Credible Inference for the Heterogeneous Returns to Schooling

Jonathan L. Gu

May 22, 2020

Abstract

This article estimates the heterogeneous returns to education by place-of-birth for males born between 1930 and 1940. We form intervals of credibility from the posterior distribution of our estimate. Instead of assuming the two-stage least squares estimate is distributed normal, we simulate draws from the posterior distribution of our two-stage estimate. When using weak priors, we obtain the same point estimates as the standard IV-2SLS methods, but we have much larger 95% credible intervals. When using the "best" priors as determined from cross-validation, we find that we are 95% sure that the returns to education are positive for only four out of nine regions, whereas the standard IV-2SLS approach would yield "significant" results for all nine geographic regions.

Keywords: Machine Learning, Causality, Bayesian, Weak Instrumental Variables, heterogeneous returns to schooling, Angrist-Krueger data.

Contents

1 Introduction

There is a long history of using instrumental variables to determine the causal effect of schooling on education. Angrist and Krueger (1991) finds that each additional year of schooling can boost earnings by about 8%. The discussion of methods for determining our confidence in these point estimates has spurred a lot of literature. Current empirical applications of instrumental variables face must first pass a binary test of whether the instruments are "strong" enough for us to assume that the likelihood function is quadratic, and therefore the estimate is distributed normal with variance as derived from the delta method. This article is the first to estimate and make statements of confidence about the heterogeneous returns to education by region of birth using the same 1980 Census data used by Angrist and Krueger. We do not suffer from "weak instruments" by avoiding using an asymptotic approximation (Leamer, 2010) of the instrumental variables two-stage least squares (IV-2SLS) estimate.

Instead, we sample from the posterior distribution of our two-stage estimate and form point estimates and measures of confidence directly from the samples. When using weak priors, we obtain the same point estimates as the standard IV-2SLS methods, but we have much larger 95% credible intervals. When using the "best" priors as determined from cross-validation, we find that we are 95% sure that the returns to education are positive for only four out of nine regions, whereas the standard IV-2SLS approach would yield "significant" results for all nine geographic regions.

Bound et al. (1995) uses simulated (fake) instruments to obtain similar results of "significance" to the results obtained by (Angrist and Krueger, 1991). In the case of IV-2SLS, standard definitions of significance is defined by whether zero lies in a 95% confidence interval formed by applying the delta method to the two-stage least squares estimate. The delta method is an appeal to the large sample asymptotics of the IV-2SLS estimate. However, we know that the just-identified IV-2SLS estimate does not have finite moments (Nelson and Startz, 1990; Phillips, 2009) [1], and therefore the appeal to asymptotics in the case of IV-2SLS analysis is diminished.

Chamberlain and Imbens (1996) explains that sampling from the posterior instead of using an asymptotic approximation leads to tight posterior intervals when using the data, and wide posterior intervals when using randomly simulated instruments. Lopes and Polson (2014) and Hoogerheide et al. (2007) discuss a variety of priors to further analyze the causal returns to education in the same setting. This article is the first to add the analysis of heterogeneous treatment effects to the tradition of applying Bayesian credible

[1] The number of finite moments depends on the degree of over-identification.

inference methods towards empirical data.

Due to the increased availability of big data and large scale AB tests in the tech industry, readily available code packages (EconML, 2019) inspired by (Newey and Powell, 2003; Hartford et al., 2017) have surfaced that advertise the ability to determine heterogeneous treatment effects in a nonparametric manner. For the initial impact of the AB test on an initial outcome, the method for discovering causal heterogeneous effects is quite straightforward. However, once we ask for the causal effect of the initial outcome on a secondary outcome, we begin to confront the issue of weak instruments. Staiger and Stock (1994); Stock and Yogo (2002) discuss conditions under which the instrument is "strong enough" to assume the IV-2SLS estimate is normal.

Their approach requires the source of heterogeneity to be non-stochastic, or independent from the omitted factors that determine the treatment. In the case of heterogeneous returns to schooling by region-of-birth, this means that we require the region-of-birth to be independent from the omitted variables that determine the years of schooling. Instead, we implement a strategy that avoids this pitfall at the cost of including estimating additional first-stage equations.

Section 2 summarizes the data, Section 3 discusses the mathematical foundations and the reason why we have many first-stage equations, Section 4 discusses the method of sampling from the posterior, Section 5 summarizes our results, and Section 6 concludes.

2 Data

Following (Angrist and Krueger, 1991), we use the 5 percent Public Use sample (the A Sample) of the US population as of April 1, 1980. Following Chamberlain and Imbens (1996), we filtered down to native-born African American/Black or Caucasian/white males with birthdays in the first or fourth quarter of a year between 1930 and 1939 (inclusive). The key insight for our identification strategy is that men born in the fourth quarter tend to have about one more year of schooling. Figure 1 shows the interquartile ranges of the years of education against quarter-of-birth. For census participants born in the fourth quarter, we see that the first quartile of the level of schooling is equal to the median. This suggests there is a binding lower bound to the number of years of education for census participants born in the fourth quarter.

I limit the sample to men with positive wage and salary earnings, and positive weeks worked in 1979, leaving 202,859 participants in the sample. We compute weekly earnings by dividing annual earnings by weeks worked. Table 1 shows that the average man was 45 years old and earned $430 dollars per week in 1979. Eighty four percent of these men were married, and nine percent of these men were African American/Black. The level of schooling is determined by the years of completed schooling, where twelve corresponds to completing high school. The original paper from 1991 conditioned on some concomitant variables such as place of current residence, and marital status, however conditioning on concomitant variables can corrupt the causal interpretation of the data, so I remove these controls. For this article, we weight each census respondent equally, and we leave extensions to weighted calculations for future research.

Since some of the states are much smaller than the others, we follow the original paper in using census defined geographic regions as the source of potentially heterogeneous treatment effects. Since the region of birth is determined chronologically before any educational and employment decisions we can worry less about the issue of concomitant variables. Figure 2 displays the census regions. Table 2 lists out the states that make up each region.

Figure 3 shows the relationship between the average log weekly earnings against the average years of education for each census region. We see a general positive trend, but more importantly, we see that each region has differing average levels of weekly earnings and education. This article checks if the returns to education differs across these geographic regions.

The quarter-of-birth is our instrument. We follow the original paper in creating more instruments by interacting the quarter of bith with the census regions and the years of birth. Adding all these instruments would make us worry about adding "weak" instruments, which would negatively affect inference in the IV-2SLS estimation approach because weak instruments make the IV-2SLS more biased and further away from being normally distributed. Due to this issue, we don't assume that the final estimate is distributed normally. Instead we sample from the posterior of the our instrumental variables estimate.

3 Model

We would like to estimate the heterogeneous impact of schooling on log weekly earnings. The heterogeneity is with respect to the region of birth [2]. Throughout this paper, data for each individual, i, is in row form, and N in the subscript denotes vertically stacked data

[2] For region definitions, see Figure 3. We leave the extension to random treatment effects for future research.

for all the participants.

$$Y_i = S_i \beta_S + S_i G_i \beta_{SG} + G_i \beta_G + X_i \beta_X + \beta_c + \varepsilon_i \tag{1}$$

Y_i is log weekly earnings, S_i is years of schooling, $\underset{1 \times (K_G - 1)}{G_i}$ is a vector of indicators for each region of birth, excluding one to prevent over-saturation. We have a $K_G = 9$ geographic regions. The controls X_i include race and indicators for each age between 1930 and 1940. We have ten controls after leaving out one age indicator and one race indicator to prevent over-saturating the model. Let $K_X = 10$ represent the number of additional controls.

