8. Bayes Estimation For Categorical Variables

Categorical And Continuous Variables

Bayesian estimation is ideally suited for mixtures of categorical and continuous variables

Maximum likelihood is far less flexible because it typically assumes multivariate normality

This assumption is problematic for incomplete predictors, where such mixtures are common

Complete Categorical Variables

Complete categorical variables can serve as predictors in the analysis model

Nominal variables must be dummy coded (Blimp's NOMINAL command automates this)

Ordinal variables can be left as-is or coded

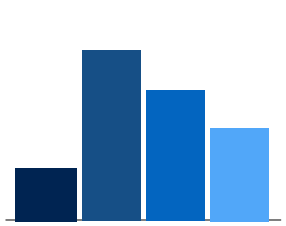
Latent Variable Formulation

The latent variable formulation for categorical variables is based on probit regression

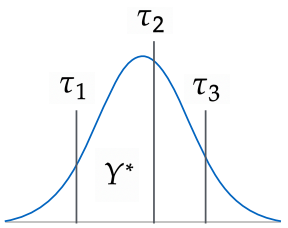
Discrete responses arise from one or more underlying normal latent variables (Y^* variables)

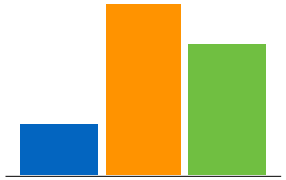
The latent variable distribution for each case is scaled as a z score

Latent Variable Transformations

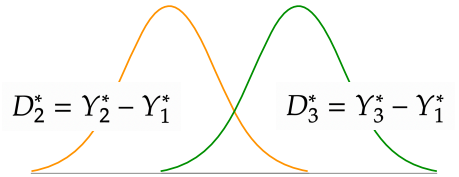


Ordinal





Nominal



Leader-Member Exchange Data

Work-related data for 630 employees nested in 105 different workgroups

The data include work-related variables such as employee empowerment, job satisfaction, turnover intentions, employee-supervisor relationship quality, organizational climate

lmxquality.dat

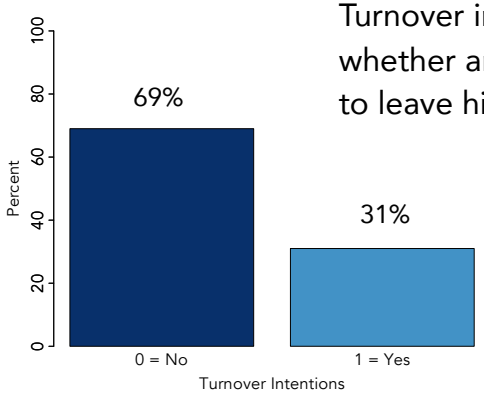
Variable	Name	Missing %	Scaling
Employee identifier	EMPLOYEE	0	Integer index
Team identifier	TEAM	0	Integer index
Turnover intentions	TURNOVER	5.1	0 = intend to stay, 1 = intend to leave
Gender	MALE	0	0 = female, 1 = male
Employee empowerment	EMPOWER	16.2	Continuous
Leader-member exchange	LMXQUALITY	4.1	Continuous
Job satisfaction	JOBSAT	4.8	7-point ordinal scale
Organizational (team) climate	CLIMATE	9.5	Continuous
Organization size	ORGSIZE	5.7	6-point ordinal scale

Binary And Ordinal Variables

The latent formulation for ordinal variables also applies to any binary variable, regardless of whether the categories are ordered (e.g., a gender dummy code)

Binary variables can also be treated as nominal

Turnover Intentions

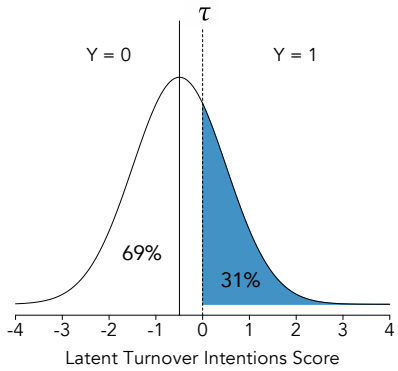


Turnover intentions captures whether an employee plans to leave his or her position

Latent Variable Distribution

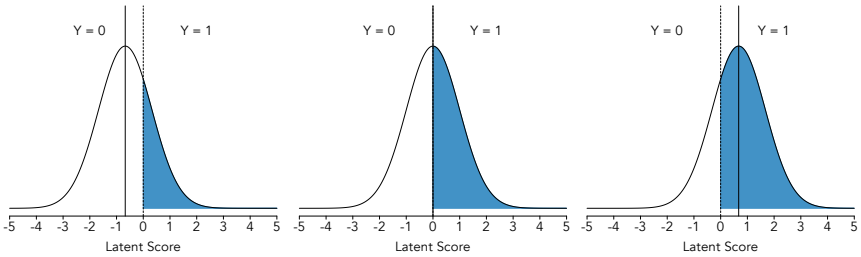
The propensity to quit can be viewed as an underlying normal latent variable

A threshold parameter (z score) separates the upper 31% and lower 69% of the latent scores



Latent Mean And Threshold

The threshold (transition) is fixed at zero, and the latent mean depends on the proportion of ones



p = .25 p = .50 p = .75

Latent Variable Regression Model

Latent variable scores are normally distributed around predicted values

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y^*) + \varepsilon_i$$
$$Y_i^* \sim N(E(Y^*), 1)$$

Residual variance is fixed at one to establish a z score metric for the latent variables

Substantive Example

Leader-member exchange (LMX) measures employee-supervisor relationship quality

$$TURNOVER_i^* = \beta_0 + \beta_1(LMX_i) + \varepsilon_i$$

Probit regression models the association between relationship quality and latent turnover intentions

Latent Scores Are Missing Data

Latent scores are missing for the entire sample!

