Misclassification in Automated Content Analysis Causes Bias in Regression:

Can We Fix It? Yes We Can!

Nathan TeBlunthuis (nathante@umich.edu)[1]
Valerie Hase [3]
Chung-Hong Chan [4]

August 19 2023

University of Michigan[1]
Northwestern University [2]
LMU Munich[3]
GESIS[4]

Introduction

I'm an *computational social scientist* doing interdisciplinary research about how people organize to produce *public goods* and do *collective action* online and how technology can support them.

2023: Postdoc at the School of Information at the University of Michigan

Introduction

I'm an *computational social scientist* doing interdisciplinary research about how people organize to produce *public goods* and do *collective action* online and how technology can support them.

2023: Postdoc at the School of Information at the University of Michigan

2022: Postdoc in Communication Studies at Northwestern University

Introduction

I'm an *computational social scientist* doing interdisciplinary research about how people organize to produce *public goods* and do *collective action* online and how technology can support them.

2023: Postdoc at the School of Information at the University of Michigan

2022: Postdoc in Communication Studies at Northwestern University

2021: PhD Communication at the University of Washington

Introduction

I'm an *computational social scientist* doing interdisciplinary research about how people organize to produce *public goods* and do *collective action* online and how technology can support them.

2023: Postdoc at the School of Information at the University of Michigan

2022: Postdoc in Communication Studies at Northwestern University

2021: PhD Communication at the University of Washington

Our paper

arXiv > cs > arXiv:2307.06483

Search...

Help | Advanced

Computer Science > Machine Learning

[Submitted on 12 Jul 2023]

Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can!

Nathan TeBlunthuis, Valerie Hase, Chung-Hong Chan

Automated classifiers (ACs), often built via supervised machine learning (SML), can categorize large, statistically powerful samples of data ranging from text to images and video, and have become widely popular measurement devices in communication science and related fields. Despite this popularity, even highly accurate classifiers make errors that cause misclassification bias and misleading results in downstream analyses-unless such analyses account for these errors. As we show in a systematic literature review of SML applications, communication scholars largely ignore misclassification bias. In principle, existing statistical methods can use "gold standard" validation data, such as that created by human annotators, to correct misclassification bias and produce consistent estimates. We introduce and test such methods, including a new method we design and implement in the R package misclassificationmodels, via Monte Carlo simulations designed to reveal each method's limitations, which we also release. Based on our results, we recommend our new error correction method as it is versatile and efficient. In sum, automated classifiers, even those below common accuracy standards or making systematic misclassifications, can be useful for measurement with careful study design and appropriate error correction methods.

Misclassification Bias in Automated Content Analysis (ACA)

Automated content analysis (ACA): Use supervised machine learning to measure variables in text or other human-interpretable high-dimensional data.

Misclassification Bias in Automated Content Analysis (ACA)

Automated content analysis (ACA): Use supervised machine learning to measure variables in text or other human-interpretable high-dimensional data.

Machine learning misclassification causes bias in statistical analysis of such variables. *But this is rarely acknowledged!*

Misclassification Bias in Automated Content Analysis (ACA)

Automated content analysis (ACA): Use supervised machine learning to measure variables in text or other human-interpretable high-dimensional data.

Machine learning misclassification causes bias in statistical analysis of such variables. *But this is rarely acknowledged!*

Current best practice is transparency (e.g., precision; recall; F1 scores).

Misclassification Bias in Automated Content Analysis (ACA)

Automated content analysis (ACA): Use supervised machine learning to measure variables in text or other human-interpretable high-dimensional data.

Machine learning misclassification causes bias in statistical analysis of such variables. *But this is rarely acknowledged!*

Current best practice is transparency (e.g., precision; recall; F1 scores).

We can do better!
Our statistical methodology can use validation data to correct misclassification bias.

I'm going to show the only method that works well for random and non-random errors in dependent and independent variables.

What's wrong?

We want to estimate $Y \sim X\beta_X + Z\beta_Z$, but we use $W = \begin{cases} X & \text{if classifier is right} \\ \neg X & \text{if classifier is wrong} \end{cases}$.

In general, $\beta_W \neq \beta_X$.

Real Data Example

Civil comments: Dataset of 448,000 comments annotated for *toxicity* and *identity disclosure*, additional info on *likes*.

Real Data Example

Civil comments: Dataset of 448,000 comments annotated for *toxicity* and *identity disclosure*, additional info on *likes*.

Jigsaw Perspective API has F1 = 0.79.

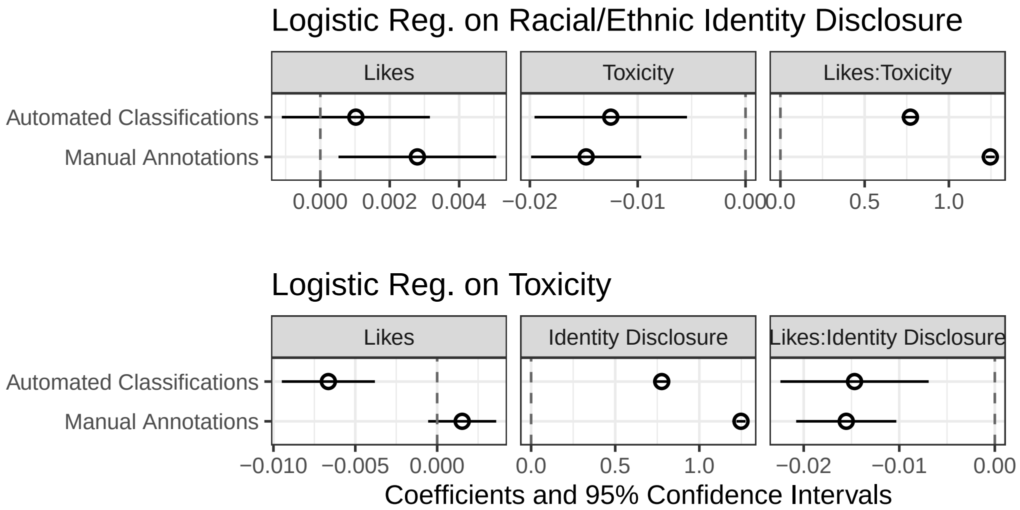
Real Data Example

Civil comments: Dataset of 448,000 comments annotated for *toxicity* and *identity disclosure*, additional info on *likes*.

Jigsaw Perspective API has F1 = 0.79.

We test logistic regression models with *toxicity* predicted or annotated.

Real Data Example



Transparency about Misclassification is Not Enough!

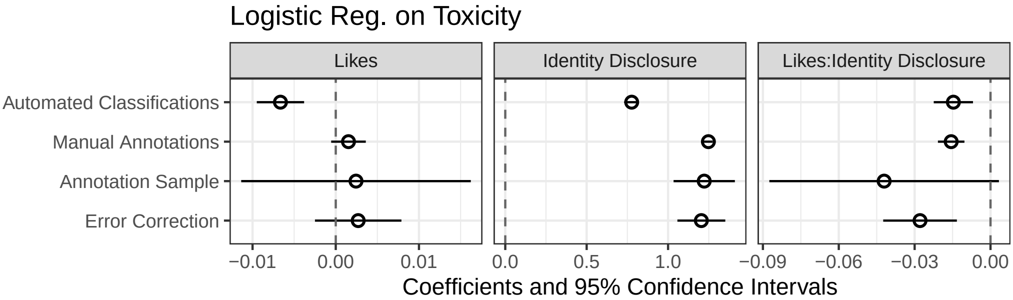

CAN WE
FIX IT?

