A blinded, controlled comparison of methods for adjusting for covariate measurement error in Regression Modelling

A joint project of Topic Groups 2 (Selection of Variables and Functional Forms) and TG4 (Measurement Error and Misclassification) of the STRATOS Initiative

Data Generation	Bayes	Regression Calibration Multiple Imputation	SIMEX
Laurence Freedman Anne Thiebaut Aris Perperoglou Mohammed Sedki	Paul Gustafson Raymond Carrol Frank Harrell Jr Nadja Klein	Victor Kipnis Doug Midthune Amer Moosa Chaloux Matthew Brian Barrett	Michal Abrahamowicz Steve Ferreira Guerra

Outline

- TG2 – TG4 Partnership / Functional Forms & Measurement Error
- The project protocol
- Results of Stages 1 & 2
- Discussion

A joint project between TG2 and TG4

TG2

Selection of variables and functional forms in multivariable analysis

Aim: Derive guidance for variable and function selection in multivariable analysis.

Main focus: identify influential variables and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) main challenges are **selection of variables** for inclusion in a multivariable explanatory model, and **choice of functional forms** for continuous variables

TG4

Measurement error and misclassification

Aim: Increase awareness of problems caused by **measurement error and misclassification** in statistical analyses and remove barriers to use statistical methods that deal with such problems.

Key messages: Only a minority of published papers present estimates that are adjusted for measurement error.

Considering measurement error is necessary because it may have an impact on the study results.

Special statistical methods are used to account for measurement error.

Additional information is required about the type and size of the measurement error to adjust for measurement error.

Key publications

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and
Prognostic Research

COMMENTARY Open Access

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Check for updates

Willi Sauerbrei[1*], Aris Perperoglou[2], Matthias Schmid[3], Michal Abrahamowicz[4], Heiko Becher[5], Harald Binder[1], Daniela Dunkler[6], Frank E. Harrell Jr[7], Patrick Royston[8], Georg Heinze[6] and for TG2 of the STRATOS initiative

1. Investigation and comparison of properties of **variable selection strategies**
2. **Comparison of spline procedures** in univariable & multivariable contexts
3. How to model one or more variables with a ‚**spike-at-zero**'?
4. Comparison of **multivariable procedures for model and function selection**
5. **Role of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

Statistics in Medicine

TUTORIAL IN BIOSTATISTICS

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment

Ruth H. Keogh, Pamela A. Shaw, Paul Gustafson, Raymond J. Carroll, Veronika Deffner, Kevin W. Dodd, Helmut Küchenhoff, Janet A. Tooze, Michael P. Wallace, Victor Kipnis, Laurence S. Freedman

First published: 03 April 2020 | https://doi.org/10.1002/sim.8532 | Citations: 56

TUTORIAL IN BIOSTATISTICS

STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics

Pamela A. Shaw, Paul Gustafson, Raymond J. Carroll, Veronika Deffner, Kevin W. Dodd, Ruth H. Keogh, Victor Kipnis, Janet A. Tooze, Michael P. Wallace, Helmut Küchenhoff, Laurence S. Freedman

First published: 03 April 2020 | https://doi.org/10.1002/sim.8531 | Citations: 28

Measurement Error in regression modelling

We are interested in learning the regression relationship between an outcome variable $\boldsymbol{Y}$ and a covariate(s) $\boldsymbol{X}$: $E(Y|X) = \beta_0 + \beta_X X$

- Classical Measurement Error Model (CME)

  $\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{U}$, where $\boldsymbol{U}$ is random variable with mean 0, independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

- **Impact on the regression relationship**
  - **Attenuation Bias:** Measurement error leads to attenuation of the estimated regression coefficients. The estimated coefficient is biased towards zero, reducing its magnitude.
  - **Loss of Precision**: Increased variance in the estimates, making them less precise. Effective sample size is reduced due to the error variance.
- **When X is not linearly related with Y**: E(Y|X)=f(X).
  - Function f() is unknown, requiring **flexible estimation methods**
  - Observing X^* measured with error distorts the identification of the functional form
- Estimation methods are affected, potentially leading to incorrect inferences about the nature of the relationship.

