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 into a csv on a Linux command line using a bash script. 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 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
```

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.csv, non-physical.txt, and physical.txt. 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	51	5.90	42.60	0.70	0.000	3.80
1st Qu.	0.0000	150	24.40	74.50	7.40	3.125	11.62
Median	0.0000	413	36.45	79.10	10.25	5.300	15.05
Mean	0.2591	1737	39.86	79.36	11.26	6.050	16.68
3rd Qu.	1.0000	1312	52.90	85.00	14.18	7.400	19.98
Max.	1.0000	35043	93.80	96.70	35.60	47.200	47.10

TBL. 1 Data set summary

The summary provided in Tbl. 1 describes a handful of significant data points that illustrate meaningful information about this data set. In tabulated form below are the interesting data points (Tbl 1.1).

Table 1.1

	Minimum		Maximum	
	Industry	**Value**	**Industry**	**Value**
Total Employed	Aerospace product and parts manufacturing	51	Education and health services	35043
Women	Logging	5.90%	Child day care services	93.8%
White	Nail salons and other personal care services	42.6%	Farm product raw material merchant wholesalers	96.7%
Black	Aerospace product and parts manufacturing	0.7%	Bus service and urban transit	35.6%
Asian	Logging	0%	Nail salons and other personal care services	47.2%
Hispanic	Logging	3.8%	Cut and sew apparel manufacturing	47.1%

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.

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

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.562901	1.755448	0.890	0.374	
log(tot_emp)	-0.001166	0.017502	-0.067	0.947	
women	-0.008945	0.001330	-6.727	1.05e-10	***
white	-0.009551	0.017796	-0.537	0.592	
black	-0.018329	0.019363	-0.947	0.345	
asian	-0.002437	0.018108	-0.135	0.893	
hispanic	0.002349	0.003444	0.682	0.496	

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 -0.008945 , which means that a unit change in *women* (1%) will correspond to a -0.008945 change in *phy_lbr*. A less literal and more

useful interpretation requires an understanding that the Boolean in this model represents the “extent to which an industry is classified as physical labor”. Meaning 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 estimate of -0.008945 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 0.89% less likely to require physical labor. These findings correspond with the prediction from Section 1.

Table 3

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	120.2058	74.4627	1.614	0.108	
log(tot_emp)	2.8891	0.7236	3.993	8.45e-05	***
phy_lbr	-16.2032	2.4086	-6.727	1.05e-10	***
white	-1.0393	0.7552	-1.376	0.170	
black	-0.6216	0.8246	-0.754	0.452	
asian	-0.2376	0.7706	-0.308	0.758	
hispanic	-0.1865	0.1463	-1.275	0.203	

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 16.2 decrease in *women* (percentage of women represented in the profession). The most surprising parameter estimate however is the estimate for *ln(tot_emp)* of 2.8891. 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 3% increase in percentage of women in that industry.

Section 3.2 F-Tests

An F-test was conducted on M1 to test the joint-significance of the explanatory variables: log(tot_emp), white, black, asian, and hispanic. The results can be seen in Fig 1 below

Fig 1

```
Linear hypothesis test

Hypothesis:
log(tot_emp) = 0
white = 0
black = 0
asian = 0
hispanic = 0

Model 1: restricted model
Model 2: phy_lbr ~ log(tot_emp) + women + white + black + asian + hispanic

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    272 43.313
2    267 42.289  5    1.0246 1.2937 0.2668
```

Figure 1: Ftest(log(tot_emp), white, black, asian, and hispanic)

The F statistic for the joint restriction that the mentioned 5 variables have a parameter value of zero is 0.2668. This value is relatively large which indicates that in addition to being individually insignificant these variables are jointly insignificant. We fail to reject the hypothesis of joint insignificance. Their insignificance being in the context of the degree to which their values are relevant predictors of the physical labor classification of an industry.

A second F test was done to analyze M2. A linear restriction was placed upon M2 such that asain equals 1. The result is depicted below in Fig 2.

Fig 2

```
Hypothesis:
asian = 1

Model 1: restricted model
Model 2: phy_lbr ~ log(tot_emp) + women + white + black + asian + hispanic

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    268 527.65
2    267  42.29  1    485.36 3064.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Linear restriction hypothesis asain=1

The F statistic for the linear restriction that the “asian” explanatory variable has a parameter value of zero is < 2.2e-16. This is very small. Thus we reject the hypothesis and the linear restriction cannot be reliably applied to this model.

