ONLINE Appendix to "Correcting the Measurement Errors of AI-assisted Labeling in Image Analysis using Design-based Supervised Learning"

Alessandra Rister Portinari Maranca* Jihoon Chung† Musashi Hinck‡

Adam D. Wolsky§ Naoki Egami¶ Brandon M. Stewart‖

March 5, 2025

A Model Details

A.1 Surrogate Model specifications

A.1.1 Won et al (2017) application

In this application, we have used the 50-layer ResNet model trained by (Won, Steinert-Threlkeld and Joo, 2017), as a surrogate for the percieved level of violence in the images. This multi-task convolutional neural network was trained on a separate dataset of human annotations on the level of violence of images. Their model architecture consists of 50 convolutional layers with batch normalization and ReLU layers. The authors use mean squared error to train violence dimension.

A.1.2 Casas et al (2019) application

We use GPT-4o in order to generate surrogates for the emotional reactions elicited in the images shared on Twitter OpenAI et al. (2024). We sample 100 images for few-shot training from the human annotations in Casas and Webb Williams (2019) where those 100 images were not used in the final analysis. We used GPT-4o in the 2024-05-13 deployment version. Due to context length constraints, for each query, we randomly sample 20 pairs from the few-shot training set and feed it as context learning. In the style of Maaz et al. (2024), we set the system prompt as:

*Graduate Student in Sociology, Princeton University

†Graduate Student in Computer Science, Princeton University.

‡AI Research Scientist, Intel Labs.

§Senior Research Specialist, Princeton University.

¶Corresponding Author. Assistant Professor, Department of Political Science, Columbia University. Email: naoki.egami@columbia.edu. URL: https://naokiegami.com.

‖Corresponding Author. Associate Professor, Department of Sociology and the Office of Population Research, Princeton University. Email: bms4@princeton.edu. URL: https://brandonstewart.org.

```
You are an intelligent chatbot designed to predict the human emotion evoked
when looking at an image. Your task is to look at an image and predict the type
and intensity of the human emotion it evokes.:
------
INSTRUCTIONS:
- Closely investigate the image.
- Explain what kind of emotion you will evoke from the given image.
```

And set the user prompt as :

```
Please look at the image and indicate the extent to which an image evoked each
of the five emotions: Anger, Enthusiasm, Fear, Sadness, and Disgust
Provide your amount of emotion only as an integer value between 0 and 10,
with 10 indicating the most evoked. Please generate the response in the form of
a Python dictionary string with keys being the emotion.
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python
dictionary string. For example, your response should look like this:
{'anger': 5.0, 'enthusiasm': 4.5, 'fear': 5, 'sadness': 0.5, 'disgust':10.0}.
```

A.1.3 Torres (2024) application

We use the Transformers implementation of CLIP Radford et al. (2021), calculating the score between the image and the text "a photo of a crowd", where higher score means that the image is close to the text. Since CLIP is highly successful with image/text matching Barraco et al. (2022), we are using the CLIP-generated score as a surrogate for the proportion of the topic "crowd" in a given image (similarly to Torres (2024)'s conceptualization). As CLIP-generated scores are unbounded, we rescale them between 0 and 1 in order to match the original paper (using $f(x) = \frac{x - \min(x)}{\max(x) - \min(x)}$).

A.2 Downstream regression specifications

A.2.1 Won et al (2017) application

Unlike the two other applications, this analysis is not in the results published by (Won, Steinert-Threlkeld and Joo, 2017), however, they represent a simple social scientific question that could be derived from their data. For this application, we ran a linear robust model with the following formula `violence` $\sim$ `sign + photo + fire + police + children + group_20 + flag + night + shouting`. Here, `violence` represents the perceived level of violence in a given image, which is both human annotated by (Won, Steinert-Threlkeld and Joo, 2017), and predicted by the author's application of ResNet. The standard error type utilized here is "HC0".

