Comparisons of Several Multivariate Means

Lecture 6
September 28, 2005
Multivariate Analysis
Today’s Lecture

- Comparisons of Two Mean Vectors (Chapter 5).
  - Paired comparisons/repeated measures for two time points.
  - Comparing mean vectors from two populations.

- Review of univariate ANOVA in preparation for next week’s lecture.

- Note: Homework # 4 due in my mailbox on Thursday (by 5pm).
Schedule Change

Because of a couple of developments, we will have a slight change in our schedule:

Today: We cover one part of several mean comparisons (Chapter 6, Sections 1-3).

10/5: We cover MANOVA in *all* its “glory.”

10/12: We cover multivariate multiple regression (and the midterm is handed out, making it due 10/26).

12/7: No class (I will be at the National Council on Responsible Gaming conference in Las Vegas).

I will amend the topics list as the course progresses.
Tests for Two Mean Vectors

- We begin our discussion of multiple multivariate mean vector comparisons by generalizing our results from the previous lecture.
  - Using $T^2$ for comparisons of two mean vectors.

- Differing types of comparisons are all incorporated into a general framework for performing statistical hypothesis tests:
  - Paired comparisons.
  - Repeated measures.

- Much like the univariate t-test, we will also see how comparisons are performed when we have samples from two different populations.
Paired Comparisons

- A research design that can be used to investigate the effects of a single treatment is a paired-comparison study.

- A basic version of a paired comparison study involves the creation of two groups (treatment one and treatment two), where individuals from each group are matched on a set of variables.

- Individuals in the first treatment group are given the first treatment, individuals in the second treatment group are given the second treatment.

- Data is collected from the two groups.

- The null hypothesis is that there are not differences between the groups.

- Paired comparison tests are needed to assess the efficacy of a treatment or (as we will see) multiple treatments.
If a single response is elicited from individuals in both groups, univariate methods can be used to test for differences in the means of both conditions.

One way to investigate the difference between groups is to create a difference score:

\[ D_j = X_{j1} - X_{j2} \]

Where:

- \( X_{j1} \) is the response to treatment one for individual \( j \).
- \( X_{j2} \) is the response to treatment two for “matched” individual \( j \).
- \((X_{j1}, X_{j2})\) are measurements recorded on paired/matched/like units \( j \).
Univariate Statistics

- If a single response is elicited from individuals in both groups, univariate methods can be used to test for differences in the means of both conditions.

- One way to investigate the difference between groups is to create a difference score:

\[ D_j = X_{j1} - X_{j2} \]

Where:

- \( X_{j1} \) is the response to treatment one for individual \( j \).
- \( X_{j2} \) is the response to treatment two for “matched” individual \( j \).
- \((X_{j1}, X_{j2})\) are measurements recorded on paired/matched/like units \( j \).
If the differences, $D$, are from a normal distribution with mean $\delta$ and variance $\sigma_d^2$, or $N(\delta, \sigma_d^2)$, then a hypothesis test can be formed based on $\delta$:

$$H_0 : \delta = 0$$

$$H_1 : \delta \neq 0$$

This hypothesis test can be evaluated using the test statistic:

$$t = \frac{\bar{D} - \delta}{s_d/\sqrt{N}}$$

$t$ has $n - 1$ df (where $n$ is the number of matched pairs).
Univariate Statistics

- Of course, the $100 \times (1 - \alpha)\%$ confidence interval for $\delta$ is given by:

$$\bar{d} - t_{n-1}(\alpha/2) \frac{s_d}{\sqrt{n}} \leq \delta \leq \bar{d} + t_{n-1}(\alpha/2) \frac{s_d}{\sqrt{n}}$$
Multivariate Paired Comparisons

- As you could probably guess, the multivariate analog of the paired-comparison t-test is a straightforward extension of the methodology.

- The research design of the paired comparison is virtually the same as previously discussed, however, multiple observations are taken from each group:

\[
\begin{align*}
X_{1j1} &= \text{variable 1 under treatment 1} \\
X_{1j2} &= \text{variable 2 under treatment 1} \\
\vdots &= \\
X_{1jp} &= \text{variable p under treatment 1} \\
X_{2j1} &= \text{variable 1 under treatment 2} \\
X_{2j2} &= \text{variable 2 under treatment 2} \\
\vdots &= \\
X_{2jp} &= \text{variable p under treatment 2}
\end{align*}
\]
Multivariate Paired Comparisons

- Similar to the univariate case, the method of testing for differences between the mean vectors of the two groups utilizes difference scores.

