Fieger:

Modified First Order Regression, eine Simulationsstudie

Sonderforschungsbereich 386, Paper 62 (1997)

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

Modified First Order Regression, eine Simulationsstudie

Andreas Fieger
Institut für Statistik
Akademiestr. 1
80799 München
andreas@stat.uni-muenchen.de

26. August 1997

Abstract. In diesem Bericht werden verschiedene Imputationsmechanismen für fehlende Kovariablen in einem linearen Regressionsmodell mit zwei Kovariablen untersucht. Hierbei ist eine der Kovariablen vollständig beobachtet, die andere nur teilweise. Die betrachteten Imputationsmechanismen sind Zero Order Regression (ZOR), First Order Regression (FOR), First Order Regression plus random noise (FOR+) und Modified First Order Regression (MFOR).

Keywords. C++ Klassen, Complete Case Schätzung, fehlende Werte, First Order Regression, Gewichtung, Imputationsmechanismen, Lineare Regression, Simulationsstudie, Zero Order Regression.

1 Einleitung

In dieser Studie werden lineare Regressionsmodelle bei fehlenden Kovariablen betrachtet. Die Ergebnisse dieser Simulationsstudien sollen einen ersten Eindruck vom Verhalten der untersuchten mixed Schätzer Schätzer vermitteln, dabei bestehende Probleme aufzeigen und mögliche Lösungsansätze auf Durchführbarkeit hin überprüfen. Die Simulationsprogramme sind in C++

unter Verwendung von Templateklassen zur linearen Algebra (Fieger, Heumann, Kastner und Watzka, 1997) und Klassenbibliotheken zur linearen Regresison (Fieger, 1997) erzeugt.

1.1 Daten

In den Simulationsstudien des vorliegenden Berichts betrachten wir stets eine Datenmatrix $X = (1, x_1, x_2)$ mit folgendender Struktur: der Kovariablenvektor x_1 ist vollständig beobachtet, der Kovariablenvektor x_2 ist nicht vollständig beobachtet. Die Aufteilung in ein 'complete' Modell (Index c) und ein 'missing' Modell (Index $*$) ergibt

$$X = \begin{pmatrix} 1 & x_{1(\mathrm{c})} & x_{2(\mathrm{c})} \\ 1 & x_{1(*)} & x_{2(*)} \end{pmatrix}$$

Die fehlenden Daten beschränken sich also auf den Vektor $x_{2(*)}$, der vollständig unbeobachtet ist.

Für die vorliegende Studie wurden zwei Kovariablen gewählt, da sich so die Korrelationsstruktur durch einen einzigen Parameter beschreiben und in Grafiken nach diesem abtragen läßt. Die betrachteten Verfahren sind jedoch nicht auf diese Situation beschränkt.

1.2 Modell

Betrachten wir ein klassisches lineares Regressionsmodell $y = X\beta + \epsilon$ mit oben beschriebener Kovariablenmatrix X, so erhalten wir nach Partitionierung das sogenannte Mixedmodell (vgl. Rao und Toutenburg, 1995)

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

1.3 Fehlendmechanismus

Die 'Erzeugung' von fehlenden Werten in der betrachteten Studie geschieht stets derart, daß die Werte 'missing completely at random' (MCAR) sind, d. h. daß das Fehlen eines Kovaraiblenwertes nicht von den Daten abhängt (vgl. Little und Rubin, 1987).

2 Untersuchte Schätzer

Im folgenden werden verschiedene Ansätze verglichen, die das Fehlen von Kovariablenwerten bei der Schätzung der Regressionsparameter berücksichtigen. Zugrundelegend ist stets die Aufteilung des Modells in das Mixedmodell (1).

2.1 Complete Case Schätzer

Die einfachste Möglichkeit, die fehlenden Daten zu behandeln, ist die complete case Methode. Fälle die nicht vollständig beobachtet sind werden ausgeschloßen. Das betrachtete Modell reduziert sich damit zu

$$y_{\mathrm{c}} = X_{\mathrm{c}}\beta + \epsilon_{\mathrm{c}}$$

Bei allen weiteren Schätzern wird das Mixed Modell (1) vollständig verwendet. Die nicht beobachteten Werte in $x_{2(*)}$ müssen dazu durch geschätzte Werte ersetzt werden um die so vervollständigte Datenmatrix

$$\begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix}$$

im mixed Schätzer

$$\hat{\beta}^{\mathrm{mixed}} = (X_{\mathrm{c}}'X_{\mathrm{c}} + X_{\mathrm{R}}'X_{\mathrm{R}})(X_{\mathrm{c}}'y_{\mathrm{c}} + X_{\mathrm{R}}'y_{*}) \tag{2}$$

zu verwenden.

2.2 ZOR Schätzer

Ersetze fehlende Werte in der Kovariablenmatrix X durch das Spaltenmittelwerte $\bar{x}_i$ das aus den Daten in X_{c} bestimmt wird. Die Zero Order Regression Methode wird deshalb auch unconditional mean imputation genannt.

2.3 FOR Schätzer

Verwenden wir die Korrelationsstruktur der Kovariablen X, so können wir einen fehlenden Kovariablenwert x_{ij} durch eine Regression auf die restlichen Kovariablen prognostizieren und als Ersatzwert verwenden. Die Regressionskoeffizienten dieser 'Hilfsregressionen' werden aus den vollständigen Fällen geschätzt. Die First Order Regression Methode wird deshalb auch als conditional mean imputation bezeichnet.

