Multiple Imputation with Categorical Variables

Stats Camp 2018: Missing Data Analysis

Kyle M. Lang

Department of Methodology & Statistics
Tilburg University

19–21 October 2018

Outline

- Review of generalized linear models
- Discuss MI for categorical variables

General Linear Model

As we've seen, MI is basically just a slightly more complicated flavor of out-of-sample prediction using regression.

- So far, we've only considered imputation using linear regression models like the following:

$$
\begin{aligned}
Y &= \mathbf{X}\beta + \varepsilon \\
&= \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \varepsilon
\end{aligned}
$$

This type of model is known as the *General Linear Model*.

- All flavors of linear regression are general linear models.
  - ANOVA
  - ANCOVA
  - Multilevel linear regression models

Components of the General Linear Model

We can break our model into pieces:

$$\eta = \mathbf{X}\beta$$
$$Y = \eta + \varepsilon$$

Because $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, we can also write:

$$Y \sim \mathrm{N}(\eta, \sigma^2)$$

In this representation:

- η is the *systematic component* of the model
- The normal distribution, $\mathrm{N}(\cdot, \cdot)$, is the model's *random component*.

Components of the General Linear Model

The purpose of general linear modeling (i.e., regression modeling) is to build a model of the outcome's mean, μ_Y.

- In this case, $\mu_Y = \eta$.
- The systematic component defines the mean of Y.

The random component quantifies variability (i.e., error variance) around μ_Y.

- In the general linear model, we assume that this error variance follows a normal distribution.
- Hence the normal random component.

Generalized Linear Model

We can generalize the models we've been using in two important ways:

1. Allow for random components other than the normal distribution.
2. Allow for more complicated relations between μ_Y and η.
    - Allow: $g(\mu_Y) = \eta$

These extensions lead to the class of *Generalized Linear Models* (GLMs).

Components of the Generalized Linear Model

The random component in a GLM can be any distribution from the so-called *exponential family*.

- The exponential family contains many popular distributions:
  - Normal
  - Binomial
  - Poisson
  - Many others...

The systematic component of a GLM is exactly the same as it is in general linear models:

$$\eta = \mathbf{X}\beta = \beta_0 + \sum_{p=1}^{P} \beta_p X_p$$

Link Functions

In GLMs, η does not directly describe μ_Y.

- We first transform μ_Y via a *link function*.
- $g(\mu_Y) = \eta$

The link function allows GLMs for outcomes with restricted ranges without requiring any restrictions on the range of the $\{X_p\}$.

- For strictly positive Y, we can use a *log link*:

$$\ln(\mu_Y) = \eta.$$

- The general linear model employs the *identity link*:

$$\mu_Y = \eta.$$

Components of the Generalized Linear Model

Every GLM is built from three components:

1. The systematic component, η.
   - A linear function of the predictors, $\{X_p\}$.
   - Describes the association between $\mathbf{X}$ and Y.

2. The link function, $g(\mu_Y)$.
   - Transforms μ_Y so that it can take any value on the real line.

3. The random component, $P(Y|g^{-1}(\eta))$
   - The distribution of the observed Y.
   - Quantifies the error variance around η.

General Linear Model $\subset$ Generalized Linear Model

The general linear model is a special case of GLM.

1. Systematic component:

$$\eta = \mathbf{X}\beta$$

2. Link function:

$$\mu_Y = \eta$$

3. Random component:

$$Y \sim \mathrm{N}(\eta, \sigma^2)$$

Logistic Regression

Logistic Regression

So why do we care about the GLM when linear regression models have worked thus far?

- In a word: Classification.

In the classification task, we have a discrete, qualitative outcome.

- We will begin with the situation of two-level outcomes.
  - Alive or Dead
  - Pass or Fail
  - Pay or Default