Each iteration consists of an imputation step and an estimation sequence

MCMC first generates a sample of latent scores or "imputations", then it uses the updated latent scores to estimate the model parameters

MCMC Recipe

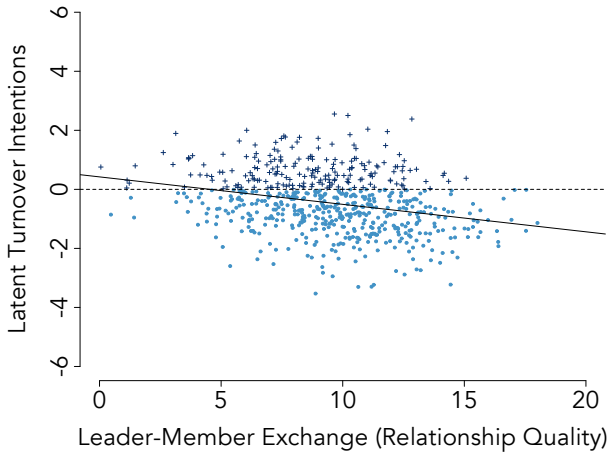
Do for t = 1 to T iterations

1. Estimate the latent scores, given the current regression coefficients and the residual variance fixed at one

2. Estimate the regression coefficients, given the current latent scores and the residual variance fixed at one

Repeat

Latent Variable Regression Model



Imputing The Latent Data

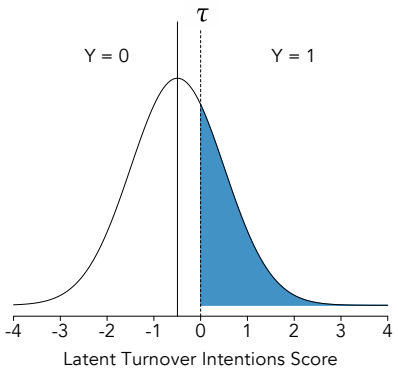
Latent scores are missing for the entire sample and must be imputed at each iteration

If Y is complete, latent scores are constrained to the region of the normal curve above or below the threshold

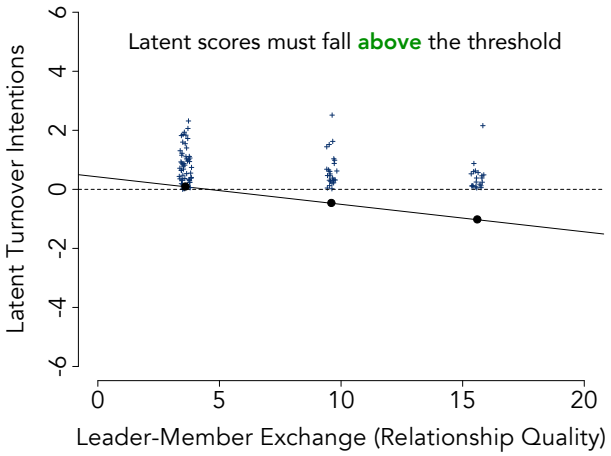
If Y is missing, latent scores are unconstrained, and their location relative to the threshold gives discrete imputes

Latent Scores When Y Is Observed

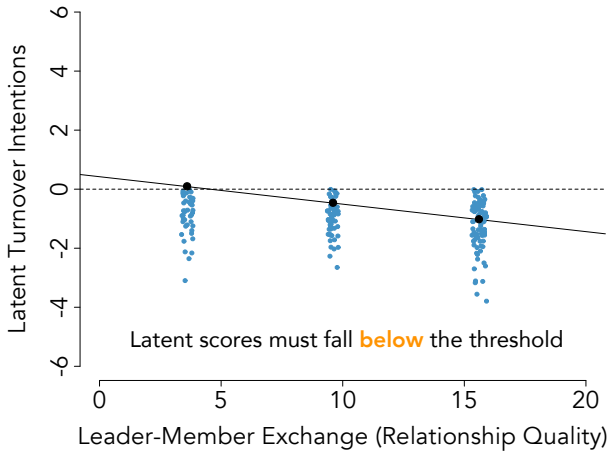
Cases with Y = 0 must have latent scores below the threshold, and cases with Y = 1 must have scores above the threshold



Distributions Of Latent Scores (Y = 1)



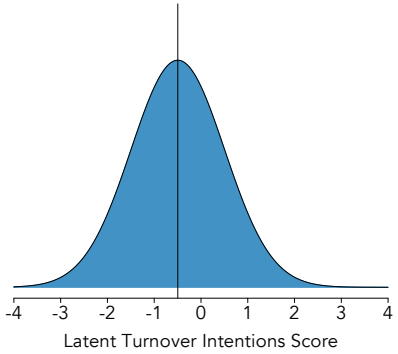
Distributions Of Latent Scores (Y = 0)



