Toutenburg, Fieger:

Using diagnostic measures to detect non-MCAR processes in linear regression models with missing covariates

Sonderforschungsbereich 386, Paper 204 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Using diagnostic measures to detect non-MCAR processes in linear regression models with missing covariates

H. Toutenburg [*] A. Fieger [†]

June 6, 2000

Abstract

This paper presents methods to analyze and detect non-MCAR processes that lead to missing covariate values in linear regression models. First, the data situation and the problem is sketched. The next section provides an overview of the methods that deal with missing covariate values. The idea of using outlier methods to detect non-MCAR processes is described in section 3. Section 4 uses these ideas to introduce a graphical method to visualize the problem. Possible extensions conclude the presentation.

1 Data and problem

We consider the classical linear regression model

$$y(n \times 1) = X(n \times p)\, \beta(p \times 1) + \epsilon(n \times 1)$$

with missing data in the $n \times p$ covariate matrix X. Reorganization of the n rows of the data matrix X, the corresponding elements of the response y and the error term ϵ leads to the following structure

$$\begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{mis}} \end{pmatrix} = \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{\mathrm{c}} \\ \epsilon_{\mathrm{mis}} \end{pmatrix} \tag{1}$$

The index c indicates the completely observed submodel whereas the index mis corresponds to the submodel with missing values in the covariate matrix X_{mis} (note that y_{mis} is completely observed).

Using the missing data indicator matrix R introduced by (Rubin, 1976)

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ is observed} \\ 0 & \text{if } Z_{ij} \text{ is missing} \end{cases}$$

with data $Z_{ij} = (y_i, X_{ij})$, the missing mechanism can be characterized by the conditional distribution $f(R|Z, \phi)$ of R given the data Z and an unknown parameter ϕ. The $n \times (p+1)$ matrix Z consists of observed data Z_{obs} and unobserved values Z_{mis}.

The data are said to be missing completely at random (MCAR) if the distribution of R given Z and ϕ only depends on the unknown parameter ϕ for any Z, i.e.

$$f(R|Z, \phi) = f(R|\phi) \quad \forall Z\,.$$

If the conditional distribution of R depends on Z only via the observed values Z_{obs} (for all Z_{mis}, i.e.

$$f(R|Z, \phi) = f(R|Z_{\mathrm{obs}}, \phi) \quad \forall Z_{\mathrm{mis}}\,,$$

the data are called missing at random (MAR).

[*]Institute of Statistics, Ludwig-Maximilians University of Munich, Germany, email: toutenb@stat.uni-muenchen.de

[†]Institute of Statistics, Ludwig-Maximilians University of Munich, Germany, email: andreas@stat.uni-muenchen.de

The optimal estimator is the Gauss-Markov estimator b applied to the data in (1):

$$\begin{aligned} b &= \left(\begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix}' \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix} \right)^{-1} \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix}' \begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{mis}} \end{pmatrix} \\ &= (X_{\mathrm{c}}'X_{\mathrm{c}} + X_{\mathrm{mis}}'X_{\mathrm{mis}})^{-1}(X_{\mathrm{c}}'y_{\mathrm{c}} + X_{\mathrm{mis}}'y_{\mathrm{mis}})\,. \end{aligned} \tag{2}$$

Due to the unknown values in X_{mis}, of course this estimator can not be used directly. There are a variety of methods that deal with this problem.

2 Dealing with missing values

A simple and often used method is to discard all the information available in $(y_{\mathrm{mis}}, X_{\mathrm{mis}})$ and to use the completely observed data in $(y_{\mathrm{c}}, X_{\mathrm{c}})$ only:

$$b_{\mathrm{c}} = (X_{\mathrm{c}}'X_{\mathrm{c}})^{-1}X_{\mathrm{c}}'y_{\mathrm{c}}\,.$$

Using the available information in Z by estimating

$$\hat{\beta} = \hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy}\,,$$

where the elements of $\hat{\Sigma}_{xx}$ and $\hat{\Sigma}_{xy}$ are formed by using all jointly observed pairs $x_{ij}, x_{i'j'}$ and $x_{ij}, y_{i'}$ is called the available case method.