Research objectives and project setup

- **Project Aim:** Evaluate and compare different methods of estimating the true relationship between an outcome variable (Y) and a covariate (X) when X is measured with error.
- **Framework of investigation:** Project will be conducted by four teams in the following workflow

Data Generation Team → shares data → Methods Team → shares code → Data Generation Team

Simulates data and defines functions

- Cubic b-splines with 1 interior knot at median of observed X*.
- P-splines with 10 interior knots. Penalty optimised by methods team.
- Fractional Polynomials of second degree (4df). Best FP2 power selected by methods team.

Creates code to apply methods

- SIMEX
- Regression Calibration & Multiple Imputation
- Bayesian

Standardise and runs code. Evaluation and presentation of results.

Phase 1: Data, code creation and evaluation

Data Generation: 5 Datasets

- Data from $\mathrm{logit}(\mathrm{P}(\mathrm{Y}=1|X)) = f(X)$ with undisclosed distribution of X and $f(X)$
- Main Study N=15000 independent realizations of a Y **binary outcome** and a **continuous covariate measured with error** X^*
- Replication substudy sample size: 250
- Measurement error variance: k*var(X)
- Error distribution: Unknown to methods team

Data Team

Code generation from methods teams on distributed “blind data”

Methods Teams

Data Team

Evaluation Mean Squared Error

- Let $f(x_1), f(x_2), \ldots, f(x_m)$ the true values of the function and $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_m)$ estimated values
- MSE computed over a limited range of X values corresponding to the 95% central portion of the distribution of X defined as

$$\text{MSE}= \sum_i \frac{\{f(x_i)-\hat{f}(x_i)\}^2}{x_{high}-x_{low}+1}$$

Data Team

Blinded results from phase 1 & Benchmarks

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
A	0.0051	0.00122	0.00518	0.0033	0.0084	0.0046
B	0.0034	0.00149	0.00454	0.0039	0.0103	0.0047
C	0.0078	0.00264	0.00278	0.0033	0.0156	0.0064
D	0.0089	0.00250	0.00400	0.0038	0.0143	0.0067
E	0.0058	0.00161	0.00822	0.0065	0.0130	0.0070
F	0.0054	0.00159	0.00893	0.0069	0.0137	0.0073
G	0.0068	0.00236	0.00430	0.0052	0.0223	0.0082
H	0.0081	0.00238	0.00576	0.0043	0.0257	0.0092
J	0.0074	0.00094	0.01079	0.0127	0.0141	0.0092
K	0.0067	0.00098	0.01078	0.0142	0.0131	0.0092
L	0.0082	0.00111	0.00550	0.0161	0.0181	0.0098
M	0.0111	0.00591	0.00445	0.0096	0.0190	0.0100
N	0.0083	0.00088	0.00663	0.0167	0.0184	0.0102
P	0.0106	0.00452	0.00440	0.0140	0.0182	0.0103
Q	0.0101	0.00080	0.00722	0.0150	0.0200	0.0106
R	0.0108	0.00040	0.00683	0.0157	0.0209	0.0109
S	0.0099	0.00073	0.00840	0.0165	0.0207	0.0112
T	0.0108	0.00047	0.00699	0.0160	0.0220	0.0113
U	0.0127	0.00090	0.00555	0.0170	0.0261	0.0124
V	0.0064	0.00097	0.00919	0.0188	0.0339	0.0139
W	0.0060	0.00102	0.01012	0.0166	0.0369	0.0141
X	0.0139	0.00135	0.01397	0.0326	0.0161	0.0156
Y	0.0137	0.00141	0.01457	0.0322	0.0167	0.0157
Z	0.0234	0.00345	0.01085	0.0447	0.0238	0.0212
AA	0.0318	0.00057	0.00597	0.0545	0.0171	0.0220
AB	0.0266	0.00057	0.00596	0.0634	0.0169	0.0227
AC	0.0320	0.00129	0.01277	0.0543	0.0135	0.0228
AD	0.0368	0.00177	0.01193	0.0531	0.0289	0.0265
AE	0.0448	0.00112	0.01355	0.0580	0.0160	0.0311
AF	0.0812	0.00359	0.00627	0.0697	0.0360	0.0394
AG	0.0626	0.00045	0.00646	0.1515	0.0339	0.0518
AH	0.0688	0.00417	0.01189	0.2070	0.0400	0.0664
AJ	0.0134	0.00187	0.14832	0.1047	0.0868	0.0710
AK	0.0130	0.00210	0.38618	0.1102	0.1093	0.1242

