Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Should Bank Stress Tests Be Fair?

Paul Glasserman, Mike Li

To cite this article:
Paul Glasserman, Mike Li (2025) Should Bank Stress Tests Be Fair?. Management Science 71(1):262-278. https://doi.org/10.1287/mnsc.2022.02060

informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 71, No. 1, January 2025, pp. 262–278
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Should Bank Stress Tests Be Fair?

Paul Glasserman,[a] **Mike Li**[a,*]

[a] Columbia Business School, New York, New York 10027

*Corresponding author

Contact: pg20@columbia.edu, https://orcid.org/0000-0002-9577-0205 (PG); mli24@gsb.columbia.edu,
https://orcid.org/0000-0002-4895-5047 (ML)

Received: July 10, 2022
Revised: May 2, 2023; September 21, 2023
Accepted: September 25, 2023
Published Online in Articles in Advance:
April 2, 2024

https://doi.org/10.1287/mnsc.2022.02060

Copyright: © 2024 INFORMS

Abstract. Regulatory stress tests have become one of the main tools for setting capital requirements at the largest U.S. banks. The Federal Reserve uses confidential models to evaluate bank-specific outcomes for bank-specific portfolios in shared stress scenarios. As a matter of policy, the same models are used for all banks, despite considerable heterogeneity across institutions; individual banks have contended that some models are not suited to their businesses. Motivated by this debate, we ask, what is a fair aggregation of individually tailored models into a common model? We argue that simply pooling data across banks treats banks equally but is subject to two deficiencies: it may distort the impact of legitimate portfolio features, and it is vulnerable to implicit misdirection of legitimate information to infer bank identity. We compare various notions of regression fairness to address these deficiencies, considering both forecast accuracy and equal treatment. In the setting of linear models, we argue for estimating and then discarding centered bank fixed effects as preferable to simply ignoring differences across banks. We also discuss extensions to nonlinear models.

History: Accepted by Kay Giesecke, finance.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.02060.

Keywords: finance • financial institutions: banks • economics: econometrics • statistics

1. Introduction

In the aftermath of the 2008 financial crisis, U.S. banking regulators adopted stress testing as a primary tool for monitoring the capital adequacy of the largest banks. For each round of annual stress tests, the Federal Reserve announces a “severely adverse stress scenario,” defined by a hypothetical path of economic variables over the next several quarters. A typical path includes an increase in unemployment, a decline in gross domestic product (GDP), and projections for the level and volatility of the stock market, among other variables. The largest banks provide the Fed with detailed information about their loan portfolios and other assets. The Fed then applies internally developed models to project revenues and losses for each bank through the stress scenario. Banks are required to have sufficient capital to weather the projected losses.

The Fed does not disclose details of the models it uses to project revenues and losses. The Fed makes clear that to ensure consistent treatment for different banks, it uses “industry models” as opposed to models tailored to individual banks. As a matter of policy, the same models are applied to all banks. Quoting Board of Governors (2021, p. 3), “two firms with the same portfolio receive the same results for that portfolio.” We will refer to this statement as the Fed’s principle of equal treatment.

Banks have countered that the Fed’s models fail to capture bank-specific features that could lower projected losses. They have made these arguments in requests for reconsideration of stress test results. Of course, banks are not objective critics of the Fed’s supervision; but significant heterogeneity among the largest banks is indisputable. The banks subject to annual stress testing include universal banks, investment banks, large regional banks, the U.S. subsidiaries of certain foreign banks, and a variety of more specialized financial firms. It is certainly possible that bank-specific models would produce more accurate forecasts than a single industry model, in which case using a single model entails a trade-off between forecast accuracy and consistency across banks. Indeed, a recent industry article (Baer and Hopper 2023) argues that “banks’ own models are trained on each bank’s specific experience, rather than relying on a one-size-fits-all assumption,” and that “[t]here is a strong argument that banks’ own models generate far more accurate results than the Fed’s.”

The heterogeneity among large banks motivates the questions we study. What is the best way to aggregate bank-specific models into an industry model? How should the Fed's principle of equal treatment be interpreted and implemented? Is simply ignoring bank identity in estimating and applying models the best way to achieve fairness? To what extent is fairness at odds with accuracy? Although the heterogeneity of large banks is widely recognized, we know of no prior work that seeks to address this property within the constraints of the Fed's policy of equal treatment. We will argue that addressing heterogeneity is preferable to ignoring it.

The question of fairness in algorithms and models has received a great deal of renewed interest in recent years, in some cases reviving earlier debates over fairness in testing and related policies that were not explicitly "algorithmic"; see, for example, the overviews in Barocas et al. (2019) and Hutchinson and Mitchell (2019). We draw on this literature, but our setting differs in important ways from most discussions of fairness.

Algorithmic fairness is usually concerned with ensuring that certain protected attributes—race or gender, for example—do not influence outcomes such as hiring decisions or loan approvals. Different methods can be compared based on alternative measures of influence and the degree to which sensitive attributes are indeed protected.

The counterpart of a protected attribute in our setting is a bank's identity; but this attribute is not so much protected (in the sense that race and gender are) as inadmissible for the Fed's purpose. In stating that "two firms with the same portfolio receive the same results for that portfolio," the Fed is stating that bank identity is not a legitimate predictor of losses. Perhaps, then, fairness is achieved as long as the Fed uses the same model for all banks. In other words, perhaps "fairness through unawareness," paraphrasing Dwork et al. (2012), is sufficient in this setting. Moreover, in questioning whether the Fed's models apply to them, banks are not claiming discrimination; on the contrary, they are asking for discrimination—asking that the Fed change its models to recognize ways in which an individual bank differs from other banks.

To investigate these issues, we focus primarily on a simple setting in which the "true" loss rate for each bank is described by a bank-specific regression on portfolio features and scenario features. The regulator's goal is to aggregate these bank-specific models into a single model. A natural interpretation of an "industry" model in this setting is a pooled regression based on combining results across banks. The pooled model treats banks equally, but we show that it has at least two significant deficiencies: When applied to heterogeneous banks, it can produce poor measures of the marginal impact of individual features, even resulting in the wrong sign; and it implicitly misdirects legitimate information in portfolio features to infer (or proxy for) bank identity in forecasting losses. The second of these deficiencies works against the spirit of equal treatment of banks, even if bank identity is not explicitly used in the model.

We then investigate the application of ideas from algorithmic fairness in our setting. The fairness literature has mainly focused on classification problems (hiring decisions and credit approvals, for example), with regression problems getting somewhat less attention. Chzhen et al. (2020) and Le Gouic et al. (2020) developed a method of particular importance for regression that Le Gouic et al. (2020) call "projection to fairness." This method produces optimal forecasts (in the least-squares sense) subject to a fairness constraint known as *demographic parity*. We examine the application of this approach in our setting and conclude that it goes too far in leveling results across banks.

The pooled method ignores fairness and the projection method goes too far in imposing fairness, so we seek an intermediate solution. Johnson et al. (2020) introduce a variety of methods for introducing fairness considerations in regression. These include methods they call "full equality of opportunity" (FEO) and "substantive equality of opportunity" (SEO). We examine these methods in our setting and conclude that the FEO method provides an attractive solution. In particular, we show that it addresses the two deficiencies of the pooled method highlighted previously: It removes the distortion in the pooled coefficients that results from bank heterogeneity, and it prevents the misdirection of legitimate information to infer bank identity. Indeed, we show that the only way to achieve lower forecast errors than the FEO method is through such misdirection, a result that sheds light on the tradeoff between accuracy and fairness.

Moreover, the method is easy to interpret and implement: fit a pooled model with centered bank fixed effects and then *discard* the centered fixed effects to forecast losses. Including the fixed effects prevents misdirection of legitimate information; discarding them is necessary to treat banks equally and centering ensures that the overall mean forecast remains unchanged. Although we mainly work with linear models, we show that these ideas can be extended to nonlinear models as well. We also derive an extension of FEO to remove certain interaction effects, as opposed to just fixed effects.

To help position our work, we briefly discuss some other research on bank stress tests. Covas et al. (2014), Kapinos and Mitnik (2016), and Kupiec (2020) find strong evidence of heterogeneity in banks' responses to macroeconomic shocks, and Kapinos and Mitnik (2016) argue that ignoring heterogeneity can substantially underestimate projected capital requirements. The related models of Hirtle et al. (2016) and Guerrieri and Welch (2012) forecast aggregate results and are therefore not concerned with differences among banks. Heterogeneity in the accuracy of the Fed's models for different banks is suggested by the comparisons in Agarwal et al. (2020), Bassett and Berrospide (2018), and Flannery et al. (2017) between the Fed's results and results based on the banks' own models.

A separate line of research considers the design of stress scenarios. Several studies (Breuer et al. 2009, Flood and Korenko 2015, Glasserman et al. 2015, Pritsker 2017, Schuermann 2020) have advocated the use of multiple scenarios to capture different combinations of risk factors. Cope et al. (2022) and Flood et al. (2022) recommend designing scenarios to reflect bank heterogeneity. Parlatore and Philippon (2022) propose a theoretical framework for scenario design as a problem of optimal information acquisition.

Several studies have investigated the information content of stress test results, either through market responses (Morgan et al. 2014, Glasserman and Tangirala 2016, Flannery et al. 2017, Georgescu et al. 2017, Fernandes et al. 2020, Sahin et al. 2020, Guerrieri and Modugno 2021) or through subsequent bank performance (Philippon et al. 2017, Kupiec 2020). Flannery (2019) discusses just how much information the Fed should disclose about stress testing procedures and outcomes. For perspectives on the effectiveness of the Fed's stress tests, see Kohn and Liang (2019) and Schuermann (2020).

We provide additional background on the Federal Reserve's stress tests in Section 2. Section 3 lays out our modeling framework and analyzes the pooled industry model within this framework. Section 4 analyzes various ways to introduce fairness considerations, including the projection-to-fairness and FEO methods. Section 5 considers nonlinear models. Proofs of our main results appear in Appendix A. Additional supporting theoretical (Sections EC.1–EC.3) and empirical (Sections EC.4–EC.7) material is included in the online appendix. Most of our discussion considers loss models, but we consider revenue models in Section EC.7.

2. Background

This section provides background on the Federal Reserve's stress testing process and on the heterogeneity of the participating banks.

2.1. Regulatory Bank Stress Tests

In early 2009, in the depths of the Global Financial Crisis, the Federal Reserve launched a stress test of the 19 largest U.S. bank holding companies to gauge how much more capital they would need if economic conditions continued to worsen. The results of the stress test were made public, and the transparency and credibility of the process have been credited with restoring public confidence and helping to end the crisis.

The Dodd-Frank Act, the package of reforms that followed the crisis, codified the use of stress testing for bank supervision. The number of banks subject to DFAST (Dodd-Frank Act Stress Tests) has varied over time. The current requirement applies annually to banks with over $250 billion in assets and every other year to banks with assets between $100 billion and $250 billion. The 2022 DFAST covered 34 banks. We refer to the participating firms as "banks," but they are more precisely holding companies, including the U.S. subsidiaries of some foreign banks.

The inputs to the stress test analysis are the stress scenario, which is common to all banks, and bank-specific balance sheet information. A scenario is specified through a hypothetical path of economic variables over the next 13 quarters. The 2022 DFAST specified paths for 28 variables, including GDP, inflation, unemployment, stock market and real estate indexes, interest rates, exchange rates, and measures of overseas economic activity. Each bank submits detailed information on its loans and other assets.

The Fed uses 21 models to integrate the stress scenarios with bank-level information to make bank-level projections. For example, one model applies to credit cards, one to first lien residential mortgages, one to commercial real estate loans, and another to commercial and industrial loans. These models project losses in each of these portfolios. Some other models project revenues.

The Fed does not disclose details of its models, either to banks or the general public. But it does describe its general modeling approach in public documents. At a high level, a model assigns a loss rate to a set of bank-specific loan portfolio features x and a common set of scenario variables z through a function $f(x, z)$. The function f is estimated from past observations of the macro variables and portfolio features for multiple banks. Thus, f is estimated as an industry-wide model and then applied individually to each bank.

This approach is described, for example, on p. 3 of Board of Governors (2021), where we read, "The Federal Reserve generally develops its models under an industry-level approach calibrated using data from many financial institutions. … The Federal Reserve models the response of specific portfolios and instruments to variations in macroeconomic and financial scenario variables such that differences across firms are driven by differences in firm-specific input data, as opposed to differences in model parameters and specifications. As a result, two firms with the same portfolio receive the same results for that portfolio in the supervisory stress test, facilitating the comparability of results."

