THREE ESSAYS ON MIGRATION, OCCUPATIONAL SORTING, AND DEGREE CHOICE: ANALYSES OF SPATIAL AUTOCORRELATION, INCOME, AND THE RACE WAGE GAP

by

Thomas F. Lanier, III

APPROVED BY SUPERVISORY COMMITTEE:

Patrick T. Brandt, Co-Chair

Rodney J. Andrews, Co-Chair

Daniel G. Arce M.

Kurt J. Beron

This dissertation is dedicated

to my wife Laura, and my children Thomas, Matthew and Lucy,

and to my parents Frank and Judy.

Without their support,

it would not have been possible.

In memory of Thomas F. Lanier, Sr.

THREE ESSAYS ON MIGRATION, OCCUPATIONAL SORTING, AND DEGREE CHOICE: ANALYSES OF SPATIAL AUTOCORRELATION, INCOME, AND THE RACE WAGE GAP

by

THOMAS F. LANIER, III, BS, MA, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

ECONOMICS

THE UNIVERSITY OF TEXAS AT DALLAS

December 2018

ACKNOWLEDGMENTS

I would like to thank Patrick T. Brandt, PhD, for his continued support in the writing of this dissertation and for agreeing to be the advisor and co-chair of the committee. He was instrumental in my academic development and deserves the utmost gratitude for his patience and assistance with the development of this dissertation. In addition, I would like to thank Rodney J. Andrews, PhD, for agreeing to co-chair the committee and his support. I express my gratitude to Daniel G. Arce M., PhD, and Kurt J. Beron, PhD, for completing the committee. The committee members are truly appreciated. I would like to thank Donggyu Sul, PhD, for his guidance and counsel throughout the program, and Judy Du for all of her assistance throughout the program. I would like to thank Nathan Berg, PhD, and James Murdoch, PhD, for guidance and assistance in the early development of the first chapter, and Dr. Berg for the opportunity to attend The University of Texas at Dallas.

I would like to thank John E. Scarbrough, PhD, Christopher L. Bartlett, PhD, and all at Litigation Analytics, Inc. for the great and fulfilling employment opportunity. I would like to thank Benjamin D. Hillman, MS, for his work and collaboration in the compilation of the American Community Survey data used in Chapters 2 and 3 of this dissertation. Further, I would like to thank my wife, Laura Lanier, for her help in proofreading initial drafts.

Finally, I would like to thank my family, wife Laura, children and parents Frank and Judy for their continued support and love throughout the PhD program at The University of Texas at Dallas and prior education. Their support was fundamental in the success and completion of the program. It is with sincere gratitude and appreciation for all the support and guidance by all those listed above and those unnamed that had an impact on me throughout my life.

September 2018

THREE ESSAYS ON MIGRATION, OCCUPATIONAL SORTING, AND DEGREE CHOICE: ANALYSES OF SPATIAL AUTOCORRELATION, INCOME, AND THE RACE WAGE GAP

Thomas F. Lanier, III, PhD
The University of Texas at Dallas, 2018

Supervising Professors: Patrick T. Brandt, Co-Chair
Rodney J. Andrews, Co-Chair

This dissertation consists of three essays concerning migration, occupational sorting, and college major choice. They each examine income, Chapter 1 concerning aggregate income at the county level, while Chapters 2 and 3 examine income at the individual level.

Chapter 1 investigates the effects of migration on income. Using migration data from the U.S. Census Bureau and income data from the IRS, the chapter examines adjusted gross income in a county and the effect that migration has on it. It is shown that spatial dependence is present in the data, thus spatial models are applied to the data. A pooled panel spatial Durbin model is used and it is shown that this is the more appropriate of the models considered. The conclusion is that migration positively effects the county experiencing the in-migration, but negatively effects the neighboring counties. The negative effect on the neighboring counties requires further investigation.

In Chapter 2, the race wage gap for high school graduates is studied to determine the effect of occupational sorting. The American Community Survey 1-year PUMS data over the years 2005 to 2012 are used. Sub-samples of 20,000 high school graduates are drawn without restrictions from each of white and black high school graduates for control groups. Using Bayesian methods, marginal posteriors of the parameters are drawn for a quartic specifica-

tion of *Human Capital Earnings Function (HCEF)* for each control group. Then, 20,000 white high school graduates are drawn using the black workers' occupational probabilities as the treatment group. The treatment group's HCEF, along with a posterior predictive distribution generated by the marginal posterior draws, are compared to the control groups. It is shown that occupational sorting accounts for approximately 39 percent of the race wage gap seen in the data, giving credence that occupational sorting as groups exacerbates the race wage gap.

For Chapter 3, the focus shifts from occupational sorting to college major choice. The race wage gap of college graduates with a bachelor's degree as their terminal degree and the effect of college major choice for differing races is examined. The data used in this chapter is the ACS 1-year PUMS from the years 2010 to 2016. Differing groupings of college majors are analyzed by first drawing control groups from white, black and Hispanic college graduates, then drawing treatment groups based on probabilities of college majors from another race/ethnicity. Further, once the quartic specification of the HCEF is estimated using Bayesian methods, the *Oaxaca-Blinder* decomposition is used to evaluate the differences between control and treatment groups. First, all college majors in the data were examined, and no effect on the race wage gap was found. STEM degrees also show no effect on the race wage gap. Business and non-business degrees show competing effects. Non-business degrees actually showed that if white graduates chose college majors like black or Hispanic graduates that their earnings would increase, while in business majors white graduates' earnings decrease when they choose business majors like black and Hispanic graduates. The Oaxaca-Blinder decomposition showed that all of this negative effect was due to the treatment effect while the composition effect was positive.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

MIGRATION AND INCOME ACROSS COUNTIES: EVIDENCE OF MIGRATION'S IMPACT ON TAXABLE INCOME OF A COUNTY USING A POOLED PANEL SPATIAL DURBIN MODEL

1.1 Introduction

Migration is a fundamental aspect of the national economy, also that of states and smaller geographic regions such as counties and towns. The research of this paper seeks to find the impact that migration has on incomes of counties. Determinants of migration have been studied in great detail. Bartolucci et al. (2018) find important factors inducing migration are "differential returns to unobserved ability and differences in employment opportunities between regions." The "dollar value of nonpecuniary gains" from amenities and location match for an individual migrating are higher than gains from migrating to states with higher wages, as found in Kennan and Walker (2011). Kennan and Walker (2011) also state, "Interstate migration is a relatively rare event and...many of the moves that do occur are motivated by something other than income gains." For rurual-to-urban migration, Katz and Stark (1986) find that this migration can be rational, "even if urban expected income is lower than rural income." Haque and Kim (1995) assume that migration decisions are in response to higher rates of return for their human capital in the destination country, stating "The differences in rates of return may arise out of differing government policies or technology and persist even if there is a preference for staying at home." The assumption for this paper is that migration is affected by the structure of the economy following Tiebout (1956) and Diamond (2016) in which "desirable wage and amenity growth" influences "in-migration," while the current economic condition or aggregate income of the county in the period of the migration decision is less of a causal factor. The research questions are: what effect does

population change (focus on migration) have on the income of the county? Is this condition of the county greatly affected by migration? Is this effect economically significant?

Spatial autocorrelation of migration and income with the spatial interaction of the two taken into account is also an interest when considering these variables. This leads to another question: Is spatial dependence a factor in these effects? If so, what is the structure of the model? Spatial dependence occurs when data at one location or geographical unit depend on data from a neighboring location or geographical unit, or when observations are related through spatial networks.[1] With migration and income data of counties, it is a fair assumption that some form of spatial dependence is present and according to the Moran's I measure, that is the case. The data are panel data. Special steps must be taken in the case of spatial panels. Elhorst (2003) goes through specification and estimation of spatial panels, covering fixed effects and random effects versions of the spatial lag and spatial error models. Lee and Yu (2010) discuss the estimation of spatial autoregressive panels with fixed effects, which will be used in this paper for the estimation of the effects of migration on income. Spatial panels or spatio-temporal data have been a neglected area of research in econometric literature. Elhorst (2014a) pulls together specification and estimation procedures for spatial panels from the literature and covers a panel version of the spatial Durbin model. Using a panel spatial Durbin model, it is shown that migration in fact affects incomes, and incomes in neighboring counties are also effected by migration.

The remainder of this paper follows: Section 1.2 discusses the panel data used for the analysis. There is also a discussion on the normalization methods performed on certain variables of the data. Section 1.3 discusses the modelling and the spatial models considered to analyze the data. Section 1.4 presents the results, in which the spatial models better fit the data. The concluding remarks are found in Section 1.5.

[1]See: LeSage and Pace (2009) and Elhorst (2014b)

1.2 Data

The data were a panel data set from the Internal Revenue Service and U.S. Census Bureau.[2] The data are cross-sectional time series for 3140 U.S. counties from 2004-2009. The data are built from county income data from the Internal Revenue Service and population data from the U.S. Census Bureau over the six years. The IRS data are available in individual years and must be combined according to the state and county FIPS codes. The Census Bureau data contains the migration, birth and death statistics used in the study. During these six years, two "counties" were formed in Alaska: Petersburg Census Area and Hoonah-Angoon Census Area. These counties have been dropped from the data because they did not have six observations. Kalawao County, Hawaii was also dropped for extreme and unrealistic data.

County census data from ESRI ArcGIS 10 are utilized to build the spatial models. Using the State Plane 83 projection, a spatial weights matrix is created accounting for the eight nearest neighbors. The ArcGIS census data are then merged with one year of the data discussed above to match the counties in the 3 data sets and allow for the spatial weights matrix to be matched to the data.

Pooled summary statistics are presented in Table 1.1. The measure used for income in the study is adjusted gross income. Adjusted gross income is defined as the amount of taxable income in the county in a given year in dollars. This measure has a major size difference between the upper and lower bounds, as one can see in Figure 1.1, it is highly skewed due major outliers in the upper bound. Two different measures are used to standardize the variable. First, the population estimates for each county for each year are used to create AGI *per capita*. The number of tax returns received from a county in a given year are also used to standardize the AGI measure. This is essentially a measure of adjusted gross income per tax paying household. One cannot really take anything away from AGI as it is, but

[2]http://www.irs.gov/uac/SOI-Tax-Stats-County-Data
http://catalog.data.gov/dataset/current-population-survey

Table 1.1. Pooled Summary Statistics

Variable	Obs	Mean	Median	Std..Dev.	Min	Max
AGI	18,840	$2.13x10^9$	$4.08x10^8$	$7.27x10^9$	$7.10x10^5$	$2.18x10^{11}$
per capita	18,840	17710.00	16830.00	5630.00	204.20	76570.00
per household	18,840	44150.00	42020.00	11670.00	7133.00	142000.00
Net Migration	18,840	293.30	1.00	3999.00	-248100.00	110500.00
per capita	18,822	0.00	0.00	0.01	-0.10	0.10
per household	18,784	0.00	0.00	0.03	-0.20	0.20
Births	18,840	1341.00	328.00	4714.00	0.00	152000.00
per capita	18,840	0.01	0.01	0.00	0.00	0.03
per household	18,818	0.03	0.03	0.01	0.00	0.10
Deaths	18,840	777.60	258.00	2152.00	0.00	61300.00
per capita	18,840	0.01	0.01	0.00	0.00	0.03
per household	18,835	0.03	0.03	0.01	0.00	0.08
Population	18,840	95560.00	25310.00	308500.00	40.00	9848000.00

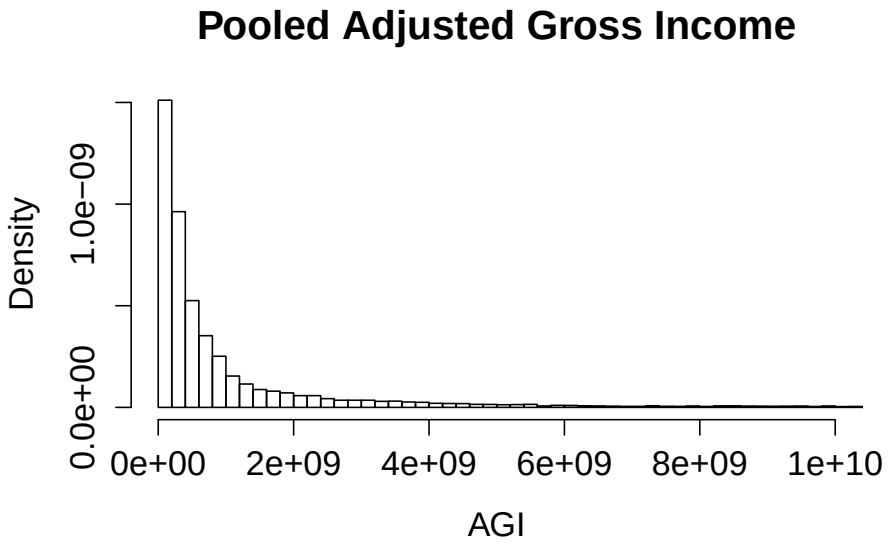

Figure 1.1. AGI

looking at AGI *per capita* and AGI per household summary statistics provide a nice insight into the average county and the ranges of counties in this study. The mean for AGI *per capita* is around $17,000, which is about where it would be expected since the denominator in the variable contains children and retirees which do not receive taxable income in most cases. AGI per household has a mean of around $44,000 with a median of $42,000. These variables are considered the dependent variables, however only AGI *per capita* will be used in the analysis section of the study. The histograms for AGI per capita for each year can be found in Figure 1.2.

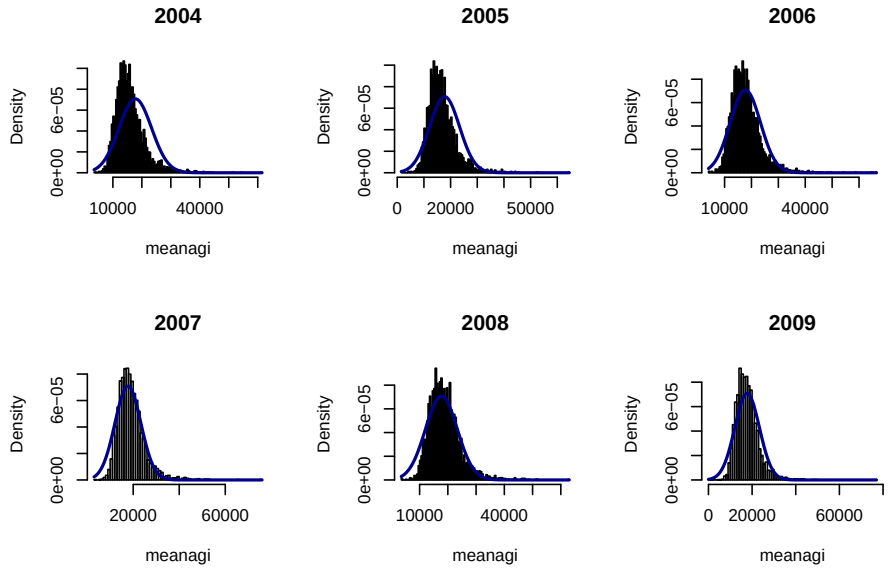

Figure 1.2. AGI per capita

The explanatory variables are presented in those three forms (levels, *per capita*, and per tax paying household), also. Net migration is the variable of concern and its effect on adjusted gross income. Net migration is simply the inflow minus the outflow of people in the county over the given year, in other words it is the net gain of people due to migration from another county and/or another state not including international immigrants. The data has net migration for each county in each given year (*netmig*), which is measure in single units (a

person). The same standardization approach is used for net migration as for adjusted gross income. Net migration as a percentage of population is the explanatory variable that will be used throughout the paper. Net migration divided by tax returns received from a county in a given year is just another standardized net migration measure, but does not allow for the same inference and interpretations as net migration *per capita* does for the study. The histograms for net migration *per capita* are found in Figure 1.3.

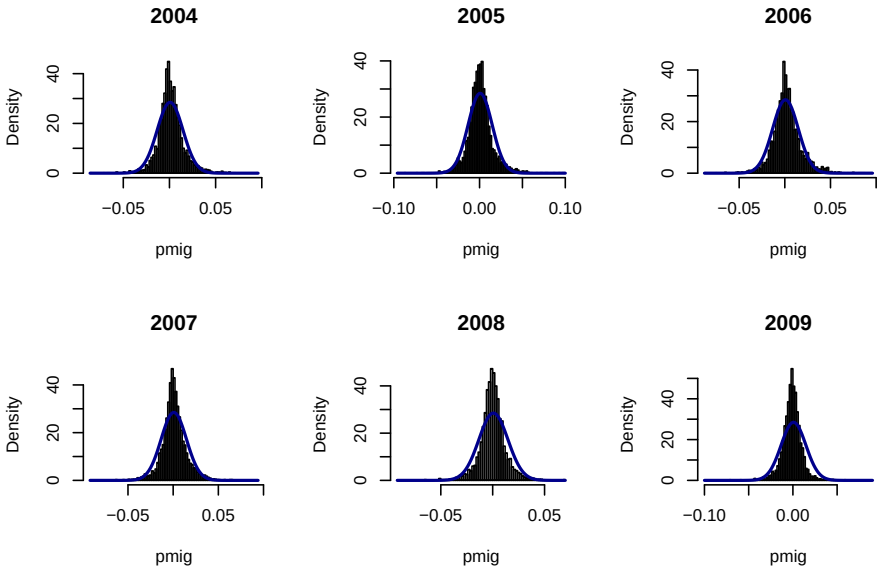

Figure 1.3. Migration flows as a percentage of population

The data also has births and deaths in a county for a given year, which are used as control variables. Births and deaths are reported in single unit level (again, a person), and divide each of these by population making births *per capita* and deaths *per capita*, respectively, as the control variables for the study. The variables births per household and deaths per household are standardized by dividing births and deaths by the number of tax returns, respectively. With these variables there were some extreme observations, e.g., after hurricane Katrina a mass exodus ensued from New Orleans. This led to net migration *per capita* needing to be bounded between [-.1,.1] because of these extreme cases. This

truncation dropped 16 observations from 18,840 total.[3] Net migration per household, births per household, and deaths per household were also truncated for extreme observations losing 8, 22, and 5 observations, respectively. For the purpose of this paper, net migration *per capita*, births *per capita*, and deaths *per capita* are the explanatory variables. This will give results one could look at from a *per capita* viewpoint. The reason for using *per capita* data is due to size distortions in counties and the multi-colinearity problem between births and deaths which is shown in Table 1.2.