The instrument for schooling is an indicator of whether the respondent was born in the fourth quarter. We have the following reduced form relationships between schooling and the instrument:

$$S_i = Z_i \pi_{S1} + G_i \pi_{G1} + X_i \pi_{X1} + \pi_{c,1} + \eta_{i1} \tag{2}$$

$$S_i G_{k,i} = Z_i \pi_{Sk} + G_i \pi_{Gk} + X_i \pi_{Xk} + \pi_{c,k} + \eta_{ik} \text{ for } k \in \{2, ..., K_G\} \tag{3}$$

Where Z_i is a $1 \times K_Z$ vector of instruments, which includes any interactions between the quarter-of-birth and any other variables. In our analyis, $K_Z = 19$, which represents interacting the quarter-of-birth with region, age, and race indicators. This is the same approach taken in the original paper. Bound et al. (1995) delves into the issue of many and weak instruments to conclude that it is not appropriate to assume the resulting two-stage least squares estimate is normally distributed due to the issue of weak instruments. Chamberlain and Imbens (1996) shows how we can overcome this issue by examining the posterior distribution instead. This article is an extension of the posterior oriented analysis that examines heterogeneous treatment effects.

3.1 Instrumental Variables Assumptions

We have the usual identifying assumptions:

Assumption 1. *The instrument is relevant, excluded, and unconfounded:*

$$\textit{Relevance: } (\pi_S' \pi_S) \text{is } \textit{is invertible} \tag{4}$$

$$\textit{Exclusion: } (Z_i \perp\!\!\!\perp Y_i) | S_i \tag{5}$$

$$\textit{Unconfoundedness: } Z_i \perp\!\!\!\perp (\varepsilon_i, \eta_i) \tag{6}$$

Where $\underset{KZ\times KG}{\pi_S}$ is the matrix of coefficients that predict the treatments from the instruments. Some sufficient conditions for $(\pi_S'\pi_S)$ to be invertible are that $KZ > KG$ (we have more instruments than treatments), and that $\pi_{S,k}$ are all linearly independent (π_S is full column rank). This second sufficient condition means that the set of instruments should impact each treatment differently, otherwise we wouldn't be able to identify the impact two two different treatments from this set of instruments.

3.2 Augmented First-Stage Reduced Form

We have augmented the initial reduced form relationship (Equation 2) with expressions for each interaction between quarter-of-birth and geographic region (Equation 3). This is a departure from widely available "nonparametric heterogeneous instrumental variables" implementations in the Python package EconML (2019). To see why, we only plug one equation (Equation 2) into Equation 1:

$$\begin{aligned}
Y_i = {} & (Z_i\pi_{S1} + G_i\pi_{G1} + X_i\pi_{X1} + \pi_{c,1} + \eta_{i1})\beta_{S1} + \\
& (Z_i\pi_{S1} + G_i\pi_{G1} + X_i\pi_{X1} + \pi_{c,1} + \eta_{i1})G_i\beta_{SG} + \\
& G_i\beta_G + X_i\beta_X + \beta_c + \varepsilon_i && (7)\\
= {} & Z_i(\pi_{S1}\beta_{S1} + \pi_{S1}G_i\beta_{SG}) + && (8)\\
& G_i(\pi_{G1}\beta_{S1} + \pi_{G1}G_i\beta_{SG} + \beta_G) + \\
& X_i(\pi_{X1}\beta_{S1} + \pi_{X1}G_i\beta_{SG} + \beta_X) + \\
& \beta_c + \pi_{c,1}(\beta_{S1}) + \pi_{c,1}G_i\beta_{SG} + \\
& \eta_{i1}\beta_{S1} + \underbrace{\eta_{i1}G_i}_{\text{interaction}}(\beta_{SG}) + \varepsilon_i && (9)
\end{aligned}$$

Due to the interaction between the first stage reduced form errors and the geographic region indicators in Equation 9, the two-stage least squares estimate will still be biased unless we assume the source of heterogeneity, G_i, is a fixed, non-stochastic vector or if we assume G_i is independent from the reduced form errors. This is like saying that any omitted factors that determine the the years of schooling for each person are independent from the place of birth!

An alternative is to instrument for the heterogeneous treatments also, which is why we include first stage reduced form representations of all the interactions between the amount of education with geography. To see this, let's first compactify the notation by representing the treatments by letting $\mathbf{S}_i$ represent all the interacted treatment effects:

$\underset{1\times K_G}{\mathbf{S}_i} \equiv [S_i \quad S_i G_i]$. Then we can stack the coefficients on the treatments β_S and the first stage coefficients on the instruments π_S:

$$\mathbf{S}_i = Z_i\pi_S + G_i\pi_G + X_i\pi_X + \pi_c + \eta_i \tag{10}$$

$$\begin{aligned} Y_i &= \mathbf{S}_i\beta_S + G_i\beta_G + X_i\beta_X + \beta_c + \varepsilon_i \qquad (11)\\ &= (Z_i\pi_S + G_i\pi_G + X_i\pi_X + \pi_c + \eta_i)\beta_S + G_i\beta_G + X_i\beta_X + \beta_c + \varepsilon_i \\ &= Z_i\pi_S\beta_S + G_i(\pi_G\beta_S + \beta_G) + X_i(\pi_X\beta_S + \beta_X) + (\pi_c\beta_S + \beta_c) + \underbrace{\eta_i\beta_S}_{\text{no interaction}} + \varepsilon_i \qquad (12)\end{aligned}$$

Equation 10 is the *first stage*, Equation 11 is the *second stage*, and Equation 12 is the *second stage reduced form*. The standard two-stage least squares estimation proceeds by predicting values for schooling, then regressing log earnings on these predicted schooling levels. Note that the first stage estimation consists of estimating a reduced form specification for each heterogenous treatment effect.

Another way to arrive at the same estimate is to regress *both* the outcomes *and* the treatments on the instruments, then "divide" the two coefficients. In matrix notation "dividing" the two coefficients involves inverting the matrix $(\pi'\pi)$. We have $\beta = E((\pi'\pi)^{-1}\pi'(\pi_{YonZ}))$. Where π_{YonZ} is the regression of Y on the right hand side from the *first-stage*. This is why the relevance condition in Assumption 1 is about the invertibility of a matrix.

In the second stage reduced form (Equation 12), the first stage errors (η_i) do not interact with the geographic indicators. This is the key reason why we choose to estimate additional first stage parameters – we don't need to assume G_i is non-stochastic, or that G_i and η_i are independent. This also simplifies both the computational and analytical calculation of the conditional posterior distributions because we don't have to deal with heterogeneous reduced form variance-covariance matrices.

By estimating more parameters in the first stage, we drastically raise the threshold of "strength of the first stage" under which we would believe that our final two-stage least squares estimate is distributed normally. Staiger and Stock (1994); Stock and Yogo (2002) display how to check the F-statistic of the first stage for conditions when we can assume the two-stage least squares estimate is asymptotically normal. Sanderson and Windmeijer (2016) carries out the same large-sample-but-worsening-instruments analysis for the case with multiple endogenous variables. As noted by many authors (Nelson and Startz, 1990; Bekker, 1994; Phillips, 2009) the actual exact distribution of the two stage least squares estimate is not normal. In the just-identified case the exact distribution of the two-stage least squares estimate doesn't even have finite moments, which would negate the use of bootstrap procedures to estimate the mean.

Instead, we work with the posterior distribution of the two-stage least squares estimate and obtain estimates of confidence from the Bayesian posterior distribution. We follow homogeneous treatment effect examples (Lopes and Polson, 2014; Wiesenfarth et al., 2014) in choosing conjugate priors that are less informative about estimate. Conjugate priors allow us to easily sample from the posterior distribution, save computational time, and write out simple expressions for the conditional posterior distributions.

3.3 Error Structure

Assumption for the Errors/Omitted Variables:

We assume that the errors are normal, and specify priors for the covariances and the coefficients.

