

***Day 1, Lecture 1:
Introduction to Dealing
with Incomplete
Longitudinal Data***

John J. McArdle

Longitudinal Research Institute

CALDAR, August 2010

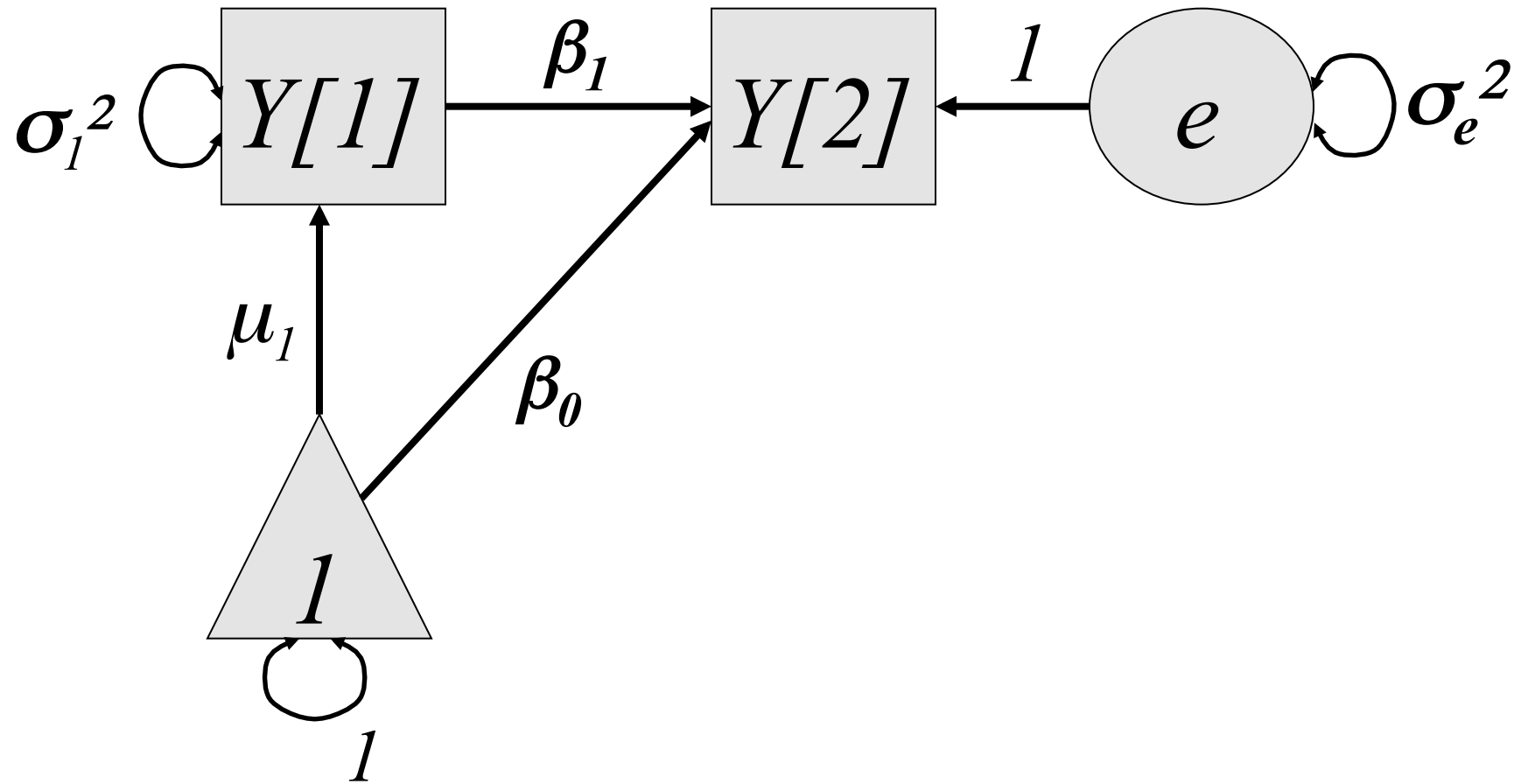
Overview

1. Common problems of Incomplete data
2. Defining Patterns of Incomplete Data
3. Simple methods for Incomplete data
4. Multiple Imputation (MI) techniques
5. Summary & Discussion

Common problems of incomplete data

- Incomplete or “missing” data are a common problem apparent in many areas of science, and the problems lead to biases and incorrect answers even in simple analyses.
- Most classical strategies rely on avoiding this problem to begin with -- get complete data on everyone (e.g., Klienbaum et al, 1998). More contemporary strategies are based on the concept of “model-based replacement.”
- A variety of sensible strategies for dealing with incomplete data now exist so there is little reason to stick with “complete cases only” analyses.

*Model A: An auto-regression model
for two repeated measures*



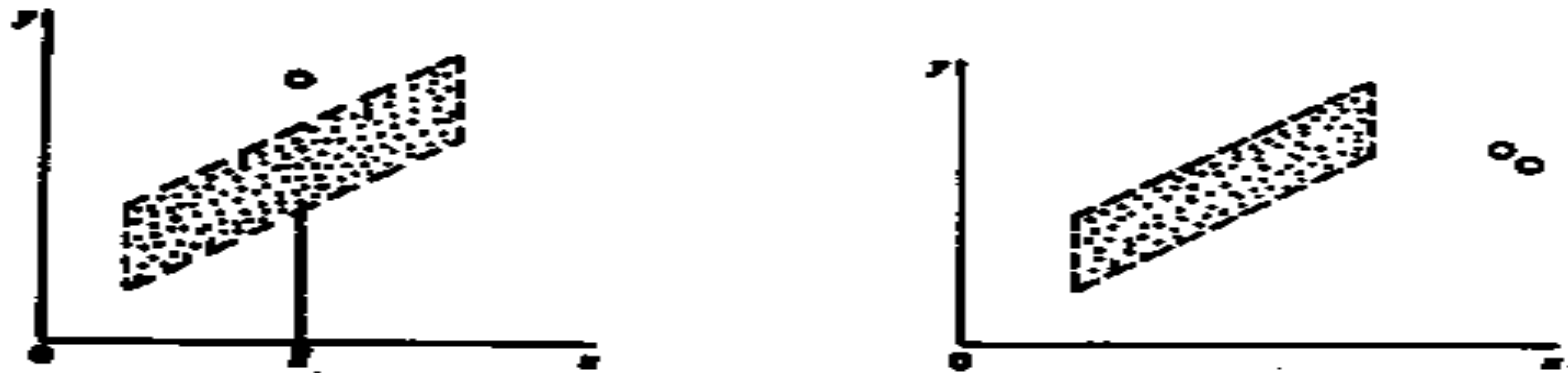


Figure 2a: Influential Observation in Regression
 (from Belsey, Kuh & Welsch, 1980, p. 3)

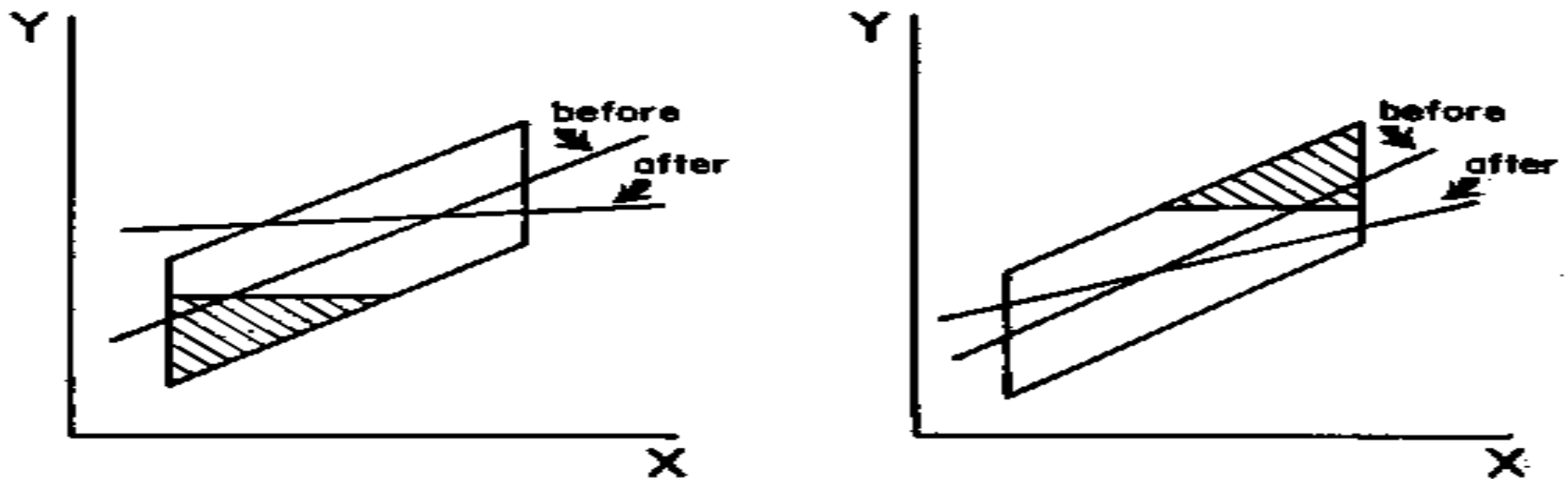


Figure 2b: Sample Selection Bias in Regression
 (from Berk, 1983, p.389)

