Should Bank Stress Tests Be Fair?

Paul Glasserman and Mike Li
Columbia Business School

June 2022

Abstract

Regulatory stress tests have become the primary tool 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 present evidence that the overall impact can be material. We also discuss extensions to nonlinear models.

1 Introduction

In the aftermath of the 2008 financial crisis, U.S. banking regulators adopted stress testing as the 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 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 [9] (p.3), "two firms with the same portfolio receive the same results for that portfolio."

Banks have countered that the Fed's models fail to capture bank-specific features that should 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.

The heterogeneity among large banks motivates the questions we study: What is the best way to aggregate bank-specific models into an industry model? 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 of banks. 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, Hardt, and Narayanan [6] and Hutchinson and Mitchell [28]. 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. [15], 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 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. [12] and Le Gouic et al. [35] developed a method of particular importance for regression that Le Gouic et al. [35] 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, Foster, and Stine [30] 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 above: 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.

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; 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 investigate the empirical relevance of these considerations. We regress the loss rates of loan portfolios (credit cards, first lien mortgages, commercial real estate, and corporate

and industrial loans) on measures of portfolio quality (past-due rates and allowances for losses) and macroeconomic variables. We document significant heterogeneity across banks in their estimated coefficients, and we show that the difference in pooled and FEO estimates can be material. This investigation is limited by the information banks make public — the Federal Reserve has access to far more granular data in forecasting losses. We cannot claim to approximate the Fed's forecasts; our goal is to provide evidence of the importance of heterogeneity.

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 presents our empirical results. Section 6 investigates cross-bank externalities that arise when bank-specific models are aggregated into a common model: a change at one bank can impact loss forecasts at other banks. Effects of this type are inevitable under model aggregation, but we argue that the effects are more transparent and less objectionable under FEO than under a pooled model. Section 7 considers nonlinear models. Additional supporting material is included in appendices.

We conclude this introduction with a brief discussion of some other research on bank stress tests. Covas, Rump, and Zakrajsek [14], Kapinos and Mitnik [32], and Kupiec [34] find strong evidence of heterogeneity in banks' responses to macroeconomic shocks, and Kapinos and Mitnik [32] argue that ignoring heterogeneity can substantially underestimate projected capital requirements. The related models of Hirtle et al. [29] and Guerrieri and Welch [26] 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. [1], Bassett and Berrospide [7], and Flannery, Hirtle, and Kovner [18] 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 (including Breuer et al. [11], Flood and Korenko [20], Glasserman et al. [22], Pritsker [40], and Schuermann [42]) have advocated the use of multiple scenarios to capture different combinations of risk factors. Cope et al. [13] and Flood et al. [19] recommend designing scenarios to reflect bank heterogeneity. Parlatore and Philippon [38] 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 (as in Fernandes, Igan, and Pinheiro [16], Flannery, Hirtle, and Kovner [18], Georgescu et al. [21], Glasserman and Tangirala [23], Guerrieri and Modugno [25], Morgan, Peristiani, and Savino [37], and Sahin, de Haan, and Neretina [41]) or through subsequent bank performance (as in Kupiec [34] and Philippon, Pessarossi, and Camara [39]). Flannery [17]

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 [33] and Schuermann [42].

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 2021 DFAST covered 23 banks, down from a peak of 35 in 2018. 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 nine quarters. The 2021 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 commercial and industrial loans, one to credit cards, one to commercial real estate loans, and another to first lien residential mortgages. 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 the bank-specific variables 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 [9], 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." We will 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. But the right notion of consistency across firms becomes less clear when portfolios vary widely. The largest U.S. banks are a highly heterogeneous group. They include universal banks (like JPMorgan Chase and Bank of America), investment banks (Goldman Sachs and Morgan Stanley), custodians (BNY Mellon and State Street), regional banks (like US Bancorp and PNC Financial), specialized banks (like American Express), and the U.S. subsidiaries of large foreign banks (such as TD Group and HSBC North America). 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 the eight U.S. Global Systemically Important Banks (G-SIBs) is illustrated in Figure 2.1. The left panel shows variation in bank size, with JPMorgan Chase (JPM) more than ten times larger than State Street (STT), as measured by total assets. The left panel also shows significant variation in the proportion of assets made up by loans. The right panel shows heterogeneity in the fractions of loans the banks hold in each of 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.

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, or 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

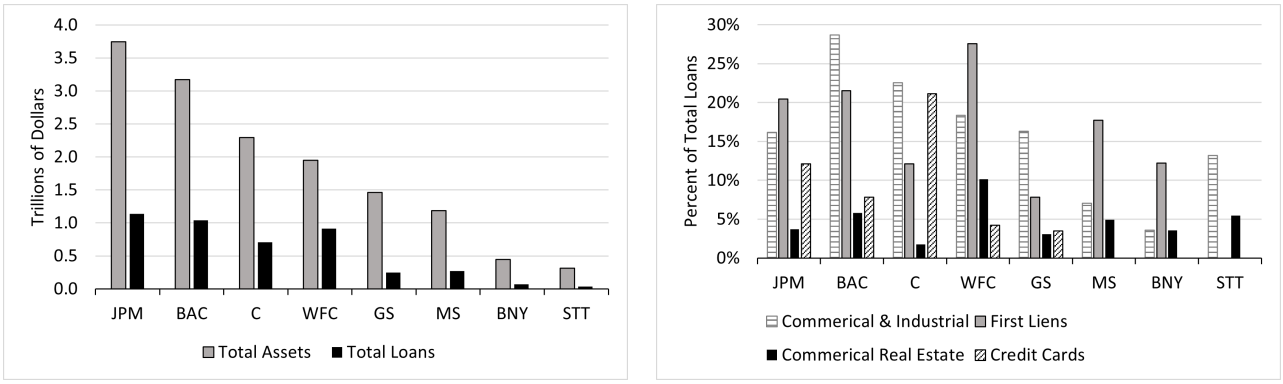

Figure 2.1: Heterogeneity among the U.S. G-SIBs. Left: Total assets and total loans for each bank. Right: The percentage of loans in each of four categories for each bank. Source: Bank Y-9C reports for Q4 2021.

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 took issue with the Fed's modeling of its trading revenues, pointing to specific features of the firm's compensation of traders. Citizens Financial claimed that the Fed's models fail to capture 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 Sources of Heterogeneity

Stepping back from the specifics of this debate, it is helpful to consider sources of heterogeneity through the lens of a linear regression of a loan portfolio's loss rate on portfolio characteristics and macro variables. In this setting, we have at least four sources of heterogeneity: the distribution of portfolio characteristics, differences in intercepts, differences in slopes, and differences in error distributions. Differences in errors mainly affect the precision of forecasts rather than the forecasts themselves, so we focus on the first three items.

[1] https://www.federalreserve.gov/supervisionreg/files/goldman-sachs-group-inc-20200904.pdf

Differences in portfolio characteristics are evident. Banks differ in their mix of corporate and retail business, their mix of regional, national, and international clients, and their focus on higher- or lower-risk borrowers. (In its 2020 annual report, Capital One reported that 31% of its credit card receivables were due from customers with FICO scores of 660 or lower. For Citigroup the proportion was less than 15% on its Citi-branded cards.)

Potential differences in slopes admit at least two possible interpretations. One possibility is that banks differ in unobserved portfolio characteristics that influence loss rates. If these unobserved characteristics are correlated with observed features, then omitting them changes the coefficients on the observed features. Unobserved portfolio characteristics could thus create differences among otherwise identical banks. A second possibility is that banks differ in their skill in managing loans, perhaps through more effective monitoring of borrowers. Differences in skill are then reflected in different coefficients linking portfolio features to loss rates. The first interpretation suggests that one bank's B-rated borrowers may be better than another bank's; the second interpretation suggests that one bank is simply better at securing repayment from B-rated borrowers.

Differences in intercepts — bank fixed effects — are commonly included in empirical work on banking. They are typically included to absorb unobserved characteristics that persist across time. As in our discussion of slopes, it is possible that some banks are better managed than others and that these differences generate predictable differences in loss rates. Regulators themselves seem to support this idea by emphasizing the importance of establishing a strong "culture" in banks.[2] However, we take the view that such differences should not confound loss forecasts in the design of an industry model for quantitative stress-testing. Equal treatment requires that stress-results be based on observable characteristics. Other parts of the overall bank supervision process (outside of stress testing) address issues like the quality of internal governance, business models, and controls. Within the Basel framework, these considerations are part of the Pillar 2 supervisory process, as described in BCBS [8].

In running the stress tests, the Fed collects highly granular data from each bank, and it has a great deal of power to collect the data it needs. It also designs the models that use the banks' data. With these considerations in mind, and recalling the statements on Fed policy quoted earlier, our analysis proceeds on the following premises: (i) The Fed considers the portfolio features used in its models to be legitimate information for forecasting losses; (ii) no obviously relevant portfolio features are excluded from the models; (iii) bank identity is not a legitimate feature for forecasting losses; (iv) the Fed seeks the "best" forecasts available based on the

[2]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.

direct impact of legitimate features; (v) legitimate features should not be used indirectly to identify banks; and (vi) the overall average level of losses projected by the Fed's industry model should be consistent with the average that would be obtained using bank-specific models.

The rest of this paper will make precise and further develop these ideas. We will investigate ways to aggregate bank-specific models into a single industry model, premised on these ideas and recognizing the heterogeneity across banks. Doing so entails balancing the goal of overall accuracy with concerns for consistent treatment of banks.

3 Pooling: Fairness Through Unawareness?

3.1 Basic Model

To capture bank heterogeneity, we consider a market with multiple banks, indexed by $s = 1, \ldots, \bar{S}$. 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 vector of predictive variables; 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 7. 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. We assume throughout that 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 loss rate $g(x, s)$ for bank s if the bank's portfolio characteristic vector is x. The forecasts should satisfy the following property, which prohibits the regulator from applying different models to different banks:

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

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

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

with the probabilities p_s summing to 1. 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 $\sum_s p_s \mathsf{E}[(\hat{Y}(X_s) - Y_s)^2]$.

We emphasize that 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 would have a panel of time series of observations of (1) 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. It would not be meaningful to refer to an estimator as unbiased or consistent until we have clearly identified the estimation target.

For the solution to (7), write

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

and

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

with

$$W_s = \Sigma_s + \mu_s \mu_s^\top - \bar{\mu}\mu_s^\top \in \mathbb{R}^{d\times d}. \tag{10}$$

Similarly,

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

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

$$\beta_{Pooled} = \mathsf{E}[W_S]^{-1} \left(\mathsf{cov}[\alpha_S, \mu_S] + \mathsf{E}[W_S \beta_S]\right) \tag{11}$$

and

$$\alpha_{Pooled} = \mathsf{E}[Y_S] - \beta_{Pooled}^\top \bar{\mu}. \tag{12}$$

Loss forecasts using α_{Pooled} and β_{Pooled} in (4) provide *fairness through unawareness*, in that they ignore bank identity. They satisfy the equal treatment principle articulated by the Federal Reserve that "two firms with the same portfolio receive the same results for that portfolio." 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_{Pooled}. 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_{Pooled} 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 above, 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_{Pooled} in (11) becomes

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

If $\mathsf{cov}[\alpha_S, \mu_S] = 0$ and if $\sigma_s^2 + \mu_s^2 - \bar{\mu}\mu_s \geq 0$, for all s, then (13) becomes a convex combination of the individual β_s. In Appendix B, 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_{Pooled}; we return to these cross-bank effects in Section 6.