A.2.2 Casas et al (2019) application

The analysis of interest here is the robust linear model given by the following formula `log(retweet_n+1)` $\sim$ `followers_count + friends_count + prev_tweets + time_control + protest + symbol + anger + fear + disgust + sadness + enthusiasm`. The standard error type utilized here is "HC0". It is important to note that here we are taking the $\log(x+1)$ of out outcome variable of interest (number of retweets) and then applying a linear model (in opposition to (Casas and

Webb Williams, 2019)'s application, where they use a negative binomial model. Other than that, the regression ran and the variables controlled for are identical. The only variables being predicted by the GPT4-o surrogate in the analysis are the emotional reactions, namely `anger, fear, disgust, sadness, enthusiasm`.

A.2.3 Torres (2024) application

The analysis performed here is a robust linear model with the formula `crow_proportion` $\sim$ `political_leaning -1`, where `crow_proportion` is predicted using CLIP scores, (Torres, 2024)'s bag of visual words model, or given by our ground truth human annotations. Moreover, `political_leaning` is the categorization made by (Torres, 2024) of the political leaning of the news outlet. The standard error type utilized here is "HC0".

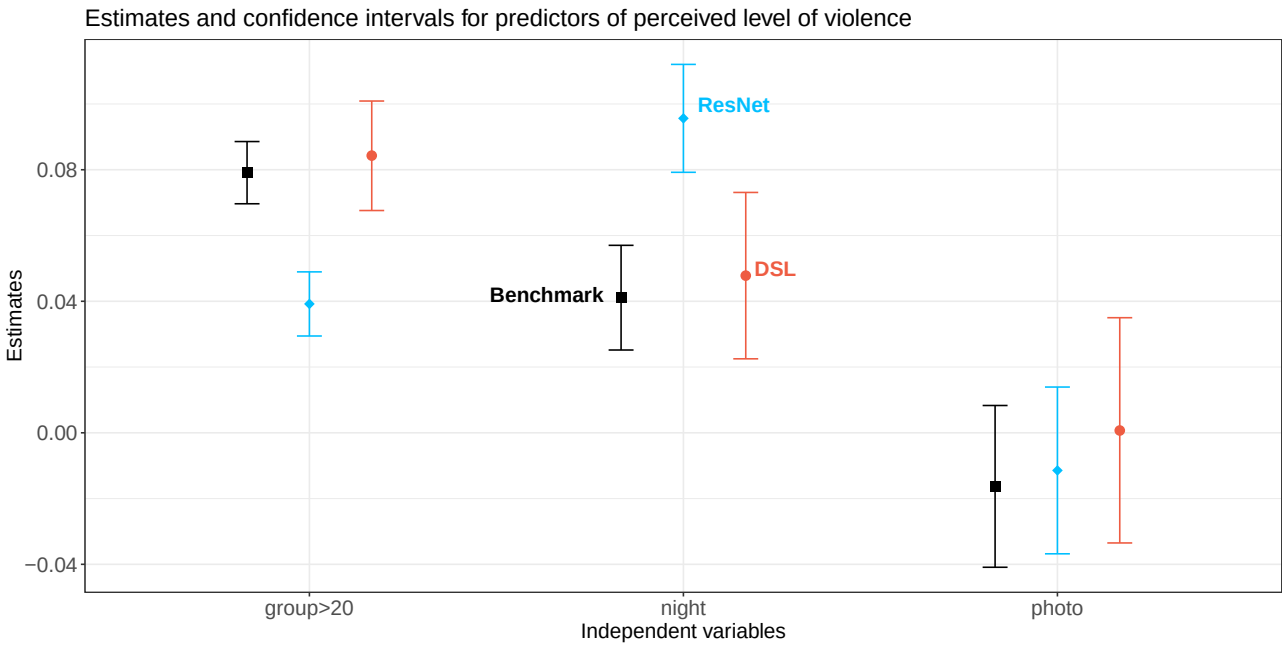

Figure 1: This shows the results of applying DSL once (rather than averaging over many results) to the Won, Steinert-Threlkeld and Joo (2017) application.

