Online Inference with Stochastic Gradient Descent: Application to the Tobit Model

Thu Pham*

Abstract

This paper applies the online inference algorithm with stochastic gradient descent developed in Lee et al. (2022b) to the case of censored data. Specifically, I show how the algorithm enhances computation speed relative to the standard maximum likelihood estimation. As an illustration, I employ this algorithm to estimate the college wage premium in an empirical analysis.

1 Introduction

In statistics, inference is a crucial research area that involves estimating model parameters by optimizing a convex objective function. Lee et al. (2021) defines a general inference problem as

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^d} Q(\beta) \tag{1}$$

where $Q : \mathbb{R}^d \to \mathbb{R}$ is convex and defined as $Q(\beta) = \mathbb{E}_{Y \sim \Pi}[q(\beta, Y)]$ with Y being a random sample from a distribution Π.

*Under advisement of Sokbae "Simon" Lee, all errors are my own.

The last few decades have seen a rapid increase in the size of data sets, with the rise of streaming data such as social media discourse or transnational data. Traditional deterministic methods have become computationally expensive and impractical, as they require storing the entire data set in memory. Consequently, a new area of research called online or dynamic inference has emerged to tackle the inference problems posed by large-scale or streaming datasets (Hoffman et al. (2013)).

One of the popular algorithms used for online inference is stochastic gradient descent or the Robbins-Monro algorithm. The SGD solution path for a data sample $Y_{i i=1}^{n}$ is given by the recursive formula:

$$\beta_i = \beta_{i-1} - \gamma_i \nabla q(\beta_{i-1}, Y_i) \tag{2}$$

where β_0 is the initial starting value, γ_i is the step size, and $\nabla q(\beta_{i-1}, Y)$ is the gradient of $q(\beta, Y)$ with respect to β at $\beta = \beta_{i-1}$.

Since this is a recursive algorithm, it performs a single update with each additional data sample and does not need to keep in memory previous iterations. As a result, SGD has a time complexity of $O(n)$ and can be well suited to inference in the online learning setting where new data comes in sequentially.

In the vanilla SGD, the solution to (2) is given by the last iterate B_n. A more widely used solution is the Polyak (1990) - Ruppert (1988) averaging estimator $\bar{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\beta_i$, also called the averaged stochastic gradient descent (AGSD). Polyak and Juditsky (1992) showed $\bar{\beta}_n$ achieved asymptotic normality under certain regularity conditions.

Despite the results in Polyak and Juditsky (1992), most work in the SGD literature focus on convergence of the objective function or

distance between the estimator and the true solution. Only recently has statistical inference, ie. constructing confidence intervals for β^* gained traction, especially in the case of online learning (eg. Chen et al. (2020), Zhu et al. (2021), and Lee et al. (2022b)). In these papers, the SGD based inference methods are exemplified using the standard linear regression model and the logistic regression model.

This paper aims to apply the SGD algorithm with random scaling in Lee et al. (2022b) to the case of censored data through estimation of a tobit model. For many datasets, values above or below a threshold may be censored or truncated. Some examples include standardized test scores, exchange rates with government intervention, and income and wealth data. Regression models with a censored or truncated outcome variable estimated using a standard linear regression model will be biased. The censored regression model, first proposed in Tobin (1958), allows for unbiased inference.

The main results are as follows: (1) the stochastic gradient descent algorithm with random scaling developed in Lee et al. (2022b) is computationally faster than the basic estimation implementation of MLE in the case of censored data, (2) we can successfully conduct statistical inference under the tobit model.

2 Literature Review

2.1 SGD in Online Inference

The literature on stochastic gradient descent is extensive, with convergence results dating back to the latter half of the 19th century (e.g. Blum (1954), Dvoretzky (1956), and Robbins and Siegmund (1971)). Recently instead of focusing on convergence of point

estimates, statistical inference problems have gained prominence. In these papers, the interest is to quantify the uncertainty of parameter estimates using confidence intervals.

A significant paper focusing on inference with SGD is Chen et al. (2020). In this paper, the authors proposed two methods, a batch-means method and plug-in method, to construct asymptotically valid confidence intervals for the AGSD estimator. While the plug-in method require estimation of the Hessian matrix, the batch-means estimator does not by dividing the iterates into batches and using the empirical covariance to estimate the asymptotic covariance. Continuing from this, Zhu et al. (2021) introduced dynamic recursive updates to the batch-means method which allows for online inference.

Concurrently, Fang et al. (2018), Fang (2019), and Lee et al. (2022b) also develop online inference procedures based on SGD. Fang et al. (2018) and Fang (2019) developed bootstrap procedures for the confidence intervals of perturbed-SGD estimates. Lee et al. (2022b) developed a method of online inference using SGD with random scaling that employs asymptotically pivotal test statistic. This method is advantageous compared to other methods for online learning as the test statistic and critical values can be updated with every SGD iterate and without resampling. Lee et al. (2022a) extend the SGD algorithm with random scaling to the quantile regression model, which is a non-convex problem. Using decennial census data which contains millions of observations, they find trends in US. college wage premium gender gaps.

2.2 Tobit Model Estimation

Tobin (1958) first developed a regression model where the dependent variable has a known upper or lower bound and proposed a starting

estimator for maximum likelihood estimation. Amemiya showed that the maximum likelihood estimator in Tobin (1958) was inconsistent and proposed a new estimator that was consistent and asymptotically normal. Since then, many different forms of the Tobit models have been developed as well as numerous different estimation methods. To my knowledge, the stochastic gradient descent algorithm has not been applied to estimate a Tobit model, nor has the model been researched in an online setting for large scale data.

3 SGD with Random Scaling

In what follows, we summarize the SGD with random scaling in Lee et al. (2022b). In this paper, the authors first consider the Polyak-Ruppert averaging estimator $\bar{\beta}_n = \frac{1}{n}\sum_{i=1}^{n}\beta_i$. It was established in Polyak and Juditsky (1992) that $\bar{\beta}_n$ is asymptotically normal under certain regularity conditions, that is

$$\sqrt{n}(\bar{\beta}_n - \beta^*) \to^d \mathcal{N}(0, \Upsilon) \tag{3}$$

where the asymptotic variance is $\Upsilon := H^{-1}SH^{-1}$, $H := \nabla^2 Q(\beta^*)$ is the Hessian matrix, and $S := \mathbb{E}[\nabla(\beta^*, Y)\nabla(\beta^*, Y)']$ is the score variance. This estimator can be computed recursively by the updating rule $\bar{\beta}_i = \bar{\beta}_{i-1}\frac{i-1}{i} + \frac{\beta_i}{i}$, however the estimation of the asymptotic variance Υ requires storing all data, so it is not well suited for online inference.

Lee et al. (2022b) proposes a test statistic via random scaling that leverages the dependence among the averages of SGD iterates. First they show that the central limit theorem in (3) can be extended to a functional central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor}(\beta_i - \beta^*) \Rightarrow \Upsilon^{1/2}W(r), \quad r \in [0,1] \tag{4}$$

where $\Rightarrow$ is weak convergence in $\ell^\infty[0,1]$ and $W(r)$ is a vector of independent standard Wiener processes on $[0,1]$. Then (3) is a special case where $r = 1$. Building on this, they studentize $\sqrt{n}(\bar{\beta}_n - \beta^*)$ and consider the following t-statistic

$$\frac{\sqrt{n}(\bar{\beta}_{n,j} - \beta_j^*)}{\sqrt{\hat{V}_{n,jj}}} \tag{5}$$

where $\hat{V}_n$ is a random scaling matrix defined as

$$\hat{V}_n := \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{j}(\beta_t - \bar{\beta}_n)\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{j}(\beta_t - \bar{\beta}_n)\right)' \tag{6}$$

.

The functional CLT theorem in (4) shows that the t-statistic is asymptotically pivotal, so that its critical values are easily calculated. Furthermore, $\hat{V}_n$ can be updated recursively from just the SGD iterates so the test statistic and critical values are completely compatible with online data.

