Macalester Journal of Economics

Volume 28 Spring 2018

Table of Contents

Published annually by the Macalester College Department of Economics and Omicron Delta Epsilon

Macalester Journal of Economics

Volume 28 *Spring 2018*

A Note from the Editors

During the spring of 2018, we had the pleasure of reading papers submitted by Macalester economics students for this edition of the Macalester Journal of Economics. These papers cover a broad range of topics, but all serve as a testament to the high level of academic excellence at Macalester, especially within the Economics Department. After much deliberation, we have selected the following five papers. We believe they best capture the quality of economics research completed by students in the Macalester Economics Department.

We would like to thank all of the faculty for their integral roles in the development of these research projects. We would also like to thank Jane Kollasch and Professor Mario Solis-Garcia for their support in the creation of this journal. Finally, we would like to congratulate Jane on her retirement and all of her contributions to Macalester and the Economics Department.
Nicoletta Peters '18 Van Anh Le '18 Jack McCarthy '18 Lily Stein '18 Elliot Cassutt '18

Foreword

The Macalester College chapter of Omicron Delta Epsilon, the international honors society in economics, proudly edits the *Macalester Journal of Economics* every year. This year's editors – Nicoletta Peters '18 (Mason City, Iowa), Van Anh Le '18 (Ho Chi Minh City, Vietnam), Jack McCarthy '18 (Evanston, Illinois), Lily Stein '18 (Chicago, Illinois), and Elliot Cassutt '18 (Minneapolis, Minnesota) – have carefully selected five papers on a variety of important topics. These papers are a sample of the research that our students produced in the last academic year.

As cities grow, the commute time from point A to point B rises as well. Los Angeles is one such place. While public transportation offers a convenient alternative to *not* be stuck in traffic, it is still an open question whether an increase in the supply of public transportation will raise the demand for it. But what does the data say? Alex Ramiller '18 (Portland, Oregon) puts forward an interesting conjecture: if public transportation is valuable, then housing that is close to a transportation hub should increase in value. Using the Expo light rail line as a natural experiment, together with micro-level data and econometric techniques that account for spatial endogeneity, Alex finds support for a positive relationship between property values and public transportation proximity.

Renewable energy – and wind in particular – has been on the rise in the last years, especially in the U.S. Midwest. What do we know about the responses of energy utilities to these developments? Jack McCarthy '18 (Evanston, Illinois) goes deep into this topic and sets forth to understand how these companies have substituted wind energy for other types of generators. Together with a detailed dataset and a well-crafted econometric design, he finds evidence that between 2013 and 2016, wind generation reduced price and increased transmission constraints when replacing coal generation.

When economists want to talk about economic growth, the starting point is usually the Solow growth model. One of the predictions of the model is that of *convergence:* other things the same, poor countries will catch up with rich countries given sufficient time. This prediction has been extensively tested at a country level – offering mixed support for the implication of the model – but how does the convergence hypothesis fare at a regional level? Chan Wang ’19 (Beijing, China) takes the Solow model to the U.S. states and asks whether the data offers support for convergence. He finds that it does: using a long sample from 1930 to 2015, U.S. states seem to converge – though at a lower rate relative to the international evidence benchmarks. Perhaps surprisingly, Chan finds that convergence breaks down in a subsample that starts in 1980 and ends in 2010, which many researchers associate to the technological innovations that brought the personal computer and the Internet.

It’s fair to say that if you’re into baseball, you’ve read Michael Lewis’s *Moneyball* (or at least watched the movie!). If this is not the case, here’s the one-sentence summary: at the turn of the century, the Oakland Athletics changed the way teams looked at prospective players by adding a heavy dose of statistical analysis, which is known in the business as “sabermetrics.” In terms of financial economics: can we find an asset (player) that is underpriced (undervalued) by the market? Matthew Yang ’19 (Bethseda, MD) takes the Moneyball book as a starting point and asks: to what extent was the Major League Baseball (MLB) labor market efficient previous to 2003, right when the Athletics started their new practices? He finds that the market was indeed inefficient, but that the inefficiencies began to dissipate after the book (and arguably, the methodologies as well) was published, as the other MLB teams caught the sabermetrics bug.

Finally, Nana Adom Mills-Robertson ’18 (Accra, Ghana) and Chan Wang ’19 (Beijing, China) ask to what extent overconfidence drives the dynamics of the entrepreneurial world. This is a question worth asking since business startups suffer from high rates of failure and low rates of economic return. Is there anything special about entrepreneurs that can account for these features? To answer this, they build a market entry experiment that incorporates many insights from behavioral economics. They find that overconfidence increases the rate at which entrepreneurs enter a market and reduces the value of profits, which seems to fit with the empirical evidence. They also show that as entrants know more about the market, the negative effects of overconfidence can be reduced.

On behalf of my colleagues in the Economics Department, I am delighted to present the research of these talented students. I am confident that you will find it enlightening and be impressed by the value of a liberal arts education.

Mario Solis-Garcia
Assistant Professor of Economics

On the Right Track?

Property Values and Rail Transit in Los Angeles County

Alex Ramiller

Introduction to Econometrics

Abstract

Previous studies have examined the relationship between public transit investment and property values, using a variety of empirical methods to produce a broad array of results. This study contributes to that literature by addressing the potential for spatial endogeneity in property values, using a generalized spatial two-stage least squares regression in order to measure the effect of proximity to light rail on property values in Los Angeles County. Focusing on the "Expo" light rail line between downtown Los Angeles and Santa Monica, this study finds that a statistically significant positive relationship between proximity to station areas and property values appeared after the line officially opened in 2012.

I. Introduction

Los Angeles County is the largest county in the United States, with a population larger than that of forty states, but it is perhaps best known for its car-centric culture and high levels of congestion. According to the State of California Department of Motor Vehicles, Los Angeles County's 10.2 million people owned 6.5 million registered automobiles and 8 million total registered vehicles in 2016 (California DMV, 2016). As a consequence, Los Angeles drivers consistently experience the worst traffic congestion in the country, spending an average of 81 hours per year waiting in traffic in 2015 (Nelson, 2016). In response, Los Angeles has increasingly invested in public transit, and particularly in high-capacity forms of transit such as subways and light rails. The Los Angeles "Metro Rail" system, which is operated by the Los Angeles County Metropolitan Transit Authority or "Metro," has grown from a single line in 1990 to six lines, with the most recent addition of the Expo Line in 2012 (see Figure 1). However, in spite of these expansions, the Los Angeles system remains much less popular than those in other cities. Although it is the second largest city in the United States, its rapid transit system is ranked 8^{th} in terms of ridership, with 25 times fewer annual rides than the New York City subway (see Figure 2) (Dickens, 2017).

Nonetheless, the rapid expansion of Metro Rail has led some to declare Los Angeles "America's next great mass transit city" (Yglesias, 2012). Residents appear to agree with this assessment, consistently supporting ballot measures to expand the County's transit system. In 2017, 71.15% of voters supported a ballot initiative known as "Measure M," providing funding for infrastructure upgrades and further transit expansions (Metro, 2016). In spite of this high level of political support, it remains an open question whether the rail system is widely valued in a city that relies on cars. In a recent op-ed, social scientist Joel Kotkin critiques the expansion of rail transit in Los Angeles, pointing out that the percentage of commuters using transit has declined as the system has expanded and arguing that the system is not sufficiently popular to justify further expense (Kotkin, 2016).

This paper examines the effect of proximity to newly built rail transit stations on property values in Los Angeles County in order to determine whether access to rail transit is valued in the Los Angeles County property market. The following analysis is divided into six sections. I begin with a review of the basic economic theory behind how property values are determined and previous empirical research into the relationship between property values and rail transit. I then establish a conceptual model of the factors influencing property values in Los Angeles, followed by a description of the data used in this analysis. I then introduce my empirical models, which are subsequently used for an econometric analysis of properties lying within a one mile radius of Expo Line stations. I conclude that a statistically significant positive relationship appears between property value and proximity to stations that began with the opening of the Expo Line in 2012.

II. Literature Review

a. Economic Theory

Unlike most goods for which prices are determined by a single attribute, properties are highly heterogeneous and have a number of different attributes that contribute to their prices. Analysis of factors affecting property values thus requires a "hedonic price model," which postulates that the observed prices of differentiated products in a given market are affected by the "implicit" or built-in prices for individual product attributes. Assuming the existence of these implicit prices, Rosen (1974) theorizes that a first-step regression analysis with the price of a product regressed on its various attributes would demonstrate the marginal willingness to pay for individual product attributes through coefficient outputs. This model can easily be applied to urban property markets, which are characterized by extreme product-level heterogeneity in both endogenous and exogenous characteristics (Hill, 2013). In most standard hedonic models, the characteristics of properties that can affect price are generally categorized by structural characteristics, locational characteristics, and neighborhood characteristics (Can, 1990).

Structural characteristics refer to individual property attributes such as lot size, age of building, or number of bedrooms, which can have an effect on price variation between properties. Locational characteristics refer to the site of a property relative to its surroundings, and include proximity to the city center or other significant sources of economic activity such as shopping centers (Sale, 2017). Finally, neighborhood-level characteristics such as affluence, racial composition, and physical configuration can influence perceptions and thereby affect property values (Baranzini et al., 2008; Rauterkus & Miller, 2011).

b. Previous Empirical Research

There have been many attempts to quantify the relationship between proximity to transit and property values over the past several decades, with many variations to the hedonic price model. As it is impractical to conduct a comprehensive review of literature on the topic, I focus on studies that are the most methodologically relevant to my own analysis.