Fixing it with `misclassificationmodels`: (IV Case)

The civil comments dataset has 488,000 observations! What if we can only afford 10,000?

```
fixed <- glm_fixit(race_disclosed ~ toxicity_coded || toxicity_pred * likes
                   data = perspective.data,
                   data2 = validation.data,
                   proxy_formula = toxicity_pred~toxicity_coded
                                                  *race_disclosed,
                   proxy_family = binomial(),
                   truth_formula = toxicity_coded ~ 1,
                   truth_family = binomial())
```

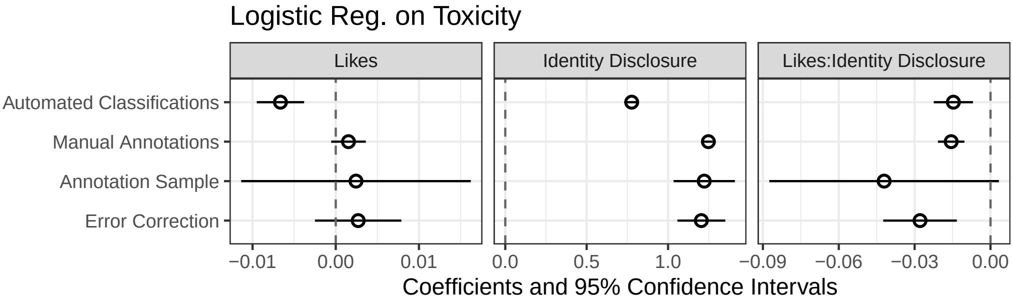
Fixing it with `misclassificationmodels`: (IV Case)



Fixing it with misclassificationmodels: (DV Case)

```
fixed <- glm_fixit_dv(toxicity_coded || toxicity_pred ~ likes*race_disclosed
                      data = perspective.data,
                      data2 = validation.data,
                      proxy_formula = toxicity_pred~toxicity_coded
                                                   *race_disclosed
                                                   *likes,
                      proxy_family = binomial(),
                      truth_formula = toxicity_coded ~ 1,
                      truth_family = binomial())
```

Fixing it with `misclassificationmodels`: (DV Case)



How does it work?

Notation

Y Dependent variable (dv) aka Outcome.

Notation

Y Dependent variable (dv) aka Outcome.

X Independent variable (iv).

Notation

Y Dependent variable (dv) aka Outcome.

X Independent variable (iv).

W Automatic classifications of either X or Y.

Notation

Y Dependent variable (dv) aka Outcome.

X Independent variable (iv).

W Automatic classifications of either X or Y.

Z Other observable variable(s). Important because misclassification of X can affect estimates of β_Z.

Notation

Y Dependent variable (dv) aka Outcome.

X Independent variable (iv).

W Automatic classifications of either X or Y.

Z Other observable variable(s). Important because misclassification of X can affect estimates of β_Z.

X The variable measured via an AC.

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.*

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.*

Use *manual annotations* to model the automatic classifications .

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.*

Use *manual annotations* to model the automatic classifications .

This requires specifying 2-3 models:

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.[*]

Use *manual annotations* to model the automatic classifications .

This requires specifying 2-3 models:

Main model: Outcome Y given X, Z (e.g., $Y = B_0 + B_1 X + B_2 Z + \varepsilon$).
Proxy Model: Automatic classifications W given X, Y, Z.
Truth Model: Annotations X (only needed in the IV case).

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.*

Use *manual annotations* to model the automatic classifications .

This requires specifying 2-3 models:

Main model: Outcome Y given X, Z (e.g., $Y = B_0 + B_1X + B_2Z + \varepsilon$).
Proxy Model: Automatic classifications W given X, Y, Z.
Truth Model: Annotations X (only needed in the IV case).

"Integrate out" missing annotations.

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Maximum Likelihood Adjustment for Misclassification

Our proposed approach implements a framework drawn from biostats.*

Use *manual annotations* to model the automatic classifications .

This requires specifying 2-3 models:

Main model: Outcome Y given X, Z (e.g., $Y = B_0 + B_1X + B_2Z + \varepsilon$).
Proxy Model: Automatic classifications W given X, Y, Z.
Truth Model: Annotations X (only needed in the IV case).

"Integrate out" missing annotations. g Jointly fit the product of the three models via maximum likelihood (MLE). If the models are valid, statistical theory promises *consistent* estimates.

More info on why this works on backup slides.

[*Carroll et al., *Measurement Error in Nonlinear Models*]

Methods from Prior Work in Social Science

Regression Calibration via Generalized Method of Moments (GMM)*

Use *manual annotations* to calibrate predictions; IV only.

Multiple Imputation (MI)†

Use predictions to impute *manual annotations*; IV or DV.

Pseudo-likelihood (PL)‡

Use *precision and recall* to model misclassification; IV or DV.

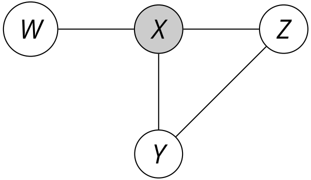
---

*(Fong and Tyler, "Machine Learning Predictions as Regression Covariates")

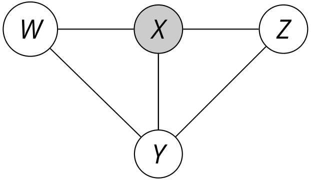
†(Blackwell, Honaker, and King, "A Unified Approach to Measurement Error and Missing Data")

‡(Zhang, *How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It*)

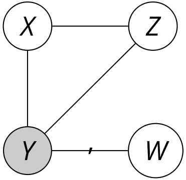
Types of Misclassification Bias



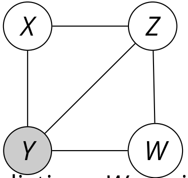
(a) When classifications *W* are independent of *Y* given *X*, a model using *W* has *non-differential error*.



(b) When classifications *W* depend on *Y* given *X*, we have *differential error*.

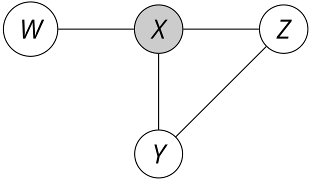


(c) An unbiased classifier measuring the outcome makes *nonsystematic* errors.

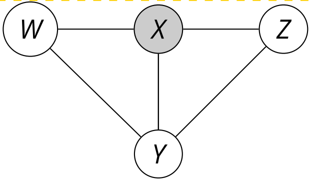


(d) When predictions *W* are independent of *Z* given *Y*, misclassification is *systematic*.

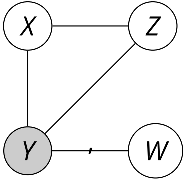
Types of Misclassification Bias



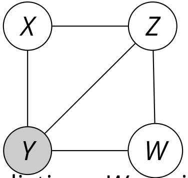
(a) When classifications W are independent of Y given X, a model using W has *non-differential error*.



