Topics for these notes:

- *Basic review of the GLM – models and assumptions*
- *Least squares – calculus, geometry and linear algebra*
- *Forms of statistical models*
  - *Effects vs. means models*
  - *Full-rank versus less-than-full-rank models*

Associated reading: Sections 1 through 3 in 'General linear models' course notes.

1 Initial notes and thoughts

- These notes highlight some of the key results in general linear models (GLM) theory. In some places, 'general linear models' are just referred to as 'linear models'. Many of the theoretical methods will be illustrated via data from a factorial experiment (the Myostatin application), introduced previously. For more complete references on general linear models, see:
  - Graybill, Franklin A. (2000). *Theory and Application of the Linear Model*.
  - Graybill, Franklin A. (2001). *Matrices with Applications in Statistics*.
  - McCulloch, Charles E. & Searle, Shayle R. (2001). *Generalized, Linear, and Mixed Models*.
  - Neter, Wasserman, Kutner, & Nachtsheim. (1996). *Applied Linear Statistical Models*.
  - Schott, James R. (2005). *Matrix Analysis for Statistics*.

- An important note: many of the concepts discussed in this chapter will also apply to longitudinal models. For example, modeling time as a class versus continuous variable, estimability, full-rank versus less-than-full-rank models are all concepts that apply to linear mixed models, which we will see later on.

- What sort of methods are associated with the general linear model?

  - *One-sample t-test*
  - *Two sample t-test (equal variance)*
  - *Simple linear and multiple regression*
  - *ANOVA*
  - *ANCOVA*

2 The Myostatin data

- 2×3 factorial treatment structure in completely randomized design.
- 2 levels of treatment: myostatin Y or N; called ‘group’ variable
- 3 levels of time: 24, 48 and 72 hours.
- Total of 24 muscle cell samples (4 replicates for each treatment).
- Outcome variable: measure of protein in the sample for the given condition (time and treatment).
  - Hypothesized that myostatin samples would have greater protein degradation than controls.
- Each sample randomly assigned to a treatment and time.
  - In the CRD, how many possible allocations of samples to treatment combinations are there? ANSWER = _____.
- The general linear model can be used to carry out the ANOVA for this experiment.

- Table for population mean leucine protein levels for group*time combinations.

Time

Group	24h	48h	72h	
C	μ_{11}	μ_{12}	μ_{13}	$\bar{\mu}_{1\bullet}$
M	μ_{21}	μ_{22}	μ_{23}	$\bar{\mu}_{2\bullet}$
	$\bar{\mu}_{\bullet 1}$	$\bar{\mu}_{\bullet 2}$	$\bar{\mu}_{\bullet 3}$	

- Write hypotheses for the following tests:
  - (i) some difference in means — H0: uij = ui'j' for all i,i',j,j'
  - (ii) main effect of Time — H0: u.1=u.2=u.3 df=2
  - (iii) main effect of Myostatin — H0: u1.=u2. df=1
  - (iv) Time×Myostatin interaction. — H0: u11-u12=u21-u22 Difference in Control and Myostatin at 24 vs 48
    H0: u12-u13=u22-u23 Difference in Control and Myostatin at 48 vs 72
    DF = 2

- Use the results from the partial ANOVA table below to make conclusions about the hypotheses:

Test	df	F	p-value
Overall (Model)	5	8.02	0.0004
Time	2	16.38	<0.0001
Myostatin	1	5.74	0.0277
Time×Myostatin	2	0.82	0.4577

- An alternative design: What if, instead of using independent samples at each time point, the same samples were measured across time points? What is the problem with using 2-way ANOVA in this case?

