DEALING WITH MISSING DATA

FALL 2015 NEBRASKA METHODOLOGY WORKSHOP

Craig Enders
UCLA Department of Psychology
cenders@psych.ucla.edu

SCHEDULE OF TOPICS

Topics
Missing Data Mechanisms
Traditional Missing Data Handling Methods
Maximum Likelihood Estimation
Maximum Likelihood Analysis Examples in Mplus
Multiple Imputation
Multiple Imputation Analysis Examples in Mplus

MISSING DATA MECHANISMS

PATTERNS VERSUS MECHANISMS

The missing data pattern describes the configuration of observed and the missing values in a data set

The pattern describes the location of the "holes" in the data but says nothing about the reasons for missingness

The mechanism describes how the propensity for missing data is related to other variables, if at all

GENERAL MISSING DATA PATTERNS

A general pattern describes missing values that are dispersed throughout the data matrix

Missingness may or may not be systematic

The methods we focus on can handle general patterns

X1	X2	Y2	Y2
?		?	
?	?		
?			?
		?	?
		?	
?			?
?			
	?		
		?	

MISSING DATA MECHANISMS

Mechanisms describe how the probability of a missing value on Y relates to other variables or to the would-be values of Y itself (Rubin, 1976)

Missing completely at random (MCAR)

Missing at random (MAR)

Not missing at random (NMAR)

MOTIVATING EXAMPLE

20 chronic pain patients enrolled in a pain management program

Respondents fill out pain severity and depression questionnaires

Pain Severity	Depression
4	11
6	19
7	14
7	11
8	6
9	7
9	11
10	12
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

MISSING COMPLETELY AT RANDOM (MCAR)

MCAR = no systematic predictors of missingness

The probability of missing data on a variable *Y* is unrelated to other measured variables and is unrelated to the would-be values of *Y*

The observed scores are a random sample of the hypothetically complete data set

MCAR EXAMPLE

MCAR requires that the probability of a missing depression score is unrelated to pain severity and to the unseen depression values

Nothing predicts missingness

Pain Severity	Depression (Hypothetical)	Depression (Observed)
4	11	?
6	19	19
7	14	?
7	11	11
8	6	6
9	7	7
9	11	11
10	12	?
10	16	16
11	9	9
12	9	9
14	14	14
14	16	16
14	21	21
15	14	14
16	14	?
16	18	18
17	19	19
18	21	21
23	18	?

MISSING AT RANDOM (MAR)

MAR = missingness predicted by observed scores

The probability of missing data on *Y* is related to other measured variables but is unrelated to the would-be values of *Y*

Scores are randomly missing after we control for the observed data

MAR EXAMPLE

Patients with mild pain are more likely to refuse the depression measure

MAR requires that missingness is unrelated to the unseen depression values after controlling for observed severity scores

Pain Severity	Depression (Hypothetical)	Depression (Observed)
4	11	?
6	19	?
7	14	14
7	11	11
8	6	?
9	7	?
9	11	11
10	12	?
10	16	16
11	9	9
12	9	9
14	14	14
14	16	16
14	21	21
15	14	14
16	14	14
16	18	18
17	19	19
18	21	21
23	18	18

NOT MISSING AT RANDOM (NMAR)

NMAR = missingness predicted by unseen scores

The probability of missing data on *Y* is related to *Y* after controlling for other observed variables

Latent (unobserved) values determine missingness

NMAR EXAMPLE

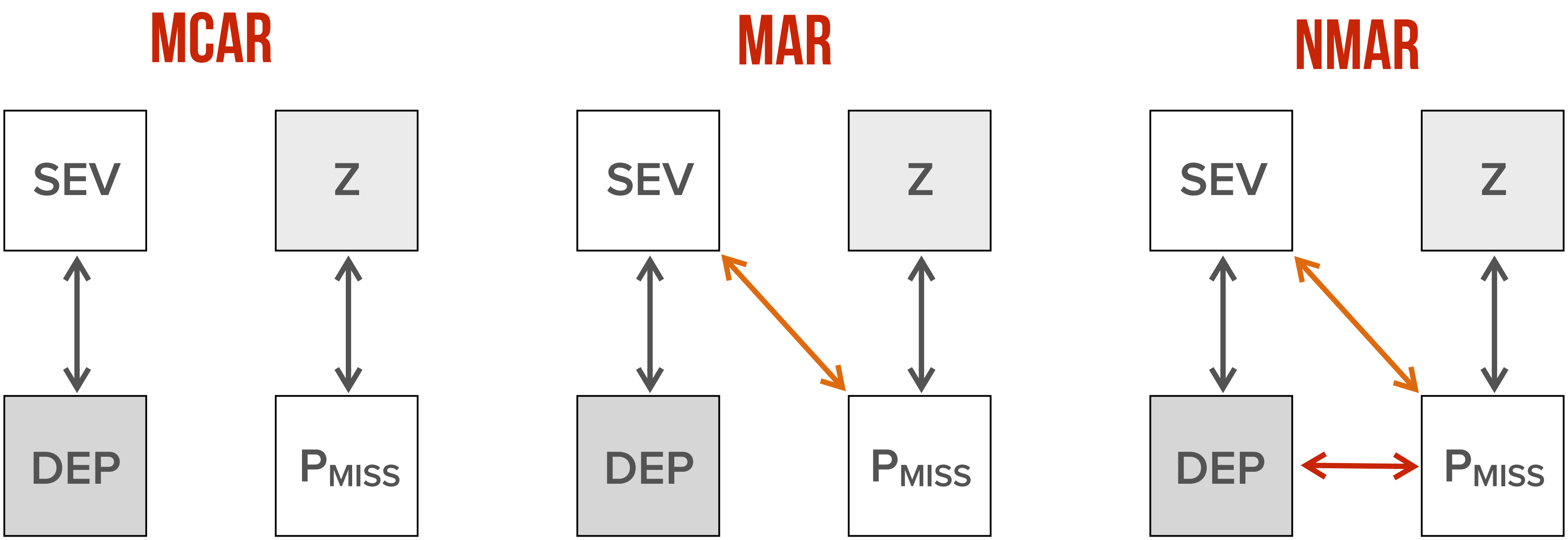
Participants with low depression scores are more likely to skip the depression measure

Unseen depression scores determine missingness, even after accounting for pain severity

Pain Severity	Depression (Hypothetical)	Depression (Observed)
4	11	11
6	19	19
7	14	14
7	11	?
8	6	?
9	7	?
9	11	11
10	12	12
10	16	16
11	9	?
12	9	?
14	14	14
14	16	16
14	21	21
15	14	14
16	14	14
16	18	18
17	19	19
18	21	21
23	18	18

DIAGRAM OF MECHANISMS

P_{MISS} = probability of missing data, Z = variables uncorrelated with SEV and DEP



WHY DO MECHANISMS MATTER?

Mechanisms are analysis assumptions

Deleting incomplete cases require MCAR

Modern approaches assume MAR (or MCAR)

Estimates are biased when assumptions are violated

ANALYSIS EXAMPLE

Complete-data means

$M_{SEV} = 12.00$

$M_{DEP} = 14.00$

Use these values to evaluate analyses under different mechanisms

Pain Severity	Depression
4	11
6	19
7	14
7	11
8	6
9	7
9	11
10	12
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

MCAR EXAMPLE

Complete-case (n = 15)

$M_{\text{SEV}} = 12.00$

$M_{\text{DEP}} = 14.07$

Maximum likelihood (N = 20)

$M_{\text{SEV}} = 12.00$

$M_{\text{DEP}} = 14.07$

Pain Severity	Depression (MCAR)
4	?
6	19
7	?
7	11
8	6
9	7
9	11
10	?
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	?
16	18
17	19
18	21
23	?

MAR EXAMPLE

Complete-case (n = 15)

M_{SEV} = 13.53

M_{DEP} = 15.00

Maximum likelihood (N = 20)

M_{SEV} = 12.00

M_{DEP} = 14.15

Pain Severity	Depression (MAR)
4	?
6	?
7	14
7	11
8	?
9	?
9	11
10	?
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

NMAR EXAMPLE

Complete-case (n = 15)

M_{SEV} = 12.87

M_{DEP} = 15.87

Maximum likelihood (N = 20)

M_{SEV} = 12.00

M_{DEP} = 15.57

Pain Severity	Depression (NMAR)
4	11
6	19
7	14
7	?
8	?
9	?
9	11
10	12
10	16
11	?
12	?
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

THE PROBLEM WITH MAR

MAR-based methods (maximum likelihood) are clearly preferable to methods that assume MCAR (deletion) but NMAR mechanisms still introduce bias

We cannot use the data to test MAR vs. NMAR

Mechanisms make different propositions about the unseen values

TESTING MECHANISMS (NOT REALLY)

Researchers often examine differences between completes and dropouts

Create a missing data indicator for each variable (0 = complete, 1 = missing) and examine mean differences on or correlations with other variables

This strategy can rule out MCAR but says nothing about MAR vs. NMAR mechanisms

MCAR EXAMPLE

M_{Comp} = 12.00, M_{Miss} = 12.00

Absence of differences suggest that severity does not predict missingness

MCAR is supported

Pain Severity	Depression (MCAR)	Missingness Indicator
4	?	1
6	19	0
7	?	1
7	11	0
8	6	0
9	7	0
9	11	0
10	?	1
10	16	0
11	9	0
12	9	0
14	14	0
14	16	0
14	21	0
15	14	0
16	?	1
16	18	0
17	19	0
18	21	0
23	?	1

MAR EXAMPLE

M_{Comp} = 13.53, M_{Miss} = 7.40

Large differences imply systematic missingness (could be MAR or NMAR)

MCAR is not plausible

Pain Severity	Depression (MAR)	Missingness Indicator
4	?	1
6	?	1
7	14	0
7	11	0
8	?	1
9	?	1
9	11	0
10	?	1
10	16	0
11	9	0
12	9	0
14	14	0
14	16	0
14	21	0
15	14	0
16	14	0
16	18	0
17	19	0
18	21	0
23	18	0

NMAR EXAMPLE

$M_{\text{Comp}} = 12.87$, $M_{\text{Miss}} = 9.40$

Large differences imply systematic missingness (could be MAR or NMAR)

MCAR is not plausible

Pain Severity	Depression (NMAR)	Missingness Indicator
4	11	0
6	19	0
7	14	0
7	?	1
8	?	1
9	?	1
9	11	0
10	12	0
10	16	0
11	?	1
12	?	1
14	14	0
14	16	0
14	21	0
15	14	0
16	14	0
16	18	0
17	19	0
18	21	0
23	18	0

PRACTICAL RECOMMENDATIONS

MAR requires logical arguments, cannot be tested

MAR-based methods are usually a good starting point, and including additional auxiliary variables can help satisfy the assumption

NMAR approaches are available but difficult to implement and require other tenuous assumptions

TRADITIONAL MISSING DATA HANDLING METHODS

COMMON APPROACHES

Deletion (listwise and pairwise)

Mean imputation

Regression imputation

Averaging the available items (questionnaire data)

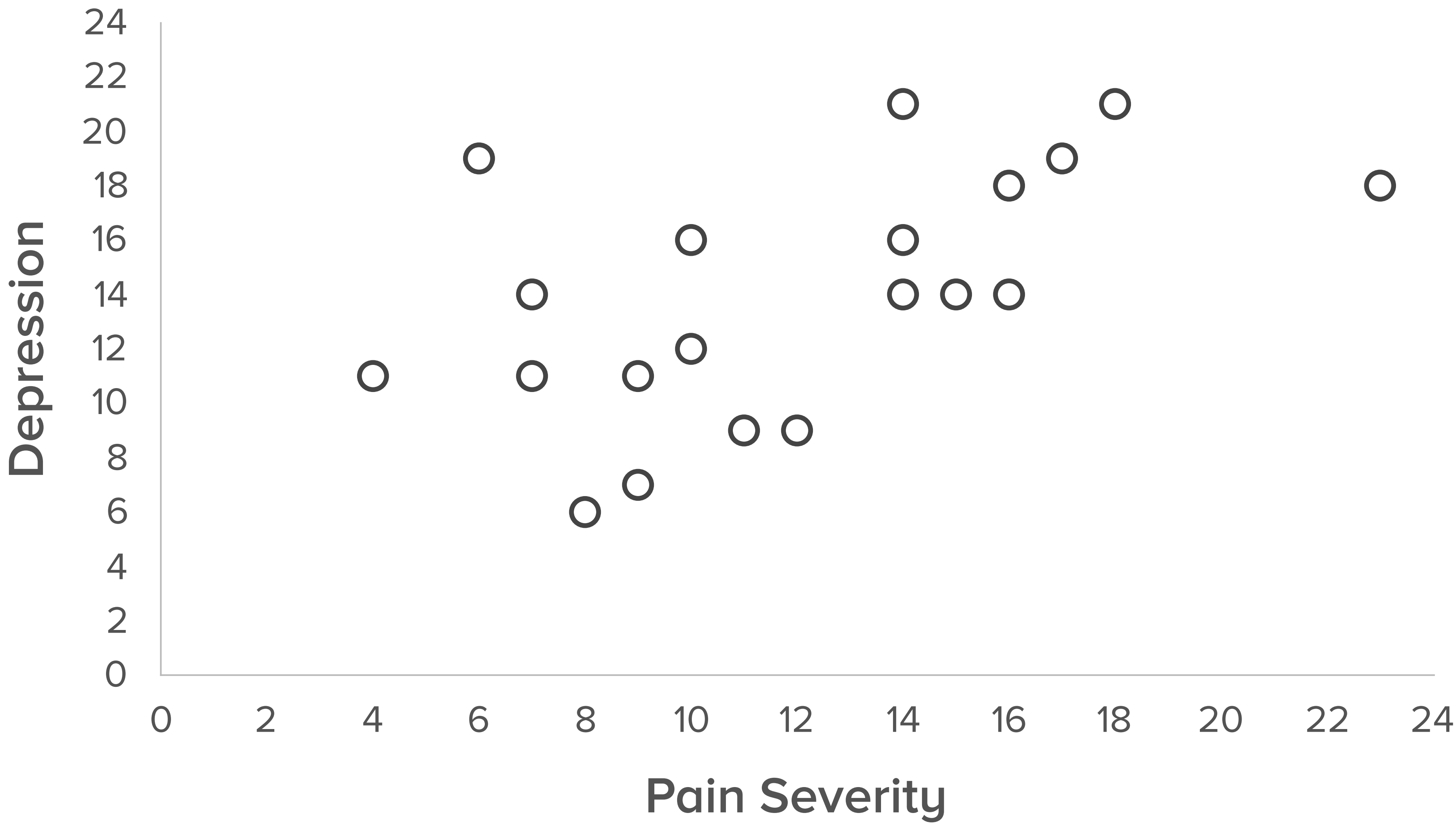
MOTIVATING EXAMPLE

20 chronic pain patients enrolled in a pain management program

Patients with mild pain are more likely to refuse the depression measure

Pain Severity	Depression
4	?
6	?
7	14
7	11
8	?
9	?
9	11
10	?
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

COMPLETE-DATA SCATTERPLOT



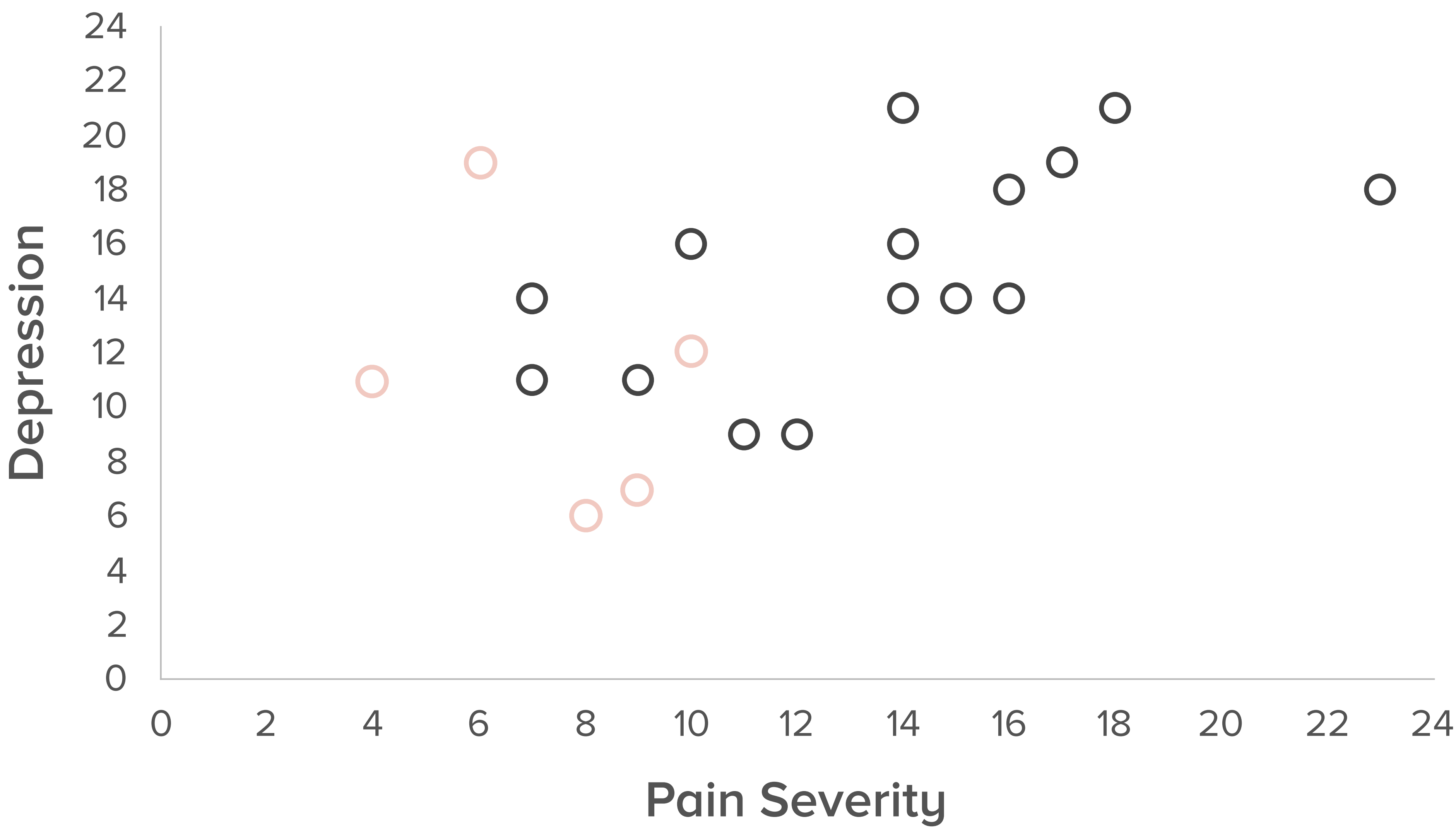
DELETION METHODS

Listwise deletion removes all incomplete data

Pairwise deletion eliminates data on an analysis-by-analysis basis (correlations based on different *N*s)

Discarding data reduces power, and deletion estimates are accurate only with MCAR mechanisms

DELETION SCATTERPLOT



MEAN IMPUTATION

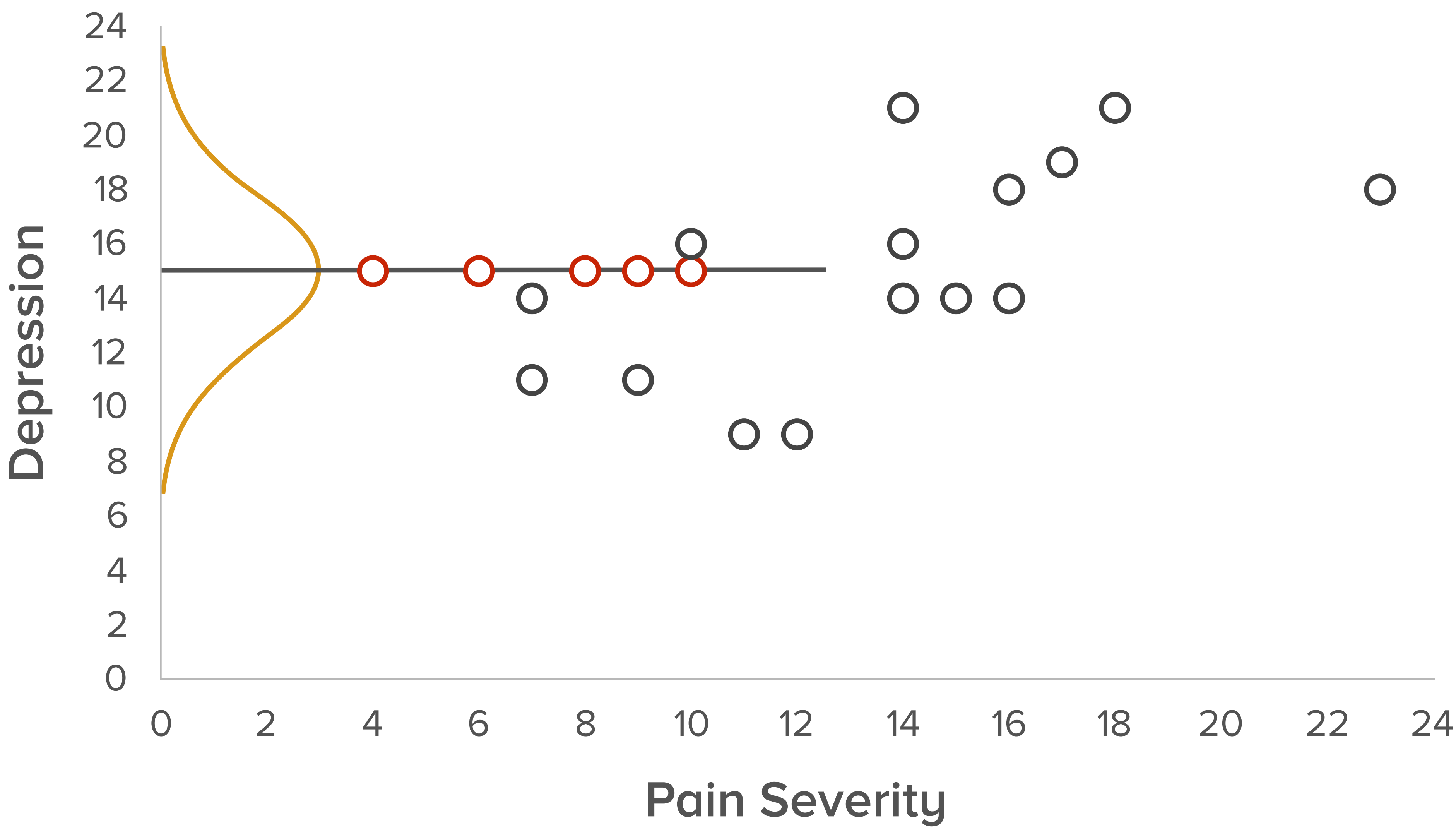
Mean imputation replaces (imputes) missing values with the average of the available scores