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 the expected coefficient,

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

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

$$\beta_{Pooled} = \frac{\sum_s p_s \sigma_s^2 \beta_s}{\sum_s p_s \sigma_s^2}. \tag{15}$$

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

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

ensures that the forecasts

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

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 this section, we turn to methods that expand the squared loss minimization objective (7) to include fairness considerations.

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. [6]. 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. [12] and Le Gouic et al. [35] 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. [35], 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 [2]. 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. For bank s, the regulator first forms the forecast $\hat{Y}_s(x) = \alpha_s + \beta_s^\top x$, using the bank-specific coefficients and the realized features $X_s = x$. Suppose $\hat{Y}_s$ falls at the 80th percentile of the forecast distribution for bank s. The regulator then takes a weighted average of the 80th percentile forecast for all of the bank-specific models. That weighted average becomes the forecast for bank s.

To make this procedure more explicit and to specialize the general framework of Chzhen et al. [12] and Le Gouic et al. [35] 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 $s = 1, \ldots, \bar{S}$, 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

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

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

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

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 $s = 1, \ldots, \bar{S}$. 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. But 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. 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, not relative to all banks. 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 loss projections to be independent of bank identity ignores relevant differences in bank's 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, Foster, and Stine [30] introduce the concept of formal equality of opportunity (FEO) in regression, based on the use of the term in political philosophy, for which they cite the review in Arneson [5]. According to Arneson [5], 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 our definition of equal treatment and 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. [30], 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 protected attributes if legitimate features (portfolio characteristics) are correlated with bank identity.

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

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

We discuss the implications of centering below. For any coefficients $\alpha, \delta_1, \ldots \delta_{\bar{S}-1} \in \mathbb{R}$ 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, $\{\delta_i, i = 1, \ldots, \bar{S} - 1\}$, 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, the $U_i(S)$, and X_S, evaluated at $S = s$ and $X_S = x$. Now drop the 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. [30]. (More precisely, steps (22)–(24) define a population counterpart of the sample formulation in [30].) 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 dummy 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

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

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. [30] to our setting. Write $U = (U_1, \ldots, U_{\bar{S}-1})^\top$ for the vector of centered dummy variables. Write $\mathsf{cov}[X_S, U(S)]$ for the $d \times (\bar{S}-1)$ 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 $\mathsf{E}[\|U(S) - \Lambda^\top(X_S - \bar{\mu})\|^2]$, so $\Lambda^\top(X_S - \bar{\mu})$ is the linear projection of the bank-identity variables $U(S)$ onto the centered portfolio features $X_S - \bar{\mu}$. The relationship between β_{Pooled} and β_F can be expressed as follows.

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

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

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

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

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

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

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

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

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

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

The term $\delta^\top[U(s) - \Lambda^\top(x - \bar{\mu})]$ in (33) changes the forecast based on information in bank identity that is orthogonal to the features x. In the terminology of Johnson et al. [30], this would be *disparate treatment*. Through "unawareness" (meaning that it has no functional dependence

on bank identity) the pooled forecast (34) drops this term, but it retains $\delta^\top \Lambda^\top (x - \bar{\mu})$, as can be seen from (30).

The term $\delta^\top \Lambda^\top (x - \bar{\mu})$ is the problematic component of the pooled method. Although it does not explicitly use bank identity (it satisfies equal treatment), 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 $\Lambda^\top (x - \bar{\mu})$ is the least-squares prediction of $U(s)$ from $x - \bar{\mu}$. In the pooled forecast (34), the regulator is implicitly "misusing" the data in the features $x - \bar{\mu}$ 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.

We will develop this idea further in Section 4.5 and conclude 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 satisfies the equal-treatment property and thus 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 "misuse" 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 $\mathsf{E}[\epsilon_s | X_s] = 0$, $s = 1, \ldots, \bar{S}$. Then, by Bayes' rule,

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

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 7, the conditional expectation $\mathsf{E}[Y_S | X_S = x]$ 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 [5], a system in which admission decisions are made through a competitive exam open to everyone achieves formal equality of opportunity; but if only the wealthy have access to the preparation required for the exam, the system fails to achieve *substantive* equality of opportunity (SEO). In the regression setting, Johnson et al. [30] 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 Arneson's [5] example, 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.

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

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

then M minimizes $\mathsf{E}[\|X_S - \bar{\mu} - M^\top U(S)\|^2]$. In accordance with Definition 2 of Johnson et al. [30], define

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

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)*

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

so the SEO forecast (38) is given by

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

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 equal treatment. 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 non-wealthy exam takers, (40) adjusts scores downward for wealthy exam takers and upward for non-wealthy 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."[3] 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 A Unified Perspective: Legitimate Information

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

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

subject to additional considerations. Table 4.1 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 the condition (3) on the errors ϵ_s.

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

	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 ([12, 35])
(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)

Table 4.1: Summary of forecast model forms and constraints.

Proposition 4.5. *In each row of Table 4.1, 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 4.1(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 4.1(iv), it enforces a requirement we call *no misdirection of legitimate information*. FEO uses all of X_s in projecting losses; but 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

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

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 $\Lambda^\top (X_S - \bar{\mu})$ is the linear projection of the centered bank identity variables $U(S)$ onto the centered portfolio features $X_S - \bar{\mu}$. 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 illustrate these ideas, consider a simple example in which some component of X_s measures exposure to beachfront real estate. 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 would not directly inform the forecasts. 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 Average Treatment Effect as an Extension of FEO

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). Although we do not recommend this choice, for reasons we return to at the end of this section, this analysis clarifies how the ATE forecast fits within a more general framework.

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 $1, U_1(S), \ldots, U_{\bar{S}-1}(S)$, $X_S - \mu_S$, $V_S - \nu_S$, $U_1(S)V_S, \ldots, U_{\bar{S}}(S)V_S$. Let β_F denote the coefficient of $X_S - \mu_F$ 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 $U_{\bar{S}}(S)V_S$ in the first step (even though we omitted $U_{\bar{S}}(S)$) 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 a constant, because $U_1(S)V_S + \cdots + U_{\bar{S}}(S)V_S = 0$. These coefficients are dropped in the second step, so their value is immaterial.

Proposition 4.7. *Suppose* $\mathsf{var}[X_s]$ *and* $\mathsf{var}[V_s]$ *have full rank and* X_s *and* V_s *are uncorrelated, for each* $s = 1, \ldots, \bar{S}$. *Then* β_F *is given by (25) and (28), and* $\gamma_F = \bar{\gamma} = \sum_s p_s \gamma_s$. *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 comparisons in Section 6.

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 bank fixed-effects. Extending this idea to include interactions is tantamount to saying that we do not want heterogeneity with respect to one portfolio feature to affect the industry coefficient for another portfolio feature. But the logic for FEO does not extend to this case: here we are dealing with two presumably legitimate features whereas the point of the FEO forecast is to remove the influence of the bank-identity variables, which are not legitimate features under equal treatment.

5 Empirical Evidence

In this section, we document empirical evidence of heterogeneity in bank-specific models of loss rates, and we examine the implications of this heterogeneity for the choice of an industry-wide model. We find strong evidence of statistically significant differences in model parameters across banks. These differences can lead to material differences between pooled and FEO coefficients in an industry model.

5.1 Data

We use two types of data and data sources: loan information for individual banks and historical macroeconomic data.

5.1.1 Loan Information for Individual Banks.

Bank holding companies publicly report financial information quarterly through the Federal Reserve's form Y-9C. We use these filings to collect information on four loan types that are treated separately in the Fed's stress tests: credit cards, first lien mortgages, commercial real

estate loans, and commercial and industrial loans. For each category, each bank, and each quarter, we collect loan balances, charge-offs, recoveries, and total amounts past due. For each bank-quarter we also collect the bank's allowances for losses; allowances are not consistently reported separately by loan category, so we use a bank's total allowances across all loan types. Allowances and amounts past due are our proxy measures of loan portfolio risk.

We collect this data from 2002 to 2021 for the twenty largest banks by total assets (as of December 2020). The banks are listed in Table D.1. In each loan category, we include only banks with at least three years (12 quarters) of data. For each quarter and loan category, we use the included banks' loan balances to determine their relative weights p_s.

We merger-adjust all bank data. For example, Truist Financial, one of the banks in Table D.1, was formed from the 2019 merger of BB&T and SunTrust, so our data for Truist in earlier years combines data from those two banks. We repeat this process as we work backwards in time. We obtain information on mergers and acquisitions from the Federal Financial Institutions Examination Council website. (We have also run our analysis without merger-adjusting the data; doing so does not change our conclusions and generally increases heterogeneity across banks.)

In each loan category, we calculate a loss rate (net charge-off rate) for each bank s and each quarter t as the ratio

$$\mathit{LossRate}_{s,t} = \frac{\mathit{Charge\text{-}offs}_{s,t} - \mathit{Recoveries}_{s,t}}{\mathit{Total\ Loans\ in\ Category}_{s,t-1}}.$$

This measure is commonly used in stress testing; see, for example, Guerrieri and Welch [26], Hirtle et al. [29], and Kapinos and Mitnik [32]. We similarly normalize the amounts past due and allowances to get a $\mathit{PastDueRate}_{s,t}$ and an $\mathit{AllowanceRate}_{s,t}$ for each bank-quarter, except that the allowance rate is normalized by the total loans in all categories. We remove values less than -100% or greater than 100% of *LossRate*, *PastDueRate* and *AllowanceRate*, and we winsorize *PastDueRate* at the upper and lower 5% levels.

Table 5.1 shows descriptive statistics for these variables. Loss rates and past due rates are shown by loan category — credit cards (CC), first liens (FL), commercial real estate (CRE), and commercial and industrial (CI). Columns 2–4 of the table summarize time-averaged values across banks. Columns 5–8 summarize observations across all banks and quarters.

	bank averages			all observations			
	min	mean	max	lower 5%	mean	upper 5%	std
Loss Rate: CC	1.51	2.82	4.21	0.76	2.87	6.72	2.06
Loss Rate: FL	0.02	0.32	1.34	-0.02	0.34	1.62	0.66
Loss Rate: CRE	-0.00	0.23	0.94	-0.04	0.16	1.02	0.42
Loss Rate: CI	0.07	0.46	1.32	0.02	0.41	1.69	0.53
Past due Rate: CC	1.80	3.18	4.30	1.43	3.28	6.32	1.50
Past due Rate: FL	1.82	5.75	10.33	0.94	7.93	19.79	6.21
Past due Rate: CRE	1.39	2.22	3.58	0.48	2.11	6.99	1.88
Past due Rate: CI	0.84	1.78	3.35	0.49	1.67	4.50	1.17
Allowance Rate: Total	1.36	2.08	3.25	0.98	2.13	4.71	1.13

Table 5.1: Descriptive statistics in percent. Columns 2-4 are calculated from banks' time averages, and columns 5-8 are calculated from all observations, with mean and standard deviation weighted by loan balance.