(b) When classifications W depend on Y given X, we have *differential error*.



(c) An unbiased classifier measuring the outcome makes *nonsystematic* errors.



(d) When predictions W are independent of Z given Y, misclassification is *systematic*.

Evaluating Methods to Correct Misclassification Bias

We want a method that provides:

- Consistent estimates.

Evaluating Methods to Correct Misclassification Bias

We want a method that provides:

- Consistent estimates.
- Accurate uncertainty quantification.

Evaluating Methods to Correct Misclassification Bias

We want a method that provides:

- Consistent estimates.
- Accurate uncertainty quantification.
- Whether an AC measures X or Y.

Evaluating Methods to Correct Misclassification Bias

We want a method that provides:

- Consistent estimates.
- Accurate uncertainty quantification.
- Whether an AC measures X or Y.
- When misclassification is *differential* or *systematic*.

Monte-carlo simulation

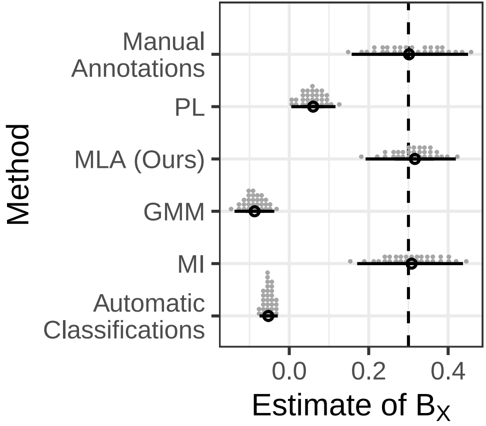
Low-medium accuracy classifier ($\sim$ 73% accuracy).

Relatively large effect sizes.

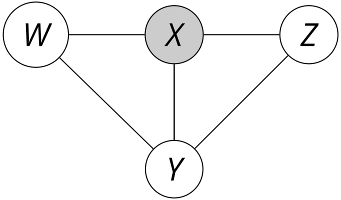
Reletively large dataset ($N = 10,000$)

Decent sample of validation data (200 labels).

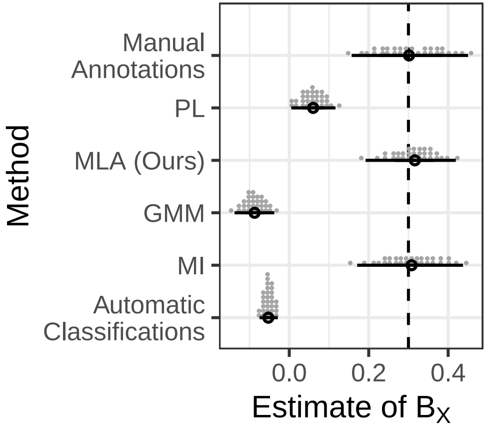
Results: IV Case



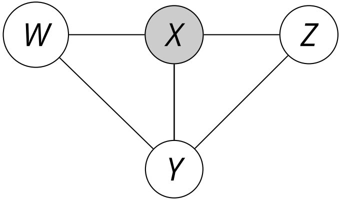
Automatic classifications are wrong and confident!



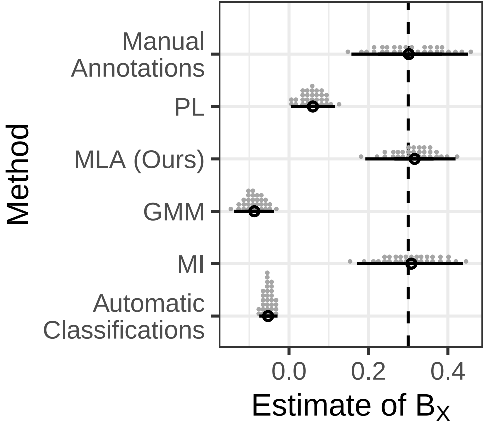
Results: IV Case



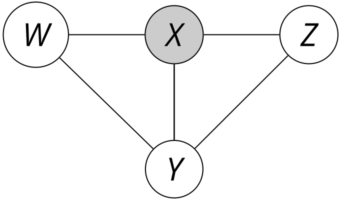
Automatic classifications are wrong and confident!
PL & *GMM* can't fix differential error.



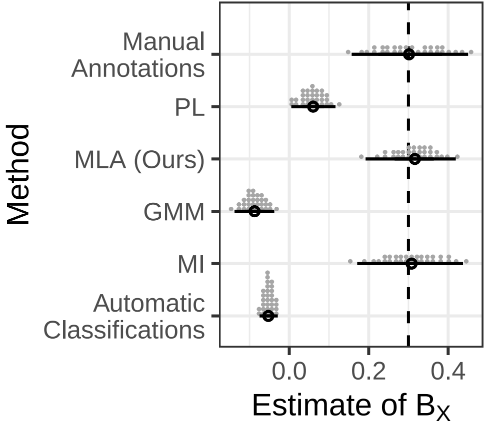
Results: IV Case



Automatic classifications are wrong and confident!
PL & *GMM* can't fix differential error.
MI is effective! But inefficient.



Results: IV Case

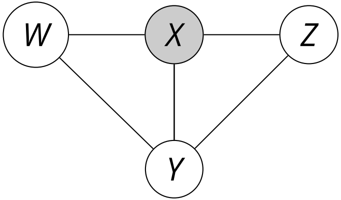


Automatic classifications are wrong and confident!

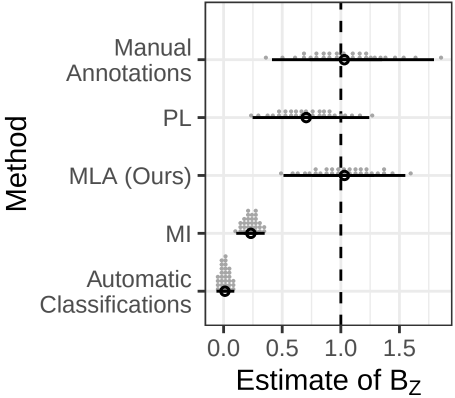
PL & *GMM* can't fix differential error.

MI is effective! But inefficient.

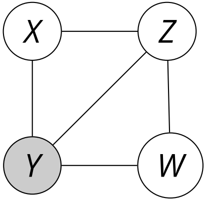
MLA (ours) is more efficient!



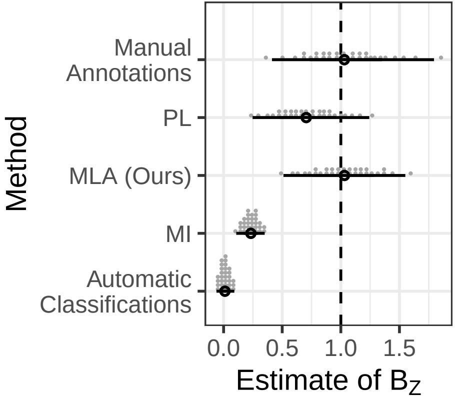
Results: DV Case



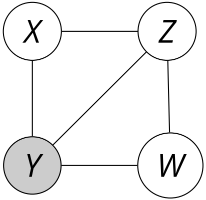
Automatic Classifications are wrong and confident!