As noted in the Introduction, we refer to the principle that banks with the same portfolio receive the same results as *equal treatment*.

2.2. Bank Heterogeneity

The appropriateness of equal treatment seems incontrovertible. However, the right notion of consistency across firms becomes less clear when portfolios vary widely, and the largest U.S. banks are a highly heterogeneous group. We may not expect a regional bank to have an investment bank's skill in the capital markets, nor do we expect the investment bank to have the regional bank's skill in making single-family residential loans.

Heterogeneity among large banks is illustrated in Figure 1. The left panel applies to the banks that participated in the Federal Reserve's 2022 stress test. It shows the distribution of the banks by their Global Industry Classification Standard subindustry classifications. This group includes diversified banks (such as JPMorgan Chase and Bank of America); regional banks (like PNC Financial and Citizens Financial); consumer finance companies (including American Express and Discover); custody banks (such as Bank of New York Mellon and State Street); investment banks (including Goldman Sachs and Morgan Stanley); and intermediate holding companies comprising the U.S. subsidiaries of foreign banks (such as TD Group and Credit Suisse USA). The distribution of banks across categories reflects important differences in their areas of specialization.

The right panel of Figure 1 applies to the U.S. Global Systemically Important Banks (G-SIBs). It shows heterogeneity in the fractions of loans the banks hold in each of the four categories. For example, for Wells Fargo (WFC) first lien mortgages are a relatively large fraction of its loans, whereas for Citigroup (C), credit cards make up a relatively large fraction. The figure suggests different areas of specialization in lending, even among the largest U.S. banks. Heterogeneity across banks is further explored empirically in the online appendix.

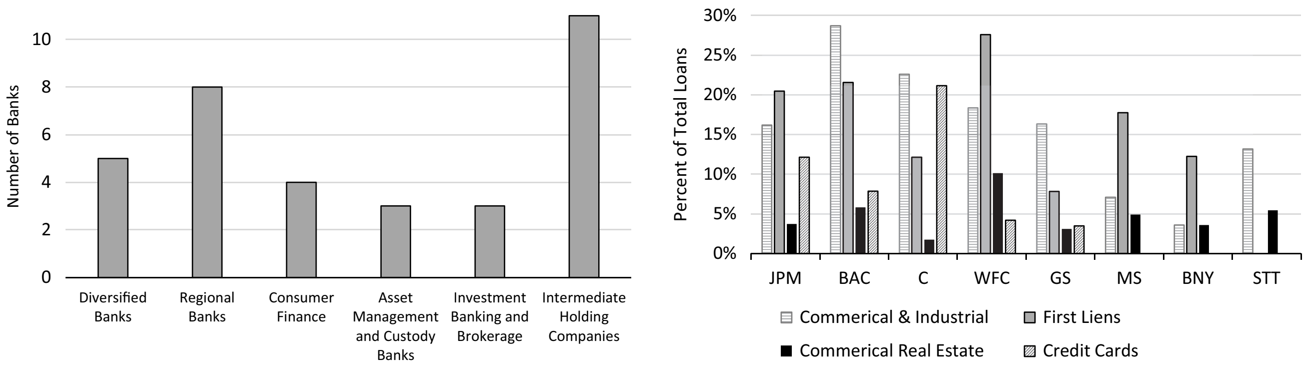
Beginning in 2020, the Federal Reserve allowed banks to submit requests for reconsideration of the stress capital buffer set by the Fed through the stress testing process. (The capital buffer is set through the Comprehensive Capital Analysis and Review (CCAR) process, which accompanies the stress test.) The banks' requests are confidential, but the Fed's responses to these requests are public. The responses show that the banks were arguing for reconsideration at least in part based on claims that the Fed's models do not capture distinctive features of the banks' businesses. For example, Regions Financial claimed that the Fed's models overlook the bank's hedging of interest rate risk. Goldman Sachs claimed that the Fed's models omit information relevant to the credit quality of the bank's mortgage loans. Citizens Financial claimed that the Fed's models overlook the bank's loss-sharing agreements in its retail portfolio.

Five firms requested reconsideration in 2020, and all five requests were rejected. In its response[1] to Goldman Sachs, the Fed wrote, "the Board has determined that it will follow its published principles for stress testing, including the principle of creating industry-level models, and not modify the existing results of these models. In particular, models used in the supervisory stress test are generally developed according to an industry-level approach, calibrated using data from many institutions." Similar statements appear in all five rejections. These exchanges point to a debate in which the banks highlight their heterogeneity and the Fed asserts the importance of consistency.

2.3. Heterogeneity and Fairness

Read narrowly, the principle of equal treatment—the Fed's statement that "two firms with the same portfolio receive the same results for that portfolio"—is easy to satisfy. It holds in any model that forecasts losses based only on portfolio features and the stress scenario, without using any other bank-specific information. Even this narrow reading has important implications. For example, the quality of a bank's information technology (IT) systems or the strength of its "culture"[2] may be important factors in determining a bank's losses under stress, but as they are not features of a loan portfolio, the Fed's modeling principle would preclude incorporating them into the Fed's models. Matters like the quality of a bank's internal governance and controls must be addressed in other parts of the overall bank supervision process, outside of stress testing.[3] Within the Basel

Figure 1. Heterogeneity Among Large Banks



Notes. (Left) Distribution of 2022 stress test banks by GIC subindustry. (Right) The percentage of loans in each of four categories for each of the U.S. G-SIBs, based on Y-9C reports for Q4 2021.

framework, these considerations are part of the Pillar 2 supervisory process, as described in BCBS (2019).

Under a broader interpretation of the principle of equal treatment, a regulatory model should also exclude indirect proxies for bank identity. Suppose, for example, that a bank with outdated IT systems had a particularly large number of loans to the energy sector. Suppose further that because of its weak IT the bank was a poor monitor of its borrowers and suffered abnormally large losses in downturns. With information about IT excluded, a predictive model of losses that uses this history would likely overstate the risk of loans to the energy sector. This outcome is arguably unfair to all banks making energy loans, in that they would be indirectly penalized for one bank's weak IT. Addressing these types of indirect effects drives our investigation. In its narrow sense, equal treatment requires an indifference to which banks hold which portfolios once a model is selected; the broader interpretation seeks to remove the influence of bank identity in the design of the model.

The Fed's stated principle implicitly responds to concerns for *disparate treatment* of banks. The broader interpretation—precluding proxies for bank identity—aligns with a concern for a particular notion of *disparate impact* used in the literature on algorithmic fairness (see, for example, chapter 6 of Barocas et al. (2019); section 3 of Lipton et al. (2018); and Prince and Schwarcz (2019)), sometimes called "proxy discrimination." The banks' objections, as reflected in their reconsideration requests, can be seen as concerns for a different type of disparate impact: even if the same model is applied to all banks, and even if the model is free of bank-identity proxies, some banks may claim to be more adversely affected than others by the model's limitations. Most of the objections raised by banks can be understood as pointing to omitted variables—features omitted from the Fed's models that a bank believes would result in a more favorable outcome if included in the models. The Fed's responses suggest a reluctance to incorporate overly narrow features into models, particularly features that might affect only a single bank. Model limitations, of the type claimed by the banks are likely inevitable, given the limited data available on bank performance in scenarios of severe stress. The Fed should strive to continue to improve its models, but our concern is not primarily for the banks' objections. Our focus is rather on how best to interpret and implement the Fed's stated principle of equal treatment, particularly under the broader interpretation that addresses elements of both disparate treatment and disparate impact, within the overarching goal of accurately forecasting stressed losses for each bank.

The fairness literature distinguishes notions of individual fairness and group fairness, where the members of a group often share a sensitive or protected attribute. More abstractly, an individual is defined by a fixed set of features, and a group is characterized by a probability distribution over features; see, for example, the characterizations of individuals and groups in sections 2 and 3 of Dwork et al. (2012). From this perspective, individual fairness is concerned with fairness conditional on a set of features, whereas measures of group fairness incorporate distributions over features. Individual fairness typically requires that individuals with similar features be treated similarly. For a portfolio loss model, this condition is satisfied if the predicted loss is a suitably smooth function of the features of individual portfolios. However, group fairness is more relevant to our setting than individual fairness because we think of each bank, with its particular mix of businesses and areas of focus, not as one portfolio but as a distribution over portfolios the bank might hold at different times. We are interested in accuracy and fairness with respect to these distributions of bank portfolio features. We therefore view each bank as a group of (and probability distribution over) individual portfolios that share the attribute of bank identity. In contrast, individual fairness would be relevant to evaluating accuracy and fairness conditional on a specific portfolio for each bank. We will make this formulation of groups and individuals more explicit in the next section after introducing our basic model.

3. Pooling: Fairness Through Unawareness?

3.1. Basic Model

To capture bank heterogeneity, we consider a market with multiple banks, indexed by $s = 1, \ldots, \overline{S}$. The loss rate (or net charge-off rate) Y_s for bank s is given by

$$Y_s = \alpha_s + \beta_s^\top X_s + \epsilon_s, \tag{1}$$

with $\alpha_s \in \mathbb{R}$ and $\beta_s \in \mathbb{R}^d$. Here, X_s is a d-dimensional random vector of predictive variables whose distribution defines bank s; at this point, we do not distinguish between portfolio characteristics and macro variables. The portfolio characteristics include information about a bank's borrowers and loan terms. We use a linear specification in (1) because it offers the simplest setting to explore the interaction of heterogeneity and fairness; we discuss nonlinear extensions in Section 5. We take (1) to be the true relationship between the loss rate Y_s for bank s over the forecast horizon and characteristics X_s known at the date the forecast is made. Loss rates are normalized by loan balances to make values of Y_s comparable across banks of different sizes.

We think of X_s as a draw from some distribution with

$$\mu_s = \mathsf{E}[X_s] \in \mathbb{R}^d, \quad \Sigma_s = \mathsf{var}[X_s] \in \mathbb{R}^{d \times d}. \tag{2}$$

The randomness in X_s can be interpreted as reflecting the variation in the characteristics for bank s (and the macro variables) over time—in particular, times of stress. As we discuss in greater detail in Remark 3.1, we think of each bank as a group, in the sense of a probability distribution over portfolios. We assume throughout that each Σ_s (hence μ_s) is finite, and each Σ_s is nonsingular. The error ϵ_s in (1) is assumed to satisfy, for each s,

$$\mathsf{E}[\epsilon_s] = 0 \quad \text{and} \quad \mathsf{cov}[X_s, \epsilon_s] = 0. \tag{3}$$

The regulator's problem is to choose a model g that forecasts the loss rate $g(x, s)$ for bank s if the bank's portfolio characteristic vector is x. The forecasts should, at a minimum, satisfy the following narrow property, which prohibits the regulator from applying different models to different banks.

Definition 3.1 (Equal Treatment). Model $g : \mathbb{R}^d \times \{1, \ldots, \overline{S}\} \to \mathbb{R}$ satisfies equal treatment if $g(x, s) = g(x, s')$, for all $x \in \mathbb{R}^d$, for all $s, s' \in \{1, \ldots, \overline{S}\}$.

As the true relationship for each bank is linear in (1), we mainly focus on the case of a linear industry-wide model. The regulator's problem is then to choose a single $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$ that it will use to form a forecast

$$\hat{Y}(x) = \alpha + \beta^\top x, \tag{4}$$

given portfolio characteristics x. The forecast (4) satisfies equal treatment because it has no functional dependence on bank identity s. The parameters of the industry model (4) may depend on the bank-specific parameters (α_s, β_s) and on the mean and variance in (2), but they should not depend on the realized features X_s.

The regulator would like the forecast loss $\hat{Y}(X_s)$ to be close to the actual loss Y_s in (1) for every bank s. To aggregate errors across banks, we introduce a random variable S that picks a bank according to a distribution

$$\mathsf{P}(S = s) = p_s, \quad s = 1, \ldots, \overline{S}, \tag{5}$$

with the probabilities p_s summing to one. In the simplest case, all banks get equal weight, and the p_s are all equal; but the p_s could also reflect relative asset sizes or other weighting schemes. When we replace a bank label s with the random variable S, we get a mixture over banks. In particular, we can combine the bank-specific models (1) into a mixture or hierarchical model by writing

$$Y_S = \alpha_S + \beta_S^\top X_S + \epsilon_S. \tag{6}$$

In choosing parameters α and β in (4), the regulator would like to make the forecast errors small for all banks. A natural way to aggregate forecast errors across banks is to consider the average squared error, in which case the regulator's problem becomes choosing α and β in (4) to solve

$$\min_{\alpha, \beta} \mathsf{E}[(\hat{Y}(X_S) - Y_S)^2]. \tag{7}$$

The objective in (7) averages squared forecast errors over banks. It can also be written as $\sum_s p_s \mathsf{E}[(\hat{Y}(X_s) - Y_s)^2]$.

