Multiple Imputation for Categorical Variables

Mixtures of Categorical and Continuous Variables

Multiple imputation is ideally suited for mixtures of categorical and continuous incomplete variables

Maximum likelihood estimation is far less flexible in this regard because it generally assumes multivariate normality

Nominal and ordinal variables can be imputed in a latent variable framework or with a logistic regression model

Complete Categorical Variables

Complete categorical variables can serve as predictors in the imputation models

Nominal variables must be first converted to dummy codes (Blimp does this automatically)

Ordinal variables can be left as-is or dummy coded

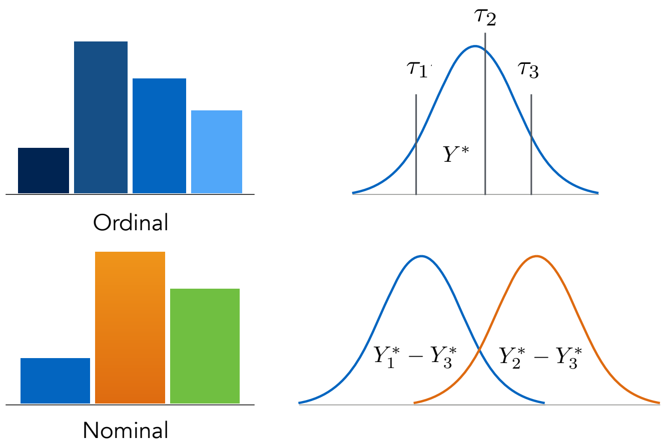
Latent Variable Formulation

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

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

The latent variable distribution for each case is centered at a predicted value and has a residual variance of one

Latent Variable Transformations



Motivating Example

Number of years smoking and number of cigarettes smoked

Participants are classified as 0 = light smokers or 1 = heavy smokers

A binary variable is a special case of an ordinal variable

Heavy Cigs	Efficacy
0	7
NA	11
0	16
0	21
0	17
0	10
0	13
NA	10
NA	11
0	13
1	11
0	16
NA	10
1	9
NA	5
0	7
1	10
0	9
0	7
0	6

Bivariate Example

The substantive analysis is a simple regression model, where the covariate is incomplete

$$Y = \beta_0 + \beta_1 (X) + e$$

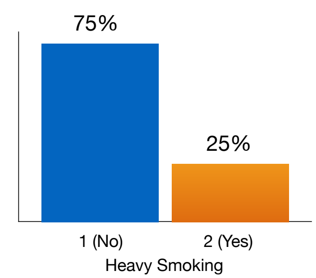
For example, efficacy to quit predicted by a heavy smoking dummy variable, which is incomplete

$$Efficacy = \beta_0 + \beta_1 (Heavy\,Cigs) + e$$

Binary Variable

The marginal distribution (ignoring covariates) has 25% heavy smokers

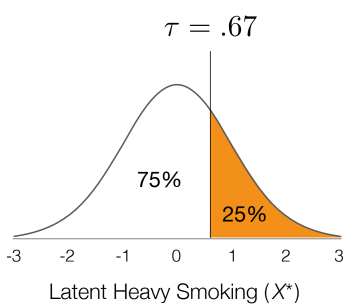
The probability of heavy smoking varies across values of predictors such as years smoking



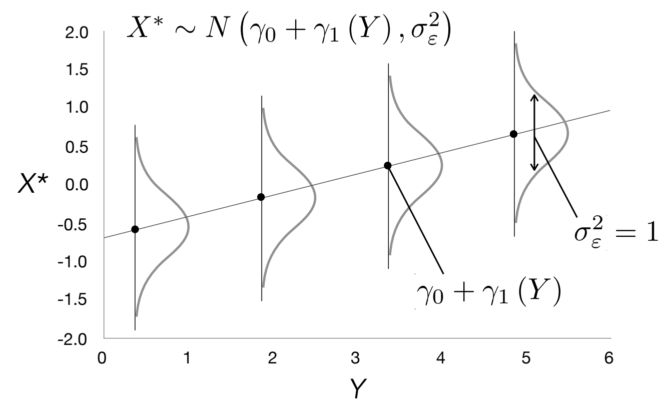
Latent Variable Distribution

The propensity for heavy smoking can be viewed as an underlying normal latent variable

