Racial Screening on the Big Screen? Evidence from the Motion Picture Industry*

Liang Zhong ⓡ
Boston University

Angela Crema ⓡ
New York University

M. Daniele Paserman ⓡ
Boston University and NBER

June 2022

Abstract

In many contexts, a decision-maker must screen applicants using only imperfect information about their quality. The decision-maker may use information about the applicant's race or gender to guide their decision, resulting in discrimination. Using data on the group differences in the output of applicants that pass the screening process, is it possible to assess the extent and nature of discrimination?

In this paper, we tackle this question in the context of the motion picture industry. The underrepresentation of non-white actors among the leading awards has raised concerns about discrimination in the industry. Using machine learning technology, we construct a new data set with racial identifiers of the cast of more than 7,000 motion pictures released in the United States between 1997 and 2017. We use this data set to test the predictions of a model of discrimination. Producers receive offers to produce movies ("scripts"), observe the racial composition of the cast implied by the script, and receive a noisy signal of a movie's expected revenue. They must then decide whether to produce and release the movie. The model nests different forms of discrimination and delivers a rich set of predictions for the difference in revenue distribution between "white" and "non-white" movies, which allows to distinguish between different forms of discrimination. Empirically, we document the following facts: a) Average box-office revenue of non-white movies is substantially higher (82%, or 60 log-points) than white movies; b) This difference is driven primarily by the left-tail of the distribution – the gap at the 10th percentile is about 80 log-points while at the 90th percentile it is about 27 log-points, suggesting that non-white movies with low box-office potential are never produced; c) These findings are robust to different definitions of non-white movies, the inclusion of an extensive battery of control variables, or using profits as the dependent variable; d) Using the difference between actual opening-weekend box office

*Preliminary and incomplete. We thank Matthew Gudgeon, Anna Weber and seminar participants at the United States Military Academy and the Boston University Labor Reading Group for helpful suggestions. ⓡ Names are in random order following Ray ⓡ Robson (2018).

revenue and predicted revenue based on the number of theaters in which the movie is initially released, we find that relative to white movies, non-white movies substantially overperform relative to expectations.

The pattern of results cannot be rationalized by customer discrimination, nor by a model of statistical discrimination where the signal embodied in non-white movies is less precise. Instead, the findings are consistent with our theoretical model if producers hold non-white movies to a higher standard or if they systematically underestimate the revenue potential of non-white movies.

Keywords: Discrimination, Machine learning, Motion picture industry
JEL codes:

1 Introduction

An employer must decide whether to hire a job applicant. An admission committee must decide whether to admit a candidate to its entering freshman class. A journal editor must decide whether to accept an article for publication. All these settings are characterized by a *decision maker* who must make an in-or-out decision about an *applicant*, having only imperfect information about the applicant's quality. The decision-maker may use information about the applicant's race or gender to guide their decision, which may result in discrimination, i.e., the unequal treatment of applicants with otherwise identical characteristics. The econometrician, however, can typically observe only the ex-post outcomes of these decisions: the worker's productivity, the student's grades, or the number of citations received by an article. If we observe differences by race or gender in the outcomes, what can we infer about the extent and nature of discrimination by the decision maker?

In this paper, we address this question in the context of the U.S. motion picture industry, where the producer is the decision maker. There are three main advantages to studying discrimination in the motion picture industry: First, this setting is of intrinsic interest because of the widespread perception of bias in the industry. For example, in the 2010s, only 7% of nominees for the Academy Awards were African Americans, far less than their proportion in the population.[1] Does this underrepresentation reflect racial bias? Second, box-office revenue can be measured precisely as productivity, which is essential for the topics of discrimination. Last but not least, as Riley (2022) pointed out, movie characters can act as role models in students' education attainment. So, discrimination in this context will potentially affect children's sense of the world and the outlook of value.

We develop a model of discrimination that allows us to interpret differences in box-office revenue, conditional on production. In the model, producers receive offers to produce movies ("scripts," akin to the applicant in the examples above). They observe the expected racial composition of the cast based on the script and receive a noisy signal of the movie's expected box-office revenue. Based on the information, they must choose whether to produce

[1] `https://www.washingtonpost.com/news/arts-and-entertainment/wp/2016/02/26/these-charts-explain-how-oscars-diversity-is-way-more-complicated-than-you-think/`, accessed on October 26, 2021.

the movie and release it to the public or not. We define a "white" movie as a movie in which the leading roles are solely played by whites and a "non-white" movie as a movie in which the leading roles include non-whites. Our model nests different forms of discrimination within it and delivers a rich set of predictions regarding the extent and nature of discrimination. We distinguish between three types of discrimination: a) *customer discrimination*, whereby moviegoers have a preference for white movies over non-white movies; b) *employer* or *taste-based discrimination*, where the producer suffers a negative utility from producing a non-white movie; and c) *statistical discrimination*, where the signal conveyed by non-white movies is less informative about the movie's true quality. We show that the moments of the distribution of box-office revenue of movies *that are produced* allow one to distinguish between the three types of discrimination.

To test the model's predictions, we construct a novel data set with racial identifiers for the cast of more than 7,000 motion pictures released in the United States between 1997 and 2017. We obtained the data by scraping the popular websites IMDB,[2] and combined it with extensive information from OpusData, a private company that specializes in providing data and information on the motion picture industry.[3] The racial identifiers are constructed using a machine learning architecture that combines a convolutional neural network (CNN) and support vector machine (SVM; Anwar and Islam, 2017). As a result, the algorithm obtains a classification accuracy of more than 95% in our validation data set, which is considered excellent in the image classification literature.

In our main analysis, we define a movie as "non-white" if two of the four top-billed performers are classified as non-white.[4] We document the following findings. First, the average box-office revenue of non-white movies is substantially *higher* than that of white movies. The raw non-white/white revenue gap is close to 300%. Inclusion of a standard set of control variables for other movie characteristics and the cast lowers the gap to between 80% and 90%, still large and highly statistically significant. Second, the box office premium of non-white movies is driven primarily by movies in the bottom half of the distribution.

[2] `http://www.imdb.com`

[3] `www.opusdata.com`

[4] The non-white category includes mostly African-Americans but may also include Asians, Hispanics, and other ethnicities.

Quantile regressions show that the adjusted gap is around 60 log points (about 82%) at the bottom quantiles of the distribution, but the gap at the upper end of the distribution shrinks to about 35 log points (about 42%). These results are robust to different definitions of "non-white" movies or different dependent variables (e.g., profit margins or profits). Third, we create a measure of the extent to which a movie's box-office revenue overperforms relative to expectations. Following Moretti (2011), we calculate this as the residual in a regression of opening weekend box-office revenue on the number of opening-weekend theaters. We find that, relative to white movies, non-white movies substantially overperform relative to expectations.

These results are not consistent with either customer discrimination or statistical discrimination. Instead, we argue that the results are consistent with taste-based discrimination, where non-white movies are held to a higher standard, i.e., they are produced only if the expected revenue surpasses a threshold that is higher than the one set for white movies. This pattern may result from either proper taste-based discrimination on the part of producers or a systematic underestimation of the box-office potential of non-white movies.

This paper contributes to the stream of the literature that aims to understand the nature of unequal treatments by either distinguishing between statistical and taste-based discrimination in the data,[5] or testing for the presence of one of the two in a specific market or context. These goals have been pursued experimentally (List, 2004; Zussman, 2013; Doleac and Stein, 2013; Bohren et al., 2019), as well as by testing theoretical predictions on already collected data (Altonji and Pierret, 2001; Knowles et al., 2001; Charles and Guryan, 2008). Close to our paper is the work by Knowles et al. (2001). They derive a test to disentangle between statistical discrimination and racial animus in the context of motor vehicle searches conducted by police officers. We contribute to this literature by focusing on a market where some salient interactions exist between employees and final customers, and customer demand drives profit maximization. Specifically, we propose a simple theoretical framework that nests not only employer taste-based and statistical discrimination but also customer racial animus, and delivers testable predictions for each source of unequal treatment. The

[5]For a review of the literature on the topic, see Guryan and Charles (2013) and Lippens et al. (2020).

paper also adds to the line of research that documents the presence of racial discrimination in the movie industry (Weaver, 2011; Fowdur et al., 2012). Closest to our work is the paper by Kuppuswamy and Younkin (2020), who find that movies with multiple African American actors enjoy a box office premium. They rule out customer racial tastes as a discrimination mechanism through an experimental approach. We confirm their conclusion on a more comprehensive data set and provide an analytical framework that can be used to interpret racial differences in the entire distribution of revenue as a function of different forms of discrimination.

The rest of the paper proceeds as follows. Section 2 describes the institutional background of the motion picture industry. Section 3 presents our theoretical model and discusses its empirical implications. Section 4 describes the data and the machine learning methodology used to classify performers by race. Section 5 presents the main empirical findings and assesses the robustness of the results to different definitions of race or dependent variables. Section 6 presents suggestive evidence of incorrect beliefs on the revenue potential of non-white movies within the industry. Section 7 discusses and concludes.

2 Institutional background of film production

Film-making is a complex industry involving a multiplicity of skills, targets, and decision makers. Each movie displayed on screen has been through three articulated macro-phases: script writing, production, and distribution. This paper studies racial discrimination at the production stage.

The key decision maker in the production phase is the producer.[6] She decides whether a script is worth being turned into a movie and, if so, raises the money (sometimes supported by one or more executive producers). The producer is then responsible for the financial and logistic aspects of the movie.[7] She oversees the hiring of the director, who is the creative soul of the movie, the cast, and the crew, and decides on the budget allocation.

