Topics for these notes:

- *Hypothesis tests and estimation in the general linear model.*
  - *t-tests*
  - *Generalized likelihood ratio F-tests*
  - *Main effect tests*
  - *Using SAS and R for custom tests and estimates*
  - *Practice Quiz*

Associated reading: *Section 5 in 'General linear models' course notes.*

5 Tests of linear hypotheses

5.1 t-tests

Using methodology presented in the Section 3, we can construct a *t*-test for H_0: $\mathbf{L\beta} = 0$ using

$$t = \mathbf{L}\hat{\boldsymbol{\beta}} / SE(\mathbf{L}\hat{\boldsymbol{\beta}})$$

which has a *t*-distribution with $n-k$ degrees of freedom under the null hypothesis.

- Note that $\mathbf{L\beta}$ is a scalar.
- Tests can be carried out in SAS using the ESTIMATE statement. Along with the *t*-test, the estimate $\mathbf{L}\hat{\boldsymbol{\beta}}$ is given along with its standard error.

Example 1: for the Myostatin data in the one-way effects model, carry out tests for (a) $H_0: \mu + \kappa_1 = 0$ and (b) $H_0: \kappa_1 - \kappa_2 = 0$. Results are below. See if you can reproduce them.

(a) $t = 6.96 / 0.3796 = 18.34$; $p<0.0001$.

(b) $t = 1.511 / 0.537 = 2.82$; $p=0.01$.

For (a), we would clearly reject H_0 and conclude that the average protein level for the Control group at 24 hours is non-zero (but the test is probably not of real interest). For (b), we would conclude that for the Control group, there is protein degradation between 24 and 48 hours, on average. (Strictly speaking, the test lets us conclude that there is a change in mean protein degradation.)

5.2 Generalized likelihood ratio F-tests

- Test statistic: $W = \dfrac{[SSE_{red} - SSE_{full}]/s}{SSE_{full}/(n-k)} = \dfrac{[\mathbf{Y}^t(\mathbf{P}_{full} - \mathbf{P}_{red})\mathbf{Y})]/s}{[\mathbf{Y}^t(\mathbf{I} - \mathbf{P}_{full})\mathbf{Y}]/(n-k)} \sim F_{s,n-k}$ under H_0.

- Notes: $k = r(\mathbf{X})$, $s = r(\mathbf{X}) - r(\mathbf{X}_{red})$; the denominator of W is the MSE; $\mathbf{P}_{full} = \mathbf{P}_{\mathbf{X}}$; *SSE*=residual sum of squares; *red*=reduced model; *full*=full model.

- Three approaches to carrying out the test:

  (1) Employ PROC GLM (SAS) or LM function (R) directly.
  (2) Fit full and reduced models separately with PROC GLM / LM function and obtain the RSS quantities to calculate *W*.
  (3) Work with projection matrices using PROC IML (or R).

- In SAS, we can conduct generalized likelihood ratio *F*-tests using the CONTRAST statement. In R, it can be carried out using the *glh.test* function that is applied to a *glm* object; the function is available via the *gmodels* package.

Example 2: Consider the Myostatin data in the 2-way model (group and time as class variables). Question: do we need the interaction term?

The null hypothesis: $H_0 : \gamma_{ij} = 0 \quad \forall\, i, j$

The projection matrices:

$\mathbf{P}_{full} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-}\mathbf{X}^t$ where $\mathbf{X}$ is defined for the Myostatin data in the 2-way effects model (including interaction).

$\mathbf{P}_{red} = \mathbf{X}_{red}(\mathbf{X}_{red}^t\mathbf{X}_{red})^{-}\mathbf{X}_{red}^t$ where $\mathbf{X}_{red}$ is same as $\mathbf{X}$ without last 6 columns.

- The full model has *group*, *time* and *group*time* as predictors, and the reduced model has just *group* and *time*. The SSE for the full and reduced models are 10.375 and 11.316, respectively; $s = r\left(\mathbf{X}_{full}\right) - r\left(\mathbf{X}_{red}\right) = 2$ (the number of degrees of freedom for the interaction), and $k = r(\mathbf{X}_{full}) = 6$.
- Thus, W = {[11.316 – 10.375] / 2} / {10.375 / (24 – 6)} = 0.82. This matches the F-statistic generated by the CONTRAST statement. The other approaches also yield W=0.82 (p=0.45). Based on the test, you could argue to drop the interaction term.

