2013

Repeated Measures ANOVA

Patricia Valdez
University of Redlands

Repeated Measures ANOVA

UNIVERSITY OF REDLANDS

Patricia Valdez

April 11, 2013

Honors Committee:
Professor Jim Bentley
Professor Rick Cornez
Professor Joanna Bieri
Professor Bryce Ryan

Contents

1 ANOVA

Analysis of variance, referred to as ANOVA, is a test for the equality of means of several different treatment groups. It is based on the partitioning of sums of squares of errors [13]. In an experiment that requires the use of ANOVA, subjects are randomly assigned to one of several treatment groups. These treatment groups can be *random*, where there are more treatment possibilities than just the ones being tested, or they can be *fixed*, where all the possible treatments are being tested. In order to test whether or not the means of the multiple treatment groups are statistically different from one another the following hypotheses are developed:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a \tag{1.1}$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one } i \neq j$$

where a is the number of treatment groups and μ_i is the i^{th} treatment mean $(i = 1, 2, ..., a)$. Instead of performing multiple t-tests to test whether rejecting the null hypothesis, H_0, is warranted, ANOVA allows us to perform a single F-test to compare two or more means.

ANOVA can be adjusted to take repeated measures and unbalanced data into account. However, before we address these special cases an understanding of the basic form of ANOVA must be developed.

1.1 One-way Classification ANOVA

We will begin with the one-way classification ANOVA with a fixed effects model. What it means for the effects to be *fixed* is that all possible treatment levels are being tested, as opposed to *random*; in which only a subset of all the possible treatments has been selected to be tested. In the fixed effects case we will have a different treatment groups with n observations under each treatment level and we begin by estimating the model

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{1.2}$$

for $i = 1, 2, ...a$ and $j = 1, 2, ..., n$ and where

- Y_{ij} is the ij^{th} observation

- μ is the overall mean
- τ_i is the i^{th} treatment effect
- ϵ_{ij} is random error.

From this we can see that μ is constant for all observations while ϵ_{ij} changes for every observation and τ_i changes for each treatment group. Also, under a balanced fixed effects model there are $N = an$ total observations. When using this ANOVA model the ϵ_{ij} are assumed to be independently and identically distributed as $N(0, \sigma^2)$ [14].

In order to test the hypotheses given in equation 1.1 we need to develop a test statistic. We now explore the total corrected sum of squares defined as

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 \tag{1.3}$$

where the dot notation indicates summation over the variable that has been replaced with a dot that is, $\sum_i y_{ij} = y_{.j}$. Thus, combining the dot notation and bar notation, which indicates an average, we have $\bar{y}_{..} = (\sum_{i=1}^{i}\sum_{j=1}^{j} y_{ij})/N$.

Variance is a measure of how far observations are spread out from the mean. Mathematically it is the expected squared differences of each observation from the population mean. Notice then, that if SS_T is divided by its degrees of freedom, $N - 1$, the result is the sample variance of the Y's [12]. Therefore, the total corrected sum of squares is a valid measure of the variability within the data. Additionally, we can rewrite it as

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}((y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..}))^2 .$$

Then, if we expand the squared term and simplify (see section A.1 of the Appendix for the steps), we obtain

$$SS_T = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2 . \tag{1.4}$$

We can see that the first half of equation 1.4 is the sum of the treatment means minus the overall mean, which is the *sum of squares due to treatments.*

This difference is a measure of the difference between treatment means and, thus often thought of as the sum of squares between treatments. The second half of this equation is the sum of the individual observations minus their corresponding treatment means, which is the *sum of squares due to error* which can be thought of as the error within treatments (see figure 1 for a graphical representation). This demonstrates that the total sum of squares

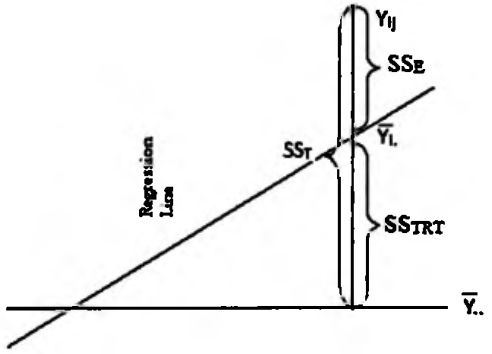


Figure 1: Graphical representation of the decomposition of the sums of squares.

is additive and we now have that

$$SS_T = SS_{TRT} + SS_E \tag{1.5}$$

where

$$SS_{TRT} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 \tag{1.6}$$

and

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2. \tag{1.7}$$

Now, since there are N total observations, the degrees of freedom for SS_T is N-1. One degree of freedom is lost because we use $\bar{y}_{..}$ to estimate the population mean, μ [13]. For the same reason, SS_{TRT} has $a-1$ degrees of freedom and SS_E has $N-a$ degrees of freedom. The reasoning for the latter is not as intuitive: we have n observations under each treatment lending

$n-1$ degrees of freedom to estimate the error within each treatment, but we have a treatments, thus we have $a(n-1) = an - a = N - a$ degrees of freedom in total. Notice that we have a similar situation here as we did with the sums of squares: the degrees of freedom are also additive ($df(SS_T) = df(SS_{TRT}) + df(SS_E)$).

Consider dividing SS_E by its degrees of freedom as we did with SS_T; when this is done we find that $SS_E/(N-a)$ is an estimate of the common variance, σ^2, within each of the a treatments [12]. Similarly, if there was no difference between the treatment means, then $SS_{TRT}/(a-1)$ could also be used to estimate σ^2. These terms are referred to as mean squared errors and we define them as

$$MS_E = \frac{SS_E}{N-a} \tag{1.8}$$

$$MS_{TRT} = \frac{SS_{TRT}}{a-1}. \tag{1.9}$$

Now consider the expected value of MS_E:

$$E(MS_E) = E\left(\frac{SS_E}{N-a}\right) = \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right) \tag{1.10}$$

$$= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2)\right)$$

$$= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - 2\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}\bar{y}_{i.} + n\sum_{i=1}^{a}\bar{y}_{i.}^2\right) \tag{1.11}$$

Looking at the middle term in equation 1.11, we can use the dot notation to indicate the summation over j and get $2\sum_{i=1}^{a} y_{i.}\bar{y}_{i.}$. Multiplying by n/n to get a $n(y_{i.}/n) = n\bar{y}_{i.}$ we have

$$E(MS_E) = \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - 2n\sum_{i=1}^{a}\bar{y}_{i.}^2 + n\sum_{i=1}^{a}\bar{y}_{i.}^2\right)$$

$$= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2 - n\sum_{i=1}^{a}\bar{y}_{i.}^2\right) \tag{1.12}$$

The last term of equation 1.12, is $\bar{y}_{i.}^2 = (y_{i.}/n)(y_{i.}/n)$. Since we can factor the n's out of the summation, one cancels with the n that is already there and we have $(1/n)\sum_{i=1}^{a} y_{i.}^2$. By converting the dot notation back to summations we get

$$E(MS_E) = \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n} y_{ij}\right)^2\right)$$

Substituting in the model from equation 1.2 on page 2 for y_{ij} we then have

$$E(MS_E) = \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n} (\mu+\tau_i+\epsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n} (\mu+\tau_i+\epsilon_{ij})\right)^2\right)$$

Additionally, since $E(\epsilon_{ij}) = 0$, $E(\epsilon_{ij}^2)$ can be replaced with σ^2 and $E(\epsilon_{i.}^2)$ replaced with $n\sigma^2$ [12]. Finally, after simplifying (see section A.2 of the Appendix for the steps) we have that

$$E(MS_E) = \sigma^2 \tag{1.13}$$

Through similar methods it can also be found that

$$E(MS_{TRT}) = \sigma^2 + n\sum_{i=1}^{a}\frac{\tau_i^2}{a-1}. \tag{1.14}$$

From these two equations it is evident that if there is no difference in the treatment means then $\sum \tau_i^2 = 0$ and $E(MS_{TRT}) = E(MS_E) = \sigma^2$. And if there is a difference in the treatment means then $E(MS_{TRT}) > E(MS_E)$. This is one reason MS_{TRT} and MS_E can be used to perform our hypothesis test. The major condition that allows the use of these values for hypothesis testing comes from Cochran's Theorem.

Cochran's Theorem. *Let Z_i be independently and identically distributed as N(0,1) for $i = 1, 2, ..., v$ and*

$$\sum_{i=1}^{v} Z_i^2 = Q_1 + Q_2 + \ldots + Q_s$$

where $s \leq v$, and Q_i has v_i degrees of freedom ($i = 1, 2, ...s$). Then the $Q_1, Q_2, \ldots, Q_s$ are independent chi-square random variables with $v_1, v_2, \ldots, v_s$ degrees of freedom, respectively, if and only if $v = v_1 + v_2 + \ldots + v_s$.

Cochran's Theorem says that since the ϵ_{ij} are independently distributed as $N(0,1)$, $SS_T = SS_{TRT} + SS_E$, and $df(SS_T) = df(SS_{TRT}) + df(SS_E)$, then $SS_{TRT}/(a-1)$ and $SS_E/(N-a)$ are independently distributed Chi-square random variables [12]. This means that if H_0 is true then

$$F_0 = \frac{MS_{TRT}}{MS_E} \tag{1.15}$$

is distributed as F with $a-1$ and $N-a$ degrees of freedom. We can use F_0 as a test statistic for performing the hypothesis test given in equation 1.1. From equations 1.8 and 1.9 we can see that if there is a difference in the treatment means, then the numerator of equation 1.15 will be larger than the denominator. Thus, the larger F_0 is, the more likely it is that there is a difference in the treatment means. More formally for a given significance level α, if

$$F_0 > F_{\alpha,\, a-1,\, N-a} \tag{1.16}$$

we can reject H_0 and conclude that at least two of the treatment means are statistically different.