Given B_n and $\hat{V}_n$, inference can be conducted by calculating the $(1-\alpha)$ confidence interval for the j-th element B_j^* of B_j:

$$\left[\bar{B}_{n,j} - cv(1-\alpha/2)\sqrt{\frac{\hat{V}_{n,j}}{n}}, \bar{B}_{n,j} + cv(1-\alpha/2)\sqrt{\frac{\hat{V}_{n,j}}{n}}\right] \tag{7}$$

Following Lee et al. (2022b), the critical value $cv(1-\alpha/2)$ comes from Abadir and Paruolo (1997). The full SGD algorithm with random scaling for online inference is given below.

Algorithm 1 Lee et al. (2022b)

Online Inference with SGD via Random Scaling

Input: function $q(\cdot)$, parameters (γ_0, a) for step size $\gamma_t = \gamma_0 t^{-a}$ for $t \geq 1$

Initialize: $\beta_0, \bar{\beta}_0, A_0$

for $i = 1, 2, ...n$ **do**

Receive: new observation Y_i

$B_i = B_{i-1} - \gamma_i \nabla q(B_{i-1}, Y_i)$

$\bar{B}_i = \bar{B}_{i-1} \frac{i-1}{i} + \frac{\beta_i}{i}$

$A_i = A_{i-1} + i^2 \bar{B}_i \bar{B}_i'$

$b_i = b_{i-1} + i^2 \bar{B}_i$

Update $\hat{V}_i$ as

$$\hat{V}_i = i^{-2}\Big(A_i - \bar{B}_{i-i}^{,} - b_i \bar{B}_i' + \bar{B}_i \bar{B}_i' \textstyle\sum_{j=1}^{i} j^2\Big)$$

end for

4 Estimation of Tobit Model

In economics, censored data is a common issue. Common examples include income, price, count and duration data. Income data can have both left and right censoring–individuals in the top income percentiles may have their actual incomes censored for privacy reasons while individuals in the lowest income percentiles may not actively report their income leading to left censored data. Prices, counts of the number of occurrences of a particular event or measures of the duration of a particular event can be bounded at zero. Additionally, prices may be censored by price ceilings such as the price ceiling for rent. Censored economic variables can present challenges for statistical inference, and appropriate methods need to be used to account for the censored nature of the data.

Tobin (1958) developed one of the first regression models to deal with the scenario where the dependent variable is censored. For the variable that is censored the distribution of the sample data will be a mixture of discrete and continuous. Let the original distribution be y^* and for convenience, let the censoring point be zero. Then we can define y as the variable transformed from y^* by the following relationship:

$$\begin{aligned} y &= 0 \quad \text{if } y^* \leq 0 \\ y &= y^* \quad \text{if } y^* > 0 \end{aligned}$$

If y^* is normally distributed such that $y^* \sim N(\mu, \sigma)$, then we have $P(y = 0) = P(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - \Phi(\mu/\sigma)$ and $y^* \sim N(\mu, \sigma)$ if $y^* > 0$.

The one sided censored regression model is as follows:

$$y_i^* = \mathbf{x_i'}\beta + \varepsilon_i \tag{8}$$

$$y_i = 0 \quad \text{if} \quad y_i^* \leq 0 \tag{9}$$

$$y_i = y_i^* \quad \text{if} \quad y_i^* > 0 \tag{10}$$

Suppose $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, the log likelihood function is

$$\ln L = \sum_{i=1}^{n} \left[-\mathbb{1}_{y_i \leq 0} \ln \Phi\Big(\frac{-x_i'\beta}{\sigma}\Big) + (1 - \mathbb{1}_{y_i \leq 0}) \Big(\ln \phi\Big(\frac{y_i - x_i'\beta}{\sigma}\Big) - \ln \sigma \Big) \right] \tag{11}$$

Under a reparameterization where $\gamma = \beta/\sigma$ and $\theta = 1/\sigma$, Olsen (1978) showed that under this reparameterization, the log-likelihood will be globally concave with respect to γ and θ.

$$\ln L = \sum_{i=1}^{n} \left[-\mathbb{1}_{y_i \leq 0} \ln \Phi(-x_i'\gamma) + (1 - \mathbb{1}_{y_i \leq 0}) \Big(\ln \phi(y_i - x_i'\gamma) - \ln \theta \Big) \right] \tag{12}$$

Since the Hessian is always negative semidefinite, We can estimate $\beta^* = \gamma^*\sigma$ as the solution to the maximum likelihood problem using the SGD algorithm with random scaling via Lee et al. (2022b).

5 Simulation

In this section, I consider a similar simulation design as in Lee et al. (2022b) and Zhu et al. (2021) to test the computational perform of the random scaling method. I am testing on a Mac using the Apple M2 chip with 16 GB memory.

The data are generated from the model

$$y_i^* = \mathbf{x_i}'\beta + \varepsilon_i \tag{13}$$

$$y_i = 0 \quad \text{if} \quad y_i^* \leq 0 \tag{14}$$

$$y_i = y_i^* \quad \text{if} \quad y_i^* > 0 \tag{15}$$

where $x_i \sim \mathcal{N}(0, I_d)$, $\varepsilon \sim \mathcal{N}(0,1)$, and $\beta \in \mathbb{R}^d$ is equispaced on $[0,1]$. For the computational comparison, I test $d \in [5, 20, 100, 500, 1000]$ and $n \in [10^5, 10^6]$. I run 10 Monte Carlo simulations for each combination of d and n and compare the random scaling method with a standard maximum likelihood estimation.

I optimize the log-likelihood from the reparameterized model following Olsen. Since $\sigma = 1$, we have $\beta = \gamma$ and $\sigma = \theta$, so the log-likelihood is

$$\ln L = \sum_{i=1}^{n} \left[-\mathbb{1}_{y_i \leq 0} \ln \Phi(-x_i'\beta) + (1 - \mathbb{1}_{y_i \leq 0}) \Big(\ln \phi(y_i - x_i'\beta) - \ln \sigma \Big) \right] \tag{16}$$

For the random scaling method, the learning rate γ_i is parameterized

by $\gamma_0 = 0.5$ and $a = 0.505$. In the random scaling algorithm, I use the gradient of the loglikelihood from equation (15). The initial value of $\beta_0 = 0$ and I burn in 1% of observations for each case of $n \in [10^5, 10^6]$.

For the maximum likelihood estimation method, I first get initial estimates for (β, σ) using OLS and then use Python's optimize function with Gurobi to optimize the log-likelihood in (15). The default non-linear optimization algorithm used is BFGS.

The running time simulation results are summarized below:

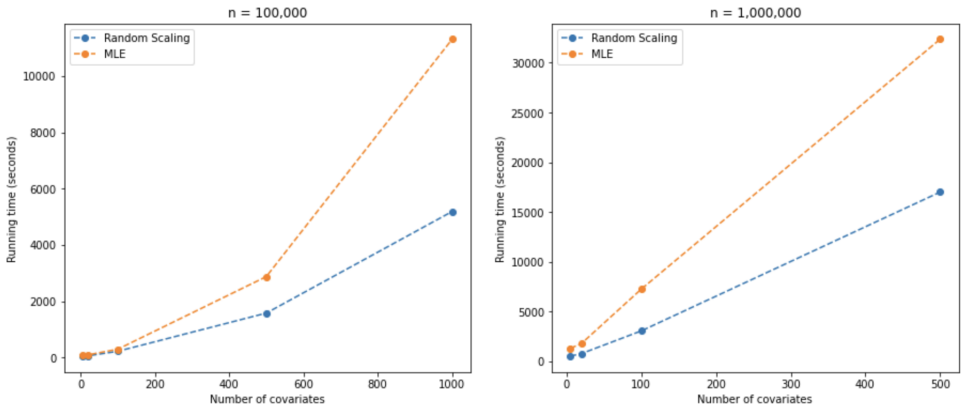

Figure 1: Comparison of Running Times

As we can see from Figure 1, the random scaling method does better than the basic maximum likelihood estimation. This is probably due to the fact that the maximum likelihood estimation uses the BFGS optimization algorithm that requires approximating the Hessian matrix which does not scale well for larger data sets. Additionally, we had to compute an initial guess using OLS which added additional computation time.