Remark 3.1. Before solving (7), we make several comments on our problem formulation.

(i) *Targets versus estimators.* The problem posed by (7), like the more general problem of choosing industry parameters in (4), is one of characterizing ideal coefficients α and β. This is a question of choosing the correct *targets* of estimation, rather than a question of choosing estimators. In particular, α and β are population quantities rather than sample quantities. In practice, the regulator may have a panel of time series of observations across banks. Estimation methods for panel data ordinarily focus on coefficients that are common to all units and exploit the panel structure to estimate these shared values. Our concern is precisely with the case of heterogeneous coefficients, where we need to identify suitable targets before we can consider their estimation.

(ii) *Groups versus individuals.* As discussed in Section 2.3, individual fairness is concerned with ensuring that if two feature vectors x and x' are close, then the predicted losses $\hat{Y}(x)$ and $\hat{Y}(x')$ are also close. Our concern is for accuracy and fairness with respect to the distributions of (X_s, Y_s), $s = 1, \ldots, \overline{S}$, and not just for individual outcomes; we do not want the choice of industry model to depend on the realization of X_s, $s = 1, \ldots, \overline{S}$. Each (X_s, Y_s) reflects a distribution over individual portfolio features and losses—individuals that share the bank identity attribute s. Each bank thus represents a group of potential individual portfolios, and we are interested in accuracy and fairness with respect to the probability distributions that define these groups.

(iii) *Stressed versus unstressed.* For the application to stress testing, it is helpful to think of the portfolio features and scenario variables in X_s and the losses Y_s in (1) as having their conditional distributions given stress conditions. The regulator is then interested in forecasting the conditional mean loss for each bank, given stress conditions. By focusing on the conditional mean, this formulation makes the squared error (7) a reasonable benchmark for studying accuracy and fairness.

(iv) *Regulator versus banks.* As discussed in Section 2.3, some of the objections raised by banks can be understood as pointing to features missing from (1) and (4), features that are rejected by the Fed as overly narrow. Our investigation assumes the bank-specific models (1)–(3) are correct; in particular, under (3) any relevant omitted features are uncorrelated with included features. We focus on the regulator's problem of how best to aggregate the bank-specific models (assuming their correctness) into an industry model, considering both accuracy and fairness; we do not address the banks' claims regarding which features should be included in the models.

For the solution to (7), write

$$\overline{\mu} = \mathsf{E}[X_S] = \mathsf{E}[\mu_S] = \sum_s p_s \mu_s \in \mathbb{R}^d, \tag{8}$$

and

$$\mathsf{var}[X_S] = \mathsf{E}[(X_S - \overline{\mu})(X_S - \overline{\mu})^\top] = \mathsf{E}[W_S] = \sum_s p_s W_s, \quad (9)$$

with

$$W_s = \Sigma_s + \mu_s \mu_s^\top - \overline{\mu}\mu_s^\top \in \mathbb{R}^{d\times d}. \quad (10)$$

Similarly,

$$\mathsf{cov}[\alpha_S, \mu_S] = \sum_s p_s \alpha_s (\mu_s - \overline{\mu}) \in \mathbb{R}^d.$$

Proposition 3.1. *Problem (7) is solved by*

$$\beta_{Pool} = \mathsf{E}[W_S]^{-1}(\mathsf{cov}[\alpha_S, \mu_S] + \mathsf{E}[W_S \beta_S]) \quad (11)$$

and

$$\alpha_{Pool} = \mathsf{E}[Y_S] - \beta_{Pool}^\top \overline{\mu}. \quad (12)$$

Loss forecasts using α_{Pool} and β_{Pool} in (4) provide *fairness through unawareness*, in that they ignore bank identity. They satisfy equal treatment in the narrow sense of Definition 3.1. Given our starting point (1), Problem (7) would seem to be the most direct interpretation of the Fed's policy of developing an "industry-level approach calibrated using data from many financial institutions."

However, the solution in (11) is not a satisfactory target. Indeed, (11) shows where heterogeneity is most problematic. If the intercepts α_s covary with the means μ_s, this effect can distort β_{Pool} through what is commonly known as Simpson's paradox. As an extreme example, consider the case that $\beta_s = 0$ for all s; in other words, none of the features in X_s is predictive of losses for any of the banks. The regulator's model (4) using β_{Pool} would nevertheless forecast losses based on these features if $\mathsf{cov}[\alpha_S, \mu_S]$ is nonzero. This covariation would create the illusion of predictability. In applying (11), we would be forecasting losses based on irrelevant features, purely as a consequence of the way we aggregated the bank-specific models.

Even in a less extreme setting in which the β_s are nonzero, the presence of the $\mathsf{cov}[\alpha_S, \mu_S]$ term in (11) reflects an indirect influence of bank identity on loss forecasts. If the bank-level mean characteristics μ_s positively covary with the bank-level intercepts α_s, then in the pooled model this covariance will lead to a higher loss forecast for a bank with a higher value of X_s. This is arguably unfair, in the sense that the loss forecast is not based on the legitimate influence of the feature X_s. We will formalize the idea that the pooled method misdirects legitimate information in Sections 4.2 and 4.5.

This effect is reminiscent of the bias incurred in panel regressions when fixed effects are present in the data but omitted from a model. As we emphasized in Remark 3.1(i), in our setting the primary objective is to define the appropriate target of estimation, given the heterogeneity in the coefficients. We cannot say the term $\mathsf{cov}[\alpha_S, \mu_S]$ introduces bias until we have decided what we are trying to estimate.

3.2. Average Treatment Effects

We can gain additional insight by considering the case of scalar X_s. In this case, the pooled coefficient β_{Pool} in (11) becomes

$$\beta_{Pool} = \frac{\mathsf{cov}[\alpha_S, \mu_S] + \sum_s p_s(\sigma_s^2 + \mu_s^2 - \overline{\mu}\mu_s)\beta_s}{\sum_s p_s(\sigma_s^2 + \mu_s^2 - \overline{\mu}\mu_s)}. \quad (13)$$

In the special case that $\mathsf{cov}[\alpha_S, \mu_S] = 0$ and $\sigma_s^2 + \mu_s^2 - \overline{\mu}\mu_s \geq 0$, for all s, (13) becomes a convex combination of the individual β_s. In Section EC.3 in the online appendix, we state some simple properties that an aggregation of the individual β_s into a single industry value should satisfy, and we show that only a convex combination satisfies these properties. Equation (13) thus shows a further potential problem with the pooled method. Even if $\mathsf{cov}[\alpha_S, \mu_S] = 0$, the coefficient on some β_s could be negative, which would mean that a reduction in β_s would increase β_{Pool}. This could mean that an improvement in risk management by one bank *increases* predicted losses at all banks. We investigate these types of cross-bank effects further in Section EC.1 in the online appendix.

We will refer to any convex combination of the β_s as a *weighted average treatment effect* or WATE parameter. This terminology is suggested by thinking of a unit increase in a portfolio characteristic X_s as a treatment, and β_s as the response to that treatment. The (ordinary) average treatment effect is the expected coefficient,

$$\beta_{ATE} = \mathsf{E}[\beta_S] = \sum_s p_s \beta_s, \quad (14)$$

but weighting the individual coefficients allows other combinations. In particular, if the μ_s are all equal, the pooled coefficient (13) becomes

$$\beta_{Pool} = \frac{\sum_s p_s \sigma_s^2 \beta_s}{\sum_s p_s \sigma_s^2}. \quad (15)$$

We will say more about these cases in subsequent sections.

To translate a WATE coefficient into a loss projection $\hat{Y}$, as in (4), we also need to specify an intercept. Setting

$$\alpha_{WATE} = \mathsf{E}[Y_S] - \beta_{WATE}^\top \overline{\mu},$$

ensures that the forecasts

$$\hat{Y}_{WATE}(X_s) = \alpha_{WATE} + \beta_{WATE}^\top X_s, \quad s = 1, \ldots, \overline{S},$$

have zero expected error, in the sense that

$$\mathsf{E}[\hat{Y}_{WATE}(X_S) - Y_S] = \sum_s p_s(\alpha_{WATE} + \beta_{WATE}^\top \mu_s) - \mathsf{E}[Y_S] = 0.$$

4. Fair Regressions

We have seen that if the regulator's sole objective is to minimize average squared forecast errors subject to equal treatment, then the solution is given by the pooled coefficients in (11) and (12). However, we have also seen that (11) has consequences that are undesirable and even unfair, in the sense that it is indirectly influenced by bank identity. In this section, we turn to methods that expand the squared loss minimization objective (7) to include fairness considerations. Because the pooled method minimizes (7), any method that addresses fairness will entail a loss of accuracy as measured by (7).

4.1. Projection to Fairness

In the literature on fairness in classification methods, *demographic parity* is among the most widely discussed fairness principles; see, for example, chapter 3 of Barocas et al. (2019). In the simplest classification setting, the counterpart of our forecast is a binary outcome $\hat{Y} \in \{0,1\}$. For example, $\hat{Y} = 1$ may indicate a hiring decision, a loan approval, or a school admission decision. The decision is to be based on certain features of a candidate that are deemed legitimate. Demographic parity requires that the event $\{\hat{Y} = 1\}$ be statistically independent of a protected attribute, such as race or gender. This objective is difficult to achieve when legitimate features covary with the protected attribute.

Chzhen et al. (2020) and Le Gouic et al. (2020) extend the notion of demographic parity to the regression setting by requiring that model predictions be independent of a protected attribute. These two articles solve the problem of finding the model that minimizes mean squared prediction errors while achieving demographic parity. We will use the term *projection to fairness* (PTF), coined in Le Gouic et al. (2020), for the method in these papers.

Both papers reduce the problem of regression fairness to one of finding the Wasserstein barycenter of a set of distributions, in the sense of Agueh and Carlier (2011). The barycenter is the distribution closest to the set of distributions in an average sense. For a squared error and one-dimensional distributions, the barycenter can be described as the distribution whose quantile function is a weighted average of the individual quantile functions. (The quantile function is the inverse of the cumulative distribution function.)

In the setting of Section 3.1, the resulting solution can be interpreted as follows. Let F_s denote the cumulative distribution function of $\hat{Y}_s(X_s)$, the forecast for bank s. Given realized features $X_s = x$, the regulator first forms the forecast $\hat{Y}_s(x) = \alpha_s + \beta_s^\top x$, using the bank-specific coefficients. The regulator then computes $q_s = F_s(\hat{Y}_s(x))$, meaning that the forecast $\hat{Y}_s(x)$ is at the q_s quantile of the distribution F_s. The PTF forecast is then achieved by taking the weighted average of the corresponding quantile of all banks' forecast distributions, $\mathsf{E}_S[F_S^{-1}(q_s)]$. If, for example, $\hat{Y}_s(x)$ falls at the 80th percentile of the forecast distribution for bank s, then the regulator takes a weighted average of the 80th percentile forecast for all the bank-specific models. That weighted average becomes the PTF forecast for bank s.

To make this procedure more explicit and to specialize the general framework of Chzhen et al. (2020) and Le Gouic et al. (2020) to our setting, we consider the case (for this section only) that each feature vector X_s has a multivariate normal distribution $N(\mu_s, \Sigma_s)$. Write $\Sigma_s^{1/2}$ for the symmetric square root of Σ_s, and define the standardized feature vectors

$$Z_s = \Sigma_s^{-1/2}(X_s - \mu_s); \tag{16}$$

each Z_s has a multivariate standard normal distribution. Write the basic identity (1) using standardized variables as

$$Y_s = \alpha_s^o + \beta_s^{o\top} Z_s + \epsilon_s,$$

with standardized coefficients

$$\beta_s^o = \Sigma_s^{1/2}\beta_s, \quad \alpha_s^o = \alpha_s + \beta_s^\top \mu_s. \tag{17}$$

Suppose $\|\beta_s^o\| \neq 0$, for all s, with $\|\cdot\|$ denoting the usual Euclidean norm. Consider the model that assigns, to each bank $s = 1, \ldots, \overline{S}$, with features $X_s = x$ the forecast

$$\hat{Y}^o(x,s) = \sum_i p_i \alpha_i^o + \sum_i p_i \|\beta_i^o\| \frac{\beta_s^{o\top} z_s}{\|\beta_s^o\|}, \quad z_s = \Sigma_s^{-1/2}(x - \mu_s). \tag{18}$$

If there exists a $\beta \in \mathbb{R}^d$ and scalars $a_s > 0$ for which

$$\beta_s^o = a_s \beta, \quad s = 1, \ldots, \overline{S}, \tag{19}$$

then we will see that (18) simplifies to the weighted average

$$\hat{Y}^o(x,s) = \overline{\alpha}^o + \overline{\beta}^{o\top} z_s, \quad \overline{\alpha}^o = \sum_i p_i \alpha_i^o, \overline{\beta}^o = \sum_i p_i \beta_i^o. \tag{20}$$

In the case of scalar X_s, (19) holds whenever all β_s have the same sign.