B Additional Results

Figures 1–3 contain additional results from the three empirical validations as described in their captions.

Figure 1 shows one application of DSL on the Won et al (2017) empirical validation instead of the avarage of the estimates (in this paper, horizontal estimates represent the result of avaraging 500 applications of DSL, whereas vertical estimates represent the result of applying the method once). We see that the DSL estimate tracks the Benchmark in all three coefficients while the ResNet surrogate underestimates, overestimates, and correctly estimates (overestimating the width of the confidence intervals) respectively for the coefficients showed.

Figure 2 shows the ratio between the DSL confidence intervals and the Oracle (Benchmark) confidence intervals in the Torres (2024) empirical validation for increasing number of labeled annotations used. Above, we see that DSL approximates the ground truth confidence interval width as we add more annotations. Below, we see that as we use more annotations, the marginal effect of using DSL in juxtapoisition to using the labeled annotations only decreases.

Figure 3 shows 50 applications of DSL in comparison to surrogate-only (ResNet), the same sub-sample of labeled annotations used for DSL, and the Benchmark (all labeled annotations) for the variable "police" in the Won et al (2017) empirical validation. We see the trend that DSL approximates the Benchmark better than the sub-sample alone or the surrogate (ResNet) alone.

Table 1 shows the effect of using additional labeled examples to fine-tune a model (and thus perhaps improving predictive capacity of the surrogate) versus using those additional labeled examples directly into DSL for the Casas et al (2019) empirical validation. We have mantained the total number of labeled examples used in the whole process (2000) constant while varying their use. We see that we get the smallest confidence intervals for all the variables in this