5.1 Inference

I also compare the inference results of both the random scaling algorithm and estimation using maximum likelihood. For each d, I run 1000 Monte Carlo trials. The results are summarized in Table 1 below. In the first estimation using random scale, I use 1% burn in-rate with $\beta_0 = 0$. In the second estimation, I employ a warm-start where I take 1% of the data and estimate β_0 using MLE assuming the tobit model. The confidence intervals are calculated according to (7), where $\alpha = 0.05$ and the critical value is 6.747.

	d=5	d=20	d=100
Random Scale			
Coverage	0.87	0.85	0.82
CI Length	0.065	0.068	0.075
Time (s)	18.1	32.9	105.8
Random Scale Warm-Start			
Coverage	0.9	0.88	0.85
CI Length	0.059	0.064	0.069
Time (s)	18.5	33.3	106.4
MLE			
Coverage	0.97	0.93	0.89
CI Length	0.03	0.034	0.036
Time (s)	23.6	54.2	310.2

Table 1: $n = 10^5$

As we can see from the coverage rates and confidence interval lengths, the random scale method may be less accurate than MLE. However, the time trade off is significant and from Figure 1, would grow at an exponential rate.

With the warm-start where we estimate β_0 using 1% of the data with

no burn-in, we have slightly better accuracy with marginal time trade off. This is intuitive as we have a more accurate starting point for β_0. Thus, a warm-start could be a useful alternative, as long as the initial sub-sample is small enough to ensure reasonable computational time.

6 Empirical Application: Estimating the College Wage Premium

The college wage premium generally refers to the difference in earnings between individuals who have completed a college degree and those who have a high school degree but no college education. Research on the college wage premium has consistently shown that individuals with a college degree earn more on average than those without a college degree. The exact size of the college wage premium varies depending on the study and the time period analyzed, but most estimates suggest that college graduates earn between 40% and 80% more than those without a college degree (James (2012)). Moreover, the college wage premium has been shown to be increasing over time, indicating that a college degree is becoming increasingly valuable in the labor market. In addition to the base college wage premium, there is significant heterogeneity in the college wage premium across different socio-characteristics such as sex, age, and race (see eg. Grogger and Eide (1995), Goldin et al. (2006), Taniguchi (2005)).

Most prominent estimates on the college wage premium rely on data from the United States Census Bureau. The Census Bureau collects income estimates from several major surveys: the decennial census, American Community Survey (ACS) and the Annual Social and Economic Supplement to the Current Population Survey (CPS ASEC). These surveys employ various forms of censoring to protect

the privacy of survey respondents while ensuring that accurate data is collected. The primary form of censoring is top coding, where values above a certain threshold are replaced with the threshold value. Additionally, the Census Bureau may use suppression, where data cells with fewer than a certain number of responses are withheld to prevent the identification of individual survey respondents.

Top coding is a form of censoring and it is clear that inference estimates of the college wage premium could be unbiased if appropriate statistical methods or adjustments are not used. Hubbard (2011) shows that after appropriately accounting for censoring of income data, there is negligible gender difference in the wage premium. While topcoding is a widely recognized issue for census data and has been prevalent in public use census files since 1967, previous research prior to Hubbard (2011) did not account for topcoding bias in the college wage premium.

In his paper, Hubbard uses the Integrated Public Use Microdata Series (IPUMS) CPS wage series from 1970 to 2008 to estimate the college wage premium and notes that the top code has intermittently increased from 50,000 nominal US dollars in 1967 to 200,000 nominal US dollars in 2000. However starting in 2003, the top-coded threshold actually varies by state and is set at the 99.5th percentile of the wage distribution at the state level. Additionally, the way values over the threshold are imputed changed starting in 1990. Specifically, the top code was the imputed value for incomes above the threshold before 1990. From 1990-2002, the median above the threshold income for each state was the imputed value for incomes above the threshold. Starting in 2003, the mean above the threshold income for each state was the imputed value. The thresholds and imputing method of the IPUMS INCWAGE series are given in Table 2.

Census	Top Code
1940	$5,001
1950	$10,000
1960	$25,000
1970	$50,000
1980	$75,000
1990	$140,000*
2000	$175,000**
ACS (2000-2002)	$200,000**
ACS (2003-onward)	99.5th Percentile in State**

Table 2: Threshold levels for the INCWAGE series in IPUMS public data.
* Higher incomes are expressed as the state medians income above the listed Top Code value for that specific Census year.
** Higher incomes are coded as the state means of values above the listed Top Code value for that specific Census year.

Hubbard (2011) uses 3 different inference methods to account for topcoding bias in both the base college wage premium and the gap in college wage premium between men and women: OLS regression with wage observations over the top-coded threshold adjusted to their expected value, tobit regression, and quantile regression. While the adjustment method in Hubbard (2011) matches the imputing method for INCWAGE for the years 2003 onwards, the author does the adjustment at the national and not state level. Additionally Hubbard (2011) controls for regional fixed effects, but does not control for state fixed effects. In Table 3 below, the threshold levels for each state varies significantly. Connecticut has the highest 99.5th percentile of income at $526,000 while South Dakota has the lowest 99.5th percentile of income at $189,000. Given this high variability in the threshold, there will most likely be bias if we assume, as Hubbard

(2011) did, that the threshold remains at $200,000 from 2000 onwards for all states and impute for values above the threshold using a national income distribution. This bias is one main motivation for the empirical application of the online inference algorithm for the tobit model.

Top 10 States	**Threshold**	**Bottom 10 States**	**Threshold**
Connecticut	526000	Wyoming	232000
District of Columbia	491000	Alaska	231000
New York	445000	Louisiana	227000
New Jersey	418000	Idaho	226000
Maryland	388000	Oklahoma	218000
Massachusetts	382000	Montana	195000
Illinois	359000	New Mexico	191000
Colorado	352000	Utah	190000
Kansas	350000	South Dakota	189000

Table 3: 99.5th wage percentile for the top 10 and bottom 10 states in 2003. Values in nominal 2003 US dollars.

Another main motivation for this empirical application is the size of the IPUMS CPS data. Since 2000, each yearly sample of the IPUMS CPS data contains hundreds of thousands of observations. While this data is not continuously streaming, it is updated annually. The necessity of state fixed effects would make this dataset a large enough panel to test the Lee et al. (2022b) SGD inference algorithm. In what follows, I describe the dataset in more details and the inference results from estimating the college wage premium and the difference in college wage premium between men and women.

6.1 Data

6.1.1 Construction

I use data from IPUMS USA for the years 2000 to 2007. This dataset at the annual level is comparable to the IPUMS CPS dataset. The data is compiled from the 2000 Census and the annual American Comunity Survey. From 2000-2007, the limiting sampling size was in the year 2002 at 0.38% so I initially restricted all yearly samples to a density of 0.38%.

Then following Hubbard (2011) and the literature, I filter for observations with positive CPS sample weights and full-time full-year workers (FTFY). Full time is defined as 35 or more hours of work per week and full year is defined as 50 or more weeks a year. FTFY workers have been shown to be the most accurately reported, whereas wages for part-time, part-year workers can be substantially under reported Roemer (2000).

Next, I create a variable for years of schooling. In the IPUMS data, the educational variable is coded based on certain milestone years of completion for those with less than a completed high school degree and degree completed for those with more education. I assume that the time it takes to complete a degree is the minimum time, ie. associate would take 14 years, bachelor's would take 16 years, master's would take 18 years, a professional degree such as an MBA or law school would take 19 years, and a doctoral degree would take 21 years.