Table 1.2. Correlation Table for Levels Model

	AGI	Net Mig.	Births	Deaths	Pop.
AGI	1.00				
Net Mig.	-0.11	1.00			
Births	0.95	-0.11	1.00		
Deaths	0.95	-0.15	0.96	1.00	
Pop.	0.97	-0.13	0.99	0.98	1

1.3 Models

The non-spatial panel model with fixed effects is first estimated, and a Hausman test is used to specify between fixed effects and random effects, the null hypothesis that random effects estimators would be consistent and efficient is rejected.

$$y_{it} = \alpha_i + \beta x_{it} + \gamma_1 z_{0it} + \gamma_2 z_{1it} + \tau_t + \epsilon_{it} \tag{1.1}$$

In this model, y_{it} is AGI *per capita* for county-year, α_i is the county fixed effect for county-year, x_{it} is net migration *per capita* for county-year, $z_{\theta it}$ are the control variables births *per*

[3]Truncations - 2004: Blaine County, OK; Forest County, PA; Loving County, TX.
2005: Forest County, PA; Loving County, TX.
2006*: Pinal County, AZ; Chattahoochee County, GA; Cameron Parish, LA; Orleans Parish, LA; Plaquemines Parish, LA; St. Bernard Parish, LA; Hancock County, MS; Harrison County, MS.
2007: Orleans Parish, LA; St. Bernard Parish, LA.
2008: Kiowa County, KS; St. Bernard Parish, LA; Loving County, TX.

capita ($\theta = 0$) and deaths *per capita* ($\theta = 1$), θ is an index for control variables, and τ_t is the time fixed effects. Time dummy variables are used to substitute for the time fixed effects, which allows to easily estimate those effects. With assumptions of the fixed effects model, the error term is assumed to be $\epsilon_{it} \sim iidN(0, \sigma^2)$. Even with the spatial autoregressive nature of the data, this estimation produces consistent and unbiased results without the spatial autocorrelation carrying over into the error term. The non-spatial fixed effects model will, however, miss important spatial relationship lying in the data.

When dealing with migration from county to county a non-spatial model can leave out very important relationships between counties in migration or income in this case. This can lead to bias in the parameter estimates of the model.[4] More insight can be gained from working with a spatial model that can account for the spatial autocorrelation. LeSage and Pace (2009) begin the spatial models with the simple spatial lag model or spatial autoregressive (SAR) model where the dependent variable has an explanatory spatial lag. To build these spatial models, an object is needed that will introduce the spatial lags into the model and that is a spatial weights matrix, denoted as ω. For each county, the eight nearest neighbors will receive a 1 in the matrix and every other county will be given a 0. The advantage of using the K nearest neighbors approach is "that it ensures there will be some neighbors for every target feature, even when feature densities vary widely across the study area."[5] When dealing with counties, it essential to use a method that allows for varying densities. In addition, "The K nearest neighbors option with 8 for Number of Neighbors is the default conceptualization used with Exploratory Regression to assess regression residuals."[6] This spatial weights matrix is time invariant so building the W for time and space, the spatial

[4]It can lead to bias, however the fixed effects model stated above is not biased by spatial autocorrelation in the error according to a Moran's I test of the error term for the non-spatial fixed effects model.

[5]*Esri ArcGIS, s.v.* "Modeling spatial relationships," accessed September 18, 2018, `http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm`

[6]Ibid.

weights matrix ($N \times N$) from ArcGIS 2010 is inserted in each diagonal (3140×3140) six times to make W (18840×18840). Every other term in this W is zero. Now, the SAR model from LeSage and Pace (2009)[7]:

$$y = \rho W y + X\gamma + \epsilon \tag{1.2}$$

This model is presented in matrix form where X is a matrix (18840×8) of the explanatory variable and control variables including the time dummies and y is a vector (18840×1). Another spatial model that needs to be estimated is the spatial error model (SEM). This model has a spatial lag term in the error, which implies a different structure. The same spatial weights matrix can be used to build this particular model from LeSage and Pace (2009):

$$y = X\beta + u \tag{1.3}$$

$$u = \rho W u + \epsilon \tag{1.4}$$

Rearranging the terms produces a geometric series of lagged ϵ's.

$$y = X\beta + (I - \rho W)^{-1}\epsilon \tag{1.5}$$

In this model, the error term contained the spatial lag. These models are special cases of a more general model known as the spatial Durbin model (SDM). Not only does the spatial Durbin have the spatial lag for the dependent variable, but also the explanatory variables have spatial lags. The spatial Durbin model can be either SAR or SEM. The spatial Durbin model is found in Elhorst (2014a):

$$y = \rho W y + X\delta_1 + WX\delta_2 + \alpha + \tau + \epsilon \tag{1.6}$$

$$\delta_1 = \beta + \gamma \tag{1.7}$$

$$\delta_2 = -\rho\beta \tag{1.8}$$

$$\epsilon \sim N(0, \sigma^2 I) \tag{1.9}$$

[7]Spatial models will be expressed in matrix form

This is the model which best fits the data in this study according to specification tests suggested by Elhorst (2014a). In this model, α is the county fixed effects and τ is the time fixed effects. Since, this is a more general approach to the earlier spatial models, the spatial Durbin model can be further specified to the prior models. If $\beta = 0$ in the spatial Durbin model, it becomes the SAR, and if $\gamma = 0$, then it becomes an SEM model.

Elhorst (2014a) provides a specification testing procedure to determine which of these models best fits the data,

> First, the non-spatial model is estimated to test it against the spatial lag and the spatial error model (specific-to-general approach). In case the non-spatial model is rejected, the spatial Durbin model is estimated to test whether it can be simplified to the spatial lag or the spatial error model (general-to-specific approach). If both tests point to either the spatial lag or the spatial error model, it is safe to conclude that model best describes the data. By contrast, if the non-spatial model is rejected in favor of the spatial lag or the spatial error model while the spatial Durbin model is not, one better adopts this more general model. (Elhorst, 2014a)

These methods are used to specify the model that best fits the data with likelihood ratio tests. Once these specification tests lead to the best fit for the data, just looking at the coefficients will not be sufficient, so it is a necessity to calculate the direct, indirect, and total effects of migration on the income of a county. These effects can be calculated by building a direct-indirect effect matrix, where the diagonal terms are the direct effects of each county and the off-diagonal terms are indirect effects of county i on county j. According to Elhorst (2014a), this matrix is independent of time, so it is equivalent to the matrix presented in LeSage and Pace (2009):

$$(I - \rho W)^{-1} \begin{bmatrix} \delta_{1k} & \cdots & w_{1N}\delta_{2k} \\ \vdots & \ddots & \vdots \\ w_{N1}\delta_{2k} & \cdots & \delta_{1k} \end{bmatrix} \tag{1.10}$$

The average direct effect is the trace of this matrix divided by N, in this case 3140. This gives us the average effect net migration has on a county's adjusted gross income. The average

indirect effect is the row or column sums minus the diagonal terms divided by N. This will show the average effect on the neighboring counties from net migration into a county. The average total effect is the row or column sums divided by N. This is the aggregate average effect that should be seen from net migration into or out of a county. These calculations are necessary to see, because unlike a non-spatial model the coefficients for the variables in the results do not tell the whole story. The results from these calculations are very interesting and will be presented in the next section of the paper.

1.4 Results

The estimation of the non-spatial model is a fixed effects panel model with independent county and time fixed effects as proposed in equation 1.1. The maximum likelihood estimator was used to check the results of the fixed effects non-spatial model. These results show that for a 1% increase in net migration there is an increase of approximately \$150-\$170 in AGI *per capita*. The fixed effects and the maximum likelihood estimators are very similar. Since the main concern is the fixed effects estimates, what does this say about the controls? With this model we get a \$304 *per capita* increase in AGI with every one percentage point in the birth rate, and a \$490 *per capita* decrease in AGI with every one percentage point in the death rate. This model is explaining 46% of the variance in AGI *per capita* according to the R-squared for the model. These results are interesting, but could be biased if there is spatial dependence at play in the data. Again, these particular results are not biased due to spatial dependence because of the fixed effects estimation, but the data has much more to tell than the results gained from the non-spatial model.

Using Moran's I for each year, there is statistically significant positive spatial dependence with the I factoring in at about .5 for each year. Not only does the Moran's I show there is spatial dependence, the non-spatial model is rejected in accordance with Elhorst (2014a) specific-to-general approach for the SAR (1.2) and SEM (1.3).

Table 1.3. Non-Spatial models results

AGI *pc*	Fixed Effect	FE w/Controls	FE w/Time Dummies	MLE
Net Mig. *pc*	9982.20	10154.7	15283.06	16957.07
	(6.61)	(6.82)	(13.56)	(15.14)
Births *pc*		214538	30418.74	9448.92
		(15.76)	(2.93)	(0.95)
Deaths *pc*		-180000	-49016.5	-93847.65
		(14.98)	(5.35)	(10.46)
2004			*Base Year*	*Base Year*
			(0.00)	(0.00)
2005			817.28	814.31
			(23.86)	(23.76)
2006			1894.89	1884.59
			(55.16)	(54.83)
2007			3198.12	3201.03
			(92.29)	(92.41)
2008			2957.42	2937.98
			(84.59)	(84.09)
2009			1861.61	1849.31
			(53.45)	(53.10)
σ_u	5344.03	5286.34	5289.15	5206.61
σ_e	1837.41	1808.95	1356.17	1357.16
ρ	0.89	0.89518	0.94	0.94
Constant	17705.53	16750.8	16015.57	16736.30
R^2	0.05	0.03	0.46	

Table 1.4. Moran's I 2004-2009

	2004	2005	2006	2007	2008	2009
I	0.52	0.51	0.49	0.49	0.52	0.50
K	9.69	9.44	10.33	10.24	7.39	10.77

Once the non-spatial model is rejected for a spatial specification, a pooled spatial Durbin model (1.6) is run and tested against the SAR (1.2) and SEM (1.3). The results for this specification test reject both the SAR and SEM in favor of the pooled SDM with statistics of 194.9491 and 470.4673, respectively. The spatial Durbin model is accepted as the best fit for the data. However, more specification tests need to be run to test if the pooled spatial Durbin model is sufficient against the spatial Durbin accounting only for county fixed effects, only

for time fixed effects, or both county and time fixed effects. According to the likelihood ratio tests, the pooled spatial Durbin model with the time dummies included (1.6) is sufficient. The results for the pooled spatial Durbin model with time dummies, the individual and time fixed effects results, and results from a simple SAR model can be found in Table 1.5.

Table 1.5. Spatial models results

AGI *pc*	Pooled SAR	SDM w/iFE & tFE	Pooled SDM	Pooled SDM w/tFE
Net Mig. *pc*	53110.276	6261.762	67377.508	67562.366
	(26.207)	(6.651)	(28.986)	(29.097)
Births *pc*	-214603.304	-4497.594	-226936.412	-228652.391
	(-21.782)	(-0.534)	(-19.999)	(-20.144)
Deaths *pc*	-374623.076	-12462.934	-346686.011	-348573.040
	(-37.358)	(-1.662)	(-27.684)	(-27.744)
Lag Mig.		23149.470	-51461.211	-50124.116
		(11.227)	(-13.035)	(-12.583)
Lag Births		108917.262	-5082.614	-28483.074
		(5.301)	(-0.262)	(-1.391)
2004				$Base\ Year$
				(0.00)
2005				211.235
				(2.194)
2006				474.426
				(4.898)
2007				1008.999
				(10.182)
2008				682.045
				(6.936)
2009				415.330
				(4.252)
ρ	0.724	0.540	0.710	0.697
Constant	11335.076		12906.997	13127.918
	(91.874)		(50.223)	(29.361)
R^2	0.539	0.962	0.541	0.542
Log-Likelihood	-183013.600	-159157.660	-182920.460	-182848.900

The pooled SDM with time dummies model gives an R-squared of .54 and a $\rho = .69$. With the results from this model, the direct-indirect matrix (1.10) must be constructed. The average direct effect is 67,100, which means that a 1% increase in net migration leads to an

increase in AGI *per capita* of $671 ($0.30) on average. The average indirect effect is -9,980, which means a 1% increase in net migration in a neighboring county will cost the county $95 ($0.24) *per capita* on average. This makes the average total effect of a 1% increase in net migration about $571 *per capita* with the spillover effect about 15% of the magnitude of the direct effect. Using a non-spatial model would likely overestate the average total effect by assuming the spatial spillover to be zero.

1.5 Conclusion

According to the specification tests the pooled spatial Durbin model is the best fit for the data with the models considered to this point. There is a significant positive effect on a counties taxable income from migration. Although the measure is in *per capita* terms, it is not certain that the effect is felt throughout the county's economy. This effect could be felt only by the migrant, a few people, or could flow through the entire county. This is an average measure, so it is quite safe to assume that not all residents of a county feel the impact from migration, however due to the neighboring counties being affected it is also safe to assume that the effect is not quarantined to only a few residents of the county. These migrants would not only increase demand in the county, but also following economic growth theory, supply increases just from the added labor input in the county, not even taking into account the physical and human capital and entrepreneurial ability the migrants might also bring in. Supply increasing in the county would lead to the "trickle-down" effect throughout the economy of the county. The counties with positive net migration are beneficiaries of what Haque and Kim (1995) refer to as "brain drain." Haque and Kim (1995) finds that "brain drain reduces the growth rate of the effective human capital that remains in the economy and hence generates a permanent reduction of *per capita* growth in the home country." The results of this paper speak to the other side of this effect where growth (retraction) is observed in the counties with positive (negative) net migration. The indirect effects of

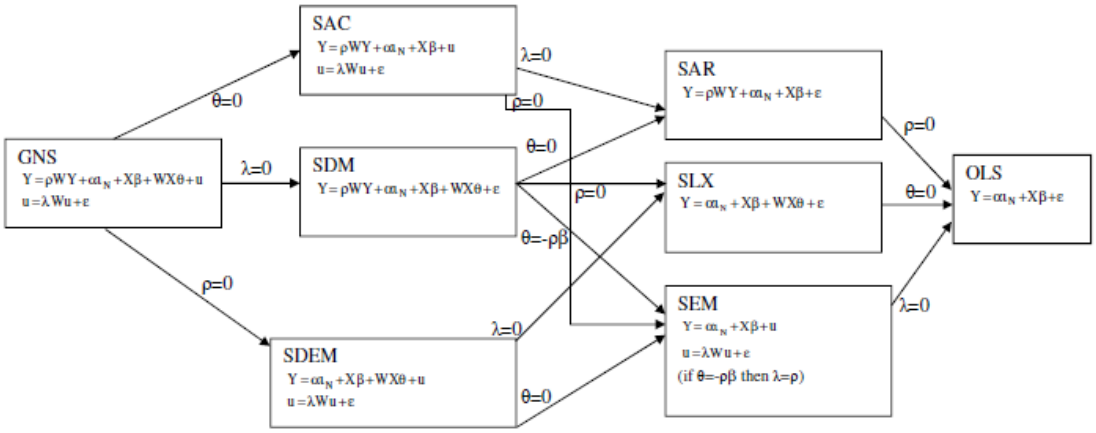

Figure 1.4. Spatial Econometric Models. *Source: Elhorst and Vega (2013)*

migration are negative for neighboring counties, although that effect needs closer scrutiny before specific conclusions can be made.

Spatial dependence is a major factor in the data, as seen from Moran's I and the results. The differing models of spatial dependence can be seen in Figure 1.4 from Elhorst and Vega (2013). Once the pooled SDM was specified, the difference can be seen between the non-spatial model estimations and the spatial model estimates. This difference is about $500 *per capita*, which is quite large speaking in economic significance terms. Also, the ρ can show that the spatial dependence is quite strong in the data. These results could still be biased according to Lee and Yu (2010) where they state that the coefficient estimates and variance estimates could be biased. Future research on the data will need to be conducted to test the validity of the pooled SDM apart from the fixed effects SDM models to ensure the indirect effects are actually negative along with modeling the causes of migration to drive the effects seen in the results of this study. As of now, it is concluded that there is an impact on a county's income and that it is economically significant with a spatial dependence of the neighboring counties.

CHAPTER 2

A BAYESIAN APPROACH TO OCCUPATIONAL SORTING AND THE EFFECT ON THE RACE WAGE GAP

2.1 Introduction

Merriam-Webster defines inequality as "the quality of being unequal or uneven; such as: a) lack of evenness, b) social disparity, or c) disparity of distribution or opportunity."[1] Income inequality is the "disparity" in the distribution of earnings, wages, and/or income. In general, income inequality refers to the disparity of income across the quantiles, from the lowest earners to the highest. The mere existence of inequality does not signal a problem that needs to be addressed because differing incentives and payoffs for differing labor market activities drive innovation and investment in human capital. As stated in Krueger (2018), the issue of income inequality has negative causes, such as when it appears for non-productive reasons. This could be due to racial, gender, class or any other demographic based discrimination, or anti-competitive activities in the market.[2] The non-productive or demographic based discriminatory causes of income inequality are not only destructive to the individual who is the victim, but also the firm, economy at large, and society.[3]

Differences in the earnings between various races continues to be a subject of concern for society and economists. Understanding these differences or gaps in earnings is essential to addressing the issue.[4] Altonji and Blank (1999) state that "race and gender differentials

[1] *Merriam-Webster, s.v.* "inequality," accessed June 30, 2018, `https://www.merriam-webster.com/dictionary/inequality`

[2] An example of anti-competitive activity could be those more well off using political capital to influence the market in their favor as Krueger (2018) notes.

[3] Lang et al. (2005) propose a model that shows discrimination reduces wages for both white and black workers and reduces total output.

[4] This is assuming that the gaps are caused by negative factors, e.g. racism.

in the labor market remain stubbornly persistent," and *prima facie*, "black and Hispanic men as well as white women earn about two-thirds of that earned by white male workers on an hourly basis." Carruthers and Wanamaker (2017) look at the role of human capital accumulation through schooling and how it affects the gap between black and white earnings. They find that if school quality reaches parity between black and white workers, the earnings gap significantly closes.[5]

The choice of occupation is a major determinant of income, and if it is truly a free choice, seen as a positive cause of income inequality. Although, there exists the possibility that racial discrimination affects occupational sorting. Even the perception alone that discrimination is present in the hiring process could cause black workers to sort to lower paying jobs.(Lang et al., 2005) This paper seeks to look at the relationship of occupational sorting and the earnings gap between black and white workers. Occupational sorting by race is thought to lead to significant differences in earnings. Carruthers and Wanamaker (2017) state that occupational sorting is "one barrier to higher wages." While specific to one firm in the study, Penner (2008) finds that occupational sorting "accounts for the substantial race and gender differences in salary at the point of hire in this firm."[6] This paper focuses on the question: Using a cross-sectional *Human Capital Earnings Function (HCEF)*, would white workers' earnings be significantly lower if their occupational distribution mimicked that of black workers?

To explore the question above, Bayesian inference and methods are employed. Bayesian methods provide many benefits to data analysis and several are key to this study.[7] For example, an individual's realized Human Capital Earnings Function will differ from the

[5]Heckman et al. (2000) acknowledges similar findings in the literature.

[6]Grodsky and Pager (2001) finds 20 percent of the wage gap is due occupational sorting. They also point out that wage gaps within occupations are heterogeneous.

[7]Gill (2014) provides an explanation of the many benefits of Bayesian approaches.

average found in this type of analysis, and the interpretation of intervals, even in a frequentist presentation, would most likely be Bayesian by most consumers of the study.[8] As will be discussed later, Bayesian methods allow posterior distributions of the parameters for a control group to be used as prior distributions of the parameters for the treatment group to enhance the balance of the matching. Further, the inspiration of the study to draw sub-samples of white workers from the black workers' occupational densities (quasi-prior distribution) is Bayesian.