Hess and Almeida (2007) provide a useful baseline for how to conduct a hedonic analysis with rail transit proximity through their study of residential property values and light rail stations in Buffalo, New York. Hess and Almeida follow the traditional model, separating the determinants of assessed property value into the three primary categories of structural characteristics, locational amenities, and neighborhood characteristics. Each of these categories contain a number of explanatory variables. Within structural characteristics, Hess and Almeida include size, building age, number of bedrooms and bathrooms, and whether or not the building has a fireplace, among other variables. For locational amenities, they include distance to the downtown Central Business District (CBD) and distance to the nearest park. For neighborhood characteristics, they look at median household income, crime rates, and occupancy change rates.

Hess and Almeida then add a fourth category representing proximity to a light rail station. This proximity is calculated using a "Nearest Neighbor Distance" formula, which determines the distance between that property and the nearest light rail station. Using this method, they find that every foot closer to a light rail station in straight-line distance increases property values by $2.31, and every foot closer along street networks increases values by $0.99.

This study reflects a cross-sectional approach to hedonic modelling, but hedonic models can also be expanded to account for changes over time by using dummy variables to represent observation years. Dummy variables for years are frequently employed in studies that look at property sales rather than assessed values, denoting the year in which a property was sold. In order to examine the effect of the Midway Line opening in Chicago in 1993, McMillen and McDonald (2004) use a "repeat-sales" model with dummy variables to separate property sales by year. They find that in sales prior to the Midway Line opening, property prices decreased by 4.2% with each additional mile from the station sites, while sale prices following the opening of the line decreased by 19.4% with each additional mile.

Similarly, Wang (2010) uses dummy variables for every year between 2001 and 2008 to look at the impact of a Shanghai subway line that began construction in 2001 and opened in 2007 on the prices of new residential properties. Focusing primarily on the period during construction, when the subway line was anticipated but not yet completed, Wang finds that property values increased almost every year. Zheng et al. (2016) further expand on this approach by interacting year dummies with dummy variables for specific districts. By interacting these temporal and spatial variables, they are able to "control for time-variant unobserved factors at the urban district level" (29).

In addition to correcting for changes over time, it is important for hedonic price models to take into account the possibility of spatial autocorrelation. Spatial autocorrelation refers to the possibility of correlation between observations that are spatially related to one another. Diao (2015) finds that failing to correct for this type of autocorrelation may result in a significant bias by distorting estimates of how much an independent variable such as transit proximity contributes to property values. Sun et al. (2015) attempt to correct for this potential for spatial autocorrelation by employing three different spatial regression techniques: the Spatial Error Model, the Spatial Autoregressive Model, and the Spatial Durbin Model. These models produce similar results to one another and do not differ substantially from OLS regressions, but are nonetheless regarded as more effective because they account for spatial autocorrelation. As Xu et al. (2016) explain, these models use a matrix that weights the spatial relationships between each observation in order to correct for the influence of spatial interactions. The weights in this matrix can be calculated using either "contiguity" – the number of observations directly adjacent to a given observation – or distance measures based on a certain threshold (225). Using these spatial regressions still produce significant results in transit proximity studies: using a spatial error model and a spatial autoregressive model, Xu et al. find a positive effect on property value within 400 meters of transit stations in Wuhan, China.

A final important consideration in hedonic price modelling is the functional form of the estimation equation, which can vary depending upon the type of analysis being performed.

McIntosh et al. (2014) note that in addition to a linear functional form that measures unit change in land value relative to unit increase in distance, it can be beneficial to partially log the estimation equation. A log-linear form, in which only the dependent variable of property value is logged, produces coefficients showing the percentage change in property value for a unit increase in distance. A log-log form, meanwhile, logs all variables and produces coefficients that show the *elasticity* of each independent variable with respect to property value. McIntosh et al. find that the most common functional forms in hedonic price models are linear and log-linear. They conclude that choice of functional form is a crucial component of hedonic price modeling, and that multiple forms should be used to ensure robustness.

These methods can be used to address a gap in the literature regarding the effects of transit on property values in a largely car-centric city like Los Angeles. The only existing study focusing on rail transit in Los Angeles was a cross-sectional hedonic price analysis published fifteen years ago, which found effects on property values that were uneven and inconsistent (Cervero & Duncan, 2002). This study does not reflect the addition of three rail lines and two Bus Rapid Transit (BRT) lines between 2002 and 2017, and did not use techniques such as multi-year regressions, spatial autoregressive models, and non-linear functional forms. This suggests that an updated analysis of Los Angeles rail transit and property values is long overdue.

III. Conceptual Model

Following previous empirical research, I consider property value as a function of individual property characteristics, amenities associated with each property, and broader neighborhood characteristics. When proximity to transit is also included as a factor, I arrive at the following functional form for the value of an individual property (V_i):

$$V_i = f\ (T_i, S_i, L_i, N_i)$$

Where T_i is the distance from that property to the nearest transit line, S_i is the sum of that individual property's structural characteristics, L_i is the sum of locational amenities associated with the property, and N_i is the sum of characteristics of the neighborhood within which the property is located. Ideally, a hedonic price analysis would include all of the variables necessary to fully explain each of these four categories, without excessive multicollinearity in the explanatory variables.

IV. Data and Measurement Issues

For this study, I examine properties that are within a one-mile radius of Expo Line stations. In order to create my dataset, I retrieved geospatial data for property tax parcels from the Los Angeles County GIS Data Portal and Metro Rail stations from the Los Angeles County Metropolitan Transit Authority. I established one mile buffers around Expo Line stations and identified any properties lying within those buffers (see Figure 3). Data concerning these properties originated from Los Angeles County Tax Assessor's Office, the Los Angeles County GIS Data Portal, and the 2015 American Communities Survey (ACS) 5-year estimate. Together, these sources provide data for each variable type in the hedonic model: property value, transit proximity, structural characteristics, locational amenities, and neighborhood characteristics.

Data on property values and structural characteristics for the years 2008, 2010, 2012, 2014, and 2016 are drawn from the Los Angeles County Tax Assessor's Office, which includes property-level information on assessed value, building square footage, number of units, number of bedrooms and bathrooms, and the year that the property most recently underwent a major remodel (also known as "effective year"). For an ideal hedonic price estimation, I would need to have accurate data on any housing characteristic that could feasibly impact valuation, including not only physical characteristics such as property size but also less tangible factors such as the view from a given property (Gillard, 1981). Unfortunately, I can only use the aforementioned variables in my analysis because additional data is not available.

Using property-level structural characteristics as explanatory variables introduces a significant possibility of multicollinearity, as many structural property characteristics are directly related to one another. For example, larger properties tend to have more bedrooms and bathrooms than smaller properties, as well as more units. In order to detect significant multicollinearity in these property characteristic variables, I examine their pairwise correlation coefficients, and find that *all* of the variables are significantly correlated except for effective year (see Figure 4). To avoid distortions resulting from multicollinearity, I therefore limit the number of explanatory variables I use for structural characteristics. Effective year can remain in the model as it does not appear to be collinear with any other structural characteristics, and square footage can be used as a proxy for the other highly collinear variables.

Data accuracy is less of a concern for the structural characteristics provided by the Tax Assessor's Office, but there are potential measurement issues with their property value data. Ideally, I would have property sale data with the precise prices and dates at which individual properties were bought and sold. Instead, the Tax Assessor's office provides property value estimates without disclosing the precise methodology behind their calculations, making it impossible to determine what factors influence the data they provide. Despite these potential flaws, however, previous studies have used assessed values in their models as well (Hess & Almeida, 2007).

Transit proximity and locational amenities are calculated using a Geographic Information System, with coordinates for individual parcels provided by the Los Angeles County GIS Data Portal. Using a "nearest-neighbor" algorithm, which calculates the straight-line distance from an observation to the nearest designated "hub," I create variables for the distance between each property and the nearest Metro station, the nearest open space, and the CBD (see Figure 5). While this provides a useful sense of how far away any given property is from these amenities, it is important to note that people do not move around cities in straight lines. These distance measurements can approximate spatial relationships, but they do not reflect lived experiences.

Finally, neighborhood characteristics are represented by median household income data at the census tract level provided through the 2015 ACS 5-year estimate. Median household income data can serve as an effective proxy for neighborhood characteristics, but it also has measurement issues due to the large margins of error in ACS data at the census tract level. In addition, the ACS provides only a single number for median household income, whereas the Assessor's Office data is available for multiple years. Therefore, changes in median income over time are not reflected in this model. This category would ideally also include other forms of data such as crime rates, but comprehensive geospatial datasets for crime were not made available by Los Angeles County.

One additional piece of information that is not available but that would likely have a significant impact upon this model would be whether or not residents of any given property own a car and whether a property includes parking. The intention of this study is to determine if the relationship between transit proximity and property values still holds given the prevalence of car ownership among Los Angeles County residents, and data on car ownership would make it possible to separate out the effects of transit proximity for car owners relative to others.