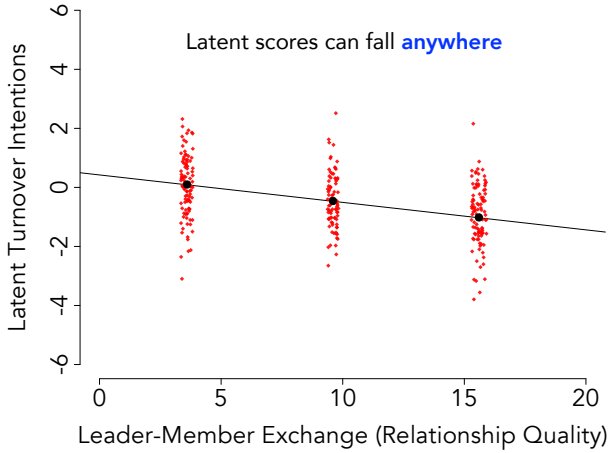
Latent Scores When Y Is Missing

The location in the latent distribution is unknown when Y is missing

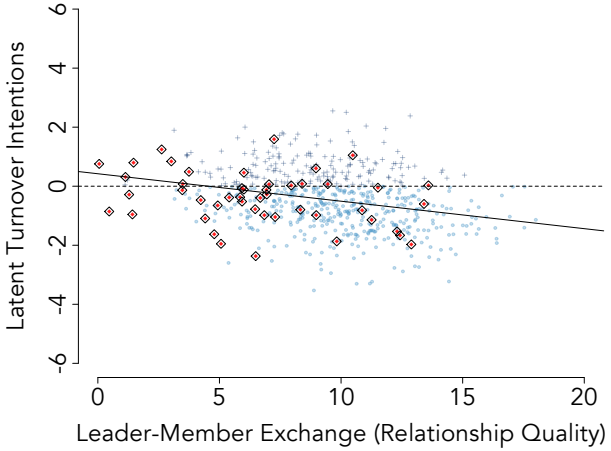
MCMC draws latent scores throughout the entire normal distribution



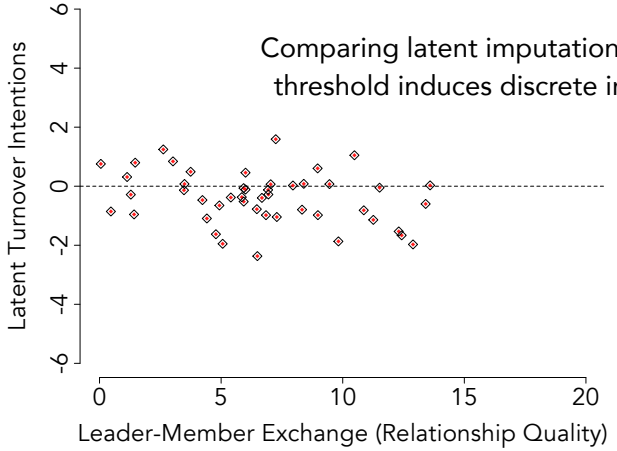
Distributions Of Latent Scores (Y = ?)



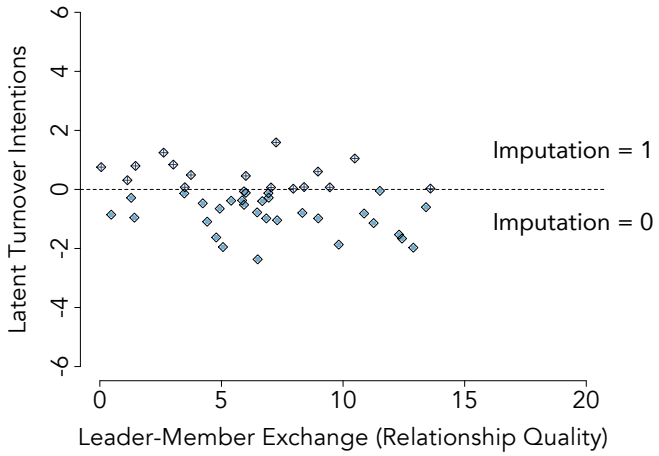
Imputed Latent Data



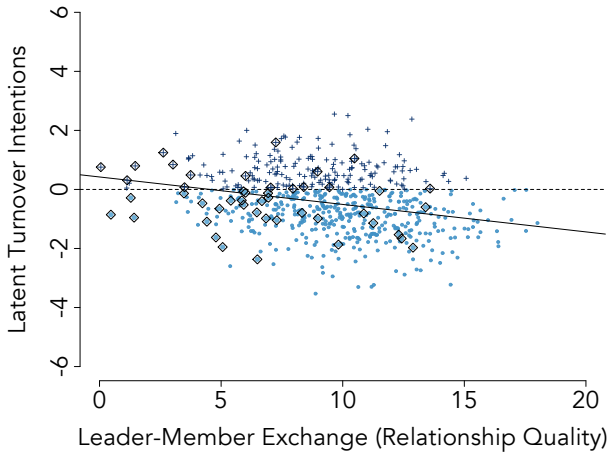
Latent Data For Missing Observations



Discrete Imputes



Full Latent Data Set



Posterior Distribution Summary

Analysis results with 10,000 MCMC iterations

	Mean	Std. Dev.	Lower 2.5%	Upper 97.5%
Intercept	-0.500	0.056	-0.611	-0.391
LMX Slope	-0.092	0.019	-0.128	-0.056
R^2	0.073	0.026	0.027	0.130

Interpretations

The intercept is the predicted latent z-score for an employee with average relationship quality

A one-point increase in relationship quality (the standard deviation is about nine) decreases latent turnover intentions by .092 z-score units

The relation is "significant" because zero is not in the 95% credible interval, and the probability that the slope is negative is $p > .975$

Blimp Bayesian Analysis Script

```
DATA: lmxquality.dat;
VARIABLES: employee team turnover male empower lmxquality jobsat
  climate orgsize;
ORDINAL: turnover;
MISSING: 999;
MODEL: turnover ~ lmxquality;
CENTER: lmxquality;
SEED: 90291;
BURN: 1000;
ITERATIONS: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;
```