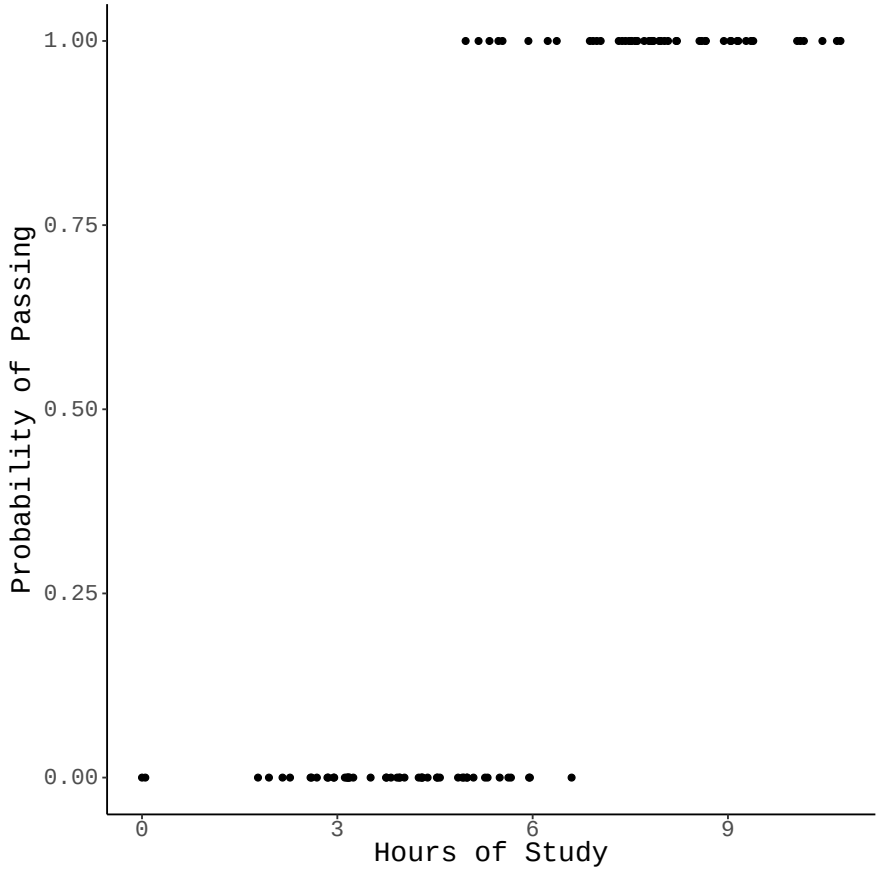
We want to build a model that predicts class membership based on some set of interesting features.

- To do so, we will use a very useful type of GLM: *logistic regression*.

Classification Example

Suppose we want to know the effect of study time on the probability of passing an exam.

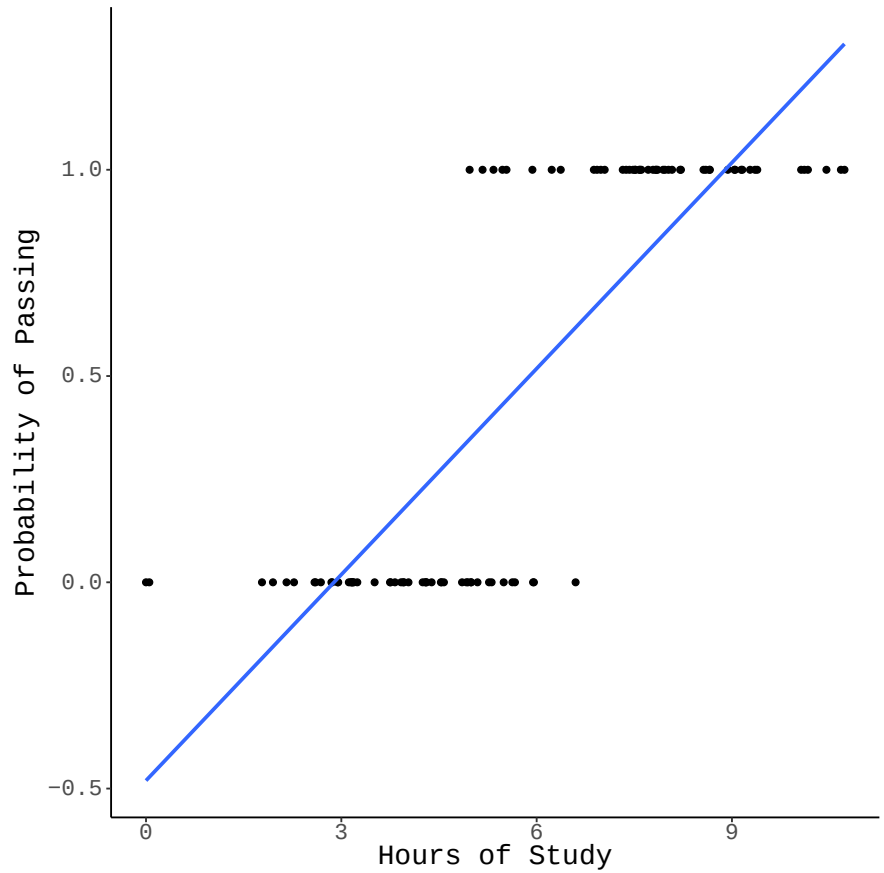
- The probability of passing must be between 0 and 1.
- We care about the probability of passing, but we only observe absolute success or failure.
  - $Y \in \{1, 0\}$



Linear Regression for Binary Outcomes?