Variability and correlations are attenuated because the imputations are constant

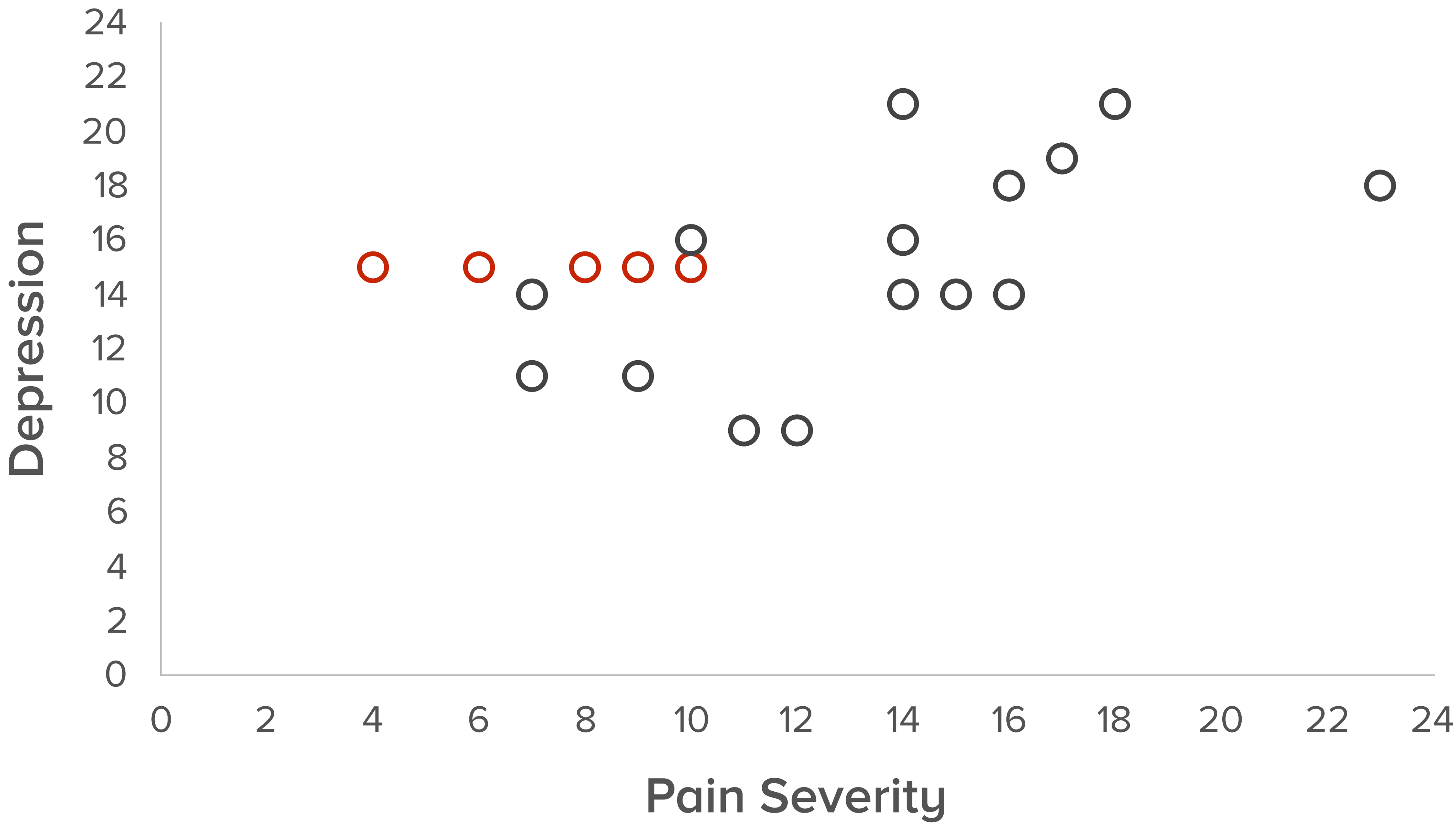
Estimates are biased under any mechanism

Mean imputation is the worst possible option

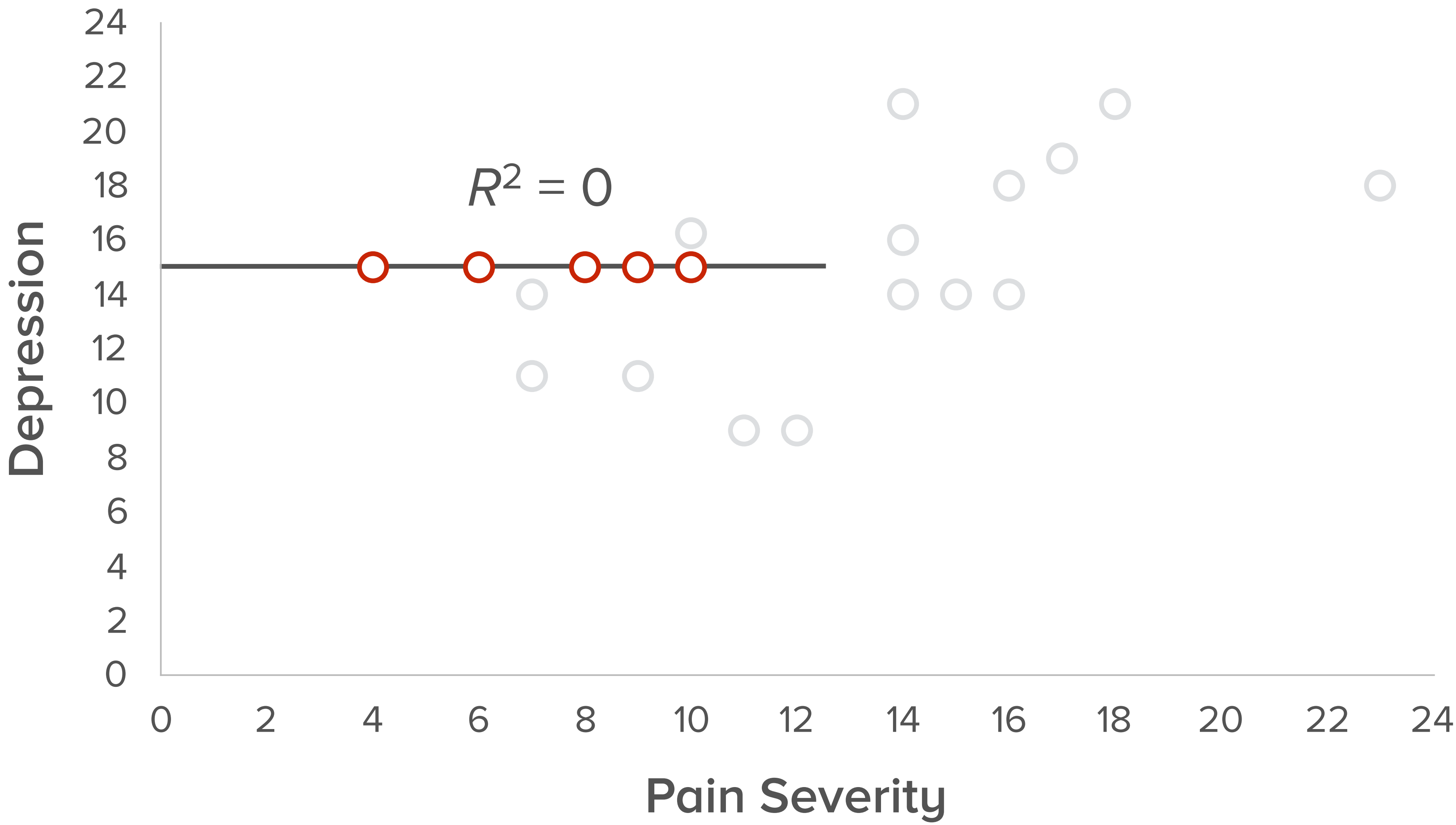
MEAN IMPUTATION



MEAN IMPUTATION SCATTERPLOT



IMPUTED VALUES



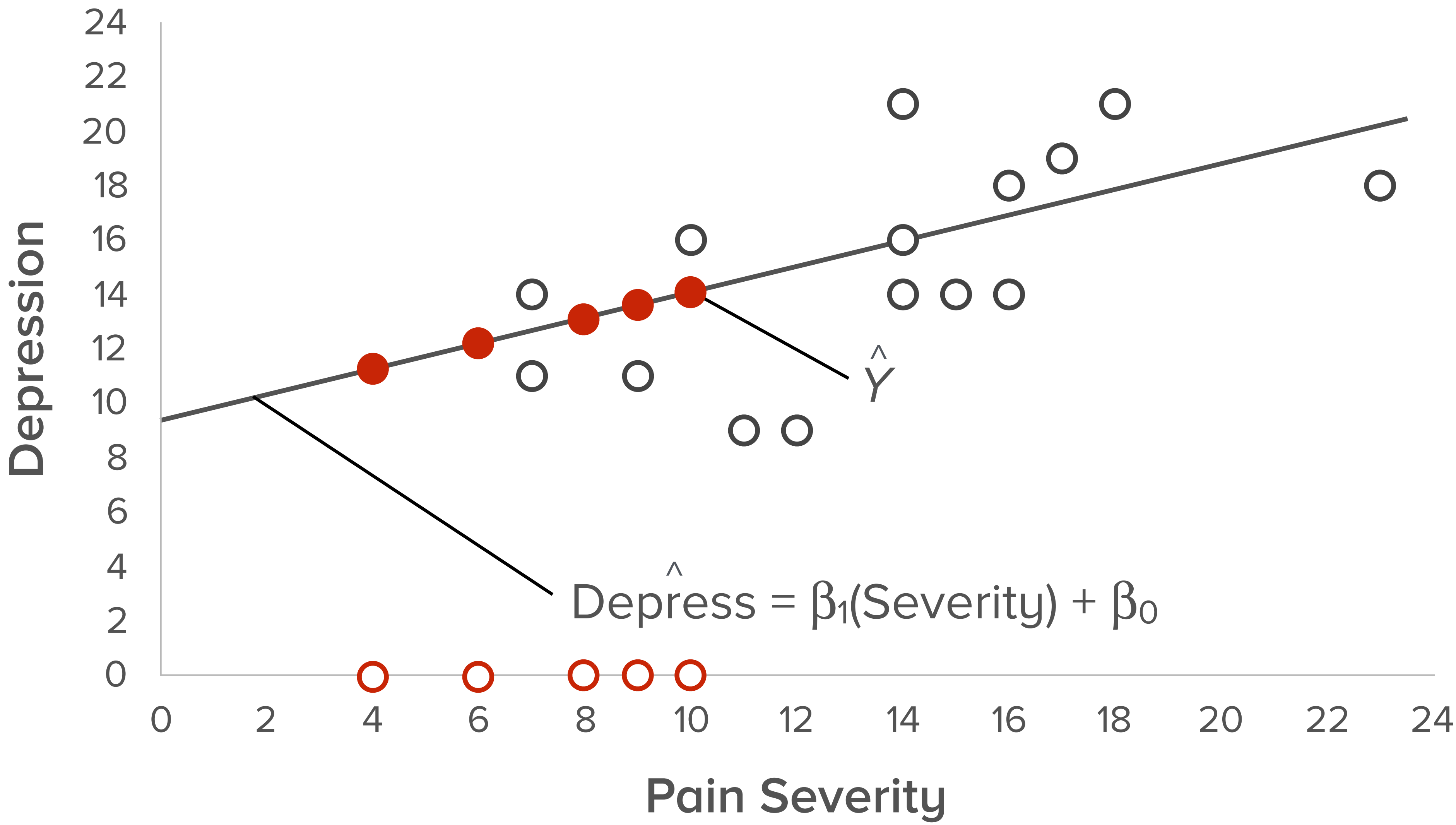
REGRESSION IMPUTATION

Regression imputation replaces missing values with predicted scores from a regression equation where complete variables predict incomplete variables

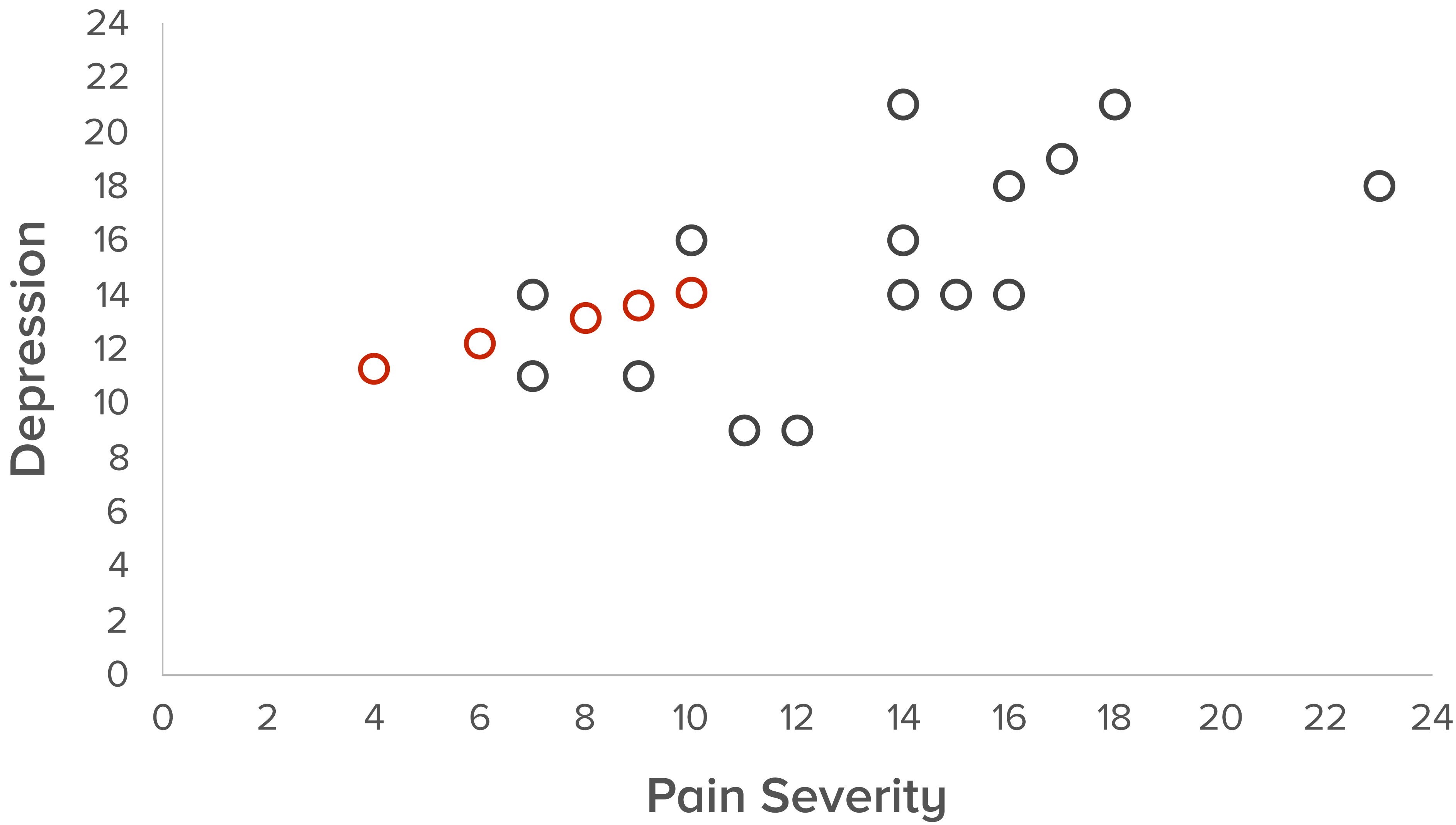
The filled-in data lack variability because the imputed values fall directly on a regression line

Measures of variation and association are biased

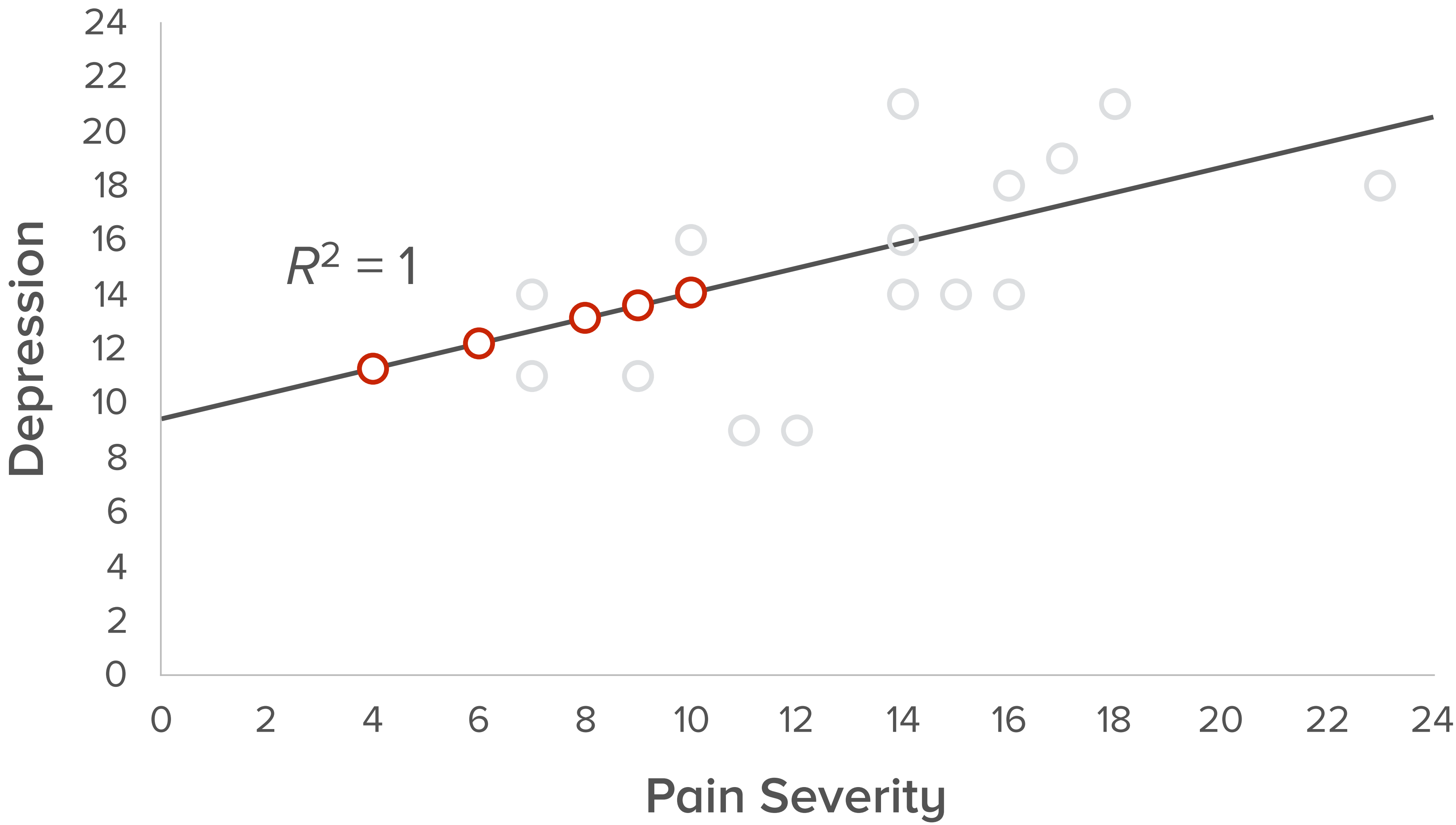
REGRESSION IMPUTATION SCHEME



REGRESSION IMPUTATION SCATTERPLOT



IMPUTED VALUES



Averaging Available Items (Proration)

Many analyses involve scales scores that sum or average a set of questionnaire items

Researchers often compute prorated scale scores scales by averaging the available items

Equivalent to imputing values with a person’s mean

EXAMPLE

PRORATED SCALE SCORE

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	?	4	4.5
3	3	2	4	3.0
4	?	3	?	3.0

PERSON-MEAN IMPUTATION

ID	Q1	Q2	Q3	Scale
1	1	2	1	1.3
2	5	4.5	4	4.5
3	3	2	4	3.0
4	3	3	3	3.0

ISSUES WITH PRORATION

Proration can work well if the mechanism is MCAR and the item means and inter-correlations are equal

Different item distributions introduce severe biases

Requires stricter conditions than deletion

MAXIMUM LIKELIHOOD ESTIMATION FOR MISSING DATA

MOTIVATING EXAMPLE

20 chronic pain patients enrolled in a pain management program

Use maximum likelihood to estimate the depression mean

Pain Severity	Depression
4	11
6	19
7	14
7	11
8	6
9	7
9	11
10	12
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

MAXIMUM LIKELIHOOD (ML) ESTIMATION

ML identifies the population parameter values that are most consistent with the raw data

A likelihood (or log likelihood) function quantifies the fit of the data to the parameters

ML requires a population distribution (normal)

PROBABILITY DENSITY FUNCTION

A density function gives the shape of the normal curve

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e\left[-.5\frac{(Y_i - \mu)^2}{\sigma^2}\right]$$

L_i (the likelihood) gives the relative probability that Y_i came from a normal distribution with a particular mean and variance

SIMPLIFYING THE LIKELIHOOD

The likelihood value is largely driven by a squared *z* score to the right of the exponent

$$L_i = \frac{1}{\sqrt{2\pi\sigma^2}} e\left[-.5\frac{(Y_i - \mu)^2}{\sigma^2}\right]$$

Small *z* score = high likelihood (probability) = close match between the data and the parameters (*Y* and μ)

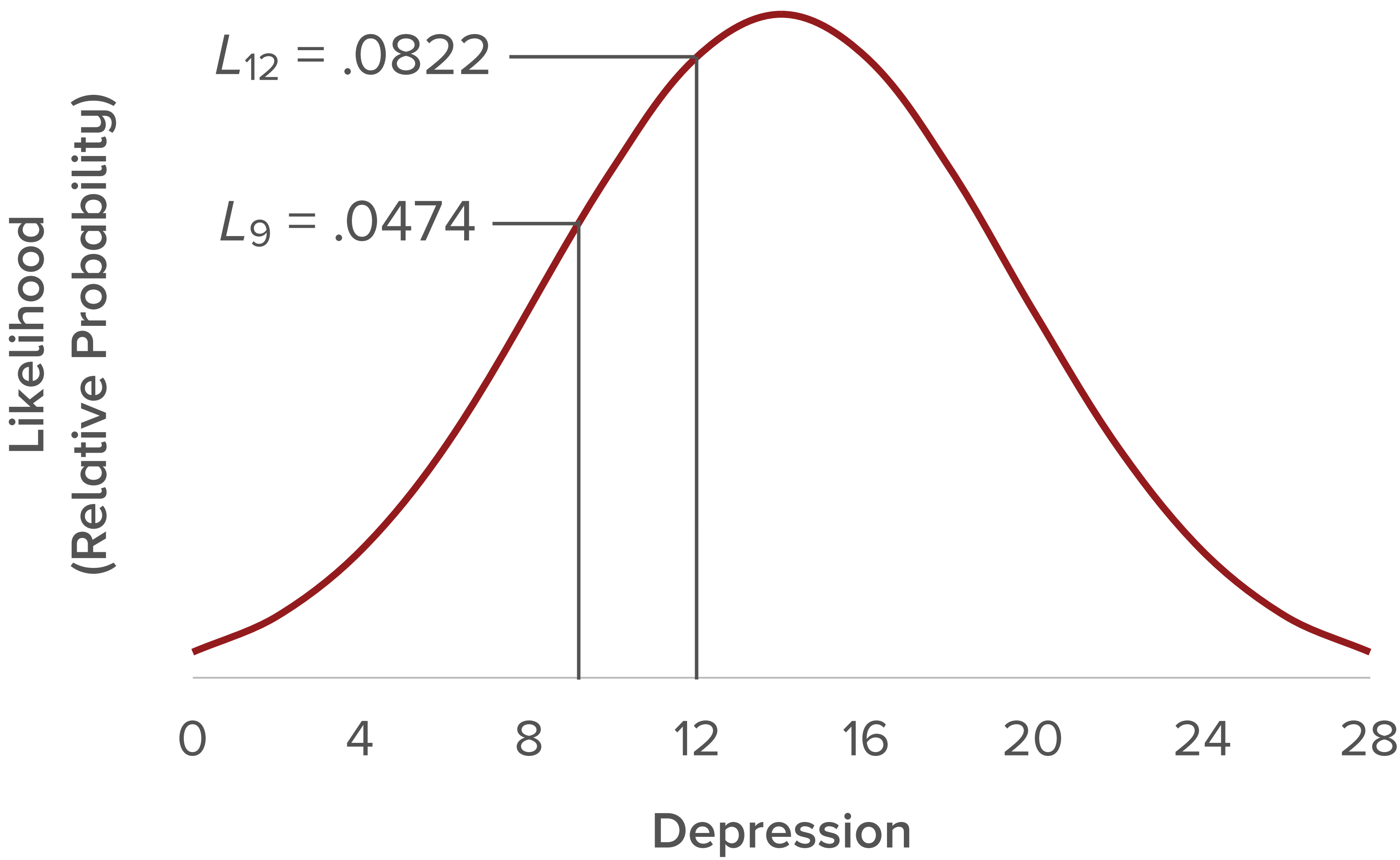
LIKELIHOOD EXAMPLE

Consider depression scores of $Y_i = 12$ and 9 and a normal distribution with $\mu = 14$ and $\sigma = 4.37$

Substituting parameters and scores into the density function gives $L_{12} = .0822$ and $L_9 = .0474$

The L_i values quantify the relative probability of obtaining each score from this normal distribution

GRAPHIC



L_{12} = .0822
L_9 = .0474
Likelihood
(Relative Probability)
0
4
8
12
16
20
24
28
Depression

EXAMPLE

$\mu = 14$ and $\sigma = 4.37$

Smaller deviations between a score and the mean produce higher likelihood values

Higher likelihood values reflect a better fit to the population parameters

Depression	Likelihood
6	0.0171
7	0.0253
9	0.0474
9	0.0474
11	0.0721
11	0.0721
11	0.0721
12	0.0822
14	0.0913
14	0.0913
14	0.0913
14	0.0913
16	0.0822
16	0.0822
18	0.0600
18	0.0600
19	0.0474
19	0.0474
21	0.0253
21	0.0253

