The Gender Wage Gap in China: Learning from Recent Longitudinal Data

Donghe Lyu *

April 30

Abstract

This paper aims to deliver the most recent estimate of the gender wage gap in China. Using the most recent longitudinal data, we find the gender wage gap in China high and persisting in the 2010s. We further discover that the raw gender wage gap underestimates the degree of discrimination faced by Chinese female workers as the gender wage gap worsens when controlled for labor endowment factors such as education and marital status. These results have important implications for the welfare of working women, who further lead the impact on the demographic composition of Chinese society and long-term economic growth.

*I sincerely thank my thesis advisor, Professor David Card, for his guidance throughout the composition of this paper. This paper would not have been possible without his tremendous support. Any shortcoming remaining is my own.

Contents

1 Introduction

The founding father of the People's Republic of China, Chairman Mao, once famously said: "women hold up half the sky." Under Mao's strict socialist regime and its egalitarian ideals, the wage gap between men and women in China was small. After Mao, Deng's Reform and Opening-up proved to be a significant economic regime change, shifting the country's economy to be market-oriented and setting it on the path to skyrocketing growth. However, studies have found that China's gender wage gap enlarged during the reforms, and consistently so into the early 2000s.

Even though policymakers always emphasize that women should be treated as equal to men, discrimination based on sex is persistent throughout Chinese history. To many, General Secretary Xi's ever-intensifying call for nationalism in the recent years has been an alarming sign of his restorative nostalgia to return to a more socially conservative era. As gender equality becomes the norm of the international community, the spill-over effects of the advocacy for feminism are felt in Chinese society, but they sometimes are labelled as "political correctness" and ridiculed upon. Some men in China believe that women in society are already enjoying the same rights as they are without discrimination; *i.e.*, China has achieved systemic gender equality.

Although preliminary statistics can easily refute such a proposition, it is instructive to illustrate the gender earnings differential and calculate how much of the earnings differential cannot be explained by observed factors. This paper aims to shine a light on the gender wage gap in China using relatively new data. We base our study on the CFPS (China Family Panel Studies) from 2010 to 2016, provided by Peking University. We seek to estimate the gender wage gap with a full set of controls, and conduct decomposition via increasingly complex model specifications. We estimate the gender wage gap as a trend by exploiting the panel data structure and the deployment of sample re-balancing strategy.

2 Literature Review

There is a rich body of literature on measuring the gender wage gap, built within the framework developed by the canonical Blinder (1973) and Oaxaca (1973) to quantitatively assess sources of wage differentials between two groups, from which we can draw reference. Blinder-Oaxaca Decomposition considers two contributing sources to wage differentials: the first is the observed difference in covariates, and the second is the unobserved difference that cannot be explained by the covariates. For example, Oaxaca finds that women on average earn less than men partially because women tend to be employed in industries in which the prevailing wage is low, e.g. hospitality and services. The inspiration for this project comes from Card and Krueger (1992), in which decomposition techniques are applied to study the wage differential between Black and White men between 1960 and 1980.

Calculating an estimate of the gender wage gap in urban China using data from 1988 and 1995, Gustafsson and Li (2000) is the first-ever in a series of efforts to quantitatively study China's gender-based earnings inequality. They use Oaxaca Decomposition to deliver the estimate, and they find that in 1988, women on average earn 15.6% less than men and in 1995 the gap is about 17.5%. About half of the wage gap can be explained by observed factors such as education, age, occupation, type of firm (State-owned or Private), and location within China. However, their estimate has limited internal validity since they use the data from the Urban Household Income Survey, which only includes information from 10 provinces/municipalities and does not provide a nationally representative sample for all the urban population.

There's another sequence of papers that base their estimation of gender wage gap on CHIP (China Household Income Project), which is considerably more nationally representative and contains wage information from 1988 to 2013. Bishop et al. (2005) re-estimates the gender wage gap in urban China in 1988 and 1995 with decomposition and quantile regression. They use quantile regression models because they believe covariates such as education have differential impacts on earning for different levels of earning. Rather than estimating the gender wage gap on aggregate level, their analysis is more focused on how each of the covariates have differential impacts for male and female workers in the urban workforce. Their decomposition shows that the unexplained portion of the gender gap falls from 1988 to 1995: in 1988, about 71% of the gender wage gap is unexplained, compared to 61% in 1995. Their results vary from the result of Gustafsson and Li (2000) because of the different data source and the usage of a different set of controls. Li et al. (2011) again provides the estimates for the trend of gender wage gap from 1995 to 2007 using Oaxaca decomposition and quantile regression, asserting that the unexplained portion of the gender wage gap enlarges during the period, suggesting worsening discrimination. The most up-to-date estimate of the gender wage gap is conducted by Song et al. (2019), in which the gender wage gap of urban workers from 1995 to 2013 is analyzed. However, CHIP data are not collected frequently: the average lag between each panel is about five years. Also, CHIP data are not panel data, therefore cannot provide tracking information for sampled households.

There are also estimates based on other data. Chen and Hamori (2008) uses the data from CHNS (China Health and Nutrition Survey) to estimate the gender wage gap in 2004 and 2006. Su and Heshmati (2012) uses the CFPS data to estimate the wage gap in 2009. However, the CFPS data in 2009 is from a pre-survey: the baseline data for CFPS is in 2010. Our analysis contributes to the body of the literature in the following ways: (1) we use the most recent CFPS data, which is far more frequent and recent than the CHIP data, which dominates the current literature. (2) We use a multinomial logistic regression model to re-balance the sample, ensuring the comparability across all rounds of the longitudinal survey, thus delivering a robust estimate of gender wage gap trend.

Our research is also the first to estimate the gender wage gap in China using data after the end of the One-Child Policy in 2015. Before 2015, a mother in the workforce would not be eligible for any additional maternity leave since women were only allowed to have one child, and such certainty benefited the employers. After 2015, employers no longer enjoy such certainty when they employ new mothers. One would expect the wage gap after 2015 to widen as the result of the new policy. Even though our analysis cannot isolate the causal effect of such policy change, this paper can illustrate the observed difference and comment on the potential impacts of such policy change.

3 Data

3.1 Overview

CFPS provides the most recent nationally representative micro-data in China. CFPS collects data at three levels (individual, family, and community) and is conducted every-two-year starting from the 2010 baseline survey. Drawing random samples from 25 largest provinces and municipalities in China, CFPS aims to be nationally representative and does not apply sample selection criterion *a priori*. At baseline, CFPS includes data for 14,960 households, which includes 33,598 adults and 8,990 youths.