2.4 FORplus Schätzer

Bei der First Order Regression wird durch zu 'glatte' Ersetzung die Residualvarianz unterschätzt. Die Einführung eines zusätzlichen Fehlerterms soll diesen Effekt ausgleichen (vgl. Simonoff, 1988).

2.5 MFOR Schätzer

Wie bei der First Order Regression wird die Korrelationsstruktur der Daten verwendet, um fehlende Beobachtungen zu ersetzen. Hier wird jedoch zusätzlich der Response y in den Hilfsregressionen verwendet.

2.6 'Wahrer' Schätzer

In den Simulationsstudien sind die fehlenden Daten künstlich aus dem Datensatz entfernt worden. Dadurch sind ihre Werte bekannt und es kann als Referenz der mixed Schätzer berechnet werden, der sich ergibt wenn die wahren Werte wieder eingesetzt werden, also quasi ein 'perfektes' Imputationsverfahren betrachtet wird.

2.7 Gewichtete Schätzer

In Little (1992) werden Gewichte für die unvollständig beobachteten Fälle betrachtet, die dazu dienen sollen, die erhöhte Residualvarianz auszugleichen. Diese Gewichte werden im gewichteten KQ-Schätzer verwendet (WLS estimation) um den Einfluß der unvollständig beobachteten Fälle zu reduzieren.

Betrachtet werden ein einfaches Gewicht

$$w^* = \frac{\sigma_{yy\cdot 2s}}{\sigma_{yy\cdot s}} = 1 - \rho^2_{2y\cdot s}$$

und ein 'verbessertes Gewicht'

$$w = \frac{(1-\rho^2_{2y\cdot s})m/n}{\rho^2_{2y\cdot s} + (1-\rho^2_{2y\cdot s})m/n}$$

3 Struktur der Simulationsstudie

Die Struktur der hier präsentierten Simulationen ist wie folgt (vergleiche auch Anhang A):

1. Erzeugung einer Datenmatrix

$$\tilde{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

   Die Zeilen von $\tilde{X}$ werden als unabhängig und identisch verteilt gemäß $N(\mu_X, \Sigma)$ erzeugt. Dabei ist $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ und $\mu_X = (0,0)'$ gesetzt. Hinzufügen einer Einsspalte (Intercept) ergibt schließlich die Datenmatrix $X = (1, \tilde{X})$. Dieser Schritt wird für bestimmte Werte von $\rho \in [-0.9, +0.9]$ (hier Schrittweite 0.1) wiederholt.

2. Aus der Datenmatrix X, dem festgelegten Parametervektor $\beta = (1,1,1)'$ und einem Fehlervektor $\epsilon \sim N(0, \sigma_\epsilon^2)$ wird der Responsevektor y gemäß $y = X\beta + \epsilon$ erzeugt. Dieser Schritt wird 200 mal wiederholt. $\sigma_\epsilon^2 = \mathrm{E}(\epsilon^2)$ wurde hierbei als 1.0 bzw. 16.0 gewählt, β wurde $(1,1,1)'$ gesetzt.

3. Fehlende Werte kommen nur in der Datenmatrix X vor, sie fehlen hier nur in der zu β_2 gehörigen Spalte. Die Wahrscheinlichkeit $P(R_{i,2} = 0)$, also die Wahrscheinlichkeit in einer gegebenen Zeile von X einen fehlenden Wert 'zu erhalten', wurde mit 0.3 bzw. 0.5 festgesetzt, gleich für alle i und unabhängig von X. Es liegt also missing completely at random (MCAR) vor.

3.1 Ergebnisse

Die im folgenden dargestellten Ergebnisse stellen mittlere Schätzwerte dar, die sich nach Aggregation über die verschiedenen verwendeten X Matrizen ergeben. Es wurden die oben vorgestellten Schätzer verwendet, die jeweils ungewichtet bzw. in der gewichteten Version bestimmt wurden. Die oben angegebenen Gewichte wurden zum Vergleich auch für die nichtstochastischen Ersetzungen (FOR, MFOR) verwendet. Bei FOR+ ist keine Gewichtung nötig, da dies bereits durch die zusätzlich eingeführte Streuung berücksichtigt wird (in den Grafiken werden bedingt durch die Struktur des Simulationsprogrammes auch hier die gewichteten Ergebnisse angegeben).

3.2 Probleme

Es ergeben sich zwei Probleme für den Schätzer $\hat{\beta}_{\text{MFOR}}$:

- die Varianz σ^2 wird unterschätzt,
- die Schätzung von β ist verzerrt.

3.3 Lösungsansätze

Als mögliche Verbesserungen von $\hat{\beta}_{\text{MFOR}}$ könnten folgende Ansätze dienen, die in einer späteren Studie untersucht werden sollen:

- Varianzkorrektur durch Imputation mit zusätzlichem Fehlerterm (existiert für FOR: dort FORplus; nicht für Bias).
- Biaskorrektur von $\hat{\beta}_{\text{MFOR}}$. Eine Schätzung des Bias von $\hat{\beta}_{\text{MFOR}}$ könnte mittels Bootstrapverfahren ermittelt werden (vgl. Abbildung 1).

