# 6. Bayesian Estimation With Missing Data Imputation



## Bayesian Estimation And Imputation

Like parameters, missing values are quantities to be estimated at each MCMC iteration

MCMC first uses filled-in data from the previous iteration to estimate model parameters

The parameters define a distribution of missing values from which imputations are drawn



## Simple Regression Analysis

Simple regression where $Y$ has missing values

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y|X) + \varepsilon_i$$

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

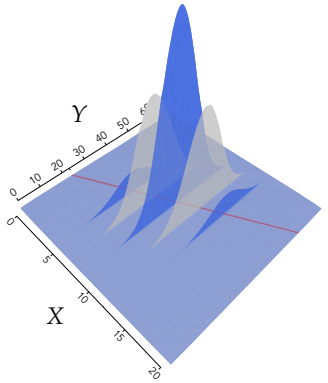
$E(Y|X)$ is a predicted value from the regression



## Distribution Of The Outcome

The regression model says that $Y$ scores are normally distributed around predicted values

The residual variance defines the spread of the $Y$ scores around the regression line




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

## 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 |



## Substantive Example

Employee empowerment regressed on employee-supervisor relationship quality

$$EMPOWER_i = \beta_0 + \beta_1(LMX_i) + \varepsilon_i$$

Empowerment score are normally distributed around the regression line



## MCMC Recipe

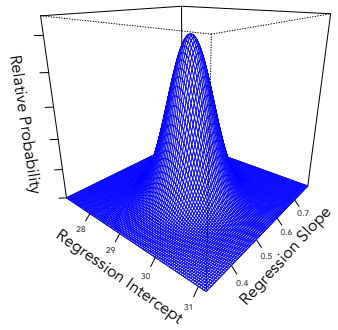
Do for $t$ = 1 to $T$ iterations

1. Estimate the regression coefficients, given the filled-in data and current value of the residual variance
2. Estimate the residual variance, given the filled-in data and current values of the coefficients
3. Estimate (impute) missing values, given the regression model parameters

Repeat

## Conditional Distribution Of The Coefficients

$$f\left(\beta|\,\sigma_\varepsilon^2, data\right) \propto f(\beta) \times f\left(data|\beta, \sigma_\varepsilon^2\right) \propto MVN\left(\hat{\beta}_{\text{OLS}}, \Sigma_{\hat{\beta}}\right)$$




The posterior distribution of $\beta$ is a multivariate normal distribution centered at OLS estimates

$$data = \left(Y_{(mis)}, Y_{(obs)}, X\right)$$



## Conceptual Explanation Of Estimation

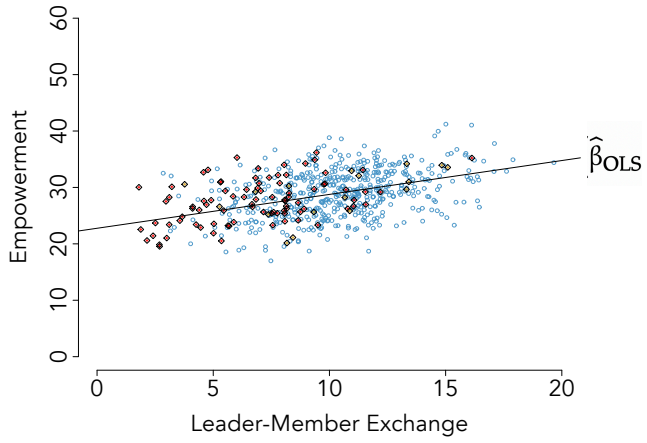
1. Compute ordinary least squares estimates from the filled-in data set

2. Add random noise terms to the OLS estimates, the magnitudes of which depend on OLS standard errors

$\beta_0$ and $\beta_1$ = OLS estimates + noise

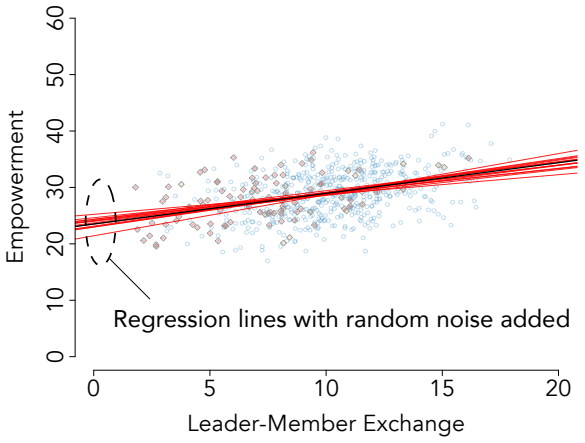


## Regression From Filled-In Data

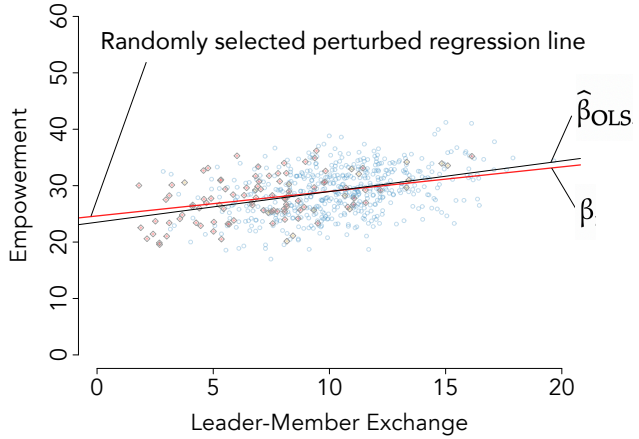





## Perturbed Regression Lines



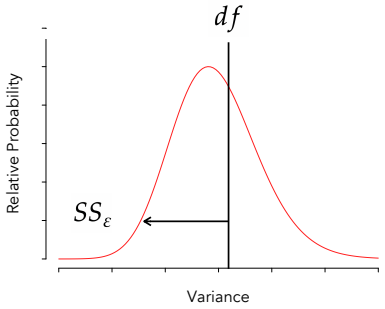
## Updated Regression Line






## Conditional Distribution Of The Variance

$$f\left(\sigma_\varepsilon^2|\beta, data\right) \propto f\left(\sigma_\varepsilon^2\right) \times f\left(data|\beta, \sigma_\varepsilon^2\right) \propto IG\left(\frac{df}{2}, \frac{\sum\left(Y_i - E(Y|X)\right)^2}{2}\right)$$