Joint Probability

From probability theory, the joint probability for a set of events is the product of individual probabilities

e.g., The probability of jointly observing two heads is (.50)(.50) = .25

Likelihood values are not probabilities, but the same rules apply

SAMPLE LIKELIHOOD

The sample likelihood is the product of the individual likelihoods

$$L = \prod \frac{1}{\sqrt{2\pi\sigma^2}} e\left[-.5\frac{(y_i - \mu)^2}{\sigma^2}\right]$$

$\prod$ is the multiplication operator

EXAMPLE

Multiplying likelihoods gives the sample likelihood

$$L = (.0171)(.0253) \ldots (.0253)(.0253)$$

$$= .00000000000000000000000000007327$$

The sample likelihood quantifies the relative probability of obtaining these 20 scores from a normal population with $\mu = 14$ and $\sigma = 4.37$

LOGARITHMS

Likelihoods are computationally difficult and introduce precision problems due to rounding error

One rule of logarithms is $\log[(a)(b)] = \log(a) + \log(b)$

Using logarithms converts a multiplication problem to an addition problem (simpler math)

LOG LIKELIHOOD VALUES

$\log L_i$ is the natural logarithm of a single likelihood

$$\log L_i = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e\left[-.5\frac{(y_i-\mu)^2}{\sigma^2}\right]\right)$$

The $\log L_i$ values also quantify relative probability, but they do so on a different metric

EXAMPLE

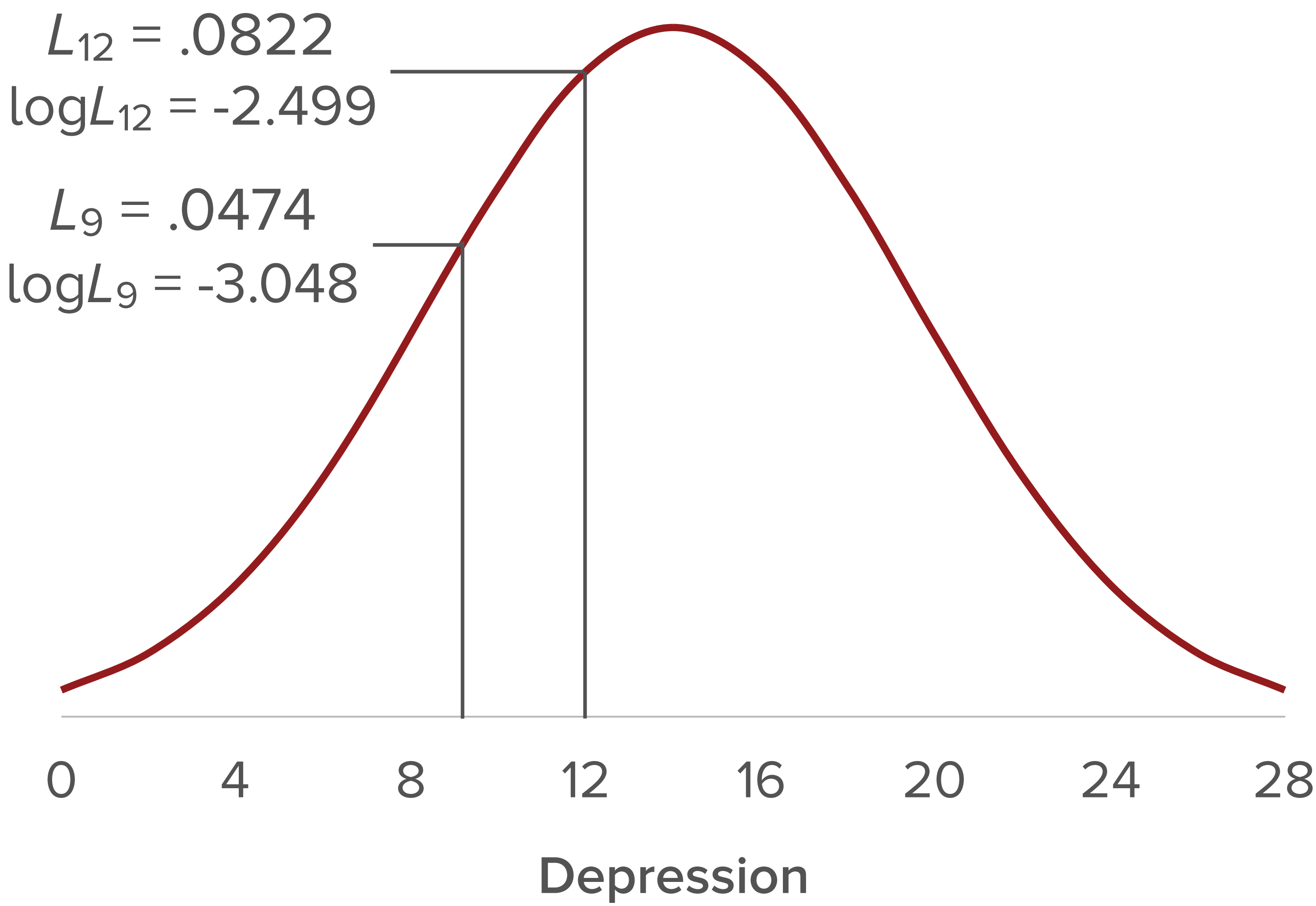
$\mu = 14$ and $\sigma = 4.37$

Smaller deviations between a score and the mean produce higher log likelihood values

Higher log likelihood values reflect a better fit to the parameters

Depression	Likelihood	log*L*
6	0.0171	-4.0692
7	0.0253	-3.6765
9	0.0474	-3.0482
9	0.0474	-3.0482
11	0.0721	-2.6294
11	0.0721	-2.6294
11	0.0721	-2.6294
12	0.0822	-2.4985
14	0.0913	-2.3938
14	0.0913	-2.3938
14	0.0913	-2.3938
14	0.0913	-2.3938
16	0.0822	-2.4985
16	0.0822	-2.4985
18	0.0600	-2.8126
18	0.0600	-2.8126
19	0.0474	-3.0482
19	0.0474	-3.0482
21	0.0253	-3.6765
21	0.0253	-3.6765

GRAPHIC



L_{12} = .0822
logL_{12} = -2.499
L_9 = .0474
logL_9 = -3.048
0
4
8
12
16
20
24
28
Depression

SAMPLE LOG LIKELIHOOD

The sample log likelihood is the sum of the individual log likelihoods

$$\log L = \sum \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e\left[-.5\frac{(y_i - \mu)^2}{\sigma^2}\right]\right)$$

The equation says to (a) compute the likelihood for each case, (b) take the natural log of each value, and (c) sum the individual log likelihoods

EXAMPLE

Adding $\log L_i$ values gives the sample log likelihood

$$\log L = (-4.0692) + (-3.6765) + \ldots + (-3.6765) = -57.8757$$

The sample log likelihood quantifies the relative probability of obtaining these 20 scores from a normal population with $\mu = 14$ and $\sigma = 4.37$

Interpreting the Log Likelihood

The log likelihood quantifies the fit between the sample data and the population parameters

No absolute criterion for a good or a bad value

The log*L* depends on the sample size, number of variables, number of parameters in the model, missing data, etc.

ESTIMATION STRATEGY

The sample log likelihood provides a mechanism for identifying unknown parameter values

Compute the log likelihood for different values of μ

Identify the value of μ that produces the highest log likelihood (highest probability, best fit to the data)

POPULATION $\mu = 12$

$\log L = (-3.336) + (-3.048) + \ldots + (-4.514) + (-4.514) = -59.967$

Depression	logL
6	-3.336
7	-3.048
9	-2.629
9	-2.629
11	-2.420
11	-2.420
11	-2.420
12	-2.394
14	-2.498
14	-2.498

Depression	logL
14	-2.498
14	-2.498
16	-2.813
16	-2.813
18	-3.336
18	-3.336
19	-3.677
19	-3.677
21	-4.514
21	-4.514

POPULATION μ = 13

$\log L = (-3.677) + (-3.336) + \ldots + (-4.069) + (-4.069) = -58.399$

Depression	logL
6	-3.677
7	-3.336
9	-2.813
9	-2.813
11	-2.498
11	-2.498
11	-2.498
12	-2.420
14	-2.420
14	-2.420

Depression	logL
14	-2.420
14	-2.420
16	-2.629
16	-2.629
18	-3.048
18	-3.048
19	-3.336
19	-3.336
21	-4.069
21	-4.069

POPULATION μ = 14

$\log L = (-4.069) + (-3.677) + \ldots + (-3.677) + (-3.677) = -57.876$

Depression	logL
6	-4.069
7	-3.677
9	-3.048
9	-3.048
11	-2.629
11	-2.629
11	-2.629
12	-2.499
14	-2.394
14	-2.394

Depression	logL
14	-2.394
14	-2.394
16	-2.499
16	-2.499
18	-2.813
18	-2.813
19	-3.048
19	-3.048
21	-3.677
21	-3.677

POPULATION $\mu = 15$

$$\log L = (-4.514) + (-4.069) + \ldots + (-3.336) + (-3.336) = -58.399$$

Depression	logL
6	-4.514
7	-4.069
9	-3.336
9	-3.336
11	-2.813
11	-2.813
11	-2.813
12	-2.629
14	-2.420
14	-2.420

Depression	logL
14	-2.420
14	-2.420
16	-2.420
16	-2.420
18	-2.629
18	-2.629
19	-2.813
19	-2.813
21	-3.336
21	-3.336

ESTIMATION SUMMARY

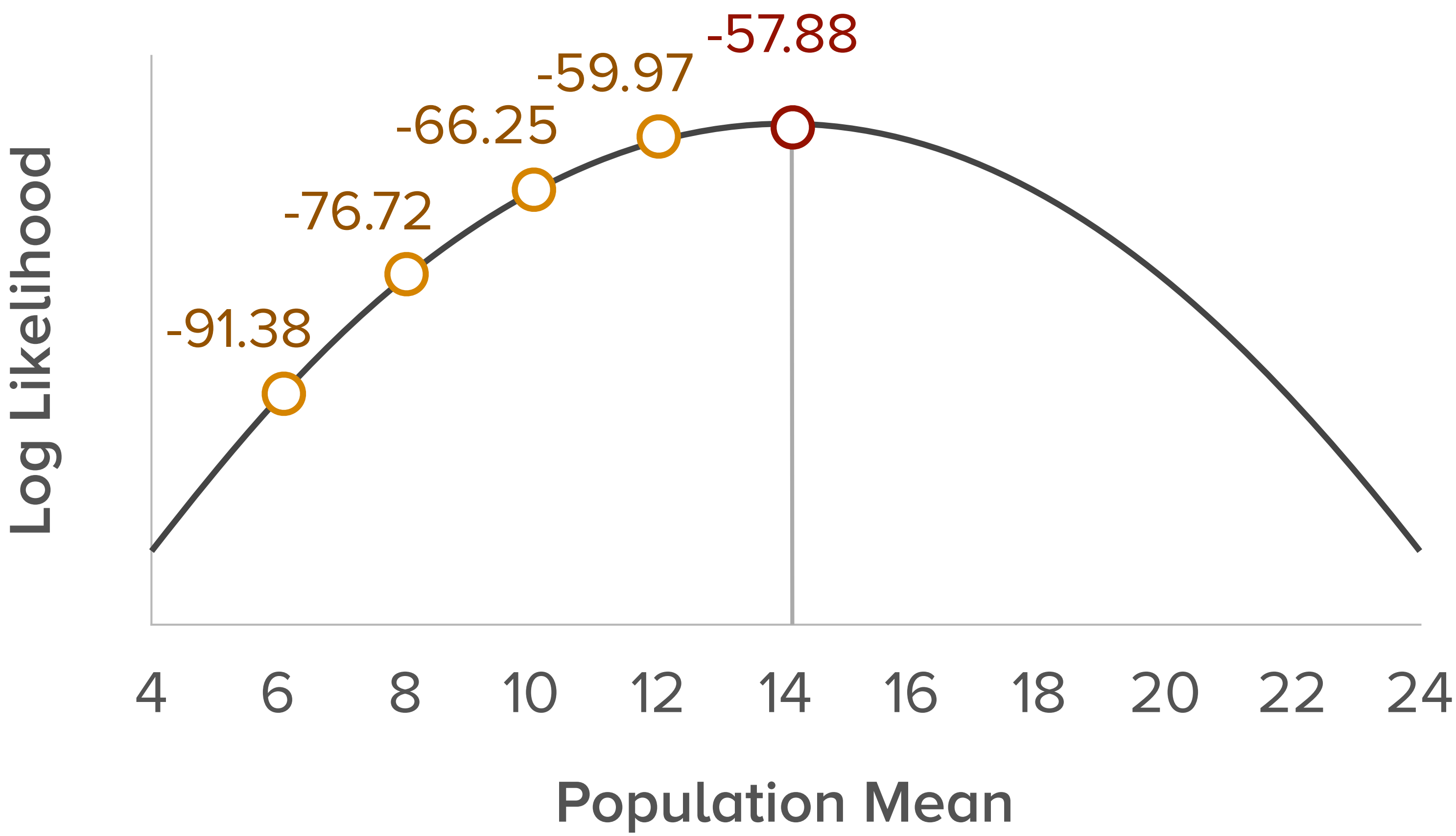
μ = 14 maximizes the probability of sampling these 20 cases

μ = 14 is the maximum likelihood estimate

Population Mean	log*L*
12	-59.967
13	-58.399
14	-57.876
15	-58.399

LOG LIKELIHOOD FUNCTION

The log likelihood function describes how the sample log likelihood changes between μ values of 4 and 24



MULTIVARIATE NORMAL DISTRIBUTION

Multivariate normal distribution

$$L_i = \frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{.5}} e\left[-.5(\mathbf{Y}_i - \boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})\right]$$

L_i is the relative probability of a set of *Y* values, given the parameter estimates in μ and Σ

SIMPLIFYING THE LIKELIHOOD

The multivariate likelihood value is still driven by a squared *z* score to the right of the exponent

$$L_i = \frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{.5}} e\left[-.5(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})\right]$$

Small *z* score = high likelihood (probability) = close match between the data and the parameters (*Y* and μ)

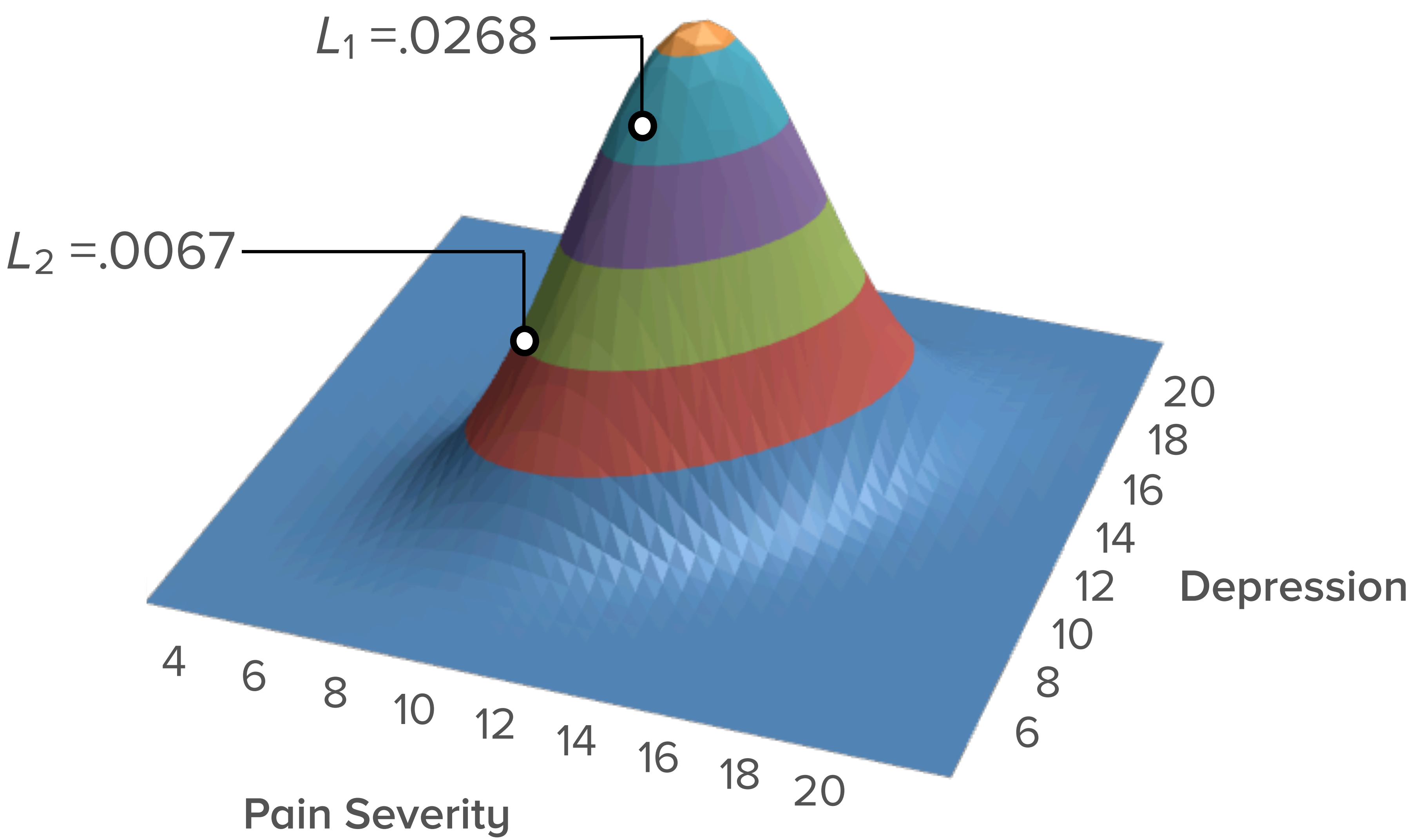
LIKELIHOOD EXAMPLE

Two pairs of depression and severity scores and parameters fixed at their sample values

$$\mathbf{Y}_1 = \begin{bmatrix} 10 \\ 12 \end{bmatrix} \quad \mathbf{Y}_2 = \begin{bmatrix} 7 \\ 11 \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} 12 \\ 14 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 21.4 & 19.5 \\ 19.5 & 19.1 \end{bmatrix}$$

Substituting parameters and scores into the density function gives $L_1 = .0268$ and $L_2 = .0067$

GRAPHIC



L_1 =.0268
L_2 =.0067
4
6
8
10
12
14
16
18
20
Pain Severity
20
18
16
14
12
10
8
6
Depression

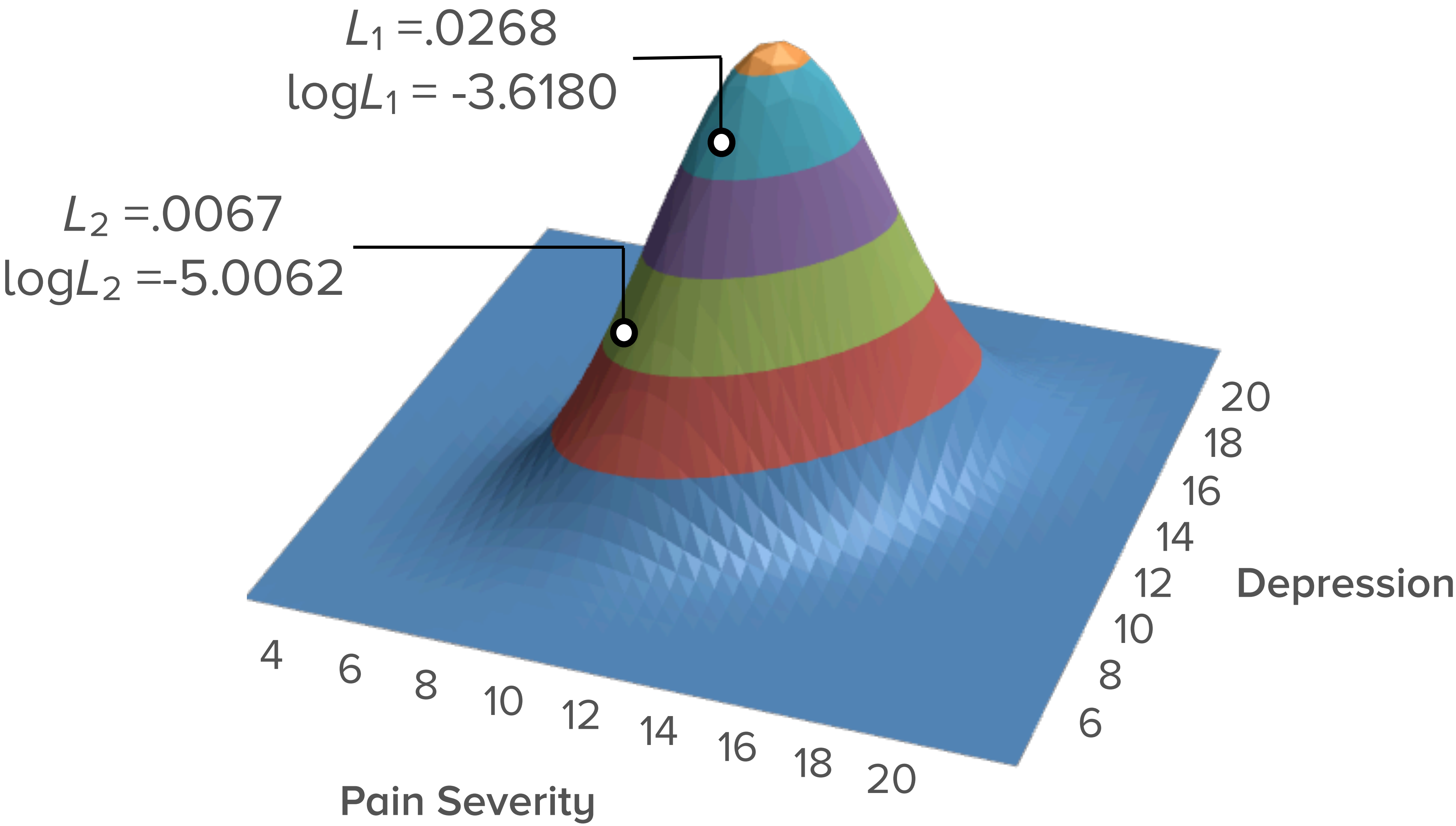
MULTIVARIATE NORMAL LOG LIKELIHOOD

The log likelihood quantifies relative probability, but on a different metric (same as before)

$$\log L_i = \log\left(\frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{.5}} e\left[-.5(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})\right]\right)$$

$$= -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\left|\Sigma\right| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu})$$

GRAPHIC



L_1 =.0268
logL_1 = -3.6180
L_2 =.0067
logL_2 =-5.0062
20
18
16
14
12
10
8
6
Depression
4
6
8
10
12
14
16
18
20
Pain Severity

MISSING DATA LOG LIKELIHOOD

Complete-data log likelihood

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log\left|\Sigma\right| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})$$

Missing-data log likelihood

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\left|\Sigma_i\right| - \frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

WHAT IS DIFFERENT?

