**STAT 504** **Solutions to Assignment 5**

1. The model is

$$y_{ijk} \;=\; \mu + \alpha_i + \beta_j + \gamma_{ij}\,, \quad \epsilon_{ijk} \sim NID(0, \sigma^2).$$

   (a) Least squares estimates and interpretations for the quantities below are based the base line restrictions specified by options( contrasts=c("contr.treatment", "contr.ploy")) in the following code:

$$\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0; \qquad 1 \le i \le 4, 1 \le j \le 3.$$

   specified by options( contrasts=c("contr.treatment", "contr.ploy")) in the following code:

```
>  dogs         <- read.table("dogs.dat", col.names=c("Drug","Disease","Y"))
>  dogs$Drug    <- as.factor(dogs$Drug)
>  dogs$Disease <- as.factor(dogs$Disease)
>  options( contrasts=c("contr.treatment", "contr.ploy") )
>
>  lm.out1 <- lm( Y ~ Drug*Disease, data=dogs )
>  lm.out1$coef
   (Intercept)          Drug2          Drug3          Drug4
    29.3333333     -3.3333333    -13.0000000    -15.7333333
      Disease2       Disease3 Drug2:Disease2 Drug3:Disease2
    -1.0833333     -8.9333333      8.5833333    -10.8500000
Drug4:Disease2 Drug2:Disease3 Drug3:Disease3 Drug4:Disease3
     0.3166667      0.9333333      1.1000000      9.5333333
>
>  dummy.coef(lm.out1)
Full coefficients are

(Intercept):        29.33333
Drug:                      1           2           3           4
                    0.000000   -3.333333  -13.000000  -15.733333
Disease:                   1           2           3
                    0.000000   -1.083333   -8.933333
Drug:Disease:            1:1         2:1         3:1         4:1
                   0.0000000   0.0000000   0.0000000   0.0000000
```

```
(Intercept):
Drug:

Disease:

Drug:Disease:          1:2         2:2         3:2         4:2
                 0.0000000   8.5833333 -10.8500000   0.3166667

(Intercept):
Drug:

Disease:

Drug:Disease:          1:3         2:3         3:3         4:3
                 0.0000000   0.9333333   1.1000000   9.5333333
```

(b) Least squares estimates of parameters resulting from using the restrictions to solve the normal equations are:

$$\begin{aligned}
\hat{\mu} &= 29.333 \\
\hat{\alpha}_1 &= 0 \\
\hat{\beta}_3 &= -8.933 \\
\hat{\gamma}_{23} &= 0.933 \\
\hat{\alpha}_2 - \hat{\alpha}_3 &= -3.333333 - (-13) \\
&= 9.667 \\
\hat{\gamma}_{22} - \hat{\gamma}_{23} - \hat{\gamma}_{32} + \hat{\gamma}_{33} &= 8.583333 - 0.9333333 - (-10.85) + 1.1 = 19.6 \\
\hat{\mu} + \hat{\alpha}_2 + \hat{\beta}_3 + \hat{\gamma}_{23} &= 29.33333 + (-3.333333) + (-8.933333) + 0.93333 = 18.0 \\
\hat{\alpha}_2 - \hat{\alpha}_3 + \frac{1}{3}(\hat{\gamma}_{21} + \hat{\gamma}_{22} + \hat{\gamma}_{23} & \\
-\hat{\gamma}_{31} - \hat{\gamma}_{32} - \hat{\gamma}_{33}) &= 9.667 + \frac{1}{3}(0 + 8.583 + 0.933 - 0 + 10.85 - 1.1) \\
&= 16.089
\end{aligned}$$

Note that by the restriction the cell means are represented in terms of the parameters as:

| | Disease 1 | Disease 2 | Disease 3 |
|---|---|---|---|
| Drug 1 | $\mu_{11} = \mu$ | $\mu_{12} = \mu + \beta_2$ | $\mu_{13} = \mu + \beta_3$ |
| Drug 2 | $\mu_{21} = \mu + \alpha_2$ | $\mu_{22} = \mu + \alpha_2 + \beta_2 + \gamma_{22}$ | $\mu_{23} = \mu + \alpha_2 + \beta_3 + \gamma_{23}$ |
| Drug 3 | $\mu_{31} = \mu + \alpha_3$ | $\mu_{32} = \mu + \alpha_3 + \beta_2 + \gamma_{32}$ | $\mu_{33} = \mu + \alpha_3 + \beta_3 + \gamma_{33}$ |
| Drug 4 | $\mu_{41} = \mu + \alpha_4$ | $\mu_{42} = \mu + \alpha_4 + \beta_2 + \gamma_{42}$ | $\mu_{43} = \mu + \alpha_4 + \beta_3 + \gamma_{43}$ |

Then, an interpretation of each quantity for the restricted model is given as follows.

$\mu$ is the mean increase of systolic blood pressure with Disease 1 and Drug 1, i.e., $\mu = E(Y_{11k})$.
Here, the least squares estimate is $\hat{\mu} = \bar{y}_{11\cdot}$.

$\alpha_1$ is restricted to be zero.