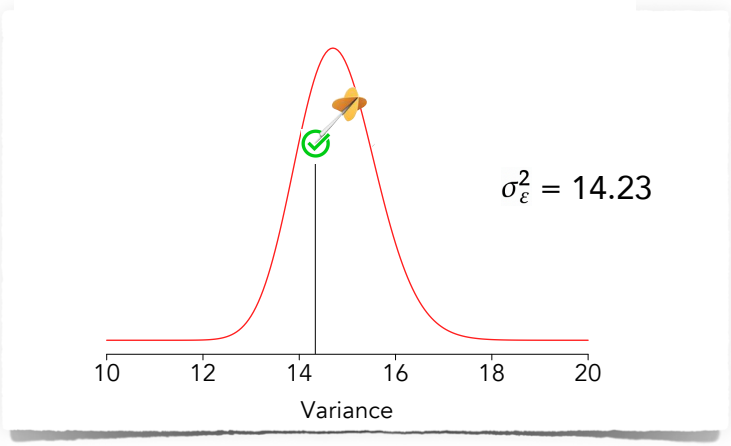



The posterior distribution of $\sigma_\varepsilon^2$ is a positively skewed inverse gamma distribution

$$data = \left(Y_{(mis)}, Y_{(obs)}, X\right)$$



## Monte Carlo Dartboard



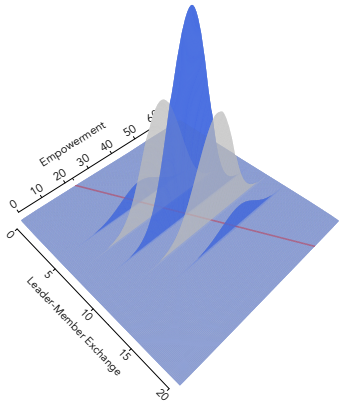



## Distribution Of Imputations

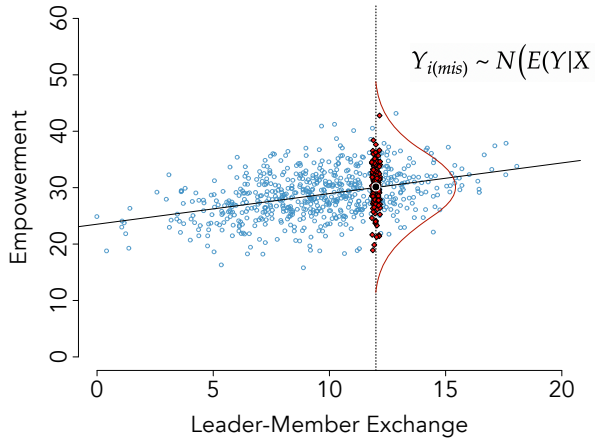
The regression parameters define the distribution of missing values