- Instead of a single difference score, \( p \) difference scores are created:

\[
D_{j1} = X_{1j1} - X_{2j1}
\]
\[
D_{j2} = X_{1j2} - X_{2j2}
\]
\[
\vdots \quad \vdots
\]
\[
D_{jp} = X_{1jp} - X_{2jp}
\]

- The mean vector of the set of difference scores can be denoted by: \( E(D) = \delta \).

- The covariance matrix of the set of difference scores can be denoted by: \( \text{Cov}(D) = \Sigma_d \).
Multivariate Paired Comparisons

- If the matrix of differences, \( D \), is from a MVN distribution with mean vector \( \delta \) and covariance matrix \( \Sigma_d \), or \( N_p(\delta, \Sigma_d) \), then a hypothesis test can be formed based on \( \delta \):

\[
\begin{align*}
H_0 : \delta &= 0 \\
H_1 : \delta \neq 0
\end{align*}
\]

- This hypothesis test can be evaluated using the test statistic:

\[
T^2 = n(\bar{D} - \delta)'S_d^{-1}(\bar{D} - \delta)
\]

- Where:

\[
\bar{D} = \frac{1}{n} \sum_{j=1}^{n} D_j \\
S_d = \frac{1}{n-1} \sum_{j=1}^{n} (D_j - \bar{D})(D_j - \bar{D})'
\]
Multivariate Paired Comparisons

- SAS Example #1...The first of what will be a parade of bad data examples.

- From page 275 in Johnson and Wichern:

  Consider a municipal wastewater treatment plant desiring to monitor water discharges into rivers and streams. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to two laboratories for testing. One half of each sample was sent to the Wisconsin State Laboratory of Hygiene, and one half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for \( n = 11 \) sample splits from the two labs.
WARNING
Snow Contains:
Wastewater Effluent, Endocrine Disruptors, Pharmaceuticals, Pathogens, Hormones, Steroids, Caffeine...
(up to 38 such contaminants)

“What will happen if children eat the snow made from reclaimed water?”
“Nothing.”
– Arizona Snowbowl’s website FAQ

“A study on the city of Flagstaff’s reclaimed water has found trace levels of organic contaminants such as pharmaceuticals and hormones.”
– Arizona Daily Sun 02/22/2004

USGS studies are still trying to understand “their ultimate overall effect on human health and the environment.”

WILL THE SNOWBOWL DRINK THEIR WORDS?

www.savetheneaks.org
Multivariate Paired Comparisons

- The $T^2$ statistic is what we have worked with in the previous week, meaning the value of $T^2$ that would lead to a rejection of $H_0$ (with Type I error rate $\alpha$) is found by:

$$
\frac{(n - 1)p}{(n - p)} F_{p, n-p}(\alpha)
$$

- This also means that $100 \times (1 - \alpha)\%$ simultaneous confidence intervals are formed by:

$$
\bar{d}_i \pm \sqrt{\frac{(n - 1)p}{(n - p)}} F_{p, n-p}(\alpha) \sqrt{\frac{s^2_{d_i}}{n}}
$$
Multivariate Paired Comparisons

- One thing to note (because this will be a recurring method) is that both $\bar{d}$ and $S_d$ (and $T^2$) can be formed by matrix equations.

- First, consider the following matrices formed by the raw (unsubtracted) observations:

$$\bar{x}_{(2p \times 1)} = \begin{bmatrix} \bar{x}_{11} \\ \bar{x}_{12} \\ \vdots \\ \bar{x}_{1p} \\ \bar{x}_{21} \\ \vdots \\ \bar{x}_{2p} \end{bmatrix}$$

$$S_{(2p \times 2p)} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

- $S_{aa'}$ is the covariance matrix between all variables for group $a$ with all variables in group $a'$. 
Multivariate Paired Comparisons

- We can construct a new matrix of constants that will aid in our computation of $T^2$:

$$C_{(p \times 2p)} = \begin{bmatrix}
  1 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 \\
  0 & 1 & \ldots & 0 & 0 & -1 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & 1 & 0 & 0 & \ldots & -1
\end{bmatrix}$$

- The $T^2$ statistic testing for $\delta = 0$ can then be formed by:

$$T^2 = n\bar{x}'C'(CSC')^{-1}C\bar{x}$$
Multivariate Paired Comparisons

- How is $T^2 = n\bar{x}'C'(CSC')^{-1}C\bar{x}$ equivalent to the $T^2$ phrased in terms of difference scores?