Proposition 4.1. *Suppose that the X_s are multivariate normal and $\|\beta_s\| \neq 0$, for all $s = 1, \ldots, \overline{S}$. Then (18) is the projection-to-fairness of the bank-specific models (1), meaning that (18) minimizes $\mathsf{E}[(\hat{Y}^o(X_S, S) - Y_S)^2]$ among all models (whether linear or not) that satisfy demographic parity. If (19) holds, the projection-to-fairness is given by (20).*

We can see from (18) that the PTF model does not satisfy equal treatment: To calculate the loss forecast for a bank, we need to know its identity s. We have included the special case of (20) because it more nearly parallels the type of model we seek in (4). The coefficients in (20) are weighted averages of bank-specific coefficients. The model in (20) satisfies equal treatment with respect to the standardized features Z_s rather than the raw features X_s: Two banks with the same standardized features will receive the same forecasts. However, the means for the two banks could be very different—the standardization is done separately for each bank—indicating that one bank's portfolio may be much riskier than the other bank's portfolio. In treating standardized characteristics for different banks as comparable, the PTF model implicitly evaluates the riskiness of each bank relative to the distribution for that bank. The suitability of PTF in our setting is therefore questionable.

The root of the problem is that demographic parity is too strong a property for our setting. Ensuring that a hiring decision is independent of race or gender is important, but forcing the distribution of loss projections to be independent of bank identity ignores relevant differences in banks' portfolios. Whereas the pooled model (11)–(12) does too little to address heterogeneity across banks, the PTF model goes too far in leveling differences. The next section provides a better balance.

4.2. Formal Equality of Opportunity

Johnson et al. (2020) introduce the concept of FEO in regression, based on the use of the term in political philosophy, for which they cite the review in Arneson (2015). According to Arneson (2015), FEO means that "positions and posts that confer superior advantages should be open to all applicants. Applications are assessed on their merits."

In adapting this idea to our setting, it is helpful to make a contrast with the previous section: Whereas demographic parity requires that loss forecasts be independent of bank identity, FEO allows bank dependence, but only through legitimate portfolio characteristics—through the bank's "merits." This notion aligns well with the Fed policy, quoted earlier, that "two firms with the same portfolio receive the same results." The objective of FEO in regression, as developed by Johnson et al. (2020), is to ensure that a protected attribute (for us, bank identity) has no direct or "causal" impact on a model's predictions. The predictions may be correlated with bank identity if different banks tend to have different levels of exposure to legitimate portfolio features.

To develop this idea in our setting, we introduce the centered dummy variables

$$U_i(s) = \mathbf{1}\{s = i\} - p_i, \quad i = 1, \ldots, \overline{S} - 1,\ s = 1, \ldots, \overline{S}. \tag{21}$$

We discuss the implications of centering later. For any coefficients $\alpha, \delta_1, \ldots \delta_{\overline{S}-1} \in \mathbb{R}$ and $\beta \in \mathbb{R}^d$, and any $x \in \mathbb{R}^d$, let

$$\hat{Y}(x, s) = \alpha + \sum_i \delta_i U_i(s) + \beta^\top x. \tag{22}$$

We have included the bank label s as an argument of $\hat{Y}$ because U_i depends on s. Let α_F, $\{\delta_i, i = 1, \ldots, \overline{S} - 1\}$, and β_F solve the error minimization problem

$$\min_{\alpha, \{\delta_i\}, \beta} \mathsf{E}[(\hat{Y}(X_S, S) - Y_S)^2]. \tag{23}$$

With the coefficients that minimize (7), (22) becomes the linear projection of Y_S onto the span of $\{1, U_1(S), \ldots, U_{\overline{S}-1}(S), X_S\}$, evaluated at $S = s$ and $X_S = x$. Now drop the centered dummy variables U_i and define

$$\hat{Y}_F(x) = \alpha_F + \beta_F^\top x. \tag{24}$$

The FEO loss forecast for bank s is $\hat{Y}_F(X_s)$.

Steps (22)–(24) result from applying the definition of an impartial estimate (their definition 2) in Johnson et al. (2020). (More precisely, Steps (22)–(24) define a population counterpart of the sample formulation in Johnson et al. 2020.) The procedure in (22)–(24) can be interpreted as follows: pool losses and portfolio features across banks; regress losses on portfolio features with bank fixed effects included; throw away the fixed effects in forecasting future losses. The resulting model (24) is an equal-treatment model, with no explicit dependence on bank identity. Centering the discarded variables U_i ensures that $\mathsf{E}[\hat{Y}_F(X_S)] = \mathsf{E}[Y_S]$, so dropping the fixed effects does not introduce an overall bias.

We will say more about the implications of this approach, but we first show that our setting allows an explicit expression for the FEO coefficients.

Proposition 4.2. (i) *The FEO coefficients are given by*

$$\beta_F = \mathsf{E}[\Sigma_S]^{-1} \mathsf{E}[\Sigma_S \beta_S], \tag{25}$$

and

$$\alpha_F = \mathsf{E}[Y_S] - \beta_F^\top \overline{\mu}. \tag{26}$$

In particular, in the scalar case,

$$\beta_F = \frac{\sum_s p_s \sigma_s^2 \beta_s}{\sum_s p_s \sigma_s^2}. \tag{27}$$

(ii) *We also have*

$$\beta_F = \mathsf{var}[X_S - \mu_S]^{-1} \mathsf{cov}[X_S - \mu_S, Y_S], \tag{28}$$

so $\beta_F^\top (X_S - \mu_S)$ is the linear projection of $Y_S - \mathsf{E}[Y_S]$ onto $X_S - \mu_S$.

We encountered (27) in (15) as a special case of the pooled coefficient when the bank means μ_s are constant. The general case in (25) similarly coincides with the pooled coefficient in (11) when the means are constant. In other words, introducing the bank-level fixed effects in (22) purges β_F of the effect of different feature means across banks; dropping these fixed effects in (24) ensures that the regulator's model has no explicit dependence on bank identity and satisfies equal treatment.

In what sense is this procedure fair? We adapt the interpretation in Johnson et al. (2020) to our setting. Write $U = (U_1, \ldots, U_{\overline{S}-1})^\top$ for the vector of centered dummy variables. Write $\mathsf{cov}[X_S, U(S)]$ for the $d \times (\overline{S} - 1)$ matrix of covariances between the components of X_S and $U(S)$. Let

$$\Lambda = (\mathsf{var}[X_S])^{-1} \mathsf{cov}[X_S, U(S)]. \tag{29}$$

This matrix minimizes $\mathsf{E}[\| U(S) - \Lambda^\top (X_S - \overline{\mu}) \|^2]$, so $\Lambda^\top (X_S - \overline{\mu})$ is the linear projection of the bank-identity variables $U(S)$ onto the centered portfolio features $X_S - \overline{\mu}$. The relationship between β_{Pool} and β_F can be expressed as follows.

Proposition 4.3. *The coefficients* β_{Pool} *and* β_F *satisfy*

$$\beta_{Pool} = \beta_F + \Lambda\delta, \tag{30}$$

where $\delta = (\delta_1, \dots, \delta_{\overline{S}-1})^\top$ *is the vector of coefficients from* (22)–(23). *In particular,*

$$\delta_s = (\alpha_s + \beta_s \mu_s) - (\alpha_{\overline{S}} + \beta_{\overline{S}} \mu_{\overline{S}}) - \beta_F^\top(\mu_s - \mu_{\overline{S}}),\ \ s = 1, \dots, \overline{S} - 1. \tag{31}$$

We can write the forecast in (22), using the optimal coefficients from (23) as

$$\hat{Y}(x, s) = \mathsf{E}[Y_S] + \delta^\top U(s) + \beta_F^\top (x - \overline{\mu}); \tag{32}$$

This is the linear projection of Y_S onto $(1, U(S), X_S)$, evaluated at $S = s$, $X_s = x$. Let $\hat{Y}_P(x) = \alpha_{Pool} + \beta_{Pool}^\top x$ denote the forecast based on the pooled coefficients (11) and (12). Decomposing $U(S)$ into its projection onto $X_S - \overline{\mu}$ and an orthogonal component leads to the following contrast of these forecasts:

$$\hat{Y}(x, s) = \mathsf{E}[Y_S] + \delta^\top \Lambda^\top (x - \overline{\mu}) + \delta^\top [U(s) - \Lambda^\top (x - \overline{\mu})] + \beta_F^\top (x - \overline{\mu}) \tag{33}$$

$$\hat{Y}_P(x) = \mathsf{E}[Y_S] + \delta^\top \Lambda^\top (x - \overline{\mu}) + \beta_F^\top (x - \overline{\mu}), \tag{34}$$

$$\hat{Y}_F(x) = \mathsf{E}[Y_S] \qquad\qquad\qquad + \beta_F^\top (x - \overline{\mu}). \tag{35}$$

The term $\delta^\top [U(s) - \Lambda^\top (x - \overline{\mu})]$ in (33) affects the forecast through information in bank identity that is orthogonal to the legitimate features x. This would be *disparate treatment*, as in Johnson et al. (2020). Through "unawareness" (meaning that it has no functional dependence on bank identity) the pooled forecast (34) drops this term, but it retains $\delta^\top \Lambda^\top (x - \overline{\mu})$, as can be seen from (30).

The term $\delta^\top \Lambda^\top (x - \overline{\mu})$ is the problematic component of the pooled method. Although it does not explicitly use bank identity, this term relies on the fact that bank identity is to some extent predictable from portfolio features. Imagine the regulator forming loss forecasts from blinded data—the regulator does not know the identity of the bank. The term $\Lambda^\top (x - \overline{\mu})$ is the least-squares prediction of $U(s)$ from $x - \overline{\mu}$. In the pooled forecast (34), the regulator is implicitly "misdirecting" the data in the features $x - \overline{\mu}$ to try to identify the bank and then to adjust the forecast based on the inferred identity. The FEO forecast (35) removes this effect and retains only the direct effect of portfolio features on the loss rate.

In the terminology of Section 2.3, dropping $\delta^\top [U(s) - \Lambda^\top (x - \overline{\mu})]$ ensures the narrow sense of equal treatment —that loss forecasts not depend explicitly on bank identity. Dropping $\Lambda^\top (x - \overline{\mu})$ ensures a broader sense of equal treatment—that loss forecasts not depend on proxies for bank identity. Johnson et al. (2020) refer to their counterpart of $\Lambda^\top (x - \overline{\mu})$ as *disparate impact*, which is consistent with the notion of "proxy discrimination" as a particular type of disparate impact (Prince and Schwarcz 2019). In our setting, as noted in Section 2.3, the disparate impact of most immediate concern to banks is the omission of features from the Fed's models that might otherwise benefit individual banks. Omitted features may contribute to $\Lambda^\top (x - \overline{\mu})$, but dropping this term does not necessarily dispel banks' complaints. The banks' disagreements with the Fed concern the scope of portfolio features that should be modeled. We therefore prefer to associate $\delta^\top [U(s) - \Lambda^\top (x - \overline{\mu})]$ and $\Lambda^\top (x - \overline{\mu})$ with narrow and broad interpretations of the Fed's own principle of equal treatment (as discussed in Section 2.3) rather than with separate concerns for disparate treatment and disparate impact by the Fed and the banks. We emphasize the interpretation of $\Lambda^\top (x - \overline{\mu})$ as a misdirection of legitimate information, rather than as a contributor to disparate impact.

The FEO method offers a further advantage over the pooled method. Recall again from Section 2.3 (and Remark 3.1(iv)) that we interpret the banks' objections as calls for the inclusion of features that are omitted from the Fed's models. Under the condition $\mathsf{cov}[X_s, \epsilon_s] = 0$ in (3), the FEO coefficients of included features are unaffected by the omission of other features. The pooled coefficients do not in general have this property.