$$Y_{i(mis)} \sim N\left(E(Y|X), \sigma_\varepsilon^2\right)$$

Imputations are normally distributed around predicted values

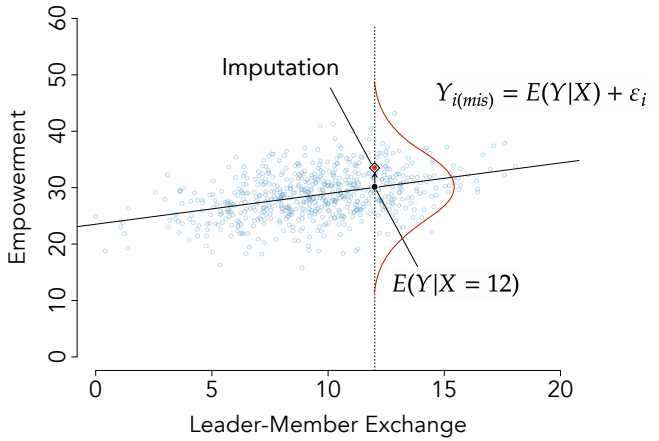


## Distribution Of Imputations



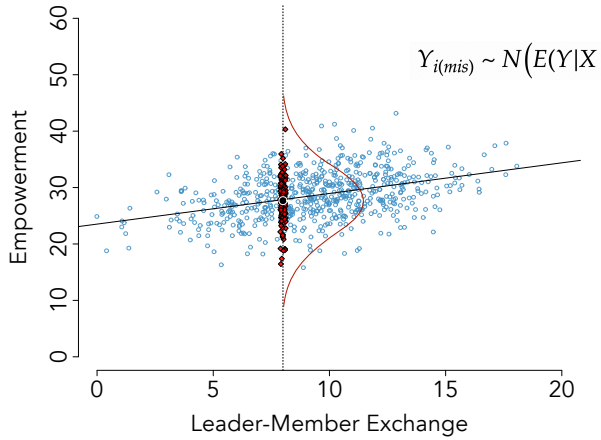

17

## Imputation = Predicted Score + Noise



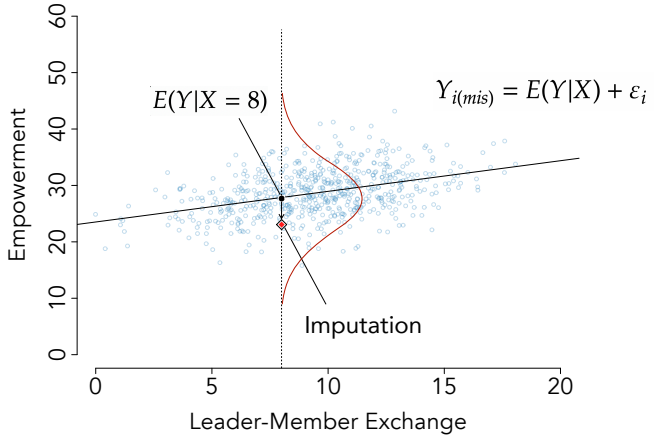

18

## Distribution Of Imputations



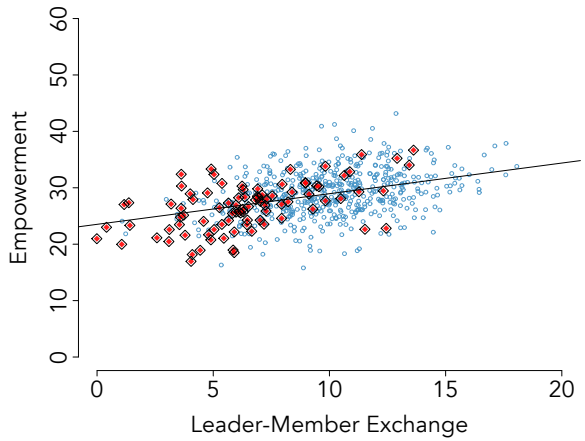

19

## Imputation = Predicted Score + Noise




20

## Imputed Data Set






## Incomplete Predictors

An incomplete predictor variable requires a distribution and its own regression model

Imputations are more complex because the predictor appears in two models — as a covariate in the substantive analysis and a dependent variable in its own model



## Predictor Distribution

Bayes' rule says that the conditional distribution of $X$ is the product of two distributions (models)

$$p(X|Y) = \frac{p(Y|X) \times p(X)}{p(Y)} \propto p(Y|X) \times p(X)$$

Predictor model

Substantive model



## Simple Regression Analysis

Substantive model

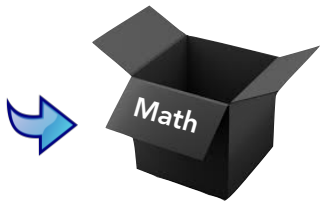
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i = E(Y|X) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y|X), \sigma_\varepsilon^2\right)$$

Predictor model

$$X_i = \gamma_0 + r$$
$$X_i \sim N\left(E(X), \sigma_r^2\right)$$

## Multiplying Two Normal Distributions …

$$f(X|Y) \propto \exp\left\{-\frac{1}{2}\frac{\left(Y_i - E(Y|X)\right)^2}{\sigma_\varepsilon^2}\right\} \times \exp\left\{-\frac{1}{2}\frac{\left(X_i - E(X)\right)^2}{\sigma_r^2}\right\}$$



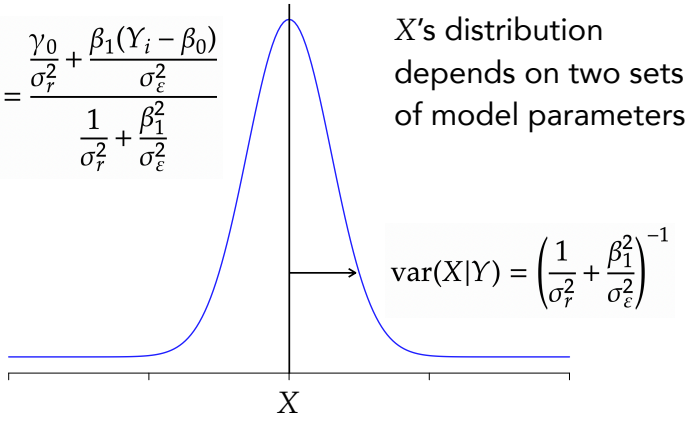



## Gives An Ugly Normal Distribution

$$E(X|Y) = \frac{\frac{\gamma_0}{\sigma_r^2} + \frac{\beta_1(Y_i - \beta_0)}{\sigma_\varepsilon^2}}{\frac{1}{\sigma_r^2} + \frac{\beta_1^2}{\sigma_\varepsilon^2}}$$

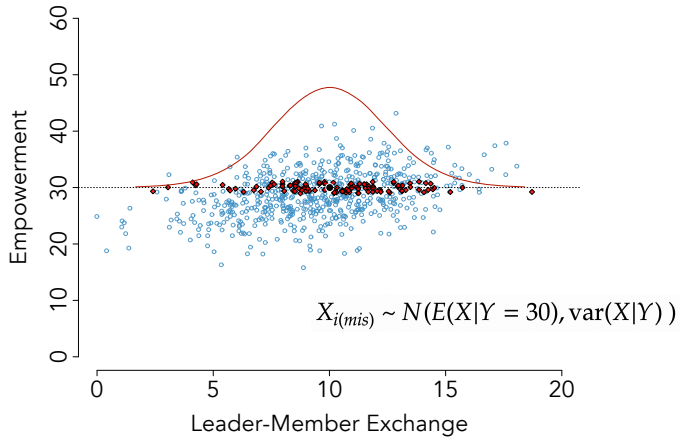
$X$'s distribution depends on two sets of model parameters

$$\text{var}(X|Y) = \left(\frac{1}{\sigma_r^2} + \frac{\beta_1^2}{\sigma_\varepsilon^2}\right)^{-1}$$