The missing data log likelihood has an *i* (individual) subscript on μ and Σ

The subscript indicates that the number of parameters in the matrices depends on the missing data pattern

The squared *z* score is computed using all available data and the parameters for which a case has data

PAIN DATA

20 chronic pain patients enrolled in a pain management program

Patients with mild pain are more likely to refuse the depression measure

Pain Severity	Depression
4	?
6	?
7	14
7	11
8	?
9	?
9	11
10	?
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

COMPLETE-DATA CALCULATIONS

The squared *z* score for the 15 complete cases uses the entire collection of parameters

$$z^2 = (\mathbf{Y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

$$= \left(\begin{bmatrix} Y_{\text{Sev}} \\ Y_{\text{Dep}} \end{bmatrix} - \begin{bmatrix} \mu_{\text{Sev}} \\ \mu_{\text{Dep}} \end{bmatrix} \right)^{\mathrm{T}} \begin{pmatrix} \sigma^2_{\text{Sev}} & \sigma_{\text{Sev,Dep}} \\ \sigma_{\text{Dep,Sev}} & \sigma^2_{\text{Dep}} \end{pmatrix}^{-1} \left(\begin{bmatrix} Y_{\text{Sev}} \\ Y_{\text{Dep}} \end{bmatrix} - \begin{bmatrix} \mu_{\text{Sev}} \\ \mu_{\text{Dep}} \end{bmatrix} \right)$$

EXAMPLE

Squared z score computation for the case with severity and depression scores of 7 and 11

$$z^2 = (\mathbf{Y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

$$= \left(\begin{bmatrix} 7 \\ 11 \end{bmatrix} - \begin{bmatrix} 12 \\ 14 \end{bmatrix} \right)^{\mathrm{T}} \begin{pmatrix} 21.4 & 19.5 \\ 19.5 & 19.1 \end{pmatrix}^{-1} \left(\begin{bmatrix} 7 \\ 11 \end{bmatrix} - \begin{bmatrix} 12 \\ 14 \end{bmatrix} \right) = 2.987$$

MISSING-DATA CALCULATIONS

The squared z score for the 5 incomplete cases uses only the severity parameters

$$z^2 = (\mathbf{Y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

$$= \frac{(Y_{\mathrm{Sev}} - \mu_{\mathrm{Sev}})^2}{\sigma^2_{\mathrm{Sev}}}$$

EXAMPLE

Squared z score computation for the case with a severity score of 10 and a missing depression score

$$z^2 = \frac{(Y_i - \mu_i)^2}{\sigma_i^2} = \frac{(10 - 12)^2}{21.4} = .187$$

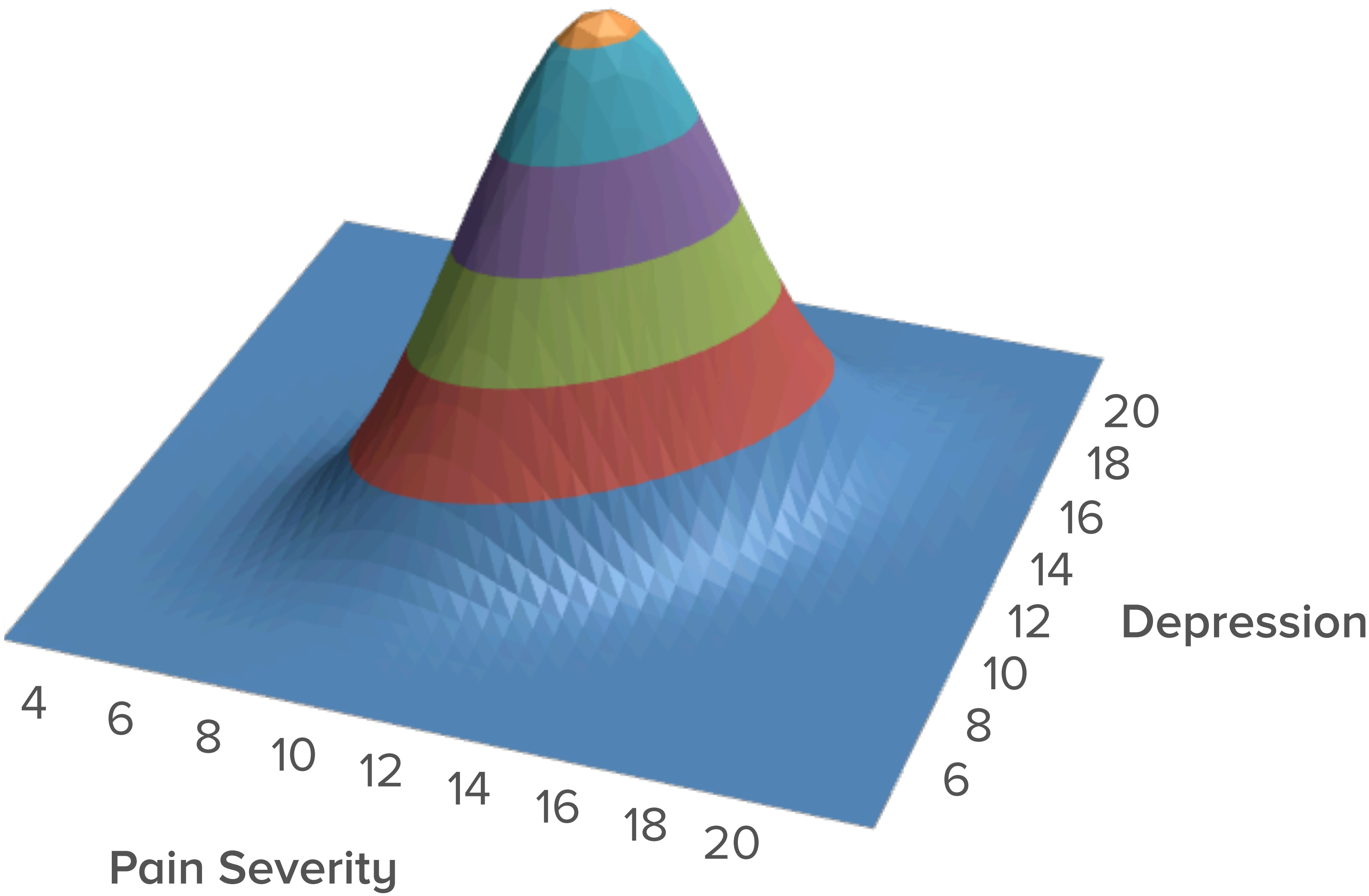
HOW DOES THIS HELP?

Maximum likelihood uses all available data to estimate parameters

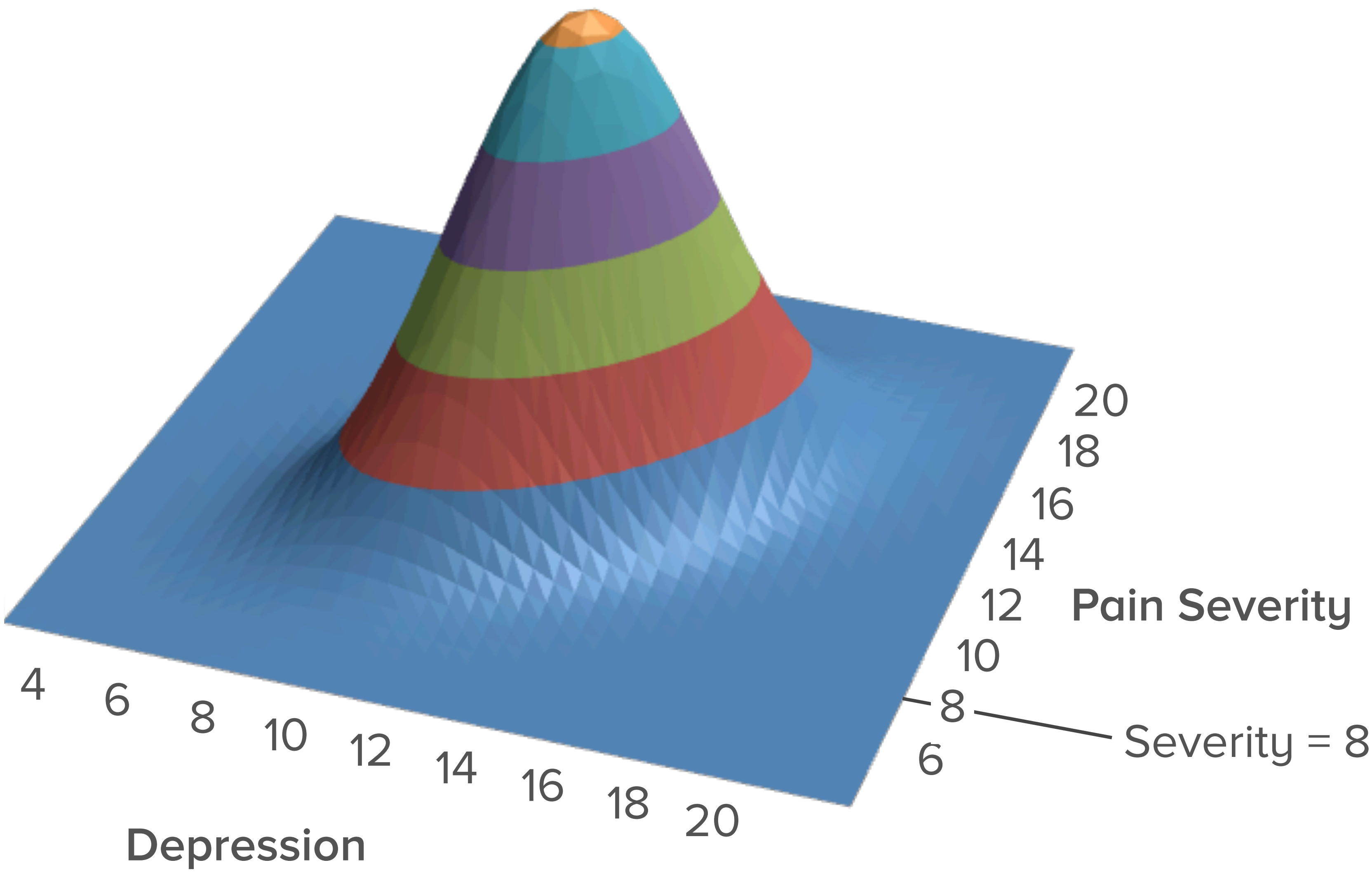
The procedure can be viewed as implicit imputation because the observed data imply plausible values for the missing scores

The normal distribution is key because it defines a range of plausible scores for the missing data

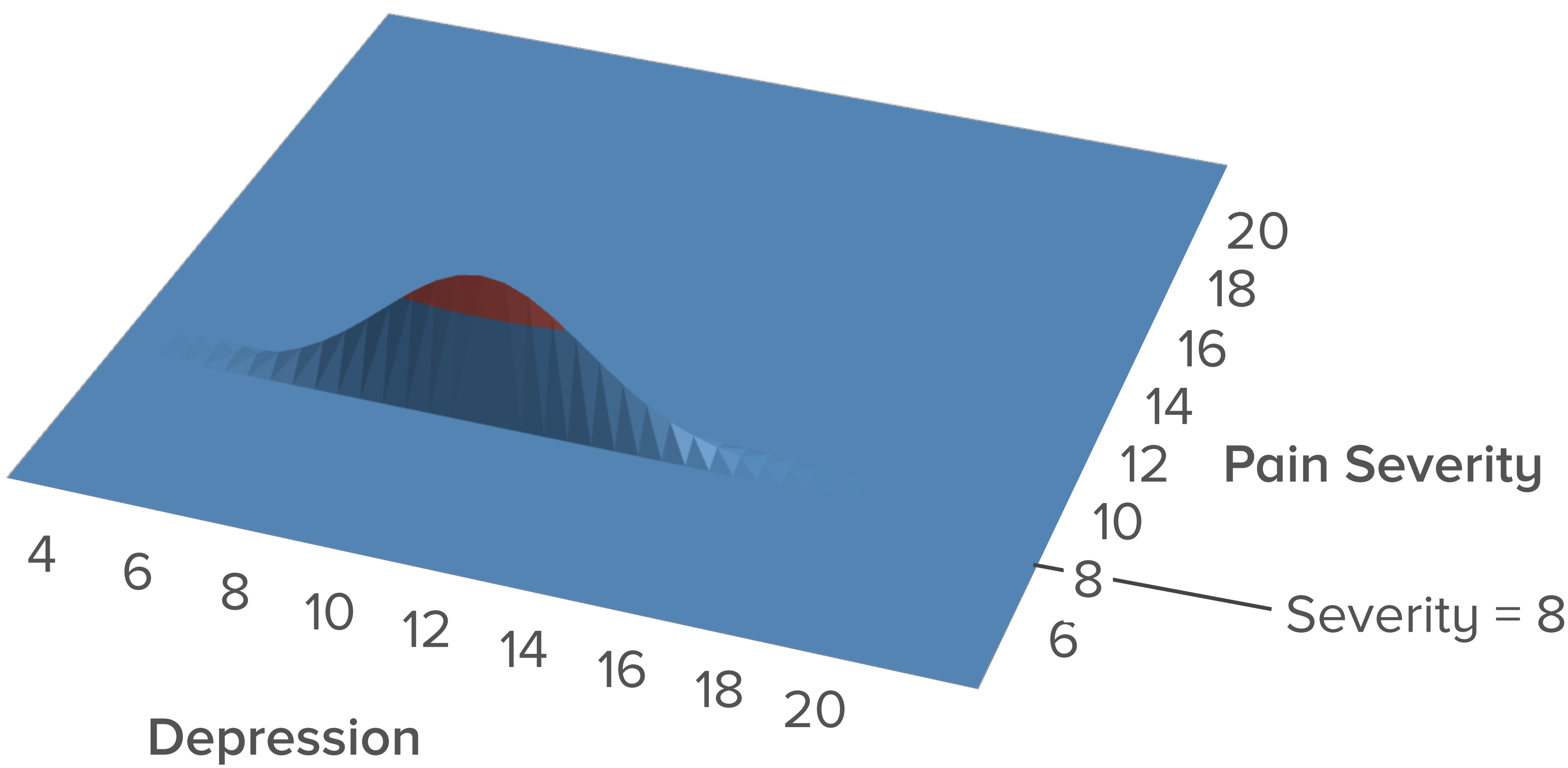
NORMAL DISTRIBUTION



SEVERITY = 8, DEPRESSION = ?

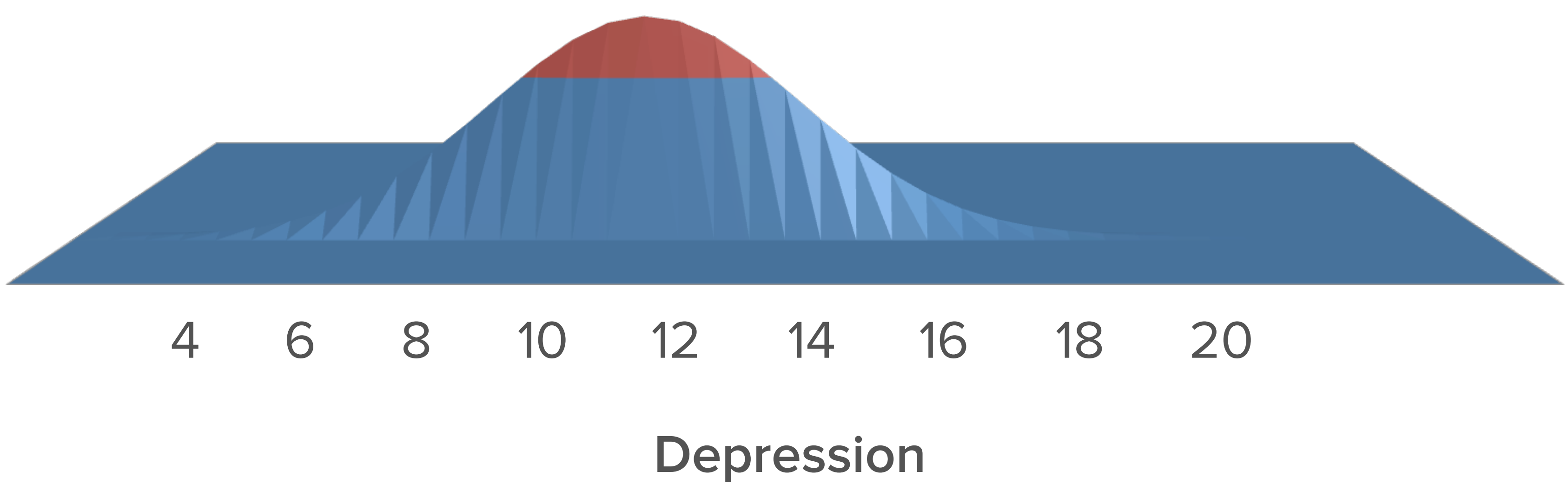


NORMAL CURVE SLICE SEVERITY = 8



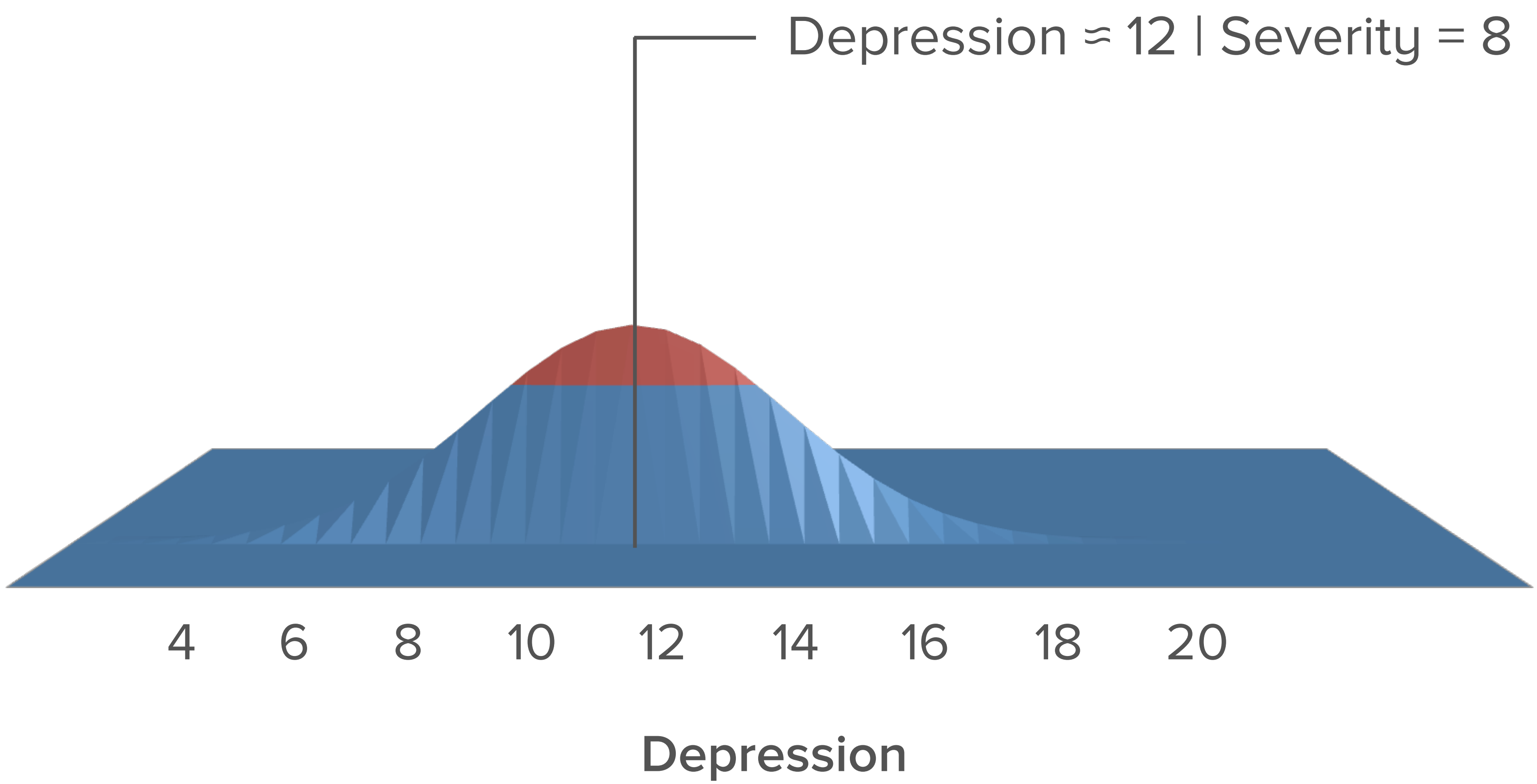
CONDITIONAL DISTRIBUTION

Distribution of plausible depression scores for a case with a severity score of 8



IMPLICIT IMPUTATION

Given a severity score of 8, the most likely value of the missing depression scale is ≈ 12



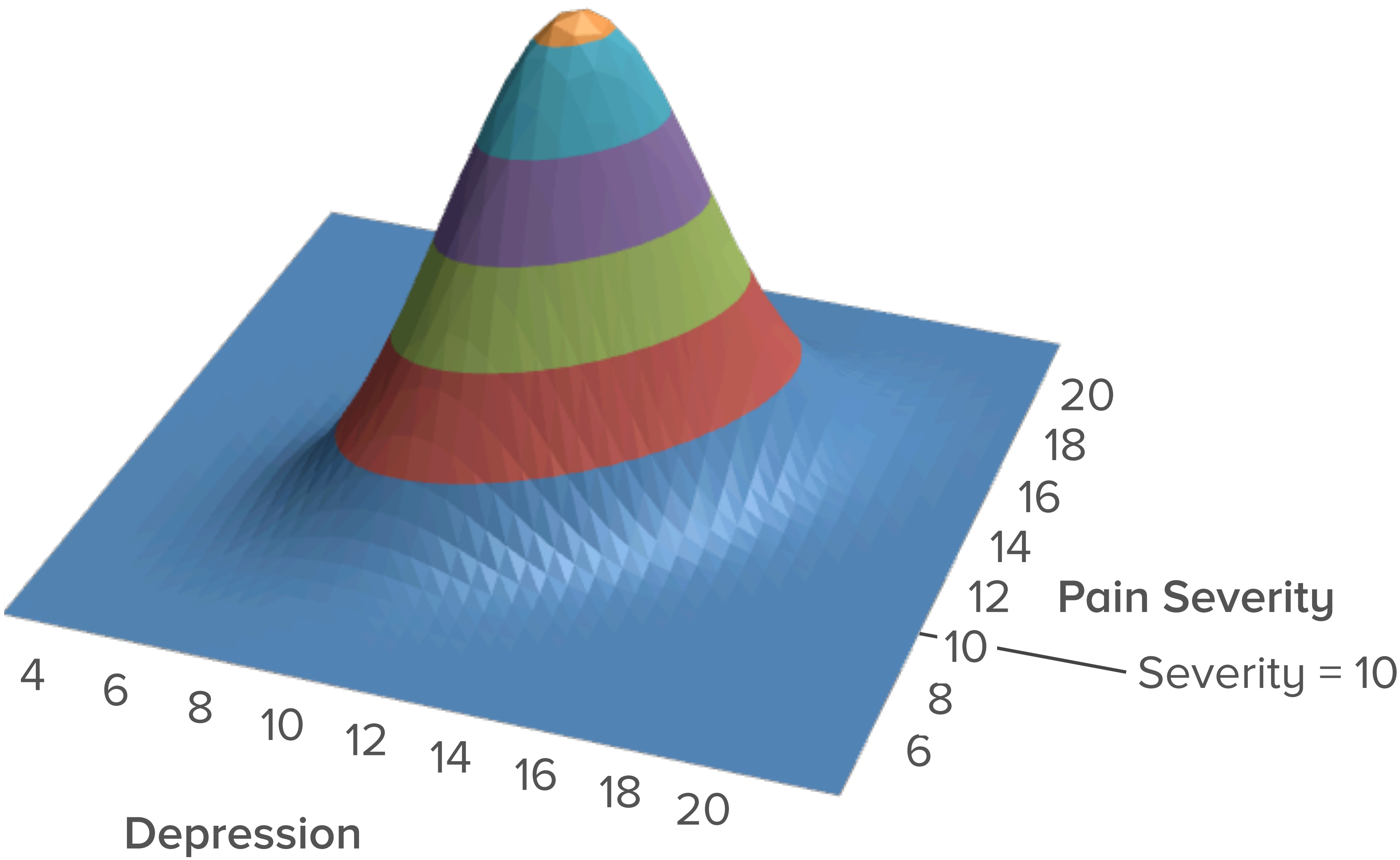
WHAT HAPPENS TO THE MEAN?

The complete cases produced an depression average of 15 (too high relative to the true estimate)

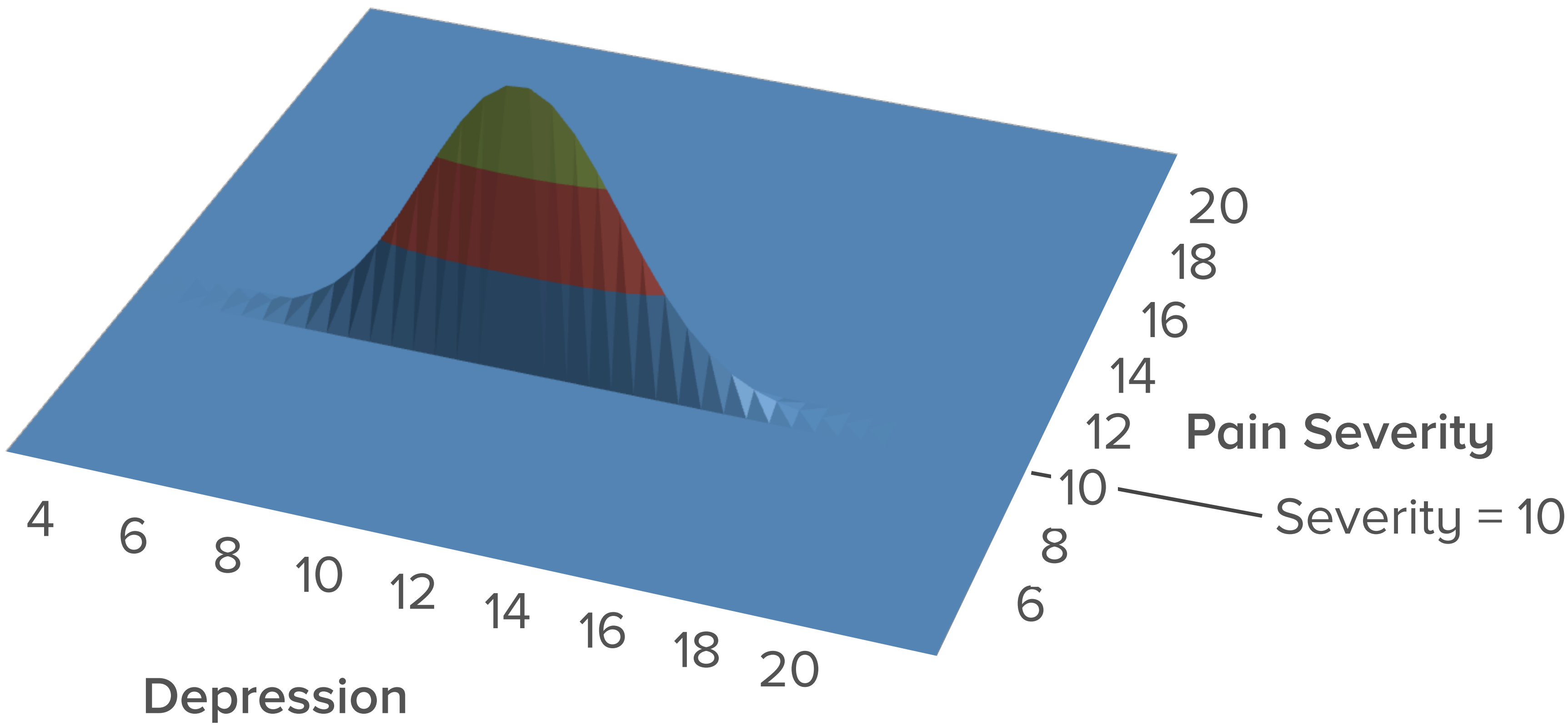
A case with an severity = 8 should have a parenting score of roughly 12

Adjusting the mean downward to account for the plausible (but missing) value brings the estimate closer to its true value of 14

SEVERITY = 10, DEPRESSION = ?

