Regression with Missing Covariates

Valentino Dardanoni, University of Palermo,
based on joint work with Giuseppe De Luca, Salvatore Modica and Franco Peracchi

*IT*9 Meeting, Canazei, January 16th, 2014

The problem

- Empirically relevant case where:
  1. the response variable of interest is observed;
  2. the values of some covariates are missing for some observations;
  3. imputations are available to fill-in the missing covariate values.
- This situation is becoming quite common, as:
  1. public-use data increasingly include imputations of key variables affected by missing data problems;
  2. specialized software for carrying out imputations directly is becoming increasingly available.
- We will not focus much on *how* to impute missing data but on *what to do* with available imputations.

Standard approaches

- Complete-case (CC) approach drops all cases with missing covariate values, ignoring imputations.
- Using a dummy variable for the missing values of each covariate.
- Fill-in (FI) approach replaces missing values with imputations and then uses all the data without distinguishing between observed and imputed values.
- NOTE: sometimes you use the FI approach unknowingly! (data agency often impute missing variables...)

How relevant is the problem? Abrevaya and Donald, 2013

Table 1: Data missingness in economics journals, 2006-2008

Journal	Empirical papers	Papers with missing data (% of empirical papers)	Method of handling missing data[a] (% of missing-data papers in parentheses)		
			Drop observations	Use indicator variables for missingness	Use an imputation method[b]
American Economic Review[c]	191	55 (28.8%)	40 (72.7%)	9 (16.4%)	14 (25.5%)
Journal of Human Resources	94	40 (42.6%)	26 (65.0%)	10 (25.0%)	6 (15.0%)
Journal of Labor Economics	52	26 (50.0%)	18 (69.2%)	4 (15.4%)	5 (19.2%)
Quarterly Journal of Economics	79	41 (51.9%)	29 (70.7%)	8 (19.5%)	10 (24.4%)
Total	416	162 (38.9%)	113 (69.8%)	31 (19.1%)	35 (21.6%)

[a] A given paper may use more than one method, so the percentages add up to more than 100%.
[b] This column includes any type of imputation methods (regression-based, using past/future values, etc.).
[c] Includes *Papers & Proceedings* issues.

Problems with standard approaches

- CC is popular because (under some conditions on the missing covariates process) gives consistent estimates. It may loose a lot of information!
- Using a dummy variable for the missing values of each covariate gives inconsitent estimates (Jones, 1996).
- FI approach replaces missing values with imputations.
- Uses all the data without distinguishing between observed and imputed values.
- Results depend on the validity of imputations.

Imputations

- In general, distinguish between
  1. Non model based imputations;
  2. Model based imputations.
- *Non model based imputations* are, for example,
  1. Mean imputations: typically downward bias in the variance of estimated coefficients.
  2. Simple hot deck: uses observed values to draw missing values; distorts covariances.
- *Model based imputations* makes draws from the estimated distribution based on postulated observed data and missing models.
- *Multiple imputations* handle missing data uncertainty by multiple draws. For each draw you estimate parameters, and then combine them appropriately.

Missing Data Assumptions

- The seminal Rubin (1976) paper defines two key missing data mechanisms:
  1. Missing Completely at Random (MCAR)
  2. Missing at Random (MAR)
- Data are MCAR if the probability of an observation being missing does not depend on observed or unobserved data, and occur entirely at random.
- An example of MCAR would be that a laboratory sample is dropped, so the resulting observation is missing.
- When data are MCAR, the analyses performed on the data are unbiased.
- Data are rarely MCAR.

MAR

- Data are MAR if the probability of an observation being missing does not depend on the true value of the unobserved data.
- An example of MAR would be that two measurements of the same variable are taken at the same time. If they differ by more than a given amount a third is taken. This third measurement is missing for those that do not differ by the given amount.
- Under MAR, likelihood based inference are generally valid; non-likelihood methods (e.g. GMM) can be 'fixed up'.
- It is important to note that:
  1. MAR is not testable.
  2. Both MCAR and MAR are defined for the *whole* data (both dependent variables and covariates).