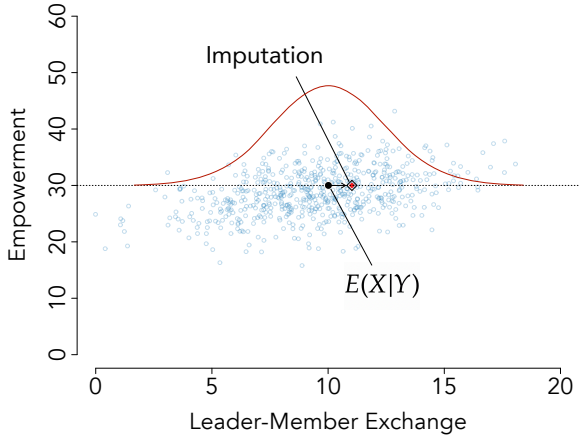



## Distribution Of Imputations

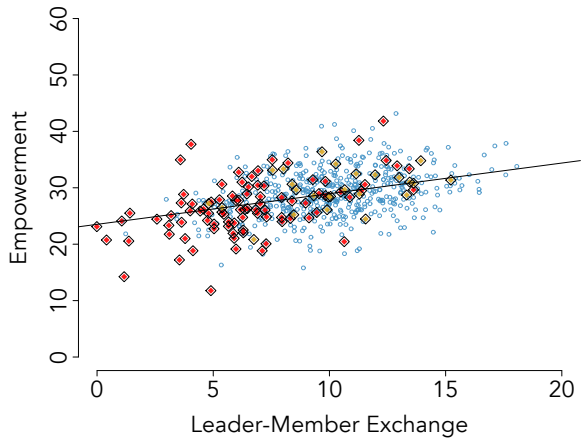





## Imputation = Predicted Score + Noise



# Imputed Data Set






# MCMC Recipe

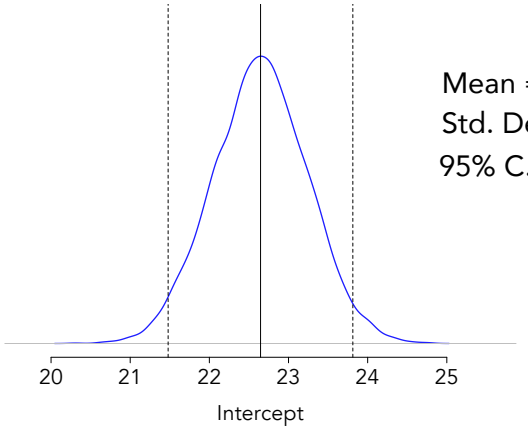
Do for $t$ = 1 to $T$ iterations

1. Estimate substantive model parameters (regression coefficients, residual variance), given the filled-in data
2. Impute missing $Y$ values, given the regression model parameters
3. Estimate predictor model parameters (mean and variance), given the filled-in data
4. Impute missing $X$ values, given both sets of model parameters

Repeat



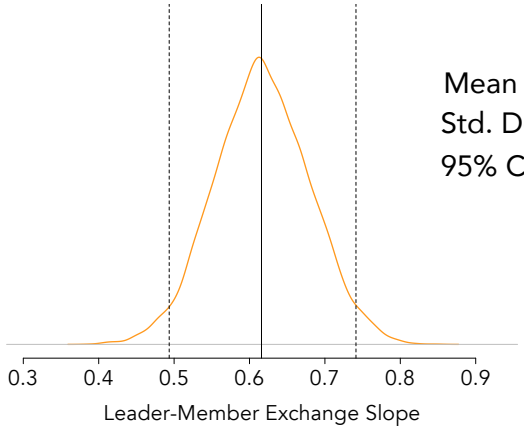
# Posterior Distribution Of The Regression Intercept




Mean = 22.65
Std. Dev. = .60
95% C.I. = (21.48, 23.81)



# Posterior Distribution Of The Leader Member Exchange Slope Coefficient




Mean = .62
Std. Dev. = .06
95% C.I. = (.49, .74)

## Interpretations

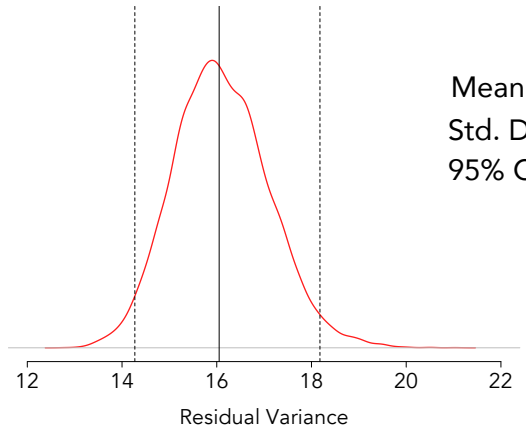
A one-point increase in relationship quality increases empowerment by .62, on average

We are 95% certain that the parameter falls between .49 and .74

The probability that the parameter is greater than zero is at least .975 (virtually 100%)



## Posterior Distribution Of The Residual Variance




Mean = 16.1
Std. Dev. = 1.00
95% C.I. = (14.27, 18.18)