2 Repeated Measures ANOVA

Repeated Measures ANOVA extends the basic concepts of one-way classification ANOVA but differs from it in a couple of ways. First, subjects must be measured at every level of a treatment — this is what makes the use of repeated measures techniques necessary. As a result of this property, the assumption of independence of the individual observations is no longer valid. By knowing which subject is being considered we are given information about that observation from the subject's other observations; thus making each subject's observations dependent upon one another. How we deal with the loss of independence will be discussed later. An advantage of using a repeated measures design is that we can now see how a single person reacts to each of the different treatment levels, instead of just one level. This allows for a better comparison of the different treatment levels.

2.1 Design

In a repeated measures design subjects must be crossed with at least one treatment and can also be nested within treatments. Treatments that subjects are considered to be *crossed* with are those that they are repeatedly measured across; each subject receives every level of that factor. Those factors in which subjects are *nested* are factors subjects receive only one level of (the non-repeated treatments). For example, say we split the subjects into 2 groups and put one group in classroom A and the other in classroom B, and then, within both of the classrooms, students take each of 3 different tests. In this design the subjects are nested in the classrooms and crossed with the tests. In the one-way classification ANOVA subjects were only nested in treatments. In the repeated measures case they must be crossed with at least one treatment.

A basic repeated measures design has J subjects crossed with one treatment, with I levels of that treatment. The model takes on the form

$$Y_{ij} = \mu + A_i + s_j + As_{ij} + \epsilon_{ij} \tag{2.1}$$

for $i = 1, 2, ...I$; $j = 1, 2, ..., J$ where

- Y_{ij} is the observed value,

- μ is the grand mean,
- A_i is the treatment (fixed) effect for the i^{th} treatment level,
- s_j is the j^{th} subject effect,
- As_{ij} is the interaction effect of the j^{th} subject with the i^{th} treatment level, and
- ϵ_{ij} is random error.

This is referred to as a Treatment by Subjects design. Table 1 shows the respective degrees of freedom (df), expected mean squares ($E(MS)$), and the appropriate F_0 ratio for testing the significance of each of the terms in equation 2.1.

Since there are I treatment levels and J subjects, we have IJ total degrees of freedom. As in the simple ANOVA, we lose one degree of freedom for estimating the mean treatment effect and the mean subject effect. Hence, we have $I-1$ and $J-1$ in the df column of table 1 for A and S, respectively. We use these degrees of freedom to estimate mean squares. After estimating the mean squares for the treatments, subjects, and treatment by subject interaction we have used $IJ-1$ of the initial IJ total degrees of freedom since $(I-1)+(J-1)+(I-1)(J-1) = IJ-1$. We then lose one more degree of freedom for estimating the mean error and because there are 0 degrees of freedom left to use we cannot estimate the mean square error. This results in there being no F-test for the subject effect [2].

Source	*df*	$E(MS)$	F_0
A	$I-1$	$\sigma_\epsilon^2 + J\sigma_A^2 + \sigma_{AS}^2$	MS_A/MS_{AS}
S	$J-1$	$\sigma_\epsilon^2 + I\sigma_S^2$	
$A*S$	$(I-1)(J-1)$	$\sigma_\epsilon^2 + \sigma_{AS}^2$	
Error	0	σ_ϵ^2	

Table 1: Summary table for a Treatments x Subjects Design.

Subjects can be exposed to any combination of crossed and nested treatments but, of course, as we add more treatments the models and calculations

get more complex very quickly. For instance, by adding one more crossed treatment, B, with K levels the model becomes

$$Y_{ijk} = \mu + s_j + A_i + As_{ij} + B_k + Bs_{kj} + AB_{ik} + ABs_{ijk} + \epsilon_{ijk} \qquad (2.2)$$

for i and j as before and $k = 1, 2, \ldots, K$. And adding a nested treatment, D, with M levels, to the design in equation 2.2 gives us

$$\begin{aligned} Y_{ijkm} \;=\; & \mu + D_m + s_{j(m)} + A_i + AD_{im} + As_{ij(m)} + B_k + BD_{km} \\ & + Bs_{kj(m)} + AB_{ik} + ABD_{ikm} + ABs_{ikj(m)} + \epsilon_{ijkm} \end{aligned} \qquad (2.3)$$

for $m = 1, 2, \ldots, M$. The parentheses in the subscripts of terms involving s indicate that s is nested in D because m is the corresponding subscript for D. This is important because we now have J subjects under each of the M treatment levels of D, giving us MJ total subjects. Thus, $s_{j(m)}$ indicates the j^{th} subject under the m^{th} treatment level of D.

When models have both crossed and nested treatments they are called mixed designs. We can split the treatments up into ones that create between-subject variation and ones that create within-subject variation. *Between-subject* variation occurs between the different subjects (*e.g.* the variation from subject 1 to subject 2) and it comes from the subjects themselves and the treatments the subjects are nested in. *Within-subject* variation is due to the inherent variation of a given subject's scores; subjects don't always have the same exact reaction to the same treatment so of course there will be differences from treatment to treatment. The sources of within-subject variation consist of the treatments the subjects are crossed with, the interactions of those crossed treatments with the subjects, and the interactions of the crossed treatments with the nested treatments.

These two categorizations are useful for organizing the different treatments and to see which interactions we should be considering. Table 2 shows how a summary table is set up for the model in equation 2.3. You can see that the between-subject sources of error are listed on the top and all of the within-subject sources of error are below. Within these sections we group the interactions of treatments. Notice that each term in the table coincides with a term in Equation 2.3 and, as before, because of the lack of degrees of freedom there are still no F-tests for the terms involving the subjects.

In the case of two nested treatments, the interaction of those two treatments would be listed under the between-subjects sources. Consider treatments A, D, and C which have I, J, and K levels, respectively. Then the

Source	*df*	$E(MS)$	F_0
Between	$JM-1$		
D	$M-1$	$\sigma^2_\epsilon + IJK\sigma^2_D + IK\sigma^2_S$	MS_D/MS_S
S	$M(J-1)$	$\sigma^2_\epsilon + IK\sigma^2_S$	
Within	$IJKM-JM$		
A	$I-1$	$\sigma^2_\epsilon + JKM\sigma^2_A + K\sigma^2_{AS}$	MS_A/MS_{AS}
AD	$(I-1)(M-1)$	$\sigma^2_\epsilon + JK\sigma^2_{AD} + K\sigma^2_{AS}$	MS_{AD}/MS_{AS}
AS	$(I-1)M(J-1)$	$\sigma^2_\epsilon + K\sigma^2_{AS}$	
B	$K-1$	$\sigma^2_\epsilon + IJM\sigma^2_B + I\sigma^2_{BS}$	MS_B/MS_{BS}
BD	$(K-1)(M-1)$	$\sigma^2_\epsilon + IJ\sigma^2_{BD} + I\sigma^2_{BS}$	MS_{BD}/MS_{BS}
BS	$(K-1)M(J-1)$	$\sigma^2_\epsilon + I\sigma^2_{BS}$	
AB	$(I-1)(K-1)$	$\sigma^2_\epsilon + JM\sigma^2_{AB} + \sigma^2_{ABDS}$	MS_{AB}/MS_{ABDS}
ABD	$(I-1)(K-1)(M-1)$	$\sigma^2_\epsilon + J\sigma^2_{ABD} + \sigma^2_{ABDS}$	MS_{ABD}/MS_{ABDS}
$ABDS$	$(I-1)(K-1)M(J-1)$	$\sigma^2_\epsilon + \sigma^2_{ABDS}$	

Table 2: Summary table for a design with Subjects nested in D and crossed with A and B.

model for a design with subjects nested in treatments C and D and crossed with treatment A where there are M subjects under each nested treatment level would be

$$\begin{aligned} Y_{ijkm} &= \mu + D_j + C_k + DC_{jk} + s_{m(j,k)} + A_i + AD_{ij} \\ &\quad + AC_{ik} + As_{im(j,k)} + ADC_{ijk} + \epsilon_{ijkm}. \end{aligned} \tag{2.4}$$

Combining the structures of equations 2.2, 2.3, and 2.4 one can see how to create a model for any combination of crossed and nested treatments. Or, an easier method may be to first create the summary table and then use it to write the corresponding model.

2.2 Independence

As mentioned earlier, because we have multiple observations for each subject, the assumption of independence of the observations is no longer valid. It was this assumption that allowed the use of F-tests to test the hypotheses. *Sphericity*, which is the requirement that the variability of the differences of a single subject's scores for every pair of treatment levels have equal variability, is often upset by the observations not being independent. These

problems ultimately result in the mean square ratios not having an exact F-distribution with the previously defined degrees of freedom. In an effort to clarify sphericity, we will first discuss what can be done with the degrees of freedom.

Consider the Treatment by Subject design in equation 2.1 on page 8. Box has shown that because of the lack of independence of the observations [3], F-tests performed using the original degrees of freedom are no longer valid. However, by using a multiplicative factor that measures the departure from sphericity, ε, on our degrees of freedom, a corrected F ratio may be found. The new degrees of freedom are $\varepsilon\,(I-1)$ and $\varepsilon\,(I-1)\,(J-1)$ [10]. This ε is a function of the population variances and covariances and can be estimated using the sample variances and covariances but it is not always necessary to calculate an estimate. With that said, there are a few ways to deal with how ε might change the decision of whether or not H_0 should be rejected but before we go into any more detail we must introduce variance of difference scores.

Given treatment levels i and i' of treatment A, the *difference score* for the j^{th} subject is defined as $d_{jii'} = y_{ij} - y_{i'j}$. Then the *variance of the difference scores* for treatment levels i and i', denoted $\sigma^2_{dii'}$, is the variance of the difference scores of treatment levels i and i' for every subject. If $\sigma^2_{dii'} = \sigma^2_{dll'}$ for every pair of levels i and i' and l and l' of treatment A, then the variances of the difference scores are said to be *homogeneous.*

The level of homogeneity that is present within the variances of the difference scores affects the value of ε. If the variances are completely homogeneous then $\varepsilon = 1$, thus giving us the original degrees of freedom $(I-1)$ and $(I-1)(J-1)$ [9]. Under the most extreme case of heterogeneity $\varepsilon = 1/(I-1)$, resulting in 1 and $(J-1)$ degrees of freedom [2]. From this we can see that no matter the amount of homogeneity that is present within the variances of the difference scores ε ranges from 1 to $1/(I-1)$ which tell us that the adjusted degrees of freedom range from $I-1$ and $(I-1)(J-1)$ to 1 and $J-1$.