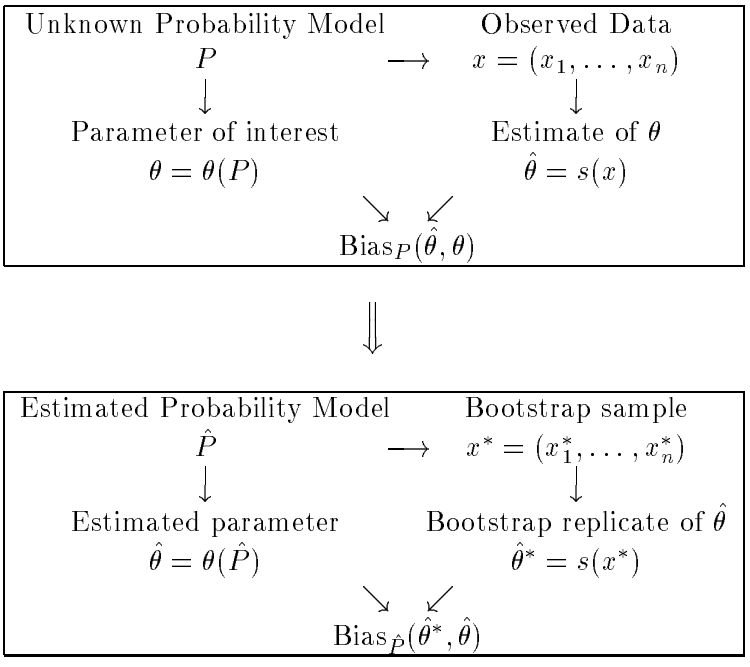

Abbildung 1: Allgemeines Diagramm der Biasschätzung. $\text{Bias}_{\hat{P}}(\hat{\theta}^*, \hat{\theta})$ ist ein allgemeines Bias Maß und muß üblicherweise mit Monte Carlo Methoden approximiert werden. (Abbildung ist aus Efron und Tibshirani (1993) entnommen.)

4 Grafische Darstellung der Ergebnisse

In den folgenden Grafiken sind mit

$$\hat{\beta}_i^{\text{Schätzer}_{\text{Gewicht}}}$$

Mittelwert bzw. Varianz der Replikationen des Schätzers für $\hat{\beta}_i$ bei gegebener Korrelationsstruktur nach der Methode 'Schätzer' (d.h. CC, ZOR, FOR, FOR+ oder MFOR) bezeichnet. Die verwendeten Gewichtungen sind 'w' (gewichtete Version des jeweiligen Schätzers) 'iw' ('verbesserte' gewichtete Version des jeweiligen Schätzers).

4.1 Alle Schätzer

In Abbildungen 2, 7 und 8 werden die ungewichteten Versionen aller betrachteten Schätzverfahren verglichen. Der 'wahre', der Complete Case und der First Order Regression Schätzer sind unverzerrt. Die Zero Order Regression liefert nur für unkorrelierte Kovariablen unverzerrte Schätzergebnisse, ansonsten ist ZOR den am stärksten verzerrte Schätzer. Der FOR+ Schätzer und der MFOR Schätzer sind bezüglich des Bias gegenläufig. Für die vollständig beobachtete Kovariable (hier X_1) überschätzt MFOR den Paramater für negative Korrelation und unterschätzt den Regressionsparameter für positive Korrelation. Für die unvollständig beobachtete Kovariable überschätzt MFOR den Paramater immer, FOR+ unterschätzt den Regressionsparameter stets.

Der Complete Case Schätzer besitzt im Vergleich zu den anderen betrachteten Verfahren die größte Varianz. Der stark verzerrte ZOR Schätzer besitzt die kleinste Varianz. Der FOR+ Schätzer besitzt im wesentlichen die gleiche Varianz wie der 'wahre' Schätzer, ist jedoch verzerrt. Der Vergleich MFOR / FOR ergibt in etwa die gleichen Varianzen für den Parameter der vollständig beobachteten Variablen, jedoch eine geringere Varianz von MFOR im Fall der unvollständig beobachteten Variablen.

4.2 Zero Order Regression

In Abbildung 3 werden der ungewichtete und die gewichteten ZOR Schätzer verglichen. Die Gewichtung reduziert die Verzerrung (das 'verbesserte' Gewicht reduziert den Bias stärker). Die Varianz wird durch die Gewichtung erhöht und ist etwa der der Complete Case Schätzung gleich.