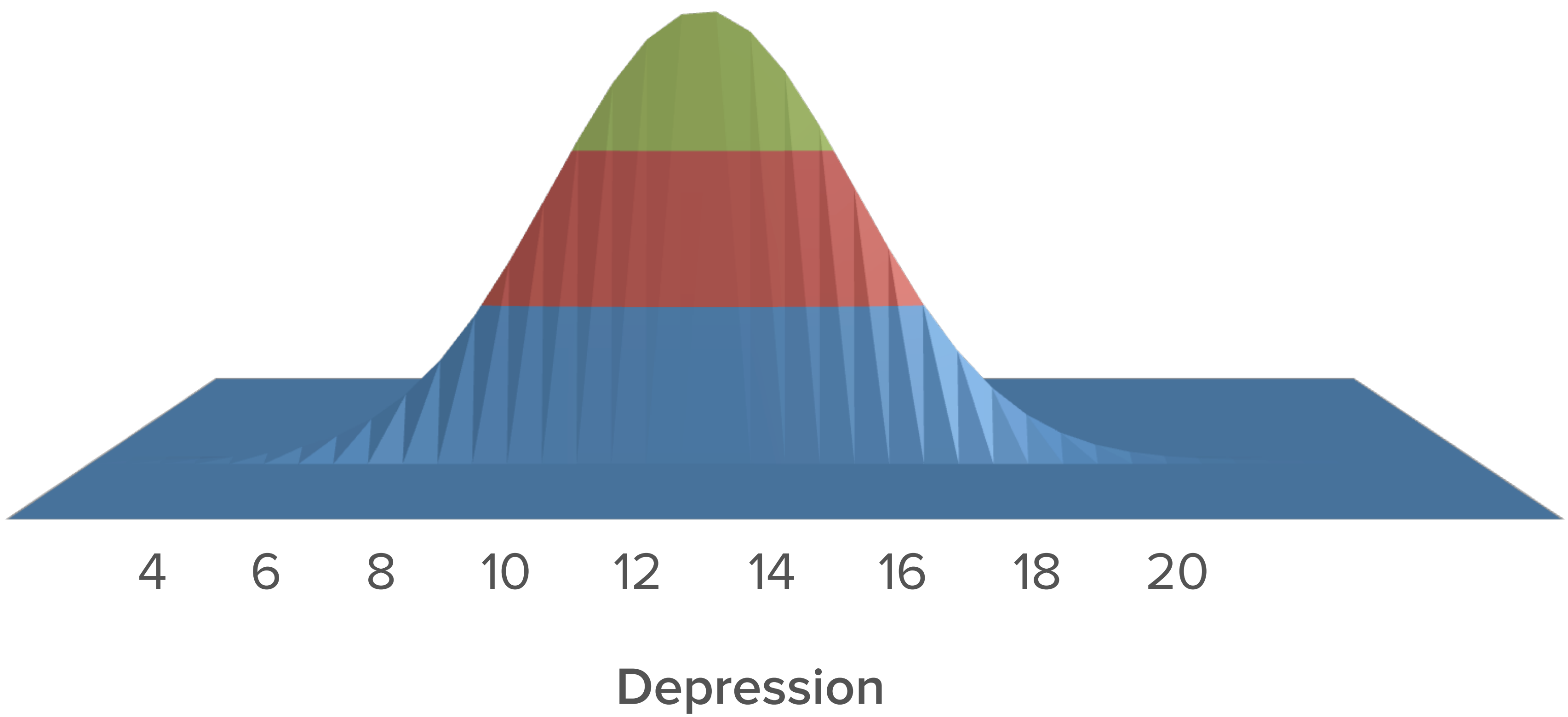


NORMAL CURVE SLICE SEVERITY = 10



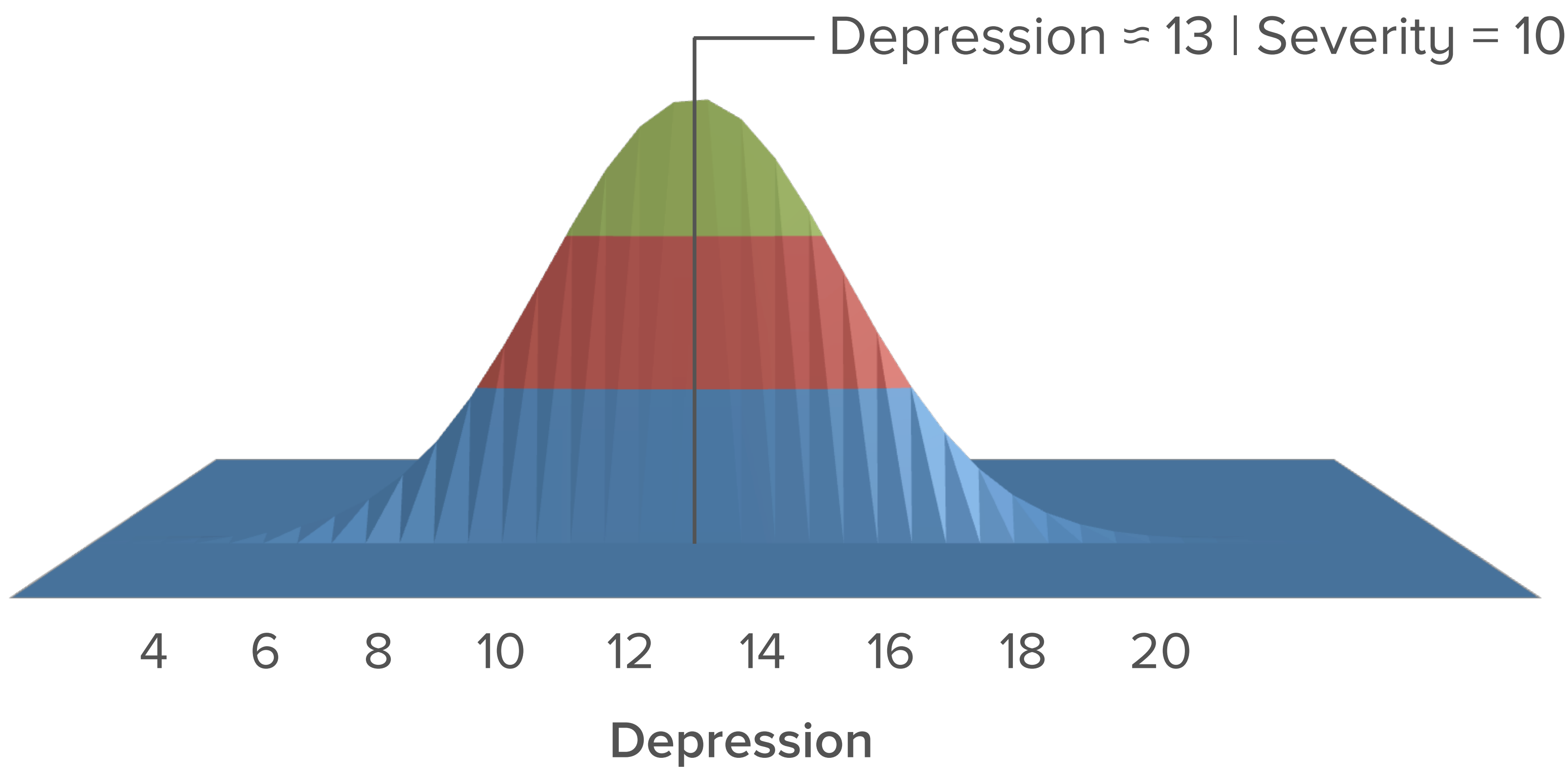
CONDITIONAL DISTRIBUTION

Distribution of plausible depression scores for a case with a severity score of 10



IMPLICIT IMPUTATION

Given a severity score of 10, the most likely value of the missing depression scale is ≈ 13



WHAT HAPPENS TO THE MEAN?

The complete cases produced an depression average of 15 (too high relative to the true estimate)

A case with an severity = 10 should have a parenting score of roughly 13

Again, adjusting the mean downward to account for the plausible (but missing) value brings the estimate closer to its true value of 14

ITERATIVE ESTIMATION

Begin with initial guesses about the parameters

Step 1: “Impute” missing values

Step 2: Update parameters based on imputations

Repeat 1 and 2 until estimates no longer change

ESTIMATION EXAMPLE

Use maximum likelihood to estimate the severity and depression means with missing data

Every combination of the two parameter values gives a log likelihood that represents fit to the data

The goal is to identify the parameter values that maximize the log likelihood (and thus fit to the data)

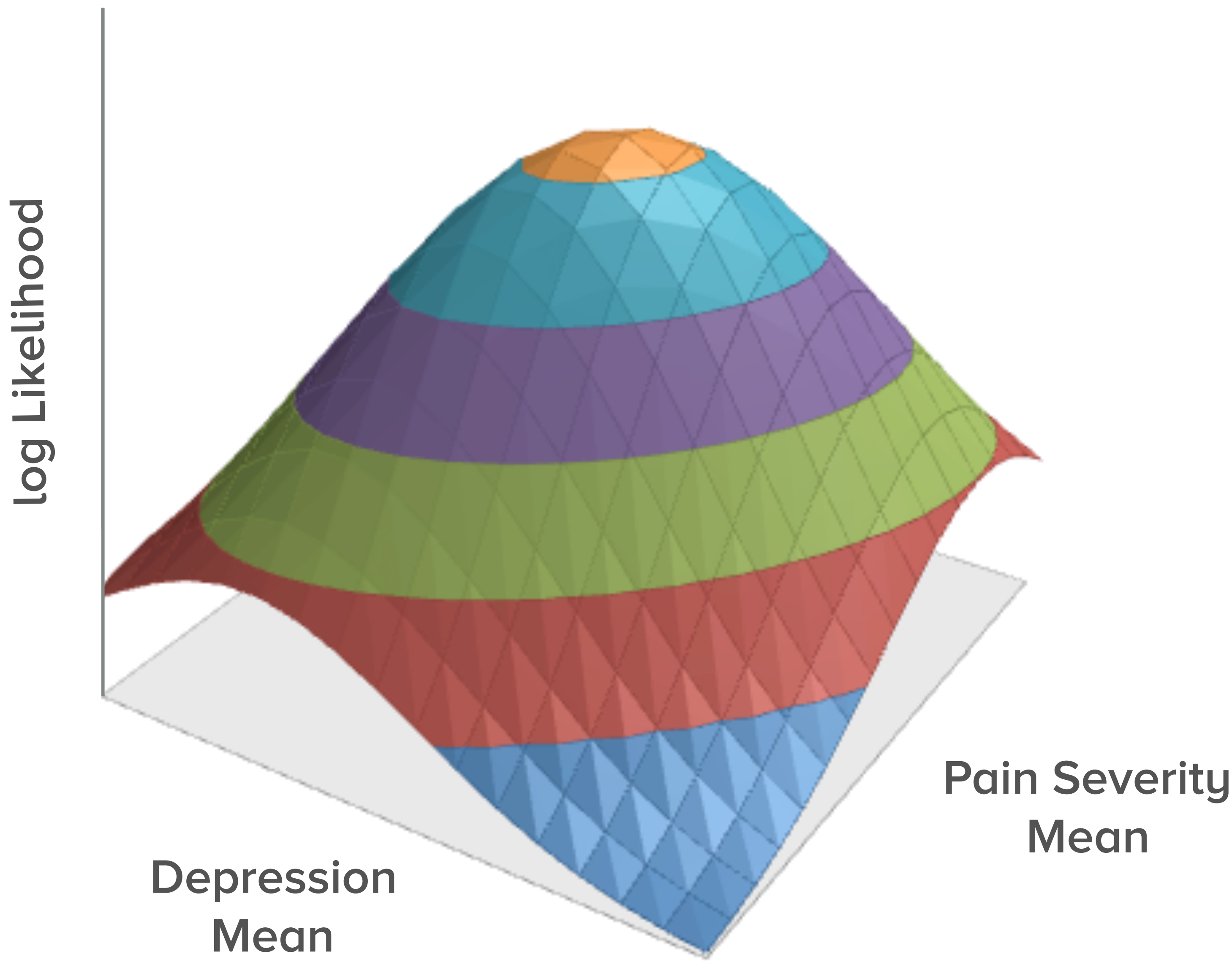
LOG LIKELIHOOD SURFACE

The log likelihood function for multiple parameters is a 3D surface that depicts the fit of different combinations of parameter values

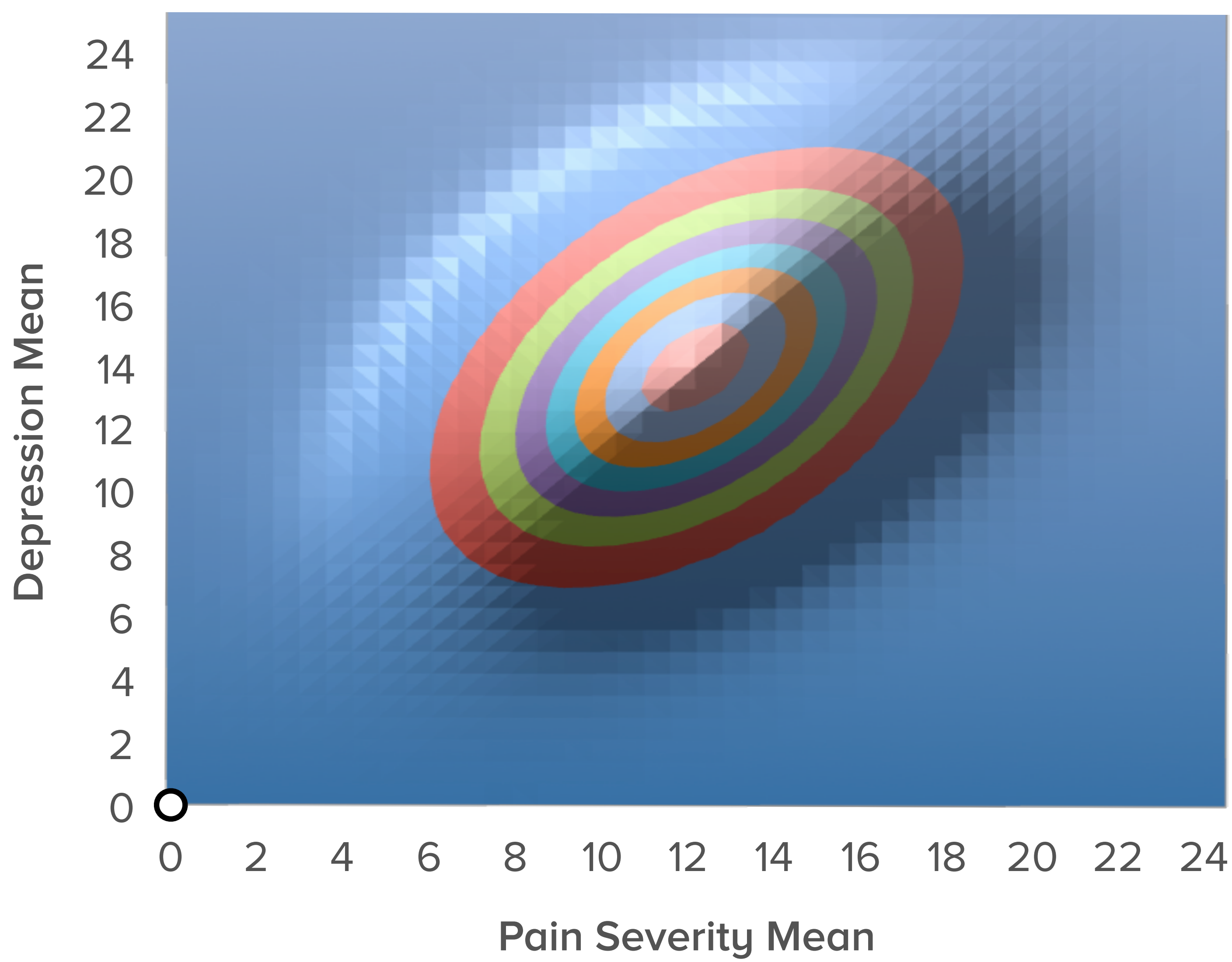
The goal of estimation is to climb to the top of the surface (identify the highest log likelihood)

The log likelihood is the altimeter for the climb

GRAPHIC



ESTIMATION GRAPHIC



Cycle	Sev. Mean	Dep. Mean
0	0	0
1	12	12.74
2	12	13.41
3	12	13.75
4	12	13.94
5	12	14.04
...	...	...
19	12	14.15

ESTIMATION SUMMARY

Including the incomplete cases gives estimates that better match those of the complete data

ML borrows information from the severity scores to adjust the depression estimates

Method	Severity Mean	Depression Mean
Complete	12.00	14.00
Deletion	13.53	15.00
ML	12.00	14.15

MAXIMUM LIKELIHOOD ESTIMATION IN MPLUS

WHY SEM SOFTWARE?

General-use software packages have a very limited capacity for ML missing data handling

Missing data is allowed only on outcomes, if at all

SEM software packages are extremely flexible, and any program can implement ML missing data handling for a wide variety of analyses

WISC DATA

WISC performance scores from 204 children

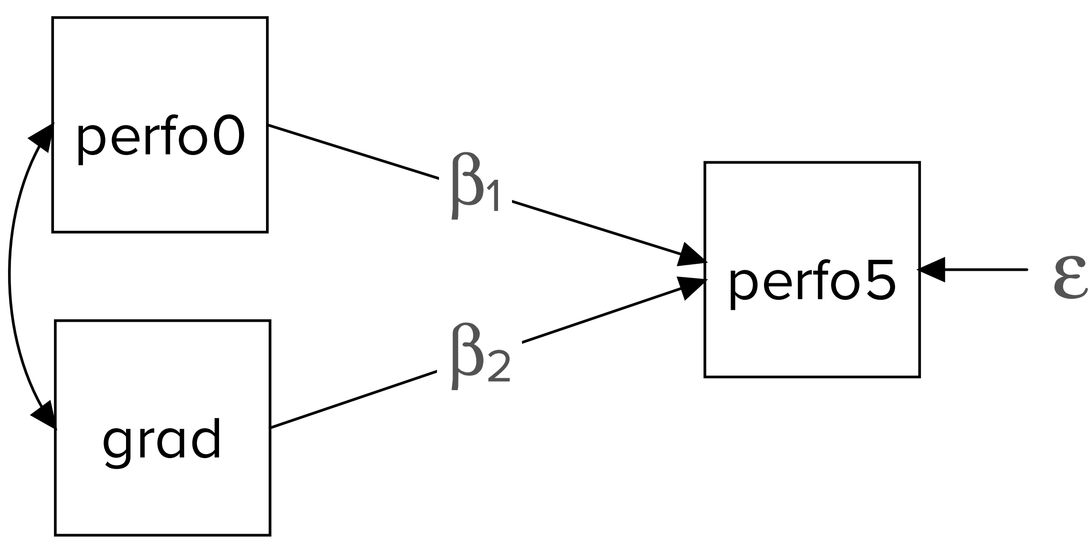
Students were tested in the spring prior to first grade (baseline), at the end of first grade (1 year later), at the end of 3rd grade (3 years later), and at the end of 5th grade (5 years later)

3rd and 5th grade scores and parent demographics are incomplete

ANALYSIS MODEL AND DIAGRAM

Mother's high school graduation status and kindergarten performance predicting 5th grade performance

$$perfo5 = \beta_0 + \beta_1(perfo0) + \beta_2(grad) + \varepsilon$$



WHAT DOES MAR REQUIRE?

Missingness on PERFO5 is completely explained by the observed values of PERFO1 or GRAD

Missingness on GRAD is completely explained by the observed values of PERFO1 or PERFO5

MPLUS COMMANDS

DATA: specify location of input text file

VARIABLE: provide variable names, select variables

DEFINE: create new or modify existing variables

ANALYSIS: specify estimator and analysis options

MODEL: specify analysis model

MODEL TEST: perform custom hypothesis tests

OUTPUT: control printing options

A FEW MPLUS RULES

Commands end in :

Subcommands end in ;

Capitalization does not matter

Variable names must be 8 characters or less

Command lines must be less than 80 characters

Use ! to specify a line that the program ignores

ML EX 1A - REGRESSION.INP

```
DATA:
file = wisc.dat;
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
  info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = perfo0 grad perfo5;
missing = all(-99);
ANALYSIS:
estimator = ml;
MODEL:
perfo0 grad;
perfo5 on perfo0 grad;
OUTPUT:
sampstat standardized(stdyx) patterns;
```

DATA COMMAND

The DATA command specifies the location of the input data file

Free format requires spaces, commas, or tabs as delimiters and a missing value code

```
DATA:
file = '/users/craig/desktop/wisc.dat';
```

Alternate Data Command

A file path is not required when the Mplus syntax file and the data are in the same directory

```
DATA:
file = wisc.dat;
```

VARIABLE COMMAND

The VARIABLE command (a) gives the order of the variables in the data file, (b) selects variables for analysis, (c) specifies missing value codes, and (d) defines special variables (categorical, grouping)

```
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
 info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = perfo0 grad perfo5;
missing = all(-99);
```

ANALYSIS COMMAND

Specify estimator and other special analysis options

Maximum likelihood (ML) is the default (no need to specify default options)

```
ANALYSIS:
estimator = ml;
```

MODEL COMMAND

ON denotes regression, WITH denotes covariance or correlation, and BY denotes a factor loading

A variable name by itself denotes a variance or residual variance and a name in [ ] specifies a mean or intercept

```
MODEL:
perfo5 on perfo0 grad;
```

Fixed Predictors

In line with OLS regression, Mplus treats predictor variables as fixed (no distributional assumptions)

Missing data handling requires a distribution

Cases with missing predictor scores are excluded

ANALYSIS SUMMARY

```
*** WARNING
  Data set contains cases with missing on x-variables.
  These cases were not included in the analysis.
  Number of cases with missing on x-variables:  14
*** WARNING
  Data set contains cases with missing on all variables except
  x-variables.  These cases were not included in the analysis.
  Number of cases with missing on all variables except x-variables:  47
   2 WARNING(S) FOUND IN THE INPUT INSTRUCTIONS


SUMMARY OF ANALYSIS

Number of groups                                                  1
Number of observations                                          143

Number of dependent variables                                     1
Number of independent variables                                   2
Number of continuous latent variables                             0
```

ASSIGNING A DISTRIBUTION

Specifying variances for the predictors triggers Mplus to treat predictors as random variables

A normal distribution is assumed, even for categorical variables

Necessary evil for missing data handling ....