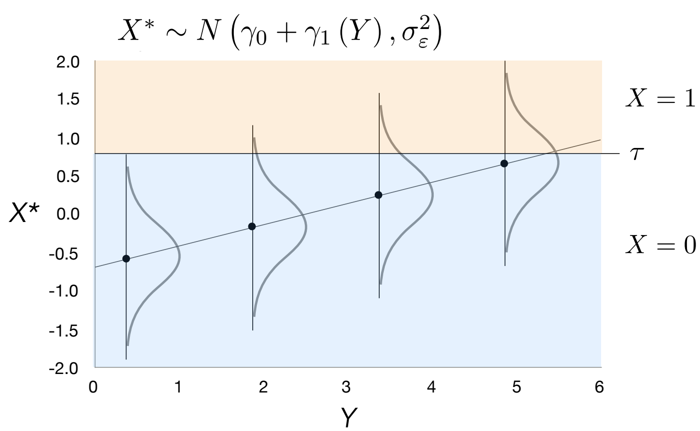
A threshold parameter (z-score) separates the upper 25% of the distribution from the lower 75%



Latent Variable Distribution



Latent Variable Threshold



Latent Variable Scores are Missing Data

Latent variable scores are missing data, and they are missing for the entire sample

MCMC draws latent variable scores for the entire sample, after which it uses the continuous values as real data and updates the regression coefficients using MCMC for linear regression

Discrete imputes are generated by comparing the latent scores to the threshold parameter

Sampling Latent Variable Scores

The threshold parameter divides the latent distributions into two segments

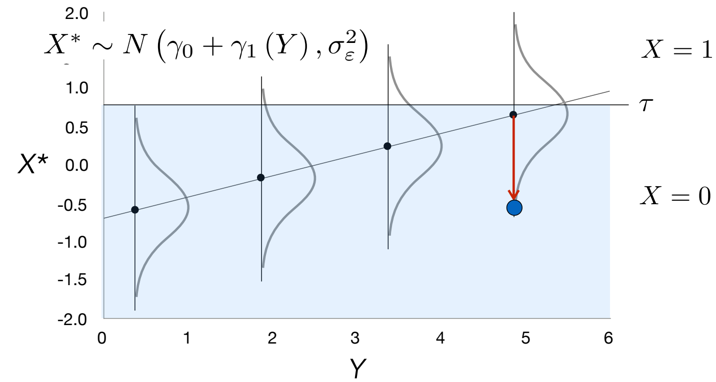
When smoking status is observed, the latent variable score must be constrained to a particular region of the distribution (e.g., heavy smokers must have latent scores above the threshold)

