Analysis of participation in physical labor industries across gender and race.

Erez Binyamin
Rochester Institute of Technology

TLDR;
Major findings:

- If the percentage of women working in an industry is **high**, that is a good predictor that the industry would **not** be classified as "physical labor"
- The percentage of Whites, Blacks, Hispanics, Asians, and the total number of people employed in an industry are **irrelevant** in predicting if the industry will be classified as "physical labor".
- Women are more likely to work in industries with larger populations.

Section 1: Abstract

This paper analyses participation in physical labor industries across gender and race. The data set used contains 300 industries and the total number of people employed, percentage of women employed, percentage of White people employed, percentage of Black people employed, percentage of Asian people employed, and the percentage of Hispanic people employed. The 300 industries in the data-set were converted from a string variable (industry name) to a Boolean variable (0 or 1). This was done using a scoring algorithm that scored "physical labor" industries as a 1 and "non-physical labor" industries as a 0.

The logic prompting this study is that discrimination in the workplace may favor certain races/genders in the "physical labor industries", and that a study of this data would reveal such discrimination. It is expected that a model of physical labor as a function of *race and gender demographics* and *total employed* will show that the only statistically significant explanatory variable will be *percentage involvement of women*. It is expected that a model of *percentage involvement of women* as a function of *race demographics*, *physical labor*, and *total employed* will show that the only statistically relevant explanatory variable will be the physical labor Boolean. Due to social bias, tradition, personal interest, and malicious hiring bias it is expected that there will be a negative linear correlation between *percentage of women involved* in a given industry and its classification as *physical labor* in both models.

Section 2: Data Analysis

Before discussing the models used and what their parameter estimates might mean, the source of the data, and the scoring of the industries as either "non-physical labor industries" or "physical labor industries" must be explained. The data set used was obtained from the US Bureau of Labor Statistics and parsed from raw html[1] into a csv[2] on a Linux command line using a bash script[3]. The first column of that csv file contained the industry category, defined qualitatively as a string. A scoring algorithm converted that column from string data to Boolean data using a keyword search. The following snippet of pseudo code and the attached bash script[4] show the scoring.

```
For each line in "bls.csv":
        IND=(industry field of line)
        if(IND contains NON-PHYSICAL_keyword):
                replace IND with a 0
        else if (IND contains PHYSICAL_keyword):
                replace IND with a 1
        else
                replace IND with a 0
```

1 Raw HTML; src/data/bls/out/raw_bls.html
2 First CSV; src/data/bls/out/raw_bls.txt
3 HTML to CSV parser; src/data/bls/out/raw_bls.csv
4 Scoring Algorithm; src/data/bls/get_parse_score.sh

The keywords used for NON-PHYSICAL were: *administration, support, sale, management, service, health, dealer, distribution, and transportation*

The keywords used for PHYSICAL were: *landscaping, building, mining, milling, manufacturing, logging, hunting, production, construction, forestry, oil, coal, petroleum, wood*

The pseudo-code snippet above first does a check for the NON-PHYSICAL keywords. This is because there are industries that support, sell, administrate ...etc and are related to and may contain PHYSICAL industry keywords. It is important that the non-physical industries that share keywords with the physical ones are not counted as physical. It is important to mention that this algorithm has been found to slightly under-estimate physical industries. This means that jobs that are in fact *physical* in nature are occasional deemed *non-physical* by this algorithm. This error was discovered when examining the output artifacts from this algorithm which are three files: bls_scored.csv[5], non-physical.txt[6], and physical.txt[7]. These files are a scored csv file ready for processing, a list of industries deemed "non-physical", and a list of industries deemed "physical" respectively.

The summary for the scored data set analyzed in this paper can be found in Tbl. 1 below.

Table 1.

	phy_lbr	tot_emp	women	white	black	asian	hispanic
Min.	0.0000	6	0.00	0.00	0.000	0.000	0.0
1st Qu.	0.0000	106	19.10	71.90	5.800	2.300	9.8
Median	0.0000	301	32.50	78.00	9.500	4.800	13.8
Mean	0.2843	2025	35.15	69.95	9.949	5.337	14.7
3rd Qu.	1.0000	1043	51.40	84.50	13.500	7.200	19.3
Max.	1.0000	155761	93.80	96.70	35.600	47.200	47.1

TBL. 1 Data set summary

Two models were used in this analysis both containing non-linearity.