[6]See for reference Crimson Engine, 2018.

[7]The Producers Guild of America (P.G.A.) has established that the producer's name in the film credits can be followed by the *p.g.a.* certification mark only if the producer has performed a significant portion of the producing duties, which includes being physically present on set for a substantial fraction of the production time (P.G.A., n.d.).

In our conceptual framework, we assume that the movie script itself determines the racial composition of the leading characters in a movie. While the producer and the casting team[8] might have some latitude in choosing the supporting characters, we think it is plausible that the race of the main characters can be inferred directly from the script. In fact, casting notices for actors typically specify features such as race and ethnicity (and other aspects of physical appearance) for specific roles.

Our model, presented below, describes the producer's decision about whether to produce the movie after she has seen the movie's script, and observed the racial composition of the cast and a signal of the movie's quality.

3 A Model of the screening process

We propose a theoretical framework that helps us understand how observed box-office revenue can inform us about the extent and nature of discrimination in the industry. We assume that the movie production process is modeled according to the following timeline.

Step 1: Script arrival

There are two types of movies: white movies, denoted by w, and non-white movies, denoted by b. A risk-averse producer wishes to maximize *log* revenue, denoted by π. The producer receives a script and perfectly observes its type t. However, box office revenues are not observed. We assume that ex-ante, box-office revenues of a movie of type t follow a log-normal distribution with type-specific parameters μ_t and $\sigma^2_{\pi t}$:

$$\pi \mid t \sim N(\mu_t, \sigma^2_{\pi t}), \qquad \forall t \in \{w, b\}$$

.

Step 2: Signal and prior updating

Based on the script, the producer updates her prior about the movie's box-office revenue. Formally, we can think of the producer observing a signal (y) of the movie's expected box-office revenue. The signal is normally distributed and is well-calibrated, meaning that in

[8]While the producer can be correctly thought of as the primary decision maker in the production process, casting decisions are typically shared among multiple roles.

expectation it is equal to the movie's actual (log) box-office revenue, but it is noisy. Critically, we assume that the precision of the signal may differ by type. Therefore:

$$y \mid \pi \sim N(\pi, \sigma_{yt}^2).$$

Given this setup, it is straightforward to calculate the posterior mean of log box-office revenue, conditional on the signal and the movie's type:

$$E(\pi \mid y, t) = \frac{\sigma_{\pi t}^2}{\sigma_{\pi t}^2 + \sigma_{yt}^2} y + \frac{\sigma_{yt}^2}{\sigma_{\pi t}^2 + \sigma_{yt}^2} \mu_t. \tag{1}$$

Step 3: Production decision

Producers will produce a movie and release it to the public if the expected log box-office revenue, conditional on the movie's type and signal, exceeds a given threshold. This threshold (the *revenue threshold*) is exogenously given. We can think of it as the reservation revenue from a sequential search model, i.e., the value of the revenue that makes the producer indifferent between producing the movie or waiting for a better script. We denote this revenue threshold π_{0t}, making the critical assumption that the threshold is type-specific. For example, this could result from the producer having a taste for producing movies of a given type.

The movie will be produced if

$$E(\pi|y, t) > \pi_{0t}, \tag{2}$$

This is equivalent to saying that the movie will be produced only if the signal y exceeds a given threshold (the *signal threshold*). Based on equation 1 and condition 2, it is easy to show that the signal threshold is

$$\bar{y}_t = \pi_{0t} + (\pi_{0t} - \mu_t)\frac{\sigma_{yt}^2}{\sigma_{\pi t}^2}.$$

In other words, the signal threshold is type-specific, and depends on the revenue threshold, the parameters of the prior distribution, and the precision of the signal.

This threshold, together with the statistical features of the ex-ante distribution of box-office revenue and the distribution of revenue conditional on the signal, determines the ex-post distribution of box-office revenue. The following proposition establishes the comparative statics of the signal threshold with respect to the parameters of the model.

Proposition 1 *The following comparative statics results hold:*

(a) $\bar{y}_t$ *decreases in* μ_t.

(b) $\bar{y}_t$ *increases in* π_{0t}.

(c) If $\pi_{0t} > \mu_t$, $\bar{y}_t$ *increases in* σ^2_{yt}.

(d) If $\pi_{0t} < \mu_t$, $\bar{y}_t$ *decreases in* σ^2_{yt}.

Proof. See Appendix A ■

The first two statements in Proposition 1 are straightforward and intuitive. If ex-ante expected (log) revenue is higher (a high μ_t), the movie will still be produced even if the signal is not very good. Similarly, when the revenue threshold (π_{0t}) is high, the signal must be excellent to produce the movie. The third and fourth items in the Proposition are more involved but are familiar from the literature on statistical discrimination (Aigner and Cain, 1977; Lundberg and Startz, 1983; Neumark, 2012). Intuitively, if the signal is less precise (a high value of σ^2_{yt}) and the producer wants to produce only high revenue movies, she will have to set a high signal threshold to make sure she only picks the right tail of the revenue distribution (item (c) in Proposition 1); on the other hand, if the producer only wants to cull out very low revenue movies and the signal is uninformative, the threshold must be set at a low value to ensure that only the very worst (i.e., lowest-revenue) movies are weeded out (item (d) in the proposition).

3.1 Predictions for empirical work

Proposition 1 characterizes the properties of the signal threshold that determines whether a movie is produced. In practice, we do not observe the signal threshold, so the results are not useful for empirical analysis. However, we do observe the box office revenue of movies that are actually produced and released to the public. The mean and variance of log box-office revenue, conditional on production, are:

$$E(\pi \mid y > \bar{y}_t) = \mu_t + \sigma \frac{\phi(\frac{\pi_0 - \mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0 - \mu_t}{\sigma})}$$

$$Var(\pi \mid y > \bar{y}_t) = \sigma^2 \left(1 + \sigma_{yt}^2 + \lambda(\frac{\pi_0 - \mu_t}{\sigma}) \left(\frac{\pi_0 - \mu_t}{\sigma} - \lambda(\frac{\pi_0 - \mu_t}{\sigma})\right)\right),$$

where $\sigma = \frac{\sigma_{\pi t}^2}{\sqrt{\sigma_{\pi t}^2 + \sigma_{yt}^2}}$ and $\lambda(x) = \frac{\phi(x)}{(1-\Phi(x))}$.

We can then formulate our central proposition, which enables us to predict how different types of discrimination affect box-office revenues of white and non-white movies produced.

Proposition 2 *Let $E_t \equiv E(\pi|y > \bar{y}_t)$ and $Var_t \equiv Var(\pi|y > \bar{y}_t)$ be the mean and variance of log box-office revenue conditional on production. Then*

$$E_t = \mu_t + \sigma \frac{\phi(\frac{\pi_0 - \mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0 - \mu_t}{\sigma})}$$

$$Var_t = \sigma^2 \left(1 + \sigma_{yt}^2 + \lambda(\frac{\pi_0 - \mu_t}{\sigma}) \left(\frac{\pi_0 - \mu_t}{\sigma} - \lambda(\frac{\pi_0 - \mu_t}{\sigma})\right)\right),$$

where $\sigma = \frac{\sigma_{\pi t}^2}{\sqrt{\sigma_{\pi t}^2 + \sigma_{yt}^2}}$ and $\lambda(x) = \frac{\phi(x)}{(1-\Phi(x))}$. The following comparative statics results hold:

(a) E_t and Var_t increase in μ_t.

(b) E_t increases in π_0, Var_t decreases in π_0.

(c) E_t decreases in σ_{yt}^2, Var_t increases in σ_{yt}^2.

Proof. See Appendix A ■

We focus first on the intuition for the comparative statics of E_t with respect to the parameters. The intuition for the first two results is straightforward. Expected revenue conditional on production is higher the more shifted to the right is the prior distribution of revenue (result (a)), and the higher the revenue threshold (result (b)). The third result says that expected revenue conditional on production is always higher the more precise the signal. This result is slightly counterintuitive, because the signal threshold can either

increase or decrease with σ_{yt}^2 (Proposition 1, results from (c) and (d)). To gain intuition on this result, it is useful to consider the extreme cases of a perfectly informative ($\sigma_{yt}^2 = 0$) or perfectly uninformative signal ($\sigma_{yt}^2 \to \infty$). If the signal is perfectly informative, the movie is produced only if the signal (which is exactly equal to box-office revenue) is above the revenue threshold. This implies that expected revenue conditional on production is strictly greater than μ_t because some movies will be below the threshold and are not produced. On the other hand, if the signal is perfectly uninformative, whether a movie exceeds the signal threshold conveys no information about its revenue – the expected revenue conditional on production is, therefore, μ_t.

To gain intuition on the variance results, it is helpful to consider the case of a perfectly informative signal. The distribution of revenue conditional on production is a truncated normal distribution, with the truncation point equal to the revenue threshold π_0. If the whole distribution is shifted to the right and the threshold remains the same, it is easy to see that the variance also increases (result (a)). If the revenue threshold π_0 increases, the truncation point shifts to the right, and the distribution variance decreases (result (b)). As for the third result, it is again useful to consider the two polar cases: with a perfectly informative signal, the distribution of box-office revenue is a truncated normal distribution, which necessarily has a smaller variance than the untruncated distribution that results if the signal is perfectly uninformative.