SAS program and output summary:

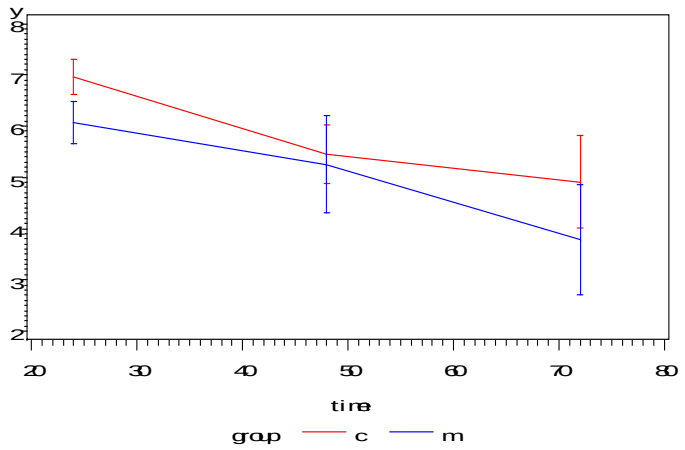
```
data myostatin;
input leucine group $ time @@;
y=leucine/1000; cards;
6568 c 24 6802 c 24 7198 c 24 7280 c 24
4992 c 48 5242 c 48 5285 c 48 6284 c 48
4092 c 72 4331 c 72 5135 c 72 6087 c 72
5516 m 24 6023 m 24 6334 m 24 6400 m 24
4512 m 48 4706 m 48 5175 m 48 6612 m 48
3076 m 72 3209 m 72 3462 m 72 5364 m 72
;
proc means data=myostatin noprint;
by group time; var y;
output out=myo_out mean=my stddev=sy n=ny;
run;
proc print data=myo_out;
var group time my sy ny; run;

proc gplot data=myostatin;
plot y*time=group / vaxis= 2 to 8;
symbol1 i=stdlmtj mode=include c=red;
symbol2 i=stdlmtj mode=include c=blue; run;
```

The output from PROC MEANS:

group	time	my	sy	ny
c	24	6.962	0.335	4
c	48	5.451	0.570	4
c	72	4.911	0.902	4
m	24	6.068	0.403	4
m	48	5.251	0.949	4
m	72	3.778	1.070	4

Graph from PROC GPLOT:



```
proc glm data=myostatin; class group time;
model y = group|time / solution; run;
```

```
The GLM Procedure

Dependent Variable: y

                                   Sum of
Source                     DF       Squares     Mean Square    F Value    Pr > F
Model                       5     23.12640221      4.62528044       8.02    0.0004
Error                      18     10.37454375      0.57636354
Corrected Total            23     33.50094596

      R-Square     Coeff Var      Root MSE        y Mean
      0.690321      14.04979      0.759186      5.403542

Source              DF     Type III SS    Mean Square    F Value    Pr > F
group                1      3.30561037      3.30561037       5.74    0.0277
time                 2     18.87957908      9.43978954      16.38    <.0001
group*time           2      0.94121275      0.47060637       0.82    0.4577
```

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		3.777750000	B	0.37959305	9.95	<.0001
group	c	1.133500000	B	0.53682564	2.11	0.0490
group	m	0.000000000	B	.	.	.
time	24	2.290500000	B	0.53682564	4.27	0.0005
time	48	1.473500000	B	0.53682564	2.74	0.0133
time	72	0.000000000	B	.	.	.
group*time	c 24	-0.239750000	B	0.75918610	-0.32	0.7558
group*time	c 48	-0.934000000	B	0.75918610	-1.23	0.2344
group*time	c 72	0.000000000	B	.	.	.
group*time	m 24	0.000000000	B	.	.	.
group*time	m 48	0.000000000	B	.	.	.
group*time	m 72	0.000000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

First impression may be that the fit is not correct but that is not the case.
Just need to use/interpret the model estimates correctly.

- The NOTE above follows since estimates of **β** elements are not unique. In particular, the highest levels of each factor were set as reference groups (along with levels of interactions involving highest levels of group or time). If different levels were used as reference groups, all of the estimates would be different. This issue will be discussed in more detail in further sections.