If we perform an F-test with the original degrees of freedom without adjusting by ε, the test statistic obtained from an F table may be biased upward and thus increase the probability of type I error [2]. Meaning we would reject H_0 when H_0 is indeed true more often than our chosen level of significance indicates. That is why adjusting the degrees of freedom by ε is

necessary. However, because we know the maximum and minimum values for ε, we do not have to actually estimate ε every time. Greenhouse and Geisser suggest performing two F-tests: one with $I-1$ and $(I-1)(J-1)$ degrees of freedom and the other with 1 and $J-1$ degrees of freedom. There are three possible outcomes for the two tests:

Case 1: H_0 is *not* rejected with $I-1$ and $(I-1)(J-1)$ degrees of freedom.

If $F_0 < F_{\alpha,\, I-1,\, (I-1)(J-1)}$ then $F_0 < F_{\alpha,\, 1,\, (J-1)}$. Since the degrees of freedom used for the first test are larger than those used for the second test, the critical F value used in the second test will be larger than the one used in the first test (see section A.3 of the Appendix). Thus, if F_0 is smaller than the first value, it will definitely be smaller than the second (larger) value. Showing that no matter what ε is, $F_0 < F_{\alpha,\, \varepsilon(I-1),\, \varepsilon(I-1)(J-1)}$ when we have this initial result. [1]

Case 2: H_0 is rejected with 1 and $J-1$ degrees of freedom.

If $F_0 > F_{\alpha,\, 1,\, (J-1)}$ then $F_0 > F_{\alpha,\, I-1,\, (I-1)(J-1)}$. Since 1 and $J-1$ are less than $I-1$ and $(I-1)(J-1)$, the critical value obtained from the F-table in the second test will be smaller than the one obtained in the first test (see section A.3 of the Appendix). And, similarly to the first case, since F_0 is greater than the first value it will be greater than the second (smaller) value also. So, since the actual degrees of freedom range from 1 and $J-1$ to $I-1$ and $(I-1)(J-1)$ and H_0 is rejected at both of these extreme points. No matter what the degrees of freedom are supposed to be H_0 will be rejected in this case.

Case 3: H_0 is rejected under the first test ($df = I-1$, $(I-1)(J-1)$) but *not* rejected under the second test.

In this case, since the actual degrees of freedom are somewhere in the range of the two sets used, we cannot make an accurate decision. The critical value in the first test may be biased upward and the critical value in the second test may be biased downward, but we have no way of knowing how much. The only way to get rid of the ambiguity is to actually compute ε.

[1] Note that since $F_0 = MS_A/MS_{AS} = \frac{(SS_A/df_A)}{(SS_{AS}/df_{AS})}$, ε will cancel out and thus does not have an effect on F_0.

We can see that when the two F-tests are consistent, we can use the decisions that the tests give us. But if the two tests yield different results, we must estimate ε in order to get a valid result.

2.3 Estimating ε

In order to estimate epsilon we will need to use variances and covariances so we must first discuss them.

2.3.1 Variance and Covariance

The *variance* of a given variable is the expected value of the squared differences of each of the observed values for that variable from the expected value of that variable. It is denoted by σ^2 and can be expressed as $E[(X-E(X))^2]$. For our purposes we use the following definition of population variance.

$$\sigma_X^2 = \frac{\sum_{j=1}^{N}(X_j - \mu_X)^2}{N}$$

where μ_X is the population mean of the variable X and X_j is the j^{th} observation out of N total observations of X. This formula can easily be rewritten as

$$\sigma_X^2 = \frac{\sum_{j=1}^{N}(X_j - \mu_X)(X_j - \mu_X)}{N}. \tag{2.5}$$

Variances can be computed for each variable that is present in a design. We can also change equation 2.5 ever so slightly to get the *covariance* between variable X and variable Y:

$$Cov_{X,Y} = \frac{\sum_{j=1}^{N}(X_j - \mu_X)(Y_j - \mu_Y)}{N} \tag{2.6}$$

From equations 2.5 and 2.6 we can see that what distinguishes a variance from a covariance is the use of two different variables instead of just one.

When variances and covariances are estimated for sample populations the equation must be altered slightly. $S_{ii'}$ is used to indicate an estimated variance or covariance and, since we don't know the actual population mean, we must use our present data to estimate the means. We define the sample covariance to be

$$S_{ii'} = \frac{\sum_{j=1}^{N}(y_{ij} - \overline{y_{i\cdot}})(y_{i'j} - \overline{y_{i'\cdot}})}{N-1} \tag{2.7}$$

where i and i' are two variables, y_{ij} and $y_{i'j}$ are the j^{th} observations for their respective variables, and $\overline{y_{i.}}$ and $\overline{y_{i'.}}$ are the i^{th} and $(i')^{th}$ variable means. Also, we now have $N-1$ in the denominator instead of N because we lost one degree of freedom for estimating the mean of the i_{th} variable. It can be seen that if $i = i'$ then $S_{ii'}$ is a variance estimate, and if $i \neq i'$, $S_{ii'}$ is an estimate of the covariance between variables i and i'.

When we apply these computations to the situation of a Treatment by Subjects repeated measures design with I treatments and J subjects, i and i' are two levels of a specific factor. We let y_{ij} be the observation of the j^{th} subject under the i^{th} treatment level and, similarly, $\overline{y_{i.}}$ is the i^{th} treatment mean. Furthermore, because we are summing over the subjects we can replace N with J. A given treatment level will have one variance and $I-1$ covariances – one with each of the other $I-1$ treatment levels.

If we expand the product in equation 2.7 we can simplify the computation to one that is much easier to deal with. We have (using J instead of N)

$$S_{ii'} = \frac{\sum_{j=1}^{J}(y_{ij}y_{i'j} - y_{ij}\overline{y_{i'.}} - y_{i'j}\overline{y_{i.}} + \overline{y_{i.}}\,\overline{y_{i'.}})}{J-1} \tag{2.8}$$

Since summations are distributive over addition and terms that are constant can be factored out we have

$$\begin{aligned} S_{ii'} &= \frac{\sum_{j=1}^{J} y_{ij}y_{i'j} - \sum_{j=1}^{J} y_{ij}\overline{y_{i'.}} - \sum_{j=1}^{J} y_{i'j}\overline{y_{i.}} + \sum_{j=1}^{J} \overline{y_{i.}}\,\overline{y_{i'.}}}{J-1} \\ &= \frac{\sum_{j=1}^{J} y_{ij}y_{i'j} - \overline{y_{i'.}}\sum_{j=1}^{J} y_{ij} - \overline{y_{i.}}\sum_{j=1}^{J} y_{i'j} + J\overline{y_{i.}}\,\overline{y_{i'.}}}{J-1} \\ &= \frac{\sum_{j=1}^{J} y_{ij}y_{i'j} - \overline{y_{i'.}}y_{i.} - \overline{y_{i.}}y_{i'.} + J\overline{y_{i.}}\,\overline{y_{i'.}}}{J-1} \end{aligned} \tag{2.9}$$

In the last equality the dot notation has been used to indicate a summation over j: $\sum_{j=1}^{J} y_{ij} = y_{i.}$. Now, if we multiply the second and third terms by $\frac{J}{J}$ (a fancy 1) we have $\frac{J}{J}\overline{y_{i'.}}y_{i.} = J\overline{y_{i'.}}\frac{y_{i.}}{J} = J\overline{y_{i'.}}\,\overline{y_{i.}}$ for the second term and, similarly, the third term becomes $J\overline{y_{i.}}\,\overline{y_{i'.}}$. Substituting these into equation

i	1	2	...	I
1	S_{11}	S_{12}	...	S_{1I}
2	S_{21}	S_{22}		S_{2I}
$\vdots$	$\vdots$		$\ddots$	
I	S_{I1}	S_{I2}		S_{II}

Table 3: Variance-Covariance Matrix for I treatment levels

2.9 we have

$$\begin{aligned} S_{ii'} &= \frac{\sum_{j=1}^{J}(y_{ij}y_{i'j}) - J\overline{y_{i'.}}\,\overline{y_{i.}} - J\overline{y_{i.}}\,\overline{y_{i'.}} + J\overline{y_{i.}}\,\overline{y_{i'.}}}{J-1} \\ &= \frac{\sum_{j=1}^{J}(y_{ij}y_{i'j}) - J\overline{y_{i.}}\,\overline{y_{i'.}}}{J-1} \\ &= \frac{\sum_{j=1}^{J}(y_{ij}y_{i'j}) - J\frac{y_{i.}}{J}\frac{y_{i'.}}{J}}{J-1} \\ &= \frac{\sum_{j=1}^{J}(y_{ij}y_{i'j}) - (y_{i.}\,y_{i'.})/J}{J-1} \end{aligned} \tag{2.10}$$

After computing the variances and covariances it is helpful to arrange them into what is called a variance-covariance matrix. Table 3 demonstrates this and we can see that the variances are down the diagonal and the off diagonal values are the covariances between the corresponding pairs of treatment levels.

Before the estimation of ϵ is introduced, it is useful to go over some notation. Once again the dot notation will be used to indicate summation over the variable that has been replaced by a dot. We define $S_{i.}$ to be the average of the i_{th} factor level's variance and the $I-1$ covariances. Relating this to the variance-covariance matrix, $S_{i.}$ would be the average of the values in the i^{th} row (or column since the i^{th} row equals the i^{th} column). Similarly, $S_{..}$ is the average of all of the variances and covariances; which is equivalent to the average of the $S_{i.}$.