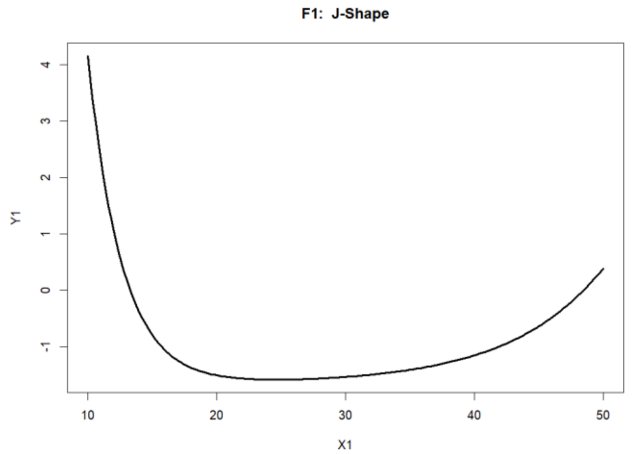
Compute MSE on X and X* values (unadjusted)

Method	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average
Bench-B X	0.0029	0.00160	0.00203	0.0034	0.0040	0.0028
Bench-P X	0.0035	0.00008	0.00280	0.0029	0.0035	0.0026
Bench-B X*	0.0124	0.00449	0.00594	0.0028	0.0311	0.0113
Bench-P X*	0.0101	0.00418	0.00850	0.0023	0.0314	0.0113

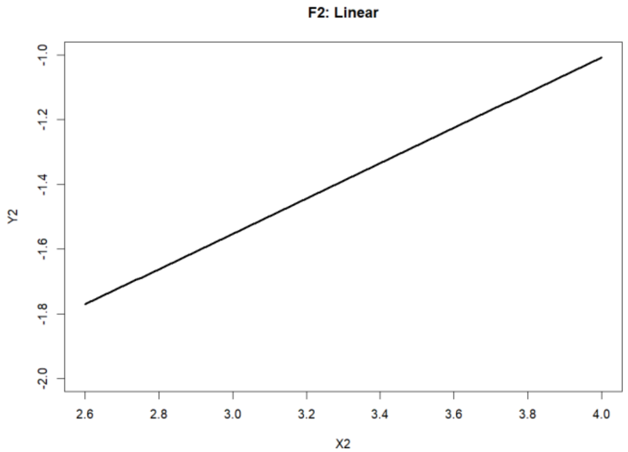
Phase 2 Scenarios

- 5 forms of Y-X relationships: logit(P(Y=1|X))=f(X)
- Main sample sizes: 15000, 30000
- Replication substudy sample sizes: 250, 750
- Measurement error variances: 0.5*var(X), 1.0*var(X)
- Error distribution: Normal, Gamma (shape parameter 3) adjusted to have mean 0
- All combinations of above, except the Phase 1 combination, leading to 15 x 5 = 75 datasets: 15 for each of the 5 forms of relationship
- Code from Phase 1 used by Data Generation and Evaluation Team to run on all 75 dataset