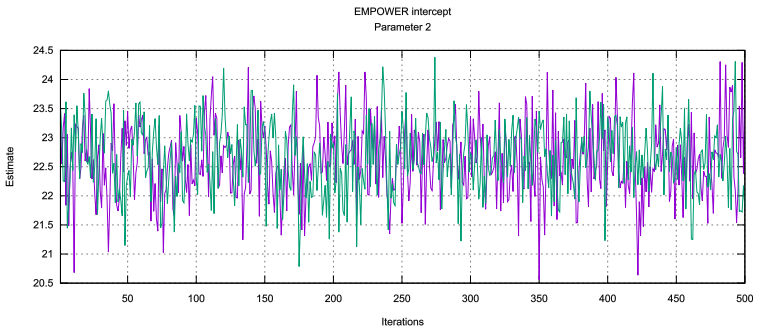
## Blimp Bayesian Analysis Script

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



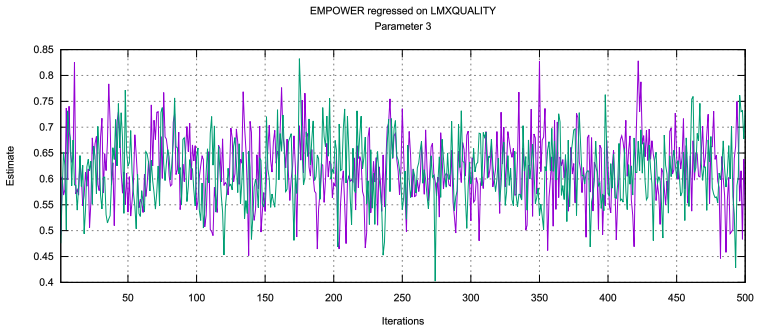
## Blimp Trace Plots

Intercept estimates from 500 iterations



# Blimp Trace Plots

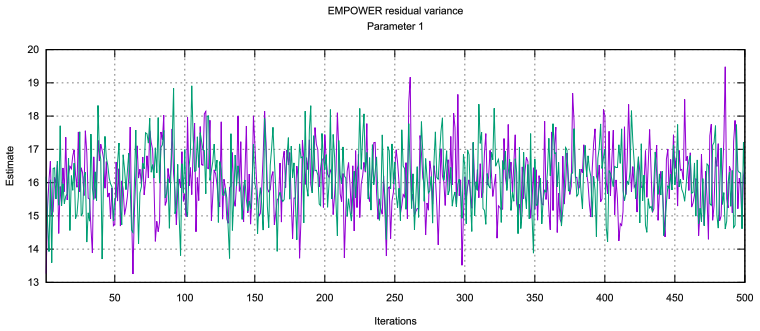
Slope estimates from 500 iterations




37

# Blimp Trace Plots

Residual variance estimates from 500 iterations




38

# Blimp Output

```
ANALYSIS MODEL ESTIMATES:

Missing outcome: empower
```

| Parameters | Mean | Median | StdDev | Lower 2.5 | Upper 97.5 |
|---|---|---|---|---|---|
| Variances: | | | | | |
| Residual Var. | 16.111 | 16.073 | 1.010 | 14.250 | 18.222 |
| Coefficients: | | | | | |
| Intercept | 22.649 | 22.651 | 0.600 | 21.463 | 23.841 |
| lmxquality | 0.616 | 0.616 | 0.063 | 0.491 | 0.739 |
| Standardized Coefficients: | | | | | |
| lmxquality | 0.421 | 0.421 | 0.037 | 0.343 | 0.490 |
| Proportion Variance Explained | | | | | |
| by Fixed Effects | 0.178 | 0.178 | 0.031 | 0.118 | 0.240 |
| by Residual Variation | 0.822 | 0.822 | 0.031 | 0.760 | 0.882 |

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

39

# Blimp Output (Predictor Distribution)

```
Covariate Models
----------------

Missing covariate:  lmxquality
```

| Parameters | Mean | Median | StdDev | Lower 2.5 | Upper 97.5 |
|---|---|---|---|---|---|
| Grand Mean | 9.633 | 9.633 | 0.123 | 9.393 | 9.876 |
| Level 1: | | | | | |
| Residual Var. | 9.145 | 9.129 | 0.532 | 8.154 | 10.242 |

40

## Mplus Bayesian Analysis Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality jobsat climate orgsize;
usevariables = empower lmxquality;
ANALYSIS:
estimator = bayes;
bseed = 90291;
fbiterations = 10000;
MODEL:
lmxquality;
empower on lmxquality;
OUTPUT:
stdyx tech8 patterns;
```



## Mplus Output

```
MODEL RESULTS

                                    Posterior  One-Tailed          95% C.I.
                  Estimate            S.D.       P-Value    Lower 2.5%  Upper 2.5%

EMPOWER    ON
   LMXQUALITY        0.617          0.062      0.000       0.495       0.737

Means
   LMXQUALITY        9.629          0.124      0.000       9.389       9.873

Intercepts
   EMPOWER          22.648          0.588      0.000      21.486      23.798

Variances
   LMXQUALITY        9.178          0.531      0.000       8.229      10.294

Residual Variances
   EMPOWER          16.126          1.010      0.000      14.285      18.281
```



## Mplus Output

```
STDYX Standardization

                                    Posterior  One-Tailed          95% C.I.
                  Estimate            S.D.       P-Value    Lower 2.5%  Upper 2.5%

EMPOWER    ON
   LMXQUALITY        0.359          0.043      0.000       0.272       0.439
   JOBSAT            0.155          0.044      0.000       0.068       0.239
LMXQUALI WITH
   JOBSAT            0.422          0.034      0.000       0.352       0.486
Residual Variances
   EMPOWER           0.799          0.032      0.000       0.733       0.859