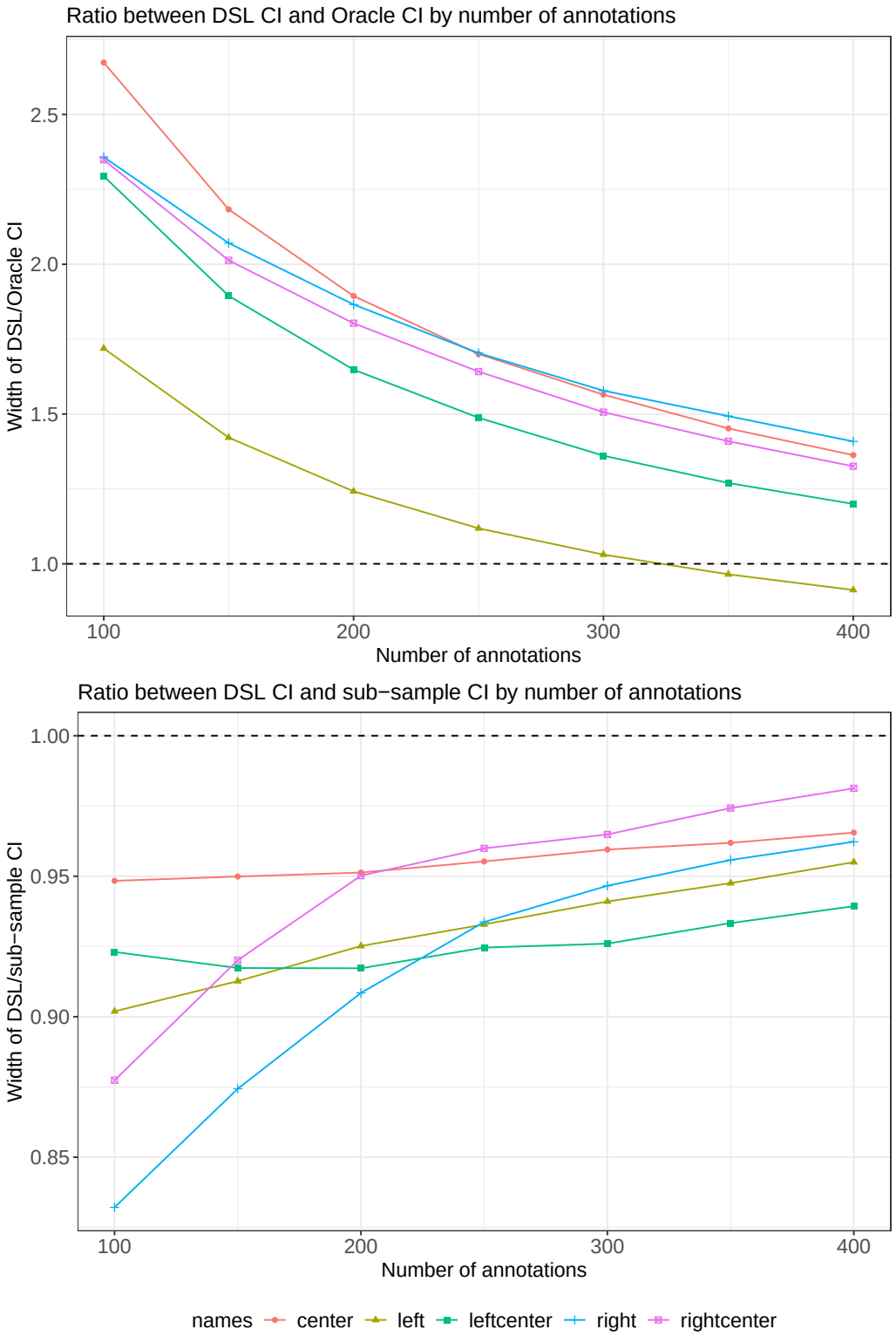

Figure 2: Confidence intervals in Torres (2024) by comparison to the oracle confidence interval and the sub-sample only confidence interval. Over 500 iterations.