Figure 2: Some Sampling Effects on Regression Estimates

Incomplete data regression analysis

- The main goal is to get accurate answers even though some of the data (X, Y) are incomplete
- We hope to counteract some confounds in *subject sampling* in order to make a *more appropriate inference* to a population of interest
- Since 1975, a host of new methods have been developed to deal with these problems; Some of the methods are very simple, but others are not.
- The choice among these techniques is not so easy, so multiple methods are often examined.

Standard statistical corrections are OK!

- There are a variety of simple statistical techniques that should not be overlooked:
- *Deletion* -- either Casewise and Pairwise adjustment of the summary statistics
- *Weighting* -- using sample weights to adjust for any known sample biases
- *Imputation* -- using sample information to adjust for “missing” data, including using the mean
- *Multiple Imputation* – creating multiple estimates of the incomplete data to obtain a more realistic view of data.
- *Likelihood-Based Estimation* – using all available data to estimate the model parameters.

- The first 3 techniques should be used as the baseline against which to judge any newer methods (i.e., MI, FIML, combinations, etc.)

Judged by a theory of “incompleteness”

- Donald Rubin (1977-1987) wrote a series of theorems on “incompleteness” that have become influential in all modern approaches:
- **MCAR** or “Missing Completely at Random” is the standard assumption used when no info is available and “complete cases” are advocated.
- **NIM** or “Non-Ignorable” implies the data are not missing completely at random and the pattern of incompleteness is not understood or measured.
- **Most usefully** → **MAR** or “Missing at Random” implies the data are not MCAR but clues as to why this happens are embedded in the measured variables.

2. Defining Patterns of Incomplete Data

Defining patterns of “incompleteness”

- It is most important that the patterns of missing data be described as well as possible.
- This can be done using a number of simply descriptive displays for two groups -- i.e., define groups of people based on the available data.
- The goal of these analyses are to describe the differences between (1) the persons who have been measured completely and (2) those persons who have not
- Here we ask the question -- “How are the available data different than we expected from our sampling?”

Typical regression input scripts

(i.e., cesdplot.SAS)

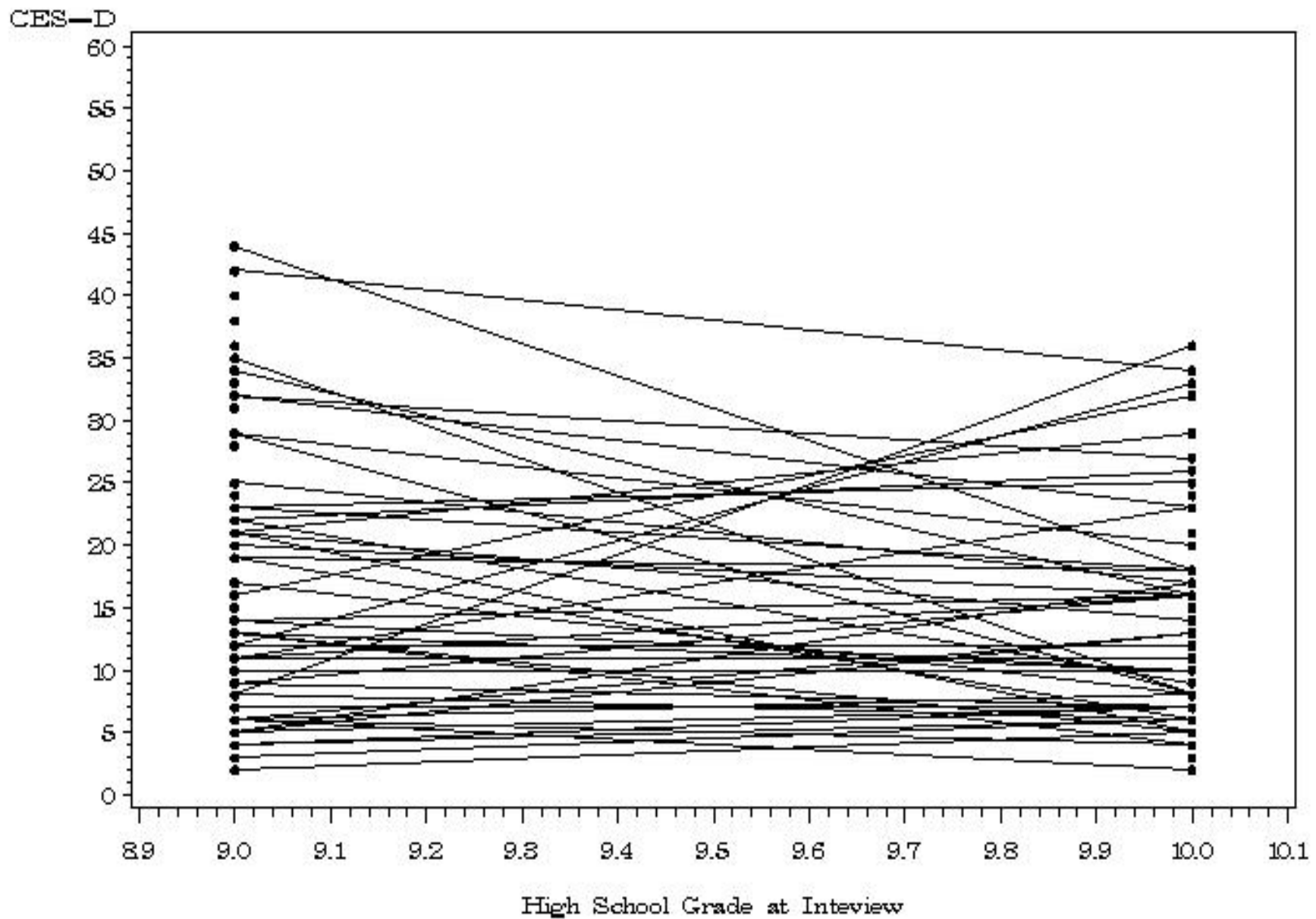
```
TITLE 'Auto-Regression models';  
PROC REG DATA=infile SIMPLE CORR;  
    model1: MODEL cesd_10 =cesd_09 / STB;  
    model2: MODEL cesd_10 =cesd_09 sex / STB;  
RUN;
```

```
TITLE 'Change-Score models;  
DATA cesd; SET cesd; change_1=cesd_10-cesd_09;  
  
PROC REG DATA=cesd SIMPLE CORR;  
    model1: MODEL change_1 = / STB;  
    model2: MODEL change_1 = sex / STB ;  
RUN;
```

Longitudinal CES-D Scores

2 Waves (Ninth & Tenth Grade)

Model 1b: Linear Time + Corr with Test



SAS input for the CESD analysis

(CESDMISS)

```
TITLE1 'Initial Incomplete Data Description for two variables';
```

```
DATA hawaii.cesd_new;  
    SET hawaii.cesd_raw;  
    miss_09 = 0; IF (cesd_09 = .) THEN miss_09=1;  
    miss_10 = 0; IF (cesd_10 = .) THEN miss_10=1;  
    cesdmiss = 0;  
    IF (cesd_09 = .) THEN cesdmiss=cesdmiss+1;  
    IF (cesd_10 = .) THEN cesdmiss=cesdmiss+10;  
    RUN;
```

```
PROC FREQ;  
    TABLE miss_09 miss_10 cesdmiss;  
    RUN;
```

```
PROC SORT;  
    BY cesdmiss;  
    RUN;
```

```
PROC CORR;  
    VAR cesd_09 cesd_10;  
    BY cesdmiss;  
    RUN;
```