Figure 5.1 plots the mean past due rate (± 1.96 standard errors) for each bank in each loan category. The banks are identified by their stock tickers (except for USAA, the United Services Automobile Association, which is not publicly traded). The figure illustrates substantial heterogeneity across banks in their loan portfolios. For example, in the credit card category Capital One (COF) has among the highest past due rates, but in the commercial real estate category it has among the lowest. This type of pattern is consistent with the idea that banks have different areas of specialization and may target different markets.

The widths of the bars in Figure 5.1 show differences across loan categories and banks in the volatility of their past due rates. The volatility for commercial real estate is particularly high, due primarily to a spike in delinquencies during the global financial crisis.

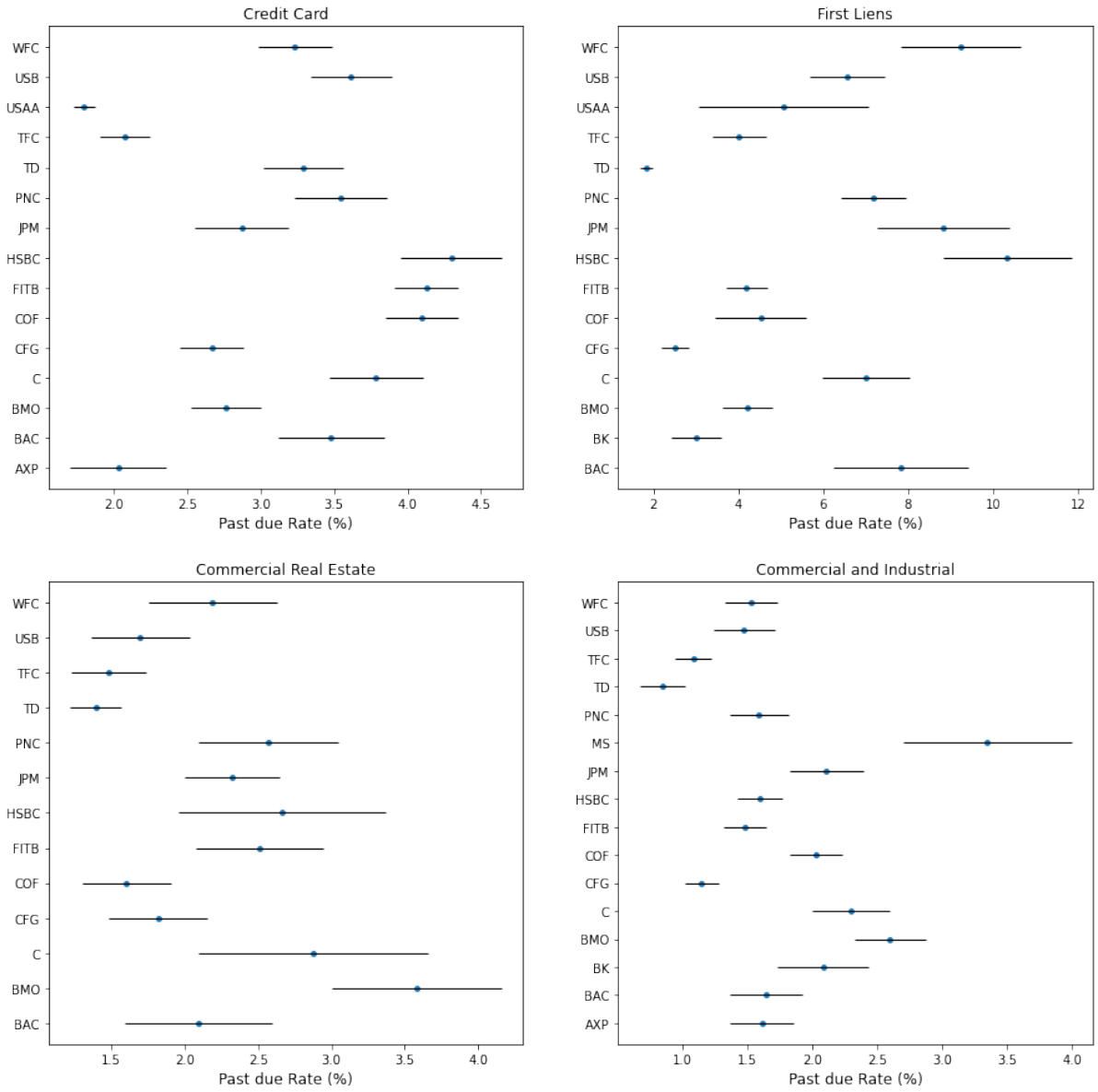

Figure 5.1: Past due rates (winsorized) by bank and loan category. The dots show mean values and each horizontal bar corresponds to ± 1.96 standard errors.

5.1.2 Macroeconomic Data

We use data on seven of the macro variables used in the Federal Reserve's stress tests: real disposable income growth, real GDP growth, house price index level, inflation rate, unemployment rate, Dow Jones total stock index level, and the Treasury spread. The Federal Reserve provides historical data on its website for all variables used in forming stress scenarios, including these. We use the values reported by the Fed for these variables in the June 2020 stress test; these values run from 1990 through 2019.

We aggregate these variables into a single macro variable by taking the first principal component of their correlation matrix. Table 5.2 shows the corresponding loadings. We see that an increase in the principal component corresponds to decreases in income growth and GDP growth and an increase in unemployment, suggesting that this composite variable serves as a reasonable measure of overall economic conditions. Figure 5.2 plots the level of this variable over time and shows a sharp climb around 2008 and 2020.[4]

Our main results use data through 2021. As a robustness check, we also run our analysis using data through 2019. This truncation serves two purposes. It ensures that our conclusions are not driven by a few extreme values during the COVID period 2020–2021, and it accounts for a change in how how banks measure allowances (the Current Expected Credit Losses methodology) that took effect at the end of 2019. See Appendix E.

Macro Factor	PC1 Loading
Real disposable income growth	-0.229
Real GDP growth	-0.525
Change House Price Index	-0.467
CPI inflation rate	-0.079
Change unemployment	0.529
Change Dow	-0.293
Change Treasury Spread	0.287

Table 5.2: Loadings of first principal component on macro variables.

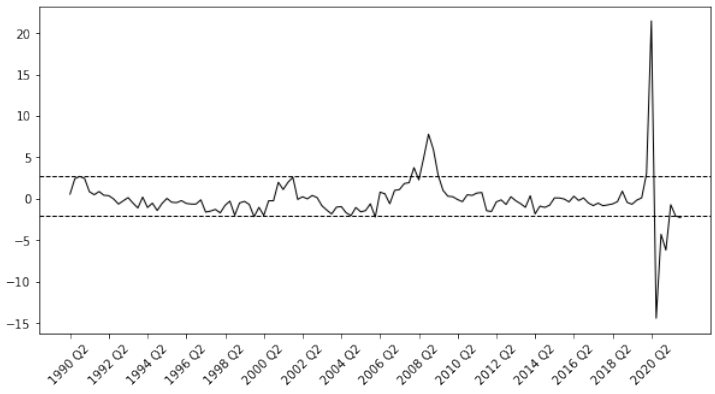

Figure 5.2: First principal component (PC1) of macro variables from 1990 Q2 to 2021 Q4. The dashed lines correspond to the 5th and 95th percentiles of PC1.

[4]The loadings in Table 5.2 are calculated using data through 2019, and we use these loadings to extend PC1 through the end of 2021. When we include the COVID period in the calculation of the principal components, PC1 becomes harder to interpret. For example, the coefficients for income growth and unemployment have the same sign.

5.2 Heterogeneity in Slopes and Intercepts

We use the bank data to approximate our theoretical framework through the specification

$$LossRate_{s,t} = \alpha_s + \beta_s^{PDR} PastDueRate_{s,t-l} + \beta_s^{AR} AllowanceRate_{s,t-l} + \gamma_s MacroPC_{t-l} + \epsilon_{s,t}, \quad (44)$$

for bank s in quarter t, where $MacroPC$ is the principal component of the macro variables introduced in Section 5.1.2. The lag l is either one quarter or one year, to mimic the stress-testing framework. We estimate separate coefficients for each of the four loan categories, for each bank. Because these are bank-specific regressions, it would not be meaningful to include bank-specific controls.

For each loan category, we want to test for heterogeneity in parameters across banks. When we test for heterogeneity, the null hypothesis states that slopes for all banks are equal,

$$H_0 : \beta_1 = \cdots = \beta_{\bar{S}}, \quad (45)$$

or that the intercepts are equal,

$$H_0 : \alpha_1 = \cdots = \alpha_{\bar{S}}. \quad (46)$$

The alternative hypothesis in each case states that the indicated parameters are not identical across banks. We will run these tests with different subsets of the variables in (44) included and interpret the coefficients in (45) accordingly.

To test these hypotheses for a particular loan category, let X_s be the n_s by p data matrix for bank s, where n_s is the number of observations for bank s in the loan category, and p =1, 2, or 3, is the number of variables included on the right side of (44). Let $\tilde{X}_s = (1, X_s)$ be X_s concatenated with a column of 1s, and let X^* be the diagonal block matrix $X^* = \text{diag}(\tilde{X}_1, .., \tilde{X}_S)$. Let $\theta^* = (\alpha_1, \beta_1^\top, ..., \alpha_S, \beta_S^\top)^\top, \epsilon^* = (\epsilon_1, ..., \epsilon_S)$, where ϵ_s is a column vector of length n_s. Our unrestricted model can be written as

$$Y = X^* \theta^* + \epsilon^*, \quad (47)$$

and the restrictions in (45) and (46) impose linear constraints on the parameter θ^*.

We apply the Wald test to test linear constraints on θ^* in (47) under various assumptions on the error covariance matrix. (i) As a baseline, we allow bank heteroskedasticity: the error variance is constant through time but varies across banks; errors are assumed uncorrelated across time and across banks. (ii) We combine (i) with clustering by time, which allows correlation in errors across banks in each quarter, but no correlation across quarters. (iii) We combine (i) with single-lag Newey-West standard errors, which allows serial correlation in errors within each bank, but no correlation across banks.

Tables 5.3 and 5.4 report p-values for the tests when different subsets of variables are included in the right side of (44), for forecast horizons of one quarter and one year, respectively. Almost all the tests indicate strong evidence of heterogeneity in the intercepts and in the coefficients for past due rates. The evidence is more mixed for allowance rates and the macro variable. These variables are also less predictive of losses than the past due rates.