The CFPS data is also relatively new. Compared to other projects (such as CHIP, which began data collection as early as 1988), CFPS only began data collection in 2010, making it effectively one of the most novel micro-level panel data available on the Chinese economy. When compared to international equivalent of longitudinal surveys such as the Panel Study on Income Dynamics in the United States, CFPS is not as consistent between each round of the survey. Some adjustments are required in the analysis, which is detailed in the sample eligibility section. Nevertheless, CFPS is the most extensive academic longitudinal data project with the highest frequency of data collection. With minor adjustments, CFPS is able to yield credible identification of the gender wage gap in China.

The CFPS data is nationally representative, which means that our sample include urban as well as rural-to-urban and rural households. However, our analysis focus mainly on the urban households, since the determination of wage is systemically different in non-urban households. The CFPS data also includes different surveys for adults and children. For the purpose of our analysis, we omit all data in the children survey since we assume children do not earn wage. In practice, determining the sample eligibility requires more attention. For example, not all non-urban individuals hold jobs where wages are not consistently paid, so a blanket omission will likely fail to capture as much representation of the population as possible. The next section details our method of creating a sample eligibility indicator with the complete CFPS data. We then conduct the analysis on the eligible sample.

3.2 Sample Eligibility

The goal of our analysis is to analyze the gender wage gap in China. CFPS is a nationally representative sample, so it is important to notice that an outstanding portion of Chinese population does not hold jobs that pay wages. Specifically, if the sampled individual lives in a rural household, it is likely that they work in some sectors of agriculture, and wages would not be paid. Thus, we need to first establish an eligible sample that is suitable for our analysis. The ideal sample selection should be consistent throughout all rounds of the longitudinal surveys and the selected samples should be comparable. In the panel setting, it is also desirable to focus the analysis solely on specific individuals as we can track their wage path throughout the years. However, due to the frequent inclusions of new cohorts of workers as well as miscellaneous data quality issues, such tracking does not yield robust results as we would expect. Further discussions on consistency and tracking throughout the years is in the robustness check section.

The following decisions are made to ensure the comparability across all rounds of surveys. First, we notice that the hourly wage information is only available at the baseline but not in the subsequent rounds. We then cross verify the data quality with the searchers at Peking

University and decide to run the analysis on *income*, one of the composite variables. This is because the composite variable is highly correlated with the wage information across all levels of the distribution. The calculation for the composite variables ensure that using income in place of the wage information will provide a valid estimate. The *income* composite variables are calculated for all rounds in CFPS and carry comparable information.

Second, we combine *a priori* sample selection with *a posteriori* sample selection. Namely, our sample eligibility is co-determined based on the individual's type of work and his/her income level. The *a priori* sample selection is intuitive as we only wish to study the population currently in the workforce. For all rounds of CFPS, we locate two variables of interest: the first variable is the current employment status, which records whether an individual holds a job or not, and the second variable is the current work category, which records the overall type of the current job, taking three distinct values: self-employed, at a company, or agriculture. We restrict the sample to individuals who currently hold a job and are self-employed or working at a company. Agriculture jobs are usually not associated wage data, thus need to be omitted in our analysis. Then, we choose to include the income level as the *a posteriori* sample selection to circumvent the issue with logarithmic transformation. Also, a portion of our samples are self-employed: the *a posteriori* selection helps to rule out the self-employed workers whose businesses are idle.

3.3 Key Variables

CFPS provides individual level data for most of the standard variables used in labor economics models. As described in the sample eligibility section, this paper uses income as a proxy for wage, the main dependent variable. Education data is provided in years. However, it is generally not a good idea to measure education in years, as the highest degrees obtained by individuals could have differential impacts that cannot be captured in linear models. As such, we impute dummy variables for each level of education based on the highest degree an individual received or expected to have received. It is also notable that a portion of our sample is illiterate, which means they receive 0 year of formal education. Such population is not rare in CFPS sample, which is nationally representative. People above 60 years old at baseline are particularly likely to have 0 year of formal education due to severe disruptions caused by the Civil War prior to the foundation of the People's Republic. In regression models, we choose illiteracy to be the baseline to ensure easiest possible interpretation.

Other variables are either imputed or encoded according to standard analysis. Experience variable refers to "potential experience," which is imputed by $expr = \max\{age - educ - 6, 0\}$. We consider the premium that Chinese pupils are not allowed to enter primary schools before their sixth birthday. Age and gender information are readily available in CFPS. The marital variable takes multiple levels: single, married (having one spouse), cohabitation, divorced, or widowed. We group all individuals with marital history as married to simplify the analysis and increase statistical robustness of identification.

The *hukou* variable, which is special to labor economics analysis in China, also takes multiple values: agricultural, non-agricultural, no registration, or foreign national. In CFPS, the frequency of no registration or foreign national is extremely low (below 1% in our eligible sample). We thus dropped these entries. Note that the hukou status does not necessarily reflect an individual's place of residence at the survey time. For example, most of the rural-to-urban migration workers have agricultural hukou, but their place of residence is urban. Hukou status is best indicative of a person's place of birth. Hukou status also tie closely with social security and other benefits in the Chinese society, as a local hukou would grant priority access to most of the public services such as healthcare, education, vehicle registration, and real estate investments. However, in CFPS, the actual hukou registration (whether the hukou is local to the current place of residence) information is not collected due to privacy concerns. For the current place of residence, we refer to the province variable. This variable carries importance as the prevailing wages in different provinces in China are different due to difference in living costs, analogous to the United States economy. This variable however cannot distinguish whether the individual lives in an urban area or a rural area. The urban versus rural indicator is provided by CFPS, but it carries little power in the analysis as most of our eligible sample comes from urban background. Thus, we only control hukou and province in our analysis.

Two other standard variables, industry and firm category, are collected in CFPS. However, we choose to omit these from the initial analysis due to somewhat poor data quality. In CFPS, the industry variable is a factor with 28 distinctive value, each representing a specific sector of jobs. However, in 2012, the work module in the CFPS survey is divided into two versions, making the industry variable corrupted for that year. The firm category variable in CFPS is specific to the Chinese economy. It encodes information regarding the type of firm for individuals with a job at a specific company, whether it is a state owned enterprise (SOE), a private Chinese company, a private multinational company, or self-employment. The complete encoding is much more granular. This variable suffers from the same problem in 2012, rendering the usage infeasible. The granular nature of these two variables also pose threat to robustness as they greatly increase the model complexity. Thus, in initial analysis, we assume that the impacts on gender wage gap based on industry and occupation are partially absorbed by other factors.

3.4 Sample Characteristics

One of the most important covariates in our analysis is education. As mentioned above, the education data is provided in years in CFPS, and we segregate the education to respective highest degree to deliver a more precise analysis. Table 1 records the mean characteristics for each education dummy variable in the eligible sample, for each round of the survey and two genders.

