.473  < 2e-16 ***
PoisonII       7.811      2.906   2.688  0.01026 *
PoisonIII     33.274      2.906  11.451 1.69e-14 ***
TreatmentB   -27.623      3.355  -8.233 2.66e-10 ***
TreatmentC    -9.536      3.355  -2.842  0.00689 **
TreatmentD   -22.639      3.355  -6.747 3.35e-08 ***

Residual standard error: 8.219 on 42 degrees of freedom
Multiple R-squared:  0.8441,    Adjusted R-squared:  0.8255
F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

Poison and Treatment are character strings in the data set, not factors. The `lm()` routine automatically converts, but `glht()` does not.

```
> poison.Poison.mc <- glht(poisont.lm,linfct=mcp(Poison="Tukey"))
> poison.Treatment.mc <- glht(poisont.lm,linfct=mcp(Treatment="Tukey"))
> summary(poison.Poison.mc)

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
II - I == 0       7.811      2.906   2.688   0.0272 *
III - I == 0     33.274      2.906  11.451   <0.001 ***
III - II == 0    25.463      2.906   8.763   <0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)

> summary(poison.Poison.mc,test=adjusted(type="none"))

Linear Hypotheses:
               Estimate Std. Error t value Pr(>|t|)
II - I == 0       7.811      2.906   2.688   0.0103 *
III - I == 0     33.274      2.906  11.451 1.69e-14 ***
III - II == 0    25.463      2.906   8.763 4.96e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- none method)
```

```
> summary(poison.Treatment.mc)
Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0  -27.623      3.355  -8.233   <0.001 ***
C - A == 0   -9.536      3.355  -2.842   0.0332 *
D - A == 0  -22.639      3.355  -6.747   <0.001 ***
C - B == 0   18.088      3.355   5.391   <0.001 ***
D - B == 0    4.984      3.355   1.485   0.4551
D - C == 0  -13.103      3.355  -3.905   0.0018 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)

> summary(poison.Treatment.mc,test=adjusted(type="none"))
Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0  -27.623      3.355  -8.233 2.66e-10 ***
C - A == 0   -9.536      3.355  -2.842 0.006893 **
D - A == 0  -22.639      3.355  -6.747 3.35e-08 ***
C - B == 0   18.088      3.355   5.391 2.97e-06 ***
D - B == 0    4.984      3.355   1.485 0.144882
D - C == 0  -13.103      3.355  -3.905 0.000336 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- none method)
```

```
> confint(poison.Poison.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = invsurv ~ Poison + Treatment, data = poison)

Quantile = 2.4293
95% family-wise confidence level


Linear Hypotheses:
               Estimate lwr     upr
II - I == 0     7.8107   0.7514 14.8700
III - I == 0   33.2737  26.2145 40.3330
III - II == 0 25.4631  18.4038 32.5223
```

```
> confint(poison.Treatment.mc)

         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = invsurv ~ Poison + Treatment, data = poison)

Quantile = 2.6753
95% family-wise confidence level


Linear Hypotheses:
            Estimate lwr      upr
B - A == 0 -27.6234 -36.5999 -18.6468
C - A == 0  -9.5356 -18.5122  -0.5590
D - A == 0 -22.6390 -31.6155 -13.6624
C - B == 0  18.0878   9.1112  27.0644
D - B == 0   4.9844  -3.9922  13.9610
D - C == 0 -13.1034 -22.0799  -4.1268
```

Scheffé Tests and Intervals

The examples to date have controlled for a specific number of comparisons; for example, with four treatments, there are six possible comparisons. The F-test in `drop1()` tests the hypothesis that all the group means are the same, which implies that any linear contrast of factor levels has a theoretical value of zero. Tests and intervals can be based on this idea, that we need to be protected from false positives in any (linear) test suggested by the results of an analysis.

Suppose we have a factor with r levels and an effect μ_i associated with each level. This could be coefficients in a regression in which the coefficient for level i is already a comparison between level i and level 1. The assertion that the factor has no effect in either case is the hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_r$. In the coefficient case, $\mu_1 = 0$ so then all the values of $\mu_i = 0$, but in any case, if we have a contrast L, then $L^\top M = 0$, where M is the vector $(\mu_1, \mu_2, \ldots, \mu_r)$. We have the (infinite) collection of contrasts and we want a test/interval such that when the total null hypothesis on the factor is true, then the chance that any test will be significant is less than or equal to α.

The *Scheffé* method uses the estimated value of the contrast and the standard error, but instead using the t-statistic, one uses instead $\hat{C} \pm s_{\hat{C}} \sqrt{(r-1) F_{\alpha;r-1;df}}$ where df is the residual degrees of freedom. So with six types of barley in an experiment with 30 data points, the multipier is $\sqrt{5 F_{.05,5,20}} = 3.68$ instead of $t_{20} = 2.086$. Generally, this level of protection is achieved at too high a cost. If differences are the inferential target, the Tukey HSD is better (with multiplier 2.518). And the protected LSD is defensible, though it does risk nominating differences as significant that are not truly different. Scheffé confidence intervals can be computed for pairwise intervals using the `calpha` argument and for other contrasts can be done "by hand."