Next we examine the impact of heterogeneity. Table 5.5 compares pooled and FEO coefficients for *PastDueRate*, *AllowanceRate*, and *MacroPC* using a one-quarter lag and four specifications of which variables are included in (44). Table 5.6 shows corresponding results with a one-year lag. We estimate β_{Pooled} in a pooled panel regression, and β_F in a panel regression with bank fixed effects included. Both regressions are weighted by asset balances. In each table, the columns labeled "% diff" show the percentage difference $100\% \times (\beta_F - \beta_{Pooled})/\beta_{Pooled}$. This is a measure of the impact of addressing heterogeneity in choosing an industry model. We also report confidence intervals for these ratios. These are bootstrap confidence intervals, estimated using the cross-sectional method in Kapetanios [31], based on resampling independently from the panel of banks.

The percentage differences between β_{Pooled} and β_F vary substantially, and the associated confidence intervals are quite wide. But we see many cases in which the percentage difference in estimated coefficients is roughly 5-10% or even larger. In the context of setting capital requirements for banks, these differences could be material.

Consider, for example, the case of first lien mortgages, with coefficients $\beta_{Pooled} = 0.053$ and $\beta_F = 0.060$ in the top panel of Table 5.5. From Table 5.1, we see that the average bank has a past due rate of 5.75% on FL loans. The difference $(0.060 - 0.053) \times 5.75\% = 0.04\%$ is 13% of the average FL loss rate of 0.32% in Table 5.1. The additional capital required to offset the higher predicted loss rate would be 13% of the capital required to offset the average loss rate.

Covariance	α				β_{PDR}				β_{AR}				γ			
Estimation	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank heteroskedasticity	0.01	0.00	0.00	0.21	0.00	0.00	0.00	0.00								
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
bank heteroskedasticity	0.00	0.00	0.00	0.11	0.02	0.00	0.00	0.00	0.05	0.00	0.00	0.00				
time clustered	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.01	0.00				
Newey-West HAC	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.00				
bank heteroskedasticity	0.01	0.00	0.01	0.22	0.00	0.00	0.00	0.00					1.00	0.94	0.00	0.75
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					0.96	0.52	0.00	0.00
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					0.93	0.38	0.00	0.00
bank heteroskedasticity	0.00	0.00	0.00	0.12	0.03	0.00	0.00	0.00	0.06	0.00	0.00	0.00	1.00	0.99	0.01	0.90
time clustered	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.06	0.00	0.87	0.80	0.00	0.00
Newey-West HAC	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.04	0.00	0.83	0.87	0.00	0.00

Table 5.3: *P*-values for heterogeneity tests when forecast horizon is one quarter. Every three rows correspond to tests with different assumptions on the error covariance matrix for four loan portfolios and different sets of variables included in the regression: (i) *PastDueRate* only; (ii) *PastDueRate* and *AllowanceRate*; (iii) *PastDueRate* and *MacroPC*; (iv) *PastDueRate*, *AllowanceRate*, and *MacroPC*.

Covariance Estimation	α				β_{PDR}				β_{AR}				γ			
	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank heteroskedasticity	0.00	0.00	0.46	0.23	0.00	0.00	0.00	0.00								
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
bank heteroskedasticity	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00				
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
bank heteroskedasticity	0.00	0.00	0.35	0.30	0.00	0.00	0.00	0.00					0.98	0.00	0.00	0.02
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					0.32	0.00	0.39	0.60
Newey-West HAC	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00					0.07	0.00	0.53	0.24
bank heteroskedasticity	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.97	0.00	0.01	0.04
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.76	0.00	0.75	0.56
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.83	0.54

Table 5.4: P-values for heterogeneity tests when forecast horizon is one year. Every three rows correspond to tests with different assumptions on the error covariance matrix for four loan portfolios and different sets of variables included in the regression: (i) *PastDueRate* only; (ii) *PastDueRate* and *AllowanceRate*; (iii) *PastDueRate* and *MacroPC*; (iv) *PastDueRate*, *AllowanceRate*, and *MacroPC*.

Loan Type	*Past Due Rate*				*Allowance Rate*				*Macro PC*			
	β_{Pooled}	β_F	% diff	90% CI	β_{Pooled}	β_F	% diff	90% CI	γ_{Pooled}	γ_F	% diff	90% CI
CC	0.738	0.770	4.3	(-3.0, 14.3)								
FL	0.053	0.060	12.3	(-0.8, 26.9)								
CRE	0.127	0.127	-0.3	(-3.9, 5.1)								
CI	0.290	0.311	7.3	(-2.1, 10.1)								
CC	0.419	0.392	-6.5	(-28.0, 10.7)	0.620	0.712	14.9	(5.1, 38.2)				
FL	0.007	0.015	111.6	(-557.9, 314.7)	0.388	0.347	-10.7	(-25.1, 21.3)				
CRE	0.084	0.072	-14.8	(-31.1, 0.5)	0.126	0.155	22.7	(-10.0, 208.2)				
CI	0.234	0.250	7.2	(-0.6, 13.9)	0.120	0.146	21.0	(-13.9, 62.7)				
CC	0.724	0.768	6.1	(-4.9, 13.9)					0.366	0.050	-86.4	(-182.8, 181.9)
FL	0.051	0.058	13.0	(-0.3, 33.7)					0.443	0.409	-7.8	(-42.7, 3.4)
CRE	0.126	0.126	-0.3	(-4.5, 4.8)					0.062	0.068	8.9	(-106.9, 112.2)
CI	0.278	0.297	6.7	(-1.1, 9.5)					0.579	0.571	-1.3	(-6.2, 22.1)
CC	0.399	0.380	-4.6	(-30.8, 18.9)	0.621	0.713	14.8	(5.4, 38.1)	0.482	0.224	-53.5	(-136.6, 75.5)
FL	0.007	0.016	119.6	(-97.3, 693.1)	0.384	0.339	-11.7	(-26.1, 23.6)	0.144	0.152	5.4	(-132.6, 66.2)
CRE	0.084	0.071	-15.0	(-32.3, -1.8)	0.126	0.156	23.5	(-1.9, 204.4)	0.010	-0.009	-189.0	(-361.3, 310.6)
CI	0.230	0.246	6.8	(-1.2, 13.3)	0.106	0.129	22.0	(-21.6, 47.2)	0.487	0.453	-6.9	(-11.4, 20.2)

Table 5.5: Comparison of coefficients for one-quarter forecasts. We regress *LossRate* on (i) *PastDueRate*, (ii) *PastDueRate* and *AllowanceRate*, (iii) *PastDueRate* and *MacroPC*, and (iv) *PastDueRate*, *AllowanceRate*, and *MacroPC*. Percent difference is calculated as $100^*(\beta_F$ - $\beta_{Pooled})/\beta_{Pooled}$. 90% confidence intervals are constructed using 100 bootstrap samples.

Loan Type	*Past Due Rate*				*Allowance Rate*				*Macro PC*			
	β_{Pooled}	β_F	% diff	90% CI	β_{Pooled}	β_F	% diff	90% CI	γ_{Pooled}	γ_F	% diff	90% CI
CC	0.824	0.855	3.8	(-4.6, 10.5)								
FL	0.046	0.052	13.4	(-2.9, 29.4)								
CRE	0.104	0.100	-3.3	(-7.8, 1.3)								
CI	0.178	0.182	2.3	(-7.5, 6.1)								
CC	0.827	0.860	3.9	(-6.4, 13.7)	-0.006	-0.008	35.2	(-897.1, 156.3)				
FL	-0.003	0.005	-287.2	(-971.9, 309.1)	0.417	0.373	-10.4	(-22.2, 3.1)				
CRE	0.057	0.046	-19.7	(-60.4, 66.2)	0.138	0.155	12.3	(-51.9, 299.6)				
CI	0.137	0.142	4.0	(-12.4, 13.2)	0.089	0.098	10.0	(-23.0, 35.3)				
CC	0.658	0.662	0.5	(-6.8, 7.9)					4.034	3.873	-4.0	(-6.3, 0.9)
FL	0.041	0.047	15.4	(-3.2, 36.7)					1.297	1.267	-2.3	(-8.9, 2.9)
CRE	0.096	0.093	-3.3	(-8.6, 3.1)					0.665	0.679	2.0	(-3.3, 12.2)
CI	0.143	0.139	-2.9	(-11.9, 1.7)					1.758	1.782	1.4	(0.1, 4.6)
CC	0.656	0.658	0.4	(-10.4, 11.6)	0.005	0.007	33.8	(-259.0, 287.1)	4.035	3.875	-4.0	(-7.6, 0.5)
FL	-0.003	0.007	-309.4	(-594.4, 537.7)	0.387	0.321	-17.1	(-25.1, 18.7)	1.002	1.026	2.5	(-3.5, 7.9)
CRE	0.056	0.056	-0.7	(-46.4, 183.2)	0.119	0.106	-10.5	(-64.1, 194.5)	0.617	0.626	1.5	(-9.7, 17.7)
CI	0.127	0.126	-0.3	(-10.0, 7.7)	0.037	0.034	-9.9	(-139.5, 98.7)	1.726	1.752	1.5	(-0.0, 5.6)

Table 5.6: Comparison of coefficients for one-year forecasts. We regress *LossRate* on (i) *PastDueRate*, (ii) *PastDueRate* and *AllowanceRate*, (iii) *PastDueRate* and *MacroPC*, and (iv) *PastDueRate*, *AllowanceRate*, and *MacroPC*. Percent difference is calculated as 100*(β_F - β_{Pooled})/β_{Pooled}. 90% confidence intervals are constructed using 100 bootstrap samples.

6 Cross-Bank Parameter Externalities

As a consequence of aggregating bank-specific results into a single industry model, changes at one bank can affect loss forecasts at other banks, and the results are sometimes counterintuitive. In this section, we argue that these cross-bank externalities are generally more reasonable under FEO forecasts than under the pooled method.

For simplicity, we consider a setting with a single scalar feature x. More generally, we can think of this as a feature that is uncorrelated with all other features. We adopt the convention that this feature is nonnegative, and that higher values of x are associated with higher losses. Thus, for each bank s we assume $\mu_s \geq 0$ and $\beta_s \geq 0$. In reducing μ_s, a bank improves its portfolio quality; in reducing β_s, a bank improves its ability to manage portfolio risk; and in reducing α_s, a bank improves unobserved features to reduces its losses. We examine how these improvements — reductions in μ_s, α_s, and β_s — affect stress test results for bank s and other banks l.

We can write the FEO loss forecast (24) for bank l evaluated at $X_l = x$ as

$$\hat{Y}_{F,l}(x) = \hat{Y}_F(x) = \sum_s p_s(\alpha_s + \beta_s \mu_s) + \beta_F(x - \bar{\mu}), \tag{48}$$

with $\beta_F = \sum_i p_i \sigma_i^2 \beta_i / \sum_i p_i \sigma_i^2$, as in (27). The forecast is the same for all banks l because FEO satisfies equal treatment. It is now easy to see that

$$\frac{\partial \hat{Y}_F(x)}{\partial \mu_s} = p_s \beta_s - p_s \beta_F \geq 0, \quad \text{if and only if } \beta_s \geq \beta_F; \tag{49}$$

$$\frac{\partial \hat{Y}_F(x)}{\partial \alpha_s} = p_s \geq 0; \tag{50}$$

and

$$\frac{\partial \hat{Y}_F(x)}{\partial \beta_s} = p_s \mu_s + (x - \bar{\mu}) p_s \sigma_s^2 / \sum_i p_i \sigma_i^2 \geq 0, \quad \text{if } x > \bar{\mu}. \tag{51}$$

In (49) we see that if bank s has above-average (relative to β_F) sensitivity to feature x, then reducing its average exposure to that feature μ_s reduces loss forecasts for all banks. Equation (50) shows a similar overall benefit if bank s improves on the other dimensions captured by α_s. In (49) we see that an improvement in risk management at bank s, corresponding to a reduction in β_s, reduces loss forecasts at above-average levels of x. If x is part of the stress scenario, then large values of x are particularly relevant.