The codebook for the education dummy variables is in the Appendix. From Table 1, we

Table 1: Mean Characteristics of Education, unweighted

	illiterate	primary	middle	high	associate	college	graduate
2010 Male	0.047	0.119	0.390	0.237	0.126	0.075	0.007
2010 Female	0.052	0.113	0.326	0.234	0.166	0.103	0.006
2012 Male	0.135	0.258	0.343	0.166	0.059	0.037	0.002
2012 Female	0.173	0.217	0.295	0.177	0.093	0.042	0.003
2014 Male	0.045	0.129	0.411	0.216	0.115	0.078	0.006
2014 Female	0.071	0.115	0.325	0.222	0.161	0.098	0.008
2016 Male	0.039	0.109	0.370	0.226	0.140	0.103	0.013
2016 Female	0.071	0.115	0.325	0.222	0.161	0.098	0.008

can see that despite the fact that CFPS is designed as a longitudinal survey, the eligible samples suffer imbalances across each round of the survey. This is due to the fact that CFPS is relatively new and the data collection, as well as the survey design, is not yet stabilized and able to yield consistent results. To circumvent this issue, two approaches are considered by this paper. The first approach is to exploit the nature of panel dataset and restrict our sample further to individuals that are in the eligible sample for all rounds of the survey. However, this approach fails to stay representative of the gender wage gap in China after the baseline as it innately omits all new cohorts of workers in subsequent rounds of the survey. Thus, we only include it as a robustness check. The second approach is to re-balance the eligible sample so that the characteristics are comparable across rounds. We attempt this via a multinomial logistic regression model. The details are discussed in section 4.2.

The other notable characteristics is the distribution of places of residence, which is indicated by the *province* variable. Since there are 25 provinces in the eligible sample, we choose to visualize the distribution using barplots rather than a table. Note that the province here stands for any provincial identity, which includes provinces, municipalities, and autonomous regions. China has 32 such regions in total (not counting 2 special administrative regions) and CFPS omits sampling from the least populous regions. The visualization is included in the Appendix. Note that some individuals in the panel move between rounds of the survey, and a selected few (less than 0.1% of the sample size) will locate to a region that is not sampled in CFPS. We consider these entries as outliers and drop these observations to eliminate their influence.

4 Models

4.1 Ordinary Least Squares

We consider this classical model that follows Mincer in explaining wage with province fixed effects:

$$\begin{aligned}\ln(income_i) = \beta_0 + \beta_1 Male_i \\ + \beta_2\ primary_i + \beta_3\ middle_i + \beta_4\ high_i + \beta_5\ associate_i + \beta_6\ college_i + \beta_7\ graduate_i \\ + \beta_8\ expr_i + \beta_9\ expr_i^2 + \beta_{10}\ age_i + \beta_{11}\ age_i^2 \\ + \beta_{12}\ hukou_i + \beta_{13}\ married_i + \sum_{p=1}^{24} \gamma_p\ prov_{p,i} + u_i \end{aligned} \quad (1)$$

Note that we leave one dummy variable out in the province fixed effects to avoid co-linearity. We estimate this model (Equation 1) for all four rounds of our data. We do not use a pooled OLS model here because we do not assume that covariates such as education have consistent influences on individual's income across all years.

Our parameter of interest in Equation 1 is β_1, the coefficient associated with the gender dummy $Male_i$. We can interpret this as the male wage premium in percentage points. It is important to note that the negative of these point estimates do not reflect how much on average a female worker earns less than her male counterpart. To get the conversion, we need to use:

$$\hat{\beta}_{Female} = \frac{\hat{\beta}_{Male}}{1 + \hat{\beta}_{Male}} \quad (2)$$

where $\hat{\beta}_{Male}$ is our estimated β_1 associated with $Male_i$ in Equation 1, and $\hat{\beta}_{Female}$ is the estimated coefficient should β_1 in Equation 1 is associated with $Female_i$ instead of $Male_i$. Both $\hat{\beta}_{Male}$ and $\hat{\beta}_{Female}$ will be reported in our discussion of results.

4.2 Re-balancing: Multinomial Logit and Weighted Least Squares

As mentioned in section 3.4, our data suffers from sample imbalances across rounds of the longitudinal survey. To ensure comparability across all four rounds of our data, we choose to use a weighted least squares regression with weights constructed from a multinomial logistic regression model to re-balance the samples.

The estimation of the multinomial logistic regression model proceeds as the following. We first pool observations in the eligible samples from all rounds of the survey together. Then, given one observation, the multinomial logistic regression model will give us the predicted probability of this observation comes from year 2010, 2012, 2014, and 2016. The actual specification is as follows:

$$
\begin{aligned}
\ln\left(\frac{P(t_j = y)}{P(t_j = 2010)}\right) &= b_{0,y} + b_{1,y} Male_j \qquad \forall y \in \{2012, 2014, 2016\} \\
&+ b_{2,y} primary_j + b_{3,y} middle_j + b_{4,y} high_j + b_{5,y} associate_j + b_{6,y} college_j + b_{7,y} graduate_j \\
&+ \eta_{1,y} Male_j \quad high_j + \eta_{2,y} Male_j \quad associate_j + \eta_{3,y} Male_j \quad college_j + \eta_{4,y} Male_j \quad graduate_j \\
&+ b_{8,y}\ expr_j + b_{9,y}\ expr_j^2 + b_{10,y}\ age_j + b_{11,y}\ age_j^2 \\
&+ b_{12,y}\ hukou_j + b_{13,y}\ married_j + \sum_{p=1}^{24} \zeta_{p,y}\ prov_{p,j} + \varepsilon_j \qquad (3)
\end{aligned}
$$

where y represents $year$, $b_{\ ,y}$ are coefficients associated with the main effects of the covariates in Equation 1, and $\eta_{\ ,y}$ are coefficients associated with the marginal effects of interaction terms between gender and education dummies. We include the marginal effects because the sample imbalance occur differentially between gender and across years according to our observation of Table 1. This model (Equation 3) essentially predicts the "log-odds" for each subsequent round of the survey against the baseline using a linear model. Note that we use index j here to differentiate from i in Equation 1 as the estimation happens with the pooled data rather than each year's data. The province fixed effects are also present in Equation 3 as they capture valuable characteristics for the eligible sample.

With the estimation of Equation 3 for all subsequent rounds of the survey, we can use the predicted probabilities and construct the regression weights as:

$$w_{i,t} = \frac{\hat{P}(year = 2010)}{\hat{P}(year = t)} \tag{4}$$

The construction in Equation 4 essentially forces all observations to resemble the baseline data, ensuring credible comparison across all rounds of the survey. Aside from the weighted regression, these weights can also be used to construct weighted mean characteristics for the sample. Table 2 below records the education characteristics computed using the multinomial logit weights.

