AEA Papers and Proceedings 2026, 116: 92–97
https://doi.org/10.1257/pandp.20261020

Performing Valid Inference with AI/ML-Generated Covariates: A Guide for Empirical Practice†

By Timothy Christensen and Stephen Hansen*

The increasing availability of unstructured data has enabled economists to generate new variables that measure important but difficult-to-quantify phenomena. Prominent examples include measures of economic policy uncertainty (Baker, Bloom, and Davis 2016), geopolitical risk (Caldara and Iacoviello 2022), firm-level exposure to political risk (Hassan et al. 2019), central bank communication (Hansen and McMahon 2016; Gorodnichenko, Pham, and Talavera 2023), and media sentiment (Shapiro, Sudhof, and Wilson 2022). Advances in machine learning (ML) and artificial intelligence (AI) are further accelerating this trend.

The creation of such variables is usually the first step in an empirical pipeline. Researchers often use the generated variables to estimate models like

$$Y_i = \beta_1' \theta_i + \beta_2' X_i + \varepsilon_i, \tag{1}$$

where θ_i are (vectors of) latent variables of economic interest that are proxied by the AI/ML-generated variables, and X_i are (vectors of) observed controls. That is, because θ_i is latent, researchers typically estimate (1) by regressing Y_i on the generated variable $\hat{\theta}_i$ and X_i, treating $\hat{\theta}_i$ as regular numeric data when performing inference. We refer to this as the *two-step strategy*.

A growing literature questions the statistical validity of this strategy. In statistics, Angelopoulos et al. (2023) introduces a framework, *prediction-powered inference* (PPI), that recognizes the bias that arises from ignoring measurement error in $\hat{\theta}_i$ and proposes a correction that relies on validation data. The validation data consists of m observations of $\left(Y_i, \theta_i, \hat{\theta}_i, X_i\right)$; that is, validation data is "complete" in the sense of pairing the true θ_i and prediction $\hat{\theta}_i$, along with Y_i and X_i. But in this case, model (1) could be estimated from the validation data alone; the generated variables only help improve efficiency. Moreover, in the PPI framework, the size of the validation data m is typically assumed to be asymptotically comparable to the original sample size n.

PPI is an important step in developing robust inference approaches. Recently, though, Battaglia et al. (2025)—henceforth, BCHS—argues that complete validation data are often not available in many economic applications because θ_i is never observable. For instance, variables like uncertainty, risk, and sentiment are inherently latent, even to domain experts. Moreover, in the rare cases that validation data can be constructed, it is often through costly human annotation, resulting in a much smaller validation sample size m than the original (unlabeled) sample size n.

BCHS make progress by assuming that the measurement error is of the same asymptotic order as sampling uncertainty in the regression model. This assumption ensures that their asymptotic results provide useful approximations to the finite-sample distribution of regression estimators encountered in practice, where both forces are present. It is also a more robust starting point than the two-step

*Christensen: Yale University (email: timothy.christensen@yale.edu); Hansen: University College London (email: stephen.hansen@ucl.ac.uk). We thank Konrad Kurczynski for excellent research assistance in preparing the software packages in this paper. This material is based upon work supported by the NSF under award 2521471 (Christensen) and by ERC Consolidator Grant 864863 (Hansen).

†Go to https://doi.org/10.1257/pandp.20261020 to visit the article page for additional materials and author disclosure statement(s).

```
from ValidMLInference import ols, ols_bcm, remote_work_data
import numpy as np

# load data and transform salary to log scale
SD_data = remote_work_data()
SD_data['salary'] = np.log(SD_data['salary'])

# regression formula (with occupation and employment-type fixed effects)
formula = "salary ~ wfh_rwo + C(soc_2021_2) + C(employment_type_name)"

# two-step OLS
mod_2step = ols(formula=formula, data=SD_data)

# bias-corrected two-step
mod_2step_bc = ols_bcm(formula=formula, generated_var="wfh_rwo",
                       data=SD_data, fpr=0.009, m=1000)
```

FIGURE 1. BIAS-CORRECTED REGRESSION USING THE VALIDMLINFERENCE PACKAGE (IMPUTED BINARY LABELS)

Notes: The function `ols_bcm` implements the multiplicative bias correction for a two-step regression with a covariate generated by a binary label classifier. The argument `fpr` is the estimated false-positive rate and `m` is the sample size used to estimate it.

strategy, which implicitly assumes that the measurement error is asymptotically of smaller order. BCHS derive the asymptotic distribution of the OLS estimator in the two-step strategy and show it has a first-order bias that can be analytically characterized and consistently estimated in several leading economic use cases.[1] They also show the usual OLS standard errors are correct. Taken together, their results enable researchers to perform valid inference on (1) without complete validation data.