The directional effects in (49)–(51) are fairly simple and reasonable, considering that cross-bank effects are inevitable in an industry model. If the industry improves its performance (perhaps because of improvements at one bank) we generally expect loss forecasts to decrease.

(A decrease in a forecast corresponds to a positive derivative because we are considering a decrease μ_s, α_s, or β_s.) Counterparts to (49)–(51) continue to hold if we replace β_F in (48) with any convex combination of the β_s, as in the WATE model. However, the pooled method behaves quite differently.

The pooled forecast $\hat{Y}_P(x)$ can be written in the same form as (48) but with β_F replaced by β_{Pooled} in (11). We now get

$$\frac{\partial \hat{Y}_P(x)}{\partial \mu_s} = p_s(\beta_s - \beta_{Pooled}) + (x - \bar{\mu})\frac{\partial \beta_{Pooled}}{\partial \mu_s}.$$

The sign of the last term is not determined by a simple condition, so the overall directional effect is difficult to predict. The sign of

$$\frac{\partial \hat{Y}_P(x)}{\partial \alpha_s} = p_s + p_s \frac{(\mu_s - \bar{\mu})(x - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \mathsf{var}(\mu_S)}\beta_s,$$

depends on the magnitudes of μ_s and x, relative to $\bar{\mu}$. For the sensitivity to β_s, we can write

$$\frac{\partial \hat{Y}_P(x)}{\partial \beta_s} = p_s\mu_s + (x - \bar{\mu})\frac{\partial \beta_{Pooled}}{\partial \beta_s}, \quad \frac{\partial \beta_{Pooled}}{\partial \beta_s} = \frac{p_s(\sigma_s^2 + \mu_s(\mu_s - \bar{\mu}))}{\sum_i p_i \sigma_i^2 + \mathsf{var}(\mu_S)}.$$

Among the most troubling aspects of the pooled model is that the last term could be negative: a reduction in β_s could produce an increase in β_{Pooled}. In particular, $\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})$ is negative for a bank with below-average exposure to feature x (so $\mu_s < \bar{\mu}$) and low variability σ_s^2 in this exposure. Under the pooled model, it is therefore possible for an improvement in risk management at one bank (a reduction in β_s) to produce an *increase* in loss forecasts at all banks.

The top panel of Table 6.1 shows sufficient conditions for positive sensitivities of $\hat{Y}_F(x)$ and $\hat{Y}_P(x)$. The middle and bottom panels show corresponding results for the expected forecasts $\mathsf{E}[\hat{Y}_l] = \mathsf{E}[\hat{Y}(X_l)]$ and for the bias $\mathsf{E}[\hat{Y}(X_l) - Y_l]$. Supporting details for the second and third cases are provided in Appendix C. We have tried to provide simple sufficient conditions, and in most cases the conditions are not necessary. All of the conditions for FEO extend to WATE with β_F replaced by the weighted average coefficient.

Some counterintuitive and undesirable cases can arise at empirically plausible parameter values. For example, in equation (68) of the appendix we derive an expression for $\partial \mathsf{E}[\hat{Y}_P(X_l)]/\partial \alpha_s$. Using estimated parameters for the credit card data in Section 5, we find that this derivative is negative when l is Citigroup and s is American Express or JPMorgan Chase. In other words, an improvement at either of these two banks would result in a higher expected loss forecast at Citigroup under the pooled model.

The bias sensitivities in Table 6.1 are more complicated than the other cases because the bias involves the difference between the predicted and actual loss rates. A reduction in the

$\hat{Y}(x)$	FEO	Pooled
$\mu_s \downarrow$	$\downarrow$ iff $\beta_s > \beta_F$	no simple rule
$\alpha_s \downarrow$	$\downarrow$	$\downarrow$ if $(\mu_s - \bar{\mu})(x - \bar{\mu}) > 0$
$\beta_s \downarrow$	$\downarrow$ if $x > \bar{\mu}$	$\downarrow$ if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](x - \bar{\mu}) > 0$
$\mathsf{E}[\hat{Y}(X_l)]$		
$\mu_s \downarrow$	$l = s$: $\downarrow$	no simple rule
	$l \neq s$: $\downarrow$ iff $\beta_s > \beta_F$	
$\alpha_s \downarrow$	$\downarrow$	$l = s$: $\downarrow$
		$l \neq s$: $\downarrow$ if $(\mu_s - \bar{\mu})(\mu_l - \bar{\mu}) > 0$
$\beta_s \downarrow$	$\downarrow$ if $\mu_s + \mu_l > \bar{\mu}$	$\downarrow$ if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$
		or if μ_s sufficiently large
$\mathsf{bias}(l)$		
$\mu_s \downarrow$	$l = s$: $\downarrow$ iff $\beta_s < \beta_F$	no simple rule
	$l \neq s$: $\downarrow$ iff $\beta_s > \beta_F$	
$\alpha_s \downarrow$	$l = s$: $\uparrow$	$l = s$: no simple rule
	$l \neq s$: $\downarrow$	$l \neq s$: $\downarrow$ if $(\mu_s - \bar{\mu})(\mu_l - \bar{\mu}) > 0$
$\beta_s \downarrow$	$l = s$: $\uparrow$ if $\mu_s < \bar{\mu}$	no simple rule
	$l \neq s$: $\downarrow$ if $\mu_s + \mu_l > \bar{\mu}$	$\downarrow$ if $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$

Table 6.1: Sensitivity of results for bank l in response to a decrease in parameter μ_s, α_s, or β_s for bank s. Sensitivities shown are for predicted loss $\hat{Y}_l(x)$ (top), mean predicted loss $\mathsf{E}[\hat{Y}(X_l)]$ (middle), and the bias $\mathsf{E}[\hat{Y}(X_l) - Y_l]$.

predicted loss rate can increase or decrease bias, depending on whether the initial forecast is too low or too high.

7 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, rather than work with the loss rate Y_s, we could specify a linear model for its logit transformation $\log(Y_s/(1-Y_s))$.

But 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

$$Y_S = g(S, X_S) + \epsilon_S, \quad \mathsf{E}[\epsilon_S | S, X_S] = 0. \tag{52}$$

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

$$f_{Pooled}(x) \equiv \mathsf{E}[Y_S | X_S = x] = \mathsf{E}[g(S, X_S) | X_S = x].$$

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 (52) admits a decomposition of the form

$$Y_S = f_0 + f_1(S) + f_2(X_S) + \epsilon, \quad \mathsf{E}[\epsilon|S] = \mathsf{E}[\epsilon|X_S] = 0, \tag{53}$$

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

$$\mathsf{E}[Y_S - f_0 - f_1(S)|X_S] = f_2(X_S) \tag{54}$$

$$\mathsf{E}[Y_S - f_0 - f_2(X_S)|S] = f_1(S), \tag{55}$$

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

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

Equations (54)–(55) are population versions of the backfitting algorithm in Hastie and Tibshirani [27], which is a special case of the alternating conditional expectations algorithm of Breiman and Friedman [10]. Given an initial choice of f_1 (and known f_0), (54) defines an initial choice of f_2 through the regression of the residual $Y_S - f_0 - f_1(S)$ on X_S. Equation (55) then defines an upated choice of f_1. The algorithm iterates over (54) and (55). In writing (53), 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 [4].

We now introduce

$$\hat{Y}_F(x) = f_0 + f_2(x) \tag{56}$$

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 Section 4.2 and 4.5. To state the result, consider forecasts of the form

$$\hat{Y}_\gamma(x) = f_0 + f_2(x) + \gamma(x), \tag{57}$$

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

Proposition 7.1. *The nonlinear FEO forecast (56) satisfies*

$$\mathsf{cov}[\hat{Y}_F(X_S) - Y_S, X_S - \mathsf{E}[X_S|S]] = 0. \tag{58}$$

For $\hat{Y}_\gamma$ as in (57), 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 it misdirects legitimate information, in the sense that

$$\mathsf{cov}[\gamma(X_S), \mathsf{E}[f_1(S)|X_S]] > 0 \tag{59}$$

and

$$\mathsf{cov}[\mathsf{E}[\gamma(X_S)|S], f_1(S)] > 0. \tag{60}$$

Property (58) parallels the condition in row (iv) of Table 4.1 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 $X_S - \mathsf{E}[X_S|S]$, which is the component of X_S orthogonal to bank identity S. Properties (59)–(60) parallel conditions (i) and (ii) in Proposition 4.6. In particular, in (59), $\mathsf{E}[f_1(S)|X_S]$ 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.

Proposition 7.1 shows that applying ideas from previous sections to nonlinear models is a computational rather than a conceptual matter. We leave a fuller investigation into the application of this result for future work.

We briefly contrast our FEO forecast in (56) 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 Grűnewälder and Khaleghi [24] and Madras et al. [36]. 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. But we argued previously that demographic parity is too strong a condition for our setting. Our FEO forecast in (56) 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.

8 Concluding Remarks

The current practice of regulatory stress testing ignores bank heterogeneity as a matter of policy and principle. We have 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 have 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 have 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.

References

[1] Agarwal, S., An, X., Cordell, L., and Roman, R.A. (202) Bank stress test results and their impact on consumer credit markets. Working paper 20-30, Federal Reserve Bank of Philadelphia.

[2] Agueh, M., and Carlier, G. (2011) Barycenters in the Wasserstein space, *SIAM Journal on Mathematical Analysis* 43(2), 904–924

[3] Angrist, J.D. and Pischke, J.S. (2008) *Mostly Harmless Econometrics*, Princeton University Press, Princeton, New Jersey.

[4] Ansley, C.F., and Kohn, R. (1994) Convergence of the backfitting algorithm for additive models, *Journal of the Australian Mathematical Society (Series A)* 57, 316–329.

[5] Arneson, R. (2015) Equality of opportunity, *Stanford Encyclopedia of Philosophy* (Summer 2015 edition), Edward N. Zalta, ed.

[6] Barocas, S., Hardt, M., and Narayanan, A. (2017) *Fairness in Machine Learning*, https://fairmlbook.org/

[7] Bassett, W.F., and Berrospide, J.M. (2018) The impact of post stress test capital on bank lending. Working paper 2018-097, Federal Reserve Board, Washington, D.C.

[8] BCBS (2019) Overview of Pillar 2 supervisory review practices and approaches, Bank for International Settlements, Basel, Switzerland.

[9] Board of Governors (2021) Dodd-Frank Act Stress Test 2021: Supervisory Stress Test Methodology. Federal Reserve System, Washington, D.C.