Two-Occasion HHS Longitudinal Data

Variable	N	Mean	Std Dev	Minimum	Maximum
CESD_09	992	16.13609	10.90879	0	58.00000
CESD_10	672	14.32589	10.42149	0	56.00000
Change1	560	-1.27857	10.35096	-42.00000	38.00000

Pearson Correlation Coefficients Number of Observations

	CESD_09	CESD_10	Change_1
CESD_09	1.00000 992	0.52369 560	-0.49851 560
CESD_10	0.52369 560	1.00000 672	0.47744 560
Change_1	-0.49851 560	0.47744 560	1.00000 560

SAS output for the CESD analysis

The FREQ Procedure

miss_09	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	992	89.86	992	89.86
1	112	10.14	1104	100.00

miss_10	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	672	60.87	672	60.87
1	432	39.13	1104	100.00

cesdmiss	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	560	50.72	560	50.72
1	112	10.14	672	60.87
10	432	39.13	1104	100.00

SAS output for the CESD group MISS=0

----- cesdmiss=0 -----

The CORR Procedure

2 Variables: CESD9 CESD10

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CESD_09	560	15.55893	10.67608	8713	0	56.00000
CESD_10	560	14.28036	10.53294	7997	0	56.00000

Pearson Correlation Coefficients, N = 560

Prob > |r| under H0: Rho=0

	CESD_09	CESD_10
CESD_09	1.00000	0.52369 <.0001
CESD_10	0.52369 <.0001	1.00000

SAS output for the CESD group MISS=10

----- cesdmiss=10 -----

The CORR Procedure
2 Variables: CESD9 CESD10
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CESD_09	432	16.88426	11.17133	7294	0	58.00000
CESD_10	0

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	CESD_09	CESD_10
CESD_09	1.00000	.
		.
	432	0
CESD_10	.	.
	.	.
	0	0

Testing patterns of “incompleteness”

- Once the patterns of complete data groups are defined, a statistical model can be used with *variables that are measured on all groups*.
- This provides a data structure for a formal test of the question -- “Are the data missing completely at random?”
- This test can be formalized for one variable at a time using *ANOVA-based* tests for continuous variables OR
- This test can be formalized for many variables using *logit regression* with the data group patterns as the outcome and the observed as the IV.

SAS input script for the CESD analysis

(CESDMISS)

```
TITLE1 'Initial statistical models -- Is it likely to  
make a difference in the outcome?';
```

```
PROC REG;  
    MODEL cesd_09 = miss_10;  
    RUN;
```

```
PROC REG;  
    MODEL cesd_10 = miss_09;  
    RUN;
```

```
TITLE2 'Initial Selection Model';  
PROC LOGISTIC ORDER=DATA;  
    MODEL miss_10 = cesd_09 /  
        RSQUARE LACKFTS RISKLIMITS; RUN;
```

```
TITLE2 'More Complete Selection model';  
PROC LOGISTIC ORDER=DATA;  
    MODEL miss_10 = cesd_09 gpa_09 sex ethnicity /  
        RSQUARE LACKFTS RISKLIMITS; RUN;
```

SAS output for the MISS_10 Logit

Response Profile

Ordered Value	miss_10	Total Frequency
1	0	560
2	1	432

Probability modeled is miss10=0.

NOTE: 112 observations were deleted due to missing values for the response or explanatory variables.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
-2 Log L	1358.642	1355.052

R-Square 0.0036 Max-rescaled R-Square **0.0048**

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.4394	0.1149	14.6265	0.0001
CESD_09	1	-0.0111	0.00586	3.5861	0.0583

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
CESD_09	0.989	0.978 1.000

Useful to have a “selection” model

- It is useful to have a statistical model based on the variables that predict “missing-ness” at a later time (after Heckman, 1979).
- This provides a data structure for a formal test of the question -- “Are the data missing completely at random?” (using *logit regression*). This test should fail because most attrition is assumed to be non-random!
- From this perspective, it helps the researcher both (a) understand the reasons for non-random selection, and (b) highlights measured variables that can be used in subsequent “corrective” analyses (i.e., to meet the MAR assumptions). Of course, many variables may be needed.

3. Simple Methods for Dealing with Incomplete Data

Fix problems by deleting some data

- Purpose is to use available information in simple ways.
- “Casewise” deletion of cases which are not complete and using only complete in regression
- “Pairwise” deletion cases in forming statistics to be used in regression.
- Benefits include: (a) simplicity, (b) clarity, and (c) wide-spread usage
- Limits include: (a) loss of subjects, (b) increased standard errors, and (c) bias if not MCAR

REG output for the COMPLETE CASE

The REG Procedure

Model: MODEL1

Dependent Variable: CESD_10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	17008	17008	210.86	<.0001
Error	558	45009	80.66047		
Corrected Total	559	62017			

Root MSE	8.98112	R-Square	0.2743
Dependent Mean	14.28036	Adj R-Sq	0.2730
Coeff Var	62.89140		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	6.24150	0.67120	9.30	<.0001	0
CESD_09	1	0.51667	0.03558	14.52	<.0001	0.52369

Replacing Incomplete Data with Means

- Purpose is to use the available information in simple ways, and this is a very simple idea.
- “Mean substitution” is based on (1) calculating summary statistics using the complete data, (2) plugging the mean values back into the data in the positions where the data are incomplete, and (3) running the regression on all persons.
- Benefits include: (a) simplicity, (b) clarity, (c) widespread usage, and (d) use of all data.
- Limits include: (a) loss of true means, (b) incorrect standard errors and DFs, and (c) bias.

SAS input script for the MEANS analysis (CESDMISS)

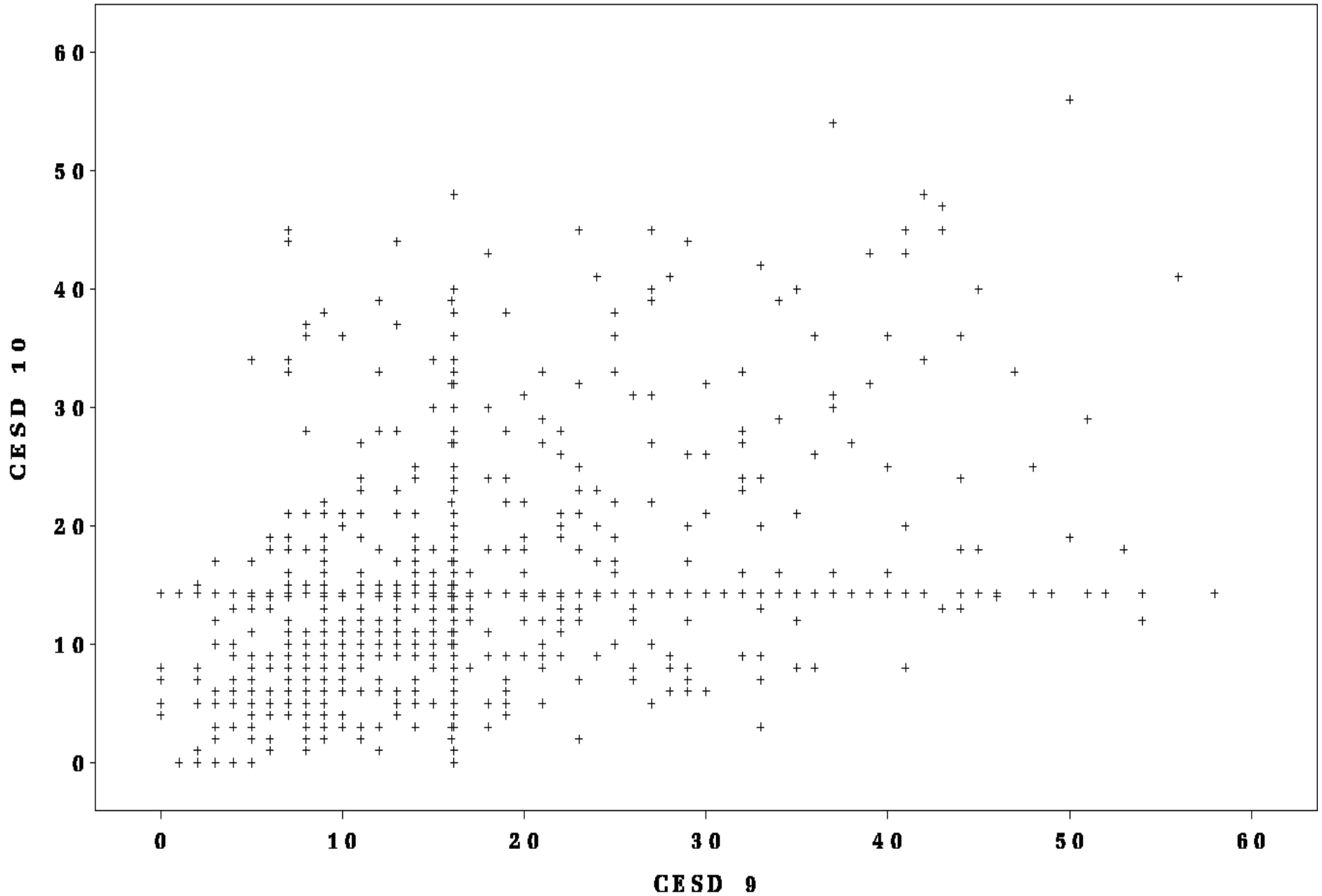
```
TITLE1 'Type 1: Mean Substitution';
DATA hawaii.cesd_new;
    SET hawaii.cesd_raw;
    IF (cesd_09 = .) THEN cesd_09 = 16.13609;
    IF (cesd_10 = .) THEN cesd_10 = 14.32589;
    Change_1 = cesd_10 - cesd_9;
    RUN;

PROC CORR NOPROB
    DATA=hawaii.cesd_new;
    VAR cesd_09 cesd_10 Change_1;
    RUN;

TITLE2 'Auto-Regression Model of Change';
PROC GPLOT
    DATA=hawaii.cesd_new;
    PLOT cesd_10 * cesd_09;
    RUN;

PROC REG
    DATA=hawaii.cesd_new;
    MODEL cesd_10 = cesd_09 / STB;
    RUN;
```