Assumption 2. *The errors are multivariate normal with mean zero and variance* Σ*:*

$$(\underset{1\times 1}{\varepsilon_i}, \underset{1\times(K_G)}{\eta_i}) \sim N(\vec{0}, \Sigma)$$

Assuming that the errors are normal is not as strong as assuming the final two-stage least squares estimate is distributed normally. As an example, recall that the most basic two-stage least squares estimate in the single treatment, single instrument case reduces to the ratio of two estimated coefficients. If the errors in the first and second stage are normal, then the two-stage least squares estimate is the ratio of two correlated normal random variables. Fieller (1932) calculates the exact probability density function of the ratio of two correlated normal random variables. In the just-identified case the two-stage least squares estimate doesn't have finite moments (Nelson and Startz, 1990).

3.4 Priors

Prior for the covariance:

We assume the variance Σ is distributed Inverse Wishart (the conjugate prior), with relatively uninformative choices for its parameters:

Prior 1. *Our prior for Σ is Inverse Wishart:*

$$\Sigma \sim IW(n_0 = 0.01, \Sigma_0 = (0.01) * \boldsymbol{I}_{(K_G+1)})$$

Where $\mathbf{I}$ represents th identity matrix, and recall that K_G is the number of geographic regions. Choosing small values for n_0 and Σ_0 are uninformative because if we were to observe N additional data: $\underset{N\times(K_G+1)}{e_N} = (\varepsilon_i, \eta_i)_{i=\{1,\ldots,N\}}$, then the posterior distribution for the variance would be $(\Sigma|e_N) \sim IW(n_0 + N, \Sigma_0 + N\hat{\Sigma}_N)$, where $\hat{\Sigma}_N$ is an estimate of the covariance from the data. Therefore, the small initial values for the shape parameters allow the data have a stronger impact on the posterior distribution of Σ.

We can also represent the variance-covariance matrix for the errors in reduced form:

$$\begin{aligned}
\Omega &= \mathbf{B}\Sigma\mathbf{B}' \\
&= \begin{pmatrix} 1 & \beta_S' \\ \vec{0} & \mathbf{I}_{K_G} \end{pmatrix} \Sigma \begin{pmatrix} 1 & \vec{0}' \\ \beta_S & \mathbf{I}_{K_G} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \beta_S' \\ \vec{0} & \mathbf{I}_{K_G} \end{pmatrix} \begin{pmatrix} \sigma_{\varepsilon\varepsilon} & \Sigma_{\varepsilon\eta} \\ \Sigma_{\eta\varepsilon} & \Sigma_{\eta\eta} \end{pmatrix} \begin{pmatrix} 1 & \vec{0}' \\ \beta_S & \mathbf{I}_{K_G} \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{\varepsilon\varepsilon} + 2\Sigma_{\varepsilon\eta}\beta_S + \beta_S'\Sigma_{\eta\eta}\beta_S & \Sigma_{\varepsilon\eta} + \beta_S'\Sigma_{\eta\eta} \\ \Sigma_{\eta\varepsilon} + \Sigma_{\eta\eta}\beta_S & \Sigma_{\eta\eta} \end{pmatrix} && (13) \\
&= \begin{pmatrix} \omega_{\eta_0\eta_0} & \Omega_{\eta_0\eta} \\ \Omega_{\eta\eta_0} & \Omega_{\eta\eta} \end{pmatrix} && (14)
\end{aligned}$$

Where I have expanded the variance-covariance matrices of the errors into convenient block form, and η_0 is the error from the reduced form regression of Y_i on the right hand side from the first stage. We have a choice of forming our prior for the covariance matrix in terms of Σ or Ω. Chao and Phillips (1998); Dreze (1976) discuss the 1-to-1 correspondence between priors on Σ and priors on Ω. Our current prior puts an identity matrix as the initial shape of the Inverse-Wishart distribution, which corresponds to an initial shape of $\mathbf{BB}'$ on the prior of the reduced form errors.

Prior for the coefficients:

I use a normal prior for the coefficients:

Prior 2. *Our prior for the coefficients is Multivariate Normal:*

$$(\beta, \pi) \sim N(\vec{0}, \mathit{Diag}([\lambda_B, \lambda_P])^{-1}) \tag{15}$$

Where (β, π) are the set of coefficients in the structural model, and $\mathit{Diag}(\nu)$ corresponds to a matrix with the vector ν on the diagonal and zero elsewhere. The main advantage of using a normal prior is because it is the conjugate prior to the specification that the errors are normal.

The normal prior with a diagonal covariance matrix corresponds to using ridge regression with the regularization penalties scaled by λ_B and λ_P. Recall that ridge regression minimizes the squared residuals plus a weighted sum of squares of the coefficients. If we were to simply use ridge regression of Y on X with constant penalties λ, then the analytical solution is: $\beta_{X,\text{ridgereg}} = (X'X + \mathbf{I}\lambda)^{-1}X'Y$, which is the standard analytical OLS solution with the λ added to the "denominator"[3]. Taking the regularization penalties towards zero is the same as increasing the variance of the normal priors to infinity, and this will lead us to the standard OLS analytical solution. Throughout this article, I refer to a "weak" prior as one that has variance set to 1 million, setting the ridge regression regularization penalties to $\frac{1}{(1 \text{ million})}$

Even without a Bayesian interpretation, some amount of regularization is still justified as the dimension of the treatment increases. If we were to attempt to estimate nonparametric returns to schooling, we would need to use regularization even if we did not want to have a Bayesian interpretation of the results because nonparametric instrumental variables are plagued by an "ill-posed inverse" problem. The estimation becomes the issue of regressing earnings on an infinite set of functions of our treatment (schooling). Unfortunately, our instrument doesn't impact all possible functions of the treatment. Newey and Powell (2003) shows that a nonparametric instrumental variables strategy can be consistent when one conducts ridge regression for the outcome on the predicted levels of schooling (the second stage), and reformulates ridge regression as a constrained maximization problem, for which the constraint on the magnitude of the coefficients is assumed to not bind.

4 Estimation

In order to obtain estimates of our confidence in the parameters, we sample from the conditional posterior distributions of each parameter (this is Gibbs sampling). Since all

[3] This is why ridge regression is also called a shrinkage estimator.

of our priors are conjugate distributions, the conditional posteriors take the same shape. First we have the posterior distributions of the variance-covariance matrices:

4.1 Conditional Posterior for the Covariance Structure

$$
\begin{aligned}
(\Sigma | \beta, \pi, RH_{1N}, RH_{2N}, Y_N) &\sim IW(n_0 + N, \Sigma_0 + N\hat{\Sigma}_N), && (16)\\
(\Omega | \Sigma, \beta, \pi, RH_{1N}, RH_{2N}, Y_N) &= \mathbf{B}\Sigma\mathbf{B}' && (17)\\
\underset{N\times(2K_G+K_X)}{RH_{2N}} &\equiv [\underset{N\times K_G}{\mathbf{S}_N}, \underset{N\times(K_G-1)}{G_N}, \underset{N\times K_X}{X_N}, 1] && (18)\\
\underset{N\times(K_Z+K_G+K_X)}{RH_{1N}} &\equiv [\underset{N\times K_Z}{Z_N}, \underset{N\times(K_G-1)}{G_N}, \underset{N\times K_X}{X_N}, 1] && (19)\\
\hat{\Sigma}_N &= [\hat{\varepsilon}_N, \hat{\eta}_N]'[\hat{\varepsilon}_N, \hat{\eta}_N] && (20)\\
\hat{\varepsilon}_N &= Y_N - (RH)_{2N}\beta && (21)\\
\hat{\eta}_N &= Y_N - (RH)_{1N}\pi && (22)
\end{aligned}
$$

Where Y_N is the column of log weekly earnings and RH_{1N} and RH_{2N} are stacked rows of right hand side variables in the first and second stages respectively. We calculate Ω at each step of the Gibbs sampling process because this will make subsequent draws from the conditional posteriors of β and π much easier. As noted above, $\hat{\Sigma}_N$ is the estimated covariance matrix where $[\hat{\varepsilon}_N, \hat{\eta}_N]$ is an $N \times 2$ matrix of the two vectors of predicted errors horizontally stacked together.