MAR with missing covariates

- MAR has not always been consistently defined and used.
- Seaman *et al.* (2013), give clear and simple definitions. We adapt it to regression with missing covariates.
- Suppose n units, one $\mathbf{y}$ and K covariates $\mathbf{x}_1, \ldots, \mathbf{x}_K$.
- $\mathbf{z}$ array the data, $\mathbf{z} = [\mathbf{y}', \mathbf{x}']'$.
- Let $\mathbf{m}$ be a $0/1$ vector of missingness indicators and $O(\mathbf{z}; \mathbf{m})$ be the observed subvector $\mathbf{z}$.
- DEFINITION: Data are MAR if for all $\mathbf{m}, \mathbf{z}, \mathbf{z}^*$ such that $O(\mathbf{z}; \mathbf{m}) = O(\mathbf{z}^*; \mathbf{m})$

$$P(\mathbf{m} \,|\, \mathbf{z}) = P(\mathbf{m} \,|\, \mathbf{z}^*)$$

MAR and regression with missing covariates

- MAR is possibly not the best assumption to make in the case of regression with missing covariates.
- For example, MAR does not justify using CC analysis: regression coefficients may not be consistent under MAR.
- Consider the following 'folks' assumption: **ASSUMPTION 1 (A1): y** and **m** are conditionally independent given **x**.
- This says:
  1. **y** may depend on **x** and not on **m**
  2. **m** may depend on **x** but not on **y**
- Should not be confused with MAR.

MAR and Assumption 1: An example

- Suppose there are two units, and suppose want to study relationship between income x and health y.
- Data are (y_1, x_1) and (y_2, x_2); suppose x may be missing on first unit, say.
- MAR says:
  $P([1,0,1,1] \mid [y_1, x_1^a, y_2, x_2]) = P([1,0,1,1] \mid [y_1, x_1^b, y_2, x_2])$
  for all $y_1, x_1^a, x_1^b, y_2, x_2$.
- A1 says:
  $P([1,0,1,1] \mid [y_1^a, x_1, y_2^a, x_2]) = P([1,0,1,1] \mid [y_1^b, x_1, y_2^b, x_2])$
  for all $y_1^a, y_1^b, x_1, y_2^a, y_2^b, x_2$.
- Note that they are logically unrelated:
  1. if $m_2 = f(x_1)$, A1 is OK but MAR not;
  2. if $m_2 = g(y_1)$, MAR is OK but A1 is not.

Complete-case analysis

- CC aestimates regression coefficients on the subsample $O(\mathbf{z}, \mathbf{m})$.
- It is the mostly used approach in economics, and is a useful benchmark because of the following "Folk Theorem" (Glynn and Laird, unpublished 1980; Jones, JASA 1996; Wooldridge, 2002; Dardanoni *et al*, 2011, 2012 and 2013):
- **THEOREM:** Under Assumption 1 estimation of the regression coefficients under CC is consistent.
- Theorem provides the main justification for CC.
  One cannot ignore the severe loss of information that may result when the fraction of missing data is not small.
- Note that violations of A1 lead to inconsistent estimates.

Fill-in approach

- FI estimates regression parameters replacing **X** by the filled-in design matrix **W**, which replaces missing observations with imputations.
- This approach requires two conditions:
  1. The imputation model is correctly specified (assumptions on the posited missing data mechanism, function forms, choice of predictors, etc.).
  2. The imputation model and the regression model must be "congenial" in the sense of Meng (1994) (i.e., the imputation model cannot be more restrictive than the model used to analyze the filled-in data).
- We say that the FI and the imputations are valid when these two conditions hold.

Remarks

- If the FI approach is valid, its estimator of β is generally more precise than the CC estimator.
- If the imputation model is incorrectly specified or the congeniality assumption does not hold, then the FI estimator of β is likely to be inconsistent.
- This looks like a trade-off problem.

The generalized missing-indicator approach

- Dardanoni *et al.* (2011, 2012, 2013) propose a GMI approach for regressions with imputed covariates.
- Dardanoni *et al.* (2011, 2012) consider linear regression; Dardanoni *et al.* (2013) consider GLMs (e.g. logit, probit, Poisson and negative binomial regression, ordered or multinomial logit and probit models etc.).
- They propose a "grand model" with two sets of regressors: observed/imputed covariates (focus), plus a missing-data indicators with interactions (auxiliary).
- This extends the simple indicator model which gives inconsistent estimates.
- It handles the trade-off between bias and precision in the estimation of the regression parameters.

Missing-data patterns

- Assume that some covariate values are missing for some observations. and imputations are available to fill-in the missing covariate values.
- A subsample with incomplete data is a set of observations with one or more missing covariates.
- If β contains the constant term and has dimension K, since the constant is always observed, the number of possible subsamples with missing covariates is equal to $2^{K-1} - 1$.