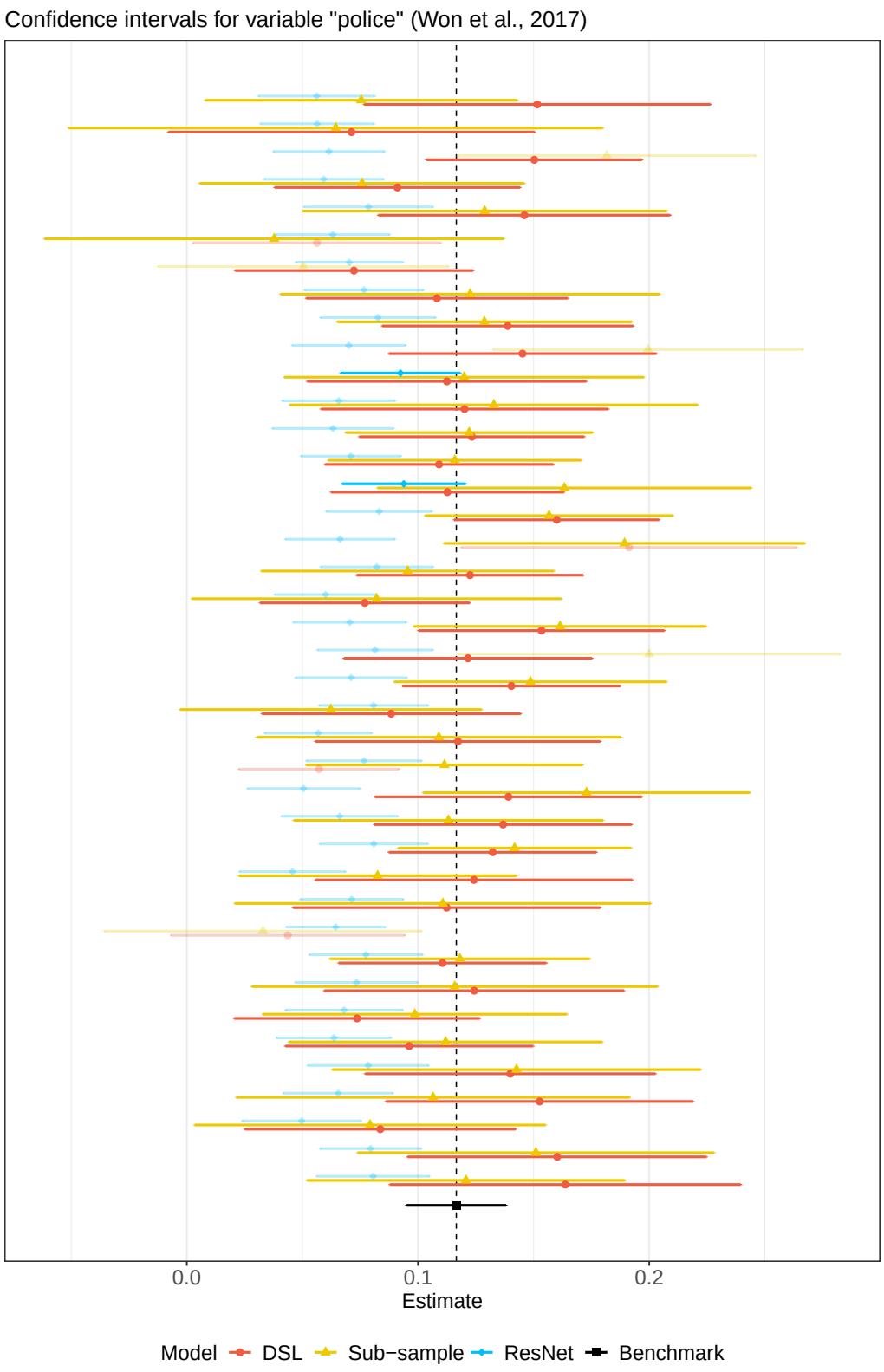

Figure 3: Results of 50 simulations for the variable 'police' in the Won et al empirical validation, for the model using DSL, with the labeled sub-sample observations without using DSL, and compared to ResNet (no labeled observations) and the benchmark (all 2343 labeled observations) applying with 200 labeled observations. In this plot, confidence intervals had their transparency increased if they did not contain the Benchmark estimate.

Type	1000DSL/1000FT	1500DSL/500FT	1900DSL/100FT
average CI width (protest)	0.16	0.13	0.13
average CI width (symbol)	0.21	0.18	0.17
average CI width (anger)	0.66	0.47	0.46
average CI width (fear)	0.70	0.56	0.51
average CI width (disgust)	0.60	0.45	0.43
average CI width (sadness)	0.47	0.34	0.31
average CI width (enthusiasm)	0.42	0.30	0.29
surrogate accuracy (anger)	0.63	0.49	0.70
surrogate accuracy (fear)	0.42	0.41	0.46
surrogate accuracy (disgust)	0.63	0.62	0.68
surrogate accuracy (sadness)	0.63	0.59	0.68
surrogate accuracy (enthusiasm)	0.55	0.72	0.63

Table 1: Results of experiment of maintaining 2000 labeled annotations constant in the Casas et al (2019), and testing whether it is more effective to use them for fine-tuning (FT) to generate more accurate surrogate annotations or to use them directly in DSL.

empirical validation when we minimize the amount of fine-tuning for GPT4-o and maximize the number of labeled examples into DSL (1900 benchmark annotations used for DSL, and 100 benchmark annotations used for fine-tuning). We also see that the surrogate accuracy does not increase linearly with adding more labeled annotations.

C Additional Details on Simulation

C.0.1 Simulation setup

For generating the simulations for Figures **??** and **??** which show the effect of surrogate accuracy on bias, RMSE, and coverage, and the effect of expert annotation accuracy on bias, RMSE, and coverage of DSL estimator (respectively), we have generated synthetic data. This data was generated analogously[1] to Wager and Athey (2018), with $n = 5000$ ('images', in our case), where Y was a smooth function supported on the first two features such as in Equation 1 below, however, we have also added normal error and a linear term of the other covariates.