4.2 Conditional Posterior for the Coefficients

Since we have chosen conjugate priors, the posterior distribution for β is also normal:

$$
\begin{aligned}
(\beta | \pi, \Sigma, \Omega, RH_{1N}, \widehat{RH}_{2N}, Y_N) &\sim N(b_1, B_1) && (23)\\
\widehat{RH}_{2N} &\equiv [RH_{1N}\pi, G_N, X_N, 1]
\end{aligned}
$$

Now we need to specify what the parameters b_1, B_1 are. Note that the posterior for β only conditions on predicted schooling because we want use the variation induced in schooling by differences in the quarter-of-birth to identify β. This is the same reasoning as in any instrumental variables estimation process.

The posterior parameters for β are:

$$\begin{aligned} B_{1,hom}^{-1} &= Diag(\lambda_B) + \widetilde{\widehat{RH}}_{2N}' \widetilde{\widehat{RH}}_{2N} && (24) \\ B_{1,hom}^{-1} b_{1,hom} &= \widetilde{\widehat{RH}}_{2N}' \widetilde{Y}_N && (25) \\ \widetilde{\widehat{RH}}_{2N} &= \widehat{RH}_{2N} \omega_{\eta_0\eta_0|\eta}^{-0.5} && (26) \\ \widetilde{Y}_N &= (Y_N - \underbrace{(\mathbf{S}_N - RH_{1N}\pi)}_{\hat{\eta}_N} \Omega_{\eta\eta}^{-1}\Omega_{\eta\eta_0})\omega_{\eta_0\eta_0|\eta}^{-0.5} && (27) \\ \omega_{\eta_0\eta_0|\eta} &= (\omega_{\eta_0\eta_0} - \Omega_{\eta_0\eta}\Omega_{\eta\eta}^{-1}\Omega_{\eta\eta_0}) && (28) \end{aligned}$$

The term $(\mathbf{S}_N - RH_{1N}\pi)$ our estimate of the reduced form errors.

Sampling for π looks quite similar to sampling from β, however we should should break π into the coefficients for their individual first stage equations. Recall from Equations (2,3) that we have one first stage equation for each interaction between the schooling and the geographic indicator because we want to avoid having our first stage reduced form errors interacting with the source of heterogeneity. Then we have K_G first stage equations, one for each geographic region, and π is a matrix of coefficients with dimension $(KZ + KG + KX) \times KG$. Let π_k represent a single column of π, then $\pi = [\pi_1, \pi_2, ..., \pi_{K_G}]$. Let $\pi_{(-k)}$ represent all the columns of π except for π_k. We can sample from the conditional posterior of π_k conditioning on $\pi_{(-k)}, \beta, \Omega, Y, RH1_N, R\hat{H}2_N$:

$$(\pi_k|\pi_{(-k)}, \beta, \Sigma, \Omega, RH_{1N}, \widehat{RH}_{2N}, Y_N) \sim N(b_1, B_1) \tag{29}$$

Since $\underset{(KZ+KG+KX)\times KG}{\pi}$ is a matrix of parameters, it will be easier to sample from the posterior distributions of each individual treatment. Let $\underset{(KZ+KG+KX)\times 1}{\pi_k}$ represent the kth column of π, and let π_{-k} represent the other columns. For example, π_1 represents the first stage coefficients that predict the first element of $\mathbf{S}_i$ from RH1.

Now it is easier to express the conditional posterior distribution of the one dimensional vector π_k.

$$(\pi_k|\pi_{-k}, \beta, \Sigma, \Omega, RH_{1N}, \widehat{RH}_{2N}, Y_N) \sim N(p_k, P_k) \tag{30}$$

We can express the posterior distribution of π are

$$
\begin{aligned}
P_k^{-1} &= Diag(\lambda_{Pk}) + \widetilde{RH}_{1kN}^{'}\widetilde{RH}_{1kN} && (31)\\
P_k^{-1}p_k &= \widetilde{RH}_{1kN}^{'}\widetilde{S}_{kN} && (32)\\
\widetilde{RH}_{1kN} &= RH_{1N}\omega_{\eta_k\eta_k|\eta_{(-k)}}^{-0.5} && (33)\\
\widetilde{S}_{kN} &= (S_k - \underbrace{([Y_N, \mathbf{S}_{(-k)}] - [\widehat{RH}_{2N}\beta, RH_{1N}\pi_{(-k)}])}_{[\hat{\eta}_0,\hat{\eta}_{-k}]}\Omega_{\eta_{(-k)}\eta_{(-k)}}^{-1}\Omega_{\eta_{(-k)}\eta_k})\omega_{\eta_k\eta_k|\eta_{(-k)}}^{-0.5} && (34)\\
\omega_{\eta_k\eta_k|\eta_{-k}} &= \omega_{\eta_k\eta_k} - \Omega_{\eta_k\eta_{(-k)}}\Omega_{\eta_{(-k)}\eta_{(-k)}}^{-1}\Omega_{\eta_{(-k)}\eta_k} && (35)
\end{aligned}
$$

Where $\mathbf{S}_{(-k)}$ are the treatments, leaving out the kth treatment, and $\omega_{\eta_k\eta_k}$ refers to the variance of the error in the kth first stage estimating equation. Since all the errors come from the same set of simultaneous equations, it makes sense that the posteriors take the exact same form. Once again, the term $([Y_N, \mathbf{S}_{(-k)}] - [\widehat{RH}_{2N}\beta, RH_{1N}\pi_{(-k)}])$ represents our estimates of the error from the reduced form second stage, and the errors from the first stages excluding the kth error.

The Gibbs sampling procedure using the Julia programming language. I drew $10,000$ samples from the Markov Chain Monte Carlo procedure, using the 2SLS estimates as the initial values and using $2,000$ initial burn-in simulations to allow the draws to stabilize. Although MCMC samples produces autocorrelated, we do not perform any thinning (only taking every nth observation), as measures of percentiles, medians, means, and variances are more precise without thinning (MacEachern and Berliner, 1994; Link and Eaton, 2012).

When we use relatively uninformative priors, we should obtain the same point estimates as the standard 2SLS approach. However, there is a literature on allowing the data to inform us about about the priors (Zellner et al., 2014; Zellner, 1978). Following recent methods Hartford et al. (2017), we use a k-fold cross validation determine the optimal variances in the normal prior (or the optimal regularization parameters.) K-fold cross validation consists of splitting the data into a training and validation set, then training the model on the training set and evaluates the mean squared error of the predicted outcomes on the validation set. I use Julia to discover the regularization penalties that minimize the out-of-sample mean-squared error. Recall that ridge regression regularization penalties correspond to variances for a normal prior.

5 Results

Table 3 summarizes the results of our exercise. We discuss each column from left to right. The first two columns are point estimates and 95% confidence intervals obtained by using standard techniques. All results are "significant". The OLS results indicate that there is a positive empirical relationship between education and earnings, although this relationship cannot be interpreted as causal. The positve point estimates for the IV estimates indicates that there might be a positive causal relationship between education and earnings. A standard interpretation of the 95% confidence interval as derived from the asymptotic approximation of the distribution of the IV-2SLS estimate would lead us to believe that the returns to education are "significantly" positive for all Census Regions.

The third column is the mean and 95% credible interval are gathered by Gibbs sampling 10,000 times from the posterior of the instrumental variables estimate. We use a "weak" prior that assumes the coefficients have variance 1 million (the ridge regression penalty is 1/(1 million)). The point estimates are the mean of the Gibbs samples, while the 95% credible intervals have endpoints determined by the 2.5 and 97.5 percentiles of the samples. Note that the point estimates for the case with weak priors are quite similar to the "standard" IV point estimates, but the 95% credible intervals are so much wider that only one out of nine intervals doesn't include zero. The wider intervals lead us to conclude that it is inappropriate to assume that the "standard" 2SLS estimate is normally distributed with the variance as derived from the delta method.