$\beta_3$ is the deviation of the mean increase in systolic blood pressure for Disease 3 treated with Drug 1 from that for Disease 3 treated with Drug 1., i.e., $\beta_3 = E(Y_{13k}) - E(Y_{11k})$.
The least squares estimate is $\hat{\beta}_3 = \bar{y}_{13\cdot} - \bar{y}_{11\cdot}$.

$\gamma_{23}$ represents an interaction contrast. It is "the difference in the mean increases in systolic blood pressure for Disease 1 and Disease 3 when Drug 1 is given" minus "the difference in mean increases in systolic blood pressure difference Disease 1 and Disease 3 when Drug 3 is given," i.e., $\gamma_{23} = \gamma_{11} - \gamma_{13} - \gamma_{21} + \gamma_{23} = E(Y_{11k}) - E(Y_{13k}) - E(Y_{21k}) + E(Y_{23k})$
The least squares estimate is $\hat{\gamma}_{23} = \bar{y}_{11\cdot} - \bar{y}_{13\cdot} - \bar{y}_{21\cdot} + \bar{y}_{23\cdot}$.

$\alpha_2 - \alpha_3$ is the difference in mean increases of systolic blood pressure between treating Disease 1 with Drug 2 or Drug 3, i.e., $\alpha_2 - \alpha_3 = E(Y_{21k}) - E(Y_{31k})$.
The least squares estimate is $\hat{\alpha_2} - \hat{\alpha_3} = \bar{y}_{21\cdot} - \bar{y}_{31\cdot}$..

$\gamma_{22} - \gamma_{23}$ $-\gamma_{32} + \gamma_{33}$
is an interaction contrast. It is "the difference in mean blood pressure increases when Drug 2 is used to treat Disease 2 or Disease 3 " minus "the difference in mean blood pressure increases when Drug 3 is used to treat Disease 2 or Disease 3 " i.e., $\gamma_{22} - \gamma_{23}]$
$-\gamma_{32} + \gamma_{33} = E(Y_{22k}) - E(Y_{23k}) - E(Y_{32k}) + E(Y_{33k})$
The least squares estimate is $\hat{\gamma}_{22} - \hat{\gamma}_{23} - \hat{\gamma}_{32} + \hat{\gamma}_{33} = \bar{y}_{22\cdot} - \bar{y}_{23\cdot} - \bar{y}_{32\cdot} + \bar{y}_{33\cdot}$.

$\mu + \alpha_2$ $+\beta_3 + \gamma_{23} = \mu_{23}$
is the mean increase in blood pressure when Drug 2 is used with Disease 3, and the OLS estimator is $\bar{y}_{23\cdot}$.

$\alpha_2 - \alpha_3$ $+\frac{1}{3}(\gamma_{21} + \gamma_{22} + \gamma_{23} - \gamma_{31} - \gamma_{32} - \gamma_{33}) \;=\; \frac{1}{3}\sum_{j=1}^{3}\mu_{2j} - \frac{1}{3}\sum_{j=1}^{3}\mu_{3j}$
is the difference between the mean increase in blood pressure when Disease 2 is treated and the mean increase in blood pressure when Disease 3 is used, averaging across the drugs giving equal weight to each drug.
The OLS estimator is $\frac{1}{3}\sum_{j=1}^{3}\bar{y}_{2j\cdot} - \frac{1}{3}\sum_{j=1}^{3}\bar{y}_{3j\cdot}$.

(c) Since $\mu$, $\alpha_1$, $\beta_3$, $\alpha_2 - \alpha_3$, and $\gamma_{23}$ are not estimable for the unrestricted model, the interpretation of these parameters and the values of the corresponding OLS estimators depend on the particular restrictions placed on the model. On the other hand, $\gamma_{22} - \gamma_{23} - \gamma_{32} + \gamma_{33}$, $\mu + \alpha_2 + \beta_3 + \gamma_{23} = \mu_{23}$, and $\beta_2 - \beta_3 + \frac{1}{3}(\gamma_{12} + \gamma_{22} + \gamma_{32} - \gamma_{13} - \gamma_{23} - \gamma_{33})$ are estimable for the unrestricted model, and their interpretations and the values of the OLS estimators do not depend on the restrictions imposed on non-estimable functions of parameters (or equivalently, they do not depend on which generalized inverse is used to solve the normal equations).

You can directly show that the last three quantities are estimable by showing that they are expectations of linear combinations of observed bread volumes. For example,

$$\begin{aligned} E(\bar{y}_{23\cdot}) &= \mu_{23} = \mu + \alpha_2 + \beta_3 + \gamma_{23} \\ E(\bar{y}_{22\cdot} - \bar{y}_{23\cdot} - \bar{y}_{32\cdot} + \bar{y}_{33\cdot}) &= \mu_{22} - \mu_{23} - \mu_{32} + \mu_{33} = \gamma_{22} - \gamma_{23} - \gamma_{32} + \gamma_{33} \\ E\left(\frac{1}{3}\sum_j \bar{y}_{2j\cdot} - \frac{1}{3}\sum_j \bar{y}_{3j\cdot}\right) &= \frac{1}{3}\sum_j \mu_{2j} - \frac{1}{3}\sum_j \mu_{3j} \\ &= \left(\alpha_2 + \frac{1}{3}\sum_j \gamma_{2j}\right) - \left(\alpha_3 + \frac{1}{3}\sum_j \gamma_{3j}\right) \end{aligned}$$