We will conclude in Section 4.5 that the FEO forecast is, in a precise sense, the best way to aggregate the bank-specific models into a single regulatory model. The FEO forecast has no direct dependence on bank identity; but it also removes the indirect dependence that results when bank identity is partly predictable from portfolio features. We discuss other methods for comparison.

4.3. Conditional Expectation Model

A similar misdirection of information occurs if we project the bank-specific models to an industry model in the sense of conditional expectation, rather than least squares. Suppose X_s has density g_s, and suppose $\mathsf{E}[\epsilon_s \mid X_s] = 0$, $s = 1, \dots, \overline{S}$. Then, by Bayes' rule,

$$\hat{Y}_C(x) \equiv \mathsf{E}[Y_S \mid X_S = x] = \frac{\sum_s p_s g_s(x)(\alpha_s + \beta_s^\top x)}{\sum_s p_s g_s(x)}. \tag{36}$$

This model satisfies equal treatment—$\hat{Y}_C(x)$ depends on the portfolio features x but not on a bank's identity. However, the point of the weights $p_s g_s(x)$ is to infer the identity of the bank from the features. Indeed, as discussed in Section 5, the conditional expectation $\mathsf{E}[Y_S \mid X_S = x]$ can be viewed as a nonlinear generalization of the pooled method, with some of the same shortcomings.

4.4. Substantive Equality of Opportunity

As discussed in Arneson (2015), a system in which admission decisions are made through a competitive exam open to everyone achieves formal equality of opportunity; however, if only the wealthy have access to the preparation required for the exam, the system fails to achieve SEO. In the regression setting, Johnson et al. (2020) interpret SEO to mean that any influence of protected attributes should be removed from other variables included in a regression model. In the analogy with the example of Arneson (2015), SEO would seek to remove the effect of economic status from performance on the exam, whereas FEO would accept exam scores as a legitimate basis for decision making. (Our use of SEO follows Johnson et al. (2020). For a broader interpretation of substantive equality in algorithmic fairness, see Green (2022).)

To apply these ideas to our setting, define the $(\overline{S}-1)\times d$ matrix

$$M=\mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S),X_S]; \qquad (37)$$

Then, M minimizes $\mathsf{E}[\|X_S-\overline{\mu}-M^\top U(S)\|^2]$. In accordance with definition 2 of Johnson et al. (2020), define

$$\hat{Y}_{SEO}(x,s)=\alpha_F+\beta_F^\top(x-M^\top U(s)), \qquad (38)$$

with α_F and β_F defined by (23). The SEO forecast adjusts the portfolio features x to remove the linear projection onto the centered bank dummy variables U. We can write (38) somewhat more explicitly as follows.

Proposition 4.4. *With M as in* (37)

$$M^\top U(s)=\sum_i(\mu_i-\mu_{\overline{S}})U_i(s)=\mu_s-\overline{\mu}, \qquad (39)$$

so the SEO forecast (38) *is given by*

$$\hat{Y}_{SEO}(x,s)=\alpha_F+\beta_F^\top(x-\mu_s+\overline{\mu}). \qquad (40)$$

The SEO forecast is the linear projection of Y_S *onto a constant and* $X_S-\mu_S$.

Recall from Section 4.1 that a model satisfies demographic parity if its forecasts are independent of bank identity. Let us say that a model satisfies *weak* demographic parity if its forecasts are *uncorrelated* with the bank identity variables $U_i(S)$. The centered features $X_S-\mu_S$ are uncorrelated with the $U_i(S)$. It therefore follows from Proposition 4.4 that SEO forecasts are uncorrelated with the $U_i(S)$. In other words, we have the following result.

Corollary 4.1. *The SEO forecast satisfies weak demographic parity.*

Under additional conditions, we get a stronger conclusion.

Corollary 4.2. *If the covariance matrix* Σ_s *and the distribution of* Z_s *in* (16) *are the same for all s, then the SEO model coincides with the standardized model* (20), *and both satisfy demographic parity.*

Under the conditions in the corollary, the mean adjustment in (40) is sufficient to give $\hat{Y}_{SEO}(X_s,s)$ the same distribution for all s. Put differently, PTF considers only the quantile of $\alpha_s+\beta_s^\top X_s$, relative to the distribution for bank s, to be legitimate information; SEO considers $X_s-\mu_s$ to be legitimate information. Under the conditions of the corollary, the two concepts coincide.

The mean adjustment in (40) requires knowledge of the bank identity s, so (38) does not satisfy Definition 3.1. The intent of the mean adjustment is to achieve a greater degree of equality. Consider the example with which began this section. If x represents an exam score and $\mu_1>\mu_0$ are the mean scores among wealthy and nonwealthy exam takers, (40) adjusts scores downward for wealthy exam takers and upward for nonwealthy exam takers.

Such an adjustment may be appropriate when the individuals or firms under evaluation are, in some sense, not responsible for their mean characteristic (or the mean in their peer group) and are therefore evaluated based on deviations from the mean. This type of consideration does not seem applicable to the stress-test setting, but it could arise more generally in settings where capital regulation intersects with other policy objectives.

One such example is suggested by the Paycheck Protection Program Lending Facility (PPPL) launched by the Federal Reserve early in the COVID crisis. The PPPL provided for loans to small businesses to be made by banks and guaranteed by the Small Business Administration. Under normal circumstances, the loans would increase participating banks' balance sheets and thus potentially increase their capital requirements. To promote use of the facility, banking regulators issued a rule excluding PPPL loans from capital requirements, thus "neutralizing the effects of participating in the PPPL Facility on regulatory capital requirements."[4] This "neutralizing" action is somewhat analogous to the SEO adjustment in that it removes responsibility for the larger balance sheet from the bank. The adjustments differ in that SEO adjusts for the mean whereas the PPPL adjustment removes the amount lent through the program.

4.5. Unified Perspective: Legitimate Information

All the methods we discussed can be seen as ways of choosing forecasts $\hat{Y}_s$, $s=1,\ldots\overline{S}$, (of the form $\hat{Y}(X_s)$ or $\hat{Y}(X_s,s)$) to minimize

$$\mathsf{E}[(\hat{Y}_S-Y_S)^2], \qquad (41)$$

subject to additional considerations. Table 1 summarizes the cases we have considered. In rows (i), (iv), and (v), we minimize (41) over the indicated coefficients. In (ii) and (iii), we allow g to be an arbitrary (suitably measurable) function of the indicated arguments. In (iii), we strengthen Condition (3) on the errors ϵ_s.

Table 1. Summary of Forecast Model Forms and Constraints

	Form	Constraint	Forecast
(i)	$\hat{Y}_s = \alpha + \beta^\top X_s$		Pooled (11)–(12)
(ii)	$\hat{Y}_s = g(X_s, s)$, some g	$\hat{Y}_S$ independent of S	PTF (Chzhen et al. 2020, Le Gouic et al. 2020)
(iii)	$\hat{Y}_s = g(X_s)$, some g, $\mathsf{E}[\epsilon_s \mid X_s] = 0$		Cond. exp. (36)
(iv)	$\hat{Y}_s = \alpha + \beta^\top X_s$	$\mathsf{cov}[Y_S - \hat{Y}_S, X_S - \mu_S] = 0$	FEO (24)
(v)	$\hat{Y}_s = \alpha + \lambda^\top U(s) + \beta^\top X_s$	$\mathsf{cov}[\hat{Y}_S, U(S)] = 0$	SEO (38)

Proposition 4.5. *In each row of Table 1, the squared loss (41) is minimized over forecasts of the form in the first column, subject to the constraint in the second column, by the model in the last column.*

The constraint in Table 1(v) is weak demographic parity. SEO implicitly takes the view that the only legitimate information in forecasting losses for bank s is the deviation $X_s - \mu_s$. In contrast, FEO takes the full set of features X_s as legitimate information. Through the constraint in Table 1(iv), it enforces a requirement we call *no misdirection of legitimate information*. FEO uses all X_s in forecasting losses; however, it chooses the coefficient β_F to be the coefficient in a regression of Y_S on $X_S - \mu_S$, which is the part of X_S orthogonal to bank identity. This condition ensures that the information in X_S is not misdirected to infer bank identity.

To make this idea precise, consider any model of the form (4). If we assume the intercept is chosen to match the unconditional mean, we may write the model as

$$\hat{Y}_\gamma(x) = \mathsf{E}[Y_S] + (\beta_F + \gamma)^\top (x - \overline{\mu}), \tag{42}$$

for some $\gamma \in \mathbb{R}^d$. With $\gamma = 0$, we get the FEO forecast (24).

Proposition 4.6. *If γ reduces errors in the sense that $\mathsf{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] < \mathsf{E}[(\hat{Y}_F(X_S) - Y_S)^2]$, then the forecast $\hat{Y}_\gamma$ misdirects legitimate information in the sense that*
(i) $\mathsf{cov}[\gamma^\top X_S, \delta^\top \Lambda^\top X_S] > 0$, *and*
(ii) $\mathsf{cov}[\gamma^\top M^\top U(S), \delta^\top U(S)] > 0$.

Recall that $\Lambda^\top (X_S - \overline{\mu})$ is the linear projection of the centered bank identity variables $U(S)$ onto the centered portfolio features $X_S - \overline{\mu}$. The condition in (i) therefore indicates that γ misdirects some of the legitimate information in X_S toward inferring bank identity. Thus, deviating from β_F in (42) either increases errors or misdirects information.

Property (ii) has a similar interpretation. The term $\delta^\top U(S)$ is the direct influence of bank identity on losses Y_S. The proposition states that any deviation γ that reduces forecast errors (relative to $\gamma = 0$) implicitly picks up some of the information in bank identity.

To further illustrate the contrast between FEO and SEO consider a simple example in which some component of X_s measures exposure to community development projects. Suppose for simplicity that this feature is uncorrelated with other features. In the SEO forecast, the only legitimate information from this exposure is a bank's deviation from its own mean. Years in which a bank had above average exposure would lead to higher loss forecasts, but the bank's average exposure to community development would not directly inform the forecasts—it is neutralized. In contrast, FEO treats the bank's total exposure (mean plus deviation) as legitimate information. Like SEO, in evaluating the impact of this exposure—that is, in estimating the coefficient on the exposure—it relies only on the within-bank variation. This ensures that the information in the exposure is not misdirected toward inferring the bank's identity, as could happen in the pooled regression.

4.6. Extension of FEO for Interaction Effects

Recall that the FEO forecast controls for bank fixed effects. One might similarly consider controlling for interactions between bank indicators and components of the feature vectors. This leads to a family of extensions of FEO that differ in which interactions they include. We will show that with a full set of interactions, the extended FEO model becomes the ATE model (14).

To examine this case, suppose the feature vector for each bank s is partitioned into two components: X_s and V_s. We extend FEO by including interactions with components of V_s but not with components of X_s. (Thus, in our discussion of FEO, V_s was empty.) We assume that for every bank s, the components of X_s are uncorrelated with the components of V_s. This allows a clear delineation between variables with and without interactions. Let $\nu_s = \mathsf{E}[V_s]$. The bank-specific models (1) now take the form

$$Y_s = \alpha_s + \beta_s^\top X_s + \gamma_s^\top V_s + \epsilon_s, \tag{43}$$

with ϵ_s uncorrelated with X_s and V_s.

We extend FEO to the following procedure:

(1) Project Y_S linearly onto one, $U_1(S), \ldots, U_{\overline{S}-1}(S)$, $X_S - \mu_S$, $V_S - \nu_S$, $U_1(S)V_S, \ldots, U_{\overline{S}}(S)V_S$. Let β_F denote the coefficient of $X_S - \mu_S$ and let γ_F denote the coefficient of $V_S - \nu_S$.

(2) Set $\hat{Y}_F(x, v) = \alpha_F + \beta_F^\top x + \gamma_F^\top v$, with α_F chosen so that $\mathsf{E}[\hat{Y}(X_S, V_S)] = \mathsf{E}[Y_S]$.

If V_S is empty, then we know from (28) that these steps do indeed reduce to the original FEO forecast. We have included the interaction $U_{\overline{S}}(S)V_S$ in the first step (even though we omitted $U_{\overline{S}}(S)$) to simplify the derivation of γ_F. Including this term means that the coefficients on the interactions $U_i(S)V_S$ are determined only up to constant, because $U_1(S)V_S + \cdots + U_{\overline{S}}(S)V_S = 0$. These coefficients are dropped in the second step, so their value is immaterial.