The remainder of this paper follows: Section 2.2 discusses the data and sub-samples used for the analysis. There is also a discussion on the matching balance of the samples and sub-samples between white and black workers. Section 2.3 discusses the modelling and the Bayesian methods used to analyze the data. Bayesian inference will allow for a more in depth analysis of the results. The results are presented in Section 2.4. Section 2.5 presents the conclusion and implications of the results.

2.2 Data

The data are taken from the American Community Survey (ACS) published by the United States Census Bureau. The ACS is conducted by contacting 3.5 million households per year. The ACS replaced the decennial census long form in 2010 and thereafter by collecting information similar to the long form each year throughout the decade instead of once every ten years. The data are ACS 1-year Public Use Microdata Sample taken in the survey from the years 2005 through 2012 concerning educational attainment, annual earnings, and occupation along with demographic information of the respondent, such as gender, race, age, etc.[9] The data are treated as one cross-sectional set. Further, the data are refined into

[8]Gelman et al. (2013) lists this as a psychological reason for using Bayesian methods, and Gill (2014) states, "the social sciences have been seriously harmed by [pseudo-frequentist null hypothesis significance testing]," which is related to the matter being discussed.

[9]Annual earnings in each year converted to 2010 dollars using the Consumer Price Index.

subsets. The survey sample is split for non-Hispanic black and white workers, and restricted to ages 18 to 67, male, high school graduates working 35 hours or more per week and greater than 40 weeks per year with positive annual earnings, and no imputed earnings.[10] The ACS asks the respondent what is the approximate number of hours usually worked in each week in the past twelve months, along with a categorical question of how many weeks were worked in the past twelve months. In addition, the survey asks for the highest degree or level of school the respondent has completed. High school graduates are considered those respondents who answered this question with high school diploma or some college credit, but less than one year. Occupational sorting is dependent on educational attainment, therefore the subsets were restricted to high school graduates.[11] The white worker sample with the above mentioned restrictions has 888,539 observations, while the black worker sample has 80,731. The summary statistics are presented in Table 2.1.

As shown in Table 2.1, on average the black workers in the data earn about 78 percent of the white workers. Even when discounting the outliers in earnings by looking at the medians for the two groups, black workers' median earnings are only approximately 80 percent of the white workers' median in the data. Black workers in the data are slightly younger and work slightly less hours. Earnings have long been believed to follow a log-normal distribution and the samples in this study do not deviate from that. Figure 2.1 shows the earnings distributions of black and white male high school graduates within the parameters discussed above, with a log normal density curve overlay. The vertical line in each indicates the mean earnings of each group. The mean earnings would overstate the earnings of approximately 60.9 percent and 58.8 percent of white and black workers, respectively. Further, 76.3 percent of black workers earn less than the white workers' mean earnings, and 67.4 percent earn less

[10]The hours and weeks restriction is similar to Altonji and Blank (1999) for "Full-time/full-year" workers, however the weeks restriction in this study is more relaxed.

[11]Fouarge et al. (2014) states, "Occupational choice is intimately connected with educational choice."

than the median white worker. For this study, simply using earnings would not be useful due to the skewed distribution. Therefore, the log of earnings, following the Mincerian earnings function, is taken as the dependent variable, also shown in Figure 2.1.

The number of weeks worked is a categorical variable in the ACS, and as shown in Figure 2.2, they do not differ very much at all between the two groups. The vast majority of both groups worked 50 to 52 weeks in the past year at the time of response to the survey. The class of worker, shown in Figure 2.3, does slightly differ between groups and this could be an important issue to look at since black workers in the sample have a lower rate of self-employment.[12] Black workers have a higher rate employed with the public sector (21.10 percent) than white workers (13.13 percent), including local, state, and federal government positions, while a slightly lower rate (77.11 percent and 80.71 percent, respectively) are employed privately, whether for profit or non-profit. Of the 21.10 percent of black workers employed by public institutions, 41.84 percent are employed by local governments, 21.22 percent are employed by states, and 36.94 percent by federal, while white workers government employment rates are 45.40 percent, 22.88 percent, and 31.72 percent, respectively. Neither group has a noticeable rate of employment in unpaid positions with family owned businesses or farms.

As stated earlier, the average black worker is slightly younger than the average white worker. Although the densities across the ages are not very dissimilar, the black workers have slightly higher rates amoung the younger ages and not as high of a rate in the late 40's as seen in Figure 2.4 for white workers. Both drop off sharply after the age of 50 towards the later ages in the groups. The focus of this study, also found in Figure 2.4, the occupational distributions for the two groups are noticeably different. White workers have higher densities in the upper management occupations (Census Occupation Codes 0000 through 1000), while black workers have higher densities in most service occupations (3600

[12]While self-employment rate is related to occupational sorting, it is beyond the scope of this study.

Table 2.1. Summary Statistics for High School Graduates

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
White Males	888,539					
Age		43.53	45.00	11.775	18.00	67.00
Hours per Week		45.20	40.00	8.813	35	99
AnnualEarnings		47,298.86	40,656.11	34,207.88	3.74	702,289.90
Black Males	80,731					
Age		42.73	44.00	11.606	18.00	67.00
Hours per Week		43.70	40.00	8.396	35	99
AnnualEarnings		36,699.78	32,448.81	22,725.07	4.47	550,548.10
Male HS Graduates	1,013,925					
Age		43.38	45.00	11.782	18.00	67.00
Hours per Week		45.04	40.00	8.808	35	99
AnnualEarnings		46,088.66	40,000.00	33,382.79	3.74	702,289.90

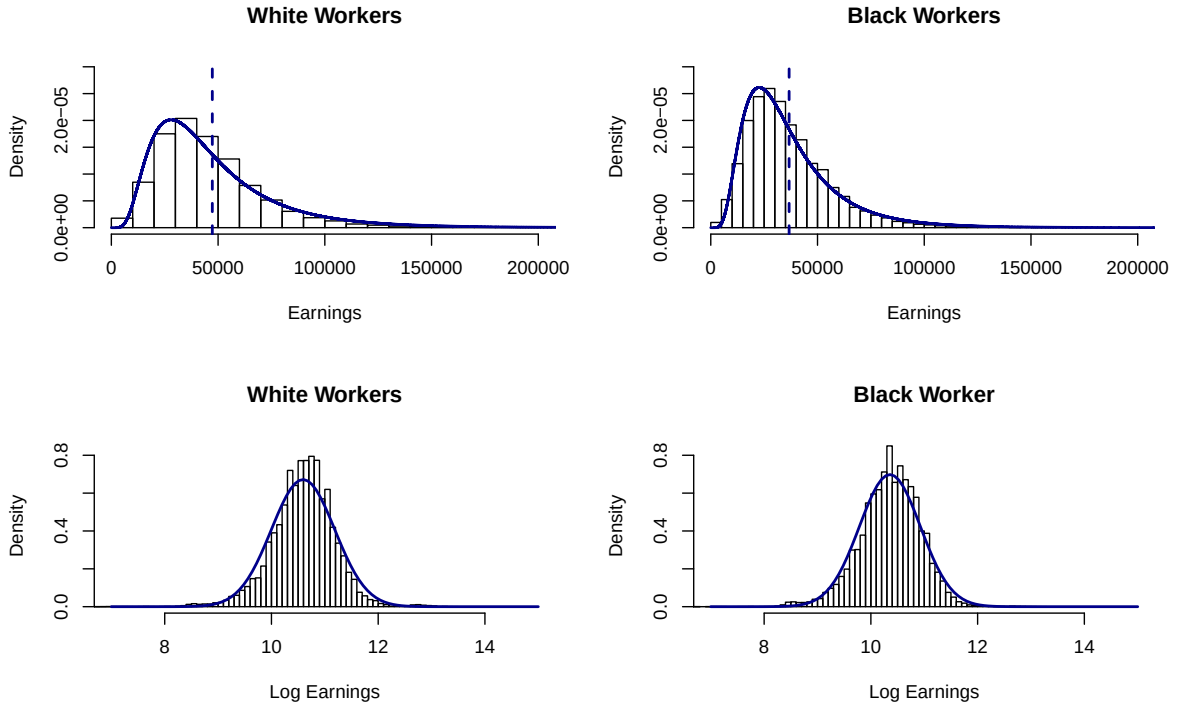

Figure 2.1. Earnings and Log Earnings

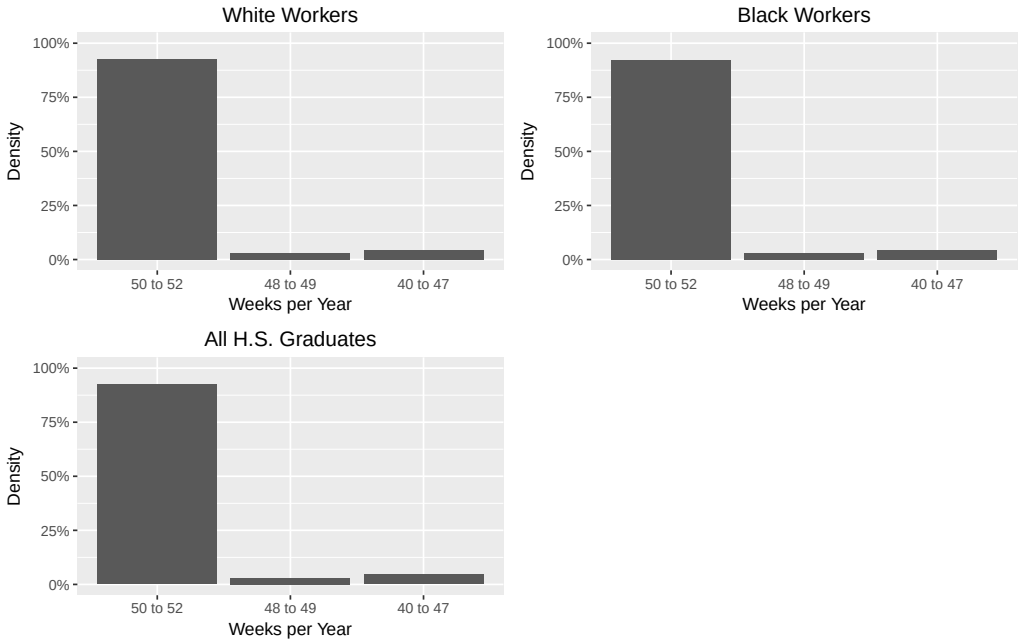

Figure 2.2. Number of Work Weeks

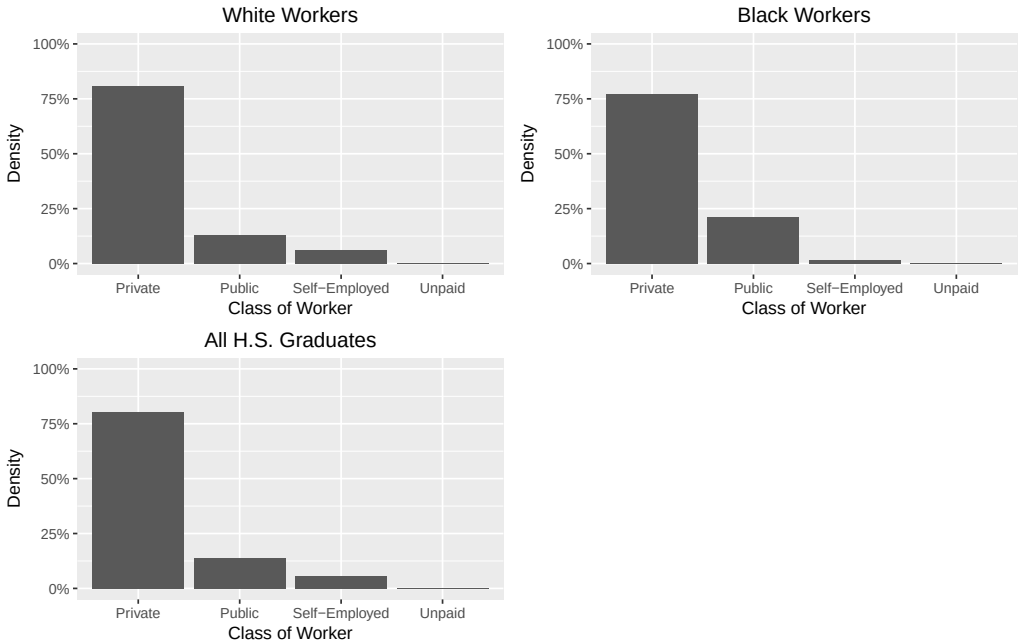

Figure 2.3. Class of Worker

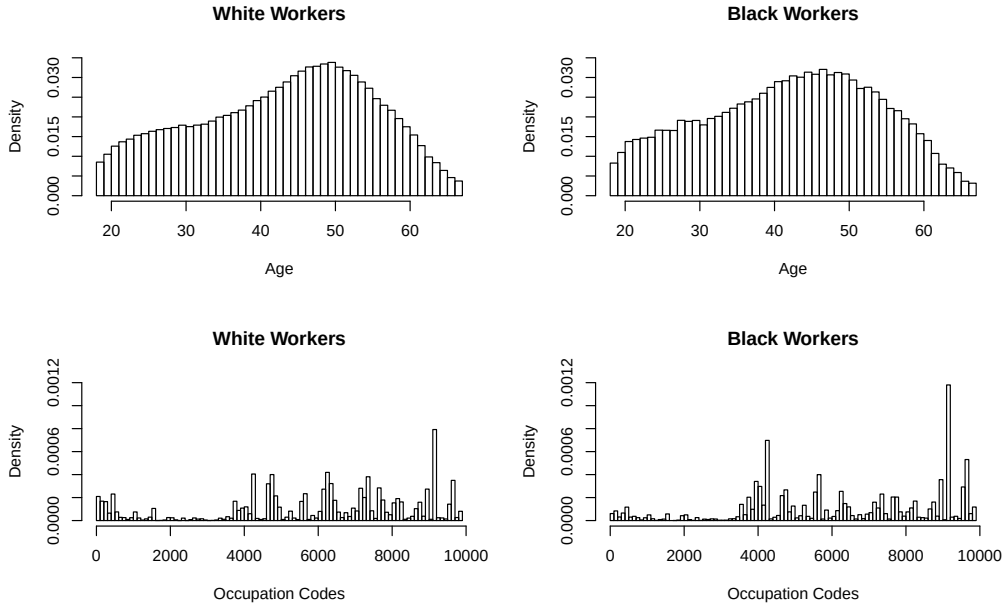

Figure 2.4. Ages and Occupations

through 4650). These attributes alone could cause a gap in the average earnings between the two groups. White workers have higher densities throughout the natural resources, construction, and maintenance occupations (6005 through 7630), for example carpenters, electricians, and plumbers are found in this range. Both groups' modal occupation is truck drivers (Code 9130).

From each of these groups, 20,000 workers are drawn using a pseudo-random sampler with a set seed. These subsets are used to construct baseline or control *human capital earnings functions (HCEFs)* for each race.

2.3 Model and Methods

The Human Capital Earnings Function was developed by Jacob Mincer in Mincer (1958) and Mincer (1974) for projecting or estimating an individual's earnings over their lifetime. The HCEF was developed to explain the phenomenon that earnings grow at a declining rate over an individual's working years. The Mincerian model has been further built upon

and refined to better fit the data observed. James J. Heckman, Ph.D. has been a critic of using the model for marginal returns to education, which is not the focus of this study.[13] In fact, Dr. Heckman has argued that marginal returns to education on log earnings should be moved toward a non-parametric specification.[14] Carneiro et al. (2011) points out with a conventional Mincer specification, the parameter for schooling (marginal return to college) is random.[15] The HCEF remains the standard for projecting or estimating an individual's earnings over their working years, and Bayesian methods allow for the parameters to be defined probabilistically, further strengthening the inference. The model used in this study is a quartic specification of the HCEF. This specification is preferred to the quadratic as Murphy and Welch (1990) presents that the quadratic understates early earnings growth, while overstating mid-career growth.[16] Earnings are known to increase with experience at a decreasing rate. In this model, log earnings (y_i) is the dependent variable. With the restrictions given above, the explanatory variable is age (x_i) as a proxy for experience. Equation (2.1) will be used in each analysis to construct and compare HCEFs between the control groups and treatment groups.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i \tag{2.1}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{2.2}$$

This study incorporates Bayesian methods for estimating the HCEFs.[17] For the two subsamples above, without knowledge of what the parameters might be, uninformative diffuse

[13]For example, Heckman et al. (2003) explains that the conditions needed for the model to produce meaningful results has been at odds with the data since the 1960 Census.

[14]See: Carneiro et al. (2011)

[15]As will be explained later, the Bayesian mindset is that unknown parameters are random in nature and must be described probabilistically.

[16]For more literature implementing the quartic specification, also see: Katz and Murphy (1992),Heckman et al. (2006),Black and Smith (2006),Autor et al. (2008).

[17]Gill (2014) provides an explanation of the many benefits of Bayesian approaches.

normal priors are used for the parameters. This is likely not a factor due to the sample size.[18] The marginal posterior samples are drawn using the MCMC algorithm *Automated Factor Slice Sampler*.(Tibbits et al., 2014) The first 1,000 iterations are discarded as burn-in. After the burn-in iterations, 32,000 iterations are run. Stationarity and other checks are performed. These conditions are satisfactory. Then, 10,000 more iterations are run in 3 parallel chains. The previously mentioned checks are satisfactory, and convergence is confirmed using Gelman and Rubin's MCMC Convergence Diagnostic.(Gelman and Rubin, 1992) The marginal posterior distributions for the white workers are shown in Figures 2.5, 2.6, and 2.7.

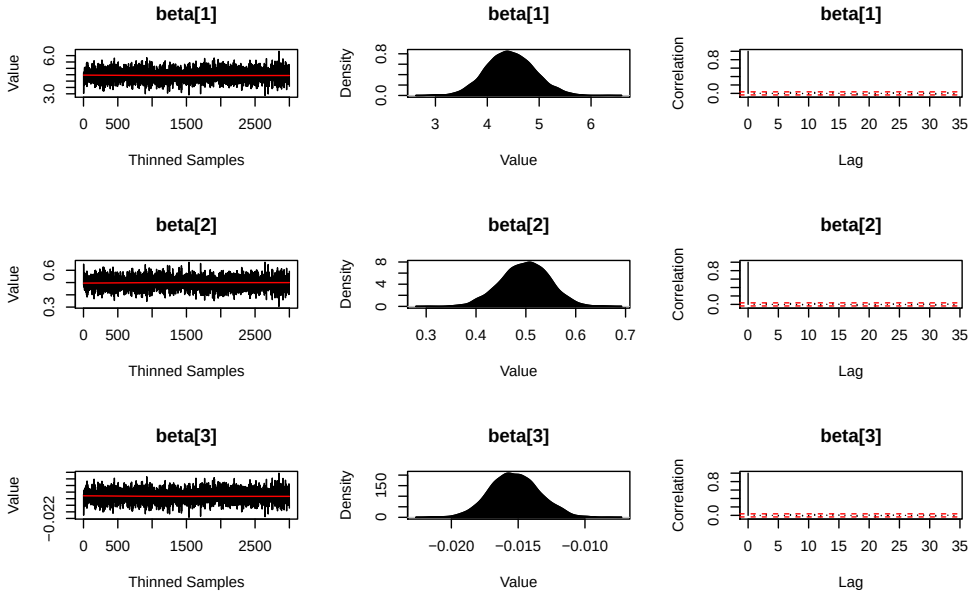

Figure 2.5. Parameters for White Workers: Betas 1 to 3

Each marginal posterior has the slightest skew, thus the medians of the marginal posteriors tend to fit the data better, and the HCEFs produced with median income at each

[18]Gill (2014) explains in this situation that the "likelihood dominates our choice of prior here." This is also shown in the results when the marginal posteriors for black and white workers are each used as priors in the treatment group.