Model 1

$$phy\_lbr_i = \beta_0 + \beta_1 \ln(tot\_emp_i) + \beta_2\, women_i + \beta_3\, white_i + \beta_4\, black_i + \beta_5\, asian_i + \beta_6\, hispanic_i + u_i$$

M.1 Physical labor as a function of race, gender, and ln(total employed)

Model 2

$$women_i = \beta_0 + \beta_1 \ln(tot\_emp_i) + \beta_2\, phy\_lbr_i + \beta_3\, white_i + \beta_4\, black_i + \beta_5\, asian_i + \beta_6\, hispanic_i + u_i$$

M.2 %women as a function of race, physical labor, and ln(total employed)

The non-linearity used was to take the natural log of the total number of people employed (tot_emp). This was done because all of the other explanatory variables were expressed as a percentage such that their parameter estimates corresponded to 1% changes in those variables. Taking the natural log of the total number of people employed in a given industry allows the parameter estimate to generate a more meaningful value with respect to the other parameter estimates.

5 Scored CSV; src/data/bls/bls.csv
6 Non-physical.txt; src/data/bls/out/non_physical.txt
7 Physical.txt; src/data/bls/out/physical.txt

The model calculated in M.1 shows the physical labor(dependent variable) as a function of race, gender and ln(total employed) (independent variables). This model will reveal which demographics are most significant when predicting the likelihood that a given industry involves physical labor given a particular percentage involvement of that demographic. A positive parameter estimate for a demographic would indicate that as the percentage involvement of that demographic in an industry increases, the likelihood that that industry involves physical labor also increases. A negative parameter estimate would indicate that as the percentage involvement of that demographic increases the likelihood of the industry they are involved in being classified as physical labor decreases. The explanatory variable ln(total employed), if found to be statistically significant will show the relationship between a percentage increase in the number of people employed in an industry and it's likelihood to be physical.

The second model (M.2) shows percentage of women involved in an industry (dependent variable) as a function of race, ln(total employed), and physical labor (independent variables). This model will reveal which explanatory variables are significant in predicting the percentage of women involved in a particular industry. For the race demographics, the parameter estimates will mean that for a unit increase in the percentage involvement of that race there will be a corresponding *P* percentage change in the percentage involvement of women (where *P* is the parameter estimate). The parameter estimate *P* for ln(total employed) will reveal the relationship between the population of a given industry and percentage of women involved in that industry. If the number of total people employed in an industry were to increase by 1% a corresponding change of female involvement by *P*% could be expected. This result will be particularly interesting in discovering weather women are more represented in high population industries or lower population industries.

The predictions from Section 1 only concern the values of one explanatory variable in each model. For M. 1 the parameter estimate for *women* is expected to be negative ($-1 < P < 0$) and the only statistically significant variable in the model. For M. 2 the parameter estimate for *phy_lbr* is expected to be negative and quite large ($-30 < P < -10$) and also the only statistically significant variable.

Section 3: Results and conclusions

Table 2

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.499408   0.088398   5.650 3.68e-08 ***
log(tot_emp) -0.003517   0.017362  -0.203    0.840
women        -0.008801   0.001364  -6.453 4.29e-10 ***
white         0.001328   0.001163   1.142    0.255
black        -0.006822   0.004401  -1.550    0.122
asian         0.008242   0.005111   1.613    0.108
hispanic      0.003124   0.003368   0.928    0.354
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TBL 2. Summary of M.1