Table 2: Mean Characteristics of Education, weighted

	illiterate	primary	middle	high	associate	college	graduate
2010 Male	0.047	0.119	0.390	0.237	0.126	0.075	0.007
2010 Female	0.052	0.113	0.326	0.234	0.166	0.103	0.006
2012 Male	0.046	0.126	0.390	0.230	0.114	0.090	0.004
2012 Female	0.050	0.093	0.346	0.240	0.167	0.098	0.006
2014 Male	0.037	0.112	0.391	0.239	0.130	0.078	0.014
2014 Female	0.056	0.118	0.318	0.243	0.154	0.102	0.009
2016 Male	0.031	0.117	0.366	0.322	0.097	0.055	0.012
2016 Female	0.056	0.118	0.318	0.243	0.154	0.102	0.009

Compared to unweighted characteristics, the weighted characteristics are significantly more balanced across each round of the survey. The most notable change is the percent of illiterate individuals in the eligible sample in 2012: after weighting, the percent decreases from 0.135 and 0.173 to 0.046 and 0.050 respectively for male and female. Similar results are observed across all sample covariate characteristics, included in the Appendix.

4.3 Decomposition

The most prominent decomposition method used in the analysis of gender wage gap is Blinder (1973) and Oaxaca (1973), where the regression is conducted for both male and female, and the average wage differential is decomposed into the "explained" part using regression coefficients of one group and the "unexplained" part weighted by the mean characteristics of the other group. While this method is canonical and able to yield direct results, the term "decomposition" is misleading. For Blinder-Oaxaca decomposition to hold, we generally assume that the wage differential between male and female is greater when we ignore control factors such as education and potential experience, which innately assumes that female workers have less endowment when compared against male workers. While such assumption is realistic in 1973, it no longer holds in 2010s.

From Table 1 and Table 2, we can see that in our sample female workers are better educated than male workers, suggesting that their endowments are in fact higher than their male counterparts. This fact is partially contributed to the sample selection process: less educated women are less likely to work. As such, Blinder-Oaxaca decomposition will likely yield misleading results. We compensate this by estimating models with increasing complexity as our main approach to decomposition, detailed in section 5.3.

5 Main Results

5.1 Regression Tables

Table 3 and Table 4 (page 15, 16) record our estimates of the OLS model using Equation 1 and the WLS model using weights from Equation 3, respectively. Note that the constant terms are omitted since they do not provide much information when the dependent variable is logarithmically transformed. The omission is also justified due to our inclusion of province fixed effects.

Table 3: OLS with Province Fixed Effects (Equation 1)

	Dependent variable:			
	ln(income)			
	(2010)	(2012)	(2014)	(2016)
Male	0.384	0.463	0.481	0.428
	(0.017)	(0.024)	(0.025)	(0.030)
primary school	0.141	0.194	0.144	0.152
	(0.042)	(0.038)	(0.061)	(0.084)
middle school	0.303	0.342	0.256	0.230
	(0.039)	(0.038)	(0.057)	(0.077)
high school	0.411	0.408	0.291	0.315
	(0.041)	(0.044)	(0.060)	(0.082)
associate	0.731	0.584	0.492	0.520
	(0.045)	(0.058)	(0.066)	(0.087)
college	1.000	0.722	0.647	0.742
	(0.049)	(0.071)	(0.072)	(0.091)
graduate	1.359	1.506	0.768	1.111
	(0.105)	(0.231)	(0.158)	(0.151)
experience	0.032	0.059	0.035	0.069
	(0.003)	(0.005)	(0.006)	(0.008)
experience^2	0.001	0.002	0.001	0.002
	(0.0001)	(0.0002)	(0.0002)	(0.0003)
age	0.052	0.002	0.156	0.084
	(0.006)	(0.006)	(0.007)	(0.009)
age^2	0.001	0.0002	0.002	0.001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
hukou: Non-agricultural	0.080	0.042	0.072	0.012
	(0.020)	(0.031)	(0.030)	(0.037)
Married	0.229	0.097	0.008	0.097
	(0.031)	(0.039)	(0.041)	(0.046)
Observations	7,078	6,618	8,000	4,113
R^2	0.358	0.215	0.257	0.200
Adjusted R^2	0.354	0.210	0.253	0.192
Residual Std. Error	0.658	0.905	1.054	0.940
F Statistic	97.911	45.103	68.830	25.433

$p<0.1$; $p<0.05$; $p<0.01$

Table 4: WLS with Province Fixed Effects

	Dependent variable:			
	ln(income)			
	(2010)	(2012)	(2014)	(2016)
Male	0.384	0.306	0.252	0.277
	(0.017)	(0.024)	(0.026)	(0.033)
primary school	0.141	0.140	0.031	0.187
	(0.042)	(0.061)	(0.068)	(0.081)
middle school	0.303	0.332	0.0001	0.160
	(0.039)	(0.055)	(0.063)	(0.075)
high school	0.411	0.328	0.033	0.176
	(0.041)	(0.058)	(0.066)	(0.079)
associate	0.731	0.603	0.139	0.025
	(0.045)	(0.064)	(0.072)	(0.088)
college	1.000	0.399	0.238	0.542
	(0.049)	(0.069)	(0.077)	(0.090)
graduate	1.359	1.343	0.137	0.715
	(0.105)	(0.166)	(0.130)	(0.163)
experience	0.032	0.050	0.010	0.025
	(0.003)	(0.004)	(0.004)	(0.006)
experience^2	0.001	0.002	0.0003	0.001
	(0.0001)	(0.0001)	(0.0001)	(0.0002)
age	0.052	0.020	0.189	0.024
	(0.006)	(0.008)	(0.008)	(0.011)
age^2	0.001	0.0001	0.003	0.0004
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
hukou: Non-agricultural	0.080	0.051	0.077	0.068
	(0.020)	(0.028)	(0.030)	(0.037)
Married	0.229	0.158	0.039	0.488
	(0.031)	(0.041)	(0.046)	(0.064)
Observations	7,078	6,618	8,000	4,113
R^2	0.358	0.149	0.218	0.317
Adjusted R^2	0.354	0.144	0.214	0.310
Residual Std. Error	0.658	0.947	1.008	1.595
F Statistic	97.911	28.886	55.581	47.228

$p<0.1$; $p<0.05$; $p<0.01$

5.2 Interpretation

5.2.1 Gender Wage Gap and Comparison with the Current Literature

The main coefficient of interest is the coefficient of $Male$. Since our dependent variable is income in its logarithmic form, the coefficient directly shows that on average how much more in percentage point a male worker earns than a female worker, controlled for our set of covariates. We can see that the gender wage gap is consistently high in China. Based on the OLS results, a male worker on average earns 38.4% more than his female counterpart in 2010, controlled for education, potential experience, age, *hukou* status, marital status, as well as the fixed effects on province location. The gender wage gap grows to 46.3% in 2012 and 48.1% in 2014, and wanes slightly to 42.8% in 2016. Based on the conversion from Equation 2, the OLS model yields the gender wage gap (female to male) estimates at -27.7%, -31.6%, -32.5%, and -30.0% for year 2010, 2012, 2014, and 2016 respectively.