The final set of estimates is also has a mean and 95% credible interval obtained by Gibbs sampling 10,000 times from the posterior of the instrumental variables estimate. We use K-Fold cross validation to determine that a prior variance of 4.45 for the second stage and upwards of 1 million for the first stage performed the best at out-of-sample prediction (minimizing mean squared error using a two-stage approach).

The rightmost values display the percent of the posterior that lies above zero. This is a measure of our confidence that the returns to education are positive. We conclude that our identification strategy leads us to be 95% confident that the returns to education are postive for the following four census regions: East North Central, East South Central, South Atlantic, West South Central. Figure 4a shows a map of the regions and corresponding proportion of the posterior that lies above zero, representing the posterior chance that the returns to education are positive. Our identification strategy leads us to be 95% confident that the returns to education are positive for the following four census regions: East North Central, East South Central, South Atlantic, West South Central.

6 Conclusion

This article estimates the heterogeneity in the returns to education. We follow Angrist and Krueger (1991) in using the quarter-of-birth as an instrument for the level of schooling obtained by men born between 1930 and 1940. The quarter-of-birth is a "weak instrument" in the sense that we cannot assume the IV-2SLS estimate is distributed normal. This assumption of normality is an appeal to the asymptotic distribution of the two-stage least squares estimate, and calculations of the "standard errors" that are displayed in common analytical programs like Stata report values calculated from applying the Delta-method.

Instead of using the asymptotic distribution, we sample from the posterior distribution of the two-stage least squares estimate. When using a weak prior, our point estimates agree with the two-stage least squares instrumental variables estimates. However our "95%" confidence intervals are much wider, leading us to reject positive returns to schooling for certain regions, while the standard approach would lead to strong, "significant" results for all the regions. We also use cross-validation to determine the priors on the coefficients (same as ridge-regression regularization penalties) that perform the best for out-of-sample prediction using a two-stage least squares approach. The cross validation tells us to use an weak prior for the first stage, but use a stronger prior in the second stage. This empirical result agrees with the observation in Newey and Powell (2003) that some regularization is needed in the second stage to solve the ill-posed inverse problem.

Our specific application is complicated by the necessity of estimating heterogeneous treatment effects. In this article, we interact our proposed treatment (education level) with geographic characteristics. We show that the currently accepted method of only have one first stage equation - which instruments for education - can lead to biased results if the source of heterogeneity is correlated with the omitted factors in the first stage specification. In our case, instrumental variables estimate would still be biased if the region of birth is correlated with any omitted factor that determined the level of schooling. Since this assumption is too strong, instead we instrument for all the heterogeneous treatments, leading us to have multiple first stage estimating equations. Instead of deriving the actual posterior distribution, we iteratively sample from the conditional posterior distributions.

We have two main findings: for men born between 1930 and 1940 and who were induced to attend an additional year of schooling due to the the interaction between compulsory schooling laws and their quarter of birth, we are 95% sure that the returns to education were positive for these four regions, and the returns to education were probably highest for this region [4].

[4] Please refer to Figure 2 for a visual display of the regions, or Table 2 for the exact definitions.

References

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, 657–681.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association 90*(430), 443–450.

Chamberlain, G. and G. Imbens (1996). Hierarchical bayes models with many instrumental variables. *NBER Technical Working Paper*.

Chao, J. C. and P. C. Phillips (1998). Posterior distributions in limited information analysis of the simultaneous equations model using the jeffreys prior. *Journal of Econometrics 87*(1), 49–86.

Dreze, J. H. (1976). Bayesian limited information analysis of the simultaneous equations model. *Econometrica: Journal of the Econometric Society*, 1045–1075.

EconML, M.-R. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML. Version 0.x.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 428–440.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017). Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1414–1423. JMLR. org.

Hoogerheide, L., F. Kleibergen, and H. K. van Dijk (2007). Natural conjugate priors for the instrumental variables regression model applied to the angrist–krueger data. *Journal of Econometrics 138*(1), 63–103.

Leamer, E. E. (2010). Tantalus on the road to asymptopia. *Journal of Economic Perspectives 24*(2), 31–46.

Link, W. A. and M. J. Eaton (2012). On thinning of chains in mcmc. *Methods in ecology and evolution 3*(1), 112–115.

Lopes, H. F. and N. G. Polson (2014). Bayesian instrumental variables: priors and likelihoods. *Econometric Reviews 33*(1-4), 100–121.

MacEachern, S. N. and L. M. Berliner (1994). Subsampling the gibbs sampler. *The American Statistician 48*(3), 188–190.

Nelson, C. R. and R. Startz (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of business*, S125–S140.

Newey, W. K. and J. L. Powell (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Phillips, P. C. (2009). Exact distribution theory in structural estimation with an identity. *Econometric Theory 25*(4), 958–984.

Sanderson, E. and F. Windmeijer (2016). A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics 190*(2), 212–221.

Staiger, D. and J. H. Stock (1994). Instrumental variables regression with weak instruments. Technical report, National Bureau of Economic Research.

Stock, J. H. and M. Yogo (2002). Testing for weak instruments in linear iv regression. Technical report, National Bureau of Economic Research.

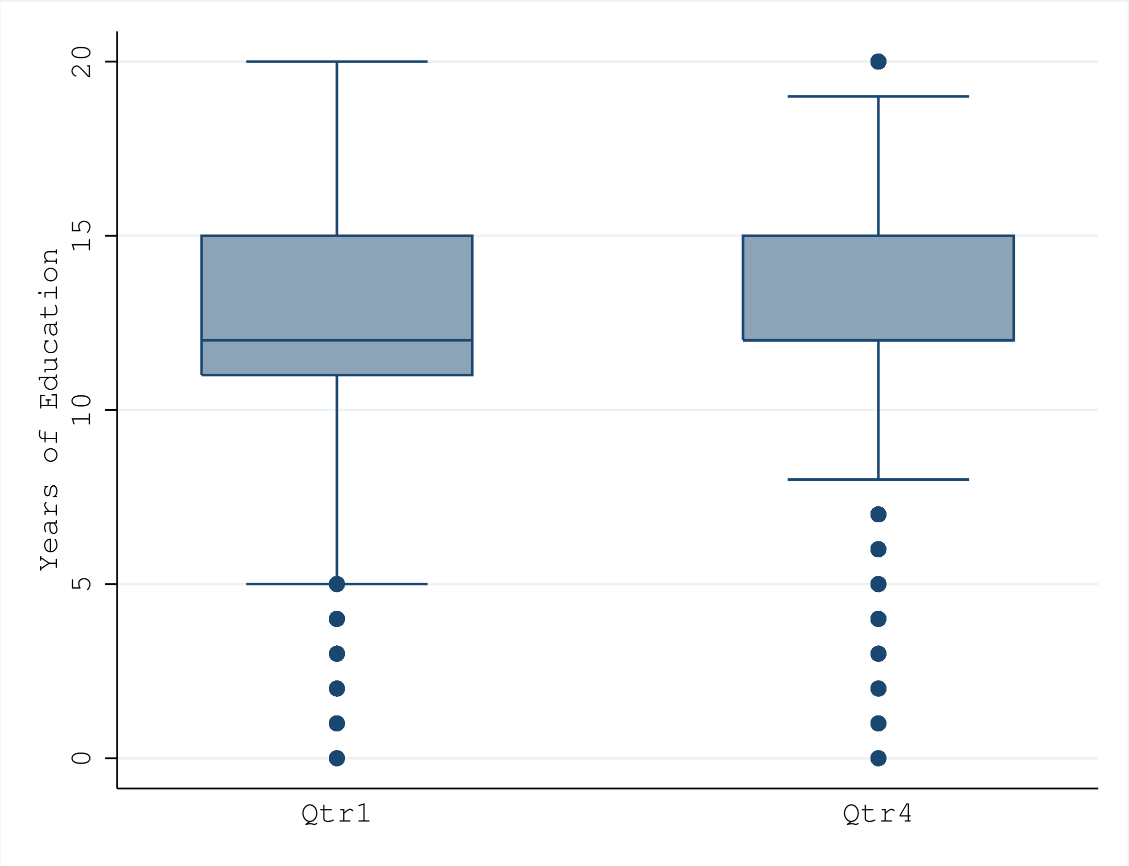
Wiesenfarth, M., C. M. Hisgen, T. Kneib, and C. Cadarso-Suarez (2014). Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. *Journal of Business & Economic Statistics 32*(3), 468–482.