The results from tbl. 2 show the parameter estimates for the beta values from M. 1. The only statistically relevant explanatory variable in the model is *women*, and it's statistical significance is very high. The parameter estimate for *women* is negative which in the context of this model means that that the higher the percentage involvement of women in a given industry, the smaller the likelihood that the industry will be classified as involving physical labor. The apparent lack of significance of the other

demographics in predicting physical labor classification should also be noted. Race demographics and the number of employed persons in a given industry are not statistically significant explanatory variables for predicting the *phy_lbr* classification of a given industry. The precise mathematical interpretation of this finding requires understanding that this model shows a Boolean (0 or 1) as a function of a percentage (0 – 100). The parameter estimate for *women* is -0.008801, which means that a unit change in *women* (1%) will correspond to a -0.008801 change in *phy_lbr*. A less literal interpretation requires an understanding the Boolean in this model represents the "extent to which an industry is classified as physical labor". This generates the conclusion that a 0 value for *phy_lbr* could be interpreted as a 0% physical labor intensive industry, while a value of 1 for *phy_lbr* could be interpreted as a 100% physical labor intensive industry. Given these assumptions a parameter estimated of -0.008801 for *women* could be interpreted to mean that a unit increase in *women* (1%) would mean that the industry for which that increase occurred is 0.88% less likely to require physical labor. These findings correspond with the prediction from Section 1.

Table 3

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.44391    3.65005  -0.670    0.504
log(tot_emp)   2.81157    0.66357   4.237 3.00e-05 ***
phy_lbr      -13.60841    2.10895  -6.453 4.29e-10 ***
white          0.20032    0.04439   4.513 9.14e-06 ***
black          0.74487    0.16844   4.422 1.36e-05 ***
asian          0.99175    0.19370   5.120 5.41e-07 ***
hispanic      -0.11861    0.13243  -0.896    0.371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

TBL 3. Summary of M.2

The results from tbl. 3 show the parameter estimates for the beta values from M. 2. All of the explanatory variables were found to be statistically significant except for the *Hispanic* variable. This partially corresponds with the prediction made in Section 1. The prediction that *phy_lbr* would be negative and statistically significant was correct, however the statistical significance of the additional explanatory variables was not anticipated. The meaning of the parameter estimate for *phy_lbr* is that for a unit change (0 to 1) there will be a 13.6 decrease in *women* (percentage of women represented in the profession). The most shocking parameter estimate however is the estimate for *ln(tot_emp)* of 2.81157. This parameter estimate means that for a unit increase in *ln(tot_emp)* which can be interpreted as a 1% increase in total number of employed persons in a given industry, there is an expected 0.028% increase in percentage of women in that industry.

The results from M. 2 also show at what rates men and women from each race join industries. If men and women were to join industries equally the parameter estimate for a unit increase of a given race would be zero (equal men and women joining would not change the representation of women). A negative parameter estimate for a race would indicate that as the representation for that race increases the representation of women decreases, or in other words more men join from that race than women. A positive parameter estimate for a given race would indicate that women join industries from that race in greater numbers than men. The statistically significant parameter estimates for *White, Black,* and *Asian* are 0.20032, 0.74487, and 0.99175 respectively. The most surprising being the Asian estimate. For Asians a 1% increase in Asian representation is expected to yield a 0.99% increase in the representation of women.

Section 4: Retrospective

Many studies that analyze female representation focus on abstract controversial high-level concepts like economic compensation, leadership representation, etc. It can be empirically proven that men and women are physically different when considering muscle mass and bone density. This analysis of industries shows that those differences can be expected to cause an advantage/disadvantage that could lead to justifiable discrimination for/against women. The overall message of this analysis is that discrimination does happen, and it is not always unjust or even counter intuitive. It would likely be to all parties determent if actions were taken to mitigate the discrepancy in female representation in physical labor professions.

An unintended discovery from this analysis was the tendency for smaller industries to be comprised of more men, and that as industries grow in population of workers the representation of women increases. The closest thing to an explanation offered for this may also explain the lack of women in dangerous professions. If one assumes that smaller industries tend to be associated with greater risk, one might believe that men are more likely to take risks, and that could explain the discrepancy in industry population and the representation of women.

The key discovery in this paper is that the difference between the male and female demographics is statistically significant in the context of physical labor and race demographics are not. This finding could be further reinforced with a better scoring algorithm for physical labor, a larger data set, and a more diverse data set across more industries and geographic areas. It is also worth noting that if economists analyzing this data were more educated that would likely discover more, explain it better, and identify inaccuracies and failings with more confidence and precision.

The moral from the story of this paper is that it is useful to prove things we think we already know. By investing the time to give justification to our theories, our opinions will hold more weight. It is one thing to say “I bet there are less women in logging because they are weaker” and it is another thing entirely to have read this report.