- Below is the analysis using R. Note that the estimates of elements of **β** are different than in the SAS analysis since R uses different reference groups, by default. Specifically, when using the 'factor' function, the first level of each factor is used as the reference group, rather than the last.

```
#read in data and name variables
myostatin <- read.csv("c:/strand_folders/teaching/longitudinal applications and simulation
programs/myostatin data/myostatin.csv")
myostatin$y=myostatin$leucine/1000;
#create a factored variable for a cell means model of all 6 levels
myostatin$gt<-factor(paste(myostatin$group,myostatin$time,sep=""))

#2-way effects model
class_fit1=lm(y ~ factor(group) + factor(time) + factor(group)*factor(time),data=myostatin)
summary(class_fit1)

> summary(class_fit1)
Call:
lm(formula = y ~ factor(group) + factor(time) + factor(group) * factor(time),
data = myostatin)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8193 -0.5470 -0.1629  0.2788  1.5862
```

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    6.9620     0.3796  18.341 4.27e-13 ***
factor(group)m                -0.8938     0.5368  -1.665  0.11325
factor(time)48                -1.5113     0.5368  -2.815  0.01146 *
factor(time)72                -2.0508     0.5368  -3.820  0.00125 **
factor(group)m:factor(time)48  0.6943     0.7592   0.914  0.37256
factor(group)m:factor(time)72 -0.2397     0.7592  -0.316  0.75579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7592 on 18 degrees of freedom
Multiple R-squared: 0.6903,     Adjusted R-squared: 0.6043
F-statistic: 8.025 on 5 and 18 DF,  p-value: 0.0003960
```

3 Basics of the general linear model and theory toolbox

3.1 General form of the model

We will look more at the mechanics of how the GLM estimates are computed. But first, we will spend a little time reviewing some important theory. For a review of matrix theory, see, for example, Chapter 1 of Graybill (2000). For the following, bolded font is used for matrices and vectors.

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

Case I: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$ Note here that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$

Case II : $\underset{n\times 1}{\boldsymbol{\varepsilon}}$ has an unspecified distribution; $E(\boldsymbol{\varepsilon})=\mathbf{0}$ and $Cov(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$

Difference between independence and uncorrelated. Independence implies uncorrelation. Something that is uncorrelated does not necessarily mean it is independent.

3.2 The least squares estimator

3.2.1 Calculus

- When we fit data to the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we must find a solution to $\boldsymbol{\beta}$ that is optimal in some sense.

- Once approach is to choose $\boldsymbol{\beta}$ that minimizes $\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Any form of $\boldsymbol{\beta}$ that satisfies $\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$ will meet this criterion, and these are often called the *normal equations*.

- The steps to carry this out are:
  - Expand the quantity $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$, resulting in $\mathbf{Y}^t\mathbf{Y}-2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{Y}+\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$.
  - Take the derivative of this expanded quantity with respect to $\boldsymbol{\beta}$ and set the new quantity to 0: $0-2\mathbf{X}^t\mathbf{Y}+2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}\equiv 0$. (We can show this change point is a minimum for the function with respect to $\boldsymbol{\beta}$ by examining the 2nd derivative of the quantity.)
  - Rework the equality to get the normal equations: $\mathbf{X}^t\mathbf{Y}=\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$.
  - If the inverse of $\mathbf{X}^t\mathbf{X}$ exists, then the solution is determined as $\hat{\boldsymbol{\beta}}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$.
  - If $\mathbf{X}$ (and hence $\mathbf{X}^t\mathbf{X}$) is not of full rank, then the regular inverse does not exist; a generalized or conditional inverse must be computed. These issues will be described more forthcoming.

X is n*p matrix. Rank of X equals to number of linearly independent columns. X is full ranked if the rank of X is equal to the minimum of n and p (which is usually p if sample size is larger than p)

3.2.2 A geometrical view, projection matrices

