Julie Legler and Paul Roback

Broadening Your Statistical Horizons

Generalized Linear Models and Multilevel Models

Contents

List of Tables

List of Figures

Preface

Placeholder

0

Review of Multiple Linear Regression

Placeholder

0.1 Learning Objectives

0.2 Introduction to Broadening Your Statistical Horizons

0.3 Ordinary Least Squares (OLS) Assumptions

0.3.1 Cases that do not violate the OLS assumptions for inference

0.3.2 Cases where the OLS assumptions for inference are violated

0.4 Review of Multiple Linear Regression

0.4.1 Case Study: Kentucky Derby

0.5 Initial Exploratory Analyses

0.5.1 Data Organization

0.5.2 Univariate Summaries

0.5.3 Bivariate Summaries

0.6 Multiple linear regression modeling

0.6.1 Simple linear regression with a continuous predictor

0.6.2 linear regression with a binary predictor

0.6.3 Multiple linear regression with two predictors

0.6.4 Inference in multiple linear regression: normal theory

0.6.5 Inference in multiple linear regression: bootstrapping

0.6.6 Multiple linear regression with an interaction term

0.6.7 Building a multiple linear regression model

0

Beyond Least Squares: Using Likelihoods to Fit and Compare Models

Placeholder

0.9 Learning Objectives

0.10 Case Study: Does sex run in families?

0.10.1 Research Questions

0.11 Model 0: Sex Unconditional Model (Equal probabilities, Independence)

0.12 Model 1: Sex Unconditional Model (Any Probability, Independence)

0.12.1 What is a likelihood?

0.12.2 Finding MLEs

0.12.2.1 Graphically approximating an MLE

0.12.2.2 Numerically approximating an MLE

0.12.2.3 MLEs using calculus (Optional)

0.12.2.4 How does sample size affect the likelihood?

0.12.3 Summary

0.12.4 Is a likelihood a probability function? (Optional)

0.13 Model 2: Sex Conditional Model (Sex Bias)

0.13.1 Model Specification

0.13.2 Application to Hypothetical Data

0.14 Case Study: Analysis of the NLSY data

0.14.1 Model Building Plan

0.14.2 Family Composition of Boys and Girls, NLSY: Exploratory Data Analysis

0.14.3 Likelihood for the Sex Unconditional Model: the NLSY data

0

Poisson Regression

Placeholder

0.27 Learning Objectives

0.28 Introduction to Poisson Regression

0.28.1 Poisson Regression Assumptions

0.28.2 A Graphical Look at Poisson Regression

0.29 Case Studies Overview

0.30 Case Study: Household Size in the Philippines

0.30.1 Data Organization

0.30.2 Exploratory Data Analyses

0.30.3 Estimation and Inference

0.30.4 Using Deviances to Compare Models

0.30.5 Using Likelihoods to fit Poisson Regression Models (Optional)

0.30.6 Second Order Model

0.30.7 Adding a covariate

0.30.8 Residuals for Poisson Models (Optional)

0.30.9 Goodness-of-fit

0.31 Least Squares Regression vs. Poisson Regression

0.32 Case Study: Campus Crime

0.32.1 Data Organization

0.32.2 Exploratory Data Analysis

0.32.3 Accounting for Enrollment

0

Generalized Linear Models (GLMs): A Unifying Theory

Placeholder

0.38 Learning Objectives

0.39 One parameter exponential families

0.39.1 One Parameter Exponential Family: Possion

0.39.2 One parameter exponential family: Normal

0.40 Generalized Linear Modeling

0.41 Exercises

0

Logistic Regression

Placeholder

0.42 Learning Objectives

0.43 Introduction to Logistic Regression

0.43.1 Logistic Regression Assumptions

0.43.2 A Graphical Look at Logistic Regression

0.44 Case Studies Overview

0.45 Case Study: Soccer Goalkeepers

0.45.1 Modeling Odds

0.45.2 Logistic Regression Models for Binomial Responses

0.45.3 Theoretical rationale for logistic regression models (Optional)

0.46 Case Study: Reconstructing Alabama

0.46.1 Data Organization

0.46.2 Exploratory Analyses

0.46.3 Initial Models

0.46.4 Tests for significance of model coefficients

0.46.5 Confidence intervals for model coefficients

0.46.6 Testing for goodness of fit

0.46.7 Residuals for Binomial Regression

0.46.8 Overdispersion

0.46.9 Summary

0.47 Least Squares Regression vs. Logistic Regression

0

Correlated Data

Placeholder

0.50 Learning Objectives

0.51 Introduction

0.52 Recognizing correlation

0.53 Case Study: Dams and pups

0.54 Sources of Variability