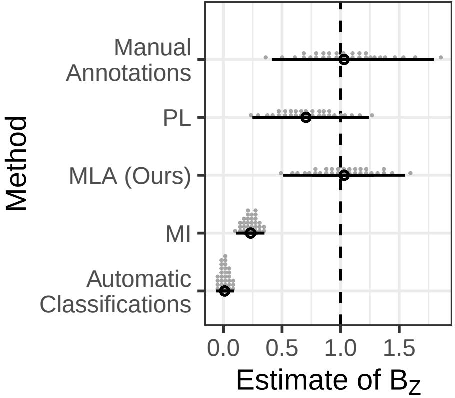
Results: DV Case



Automatic Classifications are wrong and confident!
PL can't fix systematic error.



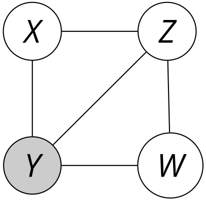
Results: DV Case



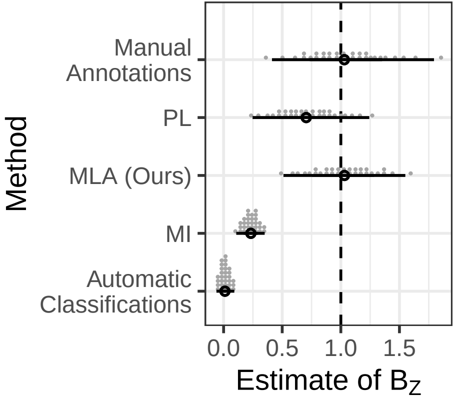
Automatic Classifications are wrong and confident!

PL can't fix systematic error.

MI doesn't work this time.



Results: DV Case

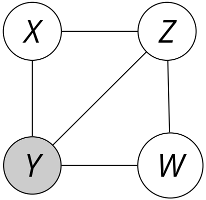


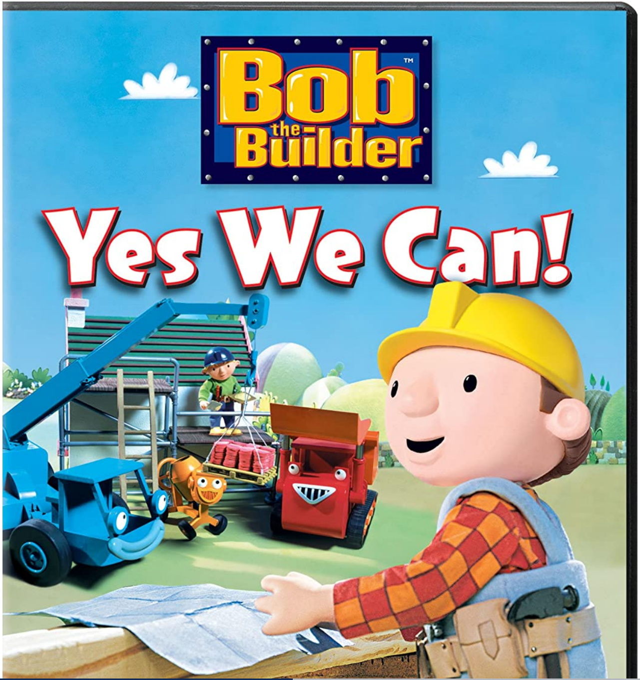
Automatic Classifications are wrong and confident!

PL can't fix systematic error.

MI doesn't work this time.

MLA (ours) is the only consistent method.




Bob the Builder
Yes We Can!

Limitations

MLA requires a correct model for W. This is possible when SML features are fully observed.

*see Bachl and Scharkow, "Correcting Measurement Error in Content Analysis"

Limitations

MLA requires a correct model for W. This is possible when SML features are fully observed.

All correction methods assume error-free ground truth, but human annotators also make errors. Future work can try to account for both sources of error. *

*see Bachl and Scharkow, "Correcting Measurement Error in Content Analysis"

Try our package:

My website:

https://teblunthuis.cc

Read our paper:

Backup Slides

MLA Derivation: IV Case

Following Carroll et al., *Measurement Error in Nonlinear Models*. Maximizing the joint likelihood given Y and W $L(Theta|Y, W)$ suffices to adjust for misclassification error. We only observe X sometimes, but we can integrate it out when missing.

$$
\begin{aligned}
P(Y, W) &= \sum_x P(Y, W, X = x) && (1)\\
&= \sum_x P(Y|W, X = x)P(W, X = x) && (2)\\
&= \sum_x P(Y, X = x)P(W|Y, X = x) && (3)\\
&= \sum_x P(Y|X = x)P(W|Y, X = x)P(X = x) && (4)
\end{aligned}
$$

MLA Specification: IV Case

Consider the regression model $Y = B_0 + B_1 X + B_2 Z + \varepsilon$ and automated classifications W of the independent variable X. We can assume that the probability of W follows a logistic regression model of Y, X, and Z and that the probability of X follows a logistic regression model of Z. In this case, the likelihood model below is sufficient to consistently estimate the parameters: $\Theta = \{\Theta_Y, \Theta_W, \Theta_X\} = \{\{B_0, B_1, B_2\}, \{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}, \{\gamma_0, \gamma_1\}\}$.

$$\mathcal{L}(\Theta; y, w) = \prod_{i=0}^{N} \sum_{x} f_{\Theta_Y}(y_i | x_i, z_i; \Theta_Y) p_{\Theta_W}(w_i | x_i, y_i, z_i; \Theta_W) p_{\Theta_X}(x_i | Z_i; \Theta_X) \quad (5)$$

$$f_{\Theta_Y}(y_i | x_i, z_i; \Theta_Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{(y_i - (B_0 + B_1 x_i + B_2 z_i))^2}{\sigma})^2} \quad (6)$$

$$p_{Theta_W}(w_i | x_i, y_i, z_i; \Theta_W) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 y_i + \alpha_2 x_i + \alpha_3 z_i)}} \quad (7)$$

$$p_{Theta_X}(x_i | z_i; \Theta_X) = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 z_i)}} \quad (8)$$

MLA Derivation: DV Case

As with the IV case, maximizing $\mathcal{L}(\Theta|Y, W)$, the joint likelihood of the parameters Θ given the outcome Y and automated classifications W measuring the dependent variable Y (Carroll et al., *Measurement Error in Nonlinear Models*). We just need to integrate out the missing Y.

$$P(Y, W) = \sum_{y} P(Y = y, W) \tag{9}$$

$$= \sum_{y} P(Y)P(W|Y) \tag{10}$$

MLA Specification: DV Case

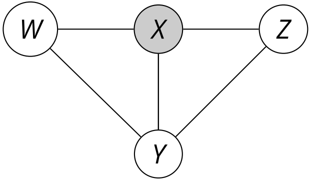
If we assume that the probability of Y follows a logistic regression model of X and Z and allow W to be biased and to directly depend on X and Z, then maximizing the following likelihood is sufficient to consistently estimate the parameters $\Theta = \{\Theta_Y, \Theta_W\} = \{\{B_0, B_1, B_2\}, \{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}\}$.