Type 1: Resulting data from Mean Substitution



REG output for the MEANS analysis

The REG Procedure

Model: MODEL1

Dependent Variable: CESD_10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9197.32272	9197.32272	159.17	<.0001
Error	1102	63678	57.78431		
Corrected Total	1103	72876			

Root MSE	7.60160	R-Square	0.1262
Dependent Mean	14.32589	Adj R-Sq	0.1254
Coeff Var	53.06196		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	9.81964	0.42417	23.15	<.0001	0
CESD_09	1	0.27927	0.02214	12.62	<.0001	0.35525

Imputing data using Regression

“Substitution” or “Imputation” by regression methods is fairly easy, and it often has better properties.

Step 1: Get weights from Complete Case Model for subjects with data

$$Y_n = \beta_0 + \beta_1 X_n + e_n$$

Step 2: Apply weights to subjects with incomplete Y to estimate a new “fixed” value for the outcome

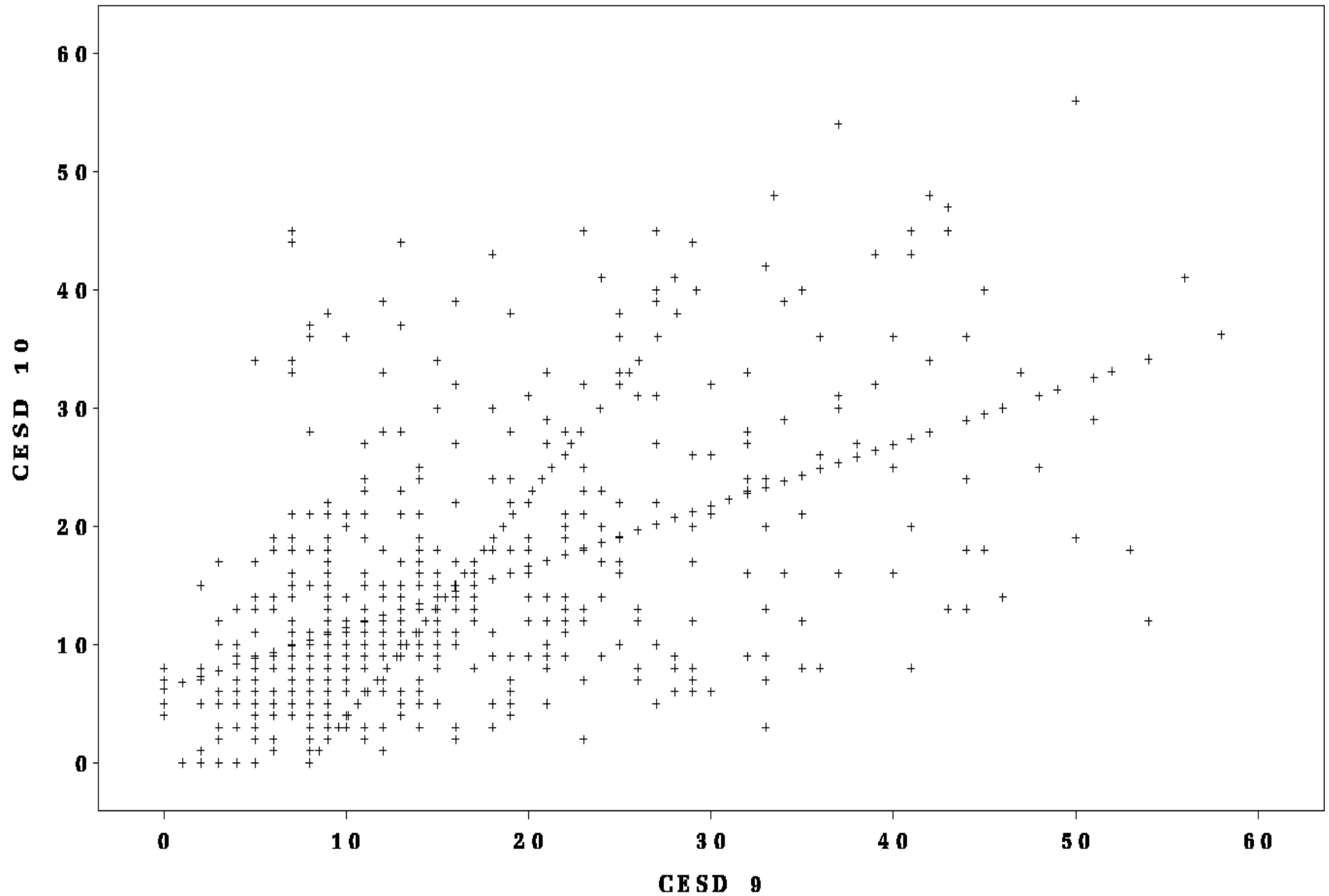
$$\textit{Estimated } \{y_m\} = \beta_0 + \beta_1 X_n$$

Step 3: Run further analyses on new outcomes **[Y: y]**

SAS input script for the CESD analysis

```
TITLE1 'Type 2: Regression Fixed Score Imputation';
PROC REG
    DATA=hawaii.cesd_raw;
    MODEL cesd10 = cesd9 ;
    MODEL cesd9 = cesd10 ;
    RUN;
DATA hawaii.cesd_new;
    SET hawaii.cesd_raw;
    b0 = 6.24150; b1 = 0.51667;
    IF (cesd_10 = .) THEN cesd_10 = b0 + b1 * cesd_09;
    a0 = 7.97877; a1 =0.53081;
    IF (cesd_09 = .) THEN cesd_09 = a0 + a1 * cesd_10;
    Change_1 = cesd_10 - cesd_09;
    RUN;
PROC CORR NOPROB
    DATA=hawaii.cesd_new;
    VAR cesd_09 cesd_10 Change_1;
    RUN;
TITLE2 'Auto-Regression Model of Change';
PROC GPLOT
    DATA=hawaii.cesd_new;
    PLOT cesd_10 * cesd_09;
    RUN;
PROC REG
    DATA=hawaii.cesd_new;
    MODEL cesd+10 = cesd_09 / STB;
    RUN;
```

Type 2: Data using Regression Imputation



REG output for the REG impute

The REG Procedure

Model: MODEL1

Dependent Variable: CESD_10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	36757	36757	800.77	<.0001
Error	1102	50584	45.90229		
Corrected Total	1103	87342			

Root MSE	6.77512	R-Square	0.4208
Dependent Mean	14.57601	Adj R-Sq	0.4203
Coef Var	46.48132		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	5.70683	0.37391	15.26	<.0001	0
CESD_09	1	0.55115	0.01948	28.30	<.0001	0.64873

Using Random Error Score imputation

- The initial regression model also gives us an estimate of the error variance (*SEE*), so this can be used as well.