Proposition 4.7. *Suppose* $\mathrm{var}[X_s]$ *and* $\mathrm{var}[V_s]$ *have full rank and* X_s *and* V_s *are uncorrelated, for each* $s = 1, \ldots, \overline{S}$. *Then* β_F *is given by* (25) *and* (28), *and* $\gamma_F = \overline{\gamma} = \sum_s p_s \gamma_s$. *In particular, if interactions with* $U(S)$ *are included for all features, the FEO vector of coefficients reduces to the average treatment effect* (14).

This result allows us to interpret the ATE forecast as a version of the FEO forecast that removes the effects of certain interactions. As a convex combination of the bank-specific coefficients, the ATE coefficient retains some of the advantages of the FEO coefficient, particularly for the cross-bank effects studied in Section EC.1 in the online appendix.

However, we do not see a compelling case for controlling for interactions between bank identity and portfolio features. When we control for the bank-identity variables in FEO, we are ensuring that the industry β for legitimate features is not affected by heterogeneity in the banks' constants (the fixed effects). This reasoning does not necessarily extend to removing the influence of heterogeneity in exposures to portfolio features.

5. Nonlinear Models

Most of the ideas developed in previous sections for linear regressions extend to generalized linear models through a transformation of the response variable. For example, instead of working with the loss rate Y_s, we could specify a linear model for its logit transformation $\log(Y_s/(1 - Y_s))$.

However, we can also extend ideas from previous sections to more fully nonlinear models. Replace the mixture model in (6) with a general representation of the form

$$Y_S = g(S, X_S) + \epsilon_S, \quad \mathrm{E}[\epsilon_S \mid S, X_S] = 0. \tag{44}$$

In other words, the loss for bank s is given by $g(s, X_s) + \epsilon_s$. We assume that $g(S, X_S)$ and ϵ_S are square-integrable. The counterpart of the pooled estimate becomes

$$f_{Pool}(x) \equiv \mathrm{E}[Y_S \mid X_S = x] = \mathrm{E}[g(S, X_S) \mid X_S = x].$$

This rule satisfies equal treatment—it has no functional dependence on S—but we argued earlier (in Section 4.3) that this forecast implicitly uses the information in the portfolio features x to infer bank identity.

To introduce a nonlinear version of the FEO forecast, we will make the relatively modest assumption that (44) admits a decomposition of the form

$$Y_S = f_0 + f_1(S) + f_2(X_S) + \epsilon, \quad \mathrm{E}[\epsilon \mid S] = \mathrm{E}[\epsilon \mid X_S] = 0, \tag{45}$$

with $f_0 = \mathrm{E}[Y_S]$, $f_1 : \{1, \ldots, \overline{S}\} \to \mathbb{R}$, $f_2 : \mathbb{R}^d \to \mathbb{R}$, and

$$\mathrm{E}[Y_S - f_0 - f_1(S) \mid X_S] = f_2(X_S), \tag{46}$$

$$\mathrm{E}[Y_S - f_0 - f_2(X_S) \mid S] = f_1(S), \tag{47}$$

$\mathrm{E}[f_1^2(S)] < \infty$, $\mathrm{E}[f_2^2(X_S)] < \infty$, and

$$\mathrm{E}[f_1(S)] = \mathrm{E}[f_2(X_S)] = 0.$$

Equations (46)–(47) are population versions of the backfitting algorithm in Hastie and Tibshirani (1986), which is a special case of the alternating conditional expectations algorithm of Breiman and Friedman (1985). Given an initial choice of f_1 (and known f_0), (46) defines an initial choice of f_2 through the regression of the residual $Y_S - f_0 - f_1(S)$ on X_S. Equation (47) then defines an updated choice of f_1. The algorithm iterates over (46) and (47). In writing (45), we are positing that this algorithm has a fixed point. Convergence of the backfitting algorithm is established under widely applicable conditions in Ansley and Kohn (1994).

We now introduce

$$\hat{Y}_F(x) = f_0 + f_2(x) \tag{48}$$

as a nonlinear counterpart of the FEO forecast. We justify this interpretation by showing that $\hat{Y}_F$ exhibits properties that are nonlinear counterparts of the key properties of the FEO forecast in Sections 4.2 and 4.5. To state the result, consider forecasts of the form

$$\hat{Y}_\gamma(x) = f_0 + f_2(x) + \gamma(x), \tag{49}$$

for some $\gamma : \mathbb{R}^d \to \mathbb{R}$ with $\mathrm{E}[\gamma(X_S)^2] < \infty$.

Proposition 5.1. *The nonlinear FEO forecast* (48) *satisfies*

$$\mathrm{cov}[\hat{Y}_F(X_S) - Y_S, X_S - \mathrm{E}[X_S \mid S]] = 0. \tag{50}$$

For $\hat{Y}_\gamma$ *as in* (49), *if* γ *reduces errors, in the sense that* $\mathrm{E}[(\hat{Y}_\gamma(X_S) - Y_S)^2] < \mathrm{E}[(\hat{Y}_F(X_S) - Y_S)^2]$, *then it misdirects legitimate information, in the sense that*

$$\mathrm{cov}[\gamma(X_S), \mathrm{E}[f_1(S) \mid X_S]] > 0 \tag{51}$$

and

$$\mathrm{cov}[\mathrm{E}[\gamma(X_S) \mid S], f_1(S)] > 0. \tag{52}$$

Property (50) parallels the condition in row (iv) of Table 1 that characterizes the FEO forecast in the linear setting. It says that the forecast error $\hat{Y}_F(X_S) - Y_S$ is uncorrelated with the legitimate information $X_S - \mathrm{E}[X_S \mid S]$, which is the component of X_S orthogonal to bank identity S. Properties (51)–(52) parallel conditions (i) and (ii) in Proposition 4.6. In particular, in (51), $\mathrm{E}[f_1(S) \mid X_S]$ is the expected impact of bank identity inferred from portfolio features; the positive covariance with $\gamma(X_S)$ thus indicates that γ misdirects some of the information in X_S to inferring S. The approach of this section is further explored numerically in the online appendix.

We briefly contrast our FEO forecast in (48) with an alternative approach to extending fairness concerns to complex, nonlinear models. The alternative seeks to strip X_S of any protected attributes before a model is estimated. Examples of this general approach include Grünewälder and Khaleghi (2021) and Madras et al. (2018). This approach is primarily concerned with ensuring demographic parity: If a model has no access—not even indirect access—to a protected attribute, its forecasts will be independent of the attribute. However, we argued previously that demographic parity is too strong a condition for our setting. Our FEO forecast in (48) treats all the information in X_S as legitimate information—even elements that could help infer S—but it ensures that the information is not in fact misdirected to infer S.

6. Concluding Remarks

The current practice of regulatory stress testing ignores bank heterogeneity in loss models as a matter of policy and principle. We argued that simply pooling banks can distort coefficients on legitimate features and is vulnerable to implicit misdirection of legitimate information to infer bank identity. We examined various ways of incorporating fairness considerations and shown that estimating and discarding centered bank fixed effects addresses the deficiencies of pooling—and it does so in an optimal sense.

Beyond this specific recommendation, the broader conclusion to be drawn from our analysis is that accuracy and equal treatment can more effectively be addressed by accounting for bank heterogeneity rather than ignoring it. Although we focused on the stress testing application, our analysis applies more generally to settings requiring the fair aggregation of individually tailored models into a single common model.

Acknowledgments

For many helpful comments, the authors thank the referees and our discussant (Matthew Plosser) at the 2022 Federal Reserve Stress Testing Conference.

Appendix A. Proofs

Proof of Proposition 3.1. Problem (7) is solved by the linear projection of Y_S onto the span of one and X_S. If $\mathsf{var}[X_S]$ is invertible, then the coefficients of the linear projection are given by (12) and

$$\beta_{Pool} = \mathsf{var}[X_S]^{-1}\mathsf{cov}[Y_S, X_S];$$

see, for example, Wooldridge (2010, p. 25). In (9)–(10) we can write

$$\mathsf{var}[X_S] = \sum_s p_s W_s = \sum_s p_s \Sigma_s + \mathsf{var}[\mu_S].$$

This matrix is positive definite because we assumed that each Σ_s is positive definite, so $\mathsf{var}[X_S] = \mathsf{E}[W_S]$ is indeed invertible. To evaluate $\mathsf{cov}[Y_S, X_S]$ for Y_S in (6), we first note that

$$\begin{aligned}\mathsf{cov}[X_S, \epsilon_S] &= \mathsf{E}[\mathsf{cov}[X_S, \epsilon_S \mid S]] + \mathsf{cov}[\mathsf{E}[X_S \mid S], \mathsf{E}[\epsilon_S \mid S]]\\ &= \mathsf{E}[0] + \mathsf{cov}[\mu_S, 0] = 0.\end{aligned}$$

It follows that

$$\begin{aligned}&\mathsf{cov}[Y_S, X_S]\\ &= \mathsf{E}[\mathsf{cov}[\alpha_S, X_S \mid S]] + \mathsf{cov}[\mathsf{E}[\alpha_S \mid S], \mathsf{E}[X_S \mid S]]\\ &\quad + \mathsf{E}[\mathsf{cov}[\beta_S^\top X_S, X_S \mid S]] + \mathsf{cov}[\mathsf{E}[\beta_S^\top X_S \mid S], \mathsf{E}[X_S \mid S]]\\ &= 0 + \mathsf{cov}[\alpha_S, \mu_S] + \mathsf{E}[\Sigma_S \beta_S] + \mathsf{E}[\mathsf{var}[\mu_S]\beta_S]\\ &= \mathsf{cov}[\alpha_S, \mu_S] + \mathsf{E}[W_S \beta_S]. \quad \square\end{aligned}$$

Proof of Proposition 4.1. By proposition 3.4 of Chzhen et al. (2020) or theorem 6 of Le Gouic et al. (2020), the expected squared error is minimized subject to demographic parity by the rule that assigns to bank s with features x the loss forecast

$$\hat{Y}_{PTF}(x, s) = \sum_i p_i F_i^{-1}(F_s(\alpha_s + \beta_s^\top x)), \tag{A.1}$$

where F_s is the cumulative distribution function of $\alpha_s + \beta_s^\top X_s$. By construction, F_s is then also the cumulative distribution function of $\alpha_s^o + \beta_s^{o\top} Z_s$, which is normal with mean α_s^o and variance $\|\beta_s^o\|^2$. Writing Φ for the standard normal distribution function, we get

$$F_s(y) = \Phi\left(\frac{y - \alpha_s^o}{\|\beta_s^o\|}\right), \quad F_i^{-1}(q) = \alpha_i^o + \|\beta_i^o\|\Phi^{-1}(q).$$

Making these substitutions in (A.1) and writing $\alpha_s^o + \beta_s^{o\top} z_s$ for $\alpha_s + \beta_s^\top x$, with $z_s = \Sigma_s^{-1}(x - \mu_s)$, we get

$$\begin{aligned}\hat{Y}_{PTF}(x, s) &= \sum_i p_i F_i^{-1}(F_s(\alpha_s^o + \beta_s^{o\top} z_s))\\ &= \sum_i p_i F_i^{-1}(\Phi(\beta_s^{o\top} z_s / \|\beta_s^o\|))\\ &= \sum_i p_i \{\alpha_i^o + \|\beta_i^o\|\Phi^{-1}(\Phi(\beta_s^{o\top} z_s / \|\beta_s^o\|))\}\\ &= \sum_i p_i \{\alpha_i^o + \|\beta_i^o\|\beta_s^{o\top} z_s / \|\beta_s^o\|\},\end{aligned}$$

which is (18). (Demographic parity holds because the distribution of $\beta_s^{o\top} Z_s / \|\beta_s^o\|$ does not depend on s.) Under (19), $\|\beta_i^o\|\beta_s^o / \|\beta_s^o\| = \|a_i\beta\|a_s\beta / \|a_s\beta\| = a_i\beta = \beta_i^o$, and we get (20). □