What happens if we try to model these data with linear regression?

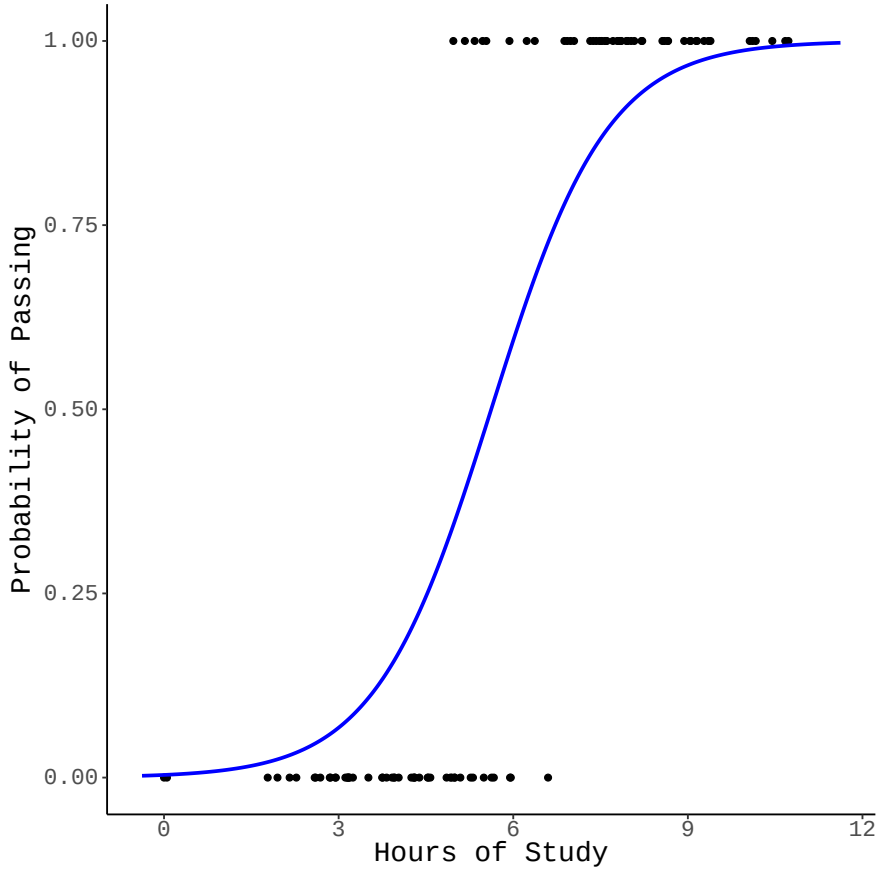
- Hmm...notice any problems?



Logistic Regression Visualized

We get a much better model using logistic regression.

- The link function ensures legal predicted values.
- The sigmoidal curve implies fluctuation in the effectiveness of extra study time.
  - More study time is most beneficial for students with around 6 hours of study.



Defining the Logistic Regression Model

In logistic regression problems, we are modeling binary data:

- Usual coding: $Y \in \{1 = \text{“Success”}, 0 = \text{“Failure”}\}$.

The *Binomial* distribution is a good way to represent this kind of data.

- The systematic component in our logistic regression model will be the binomial distribution.

The mean of the binomial distribution (with $N = 1$) is the “success” probability, $\pi = P(Y = 1)$.

- We are interested in modeling $\mu_Y = \pi$:

$$g(\pi) = \mathbf{X}\beta$$

Link Function for Logistic Regression

Because π is bounded by 0 and 1, we cannot model it directly—we must apply an appropriate link function.

- Logistic regression uses the *logit link*.
  - Given π, we can define the *odds* of success as:

$$O_s = \frac{\pi}{1 - \pi}$$

  - Because $\pi \in [0, 1]$, we know that $O_s \geq 0$.
  - We take the natural log of the odds as the last step to fully map π to the real line.