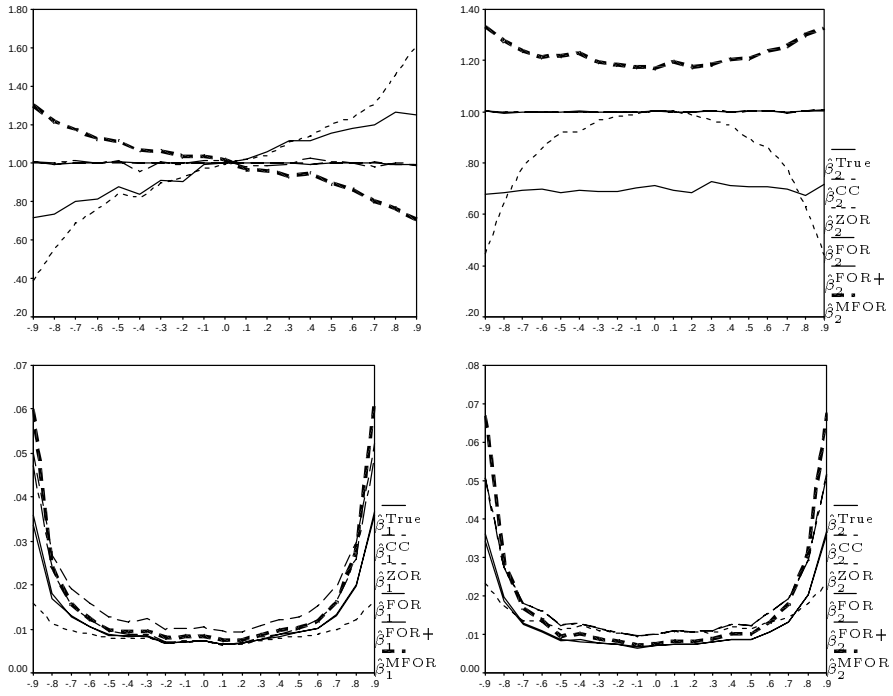

Abbildung 2: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) aller betrachteten Schätzverfahren für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 30 % Fehlwarscheinlichkeit

4.3 First Order Regression

In Abbildung 4 werden der ungewichtete und die gewichteten FOR Schätzer verglichen. Die FOR Schätzung ist unverzerrt. Die Varianz wird durch die Gewichtung erhöht. Im Falle der vollständig beobachteten Kovariablen ist sie aber geringer als die der Complete Case Schätzung. Im Falle der unvollständig beobachteten Kovariablen ist die Varianz der Complete Case, der FOR und der gewichteten FOR Verfahren gleich.

4.4 First Order Regression plus random noise

In Abbildung 5 werden der ungewichtete und die gewichteten FOR+ Schätzer verglichen. Die FOR+ Schätzung ist verzerrt. Mit steigender absoluter Korrelation zwischen den beiden Kovariablen steigt der absolute Bias. Die Gewichtung reduziert das Ausmaß der Verzerrung. Wie beim FOR Schätzer erhöht die Gewichtung die Varianz des Schätzes. Die Varianz bleibt nur im

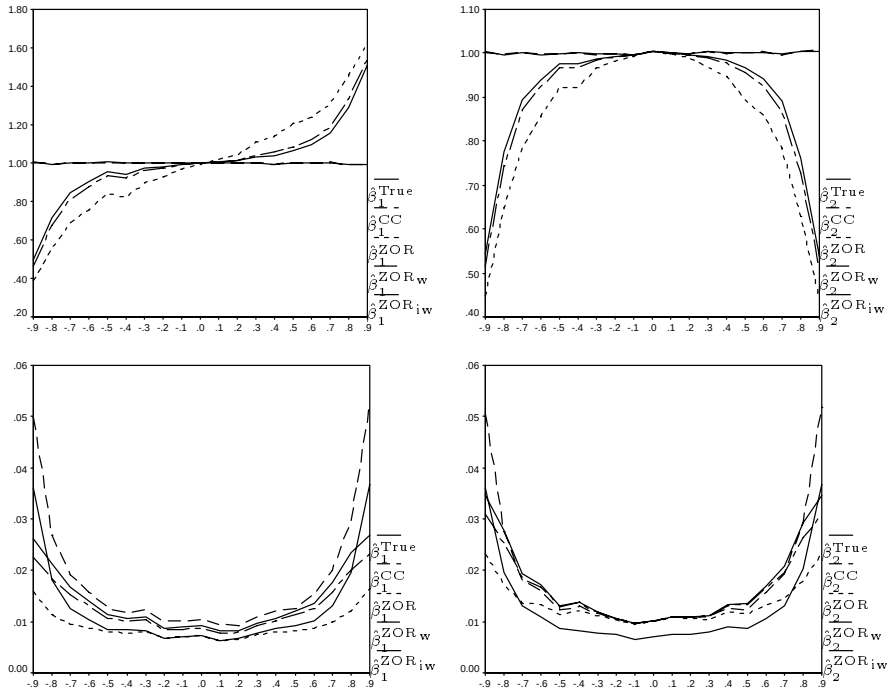

Abbildung 3: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) für Zero Order Regression (ungewichtet, gewichtet und mit modifizierten Gewichten) sowie Complete Case und 'wahre' Werte als Referenz für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 30 % Fehlwarscheinlichkeit

Fall der vollständig beobachteten Kovariablen unter der des Complete Case Schätzers. Bei der unvollständig beobachteten Kovariablen ist die Varianz des Parameterschätzers größer als im Falle des Complete Case Schätzers.

4.5 Modified First Order Regression

In Abbildung 6 werden der ungewichtete und die gewichteten MFOR Schätzer verglichen. Die MFOR Schätzung ist verzerrt. Mit steigender absoluter Korrelation zwischen den beiden Kovariablen steigt der absolute Bias, der durch die Verwendung des Gewichtungsverfahrens (das für die FOR-Ersetzung bestimmt ist) wird der Bias verringert. Die Varianz wird durch die Gewichtung im Falle der vollständig beobachteten Kovariablen erhöht, im Falle der unvollständig beobachteten Kovariablen wird sie verringert; sie bleibt jedoch bei beiden Parameterschätzungen unter der des Complete Case Schätzers.