We ran this simulation with 100 'labeled' examples and 10 covariates[2] $X_1, \ldots, X_{10}$ and introduced the prediction error in the surrogate for X_2. Note our covariates $X_1, \ldots X_{10}$ are initially drawn from a uniform distribution between 0 and 1, and then X_2 is transformed into a binary variable (1 if greater than 0.8, and 0 otherwise). Y, on the other hand, is generated by incorporating non-linear transformations of X_1 and X_2 and added linear contributions from additional covariates, alongside random normal noise. The data generating process of the outcome Y is

[1]To see the simulation this was based on, go to equation (28), p. 1238, in Wager and Athey (2018).

[2]The number of covariates used in the data generating process is 10, but we consider cases when we cannot observe all of them. We only use the first three in the downstream analysis. This shows that we do not need to assume the correct specification of the downstream model.

given by

$$Y = \left(1 + \frac{1}{(1 + e^{-20 \cdot (X_1 - 1/3)}}\right) \cdot \left(1 + \frac{1}{1 + e^{-20 \cdot (X_2 - 1/3)}}\right) + \sum_{i \in \{4,5,6,7,8,9,10\}} X_i + \varepsilon \tag{1}$$

where $\varepsilon \sim N(0, 1)$.

Then, a surrogate for X_2 is created with specified accuracy (denoted `q_acc` in the figures) by probabilistically altering the original X_2 values. Then, random binary errors in the Benchmark labeled examples are introduced to simulate expert annotation errors. For the downstream analysis showed in the figures, we are fitting the following model: $Y \sim X_1 + X_2 + X_3$ where X_2 is the binary surrogate.

C.0.2 Diagnosands

Denoting the simulation iteration as $i \in N$ and the target coefficient as $\beta_{\text{Estim.}}$, the diagnosands in our simulations are calculated as follows:

- **Mean Absolute Bias**: the average absolute difference between the estimate and true parameter value ($\frac{1}{N} \sum_i^N abs(\beta_{DSL,i} - \beta^*)$, where i denotes simulation iteration, N denotes the number of simulations and $\beta^* = \frac{1}{N} \sum_i^N \beta_{oracle,i}$)
- **Root Mean Squared Error (RMSE)**: the average RMSE
- **Coverage of 95% Intervals**: the proportion of simulations for which the true parameter value is in the confidence interval provided by the estimator

D Connection to Literature

Theoretically, our paper is most closely related to two recent papers that similarly build on the literature on doubly robust methods. Prediction-powered inference Angelopoulos et al. (2023) provides a similar framework to ours, but they have primarily focused on settings where the outcome variable is predicted while providing both asymptotic and non-asymptotic confidence intervals. Mozer and Miratrix (N.d.) focus on settings where the predicted outcome variable is used within randomized experiments. Methodologically, our paper extends these previous results in three ways. First, while these papers only cover cases of text-based outcome variables, we cover cases where any subset of the outcome and independent variables are text-based. Second, we develop a data-driven power analysis to help users determine the required number of expert annotations. Third, we derive DSL estimators for a much wider range of downstream analyses popular in the social sciences, including linear fixed effects regression and the instrumental variable method. In addition, we make practical contributions by providing new statistical software and clarifying detailed guides using two empirical applications. While they all share the same methodological foundation, each paper contains its own unique contributions tailored to different applications they focus on: our paper focusing on the social sciences, prediction-powered inference focusing mostly on the natural sciences, and model-assisted impact analysis focusing on education and randomized experiments. Given the wide applicability of the shared methodological foundation, we expect more exciting methodological developments that address various application-specific problems. Katsumata and Yamauchi (2023) also develop a framework for using predicted variables while building on a different framework of control variates (Chen and Chen, 2000).

E Power-Tuning

Following Angelopoulos, Duchi and Zrnic (2023), we implement the power-tuning that guarantees that the DSL estimator has a smaller asymptotic variance than the estimator that only uses expert annotations.

