# HUDM 5123 - Linear Models and Experimental Design
# 11 - Repeated Measures - The ANOVA Approach

## 1 Repeated Measures Data

Up to now, we have considered data sets that have only one observation per participant. This is helpful from a statistical perspective because it makes the assumption of independent observations (or residuals, after accounting for a linear function of covariates) more likely to hold. Nevertheless, research questions may necessitate the collection of data that include cases that are not independent. Three features that can lead to dependence among cases are (a) clustering of cases by groups, (b) repeated measurement of cases across factor levels, and (c) repeated measurement of cases across time.

(a) **Clustering.** An outcome variable is said to be "clustered" when, although there is only one observation per case, the cases are grouped by some other factor. An example is a measure of student math achievement. Although each student only has a single math achievement score, an assumption of independence among cases is likely to be violated because observations for students from the same school are likely to be correlated.

(b) **Repeated measures: not longitudinal.** An outcome variable is said to be a "repeated measures" variable if it is measured multiple times under different factor/level combinations. As an example, consider a perceptual psychologist interested in the effect of visual interference on letter recognition. The psychologist enrolls ten subjects for a study wherein the interference factor has four levels: no interference, mild interference, moderate interference, and severe interference. Each subject is tested for letter recognition under all four conditions, producing a total of 40 observations. Order is often randomized or systematically counterbalanced to avoid order effects.

(c) **Repeated measures: longitudinal.** An outcome variable is said to be a "longitudinal" variable if cases are measured at different time points. As an example, consider a developmental psychologist interested in the effect of playing heartbeat sounds on newborns' weight gain. Thirty newborns are randomly assigned to the heartbeat group or a control group (15 in each) and their weight is tracked weekly for six weeks, producing a total of 180 observations.

## 2 Analyzing Repeated Measures Data with Two Measurements

We will first discuss the case where there are only two repeated measurements. We separate the case of only two repeated measures because we can handle that case with methods that we already have. With more than two repeated measures, however, we will need new methodology, which will be presented below. The pre-post design is the simplest example of a repeated measures design. Suppose there is an outcome, $Y$, and some intervention, $T_i$, coded 0 for control group and 1 for intervention group, say. Also suppose that $Y$ was measured twice. At pretest, before the intervention, $Y_{i1}$, and at posttest, after the intervention, $Y_{i2}$.

## 2.1 Ignoring Pretest: ANOVA

If participants were randomly assigned to control and intervention groups, we might consider ANOVA, ignoring pretest. The ANOVA full and reduced models:

$$\text{Full: } Y_{i2} = \beta_0 + \beta_1 T_i + \epsilon_i \tag{1}$$
$$\text{Reduced: } Y_{i2} = \beta_0 + \epsilon_i. \tag{2}$$

## 2.2 Accounting for Pretest via ANCOVA

Even if the treatment was randomly assigned, we might also control for pretest to get better efficiency. If the treatment was not randomly assigned, we might wish to control for pretest to linearly adjust for any group differences on pretest at baseline; ANCOVA accomplishes this. The ANCOVA full and reduced models:

$$\text{Full: } Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 T_i + \epsilon_i \tag{3}$$
$$\text{Reduced: } Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \epsilon_i. \tag{4}$$

Recall, the ANCOVA model specifies that the slope, $\beta_1$, that governs the linear relationship between the pretest and posttest, is allowed to be freely estimated subject to the constraint that it is identical across both treatment and control groups.