- Consider the simple linear regression model:

  Subject form: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, i=1,…, n

  Matrix form: $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}}\ \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$, $\qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

- Say we observe (x,y) = {(1,1), (2,2), (3,2)}.
  Write **X** and **y**:

$$\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \beta_1 + \boldsymbol{\varepsilon}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

- Determine the rank of **X** [denoted as $r(\mathbf{X})$].

  We know $r(\mathbf{X}) \leq 2$.

  Note that $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ are linearly independent.

  Thus $r(\mathbf{X}) = 2$. A regular inverse exists.

- This tells us that the column space of **X** spans 2 dimensions. (The column space contains all linear combinations of the columns of **X**.)

- Question: is **Y** in this column space?

No, if it was then the graph would show straight line.

- We want a solution that is in the column space of $\mathbf{X}$, $C(\mathbf{X})$, but as close to $\mathbf{y}$ as possible. Let's project $\mathbf{y}$ onto $C(\mathbf{X})$. This is the least squares solution.

- When $r(\mathbf{X})=p$ (as above), we also have that $r(\mathbf{X}^t\mathbf{X})=p$, so that the inverse of $(\mathbf{X}^t\mathbf{X})$ exists. Thus, $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} = \mathbf{P}_\mathbf{X}\mathbf{y}$. $\mathbf{P}_\mathbf{X}$ denotes the *projection matrix*, where $\mathbf{y}$ is projected onto the column space of $\mathbf{X}$. The projected vector will be in $C(\mathbf{X})$, but as close to $\mathbf{y}$ as possible, in terms of least squares.

Same as Hat matrix

- Find the projection matrix for this example. $\mathbf{X}^t\mathbf{X} = \begin{pmatrix}1 & 1 & 1\\ 1 & 2 & 3\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & 2\\ 1 & 3\end{pmatrix} = \begin{pmatrix}3 & 6\\ 6 & 14\end{pmatrix}$

$$\Rightarrow \left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \frac{1}{6}\begin{pmatrix}14 & -6\\ -6 & 3\end{pmatrix} \qquad \Rightarrow \mathbf{P}_\mathbf{X} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t = \frac{1}{6}\begin{pmatrix}5 & 2 & -1\\ 2 & 2 & 2\\ -1 & 2 & 5\end{pmatrix}$$

- Determine $\mathbf{P_X y}$.

$$\mathbf{P_X y} = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 7 \\ 10 \\ 13 \end{pmatrix} = \begin{pmatrix} 1\ 1/6 \\ 1\ 2/3 \\ 2\ 1/6 \end{pmatrix}$$

- Determine $\hat{\boldsymbol{\beta}}$.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} = \frac{1}{6}\begin{pmatrix} 8 & 2 & -4 \\ -3 & 0 & 3 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 2/3 \\ 1/2 \end{pmatrix} \qquad \text{I.e., } \hat{y} = \frac{2}{3} + \frac{1}{2}x$$

- Check predicted values: e.g., for x=1, $\hat{y} = \frac{2}{3} + \frac{1}{2}(1) = 1\frac{1}{6}$. Can check others.
  Note: if **y** is in the column space of **X**, then $\mathbf{P_X y} = \mathbf{y}$ (not the case here).

3.3 Specific forms of the model: a look at the Myostatin data

- The general form for the general linear model can be written specifically for a given application, and there are alternative specific forms we can use. For the Myostatin application, there are 3 relevant forms that we will discuss, considering time and group as class variables: the two-way effects model, the one-way effects model, or the means model.

		SAS PROC GLM statement:
Two-way effects model: $Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \varepsilon_{ijk}$	group i time j replicate k	MODEL y=group|time;
One-way effects model: $Y_{ij} = \mu + \kappa_i + \varepsilon_{ij}$	group×time i replicate j	MODEL y=group*time;
Means model: $Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$	group i time j replicate k	MODEL y=group*time / noint;

Notes:

- The previous analysis in Section 2 was for the two-way effects model.
- We will consider all of these models in this GLM review.
- The parameters above are generic; you can focus on the subscript indices to help determine what they represent.
- There is no ‘right’ or ‘wrong’ model parameterization. A certain approach may make it easier or harder to get certain results of interest out of the model. It also depends somewhat on what you are more comfortable with using.

3.4 Distribution theory

3.4.1 Linear form