$$\mathcal{L}(\Theta; y, w) = \prod_{i=0}^{N} \sum_{x} p_{\Theta_Y}(y_i | x_i, z_i; \Theta_Y) p_{\Theta_W}(w_i | x_i, z_i, y_i; \Theta_W) \tag{11}$$

$$p_{\Theta_Y}(y_i | x_i, z_i; \Theta_Y) = \frac{1}{1 + e^{-(B_0 + B_1 x_i + B_2 z_i)}} \tag{12}$$

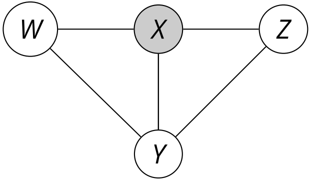
$$p_{\Theta_W}(w_i | y_i, x_i; Z_i, \Theta_W) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 y_i + \alpha_2 x_i + \alpha_3 z_i)}} \tag{13}$$

Simulation Study: IV Case



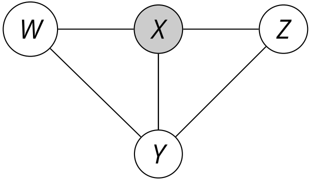
Y and Z are normally distributed.

Simulation Study: IV Case



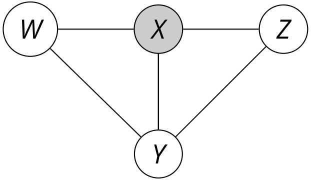
Y and Z are normally distributed.
X is binary with $P(X) = 0.5$, observed 200 times, missing at random (MAR), correlated with Z ($\rho = 0.24$).

Simulation Study: IV Case



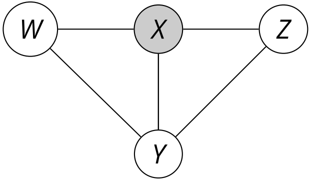
Y and Z are normally distributed.
X is binary with $P(X) = 0.5$, observed 200 times, missing at random (MAR), correlated with Z ($\rho = 0.24$).
W is a classifier predicting X with $\sim 0.73\%$ accuracy.

Simulation Study: IV Case



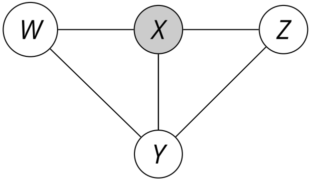
Y and Z are normally distributed.
X is binary with $P(X) = 0.5$, observed 200 times, missing at random (MAR), correlated with Z ($\rho = 0.24$).
W is a classifier predicting X with $\sim 0.73\%$ accuracy.
W's errors correlate with Y ($\rho = -0.17$).

Simulation Study: IV Case



Y and Z are normally distributed.
X is binary with $P(X) = 0.5$, observed 200 times, missing at random (MAR), correlated with Z ($\rho = 0.24$).
W is a classifier predicting X with $\sim 0.73\%$ accuracy.
W's errors correlate with Y ($\rho = -0.17$).
10,000 observations; repeat simulations 500 times.

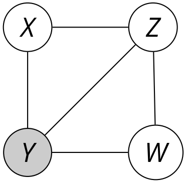
Simulation Study: IV Case



Y and Z are normally distributed.
X is binary with $P(X) = 0.5$, observed 200 times, missing at random (MAR), correlated with Z ($\rho = 0.24$).
W is a classifier predicting X with $\sim 0.73\%$ accuracy.
W's errors correlate with Y ($\rho = -0.17$).
10,000 observations; repeat simulations 500 times.

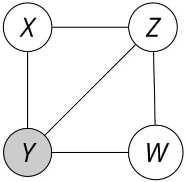
Methods tested: *GMM*, *MI*, *PL*, *MLA*, no correction (*Naïve*), annotated data only (*Feasible*).

Simulation Study: DV Case



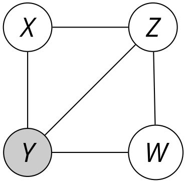
Like the IV case only:
W is a classifier predicting *Y* with $\sim 0.73\%$ accuracy, errors correlate with *Z* ($\rho = 0.2$).

Simulation Study: DV Case



Like the IV case only:
W is a classifier predicting Y with $\sim 0.73\%$ accuracy, errors correlate with Z ($\rho = 0.2$).
Y is binary ($P(Y) = 0.5$); Observed 200 times. MAR.