Section 3.3 Homoskedasticity

The Breusch-Pagan test for homoskedasticity was applied to M1 to test the hypothesis that M1 is homoskedastic. A T-test was then done on the data to reveal potentially more accurate error estimates.

```
        studentized Breusch-Pagan test

data:  bls_ols_phylbr
BP = 89.903, df = 6, p-value < 2.2e-16


t test of coefficients:

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.5629014  1.7457196  0.8953    0.3714
log(tot_emp) -0.0011656  0.0167193 -0.0697    0.9445
women        -0.0089448  0.0011714 -7.6361 3.999e-13 ***
white        -0.0095514  0.0176566 -0.5410    0.5890
black        -0.0183292  0.0189069 -0.9694    0.3332
asian        -0.0024367  0.0187636 -0.1299    0.8968
hispanic      0.0023494  0.0042706  0.5501    0.5827
```

It can be observed that the calculated probability of homoskedasticity is 2.2e-16 which is very small. Thus, we reject the null hypothesis of homoskedasticity and must assume heteroskedasticity and

recalculate our error estimates using a T-test. The outcome of the T-test shows that for women, there was *less* std. error than originally estimated, a more negative t-value, and a *higher* statistical significance. These findings only serve to reinforce the conclusions made before the heteroskedacity correction.

When analyzing M2 and testing for homoskedasticity a different conclusion was made. The Breusch-Pagan test and T-test results are available below for reference.

```
        studentized Breusch-Pagan test

data:  bls_ols_women
BP = 27.401, df = 6, p-value = 0.0001218


t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  120.20575   66.94491  1.7956   0.07369 .
log(tot_emp)   2.88905    0.64613  4.4713 1.151e-05 ***
phy_lbr      -16.20324    1.89337 -8.5579 9.171e-16 ***
white         -1.03926    0.66681 -1.5586   0.12028
black         -0.62164    0.77315 -0.8040   0.42210
asian         -0.23757    0.71398 -0.3327   0.73959
hispanic      -0.18647    0.20904 -0.8921   0.37316
```

The p-value result for M2 is significantly larger than the p-value result for M1, thus the null hypothesis cannot be rejected with as much certainty. The probability of homoskedasticity is still small, so homoskedasticity could still be rejected. Rather than focus on the rejection of the hypothesis it would be prudent to consider the implications of it's rejection by analyzing the T-test correction for hetroskedasticity. When correcting for hetroskedasticity, the std error on both statistically significant variables decreases (which supports previously held conceptions), and the statistical significance remains high for both phy_lbr and log(tot_emp) despite the fact that it increases significantly for phy_lbr and decreases slightly for log(tot_emp).

Section 4: Retrospective

This paper observers the relationship between several variables and the degree to which they can be considered reliable predictors for weather or not an industry could be classified as one involving "physical labor". It was discovered that the only such variable in this data set that can reliably predict "physical labor" was gender. Specifically it was discovered that industries with less women were more likely to be classified as involving "physical labor". An additional surprising observation generated by this analysis was that women are less likely to be employed in industries with smaller populations.

A possible explanation for the finding that "industries with less women were more likely to be classified as involving physical labor" could be that physical labor industries discriminate against women. Another possible explanation could be that women apply less to jobs that require physical labor.

A possible explanation for the finding that "women are less likely to be employed in industries with smaller populations" would be much more complex. One could suppose that industries with smaller populations of employees required more specialized skills, and one might then suppose that

most specialized skills involve more risk than generalized skills. It could also be postulated that men are less risk-averse than women and this is what generates this discrepancy.