From the years of school measure, I generate potential experience as age$-$years of school-6. Since I assumed the minimum completion time for each degree level, potential experience may be inflated but I believe this would a minor effect. As the college wage premium is generally defined in the literature between high school and college graduates, I

restrict the dataset to those with at least a high school degree. I also adjust wages using the Personal Consumption Expenditures (PCE) to 2000 dollars and drop observations with annual wages less than one-half the annual income from the minimum wage (5.15 per hour or 10,712 per year in 2000).

6.1.2 Summary Statistics

The initial sample from IPUMS contains over 9 million observations across the years 2000-2007. After filtering for FTFY workers, high school graduation, and income above half the minimum wage, the panel contains slightly more than 3.7 million observations.

Figure 2 shows the average 99.5th percentile for each state and the percentage of censored observations for the years 2000-2007. The total percentage of observations censored across all years is between 1-2% for all states. In general, states with higher income thresholds have more censorship. The average 99.5th percentile for most states hover around $300,000 and is slightly skewed right. Without states fixed effect, a Tobit model may overestimate the college wage premium due to the observations from states with higher thresholds.

Figure 3 shows the average 99.5th percentile across all US states for each year from 2000-2007. While the average percentage of censored observations across all states remain steady around 1%, there was a brief increase to almost 4% in 2004. It seems that the 99.5th percentile for the year 2004 decreased significantly across all US states. It is unclear why there is such a common shock during the year 2004 and it may perhaps be a clerical error, but I assume the reported income percentiles are correct.

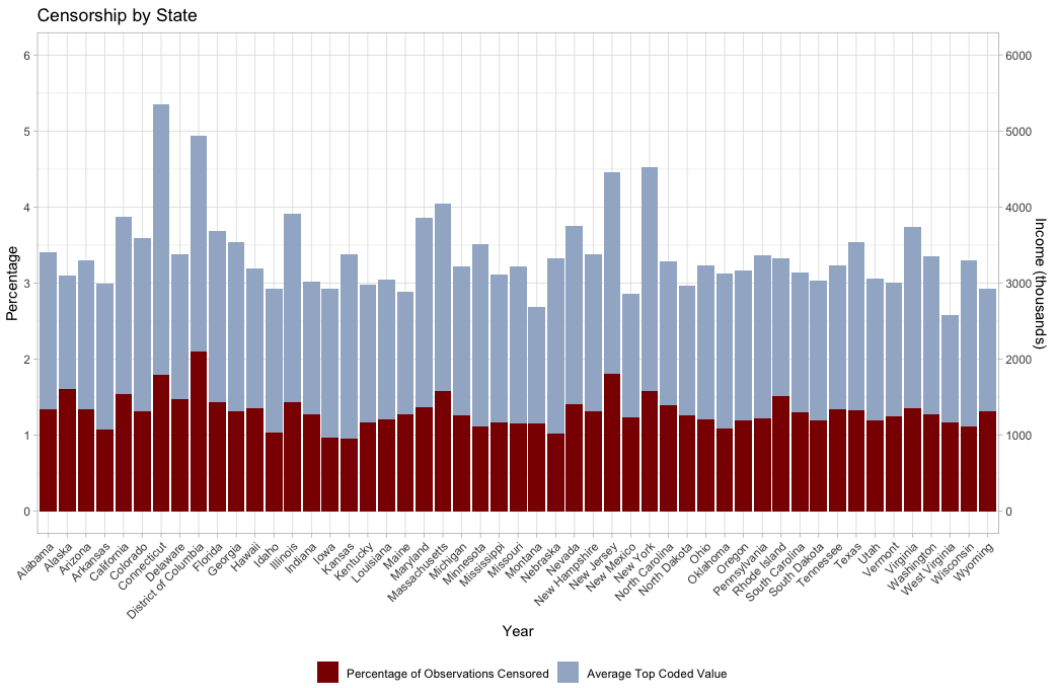

Figure 2: Top coded value is deflated using the PCE index to 2000 dollars.

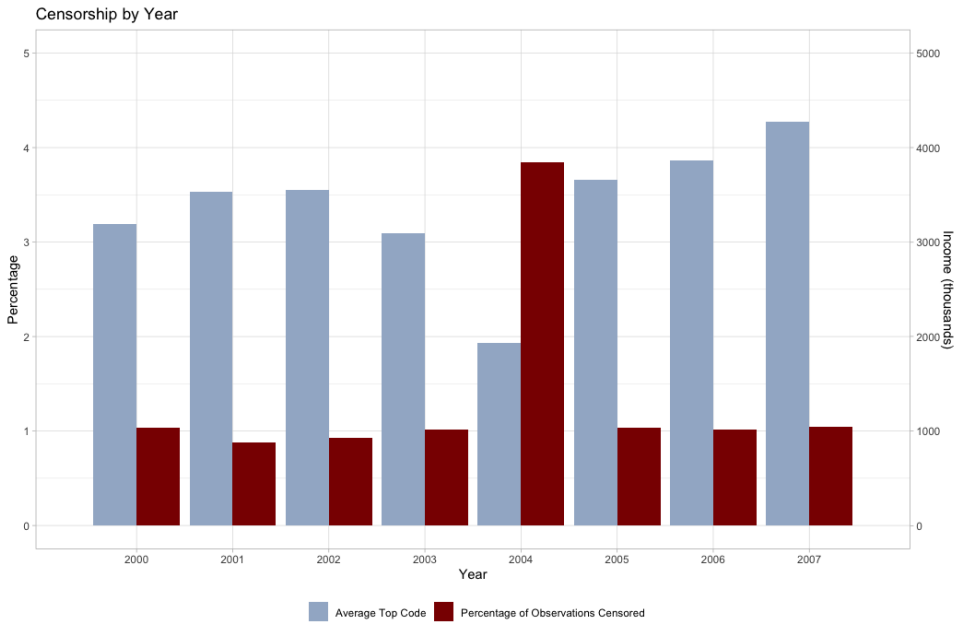

Figure 3: Wage is deflated using the PCE index to 2000 dollars.

6.1.3 Preliminary patterns

Figure 4 plots the median wage by degree achievement. Advanced degrees include professional and doctoral degrees. At preliminary examination, there seems to be a clear wage premium as educational attainment increases. In 2000, the median annual wage for high school, college, and advanced degree holders were $24,000, $37,000, and $47,000 respectively. In 2007, the median annual wage (in 2000 US dollars) for high school, college, and advanced degree holders were $32,849, $53,966 and $70,391 respectively. This gives the average annual growth rate from 2000-2007 adjusted for inflation, as 4.59%, 5.53%, and 5.94% for high school, college, and advanced degree holders respectively. Thus, the educational wage premium has also increased over time.

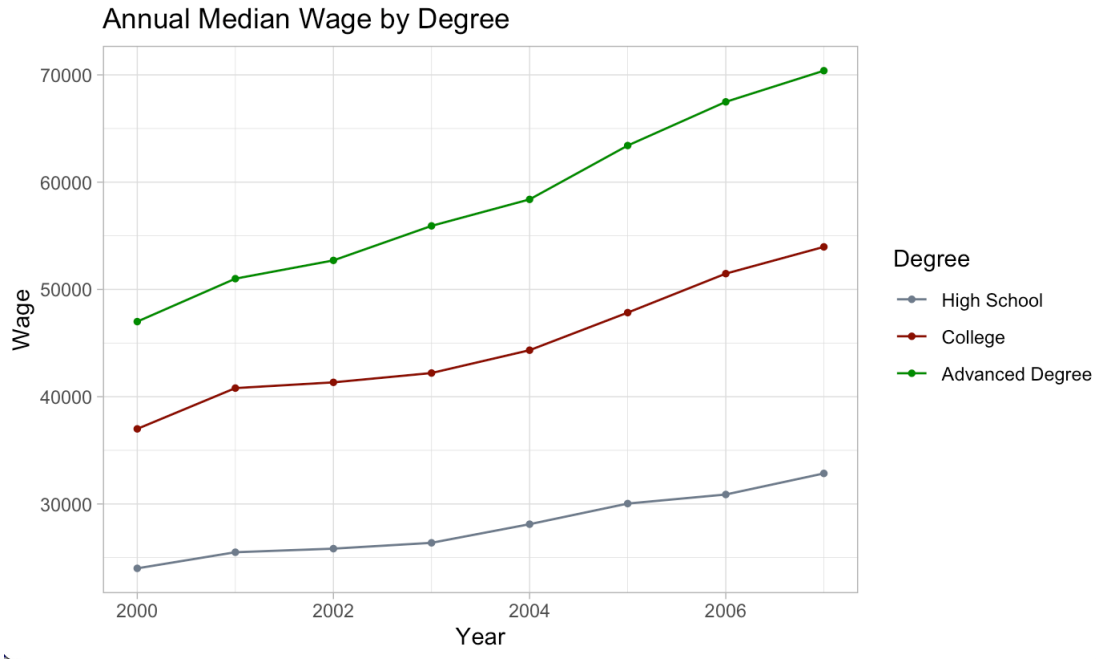

Figure 4: Wage is deflated using the PCE index to 2000 dollars.