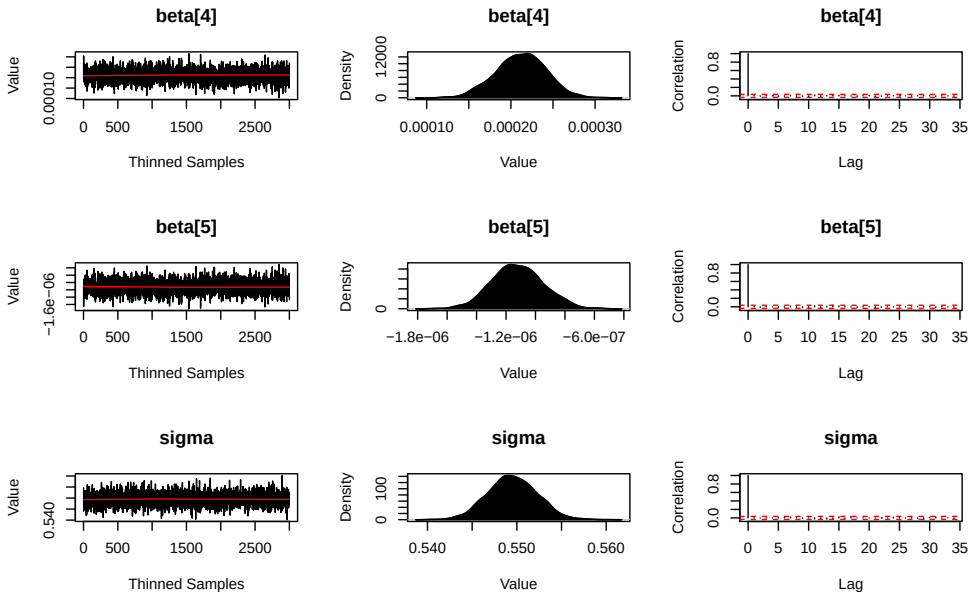

Figure 2.6. Betas 4 to 5 and Sigma

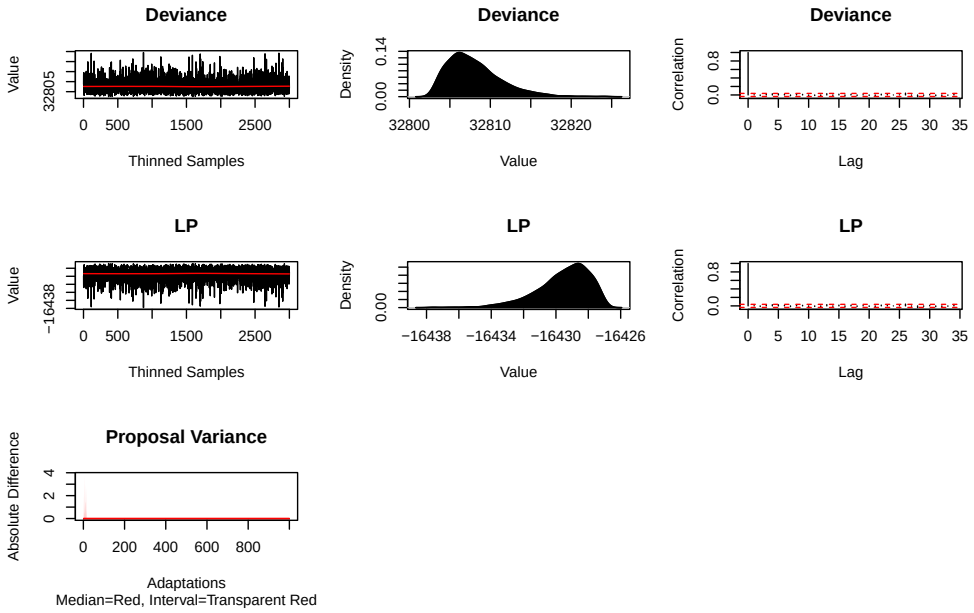

Figure 2.7. Deviance and LP

age for each group are shown in Figure 2.8. The difference in the HCEFs is apparent and significant. Although these curves are not meant to be predictive in this study, the curves

show that among younger ages the gap is much smaller than mid-career. For those workers younger than 20, the gap is almost non-existent. In later ages, the white workers appear to have a steeper decline in earnings, which begins to converge on the black workers curve and close the gap slightly. For high school graduates, the HCEFs shown in Figure 2.8 imply an earnings gap of $378,384.80 over the working ages of 18 to 67, which again shows the median black worker earning approximately 80 percent of the amount the median white worker in the study earns over a working lifetime.[19] While only showing the median parameters in Figure 2.8, a major advantage of Bayesian methods is the marginal posterior distributions, which are used to estimate posterior predictive distributions. Thus, the Human Capital Earnings Function is in fact three dimensional with densities of income at each age for the groups. These are further discussed in the Section 2.4.

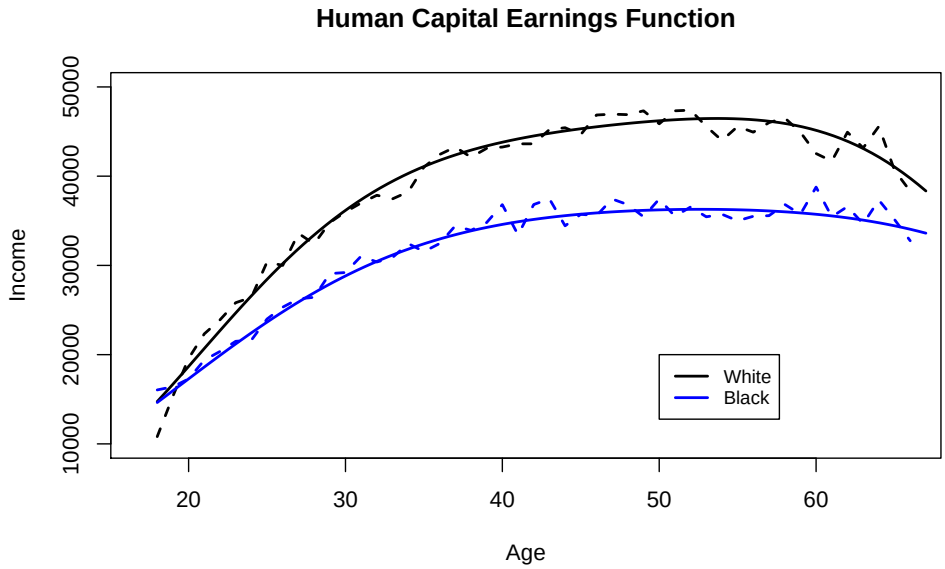

Figure 2.8. HCEFs for Control Groups

[19]This calculation is based on the integrals of the HCEFs over the ages mentioned.

In addition, with an added Bayesian approach, two more sub-samples of 20,000 are drawn from the white workers sample, however these sub-samples are drawn[20] from the black workers' occupational discrete probability distribution.[21] In other words, 20,000 white workers are drawn using the discrete probabilities from the occupational distribution of black workers. This group is referred to as the treatment group. The marginal posterior distributions from the control groups are used as priors for the HCEF parameters in the treatment group. The second sub-sample of white workers drawn from the black workers' occupational distribution is used for posterior predictive checks, mainly to test concordance against the parameters derived from the first treatment group.[22]

2.4 Results

The results are clear, when white workers sort into occupations similar to black workers their earnings are lower than the white workers in the control group. The treatment groups marginal posteriors are estimated using the control groups' marginal posteriors as priors. Figure 2.9 shows the HCEFs using the median parameters from the marginal posteriors of the control and treatment groups. The control HCEFs are the dashed lines and found in Figure 2.8. The treatment groups are consistent and using the control groups priors for treatment group 1 does not have a major impact on the outcome. The second treatment group satisfies the predictive concordance check using the model and marginal posterior distributions from the first sub-sample. It is clear that the white workers drawn from the black workers occupational distribution have a lower HCEF and are expected to earn less at

[20] Again using a pseudo-random sampler, however different seeds were used for second sub-sample for a robustness check.

[21] The thought here is that the black workers' occupational distribution is acting similar to a prior distribution.

[22] Predictive concordance described in Gelfand (1996).

Figure 2.9. HCEFs for Control and Treatment Groups

each age. Although the HCEF is significantly lower for this group, it is still higher than the black workers' curve. At the younger ages, eighteen to early twenties, the earnings according to the HCEF are similar to the white workers control group, but they begin to diverge in the mid-twenties. From this point, treatment group 1 will be used for the analysis and simply referred to as "the treatment group."

As with the control groups, the first 1000 iterations were discarded when estimating the marginal posteriors for the treatment group. After the burn-in, 32,000 iterations were run and checked for stationarity, Monte Carlo standard errors relative position to marginal posterior standard deviation, effective sample size, etc. The marginal posteriors were satisfactory in the test metrics, then 16,000 additional iterations were run in each of three parallel chains. The chains converged and were stationary. The three chains were combined making one marginal posterior for each parameter. The marginal posteriors are shown in Figures 2.10, 2.11, and 2.12, along with the value for each iteration of the left-hand side and the ACF on the right hand side as a check for stationarity. The marginal posteriors for the treatment

group are used to create a posterior predictive distribution for inference and comparison to the control groups.

Figure 2.13 shows the marginal posteriors of the treatment group with the priors used, which are the marginal posteriors of the white workers control group. As shown, the marginal posteriors differ slightly from their priors as expected. With a differing occupational distribution, the treatment groups income is affected and this effect is manifested in the marginal posteriors of the model. The effect on the median income is shown in Figure 2.14. The difference in median income for the white workers control group and the treatment group is also shown in Figure 2.14 by the pink shaded area. The younger workers' earnings in the treatment group have little to no gap, however their growth in earnings slows down quicker than the control groups' earnings. This creates the gap and it is persistent throughout the working ages. The shaded area corresponds to a difference of $147,537.20 over the working ages of the median earner. Should white workers sort into occupations similar to black workers, they could expect to earn approximately 92.21 percent of what their counterparts sorting as they do in reality. This accounts for 38.99 percent of the earnings gap between white and black workers discussed earlier. The expected earnings gap that remains between the treatment group and the black workers control group HCEFs is approximately $230,847.60 over the working ages. This translates into black workers earning approximately 86.79 percent of the earnings of white workers who have similar occupational sorting over the working ages.

Using the marginal posteriors from the treatment group, a posterior predictive distribution is generated. As shown in Figure 2.15, the posterior predictive distribution (PPD) follows a log normal distribution as expected. The treatment groups' PPD appears slightly different from each of the control groups' earnings distributions, however that slight difference results in more than $100,000 over the working ages between the treatment group and each control group. The red curves on Figure 2.15 represent the white workers control group's earnings distribution shown in Figure 2.1, while the solid red vertical lines and the