Example 3: Another test of interest may be any effect associated with time (including time or group*time). For practice: write the CONTRAST statement for this in SAS, and then compare with hand-calculated *W* statistic. The contrast can be written as follows using the means model:

```
CONTRAST 'time and group*time' group*time  1  0 -1 1  0  -1,
                               group*time  1 -1  0  1  -1 0,
                               group*time  1  0 -1 -1  0  1,
                               group*time  1 -1  0  -1  1 0;
```

This yields W=8.60, p=0.0005. Approaches 2 and 3 also yield W=8.60. This indicates that there is some effect of time (either in the main effect, interaction, or both).

For practice: try approach (1) and (3) using the 2-way effects model; results should be the same.

5.3 Main effect tests, interaction tests and more detail on CONTRAST and ESTIMATE statements

- We have discussed theory for tests of the form H_0: $\mathbf{C\beta} = \mathbf{h}$. The question is: What form of **C** and **h** are associated with tests of interest? This will depend on the data at hand, and the specific hypotheses that the researcher is interested in testing. Usually we use **h**=**0**.

- To illustrate the forms of **C**, consider the main effect tests for group and time, and the test for interaction, using the means model for the Myostatin data. Although these tests are output directly without having to specify CONTRAST or ESTIMATE statements, we will discuss how to obtain them with the latter to better understand both these particular tests as well as the statements.

- In the strict sense, a CONTRAST is a linear combination of beta elements, $\mathbf{c}_i^t\boldsymbol{\beta}$, such that $\sum \mathbf{c}_i^t = 0$. If all rows of **C** have this property, then **Cβ** is a set of contrasts, which are generally estimable. When we estimate **Lβ** using the ESTIMATE statement, elements of **L** are not constrained to sum to 0. However, if **Lβ** is not estimable for the particular **L** that you specify, then SAS will tell you that. [The **C** matrix may often be defined to have row contrasts, but generally in my notes it will not be forced to have such constraints.]

Notation for the means model

		Time	
Trt Group	μ_{11}	μ_{12}	μ_{13}
	μ_{21}	μ_{22}	μ_{23}

For this means model, $\boldsymbol{\beta} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})^t$.

Main effect test for group

The main effect test for group tests for differences in marginal means for groups. For the application above, the test can be written as H_0: $\mathbf{C\beta} = 0$, where C = (1/3 1/3 1/3 -1/3 -1/3 -1/3). This also reduces to H_0: $\bar{\mu}_{1\bullet} = \bar{\mu}_{2\bullet}$. We can add the following statements that will yield the same test results:

```
CONTRAST  'group factor' group*time 1  1  1  -1  -1 -1;
ESTIMATE  'group factor' group*time 1  1  1  -1  -1 -1 / divisor=3;
```

Notes:

- The CONTRAST statement produces an F-test, while the ESTIMATE statement produces a t-test and an estimate corresponding to the coefficients specified. In many cases these will produce the same results (if the data and coefficients are the same). Later we will discuss strengths and limitations of each approach.

- For either the CONTRAST or ESTIMATE approach, the test will be the same if the coefficients are all scaled by the same amount, since the scalar will cancel out in the test statistic numerator and denominator. (Rescaling will change the estimate but will not change the either the *t*-test invoked by the ESTIMATE statement or the *F*-test invoked by the CONTRAST statement.)

- The divisor is often used to simplify the code, but becomes important if numbers have unending decimals (e.g., 0.333…). The CONTRAST does not have the divisor option as it is not important – it is really only necessarily for the estimate in the ESTIMATE statement, not the test, due to the previous point.

- Generally, **C** will have a different form and dimensions depending on the model used (means model, one-way effects model, two-way effects model).

Main effect test for time

The main effect test for time tests for differences in marginal means for time. In this case there are 3 times, and thus the test is $H_o: \bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$. In this case we will need 2 rows in the **C** matrix for the test (one line for each equation in the hypothesis). There are different possibilities but one is:

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

The first line of $\mathbf{C\beta=0}$ is $\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22} = 0$, or $\frac{1}{2}\mu_{11} + \frac{1}{2}\mu_{21} = \frac{1}{2}\mu_{12} + \frac{1}{2}\mu_{21}$, or more simply, $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2}$. Similarly, the second line is $\bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 3}$. (Note that $\bar{\mu}_{\bullet 2} = \bar{\mu}_{\bullet 3}$ is implied through the other equalities.) We can carry out the test with the following statement in SAS:

```
CONTRAST  'time factor' group*time 1  -1   0  1  -1  0,
                        group*time 1   0  -1  1   0  -1;
```