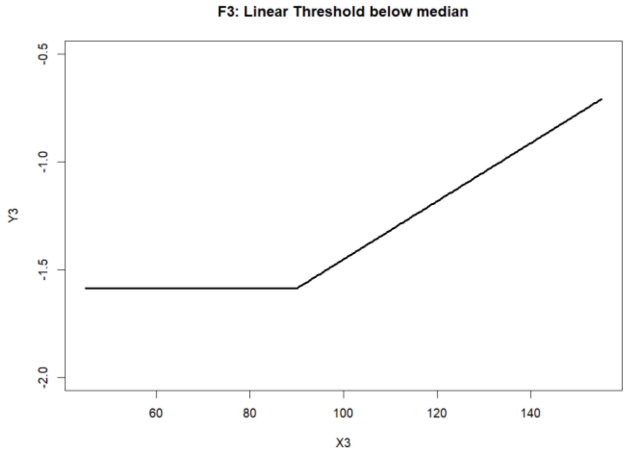
Unblinding



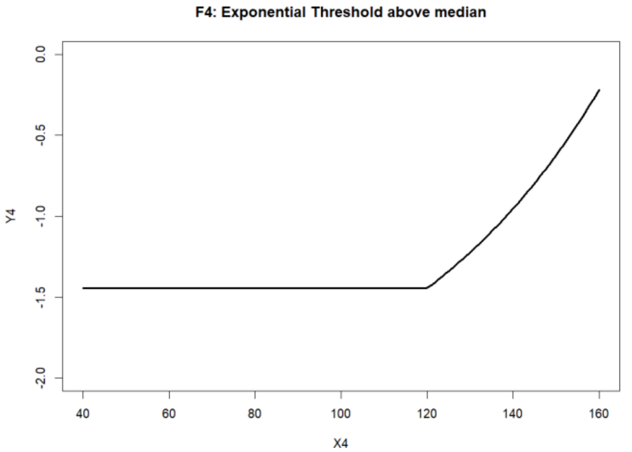
$\text{logit}[P(Y=1)] = -2.202 + 268\exp(-0.383X) + 0.00197\exp(0.139X),\ \textit{with } X \sim N(3.3, 0.25^2)$



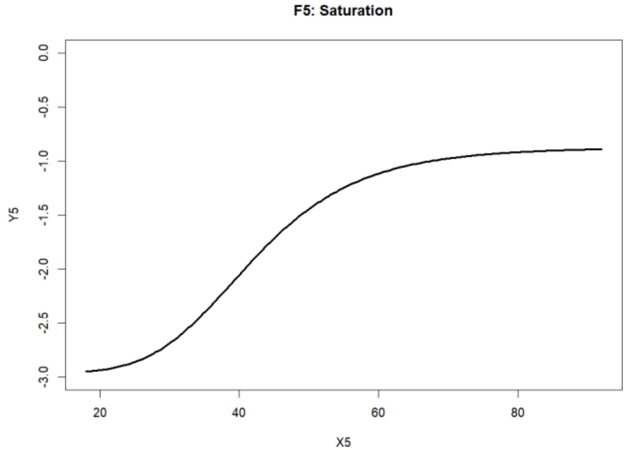
$\text{logit}[P(Y=1)] = -5.78 + 0.545\,X,$
$\textit{with } X \sim N(3.29, 0.24^2)$



$\text{logit}[P(Y=1)] = -2.2 + 0.0135T(X-90),$
$\textit{with } T(W) = W1_{\{W\geq 0\}} \textit{ and } X \sim N(100, 19.4^2)$



$\text{logit}[P(Y=1)] = -3.2 + \exp(\,0.02T(X-120),$
$\textit{with } T(W) = W1_{\{W\geq 0\}} \textit{ and } \ln(X) \sim N(4.5\ 0.23^2)$