The WLS model gives significantly different results. Choosing year 2010 as our baseline, the re-balanced samples in subsequent rounds indicate the male wage premium to be at 30.6%, 25.2%, and 27.7% for year 2012, 2014, and 2016 respectively. Through the conversion from Equation 2, the WLS model yields the gender wage gap (female to male) estimate at -27.7%, -23.4%, -20.1% and -21.7% for year 2010, 2012, 2014, and 2016 respectively.

We compare our point estimates for the gender wage gap with the current literature. Song et al. (2019), calculated the gender wage gap in 2013 to be -21.3%, which is slightly lower than our OLS estimate in magnitude, and about the same as our WLS estimate in magnitude. Other current literature unfortunately lacks comparability due to the differences in the period under study. The biggest possible source of contribution to the difference between our OLS estimates and the estimates by Song et al. (2019) is the difference between the data. Song et al. (2019) uses the data from CHIP, which is infrequent and gears towards the urban population in China when compared against CFPS. The technical papers associated with CFPS contrast the income information between the two studies. Researchers in Peking University finds that the income distribution in the CHIP sample is systematically higher than the income distribution in the CFPS sample (Xie and Hu, 2014). Both studies aim to be nationally representative, however, the different sampling design yields samples with different characteristics. The individuals in CHIP are more likely to come from urban China background and somewhat younger, since CFPS employs stratified sampling based on households rather than individuals. Among the younger population, the gender wage gap should be lower as there shouldn't be as extreme discrimination for the entry level jobs, especially in high human capital service industries.

The above rationale also partially explains why our WLS estimates are similar to the estimates in current literature. Song et al. (2019) estimates the gender wage gap in 2007 to be -25.3%, which is not statistically different from our estimate in 2010 at -27.7%. As mentioned in section 3, our data quality is not perfect. Despite the fact CHIP is collected infrequently and gears towards the urban population, CHIP is more mature when compared with CFPS, commencing in 1988 versus CFPS commencing in 2010. As the sample eligibility is co-determined with employment status and income level, our re-balancing strategy would force the samples in subsequent rounds to match the characteristics at the baseline, reducing the expressed gender wage gap in the regression model. Thus by simply re-balancing our eligible sample to resemble the characteristics at baseline, the estimates will inevitably fall in line better with the estimates in 2010. This also requires extra attention when interpreting the estimates using the WLS model, which will be further elaborated in the robustness check section.

Considering the fact that the sample is imbalanced, we think that the estimates in the WLS model give a better representation in estimating the trend of gender wage gap in 2010s, whereas the estimates in the OLS model are better interpreted as macro-level labor statistics. To elaborate, the estimates using the WLS model is best understood as the evolution of the gender wage gap starting from 2010 to 2016 for a cohort of workers with similar characteristics. The estimates using the OLS model is best understood as the gender wage gap estimated using the data from each round as is, ignoring the potential changes in the characteristics of the working population. Such changes are not insignificant due to the frequent addition of new workers as well as natural retirements, as mentioned in section 3.

5.2.2 Endowment Factors

This section discusses the endowment factors in the models. First, we examine the education. As mentioned in section 3, this paper discretizes the education and does not adopt a stepwise approach, *i.e.*, the coefficients associated with education dummy variables reflect how much more, in percentage point, an educated worker with the corresponding degree earn than an illiterate (no formal education) individual, controlled for gender, potential experience, etc. The coefficients behave according to expectation in the OLS model, as higher education gives higher wage premium. It is worth noticing that education premium seems to decrease across the tenor of this study, which is intuitive given the rise of the popularity of higher education in China. Note that the coefficients are lower for round 2014 when compared against other rounds. This is partially due to the data quality: the questionnaire associated with our data changes significantly in 2014, coupled with the influx of new cohorts and the retirement of old cohorts, we acknowledge that these point estimates associated with education in 2014 might be misleading. The robustness check section addresses this problem. This issue is more evident in our estimates using the WLS model. As we force the subsequent rounds to resemble the baseline, the coefficients associated with education in 2014 loses significance and hence interpretability. However, such loss should not discredit our point estimates as the re-weighting via propensity (Equation 4) does not structurally change the data, as relative information in each data entry is preserved: they get weighted by the same weight.

Second, we notice that both the potential experience and the age are positively associated with wage in the OLS model. They also exhibit diminishing returns to wage as well. Since these variables are measured in years, the scale of the parameters do not appear to be extremely large in absolute terms. However, the standard errors associated with these point estimates are extremely small, giving high confidence inline with the classical labor economics hypotheses. This is extended to WLS model as well, although there are some years in which the coefficients are flipped in direction or insignificant. The most notable deviation is in year 2012 with the *age* variable. This is partially due to the particular data structure in 2012, where two separate work modules are provided in the data collection process, which renders some entries missing. However, it is again safe to assume that *experience*, which is imputed from age and education, is able to absorb some of the influence from age, ensuring the interpretability of the point estimates of the gender wage gap.

Third, we notice that for rounds where the variable is statistically significant, non-agricultural *hukou* and being married both induces positive wage premium. Non-agricultural *hukou* indicates either (1) this individual comes from an urban background (2) this individual has successfully migrated to an urban area, both of which strong indicators of earning higher wage. Being married is considered being "stable" in Chinese society, which also contributes to wage positively.

5.3 Results with Increasing Model Complexity

As mentioned in section 4.3, Blinder-Oaxaca decomposition will yield misleading results as female workers earn less than male workers despite having better endowments. To compensate for this fact, we choose to estimate the gender wage gap with models of increasing complexity, and observe how new additions of control variables impact the male wage premium. We first estimate the gender wage gap (as male wage premium) using no covariates, which is simply the difference in average wage between the two groups. Then, we estimate the gender wage gap using only the education dummies as control. Then, we estimate the gender wage gap using education dummies and both linear and quadratic terms of potential experience and age. We report all the three estimates along with our estimates using the original specification in Equation 1, *i.e.*, adding *hukou* and martial status as well as province fixed effects. Figure 1 below illustrates these estimates. The left panel is the regression estimates using OLS model and the right panel is for the WLS model.