0.55 Scenario 1: No covariates

0.56 Scenario 2: Dose effect

0.57 Case Study: Tree Growth

0.57.1 Format of the data set

0.57.2 Sources of variability

0.57.3 Analysis preview: accounting for correlation within transect

0.58 Summary

0.59 Exercises

0.59.1 Conceptual Exercises

0.59.2 Guided Exercises

0.59.3 Note on Correlated Binary Outcomes

0

Introduction to Multilevel Models

0.60 Learning Objectives

After finishing this chapter, you should be able to:

- Recognize when response variables and covariates have been collected at multiple (nested) levels.
- Apply exploratory data analysis techniques to multilevel data.
- Write out a multilevel statistical model, including assumptions about variance components, in both by-level and composite forms.
- Interpret model parameters (including fixed effects and variance components) from a multilevel model, including cases in which covariates are continuous, categorical, or centered.
- Understand the taxonomy of models, including why we start with an unconditional means model.
- Select a final model, using criteria such as AIC, BIC, and deviance.

0.61 Case Study: Music Performance Anxiety

Stage fright can be a serious problem for performers, and understanding the personality underpinnings of performance anxiety is an important step in determining how to minimize its impact. **?** studied the emotional state of musicians before performances and factors which may affect their emotional state. Data was collected by having 37 undergraduate music majors from a competitive undergraduate music program fill out diaries prior to performances over the course of an academic year. In particular, study participants completed a Positive Affect Negative Affect Schedule (PANAS) before each performance. The PANAS instrument provided two key outcome measures: negative affect (a state measure of anxiety) and positive affect (a state measure of happiness). We will focus on negative affect as our primary response measuring performance anxiety.

Factors which were examined for their potential relationships with performance anxiety included: performance type (solo, large ensemble, or small ensemble); audience (instructor, public, students, or juried); if the piece was played from memory; age; gender; instrument (voice, orchestral, or keyboard); and, years studying the instrument. In addition, the personalities of study participants were assessed at baseline through the Multidimensional Personality Questionnaire (MPQ). The MPQ provided scores for one lower-order factor (absorption) and three higher-order factors: positive emotionality (PEM—a composite of well-being, social potency, achievement, and social closeness); negative emotionality (NEM—a composite of stress reaction, alienation, and aggression); and, constraint (a composite of control, harm avoidance, and traditionalism).

Primary scientific hypotheses of the researchers included:

- Lower music performance anxiety will be associated with lower levels of a subject's negative emotionality.
- Lower music performance anxiety will be associated with lower levels of a subject's stress reaction.
- Lower music performance anxiety will be associated with greater number of years of study.

0.62 Initial Exploratory Analyses

0.62.1 Data Organization

Our examination of the data from **?** in `musicdata.csv` will focus on the following key variables:

- `id` = unique musician identification number
- `diary` = cumulative total of diaries filled out by musician
- `perf_type` = type of performance (Solo, Large Ensemble, or Small Ensemble)
- `audience` = who attended (Instructor, Public, Students, or Juried)
- `memory` = performed from Memory, using Score, or Unspecified
- `na` = negative affect score from PANAS
- `gender` = musician gender
- `instrument` = Voice, Orchestral, or Piano
- `mpqab` = absorption subscale from MPQ
- `mpqpem` = positive emotionality (PEM) composite scale from MPQ
- `mpqnem` = negative emotionality (NEM) composite scale from MPQ

TABLE 0.1: A snapshot of selected variables from the first three and the last three observations in the Music Performance Anxiety case study.