The latent scores for incomplete cases can fall anywhere in the distribution

13

Sample Latent Scores: Heavy Smoking = 0

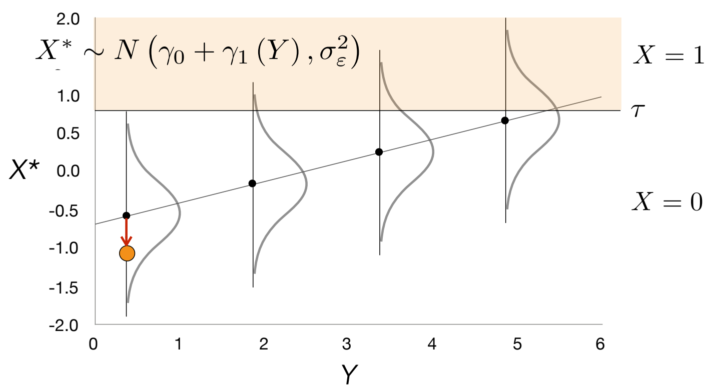
Plausible latent score, retain sample



14

Sample Latent Scores: Heavy Smoking = 1

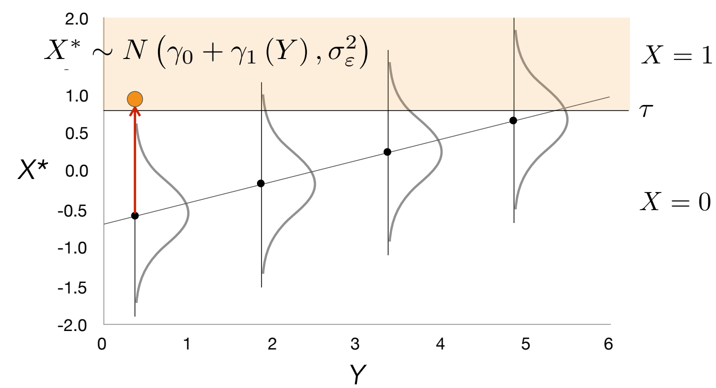
Implausible latent score, reject sample



15

Sample Latent Scores: Heavy Smoking = 1

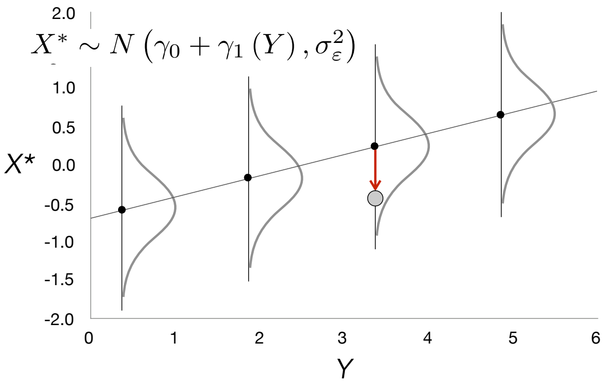
Plausible latent score, retain sample



16

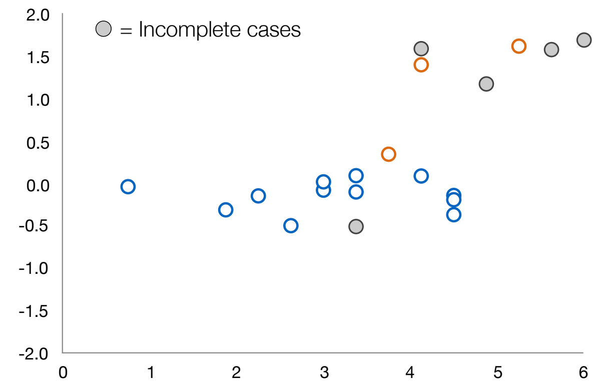
Sample Latent Scores: Heavy Smoking = NA