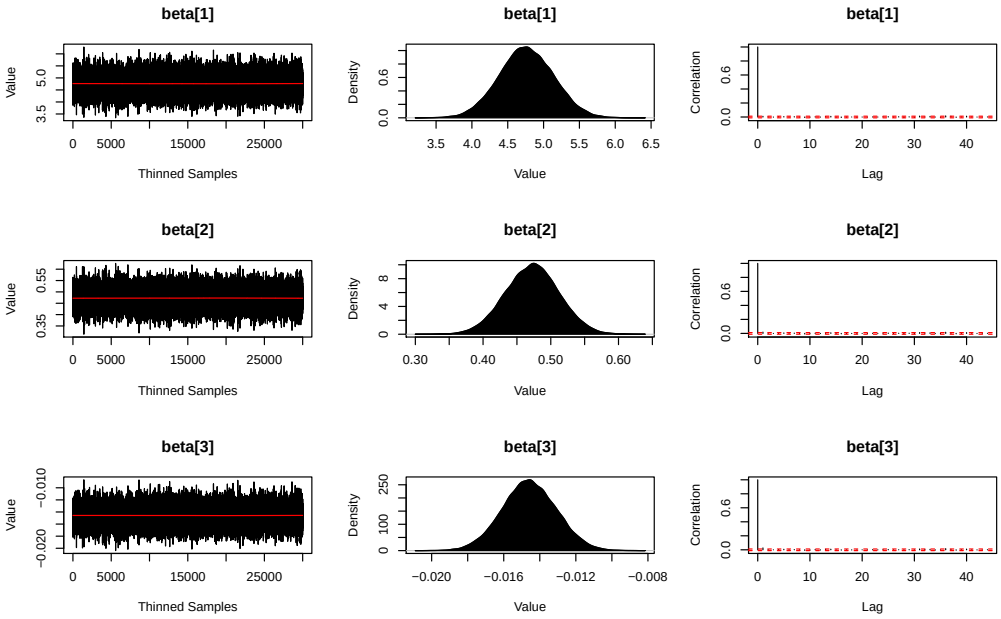

Figure 2.10. Parameters for Treatment Group: Betas 1 to 3

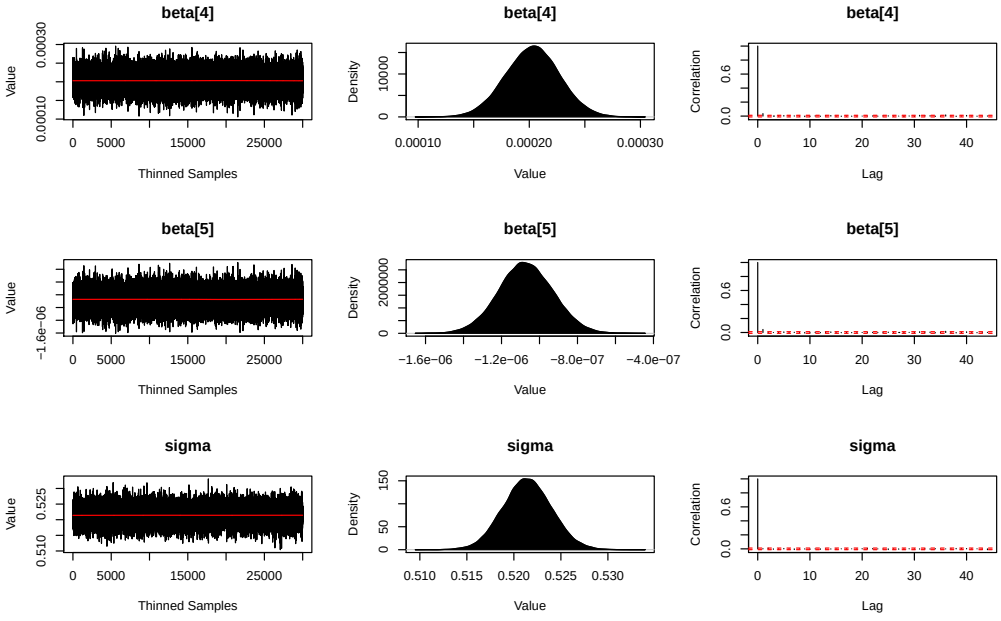

Figure 2.11. Betas 4 to 5 and Sigma

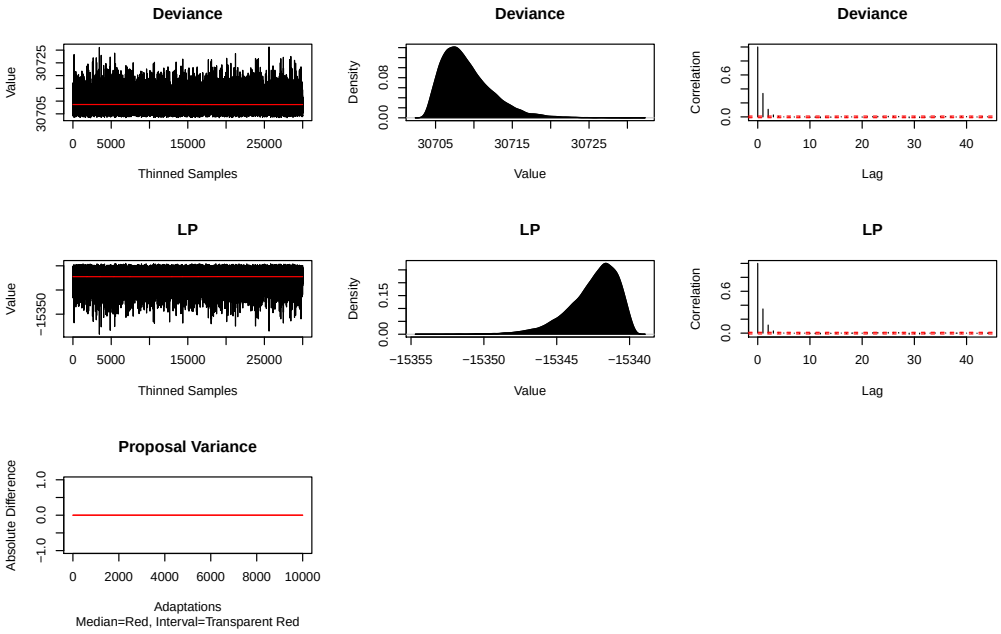

Figure 2.12. Deviance and LP

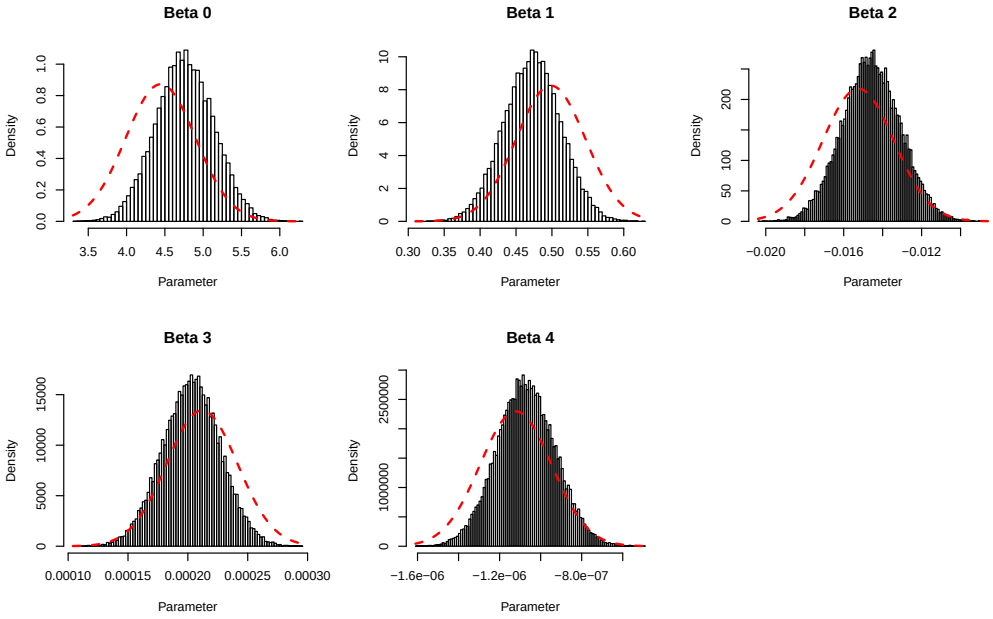

Figure 2.13. Parameter Marginal Posteriors with Priors

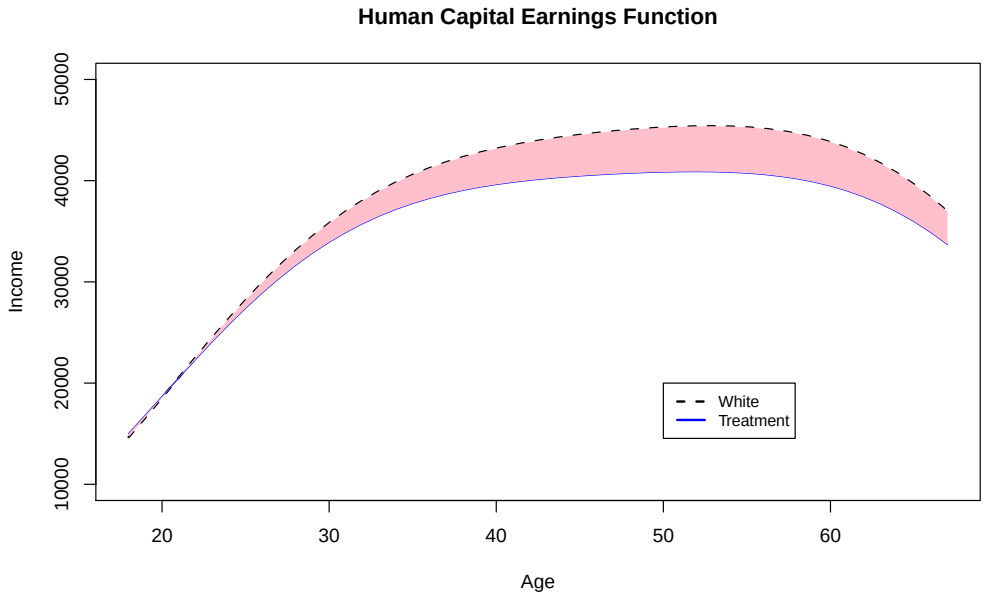

Figure 2.14. Difference in HCEFs for White Workers Control and Treatment Groups

dashed vertical lines are the median and mean of the data, respectively. The solid blue vertical lines and the dashed vertical lines are the treatment groups median and mean of the PPD, respectively. The green lines are analogous to the lines mentioned previously for white workers, but are corresponding to the black workers earnings distribution from the data.

The posterior predictive distribution of the treatment group has a mean of $42,809.87, which is in the 60.30 percentile of the distribution. The median is $37,178.33. The median earner in the treatment PPD would be in the 42.26 percentile of the white workers control distribution, while the average earnings fall in the 53.56 percentile. Again, the mean of the white workers in the data fall in the 60.86 percentile of the earnings data. Thus, the treatment group's mean falls below that of white workers by 7.30 percentage points from occupational sorting similar to black workers. The mean earnings of the white workers in the data are higher than 67.24 percent of the treatment groups PPD, while the median is higher than 56.57 percent. This is a result of the density of the PPD for the treatment group being more condensed under $50,000 and having a lower density in the upper tail. That is

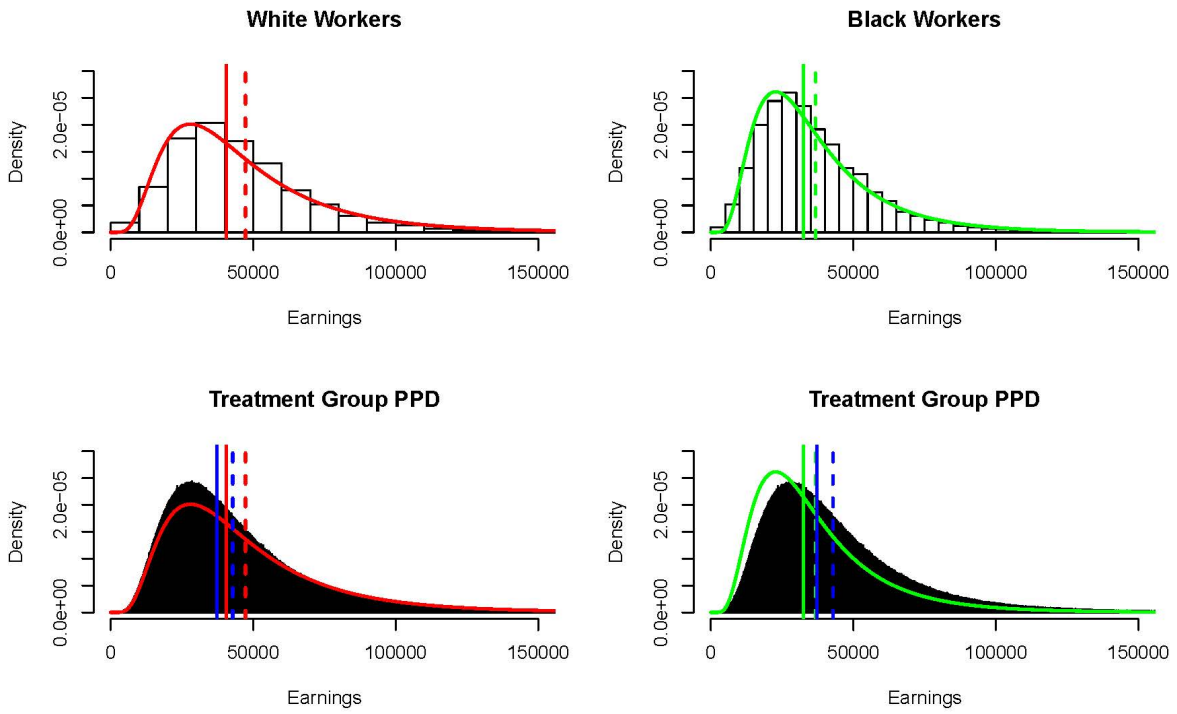

Figure 2.15. Posterior Predictive Distribution of Treatment Group

understandable with the knowledge that the occupational distribution density is lower in the upper management/executive occupations for the treatment group.

In addition, Figure 2.15 shows the treatment group PPD with an overlay of the black workers' earnings distribution from the data. With respect to the black workers' earnings distribution, the mean of the treatment groups' PPD is higher than 70.38 percent. The median of the treatment group falls at the 59.90 percentile, very near the mean of the black workers' earnings data. The median of the black workers' earnings is in the 40.10 percentile of the treatment group PPD.

Another way to look at the posterior predictive distribution for the treatment group is to plot the distribution across ages. Figure 2.16 is a two dimensional representation of a three dimensional plot with age, income/earnings, and density/frequency as the "z" axis. As with earlier figures, the red line corresponds to the white workers' data, specifically for Figure

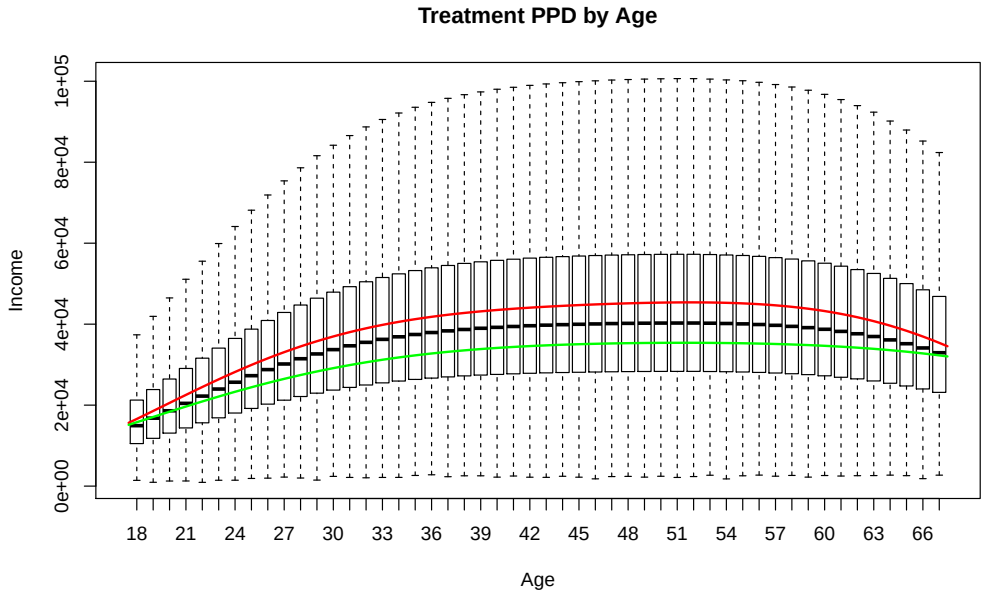

Figure 2.16. Posterior Predictive Distribution of Treatment Group Across Ages

2.16, it is the HCEF estimated earlier. The analogous curve for the black workers' HCEF is the green line. With the nature of the distributions of earnings and as shown in previous plots, it appears unlikely that one would find statistical significance in the difference in earnings and the difference doesn't look impressive in Figure 2.16, but again the differences are over $100,000 over the working years between the medians of the treatment group and the medians of each control group. In addition, a Kolmogorov-Smirnov two sample test was performed for the treatment group against white and black earnings of the control groups. The conclusion of each of the tests is that the treatment group's PPD earnings were not from the same distribution as the control groups' earnings.[23]

[23]This conclusion is reached at the 99 percent confidence level.

2.5 Conclusion

As previously stated the race wage gap has been present and persistent. For most, if not all, economists studying this issue, the final objective is clearly to undermine and eliminate all of the gap's existence, especially the portion that is due to racial discrimination. To reach that goal, a full understanding of the differences in earnings must be explored. This study sought to show how occupational sorting of the workers influences annual earnings across working ages. Occupational sorting absolutely plays a role in determining the differences in earnings between the races. According to the results of the study with the ACS data examined, approximately 39 percent of the earnings gap over the working ages are contributed to occupational sorting. Thus, should white male high school graduates, working *full-time year-round*, sort into occupations similar to black workers, the expected earnings would be approximately 92 percent of those sorting as they do in the observations. Recall that the median earners of black workers only earn approximately 80 percent of the white workers, therefore the majority of the earnings gap is not accounted for by occupational sorting. However, 39 percent is a significant portion of the earnings gap.

Meanwhile, the posterior predictive distribution shows that while it may appear that the results from the treatment are insignificant statistically, Bayesian methods are able to show that the distributions of earnings are in fact different and economically significant.[24] The white workers in the treatment group would be expected to earn in the 42.26 percentile of the white workers in the data, compared to black workers median at 33.32 percent of the white workers' earnings in the data. It is clear that occupational sorting plays a role in the race earnings gap between white and black workers. What is maybe even more clear is that the role of occupational sorting is not responsible for even half of the earnings gap.

[24]The treatment group's PPD earnings are found to be from a different distribution than the control groups' at a 99 percent confidence level, according to Kolmogorov-Smirnov tests.

Altonji and Blank (1999) states, "A key question is whether occupational and industry differences represent preferential choices or constraints." In other words, is occupational sorting due solely to choices and preferences of the workers or is it a function of discrimination? The latter is a possibility and should be explored further. The implication would be the earnings gap due to occupational sorting would not necessarily need to be "fixed" should it be driven by choices and preferences, and that portion of the earnings gap would not close unless preferences or structure of economy changed. However, if systemic discrimination effects occupational opportunities for black workers leading to lower earnings than white workers, then the methods and modes of discrimination should be exposed and eliminated.

CHAPTER 3

COLLEGE MAJOR SELECTION AND THE RACE WAGE GAP: A STUDY USING BAYESIAN METHODS

3.1 Introduction

Choices made everyday have a profound impact on lives. This holds true for economic choices that impact earnings even before an individual enters the labor market. The previous chapter sought to explore the differences in annual earnings between races, black and white, across working ages due to occupational sorting. Another topic of concern for examination in the race wage gap is college major choice for bachelor degrees. This study will build on the study of occupational sorting and look at race wage gaps between multiple races, however the concern shifts to the race wage gaps of workers with bachelor's degrees and the choice of college major. College is a costly endeavor, but in most cases, attending college is a good financial decision.[1] Assuming a 3 percent discount rate, Webber (2016) shows the present value of obtaining a college degree ranges between $85,000 and $300,000, *depending on college major*.[2] According to Neal and Johnson (1996), the race wage gap, at least among black and white workers, "primarily reflects a skill gap." College major choice can exacerbate this skill gap.(Polachek, 1978) College major choice can be determined by multiple factors, but this choice will impact future earnings. Altonji et al. (2016) state, "The evidence suggests that much of the effect of major on earnings is causal, with STEM and business-related majors leading the way."[3]

[1]Webber (2016) finds this, with STEM and business degrees leading the way, even when accounting for the possibility of not completing a degree.

[2]The lower range corresponds to Arts and Humanities degrees while the upper range STEM and business. Webber (2016) proposes more transparency of benefits and costs to help students make a more fully informed decision about their future employment prospects, similar to Baker et al. (2018).

[3]While STEM majors have one of the highest causal effects on earnings, 48 percent switch out of those particular majors or drop out according to Altonji et al. (2016).

The formation of the choice of college major begins well before the student reaches the age for college.[4] Morgan et al. (2013) find that "occupational plans of high school seniors are strong predictors of initial college majors." This finding is not in conflict with Polachek (1978), which states that an individual makes a "rational" choice to "maximize his benefit-cost ratio" in choosing a college major. Baker et al. (2018) do confirm that labor market outcomes do have an effect on major choice, but their information set on which they base their choices may be flawed.[5] Ultimately, Baker et al. (2018) find that students put the most weight on "course enjoyment" when selecting a major. In addition, Ochsenfeld (2016) finds that vocational interests and peer expectations are major determinants of major choice, while life goals or academic performance have little or no effect on gender differences in college major.

Although Polachek (1978) examines the gender wage gap, it provides useful insight to college major choice and differences in groups, such as men tended to major in physical sciences, engineering and business, while women chose liberal arts majors. The study is dated, but those differences in major choice did contribute to the gender gap in earnings.[6] The gender wage gaps are found in Lin (2010) to be smaller in the agriculture and literature fields, while larger in education, engineering, law, business and medicine.[7] Gerhart (1990) finds while growth in earnings into the career do not seem different, starting salaries are an

[4]Speer (2017) finds that differences in skill show up in test scores in the mid-teens and probably before. ASVAB test scores are found to be a very reliable indicator of gender gaps in college major, contrary to prior opinion.

[5]Students tend to overestimate their expected salaries by approximately 13 percent, while underestimating their probability of employment by approximately 25 percent, and students from lower income backgrounds are more likely to make these miscalculations. Baker et al. (2018) suggests informational interventions could positively affect major choice and years of schooling.

[6]Daymont and Andrisani (1984) found that college major choice, along with occupational role preference, accounted for approximately "one-third to two-thirds" of the gender pay gap.

[7]Lin (2010) proposes an interesting policy for the differing payoffs in these fields. Lin proposes differing tuition (price discrimination) as a way to even the relative costs and benefits.

"Important sources of men's higher starting salaries...are their higher degree attainment and concentration in different college majors."[8]

With their study on the race wage gap, Emmons and Ricketts (2017) title their paper "College Is Not Enough." The title makes it clear that even when level of education is the same, "race and ethnicity are highly predictive of family wealth." Addo et al. (2016) points out that black young adults have substantially more debt than white young adults. This difference is partially due to their family backgrounds.[9] Black students are also more likely to attend institutions that provide less aid relative to the costs, including high cost for profit colleges or universities. To exacerbate this divergence, Addo et al. (2016) also find higher drop out rates are associated with larger debt loads. Thus, given the literature on the gender pay gap and college major choice, it is imperative to study college major choice and the race wage gap. The question for this study will be: What are the differences in earnings of college graduates from different races if their college majors were distributed like those of other races? Does the race wage gap of college graduates shrink when students of a particular race/ethnicity choose majors like that of the other race/ethnicity?

This study uses cross-sectional ACS micro data and a quartic specification of the *Human Capital Earnings Function (HCEF)* that will be further discussed in the following sections.[10] The HCEF is estimated using Bayesian methods with *Markov chain Monte Carlo* sampling to derive the marginal posterior distributions of the parameters. The marginal posteriors are used for further inference into the differences from the college major choices between the races. To dig deeper into the examination of these differences, the *Oaxaca-Blinder* decomposition is employed. White, black, and Hispanic college graduates are sampled from

[8]Gerhart (1990) uses a quadratic specification of the HCEF for their analysis.

[9]Such as their family's contribution to college, differences in post-secondary educational attainment and wealth.

[10]This will be referred to the HCEF beyond this point.

the data according to their own college major distribution for the control groups, and then sampled with probability weights of the other race/ethnicity's college major distribution for the treatment groups. This study looks at all majors, where no overall treatment effect is seen.[11] For easier presentation, the majors are grouped by their more general/broad major categories, but also no treatment effect is seen. Then, the data is broken into business and non-business majors and analyzed separately. An interesting result is found in non-business majors where white workers in the treatment group relative to Hispanics fare better than the control group. For business degrees, the choice of major does negatively effect both white treatment groups relative to the control group.

The paper is ordered as follows: Section 3.2 examines the data and interesting summary statistics found within. Section 3.3 describes the model and methods used to estimate the parameters of the model and analyze the findings that the model provides. Section 3.4 presents the results of the study, including the Oaxaca-Blinder decomposition of the results. The concluding remarks and future extensions of this research are found in Section 3.5.

3.2 Data

As in the previous chapter, the United States Census Bureau's American Community Survey micro-data is used. The data are 1-year ACS Public Use Microdata Samples taken from the years 2010 through 2016 are used.[12] This data contains education achievement, annual earnings, and occupation along with demographic information of the respondent, such as gender, race, age, etc.[13] The data are pooled and assumed to be one cross-sectional set.

[11]The magnitude of the race wage gap between Hispanics and black college graduates is not very significant relative to the gap between both of them and the white college graduates, thus the white to black and white to Hispanic wage gaps are the only ones analyzed past the analysis of all majors.

[12]The questionnaire began asking about college major in 2009, however changes were made to the indicator values prior to the 2010 questionnaire, therefore 2009 was not used.

[13]Annual earnings from each year of the samples are converted to 2015 dollars using the Consumer Price Index.