We can now use Proposition 2 to characterize the mean and variance of box-office revenue for white and non-white movies under different types of discrimination.

Case 1: Customer discrimination. Customer discrimination implies that the viewing public has a preference for white movies over non-white ones. In terms of our model, this means that the entire distribution of log box-office revenue for white movies is shifted to the right relative to the distribution for non-white movies, or that $\mu_b < \mu_w$.

Then, by result 1, it follows that $E_b < E_w$, and $V_b < V_w$. We can therefore state the following prediction:

Prediction 1 *Under customer discrimination, the mean log box-office revenue for non-white movies is lower than for white movies, and the variance of log box-office revenue for non-white*

movies is lower than for white movies.

Case 2: Taste-based discrimination. We can think of taste-based discrimination as the producer suffering a negative utility from producing non-white movies. Holding everything else constant, the producer will produce a non-white movie only if expected log revenue exceeds a higher threshold than the one she sets for white movies to compensate her for the disutility of producing a non-white movie. In this case, $\pi_{0b} > \pi_{0w}$. By result 2, we have that $E_b > E_w$ and $V_b < V_w$. We can therefore state Prediction 2:

Prediction 2 *Under taste-based discrimination, the mean log box-office revenue for non-white movies is higher than for white movies. The variance of log box-office revenue for non-white movies is lower than for white movies.*

Case 3: Statistical discrimination. We classify under statistical discrimination the case where the informativeness of the signal for non-white movies is smaller than the one for white movies. We believe this assumption is plausible because it may be more difficult for producers to evaluate how successful a movie with non-white characters will be. In this case, $\sigma^2_{yb} > \sigma^2_{yw}$. By result 3, we have that $E_b < E_w$ and $V_b > V_w$. We can therefore state prediction 3:

Prediction 3 *Under statistical discrimination, the mean log box-office revenue for non-white movies is lower than for white movies. The variance of log box-office revenue for non-white movies is higher than for white movies.*

In the remainder of the paper, we use the above predictions to assess the extent and nature of discrimination in the motion picture industry.

4 Data

4.1 Facial classification

A key ingredient of the paper is creating a data set with racial identifiers for the movie's cast. We classify all performers as either "white" or "non-white"[9] using the algorithm proposed

[9]We define "non-white" as a residual category including African Americans, East and South Asians, Native Americans, etc.

by Anwar and Islam (2017)[10]. The algorithm is based on a machine learning architecture that combines a convolutional neural network (CNN) and support vector machine (SVM), described below.

Step 1. We started with a sample of more than 7000 motion pictures released in the United States between 1997 and 2017, taken from Opus Data,[11] a private company that collects data on the industry. For each movie, we took the names of the four top-billed performers. We then scraped and cropped the image appearing on each performer's page on the popular website IMDB.[12]

Step 2. We used the Visual Geometry Group[13] (V.G.G.) technique to locate the actor's face on each picture. The output of this step is a vector of information extracted from each image, or a "feature vector."

Step 3. We repeated step 2 on our training data set, the Chicago Face Database (CFD).[14] This database is intended for use in scientific research. It is useful as it contains images of 597 unique individuals (both male and female) who self-identify as White, Black, Asian, or Latino/a.

Step 4. We used CFD to train our algorithm using the Support Vector Machine (SVM) approach.[15] Intuitively, the purpose of SVM is to find the "best separation line," meaning the hyper-plane that correctly separates white from non-white performers when such performers are located in a multi-dimensional space through their feature vectors.

Step 5. We applied our trained algorithm to the pictures obtained from Steps 1 and 2. We validated our algorithm on a subsample of actors for which we manually coded the racial groups and obtained a success rate of 95%. A few examples of the outcomes of our classification algorithm are presented in Figure 1.

[10]Link: `https://arxiv.org/ftp/arxiv/papers/1709/1709.07429.pdf`.
[11]`www.opusdata.com`
[12]`www.imdb.com`
[13]See for reference `https://www.robots.ox.ac.uk/~vgg/`.
[14]The CFD is available at `https://www.chicagofaces.org/`.
[15]See for reference `https://scikit-learn.org/stable/modules/svm.html`.

4.2 Additional variables

From Opus data, we collected aggregate financial data (box office revenue, production budget, opening weekend revenue, etc.) and metadata (genre, production method) for all movies in our sample. The main variables of interest in our data set include the gross domestic box-office revenue, production costs,[16] the movie run time, Metacritic score, release date, MPAA rating, the number of theaters in which the movie was released, and the number of weeks in which the movie was in theaters. We also collected information on the gender and age of the four top-billed performers. We created a variable called "star power," equal to the cumulative box-office revenue of all movies in which each performer appeared up to the release date of the current movie.

Summary statistics are shown in Table 1. The top panel in the table shows that about 11 percent of the top-billed performers in our sample are non-white. More than three-quarters of the movies have zero non-white performers, and about 17 percent have only one non-white performer. Our baseline analysis defines a movie as non-white if at least two of the four top-billed performers are non-white. Based on this definition, about five percent of the movies in our sample are non-white. We also assess the robustness of the results to different definitions of non-white movies.

As for the other variables, we see that the distribution of box office revenue is heavily skewed to the right. Therefore, we take the logarithm of box office revenue as the main dependent variable in our baseline analysis. We collapse some of the more niche genres into broader categories so that all movies fall into one of five broad genres. For some of the variables, we only have incomplete data. For example, production costs are available for only about 55 percent of the sample,[17] while the Metacritic score is available only for 68 percent of the sample. To maximize sample size, in the empirical analysis, we replace missing values with zeros and add a dummy variable indicating that the variable is missing if the missing value is not central to the analysis.

[16] All monetary values are expressed in 2005 dollars.

[17] Probit and logit regressions suggest that movies with higher revenue have a higher probability of non-missing cost, and non-white films are more likely to have a missing cost variable.

5 Results

5.1 Non-parametric analysis

Figure 2 presents a box-whisker plot of box-office revenue by the number of non-white performers in the movie. The mean box-office revenue increases markedly with the number of non-white performers, while the dispersion of the distribution decreases. The box-whisker plot allows us to examine in detail the behavior of different quantiles of the distribution as the number of non-white performers increases. The comparison between the upper and lower quantiles is striking. The 25^{th} percentile of the distribution and the lower adjacent value increase substantially with the number of non-white performers. On the other hand, the 75^{th} percentile increases more modestly, and the upper adjacent value seems to decrease. The entire left tail of the distribution of movies with a predominantly non-white cast appears to be missing. It is consistent with the notion that non-white movies have to pass a higher bar to be produced.

Of course, this analysis does not take into account other observable differences between white and non-white movies. In the following, we assess whether the non-white premium in box-office revenue is robust to including a broad set of other movie and cast characteristics.

5.2 OLS regressions

The main regression model is the following:

$$\ln y_{it} = \beta_0 + \beta_1 Nonwhite_{it} + \beta_2 X_{it} + \delta_t + \varepsilon_{it}, \tag{3}$$

where y_{it} denotes domestic box-office revenue, in 2005 U.S. dollars, of movie i released in year t; $Nonwhite_{it}$, the key explanatory variable of interest, is a dummy variable indicating whether at least two of the four top-billed performers are non-white; X_{it} is a vector of additional control variables, including both cast (average age, gender composition, the "star power" variable described previously) and movie (the production budget, the MPAA rating, the Metacritic score, run time, genre dummies) characteristics; δ_t is a year-of-release fixed effect, and ε_{it} is the usual error term.

The results are presented in Table 2. The first column of the table shows the unadjusted difference in mean log revenue between white and non-white movies, without any controls. Mean box-office revenue of non-white movies is almost 3 times as high as that of white movies ($\exp(1.095) \approx 2.99$). In column 2, we include controls for other characteristics of the cast (average age, gender composition, and star power), and the coefficient remains almost unchanged. In column 3, we add controls for the production budget, a dummy for whether the production budget is missing, and all other movie characteristics, including genre and year-of-release fixed effects. The coefficient on the non-white indicator drops to about 0.6, implying that non-white movies earn about 82 percent more than white movies at the box office.[18]. Finally, column 4 replicates column 3, but we restrict the analysis to those movies with non-missing data on production costs. The results in this restricted sample are mostly unchanged – the coefficient on the non-white indicator rises to 0.68, implying that non-white movies earn on average about 98 percent more than white movies[19].

These initial results on the differences between white and non-white movies are not consistent with either a model of customer discrimination, where audiences prefer white movies to non-white movies nor a model of statistical discrimination, where the signal conveyed by non-white scripts is less informative about future box-office revenue. Both models predict that white movies should have on average higher box-office revenue than non-white movies, in contrast to our findings. Instead, the results are consistent with a model of taste-based discrimination, where non-white movies are held to a higher standard, i.e., they are only produced if the revenue exceeds a higher threshold than the one required of white movies. In what follows, we look at how other features of the distribution differ between white and non-white movies.

[18]The coefficient appears to be driven down primarily by the inclusion of the Metacritic score and the cost of production variables.

[19]We have data on the script languages for approximately 70% of the working sample. Within the matched sample, approximately ninety percent of the movies in our sample have English as main language, while around 95% have some parts of the script in English. Hence, our results are not driven by foreign-language movies.