Consider an $n\times 1$ random vector $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ has full rank. For an $m\times n$ matrix $\mathbf{A}$, the linear form $\mathbf{AY} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t)$ if $r(\mathbf{A})=m$, where $m\leq n$. This is an m-variate distribution since $\mathbf{AY}$ is an $m\times 1$ vector.

3.4.2 Quadratic form

For an $n\times n$ matrix $\mathbf{A}$ and $\mathbf{Y}$ as before, $\mathbf{Y}^t\mathbf{AY} \sim \chi^2_\nu(\lambda)$ where $\lambda = \tfrac{1}{2}\boldsymbol{\mu}^t\mathbf{A}\boldsymbol{\mu}$ is the noncentrality parameter and $\nu=r(\mathbf{A})$ are degrees of freedom <u>if and only if</u> any of the following conditions hold:

(i) $\mathbf{A}\boldsymbol{\Sigma}$ is idempotent [i.e., $(\mathbf{A}\boldsymbol{\Sigma})^2 = \mathbf{A}\boldsymbol{\Sigma}$],
(ii) $\boldsymbol{\Sigma}\mathbf{A}$ is idempotent [i.e., $(\boldsymbol{\Sigma}\mathbf{A})^2 = \boldsymbol{\Sigma}\mathbf{A}$], or
(iii) $\boldsymbol{\Sigma}$ is a generalized inverse of $\mathbf{A}$.

The distribution shown above is a *non-central chi-square distribution*. When $\boldsymbol{\mu}=\mathbf{0}$, we have $\lambda=0$, which is the central chi-square distribution that we're familiar with.

3.4.3 Independence of linear and quadratic forms

- For $n\times 1$ random vector $\mathbf{Y}$, let $\mathbf{Y}\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ as before. Let $\mathbf{AY}$ and $\mathbf{BY}$ be two linear forms, and let $\mathbf{Y}^t\mathbf{CY}$ and $\mathbf{Y}^t\mathbf{DY}$ be two quadratic forms.
  - $\mathbf{AY}$ and $\mathbf{BY}$ are independent if $\mathbf{A\Sigma B}^t=\mathbf{0}$
  - $\mathbf{Y}^t\mathbf{CY}$ and $\mathbf{Y}^t\mathbf{DY}$ are independent if $\mathbf{C\Sigma D}^t=\mathbf{0}$
  - $\mathbf{AY}$ and $\mathbf{Y}^t\mathbf{CY}$ are independent if $\mathbf{A\Sigma C}^t=\mathbf{0}$

3.5 Linear independence and rank of a matrix

- For the $n\times p$ matrix $\mathbf{X}$, if $r(\mathbf{X})=\min(n,p)$, then the matrix is said to be of full rank. Since usually $n>p$, $r(\mathbf{X})=p$ if $\mathbf{X}$ has full rank. If $\mathbf{X}$ does not have full rank, then some of the columns can be obtained as a linear combination of the other ones. Equivalently, rows are not all linearly independent if $\mathbf{X}$ does not have full rank. Generally, the rank of $\mathbf{X}$ is the number of linearly independent columns (or number of linearly independent rows) in $\mathbf{X}$.

- For practice, consider the following matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 1 & 4 \\ 1 & 2 & 6 \end{pmatrix}$$

- What is the rank of A? B? Is either of full rank?

- A key result involving rank of matrices is that $r(\mathbf{X}) = r(\mathbf{X}^t) = r(\mathbf{X}^t\mathbf{X}) = r(\mathbf{X}\mathbf{X}^t)$. When $\mathbf{X}$ (and hence $\mathbf{X}^t\mathbf{X}$) has full rank, the inverse exists and the solution for $\boldsymbol{\beta}$ is unique and easy to obtain. In particular, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. Otherwise, we need to either re-express the model so that $\mathbf{X}$ does have full rank, or work with generalized inverses. These approaches are essentially the same.

- Note that the message produced in SAS that '`The X'X matrix has been found to be singular`...' does not indicate that there is a problem with the model fit. However it does mean that only certain combinations of estimates are meaningful in estimating respective combinations of parameters. This issue is referred to as *estimability* and will be discussed more in coming sections.

3.6 Full-rank versus less-than-full-rank models

- Estimation in linear models requires inverting matrices (e.g., algebraic solution for Beta hat or its variance.

- When **X** and hence $\mathbf{X}^t\mathbf{X}$ are not full rank, there are two basic approaches to proceed:

    - Instead of using a regular inverse, proceed with a ‘generalized inverse’ (termed ‘conditional inverse’ by Graybill).
    - Drop linearly dependent columns by creating reference groups or levels, as commonly taught in the 6611/12 sequence.

- The two approaches above are essentially the same in some cases.

- The issue of linearly dependent columns in **X** mainly occurs when considering class/categorical variables. For example, people are either male or female. So if you include ‘male’ and ‘female’ indicator variables, information for one completely depends on the other.

