The General Linear Model (Part 1)
Lecture 10

What is GLM?

- The general linear model is a framework for statistical estimation and testing based on modeling a normally distributed outcome as a linear function of a vector of covariates
- Not to be confused with generalized linear models (including logistic regression)
- What sort of methods are associated with the general linear model?
  - One-sample *t* test
  - Two-sample *t* test
  - Simple and multiple linear regression
  - ANOVA

Example: Myostatin data

- 2×3 factorial treatment structure in completely randomized design
  - 2 levels of treatment: myostatin yes or no; 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); it was hypothesized that myostatin samples would have greater protein degradation than controls
- The general linear model can be used to carry out the ANOVA for this experiment

Effects and hypotheses

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

	time			
group	24h	48h	72h	
C	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\cdot}$
M	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	$\mu_{\cdot\cdot}$

- Write null hypotheses for the following tests:
  1. some difference in means
  2. main effect of Time
  3. main effect of Myostatin
  4. Time× Myostatin interaction

ANOVA table

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	3.31	3.31	5.74	0.0277
as.factor(time)	2	18.88	9.44	16.38	0.0001
group:as.factor(time)	2	0.94	0.47	0.82	0.4577
Residuals	18	10.37	0.58		

SAS code

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

SAS output

```
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

                                                          Standard
Parameter                 Estimate             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.
```

R code

```
myostatin <- data.frame(
leucine=c(6568 , 6802 , 7198 , 7280 ,
4992 , 5242 , 5285 , 6284 ,
4092 , 4331 , 5135 , 6087 ,
5516 , 6023 , 6334 , 6400 ,
4512 , 4706 , 5175 , 6612 ,
3076 , 3209 , 3462 , 5364),
group=rep(c(’control’,’myostatin’),each=12),
time=rep(rep(c(24,48,72),each=4),2))

mod1 <- lm(leucine/1000 ˜ group*as.factor(time),
data=myostatin,x=TRUE,y=TRUE)
summary(mod1)
anova(mod1)
```

R output

```
Call:
lm(formula = leucine/1000 ~ group * as.factor(time), data = myostatin,
    x = TRUE, y = TRUE)

Residuals:
   Min     1Q Median     3Q    Max 
-0.819 -0.547 -0.163  0.279  1.586 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        6.962      0.380   18.34  4.3e-13 ***
groupmyostatin                    -0.894      0.537   -1.66   0.1132    
as.factor(time)48                 -1.511      0.537   -2.82   0.0115 *  
as.factor(time)72                 -2.051      0.537   -3.82   0.0013 ** 
groupmyostatin:as.factor(time)48   0.694      0.759    0.91   0.3726    
groupmyostatin:as.factor(time)72  -0.240      0.759   -0.32   0.7558    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.759 on 18 degrees of freedom
Multiple R-squared:  0.69,Adjusted R-squared:  0.604 
F-statistic: 8.02 on 5 and 18 DF,  p-value: 0.000396 

Analysis of Variance Table

Response: leucine/1000
                      Df Sum Sq Mean Sq F value  Pr(>F)    
group                  1   3.31    3.31    5.74   0.028 *  
as.factor(time)        2  18.88    9.44   16.38 8.9e-05 ***
group:as.factor(time)  2   0.94    0.47    0.82   0.458    
Residuals             18  10.37    0.58                    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 
```

- The `NOTE` from SAS follows since estimates of $\boldsymbol{\beta}$ 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.
- The estimates of elements of $\boldsymbol{\beta}$ 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.
- Question: 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?

General form of the model

- We have a sample of size n, and p parameters, including the intercept, are in the model
- The model is
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
- Terms have dimensions as follows:
  - $\mathbf{Y} = n \times 1$
  - $\mathbf{X} = n \times p$
  - $\boldsymbol{\beta} = p \times 1$
  - $\boldsymbol{\epsilon} = n \times 1$
- Two cases for the error term
  1. $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$
  2. Distribution of $\boldsymbol{\epsilon}$ is unspecified except that $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$

Example with simple linear regression I

- Consider the simple linear regression model:
  - Subject form: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$
  - Matrix form: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\beta} = (\beta_0, \beta_1)^T$
- Suppose we observe $X_1 = 1, X_2 = 2, X_3 = 3$, $Y_1 = 1$, $Y_2 = 2$, $Y_3 = 2$. This translates to

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

- What is the rank of **X**?
  - Rank must be less than or equal to the number of columns in **X**, so rank here is at most 2
  - The two columns are not linearly dependent on one another $\Rightarrow$ rank is equal to 2 and a regular inverse of $\mathbf{X}^T\mathbf{X}$ exists
- When the rank of **X** is equal to its number of columns, then the rank of $\mathbf{X}^T\mathbf{X}$ is also equal to the number of columns in **X**

Example with simple linear regression II

- The predicted values based on a least-squares fit (more on this in a few slides) are $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{P_X}\mathbf{Y}$
- $\mathbf{P_X}$ is the projection matrix: technically, this means that the projected vector $\hat{\mathbf{Y}}$ will be in the column space of $\mathbf{X}$ (can be expressed as a linear combination of the columns of $\mathbf{X}$) but as close as possible to $\mathbf{Y}$ in terms of least squares
- Find the projection matrix in the example:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}$$

To find the inverse of a 2×2 matrix of the form

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we can use the formula

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Example with simple linear regression III

This means that

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{6}\begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix}$$

After a little more matrix math, we can find that

$$\mathbf{P_X} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \frac{1}{6}\begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}$$