This cannot be done for any of the first five quantities listed above in the unrestricted model and so they are not estimable. Using **Result 3.9**, this can be shown more formally by expressing each of those quantities in the form $c^T\underline{b}$ and noting that for each of the quantities a vector $\underline{d}$ can be found such that $X\underline{d} = 0$ for the unrestricted model, but $\underline{c^T d} \neq 0$.

$$\mu \rightarrow \underline{d^T} = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\alpha_1 \rightarrow \underline{d^T} = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\beta_3 \rightarrow \underline{d^T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\gamma_{23} \rightarrow \underline{d^T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

$$\alpha_2 - \alpha_3 \rightarrow \underline{d^T} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(d) The ANOVA tables are computed as follows;

```
>  # First compute the correction for the overall mean: R(mu)
>  library(MASS)
>  X       <- model.matrix(lm.out1)                          # model matrix
>  ones    <- X[,1]                                          # vector of ones
>  P1      <- ones%*%ginv(t(ones)%*%ones)%*%t(ones)          # projection matrix
>  y       <- dogs[,"Y"]                                     # responses
>  R.mu    <- t(y) %*% P1 %*% y                              # R(mu)
>  mse     <- deviance(lm.out1)                              # mse
>  F.stat  <- R.mu / mse                                     # F statistics
>  P.value <- 1 - pf(q=F.stat, df1=1, df2=lm.out1$df.resid) # P-value
>
>  data.frame( SS=R.mu, df1=1, df2=lm.out1$df.resid, F=F.stat, Pval=P.value)
        SS df1 df2        F      Pval
1 20259.59   1  46 3.995987 0.0515372
>
>  lm.out1 <- lm( Y ~ Drug*Disease, data=dogs )
```

```
>  summary.aov( lm.out1, ssType=1 )
             Df Sum Sq Mean Sq F value    Pr(>F)
Drug          3 2992.8   997.6  9.0513 8.047e-05 ***
Disease       2  365.7   182.9  1.6591    0.2015
Drug:Disease  6  737.9   123.0  1.1158    0.3680
Residuals    46 5070.0   110.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
>  lm.out2 <- lm(Y~Disease*Drug,data=dogs)
>  summary.aov(lm.out2, ssType=1)
             Df Sum Sq Mean Sq F value    Pr(>F)
Disease       2  419.8   209.9  1.9045    0.1605
Drug          3 2938.7   979.6  8.8877 9.347e-05 ***
Disease:Drug  6  737.9   123.0  1.1158    0.3680
Residuals    46 5070.0   110.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

A useful inference is that the interaction between Drug and Disease is not significant. This does not prove that interaction does not exist, but it suggests that interaction effects may be small enough for the additive model to provide a good approximation. With respect to mean increases in systolic blood pressure, there appear to be substantial differences in drugs, but no large differences among diseases.

(e) We have seen in a previous homework assignment that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-rank(X)} = \chi^2_{46}$. Furthermore, since we are assuming a normal theory Gauss-Markov model (i.e. independent errors and homogeneous variance), then $\overline{Y}_{11.} - \overline{Y}_{31.}$, $\overline{Y}_{12.} - \overline{Y}_{32.}$, and $\overline{Y}_{13.} - \overline{Y}_{33.}$ are mutually independent because each difference is obtained from a completely different set of observations. We also have

$$\overline{Y}_{1j.} - \overline{Y}_{3j.} = \sim N(\mu_{1j} - \mu_{3j}, \sigma^2[{n_{1j}}^{-1} + {n_{3j}}^{-1}]) \quad \text{for} \quad j = 1, 2, 3.$$

Consequently,

$$\begin{bmatrix} \overline{Y}_{11.} - \overline{Y}_{31.} \\ \overline{Y}_{12.} - \overline{Y}_{32.} \\ \overline{Y}_{13.} - \overline{Y}_{33.} \end{bmatrix} \sim N\left( \begin{bmatrix} \mu_{11.} - \mu_{31.} \\ \mu_{12.} - \mu_{32.} \\ \mu_{13.} - \mu_{33.} \end{bmatrix}, \sigma^2 \begin{bmatrix} {n_{11}}^{-1} + {n_{31}}^{-1} & 0 & 0 \\ 0 & {n_{12}}^{-1} + {n_{32}}^{-1} & 0 \\ 0 & 0 & {n_{13}}^{-1} + {n_{33}}^{-1} \end{bmatrix} \right)$$

Now express the numerator of the F-statistic as a quadratic form, i.e.,

$$\frac{1}{\sigma^2}\Sigma_{j=1}^{3}(n_{1j}^{-1}+n_{3j}^{-1})^{-1}(\overline{Y}_{1j.}-\overline{Y}_{3j})^2$$

$$=\begin{bmatrix}\overline{Y}_{11.}-\overline{Y}_{31.}\\ \overline{Y}_{12.}-\overline{Y}_{32.}\\ \overline{Y}_{13.}-\overline{Y}_{33.}\end{bmatrix}^T\begin{bmatrix}{n_{11}}^{-1}+{n_{31}}^{-1} & 0 & 0\\ 0 & {n_{12}}^{-1}+{n_{32}}^{-1} & 0\\ 0 & 0 & {n_{13}}^{-1}+{n_{33}}^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\overline{Y}_{11.}-\overline{Y}_{31.}\\ \overline{Y}_{12.}-\overline{Y}_{32.}\\ \overline{Y}_{13.}-\overline{Y}_{33.}\end{bmatrix}$$

Note that the matrix in the middle of this quadratic form is the inverse of the covariance matrix for $\begin{bmatrix}\overline{Y}_{11.}-\overline{Y}_{31.}\\ \overline{Y}_{12.}-\overline{Y}_{32.}\\ \overline{Y}_{13.}-\overline{Y}_{33.}\end{bmatrix}$. It follows from **Result 4.7** that

$$\frac{1}{\sigma^2}\Sigma_{j=1}^{3}(n_{1j}^{-1}+n_{3j}^{-1})^{-1}(\overline{Y}_{1j.}-\overline{Y}_{3j})^2 \;\sim\; \chi_3^2(\sum_j \frac{[\mu_{1j}-\mu_{3j}]^2}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]})$$

Use

$$\begin{bmatrix}\overline{Y}_{11.}-\overline{Y}_{31.}\\ \overline{Y}_{12.}-\overline{Y}_{32.}\\ \overline{Y}_{13.}-\overline{Y}_{33.}\end{bmatrix}=B\underline{Y}$$