5.3 Quantile regressions

The model described in Section 3 derived predictions not only for the mean revenue but also for the variance and the entire distribution of box-office revenue. In this subsection, we look specifically at the white-nonwhite gap at different quantiles of the distribution. Specifically, we estimate a series of quantile regressions of the following type:

$$Q_\tau(\ln y_{it}|Nonwhite, X) = \gamma_{0\tau} + \gamma_{1\tau} Nonwhite_{it} + \gamma_{2\tau} X_{it} + \delta_t,$$

where $Q_\tau(\ln y_{it}|Nonwhite, X)$ denotes the τth conditional quantile of the distribution of log box-office revenue, and $\tau \in \{0.05, 0.10, ...0.95\}$. The main coefficients of interest are the $\gamma_{1\tau}$'s, which tell us how the conditional quantiles of the distribution of log box-office revenue differ between white and non-white movies.

Figure 3 plots the quantile regression coefficients against the quantiles. As was already apparent from the box-whisker plots in Figure 2, from the 20^{th} quantile onwards there is a clear downward trend in the quantile coefficients: The white-nonwhite gap at the lower quantiles is around 60 log points, while it is only about 30 log points at the upper quantiles. This further confirms that non-white movies at the low end of the distribution of box-office revenue are never produced.

5.4 Robustness

We next investigate the robustness of our results to different definitions of movie type and different dependent variables.

Classification of non-white movies. In Table 3, we consider additional definitions of "non-white" movies. The first column in the table reproduces the results using our baseline classification of non-white movies as those in which at least two of the four top-billed performers are non-white. The first row in the table shows the OLS. results from Table 2, while the remaining rows present the quantile regression coefficients at selected quantiles. All specifications include the full set of control variables.

In column 2, we change the definition of non-white movies to include all movies in which at least *one* of the four top-billed performers is non-white. We view this as a noisier indicator of

the movie type, as a non-white actor may be cast in a supporting role in a movie that is mainly about white characters and storylines (a form of *tokenism*). Using this definition, the OLS coefficient is substantially reduced (about 24 log points) but still large and highly statistically significant. The pattern of quantile regression coefficients is also clearly downward sloping, with the gap going from about 30 log points at the 10^{th} to about 13 log points at the 90^{th} percentile. In column 3, we replace the dummy indicator for non-white movies with the share of non-whites among the four top-billed performers. The results are quantitatively and qualitatively similar to those of the baseline specification. Finally, in column 4, we classify a movie as non-white only if the top-billed performer is non-white. According to this definition, the average white-nonwhite premium is slightly smaller than in the baseline (45 log points), and the pattern of the quantile regression coefficients is also downward sloping.

On the whole, Table 3 shows that the main conclusions regarding the white-nonwhite premium and the nature of discrimination in the industry are not sensitive to the exact definition of non-white movies.

Choice of the dependent variable. In all the analysis so far, we have looked at the logarithm of box-office revenue as the primary dependent variable of interest. The main reason for this choice is that box-office revenue is readily available for almost all movies, and it has been traditionally used as the primary metric for assessing the commercial success of a movie.[20] However, producers are likely also to consider the expected costs of a movie when making production decisions, an aspect that we have ignored so far. In the Opus data set, we observe a movie's production budget for about 54% of all movies, so that we can calculate various measures of profit.[21] In Table 4 we examine whether the main results are robust to different measures of the dependent variable. We restrict attention only to those

[20] A movie's revenue also includes the income derived from international box-office sales, home video, broadcasting, and merchandise. With the rise of streaming services over the past few years, the share of incomes from domestic box-office revenue has shrunk. Nevertheless, for the period under analysis, we still think domestic box office revenue is a good measure of a movie's success.

[21] It should be noted that the production budget does not represent all of a movie's production costs, which typically also include marketing costs. Marketing costs are rarely disclosed in the industry. Also, the producer typically does not collect all of the box-office revenue, as the theaters also receive a cut, which will depend on bilateral negotiations between the distributor and the theaters, as well as on the length of time that a movie has been in theaters. Therefore, our measures should be viewed as only a coarse estimate of a movie's profits.

movies for which we observe the production budget. All specifications include the full set of control variables.

In column 1, we use the logarithm of the gross profit margin as a dependent variable, defined as the ratio of domestic box-office revenue to the production budget. The results are broadly consistent with those in the previous sections: non-white movies have on average a substantially higher profit margin, and the white-nonwhite gap becomes smaller as we move from the low to the high end of the distribution.

In column 2, we look instead at a total profit, calculated simply as the difference between box-office revenue and the production budget. It is still the case that the average non-white movie earns a higher profit than the average non-white movie (by about $11 million). However, we no longer observe a clear declining pattern in the white-nonwhite gap as we move from lower to upper quantiles in the profit distribution. In fact, the gap appears to be fairly stable at all quantiles of the distribution. This could be partly due to the shape of the profit distribution, which tends to be quite right-skewed. We confirm this in column 3, where we use the level of box-office revenue (rather than the logarithm) as the dependent variable. We find a positive premium favoring white movies, but now the pattern of quantile regression coefficients shows that the gap becomes larger as we move from the low to the high end of the distribution. We should note that, given the substantial right skew of the revenue distribution, the predictions regarding the variance of box-office revenue conditional on production, derived from a model that assumes normal distributions, no longer hold necessarily.

5.5 The white-nonwhite gap in residual variance

An alternative approach to verify our prediction is to look directly at how residual variance differs between white and non-white movies. Borrowing from the heteroskedasticity literature, we posit that the squared residuals from the OLS regression in equation 3 have the form:

$$u_{it}^2 = \exp(Z_{it}'\alpha),$$

where the vector Z_{it} contains a subset of the variables included in the main regression (potentially all of them); therefore, we estimate regressions of $\ln \hat{u}_{it}^2$ on the racial indicator and the additional control variables. The results are reported in Table 5.

In column 1, the residual variance is assumed to depend only on the racial indicator. Consistent with the results of the box-whisker plot and the quantile regressions, we find that non-white movies have a substantially lower residual variance than white movies. In columns 2 and 3, we progressively add additional controls to the variance regression. The results are essentially unchanged – the residual variance of non-white movies is 36% to 57% lower than that of white movies.

In the remaining three columns, we experiment with different definitions of non-white movies. The coefficients on the racial variables in the residual regressions are always negative but statistically significant at conventional levels only when the share of non-white performers is used.

In Table 6 we conduct the same type of analysis, but now examining robustness with respect to alternative dependent variables, as in Table 4. For all three measures, we find that the residual variance is lower for non-white movies. However, the difference is statistically significant only when using the gross profit margin (column 1). These results are broadly consistent with those of Table 4, where we found that the pattern of quantile regression coefficients exhibits a downward sloping pattern only when the profit margin is used as the dependent variable.

Overall, the results in this section, even though somewhat sensitive to the outcome measure, are broadly consistent with the predictions of the theoretical model with taste-based discrimination. Non-white movies appear to be held to a higher standard, resulting in a higher mean and lower variance of box-office revenue for the produced movies.

5.6 Heterogeneity Analysis

In Table 7 we explore heterogeneity of our results along a number of different dimensions.

First, we look at whether our results are driven by movies produced and distributed by specific segments of the industry. One concern is that perhaps our results are capturing

differences between movies produced by the major studios (the so-called "Big-6")[22], and those produced by smaller studios. It could be that the smaller box-office revenue of white movies reflects the fact that these are often produced by small independent studios, while non-white movies are passed over by these studios altogether. Columns 1 and 2 of the table, however, show that this is not the case: the non-white revenue premium is present among movies distributed by both types of studios.

We next look at differences across genre (columns 3-5 of the table). The non-white premium is more pronounced among comedies and dramas, where it is more likely that the script conveys information about the racial composition of the cast. By contrast, the non-white premium is small and not statistically significant in action/adventure movies.

Columns 6 and 7 examine heterogeneity by time period. We look separately at movies produced before and after 2007, the median year in our sample. If taste-based discrimination declines over time, either because of a change in attitudes, or because non-profit maximizing producers are weeded out by market forces, we would expect the non-white premium to shrink over time. The premium is in fact slightly larger in the pre-2007 period, but the difference is small and not statistically significant. It would be hard to conclude based on this evidence that taste-based discrimination has declined substantially over time.

Finally, in columns 8 and 9 we look at whether the results differ by the gender composition of the cast. We define "female" movies as those in which (strictly) more than 50% of the leading actors are women. The non-white premium is considerably larger among female movies, suggesting that non-white movies must pass an even higher threshold if the cast is predominantly female.

6 Alternative explanation: is the industry surprised?

The empirical results so far suggest that non-white movies are held to higher production standards than white movies. In the context of the model from Section 3, this means that $\pi_{0b} > \pi_{0w}$. We interpreted this difference as indicating that producers dislike producing non-

[22]These Big-6 studios are: Warner Bros., Paramount Pictures, Walt Disney, Sony / Columbia Pictures, Universal Studios, and 20th Century Fox. These six studios accounted for almost 90% of the US/Canadian market as of 2007.

white movies and face a disutility cost every time they produce one. As a result, π_b needs to be higher than π_w, to compensate the producer for the disutility of producing non-white movies.

An alternative, non mutually exclusive, interpretation is that the industry systematically underestimates the revenue potential of non-white movies relative to white movies. In other words, actual box-office revenue for non-white movies is π_b, but producers perceive it to be $\hat{\pi}_b = \pi_b - e_b$, with $e_b > 0$. This explanation would yield similar predictions to the ones derived from taste-based discrimination, even if the nature of discrimination in the industry is quite different.