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right)$$

Fully Specified Logistic Regression Model

Our final logistic regression model is:

$$Y \sim \text{Bin}(\pi, 1)$$
$$\text{logit}(\pi) = \mathbf{X}\beta$$

The fitted model can be represented as:

$$\text{logit}(\hat{\pi}) = \mathbf{X}\hat{\beta}$$

If we fit a logistic regression model to the test-passing data plotted above, we get:

$$\text{logit}(\hat{\pi}_{pass}) = 0.455 + 0.166 X_{study}$$

Multiple Logistic Regression

The preceding example was a *simple logistic regression*.

- Including multiple predictor variables in the systematic component leads to *multiple logistic regression*.
- The relative differences between simple logistic regression and multiple logistic regression are the same as those between simple linear regression and multiple linear regression.
  - The only important complication is that the regression coefficients become partial effects.

Multiple Logistic Regression Example

Suppose we want to predict the probability of a patient having "high" blood glucose from their age, BMI, and average blood pressure.

- We could do so with the following model:

$$\text{logit}(\pi_{hi.gluc}) = \beta_0 + \beta_1 X_{age} + \beta_2 X_{BMI} + \beta_3 X_{BP}$$

- By fitting this model to the "diabetes" data we get:

$$\text{logit}(\hat{\pi}_{hi.gluc}) = -8.663 + 0.022 X_{age} + 0.126 X_{BMI} + 0.027 X_{BP}$$

Multinomial Logistic Regression

Multi-Class Outcomes

So, what do we do if our outcome takes more than two levels?

- Voting intention = {Will vote, Won't vote, Not sure}
- Preferred caffeine source = {Coffee, Tea, Energy drink, None}
- Current mood = {Happy, Sad, Angry, Neutral}

Multi-Class Outcomes

So, what do we do if our outcome takes more than two levels?

- Voting intention = {Will vote, Won't vote, Not sure}
- Preferred caffeine source = {Coffee, Tea, Energy drink, None}
- Current mood = {Happy, Sad, Angry, Neutral}

Using a nominal variable with L response levels as a predictor requires creating $L - 1$ dummy codes.

- We could solve our problem by estimating $L - 1$ separate logistic regression models.
- Do you see any problems with that approach?

Multi-Class Outcomes

So, what do we do if our outcome takes more than two levels?

- Voting intention = {Will vote, Won't vote, Not sure}
- Preferred caffeine source = {Coffee, Tea, Energy drink, None}
- Current mood = {Happy, Sad, Angry, Neutral}

Using a nominal variable with L response levels as a predictor requires creating $L - 1$ dummy codes.

- We could solve our problem by estimating $L - 1$ separate logistic regression models.
- Do you see any problems with that approach?

We have a better way: *Multinomial Logistic Regression*.

Defining the Multinomial Logistic Regression Model

In multinomial logistic regression problems, we are modeling multi-class nominal data:

- Usual coding: $Y \in \{1, 2, \ldots, L\}$.

The *Multinomial Distribution*—a generalization of the binomial distribution—is a good way to represent this kind of data.

- The systematic component in our multinomial logistic regression model will be the multinomial distribution.

We are interested in modeling the $L - 1$ probabilities, $\pi_l = P(Y = l)$, of endorsing each response level instead of the *baseline* level.

$$g(\pi_l) = \beta_{0l} + \sum_{p=1}^{P} \beta_{pl} X_p, \quad l = 2, 3, \ldots, L$$

Full Multinomial Logistic Regression Model

Given L unique response levels for Y, our final multinomial logistic regression model is:

$$Y \sim \text{Multinom}(\Pi, \mathbf{1}), \quad \Pi = \{\pi_2, \pi_3, \ldots, \pi_L\}$$

$$\text{logit}(\pi_l) = \beta_{0l} + \sum_{p=1}^{P} \beta_{pl} X_p, \quad l = 2, 3, \ldots, L$$

The fitted model can be represented as:

$$\text{logit}(\hat{\pi}_l) = \hat{\beta}_{0l} + \sum_{p=1}^{P} \hat{\beta}_{pl} X_p, \quad l = 2, 3, \ldots, L$$

Note that we, *simultaneously*, estimate $L - 1$ separate sets of coefficients, $\{\beta_{0l}, \beta_{pl}\}$.

Example

Suppose we want to predict the probability of a patient having "high" or "moderate" blood glucose, versus "low" blood glucose, from their age, BMI, and average blood pressure.