Using the simplified notation (see more full results in Egami et al. (2024)), we consider the following general moment condition.

$$m_{\text{DSL}}(D_i, \widehat{D}_i, R_i; \beta, \pi) \coloneqq m(\widehat{D}_i; \beta) - \frac{R_i}{\pi_i}\Big(m(\widehat{D}_i; \beta) - m(D_i; \beta)\Big) \tag{2}$$

where D_i is a vector of observed variables, $\widehat{D}_i$ is a vector of predicted variables, R_i is an indicator binary whether a unit is sampled for expert annotations, and π_i is the known probability of being sampled for expert annotations. Note that some of $\widehat{D}_i$ can be observed and might not need to be predicted.

The power-tuning version (Angelopoulos, Duchi and Zrnic, 2023) has an additional tuning parameter λ.

$$m_{\text{DSL},\lambda}(D_i, \widehat{D}_i, R_i; \beta, \pi) \coloneqq \frac{R_i}{\pi_i} m(D_i; \beta) + \lambda\left(1 - \frac{R_i}{\pi_i}\right) m(\widehat{D}_i; \beta) \tag{3}$$

The asymptotic variance of this estimator is

$$\lambda^2 \times S_V P_1 S_V - \lambda \times S_V P_2 S_V + P_3 \tag{4}$$

where

$$\begin{aligned}
S_V &= \mathbb{E}\left(\frac{\partial m(D_i; \beta^*)}{\partial \beta}\right)^{-1} \\
P_1 &= \mathbb{E}\left(\frac{1}{\pi_i} m(\widehat{D}_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(\widehat{D}_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\Big) \\
P_2 &= \mathbb{E}\left(\frac{1}{\pi_i} m(\widehat{D}_i; \beta^*) m(D_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(\widehat{D}_i; \beta^*) m(D_i; \beta^*)^\top\Big) \\
&\quad + \mathbb{E}\left(\frac{1}{\pi_i} m(D_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\right) - \mathbb{E}\Big(m(D_i; \beta^*) m(\widehat{D}_i; \beta^*)^\top\Big) \\
P_3 &= \mathbb{E}\left(\frac{1}{\pi_i} m(D_i; \beta^*) m(D_i; \beta^*)^\top\right).
\end{aligned}$$

We tune λ to minimize the trace of the asymptotic variance.

$$\lambda^* = \frac{\text{Tr}\{S_V P_2 S_V\}}{2\text{Tr}\{S_V P_1 S_V\}}. \tag{5}$$

References

Angelopoulos, Anastasios N, John C Duchi and Tijana Zrnic. 2023. "PPI++: Efficient Prediction-Powered Inference." *arXiv preprint arXiv:2311.01453* .

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic. 2023. "Prediction-powered inference." *Science* 382(6671):669–674.

Barraco, Manuele, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi and Rita Cucchiara. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4662–4670.

Casas, Andreu and Nora Webb Williams. 2019. "Images that matter: Online protests and the mobilizing role of pictures." *Political Research Quarterly* 72(2):360–375.

Chen, Yi-Hau and Hung Chen. 2000. "A Unified Approach to Regression Analysis under Double-Sampling Designs." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62(3):449–460.

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024. "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses." Working paper.

Katsumata, Hiroto and Soichiro Yamauchi. 2023. "Statistical Analysis with Machine Learning Predicted Variables." Working Paper.

Maaz, Muhammad, Hanoona Rasheed, Salman Khan and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Mozer, Reagan and Luke Miratrix. N.d. "More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials." *Annals of Applied Statistics*. Forthcoming.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt and Sam Altman et al. 2024. "GPT-4 Technical Report." *arXiv:2303.08774*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR pp. 8748–8763.

Torres, Michelle. 2024. "A framework for the unsupervised and semi-supervised analysis of visual frames." *Political Analysis* 32(2):199–220.

Wager, Stefan and Susan Athey. 2018. "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association* 113(523):1228–1242.

Won, Donghyeon, Zachary C Steinert-Threlkeld and Jungseock Joo. 2017. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*. pp. 786–794.