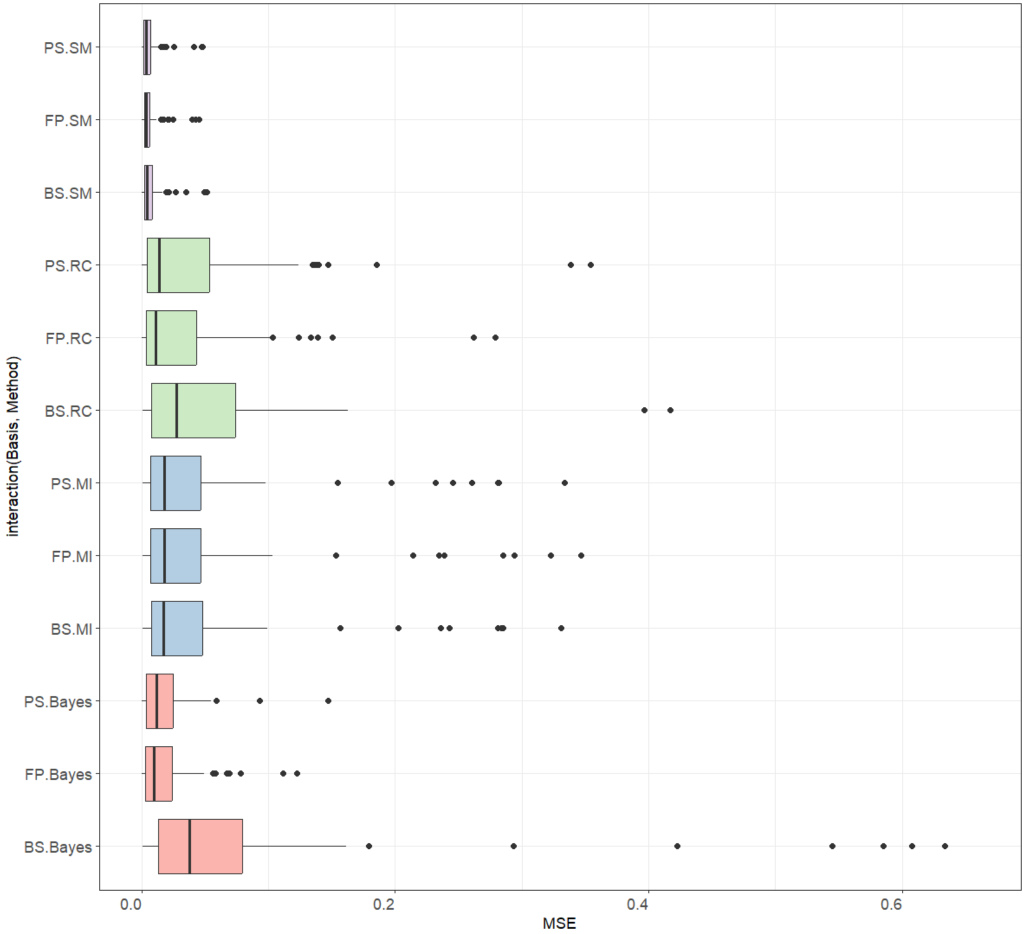


$$logit[P(Y=1)] = \ln\left(\frac{3X^6}{7} + \frac{50^6}{19}\right) - \ln(50^6 + X^6)$$
$$\textit{with } X \sim U(30{,}80)$$

Results Phase 2: MSE

Methods: Multiple Imputation (MI), Regression Calibration (RC), Bayes logit of posterior mean, Pointwise SIMEX.
Findings: SIMEX methods were most accurate, followed by Bayes FP methods, with MI and RC performing similarly. Bayes BS have shown some outliers.