V. Empirical Strategy

Unfortunately, it is nearly impossible to create a hedonic model that takes into account every determinant of property values due to the difficulty of gathering data on all potentially relevant variables. Instead, I use the available data to create an estimation equation that incorporates factors that are likely to have the largest influence on property value. As discussed previously, certain property-level variables such as the number of bathrooms, number of bedrooms, and number of units are excluded as a result of multicollinearity. This results in the following estimation equation:

$$\textit{Property Value}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\textit{Nearest Station Distance}_i + \boldsymbol{\beta_2}\,\textit{Property Size}_i + \boldsymbol{\beta_3}\,\textit{Effective Year}_i + \boldsymbol{\beta_4}\,\textit{Nearest Park Distance}_i + \boldsymbol{\beta_5}\,\textit{CBD Distance}_i + \boldsymbol{\beta_6}\,\textit{Median Income}_i + \boldsymbol{\varepsilon_i}$$

In this model, the coefficient on station distance (β_1) should be negative because previous empirical research suggests that shorter distances to rail stations are associated with higher property values. The coefficient on property size (β_2) should be positive, because larger properties should generally have higher values. The coefficient on the building's "effective year" (β_3) should also be positive, because buildings that have been remodeled more recently are likely to be higher in value. The coefficient on distance to the nearest park (β_4) should be negative, because increased proximity to an environmental amenity such as a park would theoretically lead to an increase in property values (Gibbons et al., 2014). The coefficient on distance to the CBD (β_5) should be negative for the same reason, but this is less certain because the Los Angeles metropolitan area is polycentric and access to the downtown may not be as valuable to residents

(Gordon et al., 1986). Finally, the coefficient on median household income ($ß_6$) should be positive, because a more affluent neighborhood is likely to have higher property values.

I use this basic equation to test the data in three primary ways. First, I run a standard OLS regression for property value observations in 2016. Second, I employ an autoregressive "generalized spatial two-stage least squares" (GS2SLS) model to correct for the possible presence of spatial autocorrelation. Due to operating constraints, the dataset is reduced to only properties within a half mile of a station and the GS2SLS model is chosen over the more computationally complex "maximum likelihood" model (Kelejian & Prucha, 1998). Finally, following previous studies such as McMillen and McDonald (2004), I also create dummy variables for each year present in the data (2008, 2010, 2012, 2014, and 2016) and spatial interaction variables that multiply those dummy variables by the station distance to create the following estimation equation:

$$\begin{aligned} \textit{Property Value}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\textit{Nearest Station Distance}_i + \boldsymbol{\beta_2}\,\textit{Property Size}_i + \boldsymbol{\beta_3}\,\textit{Effective Year}_i \\ + \boldsymbol{\beta_4}\,\textit{Nearest Park Distance}_i + \boldsymbol{\beta_5}\,\textit{CBD Distance}_i + \boldsymbol{\beta_6}\,\textit{Median Income}_i \\ + \boldsymbol{\beta_7}\,\textit{Year Dummies}_i + \boldsymbol{\beta_8}\,\textit{Years*Station Distance}_i + \boldsymbol{\varepsilon_i} \end{aligned}$$

I expect the signs on the year dummy variables to steadily increase, reflecting the increases in property values across the study area (see Figure 6). If theory holds and proximity to a transit station has a significant effect on property value, I would expect to see the coefficients on the interaction variables move in a negative direction following the opening of the Expo Line in 2012.

For robustness, I modify the property value and station distance variables in each of these three models to confirm the strength of the relationship between the two variables. Following McIntosh et al. (2014), I use both a linear model and a log-linear model to correct for extreme outliers (such as a single 913-unit property worth more than $400 million) and any other non-linearities in the data. In addition, due to the fact that shorter distances to rail stations are likely to be exponentially more valuable than greater distances, I also try squaring the station distance variable. These modifications appear to improve the distribution of the property value data in relation to station distance. Logging the property value significantly improves the distribution of observations, while squaring the station distance makes a much smaller difference (see Figure 7).

In summary, I use three separate models: 1) a "simple" OLS for 2016 observations, 2) an autoregressive GS2SLS for 2016 observations within a half-mile radius, and 3) a "complex" OLS with year and interaction dummy variables for 2008-2016. In each of those models I conduct four separate regressions: 1) the original estimation, 2) logged property value, 3) squared distance to nearest station, and 4) both logged property value and squared distance.

VI. Results

a. Simple OLS

The results of the OLS regressions on 2016 property values are summarized in Table 1. The signs of the coefficients for the two linear regressions largely match expectations, with the exceptions of effective year and distance to the CBD. While it is difficult to explain the negative coefficient on the effective year of the building, the positive coefficient on distance from the CBD could be explained by the fact that properties further from the downtown and closer to the beach are more desirable. The strong negative coefficients for both station distance and squared station distance indicate that this model largely fits with the findings of previous empirical research: property values decrease by $200,914 per mile from the station, translating to a decrease of $38 per additional foot of distance. By contrast, while the log-linear models make the sign on effective year more consistent with theory, they do not find statistically significant negative signs on the coefficients. This heterogeneity between the linear and log-linear models may be due to the high level of heteroscedasticity in the data. Given this, I focus on the output of the linear OLS regressions, as the higher R^2 value suggests that they do a better job of explaining variation in property value data. Neither model explains property value variance well, however, and the residuals for the linear and log-linear models are both highly clustered for property values and station distance (see Figures 8 and 9).

Table 1. Results of linear and log-linear OLS regressions for 2016 observations (1-mile radius)

VARIABLES	(1) Linear OLS	(2) Linear OLS	(3) Log-Linear OLS	(4) Log-Linear OLS
Station Distance	-200,914***		-0.00495	
	(-8.118)		(-0.316)	
Square Feet	207.8***	207.8***	2.85e-05***	2.85e-05***
	(352.2)	(352.1)	(76.39)	(76.37)
Effective Year	-256.4***	-257.1***	0.000614***	0.000613***
	(-12.57)	(-12.60)	(47.59)	(47.49)
Park Distance	-51,040*	-50,281*	-0.245***	-0.244***
	(-1.715)	(-1.689)	(-13.01)	(-12.98)
CBD Distance	14,103***	13,307***	0.0491***	0.0485***
	(8.221)	(7.783)	(45.24)	(44.84)
Income	5.819***	5.863***	7.63e-06***	7.65e-06***
	(23.09)	(23.27)	(47.88)	(48.03)
Station Distance^2		-129,159***		0.0429***
		(-6.250)		(3.283)
Constant	209,233***	146,517***	10.71***	10.70***
	(4.883)	(3.534)	(395.3)	(408.2)
Observations	70,039	70,039	70,039	70,039
R-squared	0.641	0.640	0.218	0.218
Adjusted R-squared	0.640	0.640	0.218	0.218
F	20797	20785	3260	3262
RSS	1.620e+17	1.620e+17	64734	64724

t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

b. Spatial Autoregressive Model (GS2SLS)

This inconsistency between the signs of the station distance coefficients is resolved in the spatial autoregressive models (see Table 2). After correcting for spatial autocorrelation in the data, the GS2SLS model consistently finds even larger negative coefficients on station distance than the OLS model: the effect of station distance on property values is four times as large, with property values decreasing $800,000 or 50% with a one-mile increase from the station. The other signs also remain largely consistent with theory, with the exception of the positive coefficient on distance to the nearest park in the linear models. This would suggest that proximity to park space is not desirable for Los Angeles residents, perhaps because the ability to travel by car would eliminate the importance of nearest-neighbor proximity.

Table 2. Results of linear and log-linear GS2SLS regressions for 2016 observations (0.5-mile radius)

VARIABLES	(1) Linear GS2SLS	(2) Linear GS2SLS	(3) Log-Linear GS2SLS	(4) Log-Linear GS2SLS
Station Distance	-817,887***		-0.503***	
	(-4.448)		(-6.483)	
Square Feet	270.0***	270.0***	1.93e-05***	1.93e-05***
	(203.7)	(203.7)	(34.17)	(34.17)
Effective Year	-536.8***	-537.7***	0.000887***	0.000887***
	(-8.582)	(-8.596)	(33.80)	(33.81)
Park Distance	203,270**	206,132**	-0.0418	-0.0391
	(2.261)	(2.294)	(-1.104)	(-1.033)
CBD Distance	33,000***	33,413***	0.0440***	0.0438***
	(4.347)	(4.394)	(12.13)	(12.08)
Income	9.300***	9.298***	6.35e-06***	6.35e-06***
	(9.512)	(9.506)	(18.26)	(18.26)
Station Distance^2		-1.298e+06***		-0.833***
		(-4.483)		(-6.784)
Constant	1.286e+06***	1.175e+06***	11.42***	11.37***
	(7.587)	(7.412)	(139.1)	(146.7)
Lambda	-2.79e-06***	-2.80e-06***	-1.21e-07***	-1.22e-07***
	(-8.649)	(-8.649)	(-22.59)	(-22.66)
Observations	16,677	16,677	16,677	16,677

z-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

c. *Complex OLS*

My final model uses dummy variables for different years and finds that the opening of the Expo Line in 2012 had a statistically significant effect on the relationship between property values and proximity to station sites (see Table 3). The coefficients on the year dummy variables reflect that fact that property values increased throughout the study area between 2008 and 2016. More significantly, the coefficients on the interaction variables for 2014 and 2016 consistently show coefficients that are more negative than the coefficients for 2008 and 2010. The models with normal distance showed no statistically significant relationship between property value and proximity to the station site in the years before the opening of the Expo Line, but property values decreased by $70,000 per mile in 2014 and $132,000 in 2016. The models with squared distances show that a negative relationship already existed, but it more than doubled in 2014 and 2016. These results show that the Expo Line had a significant effect on property values in relation to the distance from station sites.

