Missing Data Workshop Lab Exercises

pain.dat

The data set is from a study of $N = 275$ chronic pain patients. The data set includes psychological correlates of pain severity such as depression, pain interference with daily life, perceived control over pain, stress, and psychosocial disability. Psychosocial disability measures pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc. All missing values are coded as 999.

Variable Name	Variable Definition	Missing %	Response Scale
ID	Patient identifier	0	Integer index
MALE	Gender	0	0 = female, 1 = male
AGE	Age	0	Continuous
EDUGROUP	Education level	0	1 = Some college or less, 2 = college, 3 = Post-BA
WORKHRS	Work hours per week	11.7	Continuous
EXERCISE	Exercise	1.7	8-point ordinal scale
PAIN	Pain intensity rating	7.3	1 = none/little, 2 = moderate, 3 = severe/very severe
ANXIETY	Anxiety	6.0	Continuous
STRESS	Stress	0	7-point ordinal scale
CONTROL	Perceived control over pain	0	Continuous
INTERFERE	Pain interference with life	13.3	Continuous
DEPRESS	Depression	13.3	Continuous
DISABILITY	Psychosocial disability	3.0	Continuous

depression.dat

The data set is from a study of an online chronic pain management treatment program. A total of 275 individuals with chronic pain were randomly assigned to either a treatment (n = 133) or a wait-list control condition (n = 142). Researchers measured a number of background variables at the start of the study (e.g., age, frequency of exercise, gender, depression), and they administered a 6-item pain interference with daily life questionnaire at the conclusion of the intervention period. All missing values are coded as 999.

Variable Name	Variable Definition	Response Scale
ID	Respondent identifier	
FEMALE	Respondent gender (0 = male, 1 = female)	0-1
AGE	Respondent age	
TXGROUP	Treatment condition (0 = control, 1 = treatment group)	0-1
EXERCISE*	Frequency of exercise	1-20
	Depression Scale	
DEPRESS1*	I couldn't seem to experience any positive feeling at all.	1-4
DEPRESS2	I just couldn't seem to get going.	1-4
DEPRESS3*	I felt that I had nothing to look forward to.	1-4
DEPRESS4	I felt sad and depressed.	1-4
DEPRESS5*	I felt that I had lost interest in just about everything.	1-4
DEPRESS6*	I felt I wasn't worth much as a person.	1-4
DEPRESS7	I felt that life wasn't worthwhile.	1-4
	Pain Interference	
INTERFERE1*	Pain interferes with enjoyable activities.	1-8
INTERFERE2*	Pain interferes with responsibilities at home.	1-8
INTERFERE3	Pain interferes with relationships.	1-8
INTERFERE4*	Pain interferes with personal goals.	1-8
INTERFERE5	Pain interferes with self-care.	1-8
INTERFERE6*	Pain interferes with thinking clearly or concentrating.	1-8

* Incomplete variable.

lmxquality.dat

This data set is from an organizational research study with 630 employees grouped into 105 teams (i.e., a two-level data structure). A focal predictor for the study is a construct known as leader-member exchange (LMX), which measures the quality of an employee's relationship with his or her supervisor. It is hypothesized that employees who report higher LMX will feel more empowered, and individual empowerment is measured by a scale score capturing autonomy and capability to perform meaningful work that can impact the organization. Employee-level variables also include gender, a binary item asking whether the employee intends to quit within the next six months, and a 7-point job satisfaction rating scale. The team-level (i.e., level-2) variables include a continuous measures of group-related leadership climate, a measure of leadership behaviors directed specifically at the team. Finally, organization size is measured by six ordered categories. All missing values are coded as 999.

Variable Name	Variable Definition	Missing %	Response Scale
EMPLOYEE	Employee identifier	0	Integer index
TEAM	Team identifier	0	Integer index
TURNOVER	Intentions to quit job	5.1	0 = won't quit, 1 = might quit
MALE	Gender dummy code	0	0 = female, 1 = male
EMPOWER	Individual empowerment scale score	16.2	Continuous
LMXQUALITY	Leader-member exchange (relationship quality) scale score	4.1	Continuous
JOBSAT	Job satisfaction rating scale	4.8	1-7
CLIMATE	Team climate scale score	9.5	Continuous
ORGSIZE	Organization size	5.7	1-6

Exercise 1: Maximum Likelihood

Use the lmxquality.dat data set for this example. Write a maximum likelihood estimation script that fits a multiple regression model where employee empowerment is regressed on leader-member exchange (i.e., employee-supervisor relationship quality) and job satisfaction, as shown below. Note that job satisfaction is an ordinal scale, but it is necessary to specify normal distributions for the predictors. Additionally, use the Wald test to specify a significance test for the set of predictors (i.e., a two degree of freedom omnibus test similar to an *F* statistic in ordinary least squares).

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \beta_2(JOBSAT_i) + \varepsilon$$

1. Provide an interpretation of the Wald test of the entire mnodel.
2. Provide a substantive interpretation of the two slope coefficients and their standard errors.

Exercise 2: Maximum Likelihood with Auxiliary Variables

Use the lmxquality.dat data set for this example. Write a maximum likelihood estimation script that fits a multiple regression model where employee empowerment is regressed on leader-member exchange (i.e., employee-supervisor relationship quality) and job satisfaction, as shown below. Use turnover intentions (TURNOVER) and group climate (CLIMATE) as auxiliary variables in the analysis.