to express the quadratic form in the numerator of the F-statistic as a function of the vector of observations $\underline{Y}$. Then, $\frac{1}{\sigma^2}\Sigma_{j=1}^{3}(n_{1j}^{-1}+n_{3j}^{-1})^{-1}(\overline{Y}_{1j.}-\overline{Y}_{3j})^2$

$$=\underline{Y}^TB^T\begin{bmatrix}{n_{11}}^{-1}+{n_{31}}^{-1} & 0 & 0\\ 0 & {n_{12}}^{-1}+{n_{32}}^{-1} & 0\\ 0 & 0 & {n_{13}}^{-1}+{n_{33}}^{-1}\end{bmatrix}^{-1}B\underline{Y},$$

**Result 4.8** can be used to show that this quadratic form is distributed independently of $S^2$. Consequently, the statistic has an F-distribution with $(3, 46)$ degrees of freedom.

This result can be established in other ways. For example, let $A=(\frac{1}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}])}$ be the $1x1$ matrix in the middle of the following quadratic form:

$$[\overline{Y}_{1j.}-\overline{Y}_{3j.}](\frac{1}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]})[\overline{Y}_{1j.}-\overline{Y}_{3j.}].$$

Now apply **Result 4.7** with $\Sigma=\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]$. Then, $A\Sigma=(\frac{1}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]})\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}])=1$
which is an idempotent $1x1$ matrix. Since both $A$ and $\Sigma$ are symmetric and positive definite, the conditions of **Result 4.7** are satisfied and

$$[\overline{Y}_{1j.}-\overline{Y}_{3j.}](\frac{1}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]})[\overline{Y}_{1j.}-\overline{Y}_{3j.}] \;\sim\; \chi_1^2(\delta_j^2),$$

where $\delta_j^2=\mu_j^TA\mu_j=\frac{[\mu_{1j}-\mu_{3j}]^2}{\sigma^2[{n_{1j}}^{-1}+{n_{3j}}^{-1}]}$. Now apply **Theorem 5.3C** in Rencher to show that the sum of independent quadratic forms with chi-square distributions also has a chi-square distribution, i.e.,

$$\frac{1}{\sigma^2}\sum_{j=1}^{3}(n_{1j}^{-1}+n_{3j}^{-1})^{-1}(\overline{Y}_{1j.}-\overline{Y}_{3j.})^2 \;\sim\; \chi_3^2(\sum_j\delta_j^2)\;.$$

There are other ways to establish this result.

Next show that $S^2$ is independent of the quadratic form in the numerator of the F-statistic. One way to do this is to use **Result 4.8**. Write $\overline{Y}_{1j.} - \overline{Y}_{3j.}$ as a quadratic form in terms of $\underline{y}$. This can be done by noting that (let j=1 for example):

$$\begin{aligned}
\overline{Y}_{11.} - \overline{Y}_{31.} &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \overline{Y}_{11.} \\ \overline{Y}_{12.} \\ \overline{Y}_{13.} \\ \overline{Y}_{21.} \\ \overline{Y}_{22.} \\ \overline{Y}_{23.} \\ \overline{Y}_{31.} \\ \overline{Y}_{32.} \\ \overline{Y}_{33.} \\ \overline{Y}_{41.} \\ \overline{Y}_{42.} \\ \overline{Y}_{43.} \end{bmatrix} \\
&= \underline{a}^T \underline{y}
\end{aligned}$$

where

$$\underline{a}^T = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & \cdots & 0 & 1/3 & 1/3 & 1/3 & 0 & \cdots & 0 \end{bmatrix}$$

Now we can write $[\overline{Y}_{11.} - \overline{Y}_{31.}]^2 = (\underline{a}^T\underline{y})(\underline{a}^T\underline{y}) = \underline{y}^T\underline{a}\underline{a}^T\underline{y} = \underline{y}^T A_1 \underline{y}$. The residual sum of squares is $(n - rank(X))S^2 = \underline{y}^T(I - P_X)\underline{y}$. Then, $A_1\Sigma(I - P_X) = \underline{a}\underline{a}^T(\sigma^2 I)(I - P_X) = \sigma^2\underline{a}\underline{a}^T(I - P_X) = \mathbf{0}$, because $\underline{a}$ is in the space spanned by the columns of the model matrix $X$. Verifying this last equality is easily done (you could use R to numerically verify this, for example, but the details are omitted here to save space). Similarly, we can prove that $[\overline{Y}_{1j.} - \overline{Y}_{3j.}]^2$ is independent of the residual sum of squares for j=1,2,3. Consequently,

$$[\overline{Y}_{1j.} - \overline{Y}_{3j.}](\frac{1}{\sigma^2[{n_{1j}}^{-1} + {n_{3j}}^{-1}]})[\overline{Y}_{1j.} - \overline{Y}_{3j.}]$$

is independent of the residual sum of squares and we know shown that
$F = \frac{\sum_{j=1}^{3}(n_{1j}^{-1} + n_{3j}^{-1})^{-1}(\overline{Y}_{1j.} - \overline{Y}_{3j.})^2}{3S^2} \sim F_{(3,46)}(\sum_j \delta_j^2)$,

Note that the non-centrality parameter is $\delta^2 = \frac{1}{3}\sum_j \left[\mu_j^T A \mu_j\right] = \frac{1}{3}\sum_j \left[\frac{[\mu_{1j} - \mu_{3j}]^2}{\sigma^2[{n_{1j}}^{-1} + {n_{3j}}^{-1}])}\right]$, and this is zero if and only if $\mu_{1j} - \mu_{3j} = 0$, for j=1,2,3. It follows that the null hypothesis is

$$H_o : \mu_{1j} = \mu_{3j} \quad \text{for } j = 1, 2, 3,$$

i.e., there is no difference between the average increases of systolic blood pressure for Drugs 1 and 3 for any of the diseases.

(f) The F-statistic is 5.796863 with (3,46)d.f. and p-value=0.002. This is significant at the .05 level and the null hypothesis is rejected, i.e. for at least one disease, the average blood pressure increases are not equal for Drug 1 and Drug 3. The value for this F-statistic can be obtained by modifying the R code used to compute F-tests for Type III sums of squares.

(g)
```
> # Compute Type III sums of squares and F-tests.
> # Use the library "car" and "MASS"
>
> library(MASS)
> library(car)
> options(contrasts=c("contr.helmert","contr.poly"))
> lm.out3 <- aov(Y~Drug*Disease,data=dogs)
> Anova(lm.out3,type="III")
Anova Table (Type III tests)