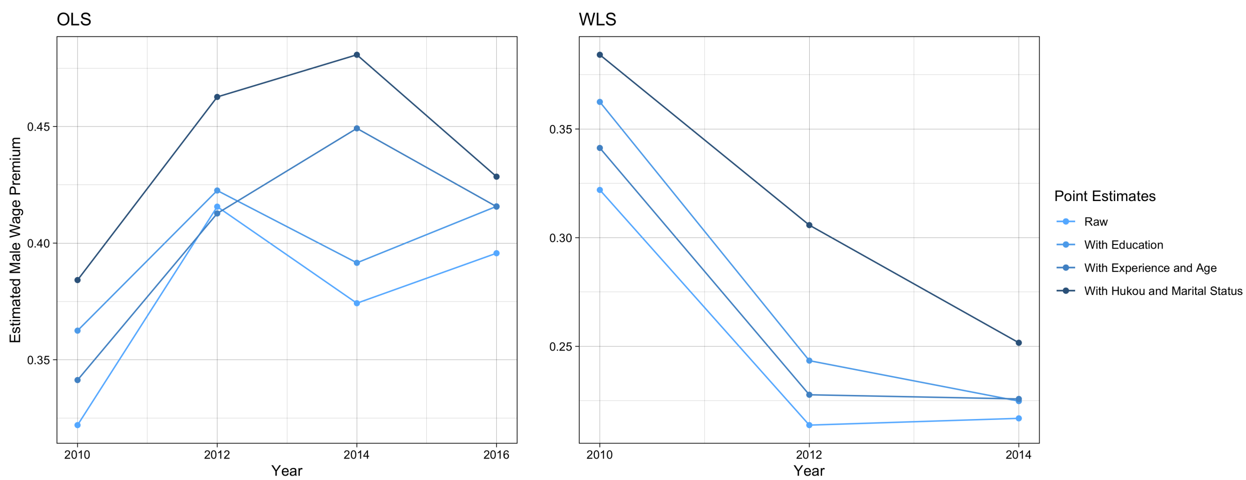

Figure 1: Estimated gender wage gap as male wage premium using models with increasing complexity, for both OLS and WLS model. Note that the province fixed effects are only controlled in the final model.

Note that we do not report the estimates in year 2016 for the WLS data due to data quality concerns that makes re-weighting misleading without full panel of control variables. Specifically, the coefficients associated with province fixed effects, hukou and martial status are significant in the multinomial logistic regression model (Equation 3) when predicting the log-odds for any given entry to be in round 2016. Such significance implies that leaving out the covariates while using the weights constructed using the same covariates will likely lead to misleading results.

We first examine the right panel in Figure 1, the estimates using WLS. As mentioned in section 5.2.1, these estimates are best understood as to reflect the evolution of the gender wage gap starting from 2010 to 2014 for a cohort of workers with similar characteristics. The increasing complexity is reflected by the darker lines in the plot. We see that the inclusion of education dummies variables actually increases the estimated male wage premium, which matches our observation that female workers have higher education than male workers in our eligible sample. The inclusion of experience and age decreases the perceived gender wage gap, and the inclusion of *hukou* and marital status, coupled with province fixed effects, widens the gap again.

The left panel in Figure 1, which is estimated using OLS, shows the similar result. The inclusion of education controls enlarges the gender wage gap, whereas the experience and age variables pull back the male wage premium. *Hukou* and marital status, along with province fixed effects, enlarges the gap again. This is expected as the selection process when determining sample eligibility favors more economically developed areas, and women in less developed areas are less likely to work. It is important to note that the results in this section appear to contradict the findings in other literature. We attribute this to an actual change in the population. The return to education is diminishing in recent years, and our analysis shows that the return to education is even worse for women towards the higher end of education. The same story goes for location indicators. Even if a female individual comes from urban background and is considered stable in her career, a male worker with similar background earns much more.

To conclude, our analysis shows that women are more discriminated than the average wage gap would suggest. The observed raw gender wage gap appears small due to the fact that female workers have better endowments than male workers. Controlled for endowment factors, the gender wage gap widens.

6 Robustness Checks

This section discusses the consistency of estimation and tracking throughout rounds of the longitudinal study. As mentioned in section 3.2, in panel setting, we are able to focus the analysis solely on specific individuals as we can track their wage path throughout the years. This will deliver an estimate of the evolution of the gender wage gap that is most accurate to that particular cohort. The problems with such analysis is also clear: by focusing the analysis on one specific cohort, the estimated wage gap would not extend to the entire population. This paper uses the methods discussed in section 4.2 to deliver estimates simultaneously accurate to represent the entire population and robust to be interpreted as a trend. The estimation via cohort tracking is presented here as a check.

We determine the cohort via joining together the four eligible samples. In joining the samples, we discover that the the data from 2016 should not be included due to a significant change in the data structure that impacts the size of the traceable cohort, thus omitted. We also discover that the data records the employment indicators erratically. We compensate this by looking back at the full sample and aims to extract a cohort from the full sample. We finally choose a cohort that satisfies the following conditions: individuals should be employed, either by a company or self-employed, for all rounds of the survey up to and including year 2014. They must also earn a positive income in year 2014 (only), as we are more interested in recent estimates rather than old estimates. Interestingly, for this cohort, the income information seems be missing or inaccurately recorded in years preceding 2014. It is required for the purpose of statistical robustness to drop these entries, and that leads to unequal number of observations across all years.

Table 5 (page 24) records the OLS estimate with this specific cohort from 2010 to 2014. Note that even in this analysis we don't have the same number of observations across all three years. As mentioned above, this is because some of the income information is missing in this cohort, thus need to be omitted from the regression.

Table 5: OLS estimates with a specific cohort starting from baseline

	Dependent variable:		
	log(income)		
	(1)	(2)	(3)
Male	0.492	0.381	0.577
	(0.033)	(0.032)	(0.042)
primary school	0.124	0.221	0.250
	(0.072)	(0.055)	(0.082)
middle school	0.457	0.298	0.504
	(0.068)	(0.052)	(0.075)
high school	0.738	0.411	0.709
	(0.072)	(0.056)	(0.081)
associate	1.218	0.661	0.988
	(0.080)	(0.066)	(0.097)
college	1.428	0.665	1.262
	(0.090)	(0.077)	(0.112)
graduate	1.537	1.346	1.326
	(0.243)	(0.226)	(0.324)
experience	0.016	0.033	0.057
	(0.004)	(0.005)	(0.009)
experience^2	0.001	0.001	0.001
	(0.0001)	(0.0002)	(0.0002)
age	0.095	0.003	0.065
	(0.010)	(0.009)	(0.010)
age^2	0.001	0.0002	0.001
	(0.0001)	(0.0001)	(0.0001)
Observations	3,005	3,286	4,392
R^2	0.360	0.234	0.435
Adjusted R^2	0.352	0.226	0.431
Residual Std. Error	0.839	0.843	1.334
F Statistic	47.665	27.624	93.235
Note:			$p<0.1$; $p<0.05$; $p<0.01$

Notice that the point estimates are somewhat higher than our original estimates. We contribute this mainly to random error. According to this estimate, the male wage premium is 49.2%, 38.1%, and 57.7% respectively for year 2010, 2012 and 2014. Using the transformation in Equation 2, that puts the gender wage gap (female to male) at -32.9%, -27.5%, and -36.5% for 2010, 2012 and 2014 respectively.