Table 3.1. Summary Statistics for College Graduates

Variable	Obs	Mean	Median	Std. Dev.	Min	Max
White Males	543,638					
Age		43.98	44.00	11.792	22.00	67.00
Hours		45.98	43.00	8.335	35	99
AnnualEarnings		94,352.93	73,086.65	82,649.85	3.95	705,105.00
Black Males	30,130					
Age		43.58	44.00	11.317	22.00	67.00
Hours		44.35	40.00	8.434	35	99
AnnualEarnings		66,153.06	55,000.00	48,805.33	29.63	613,000.00
Hispanic Males	41,474					
Age		41.16	40.00	10.944	22.00	67.00
Hours		44.822	40.00	8.283	35	99
AnnualEarnings		72,470.92	59,252.52	60,562.51	20.00	671,502.60
Male College Graduates	678,268					
Age		43.57	44.00	11.699	22.00	67.00
Hours		45.646	40.00	8.344	35	99
AnnualEarnings		90,570.68	70,652.06	79,023.97	3.95	705,105.00

A subset of the data is chosen for non-Hispanic black, non-Hispanic white, and Hispanic workers; restricted to ages 22 to 67, male, college graduates working 35 hours or more per week and greater than 40 weeks per year with positive annual earnings.[14] The white worker sample with the above mentioned restrictions has 543,638 observations, while the black worker sample has 30,130 and the Hispanic worker sample has 41,474. The summary statistics are presented in Table 3.1.

As shown in Table 3.1, the black college graduates in the data earn about 70.11 percent on average of the white college graduates, while Hispanics earn approximately 76.81 percent of white college graduates on average. The median black and Hispanic workers earn approximately 75.25 and 81.07 percent, respectively, of the median white worker's earnings.

[14] Altonji and Blank (1999) restricts to 35 hours or more per week for 48 weeks or more for "Full-time/full-year" workers, however the weeks restriction in this study is relaxed to 40 or more. In addition, College graduates are taken as those with a bachelor's degree as the terminal degree.

Hispanic workers in the data are slightly younger than white workers while black workers' ages are not much different from white workers. Both black and Hispanic workers work slightly less hours than white workers. Figure 3.1 shows the earnings distributions of black, Hispanic and white male college graduates within the parameters discussed above most assuredly follow a log normal distribution. The mean earnings for each group are indicated by the vertical line in their respective earnings distributions in Figure 3.1. Mean earnings would overstate the earnings of approximately 65.73, 62.23 and 63.46 percent of white, black and Hispanic workers, respectively. The average earnings of the white workers sample falls in the 82.43 percentile of the black workers earnings distribution and in the 78.34 percentile of the Hispanic workers earnings distribution, while the median white worker earns more than 67.97 and 63.74 percent of the black and Hispanic workers, respectively. The median black and Hispanic workers would fall in the 33.55 and 37.17 percentile, respectively, of the white workers earnings distribution. Also shown in Figure 3.1, the log of earnings with a normal density overlay is used as the dependent variable following the Mincerian earnings function or HCEF.

Respondents to the ACS are asked how many weeks in the past year have they worked, and the ACS categorizes the responses, and as shown in Figure 3.2 there is not much divergence between the three groups of interest. Within the constraints of the study, most respondents from each group worked 50 to 52 weeks in the past year. The classes of worker, shown in Figure 3.3, do differ slightly between the three groups and similar to the previous chapter looking at high school graduates, black college graduates have a lower rate of self employment and a higher rate of employment in the public sector. Black college graduates are employed with the public sector at a rate of 29.16 percent while white and Hispanic college graduates are employed in the public sector at rates of 16.02 and 21.25 percent, respectively. This includes local, state, and federal government positions. White, black and Hispanic college graduates are employed privately at rates of 76.57, 67.60, and 73.88 percent,

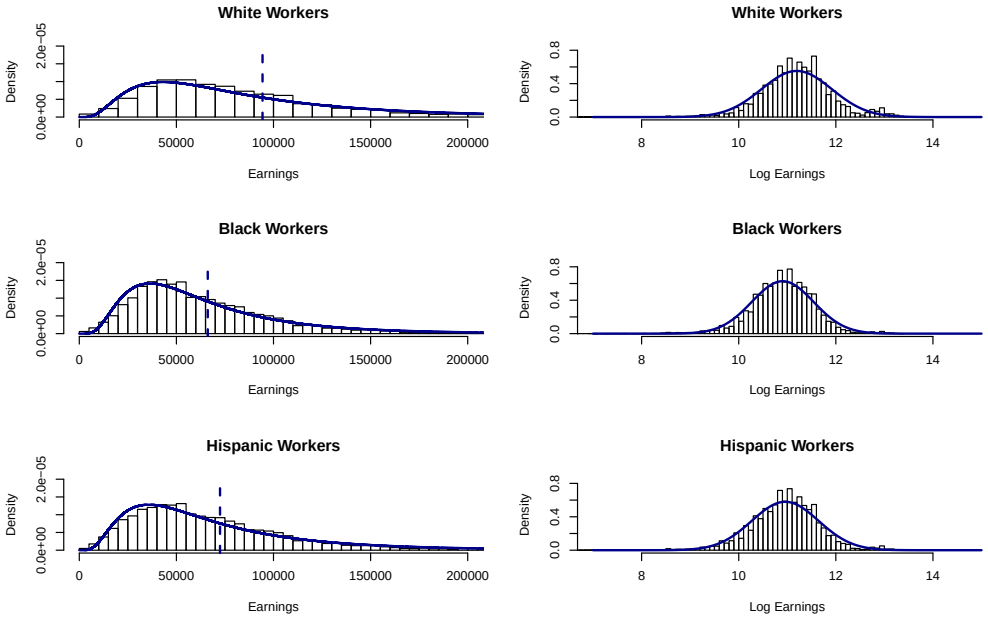

Figure 3.1. Earnings and Log Earnings

respectively. None of the groups have a noticeable rate of employment in unpaid positions with family owned businesses or farms.

As stated earlier, Hispanic workers on average are slightly younger than the other two groups. This can be seen clearly in Figure 3.4, where the Hispanic workers age distribution has noticeably higher densities at the younger ages. Occupational distributions, also found in Figure 3.4, for the three groups are apparently different. White and Hispanic workers have higher densities in the upper management occupations (Census Occupation Codes 0000 through 1000), although all three groups modal occupation is in management. The three groups also have substantial densities in the Sales and Office Occupations (Codes 4700 through 5940), while white workers have higher densities in the supervisor positions of these occupations than the other two groups. Each of these previously mentioned attributes of age and occupational distributions alone could cause gaps in the average earnings between the three groups.

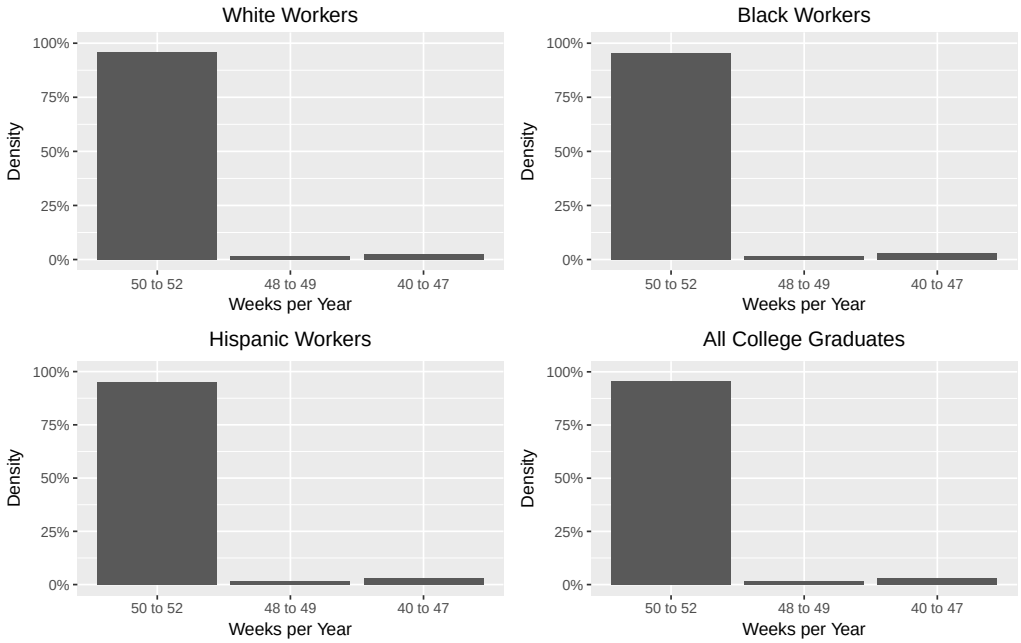

Figure 3.2. Number of Work Weeks

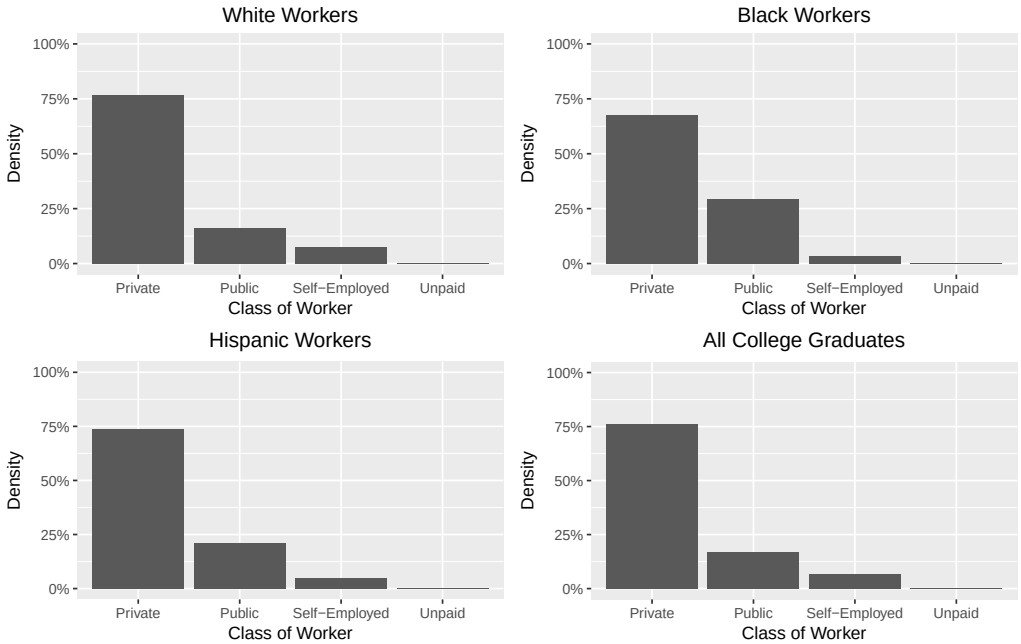

Figure 3.3. Class of Worker

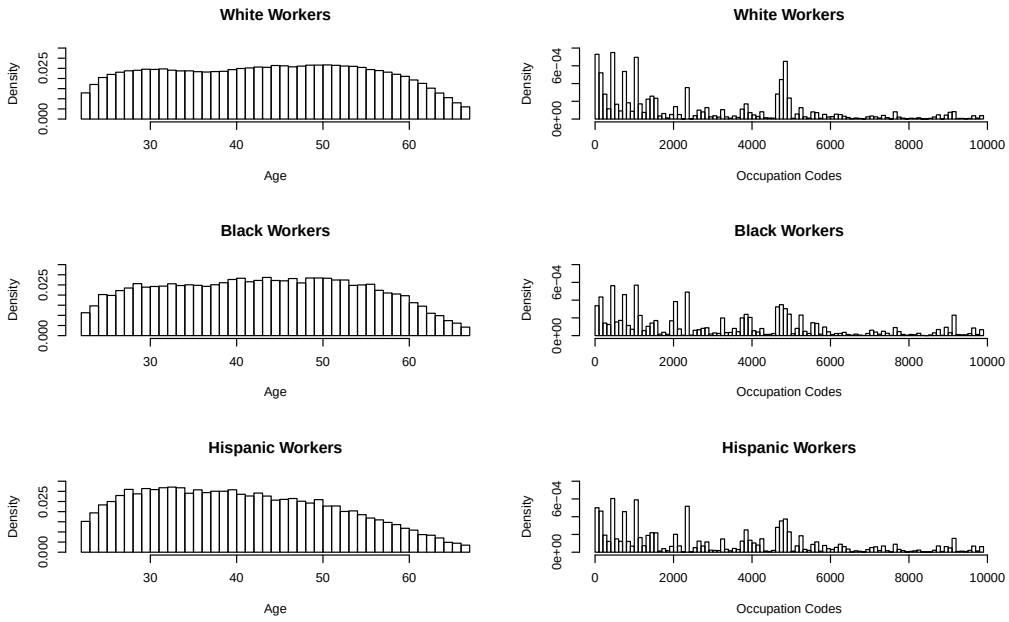

Figure 3.4. Ages and Occupations

The focus of this study is college majors (or field of degree). The distributions of the 173 different degree fields for each group are shown in Figure 3.5. Business degrees (Codes 6200 through 6299) are the most prevalent degrees among the three groups, Business Management and Administration in particular is the modal degree in each group. Another interesting characteristic of the distributions is the rates of graduates in the engineering degrees (Codes 2400 through 2599). Hispanic college graduates have the highest densities in the engineering fields, almost doubling the densities of the black workers in these fields. White college graduates have about 75 percent of the Hispanic graduates density in engineering. White college graduates have higher densities in the agricultural degree fields (Codes 1100 through 1199) than both black and Hispanics graduates, while each group has noticeable but seemingly not too different densities in the Hard and Social Sciences (Codes 3600 through 5098 and 5200 through 5599, respectively). A more presentable look at the fields of degrees is presented in Table 3.2, where the degrees are grouped into their more broad categories. As stated earlier, business degrees are by far the most prevalent fields in each group. For black graduates,

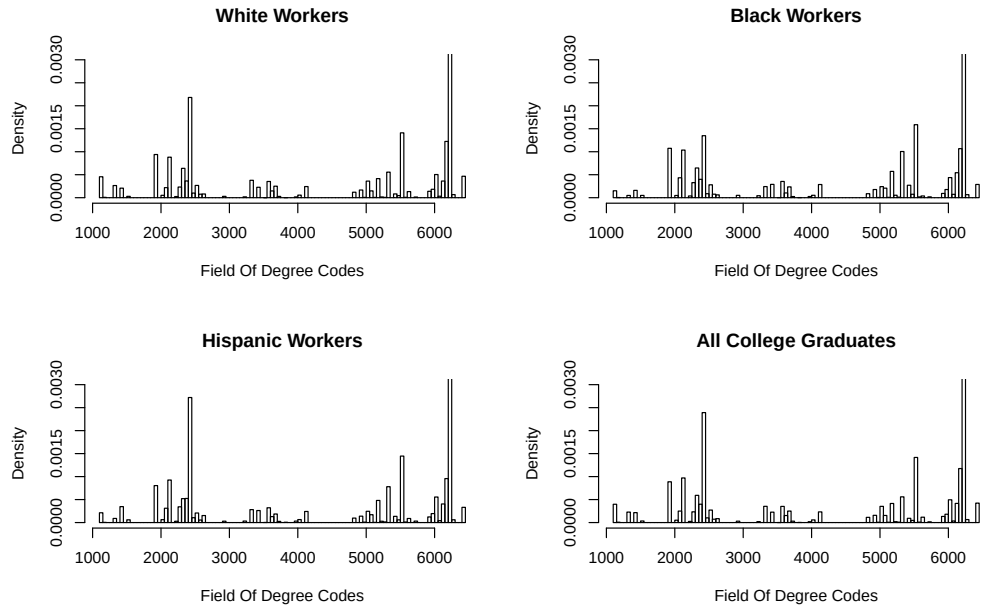

Figure 3.5. College Majors

social sciences are the second most category of degree obtained, while engineering is second for white and Hispanic graduates. As shown in Table 3.2, the top five degree categories (Business, Social Sciences, Engineering, CIS, and Education) account for approximately 60 percent of the graduates from each race/ethnicity.

From each of these samples, 20,000 workers are drawn using a pseudo-random sampler with a set seed. These are drawn for each of the analyses looking at all majors, all major broad categories and non-Business majors.[15] When looking at business majors then STEM degrees, 8,000 workers are sampled from each group based on the number of available data points in these sub-groups. These subsets are used to construct baseline or control Human Capital Earnings Function for each race/ethnicity to allow for comparison to the treatment groups. The process for drawing the treatment groups is discussed further in the next section.

[15]While non-business majors vary widely and the ACS has many observations on the whole, the sample sizes become extremely restrictive with any more specificity. Thus, STEM degrees were the most specific group studied within the non-Business majors.

Table 3.2. Percentage of Degrees by Race/Ethnicity

	Degree	Black	White	Hispanic
1	Business	30.04	30.16	27.50
2	Social Sciences	8.40	7.32	7.58
3	Engineering	8.36	12.57	16.28
4	Computer and Information Systems	7.35	5.50	6.18
5	Education	5.44	4.83	4.57
6	Comms and Journalism	5.37	4.70	4.02
7	Criminal Justice	5.02	2.78	3.90
8	Psychology	3.19	2.19	2.62
9	Health Services	3.14	1.98	2.27
10	Fine Arts	3.10	3.47	3.80
11	Biological Sciences	2.31	2.62	2.31
12	Physical Sciences	2.27	2.58	2.12
13	Engineering Tech	2.04	1.89	1.53
14	LA and Humanities	1.45	1.14	1.31
15	History	1.44	2.33	1.66
16	Parks and Rec	1.43	1.21	1.21
17	Public Admin	1.35	0.41	0.68
18	Mathematics	1.28	1.27	1.01
19	Literature	1.21	1.89	1.41
20	Theology	0.89	0.83	0.69
21	Architecture	0.80	1.04	1.74
22	Agriculture	0.77	2.32	1.08
23	Transportation Services	0.46	0.70	0.61
24	Philosophy	0.43	0.60	0.49
25	Multi/Interdisciplinary Studies	0.39	0.37	0.48
26	Linguistics and Lit	0.31	0.40	0.76
27	Family Science	0.26	0.15	0.15
28	Civilization Studies	0.25	0.15	0.30
29	Environmental Science	0.25	1.32	0.45
30	Comm Tech	0.24	0.26	0.32
31	Law Studies	0.22	0.09	0.16
32	Cosmetology	0.19	0.13	0.14
33	Construction services	0.19	0.66	0.44
34	Electrical Tech	0.09	0.07	0.16
35	Nuclear Tech	0.04	0.03	0.02
36	Library Science	0.01	0.00	0.01
37	Military Tech	0.01	0.01	0.03

3.3 Model and Methods

Earnings typically increase with experience at a decreasing rate. As in the previous chapter, the model is a quartic specification of the HCEF. Although Gerhart (1990) uses a quadratic specification, the quartic specification is preferred as Murphy and Welch (1990) shows the quadratic understates early earnings growth, and overstates mid-career growth.[16] Given the restrictions to the subsets, the explanatory variable is age (x_i) as a proxy for experience while log earnings (y_i) is the dependent variable. The model in this study follows (3.1).

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \beta_5 x_i^4 + \epsilon_i \tag{3.1}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{3.2}$$

HCEFs will be constructed to compare earnings by age between the groups. While OLS estimation would impose normality with mean zero on the error term (ϵ_i), Bayesian methods do not. This fact will be important when discussing the Oaxaca-Blinder decomposition.