Response: Y
              Sum Sq Df  F value    Pr(>F)
(Intercept)  19657.8  1 178.3553 < 2.2e-16 ***
Drug          2851.1  3   8.6226 0.0001194 ***
Disease        371.7  2   1.6863 0.1964555
Drug:Disease   737.9  6   1.1158 0.3680099
Residuals     5070.0 46
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

i. $H_0:\ \mu_{ij} - \mu_{il} - \mu_{kj} + \mu_{kl} = 0$ for all $(i, j)$ and $(k, l)$.
$\Rightarrow$ Either interactions between factor1(Drug) and factor2(Disease) do not exist, or they are small relative to the error variance.

ii. The null hypothesis is

$$H_0 : \frac{1}{3}\sum_{j=1}^{3}\mu_{1j} = \frac{1}{3}\sum_{j=1}^{3}\mu_{2j} = \frac{1}{3}\sum_{j=1}^{3}\mu_{3j} = \frac{1}{3}\sum_{j=1}^{3}\mu_{4j}$$

Or, equivalently,

$$H_0 : \alpha_1+\frac{1}{3}(\gamma_{11}+\gamma_{12}+\gamma_{13}) = \alpha_2+\frac{1}{3}(\gamma_{21}+\gamma_{22}+\gamma_{23}) = \alpha_3+\frac{1}{3}(\gamma_{31}+\gamma_{32}+\gamma_{33}) = \alpha_4+\frac{1}{3}(\gamma_{41}+\gamma_{42}+\gamma_{43})$$

Then, the C matrix for $H_0 : C\underline{\beta} = 0$ may be specified as:

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & -\frac{1}{3} & 0 & \frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 & \frac{1}{3} & -\frac{1}{3} \end{bmatrix}$$

From the ANOVA table, the test statistic is 8.62 with degrees of freedom (3, 46) and p-value 0.0001. This indicates the mean increases in systolic blood pressure, averaging with

equal weights across the Diseases, are not the same for all Drugs. Investigating further, for example, the mean increase in systolic blood pressure is lowest with Drug 3.

iii. The null hypothesis is

$$H_0 : \frac{1}{4}\sum_{i=1}^{4}\mu_{i1} = \frac{1}{4}\sum_{i=1}^{4}\mu_{i2} = \frac{1}{4}\sum_{i=1}^{4}\mu_{i3}.$$

Then, the C matrix for $H_0 : C\underline{\mu} = 0$ may be specified as:

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{bmatrix}$$

From the ANOVA table, the test statistic is 1.686 with degrees of freedom (2, 46) and p-value 0.196. This indicates the mean increases in systolic blood pressure, averaging with equal weights across the Drugs, are nearly the same for all three Diseases (i.e., we cannot reject the null hypothesis).