Obs	id	diary	perf_type	memory	na	gender	instrument	mpqab	mpqpem	mpqnem
1	1	1	Solo	Unspecified	11	Female	voice	16	52	16
2	1	2	Large Ensemble	Memory	19	Female	voice	16	52	16
3	1	3	Large Ensemble	Memory	14	Female	voice	16	52	16
495	43	2	Solo	Score	13	Female	voice	31	64	17
496	43	3	Small Ensemble	Memory	19	Female	voice	31	64	17
497	43	4	Solo	Score	11	Female	voice	31	64	17

Sample rows containing selected variables from our data set are illustrated in Table **??**; note that each subject (id) has one row for each unique diary entry.

As with any statistical analysis, our first task is to explore the data, examining distributions of individual responses and predictors using graphical and numerical summaries, and beginning to discover relationships between variables. With multilevel models, exploratory analyses must eventually account for the level at which each variable is measured. In a two-level study such as this one, **Level One** will refer to variables measured at the most frequently occurring observational unit, while **Level Two** will refer to variables measured on larger observational units. For example, in our study on music performance anxiety, many variables are measured at every performance. These "Level One" variables include:

- negative affect (our response variable)
- performance characteristics (type, audience, if music was performed from memory)
- number of previous performances with a diary entry

However, other variables measure characteristics of study participants that remain constant over all performances for a particular musician; these are considered "Level Two" variables and include:

- demographics (age and gender of musician)
- instrument used and number of previous years spent studying that instrument
- baseline personality assessment (MPQ measures of positive emotionality, negative emotionality, constraint, stress reaction, and absorption)

0.62.2 Exploratory Analyses: Univariate Summaries

Because of this data structure—the assessment of some variables on a performance-by-performance basis and others on a subject-by-subject basis—we cannot treat our data set as consisting of 497 independent observations. Although negative affect measures from different subjects can reasonably be assumed to be independent (unless, perhaps, the subjects frequently perform in the same ensemble group), negative affect measures from different performances by the same subject are not likely to be independent. For example, some subjects tend to have relatively high performance anxiety across all performances, so that knowing their score for Performance 3 was 20 makes it more likely that their score for Performance 5 is somewhere near 20 as well. Thus, we must carefully consider our exploratory data analysis, recognizing that certain plots and summary statistics may be useful but imperfect in light of the correlated observations.

First, we will examine each response variable and potential covariate individually. Continuous variables can be summarized using histograms and summaries of center and spread; categorical variables can be summarized with tables and possibly bar charts. When examining Level One covariates and responses, we will begin by considering all 497 observations, essentially treating each performance by each subject as independent even though we expect observations from the same musician to be correlated. Although these plots will contain dependent points, since each musician provides data for up to 15 performances, general patterns exhibited in these plots tend to be real. Alternatively, we can calculate mean scores across all performances for each of the 37 musicians so that we can more easily consider each plotted point to be independent. The disadvantage of this approach would be lost information which, in a study such as this with a relatively small number of musicians each being observed over many performances, could be considerable. In addition, if the sample sizes varied greatly by subject, a mean based on 1 observation would be given equal weight to a mean based on 15 observations. Nevertheless, both types of exploratory plots typically illustrate similar relationships.

In Figure **??** we see histograms for the primary response (negative affect); plot (a) shows all 497 (dependent) observations, while plot (b) shows the mean negative affect for each of the 37 musicians across all their performances. Through plot (a), we see that performance anxiety (negative affect) across all performances follows a right skewed distribution with a lower bound of 10 (achieved when all 10 questions are answered with a 1). Plot (b) shows that mean negative affect is also right-skewed (although not as smoothly decreasing in frequency), with range 12 to 23.