Table 3. Results of linear and log-linear OLS regressions for 2008-2016
(1-mile radius)

VARIABLES	(1) Linear OLS	(2) Linear OLS	(3) Log-Linear OLS	(4) Log-Linear OLS
Station Distance	-46,866***		0.137***	
	(-2.594)		(8.992)	
Square Feet	125.2***	125.2***	2.80e-05***	2.80e-05***
	(636.6)	(636.6)	(168.6)	(168.6)
Effective Year	-86.81***	-87.21***	0.000657***	0.000656***
	(-12.25)	(-12.31)	(109.8)	(109.6)
Park Distance	-77,914***	-77,724***	-0.238***	-0.238***
	(-7.865)	(-7.846)	(-28.49)	(-28.51)
CBD Distance	15,506***	15,338***	0.0552***	0.0551***
	(26.94)	(26.71)	(113.5)	(113.7)
Income	3.944***	3.948***	6.65e-06***	6.64e-06***
	(47.03)	(47.08)	(93.87)	(93.69)
Station Distance^2		31,722**		0.259***
		(2.265)		(21.91)
2008	-37,109**	14,192	-0.0441***	0.0600***
	(-2.167)	(0.855)	(-3.048)	(4.276)
2010	-30,469*	20,260	-0.0430***	0.0574***
	(-1.831)	(1.266)	(-3.055)	(4.249)
2014	104,210***	154,947***	0.212***	0.313***
	(6.278)	(9.708)	(15.15)	(23.22)
2016	215,734***	266,460***	0.329***	0.430***
	(13.00)	(16.70)	(23.50)	(31.92)
2008*Distance	14,177	-69,982***	-0.0210	-0.191***
	(0.543)	(-2.784)	(-0.951)	(-8.997)
2010*Distance	11,291	-72,089***	-0.0112	-0.176***
	(0.443)	(-2.959)	(-0.519)	(-8.562)
2014*Distance	-69,999***	-153,387***	-0.148***	-0.313***
	(-2.752)	(-6.312)	(-6.903)	(-15.27)
2016*Distance	-131,722***	-215,096***	-0.156***	-0.321***
	(-5.181)	(-8.855)	(-7.280)	(-15.67)
Constant	-11,696	-51,984***	10.32***	10.30***
	(-0.644)	(-3.304)	(672.8)	(774.8)
Observations	345,082	345,082	345,082	345,082
R-squared	0.544	0.544	0.231	0.232
Adjusted R-Squared	0.544	0.544	0.231	0.232
F	29365	29364	7400	7437
RSS	4.360e+17	4.370e+17	311684	311323

t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

VII. Conclusion

a. Summary

This paper examined the effect a new light rail transit line on property values in Los Angeles County in order to determine whether residents valued newfound access to rail transit. Consistent with previous empirical research, I find that there generally appears to be a positive relationship between rail transit proximity and property values, particularly after correcting for spatial autocorrelation. I also find that the opening of the Expo Line in 2012 had an effect on property values in subsequent years, strengthening the relationship between rail station proximity and property value.

b. Limitations

This preliminary analysis was limited primarily by a lack of adequate and accurate data. The data provided by the Los Angeles County Tax Assessor's Office, for example, used *assessed* property values rather than the actual prices at which properties were sold. This potentially distorted the results, as estimates of property values cannot possibly be entirely accurate relative to the actual functioning of the property market. In addition, while the calculations for distance to the nearest station, nearest park, and the CBD were geographically accurate, they were straight-line distances that did not take into account the actual ways that people move throughout the city. Using "network distance" that takes the street network into account (e.g. Hess & Almeida, 2007) would potentially provide a more realistic measure of proximity to a transit station. Overall, all of these models suffered from a lack of effective explanatory variables, with no model seeming to explain more than 64% of the variation in property value data. More comprehensive data from the ACS and Los Angeles County would potentially help to fill this gap.

Another limitation to this study was a lack of processing power, which prevented me from running spatial autoregressive models that are more computationally complex. Given that the initial size of the dataset for 2016 property value data was over 70,000 observations, Stata was unable to create a matrix of the size necessary to compute spatially weighted values for each property. Therefore, I had to reduce my sample size to properties within a half mile of the Expo Line stations, which potentially obscured the effects of properties at distances between 0.5 and 1 miles. Computational complexity also prevented me from running a "maximum likelihood" regression, which meant that I had to rely on a single spatial autoregressive model (GS2SLS) to correct for spatial autocorrelation. Finally, lack of computing power prevented me from including a spatial error term in my spatial models, which potentially affected the output.

c. Future Research

The relationship between rail transit and property values in Los Angeles County remains an interesting question, particularly in the context of a continuously expanding rail transit network. While some would argue that rail transit is not valuable in a metropolitan area that is synonymous with sprawl and automobile transportation, these initial findings suggest that residents may in fact find value in proximity to rail. Given the relative lack of previous studies in Los Angeles, it would be worth expanding on this preliminary research to look at the effects of other transit lines in the metropolitan area. Other Metro Rail lines, as well as other forms of rapid transit such as the commuter train network Metrolink and the county's two Bus Rapid Transit (BRT) lines could be fruitful areas for inquiry. Additionally, as the Metro Rail network continues to expand, it would be interesting to follow the lead of Wang (2010) to see if the *anticipated* opening of a new rail line has any effect on property values before it officially opens.

References

Baranzini, A., Ramirez, J., Schaerer, C., & Thalmann, P. (2008). *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*. New York: Springer. Retrieved from https://search.proquest.com/econlit/docview/56752552/D560378253984198PQ/1

California DMV. (2016). *California DMV Estimated Fee Paid Vehicle Registrations by County*. State of California Department of Motor Vehicles. Retrieved from https://www.dmv.ca.gov/portal/dmv/detail/pubs/media_center/statistics

Can, A. (1990). The Measurement of Neighborhood Dynamics in Urban House Prices. *Economic Geography*, *66*(3), 254–272. https://doi.org/10.2307/143400

Cervero, R., & Duncan, M. (2002). *Land Value Impacts of Rail Transit Services in Los Angeles County*. National Association of Realtors, Urban Land Institute.

Diao, M. (2015). Selectivity, spatial autocorrelation and the valuation of transit accessibility. *Urban Studies*, *52*(1), 159–177. https://doi.org/10.1177/0042098014523686

Gibbons, S., Mourato, S., & Resende, G. M. (2014). The Amenity Value of English Nature: A Hedonic Price Approach. *Environmental and Resource Economics*, *57*(2), 175–196. http://dx.doi.org/10.1007/s10640-013-9664-9

Gillard, Q. (1981). The Effect of Environmental Amenities on House Values: The Example of a View Lot. *The Professional Geographer*, *33*(2), 216–220. https://doi.org/10.1111/j.0033-0124.1981.00216.x

Gordon, P., Richardson, H. W., & Wong, H. L. (1986). The Distribution of Population and Employment in a Polycentric City: The Case of Los Angeles. *Environment and Planning A*, *18*(2), 161–173. https://doi.org/10.1068/a180161

Hess, D. B., & Almeida, T. M. (2007). Impact of Proximity to Light Rail Rapid Transit on Station-Area Property Values in Buffalo, New York. *Urban Studies*, *44*(5–6), 1041–1068.

Hill, R. J. (2013). Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy. *Journal of Economic Surveys*, *27*(5), 879–914.

Kelejian, H. H., & Prucha, I. R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *The Journal of Real Estate Finance and Economics*, *17*(1), 99–121. https://doi.org/10.1023/A:1007707430416

Kotkin, J. (2016, March 14). Mass Transit Expansion Goes Off The Rails In Many U.S. Cities. *Forbes*. Retrieved from https://www.forbes.com/sites/joelkotkin/2016/03/14/mass-transit-expansion-goes-off-the-rails-in-many-u-s-cities/

McIntosh, J., Trubka, R., & Newman, P. (2014). Can Value Capture Work in a Car Dependent City? Willingness to Pay for Transit Access in Perth, Western Australia. *Transportation Research: Part A: Policy and Practice*, *67*, 320–339. http://dx.doi.org.ezproxy.macalester.edu/10.1016/j.tra.2014.07.008

McMillen, D. P., & McDonald, J. (2004). Reaction of House Prices to a New Rapid Transit Line: Chicago's Midway Line, 1983-1999. *Real Estate Economics*, *32*(3), 463–486.

Metro. (2016, November 9). Measure M: Metro's Plan to Transform Transportation in LA. Retrieved November 9, 2017, from http://theplan.metro.net

Nelson, L. J. (2016, March 15). Los Angeles area can claim the worst traffic in America. Again. *Los Angeles Times*. Retrieved from http://www.latimes.com/local/lanow/la-me-ln-la-worst-traffic-20160314-story.html

Rauterkus, S. Y., & Miller, N. G. (2011). Residential Land Values and Walkability. *Journal of Sustainable Real Estate*, *3*(1), 23–43.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55.

Sale, M. C. (2017). The Impact of a Shopping Centre on the Value of Adjacent Residential Properties. *Journal for Studies in Economics and Econometrics*, *41*(1), 55–71.

Sun, W., Zheng, S., & Wang, R. (2015). The Capitalization of Subway Access in Home Value: A Repeat-Rentals Model with Supply Constraints in Beijing. *Transportation Research: Part A: Policy and Practice*, *80*, 104–115. http://dx.doi.org/10.1016/j.tra.2015.07.015

Wang, L. (2010). Impact of Urban Rapid Transit on Residential Property Values: An Empirical Analysis of the Shanghai Housing Market. *Chinese Economy*, *43*(2), 33–52.

Xu, T., Zhang, M., & Aditjandra, P. T. (2016). The Impact of Urban Rail Transit on Commercial Property Value: New Evidence from Wuhan, China. *Transportation Research: Part A: Policy and Practice*, *91*, 223–235.