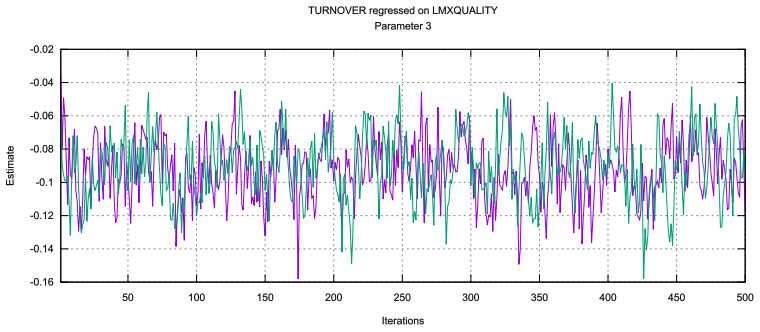

Blimp Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains   | Highest PSR | Parameter # |
                              51 to 100  |       1.131 |           2 |
                             101 to 200  |       1.060 |           5 |
                             151 to 300  |       1.114 |           5 |
                             201 to 400  |       1.009 |           5 |
                             251 to 500  |       1.016 |           2 |
                             301 to 600  |       1.013 |           5 |
                             351 to 700  |       1.009 |           2 |
                             401 to 800  |       1.014 |           5 |
                             451 to 900  |       1.012 |           5 |
                             501 to 1000 |       1.014 |           5 |
```

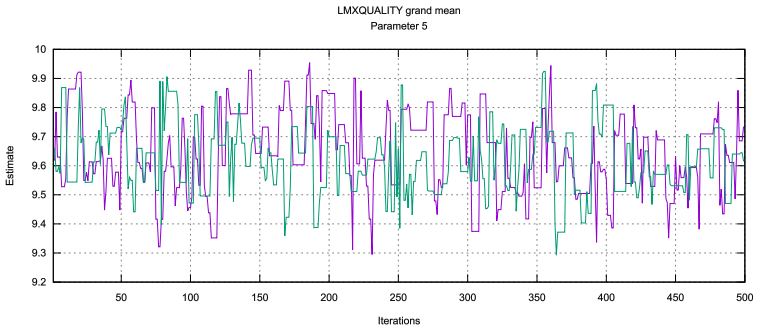

Blimp Trace Plots

Regression slope estimates from 500 iterations



Blimp Trace Plots

Predictor variable mean estimates from 500 iterations



Blimp Output

```
ANALYSIS MODEL ESTIMATES:

Missing outcome: turnover
Grand Mean Centered: lmxquality
                                 ------------------------------------------------------------
Parameters                       |   Mean    |  Median   |  StdDev   |Lower 2.5 |Upper 97.5|
                                 ------------------------------------------------------------
Variances:                       |           |           |           |          |          |
  Residual Var.                  |     1.000|     1.000|     0.000|     1.000|     1.000|
Coefficients:                    |           |           |           |          |          |
  Intercept                      |    -0.500|    -0.500|     0.056|    -0.611|    -0.391|
  lmxquality                     |    -0.092|    -0.091|     0.019|    -0.128|    -0.056|
Thresholds:                      |           |           |           |          |          |
  Tau 1                          |     0.000|     0.000|     0.000|     0.000|     0.000|
                                 |           |           |           |          |          |
                                 |-----------|-----------|-----------|----------|----------|
Proportion Variance Explained    |           |           |           |          |          |
  by Fixed Effects               |     0.073|     0.071|     0.026|     0.027|     0.130|
  by Residual Variation          |     0.927|     0.929|     0.026|     0.870|     0.973|
                                 |           |           |           |          |          |
                                 ------------------------------------------------------------
                                 Summaries based on 10000 iterations using 4 chains
```


Mplus Bayesian Analysis Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality jobsat climate orgsize;
usevariables = turnover lmxquality;
categorical = turnover;
missing = all(999);
DEFINE:
center lmxquality (grandmean);
ANALYSIS:
estimator = bayes;
bseed = 90291;
fbiterations = 10000;
MODEL:
lmxquality;
turnover on lmxquality;
OUTPUT:
stdyx tech8;
```


Mplus Diagnostic Output

```
TECHNICAL 8 OUTPUT

  TECHNICAL 8 OUTPUT FOR BAYES ESTIMATION

    CHAIN     BSEED
    1         90291
    2         255458

                    POTENTIAL       PARAMETER WITH
    ITERATION     SCALE REDUCTION       HIGHEST PSR
    100               1.003                 1
    200               1.000                 1
    300               1.005                 3
    400               1.003                 3
    500               1.008                 2
    600               1.002                 4
    700               1.000                 1
    800               1.002                 4
    900               1.000                 3
    1000              1.000                 3
```


Mplus Output

```
MODEL RESULTS
                                                    Posterior  One-Tailed          95% C.I.
                                Estimate       S.D.       P-Value    Lower 2.5%  Upper 2.5%
 TURNOVER   ON
    LMXQUALITY                    -0.092       0.019      0.000      -0.128      -0.056
 Means
    LMXQUALITY                     0.000       0.123      0.500      -0.250       0.237
 Thresholds
    TURNOVER$1                     0.501       0.054      0.000       0.395       0.604
 Variances
    LMXQUALITY                     9.207       0.542      0.000       8.222      10.360
```