Here is the same test using R, with actual output:

```
library(gmodels)
glm2<-glm(y~gt-1,data=myostatin) #Note! No-intercept model
summary(glm2)
  #F-test for Time
    C<-rbind(c(1,0,-1,1,0,-1),c(1,-1,0,1,-1,0))
    mycontrast<-glh.test(glm2,C)
    summary(mycontrast)

C matrix:
     gtc24 gtc48 gtc72 gtm24 gtm48 gtm72
[1,]     1     0    -1     1     0    -1
[2,]     1    -1     0     1    -1     0

C %*% Beta-hat:
[1] 4.34125 2.32825

F = 16.3782, df1 =  2, df2 = 18, p-value = 8.872e-05
```

Notes:

- Since there are 2 d.f., the test cannot be carried out using the ESTIMATE statement. Other forms of **C** may yield the same test. This can be explained by the *Full Rank Reparameterization Theorem* which states that

$$SS\left(\underset{q\times p}{\mathbf{C}}\ \underset{p\times 1}{\hat{\boldsymbol{\beta}}}\right) = SS\left(\underset{q\times q}{\mathbf{D}}\ \underset{q\times p}{\mathbf{C}}\ \underset{p\times 1}{\hat{\boldsymbol{\beta}}}\right) \text{ for any nonsingular } \underset{q\times q}{\mathbf{D}}.$$

- Key note: CONTRAST statements may have multiple rows, but ESTIMATE statements are restricted to one row.

Time*group interaction

Does the difference between group means depend on time? If so, then there is interaction. Similarly, you can ask the question whether differences over time are similar between groups, the test will be the same.

If the differences between group means is in fact the same at each time, then there is no interaction and this would comprise the null hypothesis:
$H_0: \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$. The **C** matrix associated with this hypothesis is:

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

As with the main effect tests, the interaction test will be part of the default output. The test can also be carried out with the following added statement:

```
CONTRAST 'interact' group*time 1 -1 0 -1 1 0, group*time 1 0 -1 -1 0 1;
```

Again, d.f.>1 so cannot get the test with an ESTIMATE statement.

Orthogonality of contrasts

You will notice that for the **C** matrices above, pairs of rows are *orthogonal* (any pair, considering all 5 rows), meaning that for row vectors **c** and **d**, $\sum_j c_j d_j = 0$, where j denotes the j^{th} element of **c** or **d**. As a consequence, Type III sums of squares for the 3 factors will add up nicely to the total sum of squares, for these data since sample sizes are equal across treatments (or cells). Independence of tests also follows from orthogonality of the contrasts.

Full SAS code to carry out all the tests described above, for the means model:

```
proc glm; class group time;
model size = group*time / XPX inverse E;
contrast 'group'       group*time 1 1 1 -1 -1 -1;
contrast 'time'        group*time 1 0 -1 1 0 -1,
                       group*time  1 -1  0  1 -1  0 ;
contrast 'interaction' group*time  1  0 -1 -1  0  1,
                       group*time  1 -1  0  -1  1 0 ; run;
```

Writing ESTIMATE and CONTRAST-type statements in R, using the means model:

Here is how to perform a CONTRAST-type statement in R. This output is the same as shown previously under main effect for time:

Code	Streamlined output

`#means model`
`  glm2<-glm(y~gt-1,data=myostatin)`
`  summary(glm2)`
`#F-test for Time`
`  C<-rbind(c(1,0,-1,1,0,-1),`
`  c(1,-1,0,1,-1,0))`
`  mycontrast<-glh.test(glm2,C)`
`  summary(mycontrast)`
	`C %*% Beta-hat:`
	`[1] 4.34125 2.32825`
	
	`F = 16.3782, df1 = 2, df2 = 18,`
	`p-value = 8.872e-05`

Here is how to perform an ESTIMATE-type statement in R (based on the *glm2* object previously defined). One difference, relative to SAS, is that a z-approximation is used, instead of the t-distribution. In this case, results are not much different for the 2 approaches:

Code	Streamlined output

`#Estimate for time 1 to time 3, and time 1 versus time 2; these are like ESTIMATE statements.`
 	
`t <- glht(glm2, linfct = C)`
`summary(t)`
	`Simultaneous Tests for General Linear Hypotheses`
	`Fit: glm(formula = y ~ gt - 1, data = myostatin)`
	
	`Linear Hypotheses:`
	`Estimate Std. Error z value Pr(>|z|)`
	`1 == 0 4.3412 0.7592 5.718 2.15e-08 ***`
	`2 == 0 2.3282 0.7592 3.067 0.00421 **`
	`---`
	`Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1`
	`(Adjusted p values reported -- single-step method)`