Figure 5 plots the median college wage gap–the difference between the

median wage of a college graduate and a high school graduate–by gender. From 2000-2007, the college wage premium was consistently higher for men then women and increased over time. In 2000, the median annual wage for high school, college, and advanced degree holders were $24,000, $37,000, and $47,000 respectively. In 2007, the median annual wage (in 2000 US dollars) for high school, college, and advanced degree holders were $32,849, $53,966 and $70,391 respectively. This gives the average annual growth rate from 2000-2007 adjusted for inflation, as 4.59%, 5.53%, and 5.94% for high school, college, and advanced degree holders respectively. Thus, the educational wage premium has also increased over time.

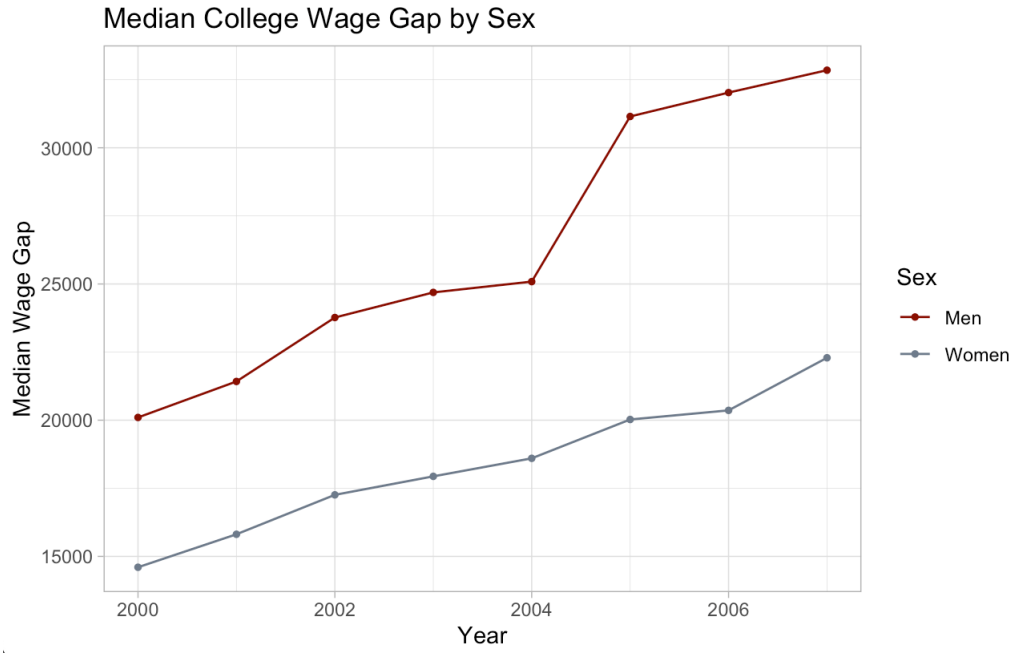

Figure 5: The wage gap is measured as the difference between the median annual wage for college graduates and the median annual wage for high school graduates.

6.2 Model

Following the literature and Hubbard, I estimate the following model for the college wage premium:

$$ln(y_i) = \alpha + \gamma Educ_i + \theta_1 Exp_i + \theta_2 Exp_i^2 + \Phi X_i + \varepsilon_i \quad (17)$$

where y_i is the annual wage income adjusted for inflation, $Educ_i$ is a dummy for college graduate (note that all observations have at least a high school degree), Exp_i is measured as age $-$ years of school $-$ 6, and X_i contains state dummies. The college wage premium is measured by the coefficient γ.

I estimate the following model for the gender gap in the college wage premium:

$$ln(y_i) = \alpha + \beta Fem_i + \gamma Educ_i + \theta_1 Exp_i + \theta_2 Exp_i^2 + \Phi X_i + \varepsilon_i \quad (18)$$

where Fem_i is a dummy for the female sex and X_i contains state dummies all interacted with the female dummy. The difference between the male and female college wage premium is measured by the coefficient δ on the interaction term.

For the base model, I first show the difference between the Census region fixed effects used in Hubbard versus state fixed effects for two specifications: (1) OLS regression when annual wages are recensored at 100,000 nominal USD as in Card and DiNardo (2002) and (2) Tobit regression.

Then for both models, I run 3 specifications with state fixed effects: (1) OLS regression when annual wages are recensored at 100,000 nominal USD as in Card and DiNardo (2002), (2) OLS regression with no adjustment to the data, (3) Tobit regression. The results are subsequently reported.

6.3 Results

6.3.1 College Wage Premium

One potential issue with the model in Hubbard (2011) is that it does not account for variable top code imputation values across states starting in 2000 and instead uses Census region fixed effects. I compare two estimates - OLS estimates with recensored wages at 100,000 nominal USD as in Card and DiNardo (2002) and Tobit estimates under the case of either Census region fixed effects or state fixed effects. The time-varying estimates for the years 2000 to 2007 are shown in Figure 6 below.

In both cases, we see that the change from Census region fixed effects to state fixed effects lowers the estimated college wage premium. When we control for state fixed effects instead of regional fixed effects, we control for more differential variation. In table 3, we can see that while many of the 10 states with the highest income thresholds are in the northeast, there is still variation in region. If we only control for region fixed effects, there is less control for regional variations and the college wage premium is biased upwards by income levels in higher cost of living/metropolitan states.

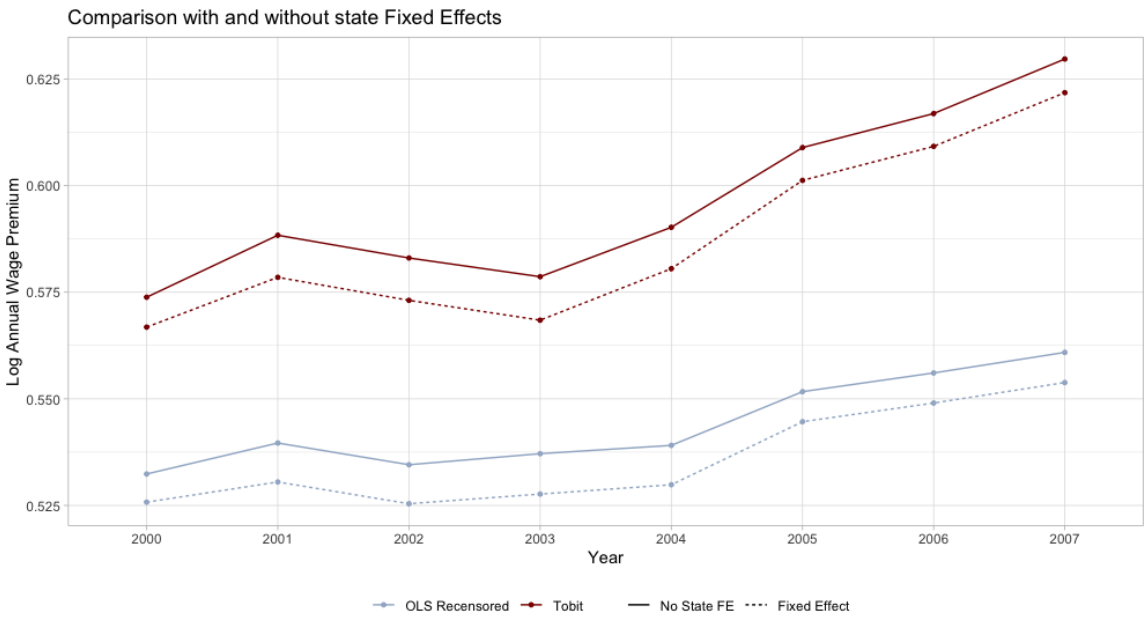

Figure 6: Wage is deflated using the PCE index to 2000 dollars.