Yglesias, M. (2012, September 17). L.A.'s Transit Revolution. *Slate*. Retrieved from http://www.slate.com/articles/business/the_pivot/2012/09/l_a_metro_how_los_angeles_is_becoming_america_s_next_great_mass_transit_city_.html

Zheng, S., Xu, Y., Zhang, X., & Wang, R. (2016). Transit Development, Consumer Amenities and Home Values: Evidence from Beijing's Subway Neighborhoods. *Journal of Housing Economics*, *33*, 22–33. http://dx.doi.org/10.1016/j.jhe.2016.05.003

Data Sources

Dickens, M. (2017). *Public Transportation Ridership Report, Fourth Quarter 2016*. American Public Transportation Association. Retrieved from http://www.apta.com/resources/statistics/Pages/ridershipreport.aspx

Hartline, R. (2017a). Assessor Parcels – 2016 Tax Roll. Retrieved December 18, 2017, from https://egis3.lacounty.gov/dataportal/2017/12/12/assessor-parcels-2016-tax-roll/

Hartline, R. (2017b). Assessor Parcels Data - 2006 thru 2017. Retrieved December 18, 2017, from https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2017/9trm-uz8i

Metro. (2010, June 9). Download GIS Data. Retrieved December 18, 2017, from https://developer.metro.net/introduction/gis-data/download-gis-data/

U. S. Census Bureau. (2015). *Median Household Income in the Past 12 Months (in 2015 Inflation-Adjusted Dollars)* (No. S1903). Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_B19013&prodType=table

Appendix

Figure 1. Metro Rail system with the year that each line began service (Metro, 2010)

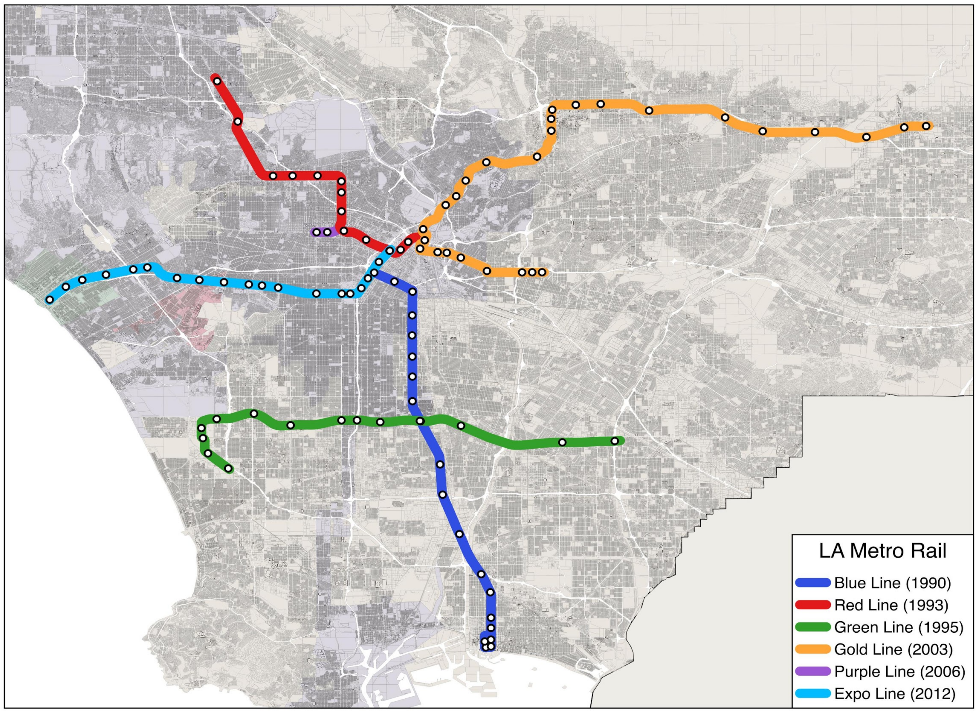


Figure 2. Ten largest rapid transit systems in the US by annual ridership (Dickens, 2017)

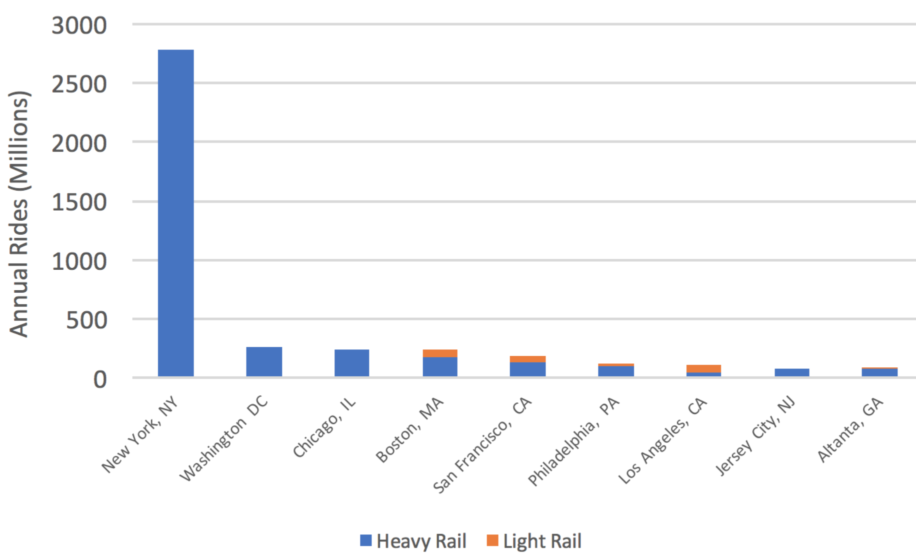


Figure 3. Properties within a one-mile radius of stations along the Expo Line (Hartline, 2017a)

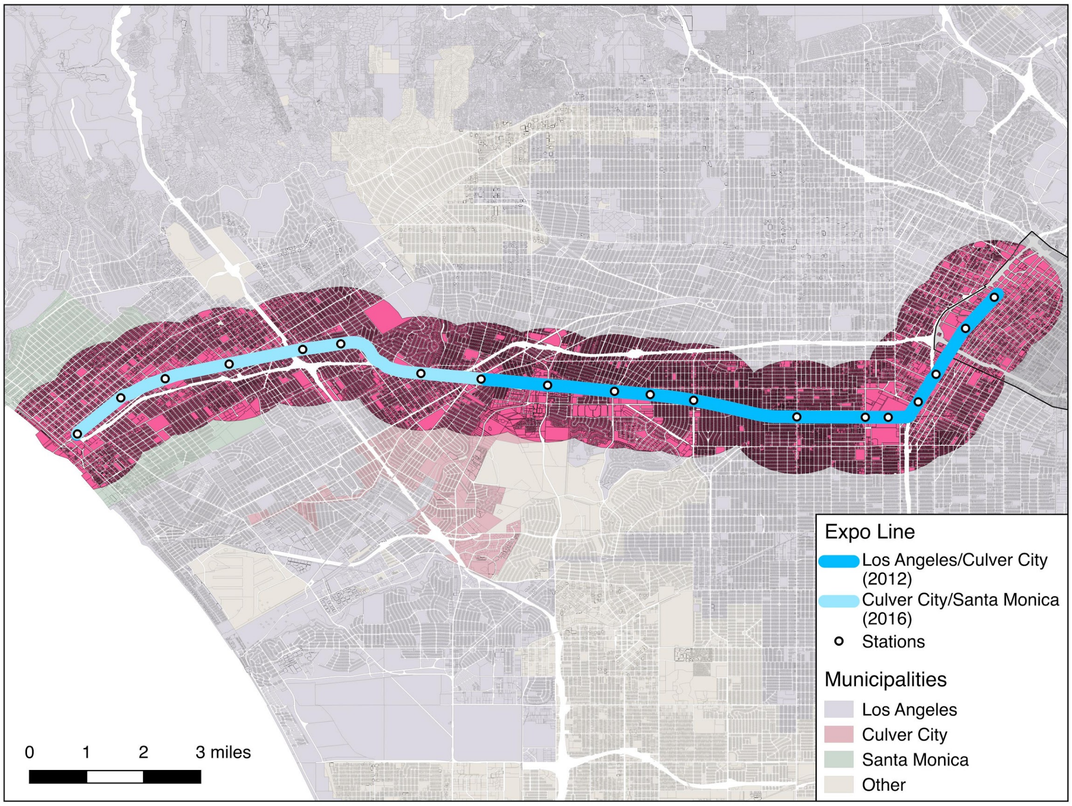


Figure 4. Pairwise correlation coefficients for structural property characteristic variables

	Sq. ft.	Bedrooms	Bathrooms	Units	Effective Year
Sq. ft.	1				
Bedrooms	0.4458*	1			
Bathrooms	0.4663*	0.8892*	1		
Units	0.9286*	0.4739*	0.5306*	1	
Effective Year	0.0313*	0.0818*	0.0769*	0.0253*	1

* = 5% level of significance

Figure 5. Nearest-Neighbor Distances for Expo Line stations and public parks (Hartline, 2017a)