## 2.3 Accounting for Pretest via Difference Scores

Another approach that is often considered is to use "gain" or "difference" scores. That is, define $D_i = Y_{i2} - Y_{i1}$. Then run an ANOVA on the difference scores with the following full and reduced models:

$$\text{Full: } D_i = \beta_0 + \beta_1 T_j + \epsilon_i \tag{5}$$
$$\text{Reduced: } D_i = \beta_0 + \epsilon_i. \tag{6}$$

The difference score full model may be written as follows:

$$D_i = \beta_0 + \beta_1 T_j + \epsilon_i \tag{7}$$
$$Y_{i2} - Y_{i1} = \beta_0 + \beta_1 T_j + \epsilon_i \tag{8}$$
$$Y_{i2} = \beta_0 + \beta_1 T_j + 1 \times Y_{i1} + \epsilon_i \tag{9}$$

The only difference between Equations 3 and 9 is that the slope on the pretest is constrained to be equal to 1 for the gain scores model. That is, $\beta_2$, in the ANCOVA model is free to be estimated, whereas $\beta_2$ in the gain score model is fixed to be 1. Thus, the assumption under the gain scores model is that there is *perfect correlation* between pretest and posttest. In other words, a one unit change in pretest corresponds with a one unit change in posttest. This assumption is overly restrictive. Even if the pretest and posttest measures are exactly the same, they will likely be correlated, but not perfectly so, due to random error.

The null hypotheses tested by gain score analysis:

- $H_0 : \beta_0 = 0$ tests if the average change is equal to zero for the control group.
- $H_0 : \beta_1 = 0$ tests if the average change is equal for the two groups.

The null hypotheses tested by ANCOVA:

- $H_0 : \beta_0 = 0$ tests if average posttest = 0 for control group subjects with 0 on pretest.
- $H_0 : \beta_1 = 0$ tests if posttest is equal for the two groups, given same value of pretest.
- $H_0 : \beta_2 = 0$ tests if posttest is related to pretest, controlling for group.

We go back to the acupuncture data for an example. Pretest and post-acupuncture measures of headache frequency are, respectively, $pk1$ and $pk5$. First, create a variable called "diffs" by taking the post - pre difference for each case: `dat$diffs <- dat$pk5 - dat$pk1`. Then, for the change score analysis, regress diffs on group.

```
lm(formula = diffs ~ group, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.3673     0.9883  -4.419 1.39e-05 ***
group1       -3.9620     1.3513  -2.932  0.00363 **
```

The average change is -3.96 and is significant, meaning that the average change in headache frequency was a decrease of about 4 points. The ANCOVA output:

```
lm(formula = pk5 ~ group + pk1, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.45098    1.41672   2.436 0.015441 *
group1      -4.58684    1.25177  -3.664 0.000294 ***
pk1          0.70730    0.04055  17.444  < 2e-16 ***
```

The posttest difference in groups, controlling for pretest, is significant and about -4.6 points. Notice that the p-value on group is .0003 for ANCOVA and .003 for gain score analysis. This highlights, at least for this one case, what tends to be true in general: the ANCOVA approach is more powerful than the gain score approach for detecting a difference.

# 3 The Univariate Approach to Repeated Measures ANOVA

Consider some fabricated data generated to correspond with the aforementioned perceptual psychologist's experiment to research the relationship between the ability to recognize letters and interfering visual stimuli. Subjects are told that they will see either the letter "T" or the letter "I" on the screen in front of them. In some trials the letter appears by itself. In others it appears embedded in a group of other letters. This factor is called *noise*, because it refers to visual noise. The amount of noise interference is varied across four groups. The outcome of interest is the reaction time until the subject recognizes the position of the target letter on the screen.

Table 1: Letter recognition time (ms)

| Subject | No | Mild | Mod | Sev |
|---|---|---|---|---|
| 1 | 420 | 480 | 600 | 780 |
| 2 | 420 | 480 | 480 | 600 |
| 3 | 480 | 540 | 660 | 780 |
| 4 | 420 | 540 | 780 | 900 |
| 5 | 540 | 540 | 660 | 720 |
| 6 | 360 | 420 | 480 | 540 |
| 7 | 530 | 540 | 720 | 840 |
| 8 | 480 | 540 | 720 | 900 |
| 9 | 540 | 540 | 720 | 780 |
| 10 | 420 | 540 | 660 | 780 |