The "grand model"

- The jth subsample contains N_j observations, K_j observed covariates and $K - K_j$ missing covariates.
- For each subsample $j = 1, \ldots, J$, $\mathbf{W}_j$ is the $N_j \times K$ filled-in design matrix containing the values of the observed covariates and the imputed values of the missing covariate.
- Define the $N \times JK$ matrix of auxiliary regressors

$$\mathbf{Z} = \begin{bmatrix} 0 & \cdots & 0 \\ \mathbf{W}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{W}_J \end{bmatrix}.$$

- Our model is an augmented regression (the grand model) with linear predictor $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}$.

Main result

- **(Equivalence theorem)** For any choice of the imputations in $\mathbf{W}$, the ML estimator of $\boldsymbol{\beta}$ in the grand model $\boldsymbol{\eta} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta}$ is equal to the CC ML estimator of $\boldsymbol{\beta}$.
- EXAMPLE: Linear model with one covariate ($J = K = 2$).
- The grand model is $y_i = \beta_0 + \beta_1 w_i + \delta_{\beta_0} + \delta_{\beta_1}\beta_1 w_i + \epsilon_i$.
- There are $J \times K = 4$ possible models to estimate the parameters of interest (β_0, β_1):
  1. $\delta_{\beta_0} = \delta_{\beta_1} = 0$ (the FI estimator);
  2. No constraint on δ's (the CC estimator);
  3. $\delta_{\beta_0} = 0$;
  4. $\delta_{\beta_1} = 0$;

GMI and the trade-off between bias and precision

- The GMI approach handles the bias/precision trade-off in the estimation of β considering all intermediate models obtained from the grand model.
- The original bias/precision trade-off in the estimation of β is transformed into a problem of uncertainty about a subset of covariates of the grand model.
- Any intermediate model in the expanded model space may play a role in finding the "best" available estimator of β.
- A further advantage is that all missing data patterns are coefficients' subsets of the *same* data set.
- You can compare models which are not comparable because of different number of observations.

Estimation under model uncertainty

Model uncertainty can be handled by either model selection or model averaging.

- Model selection involves first selecting the "best" model out and then estimating β conditional on the selected model.
- Pre-testing problem: uncertainty arising from the model selection step is ignored.
- In model averaging, one estimates β conditional on each model, then uses a weighted average of these conditional estimates.
- Model averaging is more coherent because it takes into account the uncertainty due to both the estimation and the model selection steps.

Model averaging

- Let $\mathcal{M} = \{M_1, \ldots, M_R\}$ denote the set of models being considered.
- The linear predictor for the rth model M_r is

$$\boldsymbol{\eta}_r = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}_r\boldsymbol{\delta}_r,$$

  for each subset $0 \leq r \leq JK$ of the auxiliary covariates.
- The model averaging estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are of the form

$$\widehat{\boldsymbol{\beta}} = \sum_{r=1}^{JK} \lambda_r \widehat{\boldsymbol{\beta}}_r, \qquad \widehat{\boldsymbol{\delta}} = \sum_{r=1}^{JK} \lambda_r S_r \widehat{\boldsymbol{\delta}}_r,$$

  where λ_r are nonnegative weights that add up to one.

Bayesian Model Averaging

- In BMA, $\widehat{\boldsymbol{\beta}}_r$ and $\widehat{\boldsymbol{\delta}}_r$ are weighted by the posterior probability of the rth model

$$\lambda_r = p(M_r \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid M_r)\, p(M_r)}{\sum_{r=1}^{R} p(\mathbf{y} \mid M_r)\, p(M_r)}, \qquad r = 1, \ldots, R,$$

where $p(M_r)$ is the prior on model M_r and

$$p(\mathbf{y} \mid M_r) = \int p(\mathbf{y} \mid \theta_r, M_r)\, p(\theta_r \mid M_r)\, d\theta_r$$

where $p(\mathbf{y} \mid \boldsymbol{\theta}_r, M_r)$ is the likelihood and $p(\boldsymbol{\theta}_r \mid M_r)$ is the prior on $\boldsymbol{\theta}_r = (\boldsymbol{\beta}, \boldsymbol{\delta}_r)$.

- In this setting, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\delta}}$ can be interpreted as the posterior means of the distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

Problems with BMA

- The priors $p(M_r)$ and $p(\boldsymbol{\theta}_r \mid M_r)$ often only chosen for convenience or following some convention.
- The marginal likelihoods $p(\mathbf{y} \mid M_r)$ usually do not have a closed form expression, so some approximation to the posteriors is needed.
- Even for moderate J and K, 2^{JK} may be very large, so exploring all model may be computationally expensive or even unfeasible.
- We use a "block BMA" assuming cofficients are homogenous withing each missing pattern j, reducing the number of models to 2^J.