- If a model includes an intercept term and a class variable, and each level of the class variable is given a column in the X matrix, then one of those columns will be linearly dependent on the other ones. We call the associated model a *less-than-full-rank model*.

- A model that has an associated X matrix without linearly dependent columns is a *full-rank model*.

- In order to estimate $\boldsymbol{\beta}$, linear dependencies in the X matrix need to be accounted for by either removing them up front or using a generalized inverse for $\mathbf{X}^t\mathbf{X}$, as previously noted.

3.6.1 Creating a full-rank model by employing restrictions

- One simple approach to deal with linear dependencies is to rewrite the model so that a class variable with *c* levels has *c*–1 indicator variables in the model, and hence *c*–1 columns in associated $\mathbf{X}$ matrix. A couple of ways to do this are to use *set-to-zero* or *sum-to-zero* restrictions on parameters. This new model is a 'full-rank' model since the associated X matrix has full rank, and consequently $\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$ can be computed.

- If gender is a predictor in the model, then only an indicator for ‘Female’ is needed; so we could use a variable that codes ‘l’ for Females, and ‘0’ for Males. This is a set-to-0 approach where the indicator for ‘Male’ is dropped. (Alternatively, one could drop the ‘Female’ indicator.)

- The same is true for each class variable in the model. Thus, if there are two class-level predictors, one with c_1 levels and the other with c_2 levels, then only c_1–1 indicator variables are needed for the first, and c_2–1 for the second. Consequently, only $(c_1–1)(c_2–1)$ are needed for the interaction term between these two predictors (if the interaction term is included in the model).

- If restriction are imposed beforehand, then it is up to the researcher to understand how to interpret the parameter estimates. For example, using one indicator variable for Females means that the parameter associated with the variable represents the difference between Females and Males, since Males are essentially being treated as the reference group.

- A drawback of this approach is that it may require manual creation of indicator variables and more computer code to get certain estimates/tests of interest.

- The *set-to-zero* restriction, using the highest level(s) of factor(s) as reference levels is equivalent to the fitting of the data using the generalized inverse with SAS's approach because of the way the generalized inverse is computed (which is to drop linearly dependent columns moving from left to right, which are the columns associated with the highest levels of factors).

- Using *sum-to-zero* restrictions is another way to specify the model that will allow a reduction of **X** so that it has full rank. For the Myostatin application and the two-way effects model with interaction, this would be: $\sum \alpha_i = 0$, $\sum \tau_j = 0$, $\sum_i \gamma_{ij} = 0$ for fixed *j*, and $\sum_j \gamma_{ij} = 0$ for fixed *i*.

3.6.2 Fitting less-than-full-rank models using generalized inverses

- A second approach to dealing with linear dependencies is to keep the less-than-full-rank model and have computer software work with the linear dependencies directly.

- In this case, a column is included in the X matrix for each level of each class variable that is included in the model.

- So, for gender, one column would be included to indicate ‘Female’, and another would be included for ‘Male’.

- The linear dependency here is obvious since the entries for the Male column are just the Intercept column minus the Female column entries. If the linear dependency is to remain in the model, a *generalized inverse* must be employed in order to calculate estimates. This is done easily using statistical software. Since generalized inverses are not unique, only estimable functions of parameters must be considered, rather than estimates of individual elements of $\boldsymbol{\beta}$. This will be discussed in more detail later.

3.6.2.1 Some theory of generalized inverses

- For an $m \times n$ matrix $\mathbf{A}$, a generalized inverse, denoted $\mathbf{A}^{-}$, satisfies $\mathbf{A}\,\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$.

- Each matrix has at least one generalized inverse.

- Generalized inverses are not necessarily unique.

3.6.2.2 Computations, an example

- One way to compute a generalized inverse is to drop linearly dependent columns as you move from left to right (SAS's approach). For example, consider the matrix