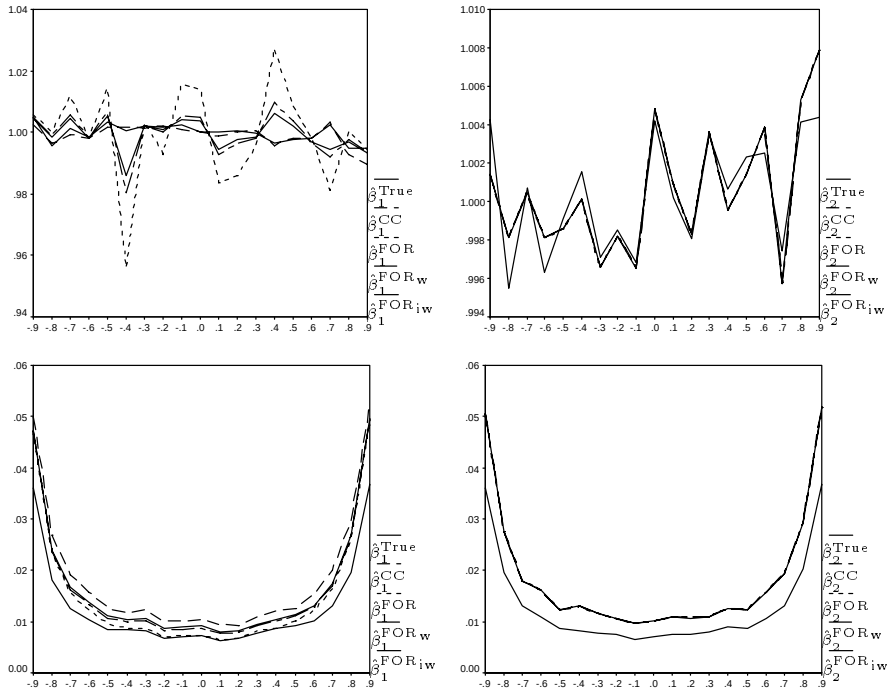

Abbildung 4: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) für First Order Regression (ungewichtet, gewichtet und mit modifizierten Gewichten) sowie Complete Case und 'wahre' Werte als Referenz für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 30 % Fehlwarscheinlichkeit

5 Vergleiche mit anderen Einstellungen

Andere Vorgaben für σ_ϵ^2 und die Fehlendwahrscheinlichkeit ergeben tendenziell die gleichen Ergebnisse. Ein höherer Fehlendanteil erhöht das Ausmaß der Verzerrung, vgl. Abbildung 7. Eine höhere Fehlervarianz σ^2 erhöht durchwegs die Varianz der Parameterschätzungen; das oben beschriebenem Verhalten der jeweiligen Schätzer ändert sich jedoch nicht (vgl. Abbildung 8).

6 Was soll noch kommen

Diese Studie ist nur eine erste Untersuchung verschiedener Ersetzungsverfahren für fehlende Werte im linearen Regressionsmodell. Mit den in (Fieger et al., 1997) und (Fieger, 1997) beschriebenen Werkzeugen ist es jedoch rela-

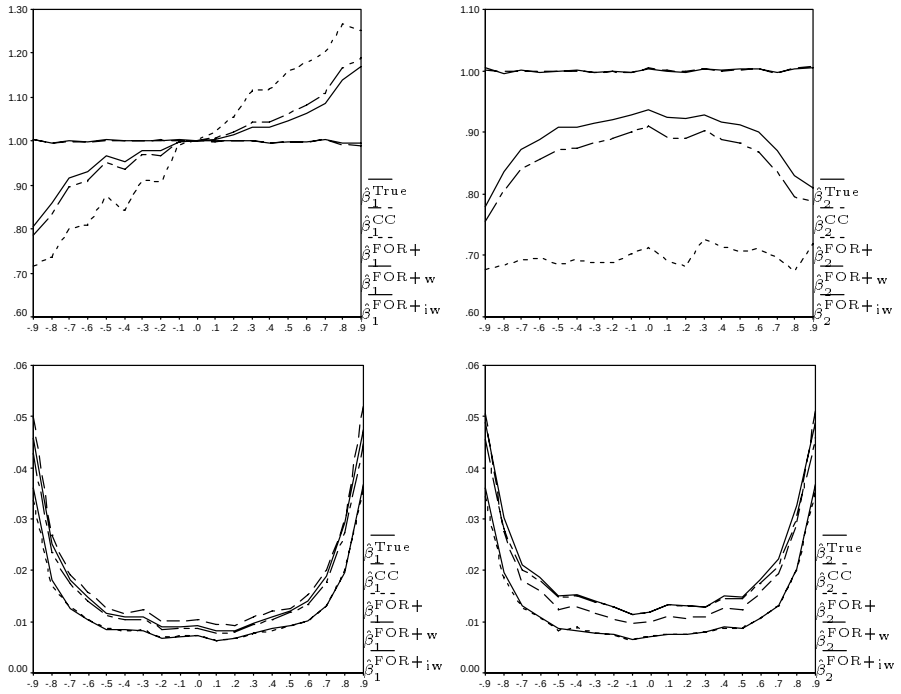

Abbildung 5: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) für First Order Regression plus random noise (ungewichtet, gewichtet und mit modifizierten Gewichten) sowie Complete Case und 'wahre' Werte als Referenz für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 30 % Fehlwarscheinlichkeit

tiv leicht möglich neue Verfahren zu implementieren und andere Situationen zu untersuchen. Interessante Fragestellungen sind

- andere Fehlendmechenismen nicht MCAR oder MAR,
- Zusamenhang zu den in Walbrunn (1997) beschriebenen Diagnosemaßen,
- Einbeziehung diskreter und stetiger Variablen,
- Untersuchung fehlender Werte in diskreten Variablen (binär, mehrkategoriell), d. h. neue, bzw. modifizierte Ersetzungsmechanismen (entwickeln und) implementieren,
- weitere Imputationsmethoden (Fehlerüberlagerung, multiple Imputation, ...).