Next, I compare 3 specifications all with states fixed effects: (1) OLS regression when annual wages are recensored at 100,000 nominal USD, (2) OLS regression with no adjustment to the data, (3) Tobit regression. The first specification is used in one of the prominent papers on wage inequality, Card and DiNardo (2002) which avoids large jumps in estimated wages when topcodes change. The second specification uses the original data which has imputed values at the state median or mean above the threshold. This is similar to the adjustment method used in Hubbard (2011) in which he adjusts wages over the top code to the expected average wage above the top code threshold assuming wages follow a Pareto distribution–here the difference is that Hubbard assumes a single Pareto distribution across all wages in the US, and does not discern at the state level. Finally, the third specification uses the Tobit algorithm extended from Lee et al. (2022b). The results are reported in Table 4 and Figure 7 below.

The log annual wage premium for the years 2000-2007 is in the range of 0.5 to 0.6, similar to the results inHubbard (2011), although in his paper he only reports the male college wage premium. This implies a college degree increases wages by approximately 50 to 60%.

As we can see the college wage premium when wages are recensored at 100,000 nominal USD is much lower. This is because the college wage premium is biased downwards as the actual observations above the threshold are likely to be college graduates with (much) higher income. The OLS estimates when wages above the threshold are imputed at the state median/mean of wages above the threshold versus the Tobit estimates are quite similar. We can see that starting in 2003 when the thresholds vary by state that there is a significant departure between the Tobit and OLS estimates.

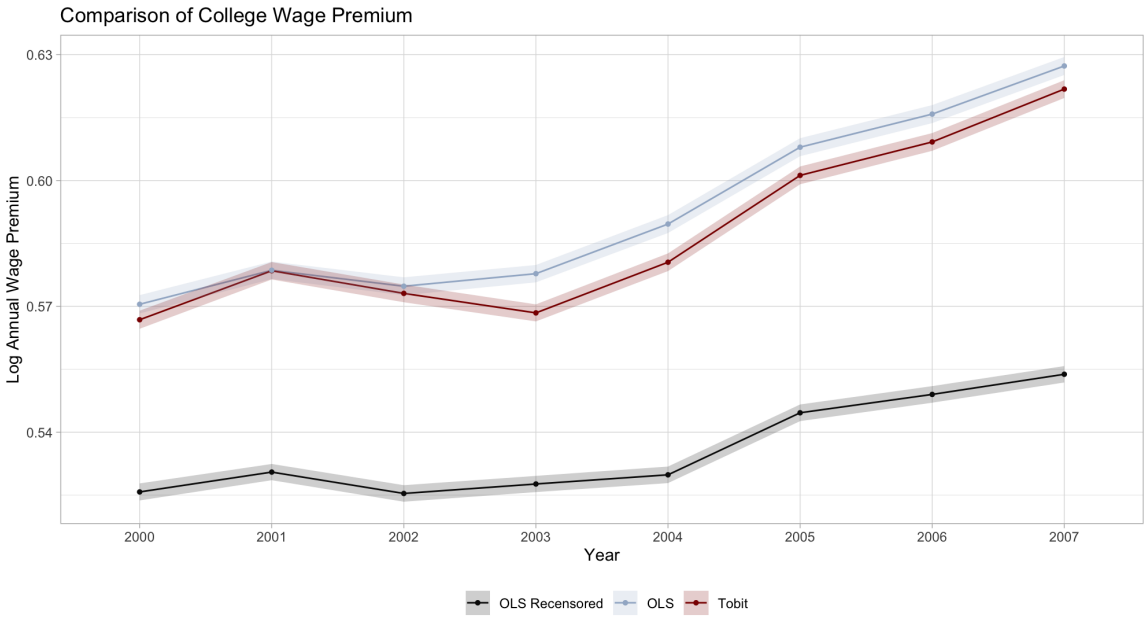

Figure 7: Wage is deflated using the PCE index to 2000 dollars.

Table 4: College Wage Premium

		Model		
	Year	OLS Recensored	OLS	Tobit
γ	2000	0.561***	0.570***	0.567***
		(0.00213)	(0.00217)	(0.00215)
	2001	0.574***	0.579***	0.578***
		(0.00206)	(0.00207)	(0.00207)
	2002	0.569***	0.575***	0.573***
		(0.00209)	(0.00212)	(0.00211)
	2003	0.560***	0.578***	0.568***
		(0.00200)	(0.00206)	(0.00203)
	2004	0.557***	0.590***	0.581***
		(0.00203)	(0.00212)	(0.00211)
	2005	0.595***	0.608***	0.601***
		(0.00210)	(0.00214)	(0.00212)
	2006	0.605***	0.616***	0.609***
		(0.00210)	(0.00214)	(0.00211)
	2007	0.605***	0.627***	0.622***
		(0.00210)	(0.00214)	(0.00212)

Note: *p<0.1; **p<0.05; ***p<0.01

6.3.2 Gender Gap in College Wage Premium

Just as in Hubbard (2011), I also investigate the difference in the college wage premium between men and women. I again compare 3 estimates of the model in equation (18) with states fixed effects: (1) OLS regression

when annual wages are recensored at 100,000 nominal USD, (2) OLS regression with no adjustment to the data, (3) Tobit regression. The results are reported in Table 5 and Figure 8 below.

The difference between the annual wage premium of men and women for the years 2000-2007 is in the range of -0.04 to -0.07 for the Tobit and median/mean imputed wages similar to the results in Hubbard (2011). That is on average, women gain 4-7% less in wages from having a college education compared to men. When the annual wages are censored to 100,000 nominal USD, we get the results that the college wage premium is higher for women which was the common stylized fact in the early 2000s. As noted in Hubbard (2011), there is downward bias when annual wages are censored as the proportion of censored observations above the thresholds tend to be predominantly men.

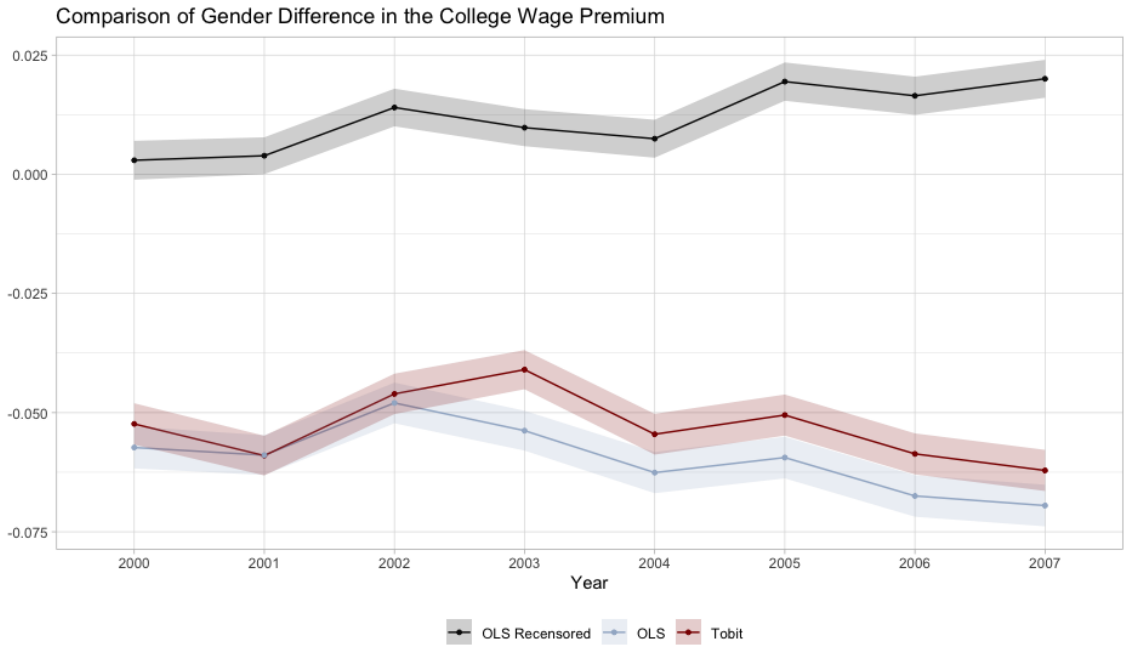

Figure 8: Wage is deflated using the PCE index to 2000 dollars.