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

C(1) - C(2) C(1)-C(4)
Take out the columns 3 and 5.

Find the (X'X)^-1
(X'X)^- = put 0's in the 3rd and 5th columns

	1	2	3	4	5
1			0		
2			0		
3	0	0	0	0	0
4			0		
5			0		

- The X matrix above corresponds to a two-way ANOVA model without interaction, with only 1 replicate per treatment combination.

- The g-inverse approach in the example above is essentially the same as setting the highest levels of factors to 0.

- Generalized inverses and hence beta estimates are not unique. E.g., you could have dropped linearly dependent columns moving from right to left (making the lowest levels the references).

- Another common generalized inverse is called the Moore-Penrose inverse. This has unique properties that are discussed in more detail in the course notes. Using the MP inverse does not correspond to setting the level of one or more factors to 0. (An example is forthcoming!)

- For the two-way ANOVA model with interaction, SAS's g-inverse approach is essentially equivalent to setting highest levels of factors to 0 and setting the levels of the interaction involving either the highest level of factor A or the highest level of factor B to 0.

3.6.2.3 Implications in estimating $\boldsymbol{\beta}$

- When $\mathbf{X}$ does not have full rank and hence $\mathbf{X}^t\mathbf{X}$ does not have an inverse, then the least squares estimate for $\boldsymbol{\beta}$ can be expressed as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-}\mathbf{X}^t\mathbf{Y} . \qquad (1)$$

- Note that equation (1) with $(\mathbf{X}^t\mathbf{X})^{-1}$ in place of $(\mathbf{X}^t\mathbf{X})^{-}$ is a special case of the formula.

- In (1), $\tilde{\boldsymbol{\beta}}$ is used in place of $\hat{\boldsymbol{\beta}}$ to indicate that the solution is not necessarily unique. (In some cases later on we may relax this and just go back to $\hat{\boldsymbol{\beta}}$ even when the solution is not unique…)

3.6.2.4 Comparison of approaches

Please read the course notes for a discussion of the advantages of using the g-inverse approach as opposed to manually employing restrictions.

3.6.3 Writing full-rank versus less-than-full-rank models: examples and notation

- We have discussed the rank of **X** for general linear model applications and implications for estimation. In this subsection, we discuss different ways to specify the model in order to achieve full-rank or less-than-full-rank models. In terms of estimable functions of parameters, either type of model will provide the same results.

- A 'full-rank' and 'less-than-full-rank' terms refer to whether the associated **X** matrix has full rank or not (the 'less-than-full rank' **X** matrix has more columns than the full-rank **X**). In other courses you may have only focused on one approach.

- Full rank approach. Say we have a continuous variable such as *time*, and 3 treatment *groups* (A, B, Control); *group*×*time* will also be included in the model; **X** has 6 columns. Let x_1 = time, x_2 =1 for treatment A, 0 otherwise; x_3 =1 for treatment B, 0 otherwise (Control will be the reference group – i.e., we have imposed a set-to-0 restriction for the Control group). Let *i* denote the unique data-wide index for subject, and *j* the index for time. Since times of measurement may not be the same for subjects, I keep the *i* index on the time variable as well as *j*.

The model can then be expressed as:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1ij} \cdot x_{2i} + \beta_5 x_{1ij} \cdot x_{3i} + \varepsilon_{ij}$$

Note that we can use like notation for all predictors, but we will have to 'manually construct' the dummy variables for treatment group. We have a model with full rank **X**, so don't need to worry about generalized inverses.

- Less-than-full-rank approach: there are 8 columns in **X**; let's define the treatment effects as κ_h for *group* 1 (A), 2 (B) and 3 (Control), and let γ_h denote the *group*×*time* interaction for h=1,…,3. Since there is only one interaction term, we don't need to add another index on the effect other than one for group. The model is

$$Y_{hij} = \beta_0 + \beta_1 x_{1ij} + \kappa_h + \gamma_h x_{1ij} + \varepsilon_{ij}$$

for group h, subject i and time j. The statistical model has mixed notation, and the associated matrix has less-than-full rank (i.e., dependency in columns). This is the model that SAS fits.