Writing tests in SAS, in terms of the 2-way ANOVA model:

We can also construct ESTIMATE and CONTRASTS statements for the two-way effects model. Recall that this model is $Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$, where *i* denotes group, *j* denotes time, and *k* denotes replicate. Referring back to the less-than-full rank model, there were 12 parameters and **β** had the following form:

$$\boldsymbol{\beta}^t = \begin{pmatrix} \mu & \alpha_1 & \alpha_2 & \tau_1 & \tau_2 & \tau_3 & \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix}$$

Thus, **L** will be a 1x12 vector. When we write ESTIMATE or CONTRAST statements in SAS, the coefficients in L are broken down by factors. For example, say we want to estimate the mean for the Control group at 72 hours. For this, **L**=(1 1 0 0 0 1 0 0 1 0 0 0). However, in SAS we would write:

```
ESTIMATE 'C at 72 hrs' intercept 1 group 1 0 time 0 0 1
    group*time 0 0 1 0 0 0;
```

If certain factors do not come into play, then we do not need to include them in the ESTIMATE or CONTRAST statement. For example, say we want to compare means for treatment groups at 24 hours. The entire **L** would be (0 1 -1 0 0 0 1 0 0 -1 0 0). In SAS, the estimate statement is:

```
ESTIMATE 'group diffs at 24 hrs' intercept 0 group 1 -1 time 0 0 0
    group*time 1 0 0 -1 0 0;
```

But since we have 0's for two of the factors, we can just write:

```
ESTIMATE 'group diffs at 24 hrs' group 1 -1 group*time 1 0 0 -1 0 0;
```

Note: in order to figure out how to write L above, you might find it easier to consider each group first, and then take the difference:

	Int	Group		Time			Group*Time				
Control group at 24 hours:	**L**=(1	1	0	1	0	0	1	0	0	0	0	0)
Myostatin group at 24 hours:	**L**=(1	0	1	1	0	0	0	0	0	1	0	0)
Difference:	**L**=(0	1	-1	0	0	0	1	0	0	-1	0	0)

This shows us that intercept and time factors cancel out in the estimate. You can also use the CONTRAST statement to run the test, but there is really no advantage to doing that here; the ESTIMATE statement will run the same test as well as estimate the quantity of interest.

For practice: write out the null hypothesis for the main effect tests and interaction test based on the 2-way model.

Quiz: CONTRAST and ESTIMATE statements

Consider the GLM: $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ (Case I). We are interested in estimating $\mathbf{L}\boldsymbol{\beta}$ or testing H_0: $\mathbf{L}\boldsymbol{\beta}=0$ (vs. not equal) for some row vector $\mathbf{L}$. [An easy way to do this is to use the ESTIMATE statement in PROC GLM.] For the specific model $Y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}$, say that we have an experiment where i=1,…,3, j=1,…,3 and k=1,…, 5 (subjects per treatment).

(1) For the 'less than full rank' parameterization above, what is p (number of columns in $\mathbf{X}$)?

(2) What is the rank of $\mathbf{X}$?

(3) Write $\boldsymbol{\beta}^t$

(4) Is $\alpha_1 - \alpha_2$ estimable? Consider $\mathbf{LH}$ with form:

(L_1 | L_2 L_3 L_1–L_2–L_3 | L_5 L_6 L_1–L_5–L_6).

(5) Write the SAS PROC GLM code to fit the model and estimate $\alpha_1 - \alpha_2$ (if estimable).

(6) For a given $n \times p$ matrix $\mathbf{X}$ with $r(\mathbf{X}) < p$, answer the following for each matrix quantity. Consider $\mathbf{L}\boldsymbol{\beta}$ that is estimable. For starters, we know that the M-P inverse $\mathbf{X}^+$ has dimension $p \times n$ and it is unique; $\mathbf{X}^-$ is a conditional inverse of $\mathbf{X}$, also is $p \times n$, and is not unique. We also know that $\mathbf{P}_\mathbf{X} = \mathbf{X}\mathbf{X}^+$ is symmetric, idempotent, and invariant to choice of $(\mathbf{X}^t\mathbf{X})^-$.

Matrix quantity	Another name? (Or its importance?)	Dimension	Invariant to choice of $(\mathbf{X}^t\mathbf{X})^-$?	Other properties
$(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t$				
$(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y}$				
$\mathbf{L}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{Y}$				
$\mathbf{Y}^t(\mathbf{I} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t)\mathbf{Y}$				
$\mathbf{L}(\mathbf{X}^t\mathbf{X})^-\mathbf{L}^t$				