- We could do so with the following model:

$$\text{logit}(\pi_l) = \beta_{l0} + \beta_{l1}X_{age} + \beta_{l2}X_{BMI} + \beta_{l3}X_{BP}$$

- By fitting this model to the "diabetes" data we get:

$$\text{logit}(\hat{\pi}_{hi.gluc}) = -14.184 + 0.035X_{age} + 0.241X_{BMI} + 0.067X_{BP}$$

$$\text{logit}(\hat{\pi}_{mid.gluc}) = -6.866 + 0.016X_{age} + 0.132X_{BMI} + 0.046X_{BP}$$

Classification

Predictions from Logistic Regression

Given a fitted logistic regression model, we can get predictions for new observations of $\mathbf{X}$, $\mathbf{X}'$.

- Directly applying $\hat{\beta}$ to $\mathbf{X}'$ will produce predictions on the scale of η:

$$\hat{\eta}' = \mathbf{X}'\hat{\beta}$$

- By applying the inverse link function, $g^{-1}(\cdot)$, to $\hat{\eta}'$, we get predicted success probabilities:

$$\hat{\pi}' = g^{-1}(\hat{\eta}')$$

Predictions from Logistic Regression

In logistic regression, the inverse link function, $g^{-1}(\cdot)$, is the *logistic function*:

$$\text{logistic}(X) = \frac{e^X}{1 + e^X}$$

So, we convert $\hat{\eta}'$ to $\hat{\pi}'$ by:

$$\hat{\pi}' = \frac{e^{\hat{\eta}'}}{1 + e^{\hat{\eta}'}} = \frac{\exp\left(\mathbf{X}'\hat{\beta}\right)}{1 + \exp\left(\mathbf{X}'\hat{\beta}\right)}$$

Classification with Logistic Regression

Once we have computed the predicted success probabilities, $\hat{\pi}'$, we can use them to classify new observations.

- By choosing a threshold on $\hat{\pi}'$, say $\hat{\pi}' = t$, we can classify the new observations as "Successes" or "Failures":

$$\hat{Y}' = \begin{cases} 1 & if \quad \hat{\pi}' \geq t \\ 0 & if \quad \hat{\pi}' < t \end{cases}$$

Classification Example

Say we want to classify a new patient into either the "high glucose" group or the "not high glucose" group using the model fit above.

- Assume this patient has the following characteristics:
  - They are 57 years old
  - Their BMI is 28
  - Their average blood pressure is 92

First we plug their predictor data into the fitted model to get their model-implied η:

$$-1.347 = -8.663 + 0.022(57) + 0.126(28) + 0.027(92)$$

Classification Example

Next we convert the predicted η value into a model-implied success probability by applying the logistic function:

$$0.206 = \frac{e^{-1.347}}{1 + e^{-1.347}}$$

Finally, to make the classification, assume a threshold of $\hat{\pi}' = 0.5$ as the decision boundary.

- Because $0.206 < 0.5$ we would classify this patient into the "low glucose" group.

Predictions from Multinomial Logistic Regression

Generating predictions from a multinomial logistic regression model is nearly identical to predicting with a logistic regression model.

- The only difference is that the multinomial logistic regression model will produce L distinct estimates of $\hat{\eta}'_l$ and $\hat{\pi}'_l$:

$$\hat{\eta}'_l = \begin{cases} \hat{\beta}_{0l} + \sum_{p=1}^{P} \hat{\beta}_{pl} X'_p & \text{if } l > 1 \\ 0 & \text{if } l = 1 \end{cases}$$

$$\hat{\pi}'_l = g^{-1}(\hat{\eta}'_l)$$

Predictions from Multinomial Logistic Regression

In multinomial logistic regression, the inverse link function, $g^{-1}(\cdot)$, is the *softmax function*:

$$\text{softmax}(X_l) = \frac{e^{X_l}}{\sum_{j=1}^{L} e^{X_j}}$$

So, we convert each $\hat{\eta}'_l$ to $\hat{\pi}'_l$ by:

$$\hat{\pi}'_l = \frac{e^{\hat{\eta}'_l}}{\sum_{j=1}^{L} e^{\hat{\eta}'_j}} = \begin{cases} \frac{\exp\left(\hat{\beta}_{0l} + \sum_{p=1}^{P} \hat{\beta}_{pl} X'_p\right)}{1 + \sum_{j=2}^{L} \exp\left(\hat{\beta}_{0j} + \sum_{p=1}^{P} \hat{\beta}_{pj} X'_p\right)} & \text{if } l > 1 \\ \frac{1}{1 + \sum_{j=2}^{L} \exp\left(\hat{\beta}_{0j} + \sum_{p=1}^{P} \hat{\beta}_{pj} X'_p\right)} & \text{if } l = 1 \end{cases}$$

Classification with Multinomial Logistic Regression

Once we have computed the L predicted success probabilities, $\hat{\pi}_l'$, we can use them to classify new observations.