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```

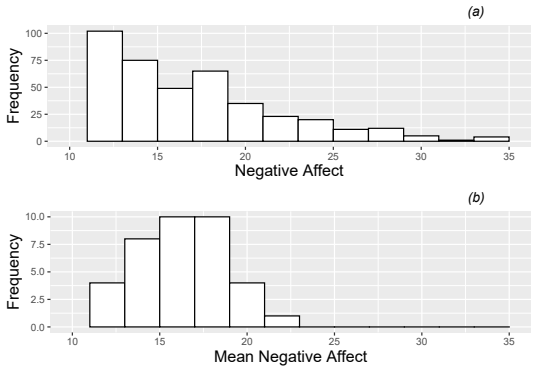

FIGURE 1: Histogram of the continuous Level One response (negative effect). Plot (a) contains all 497 performances across the 37 musicians, while plot (b) contains one observation per musician (the mean negative affect across all performances).

We can also summarize categorical Level One covariates across all (possibly correlated) observations to get a rough relative comparison of trends. 56.1% of the 497 performances in our data set were solos, while 27.3% were large ensembles and 16.5% were small emsembles. The most common audience type was a public performance (41.0%), followed by instructors (30.0%), students (20.1%), and finally juried recitals (8.9%). In 30.0% of performances, the musician played by memory, while 55.1% used the score and 14.9% of performances were unspecified.

To generate an initial examination of Level Two covariates, we consider a data set with just one observation per subject, since Level Two variables are constant over all performances from the same subject. Then, we can proceed as we did with Level One covariates—using histograms to illustrate the distributions of continuous covariates (see Figure **??**) and tables to summarize categorical covariates. For example, we learn that the majority of subjects have positive emotionality scores between 50 and 60, but that several subjects fall into a lengthy lower tail with scores between 20 and 50. A summary of categorical Level Two covariates reveals that among the 37 subjects (26 female and 11 male), 17 play an orchestral instrument, 15 are vocal performers, and 5 play a keyboard instrument.

0.62.3 Exploratory Analyses: Bivariate Summaries

The next step in an initial exploratory analysis is the examination of numerical and graphical summaries of relationships between model covariates and responses. In examining these bivariate relationships, we hope to learn: (1) if

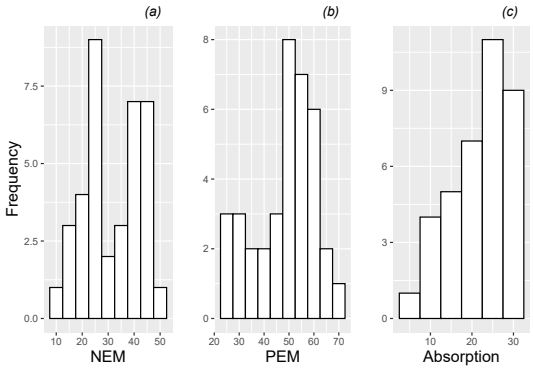

FIGURE 2: Histograms of the 3 continuous Level Two covariates (negative emotionality (NEM), positive emotionality (PEM), and absorption). Each plot contains one observation per musician.

there is a general trend suggesting that as the covariate increases the response either increases or decreases, (2) if subjects at certain levels of the covariate tend to have similar mean responses (low variability), and (3) if the variation in the response differs at different levels of the covariate (unequal variability).

As with individual variables, we will begin by treating all 497 performances recorded as independent observations, even though blocks of 15 or so performances were performed by the same musician. For categorical Level One covariates, we can generate boxplots against negative affect as in Figure **??**, plots (a) and (b). From these boxplots, we see that lower levels of performance anxiety seem to be associated with playing in large ensembles and playing in front of an instructor. For our lone continuous Level One covariate (number of previous performances), we can generate a scatterplot against negative affect as in plot (c) from Figure **??**, adding a fitted line to illustrate general trends upward or downward. From this scatterplot, we see that negative affect seems to decrease slightly as a subject has more experience.