Geisser and Greenhouse [5] have shown that an estimator for the value

of ε can be computed by

$$\hat{\varepsilon} = \frac{I^2 \left(\frac{\sum_{i=1}^{I} S_{ii}}{I} - S_{..} \right)^2}{(I-1) \sum_{i}^{I} \sum_{i'}^{I} S_{ii'}^2 - 2I \sum_{i}^{I} S_{i.}^2 + I^2 S_{..}^2} \tag{2.11}$$

The estimate for ε can be broken down into:

- $\sum_{i=1}^{I} S_{ii}$ is the sum of the variances: $S_{11} + S_{22} + ... + S_{II}$,
- $\sum_{i}^{I} \sum_{i'}^{I} S_{ii'}^2$ is the sum of the squares of each of the variances and covariances: $S_{11}^2 + S_{12}^2 + ... + S_{I(I-1)}^2 + S_{II}^2$,
- $\sum_{i}^{I} S_{i.}^2$ is the sum of the squares of the $S_{i.}$,

and lastly we have that

- $S_{..}^2$ is the square of the average of all of the variances and covariances.

Collier, Baker, Mandeville, and Hayes (1967) and Stoloff (1970) investigated the effects of using $\hat{\varepsilon}$ on significance levels. Using simulations they found that by using this adjustment on the degrees of freedom, the actual level of significance that results from use of the new critical F value comes very close to the chosen level of significance [10]. It is much closer than when no adjustment is made and heterogeneity of the variances of the differences scores exists. However, they also found that when $\hat{\varepsilon}$ is near, or above .75, $\hat{\varepsilon}$ is badly biased.

When $\hat{\varepsilon}$ is around .75 it ends up under estimating the degrees of freedom. This results in a significance level smaller than the one chosen. The reason for this bias stems from the fact that even though the population may have homogenous variances of difference scores, a sample of that population "can always be expected to evidence some heterogeneity" of variances of difference scores [10]. This case results in an unnecessary reduction of the degrees of freedom. In order to correct for this bias, Huynh and Feldt derived the following correction factor.

$$\tilde{\varepsilon} = \frac{J(I-1)\hat{\varepsilon} - 2}{(I-1)(J-1-(I-1)\hat{\varepsilon})} \tag{2.12}$$

This estimator of ε, compared to the one presented in equation 2.11, is less biased and less dependent on a large sample size when there is only

slight heterogeneity of variances of difference scores present [10]. Since it does not make sense to estimate ϵ by a value larger than 1, we make $\tilde{\epsilon}$ equal to 1 whenever it is greater than this upper bound. Furthermore, $\tilde{\epsilon} \geq \hat{\epsilon}$ for every number of treatment levels and subjects and, the equality holds when $\hat{\epsilon} = 1/(I-1)$.

For a design that also has nested factors we can compute these estimators in a similar way. First, it is assumed that the K independent groups, one for each of the K levels of the nested treatment, have the same variance-covariance matrix [10]. The formula for $\hat{\epsilon}$ is the same as before except that now corresponding values from all of the groups are being used to compute the estimates. With N being the total number of subjects, the computation for $\tilde{\epsilon}$ becomes

$$\tilde{\epsilon} = \frac{N(I-1)\hat{\epsilon} - 2}{(I-1)(N-K-(I-1)\hat{\epsilon})} \tag{2.13}$$

Huynh and Feldt found, through Monte Carlo methods, that when the parameter ϵ is around .5, $\hat{\epsilon}$ is a better estimate. In addition, their work upheld their intentions of $\tilde{\epsilon}$ being the better estimate when $\epsilon > .75$ [10]. Not surprisingly, they also found that an increase in sample size results in less bias for both estimators.

2.4 Sphericity

Sphericity is the necessary and sufficient condition for the mean square ratios to have an exact F-distribution [9] and it can be discussed in several different contexts. The definition of sphericity that fits best with the other topics of this paper is as follows. If the variances of the difference scores are homogeneous then the variance-covariance matrix is said to be *spherical*. Recalling our previous discussion of the relationship between variances of difference scores and ϵ we can see this is equivalent to ϵ being equal to 1.

Equivalent definitions are in the context of matrices. If we let Σ be the variance-covariance matrix and M be an $I-1$ by I matrix of orthonormal vectors, then the property of sphericity is met if $M\Sigma M^T = \lambda I$ where λ is a constant and I is the $I-1$ by $I-1$ identity matrix [2]. This says that, after this transformation, the variances of each factor level must all be equal and the covariances between the different factor levels must all equal 0 in order for sphericity met. Huynh and Feldt defined a *type H matrix* to be a

variance-covariance matrix that demonstrates equal variances of differences between pairs of scores. Another explanation using matrices is that a type H matrix has the sphericity property that $\Sigma = A + A^T + \lambda I_I$ where I_I is the I by I identity matrix and, once again, λ is some constant and Σ is the variance-covariance matrix [8].

We can now use ε as a measure of the extent to which the variance-covariance matrix departs from sphericity. Values of ε near 1 indicate the variance-covariance matrix is spherical. Similarly, the closer ε is to $1/(I-1)$ the more severe the departure from sphericity. Note that when there are only 2 levels of the crossed treatment sphericity is not an issue. This is because there will only be one set of difference scores and thus only one variance of difference scores. If nested treatments are present then sphericity must be present in the variance-covariance matrices for each group of subjects.

Adjusting the degrees of freedom by ε acts as a correction for a lack of sphericity. However, as previously mentioned, this correction can only help to approximate an F-distribution — it is not as good as if the assumption of sphericity were upheld. The obtained results are still valid, but once sphericity is violated, in order to get the most accurate results, it is best to move on to methods that are more flexible in dealing with a lack of sphericity.

3 Application of ANOVA Techniques to Real World Data

In order to test ANOVA techniques I will be using data on mice activity levels. This data comes from a study investigating autism in mice conducted at the University of Redlands and led by Professor Ryan. However, for the purposes of this paper, only a statistical analysis will be performed. I do not have a background in mice or autism and thus cannot interpret what the following results indicate about the autism in mice.

The autism data set consists of one crossed treatment — post-natal days (age), and two nested treatments — cage type and strains of mice; all of which are considered to be fixed. There are 3 different strains of mice in this study: C57, C58, and FVB; and 3 different cage types used in the study: new, old, and enriched. Each litter of mice is placed in a cage with a unique cage number and the type of cage is either new, old, or enriched. Also each litter comes from one of the three strains mentioned. Starting when the mice are 2 days old, data is collected every other day up until the mice are 12 days old. The collected variables include measures such as weight, distance traveled when allowed to move around, time it takes to turn over when placed on their back, and other measures of each mouse's activity levels.

The age of each mouse, denoted PND, acts as a time variable and each response is repeatedly measured over each PND "level" – making PND the crossed, or within-subject, treatment. However, because we have no way of distinguishing the mice from one another when they are in the cage together, we have no way of matching up individual mouse records from day to day. To deal with this response outcomes within each cage for each PND level were averaged (see Appendix A.4 for the SAS code used to perform this transformation). For example, cage 1 had 3 mice in it so at each PND level we have 3 measurements for each response variable. But we cannot match the observations under PND 2 to those under PND 4 so we take the 3 values under PND 2 and average them for each of the separate response variables; this is done for each PND level and for each cage of mice. Now, instead of each mouse being a subject, the unique cages are the subjects. Table 4 shows the design of the data.

By taking the averages we drop from 2,831 observations to only 429 observations. This alters the degrees of freedom available to use when performing

F-tests; for the analysis of the original data we have enough degrees of freedom to estimate the mean squared error but this is not the case for the averaged data. Because of this, the design for the analysis of the original data includes F ratios for the subject (cage) effect and for the subject by PND interaction term while the design for the analysis for the averaged data does not.

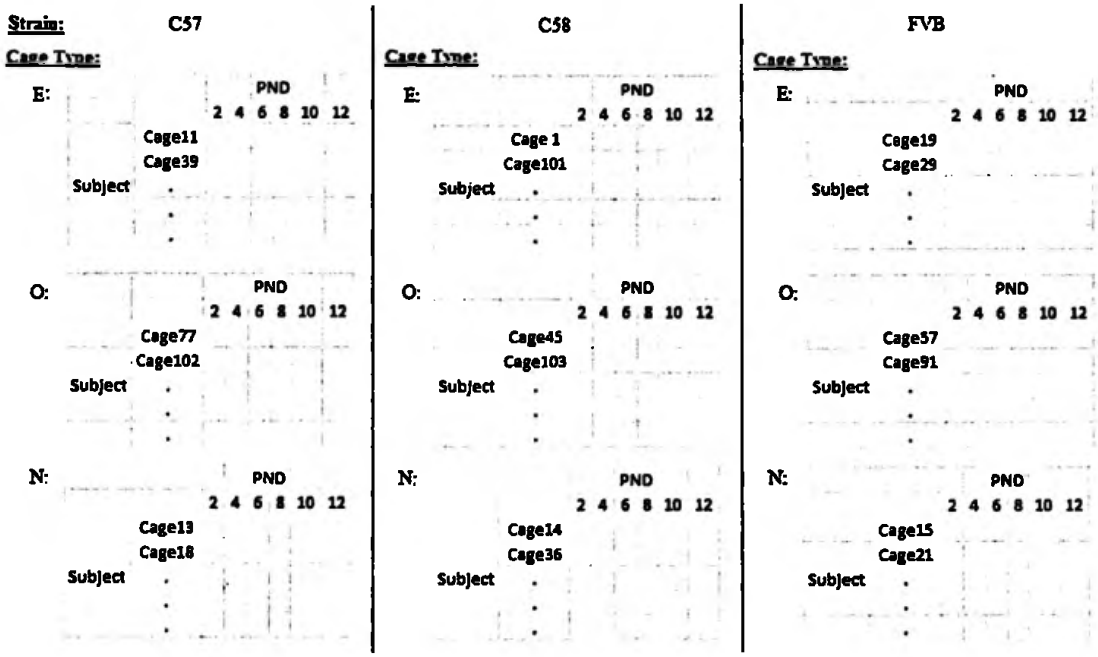


Table 4: Visual display of the design set up.