Bayesian methods are employed for estimating the HCEFs.[17] With the three sub-samples above, non-informative diffuse normal probability densities are used as priors for the parameters. The sample size renders the non-informative prior moot.[18] The *Automated Factor Slice Sampler*, as found in Tibbits et al. (2014), is used for drawing the marginal posterior samples for the parameters.[19] The iterations differ between the differing analyses. For the analyses on all majors, broad major categories and non-business majors, 1,000 iterations are run and discarded as burn-in, then 30,000 iterations are run. Stationarity and other

[16]Other literature employing the quartic specification: Katz and Murphy (1992), Heckman et al. (2006), Black and Smith (2006), Autor et al. (2008).

[17]A detailed explanation of the many benefits of Bayesian methods can be found in Gill (2014).

[18]As found in Gill (2014), it is extremely likely that the choice of prior is dominated by the data.

[19]This algorithm is very suitable for regressions when high multicolinearity exists between explanatory variables. This is important when considering a quartic regression in age.

checks are performed and each control groups marginal posteriors pass the checks. With these conditions satisfactory, 10,000 iterations are run in 3 parallel chains. The previously mentioned checks for stationarity, etc. are performed on each chain and a convergence check using Gelman and Rubin's MCMC Convergence Diagnostic is performed and found to be satisfactory for each control group.(Gelman and Rubin, 1992) For the analyses on business and STEM majors, the same burn-in iterations are run, however with the reduced sample size 50,000 iterations are run following the burn-in. Each control groups marginal posteriors are checked as mentioned previously, then 20,000 draws across three chains are run. All checks are satisfactory for these converged chains of marginal posteriors.

In addition, with a Bayesian approach, six more sub-samples of 20,000 or 8,000, corresponding to each analysis mentioned above, are drawn from the white, black and Hispanic workers samples, however these sub-samples will be drawn from the other two groups of workers' discrete probability distribution of college majors in each subset of majors analyzed. For example, 20,000 white workers are drawn from the distribution of all college majors of both black and Hispanic workers, and *vice versa*. In the previous chapter a second sample was drawn for the white treatment group to test robustness and create a posterior predictive distribution to test concordance of the first treatment group's marginal posteriors of the parameters, it was assumed that the robustness and concordance tests would be satisfied. In other words, it is assumed each sub-sample would satisfy the predictive concordance check using the model and marginal posterior distributions from the first sub-sample.[20] In addition to drawing white workers from the college major distributions of black and Hispanic workers, black and Hispanic workers will be drawn from white workers' college major distribution. The HCEFs derived from these sub-samples are compared against the control groups to see the effects of college major selection on the earnings of each group by age. While it was

[20]For a detailed explanation of predictive concordance, see: Gelfand (1996)

expected based on the gender pay gap literature that the white graduates' college major distribution would result in higher earnings or given Gerhart (1990), the expectation might be a change in earnings at the beginning of the career and not the growth in earnings over the work life, however no effect is found for the distribution of all majors, major categories, and STEM degrees. Interestingly, black and Hispanic graduates seem to have superior sorting in terms of earnings in non-business majors, while in business majors black and Hispanic sorting are shown to negatively effect earnings.

Once the marginal posteriors of the parameters in the HCEF are derived, the Oaxaca-Blinder decomposition is considered to further dissect the differences between the control and treatment groups.[21] The decomposition method starting with the papers of Blinder (1973) and Oaxaca (1973) has been used in papers studying discrimination now for decades. The standard decomposition could take two forms, one assuming that white and/or male workers parameters would prevail in a market without discrimination, meaning the opposing group in the study was the victim of negative effects of discrimination, or assuming the parameters of the opposing group were the prevalent parameters in a non-discrimatory market, thus white and/or male workers were benefactors of positive discriminatory effects on wages. Cotton (1988) argued that the assumption that one groups parameters would prevail in the absence of discrimination is not correct.[22] In fact, Cotton (1988) states,

> Separately considered, each assumption abstracts from the central reality of wage and other forms of economic discrimination: not only is the group discriminated against undervalued, but the preferred group is overvalued, and the undervaluation of the one subsidizes the overvaluation of the other. Thus, the white and black wage structures are both functions of discrimination and we would not expect either to prevail in the absence of discrimination.

Thus, Cotton proposed a β^* parameter set that represented a "nondiscriminatory wage structure." This nondiscriminatory parameter set is discussed further in Oaxaca and Ran-

[21] Lin (2010) mentioned earlier uses OB decomposition to present their results.

[22] Neumark (1988) and others provided a weighting method to correct for this issue.

som (1994) where it is shown that the assumption of one or another groups parameter set as the market structure without discrimination provides extreme estimates in terms of discrimination and provided a detailed examination of weighting methods to estimate β^*. Other work has been done on correcting for issues regarding the decomposition method.[23] That being said the Oaxaca-Blinder decomposition has been used in a fair amount of discrimination studies. Darity Jr et al. (1996) uses the decomposition method to find that racial discrimination is present even when accounting for cultural differences. Hicks et al. (2018) finds discrimination against women in "all labour" markets in Australia. Arraes et al. (2014) uses Mincerian quantile regressions with the Oaxaca-Blinder decomposition to look at labor markets in Brazil at differing levels of income. They find both glass ceiling and sticky floor effects with negative income differentials in all occupations, especially unskilled manual labor occupations.[24] Schirle (2015) looks at Canada's labor market and finds while the wage structure (unexplained) gap or the gap attributable to discrimination has generally decreased over time, it still exists between male and female. Juhn and McCue (2017) look at the gender pay gap and find that men and women typically earn the same in their early careers and marriage does not have a significant effect on the gap. The gap is greatly affected by child bearing and the gap is larger in jobs that require long hours.

This study uses the Oaxaca-Blinder decomposition in a nuanced fashion. The previously mentioned studies were using it to determine the explained differences from the observable explanatory variables and the unexplained (wage structure) differences which are considered attributable to discrimination. Here it is used to find the wage structure differences which

[23]Oaxaca and Ransom (1998) derives approximate variances and provides tests for significance. Neuman and Oaxaca (2004) look at the decomposition when using the Heckman two step selection model. Elder et al. (2010) looks at the pooled weighting from Oaxaca and Ransom (1994) and concludes that without a group indicator the decomposition overstates the contribution of observables. Bauer and Sinning (2008) extends the decomposition to non-linear models.

[24]Glass ceiling is described as largest gaps in the higher quantiles, while sticky floor effect is larger gaps in the lowest quantiles.

are attributable to the differences in college major choice between the control and treatment groups. The decomposition method is taken from Fortin et al. (2011) and is shown in (3.3) and (3.4):

$$\Delta_y = y_i^T - y_i^C \tag{3.3}$$

$$= X_i^T(\beta_i^T - \beta_i^C) + (X_i^T - X_i^C)\beta_i^C \tag{3.4}$$

The explained differences or "composition effect" from observables are found in the term $(X_i^T - X_i^C)\beta_i^C$ with the control groups wage structure being the base or market wage structure for the race/ethnicity of the control group, where superscript T and C denote the treatment and control groups, respectively. The differences in wage structure or "treatment effect" due to college major choice are found in the term $X_i^T(\beta_i^T - \beta_i^C)$. The error terms do not need to be mean zero, however to satisfy "ignorability", the differences in the error terms between the control and treatment groups should be approximately mean zero as shown in (3.5).[25]

$$(\epsilon_i^T - \epsilon_i^C) \sim N(0, \sigma^2) \tag{3.5}$$

As will be shown in Section 3.4, ignorability is a reasonable assumption for this study.[26] In addition, it is known that subtle differences in work hours, weeks worked, class of worker, etc. exist between control groups and treatment groups. To attempt to account for these differences in observable characteristics in the decomposition, the explained differences are estimated by solving for the differences in explanatory variables, seen in (3.6).

$$(X_i^T - X_i^C) = (y_i^T - y_i^C)\beta_i^{Cg} - [X_i^T(\beta_i^T - \beta_i^C)]\beta_i^{Cg} \tag{3.6}$$

In (3.6), β_i^{Cg} is the generalized (*Moore-Penrose*) inverse of the control group parameters. This augmentation results in a more precise composition effect distribution, as will be shown

[25]Ignorability discussed in detail in Fortin et al. (2011).

[26]The error term differences 95 percent credible set includes 0.

in the results section. Following the augmentation of the composition effect, the results from (3.4) will be presented. Lastly, Fortin et al. (2011) discusses the decomposition for distributional statistics and Oaxaca and Ransom (1998) proposes methods for computing variances to test significance of the estimates. Fortin et al. (2011) spends a considerable amount of discussion on this and even mentions Bayes theorem when analyzing the re-weighting methods of DiNardo et al. (1996), but does not discuss Bayesian methods or inference. With the use of the marginal posteriors and the full samples, distributions for the composition and treatment effects are derived.[27]

3.4 Results

The race wage gap is prevalent even in the earnings of college graduates. Figure 3.6 shows the *HCEFs* for each group with their respective medians. The earnings gaps are evident throughout a college graduates career between white, black and Hispanic graduates. The difference for white and black graduates is approximately $816,509 over the career and that translates to black graduates earning approximately 75 percent of white graduates earnings. Hispanic graduates earn approximately 79 percent over their careers with an earnings gap of $679,511. The focus of the study is to see if sorting differences in college majors has any effect on these gaps.

3.4.1 All Majors, STEM Majors, and Non-Business Majors

When looking at the college major choices of the control and treatment groups across all majors, the results show that overall no effect is present for any of the treatment groups. These results hold for STEM degrees as a whole as well. This is not to say there is no

[27]This allows for a much simpler and straightforward method for analyzing significance with credible intervals, where at least in all the literature mentioned, the results where confined to point estimate parameters and the averages of the data.

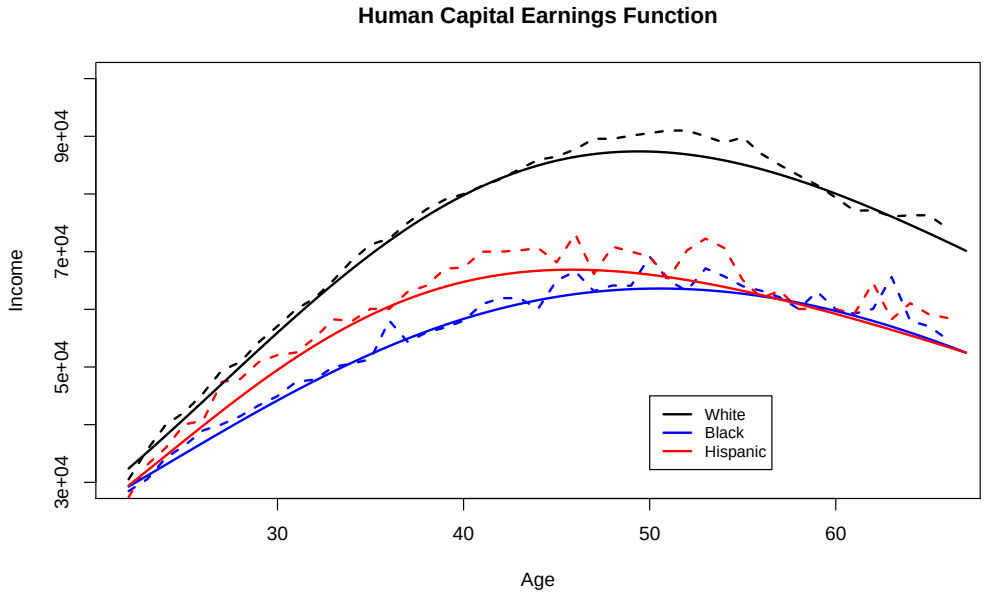

Figure 3.6. HCEFs for College Graduates

effect from college major choice, but when broken down further competing effects appear. The treatment groups for black and Hispanic graduates do not see any bump from sorting into majors like white graduates in each analysis, shown in Figures 3.7 and 3.8.[28] With the previously mentioned results, the samples are broken into more specific groups to test for any effects within those groupings. Specific business majors are considered and analyzed following a brief discussion of non-business majors.

When considering only non-business majors, the results are interesting and somewhat puzzling.[29] White graduates that sort into non-business majors like black or Hispanic grad-

[28]This could be do to the sampling where the number of samples were too high for the number of total observations in the black and Hispanic groups. The interaction between the white graduates college major probabilities and the data could be dominated by the data without replacement in sampling, thus the HCEFs will not diverge from their respective control groups.

[29]Non-business majors as a category are heterogeneous and any more specificity of majors beyond this runs into sample size issues. While not the main focus, a discussion of the results for non-business majors is appropriate for future research on the topic. Expressed another way, this is simply dropping the modal major category from all majors to view the effects of major selection independent of that category.

Figure 3.7. Black and Hispanic Treatment HCEFs - All Majors and STEM

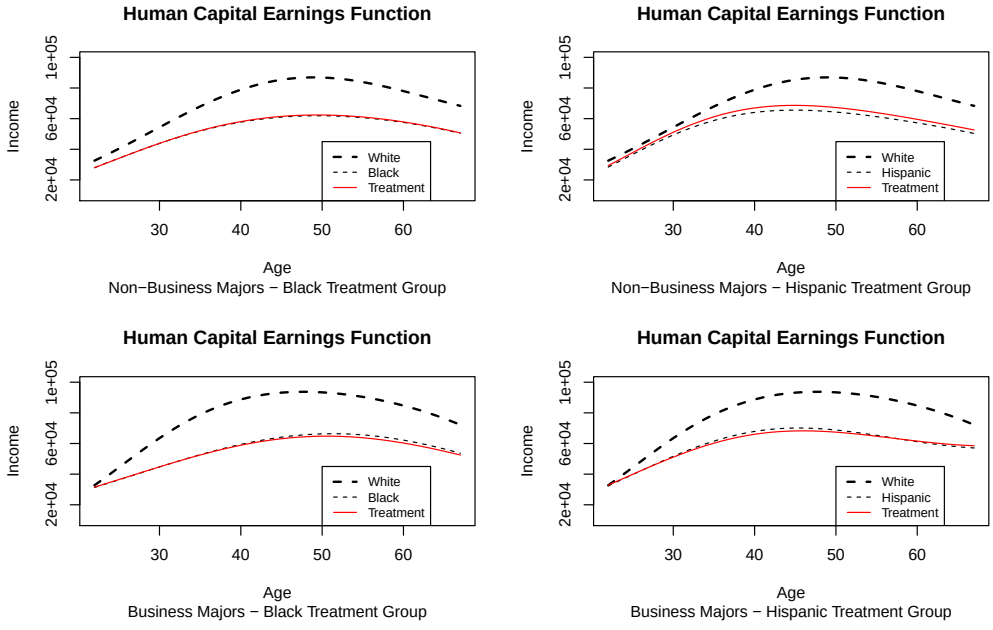

Figure 3.8. Black and Hispanic Treatment HCEFs - Business and Non-Business

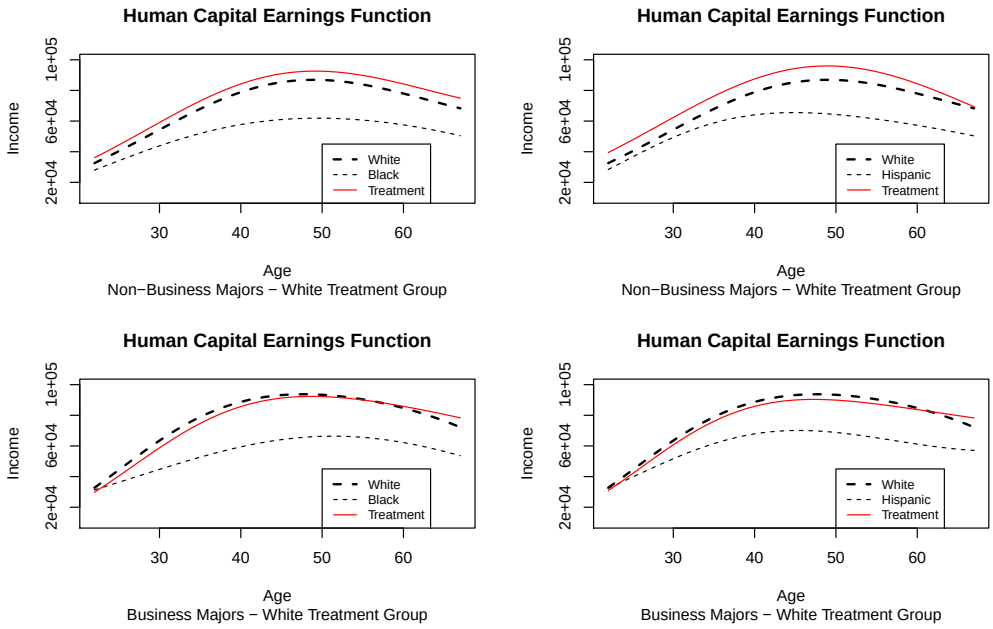

Figure 3.9. White Treatment HCEFs

uates, see an increase in earnings. This result can be seen in Figure 3.9. The white graduates treatment relative to black and Hispanic graduates see earnings increase by approximately 7.62 and 10.63 percent, respectively. This actually widens the earnings gap by 29.88 percent relative to black graduates and 36.07 percent relative to Hispanic graduates. Using the Oaxaca-Blinder decomposition for each group, the composition and treatment effects can be distinguished in these results. For the white treatment group drawn from the black non-business majors distribution, the average treatment effect is 0.083 log points and the average composition effect is -0.002 log points, thus the treatment effect is approximately all of the positive difference between the treatment group and the control group. The Oaxaca-Blinder decomposition for this group is detailed in Table 3.3.

Table 3.3. OB Decomposition for White to Black Treatment Group - Non-Business Majors

	Mean	Stan.Dev.	95% Credible Interval	Diff Attributable
Treatment Effect	0.0830	0.0066	0.0701 to 0.0959	1.0220
Composition Effects	-0.0018	0.0002	-0.0021 to -0.0014	-0.0220

For the white treatment group drawn from the Hispanic non-business majors distribution, the average treatment effect is 0.116 log points and the average composition effect is 0.005 log points, thus the treatment effect is approximately 95.95 percent of the difference between the treatment group and the control group. The Oaxaca-Blinder decomposition for this group is detailed in Table 3.4.

Table 3.4. OB Decomposition for White to Hispanic Treatment Group - Non-Business Majors

	Mean	Stan.Dev.	95% Credible Interval	Diff Attributable
Treatment Effect	0.1163	0.0067	0.1034 to 0.1293	0.9595
Composition Effects	0.0049	0.0003	0.0044 to 0.0055	0.0405

Within non-business majors the results indicate that if white graduates sorted into these majors as black and Hispanic graduates do, their earnings would actually increase. The treatment effects from the Oaxaca-Blinder decomposition show that the treatment effect is positive and significant. This is potentially strong evidence for discrimination as the earnings gap widens from the treatment. The study will conclude with business majors and the treatment effect is in opposition with non-business majors.