The purpose of this paper is to provide practical guidance for implementing these bias corrections in empirical work. BCHS's framework can be adapted to many settings; we focus on two. The first is where $\hat{\theta}_i$ is a binary variable imputed via a classifier. The second is where θ_i is continuous and measured via an AI/ML-generated index, which to date is the most influential setting in economics. For implementation we use the Python package `ValidMLInference`. Documentation and Jupyter notebooks with additional illustration are in the repository https://github.com/timothymchristensen/ValidMLInference.[2]

I. Application 1: Imputed Binary Labels

We first consider the scenario where $\theta_i \in \{0,1\}$, which is the leading case in the PPI literature and is often encountered in economics. We assume the researcher has a classifier that generates a prediction $\hat{\theta}_i \in \{0,1\}$ of θ_i. Measurement error arises from misclassification error whenever $\theta_i \neq \hat{\theta}_i$ has nonzero probability. BCHS show the asymptotic bias of the OLS estimator in the two-step strategy is proportional to $\sqrt{n} \times \text{FPR}$, where $\text{FPR} \equiv E\left[\hat{\theta}_i(1-\theta_i)\right]$ is the false-positive rate of the classifier.[3] BCHS's bias correction only requires an estimate of FPR, which can be much less demanding

[1] This is different from a standard errors-in-variables problem, since the measurement error in $\hat{\theta}_i$ can be nonclassical. This means that even the sign of the bias may be unknown, except in special cases.

[2] A complementary R package MLBC is also available with source files at https://github.com/timothymchristensen/MLBC.

[3] This holds under the assumption that $\lim_{n\to\infty} \sqrt{n} \times \text{FPR} = \kappa \in [0,\infty)$, so that measurement error and sampling uncertainty in the regression are asymptotically comparable.

TABLE 1—EFFECT OF REMOTE WORK ON LOG SALARY

	Two-step	Bias-corrected
Remote work indicator	0.364	0.641
	[0.322, 0.406]	[0.446, 0.836]
Occupation fixed effects	Yes	Yes
Employment type fixed effects	Yes	Yes

Notes: 95 percent confidence intervals in brackets. The bias-corrected column uses `ols_bcm` with FPR $= 0.009$ estimated from an external sample of $m = 1{,}000$ hand-labeled job postings.

to construct than complete validation data. When such an estimate is available externally, it can be directly incorporated for bias correction.

We illustrate this with the remote work database of Hansen et al. (2023). They use a classifier based on DistilBERT (Sanh et al. 2020) to impute whether the text of a job posting explicitly mentions the possibility to work from home at least one day per week. The database has 500 million postings. We focus on a particular cell for illustration: all postings in 2022–2023 from San Diego, CA in the North American Industry Classification System industry "Accommodation and Food Services." There are $n = 16{,}315$ observations in this cell. To estimate the FPR, we randomly sample $m = 1{,}000$ postings[4] and compare the classified ones against ground-truth labels from Amazon Mechanical Turk. Because only 26 are predicted as remote-work postings, only those require inspection; nine are false positives, implying an estimated FPR of 0.009. By contrast, complete validation data would require inspecting all 1,000 postings.

We now describe the function `ols_bcm` for bias-correcting the two-step regression in this example.[5]

`ols_bcm(formula, data, generated_var, fpr, m, intercept=True)`

`formula`	Regression formula in R/statsmodels syntax
`data`	DataFrame containing the variables
`generated_var`	Name of the AI/ML-generated binary variable
`fpr`	Estimated false positive rate of the classifier
`m`	Sample size used for fpr, for standard error adjustment[6]
`intercept`	Include intercept term (default: `True`)

Figure 1 demonstrates how to use the function to bias-correct OLS estimators computed using the two-step strategy. The outcome Y_i is log posted salary, $\hat{\theta}_i$ is the prediction that a posting offers remote work imputed by the classifier (`wfh_rwo`), and X_i includes fixed effects for two-digit SOC occupation and whether the position is part-time or full-time. `ValidMLInference` includes a convenience function `ols` for implementing two-step OLS using the same syntax as `ols_bcm`.