Mplus Output

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	Posterior S.D.	One-Tailed P-Value	95% C.I. Lower 2.5%	Upper 2.5%
TURNOVER ON					
LMXQUALITY	-0.269	0.051	0.000	-0.365	-0.167
Means					
LMXQUALITY	0.000	0.040	0.500	-0.081	0.078
Thresholds					
TURNOVER$1	0.482	0.052	0.000	0.381	0.582
Variances					
LMXQUALITY	1.000	0.000	0.000	1.000	1.000

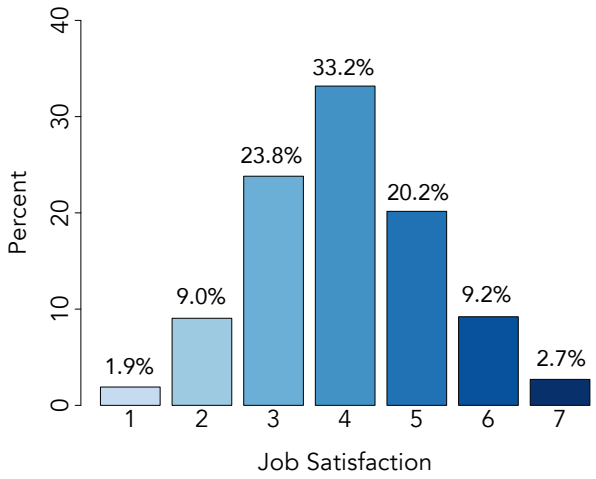
Substantive Example

Job satisfaction is a 7-point rating scale

$$JOBSAT_i^* = \beta_0 + \beta_1(LMX_i) + \varepsilon_i$$

Probit regression models the relation between relationship quality and latent job satisfaction

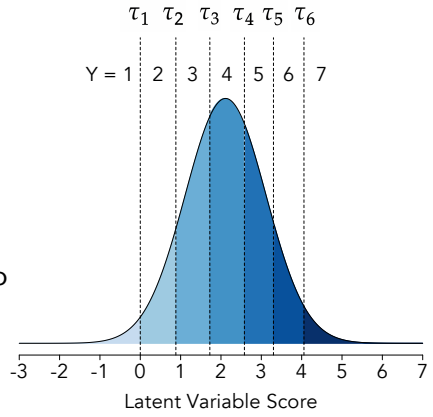
Job Satisfaction Bar Plot



Latent Variable Distribution

The propensity for job satisfaction is an underlying normal latent variable

Six thresholds (z-score cutoffs) slice the latent distribution into seven discrete categories



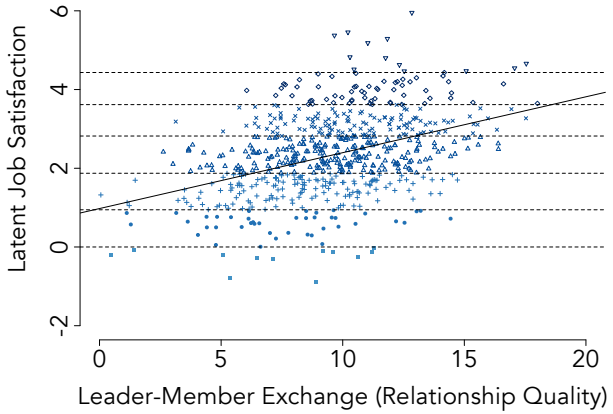
MCMC Recipe

Do for t = 1 to T iterations

1. Estimate thresholds, given the current latent scores, regression coefficients, and residual variance fixed at one
2. Estimate the latent scores, given the current thresholds, regression coefficients, and residual variance fixed at one
3. Estimate the regression coefficients, given the current latent scores and the residual variance fixed at one

Repeat

Latent Variable Regression Model



Posterior Distribution Summary

	Mean	Std. Dev.	Lower 2.5%	Upper 97.5%
Intercept	2.296	0.134	2.052	2.604
LMX Slope	0.160	0.015	0.130	0.190
Threshold 1	0.000	0.000	0.000	0.000
Threshold 2	0.945	0.123	0.725	1.206
Threshold 3	1.873	0.133	1.621	2.166
Threshold 4	2.819	0.141	2.562	3.135
Threshold 5	3.617	0.151	3.341	3.962
Threshold 6	4.433	0.183	4.104	4.837
R^2	0.191	0.029	0.134	0.248

Interpretations

The intercept is the predicted latent z-score for an employee with average relationship quality

A one-point increase in relationship quality (the standard deviation is about nine) increase latent job satisfaction by .16 z-score units

The threshold parameters are z-score cutoffs that carve the latent distribution into discrete responses

Blimp Bayesian Analysis Script

```
DATA: lmxquality.dat;
VARIABLES: employee team turnover male empower lmxquality jobsat
  climate orgsize;
ORDINAL: jobsat;
MISSING: 999;
MODEL: jobsat ~ lmxquality;
CENTER: lmxquality;
SEED: 90291;
BURN: 5000;
ITERATIONS: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;
```