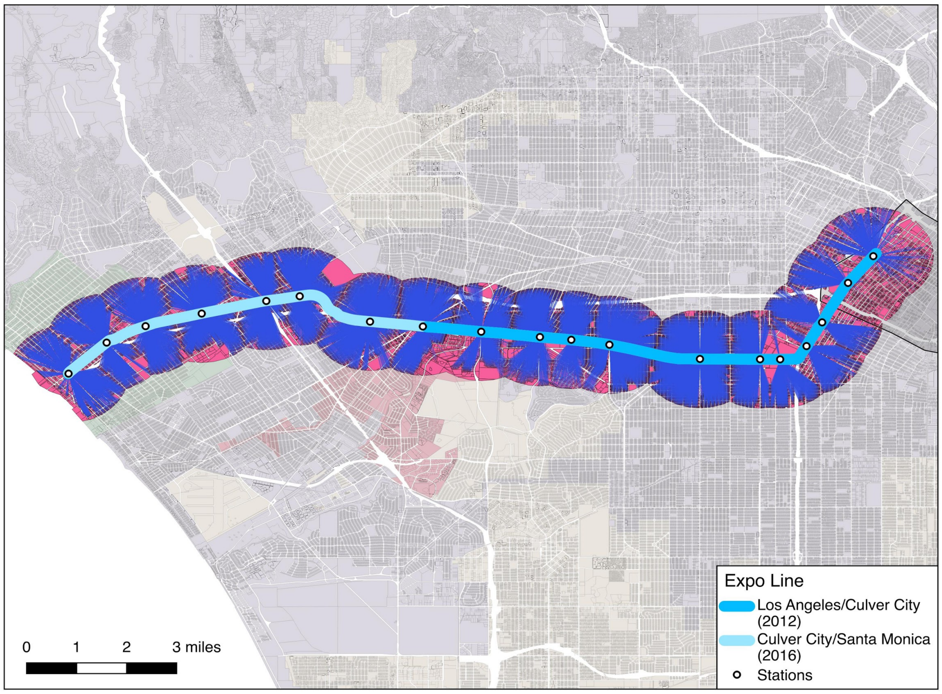


Figure 6. Mean property values in the study area between 2008 and 2016

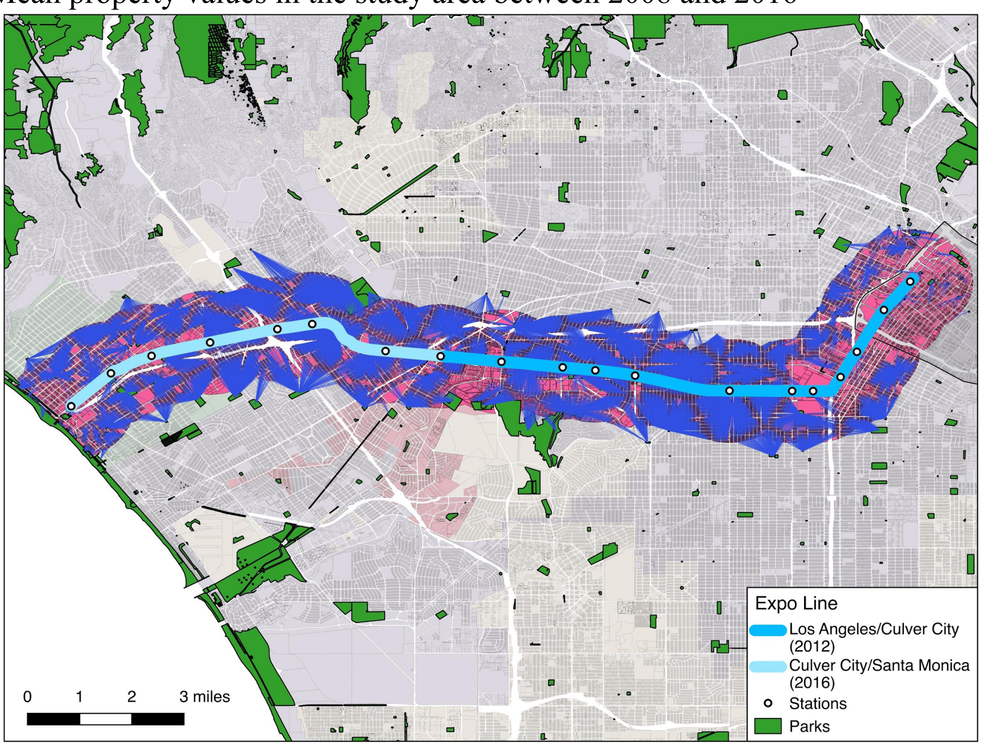


Figure 7. Scatterplots showing property value and logged property value vs. station distance and squared station distance

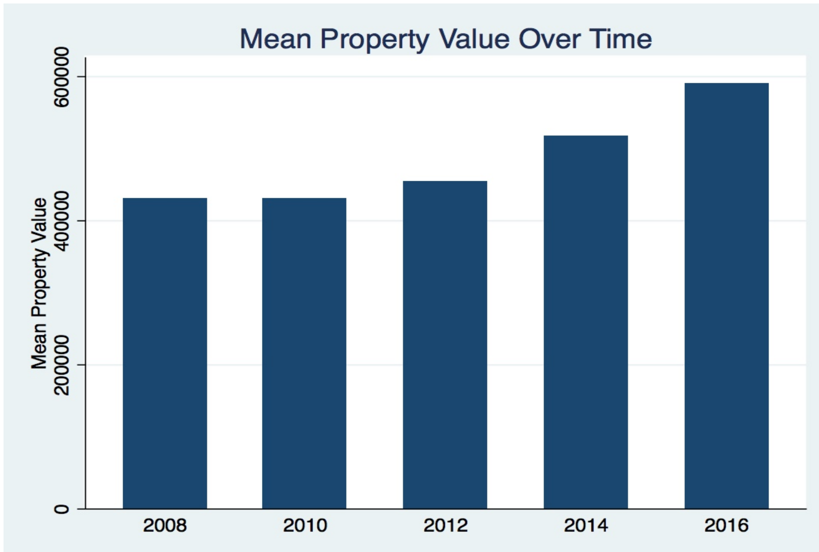


Figure 8. Residuals plots of property values (OLS)

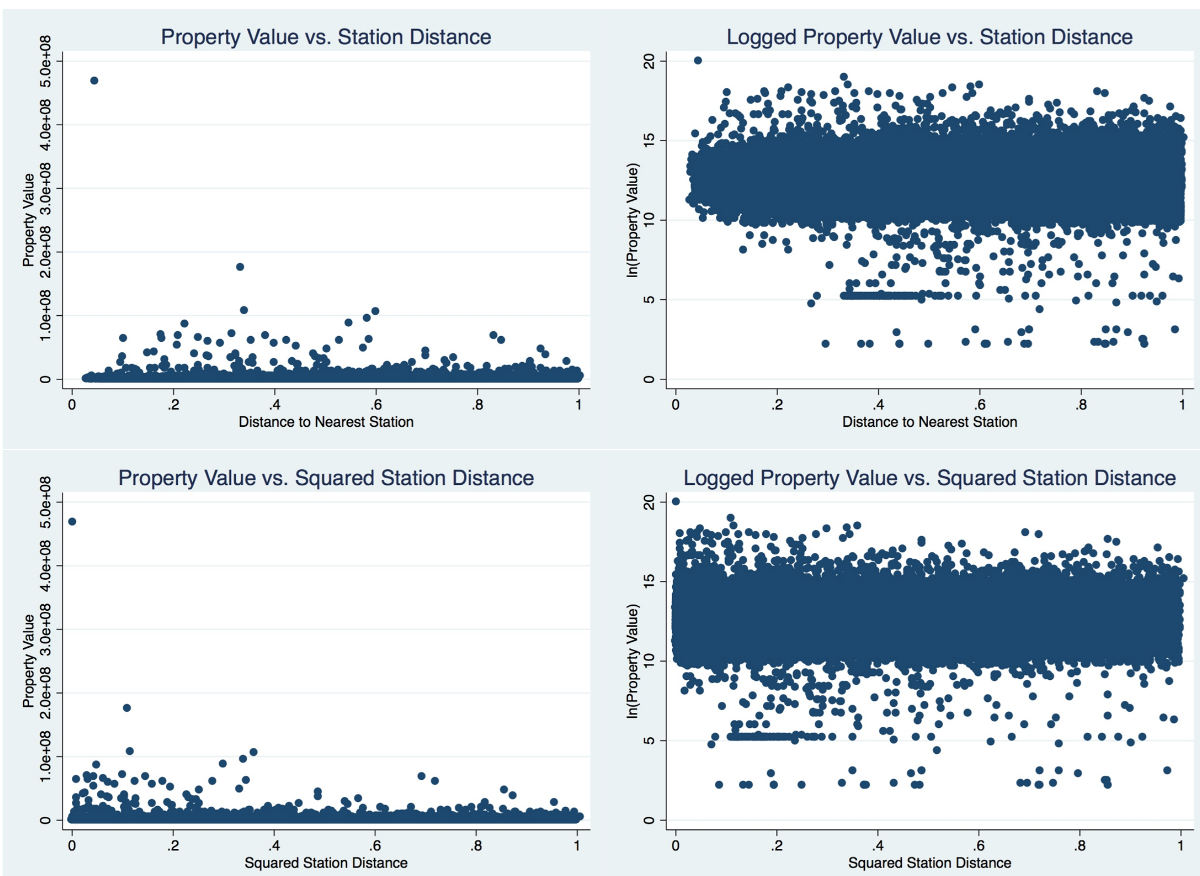
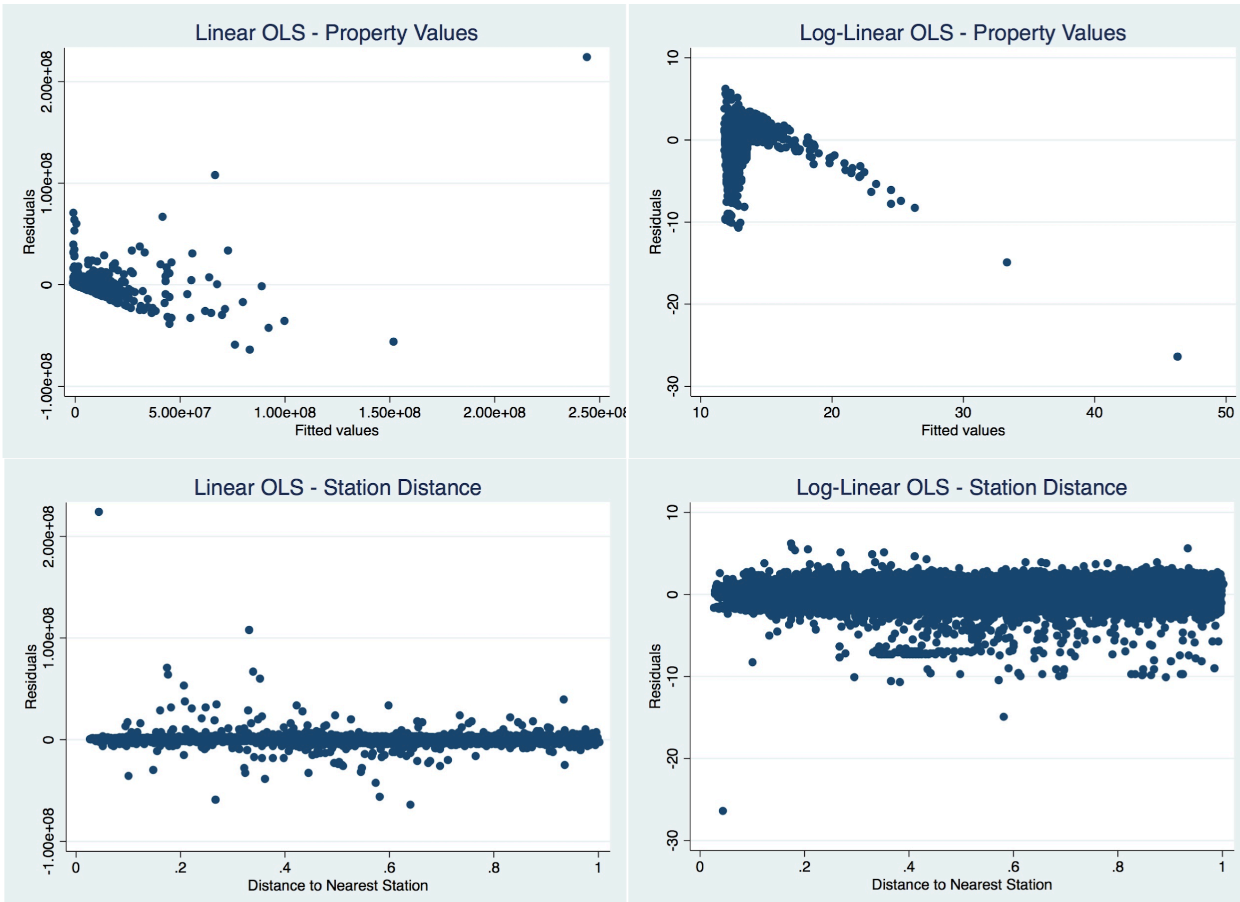