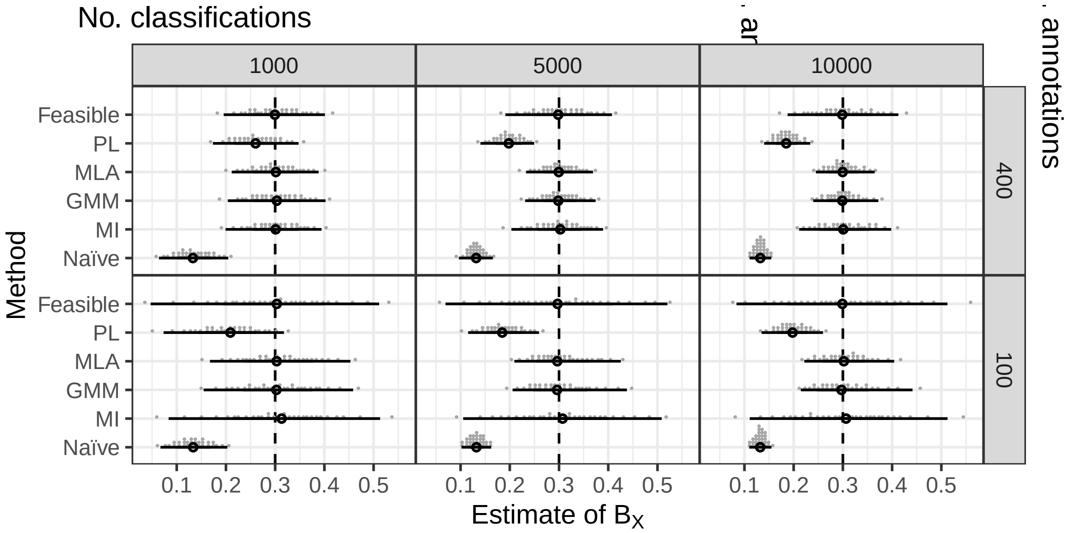
Simulation Study: DV Case



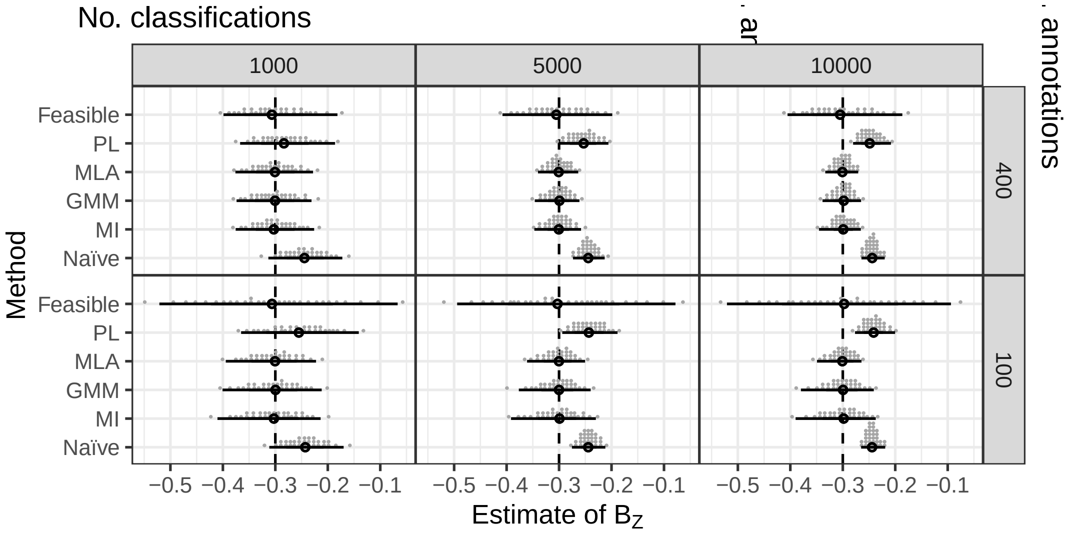
Like the IV case only:
W is a classifier predicting *Y* with $\sim 0.73\%$ accuracy, errors correlate with *Z* ($\rho = 0.2$).
Y is binary ($P(Y) = 0.5$); Observed 200 times. MAR.

Methods tested: *MI*, *PL*, *MLA*, no correction (*Naïve*), annotated data only (*Feasible*).

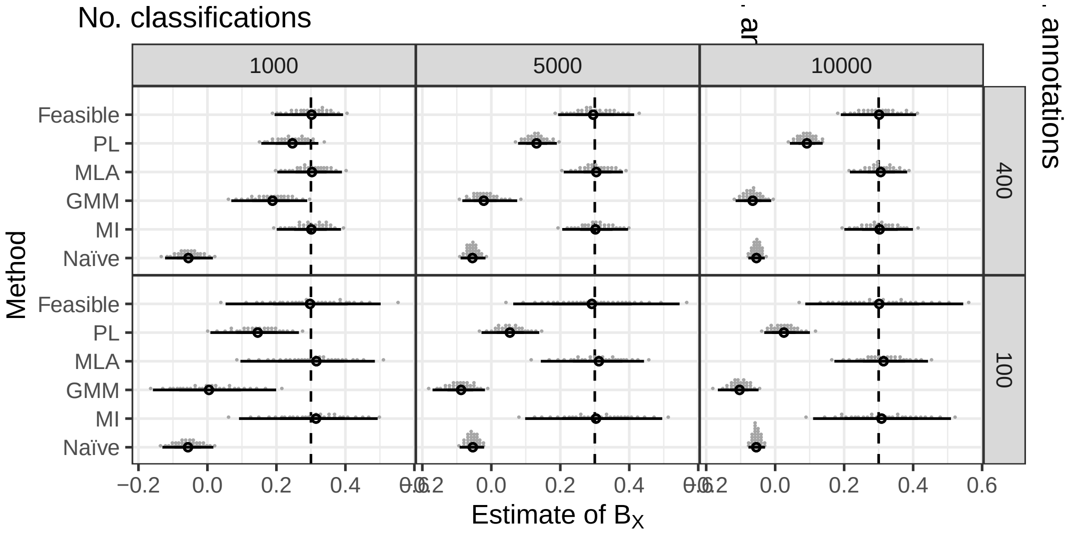
Simulation Results: Non-differential Error



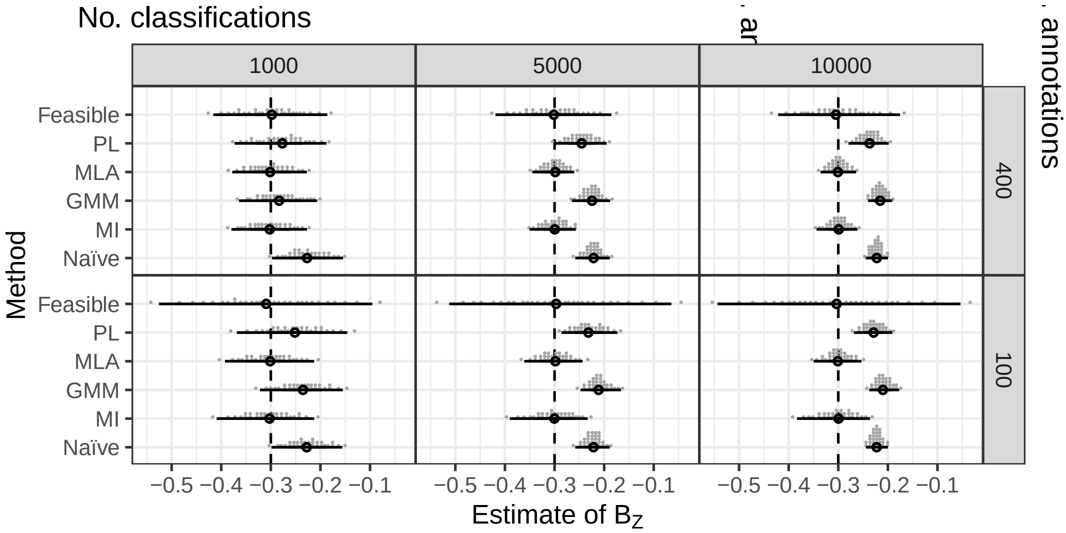
Simulation Results: Non-differential Error (Z)