To avoid the issue of dependent observations in our three plots from Figure **??**, we could generate separate plots for each subject and examine trends within and across subjects. These "lattice plots" are illustrated in Figures **??**, **??**, and **??**; we discuss such plots more thoroughly in Chapter **??**. While general trends are difficult to discern from these lattice plots, we can see the variety in subjects in sample size distributions and overall level of performance anxiety. In particular, in Figure **??**, we notice that linear fits for many subjects illustrate the same same slight downward trend displayed in the overall scatterplot in Figure **??**, although some subjects experience increasing anxiety and others exhibit non-linear trends. Having an idea of the range of individual trends will be important when we begin to draw overall conclusions from this study.

In Figure **??**, we use boxplots to examine the relationship between our pri-

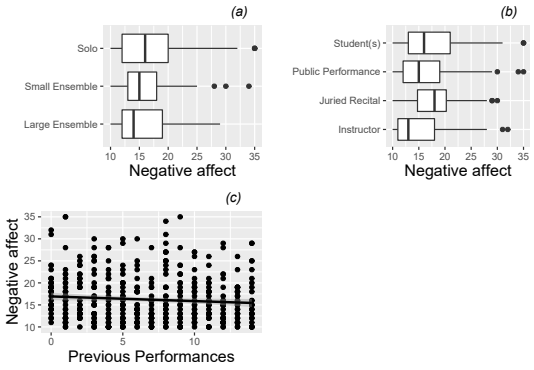

FIGURE 3: Boxplots of two categorical Level One covariates (performance type (a) and audience type (b)) vs. model response, and scatterplot of one continuous Level One covariate (number of previous diary entries (c)) vs. model response (negative affect). Each plot contains one observation for each of the 497 performances.

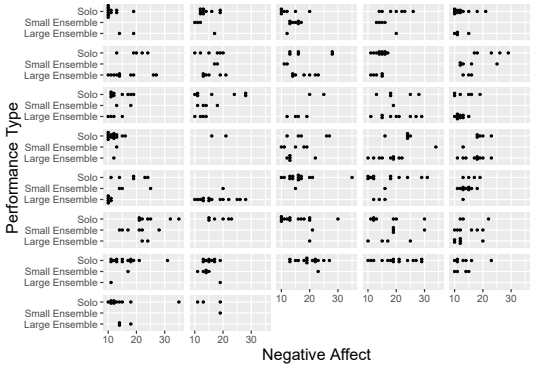

FIGURE 4: Lattice plot of performance type vs. negative affect, with separate dotplots by subject.

mary categorical Level Two covariate (instrument) and our continuous model response. Plot (a) uses all 497 performances, while plot (b) uses one observation per subject (the mean performance anxiety across all performances) regardless of how many performances that subject had. Naturally, plot (b) has a more condensed range of values, but both plots seem to support the notion that performance anxiety is slightly lower for vocalists and maybe a bit higher for keyboardists

In Figure **??**, we use scatterplots to examine the relationships between continuous Level Two covariates and our model response. Performance anxiety appears to vary little with a subject's positive emotionality, but there is some

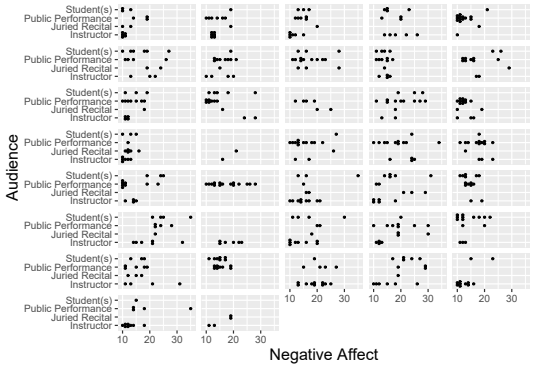

FIGURE 5: Lattice plot of audience type vs. negative affect, with separate dotplots by subject.