In this data set there are two types of unbalance present; one due to missing values over time and another due to different group sizes. Most of the variables in the original data set have missing values, however because there are some observations for the specific cage under the given PND it does not result in a missing value in the averaged data set. Also, some cages had PND's that were skipped, but in these instances the skipped PND was completely left out instead of entering missing values; this does not create missing values in either version of the data. It instead contributes to the lack of balance in the data. These skipped PND's for some cages result in

a lack of balance due to unequal group sizes; in the averaged data there is an unequal number of observations for each PND. Furthermore, because the mice are placed in the cages by litter, there is an unequal number of mice in each cage. Thus there are unequal group sizes and an unequal number of observations under each PND from cage to cage. Taking the average of the response variables by cage and PND helps correct for a lack of balance due to unequal litter sizes.

Additionally, because tests are being performed for each response variable there are issues with the experiment wide significance level, α. The chosen α level, generally .05, may be appropriate for each individual significance test however, when it is used for every test, the experiment wide significance level is actually much higher than our chosen α [15]. This is because the more tests that are performed on a set of data, the more likely we are to reject H_0 when it is true, referred to as a Type I error [1]. When tests are being performed in this manner it is referred to as a problem with *Multiple Comparisons*. A common correction for the higher probability of Type I error and the inflation of α is to divide α by the number of tests being performed, called the *Bonferroni Correction*. While other corrections exist, the Bonferroni is the simplest and most conservative correction for multiple comparisons.

The analysis of this data set will involve 7 individual significance tests because there are 7 response variables. Therefore, an experimental wide significance level of $\alpha = .05$ translates into significance levels of .00714 for the individual tests. If an experimental wide significance level of .01 is desired, then the tests for the individual variables would need to use .00143 as the significance level.

The Repeated Measures Analysis of Variance option in NCSS is used to perform the needed analysis. The output of this method presents an ANOVA summary table with a breakdown of all of the necessary values discussed in the previous parts of this paper (degrees of freedom, mean squares, sums of squares, F-ratios, and probability levels). Additionally, $\hat{\epsilon}$ and $\tilde{\epsilon}$ are computed and F-tests using the corrected degrees of freedom are performed. Running this analysis for each response variable we get information on the effect that each treatment has on the response variables. In addition to using SAS to create an averaged data set, SAS was also used to compute descriptive statistics, see Appendix A.4. The package R was also used to create relevant

plots of means by the different treatments.

In order to see how transforming the data affects the results, analyses on both the original data and the data that has been averaged over the PND levels within each cage were performed. ANOVA tables for each variable can be seen in Section A.5 of the Appendix.

3.1 Corrected Negative Geotaxis

Corrected Negative Geotaxis, denoted CORNEGGEO, measures the amount of time, in seconds, it takes a mouse to turn around when they are placed on an inclined plane with a maximum cutoff time of 30 seconds. The overall mean for this variable is 12.71 and the means by strain are approximately 11.4, 11.1, and 14.3 for C57, C58, and FVB respectively. The means by cage type are approximately 12.33 for enriched cages and 12.9 for both new and old cages. Note that there is not much difference in the means between the two data sets. There were 701 observations in which the maximum value of 30 was recorded; this drags the means up quite a bit, note that the overall median is only 7.03. From the means by PND, we can see that as the mice get older they seem to be quicker to turn around. With means of approximately 26 under PND 2 and means of approximately 6 under PND 12 in both sets of data it appears that these maximum values are coming from the younger mice.

Going from the original data to the data that has been averaged did not cause any changes in significance. Using a significance level of 0.00714 we find that strain and PND have a significant effect on the times it takes a given mouse to turn around when placed on an incline, with p-values < 0.0001 and < 0.0001 respectively. In the results for the analysis of the original data the subject (cage) effect does have a significant impact on CORNEGGEO as does the interaction of subject with PND. Referring to the means stated above, the FVB strain takes longer to turn around than the two other strains. These relationships can be seen in Plot 2(a) of section A.6 of the Appendix.

The values for $\hat{\epsilon}$, referred to as Geisser-Greenhouse Epsilon, and $\tilde{\epsilon}$, referred to as Huynh-Feldt Epsilon, also do not change from the original data to the averaged data. $\hat{\epsilon} \approx 0.666$ and $\tilde{\epsilon} \approx 0.792$ for both data sets. These high values (close to 1) indicate that there is slight heterogeneity of variances of difference scores and thus the assumption of sphericity is violated. When

F-tests are performed using the degrees of freedom that have been adjusted by epsilon the significance of the effects did not change.

3.2 Rearing

The variable REAR measures the number of times the mouse rears up on its hind legs when allowed to move around. There are 2 observations that are missing a value for this variable. Although the range of this variable is 0 to 64, the overall mean is only 0.53 and the median 0 – this is a result of 2520 out of the 2829 total observations being 0.

Examining the F-ratios for this variable we see that, at the .00714 significance level, strain, PND, and the strain by PND interaction all have a significant effect on how much a mouse rears. The p-values from the averaged data are 0.0037, < 0.0001, and 0.0035 for strain, PND, and the strain by PND interaction respectively, and from the original data the p-values are (in the same order) 0.0019, < 0.0001, and 0.0005. Although there is more of difference between the p-values of the two different versions of the data than there is with most of the other variables, there still is not a big enough difference to change the outcomes. From the results on the original data we also find there is a significant subject (cage) effect and a significant subject by PND interaction effect (with both these p-values being < 0.0001).

Now, examining the epsilon values, we have that $\hat{\epsilon} \approx 0.368$ and $\tilde{\epsilon} \approx 0.426$ for both versions of the data. These are pretty small values for ϵ and thus indicate there is quite extreme heterogeneity of the variances of difference scores present causing the requirement of sphericity to be violated. When supplementary F-tests are performed to help correct for this issue the critical F-values increase which raises the p-values. However after this adjustment, the effect of PND is still significant and the strain by PND interaction effect is not. This indicates that the strain by PND interaction effect may be easily influenced by changes in the data. However, despite the lack of sphericity, we can be confident that strain and PND are associated with how much a mouse rears.

The means broken up by strain indicate that the FVB strain rears the most; followed by the C58 strain and then C57. Also, we find that the mice rear more as they get older. There is quite a large jump in how much a mouse rears going from PND 8 to PND 10, especially with the FVB strain.

Refer to Plots 2(b) and 2(c) in section A.6 the Appendix for plots of these interactions. Since the strain by cage-type interaction effect was significant in the averaged data model, it is worth noting that, overall, the mice in enriched cages rear more than mice in the other two cage types, and mice in new cages rear more than those in old cages.

3.3 Number of Squares Crossed

The variable NUMCROSS records the number of 2 centimeter squares a given mouse crosses when allowed to move around. With 3 missing values, this variable has a mean of roughly 14.58 and a median of 8.

When looking at the F-tests for NUMCROSS we find that strain, PND, and the strain, PND interaction all have significant effects on how much a mouse moves. In addition, the strain by cage type by PND also has a significant impact on NUMCROSS in the averaged data, but one of 0.018167 in the original data set. All four of these terms have extremely low p-values: the 3-way interaction term has a p-value $< .0001$ for the averaged data while the other three terms all have p-values < 0.0001 for both versions of the data. Aside from the p-value of 0.018167, these p-values would be significant at even the most strict significance levels. In the results for the original data we also have that subject and subject, PND interaction effects are significant with p-values < 0.0001. The p-values for the terms that were not significant were quite far from the 0.00714 cut off, ranging from 0.2 to 0.6.

For this variable we have epsilon values of $\hat{\epsilon} = 0.494$ and $\tilde{\epsilon} = 0.578$ for both versions of the data. The value of $\hat{\epsilon}$ indicates strong heterogeneity of the variances of difference scores and thus a violation of sphericity. Using $\hat{\epsilon}$ to adjust the degrees of freedom does not effect the significance of the effects on NUMCROSS. Although the strain by cage type by PND interaction term is not significant in the original data, its p-value in the averaged data is very low so it appears that this term does have a significant effect on NUMCROSS.

Examining the means for the number of squares crossed broken up by strain, it is evident that the C58 strain is the most active in terms of this variable, with a mean of approximately 23.5 squares crossed. The FVB strain has the second highest activity level, with a mean of approximately 17, and the C57 strain has a mean of approximately 7 squares crossed. Additionally, the older the mice are, the more they move around (Figure 2(d) in section

A.6 of the Appendix).

3.4 Righting

The variable RIGHT measures the amount of time it takes a mouse to turn over when placed on its back, referred to as righting itself. It is measured in seconds and has a maximum cutoff value of 30 seconds. There are 2 observations that do not have a value for this variable. Righting has an overall mean of approximately 3.67 and a median of 1.41 indicating that the 117 high values of 30 are pulling the mean up.

Once again the analysis shows that the strain, the PND, and the strain by PND interaction have significant effects on the response variable. With p-values < 0.0001 these effects are significant even at the strongest significance levels. Similarly, the subject effect and the subject by PND interaction effect are also significant with p-values < 0.0001.

Examining the values of $\hat{\varepsilon}$ and $\tilde{\varepsilon}$ we see that, since they are less than .75, $\hat{\varepsilon}$ is the less biased estimate, and it indicates a lack of homogeneity of the variances of difference scores. Once again these issues result in the sphericity requirement being violated. With the original p-values being so small, there is no evident change in the p-values when efforts are made to correct for this lack of sphericity by adjusting the degrees of freedom for the F-tests. Since the relevant p-values are so small it can be concluded that the strain effect, PND effect, and the strain PND interaction effect are all significantly different from zero, even with the lack of sphericity.

The means for righting split up by strain are approximately 4.8, 6.3, and 1.8 for C57, C58, and FVB respectively. These values indicate that the mice from the C58 strain take the longest to right themselves and the mice from the FVB strain are the fastest. As the mice get older they get better at righting themselves. When looking at the means grouped by strain and PND we find that as the mice get older the differences between the strains decrease. At PND 2 there are very large differences between the means for each of the strains, however by PND 8 these differences are drastically smaller and continue to decrease (see Figure 2(e) in section A.6 of the Appendix).