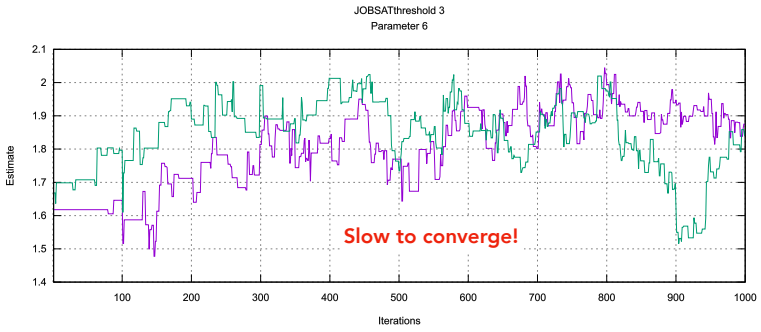

Blimp Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 4 chains   | Highest PSR | Parameter # |
                          51 to 100      |       2.446 |           7 |
                         101 to 200      |       1.798 |           7 |
                             ...         |         ... |         ... |
                        1501 to 3000     |       1.182 |           6 |
                        1551 to 3100     |       1.141 |           6 |
                        1601 to 3200     |       1.094 |           6 |
                        1651 to 3300     |       1.064 |           6 |
                        1701 to 3400     |       1.051 |           6 |
                        1751 to 3500     |       1.047 |           6 |
                        1801 to 3600     |       1.038 |           6 |
                        1851 to 3700     |       1.036 |           6 |
                        1901 to 3800     |       1.029 |           6 |
                        1951 to 3900     |       1.022 |           6 |
                        2001 to 4000     |       1.016 |           6 |
```

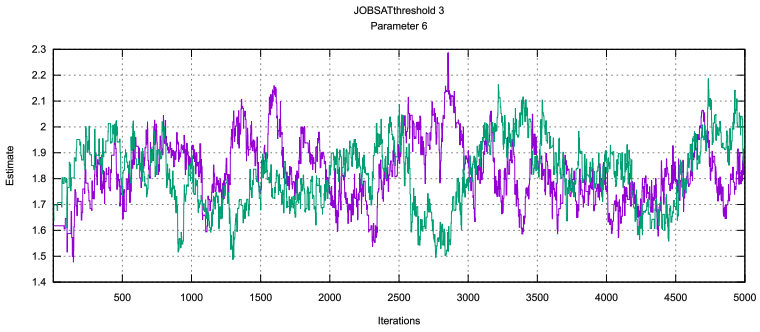

Blimp Trace Plots

Threshold estimates from 1000 iterations (PSRF = 1.299)



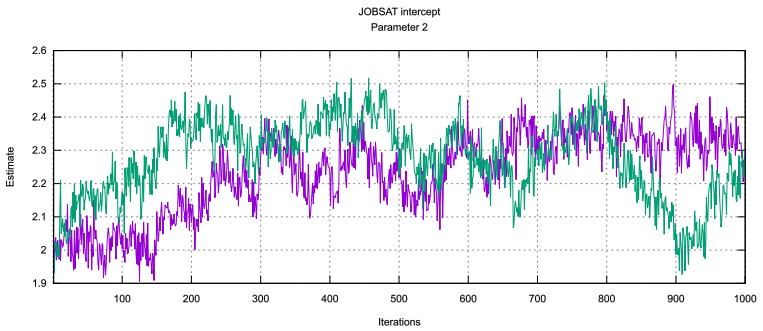
Blimp Trace Plots

Threshold estimates from 5000 iterations



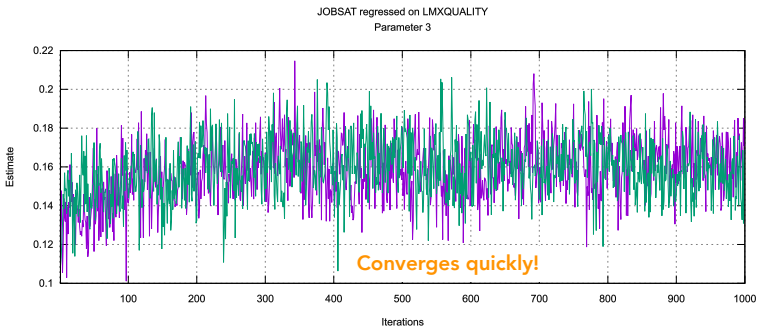
Blimp Trace Plots

Regression intercept estimates from 1000 iterations



Blimp Trace Plots

Regression slope estimates from 1000 iterations



Blimp Output

```
Missing outcome: jobsat

Grand Mean Centered: lmxquality
```

Parameters	Mean	Median	StdDev	Lower 2.5	Upper 97.5
Variances:					
Residual Var.	1.000	1.000	0.000	1.000	1.000
Coefficients:					
Intercept	2.315	2.318	0.131	2.052	2.560
lmxquality	0.161	0.161	0.015	0.130	0.191
Thresholds:					
Tau 1	0.000	0.000	0.000	0.000	0.000
Tau 2	0.952	0.951	0.118	0.727	1.192
Tau 3	1.893	1.893	0.128	1.634	2.144
Tau 4	2.838	2.842	0.139	2.549	3.092
Tau 5	3.640	3.645	0.151	3.322	3.912
Tau 6	4.459	4.457	0.186	4.088	4.811
Standardized Coefficients:					
lmxquality	0.437	0.438	0.034	0.367	0.501
Proportion Variance Explained					
by Fixed Effects	0.192	0.192	0.030	0.135	0.251
by Residual Variation	0.808	0.808	0.030	0.749	0.865

```
Summaries based on 10000 iterations using 4 chains
```


Mplus Bayesian Analysis Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality jobsat climate orgsize;
usevariables = jobsat lmxquality;
categorical = jobsat;
missing = all(999);
DEFINE:
center lmxquality (grandmean);
ANALYSIS:
estimator = bayes;
bseed = 90291;
fbiterations = 10000;
MODEL:
lmxquality;
jobsat on lmxquality;
OUTPUT:
stdyx tech8;
```