Proof of Proposition 4.2. We can rewrite $\hat{Y}(x, s)$ in (22) as

$$\hat{Y}(x, s) = \sum_{i=1}^{\bar{S}} a_i \mathbf{1}\{s = i\} + \beta^\top x,$$

for suitable a_i. Minimizing (23) over the a_i and β yields the same value for β as minimizing (23) using (22) because the indicators $\mathbf{1}\{s = i\}$ have the same span as the $U_i(s)$ and a constant. Thus, the β_F defined by (23) is the coefficient of X_S in the regression of Y_S on X_S and the indicators $\mathbf{1}\{S = i\}$. By the Frisch-Waugh-Lovell theorem (Angrist and Pischke 2008, pp. 35–36), we can therefore evaluate β_F as the coefficient in the regression of Y_S on the component of X_S orthogonal to the other variables, which in our case are the indicators. The projection of X_S onto the indicators is given by $\sum_i \mu_i \mathbf{1}\{S = i\} = \mu_S$, so the orthogonal component is $X_S - \mu_S$. We may therefore evaluate β_F as the coefficient in the regression of $Y_S - \mathsf{E}[Y_S]$ on $X_S - \mu_S$, which is (28). For the first factor in (28), we have

$$\begin{aligned}\mathsf{var}[X_S - \mu_S] &= \mathsf{E}[\mathsf{var}[X_S - \mu_S \mid S]] + \mathsf{var}[\mathsf{E}[X_S - \mu_S \mid S]]\\ &= \mathsf{E}[\Sigma_S] + 0.\end{aligned}$$

For the second factor, we similarly have

$$\begin{aligned}\mathsf{cov}[X_S - \mu_S, Y_S] &= \mathsf{E}[\mathsf{cov}[X_S - \mu_S, Y_S \mid S]]\\ &= \mathsf{E}[\mathsf{cov}[X_S - \mu_S, \beta_S^\top X_S \mid S]] = \mathsf{E}[\Sigma_S \beta_S],\end{aligned}$$

so (25) follows. The optimal α_F in (23) ensures that $\mathsf{E}[\hat{Y}(X_S, S)] = \mathsf{E}[Y_S]$, which yields (26). □

Proof of Proposition 4.3. The minimization in (23) yields coefficients α_F, δ, and β_F, with which we can write

$$Y_S = \alpha_F + \sum_{i=1}^{\overline{S}-1} \delta_i U_i(S) + \beta_F^\top X_S + u, \tag{A.2}$$

where the error u has mean zero and is uncorrelated with $U(S)$ and X_S. We thus have

$$\begin{aligned}\beta_{Pool} &= \mathsf{var}[X_S]^{-1}\mathsf{cov}[X_S, Y_S]\\ &= \mathsf{var}[X_S]^{-1}\{\mathsf{cov}[X_S, \beta_F^\top X_S] + \mathsf{cov}[X_S, \delta^\top U(S)]\}\\ &= \mathsf{var}[X_S]^{-1}\{\mathsf{var}[X_S]\beta_F + \mathsf{cov}[X_S, U(S)]\delta\}\\ &= \beta_F + \Lambda\delta,\end{aligned}$$

using the expression for Λ in (29) for the last step.

Next, we evaluate δ. Using (A.2), we can derive δ as the vector of coefficients in a regression of $Y_S - \beta_F^\top X_S$ on $U(S)$. Thus,

$$\begin{aligned}\delta &= \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), Y_S - \beta_F^\top X_S]\\ &= \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), Y_S] - \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), X_S]\beta_F.\end{aligned} \tag{A.3}$$

To evaluate $\mathsf{var}[U(S)]^{-1}$, we first note that

$$\mathsf{var}[U(S)] = \begin{pmatrix} p_1 - p_1^2 & -p_1 p_2 & \cdots & -p_1 p_{\overline{S}-1}\\ -p_2 p_1 & p_2 - p_2^2 & \cdots & -p_2 p_{\overline{S}-1}\\ \vdots & \vdots & \ddots & \vdots\\ -p_{\overline{S}-1} p_1 & -p_{\overline{S}-1} p_2 & \cdots & p_{\overline{S}-1} - p_{\overline{S}-1}^2 \end{pmatrix};$$

direct multiplication then verifies that

$$\mathsf{var}[U(S)]^{-1} = \begin{pmatrix} 1/p_1 + 1/p_{\overline{S}} & 1/p_{\overline{S}} & \cdots & 1/p_{\overline{S}}\\ 1/p_{\overline{S}} & 1/p_2 + 1/p_{\overline{S}} & \cdots & 1/p_{\overline{S}}\\ \vdots & \vdots & \ddots & \vdots\\ 1/p_{\overline{S}} & 1/p_{\overline{S}} & \cdots & 1/p_{\overline{S}-1} + 1/p_{\overline{S}} \end{pmatrix}.$$

The vector $\mathsf{cov}[U(S), Y_S]$ has elements

$$[\mathsf{cov}[U(S), Y_S]]_s = p_s(\mathsf{E}[Y_s] - \mathsf{E}[Y_S]), \quad s = 1, \ldots, \overline{S} - 1;$$

and row s of the matrix $\mathsf{cov}[U(S), X_S]$ is given by $p_s(\mu_s - \overline{\mu})^\top$. Thus, for $s = 1, \ldots, \overline{S} - 1$, we have the vector elements

$$\begin{aligned}&(\mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), Y_S])_s\\ &\quad = (\mathsf{E}[Y_s] - \mathsf{E}[Y_S]) + \sum_{i=1}^{\overline{S}-1} p_i(\mathsf{E}[Y_i] - \mathsf{E}[Y_S])/p_{\overline{S}}\\ &\quad = \mathsf{E}[Y_s] - \mathsf{E}[Y_{\overline{S}}],\end{aligned}$$

and similarly row s of the matrix $\mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), X_S]$ is given by

$$(\mu_s - \overline{\mu})^\top + \sum_{i=1}^{\overline{S}-1} p_i(\mu_i - \overline{\mu})^\top / p_{\overline{S}} = (\mu_s - \mu_{\overline{S}})^\top. \tag{A.4}$$

Combining these terms in (A.3) yields (31). □

Proof of Proposition 4.4. We derived an expression for the rows of M in (56), and (39) follows from that expression. By applying Expression (28) for β_F in (40), we see that $\hat{Y}_{SEO}$ is the claimed projection. □

Proof of Corollary 4.2. From (25), we know that if $\Sigma_s \equiv \Sigma$, then $\beta_F = \mathsf{E}[\beta_S]$. From (20), we get $\overline{\beta}^o = \sum_i p_i \beta_i^o = \sum_i p_i \Sigma^{1/2}\beta_i = \Sigma^{1/2}\mathsf{E}[\beta_S] = \Sigma^{1/2}\beta_F$. Thus, $\beta_F^\top(x - \mu_s) = \overline{\beta}^{o\top}\Sigma^{-1/2}(x - \mu_s) = \overline{\beta}^{o\top} z_s$. It follows that (20) and (40) coincide because they have the same overall mean. If the distribution of Z_s does not depend on s, then (20) satisfies demographic parity. □

Proof of Proposition 4.5. The claim for (i) simply restates Proposition 3.1. The constraint in (ii) is demographic parity, so the optimizer follows from the definition of projection to fairness. The constraint in (v) requires $\mathsf{cov}[\lambda^\top U(S) + \beta^\top X_S, U(S)] = 0$. Rearranging this equation, we get $\lambda = -(\mathsf{var}[U(S)])^{-1}\mathsf{cov}[U(S), X_S]\beta$; that is, $\lambda = -M\beta$. Making this substitution in the form of $\hat{Y}_S$ in row (v), (41) becomes

$$\begin{aligned}&\mathsf{E}[(Y_S - \alpha - \lambda^\top U(S) - \beta^\top X_S)^2]\\ &\quad = \mathsf{E}[(Y_S - \alpha - \beta^\top[X_S - M^\top U(S)])^2].\end{aligned} \tag{A.5}$$

Minimizing this expression over α and β yields the coefficients in a linear regression of Y_S on a constant $X_S - M^\top U(S)$. In light of Proposition 4.4, the optimal β in (A.5) is then the coefficient on $X_S - \mu_S$ in a regression of Y_S on a constant $X_S - \mu_S$. It follows from (28) that the optimal β in (A.5) is therefore β_F. Because $\mathsf{E}[U(S)] = 0$, the minimizing α in (A.5) is the α_F defined by (23). We have thus shown that the optimal forecast in row (v) is

$$\hat{Y}_s = \alpha_F + \beta_F^\top(X_s - M^\top U) = \alpha_F + \lambda^\top U(s) + \beta_F^\top X_s.$$

In case (iv), by applying (39), we see that the constraint requires $\mathsf{cov}[Y_S - \beta^\top X_S, X_S - M^\top U(S)] = 0$, so $\beta = (\mathsf{cov}[X_S, X_S - M^\top U(S)])^{-1}\mathsf{cov}[Y_S, X_S - M^\top U(S)]$. Using the fact that $X_S - M^\top U(S)$ is orthogonal to $U(S)$, we get

$$\begin{aligned}&\mathsf{cov}[X_S, X_S - M^\top U(S)]\\ &\quad = \mathsf{cov}[X_S - M^\top U(S), X_S - M^\top U(S)]\\ &\qquad + \mathsf{cov}[M^\top U(S), X_S - M^\top U(S)]\\ &\quad = \mathsf{var}[X_S - M^\top U(S)],\end{aligned}$$

and therefore $\beta = (\mathsf{var}[X_S - M^\top U(S)])^{-1}\mathsf{cov}[Y_S, X_S - M^\top U(S)]$. In other words, the optimal β in (iv) is the coefficient in a linear regression of Y_S on $X_S - M^\top U(S)$. As noted in the discussion of (v), this is β_F, and it follows from $\mathsf{E}[U(S)] = 0$ that the optimal α in (iv) is α_F. □

Proof of Proposition 4.6. By construction, the least-squares projection of Y_S onto a constant and X_S is given by the pooled forecast, so

$$Y_S = \mathsf{E}[Y_S] + \beta_{Pool}^\top(X_S - \overline{\mu}) + \epsilon_P,$$

for some orthogonal error ϵ_P with a variance σ_P^2 that does not depend on γ. We therefore have

$$\begin{aligned}\mathsf{E}[(\hat{Y}_\gamma(X_S)-Y_S)^2] &= \mathsf{E}[\{\hat{Y}_\gamma(X_S)-\mathsf{E}[Y_S]-\beta_{Pool}^\top(X_S-\overline{\mu})\}^2]+\sigma_P^2\\ &= \mathsf{E}[\{(\gamma-\Lambda\delta)^\top(X_S-\overline{\mu})\}^2]+\sigma_P^2,\end{aligned}$$

from which (i) follows.

Using the linear projection of Y_S onto $(1, U(S), X_S)$ in (32), we can write

$$Y_S = \mathsf{E}[Y_S]+\delta^\top U(S)+\beta_F^\top(X_S-\overline{\mu})+\epsilon,$$

for some orthogonal error ϵ with a variance σ_ϵ^2 that does not depend on γ. We therefore have

$$\begin{aligned}\mathsf{E}[(\hat{Y}_\gamma(X_S)-Y_S)^2] &= \mathsf{E}[\{\gamma^\top(X_S-\overline{\mu})-\delta^\top U(S)\}^2]+\sigma_\epsilon^2\\ &= \mathsf{E}[\{\gamma^\top(X_S-\mu_S)+\gamma^\top(\mu_S-\overline{\mu})-\delta^\top U(S)\}^2]\\ &\quad+\sigma_\epsilon^2\\ &= \mathsf{E}[\{\gamma^\top(X_S-\mu_S)+(\gamma^\top M^\top-\delta^\top)U(S)\}^2]+\sigma_\epsilon^2\\ &= \mathsf{E}[\{\gamma^\top(X_S-\mu_S)\}^2]\\ &\quad+\mathsf{E}[\{(\gamma^\top M^\top-\delta^\top)U(S)\}^2]+\sigma_\epsilon^2,\end{aligned}$$

where the third equality uses (39), and the last equality uses the orthogonality of $X_S-\mu_S$ and $U(S)$. If this expression is smaller than the corresponding value with $\gamma = 0$, then (ii) must hold. □

Proof of Proposition 4.7. We saw in the proof of Proposition 4.2 that $X_S-\mu_S$ is uncorrelated with the centered indicators $U_i(S)$. It is also uncorrelated with $V_S-\nu_S$ because

$$\mathsf{E}[(X_S-\mu_S)(V_S-\nu_S)] = \sum_s p_s\mathsf{E}[(X_s-\mu_s)(V_s-\nu_s)] = 0,$$

under our assumption that X_s and V_s are uncorrelated. Similarly,

$$\begin{aligned}\mathsf{E}[(X_S-\mu_S)U_i(S)V_S] &= p_i\mathsf{E}[(X_i-\mu_i)V_i]-p_i\mathsf{E}[(X_S-\mu_S)V_S]\\ &= 0,\end{aligned}$$

so $X_S-\mu_S$ is uncorrelated with the interaction terms. Thus, $X_S-\mu_S$ is uncorrelated with all the elements of $\mathcal{O}=\{1, U(S), V_S-\nu_S, U_1(S)V_S,\ldots,U_{\overline{S}}(S)V_S\}$.