- Any vector with missing data is filled-in using these parameters -- but here we also add a random error

$$\{e_m\} = \text{RANNOR}(r) * \text{SEE}$$

- For each *person*, any imputation is based on both a “fixed” part and a “random” part

$$\text{Estimated } \{y_m\} = \beta_0 + \beta_1 X_m + \{e_m\}$$

- The net result of analyzing [*Y: y*] is a “less optimistic” but “more realistic” view of the model parameters.

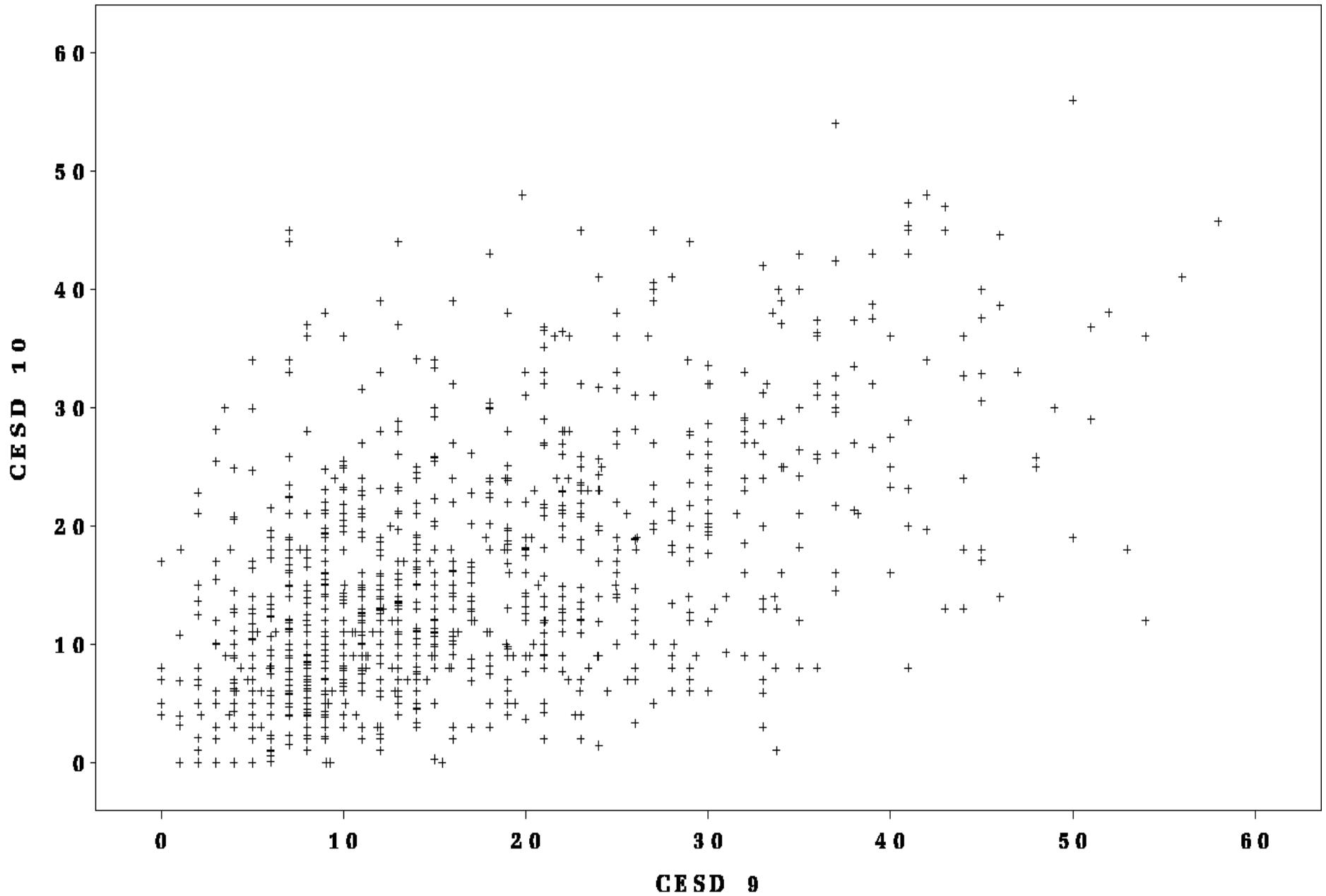
SAS input script for the RANREG

```
TITLE1 'Type 3: Regression random Score Imputation';
PROC REG
    DATA=hawaii.cesd_new;
    MODEL cesd_10 = cesd_09 ;
    MODEL cesd_09 = cesd_10 ;
    RUN;
DATA hawaii.cesd_new;
    SET hawaii.cesd_raw; seed=20020117;
    b0 = 6.24150; b1 = 0.51667; e_10 = RANNOR(seed)*8.98112;
    IF (cesd_10 = .) THEN cesd_10 = b0 + b1 * cesd_9 + e_10;

    a0 = 7.97877; a1 =0.53081; e_09 = RANNOR(seed)*9.10317;
    IF (cesd_09 = .) THEN cesd_09 = a0 + a1 * cesd_10 + e9;
    Change_1 = cesd_10 - cesd_09;
    RUN;

PROC CORR NOPROB
    DATA=hawaii.cesd_new;
    VAR cesd_09 cesd_10 Change_1 e_9 e_10;
    RUN;
TITLE2 'Auto-Regression Model of Change';
PROC GPLOT
    DATA=hawaii.cesd_new;
    PLOT cesd_10 * cesd_09;
    RUN;
PROC REG
    DATA=hawaii.cesd_new;
    MODEL cesd_10 = cesd_09 / STB;
    RUN;
```

Type 3: Using Regression Imputation + Error



REG output of Random Reg imputation

The REG Procedure

Model: MODEL1
Dependent Variable: CESD_10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	37863	37863	487.74	<.0001
Error	1102	85548	77.62937		
Corrected Total	1103	123411			

Root MSE	8.81075	R-Square	0.3068
Dependent Mean	14.86690	Adj R-Sq	0.3062
Coeff Var	59.26422		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	6.06412	0.47874	12.67	<.0001	0
CESD_09	1	0.54198	0.02454	22.08	<.0001	0.55390

Alternative RANREG with new SEED

The REG Procedure
Model: MODEL1
Dependent Variable: CESD_10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	34241	34241	441.78	<.0001
Error	1102	85414	77.50779		
Corrected Total	1103	119655			

Root MSE	8.80385	R-Square	0.2862
Dependent Mean	14.71906	Adj R-Sq	0.2855
Coeff Var	59.81259		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	6.47956	0.47316	13.69	<.0001	0
CESD_09	1	0.51024	0.02428	21.02	<.0001	0.53495

Imputing data in general

- Purpose is to counteract some known confounds in data *collection* in order to make a *more appropriate inference* to a population of interest.
- Benefits include: clarity of both (a) sample statistics and (b) population of interest.
- Limits include: (a) often hard to define population mechanisms, and (b) leads to increased standard errors and possible increase in bias (opposite of goal).

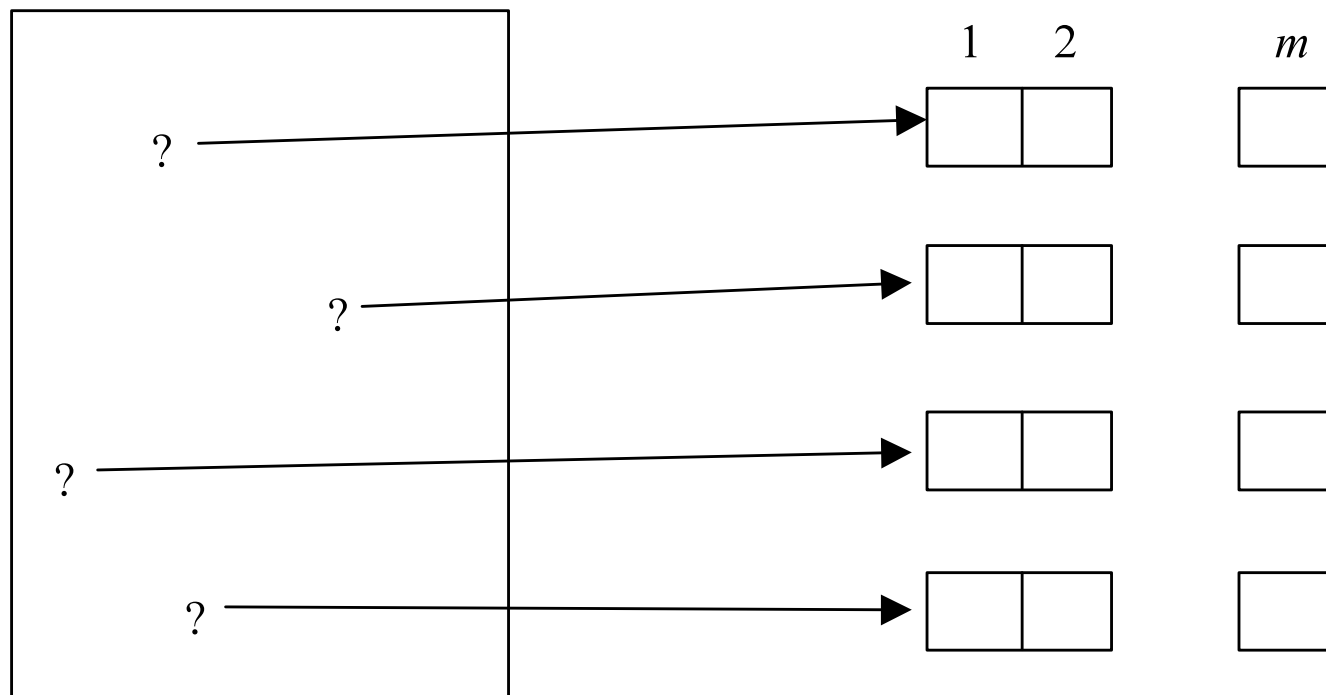
***4. Multiple
Imputation (MI)
in Repeated
Measures Analyses***