7 Conclusion

Using ordinary least squares estimation, we find that the gender wag gaps (female to male) in China are -27.7%, -31.6%, -32.5%, and -30.0% for year 2010, 2012, 2014, and 2016 respectively. We deliver the estimates using data from the China Family Panel Studies, conducted by researchers in Peking University. To validate our results and further understands the evolution of gender wage gap, we construct weights from a multinomial logistic regression model and adopt weighted least squares estimation to circumvent the sample imbalance problem. The weighted regression gives the gender wage gaps at -27.7%, -23.4%, -20.1% and -21.7% for year 2010, 2012, 2014, and 2016 respectively. Our estimates are inline with the mainstream literature (results up to 2013), and suggests that the gender wage gap still appears large in the modern Chinese society. Further decomposition shows that women are even more discriminated in the labor market, as the wage gap widens when controlled for endowment factors. Despite having better endowments, female workers earn significantly less than male workers, suggesting strong evidence of systematic gender discrimination.

We conclude by discussing some of the limitations of this study and potential future work. The main limitation of this study is of the data source. Broad longitudinal surveys are rare in China. CFPS is a great initiative, but it is not yet mature enough for robust analysis without much adjustments. It is also worth noticing that the trend of labor economics research focus more and more heavily on credible identification over what factors are related with gender and impact the gender wage gap over the course of an individual's career. Specifically, studies such as Bertrand et al. (2010) are also be identify gender specific factors that impact the widening of earnings gap. Considering the fact that women have on average higher education than men (if participating in the work force), it will be particularly illustrative to conduct a similar study in the Chinese economy, where a specific cohort of highly educated workers can be tracked over an extended period of time. Other studies focus more on major events occurred during individual's career, such as the birth of the first child. For example, Cortés and Pan (2020) uses event time of the arrival of the first child to study the impact of child bearing to wage. Such analysis will be particularly valuable in China, as the perceived child penalty is high for female workers. This paper cannot re-create the analysis due to the fact that we only have 4 rounds of data. With the arrival of new data, event time studies will become feasible.

References

[1] Bishop, John A., Feijun Luo, and Fang Wang. "Economic Transition, Gender Bias, and the Distribution of Earnings in China." Economics of Transition and Institutional Change 13, no. 2 (2005): 239–59. https://doi.org/10.1111/j.1468-0351.2005.00218.x.

[2] Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz. 2010. "Dynamics of the Gender Gap for Young Professionals in the Financial and Corporate Sectors." American Economic Journal: Applied Economics, 2 (3): 228-55. DOI: 10.1257/app.2.3.228

[3] Blinder, Alan S. "Wage Discrimination: Reduced Form and Structural Estimates." The Journal of Human Resources 8, no. 4 (1973): 436–55. https://doi.org/10.2307/144855.

[4] Card, David, and Alan B. Krueger. "School Quality and Black-White Relative Earnings: A Direct Assessment." The Quarterly Journal of Economics 107, no. 1 (1992): 151–200. https://doi.org/10.2307/2118326.

[5] Chen, Guifu, and Shigeyuki Hamori. "An Empirical Analysis of Gender Wage Differentials in Urban China." Kobe University Economic Review 54 (January 1, 2008). https://doi.org/10.1007/978-3-642-41109-0_5.

[6] Cortés, Patricia, and Jessica Pan. Children and the Remaining Gender Gaps in the Labor Market. NBER Working Paper Series, no. w27980. Cambridge, Mass: National Bureau of Economic Research, 2020.

[7] Gustafsson, Björn, and Shi Li. "Economic Transformation and the Gender Earnings Gap in Urban China." Journal of Population Economics 13, no. 2 (July 1, 2000): 305–29. https://doi.org/10.1007/s001480050140.

[8] Oaxaca, Ronald. "Male-Female Wage Differentials in Urban Labor Markets." International Economic Review 14, no. 3 (1973): 693–709. https://doi.org/10.2307/2525981.

[9] Shi, Li, Song Jin, and Liu Xiaochuan. "Evolution of the Gender Wage Gap among China's Urban Employees." Social Sciences in China 32, no. 3 (August 1, 2011): 161–80. https://doi.org/10.1080/02529203.2011.598307.

[10] Song, Jin, Terry Sicular, and Björn Gustafsson. "China's Urban Gender Wage Gap: A New Direction?" Working Paper. CHCP Working Paper, 2017. https://www.econstor.eu/handle/10419/180868.

[11] Su, Biwei, and Almas Heshmati. "Analysis of Gender Wage Differential in China's Urban Labor Market." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 8, 2012. https://papers.ssrn.com/abstract=1981208.

[12] Xie, Yu, and Jingwei Hu. "An Introduction to the China Family Panel Studies (CFPS)." Chinese Sociological Review 47, no. 1 (September 1, 2014): 3–29. https://doi.org/10.2753/CSA2162-0555470101.2014.11082908.

[13] Zhang, Junsen, Jun Han, Pak-Wai Liu, and Yaohui Zhao. "Trends in the Gender Earnings Differential in Urban China, 1988–2004." ILR Review 61, no. 2 (January 1, 2008): 224–43. https://doi.org/10.1177/001979390806100205.