Maximum likelihood procedures address the missing data problem by factoring the joint distribution

$$f(Z, R|\theta, \xi) = f(Z|\theta)f(R|Z, \xi)\,.$$

Integration over the missing data Z_{mis} yields

$$f(Z_{\mathrm{obs}}, R|\theta, \xi) = \int f(Z, R|\theta, \xi)\, dZ_{\mathrm{mis}} = \int f(R|Z, \xi)f(Z|\theta)\, dZ_{\mathrm{mis}}\,.$$

If $f(R|Z)$ depends only on the observed data Z_{obs}, i.e. the MAR assumption holds, we have

$$f(Z_{\mathrm{obs}}, R|\theta, \xi) = f(R|Z_{\mathrm{obs}}, \xi)\int f(Z|\theta)\, dZ_{\mathrm{mis}} = f(R|Z_{\mathrm{obs}}, \xi)f(Z_{\mathrm{obs}}|\theta)\,,$$

which is why the missing data mechanism is also called ignorable in this case.

Imputation procedures form a different approach to the problem. Here the missing values in X_{mis} are replaced by new values. Having done this by some procedure, the estimator (2) with X_{mis} replaced by X_R with X_R as described below becomes operational. To replace the unknown values in X_{mis} a variety of imputation methods exist: mean imputation or zero order regression (ZOR) replaces an unknown value x_{ij} by the mean $\bar{x}_j$, either formed of the complete cases in X_{c} or the available cases in X_{c} and X_{mis}.

Conditional mean imputation or first order regression (FOR) uses auxilliary regressions to find replacements for the missing values. Regressing the covariate with missing values on the remaining covaraites (with parameters estimates based on the complete cases) yields predictions of the missing values that are used as substitutes. If the response y is also used in these regressions a stochastic element is introduced (see Buck (1960) or Toutenburg and Shalabh (1998)).

Multiple Imputation (Rubin, 1987; Schafer, 1997) repeats the imputation step and averages the results. While a single imputation is too smooth, the differences between the individual imputation steps can be properly used to estimate the variance as the sum of the average variance within the imputed data sets and the between imputation variance. This strategy reflects the uncertainty about the imputation process which is ignored in a single imputation strategy.

By replacing a missing value by x_R, the model (2) becomes the mixed model

$$\begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{R}} \end{pmatrix} = \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix}\beta + \begin{pmatrix} 0 \\ \delta \end{pmatrix} + \begin{pmatrix} \epsilon_{\mathrm{c}} \\ \epsilon_{\mathrm{R}} \end{pmatrix}\,,$$

where δ addresses the difference between the true but unknown values in X_{mis} and their replacements in X_R. Using the mixed estimator (Theil and Goldberger, 1961) we have

$$\begin{aligned} b &= \left(\begin{pmatrix} X_{\text{c}} \\ X_{\text{R}} \end{pmatrix}' \begin{pmatrix} X_{\text{c}} \\ X_{\text{R}} \end{pmatrix} \right)^{-1} \begin{pmatrix} X_{\text{c}} \\ X_{\text{R}} \end{pmatrix}' \begin{pmatrix} y_{\text{c}} \\ y_* \end{pmatrix} \\ &= (X_{\text{c}}'X_{\text{c}} + X_{\text{R}}'X_{\text{R}})^{-1}(X_{\text{c}}'y_{\text{c}} + X_{\text{R}}'y_*)\,, \end{aligned}$$

The weighted-mixed-estimator introduced by Schaffrin and Toutenburg (1990) uses a weight $\lambda < 1$ for the values in (y_{mis}, X_R)

$$b(\lambda) = (X_{\text{c}}'X_{\text{c}} + \lambda X_{\text{R}}'X_{\text{R}})^{-1}(X_{\text{c}}'y_{\text{c}} + \lambda X_{\text{R}}'y_{\text{R}})\,. \tag{3}$$

This estimator may be interpreted as the familiar mixed estimator in the model

$$\begin{pmatrix} y_{\text{c}} \\ \sqrt{\lambda}y_* \end{pmatrix} = \begin{pmatrix} X_{\text{c}} \\ \sqrt{\lambda}X_{\text{R}} \end{pmatrix}\beta + \begin{pmatrix} \epsilon_{\text{c}} \\ \sqrt{\lambda}\phi \end{pmatrix}.$$

3 MCAR diagnosis with outlier measures

Popular diagnostics to detect non-MCAR processes contain the comparison of correlation or covariance matrices, the comparison of means ($\bar{y}_{\text{c}}$ vs. $\bar{y}_{\text{mis}}$) or a more general test as described by Little (????). For the situation with only one column affected by missing values Simon and Simonoff (1986) present diagnostic plots where 'envelopes' are compared.

The idea first presented by Simonoff (1988) combines the missing data problem with statistics that derive from the oulier detection field. A comparison of the values of a statistic computed with and without imputation is the comparison of the sub-samples Z_{c} and Z_{mis}.