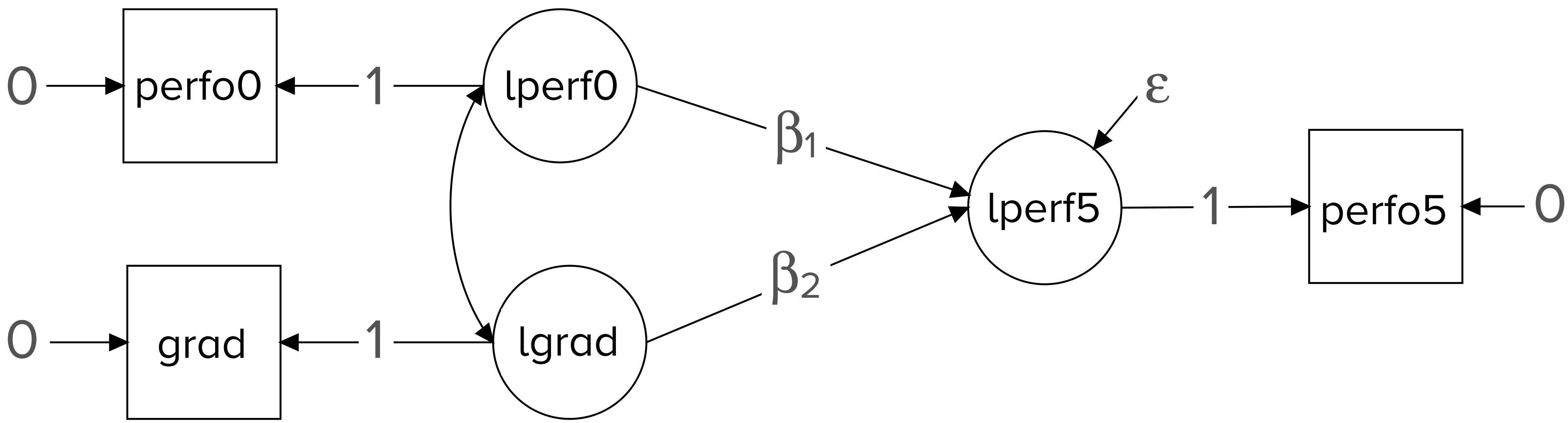
REVISED MODEL COMMAND

Specifying the variances of the predictors in the MODEL command triggers a normal distribution assumption and missing data handling for WBEING and JOBSAT

```
MODEL:
perfo0 grad;
perfo5 on perfo0 grad;
```

UNDERLYING MODEL

Predictor variables are treated as outcomes, and isomorphic latent variables replace predictors



ANALYSIS SUMMARY

INPUT READING TERMINATED NORMALLY

SUMMARY OF ANALYSIS

Number of groups	1
Number of observations	204
Number of dependent variables	1
Number of independent variables	2
Number of continuous latent variables	0

OUTPUT COMMAND

OUTPUT specifies information for the output file

SAMPSTAT gives descriptives, STANDARDIZED gives beta weights, and PATTERNS prints missing data patterns

```
OUTPUT:
sampstat standardized(stdyx) patterns;
```

MISSING DATA PATTERNS

```
SUMMARY OF MISSING DATA PATTERNS

    MISSING DATA PATTERNS (x = not missing)

              1  2  3  4
 PERFO5       x  x
 PERFO0       x  x  x  x
 GRAD         x     x

     MISSING DATA PATTERN FREQUENCIES

    Pattern   Frequency     Pattern   Frequency
          1         143           3          47
          2           9           4           5
```

COVARIANCE COVERAGE

The covariance coverage matrix gives the proportion of complete data for each variable or variable pair

```
PROPORTION OF DATA PRESENT

          Covariance Coverage
             PERFO5        PERFO0        GRAD
             ________      ________      ________
PERFO5         0.745
PERFO0         0.745         1.000
GRAD           0.701         0.931         0.931
```

DESCRIPTIVES

ESTIMATED SAMPLE STATISTICS

Means

PERFO5	PERFO0	GRAD
50.639	17.977	0.213

Covariances

	PERFO5	PERFO0	GRAD
PERFO5	160.847		
PERFO0	74.362	69.377	
GRAD	1.781	1.256	0.168

Correlations

	PERFO5	PERFO0	GRAD
PERFO5	1.000		
PERFO0	0.704	1.000	
GRAD	0.342	0.368	1.000

UNSTANDARDIZED ESTIMATES

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERFO5 ON				
PERFO0	1.018	0.094	10.837	0.000
GRAD	2.990	1.779	1.680	0.093
Intercepts				
PERFO5	31.707	1.771	17.905	0.000
Residual Variances				
PERFO5	79.842	9.163	8.713	0.000

INTERPRETATIONS

Interpret and report ML estimates in the same way as a complete-data analysis

Controlling for graduation status, a one-point increase in baseline performance results in a 1.018 increase in 5th grade performance, on average

Controlling for baseline performance, children with mothers who graduated scored 2.99 points higher at 5th grade, on average

STANDARDIZED ESTIMATES

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERFO5 ON				
PERFO0	0.668	0.047	14.160	0.000
GRAD	0.097	0.058	1.676	0.094

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERFO5	0.504	0.055	9.190	0.000

INTERPRETATIONS

Controlling for graduation status, a one standard deviation increase in baseline performance results in a .668 standard deviation increase in 5th grade performance, on average

Together, the two predictors explain 50.4% of the variance in job performance ratings

ADVANCED TACTICS: AUXILIARY VARIABLES

Researchers rarely know why data are missing

An inclusive strategy incorporates a set of auxiliary variables into the missing data handling routine

Good auxiliary variables are either (a) correlates of incomplete variables or (b) correlates of missingness

Auxiliary variables can increase power reduce bias

AUXILIARY VARIABLES

MOMED is associated with missingness on PERFO5 (mothers who did not graduate have kids with higher rates of missingness)

Including MOMED as an auxiliary variable can reduce nonresponse bias

Including PERFO3 can increase power because it is strongly correlated with PERFO5 (R = .81)

SPIDER MODEL

Graham (2003) outlined a so-called spider model for auxiliary variables

The model transmits information from the auxiliary variables via a series of correlations

The spider model does not alter the substantive interpretation of the parameter estimates

SPIDER MODEL RULES

Correlate each auxiliary variable with …

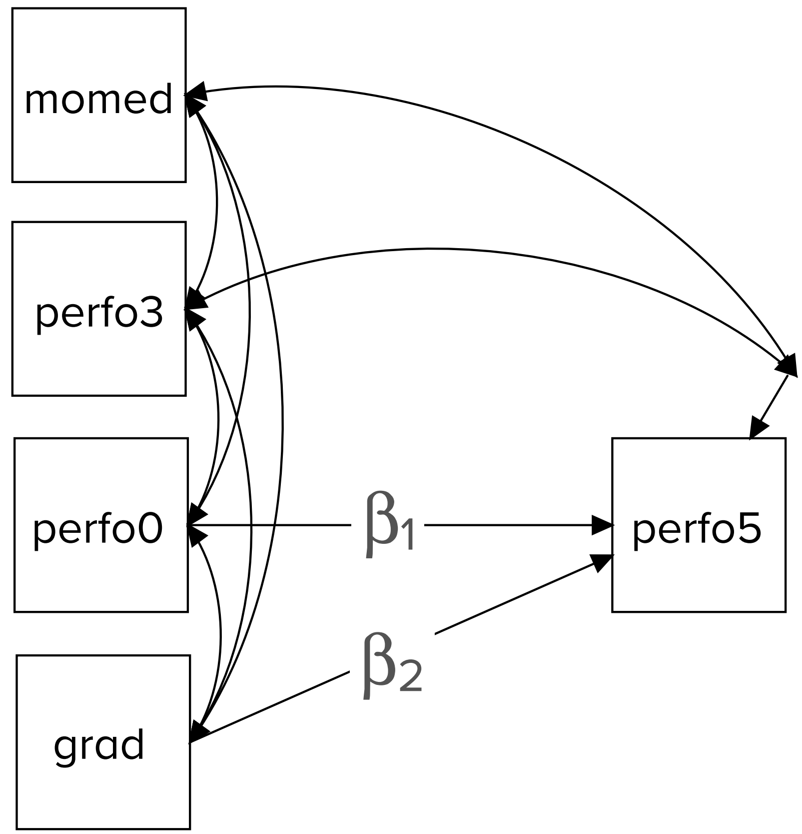
- Manifest predictor variables
- Other auxiliary variables
- The residual terms of all outcome variables

Do not correlation auxiliary variables with latents

ANALYSIS MODEL AND DIAGRAM

The interpretation of model parameters is unaffected by the presence of auxiliary variables

Interpret β_1 and β_2 in the same way as before



ML EX 1B - AUXILIARY VARIABLES.INP

```
DATA:
file = wisc.dat;
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
  info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = perfo0 grad perfo5;
missing = all(-99);
auxiliary = (m) momed perfo3;
ANALYSIS:
estimator = ml;
MODEL:
perfo0 grad;
perfo5 on perfo0 grad;
OUTPUT:
sampstat standardized(stdyx);
```

ANALYSIS SUMMARY

```
SUMMARY OF ANALYSIS

Number of groups                                                  1
Number of observations                                          204

Number of dependent variables                                     1
Number of independent variables                                   2
Number of continuous latent variables                             0

Observed dependent variables

  Continuous
   PERFO5

Observed independent variables
   PERFO0      GRAD

Observed auxiliary variables
   MOMED       PERFO3
```

UNSTANDARDIZED ESTIMATES

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERFO5 ON				
PERFO0	1.020	0.091	11.163	0.000
GRAD	2.790	1.735	1.608	0.108
Intercepts				
PERFO5	31.689	1.735	18.261	0.000
Residual Variances				
PERFO5	79.399	8.986	8.836	0.000

INTERPRETATIONS

Interpret and report ML estimates in the same way as a complete-data analysis

Controlling for graduation status, a one-point increase in baseline performance results in a 1.02 increase in 5th grade performance, on average

Controlling for baseline performance, children with mothers who graduated scored 2.79 points higher at 5th grade, on average

PRACTICAL ADVICE

Using a large number of auxiliary variables can lead to convergence problems

Identify a small number of variables with strong correlations (R > .40) with the analysis variables

Using a small number of variables with strong correlations is usually better than using a large number of variables with weak correlations

ANALYSIS EXAMPLE 2:

REPEATED MEASURES

ML EX 2A - REPEATED MEASURES.INP

```
DATA:
file = wisc.dat;
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
  info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = perfo0 perfo1 perfo3 perfo5;
missing = all(-99);
ANALYSIS:
estimator = ml;
MODEL:
[perfo0-perfo5] (mean0 mean1 mean3 mean5);
perfo0-perfo5 with perfo0-perfo5;
MODEL TEST:
mean0 = mean1; mean1 = mean3; mean3 = mean5;
OUTPUT:
sampstat patterns;
```

ANALYSIS SUMMARY

INPUT READING TERMINATED NORMALLY

SUMMARY OF ANALYSIS

Number of groups	1
Number of observations	204
Number of dependent variables	4
Number of independent variables	0
Number of continuous latent variables	0

COVARIANCE COVERAGE

PROPORTION OF DATA PRESENT

Covariance Coverage

	PERFO0	PERFO1	PERFO3	PERFO5
PERFO0	1.000			
PERFO1	1.000	1.000		
PERFO3	0.848	0.848	0.848	
PERFO5	0.745	0.745	0.745	0.745

MISSING DATA PATTERNS

```
SUMMARY OF MISSING DATA PATTERNS


     MISSING DATA PATTERNS (x = not missing)

           1  2  3
 PERFO0    x  x  x
 PERFO1    x  x  x
 PERFO3    x  x
 PERFO5    x


     MISSING DATA PATTERN FREQUENCIES

    Pattern   Frequency     Pattern   Frequency     Pattern   Frequency
          1         152           2          21           3          31
```

UNSTANDARDIZED ESTIMATES

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Means				
PERFO0	17.977	0.583	30.827	0.000
PERFO1	27.690	0.698	39.682	0.000
PERFO3	39.231	0.741	52.960	0.000
PERFO5	50.633	0.927	54.601	0.000
Variances				
PERFO0	69.377	6.869	10.100	0.000
PERFO1	99.333	9.835	10.100	0.000
PERFO3	105.107	10.873	9.667	0.000
PERFO5	157.091	16.699	9.407	0.000

UNSTANDARDIZED ESTIMATES, CONT.

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
PERFO0 WITH				
PERFO1	64.777	7.372	8.787	0.000
PERFO3	62.302	7.566	8.235	0.000
PERFO5	72.233	9.209	7.844	0.000
PERFO1 WITH				
PERFO3	80.439	9.245	8.701	0.000
PERFO5	97.959	11.401	8.592	0.000
PERFO3 WITH				
PERFO5	103.742	12.120	8.560	0.000

MODEL TEST (WALD STATISTIC)

The MODEL TEST command specifies constraints that are consistent with a hypothesis of no change (mean0 = mean1, mean1 = mean3, mean3 = mean5)

df = 3 because the Wald test posits three constraints

MODEL TEST OUTPUT

The significant chi-square, $\chi^2(3) = 2487.51$, indicates that the data are inconsistent with the null hypothesis of no change

```
Wald Test of Parameter Constraints

          Value                                2487.510
          Degrees of Freedom                          3
          P-Value                                0.0000
```

ADVANCED TACTICS: NEW PARAMETERS

Mplus provides facilities for computing and testing new parameters that are functions of estimated parameters

Pairwise comparisons and effect size estimates might be of interest in a repeated measures analysis

Label parameters and use the labels to define new parameters in the MODEL CONSTRAINT command

MODEL CONSTRAINT COMMAND

The MODEL CONSTRAINT command defines a pairwise comparison and Cohen's *d* effect size

```
MODEL:
[perfo0-perfo5] (mean0 mean1 mean3 mean5);
perfo0-perfo5 (var0 var1 var3 var5);
perfo0-perfo5 with perfo0-perfo5;
MODEL TEST:
mean0 = mean1; mean1 = mean3; mean3 = mean5;
MODEL CONSTRAINT:
new(change cohensd);
change = mean5 - mean0;
cohensd = change / sqrt(var0);
```

ML EX 2B - REPEATED MEASURES.INP

```
DATA:
file = wisc.dat;
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
  info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = perfo0 perfo1 perfo3 perfo5;
missing = all(-99);
ANALYSIS:
estimator = ml;
MODEL:
[perfo0-perfo5] (mean0 mean1 mean3 mean5);
perfo0-perfo5 (var0 var1 var3 var5);
perfo0-perfo5 with perfo0-perfo5;
MODEL TEST:
mean0 = mean1; mean1 = mean3; mean3 = mean5;
MODEL CONSTRAINT:
new(change cohensd);
change = mean5 - mean0;
cohensd = change / sqrt(var0);
```

ADDITIONAL ESTIMATES

The total mean difference between the first and last assessment is 32.65, which is equivalent to 3.92 standard deviation units (large effect size)

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
New/Additional Parameters				
CHANGE	32.656	0.701	46.563	0.000
COHENSD	3.921	0.212	18.531	0.000

MULTIPLE IMPUTATION

OVERVIEW

Multiple imputation creates several (20 or more) copies of the data, each with a different set of plausible replacement values

A single collection of imputed data sets can serve as input for many different analyses

This contrasts maximum likelihood, where missing data handling and estimation are integrated

Multiple Imputation Steps

Imputation phase

- Create copies of the data with different imputed values

Analysis phase

- Perform analyses separately on each data set

Pooling phase

- Combine estimates and standard errors

THE IDEA BEHIND IMPUTATION

Specify a distribution for the missing values

Use a regression model to sample missing values from a distribution that conditions on the complete data

Complete variables are predictors and incomplete variables are outcomes

OVERVIEW OF IMPUTATION PHASE

Markov chain Monte Carlo (MCMC) is the mathematical machinery for Bayesian estimation and imputation

A two-step MCMC algorithm repeatedly generates imputations (imputation step) and samples new regression model parameters (posterior step)

A unique set of regression parameters generates each imputed data set

MCMC CYCLE 1

Start with initial regression model parameters

Imputation Step: Sample new imputations, conditional on the initial regression parameters

Posterior Step: Sample new regression parameters, conditional on the cycle 1 imputations

End the first MCMC cycle

MCMC CYCLE 2

Start with regression parameters from the first cycle

Imputation Step: Sample new imputations, conditional on the cycle 1 regression parameters

Posterior Step: Sample new regression parameters, conditional on the cycle 2 imputations

End the second MCMC cycle

PAIN DATA

20 chronic pain patients enrolled in a pain management program

Patients with mild pain are more likely to refuse the depression measure

Pain Severity	Depression
4	?
6	?
7	14
7	11
8	?
9	?
9	11
10	?
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

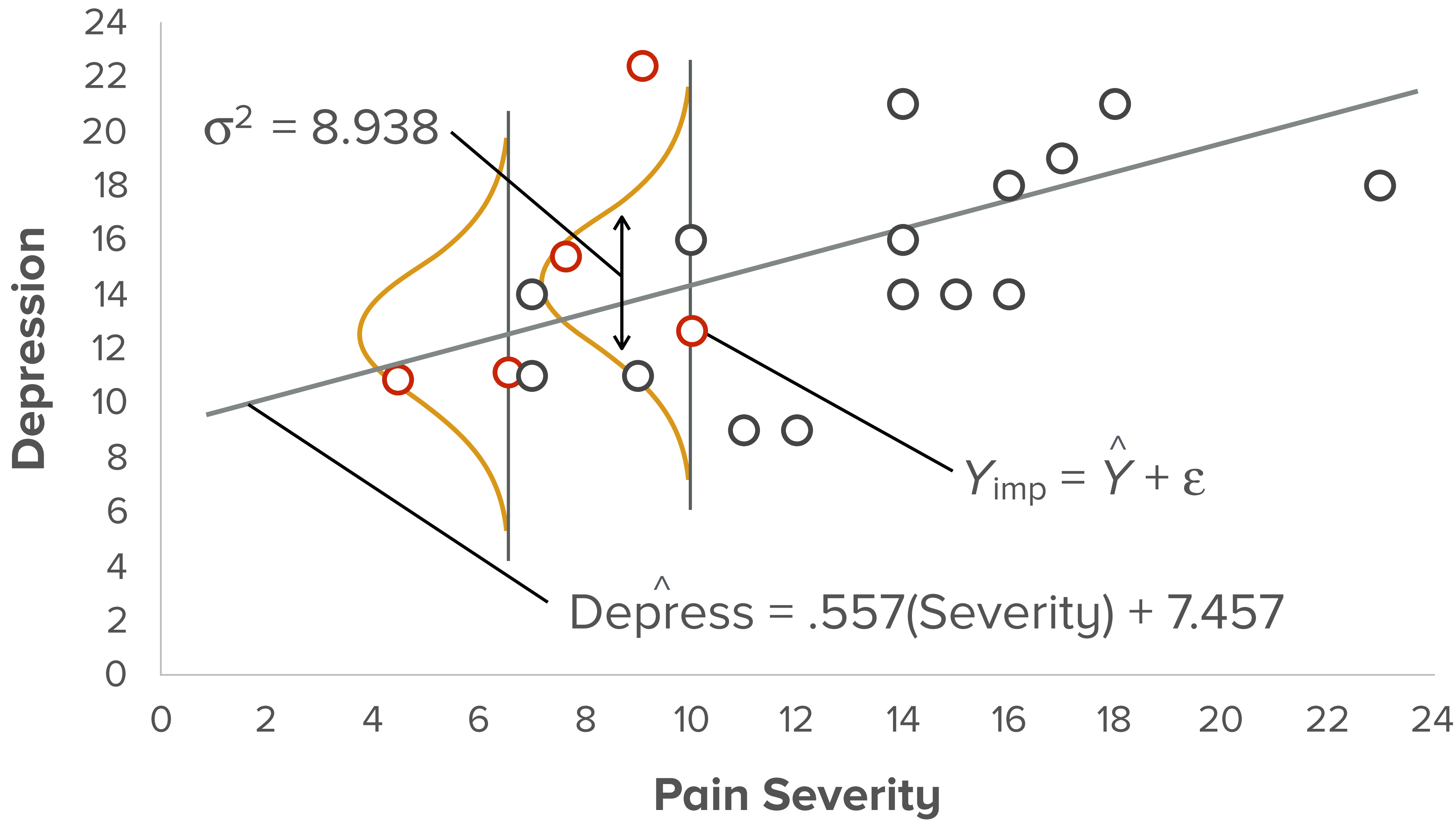
INITIAL REGRESSION PARAMETERS

The imputation regression model specifies complete pain ratings as a predictor and the incomplete depression variable as a normally distributed outcome

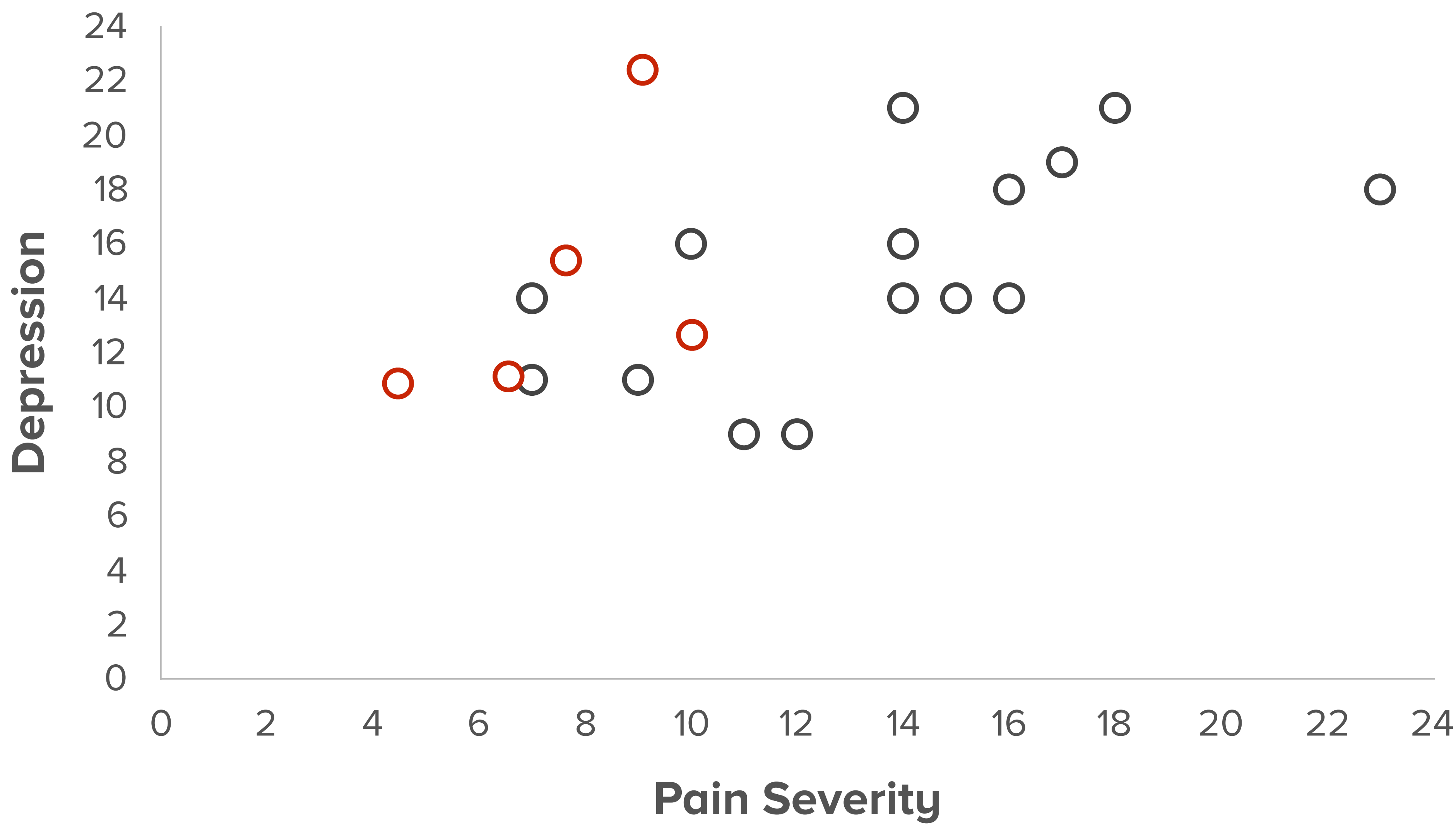
The first imputation step requires an intercept, slope, and a residual variance (e.g., from deletion)

$\beta_0 = 7.457$, $\beta_1 = .557$, $\sigma^2 = 8.938$

CYCLE 1 IMPUTATION STEP (I-STEP)