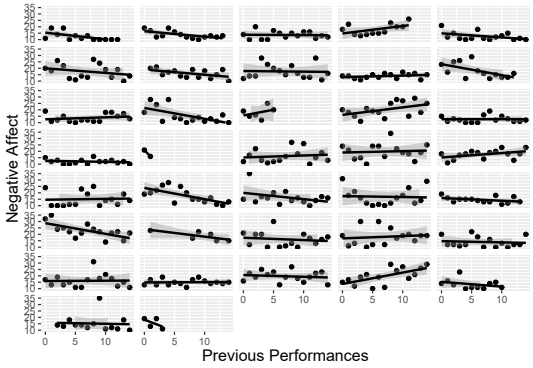

FIGURE 6: Lattice plot of previous performances vs. negative affect, with separate scatterplots with fitted lines by subject.

evidence to suggest that performance anxiety increases with increasing negative emotionality and absorption level. Plots based on mean negative affect, with one observation per subject, support conclusions based on plots with all observations from all subjects; indeed the overall relationships are in the same direction and of the same magnitude.

Of course, any graphical analysis is exploratory, and any notable trends at this stage should be checked through formal modeling. At this point, a statistician begins to ask familiar questions such as:

- which characteristics of individual performances are most associated with performance anxiety?
- which characteristics of study participants are most associated with performance anxiety?

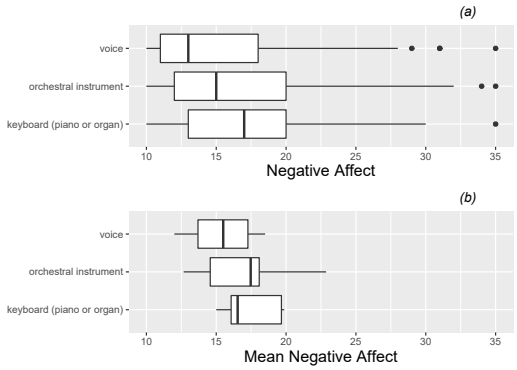

FIGURE 7: Boxplots of the categorical Level Two covariate (instrument) vs. model response (negative affect). Plot (a) is based on all 497 observations from all 37 subjects, while plot (b) uses only one observation per subject.

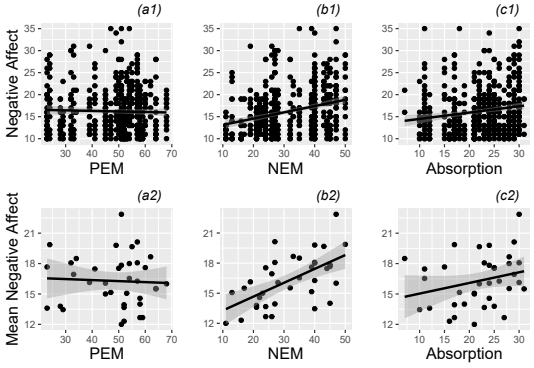

FIGURE 8: Scatterplots of continuous Level Two covariates (positive emotionality (PEM), negative emotionality (NEM), and absorption) vs. model response (negative affect). The top plots (a1, b1, c1) are based on all 497 observations from all 37 subjects, while the bottom plots (a2, b2, c2) use only one observation per subject.

- are any of these associations statistically significant?
- does the significance remain after controlling for other covariates?
- how do we account for the lack of independence in performances by the same musician?

As you might expect, answers to these questions will arise from proper consideration of variability and properly identified statistical models.