- Recall the original formula shown a minute ago (but with $\delta = 0$):

$$T^2 = n\bar{D}'S^{-1}_d\bar{D}$$

- Also recall how linear combinations of variables produce differing mean vectors and covariance matrices:

  - Imagine then that you combine all multivariate variables into a new composite variable $z_j$:

$$z_j = a_1x_{j1} + a_2x_{j2} + \ldots + a_px_{jp}$$

$$\bar{z} = a\bar{x}$$

$$s_z^2 = a'S_xa$$

- What does this imply about $\bar{D}$ and $S_d$?
Repeated Measures Comparisons

- Instead of testing for treatments across matched pairs, another method of research is to administer a set of treatments to each subject.

- For this design, we have \( j \) subjects, each receiving \( q \) treatments.

- For this case, we are only concerned with:
  - A single response variable measured for each subject \( j \) at each time point \( q \):

\[
X_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jq} \end{bmatrix}
\]

- Detecting if all treatments have equivalent means.
Repeated Measures Comparisons

- We can construct a mean vector for the mean at each time point $q$, of size $(q \times 1)$.

- We then construct a Contrast Matrix (what we used for the final result for paired comparisons).

- This contrast matrix specifies a linear combination that will perform the comparison test for equality of all means.

- Note that equality of all means can be tested by several different linear combinations:
  - Comparing all means with a given mean (such as comparing the mean of time 1 with all other time points).
  - Comparing successive means (mean of time 1 with mean of time 2; mean of time 2 with mean of time 3; and so on...).
Contrast Matrix

- Comparing all means with a given mean (such as comparing the mean of time 1 with all other time points).

\[
\begin{bmatrix}
\mu_1 - \mu_2 \\
\mu_1 - \mu_3 \\
\vdots \\
\mu_1 - \mu_q
\end{bmatrix}
= \begin{bmatrix}
1 & -1 & 0 & \ldots & 0 \\
1 & 0 & -1 & \ldots & 0 \\
1 & 0 & 0 & \ldots & -1
\end{bmatrix}
\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3 \\
\vdots \\
\mu_q
\end{bmatrix}
= C_1 \mu
\]

- Comparing successive means (mean of time 1 with mean of time 2; mean of time 2 with mean of time 3; and so on...).

\[
\begin{bmatrix}
\mu_2 - \mu_1 \\
\mu_3 - \mu_2 \\
\vdots \\
\mu_q - \mu_{q-1}
\end{bmatrix}
= \begin{bmatrix}
-1 & 1 & 0 & \ldots & 0 & 0 \\
0 & -1 & 1 & \ldots & 0 & 0 \\
0 & 0 & -1 & \ldots & -1 & 1
\end{bmatrix}
\begin{bmatrix}
\mu_1 \\
\mu_2 \\
\mu_3 \\
\vdots \\
\mu_q
\end{bmatrix}
= C_2 \mu
For either contrast matrix ($C_1$ or $C_2$), it can be shown that the resulting test statistic is equivalent.

The test statistic for the repeated measures hypothesis of equal means at each time point is:

$$T^2 = n\bar{x}'C'(CSC')^{-1}C\bar{x}$$

If this looks familiar, then the distribution of this statistic will, too.

The critical value for rejecting $H_0$ of equal means at each time point is found from:

$$\left(\frac{n - 1}{n - p}\right) F_{p,n-p}(\alpha)$$
From page 280 of Johnson and Wichern:
Improved anesthetics are often developed by first studying their effects on animals. In one study, 19 dogs were initially given the drug pentobarbital. Each dog was then administered carbon dioxide at each of two pressure levels. Next, halothane was added and the administration of CO$_2$ was repeated. The response, in milliseconds between heartbeats, was measured for the four treatment conditions.

- Treatment 1 = high CO$_2$ w/o H
- Treatment 2 = low CO$_2$ w/o H
- Treatment 3 = high CO$_2$ w/ H
- Treatment 4 = low CO$_2$ w H
There are three contrasts of interest:

1. \((\mu_3 + \mu_4) - (\mu_1 + \mu_2)\) - H

2. \((\mu_1 + \mu_3) - (\mu_2 + \mu_4)\) - CO2

3. \((\mu_1 + \mu_4) - (\mu_2 + \mu_3)\) - Interaction
Comparing Two Populations

- We have been over variations of the $T^2$ test for mean vectors for paired comparisons or for repeated measures designs - both of which were cases where the sample was from a single population.