Appendix

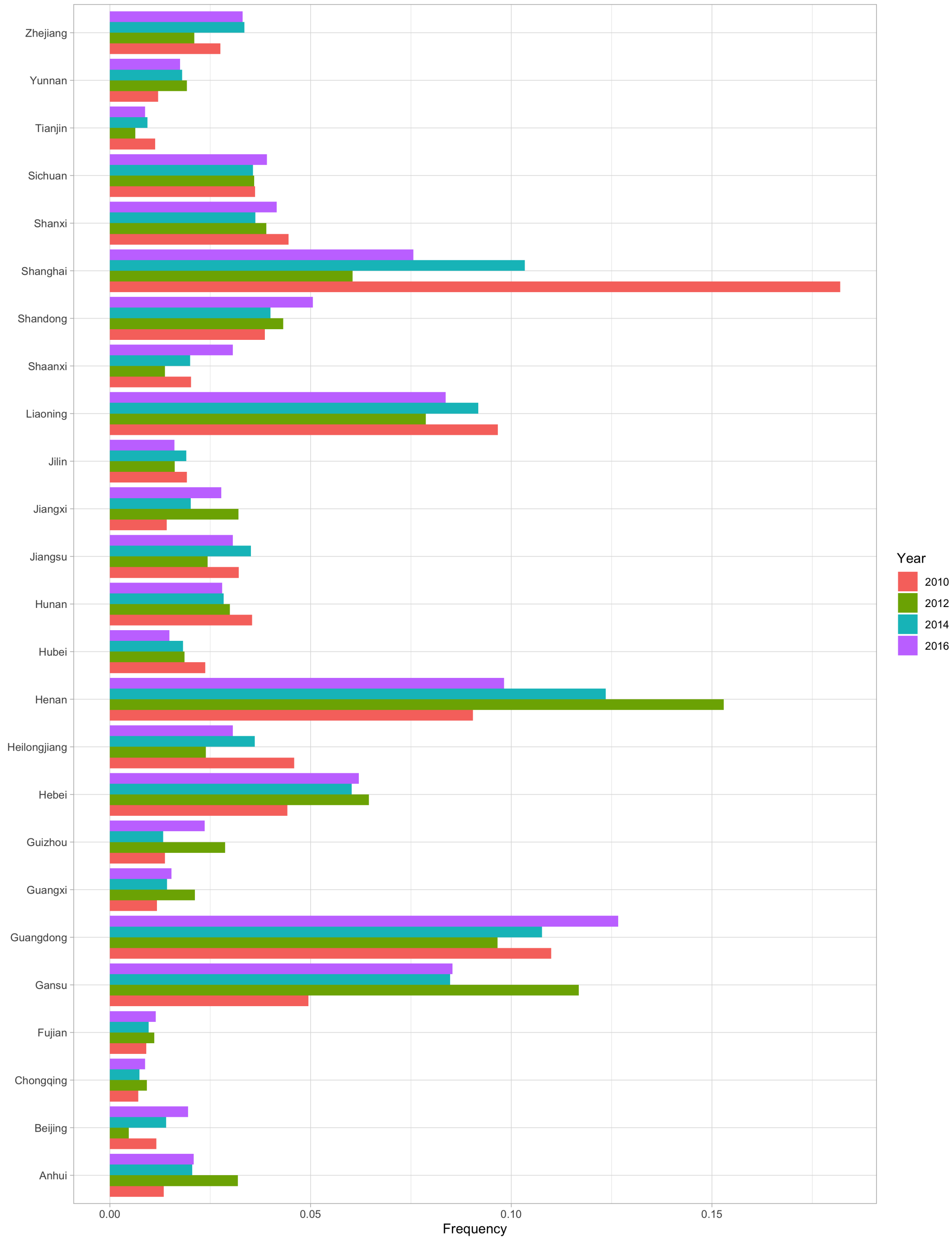

Figure 2: The distribution of province locations in the eligible samples.

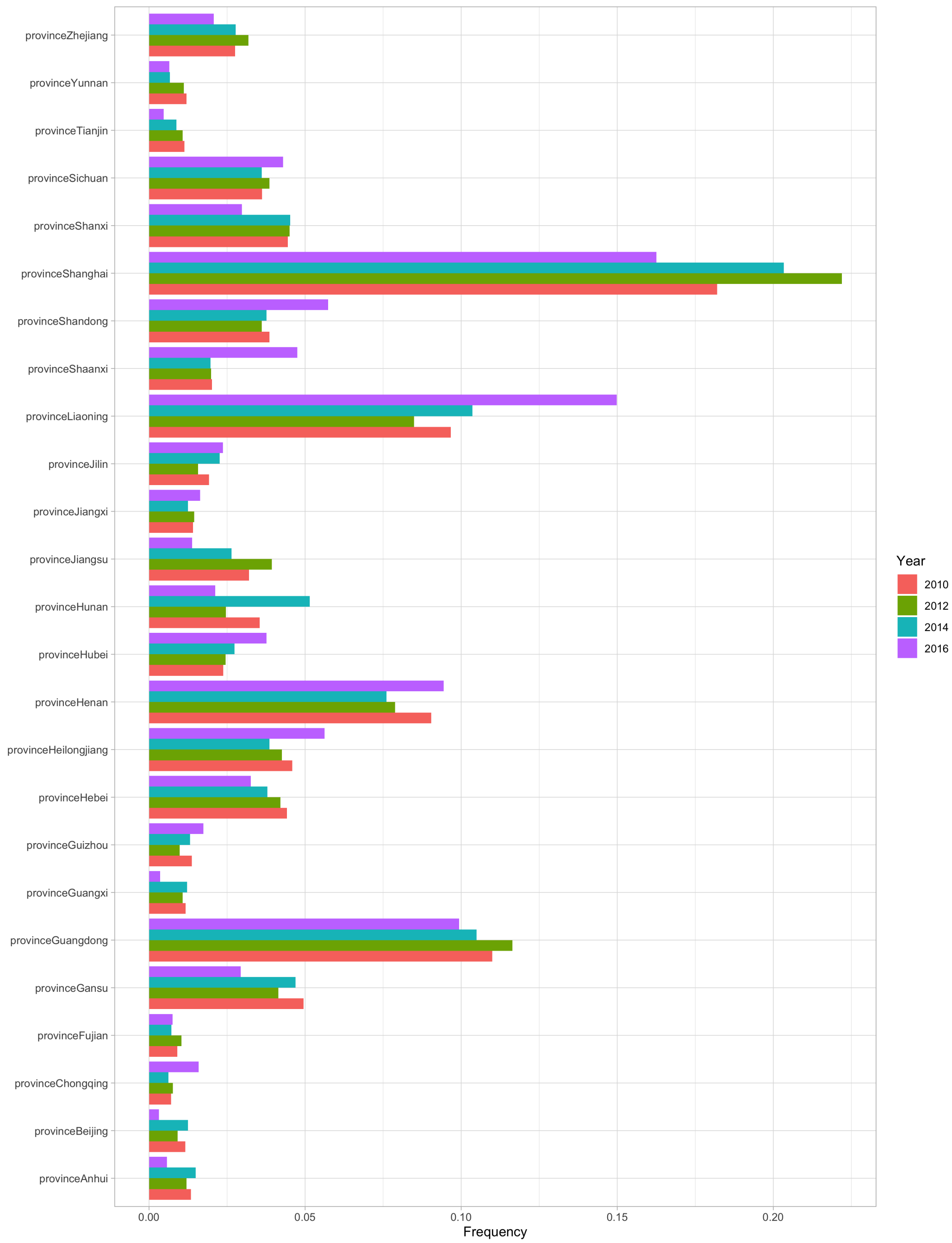

Figure 3: The distribution of province locations in the eligible samples, re-balanced using weights constructed from Equation 4. We can see that the characteristics are significantly more balanced.