0.63 Two level modeling: preliminary considerations

0.63.1 Ignoring the two level structure (not recommended)

Armed with any statistical software package, it would be relatively simple to take our complete data set of 497 observations and run an OLS multiple linear regression model seeking to explain variability in negative affect with a number of performance-level or musician-level covariates. As an example, output from a model with two binary covariates (Does the subject play an orchestral instrument? and, Was the performance a large ensemble?) is presented below. Do you see any problems with this approach?

```
lm(formula = na ~ orch + large + orch:large, data = music)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.7212     0.3591  43.778  < 2e-16 ***
orch          1.7887     0.5516   3.243  0.00126 **
large        -0.2767     0.7910  -0.350  0.72662
orch:large   -1.7087     1.0621  -1.609  0.10831
---
Residual standard error: 5.179 on 493 degrees of freedom
Multiple R-squared:  0.02782,   Adjusted R-squared:  0.0219
F-statistic: 4.702 on 3 and 493 DF,  p-value: 0.003012
```

Other than somewhat skewed residuals, residual plots (not shown) do not indicate any major problems with the OLS multiple regression model. However, another key assumption in these models is the independence of all observations. While we might reasonably conclude that responses from different study participants are independent (although possibly not if they are members of the same ensemble group), it is not likely that the 15 or so observations taken over multiple performances from a single subject are similarly independent. If a subject begins with a relatively high level of anxiety (compared to other subjects) before her first performance, chances are good that she will have relatively high anxiety levels before subsequent performances. Thus, OLS multiple linear regression using all 497 observations is not advisable for this study (or multilevel data sets in general).

0.63.2 A two-stage modeling approach (better but imperfect)

If we assume that the 37 study participants can reasonably be considered to be independent, we could use traditional OLS regression techniques to analyze

TABLE 0.2: Data from the 15 performances of Musician 22

	id	diary	perform_type	audience	na	instrument
240	22	1	Solo	Instructor	24	orchestral instrument
241	22	2	Large Ensemble	Public Performance	21	orchestral instrument
242	22	3	Large Ensemble	Public Performance	14	orchestral instrument
243	22	4	Large Ensemble	Public Performance	15	orchestral instrument
244	22	5	Large Ensemble	Public Performance	10	orchestral instrument
245	22	6	Solo	Instructor	24	orchestral instrument
246	22	7	Solo	Student(s)	24	orchestral instrument
247	22	8	Solo	Instructor	16	orchestral instrument
248	22	9	Small Ensemble	Public Performance	34	orchestral instrument
249	22	10	Large Ensemble	Public Performance	22	orchestral instrument
250	22	11	Large Ensemble	Public Performance	19	orchestral instrument
251	22	12	Large Ensemble	Public Performance	18	orchestral instrument
252	22	13	Large Ensemble	Public Performance	12	orchestral instrument
253	22	14	Large Ensemble	Public Performance	19	orchestral instrument
254	22	15	Solo	Instructor	25	orchestral instrument

data from this study if we could condense each subject's set of responses to a single meaningful outcome. Candidates for this meaningful outcome include a subject's last performance anxiety measurement, average performance anxiety, minimum anxiety level, etc. For example, in clinical trials, data is often collected over many weekly or monthly visits for each patient, except that many patients will drop out early for many reasons (e.g., lack of efficacy, side effects, personal reasons). In these cases, treatments are frequently compared using "last-value-carried-forward" methods—the final visit of each patient is used as the primary outcome measure, regardless of how long they remained in the study. However, "last-value-carried-forward" and other summary measures feel inadequate, since we end up ignoring much of the information contained in the multiple measures for each individual. A more powerful solution is to model performance anxiety at multiple levels.

We will begin by considering all performances by a single individual. For instance, consider the 15 performances for which Musician #22 recorded a diary, illustrated in Table **??**.

Does this musician tend to have higher anxiety levels when he is playing in a large ensemble or playing in front of fellow students? Which factor is the biggest determinant of anxiety for a performance by Musician #22? We can address these questions through OLS multiple linear regression applied to only Musician #22's data, using appropriate indicator variables for factors of interest.