The output of this code is in Table 1. The estimated coefficient is substantially larger after bias correction and lies outside the two-step 95 percent confidence interval. Note the usual two-step confidence interval and the bias-corrected confidence intervals do not overlap, even though the estimated FPR is less than 1 percent. Thus, even low misclassification probabilities can materially distort inference.

[4] Battaglia et al. (2025) prove that the bias correction yields valid inference under the condition that $\frac{n}{m^2} \to 0$, which is a substantially weaker condition than $\frac{n}{m} \to c > 0$, as assumed by the current literature.

[5] `ols_bcm` implements multiplicative bias correction. `ValidMLInference` supplies an alternative function `ols_bca` for implementing additive bias correction. BCHS recommend using a multiplicative correction for imputed binary labels.

[6] $m =$ `inf` can be used if one does not wish to apply this correction.

```
from ValidMLInference import ols, ols_bca_topic, fed_sentiment_data
import numpy as np

# load Fed sentiment data
data = fed_sentiment_data ()
meetings, B, kappa = data['meetings'], data['B'], data['kappa']

# outcome and control variables
Y = meetings['path_factor'].values
Q = meetings[['shadow_rate']].values

# selection matrix: hawkish and dovish (exclude neutral)
S = np.array([[1, 0, 0], [0, 1, 0]], dtype=float)

# raw classifier frequencies
counts = meetings[['hawkish', 'dovish', 'neutral']].values
P = counts / counts.sum(axis=1, keepdims=True)

# estimated true class probabilities
W = meetings[['w_hawkish', 'w_dovish', 'w_neutral']].values

# two-step (sentiment via raw frequencies)
mod_2step_1 = ols(Y=Y, X=np.column_stack ([P @ S.T, Q]))

# two-step (sentiment via estimated true class probabilities)
mod_2step_2 = ols (Y=Y, X=np.column_stack ([W @ S.T, Q]))

# bias-corrected
mod_bc = ols_bca_topic (Y=Y, Q=Q, W=W, S=S, B=B, k=kappa)
```

FIGURE 2. BIAS-CORRECTED REGRESSION USING THE VALIDMLINFERENCE PACKAGE (AI/ML-GENERATED INDICES)

Notes: The function `ols_bca_topic` implements the additive bias correction for a two-step regression onto estimated class probabilities. `B` is an estimated confusion matrix, and `kappa` is the scaling parameter $\kappa = \sqrt{n} \times \mathbb{E}\left[C_i^{-1}\right]$.

II. Application 2: AI/ML-Generated Indices

Many economic indices are built by aggregating multiple classification decisions. For example, Baker, Bloom, and Davis (2016) classify news articles as pertaining to policy uncertainty or not, then aggregate by time. After applying the classifier, the researcher observes a vector N_i containing the number of units classified in each of the J categories. This suggests a data-generating process

$$N_i \big| \left(C_i, w_i\right) \sim \text{Multinomial}\left(C_i, B^T w_i\right), \tag{2}$$

where C_i are the number of units classified for observation i, w_i is a Δ^{J-1}-valued latent random vector from which a "true" class label is drawn for each unit, and B allows for misclassification: its (i,j)th element β_{ij} denotes the probability that a unit with true label j is assigned label i. Finally, $\theta_i = Sw_i$, where S is a selection matrix. In index construction, θ_i represents an object like uncertainty, risk, or sentiment that is *never* directly observed, so building complete validation data is impossible.

We replicate the hawk/dove sentiment measure in Gorodnichenko, Pham, and Talavera (2023). We train BERT (Devlin et al. 2019) to classify human labels for hawk, neutral, and dovish sentiment ($J = 3$) and apply it to paragraphs of Federal Open Market Committeee (FOMC)

TABLE 2—EFFECT OF FED SENTIMENT ON POLICY PATH

	Two-step (raw freq.)	Two-step (estimate)	Bias-corrected
Hawkish	0.080	0.063	0.090
	[−0.006, 0.165]	[−0.012, 0.138]	[0.015, 0.165]
Dovish	−0.048	−0.046	−0.098
	[−0.159, 0.063]	[−0.145, 0.053]	[−0.197, 0.001]
Shadow rate	−0.005	−0.005	−0.009
	[−0.014, 0.004]	[−0.013, 0.004]	[−0.018, −0.001]