3.5 Pivots

The PIVOT variable measures the number of sideways movements a given mouse makes when it is allowed to move around. There are 2 observations that do not have a value recorded for this variable. Although the range of pivots is 0 to 17, there were 1036 observations in which the mouse pivoted 0 times and only 172 observations in which the mouse pivoted more than 5 times. This high concentration of small values shifts the data downward to have a mean of 1.75 and a median of 1.

Examining the F-ratios and p-values for the pivots variable we find that strain and PND have significant effects on this variable. Also, the strain by PND interaction shows a significant effect in the original data. The p-value for the strain by PND interaction effect in the original data is 0.0024 and in the averaged data the p-value is 0.0132. The p-value for the strain effect is < 0.0001 in both versions of the data and the PND effect has a p-value < 0.0001 for the averaged data and one of < 0.0001 for the original data. These values make it evident that strain and PND would have significant impacts on how much a mouse pivots at all reasonable significance levels.

Now moving to the estimates of epsilon we find that $\hat{\varepsilon} \approx .864$ and $\tilde{\varepsilon} = 1$. With what is known about how $\tilde{\varepsilon}$ is computed, it is likely that the computed estimate was greater than 1, yielding this exact value of 1. These values are higher than the others we have seen and they indicate only a slight deviation from homogeneity of the variances of difference scores. Despite the $\tilde{\varepsilon}$ value of 1, the results do indeed indicate that sphericity has been violated. Even though $\tilde{\varepsilon}$ indicates the sphericity assumption is upheld, the fact that $\hat{\varepsilon}$ is only .864 creates a chance that the actual value of ε is somewhere between the two estimate values and any deviation from 1 would cause homogeneity of variances of the difference scores (and thus sphericity) to be violated.

Since $\tilde{\varepsilon} = 1$ adjusting by it would not actually be an adjustment at all and thus the p-values would not change. Additionally, since $\tilde{\varepsilon}$ is large, adjusting by $\tilde{\varepsilon}$ will not alter the degrees of freedom drastically and thus the p-values for the F-test done with the adjustment will not differ a lot from the initial p-values. As expected, the p-values change ever so slightly and all of the above statements about significance levels are still valid.

Breaking down the mean number of pivots by strain shows that mice from the C58 strain pivot the most followed by the FVB strain and then

the C57 strain. The means for C58, C57, and FVB are 2.9, 1.6, and 1.2 respectively. Although the results show a significant PND effect, there is not a strong pattern within the PND's. The means for PND's 2 and 12 are both approximately 1.3 while the means for the middle PND's bounce around in the 1–3 range. For a plot of these relationships see Figure 2(f) in section A.6 of the Appendix.

3.6 Forward

The FORWARD variable measures of the amount of time a mouse spends traveling forward when it is allowed to move around. It is measured in seconds. With a range of 0 to 151.6 and 1,417 observed values of 0 (out of 2,828), this variable is quite skewed; having a mean of 6.04 and a median of 0. There are 3 missing values.

Interpreting the results of the F-tests, we find that the usual suspects, strain, PND, and the strain by PND interaction, once again have significant effects on the variable. Strain and PND have p-values < 0.0001 while their interaction effect has a p-value of 0.0017 in the original data and a p-value of 0.0044 in the averaged version of the data. All of these effects are found to be significant at the 0.00714 significance level.

Moving to the sphericity section of the results we find that $\hat{\varepsilon} \approx 0.644$ and $\tilde{\varepsilon} \approx 0.765$. As is expected based on these values, sphericity has been violated; we must correct for this issue by adjusting the degrees of freedom by the estimate of ε. Under the adjusted F-tests the strain by PND interaction effect is significant only in the original data and only under the Huynh-Feldt Epsilon adjustment (with a p-value of 0.0046). Because of this it appears that, in general, the strain by PND interaction does not have a significant effect on the forward variable. The strain effect and the PND effect are still significant under the adjusted F-tests.

The means for the forward variable broken up by the strains C57, C58, and FVB are 2.5, 8.2, and 8 respectively. These values show that mice from the C57 strain move forward much less than the other two strains. Looking at the means by PND it is evident that, for the most part, as the mice get older they spend more time moving forward. The means actually increase from PND 2 to PND 4 and then drop back down to levels lower than those of PND 2. From PND 6 they begin to increase again with large jumps occurring

between PND's 8 and 10 and between PND's 10 and 12. Refer to Figures 2(g) and 2(h) in section A.6 of the Appendix for plots of the different relationships for this variable.

3.7 Weight

At every PND each mouse's weight is measured in grams. With 35 missing values this variable by far has the most missing values. The range of the weights of the mice is 0.54 grams to 11.91 grams. The mean weight of all the mice is 3.88 grams and the median is 3.62, indicating the the weights are pretty evenly spread out.

Looking at the results of the F-tests we see that the PND, the strain by PND interaction, and the the cage type by PND interaction all have significant effects on the weight of the mice in both versions of the data. Additionally, in the original version of the data the subject effect and the subject by PND interaction effect are significant. Aside from the cage type by PND interaction effect, which is only significant at the 0.00714 level, the other two effects are significant at the more stringent 0.00143 significance level.

Turning to the estimates of epsilon we see that $\hat{\varepsilon} = 0.287$ and $\tilde{\varepsilon} = 0.328$. These are very low values and they suggest severe heterogeneity of the variances of the difference scores. However, the results state that the sphericity requirement is met. Note, the lower bound on epsilon, $1/(I-1) = 1/(6-1)$, is 0.2 so $\hat{\varepsilon}$ is quite close to being as small as possible. Since the estimates for epsilon are so small, adjusting the degrees of freedom by $\hat{\varepsilon}$ will decrease the degrees of freedom quite a bit, thus resulting in a larger increase in the p-values. With this increase, the cage type by PND interaction effect is no longer significant and the other effects are now only significant at the 0.00714 level.

The mean weights broken down by strain indicate mice from the FVB strain weigh more than the two other strains, and mice from the C57 strain weigh slightly more than mice from the C58 strain. A break down of the means by PND reveals that, as expected, the mice weigh more the older they are: the mean weight under PND 2 is 1.6 grams while the mean weight under PND 12 is 6.5 grams. Additionally, mice that are caged in enriched cages consistently weigh the most followed by those in old cages. These relationships

can be seen in Figures 2(i) and 2(j) in section A.6 of the Appendix.

3.8 Conclusions

It is evident that in most cases the strain that a mouse comes from, the PND under which the observation was taken, and the interaction of these two treatments all significantly effect the outcome of the response variable. These results are reasonable since it is likely that mice from the same strain share common behaviors and characteristics. Additionally, as the mice get older their behaviors and abilities are likely to change.

Furthermore, although sphericity was violated in all but one of the models, these results are valid. Most of the p-values for effects that were significant were quite low – showing that the results for those effects are fairly stable. Also, most of the p-values for the effects that were not significant were not very close to the cut off value of 0.00714. This indicates that those results also would not be greatly effected by slight changes in the data or methods. The outcomes that are most susceptible to changes in the data and the methods used are those terms which had p-values very close to 0.00714 (above and below 0.00714). Although these results are valid, if the most accurate results possible are warranted, especially for the terms with p-values close to 0.00714 more advanced methods, such as mixed models, that are more forgiving of a violation of sphericity should be used.

A Appendix

A.1 Simplification of SS_T to Show Additivity of the Sums of Squares

The corrected total sums of squares can be decomposed as

$$\begin{aligned}
SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n}((y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{y}_{..}))^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n}\left((y_{ij}-\bar{y}_{i.})^2+2\,(y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{..})+(\bar{y}_{i.}-\bar{y}_{..})^2\right) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2+2\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{..})+\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2+2\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{..})+n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2
\end{aligned} \tag{A.1}$$

Examination of the middle term in equation A.1 suggests that since there are no j's in $(\bar{y}_{i.}-\bar{y}_{..})$ it can be factored out of the summation over j and we have

$$2\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\,(\bar{y}_{i.}-\bar{y}_{..}) = 2\sum_{i=1}^{a}\left((\bar{y}_{i.}-\bar{y}_{..})\left(\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\right)\right) \tag{A.2}$$

Looking at the inner summation term in equation A.2 we see that:

$$\begin{aligned}
\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.}) &= \sum_{j=1}^{n}y_{ij}-\sum_{j=1}^{n}\bar{y}_{i.} \\
&= y_{i.}-n\bar{y}_{i.} \\
&= y_{i.}-n\frac{y_{i.}}{n} \\
&= y_{i.}-y_{i.} \\
&= 0
\end{aligned}$$

Substituting this result into equation A.2 we have

$$2\sum_{i=1}^{a}\left((\bar{y}_{i.}-\bar{y}_{..})\left(\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})\right)\right)=2\sum_{i=1}^{a}\left((\bar{y}_{i.}-\bar{y}_{..})\,(0)\right)$$
$$=0 \qquad \text{(A.3)}$$

Finally, from the result in equation A.3 we have that the middle term in equation A.1 is zero and therefore equation A.1 reduces to

$$SS_T = n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

A.2 Simplification of $E(MS_E)$

Proof that the expected mean squared error is σ^2 can be seen as follows.

$$
\begin{aligned}
E(MS_E) &= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})^2 - \frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})\right)^2\right) \\
&= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\left(\mu^2+2\tau_i\mu+2\mu\epsilon_{ij}+\tau_i^2+2\tau_i\epsilon_{ij}+\epsilon_{ij}^2\right)\right. \\
&\quad \left.-\frac{1}{n}\sum_{i=1}^{a}\left(n^2\mu^2+n^2\tau_i^2+(\epsilon_{i.})^2+2n^2\mu\tau_i+2n\mu\epsilon_{i.}+2n\tau_i\epsilon_{i.}\right)\right) \\
&= \frac{1}{N-a}E\left(N\mu^2+2n\mu\tau_{.}+2\mu\epsilon_{..}+n\sum_{i}^{a}\tau_i^2+2\tau_{.}\epsilon_{..}+\sum_{i=1}^{a}\sum_{j=1}^{N}\epsilon_{ij}^2\right. \\
&\quad \left.-N\mu^2-n\sum_{i}^{a}\tau_i^2-\frac{1}{n}\sum_{i=1}^{a}(\epsilon_{i.})^2-2n\mu\tau_{.}-2\mu\epsilon_{..}-2\tau_{.}\epsilon_{..}\right) \\
&= \frac{1}{N-a}E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij}^2-\frac{1}{n}\sum_{i=1}^{a}(\epsilon_{i.})^2\right) \\
&= \frac{1}{N-a}\left(E\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\epsilon_{ij}^2\right)-E\left(\frac{1}{n}\sum_{i=1}^{a}(\epsilon_{i.})^2\right)\right) \\
&= \frac{1}{N-a}\left(\sum_{i=1}^{a}\sum_{j=1}^{n}E\left(\epsilon_{ij}^2\right)-\frac{1}{n}\sum_{i=1}^{a}E\left((\epsilon_{i.})^2\right)\right) \qquad \text{(A.4)}
\end{aligned}
$$

Equation A.4 is true because expected values are distributive over sums — the expected value of a sum is equal to the sum of the expected values.