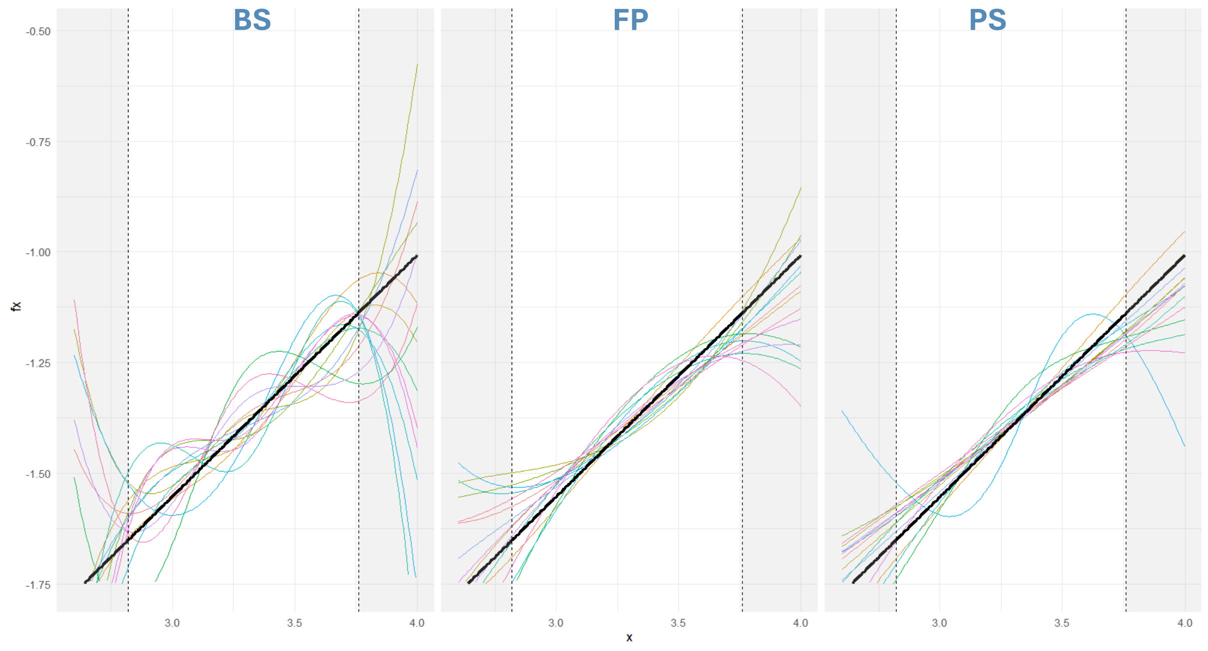
Method	Data 1	Data 2	Data 3	Data 4	Data 5	Average
SIMEX-PS	0.006	0.001	0.005	0.004	0.018	0.0065
SIMEX-FP	0.004	0.002	0.005	0.004	0.019	0.0065
SIMEX-BS	0.006	0.004	0.006	0.005	0.016	0.0073
Bayes-FP logit	0.047	0.004	0.004	0.02	0.017	0.0183
RC-FP	0.12	0.003	0.005	0.038	0.012	0.0356
RC-PS-reml	0.146	0.005	0.006	0.048	0.016	0.044
MI-PS-reml	0.068	0.035	0.023	0.036	0.072	0.0469
MI-BS	0.068	0.036	0.025	0.036	0.073	0.0473
MI-FP	0.067	0.036	0.025	0.035	0.078	0.0481
RC-BS	0.132	0.021	0.021	0.06	0.031	0.0531
Bayes-BS logit	0.227	0.261	0.051	0.069	0.084	0.1383