Mplus Diagnostic Output

```
TECHNICAL 8 OUTPUT

  TECHNICAL 8 OUTPUT FOR BAYES ESTIMATION

    CHAIN     BSEED
    1         90291
    2         255458

                    POTENTIAL        PARAMETER WITH
    ITERATION    SCALE REDUCTION     HIGHEST PSR
    100              1.836               4
    200              1.355               4
     ...               ...              ...
    3600             1.083               9
    3700             1.068               9
    3800             1.055               9
    3900             1.047               9
    4000             1.059               9
    4100             1.050               9
    4200             1.054               9
    4300             1.040               9
    4400             1.044               9
    4500             1.047               9
```


Mplus Output

```
MODEL RESULTS

                                    Posterior  One-Tailed          95% C.I.
                  Estimate           S.D.       P-Value    Lower 2.5%  Upper 2.5%
JOBSAT     ON
   LMXQUALITY        0.160          0.015      0.000       0.131       0.190
Means
   LMXQUALITY       -0.013          0.123      0.456      -0.256       0.227
Thresholds
   JOBSAT$1         -2.265          0.133      0.000      -2.539      -2.029
   JOBSAT$2         -1.355          0.074      0.000      -1.502      -1.211
   JOBSAT$3         -0.426          0.056      0.000      -0.537      -0.317
   JOBSAT$4          0.521          0.055      0.000       0.416       0.628
   JOBSAT$5          1.318          0.071      0.000       1.180       1.464
   JOBSAT$6          2.135          0.110      0.000       1.923       2.363
Variances
   LMXQUALITY        9.187          0.536      0.000       8.224      10.328
```


Mplus Output

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

                                    Posterior  One-Tailed          95% C.I.
                  Estimate           S.D.       P-Value    Lower 2.5%  Upper 2.5%
 JOBSAT     ON
   LMXQUALITY        0.436          0.035      0.000       0.365       0.502

...

R-SQUARE

                                    Posterior  One-Tailed          95% C.I.
    Variable      Estimate           S.D.       P-Value    Lower 2.5%  Upper 2.5%
    JOBSAT            0.190          0.030      0.000       0.134       0.252
```


Regression With Categorical Predictors

Linear regression with a continuous covariate and two dummy codes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{2i} + \beta_3 D_{3i} + \varepsilon_i$$

$$Y_i \sim N\left(E(Y|X, D_2, D_3), \sigma_\varepsilon^2\right)$$

The first category is the reference group, and D_2 and D_3 are dummy codes contrasting the second and third groups to the reference category

Chronic Pain Data

Pain-related data for 630 chronic pain patients

In addition to pain intensity, the data include behavioral variables such as work hours and exercise frequency and psychological variables such as perceived control over pain, depression, and pain interference with daily life

pain.dat

Variable	Name	Missing %	Scaling
Patient identifier	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Age	AGE	0	Continuous
Education level	EDUGROUP	0	3-point ordinal scale
Work hours per week	WORKHRS	11.7	Continuous
Exercise hours per week	EXERCISE	1.7	8-point ordinal scale
Pain intensity	PAIN	7.3	1 = little, 2 = moderate, 3 = severe
Anxiety	ANXIETY	6.0	Continuous
Stress	STRESS	0	7-point ordinal scale
Perceived control over pain	CONTROL	0	Continuous
Pain interference with life	INTERFERE	13.3	Continuous
Depression	DEPRESS	13.3	Continuous
Psychosocial disability	DISABILITY	3.0	Continuous

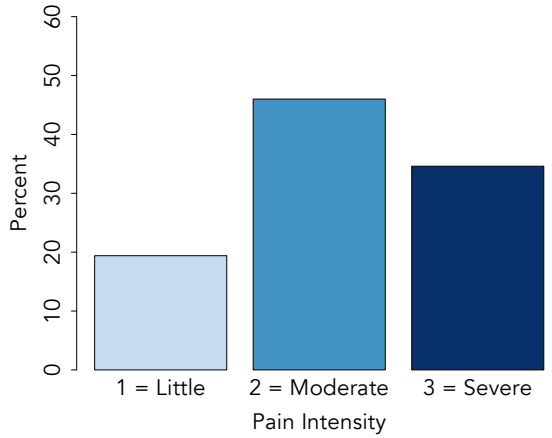
Substantive Example

Pain interference is regressed on perceived control, and a three-category pain intensity rating

$$INTERFERE_i = \beta_0 + \beta_1(CONTROL_i) + \beta_2(MODERATE_i) + \beta_3(SEVERE_i) + \varepsilon_i$$

Dummy codes contrast moderate and severe pain groups vs. the no/little pain reference category

Pain Intensity Bar Plot



Nominal Variables

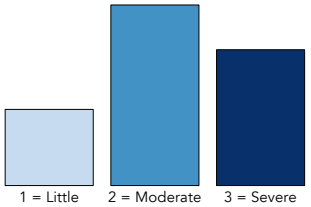
Each nominal category has an underlying latent variable, and the discrete response corresponds to the highest latent score

Using all latent scores is redundant, so difference scores contrast each latent variable relative to that of the reference category (akin to latent dummy codes)

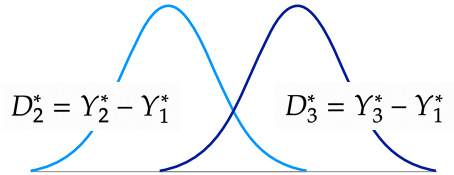
Nominal variables appear as dummy codes in the analysis model, but imputation is on the latent metric

Dual Representation Of A Categorical Predictor

Analysis model



Imputation Model



Comparison To Ordinal Probit Model

The latent formulation for nominal variables (multinomial probit) does not require threshold parameters (will usually converge faster as such)

The magnitude and rank order of the latent differences determines category membership

e.g., A response in the reference category requires all negative latent difference scores

Posterior Distribution Summary

Analysis results with 10,000 MCMC iterations

Parameter	Mean	Std. Dev.	Lower 2.5%	Upper 97.5%
Intercept	20.294	1.070	18.148	22.380
CONTROL slope	-0.468	0.090	-0.648	-0.291
MODERATE slope	6.548	1.261	4.042	9.065
SEVERE slope	11.954	1.339	9.352	14.600

Interpretations

The intercept is the predicted pain interference score for a patient with little/no pain and average perceived control (control is grand mean centered)

Controlling for pain intensity ratings, a one-unit increase in perceived control decreases pain interference by .468

Controlling for perceived control, the pain interference mean for patients with moderate pain is 6.548 points higher than that of the comparison group (little pain)

Blimp Bayesian Analysis Script