Empirical application

- We use data on the elderly European population to investigate how cognitive functioning varies with physical health and socio-economic status.
- Data are from the Survey of Health, Ageing and Retirement in Europe (SHARE), a multidisciplinary and cross-national household panel survey.
- The 1st wave, conducted in 2004, covered 15,544 households and 22,431 individuals in 11 European countries (AT, BE, CH, DE, DK, ES, FR, GR, IT, NL, SE).

Outcomes of interest

We consider two dimensions of cognitive functioning:

- The test of verbal fluency consists of counting how many distinct members of the animal kingdom the respondent can name in 1 minute. The outcome is an integer ranging from 0 to 90.
- The test of numeracy consists of four possible questions involving simple arithmetical calculations based on real life situations. In this case, the outcome is an integer ranging from 1 (no correct answer) to 5 (correct answer to the most difficult question).

Covariates

Our covariates include:

- Socio-demographic variables: age, gender, educational attainments, per-capita household income and household net worth.
- Self-reported measures of physical health: number of limitations with activities of daily living and number of chronic diseases.
- Objective measures of physical health: hand grip strength.

Imputations

- Hand grip strength, per-capita household income and household net worth have substantial item nonresponse (6%, 62% and 64%, respectively).
- SHARE data include imputations of key variables, using a multivariate iterative procedure attempting to preserve the correlation structure of the imputed data.
- Congeniality of the SHARE imputations for income and net worth might be questioned since verbal fluency, number of chronic diseases and hand grip strength are not used by the SHARE imputation model.
- We produce our own imputations for the missing values on hand grip strength using a simple hot-deck procedure.

Model specification and estimation

- We estimate a Poisson regression model for verbal fluency and an ordered probit model for numeracy.
- Each model is estimated separately by macro-region: North (DK, NL, SE), Center (AT, BE, CH, DE, FR) and South (ES, GR, IT).
- The number of missing-data patterns is $J = 7$ for all models, so our block-BMA procedure considers $R = 2^7 = 128$ models for each outcome and macro-region.
- We compare estimated coefficients and standard errors for the CC ML estimator, the FI ML estimator, and BMA estimators based on AIC, BIC and RIC priors respectively.

Poisson regression models for verbal fluency (Center)

			Block BMA		
	CCA	FIA	AIC	RIC	BIC
adl	-.0325	-.0561	-.0400	-.0591	-.0611
	(.0097)	(.0038)	(.0094)	(.0071)	(.0056)
chronic	-.0029	-.0031	-.0024	-.0037	-.0042
	(.0036)	(.0016)	(.0030)	(.0025)	(.0023)
grip strength	.0043	.0043	.0044	.0046	.0046
	(.0006)	(.0003)	(.0005)	(.0006)	(.0004)
age	-.0054	-.0058	-.0056	-.0054	-.0055
	(.0006)	(.0003)	(.0005)	(.0006)	(.0004)
male	-.0718	-.0770	-.0770	-.0848	-.0846
	(.0134)	(.0058)	(.0116)	(.0115)	(.0092)
education	.1388	.1574	.1438	.1527	.1540
	(.0095)	(.0041)	(.0086)	(.0064)	(.0060)
income	.0092	.0073	.0101	.0118	.0114
	(.0021)	(.0007)	(.0018)	(.0022)	(.0022)
net worth	-.0006	.0005	.0005	.0011	.0007
	(.0009)	(.0002)	(.0010)	(.0009)	(.0004)
N	2,575	13,635	13,635	13,635	13,635

Main findings

- Little differences in the sign of the estimated coefficients across cognitive domains, macro-regions and estimation methods.
- Some differences in the size of some estimated coefficients and their standard errors across estimation method.
- CC and FI estimates are usually quite different. CC also implies a substantial loss of precision.
- BMA estimates with BIC or RIC priors are closer to those from the more parsimonious FI, while BMA estimates with AIC are closer to those from the less parsimonious CC.

Conclusions

- The choice between CC and FI generates a bias/precision trade-off.
- Either CC or FI is unlikely to emerge as the "best" model since all intermediate models carry information about the parameters of interest.
- Our approach still assumes that we are interested in the β's as measure of the unconditional effect of **x** on **y**.
- If there is substantial heterogeneity (δ's are significantly different than zero), it means that either Assumption 1 or imputations are invalid.
- Our GMI mothod can also estimate the conditional effects of **x** on **y** in different missing data pattern
- We are currently working on that.