Our model is intrinsically unable to identify taste-based disutility costs and incorrect beliefs separately. Nevertheless, we can make some progress on the front exploiting the decision that distributors make on the number of theaters at which the movie would be displayed over the opening weekend. We argue that this decision reflects the market's rational expectation on the movie potential upon production, as distributors' decision-making is less likely to be affected by taste-based or statistical discrimination. While producers "sign" a movie as a creation of theirs and create a permanent bond with the film, studios and theater owners are more likely to make distribution choices based on purely profit-maximizing considerations once the movie has been produced. Moreover, statistical discrimination should also be of relatively less importance at the distribution stage because distributors also observe the ex-post quality of the movie, rather than just a script..

We conjecture that distributors choose the number of theaters as a function of expected customer demand. If non-white movies are displayed in fewer theaters than white movies, this indicates that distributors expect relatively smaller revenue from the non-white movies. Therefore, if non-white movies have the same level of customer demand but are displayed in fewer theaters, we conclude that distributors underestimate their revenue potential.

We can test for the hypothesis that the industry systematically underestimates the revenue potential of non-white movies using data on film distribution. Specifically, we first regress first-weekend box office revenues on the number of theaters in which movies are projected over the first weekend upon their release. Following Moretti (2011), we interpret

this as a proxy for the industry expectation of a movie's box-office revenue. The residuals from this regression can then be viewed as a measure of the industry's underestimation or overestimation of a movie's revenue potential. If the non-white mean residual is significantly larger than the white mean residual, this will suggest that the industry systematically underestimates non-white movies' revenue potential relative to white movies.

We start by running a simple bivariate regression of log first-weekend revenues on the log number of theaters.[23] The results, shown in the top panel of Table 8, suggest that the independent variable explains around 89% of the variability in the dependent (column 1). The R-squared is relatively stable as further controls are added (columns 2 through 4). The number of theaters is hence a good predictor of first-weekend revenues.

We then test whether the residuals obtained from the regressions in Table 8, column (1) are different on average across non-white and white movies. The results are presented in the bottom panel of the table. We find that the mean residual for non-white movies ranges between 15 and 20 log points, while the mean residual for white movies is close to zero. This is equivalent to saying that the industry underestimates the first-weekend success of non-white movies by around 15% on average, while white movie performance is predicted much more accurately. The difference between the white and the non-white residual is always statistically significant.

We therefore conclude that at least part of the explanation for the higher standard to which non-white movies are held lies in systematic underestimation of non-white movies' box-office potential by distributors and, likely, producers.

7 Conclusion

This paper presented a framework for detecting the extent and nature of discrimination in contexts in which decision-makers screen applicants, and the econometrician can only observe the outcomes of applicants who successfully passed this screening process. The framework

[23] We pass from 6,990 (Table 2) down to 6,276 observations (Table 8) because of missing information for opening weekend revenues (310), number of theaters (253), or both (96), or zero first weekend revenues (55). We choose the log-to-log specification over the level-to-level, log-to-level, and level-to-log specifications because it fits the data best.

nests several leading theories of discrimination and derives a rich set of testable empirical predictions.

We applied these tests in the context of racial differences in the U.S. motion picture industry. We found consistent and robust evidence that non-white movies earn a box-office premium. The gap is particularly pronounced at low quantiles of the distribution, suggesting that non-white movies with low box-office potential are never produced; in other words, those non-white movies are held to a higher standard in the production decision. This evidence is consistent with taste-based discrimination on the part of producers, who suffer a utility loss from producing non-white movies. However, we also find some evidence that the results may be driven by systematic underestimation by producers and distributors of the revenue potential of non-white movies. These results are somewhat puzzling because both taste-based discrimination and systematic underestimation of a movie's potential because of the racial composition of its cast should not be able to persist in a competitive equilibrium. At the same time, the motion picture industry is highly concentrated, with the "Big Five" studios[24] typically accounting for more than 80% of the industry's total market share. We leave the investigation of this puzzle to future research.

While the specific application in this paper looked at the motion picture industry, it is not difficult to see how our model can be readily applied to other contexts in which decision makers can use group identifiers to screen applicants, and one can observe the outcome or productivity of successful applicants. For example, Card et al. (2020) find that female-authored papers in four leading economics journals receive 25% more citations than comparable male-authored papers and argue that one possible explanation for these findings is that referees hold female authors to a higher bar. Other contexts in which this framework can be applied are a promising avenue for future research.

[24]Disney, Paramount, Sony/Columbia, Universal Pictures and Warner Bros. The group was formerly known as the "Big Six" when it also included 20$^{\text{th}}$ Century Fox, which Disney acquired in 2020.

References

[1] Aigner, D. J., and Cain, G. (1977). Statistical theories of discrimination in labor markets. Industrial and Labor Relations Review, 30(2), 175-187.

[2] Altonji, J. G., & Pierret, C. R. (2001). Employer learning and statistical discrimination. The Quarterly Journal of Economics, 116(1), 313-350.

[3] Anwar, I., & Islam, N. U. (2017). Learned features are better for ethnicity classification. Cybernetics and Information Technologies, vol. 17, no. 3, pp. 152-164.

[4] Bohren, J. A., Haggag, K., Imas, A., & Pope, D. G. (2019). Inaccurate statistical discrimination: An identification problem (No. w25935). National Bureau of Economic Research.

[5] Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2020). Are Referees and Editors in Economics Gender Neutral? Quarterly Journal of Economics, 135(1), 269-327.

[6] Charles, K. K., & Guryan, J. (2008). Prejudice and wages: an empirical assessment of Becker's The Economics of Discrimination. Journal of political economy, 116(5), 773-809.

[7] Crimson Engine (2018, February 13). What does a Producer ACTUALLY do? Link.

[8] Davidson, A. (2012). How Does the Film Industry Actually Make Money? The New York Times Magazine, link.

[9] Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. The Economic Journal, 123(572), F469-F492.

[10] Fowdur, L., Kadiyali, V., and Prince, J. (2012). Racial bias in expert quality assessment: A study of newspaper movie reviews. *Journal of Economic Behavior & Organization, 84*(1), 292-307.

[11] Fusion (2016, September 14). Hollywood's Diversity Problem: Breaking Down the Casting Process. Link.

[12] Guryan, J., & Charles, K. K. (2013). Taste-based or statistical discrimination: the economics of discrimination returns to its roots. The Economic Journal, 123(572), F417-F432.

[13] Knowles, J., Persico, N., & Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. Journal of Political Economy, 109(1), 203-229.

[14] Kuppuswamy, V., and Younkin, P. (2020). Testing the Theory of Consumer Discrimination as an Explanation for the Lack of Minority Hiring in Hollywood Films. *Management Science, 66*(3), 1227-1247.

[15] Lippens, L., Baert, S., Ghekiere, A., Verhaeghe, P. P., & Derous, E. (2020). Is labour market discrimination against ethnic minorities better explained by taste or statistics? A systematic review of the empirical evidence. IZA Discussion Paper No. 13523.

[16] List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. The Quarterly Journal of Economics, 119(1), 49-89.

[17] Lundberg, S. J. and Startz, R. (1983). Private Discrimination and Social Intervention in Competitive Labor Markets. American Economic Review, 73(3), 340-347.

[18] Moretti, E. (2011). Social learning and peer effects in consumption: Evidence from movie sales. The Review of Economic Studies, 78(1), 356-393.

[19] Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. Journal of Human Resources, 47(4), 1128-1157.

[20] Producers Guild of America (n.d.). Link.

[21] Ray, D. ⓡ Robson, R. (2018). Certified Random: A New Order for Coauthorship. American Economic Review, 108(2), 489-520.

[22] Riley, E. (2022). Role Models in Movies: The Impact of Queen of Katwe on Students' Educational Attainment. The Review of Economics and Statistics 1–48.

[23] Smith, D. (2021). Independent Producer vs. Studio Producer. Chron, link.

[24] Weaver, A. J. (2011). The role of actors' race in white audiences' selective exposure to movies. Journal of Communication, 61(2), 369-385.

[25] Zussman, A. (2013). Ethnic discrimination: Lessons from the Israeli online market for used cars. The Economic Journal, 123(572), F433-F468.

Tables and Figures

Table 1: Summary Statistics

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
PANEL A: Classification of movies by type					
Share of non-white performers	7,840	0.11	0.22	0	1
At least one non-white	7,840	0.23	0.42	0	1
At least two non-whites	7,840	0.05	0.23	0	1
Distribution of the number of non-white performers (percentages):					
0	77.2%				
1	17.3				
2	3.6				
3	1.6				
4	0.3				
PANEL B: Other variables					
Gross revenue(in Millions of 2005 Dollars)	7,205	26	54.3	$1.94 * 10^{-5}$	804
Ln (Gross revenue)	7,205	14.15	3.39	2.97	20.50
Cost(in Millions of 2005 Dollars)	3,984	37.7	45.1	$1.1 * 10^{-3}$	907
Ln(Cost)	3,984	16.73	1.45	7.01	20.63
Run time(minutes)	6,851	103.54	18.47	38	600
IMDB score	4,962	6.25	0.97	1.50	9
Metacritic score	4,962	51.55	17.08	1	100
Average age of billed performers	7,762	41.91	10.40	10	99
Star power(in Millions)	7,840	263	304	0	2,350
Ln(Star power)	7,840	17.51	4.53	0	21.58
Number of weeks	6,472	11.62	14.66	1	476
Ln(Number of screens)	6,802	4.36	3.17	0.69	8.43
Distribution of movies by genre (percentages):					
Action	16.77				
Animation	0.17				
Comedy	26.20				
Drama	36.58				
Other	20.28				

Note: Source: authors' calculations. Data sources are described in Section 4.