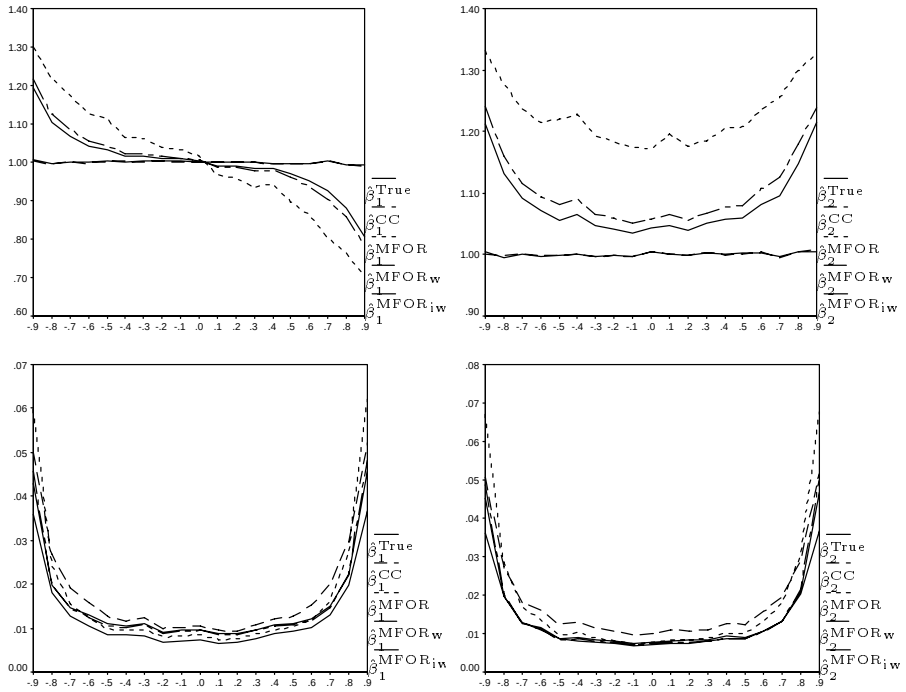

Abbildung 6: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) für modified First Order Regression (ungewichtet, gewichtet und mit modifizierten Gewichten) sowie Complete Case und 'wahre' Werte als Referenz für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 30 % Fehlwarscheinlichkeit

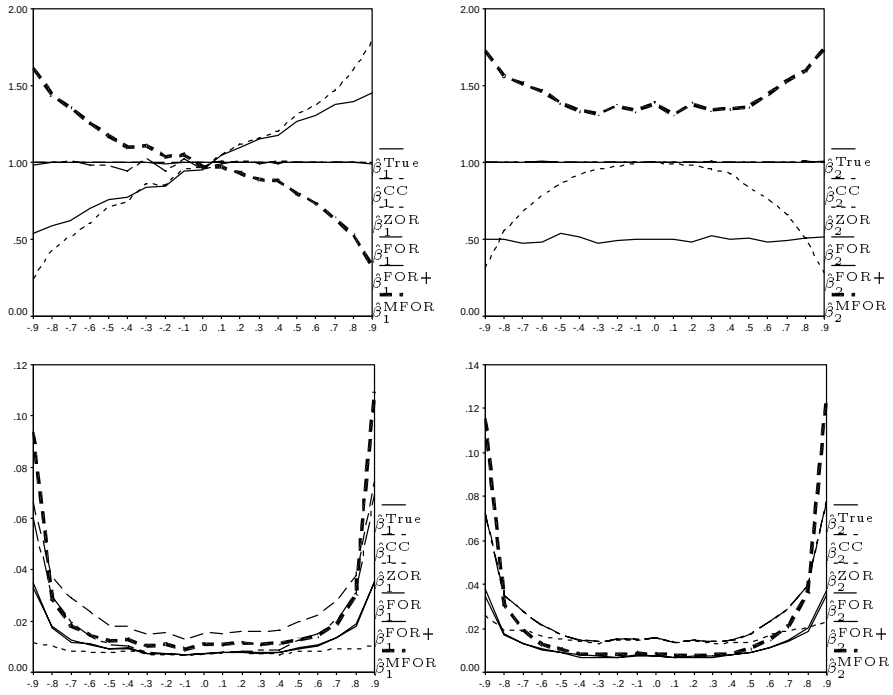

Abbildung 7: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) aller betrachteten Schätzverfahren für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 1.0$ und 50 % Fehlwarscheinlichkeit