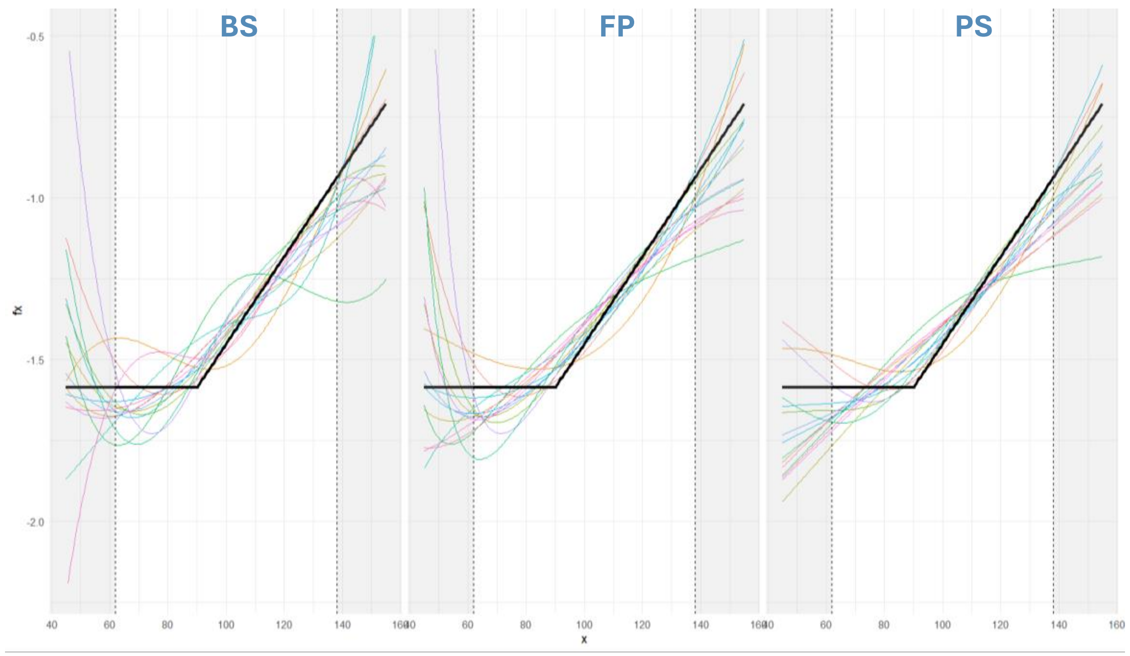


Functional forms with SIMEX

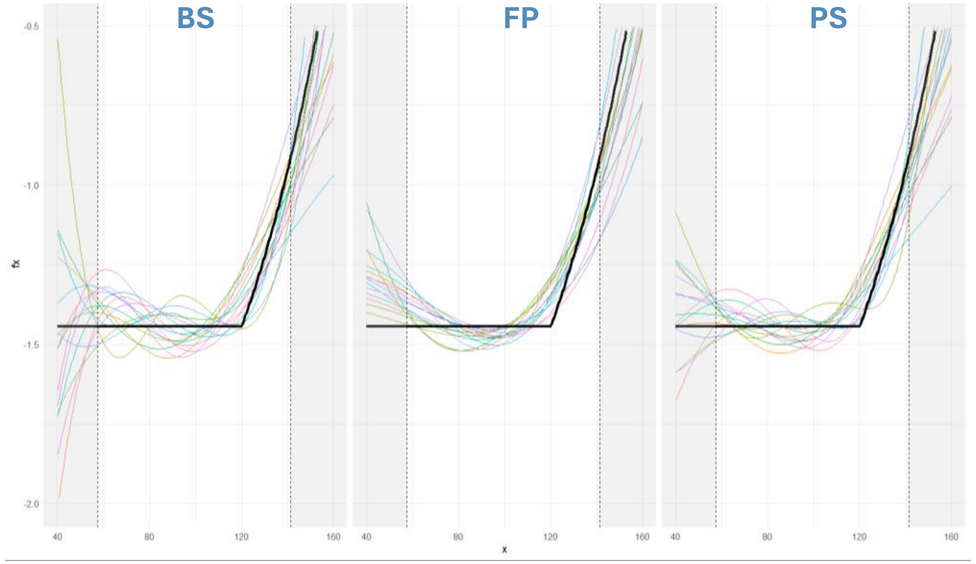
Three levels: B-spline (BS), Fractional Polynomials with 4df (FP) and P-splines (PS).

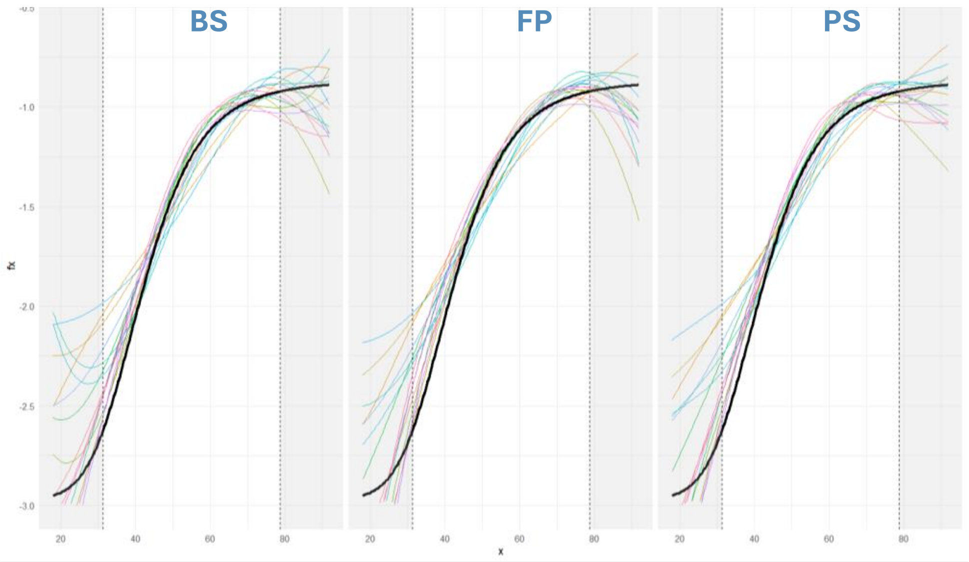
- Overall Fractional Polynomials and P-spline were more accurate than B-splines.
- Linear function (F2) had smaller MAE, followed by change-point below median (F3).