The Multiple Imputation approach

- ML estimation using all available data:
“[t]he technique that replaces each missing or deficient value with two or more acceptable values representing a distribution of possibilities” (Rubin, 1987)
- MI estimates parameters directly from the data and results in both *unbiased parameter estimates* and *smaller standard errors*. Statistical uncertainty information is obtained for clear statistical inferences
- Any vector with missing data is filled-in using these parameters -- but here we also add a random error $RANNOR(r) * SEE$. For each n , imputation is repeated R times resulting in $R * N$ vectors for further analysis

Multiple Imputation

Matrix of Multivariate Data with Missing Values
(Schafer & Olsen, 1998)



“Multiple Imputation” methods

- *Imputation* implies the use of a statistical model to “fill-in” any missing data.
- Usual imputation gives unbiased parameter estimates with smaller standard errors
- More accurate standard errors can be obtained by adding random components and repeating the imputation *multiple* times
- *References*: Rubin ‘87; Schaffer, ‘97, any recent *PMK* or *JASA*.

SAS input script for the CESD analysis

```
TITLE1 'Type 4: Multiple Imputation Approach';
TITLE2 'Creating imputed scores';
PROC MI
    DATA=hawaii.cesd_raw OUT=temp_mi
    NIMPUTE=10; /* monotone method=reg; */;
    VAR cesd_09 cesd_10;
    RUN;
PROC CORR
    DATA=temp_mi NOPROB NOMISS;
    VAR cesd_09 cesd_10 Change_1;
    BY _Imputation_;
    RUN;
TITLE2 'Auto-Regression Model of Change';
PROC REG
    DATA=temp_mi OUTEST=temp_reg COVOUT NOPRINT;
    MODEL cesd_10 = cesd_09 / STB;
    BY _Imputation_;
    RUN;
PROC PRINT
    DATA=temp_reg;
    RUN;
PROC MIANALYZE
    DATA=temp_reg;
    VAR Intercept cesd_09;
    RUN;
```

Initial SAS output for the MI analysis

The MI Procedure

Model Information

```

Data Set          HAWAII.CESD_RAW
Method           MCMC
Multiple Imputation Chain  Single Chain
Initial Estimates for MCMC  EM Posterior Mode
Number of Imputations      10
Number of Burn-in Iterations 200
Number of Iterations      100
Seed for random number generator 44925
  
```

Missing Data Patterns

-----Group Means-----

Group	CESD_90	CESD_10	Freq	Percent	CESD_09	CESD_10
1	X	X	560	50.72	15.558929	14.280357
2	X	.	432	39.13	16.884259	.
3	.	X	112	10.14	.	14.553571

EM (Posterior Mode) Estimates

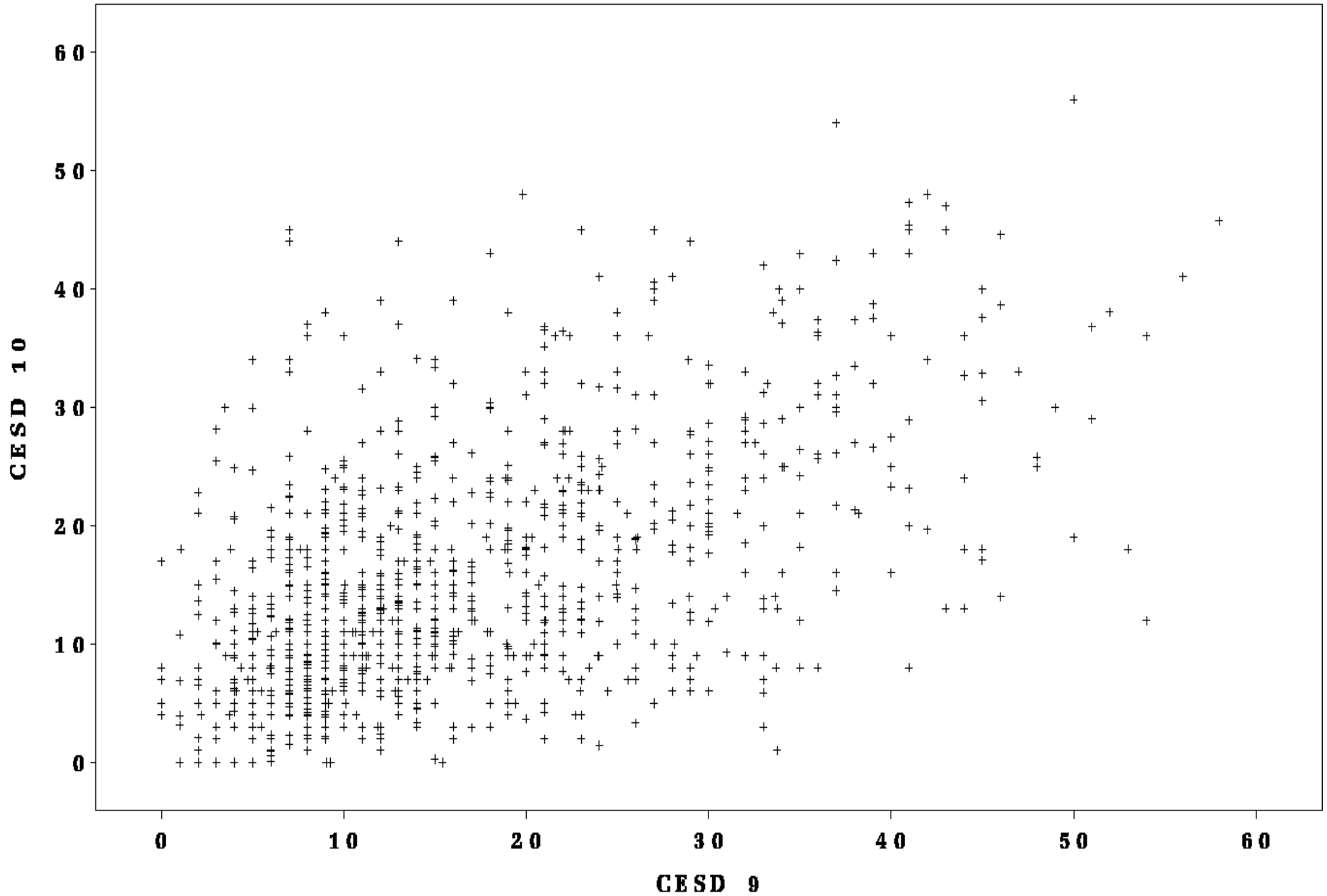
TYPE	_NAME_	CESD_09	CESD_10
MEAN		16.135037	14.570564
COV	CESD_09	118.105372	60.005642
COV	CESD_10	60.005642	109.176926

Multiple Imputation Variance Information

-----Variance-----

Variable	Between	Within	Total	DF
CESD_09	0.007408	0.106425	0.114574	649.43
CESD_10	0.031131	0.098653	0.132896	116.27