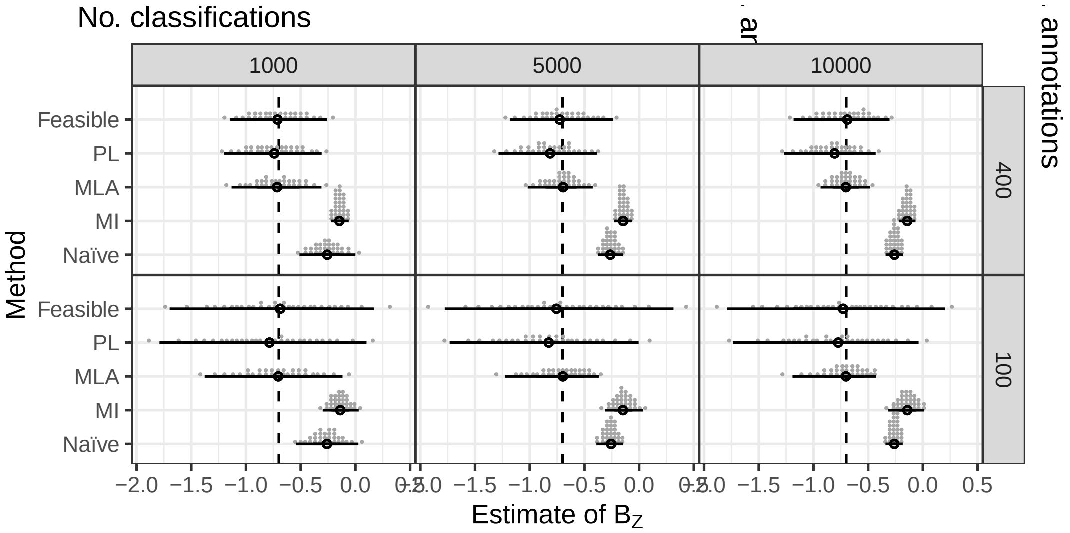
Simulation Results: Differential Error



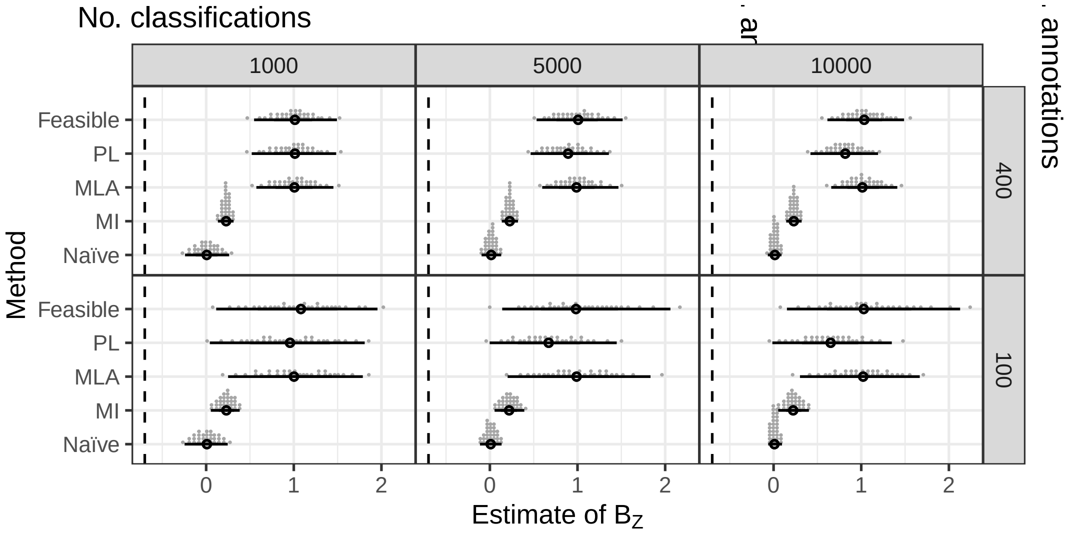
Simulation Results: Differential Error (Z)



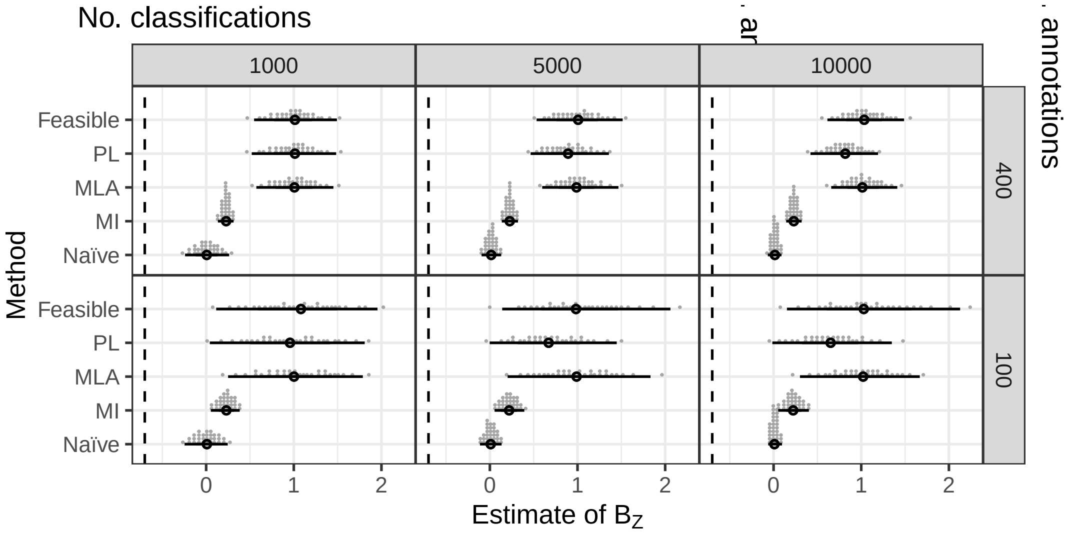
Simulation Results: Nonsystematic Misclassification



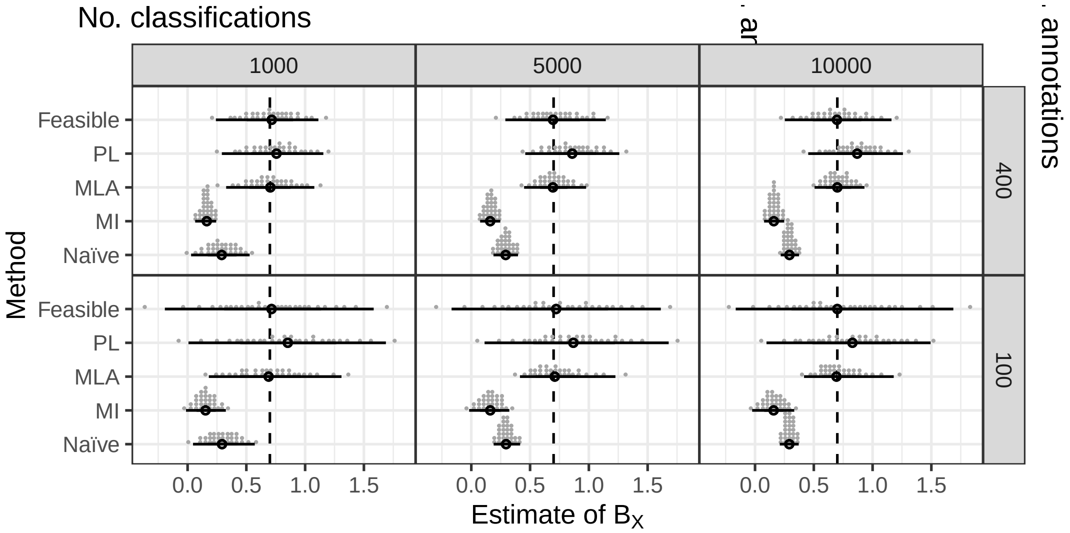
Simulation Results: Nonsystematic Misclassification (X)



Simulation Results: Systematic Misclassification



Simulation Results: Systematic Misclassification (X)