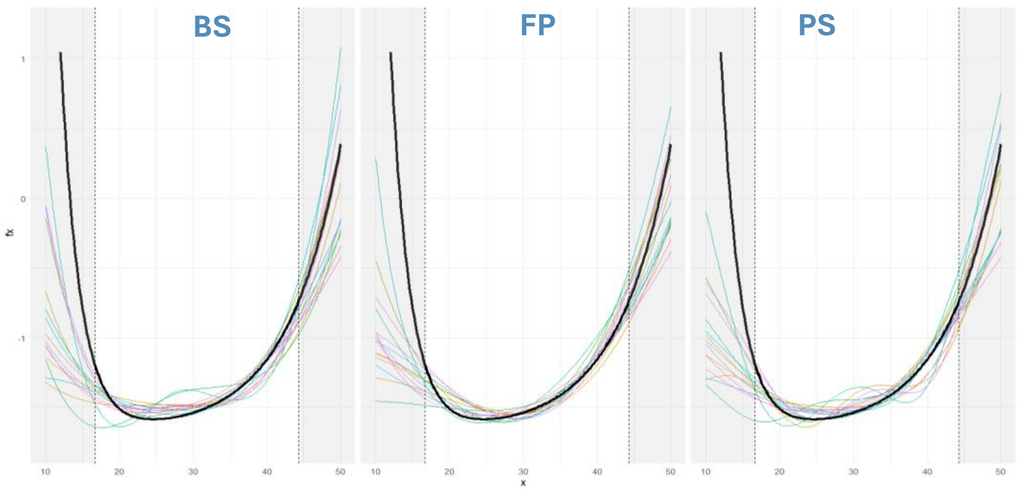




Functional forms with SIMEX







J-shape (F1) had the highest log MAE followed by saturation model (F5) and the threshold model with change-point above the median.

Further Analysis

- Used natural logarithm of Mean Absolute Error (MAE) for prediction accuracy.
- Chosen for its approximately normal distribution across 15 versions of each dataset.
- With a Linear Regression Analysis we analysed the influence of analytic methods and dataset characteristics on log(MAE)
- The model was a linear regression with multiple covariates.
- The covariates were: analytic method, spline method, X-Y relationship, sample size, replicate sample size, error magnitude, error distribution
- Interactions between analytic method and the other covariates were explored

Analysis of Methods, Dataset Characteristics & Interactions

- Measurement error method:
  - SIMEX < {MI, RC, Bayes}
- Functional Form method:
  - {P-S, FP} < B-S
- Combination:
  - SIMEX < Bayes (FP) < {MI, RC} <Bayes (B-S)

Dataset characteristics

- X-Y relationship:
  - Linear < Threshold change-point below median < {Saturation, Threshold change-point above median} < J-shape
- Other characteristics:
  - Main study sample size: 30,000 < 15,000
  - Replicate sub-study sample size: 750 < 250
  - Measurement error magnitude: 0.5*Var(X) < 1.0*Var(X)
  - Measurement error distribution: Shifted-gamma < Normal

Interactions

- While Bayes methods with FPs performed well, they were particularly competitive for the linear model.
- While the SIMEX methods performed better than other methods on most of the datasets, their superiority was less marked for the linear model and the saturation model.
- Regression calibration using p-splines or fractional polynomials exceeded its overall performance when applied to estimating the linear model and the saturation model.
- There were no interactions with main study sample size. That is, across all methods increasing sample size increased the accuracy of estimation by approximately the same order
- For multiple imputation and Bayes methods, larger replicate size increased the accuracy of estimation.
- For regression calibration the increased accuracy was less marked.
- For SIMEX methods larger replicate sample size did not improve the accuracy.

Discussion points and next steps

- The blinded controlled comparison led to unexpected results. In the design stage there was debate over whether it was even worth including the SIMEX method.
- Post-mortem as to why SIMEX performed better than other methods that have better theoretical credentials.
- Extension of the simulations and next steps to be determined on our next meetings