```
DATA: pain.dat;
VARIABLES: id male age edugroup workhrs exercise pain anxiety stress
control interfere depress disability;
NOMINAL: pain;
MISSING: 999;
MODEL: interfere ~  control pain;
CENTER: control;
SEED: 90291;
BURN: 1000;
ITERATIONS: 10000;
CHAINS: 4 processors 4;
OPTIONS: psr;
```

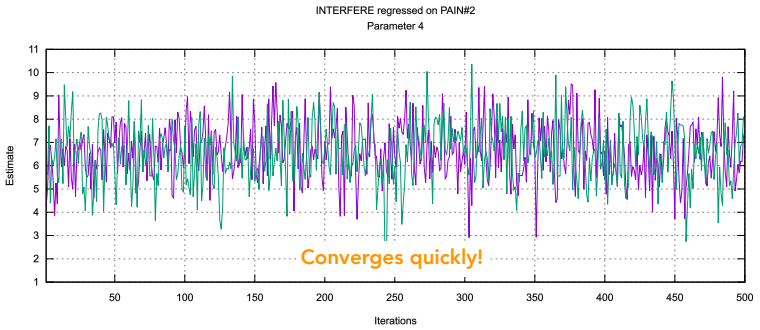

Blimp Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations across 10 chains  | Highest PSR | Parameter # |
                          51 to 100      |       1.162 |          13 |
                         101 to 200      |       1.034 |          10 |
                         151 to 300      |       1.035 |          13 |
                         201 to 400      |       1.031 |          11 |
                         251 to 500      |       1.025 |          11 |
                         301 to 600      |       1.024 |          13 |
                         351 to 700      |       1.013 |          11 |
                         401 to 800      |       1.014 |          11 |
                         451 to 900      |       1.013 |          11 |
                         501 to 1000     |       1.014 |          11 |
```

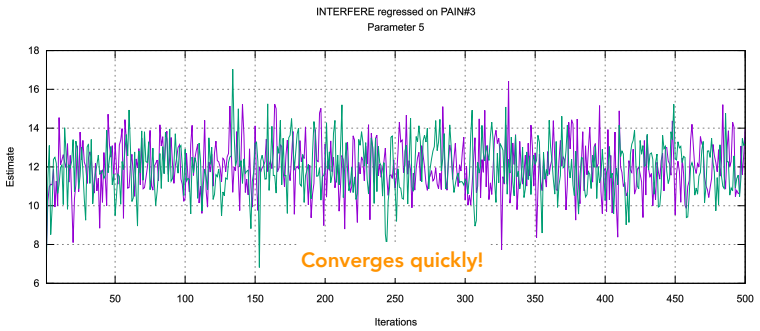

Blimp Trace Plots

Moderate dummy slope estimates from 500 iterations



Blimp Trace Plots

Severe dummy slope estimates from 500 iterations



Blimp Output

```
Missing outcome: interfere

Grand Mean Centered: control

                                 ------------------------------------------------------
Parameters                       |   Mean   |  Median  |  StdDev  |Lower 2.5 |Upper 97.5|
                                 ------------------------------------------------------
Variances:                       |          |          |          |          |          |
  Residual Var.                  |   49.996|   49.739|    4.736|   41.654|   60.337|
                                 |----------|----------|----------|----------|----------|
Coefficients:                    |          |          |          |          |          |
  Intercept                      |   20.294|   20.312|    1.070|   18.148|   22.380|
  control                        |   -0.468|   -0.468|    0.090|   -0.648|   -0.291|
  pain#2                         |    6.548|    6.539|    1.261|    4.042|    9.065|
  pain#3                         |   11.954|   11.956|    1.339|    9.352|   14.600|
                                 |----------|----------|----------|----------|----------|
Standardized Coefficients:       |          |          |          |          |          |
  control                        |   -0.277|   -0.277|    0.050|   -0.375|   -0.174|
  pain#2                         |    0.368|    0.369|    0.068|    0.231|    0.501|
  pain#3                         |    0.636|    0.638|    0.061|    0.510|    0.750|
                                 |----------|----------|----------|----------|----------|
Proportion Variance Explained    |          |          |          |          |          |
  by Fixed Effects               |    0.367|    0.368|    0.045|    0.278|    0.453|
  by Residual Variation          |    0.633|    0.632|    0.045|    0.548|    0.722|
                                 |          |          |          |          |          |
                                 ------------------------------------------------------
Summaries based on 10000 iterations using 4 chains
```