Zellner, A. (1978). Estimation of functions of population means and regression coefficients including structural coefficients: A minimum expected loss (melo) approach. *Journal of Econometrics 8*(2), 127–158.

Zellner, A., T. Ando, N. Baştürk, L. Hoogerheide, and H. K. Van Dijk (2014). Bayesian analysis of instrumental variable models: Acceptance-rejection within direct monte carlo. *Econometric Reviews 33*(1-4), 3–35.

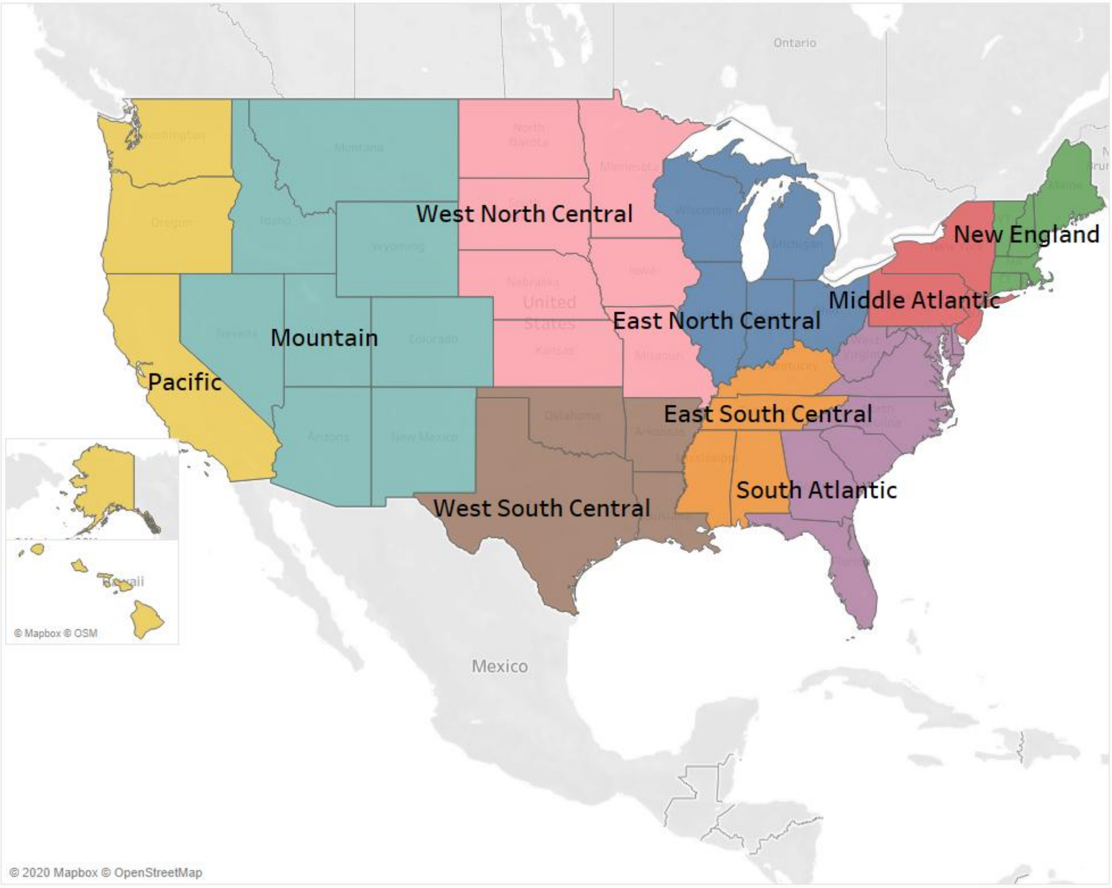
List of Figures

Figure 1: Quarter of Birth Against Years of Education



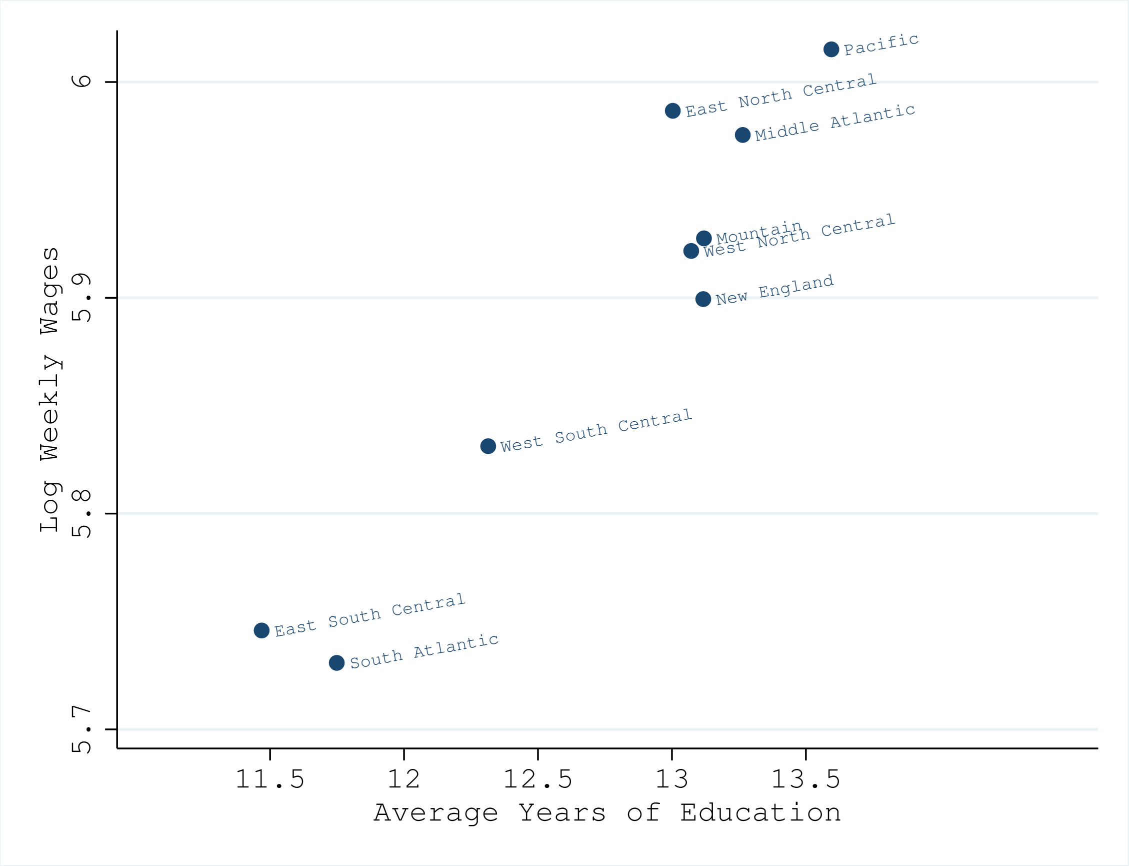
This figure shows the interquartile ranges of the years of education against quarter-of-birth. For census participants born in the fourth quarter, we see that the first quartile of the level of schooling is equal to the median. This suggests there is a binding lower bound to the number of years of education for census participants born in the fourth quarter.