3.4.2 Business Majors

Unlike non-business majors, the results for business majors are negative from the treatment. White graduates that sort into business majors like black or Hispanic graduates, see a decrease in earnings. Although not starkly apparent, this result can be seen in Figure 3.9. The white graduates treatment relative to black and Hispanic graduates see earnings decrease by approximately 2.08 and 2.63 percent, respectively. This accounts for approximately 7.23 percent of the earnings gap relative to black graduates and 11.15 percent of the earnings gap relative to Hispanic graduates. Using the Oaxaca-Blinder decomposition for each group, the composition and treatment effects can be distinguished in these results. For the white treatment group drawn from the black business majors distribution, the average treatment

effect is -0.044 log points and the average composition effect is 0.014 log points, thus the treatment effect is approximately all of the negative difference between the treatment group and the control group. The Oaxaca-Blinder decomposition for this group is detailed in Table 3.5.

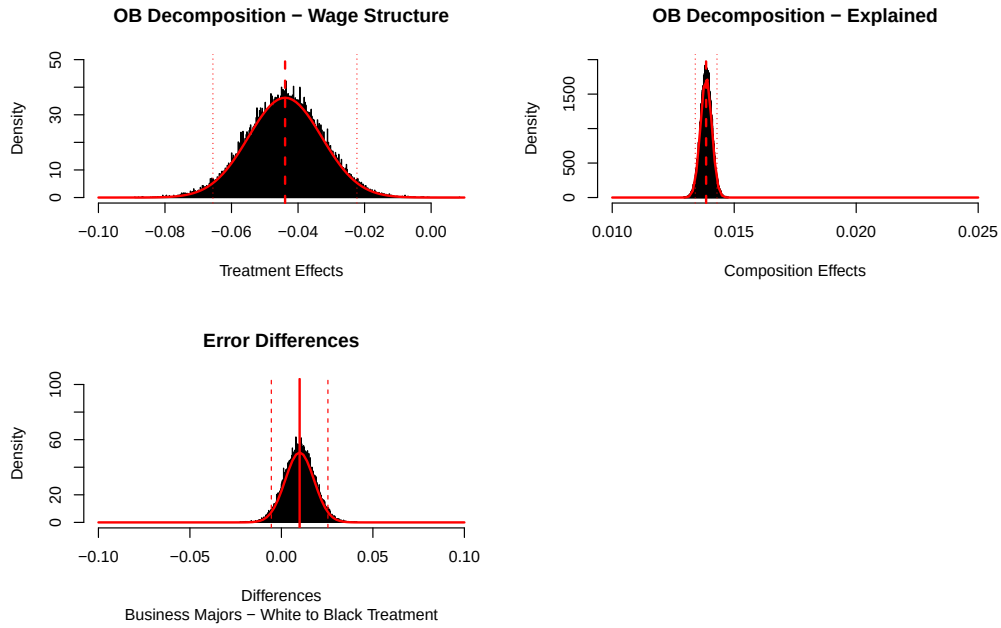

Figure 3.10. OB Decomposition for White to Black Treatment Group - Business Majors

Table 3.5. OB Decomposition for White to Black Treatment Group - Business Majors

	Mean	Stan.Dev.	95% Credible Interval	Difference Attributable
Treatment Effect	-0.0439	0.0110	-0.0656 to -0.0223	1.4609
Composition Effects	0.0138	0.0002	0.0134 to 0.0143	-0.4609

As discussed earlier, Bayesian methods allow for the Oaxaca-Blinder decomposition to be expressed in distributions. Figure 3.10 presents these results along with the differences in the error terms. The distribution of the differences in the error terms shows that ignorability is a reasonable assumption. The treatment and composition effects are significantly different from zero and show that college major sorting accounts for a portion of the earnings gap between white and black college graduates.

For the white treatment group drawn from the Hispanic business majors distribution, the average treatment effect is -0.039 log points and the average composition effect is 0.017 log points, thus the treatment effect is approximately all of the negative difference between the treatment group and the control group. The Oaxaca-Blinder decomposition for this group is detailed in Table 3.6.

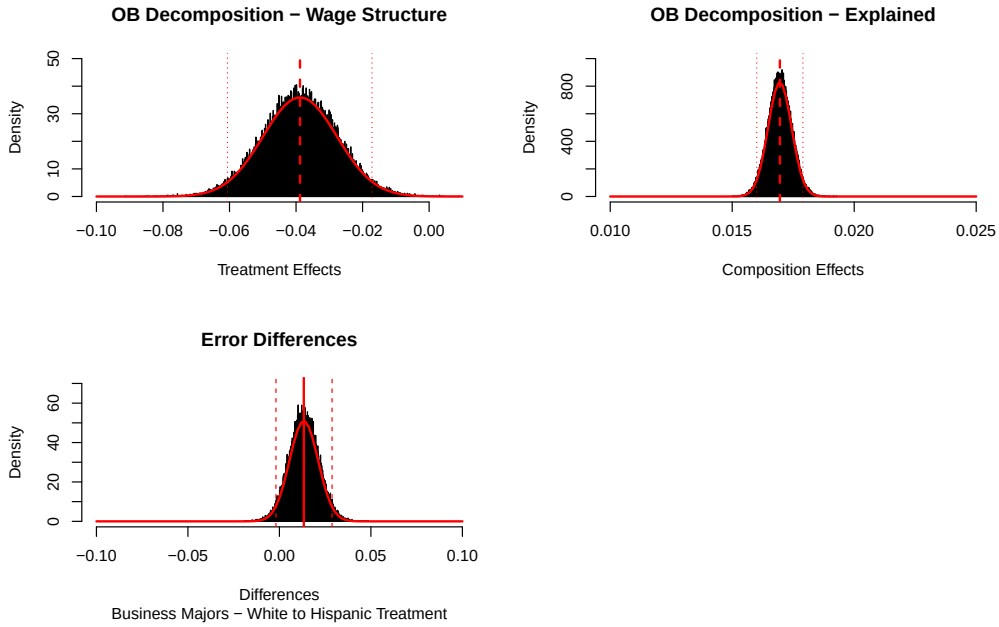

Figure 3.11. OB Decomposition for White to Hispanic Treatment Group - Business Majors

Table 3.6. OB Decomposition for White to Hispanic Treatment Group - Business Majors

	Mean	Stan.Dev.	95% Credible Interval	Difference Attributable
Treatment Effect	-0.0388	0.0111	-0.0606 to -0.0172	1.7755
Composition Effects	0.0170	0.0005	0.0160 to 0.0179	-0.7755

Looking at the distributions of the decomposition for white graduates sorting in business majors like Hispanic graduates, the treatment and composition effects are significantly different from zero and show that college major choice does account for a portion of the earnings gap between white and Hispanic college graduates.

Within business majors, the results indicate that if white graduates sorted into these majors like black and Hispanic graduates, their earnings would decrease over a standard career. The treatment effects from the Oaxaca-Blinder decomposition show that the treatment effects are negative and significant.

3.5 Conclusion

The goal of this study is to show the effect of college major selection has on the race wage gap. By drawing white workers from another group's distribution of college majors, it is shown the choices of the group on the whole reduce the HCEFs for white business majors in the treatment group, while they increase the HCEFs for non-business majors. In these regards, college major choice has an effect on the race wage gap. These competing effects explain why no difference was found when looking at all majors. In non-business degrees, the black and Hispanic graduates' college major choices appear to be superior to that of white graduates, however a sizable earnings gap still exists between white and black graduates and white and Hispanic graduates, respectively. In fact, white graduates sorting in non-business degrees appears inferior to both black and Hispanic graduates in terms of earnings, but still the gaps exist. Further examination is required to understand this gap. As in the focus of the previous chapter, could the occupational sorting of college graduates account for a portion of the gap? Could other unobservable characteristics such as family background, college debt, etc. be to blame? The glaring question is does discrimination exist at a level that earnings gaps between white graduates and black and Hispanic graduates reach $816,509 and $679,511, respectively, over a college graduate's career?

With the use of the Oaxaca-Blinder decomposition, it provided the ability to distinguish between the treatment effects of interest to the study and the composition effects of the data. The decomposition showed that all of the positive difference between the white to black treatment and the control groups in non-business majors was due to the treatment

effects, while approximately 95.95 percent of the positive difference was due to the treatment effect for the white to Hispanic treatment group in non-business majors. The fact that this effect increased the earnings gaps was interesting and puzzling. In other words, white graduates sorting in non-business majors actually decreases what the earnings gaps would be if they sorted into majors more optimally. Again, further examination is required. Then, the decomposition showed that all of the negative differences in the treatment and control groups in business majors are due to the treatment effects. Thus, college major selection by black and Hispanic graduates in business majors account for portions of the respective earnings gaps. The decomposition results showed that 7.23 and 11.15 percent of the white-to-black and white-to-Hispanic earnings gaps, respectively, are due to college major selection within business majors.

Finally, Bayesian inference also provides many tools and benefits for economists. The methods used in this study can be used in many other studies, such as gender pay gap studies like those previously mention in the literature. Using Bayesian methods allows for much more simple and straightforward inference of the differences, including the Oaxaca-Blinder decomposition. As stated for the previous chapter, the objective is to understand and eliminate any wage gap due to racial discrimination or any other prejudicial trait. Understanding the impact of college major selection on future earnings and race wage gaps will further allow those goals to be met.

REFERENCES

Addo, F. R., J. N. Houle, and D. Simon (2016). Young, black, and (still) in the red: Parental wealth, race, and student loan debt. *Race and Social Problems 8*(1), 64–76.

Altonji, J. G., P. Arcidiacono, and A. Maurel (2016). The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the Economics of Education*, Volume 5, pp. 305–396. Elsevier.

Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. *Handbook of Labor Economics 3*, 3143–3259.

Arraes, R. d. A., F. L. S. Menezes, and A. G. Simonassi (2014). Earning differentials by occupational categories: Gender, race and regions. *EconomiA 15*(3), 363–386.

Autor, D. H., L. F. Katz, and M. S. Kearney (2008). Trends in us wage inequality: Revising the revisionists. *The Review of Economics and Statistics 90*(2), 300–323.

Baker, R., E. Bettinger, B. Jacob, and I. Marinescu (2018). The effect of labor market information on community college students major choice. *Economics of Education Review 65*, 18–30.

Bartolucci, C., C. Villosio, and M. Wagner (2018). Who migrates and why? evidence from italian administrative data. *Journal of Labor Economics 36*(2), 000–000.

Bauer, T. K. and M. Sinning (2008). An extension of the blinder–oaxaca decomposition to nonlinear models. *AStA Advances in Statistical Analysis 92*(2), 197–206.

Black, D. A. and J. A. Smith (2006). Estimating the returns to college quality with multiple proxies for quality. *Journal of Labor Economics 24*(3), 701–728.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human resources 8*(4), 436–455.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 2754–81.

Carruthers, C. K. and M. H. Wanamaker (2017). Separate and unequal in the labor market: human capital and the jim crow wage gap. *Journal of Labor Economics 35*(3), 000–000.

Cotton, J. (1988). On the decomposition of wage differentials. *The Review of Economics and Statistics 70*(2), 236–243.

Darity Jr, W., D. K. Guilkey, and W. Winfrey (1996). Explaining differences in economic performance among racial and ethnic groups in the usa: the data examined. *American Journal of Economics and Sociology 55*(4), 411–425.

Daymont, T. N. and P. J. Andrisani (1984). Job preferences, college major, and the gender gap in earnings. *Journal of Human Resources 19*(3), 408–428.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review 106*(3), 479–524.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica 64*(5), 1001–1044.

Elder, T. E., J. H. Goddeeris, and S. J. Haider (2010). Unexplained gaps and oaxaca–blinder decompositions. *Labour Economics 17*(1), 284–290.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International regional science review 26*(3), 244–268.

Elhorst, J. P. (2014a). Matlab software for spatial panels. *International Regional Science Review 37*(3), 389–405.

Elhorst, J. P. (2014b). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Berlin, Heidelberg: Springer.

Elhorst, J. P. and S. H. Vega (2013). On spatial econometric models, spillover effects, and w. 53rd Congress of the European Regional Science Association: "Regional Integration: Europe, the Mediterranean and the World Economy", 27-31 August 2013, Palermo, Italy, Louvain-la-Neuve. European Regional Science Association (ERSA).

Emmons, W. and L. Ricketts (2017). College is not enough: Higher education does not eliminate racial and ethnic wealth gaps. *Federal Reserve Bank of St. Louis Review 99*(1), 7–39.

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, Volume 4, pp. 1–102. Elsevier.

Fouarge, D., B. Kriechel, and T. Dohmen (2014). Occupational sorting of school graduates: The role of economic preferences. *Journal of Economic Behavior & Organization 106*, 335–351.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, Chapter 9, pp. 145–161. London: Chapman and Hall.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. CRC Press.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science 7*(4), 457–472.

Gerhart, B. (1990). Gender differences in current and starting salaries: The role of performance, college major, and job title. *ILR Review 43*(4), 418–433.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach*, Volume 20. CRC press.

Grodsky, E. and D. Pager (2001). The structure of disadvantage: Individual and occupational determinants of the black-white wage gap. *American Sociological Review 66*(4), 542–567.

Haque, N. U. and S.-J. Kim (1995). human capital flight: Impact of migration on income and growth. *Staff Papers 42*(3), 577–607.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2003). Fifty years of Mincer earnings regressions. Technical report, National Bureau of Economic Research.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. *Handbook of the Economics of Education 1*, 307–458.

Heckman, J. J., T. M. Lyons, and P. E. Todd (2000). Understanding black-white wage differentials, 1960-1990. *The American Economic Review 90*(2), 344–349.

Hicks, J., G. Mallik, and P. Basu (2018). Earnings outcomes in metropolitan and regional labour markets? a gender-based analysis for New South Wales and Victoria. Technical report, International Institute of Social and Economic Sciences.

Juhn, C. and K. McCue (2017). Specialization then and now: Marriage, children, and the gender earnings gap across cohorts. *Journal of Economic Perspectives 31*(1), 183–204.

Katz, E. and O. Stark (1986). Labor migration and risk aversion in less developed countries. *Journal of Labor Economics 4*(1), 134–149.

Katz, L. F. and K. M. Murphy (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The quarterly journal of economics 107*(1), 35–78.

Kennan, J. and J. R. Walker (2011). The effect of expected income on individual migration decisions. *Econometrica 79*(1), 211–251.

Krueger, A. B. (2018). Inequality, too much of a good thing. In *The Inequality Reader*, pp. 25–33. Routledge.

Lang, K., M. Manove, and W. T. Dickens (2005). Racial discrimination in labor markets with posted wage offers. *American Economic Review 95*(4), 1327–1340.

Lee, L.-f. and J. Yu (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics 154*(2), 165–185.

LeSage, J. P. and R. K. Pace (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

Lin, E. S. (2010). Gender wage gaps by college major in taiwan: Empirical evidence from the 1997–2003 manpower utilization survey. *Economics of Education Review 29*(1), 156–164.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy 66*(4), 281–302.

Mincer, J. (1974). *Schooling, experience, and earnings*. Human behavior and social institutions. National Bureau of Economic Research; distributed by Columbia University Press.

Morgan, S. L., D. Gelbgiser, and K. A. Weeden (2013). Feeding the pipeline: Gender, occupational plans, and college major selection. *Social Science Research 42*(4), 989–1005.

Murphy, K. M. and F. Welch (1990). Empirical age-earnings profiles. *Journal of Labor economics 8*(2), 202–229.

Neal, D. A. and W. R. Johnson (1996). The role of premarket factors in black-white wage differences. *Journal of political Economy 104*(5), 869–895.

Neuman, S. and R. L. Oaxaca (2004). Wage decompositions with selectivity-corrected wage equations: A methodological note. *The Journal of Economic Inequality 2*(1), 3–10.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. *The Journal of Human Resources 23*(3), 279–295.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International economic review 14*(3), 693–709.

Oaxaca, R. L. and M. Ransom (1998). Calculation of approximate variances for wage decomposition differentials. *Journal of Economic and Social Measurement 24*(1), 55–61.

Oaxaca, R. L. and M. R. Ransom (1994). On discrimination and the decomposition of wage differentials. *Journal of econometrics 61*(1), 5–21.

Ochsenfeld, F. (2016). Preferences, constraints, and the process of sex segregation in college majors: A choice analysis. *Social science research 56*, 117–132.

Penner, A. M. (2008). Race and gender differences in wages: The role of occupational sorting at the point of hire. *The Sociological Quarterly 49*(3), 597–614.

Polachek, S. W. (1978). Sex differences in college major. *ILR Review 31*(4), 498–508.

Schirle, T. (2015). The gender wage gap in the Canadian provinces, 1997–2014. *Canadian Public Policy 41*(4), 309–319.

Speer, J. D. (2017). The gender gap in college major: Revisiting the role of pre-college factors. *Labour Economics 44*, 69–88.

Tibbits, M. M., C. Groendyke, M. Haran, and J. C. Liechty (2014). Automated factor slice sampling. *Journal of Computational and Graphical Statistics 23*(2), 543–563.

Tiebout, C. M. (1956). A pure theory of local expenditures. *The Journal of Political Economy 64*(5), 416–424.

Webber, D. A. (2016). Are college costs worth it? how ability, major, and debt affect the returns to schooling. *Economics of Education Review 53*, 296–310.

BIOGRAPHICAL SKETCH

Thomas F. Lanier graduated from Texas A&M University with a Bachelor of Science in Political Science in 2008. Throughout his undergraduate tenure at Texas A&M, it became more and more clear that his passion was in economics as essentially all his electives were economics courses. He went on to The University of Texas at Arlington in pursuit of a master's degree in economics. After graduating from The University of Texas at Arlington with a Master of Arts in Economics, he continued his academic endeavors at The University of Texas at Dallas in pursuit of a PhD in Economics.

While at The University of Texas at Dallas, Thomas F. Lanier held teaching and research assistant positions, from the first two years as a Teaching Assistant for Susan McElroy, PhD, to teaching his own undergraduate economics courses. In 2014, Thomas F. Lanier was hired by Litigation Analytics, Inc. as an Economist. He has remained in this position while completing his PhD program at The University of Texas at Dallas, culminating in the completion of this dissertation.

CURRICULUM VITAE

Thomas F. Lanier

October 17, 2018

Contact Information:

Litigation Analytics, Inc.
8505 Freeport Parkway, Ste. 390
Irving, TX 75063, U.S.A.

Voice: (817) 901-8508
Email: tflanier3@gmail.com

Educational History:

BS, Political Science, Texas A&M University, 2008
MA, Economics, University of Texas at Arlington, 2010
MS, Economics, University of Texas at Dallas, 2013
PhD, Economics, University of Texas at Dallas, 2018

Employment History:

Sr. Research Economist, Litigation Analytics, Inc., September 2018 – present
Economist, Litigation Analytics, Inc., September 2014 – September 2018
Teaching Assistant, University of Texas at Dallas, August 2010 – August 2014
Adjunct Professor, Tarrant County College, Fort Worth, TX, January 2012 – August 2014
Graduate Teaching Assistant, University of Texas at Arlington, May 2009 – May 2010

Professional Memberships:

American Economic Association (AEA), 2010–present
The Econometric Society, 2012–present
Society of Labor Economists, 2014–present