[10] Breiman, L., and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 580–598.

[11] Breuer, T., Jandacka, M., Rheinberger, K., and Summer, M. (2009) How to find plausible, severe, and useful stress scenarios. *International Journal of Central Banking* 5, 205–224.

[12] Chzhen, E., Denis, C., Hebiri, M., Oneto, L, and Pontil, M. (2020) Fair regression with Wasserstein barycenters, *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

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

[14] Covas, F.B., Rump, B., and Zakrajsek, E. (2014) Stress-testing US bank holding companies: A dynamic panel quantile regression approach. *International Journal of Forecasting* 30, 691–713.

[15] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012) Fairness through awareness, pp.214–226, in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

[16] Fernandes, M., Igan, D., and Pinheiro, M. (2020) March madness in Wall Street: (What) does the market learn from stress tests? *Journal of Banking and Finance* 112, 105250.

[17] Flannery, M.J. (2019) Transparency and model evolution in stress testing, Available at SSRN: https://ssrn.com/abstract=3431679.

[18] Flannery, M., Hirtle, B., and Kovner, A. (2017) Evaluating the information in the Federal Reserve stress tests. *Journal of Financial Intermediation* 29, 1–18.

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

[20] Flood, M.D. and Korenko, G.G. (2015) Systematic scenario selection: stress testing and the nature of uncertainty. *Quantitative Finance* 15, 43–59.

[21] Georgescu, O.-M., Gross, M., Kapp, D., and Kok, C. (2017) Do stress tests matter? Evidence from the 2014 and 2016 stress test. Working paper 2054, European Central Bank, Frankfurt, Germany.

[22] Glasserman, P., Kang, C., and Kang, W. (2015) Stress scenario selection by empirical likelihood. *Quantitative Finance* 15, 25–41.

[23] Glasserman, P., and Tangirala, G. (2016) Are the Federal Reserve's stress test results predictable? *Journal of Alternative Investments: Systemic Risk Special Edition* 18, 82–97.

[24] Grűnewälder, S., and Khaleghi, A. (2021) Oblivious data for fairness with kernels, *Journal of Machine Learning Research* 22, 1–36.

[25] Guerrieri, L., and Modugno, M. (2021) The information content of stress test announcements. Working paper 2012-012, Federal Reserve Board, Washington, D.C.

[26] Guerrieri, L., and Welch, M. (2012) Can macro variables used in stress test forecast the performance of banks? Working paper 2012-49, Federal Reserve Board, Washington, D.C.

[27] Hastie, T., and Tibshirani, R. (1986) Generalized additive models, *Statistical Science* 1(3), 297–318.

[28] Hutchinson, B., and Mitchell, M. (2019) 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58.

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

[30] Johnson, K., Foster, D., and Stine, R. (2020) Impartial predictive modeling: ensuring group fairness in arbitrary models, arXiv:1608.00528.

[31] Kapetanios, G. (2008), A bootstrap procedure for panel data sets with many cross-sectional units. *Econometrics Journal* 11, 277–395.

[32] Kapinos, P., and Mitnik, O.A. (2016) A top-down approach to stress-testing banks. *Journal of Financial Services Research* 49, 229–264.

[33] Kohn, D., and Liang, N. (2019) Understanding the effects of the U.S. stress tests. Brookings Institution, Washington, D.C.

[34] Kupiec, P. (2020) Policy uncertainty and bank stress testing. *Journal of Financial Stability* 51, 100761.

[35] Le Gouic, T., Loubes, J.-M., and Rigollet, P. (2020), Projection to fairness in statistical learning, arXiv:2005.11720

[36] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018) Learning adversarially fair and transferable representations, arXiv:1802.06309

[37] Morgan, D.P., Peristiani, S., and Savino, V. (2014) The information value of the stress test. *Journal of Money, Credit and Banking* 46, 1479–1500.

[38] Parlatore, C., and Philippon, T. (2022) Designing stress scenarios. Working paper w29901, National Bureau of Economic Research, Cambridge, Mass.

[39] Philippon, T., Pessarossi, P., and Camara, B. (2017) Backtesting european stress tests. Working paper w23083, National Bureau of Economic Research, Cambridge, Mass.

[40] Pritsker, M.G. (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.

[41] Sahin, C., de Haan, J., and Neretina, E. (2020) Banking stress test effects on returns and risks. *Journal of Banking and Finance* 117, 105843.

[42] Schuermann, T. (2020) Capital adequacy pre- and postcrisis and the role of stress testing. *Journal of Money, Credit and Banking* 52, 87–105.

[43] Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, Second Edition, MIT Press.

A Proofs

Proposition 3.1. Problem (7) is solved by the linear projection of Y_S onto the span of 1 and X_S. If $\mathsf{var}[X_S]$ is invertible, then the coefficients of the linear projection are given by (12) and

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

see, for example, Wooldridge [43], 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

$$\mathsf{cov}[X_S, \epsilon_S] = \mathsf{E}[\mathsf{cov}[X_S, \epsilon_S|S]] + \mathsf{cov}[\mathsf{E}[X_S|S], \mathsf{E}[\epsilon_S|S]] = \mathsf{E}[0] + \mathsf{cov}[\mu_S, 0] = 0.$$

It follows that

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

□

Proposition 4.1. By Proposition 3.4 of Chzhen et al. [12] or Theorem 6 of Le Gouic et al. [35], the expected squared error is minimized subject to demographic parity by the rule that assigns

to bank s with features x the loss forecast

$$\hat{Y}_{PTF}(x,s) = \sum_i p_i F_i^{-1}(F_s(\alpha_s + \beta_s^\top x)), \tag{61}$$

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 (61) 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). □

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 (as in Angrist and Pischke [3], 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

$$\mathsf{var}[X_S - \mu_S] = \mathsf{E}[\mathsf{var}[X_S - \mu_S|S]] + \mathsf{var}[\mathsf{E}[X_S - \mu_S|S]] = \mathsf{E}[\Sigma_S] + 0.$$

For the second factor, we similarly have

$$\mathsf{cov}[X_S - \mu_S, Y_S] = \mathsf{E}[\mathsf{cov}[X_S - \mu_S, Y_S|S]] = \mathsf{E}[\mathsf{cov}[X_S - \mu_S, \beta_S^\top X_S|S]] = \mathsf{E}[\Sigma_S \beta_S],$$

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). □

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

$$Y_S = \alpha_F + \sum_{i=1}^{\bar{S}-1} \delta_i U_i(S) + \beta_F^\top X_S + u, \tag{62}$$

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

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

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

Next, we evaluate δ. Using (62), we can derive δ as the vector of coefficients in a regression of $Y_S - \beta_F^\top X_S$ on $U(S)$. Thus,

$$\begin{aligned}
\delta &= \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), Y_S - \beta_F^\top X_S] \\
&= \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), Y_S] - \mathsf{var}[U(S)]^{-1}\mathsf{cov}[U(S), X_S]\beta_F.
\end{aligned} \tag{63}$$

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

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

direct multiplication then verifies that

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

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

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

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

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

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

$$(\mu_s - \bar{\mu})^\top + \sum_{i=1}^{\bar{S}-1} p_i(\mu_i - \bar{\mu})^\top/p_{\bar{S}} = (\mu_s - \mu_{\bar{S}})^\top. \tag{64}$$

Combining these terms in (63) yields (31). □

Proposition 4.4. We derived an expression for the rows of M in (64), 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. □

Corollary 4.2. From (25) we know that if $\Sigma_s \equiv \Sigma$ then $\beta_F = \mathsf{E}[\beta_S]$. From (20), we get $\bar{\beta}^o = \sum_i p_i\beta_i^o = \sum_i p_i\Sigma^{1/2}\beta_i = \Sigma^{1/2}\mathsf{E}[\beta_S] = \Sigma^{1/2}\beta_F$. Thus, $\beta_F^\top(x - \mu_s) = \bar{\beta}^{o\top}\Sigma^{-1/2}(x - \mu_s) = \bar{\beta}^{o\top}z_s$. 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. □

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 $\lambda = -(\mathsf{var}[U(S)])^{-1}\mathsf{cov}[U(S), X_S]\beta$; i.e., $\lambda = -M\beta$. Making this substitution in the form of $\hat{Y}_S$ in row (v), (41) becomes

$$\mathsf{E}[(Y_S - \alpha - \lambda^\top U(S) - \beta^\top X_S)^2] = \mathsf{E}[(Y_S - \alpha - \beta^\top[X_S - M^\top U(S)])^2]. \tag{65}$$

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 (65) 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 (65) is therefore β_F. Because $\mathsf{E}[U(S)] = 0$, the minimizing α in (65) 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

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

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. □

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

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

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

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

from which (i) follows.

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

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

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

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

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. □

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,

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

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

Starting from the representation of (43) as

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

we may write

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

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 $a_i = \alpha_i + \beta_i^\top \mu_i + \bar{\gamma}^\top \nu_i$ and $\bar{a} = \sum_i p_i a_i$. Then,

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

Thus, $\tilde{Y}$ is in the span of $\mathcal{O}$, and its coefficient on $V_S - \nu_S$ is $\bar{\gamma}$. 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 $\bar{\gamma}$ is uniquely determined. □

Proposition 7.1. For the first claim, we have

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

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

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

It now follows that if γ reduces the expected squared forecast error then $\mathsf{E}[\gamma(X_S)f_1(S)] > 0$, which implies (59) and (60). □

B Convex Combinations of Coefficients

Equation (13) aggregates the individual scalar slopes β_s into a single value. We can generalize this perspective and ask what properties we would like in an aggregation function, meaning a function $f : \mathbb{R}^{\bar{S}} \to \mathbb{R}$,

$$\beta_* = f(\beta_1, \ldots, \beta_{\bar{S}}),$$

that combines bank-specific coefficients β_s into an "industry" parameter β_*.

We consider the following properties:

(i) $f(kb_1, \ldots, kb_{\bar{S}}) = kf(b_1, \ldots, b_{\bar{S}})$, for all k, $b_1, \ldots, b_{\bar{S}} \in \mathbb{R}$;

(ii) $f(b, \ldots, b) = b$, for at least one nonzero $b \in \mathbb{R}$;

(iii) $b_s > 0$, for all s, implies $f(b_1, \ldots, b_{\bar{S}}) \geq 0$;

(iv) f is differentiable at zero.