Notice in the data in Table 1 that each subject was measured four times: once for each level of the noise factor. To begin thinking about how to model these data, suppose, for a moment, that instead of 10 subjects measured four times each, 40 subjects had been randomly assigned to be in one of the four groups such that there was only one measurement per subject. Then a test of the one-way ANOVA omnibus null hypothesis, $H_0 : \mu_{\text{No}} = \mu_{\text{Mild}} = \mu_{\text{Mod}} = \mu_{\text{Sev}}$, could be carried out with the following full and reduced models:

$$\text{Full: } Y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \epsilon_i, \tag{10}$$

$$\text{Reduced: } Y_i = \beta_0 + \epsilon_i, \tag{11}$$

where $D_{i1}$, $D_{i2}$, and $D_{i3}$ are either dummy- or effect-coded based on group membership and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$. Now, back to the repeated-measures data. When there are multiple observations per subject, the notation may be extended by using an additional subscript. Let $i$ represent subjects ($i = 1, 2, \ldots, N$) and let $j$ represent measurement occasions ($j = 1, 2, \ldots, n_i$). Notice that the total number of measurement occasions may differ across subjects, hence, $j$ ranges from 1 to $n_i$, a number that depends on the subject. Rewriting the full model from Equation 10 using notation with two subscripts gives the following:

$$Y_{ij} = \beta_0 + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 D_{ij3} + \epsilon_{ij}. \tag{12}$$

A simple extension of the model given in Equations 10 and 12 to allow for the influence of each individual on their repeated outcomes can be constructed by adding a subject-specific

term to the intercept. That is,

$$Y_{ij} = \beta_0 + \tau_{0i} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 D_{ij3} + \epsilon_{ij}, \tag{13}$$

where $\tau_{0j} \overset{iid}{\sim} N(0, \sigma^2_\tau)$ is a subject-specific contribution to the intercept that is assumed to be independent of the residual variation $\epsilon_{ij}$. If there are no subject-specific influences on repeated outcomes, all of the $\tau_{0i}$ terms would be equal to 0, though this would be quite unusual. More typically, the $\tau_{0i}$ terms will deviate from 0. The model given in Equation 13 is called a *random intercept model*. It is one of the most basic *mixed-effects regression models* (MRMs), also referred to as a *hierarchical linear model* (HLM). To represent Equation 12 as an HLM, it may partitioned into a level-1 (within-subjects) model and a level-2 (between-subjects) model. The level-1 model:

$$\text{Level-1:} \quad Y_{ij} = b_{0i} + b_{1i} D_{ij1} + b_{2i} D_{ij2} + b_{3i} D_{ij3} + \epsilon_{ij} \tag{14}$$

The level-2 model:

$$\begin{aligned} \text{Level-2:} \quad b_{0i} &= \beta_0 + \tau_{0i}, && (15)\\ b_{1i} &= \beta_1, && (16)\\ b_{2i} &= \beta_2, && (17)\\ b_{3i} &= \beta_3. && (18) \end{aligned}$$

The level-1 model specifies that subject $i$'s response at time measurement occasion $j$ is influenced by his or her initial level $\beta_{0i}$ and his or her group membership, indicated by $b_{1i}$, $b_{2i}$, and $b_{3i}$. The level-2 model specifies that subject $i$'s initial level is determined by the overall population initial level $\beta_0$, plus a subject-specific contribution $\tau_{0i}$. On the other hand, the group effects are assumed to be constant for all subjects (for example, $b_{1i} = \beta_1$.)

Clearly, the random intercept model is modeling subject-specific variability with the $\tau_{0i}$'s. However, what are the implications for how the dependence among each subject's measurements are modeled? To understand that, we will need to calculate some variances and covariances. We may use Equation 13,

$$Y_{ij} = \beta_0 + \tau_{0i} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 D_{ij3} + \epsilon_{ij},$$

to determine the variance of the outcome.