If the imputation of values can be considered appropriate under MCAR and we really have MCAR (which is the null hypothesis H_0), the statistics should be 'more or less' the same. If we have something other than MCAR, the statistics should reflect this by having different values.

Simonoff (1988) uses Cook's distance, which is based on the confidence ellipsoid

$$C = \frac{(\hat{\beta}_* - \hat{\beta}_{\text{c}})(X_*'X_*)(\hat{\beta}_* - \hat{\beta}_{\text{c}})}{ps_*^2}\,,$$

the residual sum of squares $DRSS$ (Andrews and Pregibon, 1978)

$$DRSS = \frac{(RSS_* - RSS_{\text{c}})/n_{\text{mis}}}{RSS_{\text{c}}/(n_{\text{c}} - n_{\text{mis}} - p + 1)}\,,$$

and the determinant of the $X'X$ matrix DXX (Andrews and Pregibon, 1978)

$$DXX = \frac{|X_{\text{c}}'X_{\text{c}}|}{|X_*'X_*|}\,.$$

For the construction of tests the distribution of the measures under H_0 is needed. As this distribution also depends on the X values, Monte Carlo methods are used to determine it by first computing the complete case statistics, and imputing missing values under MCAR-assumption. The generation of new response values

$$y_{\text{mis}}^{\text{MC}} = \hat{X}_{\text{mis}}\hat{\beta}_{\text{c}} + \epsilon^{\text{MC}}$$

with $\epsilon^{\text{MC}} \sim N(0, s^2 I)$ generates a new data set where 'missing values' are drawn from using an MCAR mechanism.

After applying the imputation procedure to these data the diagnostic measures are computed. Repeating the 'data deletion' and imputation steps a null distribution of the diagnostic measure is generated. Finally the measure can be applied to the imputed original data and the resulting value can be compared to the null distribution.

4 Graphical diagnosis of the missing mechanism

Animated residualplots are presented in Cook and Weisberg (1989). In a stepwise procedure weights between 0 and 1 are used to include one case into the regression. The plots thus represent the influence of that single case. Park, Kim and Toutenburg (1992) present a similar approach to visualize the inclusion of a further variable into a regression model.

The adaption to the situation with missing data which shows a close relationship to the procedures of the preceeding section is described in **?**. Like in the above procedures, imputation is performed under an MCAR assumption. Having filled the gaps in X_{mis}, the weighted-mixed-estimator (3) is computed for certain values $\lambda \in [0, 1]$. Again the ideas is, that if we really have MCAR, there should not be any tendency in the residual plot, when stepwise including Z_R in the model by increasing the weight from 0 to 1.

Figure 1 shows a small program that visualizes the following procedure:

for ($\lambda = 0$; $\lambda \leq 1$; λ+=step) {
    compute regression parameters;
    compute estimated residuals;
    display residual-plot;
}

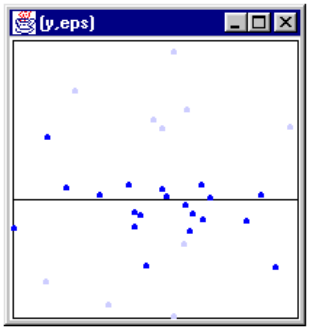

Figure 1: Java program for visualization on computer screen. Reads data for each frame of the animation and draws the single plots of the animation. See `http://www.stat.uni-muenchen.de/~andreas/`

The residual plot in figure 2 shows an example of an animated plot of $\hat{y}$ (an the X-axis) versus $\hat{\epsilon}$ (on the Y-axis) for a model $y = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \epsilon$ with missing data generated by a non MAR process where $P(R_{i2} = 0)$ (a value x_{i2} is missing) dependes on x_2.

An increasing λ gives higher weight to the imputed data in the estimation of the regression parameters. The center of the residual plot in figure 2 shifts towards the origin. For $\lambda = 1$ the imputed data have the same weight as the complete data and biased estimated result. $\lambda = 0$ (the complete case esitmator) on the other hand gives consistent estimates as the missing process is independent of the response y.

5 Possible extensions

The ideas of animated residual plots could be extended in various ways. Imagine a simultaneous plot of $\hat{\epsilon}$ vs. $\hat{y}$ vs. X_i in different windows where the windows are linked. By brushing selected points of the plot could be highlighted in all windows and their location or movement can be studied while the value of λ changes.

Univariate plots y vs. X_j for all j together with the estimated regression line $\hat{\beta}_0 + \hat{\beta}_i X_i$, where the points are static (as the imputation does not depend on the weight λ) and the estimated regression line is dynamic. Again, these plots could be linked as described above.