Figure 9. Residuals plots of station distance (OLS)



Blowing in the Wind:

Estimating the effects of Wind Energy on the Midwestern Power Market

Jack McCarthy

Introduction to Econometrics

Abstract:

This paper estimates the effect of wind energy generation on behaviors of wholesale power market in the Midwest. By correlating hourly real time LMP values in MISO with hourly generation from different generation technologies, I create a panel data set for 2013 to 2016 in multiple MISO regions. I then construct a statistical model for power market prices, transmission constraint, and volatility. A difference-in-difference regression with fixed effect finds evidence suggesting that replacing coal with wind generation lead to a reduction in LMP price, an increase in transmission constraint, and a reduction in volatility.

I. Introduction

In the past decade, wind energy has become a mainstream industry and has reliably expanded faster than policymakers and energy utilities have predicted.[1,2] As a result, national electricity markets are changing faster than utilities have planned for. A better understanding of the effects of large scale integration of wind energy generation into power grids will help electric utilities and policy makers design markets that improve consumer surplus.

Renewable energy is distinct from conventional generation due to three key attributes: it has free fuel and thus low marginal cost; it has variable generation which means that operators cannot dictate when wind or solar is generating; and it is decentralized. For this analysis, I choose to analyze wind because it was developed earlier than solar and is commonly used in large scale projects. I focus on the Midwest because it has a significant wind resource and has been the site of large scale wind development. The growth of renewable energy has been claimed to influence three elements of wholesale power market behavior: low marginal cost reduces prices, variable generation increases price volatility, and decentralized generation worsens grid curtailment in the absence of adequate transmission infrastructure.[3]

In this paper, I first review existing studies on the effects of wind on power market prices with a discussion of estimation techniques and economic theory. Then I build a conceptual model for wholesale power market prices. I then overview the best case for estimating the effects of wind. This is followed by a discussion of the actual data and estimation used. Following this I present my results and finally discuss and conclude.

---

[1] http://www.mc-group.com/the-renewable-energy-revolution/
[2] https://www.vox.com/2015/10/12/9510879/iea-underestimate-renewables
[3] Curtailment is an involuntary reduction in generation due to lack of access to transmission.

II. Literature Review

The North American electrical system was built haphazardly over a century of varying policy regimes, technological innovations, and business decisions. Seen as one entity, the electrical grid is the largest machine humanity has ever constructed, spanning much of the surface of the globe (Bakke, 2016). As such, an understanding of market structure is crucial in understanding the functioning of the electric grid. This includes the effects of policy and firm actions. Regulators must ensure reliable electrical service and firms must schedule production. Utilities, independent system operators (ISOs), regional transmission organizations (RTOs), and independent power producers (IPPs) make production decisions based on how markets are constrained and segmented. As renewable generation technologies, with low marginal cost but high variability, are incorporated into existing market structures, the efficiency of markets and the ability of individual firms to exercise market power will be affected. My goal in this paper is to study the impacts of renewable variable power generation on wholesale power market price behavior given transmission constraints. This review of the existing literature begins with a theory section followed by case studies.

a. Theory

Technologies for renewable electricity generation are known to generate significant positive externalities for society such as reducing pollution, reducing greenhouse gas emissions, and improving health. However, externalities for the energy market also exist. An example of positive market externalities is the price merit order effects, whereby the low marginal cost of renewable energy reduces the market clearing price. There are also negative market externalities, such as increased price volatility and the inability to compete in some markets due to the non-dispatchability of renewable energy (Edenhofer et al, 2013; Heal, 2009). As renewable energy capacity increases, the hour-to-hour wholesale market price becomes more volatile, reducing market efficiency and partially offsetting the positive externalities of renewables. This dynamic becomes exacerbated when wholesale power markets are constrained by the lack of sufficient grid interconnections, which prevents arbitrage between power markets. Constraints on trade between people or geographical regions, such as the limited capability of a transmission line between two regions, create sequential markets with heterogeneous market behavior between market segments. Constraints on trade between these sequential markets (between ISOs in the US, for example), reduce the allocative efficiency of commodities, preventing prices from converging, and increase firms' ability to exercise market power (Ito & Reguant, 2016). As a result of variable generation and constraints to arbitrage between markets, market nodes exhibit price variation even when controlling for variations in demand and supply (Bushnell et al., 2008).

b. Case Studies

Existing empirical studies on the constraint-driven effects of renewable generation on market prices largely corroborate theoretical models. Often the methodological challenge is separating what proportion of price variation is caused by constraints, and what proportion is caused by firms exercising market power. In the case of the San Onofre Nuclear Generating station (SONGs) closure in California and market undersupply in India, transmission constraints and monopolistic behavior increase market prices. In California, a regression-discontinuity comparison of generation merit-order before and after closure shows that lost generation from closing the southern Californian nuclear plant was mostly met by increasing generation at high cost southern Californian plants, not the lower marginal cost plants in northern California (Davis & Hausman, 2014). Merit-order is the process of prioritizing generation with the lowest marginal cost. Volatility or transmission constraints may prevent generators from producing electricity in the lowest cost order, thus reducing market efficiency. In India, despite high market demand, power plants are underutilized due to insufficient transmission, which allowed producers to set prices and reduce output. This is verified by empirically estimating a counterfactual for generation outcomes to estimate how competitive firms would behave if transmission infrastructure was able to create geographically integrated power markets (Ryan, 2017). Another empirical study of transmission constraints in Alberta finds a significant positive correlation between transmission expansion and market competition. As transmission congestion was alleviated, firms submitted offer curves increasingly closer to their marginal cost curves (Wolak, 2015).

Other studies estimate the impacts of wind and solar generation on electricity prices. These studies find natural experiments in regions with higher levels of renewable generation, including Germany, Texas, and Great Britain. In Germany, where renewables make up 20% of total generation, a merit-order effect analysis and a GARCH process on share of production from wind and seasonally adjusted day-ahead prices find that a 1% increase in the share of generation from wind caused a 1.3% reduction in prices and increased volatility (Ketterer, 2014; Würzburg et al., 2013). A study of high-resolution generation data in Texas finds a similar effect of decreased price and increased volatility (Woo et al., 2011). The same effects are found in a similar study of Great Britain, with the additional finding that non-renewable generators benefited from a significant increase in market power (Green & Vasilakos, 2010).