CYCLE 1 IMPUTATIONS



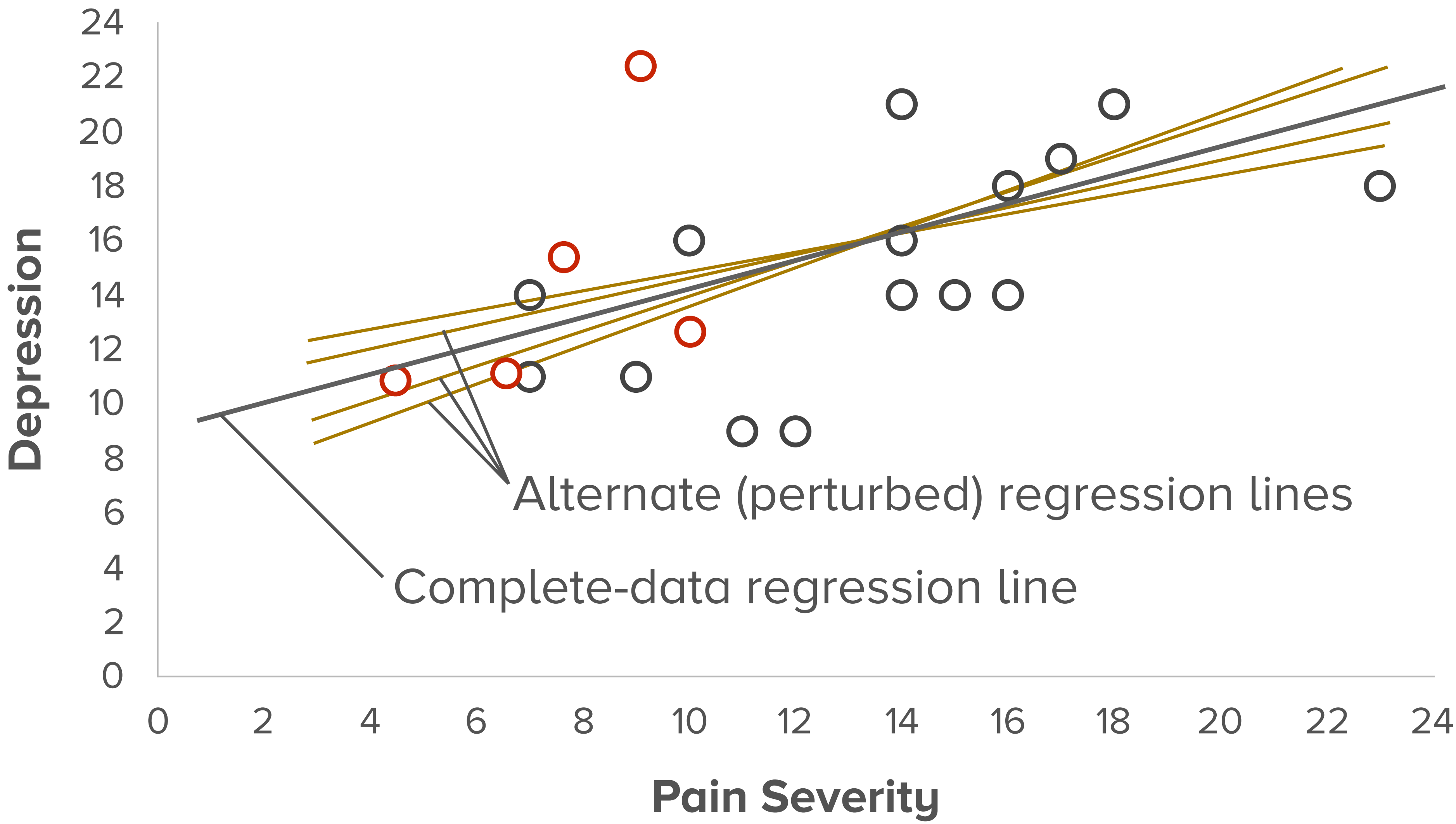
POSTERIOR STEP (P-STEP)

The next round of imputation requires a different set of regression parameters

Updated values are obtained by estimating the regression from the filled-in data and randomly perturbing the resulting estimates

Updating is performed within the Bayes framework

ALTERNATE REGRESSION LINES



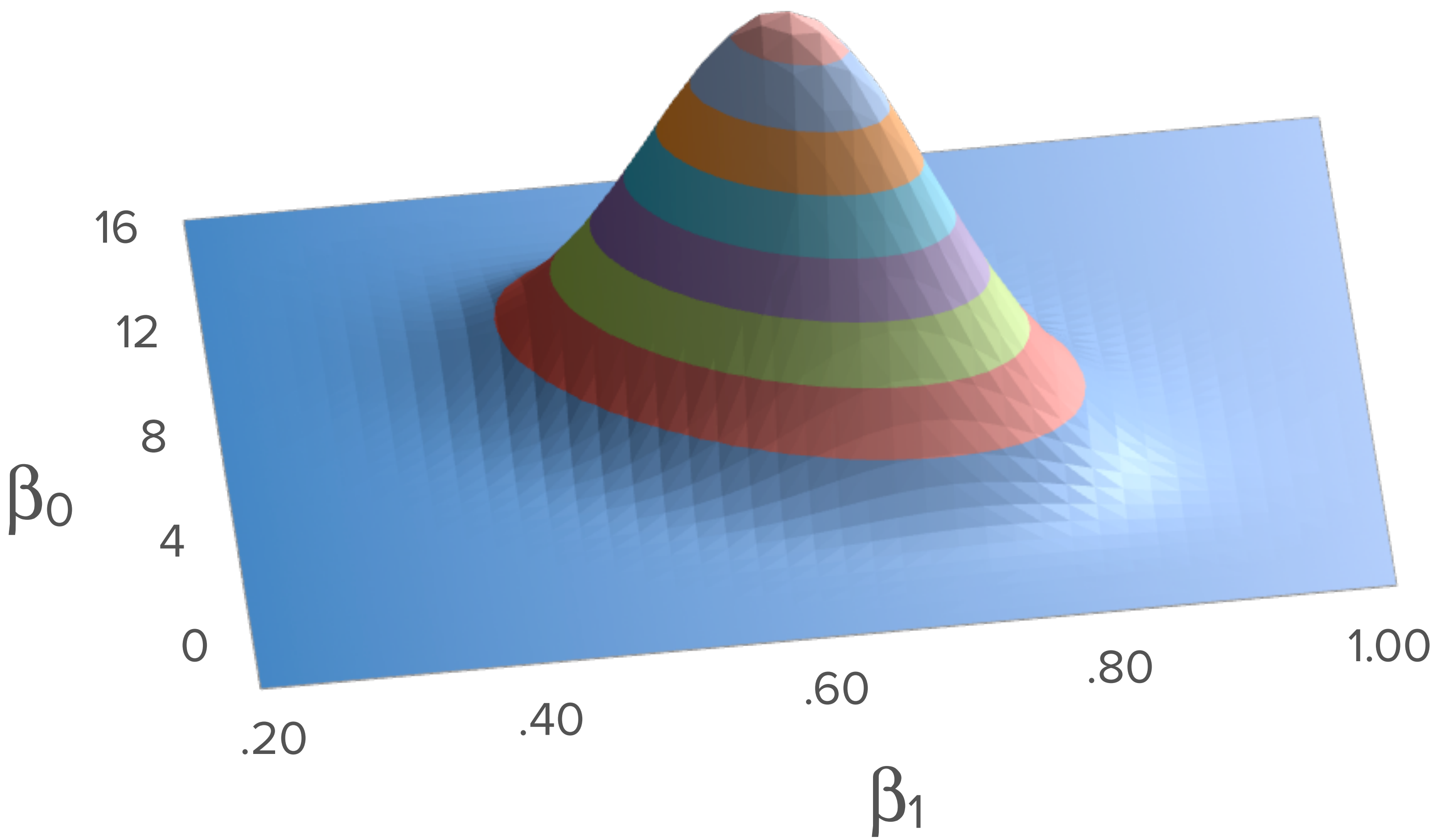
SAMPLING REGRESSION COEFFICIENTS

New βs are drawn from a multivariate normal distribution, where OLS estimates from the complete data define the mean vector and covariance matrix

$$\beta \sim MVN(\hat{\beta}, \Sigma_{\beta})$$

$$\hat{\beta} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y} \qquad \Sigma_{\beta} = \sigma_{\varepsilon}^{2}(\mathbf{X}^{T}\mathbf{X})^{-1}$$

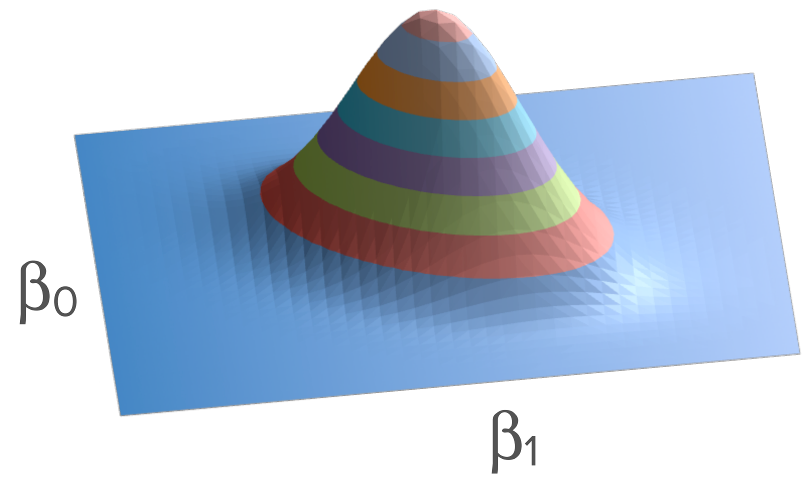
POSTERIOR DISTRIBUTION GRAPHIC

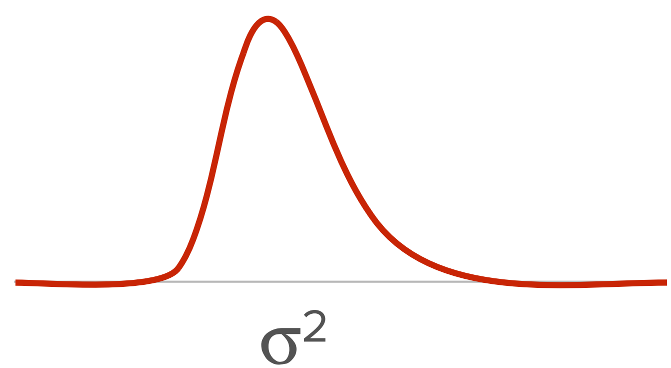


CYCLE 1 PARAMETER VALUES

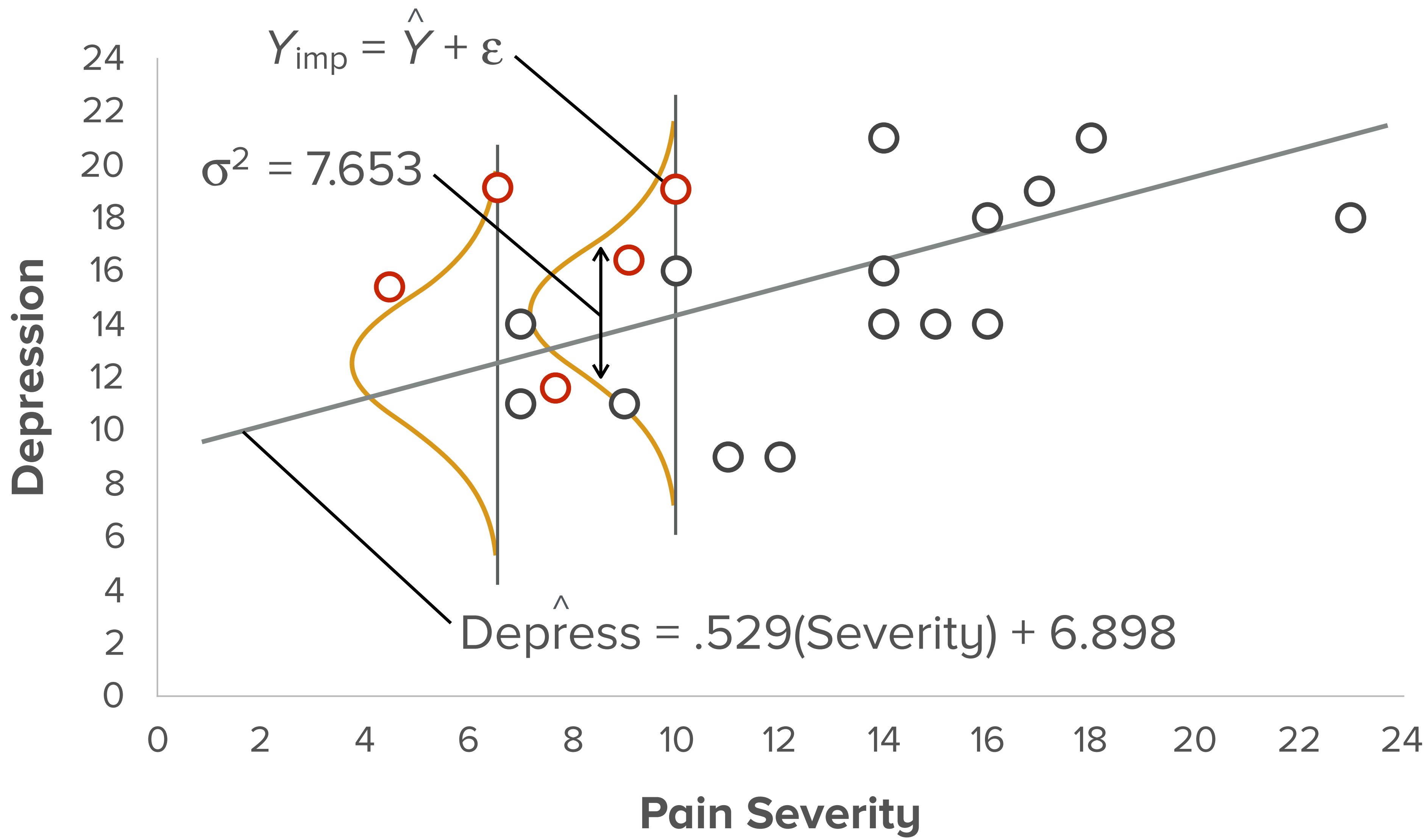
The MCMC algorithm uses random number generation to randomly update the parameter values

β_0 = 6.898, β_1 = .529, σ^2 = 7.653

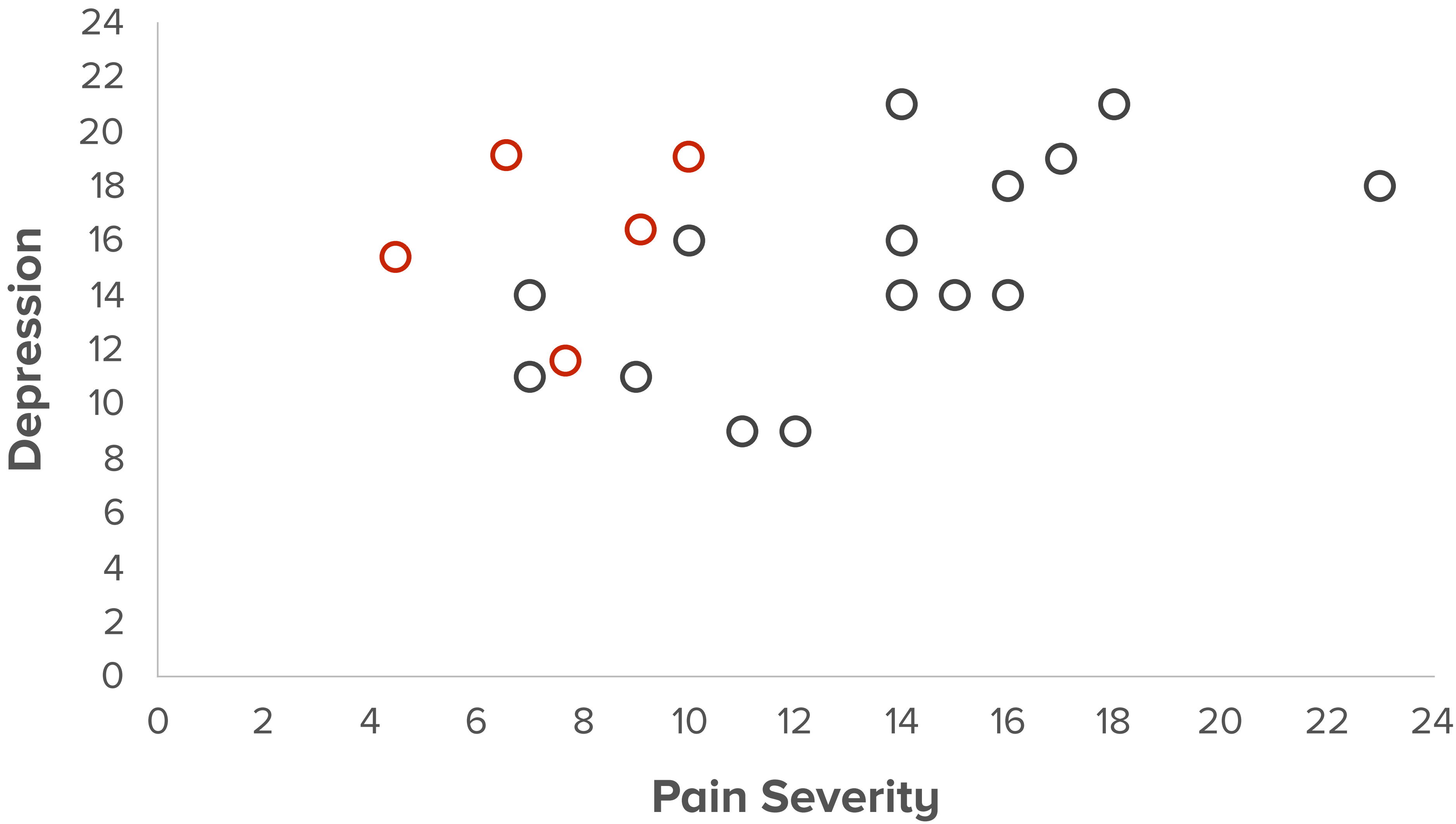




CYCLE 2 IMPUTATION STEP (I-STEP)



CYCLE 2 IMPUTATIONS



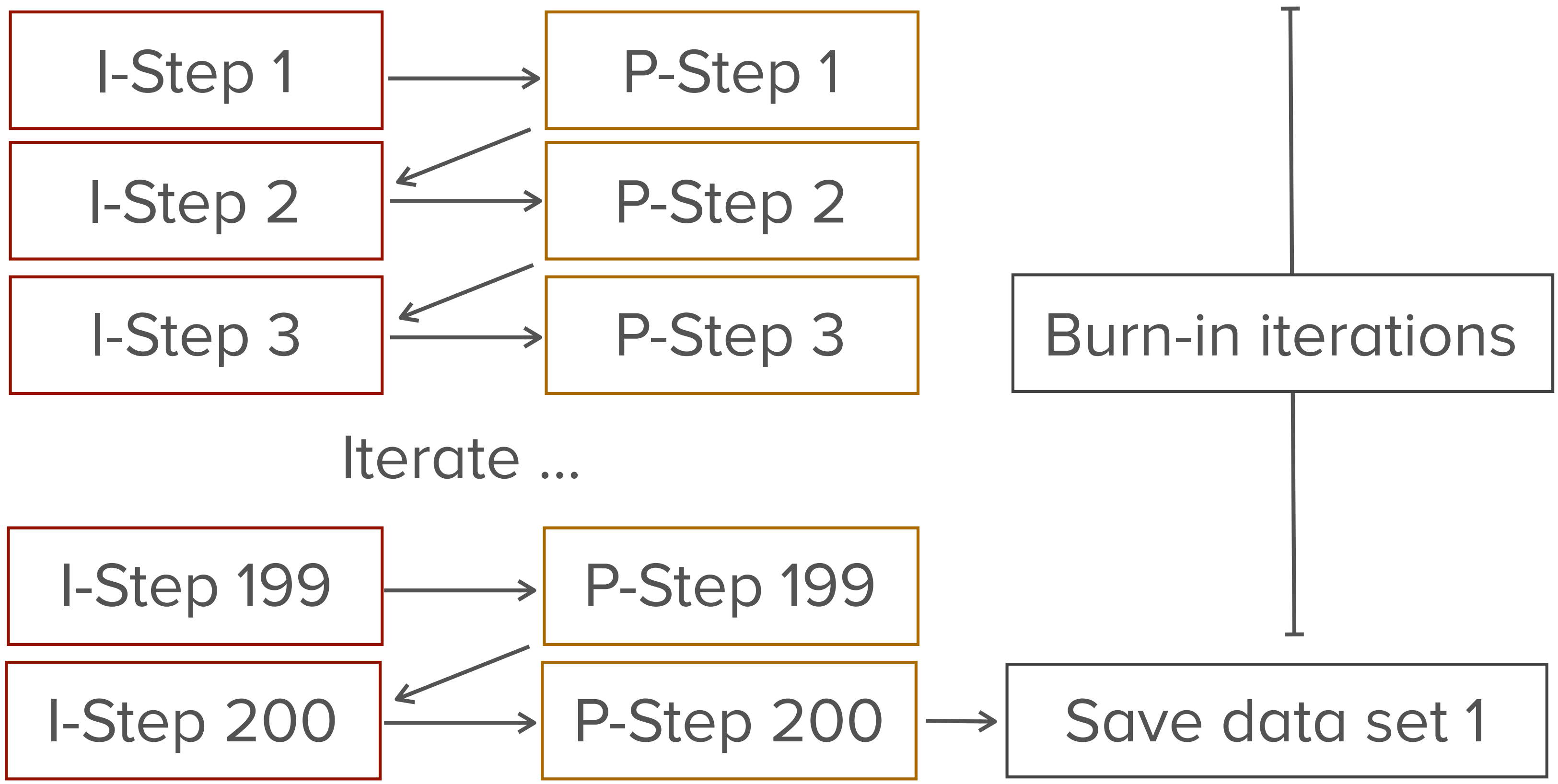
THINNING

Imputed data sets from consecutive MCMC cycles are highly correlated (too similar)

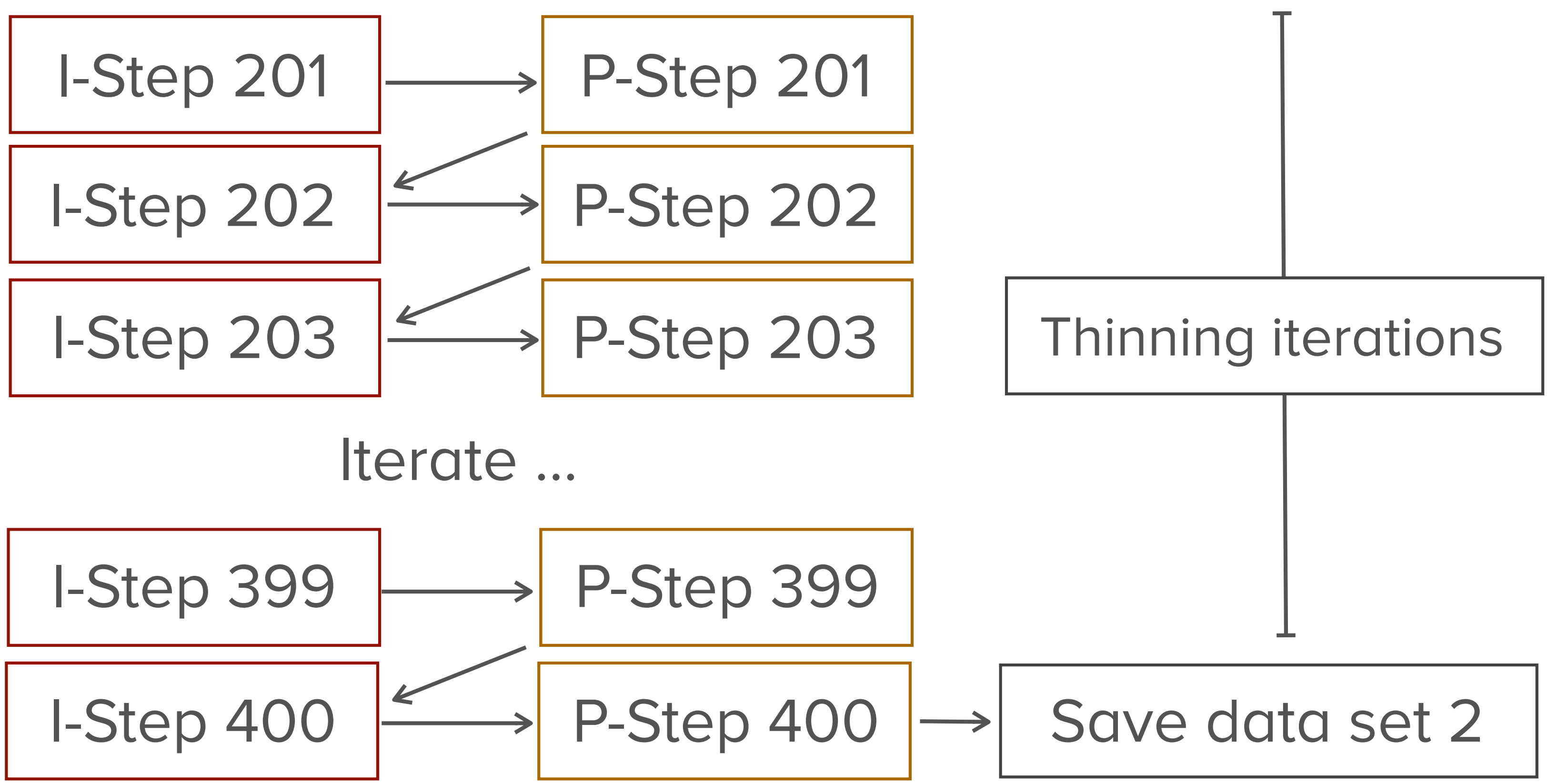
Saving imputed data sets at specified intervals in the MCMC chain (after every 200th cycle) eliminates unwanted dependencies