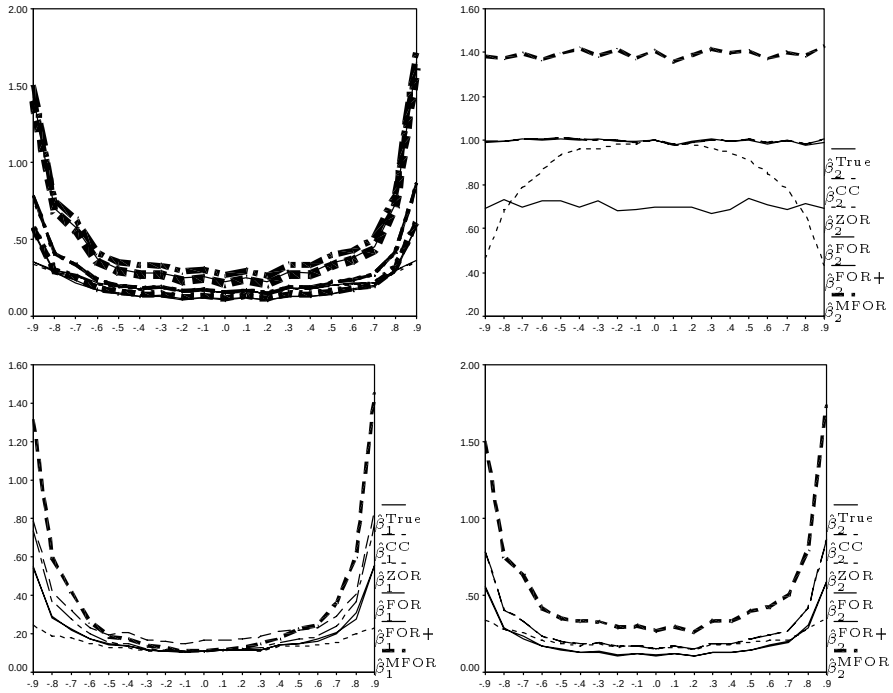

Abbildung 8: Mittelwerte (obere Bilder) und Varianzen (untere Bilder) aller betrachteten Schätzverfahren für $\hat{\beta}_1$ (linke Bilder) und $\hat{\beta}_2$ (rechte Bilder) für $\rho \in [-0.9; +0.9]$ bei $\sigma_\epsilon^2 = 16.0$ und 30 % Fehlwarscheinlichkeit