R-SQUARE

                                    Posterior  One-Tailed          95% C.I.
   Variable       Estimate            S.D.       P-Value    Lower 2.5%  Upper 2.5%
   EMPOWER           0.201          0.032      0.000       0.141       0.267
```



## Multiple Regression Analysis

Multiple regression with two predictors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i = E(Y|X) + \varepsilon_i$$

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

$Y$ is normally distributed around the regression line (predicted values) with constant variation

## Predictor Distribution

The conditional distribution of $X_1$ (or $X_2$) is again the product of two distributions (models)

Predictor model

$$p(X_1|Y,X_2) = \frac{p(Y|X_1,X_2) \times p(X_1,X_2)}{p(Y)} \propto p(Y|X_1,X_2) \times p(X_1|X_2)$$

Substantive model



## Predictor Models

Each incomplete predictor variable requires a regression model linking it to other predictors

$$X_{1i} = \gamma_{10} + \gamma_{11}X_{2i} + r_{1i}$$
$$X_{1i} \sim N\left(E(X_1|X_2), \text{var}(X_1|X_2)\right)$$

$$X_{2i} = \gamma_{20} + \gamma_{21}X_{1i} + r_{2i}$$
$$X_{2i} \sim N\left(E(X_2|X_1), \text{var}(X_2|X_1)\right)$$

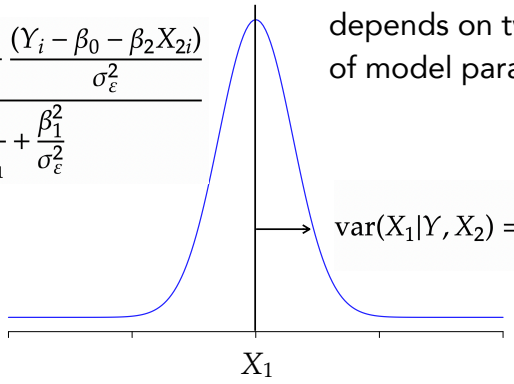


## Ugly Normal Distribution Revisited

$$E(X_1|Y,X_2) = \frac{\frac{\gamma_{10}+\gamma_{11}X_{2i}}{\sigma^2_{r_1}} + \frac{(Y_i-\beta_0-\beta_2X_{2i})}{\sigma^2_\varepsilon}}{\frac{1}{\sigma^2_{r_1}} + \frac{\beta_1^2}{\sigma^2_\varepsilon}}$$

$X_1$'s distribution again depends on two sets of model parameters

$$\text{var}(X_1|Y,X_2) = \left(\frac{1}{\sigma^2_{r_1}} + \frac{\beta_1^2}{\sigma^2_\varepsilon}\right)^{-1}$$



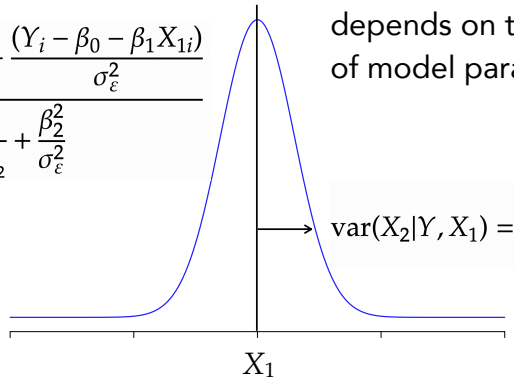



## Ugly Normal Distribution Revisited

$$E(X_2|Y,X_1) = \frac{\frac{\gamma_{20}+\gamma_{21}X_{1i}}{\sigma^2_{r_2}} + \frac{(Y_i-\beta_0-\beta_1X_{1i})}{\sigma^2_\varepsilon}}{\frac{1}{\sigma^2_{r_2}} + \frac{\beta_2^2}{\sigma^2_\varepsilon}}$$

$X_2$'s distribution also depends on two sets of model parameters

$$\text{var}(X_2|Y,X_1) = \left(\frac{1}{\sigma^2_{r_2}} + \frac{\beta_2^2}{\sigma^2_\varepsilon}\right)^{-1}$$



## MCMC Recipe

Do for $t$ = 1 to $T$ iterations

1. Estimate substantive model parameters (regression coefficients, residual variance), given the filled-in data