The Bayes literature refers to this as thinning

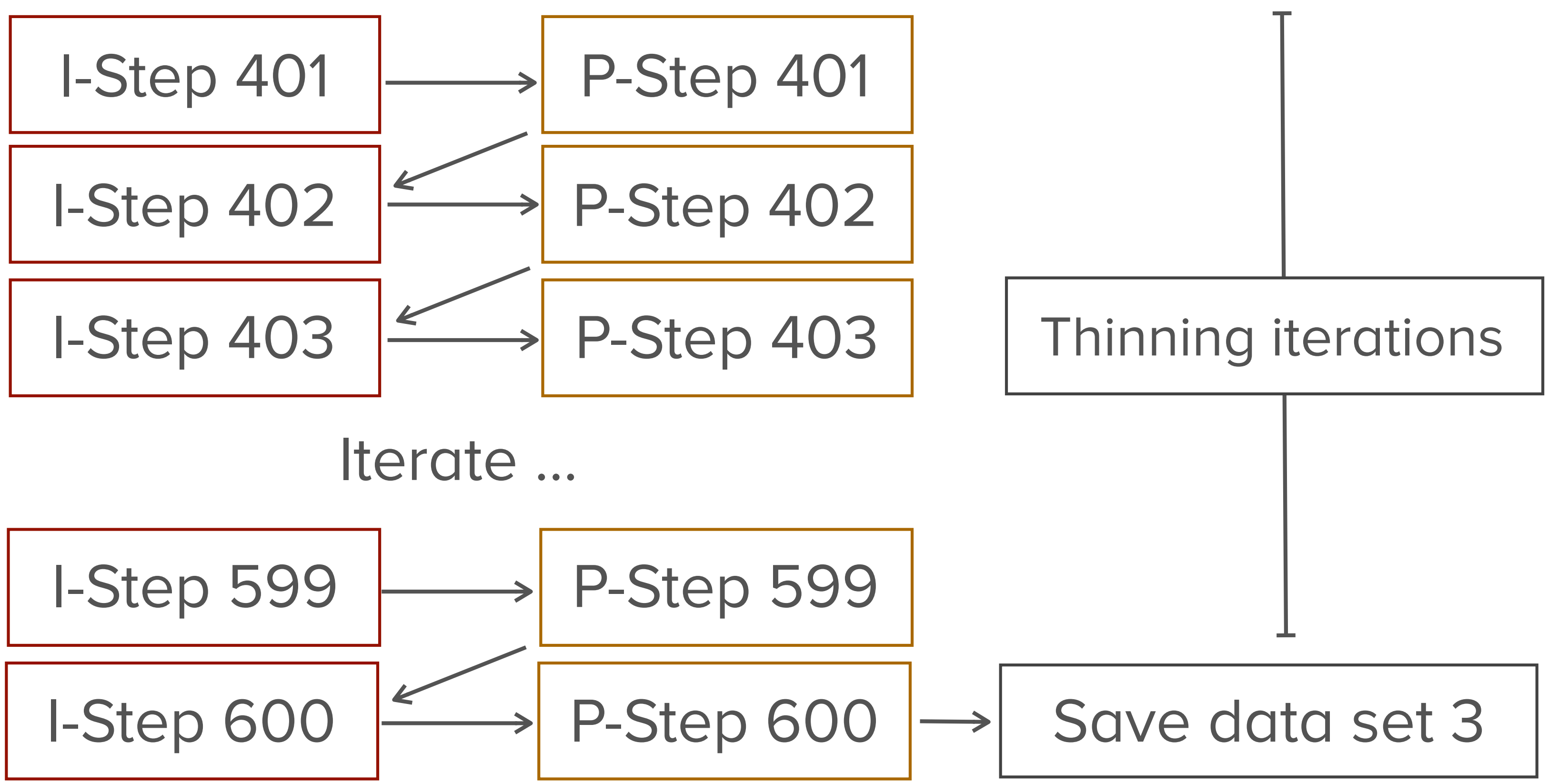
BURN-IN ITERATIONS



THINNING ITERATIONS

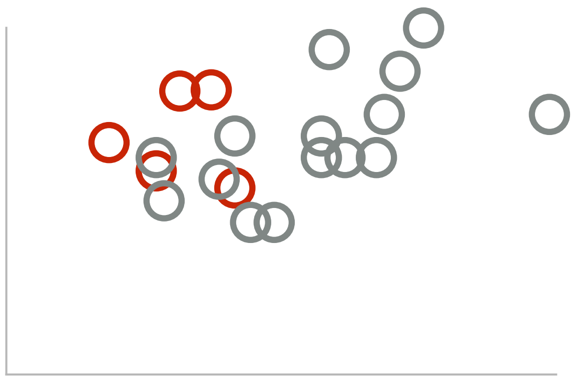


THINNING ITERATIONS

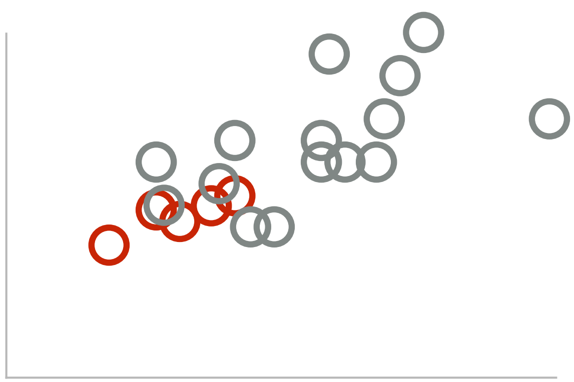


IMPUTED DATA SCATTERPLOTS

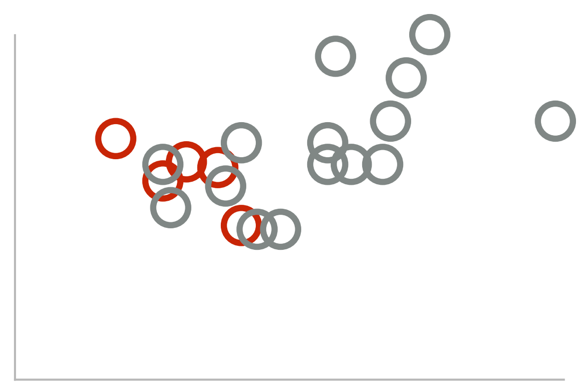
Dataset 1



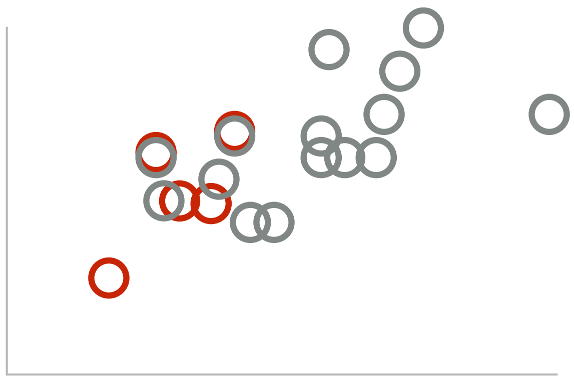
Dataset 2



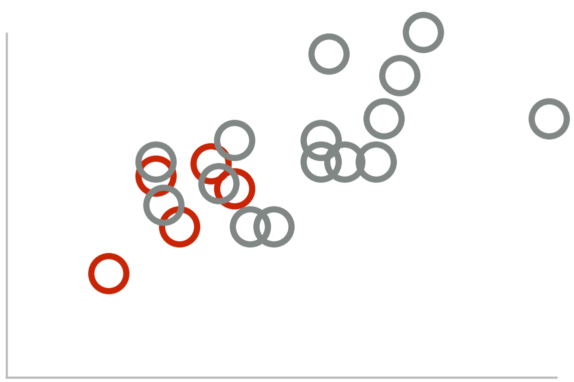
Dataset 3



Dataset 4



Dataset 5



IMPUTED DATA SETS

Pain	Depress
4	16.87
6	15.00
7	14
7	11
8	10.06
9	18.64
9	11
10	18.02
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

Pain	Depress
4	11.34
6	7.80
7	14
7	11
8	15.61
9	13.32
9	11
10	11.61
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

Pain	Depress
4	14.48
6	10.86
7	14
7	11
8	12.16
9	15.28
9	11
10	6.36
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

Pain	Depress
4	7.86
6	18.29
7	14
7	11
8	16.72
9	11.26
9	11
10	2.93
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

Pain	Depress
4	12.55
6	8.70
7	14
7	11
8	16.89
9	17.03
9	11
10	13.05
10	16
11	9
12	9
14	14
14	16
14	21
15	14
16	14
16	18
17	19
18	21
23	18

ANALYSIS AND POOLING PHASES

Following imputation, analyze each filled-in data set to get estimates and standard errors from each

The pooling phase combines the estimates and standard errors into a single set of results

Rubin (1987) gives the pooling equations

AVERAGING PARAMETER ESTIMATES

The multiple imputation point estimate is the arithmetic average of the *m* complete-data estimates

$$\theta = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m}$$

$\hat{\theta}_i$ is a parameter estimate from data set *i*

ANALYSIS RESULTS

Imputation	$M_{Depress}$	$SD_{Depress}$	$R_{Pain.Depress}$
1	15.18	3.74	0.40
2	14.23	3.87	0.66
3	14.21	4.00	0.57
4	14.10	4.73	0.50
5	14.66	3.78	0.57
MI Estimate	**14.48**	**4.02**	**0.54**

POOLING STANDARD ERRORS

Standard errors consist of two components

The within-imputation variance estimates complete-data sampling error

The between-imputation variance estimates the additional noise from missing data

ANALYSIS RESULTS

Imputation	$M_{Depress}$	SE	SE^2
1	15.18	0.836	0.699
2	14.23	0.865	0.748
3	14.21	0.894	0.799
4	14.10	1.057	1.117
5	14.66	0.844	0.713

WITHIN-IMPUTATION VARIANCE

Within-imputation variance is the average sampling variance (squared standard error) from the *m* imputed data sets

$$V_W = \frac{\sum_{i=1}^{m} SE_i}{m}$$

V_W estimates the sampling error that would have resulted had the data been complete

MISSING DATA UNCERTAINTY

Missing values do not affect the pain severity estimates

The depression parameters vary because the five data sets contain different imputations

Imputation	M_{Pain}	$M_{Depress}$
1	12.00	15.18
2	12.00	14.23
3	12.00	14.21
4	12.00	14.10
5	12.00	14.66

BETWEEN-IMPUTATION VARIANCE

Between-imputation variance quantifies variation in the parameter values caused by missing data

$$V_B = \frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}{m - 1}$$

V_B applies the usual formula for the sample variance to the *m* parameter estimates

STANDARD ERROR

The standard error combines complete-data and missing-data variation

$$SE = \sqrt{V_W + V_B + \frac{V_B}{m}}$$

$m^{-1} V_B$ is the squared standard error of the pooled parameter estimate from the V_B formula

EXAMPLE

Complete-data sampling variance

$$V_W = \frac{.699 + .748 + .799 + 1.117 + .713}{5} = .815$$

Missing-data variance

$$V_B = \frac{(15.18 - 14.48)^2 + (14.23 - 14.48)^2 + \ldots + (14.66 - 14.48)^2}{5 - 1} = .199$$

Standard error

$$SE = \sqrt{.815 + .199 + .199 / 5} = 1.027$$

SIGNIFICANCE TESTS

Significance tests use the usual *t* (or *z*) ratio

$$t = \frac{\bar{\theta} - \theta_0}{SE}$$

Degrees of freedom are complex and depend on *m*, the amount of missing data, and the correlations among the variables

SELECTING VARIABLES FOR IMPUTATION

The imputation phase must include all variables and effects (interactions, non-linear terms, special data structures) that will be part of the subsequent analyses as well as any auxiliary variables

Excluding analysis variables will bias parameter estimates toward zero

Special algorithms are needed for multilevel data

HOW MANY IMPUTATIONS?

Classic references recommend 3 to 5 data sets

Standard errors decrease as the number of imputed data sets increases (to a point)

Recent research suggests that m = 20 often yields power that is comparable to maximum likelihood (Graham, Olchowski, & Gilreath, 2007)

MULTIPLE IMPUTATION IN MPLUS

MI EX 1A - IMPUTATION.INP

```
DATA:
file = wisc.dat;
VARIABLE:
names = id verb0 verb1 verb3 verb5 perfo0 perfo1 perfo3 perfo5
  info0 comp0 simi0 voca0 info5 comp5 simi5 voca5 momed grad;
usevariables = grad perfo0 perfo1 perfo3 perfo5;
missing = all(-99);
ANALYSIS:
type = basic;
bseed = 90291;
DATA IMPUTATION:
impute = grad (c) perfo3 perfo5;
ndatasets = 50;
save = wiscimp*.dat;
thin = 200;
OUTPUT:
tech8;
```

IMPUTED DATA FORMAT

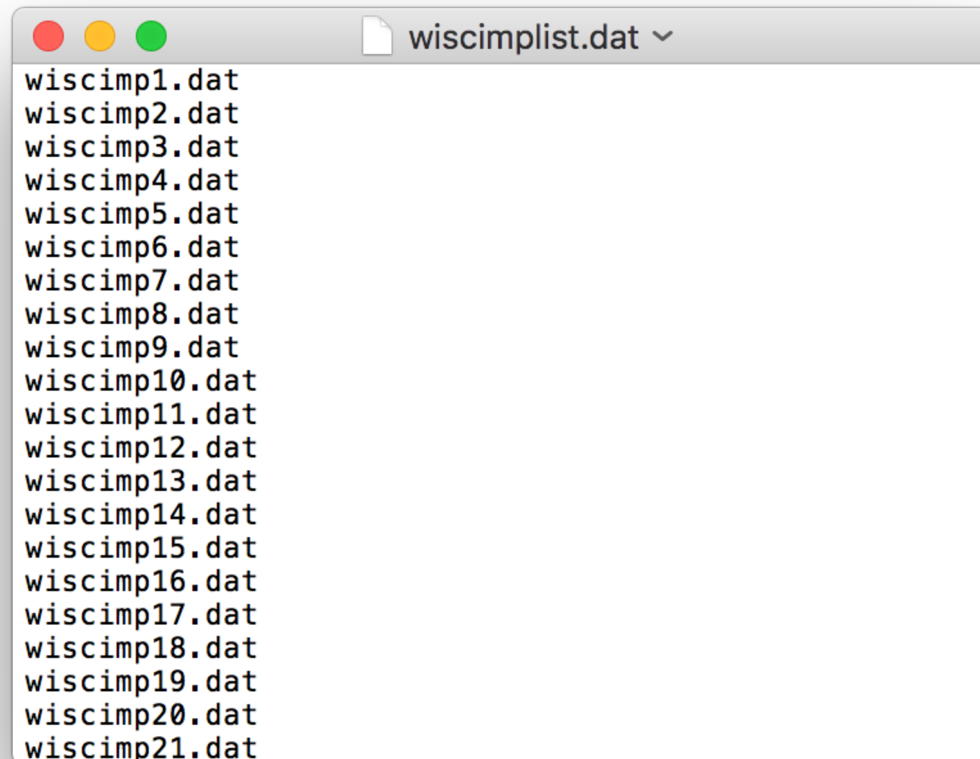
Mplus saves each imputed data set to a separate file

The file names use the prefix specified in the SAVE command (wiscimp*.dat)

Mplus creates a text file containing the data set names, and this file serves as input for the analysis

LISTING FILE

The listing file containing the data set names appends the word “list” to the prefix specified in the SAVE command


wiscimplist.dat

```
wiscimp1.dat
wiscimp2.dat
wiscimp3.dat
wiscimp4.dat
wiscimp5.dat
wiscimp6.dat
wiscimp7.dat
wiscimp8.dat
wiscimp9.dat
wiscimp10.dat
wiscimp11.dat
wiscimp12.dat
wiscimp13.dat
wiscimp14.dat
wiscimp15.dat
wiscimp16.dat
wiscimp17.dat
wiscimp18.dat
wiscimp19.dat
wiscimp20.dat
wiscimp21.dat
```

VARIABLE ORDER

Mplus lists the variable order for the imputed data sets near the bottom of the output file

```
SAVEDATA INFORMATION

  Save file
    wiscimp*.dat

  Order of variables

    GRAD
    PERFO0
    PERFO1
    PERFO3
    PERFO5
```

Analyzing Imputed Data

Mplus automates the analysis and pooling phases

Analyzing imputed data sets requires a small change to the DATA command, but the remaining commands are identical to a complete-data analysis

Many analyses can draw on the same imputations

MI EX 1B - REGRESSION ANALYSIS.INP

```
DATA:
file = wiscimplist.dat;
type = imputation;
VARIABLE:
names = grad perfo0 perfo1 perfo3 perfo5;
usevariables = grad perfo0 perfo5;
ANALYSIS:
estimator = ml;
MODEL:
perfo5 on perfo0 grad;
OUTPUT:
standardized(stdyx);
```

ANALYSIS SUMMARY

```
INPUT READING TERMINATED NORMALLY


SUMMARY OF ANALYSIS

Number of groups                                                  1
Average number of observations                                  204

Number of replications
    Requested                                                    50
    Completed                                                    50

Number of dependent variables                                     1
Number of independent variables                                   2
Number of continuous latent variables                             0
```

DESCRIPTIVES

NOTE: These are average results over 50 data sets.

SAMPLE STATISTICS

Means

PERFO5	GRAD	PERFO0
50.554	0.211	17.977

Correlations

	PERFO5	GRAD	PERFO0
PERFO5	1.000		
GRAD	0.353	1.000	
PERFO0	0.690	0.374	1.000

UNSTANDARDIZED ESTIMATES

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
PERFO5 ON					
PERFO0	0.976	0.088	11.029	0.000	0.156
GRAD	3.380	1.746	1.936	0.053	0.096
Intercepts					
PERFO5	32.296	1.684	19.182	0.000	0.200
Residual Variances					
PERFO5	80.468	9.329	8.626	0.000	0.270

INTERPRETATIONS

Interpret and report estimates in the same way as a complete-data analysis

Controlling for graduation status, a one-point increase in baseline performance results in a .976 increase in 5th grade performance, on average

Controlling for baseline performance, children with mothers who graduated scored 3.38 points higher at 5th grade, on average

STANDARDIZED ESTIMATES

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
PERFO5 ON					
PERFO0	0.649	0.047	13.759	0.000	0.152
GRAD	0.110	0.057	1.942	0.052	0.093

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
PERFO5	0.487	0.054	8.999	0.000	0.146

INTERPRETATIONS

Controlling for graduation status, a one standard deviation increase in baseline performance results in a .649 standard deviation increase in 5th grade performance, on average

Together, the two predictors explain 48.7% of the variance in job performance ratings

COMPARISON OF ESTIMATES

Maximum Likelihood	Estimate	Multiple Imputation	Estimate
PERFO5 ON		PERFO5 ON	
PERFO0	1.018	PERFO0	0.976
GRAD	2.990	GRAD	3.380
Intercepts		Intercepts	
PERFO5	31.707	PERFO5	32.296
Residual Variances		Residual Variances	
PERFO5	79.842	PERFO5	80.468

Practical Conclusions

Maximum likelihood and multiple imputation produced nearly identical results

This is typically the case, as the procedures are equivalent in large samples

Practical considerations and personal preference often dictate the choice of method

ANALYSIS EXAMPLE 2:

REPEATED MEASURES

MI EX 2 - REPEATED MEASURES.INP

```
DATA:
file = wiscimplist.dat;
type = imputation;
VARIABLE:
names = grad perfo0 perfo1 perfo3 perfo5;
usevariables = perfo0 perfo1 perfo3 perfo5;
missing = all(-99);
ANALYSIS:
estimator = ml;
MODEL:
[perfo0-perfo5] (mean0 mean1 mean3 mean5);
perfo0-perfo5 with perfo0-perfo5;
MODEL TEST:
mean0 = mean1; mean1 = mean3; mean3 = mean5;
```

UNSTANDARDIZED ESTIMATES

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
Means					
PERFO0	17.977	0.583	30.827	0.000	0.000
PERFO1	27.690	0.698	39.682	0.000	0.000
PERFO3	39.303	0.739	53.194	0.000	0.062
PERFO5	50.502	0.932	54.193	0.000	0.122
Variances					
PERFO0	69.377	6.869	10.100	0.000	0.000
PERFO1	99.333	9.835	10.100	0.000	0.000
PERFO3	104.535	10.772	9.704	0.000	0.076
PERFO5	155.723	16.276	9.568	0.000	0.102

MODEL TEST (WALD STATISTIC)

The MODEL TEST command specifies constraints that are consistent with a hypothesis of no change (mean0 = mean1, mean1 = mean3, mean3 = mean5)

df = 3 because the Wald test posits three constraints

MODEL TEST OUTPUT

The significant chi-square, $\chi^2(3)= 3101.989$, indicates that the data are inconsistent with the null hypothesis of no change

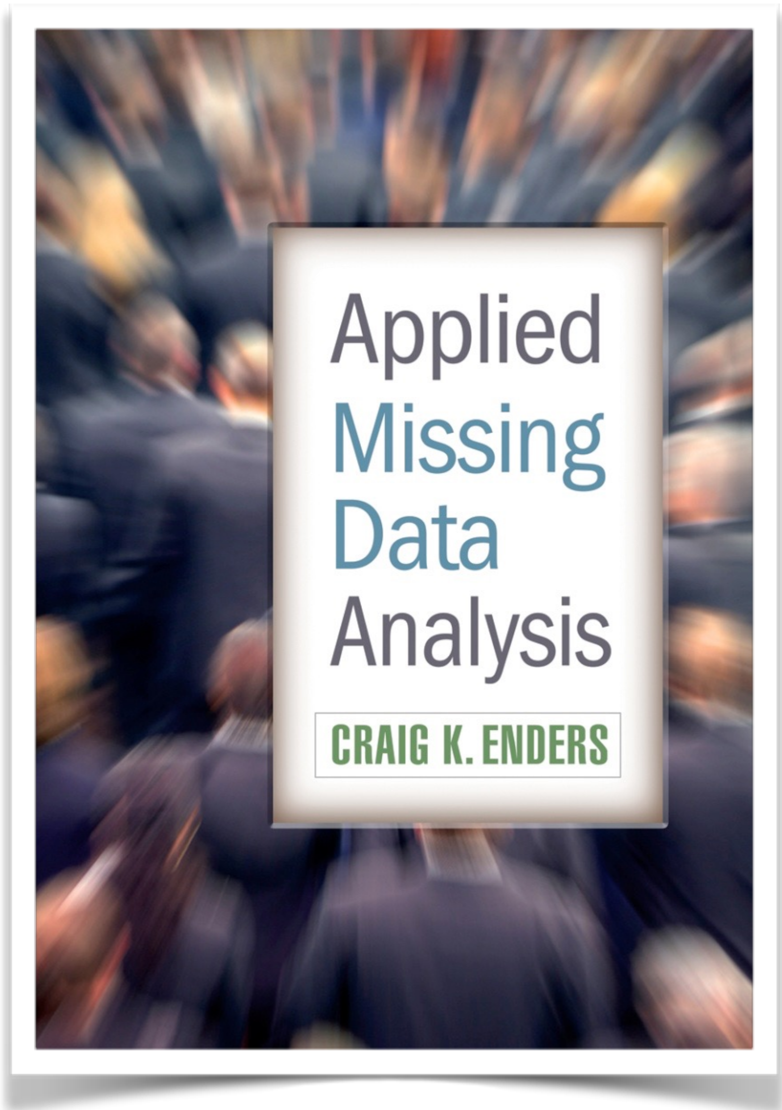
```
Wald Test of Parameter Constraints

          Value                               3101.989
          Degrees of Freedom                         3
          P-Value                               0.0000
```

ADDITIONAL RESOURCES

www.appliedmissingdata.com

Data sets and Mplus program files from the book

Many additional data sets and Mplus scripts



QUESTIONS?

THANK YOU!