2. **(a)** ANOVA table:

| source | DF | SS | MS | F | p-value |
|---|---|---|---|---|---|
| method | 2 | 1.9489 | 0.9744 | 20.01 | <0.0001 |
| plants | 3 | 343.31 | 114.43 | 4.89 | <0.05 |
| leaves(plants) | 8 | 186.90 | 23.36 | 479.84 | <0.0001 |
| error | 22 | 1.0711 | 0.04869 | | |

**(b)** Formulas for expectations of mean squares:

| source | expected mean squares |
|---|---|
| plants | $\sigma_e^2 + 3\sigma_\gamma^2 + 9\sigma_\beta^2$ |
| leaves(within plants) | $\sigma_e^2 + 3\sigma_\gamma^2$ |
| method | $\sigma_e^2 +$ (quadratic form involving method effects) |
| error(within leaves) | $\sigma_e^2$ |

**(c)** REML estimates for the variance components:
$\hat{\sigma}_\beta^2 = 10.1193,\ \hat{\sigma}_\gamma^2 = 7.7711,\ \hat{\sigma}_e^2 = 0.04869.$
The largest source of random variation is from plant variation.

**(d)** $\bar{Y}_{3..} = 14.5917$
$E(\bar{Y}_{3..}) = \mu + \alpha_3$
$var(\bar{Y}_{3..}) = 1/12(\sigma_e^2 + \sigma_\gamma^2 + 3\sigma_\beta^2)$
s.d of $\bar{Y}_{3..} = \sqrt{1/12(\hat{\sigma}_e^2 + \hat{\sigma}_\gamma^2 + 3\hat{\sigma}_\beta^2)} = 1.7837$
A 95% CI for $\mu + \alpha_3$ is: $14.5917 \pm t_{3.01,.975}1.7837 = [8.9152, 20.2682]$
where 3.01 is the Cochran-Satterhwaite degrees of freedom.

**(e)** $E(\bar{Y}_{3..} - \bar{Y}_{1..}) = \alpha_3 - \alpha_1$
$\bar{Y}_{3..} - \bar{Y}_{1..} = 14.5917 - 14.0750 = 0.5167.$

$var(\bar{Y}_{3..} - \bar{Y}_{1..}) = \frac{1}{6}\sigma_e^2$
s.d. of $\bar{Y}_{3..} - \bar{Y}_{1..} = \sqrt{\frac{1}{6}\hat{\sigma}_e^2} = 0.0901$

A 95% CI for $\alpha_3 - \alpha_1$ is: $0.5167 \pm t_{22,0.975} * 0.0901 = [0.3299, 0.7035]$

**(f)** The mean acid concentration measurement provided by method C is significantly higher than mean acid concentrations measurement provided by methods A and B( adjusted Tukey p-values are less than 0.0001 in both comparisons), but there is no significant difference between mean acid concentrations provided by methods A and B ( p-value=0.845).

**(g)** $corr(Y_{ijk}, Y_{sjm}) = \frac{cov((Y_{ijk}, Y_{sjm})}{\sqrt{var(Y_{ijk} var Y_s jm}} = \frac{\sigma_\beta^2}{\sigma_e^2+\sigma_\beta^2+\sigma_\gamma^2}$ is estimated as $\hat{\rho} = 10.1193/17.939 = 0.564$

**(h)** $corr(Y_{ijk}, Y_{sjk}) = \frac{cov(Y_{ijk}, Y_{sjk})}{\sqrt{var(Y_{ijk} var(Y_{sjk})}} = \frac{\sigma_\beta^2+\sigma_\gamma^2}{\sigma_e^2+\sigma_b^2 eta+\sigma_\gamma^2}$ is estimated as $\hat{\rho} = 17.890/17.939 = 0.997$

**(i)** The new model is: $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, where $Y_{ij}$ is the observed acid concentration made by the i-th method on a leaf from the j-th plant, $\beta_j \sim NID(0, \sigma_\beta^2)$, $\epsilon_{ij} \sim NID(0, \sigma_\epsilon^2)$, and $\sigma_\epsilon^2 = \sigma_\gamma^2 + \sigma_e^2$. The new experiment would result in less precise comparisons of the three methods, because different methods are applied to different leaves instead of being applied to the same leaf. Variability in the difference between estimated means for any pair of methods now involves variation among leaves in addition to random measurement error. In estimating the difference between the mean acid concentrations determined by two methods, we now have

$$var(\bar{Y}_{i..}) - var(\bar{Y}_{k..}) = \frac{1}{6}\sigma_\epsilon^2 = \frac{1}{6}(\sigma_\gamma^2 + \sigma_e^2)$$

For the original experiment (see part E) we have

$$var(\bar{Y}_{i..}) - var(\bar{Y}_{k..}) = \frac{1}{6}\sigma_e^2$$

The ratio of these two variances is the relative efficiency of the two experiments with respect to the estimation of differences in mean acid concentration determined by two different methods, i.e.

$$\text{efficiency of the new experiment} = \frac{\sigma_\gamma^2 + \sigma_e^2}{\sigma_e^2}$$

Substituting the estimates of the variance components obtained in part C, the estimated efficiency of the new experiment relative to the original experiment is $.04869/(7.7711 + .04869) = .0062265$. Consequently, the number of observations made on each method in the new experiment would have to be about $1/(.0062265)=160$ times greater than the number of observations made in the original experiment to estimate the difference in mean acid concentration by provide by two different methods with about the same accuracy.