- To make the method of testing more general, consider the comparison of mean vectors from two populations.

- As with every multivariate test presented up to this point, we again turn our attention to similar tests for univariate cases.
Univariate Two-Population t-tests

- Recall from univariate statistics situations where samples were taken from two different populations.

- For instance, reliving our height example from the past week, imagine that I am interested in comparing the heights of KU grad students with the heights of grad students at Illinois (the real reason Bill Self left).

- Let $x_{1j}$ represent the height of a grad student at Illinois.

- Let $x_{2j}$ represent the height of a grad student at KU.

- We can then construct the following summary statistics for each sample:
  - The mean heights: $\bar{x}_1$ and $\bar{x}_2$.
  - The variances: $s_1^2$ and $s_2^2$. 

Overview

Tests for Two Mean Vectors
- Paired Comparisons
- Multivariate Comparisons
- Repeated Measures

Comparing Two Populations

Univariate t-tests
- Multivariate t-tests

1-Way ANOVA

Wrapping Up
Univariate Two-Population t-tests

- We seek to test the following hypothesis:

  \[ H_0 : \mu_1 = \mu_2 \]

  \[ H_1 : \mu_1 \neq \mu_2 \]

- One way of doing this is to compute a two-sample t-test.

- For small samples, we must assume that \( s_1^2 = s_2^2 \), leading to the following test:

  \[ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \]

- Here, this test is compared against a t-value for Type I error rate \( \alpha \) with \( n_1 + n_2 - 2 \) degrees of freedom.
Univariate Two-Population t-tests

■ If the assumption of equivalent variances is used \( s_1^2 = s_2^2 \), the pooled estimate of variance is:

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}
\]

■ If this assumption is violated (and sample sizes are large), the pooled variance estimate is sometimes computed by:

\[
s_p^2 = \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2
\]

■ Now we will see how these values generalize to their multivariate counterparts.
For our multivariate test, consider taking a sample of more than one variable from two populations.

This would be equivalent to our example of height and forearm length from last week.

Now we wish to test the hypothesis that the mean vectors are equal against the alternative that the mean vectors are not equal:

\[ H_0 : \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = \delta_0 \]

\[ H_1 : \mu_1 \neq \mu_2 \iff \mu_1 - \mu_2 \neq \delta_0 \]
Assumptions

Prior to conducting our hypothesis test, we must first note the assumptions of the test:

1. The sample $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of size $n_1$ from a $p$-variate population with mean vector $\mu_1$ and covariance matrix $\Sigma_1$.

2. The sample $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of size $n_2$ from a $p$-variate population with mean vector $\mu_2$ and covariance matrix $\Sigma_2$.

3. $X_{11}, X_{12}, \ldots, X_{1n_1}$ are independent of $X_{21}, X_{22}, \ldots, X_{2n_2}$.

4. If $n_1$ and $n_2$ are small: both populations are MVN.

5. If $n_1$ and $n_2$ are small: $\Sigma_1 = \Sigma_2$. 

Assumptions
Test Statistic

- For small samples satisfying assumptions #1-#5, the test statistic is:

\[ T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_p \right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) \]

where the pooled covariance matrix is formed by:

\[ S_p = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2 \]

- The critical value for rejecting \( H_0 \) of equal population mean vectors is found from:

\[ \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - 1 - p)} F_{p, n_1+n_2-1-p}(\alpha) \]
Confidence Regions

- Similar to all tests up to this point, we can construct confidence regions for the difference in means.

- Here, the region is centered at:

  \[ \bar{x}_1 - \bar{x}_2 \]

- The region has axes that are formed by:

  \[ \bar{x}_1 - \bar{x}_2 \pm \sqrt{\lambda_i} \sqrt{\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - 1 - p)}} F_{p,n_1+n_2-1-p(\alpha)} e_i \]

- Where the eigenvalues and eigenvectors come from the pooled covariance matrix \( S_p \).
Again, similar to all tests up to this point, we can construct a set of simultaneous univariate confidence intervals for the difference in means.

The $100 \times (1 - \alpha)\%$ confidence interval for $\mu_z = a\mu_x$ is given by:

$$a'(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - 1) - p}} F_{p, n_1 + n_2 - 1 - p} \left( \frac{1}{n_1} + \frac{1}{n_2} \right) a'S_p a$$

SAS Example #3...