- Determine $\mathbf{P_X Y}$

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

Example with simple linear regression IV

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

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \frac{1}{6}\left(\begin{array}{rrr} 8 & 2 & -4 \\ -3 & 0 & 3 \end{array}\right)\left(\begin{array}{c} 1 \\ 2 \\ 2 \end{array}\right) = \frac{1}{6}\left(\begin{array}{c} 4 \\ 3 \end{array}\right)$$

This is the same as saying that $\hat{Y}_i = \frac{2}{3} + \frac{1}{2}X_i$

Special cases of GLM

- $\mathbb{E}(Y_1) - \mathbb{E}(Y_2) = \beta_0$ where Y_1 and Y_2 are paired by subject. Test for $H_0 : \beta_0 = 0$ is equivalent to what simple test we know?
- $\mathbb{E}(Y) = \beta_0 + \beta_1 \texttt{group}_i$ where $\texttt{group}_i = 0$ if i is control and 1 if i is treatment. Test for $H_0 : \beta_1 = 0$ is equivalent to what simple test we know?
- Fit Y versus continuous X. What is this called?
- Fit a model for Y, with Group as a class-level predictor that has more than 2 groups. What is the test to compare groups called?
- Fit a model for Y versus Group (as above), and add a continuous predictor. You can optionally add interactions between group and the continuous predictor.
- Fit Y versus a continuous X, plus other predictors that may be continuous or categorical variables. (This encompasses other methods.)

Least-squares

- When we fit the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to data, we must find a solution for $\boldsymbol{\beta}$ that is optimal in some sense
- One approach is to choose $\boldsymbol{\beta}$ that minimizes $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{Y} = \mathbf{X}\beta)^T\,(\mathbf{Y} = \mathbf{X}\beta)$. Any form of $\boldsymbol{\beta}$ that satisfies $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$ will meet this criterion, and these are often called the *normal equations*.

Estimation

- Estimates may be found in closed form
  1. $(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)=\mathbf{Y}^T\mathbf{Y}-2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\beta$
  2. Take the derivative of the expanded quantity with respect to $\boldsymbol{\beta}$ and set the new quantity equal to $\mathbf{0}$: $\mathbf{0}-2\mathbf{X}^T\mathbf{Y}+2\mathbf{X}^T\mathbf{X}\beta=\mathbf{0}$
  3. Rework the equality to get the normal equations $\mathbf{X}^T\mathbf{X}\beta=\mathbf{X}^T\mathbf{Y}$
  4. If the inverse of $\mathbf{X}^T\mathbf{X}$ exists, then the solution is determined as $\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$
- We can see that $\mathbf{X}\hat{\boldsymbol{\beta}}=\hat{\mathbf{Y}}=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}=\mathbf{P}_{\mathbf{X}}\mathbf{Y}$
  - $\mathbf{P}_{\mathbf{X}}$ is the projection matrix
  - We have seen this before as the hat matrix: this is how you take the observed $\mathbf{Y}$ and get the predicted $\hat{\mathbf{Y}}$ (it "puts the hat on $\mathbf{Y}$")

Specific forms of the model

- 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 (categorical) variables:
  - two-way effects model
  - one-way effects model
  - means model

Model	Formula	Indices
Two-way effects model	$Y_{ijk} = \mu + \alpha_i + \tau_j + \gamma_{ij} + \epsilon_{ijk}$	group i, time j, replicate k
One-way effects model	$Y_{ij} = \mu + \kappa_i + \epsilon_{ij}$	group $\times$ time i, replicate j
Means model	$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$	group i, time j, replicate k

- The previous analysis was for the two-way effects model.
- 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.

Distribution theory

Assume we have a $n \times 1$ random vector $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and an $m \times n$ full-rank matrix $\mathbf{A}$,

- $\mathbf{AY} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$
- $\mathbf{Y}^T\mathbf{AY} \sim \chi^2_m(\lambda)$
  - $\lambda = \frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$ is the noncentrality parameter, m is degrees of freedom
  - This distributional result is true if and only if any of the following conditions hold:
    1. $\mathbf{A}\boldsymbol{\Sigma}$ is idempotent, i.e., $(\mathbf{A}\boldsymbol{\Sigma})^2 = \mathbf{A}\boldsymbol{\Sigma}$
    2. $\boldsymbol{\Sigma}\mathbf{A}$ is idempotent
    3. $\boldsymbol{\Sigma}$ is a generalized inverse of $\mathbf{A}$
  - When $\boldsymbol{\mu} = \mathbf{0}$, $\lambda = 0$, giving the central chi-square distribution that we're familiar with

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

- Estimation in linear models requires inverting matrices (e.g., algebraic solution for $\hat{\beta}$ or its variance.
- Definition: a full-rank matrix only contains columns that are mutually linearly independent: no column can be a linear combination of any other columns
- Example: consider two matrices

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

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

Are either of these full rank?

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

- When $\mathbf{X}$ and hence $\mathbf{X}^T\mathbf{X}$ are not full-rank, easiest approach is to drop linearly dependent columns by creating reference groups or levels
- The issue of linearly dependent columns in $\mathbf{X}$ mainly occurs when considering class/categorical variables.
  - For example, people are either male or female
  - $\Rightarrow$ 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 $\mathbf{X}$ 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 $\mathbf{X}$ matrix without linearly dependent columns is a full-rank model.

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

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

Dealing with linear dependencies I

- 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 X matrix
  - *set-to-zero* restrictions: 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).
  - *sum-to-zero* restrictions: 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.
- If gender is a predictor in the model, then only an indicator for 'Female' is needed, so we could use a variable that codes 1 for Females, and 0 for Males.

Dealing with linear dependencies II

  - This is a set-to-zero 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).
- 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.