Any latent score is plausible, retain sample



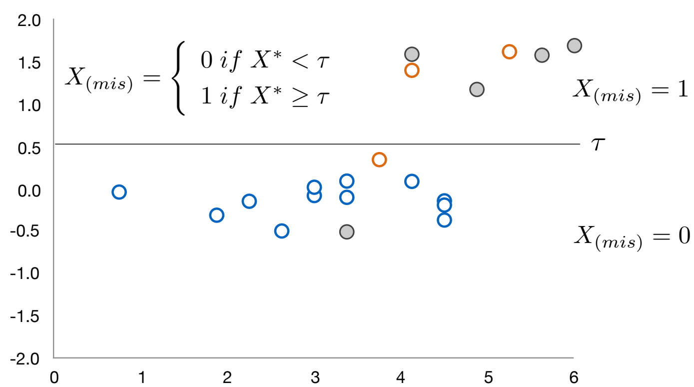
17

Latent Variable Scatterplot For Full Sample



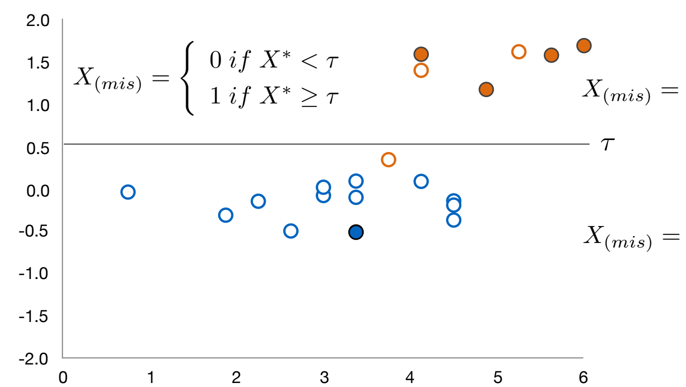
18

Compare Missing Values To Threshold



19

Generating Discrete Imputes



20

Ordinal Variables

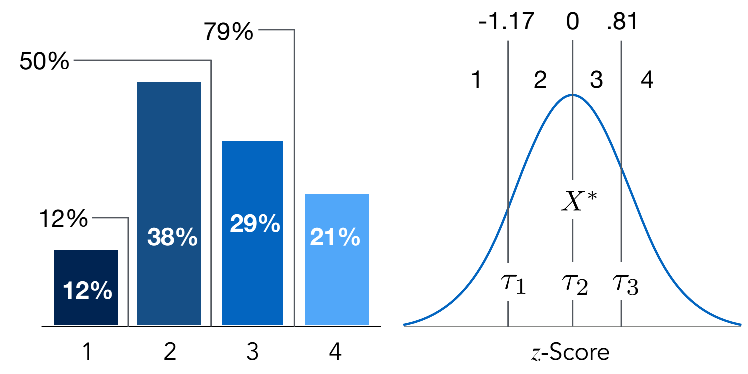
Ordinal variables follow an identical procedure but require additional threshold parameters

An ordinal variable with *K* response options requires *K*-1 threshold parameters

MCMC steps are identical to the binary case

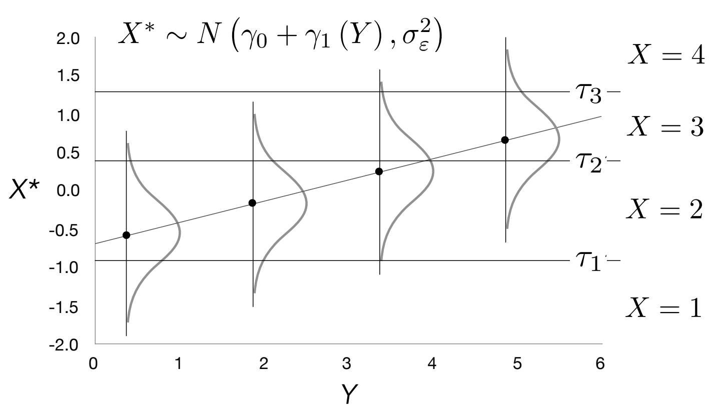
21

Marginal Distribution Example



22

Latent Variable Thresholds



23

Analysis Example

24

Analysis Model

The analysis model is a multiple regression predicting self-efficacy to quit based on heavy cigarette smoking, gender, and years smoking

Binary variables can be treated as ordinal or nominal

$$Efficacy = \beta_0 + \beta_1(Heavy\,Cigs) + \beta_2(Male) + \beta_3(Years) + e$$

Ex5.1.imp
Blimp Diagnostic Script

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
   efficacy stress;
NOMINAL: male;
ORDINAL: heavycig;
MISSING: -99;
MODEL: ~ efficacy heavycig male years;
SEED: 90291;
BURN: 6000;
THIN: 1;
NIMPS: 2;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate psr;
CHAINS: 2 processors 2;
```


Diagnostic Output

```
POTENTIAL SCALE REDUCTION (PSR) OUTPUT:

  Comparing iterations 2801 to 5600 for 2 chains.
                    ----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ----------------------------------------------------
           Max PSR  |       1.062|         nan|       1.000|        -inf|
   Missing Variable |    heavycig|            |    efficacy|            |
                    ----------------------------------------------------

  Comparing iterations 2851 to 5700 for 2 chains.
                    ----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ----------------------------------------------------
           Max PSR  |       1.054|         nan|       1.000|        -inf|
   Missing Variable |    heavycig|            |    efficacy|            |
                    ----------------------------------------------------

  Comparing iterations 2901 to 5800 for 2 chains.
                    ----------------------------------------------------
                    |     Fix Eff| Ran Eff Var|     Err Var|   Threshold|
                    ----------------------------------------------------
           Max PSR  |       1.041|         nan|       1.000|        -inf|
   Missing Variable |    heavycig|            |    efficacy|            |
                    ----------------------------------------------------
```


Ex5.2.imp
Blimp Imputation Script (Mplus Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
   efficacy stress;
NOMINAL: male;
ORDINAL: heavycig;
MISSING: -99;
MODEL: ~ efficacy heavycig male years;
SEED: 90291;
BURN: 3000;
THIN: 3000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imp*.csv;
OPTIONS: separate;
CHAINS: 2 processors 2;
```

Blimp Output

```
---------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

   id quitmeth male age years cigs heavycig efficacy stress

---------------------------------------------------------------
```


Ex5.3.inp
Mplus Analysis Script

```
DATA:
file = implist.csv;
type = imputation;
VARIABLE:
names = id quitmeth male age years cigs heavycig
  efficacy stress;
usevariables = efficacy heavycig male years;
MODEL:
efficacy on heavycig (b1)
  male (b2)
  years (b3);
MODEL TEST:
b1 = 0; b2 = 0; b3 = 0;
OUTPUT:
standardized(stdyx);
```


Mplus Analysis Output

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
EFFICACY ON					
HEAVYCIG	-1.647	2.516	-0.655	0.513	0.151
MALE	1.780	2.022	0.881	0.379	0.245
YEARS	-0.610	0.249	-2.450	0.014	0.059
Intercepts					
EFFICACY	17.884	3.135	5.705	0.000	0.123
Residual Variances					
EFFICACY	11.248	4.479	2.511	0.012	0.347

Ex5.4.imp
Blimp Imputation Script (R, SAS, SPSS, and Stata Format)

```
DATA: ~/desktop/examples/smoking.csv;
VARNAMES: id quitmeth male age years cigs heavycig
   efficacy stress;
NOMINAL: male;
ORDINAL: heavycig;
MISSING: -99;
MODEL: ~ efficacy heavycig male years;
SEED: 90291;
BURN: 3000;
THIN: 3000;
NIMPS: 20;
OUTFILE: ~/desktop/examples/imps.csv;
OPTIONS: stacked;
CHAINS: 2 processors 2;
```

Blimp Output

```
---------------------------------------------------------------

VARIABLE ORDER IN SAVED DATA:

  imp# id quitmeth male age years cigs heavycig efficacy stress

---------------------------------------------------------------
```


Ex5.5.r
R Analysis Script

```
# Required packages
library(mitml)

# Read data
filepath <- "~/desktop/examples/imps.csv"
impdata <- read.csv(filepath, header = F)
names(impdata) <- c("imputation", "id", "quitmeth", "male", "age",
  "years", "cigs", "heavycig", "efficacy", "stress")

# Analyze data and pool estimates
implist <- as.mitml.list(split(impdata, impdata$imputation))
analysis <- with(implist, lm(efficacy ~ heavycig + male + years))
estimates <- testEstimates(analysis, var.comp = T, df.com = 17)
estimates

# Test full model with Wald test
emptymodel <- with(implist, lm(efficacy ~ 1))
testModels(analysis, emptymodel, method = "D1")
```