Fifty bars of soap are manufactured in each of two ways. Two characteristics - $X_1 = \text{lather}$ and $X_2 = \text{mildness}$, are measured. Compare the two methods to detect mean differences in either of $X_1$ or $X_2$. 
Unequal Covariance Matrices

- Often times, covariance matrices are not equal for differing populations.

- Although statistical tests have been created to detect differing covariance matrices, their results are suspect when the populations are not MVN.

- The book suggests that any element of the covariance matrices for either sample that is greater than four times the corresponding element for the other sample should cause concern.

- If \( n_1 \) and \( n_2 \) are large, the pooled covariance estimate can be formed by:

  \[ S_p = \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \]

- If sample sizes are small, the book suggests trying transformations.
All through this point we have been dealing with cases where we were comparing one mean to another.

Because of this, we were always able to simply compute a standardized distance between them and see if it was too extreme.

Now we are going to discuss what we will do if we several means (meaning two or more).
Specifically, we will begin with a scenario where we have a total of \( k \) groups and we want to test the hypothesis

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_k \]

Now because we have several mean vectors it will not make sense to simply compute a squared distance and so we will look at something else.

Similar to the ANOVA (which looks at variances) we will also focus on variances, but in a multivariate setting, making it Multivariate ANOVA, or MANOVA.
Univariate ANOVA

- We begin with a reminder of what we did when all we had was a single observation per unit and there were several groups.

- So our data looked like

\[
\begin{array}{cccc}
\text{Group 1} & \text{Group 2} & \ldots & \text{Group } k \\
\hline
x_{11} & x_{21} & \ldots & x_{k1} \\
x_{12} & x_{22} & \ldots & x_{k2} \\
\vdots & \vdots & \ddots & \vdots \\
x_{1n} & x_{2n} & \ldots & x_{kn}
\end{array}
\]

- We want to test the basic hypothesis .

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_k \]
Here we had to make the basic assumptions that:

- Each observation came from a normal distribution with mean $\mu$ and variance $\sigma^2$ (we assume equal variances across groups)

- All observations are independent

The basic linear model is $y_{ij} = \mu + \tau_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}$
From Pedhazur (1997; p. 343): “Assume that the data reported [below] were obtained in an experiment in which $E$ represents an experimental group and $C$ represents a control group.

<table>
<thead>
<tr>
<th></th>
<th>$E$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sum Y$</td>
<td>85</td>
<td>65</td>
</tr>
<tr>
<td>$\bar{Y}$</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>$\sum (Y - \bar{Y})^2 = \sum y^2$</td>
<td>26</td>
<td>34</td>
</tr>
</tbody>
</table>
Effect Coding

- Effect coding is a method of coding categorical variables compatible with the general linear model ANOVA parameterization.

- In effect coding, one creates a set of column vectors that represent the membership of an observation to a given category level or experimental condition.

- The total number of column vectors for a categorical variable are equal to one less than the total number of category levels.
If an observation is a member of a specific category level, they are given a value of 1 in that category level’s column vector.

If an observation is not a member of a specific category and is not a member of the omitted category, they are given a value of 0 in that category level’s column vector.

If an observation is a member of the omitted category, they are given a value of -1 in every category level’s column vector.
Effect Coding

- For each observation, a no more that a single 1 will appear in the set of column vectors for that variable.

- The column vectors represent the predictor variables in a regression analysis, where the dependent variable is modeled as a function of these columns.

- Because all observations at a given category level have the same value across the set of predictors, the predicted value of the dependent variable, $Y'$, will be identical for all observations within a category.

- The set of category vectors (and a vector for an intercept) are now used as input into a regression model.
### Effect Coded Regression Example

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X₁</th>
<th>X₂</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>1</td>
<td>-1</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>1</td>
<td>-1</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>1</td>
<td>-1</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>1</td>
<td>-1</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>1</td>
<td>-1</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>15</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
The General Linear Model states that the estimated regression parameters are given by:

\[ \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \]
For our second example analysis, consider the regression of $Y$ on $X_1$ and $X_2$.

$$Y = a + b_2 X_2 + e$$

- $a = 15$
- $b_2 = 2$
- $\sum y_2 = 100$
- $SS_{res} = X'X = 60$
- $SS_{reg} = 100 - 60 = 40$
- $R^2 = \frac{40}{100} = 0.4$
Effect Coded Regression - $X_1$ and $X_2$

- $a = 15$ is the overall mean of the dependent variable across all categories.