$$\begin{aligned} \text{Var}(Y_{ij}) &= \text{Var}(\beta_0 + \tau_{0i} + \beta_1 D_{ij1} + \beta_2 D_{ij2} + \beta_3 D_{ij3} + \epsilon_{ij}) \\ &= \text{Var}(\beta_0) + \text{Var}(\tau_{0i}) + \text{Var}(\beta_1 D_{ij1}) + \text{Var}(\beta_2 D_{ij2}) + \text{Var}(\beta_3 D_{ij3}) + \text{Var}(\epsilon_{ij}) \\ &= \text{Var}(\tau_{0i}) + \text{Var}(\epsilon_{ij}) \\ &= \sigma^2_\tau + \sigma^2_\epsilon \end{aligned}$$

The covariance between two observations measured at different measurement occasions:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{Ij'}) &= \text{Cov}(\tau_{0i} + \epsilon_{ij}, \tau_{0i} + \epsilon_{ij'}) \\ &= \text{Cov}(\tau_{0i}, \tau_{0i}) + \text{Cov}(\tau_{0i}, \epsilon_{ij}) + \text{Cov}(\tau_{0i}, \epsilon_{ij'}) + \text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) \\ &= \text{Cov}(\tau_{0i}, \tau_{0i}) \\ &= \text{Var}(\tau_{0i}) \\ &= \sigma^2_\tau \end{aligned}$$

Notice that neither the variance nor the covariance depend on the measurement occasion. Thus, there are implicit assumptions that (a) the variance of the outcome does not change over measurement occasions, and (b) the covariances (and hence, correlations) of the outcome are fixed and identical across measurement occasions. In particular, we can write-down the form of the variance/covariance matrix of the outcome variable across measurement occasions. If there are four repeated measures, as in the example given about in Table 1, it will look like this:

$$\Sigma = \begin{bmatrix} \sigma_\tau^2 + \sigma_\epsilon^2 & \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\epsilon^2 & \sigma_\tau^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\epsilon^2 & \sigma_\tau^2 \\ \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma_\epsilon^2 \end{bmatrix}$$

The structure shown in the variance/covariance matrix above is called *compound symmetry*. Compound symmetry implies that the variance is constant across repeated measurements ($\sigma_\tau^2 + \sigma_\epsilon^2$) and the covariances are constant across repeated measurements ($\sigma_\tau^2$). It also has implications for correlations between outcomes measured at different occasions. Recall that the correlation between two variables is defined as the ratio of their covariance to the product of their standard deviations. That is, $\text{Cor}(X, Y) = \text{Cov}(X, Y)/\sqrt{\text{Var}(X)\text{Var}(Y)}$. The correlation between responses given at different measurement occasions is called the *intraclass correlation*, and is denoted $\rho_{\text{ICC}}$. The ICC may be interpreted as the expected correlation between randomly drawn observations from the same subject.

$$\begin{aligned} \rho_{\text{ICC}} &= \text{Cor}(Y_{ij}, Y_{ij'}) = Cov(Y_{ij}, Y_{ij'})/\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{ij'})} \\ &= \sigma_\tau^2/(\sigma_\tau^2 + \sigma_\epsilon^2). \end{aligned}$$

The ANOVA framework may be used to estimate variances and coefficients associated with models like the one given in Equation 15. Although the ANOVA framework is familiar, a challenge is that all subjects are required to have the same number of repeated measurements. This is not the case when maximum likelihood is used for estimation; you will hear more about estimation using maximum likelihood next class.

## 3.1 Checking Assumptions

Huynh and Feldt (1970) showed that the compound symmetry assumption is equivalent to assuming that the population covariance matrix has a particular form, called *sphericity*. Sphericity is equivalent to asserting that the variances of the differences in outcomes are identical. For example, for four outcomes, sphericity asserts that $\sigma^2_{Y_1-Y_2} = \sigma^2_{Y_1-Y_3} = \sigma^2_{Y_1-Y_4} = \sigma^2_{Y_2-Y_3} = \sigma^2_{Y_2-Y_4} = \sigma^2_{Y_3-Y_4}$.