3. **(a)** Fixed blocking factors: none
Random blocking factors: autos and drivers
Fixed treatment factors: gas additives
Random treatment factors: none

**(b)** $Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk}$

where $Y_{ijk}$ is the oxide level produced by additive i when the k-th car is driven by the j-th driver, and $\beta_j \sim NID(0, \sigma_\beta^2)$, $\tau_k \sim NID(0, \sigma_\tau^2), \epsilon_{ijk} \sim NID(0, \sigma_e^2)$, and any $\beta_j$, $\tau_k$, $\epsilon_{ijk}$ are mutually independent.

**(c)** ANOVA table:

| source | df | SS | MS | F-value | p-value | E(MS) |
|---|---|---|---|---|---|---|
| additive | 3 | 40 | 13.33 | 5 | 0.0452 | var(error)+Q(additive) |
| auto | 3 | 24 | 8 | 3 | 0.1170 | var(error)+4var(auto) |
| driver | 3 | 216 | 72 | 27 | 0.0007 | var(error)+4var(driver) |
| error | 6 | 16 | 2.67 | | | var(error) |

**(d)** $\hat{\sigma}_\beta^2 = 17.3333, \hat{\sigma}_\tau^2 = 1.3333, \hat{\sigma}_e^2 = 2.6667$

Since it is a balanced design, the REML estimates are equal to method of moments estimators ( check it).

**(e)** $\bar{Y}_{1..} = \frac{1}{4}\sum_{j=1}^{4}\sum_{k=1}^{4} Y_{1jk} = 18$ and $var(\bar{Y}_{1..}) = \frac{1}{4}(\sigma_\beta^2 + \sigma_\tau^2 + \sigma_e^2)$

The stadard error for $\bar{Y}_{1..}$ is: $S_{\bar{Y}_{1..}} = \sqrt{(\hat{\sigma_\beta^2} + \hat{\sigma_\tau^2} + \hat{\sigma_e^2})/4} = \sqrt{\frac{2}{4}MSE + \frac{1}{4}MS_{auto} + \frac{1}{4}MS_{driver}} = 2.3094$

A 95% CI for the mean oxide reduction by method A is:
$18 \pm t_{\nu,.975} * 2.3094 = [11.6792, 24.3208]$
where $\nu$=4.15 by Cochran-Satterthwaite approximation.

**(f)** $\bar{Y}_{1..} - \bar{Y}_{2..} = 18 - 22 = -4$ and $var(\bar{Y}_{1..} - \bar{Y}_{2..}) = \frac{2}{4}\sigma_e^2$

The standard error for$\bar{Y}_{1..} - \bar{Y}_{2..}$ is: $\sqrt{MSE/2} = 1.1547$

A 95% CI for the difference in the mean oxide reductions by methods A and B is:
$-4 \pm t_{6,.975} * 1.1547 = [-6.8255, -1.1745]$

**(g)** HSD shows that method B provides significantly greater oxide reduction than either method A or D. There is no significant difference in mean oxide reduction between methods B and C. Methods A and D can be eliminated from future consideration. Any additional runs should examine the difference between methods B and C.

4. **(a)** primary (or whole plot) experimental units: trays
sub-plot units: pots
treatment factors: levels of fertilizers and moisture
blocking factors: none

**(b)** ANOVA table:

| source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| moisture | 3 | 269.19 | 89.73 | 26.34 | 0.0002 |

| | | | | | |
|---|---|---|---|---|---|
| trays(moisture) | 8 | 27.25 | 3.406 | 4.53 | 0.0019 |
| fertilizer | 3 | 297.05 | 99.018 | 131.65 | <0.0001 |
| moist*fert | 9 | 38.06 | 4.228 | 5.62 | 0.0003 |
| error | 24 | 18.05 | 0.752 | | |

| | expected mean squares |
|---|---|
| moisture | var(error)+4var(tray)+Q(moist,moist*fert) |
| tray | var(error)+4var(tray) |
| fertilizer | var(error)+Q(ferti,moist*ferti) |
| moist*fert | var(error)+Q(moist*ferti) |
| error | var(error) |

**(c)** Method of moments estimates of variance components are: $\hat{\sigma}_e^2 = 0.7521, \hat{\sigma}_\gamma^2 = 0.6635$

**(d)** $\mu + \tau_1$ is not estimable.

$\mu + \alpha_1 + \tau_1 + \delta_{11}$ is estimable since $E(\bar{Y}_{1.1}) = \mu + \alpha_1 + \tau_1 + \delta_{11}$

$\tau_1 - \tau_2$ is not estimable.