- $b_2 = 2$ is the called the effect of the experimental group.

- This effect represents the difference between the experimental group mean and the overall mean.

- For members of the $E$ category:
  \[ Y' = a + b_2X_2 = 15 + 2(1) = 17 \]

- For members of the $C$ category:
  \[ Y' = a + b_2X_2 = 15 + 2(-1) = 13 \]

- The fit of the model is the same as was found in the dummy coding from the previous class.
Effect coding is built to estimate the fixed linear effects model.

\[ Y_{ij} = \mu + \tau_j + \epsilon_{ij} \]

- \( Y_{ij} \) is the value of the dependent variable of individual \( i \) in group/treatment/category \( j \).
- \( \mu \) is the population (grand) mean.
- \( \tau_j \) is the effect of group/treatment/category \( j \).
- \( \epsilon_{ij} \) is the error associated with the score of individual \( i \) in group/treatment/category \( j \).
The fixed effects linear model states that a predicted score for an observation is a composite of the grand mean and the treatment effect of the group to which the observation belongs.

\[ Y_{ij} = \mu + \tau_j + \epsilon_{ij} \]

For all category levels (total represented by \( G \)), the model has the following constraint:

\[ \sum_{g=1}^{G} \tau_g = 0 \]
The Fixed Effects Linear Model

- This constraint means that the effect for the omitted category level (o) is equal to:

\[
\tau_o = \sum_{g \neq o} -\tau_g = -\tau_1 - \tau_2 \ldots
\]

- From the example, the effect for the control group is equal to:

\[
\tau_C = -\tau_E = -2
\]

- Just to verify:

\[
\tau_E + \tau_C = 2 + (-2) = 0
\]
Hypothesis Test of the Regression Coefficient

- Because each model had the same value for $R^2$ and the same number of degrees of freedom for the regression (1), all hypothesis tests of the model parameters will result in the same value of the test statistic.

$$F = \frac{R^2/k}{(1 - R^2)/(N - k - 1)} = \frac{0.4/1}{(1 - 0.4)/(10 - 1 - 1)} = 5.33$$

- From Excel ("=fdist(5.33,1,8)"), $p = 0.0496$.

- If we used a Type-I error rate of 0.05, we would reject the null hypothesis, and conclude the regression coefficient for each analysis would be significantly different from zero.
The thing to notice is that if the Null hypothesis is correct (i.e., all equal variances) then we have two ways to estimate the population variance.

1. We have a total of $k$ samples and we know that for any given $i^{th}$ sample, its variance $s_i^2$ is an estimate of the population variance. So we could estimate the population variance by simply pooling together the variances WITHIN each group (i.e., compute and average variance).

2. On the other hand we have $k$ groups and for each group we have a mean $\bar{y}_i$. This is a sampling distribution. We know that the variance of the means should be $\sigma^2/n$. So we could estimate the variance BETWEEN each of the groups and multiple it by $n$ to get an estimate.
The important thing that we use for our hypothesis test is that number (2) will increase if there are actual differences in the group means.

How did we estimate these computationally?

1. **WITHIN** estimate of Variance ($s_e^2$):

$$s_e^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2}{k(n-1)} = \frac{SSE}{k(n-1)} = MSE$$

2. **BETWEEN** estimate of Variance ($n s_{\bar{y}}^2$):

$$n s_{\bar{y}}^2 = \frac{n \sum_{i=1}^{k} (\bar{x}_i - \bar{x}_{..})^2}{k-1} = \frac{SST}{k-1} = MST$$
Now given our two estimates of variance we simply took the ratio

\[ F = \frac{MST}{MSE} = \frac{\frac{SST}{k-1}}{\frac{SST}{k(n-1)}} \]

Where our F-statistic, assuming the Null hypothesis is true, follows an F-Distribution with degrees of freedom equal to \( k - 1 \) and \( k(n - 1) \).

If there are true differences BETWEEN our group means then MST will over-estimate the variance and therefore F will be larger than we expect (i.e., it would be too extreme) and we would reject.
Today we covered the multivariate analog of two-sample t-tests.

Next week we will discover how these tests are incorporated into the MANOVA framework.

Please use the ANOVA section of this course as a refresher for topics we will cover next time.
Next Time

- MANOVA.
- Profile analysis.
- Growth Curves.
- Beginning of multivariate multiple regression.