- Each observation is labeled with the response level associated with the largest $\hat{\pi}_l'$
- For example:
  - Given the response options $Y \in \{\text{Coffee, Tea, Energy Drinks, None}\}$
  - And corresponding success probabilities $\hat{\pi}_l \in \{0.45, 0.2, 0.15, 0.2\}$
  - We would assign the observation to the "Coffee" group

Classification Example

Let's re-classify our patient into either the "high glucose", "moderate glucose", or "low glucose" group using the model fit above.

- First we plug their predictor data into the fitted model to get their set of model-implied η_l values:

$$\text{logit}\left(\frac{\pi_{low.gluc}}{\pi_{low.gluc}}\right) = 0 + 0(57) + 0(28) + 0(92) = 0$$

$$\text{logit}\left(\frac{\pi_{mid.gluc}}{\pi_{low.gluc}}\right) = -6.866 + 0.016(57) + 0.132(28) + 0.046(92) = 1.929$$

$$\text{logit}\left(\frac{\pi_{hi.gluc}}{\pi_{low.gluc}}\right) = -14.184 + 0.035(57) + 0.241(28) + 0.067(92) = 0.762$$

Classification Example

- Next we apply the softmax function to convert the predicted η_l values into model-implied success probabilities:

$$\hat{\pi}_{low.gluc} = \frac{1}{1 + e^{1.929} + e^{0.762}} = 0.1$$

$$\hat{\pi}_{mid.gluc} = \frac{e^{1.929}}{1 + e^{1.929} + e^{0.762}} = 0.687$$

$$\hat{\pi}_{hi.gluc} = \frac{e^{0.762}}{1 + e^{1.929} + e^{0.762}} = 0.214$$

- Finally, to make the classification, we find the largest $\hat{\pi}'_l$:
  - Because $\hat{\pi}_{mid.gluc} = 0.687$ is the largest, we would classify this patient into the "moderate glucose" group.

From Classification to Imputation

We can replace the missing data in categorical variables with the classifications produced by Bayesian generalized linear models.

- We need to extend the frequentist models demonstrated above to allow distributions of the model parameters.
- As with linear regression-based imputation, we need to account for uncertainty in the imputation model parameters.

From Classification to Imputation

We can replace the missing data in categorical variables with the classifications produced by Bayesian generalized linear models.

- We need to extend the frequentist models demonstrated above to allow distributions of the model parameters.
- As with linear regression-based imputation, we need to account for uncertainty in the imputation model parameters.

When employing the FCS approach, different flavors of GLM can easily be mixed and matched as required by the particular set of incomplete variables.

- This flexibility is a primary reason why FCS is currently the preferred imputation framework.

Alternatives to GLM?

An old recommendation called for imputing categorical items using normal-theory regression models and subsequently rounding the real-valued imputations to integer codes (Allison, 2002; Honaker & King, 2010).

- This approach tends to perform poorly (Lang & Wu, 2017).
- When dealing the *ordinal* variables that will be analyzed as continuous, *un-rounded* normal-theory imputations can work well (Wu, Jia, & Enders, 2015).

Alternatives to GLM?

Predictive mean matching can work well for ordinal variables.

- To lesser extent for *binary* nominal variables, too.

For multi-category nominal variables, we have to use GLM or some fancy-pants method.

- Tree-based methods
- Machine learning algorithms
- Rule-based donor methods

References

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publictions.

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. doi: 10.1111/j.1540-5907.2010.00447.x

Lang, K. M., & Wu, W. (2017). A comparison of methods for creating multiple imputations of nominal variables. *Multivariate Behavioral Research*, *52*(3), 290-304. doi: 10.1080/00273171.2017.1289360

Wu, W., Jia, F., & Enders, C. (2015). A comparison of imputation strategies for ordinal missing data on likert scale variables. *Multivariate Behavioral Research*, *50*(5), 484-503. doi: 10.1080/00273171.2015.1022644