Table 2: The non-white revenue premium

Sample:	(1) Full Ln(Gross Revenue)	(2) Full Ln(Gross Revenue)	(3) Full Ln(Gross Revenue)	(4) Non-missing cost variable Ln(Gross Revenue)
Race: At least two non-white	1.093***	1.058***	0.560***	0.628***
	(0.169)	(0.159)	(0.094)	(0.096)
Share of female		-1.078***	0.174**	0.128
		(0.142)	(0.087)	(0.107)
ln(Star Power)		0.234***	0.023***	-0.034***
		(0.008)	(0.006)	(0.008)
Average age		-0.065***	-0.004	-0.012***
		(0.004)	(0.002)	(0.003)
ln(Cost)			0.556***	0.726***
			(0.027)	(0.025)
=1 if ln(Cost) is missing			6.268***	omitted
			(0.432)	
Movie's control			Y	Y
N	6943	6943	6943	3853
R^2	0.006	0.125	0.697	0.597

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Robustness with respect to different definitions of "non-white" movies

	(1)	(2)	(3)	(4)
Race:	At least two non-white Ln(Gross Revenue)	At least one non-white Ln(Gross Revenue)	Share of non-white Ln(Gross revenue)	Leading role is non-white Ln(Gross revenue)
Race	0.560***	0.221***	0.623***	0.436***
	(0.094)	(0.053)	(0.101)	(0.075)
Q10	0.574***	0.191**	0.642 ***	0.417***
	(0.164)	(0.097)	(0.178)	(0.135)
Q25	0.613***	0.227***	0.740 ***	0.430***
	(0.131)	(0.075)	(0.139)	(0.107)
Q50	0.492***	0.184***	0.576***	0.315***
	(0.121)	(0.069)	(0.135)	(0.097)
Q75	0.344***	0.181***	0.424***	0.312***
	(0.111)	(0.064)	(0.121)	(0.087)
Q90	0.350***	0.217**	0.509***	0.325***
	(0.134)	(0.073)	(0.139)	(0.109)
Cast's control	Y	Y	Y	Y
Movie's control	Y	Y	Y	Y
N	6943	6943	6943	6943

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Robustness to different dependent variables

Sample:	(1) Non-missing cost variable Ln(Profit Margin+1)	(2) Non-missing cost variable Profit(in million)	(3) Non-missing cost variable Revenue(in million)
Race: At least two non-white	0.660***	10.57***	4.76
	(0.098)	(3.19)	(3.33)
Q10	0.591**	7.09***	3.96***
	(0.231)	(2.64)	(1.25)
Q25	0.524***	8.11***	6.31 ***
	(0.142)	(2.00)	(1.86)
Q50	0.576***	9.42***	5.64**
	(0.092)	(1.99)	(2.68)
Q75	0.450***	12.56***	8.40*
	(0.088)	(3.56)	(4.71)
Q90	0.399***	13.21*	9.18
	(0.116)	(7.93)	(8.96)
Cast's control	Y	Y	Y
Movie's control	Y	Y	Y
N	3853	3853	3853

Note: Data sources and specification are described in Sections 4 and 5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Conditional residual variance regressions:
robustness with respect to different definitions of non-white movies

	(1)	(2)	(3)	(4)	(5)	(6)
Race definition	At least two	At least two	At least two	At least one	Share	Leading role
	Dependent variable: Ln(residual square)					
Race	-0.493***	-0.490***	-0.340***	-0.074	-0.334***	-0.104
	(0.116)	(0.116)	(0.110)	(0.061)	(0.118)	(0.088)
Cast's Control		Y	Y	Y	Y	Y
Movie's Control			Y	Y	Y	Y
N	6943	6943	6943	6943	6943	6943
R^2	0.003	0.012	0.121	0.122	0.123	0.120

Note: Data sources and specification are described in Sections 4 and 5.5. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Conditional residual variance regressions:
robustness with respect to different dependent variables

	(1) Ln(Profit Margin+1)	(2) Profit(in million)	(3) Revenue(in million)
Race: At least two non-white	-0.289*	-0.124	-0.053
	(0.149)	(0.152)	(0.145)
Cast Controls	Y	Y	Y
Movie Controls	Y	Y	Y
N	3853	3853	3853
R^2	0.133	0.132	0.144

Note: Data sources and specification are described in Sections 4 and 5.5. In all specifications, the sample is restricted to observations with non-missing data on production costs. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Heterogeneity Analysis

	(1) Distributor Not Big-6	(2) Distributor Big-6	(3) Genre: Action/Adventure	(4) Genre: Comedy	(5) Genre: Drama
Race: At least two non-white	0.719***	0.450***	0.153	0.860***	0.607***
	(0.139)	(0.0920)	(0.188)	(0.173)	(0.142)
Cast Controls	Y	Y	Y	Y	Y
Movie Controls	Y	Y	Y	Y	Y
N	4770	2173	1136	1881	2593

	(6) Period: Pre-2007	(7) Period: Post-2008	(8) Gender: ≤50% female	(9) Gender: >50% female
Race: At least two non-white	0.588***	0.452***	0.544***	1.015***
	(0.131)	(0.131)	(0.104)	(0.315)
Cast Controls	Y	Y	Y	Y
Movie Controls	Y	Y	Y	Y
N	2774	4169	5932	1011

Note: Data sources and specification are described in Sections 4 and 5.5. In all specifications, the sample is restricted to observations with non-missing data on production costs. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Regressions of first-weekend theaters on number of theaters

VARIABLES	(1) First-weekend revenues, log	(2) First-weekend revenues, log	(3) First-weekend revenues, log
First-weekend # theaters, log	0.990*** (0.004)	0.980*** (0.005)	0.855*** (0.007)
Cast's control	N	Y	Y
Movie's control	N	N	Y
Residuals: white vs non-white			
Average difference	-0.153	-0.148	-0.200
p-value of t-test (two-sided)	0.011	0.013	0.000
N	6,276	6,276	6,276
R^2	0.889	0.890	0.926

Note: Data sources and specification are described in Sections 4 and 6. Cast control variables include the share of females, the average age of the four top-billed performers, and "star power" (defined as the log of performers' cumulative box office revenues up to the movie release date). Movie control variables include indicators for movie genre, indicator of whether the movie is from the "Big 6", run time, Metacritic score, MPAA rating, year fixed effects, and indicators for missing run time, Metacritic score, or MPAA rating. The movie budget cost (in the log) is included among the control variables when revenue is the dependent variable. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