The difference between the OLS estimates with median/mean imputed

wages and Tobit estimates are not significantly different besides the year 2003 when the shift to state percentile thresholds occurred. The Tobit estimates are slightly smaller in magnitude post-2003 which could be due less sensitivity to the variation in thresholds.

Table 5: Gender Gap in College Wage Premium

		Model		
	Year	OLS Recensored	OLS	Tobit
δ	2000	-0.0438***	-0.0573***	-0.0524***
		(0.00429)	(0.00438)	(0.00434)
	2001	-0.0530***	-0.0589***	-0.0590***
		(0.00412)	(0.00416)	(0.00416)
	2002	-0.0402***	-0.0480***	-0.0461***
		(0.00420)	(0.00425)	(0.00424)
	2003	-0.0293***	-0.0538***	-0.0410***
		(0.00404)	(0.0017)	(0.00410)
	2004	-0.0235***	-0.0626***	-0.0545***
		(0.00411)	(0.00431)	(0.00427)
	2005	-0.0418***	-0.0594***	-0.0505***
		(0.00426)	(0.00435)	(0.00431)
	2006	-0.0526***	-0.0675***	-0.0587***
		(0.00427)	(0.00435)	(0.00430)
	2007	-0.0571***	-0.0695***	-0.0621***
		(0.00429)	(0.00435)	(0.00431)

Note: *p<0.1; **p<0.05; ***p<0.01

Figures 9-11 show the individual college wage premium for men and women for each of the three estimates. When wages are recensored at 100,000 nominal USD threshold, we see the downward bias for the college wage premium for men and an upward bias for the college wage premium for women.

Figures 10 and 11, showing the OLS and Tobit estimates respectively, look remarkably similar except for 2003 when the change to state varying thresholds occur when the male wage premium estimates slightly dip. This could be due to less sensitivity in the Tobit estimation to the variation in thresholds.

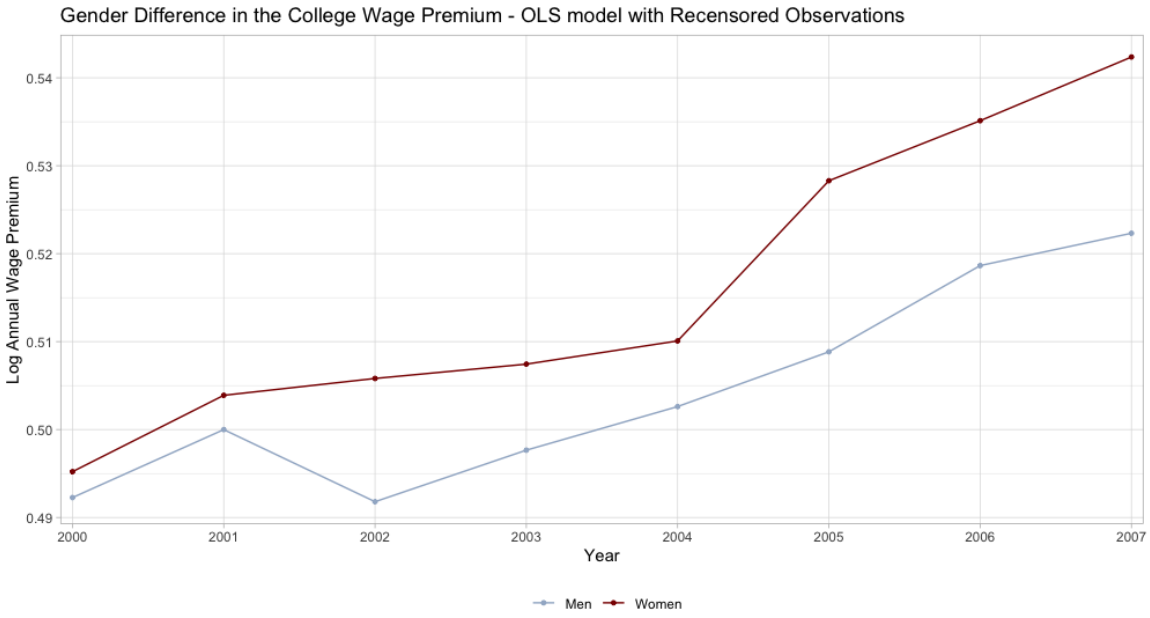

Figure 9: Wage is deflated using the PCE index to 2000 dollars.

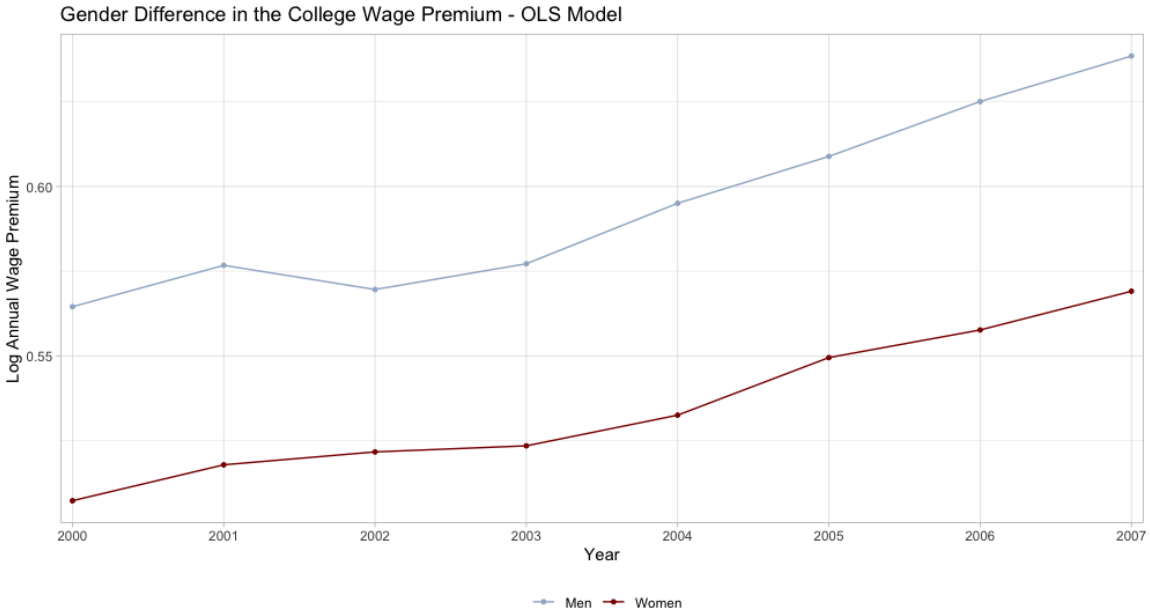

Figure 10: Wage is deflated using the PCE index to 2000 dollars.