A Quelltext

Im folgenden ist der den Simulationsprogrammen zugrundeliegende Quelltext angegeben um den genauen Ablauf der Studien zu beschreiben. Quelltextstellen die hierfür nicht von bedeutung sind (wie z. B. das Schreiben der Ergebnisse in eine Datei) wurden ausgelassen und duch '[...]' gekennzeichnet.

```
/* [ ... ] */
double
  truesigma2 = 16.0,          /* E(ε²) = σ² */
  sigma2_Xvalues = 1.0;
matrix truebeta( K, 1, 1.0 );  /* β */
/* [ ... ] */
/* matrices containing the data */
matrix y, X;
/* matrices containing the submodel data after deleting values */
matrix Xmis, ymis, Xc, yc;
intMatrix  R;

/* run one step of the simulation */
matrix SimuStep() {
  /* [ ... ] */
  /* create a model that has all the information,
     i.e. X before deleting some values */
  af_Classical_Linear_Regression_Model mytruemodel( y, X );
  matrix hatbetaTrue = mytruemodel.get_hatbeta();
  /* create a 'Model-Object' for the complete case data */
  af_Classical_Linear_Regression_Model mymodel( yc, Xc );
  matrix hatbetaCC = mymodel.get_hatbeta();
  /* create a new ZOR model */
  af_ZOR_Model myZORmodel( yc, Xc, ymis, Xmis, R );
  // unweighted
  matrix hatbetaallZOR = myZORmodel.get_hatbetaall();
  // weighted
  myZORmodel.set_useweights( 1 );
  myZORmodel.set_whichweight( 0 );
  matrix hatbetaallwZOR = myZORmodel.get_hatbetaall();
  // weighted (improved weights)
  myZORmodel.set_whichweight( 1 );
  matrix hatbetaalliwZOR = myZORmodel.get_hatbetaall();
  /* create a new FOR model */
  af_FOR_Model myFORmodel( yc, Xc, ymis, Xmis, R );
  // unweighted
  matrix hatbetaallFOR = myFORmodel.get_hatbetaall();
  // weighted
  myFORmodel.set_useweights( 1 );
  myFORmodel.set_whichweight( 0 );
  matrix hatbetaallwFOR = myFORmodel.get_hatbetaall();
```

```
  // weighted (improved weights)
  myFORmodel.set_whichweight( 1 );
  matrix hatbetaalliwFOR = myFORmodel.get_hatbetaall();
  /* create a new FORplus model */
  af_FORplus_Model myFORplusmodel( yc, Xc, ymis, Xmis, R );
  // unweighted
  matrix hatbetaallFORplus = myFORplusmodel.get_hatbetaall();
  // weighted
  myFORplusmodel.set_useweights( 1 );
  myFORplusmodel.set_whichweight( 0 );
  matrix hatbetaallwFORplus = myFORplusmodel.get_hatbetaall();
  // weighted (improved weights)
  myFORplusmodel.set_whichweight( 1 );
  matrix hatbetaalliwFORplus = myFORplusmodel.get_hatbetaall();
  /* create a new MFOR model */
  af_MFOR_Model myMFORmodel( yc, Xc, ymis, Xmis, R );
  // unweighted
  matrix hatbetaallMFOR = myMFORmodel.get_hatbetaall();
  // weighted
  myMFORmodel.set_useweights( 1 );
  myMFORmodel.set_whichweight( 0 );
  matrix hatbetaallwMFOR = myMFORmodel.get_hatbetaall();
   // weighted (improved weights)
  myMFORmodel.set_whichweight( 1 );
  matrix hatbetaalliwMFOR = myMFORmodel.get_hatbetaall();

  /* copy the results to one single matrix (one row)
     NoOfEstimates with K components + NoOfStatistics * K values */
  matrix results(1, NoOfEstimates*(K+NoOfStatistics), 0.0 );
  /* [ ... ] */
  return results;
}

/* create new data */
void create_data ( void ) {
  /* [ ... ] */
  /* create (X_1, X_2) ~ N(μ, Σ) with Σ determined by Rho and variances */
  af_Normal_Data DC1( mu, variances, Rho, N );
  matrix X1 = DC1.get_y();
  matrix X2 = DC1.get_X();
  /* now use hcat() to build the matrix X = (1, X_1, X_2) */
  matrix columnofones( N, 1, 1.0 );
  X = hcat( columnofones, hcat( X1, X2 ) );
  /* construct β according to the model  NOTE: this y will never be used, but
     we need to create it, as the definition of af_MCAR_Mechanism
      TheMissingMechanism( y, X, themissprobs ); needs an (arbitrary but valid) y. */
  af_LinearRegression_Data my_DC2( X, truebeta, N, truesigma2 );
  y = my_DC2.get_y();
}
```

```
int main ( void ) {
  /* [ ... ] */
  /* iterate ρ ∈ [−0.9, 0.9] */
  for ( int rhosteps=−9; rhosteps<10; rhosteps++ ) {
    rho = double(rhosteps) * 0.1;
    /* use the same correlation structure to construct multiple X's */
    for ( int X_reps=0; X_reps<NoOfDifferentX; X_reps++ ) {
      /* create the data using ρ, etc. */
      create_data();
      /* now we have X with a specified covariance structure */
      /* run the estimations several times for the current correlation structure.
         Use X from above, deleting some values in X and use different
         ε-values to create new y-values in every iteration
         (Note deleting values in X only means marking their position in R,
         i.e. their true value is still available as long as no replacement procedure
         has been called). Create data for the complete and 'missing' submodels and
         get the submodel data (we don't need the y's, they will be created later) */
      af_MCAR_Mechanism TheMissingMechanism( y, X, themissprobs );
      TheMissingMechanism.reset();
      Xmis = TheMissingMechanism.get_Xast();
      R = TheMissingMechanism.get_R();
      Xc = TheMissingMechanism.get_Xc();
      /* don't forget to reorganize the rows of X according to y_c and y_* */
      X = vcat(Xc, Xmis);
      /* create a list that will contain the matrices with
         the results of the steps in the following loop */
      LinkedList<matrix> XRuns_List;
      /* create a list iterator and link to the list */
      LinkedListIterator<matrix> XRuns_ListIter(XRuns_List);
      for ( int runs=0; runs<NoOfXRuns; runs++ ) {
        /* create new errors ε ∼ N(0, σ²) */
        matrix cc_errors( Xc.rows(), 1 );
        matrix mis_errors( Xmis.rows(), 1 );
        for ( unsigned eps_count=0; eps_count<cc_errors.rows(); eps_count++ ) {
          cc_errors.put( eps_count, 0, Normal( 0.0, truesigma2 ) );
        }
        for ( eps_count=0; eps_count<mis_errors.rows(); eps_count++ ) {
          mis_errors.put( eps_count, 0, Normal( 0.0, truesigma2 ) );
        }
        /* use the errors to create new y's */
        yc = Xc * truebeta + cc_errors;
        ymis = Xmis * truebeta + mis_errors;
        y = vcat(yc, ymis);
        /* run one step of the simulation and store the results in a list */
        XRuns_List.insert( SimuStep() );
      }
      /* compute statistics from all results in the list
         Size of the matrices as the result of SimuStep() */
```

```
      /* [ ... ] */
      /* print the reults to file */
      /* [ ... ] */
    }
  }
  /* terminate program */
  return 0;
}
```

Literatur

Efron, B. und Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

Fieger, A. (1997). C++ Klassen zur Linearen Regression bei fehlenden Kovariablen, *SFB386 – Discussion Paper 61*, Universität München.

Fieger, A., Heumann, C., Kastner, C. und Watzka, K. (1997). Generische Bibliothek zur Linearen Algebra und zur Simulation in C++, *Technical Report 63*, Universität München.

Little, R. J. A. (1992). Regression with missing X's: a review, *Journal of the American Statistical Association* **87**: 1227–1237.

Little, R. J. A. und Rubin, D. B. (1987). *Statistical analysis with missing data*, Wiley, New York.

Rao, C. R. und Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives*, Springer, New York.

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

Walbrunn, D. (1997). *Regressionsdiagnostik zur Identifizierung von nicht-MCAR Prozessen*, Diplomarbeit, Institut für Statistik, Universität München, Ludwigstr. 33, 80535 München, Germany.