A good first step is to examine the estimated variance/covariance and correlation matrices to see if compound symmetry seems plausible. For the letter recognition data, the variance/covariance matrix:

```
No  Mil   Mod   Sev
```

```
No  3360 1770  3280  2720
Mil 1770 1760  3680  3920
Mod 3280 3680 10240 10960
Sev 2720 3920 10960 13640
```

and the correlation matrix:

```
      No  Mil  Mod  Sev
No  1.00 0.72 0.56 0.40
Mil 0.72 1.00 0.87 0.80
Mod 0.56 0.87 1.00 0.93
Sev 0.40 0.80 0.93 1.00
```

The variances appear to differ across measurement occasions (13640 is about 8 times larger than 1760). The covariances appear to differ as well (10960 is about 6 times larger than 1770). Looking at the correlations, we see what appears to be a higher correlation between adjacent measurement occasions than between distant ones. This is expected, even though it violates the compound symmetric structure. Unfortunately, the ANOVA $F$ test of significance in a repeated measures design is not robust to violation of sphericity. Mauchly (1940) developed a test for violation of sphericity. The null hypothesis of the test is that sphericity is satisfied. If the null hypothesis is rejected, we have evidence to believe sphericity is violated.

```
mauchly.test(object = mlm1, X = ~1, idata = idata)
Mauchly's test of sphericity
Contrasts orthogonal to ~1
data:  SSD matrix from lm(formula = rmd ~ 1)
W = 0.11934, p-value = 0.00617
```

In the output above, note that the p-value of .006 $<$ .05, so we reject the null hypothesis that the sphericity assumption is satisfied and conclude that it is not. So what do we do now? Box (1954) showed that a correction can be made by adjusting the numerator and denominator degrees of freedom of the F test by multiplying both by a constant, called *epsilon*. Four degrees of correction for violation of sphericity are sometimes implemented in software packages that handle the univariate ANOVA approach to repeated measures:

- Sphericity assumed (no adjustment): $\epsilon = 1$. Probably too lenient.
- Huynh-Feldt's correction: $\tilde{\epsilon}$ ("epsilon tilde").
- Box's correction (also called Greenhouse-Geisser correction): $\hat{\epsilon}$ ("epsilon hat").
- Geisser-Greenhouse correction (lower-bound): $\epsilon = 1/(a-1)$. Probably too stringent.

Huynh-Feldt's correction.

$$\tilde{\epsilon} = \frac{n(a-1)\hat{\epsilon} - 1}{(a-1)[(n-1) - (a-1)\hat{\epsilon}]}$$

Box's correction.

$$\hat{\epsilon} = \frac{a^2(\bar{E}_{ll} - \bar{E})^2}{(a-1)\left[(\sum\sum E_{lm}^2) - \left(2a\sum \bar{E}_l^2\right) + (a^2\bar{E}^2)\right]}$$

where $E_{lm}$ is the element in row $l$ and column $m$ of the sample covariance matrix, $\bar{E}_{ll}$ is the mean of the diagonal entries in the sample covariance matrix, $\bar{E}_l$ is the mean of the entries in the $l$th row of the sample covariance matrix, and $\bar{E}$ is the mean of all entries in the sample covariance matrix.

Relationship between the four estimates for $\epsilon$:

$$\frac{1}{a-1} \leq \hat{\epsilon} \leq \tilde{\epsilon} \leq 1$$

$\hat{\epsilon}$ tends to underestimate $\epsilon$ and $\tilde{\epsilon}$ tends to overestimate $\epsilon$. When there are two groups ($a = 2$), they are all the same.

For the letter recognition data, output, including Huyn-Feldt and Greenhouse-Geisser corrections are presented below:

```
Greenhouse-Geisser epsilon: 0.4528
Huynh-Feldt epsilon:        0.5054

  Res.Df Df Gen.var.      F num Df den Df     Pr(>F)     G-G Pr     H-F Pr
1      9      888.07
2     10  1 1782.73 65.494      3     27 1.6298e-12 1.0464e-06 2.8685e-07
```

Note that the result is significant no matter which correction is used.