Type 4.1: Using Regression Imputation + Error1



Results from Imputation #1

----- Imputation Number=1 -----

The REG Procedure
 Model: MODEL1
 Dependent Variable: CESD_10

Number of Observations Read 1104
 Number of Observations Used 1104

Analysis of Variance

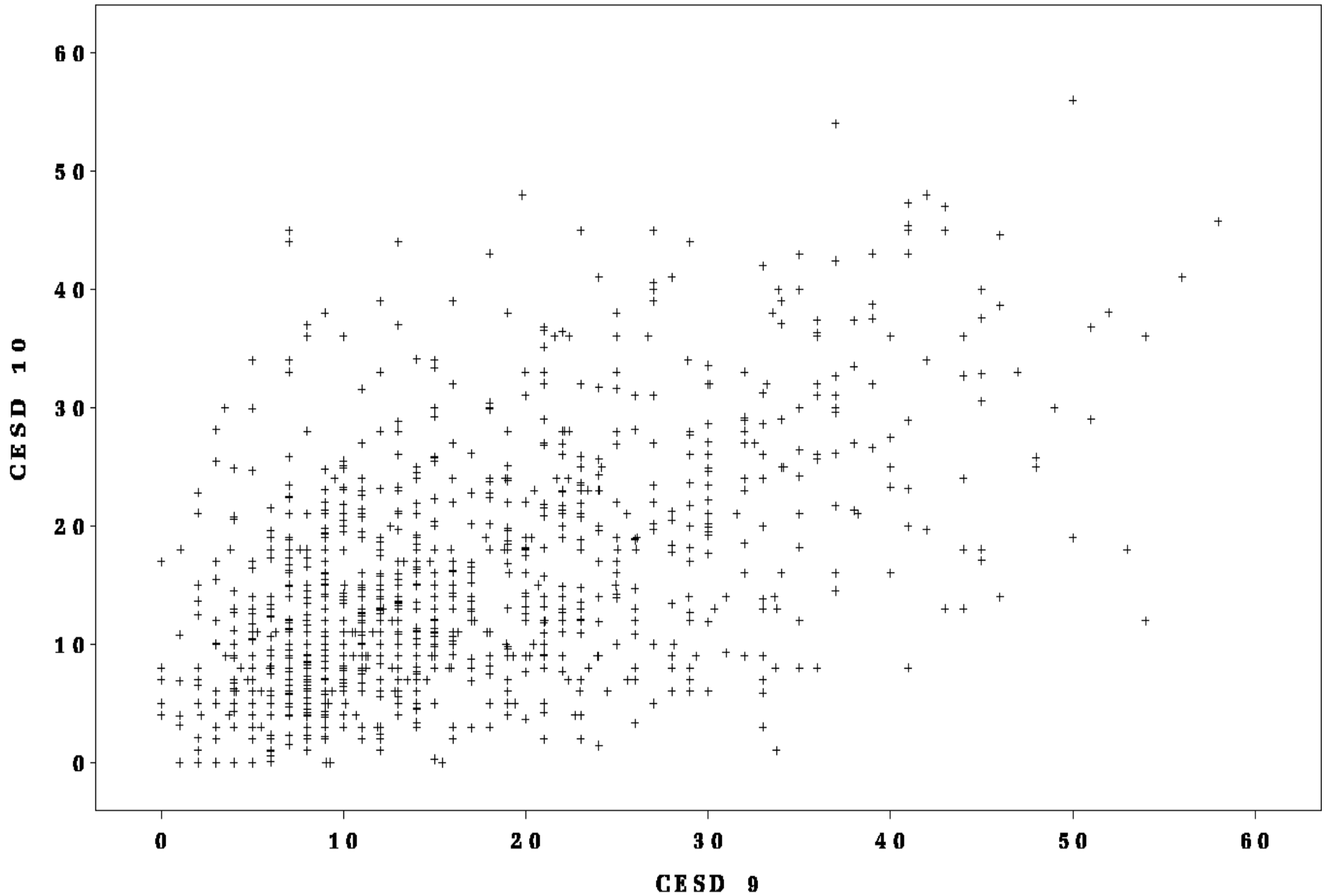
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	27263	27263	332.39	<.0001
Error	1102	90385	82.01949		
Corrected Total	1103	117648			

Root MSE 9.05646 R-Square 0.2317
 Dependent Mean 14.77267 Adj R-Sq 0.2310
 Coeff Var 61.30551

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	7.34549	0.49015	14.99	<.0001	0
CESD_09	1	0.45584	0.02500	18.23	<.0001	0.48139

Type 4.2: Using Regression Imputation + Error2



Results from Imputation #2

----- Imputation Number=2 -----

The REG Procedure
Model: MODEL1
Dependent Variable: CESD_10

Number of Observations Read 1104
Number of Observations Used 1104

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	39524	39524	514.46	<.0001
Error	1102	84664	76.82741		
Corrected Total	1103	124188			

Root MSE	8.76512	R-Square	0.3183
Dependent Mean	14.67809	Adj R-Sq	0.3176
Coeff Var	59.71571		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate
Intercept	1	5.87412	0.46931	12.52	<.0001	0
CESD_09	1	0.54948	0.02423	22.68	<.0001	0.56415

REG results for each imputation

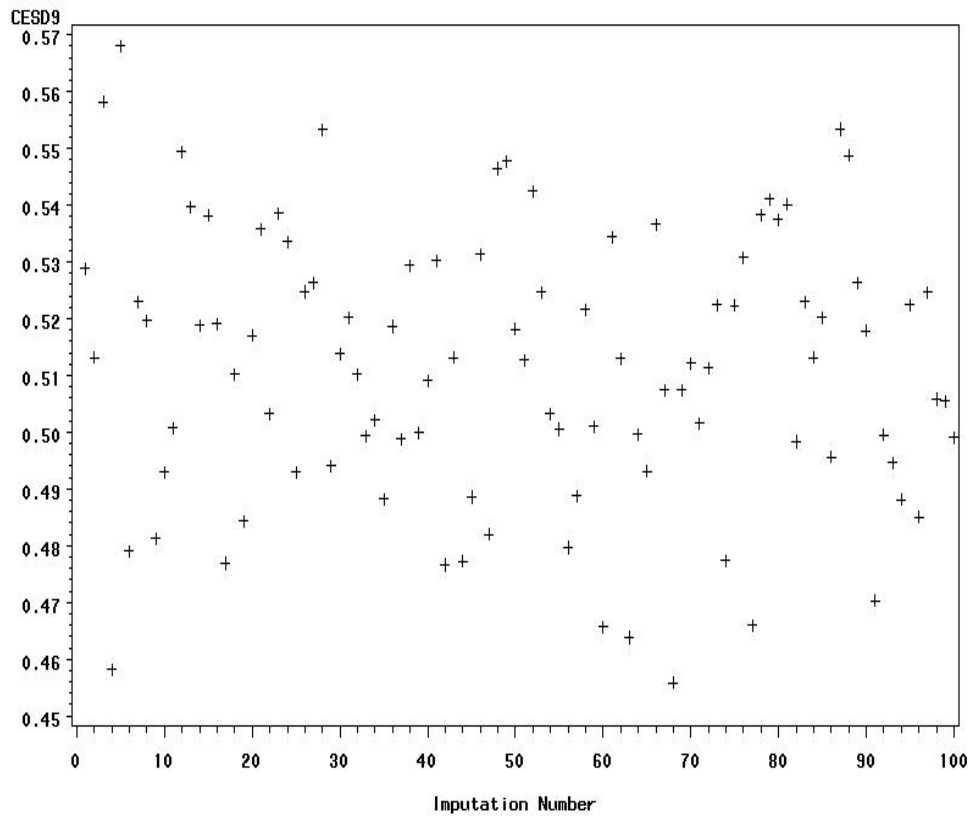
<u>_Imputation_</u>	<u>Intercept</u>	<u>CESD9</u>	<u>_RMSE_</u>
1	5.92245	0.52082	8.69681
2	6.60343	0.51281	8.84086
3	6.67435	0.50953	8.97294
4	6.42215	0.52255	8.98625
5	6.34208	0.52561	8.99383
6	5.57797	0.56303	8.71061
7	6.98762	0.47630	8.90088
8	5.96554	0.54462	8.74782
9	5.65779	0.53433	9.01019
10	6.03516	0.52856	8.93192

Basic Statistical Measures (based on **NI=100**)