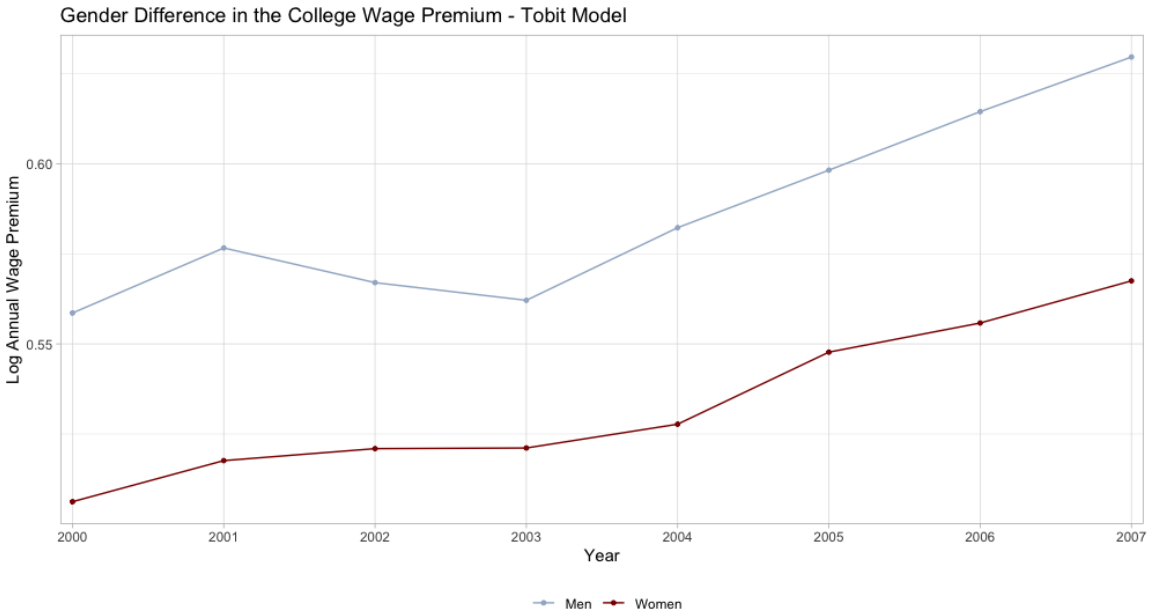

Figure 11: Wage is deflated using the PCE index to 2000 dollars.

6.3.3 Computational Results

Below I compare the average computational time to estimate the models in equation (17) and (18). Additionally, I compare the confidence interval length of the college wage premium and male college wage premium for each model respectively. Across 8 years, each panel has approximately 500,000 observations. Model 1 is the base model with no control for gender. Model 2 has additional controls and interaction terms with gender.

	Model 1 d = 52	Model 2 d = 57
Random Scale		
CI Length	0.00617	0.00696
Time (s)	49.5	52.9
Random Scale Warm-Start		
CI Length	0.00592	0.00619
Time (s)	52.3	57.1
MLE		
CI Length	0.00421	0.00579
Time (s)	74.2	79.2

Table 6: $n = 5 * 10^5$

As we can see from the confidence interval lengths, the random scale method are less accurate than MLE while the time trade off is positive.

7 Conclusion

In this paper, I showed that the stochastic gradient descent algorithm with random scaling developed in Lee et al. (2022b) for online inference

is computationally faster than the basic estimation implementation of MLE in the case of censored data. Additionally, I applied the algorithm to estimate the college wage premium in an empirical example. The results are consistent with past estimates from Hubbard (2011). Future work in this research would involve additional tests of robustness for the random scaling algorithm as well as testing on larger datasets and computational clusters.

References

Abadir, Karim M. and Paolo Paruolo (1997) "Two Mixed Normal Densities from Cointegration Analysis," *Econometrica*, 65 (3), 671–680, http://www.jstor.org/stable/2171758.

Blum, Julius R. (1954) "Multidimensional Stochastic Approximation Methods," *The Annals of Mathematical Statistics*, 25 (4), 737 – 744, 10.1214/aoms/1177728659.

Card, David and John E. DiNardo (2002) "Skill-biased technological change and rising wage inequality: Some problems and puzzles," *Journal of Labor Economics*, 20 (4), 733–783, 10.1086/342055.

Chen, Xi, Jason D. Lee, Xin T. Tong, and Yichen Zhang (2020) "Statistical inference for model parameters in stochastic gradient descent," *The Annals of Statistics*, 48 (1), 251 – 273, 10.1214/18-AOS1801.

Dvoretzky, Aryeh (1956) "Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 3.1 (7), 39–55.

Fang, Yixin (2019) "Scalable statistical inference for averaged implicit stochastic gradient descent," *Scandinavian Journal of Statistics*, 46 (4), 987–1002, 10.1111/sjos.12378.

Fang, Yixin, Jinfeng Xu, and Lei Yang (2018) "Online Bootstrap Confidence Intervals for the Stochastic Gradient Descent Estimator," *Journal of Machine Learning Research*, 19 (78), 1–21, http://jmlr.org/papers/v19/17-370.html.

Goldin, Claudia, Lawrence F Katz, and Ilyana Kuziemko (2006) "The Homecoming of American College Women: The Reversal of the

College Gender Gap," *Journal of Economic Perspectives*, 20 (4), 133–156, https://doi.org/10.1257/jep.20.4.133.

Greene, William H. (2003) *Econometric Analysis*: Pearson Education.

Grogger, Jeff and Eric Eide (1995) "Changes in College Skills and the Rise in the College Wage Premium," *The Journal of Human Resources*, 30 (2), 280–310, https://doi.org/10.2307/146120.

Hoffman, Matthew D., David M. Blei, Chong Wang, and John Paisley (2013) "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14 (40), 1303–1347, `http://jmlr.org/papers/v14/hoffman13a.html`.

Hubbard, William H. (2011) "The phantom gender difference in the college wage premium," *Journal of Human Resources*, 46 (3), 568–586, 10.1353/jhr.2011.0030.

James, Jonathan (2012) "The College Wage Premium," *Economic Commentary (Federal Reserve Bank of Cleveland)*, 1–4, https://doi.org/10.26509/frbc-ec-201210.

Lee, Sokbae, Yuan Liao, Myung Hwan Seo, and Youngki Shin (2022a) "Fast Inference for Quantile Regression with Tens of Millions of Observations," 10.48550/ARXIV.2209.14502.

——— (2022b) "Fast and Robust Online Inference with Stochastic Gradient Descent via Random Scaling," *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (7), 7381–7389, 10.1609/aaai.v36i7.20701.

Olsen, Randall J. (1978) "Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model," *Econometrica*, 46 (5), 1211–1215, `http://www.jstor.org/stable/1911445`.

Polyak, Boris (1990) "New stochastic approximation type procedures," *Avtomatica i Telemekhanika*, 7, 98–107.

Polyak, Boris and Anatoli Juditsky (1992) "Acceleration of Stochastic Approximation by Averaging," *SIAM Journal on Control and Optimization*, 30 (4), 838–855, 10.1137/0330046.

Robbins, Herbert and Sutton Monro (1951) "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22 (3), 400 – 407, 10.1214/aoms/1177729586.

Robbins, Herbert and David Siegmund (1971) "A convergence theorem for non negative almost supermartingales and some applications," *Optimizing Methods in Statistics*, 233–257, https://doi.org/10.1016/B978-0-12-604550-5.50015-8.

Roemer, Marc I (2000) "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990 - 1996."

Ruppert, David (1988) "Efficient Estimations from a Slowly Convergent Robbins-Monro Process,"Technical Report 781, Cornell University Operations Research and Industrial Engineering.

Taniguchi, Hiromi (2005) "The Influence of Age at Degree Completion on College Wage Premiums," *Research in Higher Education*, 46 (8), 861–881, https://doi.org/10.1007/s11162-005-6932-8.

Tobin, James (1958) "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26 (1), 24–36, `http://www.jstor.org/stable/1907382`.

Zhu, Wanrong, Xi Chen, and Wei Biao Wu (2021) "Online Covariance Matrix Estimation in Stochastic Gradient Descent," *Journal of the American Statistical Association*, 0 (0), 1–12, 10.1080/01621459.2021.1933498.