Figure 2: Census Regions



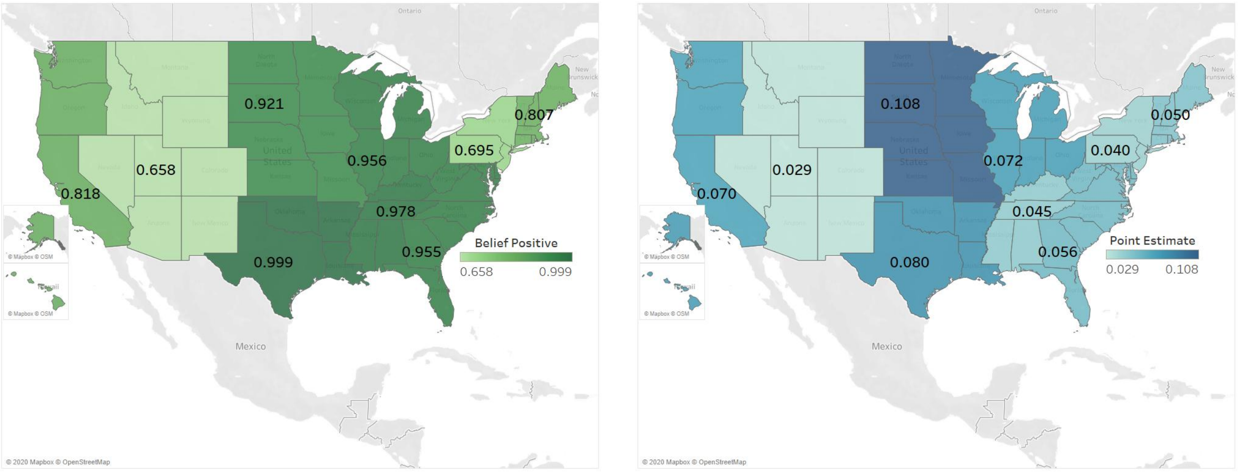
This figure shows the census defined geographic regions. Table 2 lists out the states that make up each region.

Figure 3: Income Against Education by Census Region



This figure shows the relationship between the average log weekly earnings against the average years of education for each census region. We see a general positive trend, but more importantly, we see that each region has differing average levels of weekly earnings and education. This article checks if the returns to education differs across these geographic regions.

Figure 4: The Heterogenous Returns to Education



(a) Percent of Posterior Distribution Positive

(b) Mean of Posterior Distribution

These two maps show summary measures of the posterior distribution of the causal returns to education. We use K-Fold cross validation to determine that a prior variance of 4.5 for the second stage and a large prior variance upwards of 1000 for the first stage performed the best at out-of-sample prediction (using a two-stage approach) for log weekly earnings.

Figure 4a shows the proportion of the posterior that lies above zero, representing the posterior chance that the returns to education are positive. Our identification strategy leads us to be 95% confident that the returns to education are positive for the following four census regions: East North Central, East South Central, South Atlantic, West South Central.

Figure 4b shows the mean of the posterior distribution. Note that a larger mean does not correspond to more confidence in positive returns to education. Although West North Central had the largest point estimate, we aren't even 95% sure that the returns to education are positive for West North Central.

List of Tables

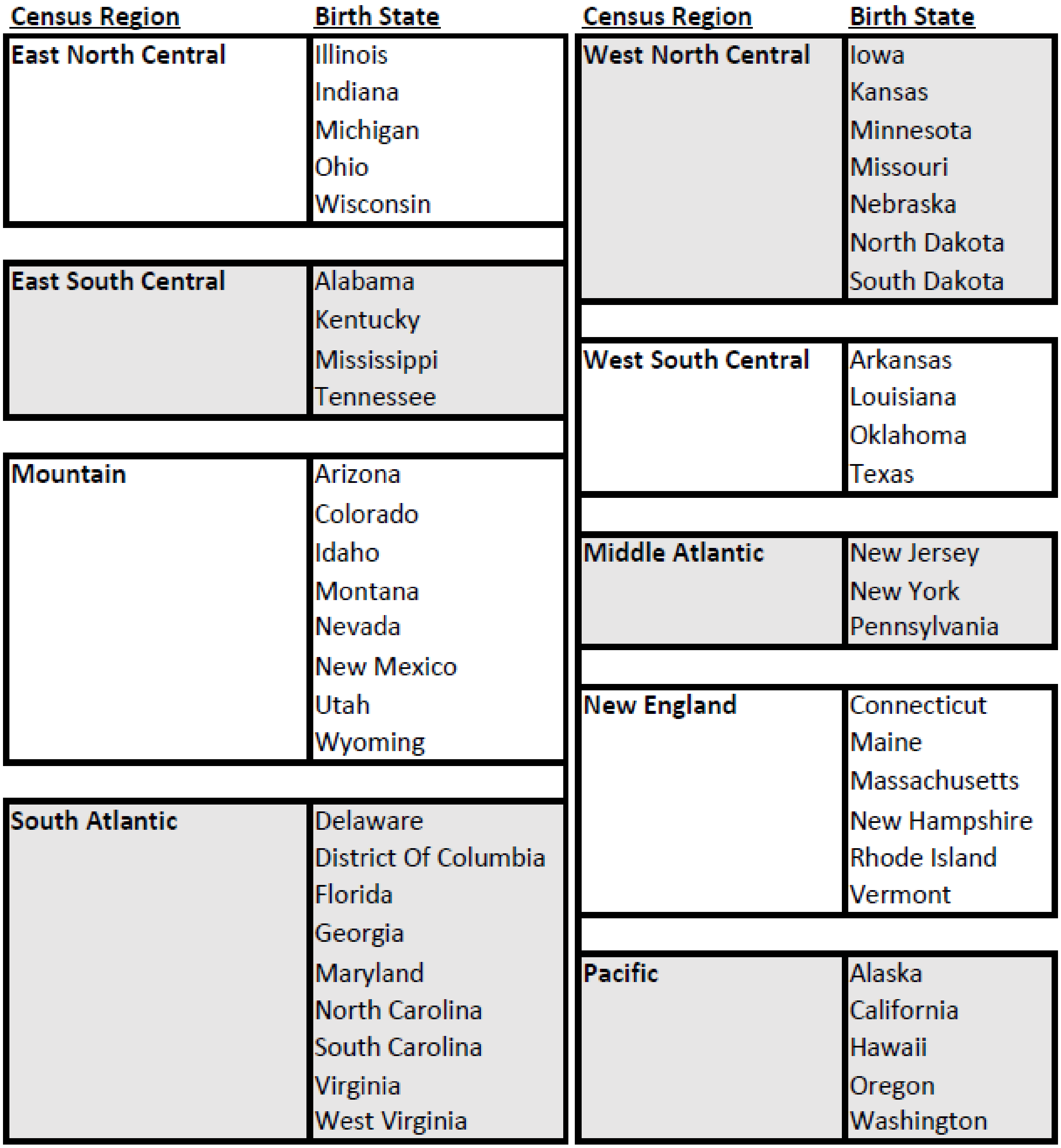
Table 1: Descriptive Statistics

	Mean	SD	10th Pct	90th Pct
Weekly Wages	427.4837	253.9709	179.4828	692.4039
Years of Education	12.68296	3.301034	8	17
Age	44.90577	2.948244	41	49
Married	.8406331	.3660187	0	1
African America	.0946618	.292748	0	0

N = 202,859 individuals in dataset.