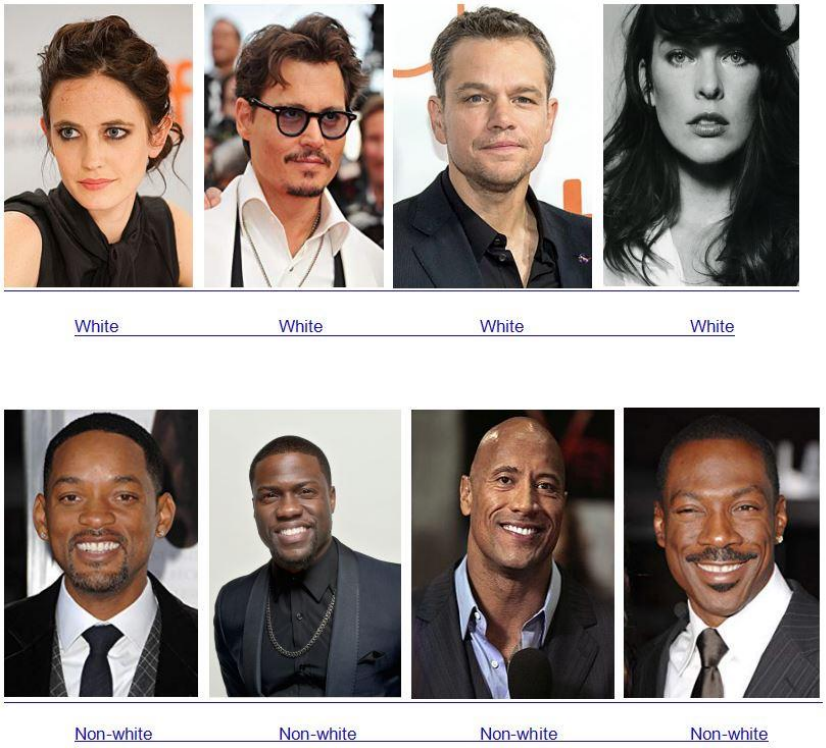

Figure 1: Output of facial classification

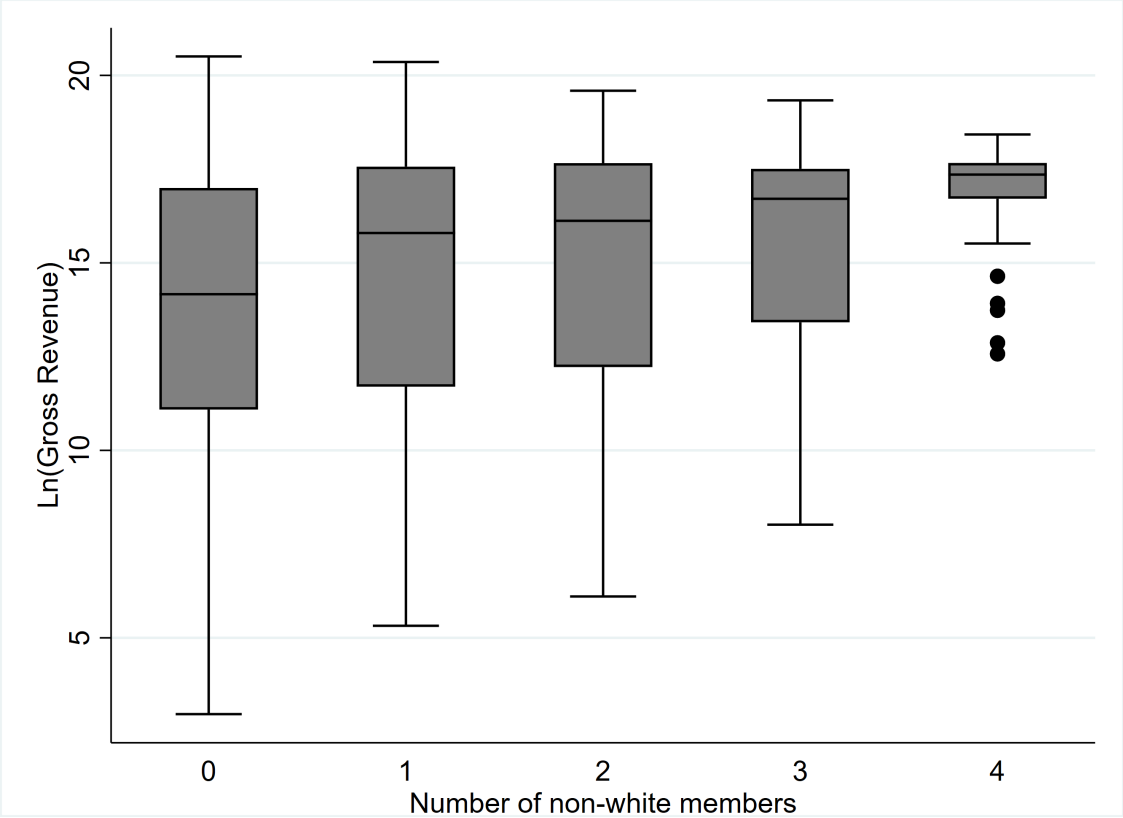

Figure 2: Revenue distribution by number of non-white members

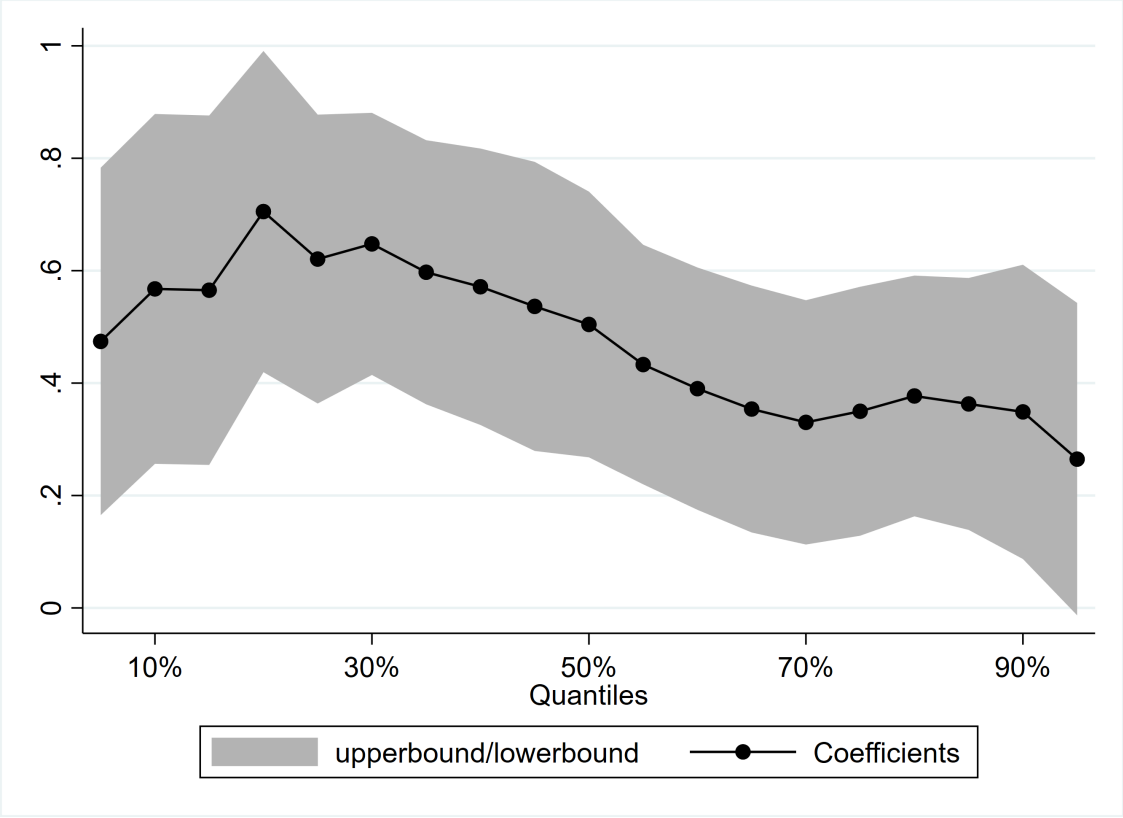

Figure 3: Coefficients are decreasing over Quantiles

A Appendix: Proofs

Lemma 1: (Inverse Mills ratio). If X is a normally distributed random variable with mean μ and variance σ^2, then

$$E(X \mid X > \alpha) = \mu + \sigma \frac{\phi(\frac{\alpha-\mu}{\sigma})}{1 - \Phi(\frac{\alpha-\mu}{\sigma})}$$

where ϕ and Φ are the p.d.f. and c.d.f. of the Normal, respectively.

Proof of Proposition 2:

Part 1: mean and variance of box-office revenue conditional on production

(i) Given two normal distributions π and y, $y_t \sim N(\mu_t, \sigma^2_{\pi t} + \sigma^2_{yt})$, so:

$$E(\pi|y) = \mu_\pi + \rho \frac{\sigma_\pi}{\sqrt{\sigma^2_{\pi t} + \sigma^2_{yt}}}(y - \mu_y)$$

$$Var(\pi|y) = \sigma^2_\pi(1 - \rho^2)$$

where ρ is the correlation between π and y: $\rho = \frac{\sigma_\pi}{\sqrt{\sigma^2_{\pi t}+\sigma^2_{yt}}}$.

Hence, $E(\pi_t \mid y_t) = \frac{\sigma^2_{\pi t}}{\sigma^2_{\pi t}+\sigma^2_{yt}} y_t + \frac{\sigma^2_{yt}}{\sigma^2_{\pi t}+\sigma^2_{yt}} \mu_t \sim N(\mu_t, \sigma^2)$, where $\sigma = \frac{\sigma^2_{\pi t}}{\sqrt{\sigma^2_{\pi t}+\sigma^2_{yt}}}$

Then by Lemma 1 and the law of total expectation,

$$\begin{aligned} E(\pi_t|y_t > \bar{y}_t) &= E(\pi_t \mid E(\pi_t \mid y_t) > \pi_0) \\ &= E(E(\pi_t \mid y_t) \mid E(\pi_t \mid y_t) > \pi_0) \\ &= \mu_t + \sigma \frac{\phi(\frac{\pi_0-\mu_t}{\sigma})}{1 - \Phi(\frac{\pi_0-\mu_t}{\sigma})} (3) \end{aligned}$$

END.

(ii) Now for variance:

$$
\begin{aligned}
Var(\pi_t|y_t > \bar{y}_t) &= Var(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E(\pi_t^2|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E(E(\pi_t^2|y_t)|E(\pi_t|y_t) > \pi_0) - E^2(\pi_t|E(\pi_t|y_t) > \pi_0) \\
&= E([Var(\pi_t|y_t) + E^2(\pi_t|y_t)]|E(\pi_t|y_t) > \pi_0) - E^2(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0) \\
&= \frac{\sigma_{\pi_t}^2 \sigma_{y_t}^2}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2} + E\Big(E^2(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big) - E^2\Big(E(\pi_t|y_t)|E(\pi_t|y_t) > \pi_0\Big) \quad (4)
\end{aligned}
$$

For a standard normal distribution, $z \sim N(0,1)$.

$$
\begin{aligned}
E(z^2|z > c) &= \frac{1}{1-\Phi(c)} \int_c^\infty \frac{z^2}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) dz \\
&= \frac{1}{1-\Phi(c)} \int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg) dz \\
&= \frac{1}{1-\Phi(c)} \int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg) dz \\
&= \frac{1}{1-\Phi(c)} \int_c^\infty \Bigg(\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big) - \Big(\frac{z}{\sqrt{2\pi}} exp\Big(-\frac{z^2}{2}\Big)\Big)'\Bigg) dz \\
&= 1 + \frac{c\phi(c)}{1-\Phi(c)}
\end{aligned}
$$

So, for $x \sim N(\mu, \sigma^2)$

$$
\begin{aligned}
1 + \frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} &= E\Bigg(\Big(\frac{x-\mu}{\sigma}\Big)^2 | \frac{x-\mu}{\sigma} > \frac{c-\mu}{\sigma}\Bigg) \\
&= \frac{1}{\sigma^2}\Big(E(x^2|x > c) - 2\mu E(x|x > c) + \mu^2\Big)
\end{aligned}
$$