Fixing it with `misclassificationmodels`

```
kable(nrow(research.data))
```

x
4900

```
kable(research.data[1:5,],align='c')
```

y	z	w
-0.16	0	1
-0.33	1	1
0.59	1	0
-0.05	0	0
-0.12	1	1

Fixing it with `misclassificationmodels`

```r
kable(nrow(validation.data))
```

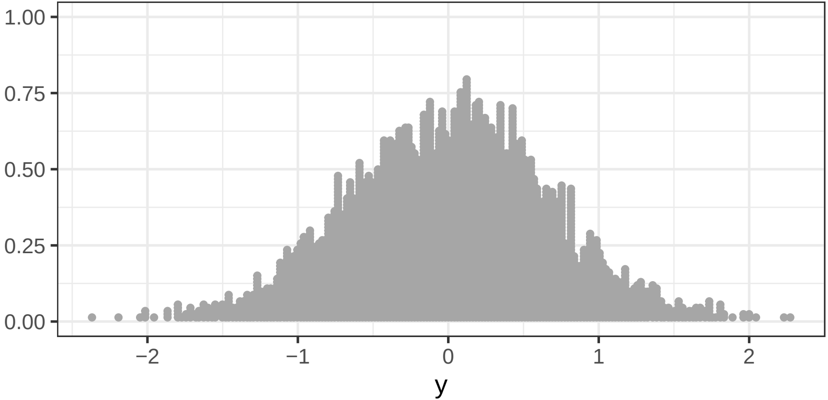
x
100

```r
kable(validation.data[1:5,],align='c')
```

x	y	z	w
0	0.54	0	0
0	-0.51	0	1
1	-0.27	1	1
1	-0.36	1	1
1	0.13	0	1

Fixing it with `misclassificationmodels`

```r
library(ggdist)
ggplot(research.data, aes(x=y)) + geom_dots(binwidth=unit(0.01,"npc"),
                                            overflow='compress')
```



Fixing it with `misclassificationmodels` (Normal data)

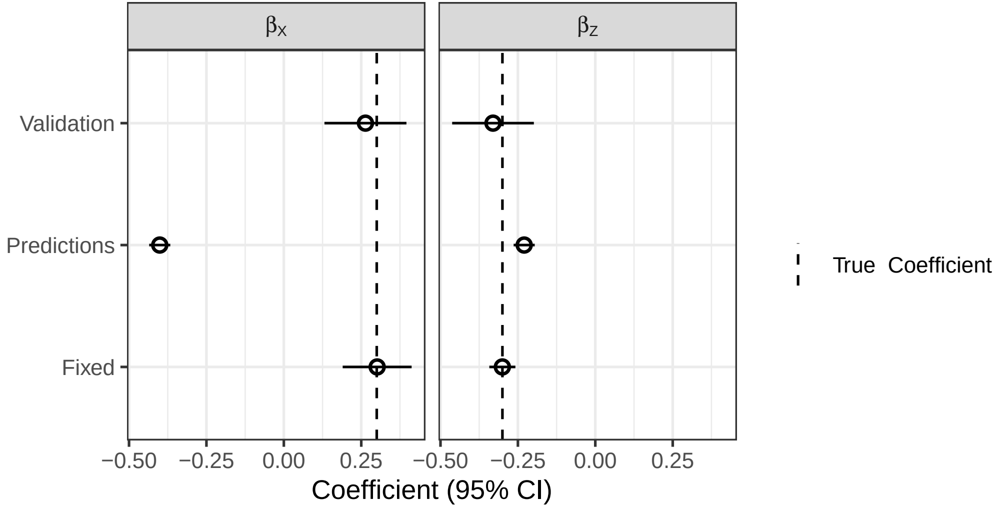
```r
tab <- table(w = validation.data[["w"]], x = validation.data[["x"]])
conmat <- caret::confusionMatrix(tab,mode = "everything",positive = "1")
kable(conmat[["byClass"]][c("Precision",
                            "Recall",
                            "F1",
                            "Sensitivity",
                            "Specificity")])
```

	x
Precision	0.60
Recall	0.56
F1	0.58
Sensitivity	0.56
Specificity	0.57

Fixing it with `misclassificationmodels` (Normal data)

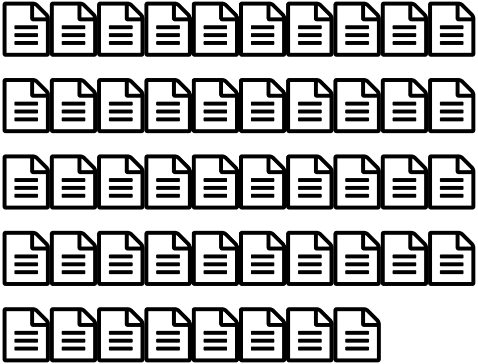
```
validation <- lm(y~x+z,data=validation.data)

predictions <- lm(y~w+z,data=research.data)

fixed <- glm_fixit(formula = y ~ x || w + z,
                   data = research.data,
                   data2 = validation.data,
                   proxy_formula = w ~ x*y*z,
                   proxy_family = binomial(),
                   truth_formula = x ~ z,
                   truth_family = binomial())
```

Fixing it with `misclassificationmodels` (Normal data)



Systematic Literature Review

Systematic Literature Review

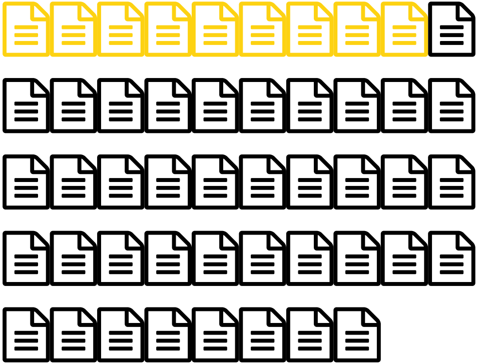


SML-based text-as-data studies (N=48) identified in prior reviews.*

[* Baden et al., “Three Gaps in Computational Text Analysis Methods for Social Sciences”
Hase, Mahl, and Schäfer, “Der „Computational Turn“”
Jünger, Geise, and Hännelt, “Unboxing Computational Social Media Research From a Datahermeneutical Perspective: How Do Scholars Address the Tension Between Automation and Interpretation?”

Song et al., “In Validations We Trust?”]

Systematic Literature Review



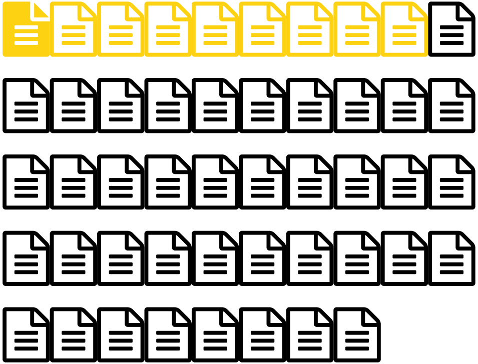
SML-based text-as-data studies (N=48) identified in prior reviews.*

9 mention misclassification as a threat to validity.

[* Baden et al., "Three Gaps in Computational Text Analysis Methods for Social Sciences"
Hase, Mahl, and Schäfer, "Der „Computational Turn""
Jünger, Geise, and Hännelt, "Unboxing Computational Social Media Research From a Datahermeneutical Perspective: How Do Scholars Address the Tension Between Automation and Interpretation?"

Song et al., "In Validations We Trust?"]

Systematic Literature Review



SML-based text-as-data studies (N=48) identified in prior reviews.*

9 mention misclassification as a threat to validity.

1 employs error correction methods.

[* Baden et al., “Three Gaps in Computational Text Analysis Methods for Social Sciences”
Hase, Mahl, and Schäfer, “Der „Computational Turn“”
Jünger, Geise, and Hännelt, “Unboxing Computational Social Media Research From a Datahermeneutical Perspective: How Do Scholars Address the Tension Between Automation and Interpretation?”

Song et al., “In Validations We Trust?”]

Misclassification is a largely ignored threat.