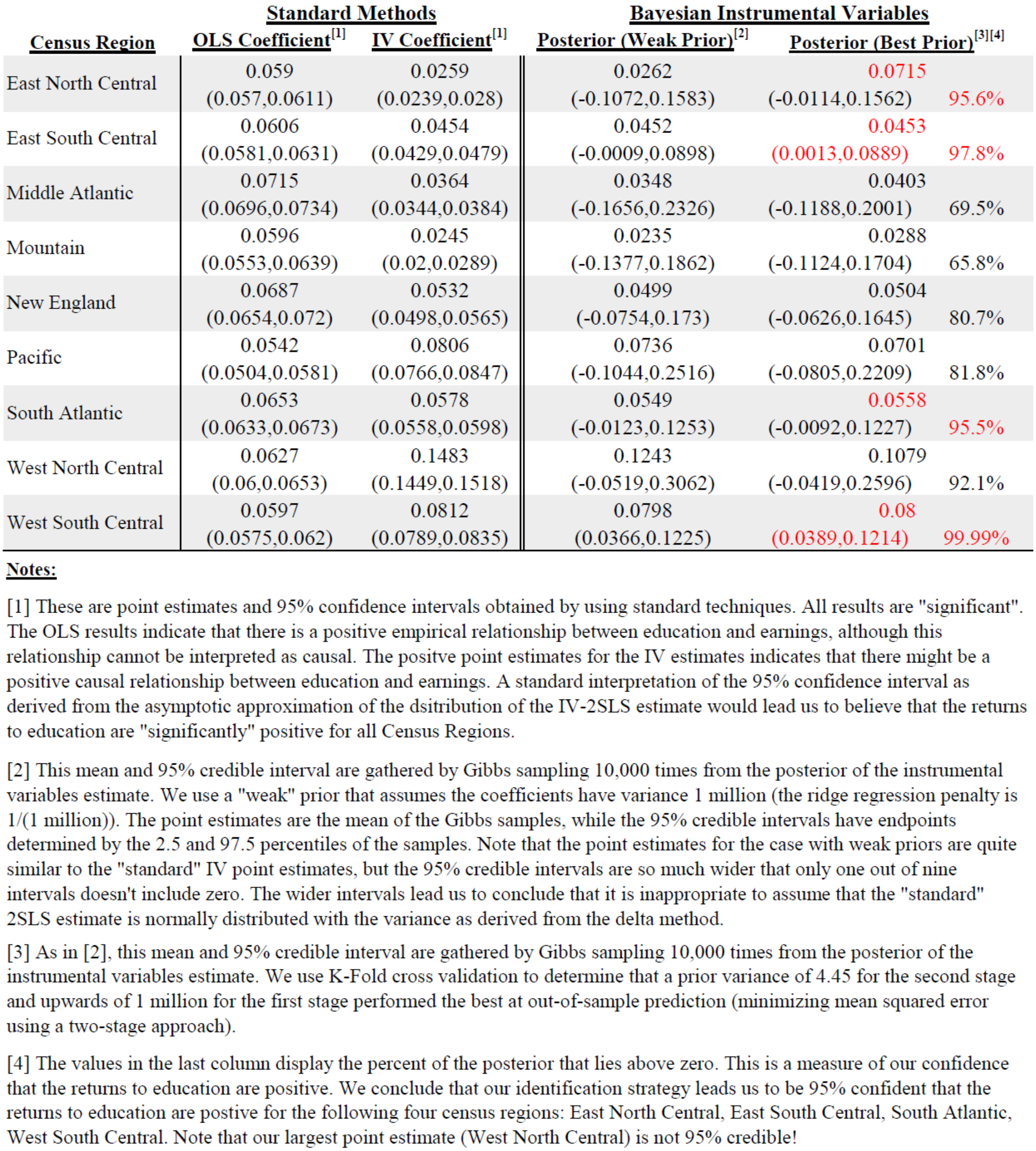
Source: 1980 Census (5% Public Use Sample A). We filtered down to native born African American/Black or Caucasian/white males with birthdays in the first or fourth quarter of a year between 1930 and 1939 (inclusive) with positive wage and salary earnings, and positive weeks worked in 1979.

Table 2: Geographic Regions



Census Region	Birth State
East North Central	Illinois Indiana Michigan Ohio Wisconsin
East South Central	Alabama Kentucky Mississippi Tennessee
Mountain	Arizona Colorado Idaho Montana Nevada New Mexico Utah Wyoming
South Atlantic	Delaware District Of Columbia Florida Georgia Maryland North Carolina South Carolina Virginia West Virginia
West North Central	Iowa Kansas Minnesota Missouri Nebraska North Dakota South Dakota
West South Central	Arkansas Louisiana Oklahoma Texas
Middle Atlantic	New Jersey New York Pennsylvania
New England	Connecticut Maine Massachusetts New Hampshire Rhode Island Vermont
Pacific	Alaska California Hawaii Oregon Washington

Table 3: Results



	Standard Methods		Bayesian Instrumental Variables		
Census Region	OLS Coefficient[1]	IV Coefficient[1]	Posterior (Weak Prior)[2]	Posterior (Best Prior)[3][4]	
East North Central	0.059 (0.057,0.0611)	0.0259 (0.0239,0.028)	0.0262 (-0.1072,0.1583)	0.0715 (-0.0114,0.1562)	95.6%
East South Central	0.0606 (0.0581,0.0631)	0.0454 (0.0429,0.0479)	0.0452 (-0.0009,0.0898)	0.0453 (0.0013,0.0889)	97.8%
Middle Atlantic	0.0715 (0.0696,0.0734)	0.0364 (0.0344,0.0384)	0.0348 (-0.1656,0.2326)	0.0403 (-0.1188,0.2001)	69.5%
Mountain	0.0596 (0.0553,0.0639)	0.0245 (0.02,0.0289)	0.0235 (-0.1377,0.1862)	0.0288 (-0.1124,0.1704)	65.8%
New England	0.0687 (0.0654,0.072)	0.0532 (0.0498,0.0565)	0.0499 (-0.0754,0.173)	0.0504 (-0.0626,0.1645)	80.7%
Pacific	0.0542 (0.0504,0.0581)	0.0806 (0.0766,0.0847)	0.0736 (-0.1044,0.2516)	0.0701 (-0.0805,0.2209)	81.8%
South Atlantic	0.0653 (0.0633,0.0673)	0.0578 (0.0558,0.0598)	0.0549 (-0.0123,0.1253)	0.0558 (-0.0092,0.1227)	95.5%
West North Central	0.0627 (0.06,0.0653)	0.1483 (0.1449,0.1518)	0.1243 (-0.0519,0.3062)	0.1079 (-0.0419,0.2596)	92.1%
West South Central	0.0597 (0.0575,0.062)	0.0812 (0.0789,0.0835)	0.0798 (0.0366,0.1225)	0.08 (0.0389,0.1214)	99.99%

Notes:

[1] These are point estimates and 95% confidence intervals obtained by using standard techniques. All results are "significant". The OLS results indicate that there is a positive empirical relationship between education and earnings, although this relationship cannot be interpreted as causal. The positve point estimates for the IV estimates indicates that there might be a positive causal relationship between education and earnings. A standard interpretation of the 95% confidence interval as derived from the asymptotic approximation of the dsitribution of the IV-2SLS estimate would lead us to believe that the returns to education are "significantly" positive for all Census Regions.

[2] This mean and 95% credible interval are gathered by Gibbs sampling 10,000 times from the posterior of the instrumental variables estimate. We use a "weak" prior that assumes the coefficients have variance 1 million (the ridge regression penalty is 1/(1 million)). The point estimates are the mean of the Gibbs samples, while the 95% credible intervals have endpoints determined by the 2.5 and 97.5 percentiles of the samples. Note that the point estimates for the case with weak priors are quite similar to the "standard" IV point estimates, but the 95% credible intervals are so much wider that only one out of nine intervals doesn't include zero. The wider intervals lead us to conclude that it is inappropriate to assume that the "standard" 2SLS estimate is normally distributed with the variance as derived from the delta method.

[3] As in [2], this mean and 95% credible interval are gathered by Gibbs sampling 10,000 times from the posterior of the instrumental variables estimate. We use K-Fold cross validation to determine that a prior variance of 4.45 for the second stage and upwards of 1 million for the first stage performed the best at out-of-sample prediction (minimizing mean squared error using a two-stage approach).

[4] The values in the last column display the percent of the posterior that lies above zero. This is a measure of our confidence that the returns to education are positive. We conclude that our identification strategy leads us to be 95% confident that the returns to education are postive for the following four census regions: East North Central, East South Central, South Atlantic, West South Central. Note that our largest point estimate (West North Central) is not 95% credible!