Starting from the representation of (43) as

$$Y_S = \sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}\{\alpha_i+\beta_i^\top X_S+\gamma_i^\top V_S+\epsilon_i\},$$

we may write

$$\begin{aligned}Y_S &= \beta_S^\top(X_S-\mu_S)+\sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}(\alpha_i+\beta_i^\top\mu_i)+\sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}\gamma_i^\top V_S+\epsilon_S\\ &\equiv \beta_S^\top(X_S-\mu_S)+\tilde{Y}+\epsilon_S,\end{aligned}$$

which expresses Y_S as the sum of three mutually orthogonal terms. As $X_S-\mu_S$ is uncorrelated with $\mathcal{O}$, and $\tilde{Y}$ is uncorrelated with $X_S-\mu_S$, we may calculate the projection of Y_S onto the span of $X_S-\mu_S$ and $\mathcal{O}$ by projecting $\beta_S^\top(X_S-\mu_S)$ onto $X_S-\mu_S$ and projecting $\tilde{Y}$ onto $\mathcal{O}$.

We know from (28) that the projection of $\beta_S^\top(X_S-\mu_S)$ onto $X_S-\mu_S$ is $\beta_F^\top(X_S-\mu_S)$; in other words, including V_S and the interaction terms does not change β_F.

For the projection of $\tilde{Y}$ onto $\mathcal{O}$, let $a_i=\alpha_i+\beta_i^\top\mu_i+\overline{\gamma}^\top\nu_i$ and $\overline{a}=\sum_i p_i a_i$. Then,

$$\begin{aligned}\tilde{Y} &= \sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}(\alpha_i+\beta_i^\top\mu_i)+\sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}\gamma_i^\top V_S\\ &= \sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}(\alpha_i+\beta_i^\top\mu_i)+\sum_{i=1}^{\overline{S}}U_i(S)\gamma_i^\top V_S+\sum_{i=1}^{\overline{S}}p_i\gamma_i^\top V_S\\ &= \sum_{i=1}^{\overline{S}}\mathbf{1}\{S=i\}(\alpha_i+\beta_i^\top\mu_i+\overline{\gamma}^\top\nu_i)+\sum_{i=1}^{\overline{S}}U_i(S)\gamma_i^\top V_S\\ &\quad+\sum_{i=1}^{\overline{S}}p_i\gamma_i^\top(V_S-\nu_S)\\ &= \overline{a}+\sum_{i=1}^{\overline{S}-1}U_i(S)(a_i-a_{\overline{S}})+\sum_{i=1}^{\overline{S}}U_i(S)\gamma_i^\top V_S+\overline{\gamma}^\top(V_S-\nu_S).\end{aligned}$$

Thus, $\tilde{Y}$ is in the span of $\mathcal{O}$, and its coefficient on $V_S-\nu_S$ is $\overline{\gamma}$. With all $\mathsf{var}[V_s]$ having full rank, $V_S-\nu_S$ is not spanned by the other elements of $\mathcal{O}$, so its coefficient $\overline{\gamma}$ is uniquely determined. □

Proof of Proposition 5.1. For the first claim, we have

$$\begin{aligned}&\mathsf{cov}[\hat{Y}_F(X_S)-Y_S, X_S-\mathsf{E}[X_S\mid S]]\\ &= -\mathsf{E}[(f_1(S)+\epsilon)(X_S-\mathsf{E}[X_S\mid S])]\\ &= -\mathsf{E}[f_1(S)(X_S-\mathsf{E}[X_S\mid S])]-\mathsf{E}[\epsilon X_S]+\mathsf{E}[\epsilon\mathsf{E}[X_S\mid S]]\\ &= 0+\mathsf{E}[\mathsf{E}[\epsilon\mid S]\mathsf{E}[X_S\mid S]]-\mathsf{E}[\mathsf{E}[\epsilon\mid X_S]X_S]=0.\end{aligned}$$

For the second claim, we have

$$\begin{aligned}\mathsf{E}[(\hat{Y}_\gamma(X_S)-Y_S)^2] &= \mathsf{E}[(\gamma(X_S)-f_1(S)-\epsilon)^2]\\ &= \mathsf{E}[(\gamma(X_S)-f_1(S))^2]+\mathsf{E}[\epsilon^2].\end{aligned}$$

The last step uses

$$\mathsf{E}[(\gamma(X_S)-f_1(S))\epsilon] = \mathsf{E}[(\gamma(X_S)-f_1(S))\mathsf{E}[\epsilon\mid S]] = 0.$$

It now follows that if γ reduces the expected squared forecast error then $\mathsf{E}[\gamma(X_S)f_1(S)]>0$, which implies (51) and (52). □

Endnotes

[1] See https://www.federalreserve.gov/supervisionreg/files/goldman-sachs-group-inc-20200904.pdf. Accessed September 20, 2023.

[2] For a perspective on the importance of culture, see, for example, "Enhancing Financial Stability by Improving Culture in the Financial Services Industry," a speech given by then president of the Federal Reserve Bank of New York, William C. Dudley, on October 20, 2014, https://www.newyorkfed.org/newsevents/speeches/2014/dud141020a.html. Assessed September 20, 2023.

[3] The Fed's stress tests previously included a qualitative component, but this component was dropped in 2019.

[4] *Federal Register*, Vol. 85, No. 71, p. 20389, April 13, 2020.

References

Agarwal S, An X, Cordell L, Roman RA (2020) Bank stress test results and their impact on consumer credit markets. Working paper, Federal Reserve Bank of Philadelphia, Philadelphia.

Agueh M, Carlier G (2011) Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* 43(2):904–924.

Angrist JD, Pischke JS (2008) *Mostly Harmless Econometrics* (Princeton University Press, Princeton, NJ).

Ansley CF, Kohn R (1994) Convergence of the backfitting algorithm for additive models. *J. Australian Math. Soc.* 57:316–329.

Arneson R (2015) Equality of opportunity. Zalta EN, ed. *Stanford Encyclopedia of Philosophy*, summer 2015 ed.

Baer G, Hopper G (2023) The Fed's stress test models are inaccurate. Something has to change. *Risk* (September 14). https://www.risk.net/comment/7957648/the-feds-stress-test-models-are-inaccurate-something-has-to-change.

Barocas S, Hardt M, Narayanan A (2019) Fairness in machine learning. Accessed September 20, 2023, https://fairmlbook.org/.

Bassett WF, Berrospide JM (2018) The impact of post stress test capital on bank lending. Working paper, Federal Reserve Board, Washington, DC.

BCBS (2019) *Overview of Pillar 2 Supervisory Review Practices and Approaches* (Bank for International Settlements, Basel, Switzerland).

Board of Governors (2021) *Dodd-Frank Act Stress Test 2021: Supervisory Stress Test Methodology* (Federal Reserve System, Washington, DC).

Breiman L, Friedman JH (1985) Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* 80:580–598.

Breuer T, Jandacka M, Rheinberger K, Summer M (2009) How to find plausible, severe, and useful stress scenarios. *Internat. J. Central Bank.* 5:205–224.

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020) Fair regression with Wasserstein barycenters. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates Inc., Red Hook, NY), 7321–7331.

Cope D, Hsu C, Lively C, Morgan J, Schuermann T, Sekeris E (2022) Stress testing for commercial, investment, and custody banks. *Handbook of Financial Stress Testing* (Cambridge University Press, Cambridge, UK), 247–270.

Covas FB, Rump B, Zakrajsek E (2014) Stress-testing US bank holding companies: A dynamic panel quantile regression approach. *Internat. J. Forecasting* 30:691–713.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proc. 3rd Innovations Theoretical Computer Sci. Conf. (ITCS '12)* (Association for Computing Machinery, New York, NY), 214–226.

Fernandes M, Igan D, Pinheiro M (2020) March madness in Wall Street: (What) does the market learn from stress tests? *J. Bank. Finance* 112:105250.

Flannery MJ (2019) Transparency and model evolution in stress testing. Preprint, submitted August 7, https://dx.doi.org/10.2139/ssrn.3431679.

Flannery M, Hirtle B, Kovner A (2017) Evaluating the information in the Federal Reserve stress tests. *J. Financial Intermediary* 29:1–18.

Flood MD, Korenko GG (2015) Systematic scenario selection: Stress testing and the nature of uncertainty. *Quant. Finance* 15:43–59.

Flood MD, Jones J, Pritsker M, Siddique A (2022) The role of heterogeneity in scenario design for financial stability stress testing. *Handbook of Financial Stress Testing* (Cambridge University Press, Cambridge, UK), 98–127.

Georgescu O-M, Gross M, Kapp D, Kok C (2017) Do stress tests matter? Evidence from the 2014 and 2016 stress test. Working paper, European Central Bank, Frankfurt, Germany.

Glasserman P, Tangirala G (2016) Are the Federal Reserve's stress test results predictable? *J. Alternative Investments* 18:82–97.

Glasserman P, Kang C, Kang W (2015) Stress scenario selection by empirical likelihood. *Quant. Finance* 15:25–41.

Green B (2022) Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophical Tech.* 35(90):1–32.

Grünewälder S, Khaleghi A (2021) Oblivious data for fairness with kernels. *J. Machine Learn. Res.* 22:1–36.

Guerrieri L, Modugno M (2021) The information content of stress test announcements. Working paper, Federal Reserve Board, Washington, DC.

Guerrieri L, Welch M (2012) Can macro variables used in stress test forecast the performance of banks? Working paper, Federal Reserve Board, Washington, DC.

Hastie T, Tibshirani R (1986) Generalized additive models. *Statist. Sci.* 1(3):297–318.

Hirtle B, Kovner A, Vickery J, Bhanot M (2016) Assessing financial stability: The Capital and Loss Assessment under Stress Scenarios (CLASS) model. *J. Bank. Finance* 69:S35–S55.

Hutchinson B, Mitchell M (2019) 50 years of test (un) fairness: Lessons for machine learning. *Proc. Conf. Fairness Accountability Transparency (FAT* '19)* (Association for Computing Machinery, New York, NY), 49–58.

Johnson K, Foster D, Stine R (2020) Impartial predictive modeling: Ensuring group fairness in arbitrary models. Preprint, submitted October 11, https://arxiv.org/abs/1608.00528.

Kapinos P, Mitnik OA (2016) A top-down approach to stress-testing banks. *J. Financial Service Res.* 49:229–264.

Kohn D, Liang N (2019) *Understanding the Effects of the U.S. Stress Tests* (Brookings Institution, Washington, DC).

Kupiec P (2020) Policy uncertainty and bank stress testing. *J. Financial Stability* 51:100761.

Le Gouic T, Loubes J-M, Rigollet P (2020) Projection to fairness in statistical learning, Preprint, submitted June 25, https://arxiv.org/abs/2005.11720.

Lipton Z, Chouldechova A, McAuley J (2018) Does mitigating ML's impact disparity require treatment disparity? Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Proc. 32nd Conf. Neural Inform. Processing Systems*, 31 (Curran Associates Inc., Red Hook, NY).

Madras D, Creager E, Pitassi T, Zemel R (2018) Learning adversarially fair and transferable representations. Preprint, submitted October 22, https://arxiv.org/abs/1802.06309.

Morgan DP, Peristiani S, Savino V (2014) The information value of the stress test. *J. Money Credit Bank.* 46:1479–1500.

Parlatore C, Philippon T (2022) Designing stress scenarios. NBER Working Paper No. w29901, National Bureau of Economic Research, Cambridge, MA.

Philippon T, Pessarossi P, Camara B (2017) Backtesting European stress tests. NBER Working Paper No. w23083, National Bureau of Economic Research, Cambridge, MA.

Prince AE, Schwarcz D (2019) Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Rev.* 105:1257–1358.

Pritsker MG (2017) Choosing stress scenarios for systemic risk through dimension reduction. Risk and Policy Analysis Unit Paper No. RPA 17-4, Federal Reserve Bank of Boston, Boston.

Sahin C, de Haan J, Neretina E (2020) Banking stress test effects on returns and risks. *J. Bank. Finance* 117:105843.

Schuermann T (2020) Capital adequacy pre- and postcrisis and the role of stress testing. *J. Money Credit Bank.* 52:87–105.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (MIT Press, Cambridge, MA).