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \beta_2(JOBSAT_i) + \varepsilon$$

1. How do the interpretation of the slope coefficients and their standard errors change when using auxiliary variables?

Exercise 3: Bayesian Estimation

Use the pain.dat data set for this example. Write a Bayesian estimation script in Mplus or Blimp that fits a multiple regression model where psychosocial disability (pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.) is regressed on depression and gender, as shown below. If using Blimp, define the gender code as NOMINAL or ORDINAL. If using Mplus, this variable must be treated as normally distributed. Set the burn-in period to 2000 iterations and specify an additional 10,000 iterations for the parameter summary.

$$DISABILITY_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(MALE_i) + \varepsilon$$

1. Provide a substantive interpretation of the posterior means and standard deviations of the two regression slopes.
2. Use credible intervals to evaluate whether zero is a plausible parameter value for the β_1 and β_2 coefficients.

Exercise 4: Bayesian Estimation Convergence Diagnostics

Use the pain.dat data set for this example. Write a Bayesian estimation script that fits a multiple regression model where psychosocial disability (pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.) is regressed on depression and gender, as shown below. If using Blimp, define the gender code as NOMINAL or ORDINAL. If using Mplus, this variable must be treated as normally distributed.

$$DISABILITY_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(MALE_i) + \varepsilon$$

1. Use convergence diagnostics (PSRF and trace plots) to determine the number of iterations required for MCMC to converge. For these diagnostic runs, you can experiment with different burn-in periods. The number of post-burn iterations can be very small because the goal is to examine the MCMC's behavior during the initial estimation period. What are the main features of these diagnostics that informed your choice?

Exercise 5: Bayesian Estimation with Categorical Variables

Use the lmxquality.dat data set for this example. Write a Blimp Bayesian estimation script that fits a multiple regression model where employee empowerment is regressed on leader-member exchange (i.e., employee-supervisor relationship quality) and job satisfaction, as shown below. This analysis is not possible in Mplus because incomplete predictors must be treated as normally distributed variables.

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \beta_2(JOBSAT_i) + \varepsilon$$

1. For the first analysis, define job satisfaction as an ordinal variable (it is a 7-point rating scale). Use convergence diagnostics (PSRF and trace plots) to determine the number of iterations required for MCMC to converge.
2. After deciding on an appropriate burn-in period, perform an analysis by specifying an additional 10,000 iterations for the parameter summary. Provide a description of the posterior distributions of the slope coefficients.

3. For the second analysis, use Blimp's NOMINAL command to define job satisfaction as set of six dummy codes (Blimp uses the lowest category as the reference group). Use convergence diagnostics (PSRF and trace plots) to determine the number of iterations required for MCMC to converge.
4. After deciding on an appropriate burn-in period, perform an analysis by specifying an additional 10,000 iterations for the parameter summary. Provide a description of the posterior distributions of the slope coefficients.

Exercise 6: Bayesian Estimation with Interaction Effects

Use the lmxquality.dat data set for this example. Write a Bayesian estimation script that fits a moderated regression model where employee empowerment is regressed on leader-member exchange (i.e., employee-supervisor relationship quality), group climate, and the interaction of the two. The research question is whether the influence of relationship quality (leader-member exchange) differs as group climate improves or gets worse. The hypothesis is that relationship quality's influence will be more positive when group climate is high and less positive when climate is low (poor).

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \beta_2(CLIMATE_i) + \beta_3(LMX_i)(CLIMATE_i) + \varepsilon$$

1. Use convergence diagnostics (PSRF and trace plots) to determine the number of iterations required for MCMC to converge.
2. After deciding on an appropriate burn-in period, perform an analysis by specifying an additional 10,000 iterations for the parameter summary. Center both predictor variables at their grand means and request simple slopes where leader-member exchange is the focal predictor and group climate is the moderator.
3. Use the credible interval to determine whether zero is a plausible parameter value for the interaction coefficient.
4. Provide an interpretation and description of the interaction effect (i.e., How does the influence of LMX change as climate improves or gets worse?). Use the overall analysis results from the centered solution as well as the simple slopes to inform your response.

Exercise 7: Multiple Imputation

Use the pain.dat data set for this example. Write an FCS multiple imputation script that is appropriate for a multiple regression model where psychosocial disability (pain's impact on emotional behaviors such as psychological autonomy and communication, emotional stability, etc.) is regressed on depression and gender, as shown below.

$$DISABILITY_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(MALE_i) + \varepsilon$$

Use stress and perceived control over pain (STRESS and CONTROL) as auxiliary variables during imputation. Attend to categorical variables, specifying the appropriate metric in the Blimp script.

1. Implement an MCMC algorithm that generates at least 20 imputed data sets. Use convergence diagnostics (PSRF and trace plots) to determine an appropriate burn-in and thinning interval for imputation.
2. Use the software package of your choosing to analyze the imputations and pool the estimates and standard errors.
3. Provide a substantive interpretation of the two slope coefficients and their standard errors.

Exercise 8: Multiple Imputation for Questionnaire Data

Use the depression.dat data set for this example. Write an FCS multiple imputation script that is appropriate for a multiple regression model where pain interference with daily life (a scale score computed as the sum of six questionnaire items) is regressed on depression (a scale score computed as the sum of seven questionnaire items) and treatment group membership. The imputation routine should impute the questionnaire items, and the scale scores will be computed prior to analyzing the data. In addition to the questionnaire items and the treatment group indicator, use age and gender as auxiliary variables.

$$INTERFERE_i = \beta_0 + \beta_1(DEPRESS_i) + \beta_2(TXGROUP_i) + \varepsilon$$

1. Implement an MCMC algorithm that generates at least 20 imputed data sets. Use convergence diagnostics (PSRF and trace plots) to determine an appropriate burn-in and thinning interval for imputation.
2. Use the software package of your choosing to analyze the imputations and pool the estimates and standard errors.
3. Provide a substantive interpretation of the two slope coefficients and their standard errors.