2. Impute missing $Y$ values, given the regression model parameters
3. Estimate predictor model regression parameters (regression coefficients, residual variances), given the filled-in data
4. Impute missing predictors, given two sets of model parameters

Repeat



## Substantive Example

Substantive model

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

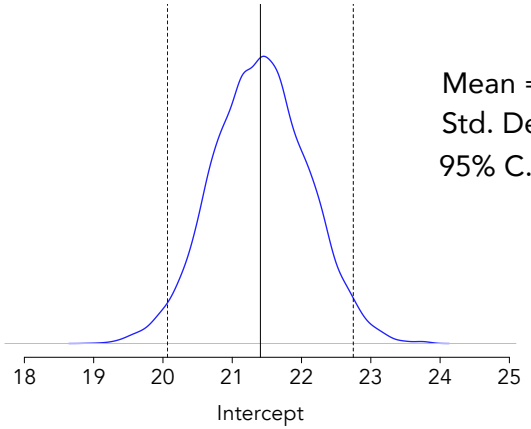
Predictor models

$$LMX_i = \gamma_{10} + \gamma_{11}(JOBSAT_i) + r_{1i}$$

$$JOBSAT_i = \gamma_{20} + \gamma_{21}(LMX_i) + r_{2i}$$



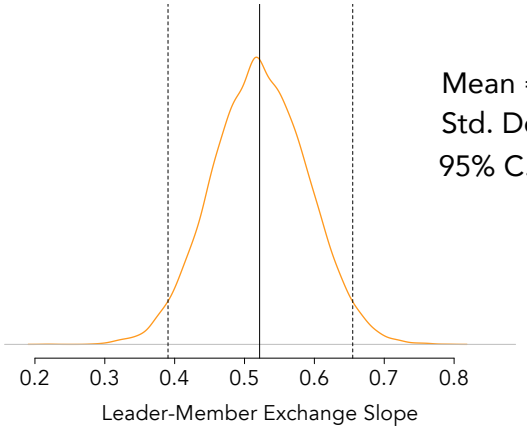
## Posterior Distribution Of The Regression Intercept




Mean = 21.39
Std. Dev. = .69
95% C.I. = (20.06, 22.78)

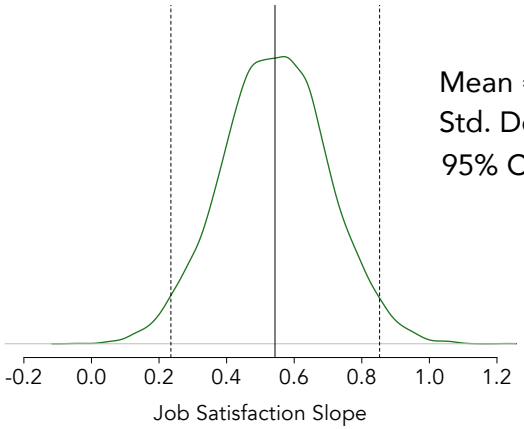


## Posterior Distribution Of The Leader-Member Exchange Slope Coefficient




Mean = .52
Std. Dev. = .07
95% C.I. = (.39, .66)

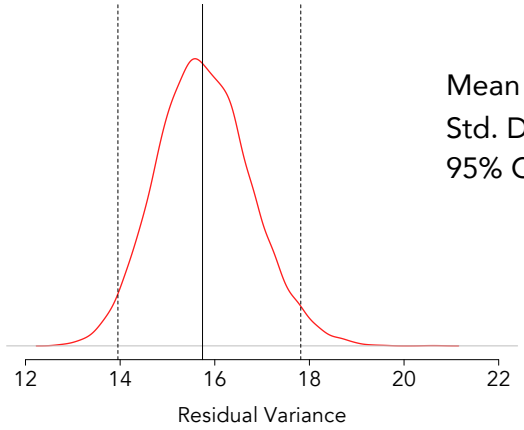
## Posterior Distribution Of The Job Satisfaction Slope Coefficient




Mean = .54
Std. Dev. = .15
95% C.I. = (.24, .84)



## Posterior Distribution Of The Residual Variance




Mean = 15.77
Std. Dev. = .99
95% C.I. = (13.96, 17.81)



## Interpretations

Controlling for job satisfaction, a one-point increase in relationship quality increases employee empowerment by .52, on average

Controlling for relationship quality, a one-point increase in job satisfaction increases employee empowerment by .54, average

Both relations are "significant" because zero is not in the 95% credible interval



## Maximum Likelihood vs. Bayesian

Point estimates and standard errors are numerically similar to posterior means and standard deviations

| | Maximum Likelihood | | Bayesian Estimation | |
|---|---|---|---|---|
| | Estimate | Std. Error | Mean | Std. Dev. |
| Intercept | 21.399 | 0.684 | 21.394 | 0.692 |
| LMX Slope | 0.524 | 0.067 | 0.523 | 0.068 |
| JOBSAT Slope | 0.541 | 0.154 | 0.543 | 0.154 |
| Variance | 15.607 | 0.965 | 15.766 | 0.985 |

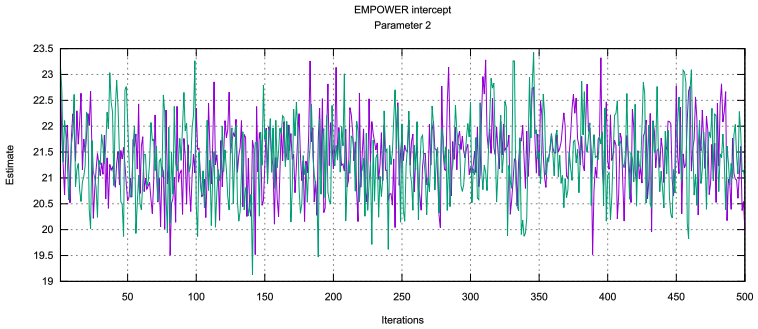
## Blimp Bayesian Analysis Script

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



## Blimp Trace Plots

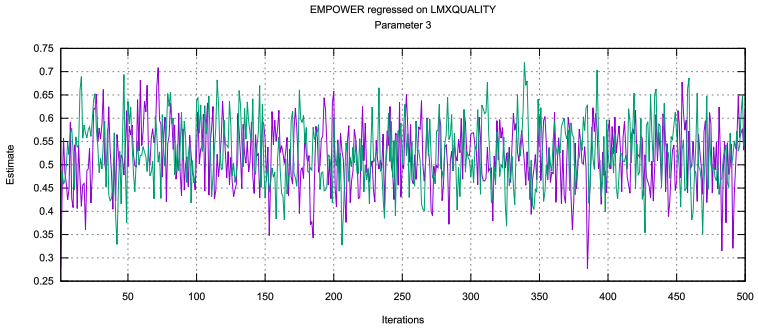
Intercept estimates from 500 iterations






## Blimp Trace Plots

Slope estimates from 500 iterations






## Blimp Output