Property (i) is needed for the aggregation to perform sensibly under a change of units in the measurement of X_s: if we divide each X_s by k, each β_s increases by a factor of k, and it is natural to require that β_* scale accordingly. Properties (ii) and (iii) are also very modest requirements. Property (iv) is harder to motivate but not unreasonable. These properties constrain the aggregation function as follows:

Proposition B.1. *If (i)–(iv) hold, then $f(\beta_1, \ldots, \beta_{\bar{S}})$ is a convex combination of its arguments.*

Proof. Fix $\beta \in \mathbb{R}^{\bar{S}}$. Let $g(t) = f(t\beta)$. By condition (iv), $g'(0) = \beta^\top f'(0)$. Condition (i) and (ii) imply $g(t) = tf(\beta)$, so $g'(t) = f(\beta)$ for any t. Thus, $f(\beta) = g'(0) = \beta^\top f'(0) = \sum_{i=1}^{\bar{S}} f_i'(0)\beta_i$. Condition (ii) now implies $\sum_{i=1}^{\bar{S}} f_i'(0) = 1$, and condition (iii) implies $f_i'(0) \geq 0$, for all i. Thus, $f(\beta) = \sum_{i=1}^{\bar{S}} f_i'(0)\beta_i$ is a convex combination of the components of β. □

The scalar FEO coefficient in (27) is a convex combination of the bank-specific coefficients β_s, but the pooled coefficient (13) is generally not. This property of the FEO model extends to the multivariate case under additional conditions. If all the bank-specific covariance matrices Σ_s, $s = 1, \ldots, \bar{S}$, coincide, then in (25) we get $\beta_F = \mathsf{E}[\beta_S] = \sum_s p_s \beta_s$. If all Σ_s are diagonal (but not necessarily identical), then the representation of the scalar FEO coefficient in (27) applies to each coordinate of β_F. If all Σ_s have the same eigenvectors, then we can transform the original features X_s into uncorrelated features using principal components. Using these transformed features, each coordinate of β_F is a convex combination of bank-specific coefficients.

C Sensitivity Analysis

This appendix provides supporting details for Section 6, particularly the conclusions summarized in the middle and bottom panels of Table 6.1. We begin with an analysis of forecast bias that is of independent interest.

C.1 Forecast Bias

If losses at different banks are described by different models, then forecast bias becomes inevitable when we apply a single model to all banks. But the distribution of bias across banks may differ under different choices of the single model.

Let $\hat{Y}_s$ be any of the forecasts for bank s in Table 4.1, and, as in (1), let Y_s denote the actual loss rate for bank s. Both $\hat{Y}_s$ and Y_s are evaluated at X_s. Define the forecast bias for bank s to be

$$\mathsf{bias}(s) = \mathsf{E}[\hat{Y}_s - Y_s]. \tag{66}$$

The expectation integrates over the distribution of the error ϵ_s in (1) and the features X_s.

Proposition C.1. *For each forecast in Table 4.1, the bias is as follows.*

(i) Pooled: $\mathsf{bias}(s) = \mathsf{E}[Y_S] - \mathsf{E}[Y_s] + \beta_{Pooled}^\top(\mu_s - \bar{\mu})$;

(ii) PTF in (18): $\mathsf{bias}(s) = \mathsf{E}[Y_S] - \mathsf{E}[Y_s]$;

(iii) Conditional expectation: $\mathsf{bias}(s) = \mathsf{E}[\hat{Y}_C(X_s)] - \mathsf{E}[Y_s]$;

(iv) FEO: $\mathsf{bias}(s) = \mathsf{E}[Y_S] - \mathsf{E}[Y_s] + \beta_F^\top(\mu_s - \bar{\mu})$;

(v) SEO: $\mathsf{bias}(s) = \mathsf{E}[Y_S] - \mathsf{E}[Y_s]$.

Proof. For (i), we have, using the definition of α_{Pooled} in (12),

$$\begin{aligned}\mathsf{E}[\hat{Y}_s - Y_s] &= \mathsf{E}[\alpha_{Pooled} + \beta_{Pooled}^\top X_s - Y_s] \\ &= (\mathsf{E}[Y_S] - \beta_{Pooled}^\top \bar{\mu}) + \beta_{Pooled}^\top \mu_s - \mathsf{E}[Y_s] \\ &= \mathsf{E}[Y_S] - \mathsf{E}[Y_s] + \beta_{Pooled}^\top(\mu_s - \bar{\mu}).\end{aligned}$$

For the PTF forecast, (17) and (18) yield

$$\mathsf{E}[\hat{Y}_s] = \bar{\alpha}^o = \sum_s p_s \alpha_s^o = \sum_s p_s \mathsf{E}[Y_s] = \mathsf{E}[Y_S],$$

and the bias in (ii) follows. The expression in (iii) holds by definition. The argument for (iv) is the same as the argument for (i). The bias in (v) follows from (iv) because we see from (38) that the SEO forecast for bank s subtracts $\beta_F^\top(\mu_s - \bar{\mu})$ from the FEO forecast. □

In every case of Proposition C.1, the average bias $\sum_s p_s \mathsf{bias}(s)$ is zero, but the methods differ in how they distribute bias across banks. We saw previously that the PTF and SEO methods go the farthest in equalizing differences; we now see that the bias for each of these methods is the difference $\mathsf{E}[Y_S] - \mathsf{E}[Y_s]$ between the average loss rate for all banks and the average for an individual bank.

Using the relationship $\beta_{Pooled} = \beta_F + \Lambda\delta$ from (30), we see that the difference between the expressions in (i) and (iv) is

$$\mathsf{bias}_{Pooled}(s) - \mathsf{bias}_{FEO}(s) = \delta^\top \Lambda^\top (\mu_s - \bar{\mu}).$$

In light of the discussion in Section 4.2, this difference is the expected disparate impact on bank s of using the pooled model.

C.2 Improvement in Intercept α_s

By taking the expectation of (48), we get

$$\mathsf{E}[\hat{Y}_F(X_l)] = \sum_s ps(\alpha_s + \beta_s \mu_s) + \beta_F(\mu_l - \bar{\mu}), \tag{67}$$

and the same holds for the expected pooled forecast with β_F replaced by β_{Pooled}. It follows that, for any banks s and l,

$$\frac{\partial \mathsf{E}[\hat{Y}_F(X_l)]}{\partial \alpha_s} = p_s > 0.$$

In other words, all expected forecasts decrease following a reduction in α_s.

In contrast, for the pooled model we get

$$\frac{\partial \mathsf{E}[\hat{Y}_P(X_l)]}{\partial \alpha_s} = p_s - \frac{\partial \mathsf{cov}(\alpha_S, \mu_S)/\partial \alpha_s}{\sum_t p_t \sigma_t^2 + \mathsf{var}(\mu_S)} \beta_s(\bar{\mu} - \mu_l) = p_s + p_s \frac{(\mu_s - \bar{\mu})(\mu_l - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \mathsf{var}(\mu_S)} \beta_s. \tag{68}$$

Bank s benefits from its reduction of α_s, in the sense that the derivative with $l = s$ is positive. For $l \neq s$, the sign of (68) does not admit a simple description. In particular, it may be negative when μ_s and μ_l are on opposite sides of $\bar{\mu}$, meaning that one bank's loans are riskier than average and the other bank's loans are less risky than average.

For the bias under FEO we have

$$\frac{\partial \mathsf{bias}_F(l)}{\partial \alpha_s} = p_s - \mathbf{1}\{l = s\}$$

It is then immediate that

$$\frac{\partial \mathsf{bias}_F(l)}{\partial \alpha_s} > 0 \text{ if } l \neq s \quad \text{and} \quad \frac{\partial \mathsf{bias}_F(s)}{\partial \alpha_s} < 0.$$

The direction of change makes sense. If the bias for a bank is positive, meaning that the industry model overestimates its losses, then improvements at other banks will reduce loss forecasts and thus reduce the bias. The bank's own improvements will increase the bias by reducing the bank's own losses by more than they reduce the model's forecasts. The situation is reversed for a bank with a negative bias.

However, for the pooled regression method,

$$\frac{\partial \mathsf{bias}_P(l)}{\partial \alpha_s} = p_s + p_s \frac{(\mu_s - \bar{\mu})(\mu_l - \bar{\mu})}{\sum_s p_s \sigma_s^2 + \mathsf{var}(\mu_S)} \beta_s - \mathbf{1}\{l = s\},$$

and the direction of change is unclear.

C.3 Improvement in Loan Quality

Now suppose bank s improves the quality of its loan portfolio, resulting in a smaller μ_s. This has no effect on β_F, which makes sense — changing one bank's loan quality should not change the sensitivity of losses to loan quality. However, it is evident from (11) that β_{Pooled} does change with μ_s.

Under FEO, the mean the mean predicted loss rate satisfies

$$\frac{\partial \mathsf{E}\hat{Y}_F(X_l)}{\partial \mu_s} = p_s \beta_s + \beta_F(\mathbf{1}\{l = s\} - p_s),$$

which is always positive if $l = s$. This means that an improvement in bank l's loan quality (a reduction in μ_l) reduces bank l's mean predicted losses. In the pooled model,

$$\frac{\partial \mathsf{E}\hat{Y}_P(X_l)}{\partial \mu_s} = p_s \beta_s + \beta_{Pooled}(\mathbf{1}\{l = s\} - p_s) + (\mu_s - \bar{\mu}) \frac{\partial \beta_{Pooled}}{\partial \mu_s};$$

this expression could be negative, even with $l = s$, meaning that a bank could be penalized (through a higher mean predicted loss rate) as a result of improving its loan quality.

The sensitivity of the bias under FEO is given by

$$\frac{\partial \mathsf{bias}_F(l)}{\partial \mu_s} = (\mathbf{1}\{l = s\} - p_s)(\beta_F - \beta_s);$$

in particular, the bias for bank l moves in opposite directions with respect to changes in μ_l and μ_s, $s \neq l$. Suppose industry model overestimates bank l's losses, in the sense that the bias is positive, and suppose the industry model overestimates bank l's sensitivity to loan quality, in the sense that $\beta_F > \beta_l$. Then bank l will benefit (in the sense of reducing the bias) from improving its loan quality by reducing μ_l.

For the pooled regression,

$$\frac{\partial \mathsf{bias}_P(l)}{\partial \mu_s} = (\beta_{Pooled} - \beta_s)(\mathbf{1}\{l = s\} - p_s) + (\mu_s - \bar{\mu})\frac{\partial \beta_{Pooled}}{\partial \mu_s}.$$

The sign of this expression does not admit a simple condition.

C.4 Improvement in Loan Management

Now suppose bank s improves its abilities in loan management, resulting in a reduction in β_s. The mean predicted loss rate under FEO satisfies

$$\frac{\partial \mathsf{E}[\hat{Y}_F(X_l)]}{\partial \beta_s} = p_s(\mu_s + \mu_l - \bar{\mu}),$$

and is positive if $\mu_s + \mu_l > \bar{\mu}$. In the pooled model

$$\frac{\partial \mathsf{E}[\hat{Y}_P(X_l)]}{\partial \beta_s} = p_s\mu_s + \frac{p_s(\sigma_s^2 + \mu_s(\mu_s - \bar{\mu}))}{\sum_i p_i(\sigma_i^2 + \mu_i(\mu_i - \bar{\mu}))}(\mu_l - \bar{\mu}),$$

so $[\sigma_s^2 + \mu_s(\mu_s - \bar{\mu})](\mu_l - \bar{\mu}) > 0$ is a sufficient condition for the sensitivity to be positive. Regardless of the value of μ_l, the sensitivity is positive for all sufficiently large μ_s.

