**PSY 653 Module 9: Categorical Predictors and Nonlinear Models**

**Gemma Wallace & Neil Yetz**

**April 8, 2020**

*Description of the datasets for the demo activity:*

1. ***“slpdata.csv”***

A team of sleep researchers sought to study the effects of a 6-week sleep intervention aimed to improve participant’s sleep hygiene. Sleep hygiene encompasses a variety of practices and habits that are necessary to have good nighttime sleep quality and full daytime alertness. The team formulated three different versions of the intervention. The first version (condition 1) provided participants with a self-help book on the topic of sleep hygiene. The second version (condition 2) brought participants together once per week in groups of 10-12 to teach the principles of sleep hygiene in a classroom setting. The final version (condition 3) also used the group-based classroom setting of condition 2, but in addition, each participant’s partner was invited to also take part in the group sessions. Six-hundred male and female adults living with an intimate partner and suffering from a sleep disorder were recruited to take part in the study, the participants were randomly assigned to one of the three conditions. The data set includes the following variables:

* **sex:** 1=male, 2=female
* **age:** Participant’s age in years
* **anxiety:** Participant’s level of general anxiety measured at the start of the study via a multi-item scale. The scale (average of all items) ranges from 1 to 7, where a higher score indicates a higher level of anxiety.
* **prior:** An indicator of whether or not the participant had previously participated in some type of sleep intervention, 1 = yes, 0 = no.
* **hygiene:** Participant’s sleep hygiene at week 6. It ranges from 0 to 10, and higher means better sleep practices.
* **support:** Participant’s perception that their partner is supportive of their struggles with sleep and their efforts to improve sleep. It is a multi-item scale that ranges from 1 to 5, where higher indicates more support.
* **sleep:** Participant’s average sleep efficiency during the month following the intervention, calculated as time spent in bed asleep (minus all the awakenings), divided by the total time spent in bed. It is expressed as a percentage.
* **lifesat:** Participant’s sense of life satisfaction measured 30 days after the completion of the intervention. It is a multi-item scale that ranges from 1 to 7, where a higher score indicates more satisfaction.
* **cond:** Treatment condition, 1 = self-help, 2 = group-based intervention, 3 = group-based plus partner participation.

1. ***“cogtest.csv”***

Researcheres were interested in the effect of time spent in practice on the performance of a visual discrimination task. Subjects were randomly assigned to different levels of practice, following which a test of visual discrimination is administered, and the number of correct responses is recorded for each subject. 40 subjects were randomly assigned to practice 0 minutes, 2 minutes, 4 minutes, 6 minutes, 8 minutes, 10 minutes, 12 minutes, or 14 minutes.

There are two variables:

* **practice:** minutes spent practicing, this was assigned by the experimenter
* **score:** the number of correct answers on the test

**Demo Activity**

1. Create a new R notebook and load the following libraries: tidyverse, psych, olsrr
2. Read in the datafile “slpdata.csv”. Use this datafile for the analyses in Parts 1-2.
3. **Part 1:** Using dummy coding, use treatment condition, sex, and their interaction to predict sleep hygiene.
   * 1. How well do these variables predict sleep hygiene?
     2. What information do the regression coefficients for treatment condition and sex give you about the differences in sleep hygiene for people of different treatment conditions and sex?
     3. What conclusions do you reach about treatment condition, sex, and their interaction?
4. **Part 2:** Redo this analysis using effect coding for treatment condition
   * 1. What information do the coefficients for job category variables give you about the differences in salary for people in different types of jobs?
5. Read in the datafile “cogtest.csv”. Use this datafile for the analyses in Part 3.
6. Plot the relationship between practice and score. What type of relationship do you think exists if any?
7. Filter the dataset to only include data from participants who were in the following levels of the practice variable: 0 minutes, 4 minutes, 8 minutes, and 12 minutes
8. **Part 3**: Use the method of orthogonal polynomial coding to test hypotheses about the relationship (its form and strength) between score and four different levels of the practice variable.
   * 1. Which type of relationship, if any, best fits the data for this research question?

**Try It Yourself Activity**

We will use global data on the global Covid-19 pandemic for this activity. This data is publicly available from Johns Hopkins University and includes data on confirmed cases, deaths, countries and regions, government restriction policies (e.g., lockdown policies, number of social distancing policies, etc.), country socioeconomic status, etc.

The following website describes a package for accessing and using the data: <https://joachim-gassen.github.io/2020/03/meet-tidycovid19-yet-another-covid-19-related-r-package/>

The variables that we are using in this activity include:

* **income**: countries are coded as one of four income category levels: Low income, Lower middle income, Upper middle income, High income
* **lockdown:** counties are coded from 0-8, with any number above 1 representing a country that has implemented a government-enforced lockdown procedure (more stringent than social distancing orders).
* **confirmed:** the number of confirmed cases confirmed in each country per day

1. Read in the datafile “covid.csv”
2. Transform the lockdown variable into a binary indicator where anything above 0 is transformed into 1. This creates a variable where any country that has implemented any form of formal lockdown is coded as 1, while countries without lockdown procedures (e.g., encouraging but not enforcing social distancing) is coded as 0.
3. Using dummy coding, use lockdown, income, and their interaction to predict confirmed cases (named “confirmed” in the dataframe).
   * 1. How well do these variables predict confirmed cases?
     2. What information do the regression coefficients for lockdown & income give you about the differences in confirmed cases for countries of different lockdown status and income?
     3. What conclusions do you reach about treatment lockdown, income, and their interaction?
4. Redo this analysis using effect coding for income category
   * 1. What information do the coefficients for income category variables give you about the differences in confirmed cases?
5. Plot the relationship between income and confirmed cases. What type of relationship do you think exists if any?
6. Use the method of orthogonal polynomial coding to test hypotheses about the relationship (its form and strength) between the four different levels of income in predicting confirmed cases.
   * 1. Which type of relationship, if any, best fits the data for this research question?