```
ANALYSIS MODEL ESTIMATES:

Missing outcome: empower

                                 ---------------------------------------------------------
Parameters                       |   Mean   |  Median  |  StdDev  |Lower 2.5 |Upper 97.5|
                                 ---------------------------------------------------------
Variances:                       |          |          |          |          |          |
  Residual Var.                  |   15.766|   15.725|    0.985|   13.964|   17.814|
                                 |----------|----------|----------|----------|----------|
Coefficients:                    |          |          |          |          |          |
  Intercept                      |   21.394|   21.398|    0.692|   20.059|   22.780|
  lmxquality                     |    0.523|    0.523|    0.068|    0.391|    0.655|
  jobsat                         |    0.543|    0.542|    0.154|    0.242|    0.843|
                                 |----------|----------|----------|----------|----------|
Standardized Coefficients:       |          |          |          |          |          |
  lmxquality                     |    0.356|    0.356|    0.042|    0.271|    0.437|
  jobsat                         |    0.155|    0.155|    0.043|    0.069|    0.238|
                                 |----------|----------|----------|----------|----------|
Proportion Variance Explained    |          |          |          |          |          |
  by Fixed Effects               |    0.200|    0.199|    0.032|    0.139|    0.263|
  by Residual Variation          |    0.800|    0.801|    0.032|    0.737|    0.862|
                                 |          |          |          |          |          |
                                 ---------------------------------------------------------
Summaries based on 10000 iterations using 4 chains
```

# Blimp Output (Predictor Distributions)

```
Covariate Models
----------------
Missing covariate:  lmxquality

                          ------------------------------------------------------
Parameters                |   Mean    |  Median   |  StdDev   |Lower 2.5 |Upper 97.5|
                          ------------------------------------------------------
Grand Mean                |     9.622|     9.620|     0.123|     9.386|     9.867|
Level 1:                  |          |          |          |          |          |
  jobsat                  |     1.004|     1.004|     0.090|     0.827|     1.180|
  Residual Var.           |     7.517|     7.506|     0.433|     6.700|     8.395|
                          |          |          |          |          |          |
                          ------------------------------------------------------
Missing covariate:  jobsat

                          ------------------------------------------------------
Parameters                |   Mean    |  Median   |  StdDev   |Lower 2.5 |Upper 97.5|
                          ------------------------------------------------------
Grand Mean                |     3.983|     3.983|     0.052|     3.881|     4.086|
Level 1:                  |          |          |          |          |          |
  lmxquality              |     0.177|     0.177|     0.016|     0.146|     0.208|
  Residual Var.           |     1.324|     1.321|     0.078|     1.181|     1.488|
                          |          |          |          |          |          |
                          ------------------------------------------------------
```



# Mplus Bayesian Analysis Script

```
DATA:
file = lmxquality.dat;
VARIABLE:
names = employee team turnover male empower lmxquality jobsat climate orgsize;
usevariables = empower lmxquality jobsat;
ANALYSIS:
estimator = bayes;
bseed = 90291;
fbiterations = 10000;
MODEL:
lmxquality with jobsat;
empower on lmxquality jobsat;
OUTPUT:
stdyx tech8 patterns;
```



# Mplus Output

MODEL RESULTS

| | | Posterior | One-Tailed | 95% C.I. | |
|---|---|---|---|---|---|
| | Estimate | S.D. | P-Value | Lower 2.5% | Upper 2.5% |
| EMPOWER ON | | | | | |
| LMXQUALITY | 0.526 | 0.067 | 0.000 | 0.391 | 0.655 |
| JOBSAT | 0.539 | 0.154 | 0.000 | 0.238 | 0.845 |
| LMXQUALI WITH | | | | | |
| JOBSAT | 1.624 | 0.175 | 0.000 | 1.300 | 1.989 |
| Means | | | | | |
| LMXQUALITY | 9.621 | 0.125 | 0.000 | 9.378 | 9.867 |
| JOBSAT | 3.982 | 0.052 | 0.000 | 3.879 | 4.083 |
| Intercepts | | | | | |
| EMPOWER | 21.387 | 0.685 | 0.000 | 20.050 | 22.737 |
| Variances | | | | | |
| LMXQUALITY | 9.198 | 0.538 | 0.000 | 8.252 | 10.357 |
| JOBSAT | 1.617 | 0.095 | 0.000 | 1.444 | 1.821 |
| Residual Variances | | | | | |
| EMPOWER | 15.759 | 0.992 | 0.000 | 13.984 | 17.878 |



# Mplus Output

STDYX Standardization

| | | Posterior | One-Tailed | 95% C.I. | |
|---|---|---|---|---|---|
| | Estimate | S.D. | P-Value | Lower 2.5% | Upper 2.5% |
| EMPOWER ON | | | | | |
| LMXQUALITY | 0.359 | 0.043 | 0.000 | 0.272 | 0.439 |
| JOBSAT | 0.155 | 0.044 | 0.000 | 0.068 | 0.239 |
| LMXQUALI WITH | | | | | |
| JOBSAT | 0.422 | 0.034 | 0.000 | 0.352 | 0.486 |
| Residual Variances | | | | | |
| EMPOWER | 0.799 | 0.032 | 0.000 | 0.733 | 0.859 |

...

R-SQUARE

| | | Posterior | One-Tailed | 95% C.I. | |
|---|---|---|---|---|---|
| Variable | Estimate | S.D. | P-Value | Lower 2.5% | Upper 2.5% |
| EMPOWER | 0.201 | 0.032 | 0.000 | 0.141 | 0.267 |