Creation of a null plot where missing values are created artificially by a known MCAR mechanism. This plot can be used as a means of comparison in order to have an idea of what the plot should look like under MCAR.

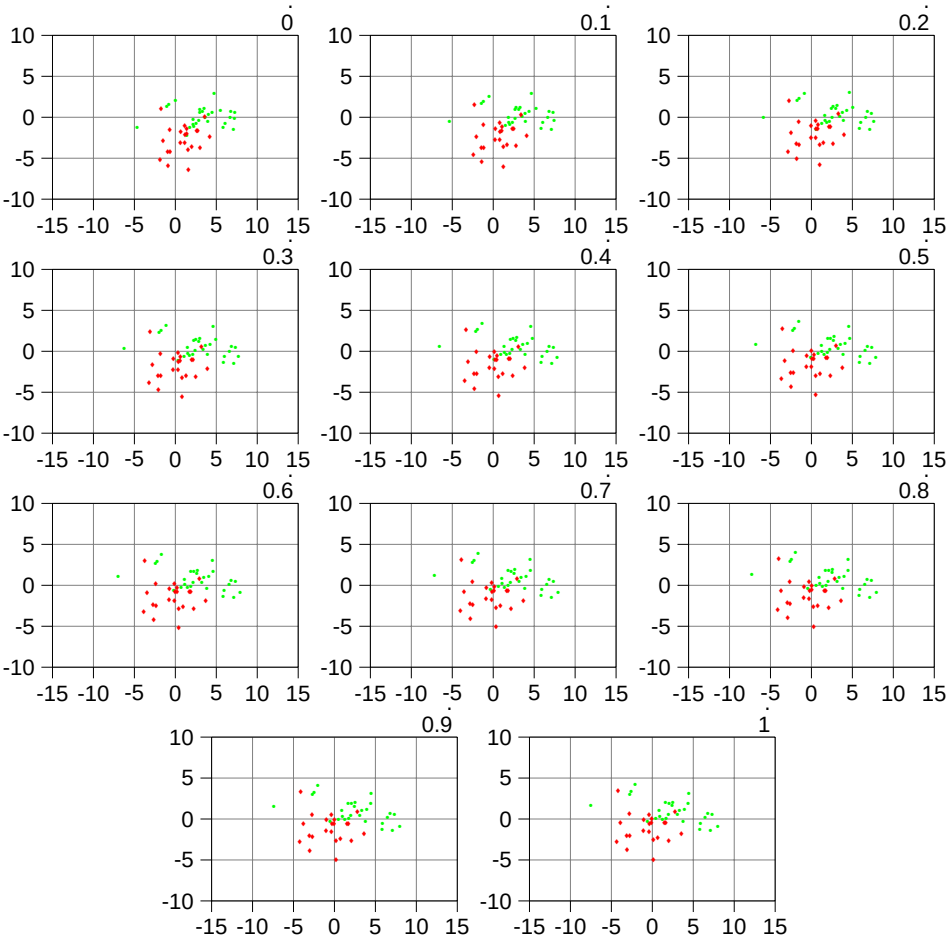

Figure 2: Residualplot of $\hat{y}$ (X-axis) versus $\hat{\epsilon}$ (Y-axis) for a model $y = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \epsilon$ with a non MAR process where $P(R_{i2} = 0)$ (a value x_{i2} is missing) dependes on x_2. λ from 0 (top left) to 1 (bottom right).

References

Andrews, D. F. and Pregibon, D. (1978). Finding outliers that matter, *Journal of the Royal Statistical Society, Series B* **40**: 85–93.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society, Series B* **22**: 302–307.

Cook, R. D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphics, *Technometrics* **31**: 277–291.

Park, S. H., Kim, Y. H. and Toutenburg, H. (1992). Regression diagnostics for removing an observation with animating graphics, *Statistical Papers* **33**: 227–240.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**: 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Sample Surveys*, Wiley, New York.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall, London.

Schaffrin, B. and Toutenburg, H. (1990). Weighted mixed regression, *Zeitschrift für Angewandte Mathematik und Mechanik* **70**: 735–738.

Simon, G. A. and Simonoff, J. S. (1986). Diagnostic plots for missing data in least squares regression, *Journal of the American Statistical Association* **81**: 501–509.

Simonoff, J. S. (1988). Regression diagnostics to detect nonrandom missingness in linear regression, *Technometrics* **30**: 205–214.

Theil, H. and Goldberger, A. S. (1961). On pure and mixed estimation in econometrics, *International Economic Review* **2**: 65–78.

Toutenburg, H. and Shalabh (1998). Prediction of response values in linear regression models from replicated experiments, *to be published* .