Examining the first expected value in Equation A.4 we see that, since $Var(\epsilon_{ij}) = \sigma^2 = E(\epsilon_{ij}^2) - (E(\epsilon_{ij}))^2$, $E(\epsilon_{ij}^2) = \sigma^2 + (E(\epsilon_{ij}))^2$. But we know that $E(\epsilon_{ij}) = 0$. Thus

$$E(\epsilon_{ij}^2) = \sigma^2 \qquad \text{(A.5)}$$

Now looking at the second expected value in Equation A.4 we have that $E((\epsilon_{i.})^2) = E((\sum_{j=1}^{n}\epsilon_{ij})^2) = nE(\epsilon_{ij}^2) + n(n-1)E(\epsilon_{ij}\epsilon_{ij'})$. Since ϵ_{ij} and $\epsilon_{ij'}$ are independent we can say $n(n-1)E(\epsilon_{ij}\epsilon_{ij'}) = n(n-1)E(\epsilon_{ij})E(\epsilon_{ij'})$ and

using $E(\epsilon_{ij}) = 0$, we have that $n(n-1)E(\epsilon_{ij})E(\epsilon_{ij'}) = 0$. Thus

$$E((\epsilon_{i.})^2) = nE(\epsilon_{ij}^2) \tag{A.6}$$

Substituting Equations A.5 and A.6 into Equation A.4 we have

$$\begin{aligned} E(MS_E) &= \frac{1}{N-a}\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\sigma^2 - \frac{1}{n}\sum_{i=1}^{a}nE\left(\epsilon_{ij}^2\right)\right) \\ &= \frac{1}{N-a}\left(N\sigma^2 - a\sigma^2\right) \\ &= \frac{1}{N-a}(N-a)\sigma^2 \\ &= \sigma^2 \end{aligned}$$

A.3 F-distribution Table

F-Distribution Table for $\alpha = .05$. Notice that the critical values are decreasing for increasing denominator ($df2$) degrees of freedom.

$df1$:	1	2	3	4	5	6	7	8
$df2$: 1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
1,000,000	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94

Latex table generated in R 2.14.1 by xtable 1.7-0 package. R code for generating the F-table:

```
df1 = 1:8
df2 = c(1:10,1000000)
library(xtable)
xtable(outer(df2, df1, function(x,y){qf(0.95, y, x)}))
```

A.4 SAS Code

```
proc import out = data.micedata
            datafile= "C:\SeniorProject\NeonatalData_DeletedOutliersandCage3.xls"
            dbms=excel replace;
     range="Sheet1$";
     getnames=yes;
     mixed=yes;
     scantext=yes;
     usedata=yes;
     scantime=yes;
RUN;

data a;
  set data.micedata;
run;

proc sort data = a;
by cage pnd strain cage_eon;
run;

proc means data = a MEAN;
var right weight corneggeo pivots forwar numcross rear;
by cage pnd strain cage_eon;
output out = "C:\SeniorProject\SAS\averagedData" mean(RIGHT weight
        corneggeo pivots forwar numcross rear) = rightavg weightavg
        corneggeoavg pivotsavg forwaravg numcrossavg rearavg;
run;

proc univariate data = a;
var right weight corneggeo pivots forwar numcross rear;
run;

proc freq data = a;
table right weight corneggeo pivots forwar numcross rear;
run;
```

A.5 ANOVA Reports for the Averaged Mice Data

Expected Mean Squares:

Source Term	DF	Term Fixed?	Denominator Term	Expected Mean Square
A: STRAIN	2	Yes	C(AB)	S+dsC+bcdsA
B: CAGE.EON	2	Yes	C(AB)	S+dsC+acdsB
AB	4	Yes	C(AB)	S+dsC+cdsAB
C(AB): CAGE	65	No	S(ABCD)	S+dsC
D: PND	5	Yes	CD(AB)	S+sCD+abcsD
AD	10	Yes	CD(AB)	S+sCD+bcsAD
BD	10	Yes	CD(AB)	S+sCD+acsBD
ABD	20	Yes	CD(AB)	S+sCD+csABD
CD(AB)	310	No	S(ABCD)	S+sCD
S(ABCD)	0	No	S	

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power α=0.05
RIGHT						
A: STRAIN	2	1184.635	592.3173	30.57	0.000000	1.000000
B: CAGE.EON	2	1.916075	0.9580374	0.05	0.951799	0.057156
AB	4	32.3933	8.098325	0.42	0.795134	0.141133
C(AB): CAGE	65	1259.575	19.37808			
D: PND	5	4787.406	957.4811	115.93	0.000000	1.000000
AD	10	2095.269	209.5269	25.37	0.000000	1.000000
BD	10	44.35108	4.435108	0.54	0.863479	0.279542
ABD	20	78.76699	3.938349	0.48	0.973870	0.361546
CD(AB)	310	2560.344	8.259176			
S	0					

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power α=0.05
WEIGHT						
A: STRAIN	2	26.50912	13.25456	4.27	0.018144	0.725956
B: CAGE.EON	2	20.87215	10.43607	3.36	0.040874	0.614837
AB	4	33.79847	8.449617	2.72	0.036998	0.721832
C(AB): CAGE	65	201.8995	3.106147			
D: PND	5	948.9018	189.7803	1225.75	0.000000	1.000000
AD	10	8.84283	0.8842829	5.71	0.000000	0.999963
BD	10	4.34323	0.434323	2.81	0.002423	0.971182
ABD	20	5.478618	0.2739309	1.77	0.023141	0.969059
CD(AB)	310	47.99664	0.1548279			
S	0					
CORNEGGEO						
A: STRAIN	2	788.8692	394.4346	13.42	0.000013	0.997046
B: CAGE.EON	2	110.2957	55.14784	1.88	0.161455	0.376715
AB	4	21.81124	5.452809	0.19	0.945167	0.087025
C(AB): CAGE	65	1911.123	29.4019			
D: PND	5	16441	3288.2	171.74	0.000000	1.000000
AD	10	358.1508	35.81509	1.87	0.048649	0.851806
BD	10	268.0713	26.80713	1.40	0.179033	0.707733
ABD	20	394.5971	19.72986	1.03	0.425501	0.764450
CD(AB)	310	5935.498	19.14677			
S	0					
PIVOTS						
A: STRAIN	2	161.8623	80.93117	24.19	0.000000	0.999996
B: CAGE.EON	2	1.041986	0.5209929	0.16	0.856099	0.073069
AB	4	3.629704	0.9074259	0.27	0.895463	0.106111
C(AB): CAGE	65	217.4413	3.345251			
D: PND	5	49.8658	9.973159	5.59	0.000060	0.991616
AD	10	40.93186	4.093186	2.29	0.013162	0.926284
BD	10	7.452648	0.7452648	0.42	0.937834	0.217749
ABD	20	43.85695	2.192847	1.23	0.228524	0.853643
CD(AB)	310	553.3202	1.784904			
S	0					

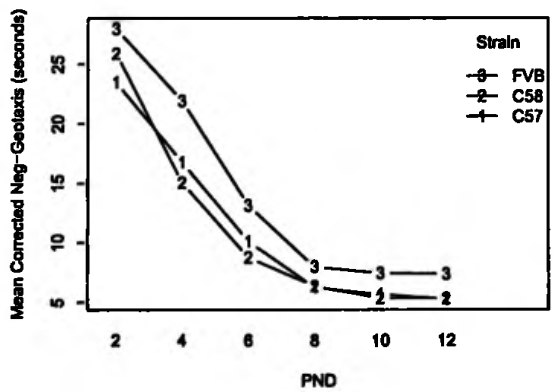
Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power α=0.05
FORWARD						
A: STRAIN	2	2224.838	1112.419	18.88	0.000000	0.999884
B: CAGE.EON	2	80.72826	40.36413	0.68	0.507707	0.160626
AB	4	106.7194	26.67985	0.45	0.770022	0.149841
C(AB): CAGE	65	3830.341	58.92833			
D: PND	5	5305.279	1061.056	20.77	0.000000	1.000000
AD	10	1341.114	134.1114	2.63	0.004439	0.959484
BD	10	322.1174	32.21174	0.63	0.787487	0.329846
ABD	20	1233.177	61.65884	1.21	0.246222	0.845449
CD(AB)	310	15836.29	51.08482			
S	0					
NUMCROSS						
A: STRAIN	2	16961.78	8480.888	27.09	0.000000	0.999999
B: CAGE.EON	2	664.0782	332.0391	1.06	0.352152	0.228019
AB	4	1119.887	279.9718	0.89	0.472544	0.268961
C(AB): CAGE	65	20348.76	313.0578			
D: PND	5	25551.62	5110.323	30.54	0.000000	1.000000
AD	10	10126.53	1012.653	6.05	0.000000	0.999985
BD	10	2162.967	216.2967	1.29	0.233499	0.664338
ABD	20	9541.68	477.084	2.85	0.000061	0.999406
CD(AB)	310	51869.55	167.3211			
S	0					
REAR						
A: STRAIN	2	20.08432	10.04216	6.21	0.003401	0.879158
B: CAGE.EON	2	2.514194	1.257097	0.78	0.463782	0.176960
AB	4	16.44737	4.111843	2.54	0.047864	0.688463
C(AB): CAGE	65	105.0956	1.616856			
D: PND	5	94.20391	18.84078	13.42	0.000000	1.000000
AD	10	37.74133	3.774132	2.69	0.003597	0.963988
BD	10	14.61899	1.461899	1.04	0.408415	0.548215
ABD	20	36.07182	1.803591	1.28	0.186882	0.873344
CD(AB)	310	435.2346	1.403983			
S	0					