$\alpha_1 + \delta_{11} - \alpha_2 - \delta_{21}$ is estimable since $E(\bar{Y}_{1.1} - \bar{Y}_{2.1}) = \alpha_1 + \delta_{11} - \alpha_2 - \delta_{21}$

$\delta_{11} - \delta_{13} - \delta_{21} + \delta 23$ is estimable since $E(\bar{Y}_{1.1} - \bar{Y}_{1.3} - \bar{Y}_{2.1} + \bar{Y}_{2.3} = \delta_{11} - \delta_{13} - \delta_{21} + \delta 23$

$(\alpha_1 + \frac{1}{4}\sum_{k=1}^4 \delta_{1k}) - (\alpha_2 + \frac{1}{4}\sum_{k=1}^4 \delta_{2k})$ is estimable since $E(\bar{Y}_{1..} - \bar{Y}_{2..}) = (\alpha_1 + \frac{1}{4}\sum_{k=1}^4 \delta_{1k}) - (\alpha_2 + \frac{1}{4}\sum_{k=1}^4 \delta_{2k})$

By the output from PROC MIXED, A 95% CI for $\mu+\alpha_1+\tau_1+\delta_{11}$ is : $3.122 \pm t_{19.3,.975} * 0.687 = [1.6856, 4.5584]$

A 95% CI for $\alpha_1 + \delta_{11} - \alpha_2 - \delta_{21}$ is: $-2.8491 \pm t_{19.3,.975} * 0.9715 = [-4.8804, -0.8177]$

A 95% CI for $\delta_{11} - \delta_{13} - \delta_{21} + \delta 23$ is: $2.6673 \pm t_{24,.975} * 1.0014 = [0.6004, 4.7341]$

A 95% CI for $(\alpha_1+\frac{1}{4}\sum_{k=1}^4 \delta_{1k})-(\alpha_2+\frac{1}{4}\sum_{k=1}^4 \delta_{2k})$ is: $-5.0510 \pm t_{8,.975} * 0.7535 = [-6.7885, -3.3135]$

**(e)** The profile plot with moisture level on the horizontal axis shows a quadratic trend for each fertilizer level. There is no indication of existence of interaction. The profile plot with fertilizer level on the horizontal axis shows a linear trend for each fertilizer level. There is no strong evidence of existence of interaction. REML estimates for $\sigma_e^2$ and $\sigma_\gamma^2$: $\hat{\sigma}_e^2 = 1.4407, \hat{\sigma}_\gamma^2 = 0.3968$

| effect | estimate | s.d. | df | t | p-value |
|---|---|---|---|---|---|
| $\beta_0$ | 10.5489 | 0.4567 | 14.6 | 23.10 | ¡0.0001 |
| $\beta_1$ | 0.1294 | 0.02246 | 9 | 5.76 | 0.0003 |
| $\beta_2$ | 1.1066 | 0.07748 | 33 | 14.28 | ¡0.0001 |
| $\beta_3$ | 0.01818 | 0.00693 | 33 | 2.62 | 0.0131 |
| $\beta_4$ | -0.01875 | 0.002512 | 9 | -7.46 | ¡0.0001 |
| $\beta_5$ | 0.04888 | 0.04331 | 33 | 1.13 | 0.2672 |

**(f)** From the output,the estimate of mean weight is 7.38 and a 95% CI is $[6.46, 8.30]$

**(g)** Because time trends and difference in time trends are within tray contrasts and comparisons, we can obtain an approximate F-test of the null hypothesis:

$$H_0\text{: the reduced model is appropriate}$$

by comparing residual sums of squares. The ANOVA table for the model in part E:

| source | df | SS |
|---|---|---|
| model | 14 | 602.06 |
| error | 33 | 47.54 |

We can perform the lack of fit test:
$F = \frac{(SSE(reduced) - SSE(full))/(33-24)}{SSE(full)/24} = \frac{(47.54-18.05))/(33-24)}{18.05/24} = 4.36$ on (9,24) df with p-value=0.002
So we reject the null hypothesis and conclude the reduced model is not appropriate.

This F-test is not entirely appropriate because the estimates of the variance components change for the two models (this invovles the variation among trays in addition to the within tray random variation). It would be better to perform a likelihood ratio test. This can be done with PROC MIXED in SAS bat adding the method=ml option to the PROC MIXED statement. It can be done in S-PLUS by adding the argument method="ML" to the lme( ) function. SAS prodcues the value of -2(log-likelihood) and S-PLUS produces the value of the log-likelihood simultaneously fitting both the fixed effects and the variance components. The results are:

| model | log-likelihood | -2(log-likelihood) |
|---|---|---|
| general effects | $-60.60$ | 121.2 |
| quadratice surface | $-78.04$ | 156.1 |

The value of the chi-square test of the null hypothesis that the quadratic surface fits the data as well as the general effects model is

$$\chi^2 = 156.1 - 121.2 = 34.9 \text{ with 10 df and p-value¡.0001}$$

The proposed quadratic surface is not adequate, look for a better model.