Combining with

$$
E(x|x > c) = \mu + \sigma \frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}
$$

we obtain

$$
E(x^2|x > c) = \sigma^2 + \sigma^2 \frac{\frac{c-\mu}{\sigma}\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})} + \mu^2 + 2\mu\sigma \frac{\phi(\frac{c-\mu}{\sigma})}{1-\Phi(\frac{c-\mu}{\sigma})}
$$

Plugging in (4) yields

$$
Var(\pi_t|E(\pi_t|y_t) > \pi_0) = \frac{\sigma_{\pi_t}^2 \sigma_{y_t}^2}{\sigma_{\pi_t}^2 + \sigma_{y_t}^2} + \sigma^2 + \sigma^2 \frac{\frac{\pi_0-\mu_t}{\sigma}\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})} - \sigma^2 \Bigg(\frac{\phi(\frac{\pi_0-\mu_t}{\sigma})}{1-\Phi(\frac{\pi_0-\mu_t}{\sigma})}\Bigg)^2 \qquad (I)
$$

Then, $(I) = \sigma^2_{\pi_t} + \sigma^2(x\lambda(x) - \lambda^2(x))$, by $\sigma^2 = \frac{\sigma^4_{\pi_t}}{\sigma^2_{\pi_t}+\sigma^2_{y_t}}, x = \frac{\pi_0-\mu_t}{\sigma}, \lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$
END.

Part 2: comparative statics

Building blocks

Lemma 2: For $\lambda(x) = \frac{\phi(x)}{1-\Phi(x)}$, $\frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2}$ for $x \in R$.

Proof: Normally, a computer can confirm this lemma. However, when $x > 7$, both the numerator and the denominator of λ are so close to 0 that the value for λ is heavily biased. Hence, this proof will only target the case where $x > 7$.

First, taking the first derivative of $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ yields $\phi^{'}(x) = -x\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = -x\phi(x)$. It follows that

$$\begin{aligned}
1 - \Phi(x) &= \int_x^\infty \phi(u)du \\
&= -\int_x^\infty \frac{\phi'(u)}{u}du \\
&= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \int_x^\infty \frac{105\phi(u)}{u^8}du \\
&= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \frac{3\phi(x)}{x^5} - \frac{15\phi(x)}{x^7} + \frac{105\phi(x)}{x^9} - \int_x^\infty \frac{945\phi(u)}{u^{10}}du
\end{aligned}$$

Then

$$\frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7} + \frac{105}{x^9}} < \frac{\phi(x)}{1-\Phi(x)} < \frac{1}{\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7}} \qquad when \quad x > 7$$

Let the left (right) term of the inequality be denoted as LHS (RHS). We first prove that when $x > 7$, $LHS > \frac{3x+\sqrt{x^2+8}}{4}$. Assume that this is true. Then

$$LHS > \frac{3x+\sqrt{x^2+8}}{4}$$

$$\iff (x^9+3x^7-9x^5+45x^3-315x)^2 > (x^2+8)(x^8-x^6+3x^4-15x^2+105)^2$$

$$\iff x^{18}+6x^{16}-9x^{14}+36x^{12}-279x^{10}-27000x^8+7695x^6-28350x^4+99225x^2 >$$

$$x^{18}+6x^{16}-9x^{14}+20x^{12}-39x^{10}+1692x^8-1545x^6+3690x^4-14175x^2+88200$$

$$\iff 16x^{12}-240x^{10}-4392x^8+9240x^6-32040x^4+113400x^2-88200>0$$

Then for $x>7$, $16x^{12} > 16*7^2x^{10}$, i.e. $16x^{12} > 784x^{10}$, which is true.

We now prove that when $x>7$, $RHS < \frac{x+\sqrt{x^2+4}}{2}$

$\iff (x^7+x^5-3x^3+15x)^2 < (x^2+4)(x^6-x^4+3x^2-15)^2$

$\iff x^{14}+2x^{12}-5x^{10}+24x^8+39x^6-90x^4+225x^2 <$

$x^{14}+2x^{12}-x^{10}-8x^8-105x^6+66x^4-135x^2+900$

$\iff x^{10}-8x^8-36x^6+39x^4-90x^2+225>0$

Then for $x>7$, $x^{10} > 7^2x^8$, i.e. $x^{10} > 49x^8$, which is also true.

A computer can easily confirm that the lemma holds also for $x<7$, which completes the proof. In addition, for $x>0$, we can show that $x < \frac{3x+\sqrt{x^2+8}}{4} < \lambda(x) < \frac{x+\sqrt{x^2+4}}{2} < x+\frac{1}{x}$.

END.

Comparative statics 2(a):

(i) Let $x = \frac{\pi_0-\mu}{\sigma}$. Then

$$\frac{\mathrm{d}(3)}{\mathrm{d}\mu} = 1+\sigma\lambda'(x) = 1+\sigma(-x\lambda(x)+\lambda^2(x))\left(-\frac{1}{\sigma}\right) = -\left(\lambda(x)-\frac{x+\sqrt{x^2+4}}{2}\right)\left(\lambda(x)-\frac{x-\sqrt{x^2+4}}{2}\right) \qquad (i)$$

By Lemma 2, $\frac{\mathrm{d}(3)}{\mathrm{d}\mu} > 0 \ \forall x \in R$.

END.

(ii) Again let $x = \frac{\pi_0 - \mu}{\sigma}$. Then

$$\begin{aligned}\frac{\mathrm{d}(I)}{\mathrm{d}\mu} &= \sigma^2(\lambda(x) + x\lambda'(x) - 2\lambda(x)\lambda'(x))\Big(-\frac{1}{\sigma}\Big) \\ &= -\sigma\Big(\lambda(x) - x^2\lambda(x) + 3x\lambda^2(x) - 2\lambda^3(x)\Big) \\ &= \sigma\lambda(x)\left(\lambda(x) - \frac{3x + \sqrt{x^2+8}}{4}\right)\left(\lambda(x) - \frac{3x - \sqrt{x^2+8}}{4}\right)\end{aligned}$$

By Lemma 2, $\frac{\mathrm{d}(I)}{\mathrm{d}\mu} > 0 \ \forall x \in R$.

END.

Comparative statics 2(b):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\pi_0} = \sigma\lambda'(x) = \sigma(-x\lambda(x) + \lambda^2(x))(\frac{1}{\sigma}) = (-x + \lambda(x))\lambda(x) > 0$$

.

END.

(ii) See the proof for 2(a), (ii).

END.

Comparative statics 2(c):

(i)

$$\frac{\mathrm{d}(3)}{\mathrm{d}\sigma} = \sigma\lambda'(x) + \lambda(x) = \Big(-x\lambda(x) + \lambda^2(x)\Big)\left(-\frac{y_0 - \mu_t}{\sigma}\right) + \lambda(x) = \lambda(x)\Big(1 + x^2 - x\lambda(x)\Big) \quad (4)$$

where, again, $\sigma = \frac{\sigma_{\pi t}^2}{\sqrt{\sigma_{\pi t}^2 + \sigma_{yt}^2}}$, $x = \frac{\pi_0 - \mu_t}{\sigma}$, $x > 0$, and $\lambda(x) = \frac{\phi(x)}{1 - \Phi(x)}$.

When $x > 0$, we can write $(4) = x\lambda(x)(\frac{1}{x} + x - \lambda(x))$. By Lemma 2 and $\frac{x + \sqrt{x^2+4}}{2} < x + \frac{1}{x}$, $(4) > 0$ holds.

When $x < 0$, $(4) > 0$ clearly holds.

Hence, if σ_{yt}^2 increases, then σ^2 decreases and (3), i.e. E_t decreases too.

END.

(ii)

$$\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} = \sigma_{yt}^4 \frac{\Big(x\lambda(x) - \lambda^2(x)\Big)'(\sigma_{\pi t}^2 + \sigma_{yt}^2) - \Big(x\lambda(x) - \lambda^2(x)\Big)}{(\sigma_{\pi t}^2 + \sigma_{yt}^2)^2}$$

$$= \sigma_{\pi t}^4 \left(\frac{\lambda(x)\Big(1 - x^2 + 3x\lambda(x) - 2\lambda^2(x)\Big)}{\sigma_{\pi t}^2 + \sigma_{yt}^2} \Big(- \frac{\pi_0 - \mu}{\sigma^2} \Big) \Big(- \frac{\sigma_{\pi t}^2}{2(\sigma_{\pi t}^2 + \sigma_{yt}^2)^{\frac{3}{2}}} \Big) - \frac{x\lambda(x) - \lambda^2(x)}{(\sigma_{\pi t}^2 + \sigma_{yt}^2)^2} \right)$$

$$= -2x \frac{\lambda(x)\sigma_{\pi t}^4}{2(\sigma_{\pi t}^2 + \sigma_{yt}^2)^2} \left(\lambda(x) - \frac{x}{2} \right) \left(\lambda(x) - x - \frac{1}{x} \right)$$

When $x > 0$, $x < \lambda(x) < x + 1/x$, which implies that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.

When $x < 0$, it is easy to see that $\frac{\mathrm{d}(I)}{\mathrm{d}\sigma_{yt}^2} > 0$.

Hence, if σ_{yt}^2 increases, (I), i.e. Var_t increases too.

END.