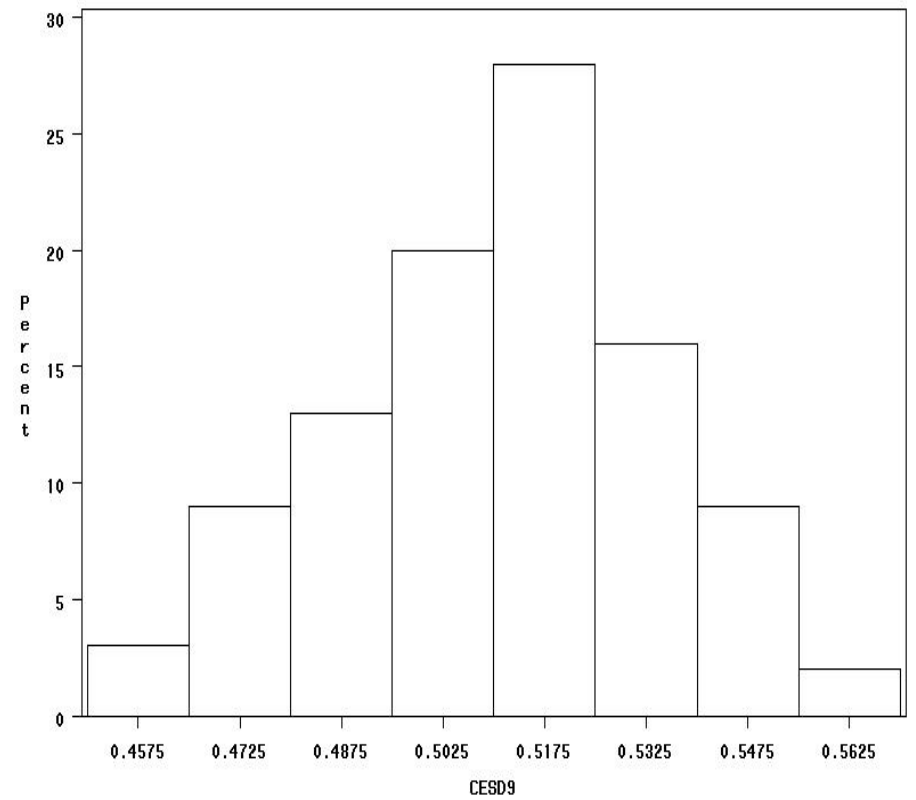
100% Max	0.574738
99%	0.567578
95%	0.545667
90%	0.539644
75% Q3	0.521607
50% Median	0.504566
25% Q1	0.485808
10%	0.472108
5%	0.465980
1%	0.439767
0% Min	0.436971

Results for β_1 -weight from MI=100

Type 4: Multiple Imputation Approach
Auto-Regression Model of Change



Type 4: Multiple Imputation Approach
Auto-Regression Model of Change



Rubin's (1987) rules for Combining MI results

- Degrees of freedom vary from $(m-1)$ to ∞ , depending on relative sizes of $(1 + m^{-1})B$ and \bar{U}
- Relative increase in variance due to non-response is estimated by

$$r = \frac{(1 + m^{-1})B}{\bar{U}}$$

- Fraction of missing information is estimated by

$$\lambda = \frac{r+2/(v+3)}{r+1}$$

- Review in Schafer (1997, ch. 4)

SAS-MI output for the CESD analysis

The MIANALYZE Procedure
Multiple Imputation Variance Information

Parameter	-----Variance-----			DF
	Between	Within	Total	
Intercept	0.122700	0.228241	0.363211	65.176
cesd_09	0.000682	0.000605	0.001355	29.36

Multiple Imputation Variance Information

Parameter	Relative	Fraction
	Increase	Missing
Parameter	in Variance	Information
Intercept	0.591346	0.390036
cesd_09	1.240441	0.581245

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits	DF
Intercept	6.271201	0.602670	5.067648 7.474754	65.176
cesd_09	0.509955	0.036810	0.434711 0.585199	29.36

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum
Intercept	5.793531	6.945164
cesd_09	0.450914	0.550327

Multiple Imputation Parameter Estimates

t for H0:

Parameter	Theta0	Parameter=Theta0	Pr > t
Intercept	0	10.41	<.0001
cesd_09	0	13.85	<.0001

Multiple Imputation Programs

- SAS PROC MI and PROC MIANALYZE
- <http://www.multiple-imputation.com/>
- NORM, CAT, MIX and PAN
- MICE
- AMELIA
- SOLAS for Missing Data Analysis 2.0
- Hmisc library (S-plus)

5. Summary & Discussion

Basic cautions about using incomplete data modeling

- It is difficult to know with any certainty that the “unbiased” answer is correct. In most cases, we would rather have all the data (including all the latent scores!). So we do not pretend these results are as good as having all the people measured all the time.
- But this is a practical solution to some big some practical problems (e.g., fatigue, attrition) that are otherwise are handled as complete case analyses with large potential for biased answers .
- There are a variety of alternative statistical techniques that should be tried (e.g., multiple imputation) to see if any empirical differences emerge.

Incomplete or ALL data modeling

- Multiple Imputation has advantages when only some data are missing (<25%) and is simple to describe.
- Likelihood methods (to be discussed) have advantages when a great deal of data are missing (>50%) and the mechanisms are clear.
- Pattern-Mixture models (Little, 1995) can be added to either approach, and these may help with interpretation.
- It is a good idea to use many of these methods and see how or when any key results are different.
- Many of the newest programs automatically use the same techniques.
- This approach also allows “intentionally incomplete” data layouts (to be discussed).

“Drop or Mask some data,” just to see if it works for you?”

```
TITLE1 'Creating Incomplete Data from Complete Data';
DATA hawaii.cesd_less;
  SET hawaii.cesd_new;
  Change_1 = cesd_10 - cesd_09;
  /* drop all incomplete case */
  IF (cesd_09 = . OR cesd_10 =.) THEN DELETE;
  /* also drop some of the complete cases */
  IF (cesd_09 GT 20) THEN DELETE;
RUN;
```

2 Variables: CESD9 CESD10						
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CESD_09	416	10.27163	4.42890	4273	0	20.00000
CESD_10	416	11.48077	8.17870	4776	0	45.00000

Pearson Correlation Coefficients, N = 416
Prob > |r| under H0: Rho=0

	CESD_09	CESD_10
CESD_09	1.00000	0.30565 <.0001
CESD_10	0.30565 <.0001	1.00000

Does the MI procedure work when we know the answer?

Number of Imputations 100

Multiple Imputation Variance Information

Parameter	-----Variance-----			DF
	Between	Within	Total	
Intercept	0.138642	0.230252	0.370280	692.25
cesd_09	0.000572	0.000608	0.001186	417.17

Multiple Imputation Variance Information

Parameter	Relative	Fraction	Relative
	Increase	Missing	Efficiency
Intercept	0.608152	0.379957	0.996215
cesd_09	0.949875	0.489588	0.995128

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits	DF
Intercept	6.370238	0.608506	5.175498 7.564977	692.25
cesd_09	0.511271	0.034441	0.443572 0.578970	417.17

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum
Intercept	5.695004	7.345490
cesd_09	0.455841	0.568115

Multiple Imputation Parameter Estimates

Parameter	t for H0:		
	Theta0	Parameter=Theta0	Pr > t
Intercept	0	10.47	<.0001
cesd_09	0	14.84	<.0001

Do incomplete data techniques work?

- Any incomplete data technique is based on a set of assumptions -- models of behaviors -- that may or may not be present in any specific dataset.
- This means that specific studies of a specific dataset are needed to determine the utility of any incomplete data technique – *so be cautious!*
- Studies of the number of MC multiple imputations required is easy and can be done without computational difficulty (e.g., Bell, 1954).
- *Question for the analyst – Do you analyze the **complete cases** or do you analyze the **complete data** ???*

References

- Little, R.J.A. (1995) Modeling the dropout mechanism in repeated-measures studies. *Jour. Amer. Stat. Ass.*, 90, 1112-1121.
- Little, R.J. & Rubin, D.B. (2002). *Statistical analysis with missing data. Second edition*. New York: John Wiley & Sons.
- Marini, M.M., Olsen, A.R. & Rubin, D.B. (1980). Maximum-likelihood estimation in panel studies with missing data. *Sociological Methodology*, 11, 314-357.
- Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
- Rubin, D.B. (1967). Inference and missing data. *Biometrika*, 63 (3), 581-592.
- Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.
- Schafer, J.L. (1999). *Multiple imputation: a primer*. *Statistical Methods in Medical Research*, in press.