Notes: 95 percent confidence intervals in brackets. Outcome is the GSS path factor measuring expected future policy rates. Sample covers 199 FOMC meetings from 1995–2023. The first column measures sentiment with raw label frequencies. The second uses estimates of hawkish and dovish labels probabilities. The third applies additive bias correction to the second.

statements for 199 meetings from 1995 through 2023. A first strategy for measuring hawkish and dovish sentiment is to form the empirical frequencies of dovish $(N_{i,D}/C_i)$ and hawkish $(N_{i,H}/C_i)$ classifications, where C_i is the number of paragraphs in the meeting i statement, and $N_{i,D}$ and $N_{i,H}$ are the number classified as dovish and hawkish. This common approach does not account for misclassification. Alternatively, we form an estimate of B using the test-set confusion matrix from training BERT. This, combined with (2), allows us to estimate $w_{i,D}$ and $w_{i,H}$, the probabilities of true dovish and hawkish labels in meeting i.

`ols_bca_topic` bias-corrects the two-step regression when estimates $\hat{\theta}_i = S\hat{w}_i$ are covariates.[7] Importantly, it allows fixed misclassification error and only requires that $\sqrt{n} \times \mathbb{E}\left[C_i^{-1}\right]$ converges to a constant $\kappa \in [0, \infty)$.[8]

`ols_bca_topic(Y, X, W, S, B, k, intercept=True)`

`Y`	Outcome variable (array of length n)
`X`	Numeric control variables (matrix of dimension $n \times q$)
`W`	Estimated probabilities of true class labels $\hat{w}_i$ (matrix of dimension $n \times J$)
`S`	Selection matrix mapping w_i to θ_i (matrix of dimension $r \times J$)
`B`	Estimated confusion matrix $\hat{B}$ (matrix of dimension $J \times J$)
`k`	Scaling parameter $\kappa = \sqrt{n} \times \mathbb{E}\left[C_i^{-1}\right]$
`intercept`	Include intercept term (default: True)

Figure 2 demonstrates bias correction when Y_i is the path factor from Gürkaynak, Sack, and Swanson (2005), a measure of long-rate changes on FOMC meeting days. Covariates are estimates of $w_{i,D}$ and $w_{i,H}$ and X_i contains the short-run interest rate.

Table 2 shows results. Bias correction increases the magnitude of point estimates and shifts confidence intervals for several coefficients to exclude zero.

REFERENCES

Angelopoulos, Anastasios N., Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. "Prediction-Powered Inference." *Science* 382 (6671): 669–74.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." *Quarterly Journal of Economics* 131 (4): 1593–636.

[7] `ols_bcm_topic` instead implements a multiplicative bias correction.
[8] The exact formula is available in BCHS.

Battaglia, Laura, Timothy Christensen, Stephen Hansen, and Szymon Sacher. 2025. "Inference for Regression with Variables Generated by AI or Machine Learning." Preprint, arXiv. https://doi.org/10.48550/arXiv.2402.15585.

Caldara, Dario, and Matteo Iacoviello. 2022. "Measuring Geopolitical Risk." *American Economic Review* 112 (4): 1194–225.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of NAACL-HLT*, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 4171–86. Association for Computational Linguistics.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera. 2023. "The Voice of Monetary Policy." *American Economic Review* 113 (2): 548–84.

Gürkaynak, Refet S., Brian Sack, and Eric T. Swanson. 2005. "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements." *International Journal of Central Banking* 1 (1): 55–93.

Hansen, Stephen, Peter John Lambert, Nicholas Bloom, Steven J. Davis, Raffaella Sadun, and Bledi Taska. 2023. "Remote Work across Jobs, Companies, and Space." NBER Working Paper 31007.

Hansen, Stephen, and Michael McMahon. 2016. "Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication." *Journal of International Economics* 99 (S1): S114–33.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2019. "Firm-Level Political Risk: Measurement and Effects." *Quarterly Journal of Economics* 134 (4): 2135–202.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." Preprint, arXiv. https://doi.org/10.48550/arXiv.1910.01108.

Shapiro, Adam Hale, Moritz Sudhof, and Daniel J. Wilson. 2022. "Measuring News Sentiment." *Journal of Econometrics* 228 (2): 221–43.