R Analysis Output

```
Final parameter estimates and inferences obtained from 20 imputed data sets.

            Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)   17.884     3.462     5.166    13.673     0.000     0.111     0.101
heavycig      -1.647     2.771    -0.594    13.279     0.562     0.140     0.124
male           1.780     2.205     0.807    11.894     0.435     0.253     0.206
years         -0.610     0.277    -2.204    14.546     0.044     0.050     0.048

                   Estimate
Residual~~Residual   14.060

Hypothesis test adjusted for small samples with df=[17]
complete-data degrees of freedom
```


R Analysis Output

```
Model comparison calculated from 20 imputed data sets.
Combination method: D1

    F.value      df1      df2    P(>F)      RIV
      2.572        3 3245.939    0.052    0.141

Unadjusted hypothesis test as appropriate in larger samples.
```

Ex5.6.sps
SPSS Analysis Script

```
data list free file = '/users/craig/desktop/examples/imps.csv'
 /imputation_ id quitmeth male age years cigs heavycig
  efficacy stress.
exe.

* Initiate pooling routines.
sort cases by imputation_.
split file layered by imputation_.

* Analysis and pooling.
regression
  /descriptives mean stddev corr sig n
  /dependent efficacy
  /method enter heavycig male years.
```


SPSS Analysis Output

Coefficients[a]

			Unstandardized Coefficients		Standardized Coefficients		
imputation_	Model		B	Std. Error	Beta	t	Sig.
1.00	1	(Constant)	16.906	3.613		4.680	.000
		heavycig	-2.154	2.894	-.176	-.744	.467
		male	2.710	2.078	.309	1.304	.211
		years	-.498	.291	-.383	-1.711	.106
...							
20.00	1	(Constant)	17.410	3.116		5.586	.000
		heavycig	-.657	2.497	-.060	-.263	.796
		male	.841	1.792	.107	.469	.645
		years	-.650	.251	-.559	-2.592	.020
Pooled	1	(Constant)	17.884	3.462		5.166	.000
		heavycig	-1.647	2.771		-.594	.552
		male	1.780	2.205		.807	.420
		years	-.610	.277		-2.204	.028

Ex5.7.do
Stata Analysis Script

```
// Import and save original data
import delimited "~/desktop/examples/smoking.csv"
rename (v1 - v9)(id quitmeth male age years cigs heavycig
  efficacy stress)
generate imp=0

// Recode missing values
foreach var of varlist id - stress {
     replace `var' = . if `var'== -99
}
save original, replace

// Import and save imputed data
clear
import delimited "~/desktop/examples/imps.csv"
rename (v1 - v10)(imp id quitmeth male age years cigs heavycig
  efficacy stress)
save imputed, replace
```


Ex5.7.do
Stata Analysis Script

```
// Append original and imputed data
use original, clear
append using imputed

// Convert to mi data
mi import flong, m(imp) id(id) imputed(quitmeth - stress) clear

// Analyze data and pool results
mi estimate, cmdok: regress efficacy heavycig male years
```

Stata Analysis Output

```
Multiple-imputation estimates                     Imputations       =         20
Linear regression                                 Number of obs     =         20
                                                  Average RVI       =     0.1333
                                                  Largest FMI       =     0.2344
                                                  Complete DF       =         16
DF adjustment:   Small sample                     DF:     min       =      11.15
                                                          avg       =      12.50
                                                          max       =      13.61
Model F test:       Equal FMI                     F(   3,   13.6)   =       2.57
Within VCE type:          OLS                     Prob > F          =     0.0969

------------------------------------------------------------------------------
    efficacy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    heavycig |  -1.647135   2.770797    -0.59   0.563    -7.660919    4.366649
        male |   1.780121   2.205269     0.81   0.436    -3.065826    6.626069
       years |  -.6100901   .2767875    -2.20   0.045    -1.205333    -.014847
       _cons |   17.88396   3.461593     5.17   0.000     10.39374    25.37419
------------------------------------------------------------------------------
```