A.6 Plots for the Mice Data

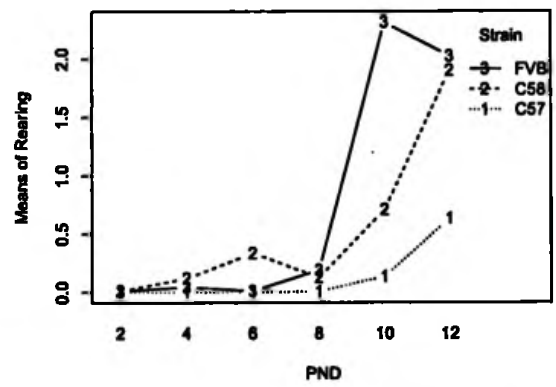
The following plots were produced in R using the interaction.plot function. Each plot was produced in the same manner changing only the variables being used. The code for Figure 2(a) is shown below.

```
interaction.plot(data$PND, data$STRAIN, data$corneggeoavg,
 xlab = "PND", ylab = "Means of Corrected Neg-Geotaxis",
 type = "b", lty=1, trace.label="Strain")
```

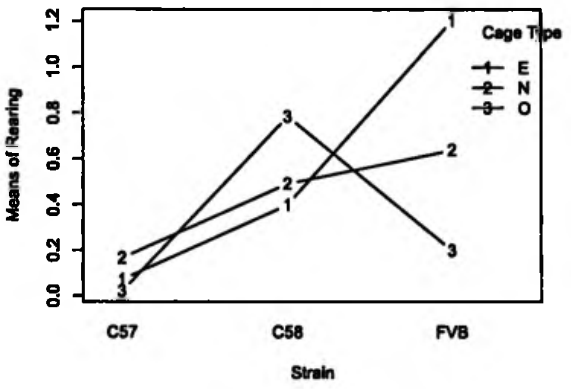
By changing relevant varible names the reader can reproduce any of the plots below.



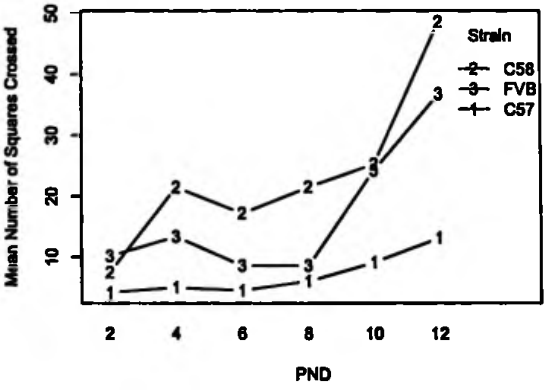
(a) Mean Corrected Negative Geotaxis by PND grouped according to Strain



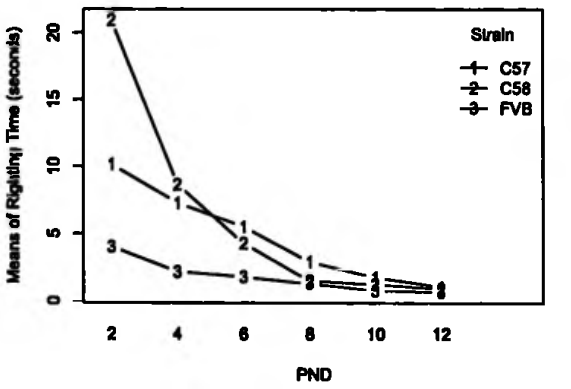
(b) Means of Rearing by PND grouped according to Strain



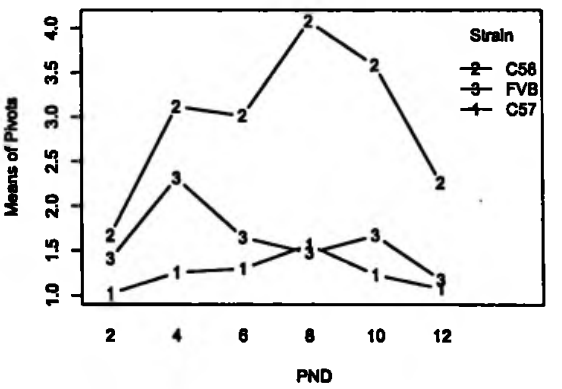
(c) Means of Rearing by Strain grouped according to Cage Type



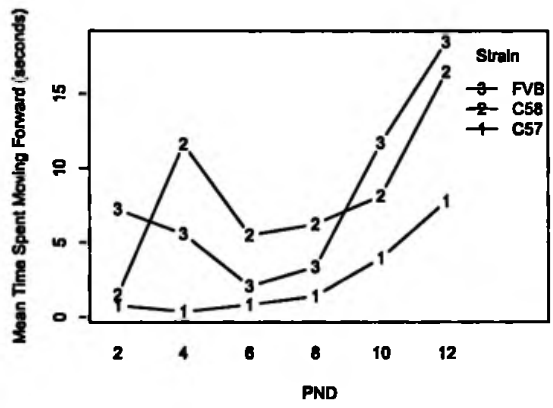
(d) Means of NUMCROSS by PND grouped according to Strain



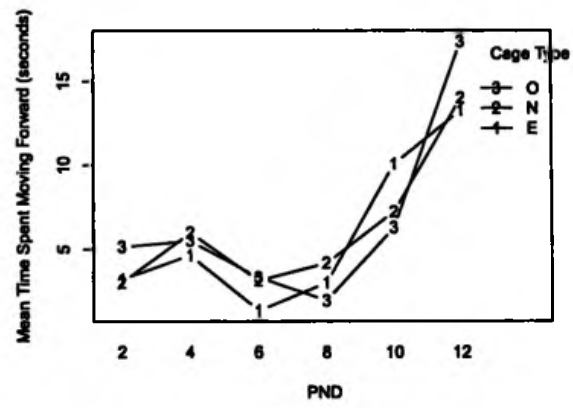
(e) Means of Righting by PND grouped according to Strain



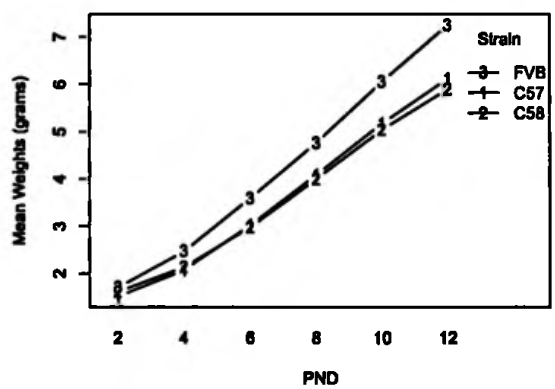
(f) Means of Pivots by PND grouped according to Strain



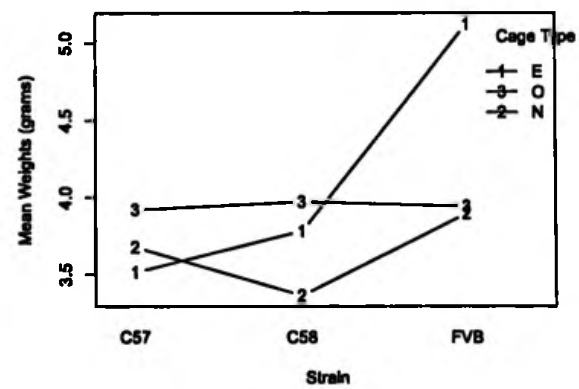
(g) Means of Forward by PND grouped according to Strain



(h) Means of Forward by PND grouped according to Cage Type



(i) Means of Weight by PND grouped according to Strain



(j) Means of Weight by Strain grouped according to Cage Type

Figure 2: Plots of Interactions

References

[1] Herve Abdi. The Bonferroni and Sidak corrections for multiple comparisons.

[2] Richard S. Bogartz. *An Introduction to the Analysis of Variance.* Praeger Publishers, Westport, Connecticut, 1994.

[3] G.E.P Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *The Annals of Mathematical Statistics*, 25:484–498, 1954.

[4] Seymour Geisser and Samuel W. Greenhouse. An extension of Box's result on the use of the F distrubution in multivariate analysis. *The Annals of Mathematical Statistics*, 29(3):885–891, September 1958.

[5] Seymour Geisser and Samuel W. Greenhouse. On methods in the analysis of profile data. *Psychometrika*, 24(2), June 1959.

[6] Ellen R. Girden. *ANOVA: Repeated Measures.* SAGE Publications, Inc, Newbury Park, California, 1992.

[7] UCLA: Statistical Consulting Group. Repeated measures ANOVA using SAS PROC GLM. 1997.

[8] J. Hintze. *NCSS.* NCSS, LLC. Kaysville, Utah, USA., 2007.

[9] Huynh Huynh and Leonard S. Feldt. Conditions under which mean square ratios in repeated measurements designs have exact F-distribution. *Journal of the American Statistical Association*, 65(332), December 1970.

[10] Huynh Huynh and Leonard S. Feldt. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1(1):69–82, 1976.

[11] Huynh Huynh and Garrett K. Mandeville. Validity conditions in repeated measures designs. *Psychological Bulletin*, 86(5):964–973, 1979.

[12] Douglas C. Montgomery. *Design and Analysis of Experiments.* John Wiley and Sons, Inc., New York, 1984.

[13] John Neter. *Applied Linear Regression Models.* Irwin, Chicago, 1996.

[14] Eric W. Weisstein. ANOVA. MathWorld–A Wolfram Web Resource.

[15] Eric W. Weisstein. Bonferroni Correction. MathWorld–A Wolfram Web Resource.