For $l \neq s$, the sensitivity of the bias for bank l with respect to β_s equals the sensitivity of the mean predicted loss because the actual expected loss $\mathsf{E}[Y_l]$ is unaffected by β_s. We therefore focus on the case $l = s$. Under FEO,

$$\frac{\partial \mathsf{bias}_F(s)}{\partial \beta_s} = (p_s - 1)\mu_s + p_s(\mu_s - \bar{\mu}),$$

which is guaranteed to be negative if $\mu_s < \bar{\mu}$. Under the pooled model, the sign of

$$\frac{\partial \mathsf{bias}_P(s)}{\partial \beta_s} = (p_s - 1)\mu_s + \frac{p_s(\sigma_s^2 + \mu_s(\mu_s - \bar{\mu}))}{\sum_i p_i(\sigma_i^2 + \mu_i(\mu_i - \bar{\mu}))}(\mu_s - \bar{\mu})$$

does not admit a simple characterization.

D Additional Information on Empirical Analysis

Table D.1 lists the bank holding companies included in our empirical analysis and the symbols we use to refer to them. The companies are listed in order of size by total assets.

Ticker	**Bank Name**
JPM	JPMORGAN CHASE & CO.
BAC	BANK OF AMERICA CORPORATION
C	CITIGROUP INC.
WFC	WELLS FARGO & COMPANY
GS	GOLDMAN SACHS GROUP, INC.
MS	MORGAN STANLEY
SCHW	CHARLES SCHWAB CORPORATION
USB	U.S. BANCORP
PNC	PNC FINANCIAL SERVICES GROUP, INC.
TFC	TRUIST FINANCIAL CORPORATION
TD	TD GROUP US HOLDINGS LLC
BK	BANK OF NEW YORK MELLON CORPORATION
COF	CAPITAL ONE FINANCIAL CORPORATION
STT	STATE STREET CORPORATION
HSBC	HSBC NORTH AMERICA HOLDINGS INC.
FITB	FIFTH THIRD BANCORP
USAA	UNITED SERVICES AUTOMOBILE ASSOCIATION
BMO	BMO FINANCIAL CORP.
CFG	CITIZENS FINANCIAL GROUP, INC.
AXP	AMERICAN EXPRESS COMPANY

Table D.1: Symbols and names of included bank holding companies.

We construct the loss rates, past due rates, and allowance rates using the entries in FR Y-9C forms outlined in Table D.2.

E Robustness Check: Pre-COVID Data

We repeat the analysis of Section 5, limiting the data to 2002–2019. This serves two purposes. It addresses the possibility that our results are driven by a few extreme values during the COVID period 2020–2021. It also accounts for a change in how how banks measure allowances (the Current Expected Credit Losses methodology) that took effect at the end of 2019. The evidence for heterogeneity and its impact is generally at least as strong using the pre-COVID data as using data through 2021.

Variables	**Loan Types**	**2007Q1 – Present**	**2003Q1 – 2006Q4**
Loan Amount	CC	BHCKB538	BHCKB538
	FL	BHDM5367	BHDM5367
	CRE	Owned: BHCKF160 Other: BHCKF161	BHDM1480
	CI	BHCK1763	BHCK1763
	Total	BHCK2122	BHCK2122
Allowance Amount	Total	BHCK3123	BHCK3123
Charge-Offs	CC	BHCKB514	BHCKB514
	FL	BHCKC234	BHCKC234
	CRE	Owned: BHCKC895 Other: BHCKC897	BHCK3590
	CI	BHCK4645	BHCK4645
Recoveries	CC	BHCKB515	BHCKB515
	FL	BHCKC217	BHCKC217
	CRE	Owned: BHCKC896 Other: BHCKC898	BHCK3591
	CI	BHCK4617	BHCK4617
Past Due: 30-89 days and accruing	CC	BHCKB575	BHCKB575
	FL	BHCKC236	BHCKC236
	CRE	Owned: BHCKF178 Other: BHCKF179	BHCK3502
	CI	BHCK1606	BHCK1606
Past Due: 90 days and accruing	CC	BHCKB576	BHCKB576
	FL	BHCKC237	BHCKC237
	CRE	Owned: BHCKF180 Other: BHCKF181	BHCK3503
	CI	BHCK1607	BHCK1607
Past Due: non-accrual	CC	BHCKB577	BHCKB577
	FL	BHCKC229	BHCKC229
	CRE	Owned: BHCKF182 Other: BHCKF183	BHCK3504
	CI	BHCK1608	BHCK1608

Table D.2: Loan variables and FR Y-9C form correspondence.

Covariance	α				β_{PDR}				β_{AR}				γ			
Estimation	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank heteroskedasticity	0.03	0.00	0.00	0.34	0.00	0.00	0.00	0.00								
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00								
Newey-West HAC	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00								
bank heteroskedasticity	0.04	0.00	0.02	0.20	0.15	0.00	0.00	0.00	0.30	0.00	0.00	0.00				
time clustered	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00				
Newey-West HAC	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.00				
bank heteroskedasticity	0.07	0.00	0.01	0.04	0.01	0.00	0.00	0.00					0.59	0.42	0.00	0.03
time clustered	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					0.04	0.00	0.00	0.00
Newey-West HAC	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00					0.01	0.00	0.00	0.00
bank heteroskedasticity	0.03	0.00	0.01	0.27	0.11	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.59	0.43	0.00	0.28
time clustered	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00
Newey-West HAC	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00

Table E.1: Heterogeneity tests using pre-COVID data and a forecast horizon of one quarter.

Covariance	α				β_{PDR}				β_{AR}				γ			
Estimation	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI	CC	FL	CRE	CI
bank heteroskedasticity	0.00	0.04	0.60	0.42	0.00	0.00	0.00	0.00								
time clustered	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00								
Newey-West HAC	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00								
bank heteroskedasticity	0.00	0.00	0.01	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
time clustered	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00				
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00				
bank heteroskedasticity	0.00	0.24	0.54	0.14	0.00	0.00	0.00	0.00					0.32	0.00	0.00	0.00
time clustered	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00					0.00	0.00	0.00	0.00
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00					0.00	0.00	0.00	0.00
bank heteroskedasticity	0.01	0.04	0.03	0.48	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.05	0.31	0.00	0.00	0.00
time clustered	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Newey-West HAC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00

Table E.2: Heterogeneity tests using pre-COVID data and a forecast horizon of one year.

Loan Type	*Past Due Rate*				*Allowance Rate*				*Macro PC*			
	β_{Pooled}	β_F	% diff	90% CI	β_{Pooled}	β_F	% diff	90% CI	γ_{Pooled}	γ_F	% diff	90% CI
CC	0.711	0.742	4.3	(-4.8, 9.7)								
FL	0.053	0.059	11.4	(-1.1, 26.0)								
CRE	0.126	0.126	-0.2	(-3.9, 4.6)								
CI	0.291	0.313	7.4	(-0.7, 10.1)								
CC	0.378	0.386	2.2	(-8.4, 13.8)	0.663	0.748	12.9	(2.2, 24.7)				
FL	0.007	0.015	105.7	(-451.2, 410.3)	0.388	0.348	-10.2	(-24.9, 20.6)				
CRE	0.083	0.069	-16.8	(-37.2, -0.9)	0.127	0.161	26.6	(-6.3, 357.2)				
CI	0.240	0.258	7.3	(-3.0, 13.1)	0.114	0.139	22.2	(-14.3, 59.3)				
CC	0.704	0.750	6.6	(-2.7, 11.8)					0.164	-0.157	-195.7	(-982.4, 467.5)
FL	0.051	0.057	11.9	(-0.8, 28.7)					0.446	. 0.410	-8.2	(-31.3, 2.5)
CRE	0.126	0.125	-0.2	(-4.3, 5.1)					0.056	0.057	1.6	(-89.2, 63.1)
CI	0.276	0.296	7.2	(-1.0, 9.8)					0.572	0.546	-4.5	(-9.5, 5.7)
CC	0.359	0.384	6.8	(-15.2, 22.1)	0.665	0.748	12.5	(3.1, 26.3)	0.391	0.047	-88.0	(-438.6, 418.1)
FL	0.007	0.015	111.7	(-656.0, 423.3)	0.384	0.341	-11.2	(-27.4, 29.8)	0.136	0.146	7.2	(-95.4, 101.4)
CRE	0.083	0.069	-17.2	(-35.0, -4.8)	0.127	0.162	28.0	(-6.3, 220.4)	0.004	-0.022	-617.0	(-262.9, 604.7)
CI	0.233	0.250	7.2	(-0.5, 13.8)	0.102	0.125	23.3	(-25.2, 49.9)	0.501	0.452	-9.8	(-13.1, 6.3)

Table E.3: Comparison of coefficients using pre-COVID data and a forecast horizon of one quarter.

Loan Type	*Past Due Rate*				*Allowance Rate*				*Macro PC*			
	β_{Pooled}	β_F	% diff	90% CI	β_{Pooled}	β_F	% diff	90% CI	γ_{Pooled}	γ_F	% diff	90% CI
CC	0.838	0.885	5.6	(-0.7, 15.5)								
FL	0.046	0.052	12.3	(-1.5, 29.3)								
CRE	0.104	0.101	-3.1	(-8.1, 4.6)								
CI	0.210	0.216	2.9	(-5.6, 6.5)								
CC	0.821	0.869	5.9	(-1.5, 15.3)	0.034	0.034	-1.2	(-162.4, 181.5)				
FL	-0.002	0.005	-386.7	(-517.0, 604.9)	0.415	0.371	-10.5	(-21.0, 16.4)				
CRE	0.058	0.045	-21.5	(-89.5, 42.8)	0.136	0.157	15.8	(-37.5, 180.4)				
CI	0.183	0.191	4.2	(-6.6, 10.5)	0.061	0.067	8.5	(-95.8, 43.3)				
CC	0.656	0.683	4.1	(-2.3, 9.5)					4.016	3.796	-5.5	(-7.5, -1.4)
FL	0.042	0.047	13.7	(-2.9, 36.4)					1.256	1.225	-2.4	(-7.2, 1.1)
CRE	0.097	0.094	-3.0	(-8.8, 4.5)					0.635	0.641	1.0	(-3.6, 6.9)
CI	0.162	0.159	-1.5	(-10.1, 2.7)					1.892	1.898	0.3	(-0.2, 2.0)
CC	0.628	0.658	4.9	(-4.3, 13.1)	0.057	0.052	-8.6	(-157.6, 129.7)	4.034	3.809	-5.6	(-9.0, -1.5)
FL	-0.002	0.008	-517.3	(-827.9, 474.8)	0.384	0.319	-17.0	(-25.1, 13.6)	0.953	0.981	2.9	(-5.4, 8.6)
CRE	0.058	0.055	-4.6	(-49.7, 134.5)	0.117	0.112	-4.1	(-153.3, 192.6)	0.587	0.586	-0.2	(-11.1, 12.8)
CI	0.155	0.156	0.7	(-10.7, 7.4)	0.016	0.009	-44.1	(-248.6, 129.8)	1.881	1.891	0.6	(-0.4, 3.1)

Table E.4: Comparison of coefficients using pre-COVID data and a forecast horizon of one year.