III. Conceptual Model

The ideal model attempts to estimate the effect of wind generation on price, volatility, and constraint. Using generalized concepts of power market behavior from Würzburg et al. (2013) yields the following conceptual model:

$$P_{it} = f(Generation\ Capacity_{it}, Energy\ Mix_{it}, Fuel\ Price_{it}, Load\ Demand_{it}, Location_i, Time_t)$$

$$V_{it} = f(Generation\ Capacity_{it}, Energy\ Mix_{it}, Fuel\ Price_{it}, Load\ Demand_{it}, Location_i, Time_t)$$

$$C_{it} = f(Generation\ Capacity_{it}, Energy\ Mix_{it}, Fuel\ Price_{it}, Load\ Demand_{it}, Location_i, Time_t)$$

Where P_{it} is electricity price, V_{it} is price volatility, and C_{it} is power grid constraint at location *i* at time *t* . To frame these ideal variables in terms of generalized economic theory: generation capacity represents supply, energy mix is the proportion of supply using different production functions, fuel prices are the marginal costs of each production function, load demand represents market demand. P_{it} , V_{it} , C_{it} are also affected by other location-specific and time-specific variables.

IV. Data

My actual data are taken from the Midcontinent Independent System Operator (MISO), which oversees power grid operations in the Midwest. While publicly available at hourly intervals for many price nodes, generation mix data are only available beginning in 2013. Therefore, I narrow my time frame to 2013 - 2016 in MISO's Northern and Southern planning regions. I choose MISO's real-time locational marginal price (LMP) as my dependent variable, as it is a proxy for electricity price that is available hourly at every MISO price node (figure 5)

(MISO, 2017a). For MISO price nodes, I select the Illinois, Indiana, Michigan, and Minnesota hubs, which are the primary price locations in MISO. These hubs do not represent individual price points, but aggregations of regional price nodes. As such they are better representatives of market-wide prices in those regions. Real time LMPs are the total of three price components: marginal energy component (MEC), which reflects generation cost and are homogenous across MISO at a given time; marginal congestion component (MCC), which is a measure of transmission constraints at the location; and marginal loss component (MLC), which reflects distance and hence the loss between generation and load.

In addition to LMP, I select MCC as a proxy for transmission constraint. MCC is negative when additional generation at that point adds constraint to the system and positive when it reduces constraint. To create a proxy for price volatility, I calculate the standard deviation of LMP at each node within a day. A graph of average price data by node can be seen in figure 7.

As a proxy for energy mix and total load I select MISO's generation mix dataset, which has data on megawatts generated from wind, nuclear, hydro, natural gas, coal, and 'other' for each hour by region (MISO, 2017b). To connect the generation and price variables, I associated the MISO price nodes with the MISO region within which they are located. Minn.hub is in the northern region while Illinois, Indiana, and Michigan are in the central region (figure 6). I calculate the percent of generation supplied by each fuel instead of megawatts in order to compare the effect of generation technologies on price across time and reduce collinearity between fuel type and total generation. A graph of average generation data by region can be seen in figure 8. Table 1 provides a summary of variables.

Table 1. Summary Statistics

Variable	Mean	Std. Deviation	Min	Max
LMP ($/MWh)	29.23	20.69	-165.53	990.08
Daily SD LMP	11.95	13.14	0.73	196.61
MCC ($/MWh)	-0.25	10.97	-233.42	269.13
% Wind	7.13	11.72	0	67.74
% Nuclear	12.77	3.69	0	29.72
% Hydro	1.37	0.84	0.11	6.45
% Gas	9.78	7.13	0	34.11
% Coal	68.19	13.08	16.8	97.58
Hourly Generation (MW)	28305	10669	7944	60176

V. Model Estimation

Given the collected data, I create a model that includes generation mix from different energy sources, supply, demand, location, and time to estimate three dependent variables which together represent market behavior. My estimation equation is:

$$LMP_{it} = \beta_0 + \beta_1 Wind_{it} + \beta_2 WindPeak_{it} + \sum \beta_3 Nonwind_{it} + \beta_5 TotalGen_{it} + \sum \beta_i Location + \sum \beta_t Datetime + \varepsilon_{it}$$

$$\delta LMP_{id} = \gamma_0 + \gamma_1 Wind_{it} + \gamma_2 WindPeak_{it} + \sum \gamma_3 Nonwind_{it} + \gamma_5 TotalGen_{it} + \sum \gamma_i Location + \sum \gamma_t Datetime + \varepsilon_{it}$$

$$MCC_{it} = \eta_0 + \eta_1 Wind_{it} + \eta_2 WindPeak_{it} + \sum \eta_3 Nonwind_{it} + \eta TotalGen_{it} + \sum \eta_i Location + \sum \eta_t Datetime + \varepsilon_{it}$$

LMP_{it} = Locational Marginal Price of electricity ($/MWh) at location i, at date-time *t*

δLMP_{it}= Standard deviation of LMP ($/MWh) at location *i* on day *d*

MCC_{it}= Marginal Congestion Charge ($/MWh) at location *i*, at date-time *t*

The three specifications are identical except for the dependent variables. *i* represents the four price locations across the north and central MISO planning regions, while *t* represents each hourly time interval.

In modelling for price, volatility, and constraint, I expect the coefficients for total generation β_5, γ_5, η_5, to be positive, positive, and negative respectively because total generation is a proxy for the quantity of electricity supplied and demanded at any given time. The nature of the power grid means that in the absence of system failure there cannot be a significant surplus or shortage of electricity at a given time and generation is prioritized by lowest marginal cost, thus prices should rise as higher cost capacity is brought online. I expect price volatility to positively correlate with generation because when prices and generation rise to high levels, by mathematical definition a larger increase in price increases price volatility. I expect higher total generation to approach the maximum capacity constraints of transmission infrastructure, thus reducing marginal congestion cost. However, because MCC also increases downstream of constraints, incentivizing generation that relieves grid congestion, the overall effect of generation may be ambiguous.

When modelling for price, volatility, and constraint, I expect signs of the coefficients on share of generation from wind, β_1, γ_1, η_1, and non-wind, β_3, γ_3, η_3, to have the same signs because all generation sources on average have prices greater than zero and an effect on volatility and constraint. Because the generation mix variables sum to 1, effects on price, volatility, and constraint from a generation source will depend on the coefficient's size relative to another generation source it is replacing. Therefore, I hypothesize in order from most to least price reducing according to levelized cost of electricity estimates from Lazard: wind, hydro, gas, coal, nuclear. In order from most to least volatility increasing I hypothesize by technology frequency of on and off cycling: wind, gas, coal, hydro, nuclear. Finally, in order of most to least transmission constraining I hypothesize wind will be most transmission constraining and the others are uncertain due to lack of detailed transmission information.

WindPeak is an interaction term of percent generation from wind and a dummy that is 1 at peak demand hours which MISO defines as weekdays 600 - 2200 EST. I anticipate WindPeak to have a positive coefficient for price because during high demand events, generators experience higher prices, a positive coefficient for volatility at peak hours, and a negative coefficient for constraint because higher demand at peak hours causes transmission to be more easily congested given a spike in wind generation.

Given my data is available 2013 - 2016 at four locations I used a panel data regression with fixed effects to test the data. In order to determine whether to use fixed or random effects in my panel data regression I conducted a Hausman test and received a χ^2 value of 2414.75 with a P value of 0, thus I reject the null hypothesis and use fixed effects. The use of fixed effects still will not account for the differential effects between locational markets, for that reason I include a factor variable that represents the different market behaviors at the market price nodes in Illinois, Indiana, Michigan, and Minnesota.

VI. Results

The regression output from the three model specifications is presented in table 2. I find coefficients of all variables to be statistically significant except hydro, gas, and coal. The r-squared values are low for all three models, suggesting that the models do not explain most of the variance in price, volatility, and grid constraint. It is noteworthy that while the within and between r-squared are similar for LMP and SD_LMP, MCC's within r-squared is about ten times greater than its between r-squared. This means that the MCC model is much better at explaining the variance in constraint at a given node across time than it does the variance between nodes. This is to be expected as the model lacks a variable for transmission capacity, which may be a primary difference influencing how different MISO planning regions behave.

When analyzing the coefficients, I find that they only partially correspond with theorized signs. Total generation coefficients are statistically significant and positive in all three models, which is consistent with theory. Generation mix coefficients are inconsistent with theory and only partially statistically significant. The observed order of generation mix from most to least price reducing in the LMP is: nuclear, wind, coal, gas (insignificant), and hydro (insignificant). This is not consistent with the theorized order. Nuclear may vary from its levelized cost of energy due to high and unaccounted for subsidy levels. Importantly, it may still be concluded that wind generation decreases prices when it is replacing coal generation. Regarding volatility, this regression concludes that volatility decreases if wind replaces gas or coal despite wind being theorized as the most volatile. Daily standard deviation of LMP may not be a good measure for volatility. Alternatively, there might exists multicollinearity between wind and total generation as wind is a higher share of generation when demand is low, which in turn is correlated with low prices and lower volatility. In the MCC model, nuclear is predicted to constrain the grid more than wind, while hydro, gas and coal are statistically insignificant. This also contradicts the theorized signs. However, wind is still statistically significant and has the correct sign relative to the majority of other generation technologies, indicating that wind does induce transmission constraints. WindPeak does not correspond with the theorized signs. The coefficients for price and volatility are positive, significant, and follow theory. However, the coefficient for transmission constraint is positive, meaning that on peak hours, wind generation reduces constraint more than at other times.

Although most coefficients are statistically significant, the model's results might indicate that the specification or the data are insufficient to accurately represent the effect of wind power on metrics of market behavior. Residuals analysis provides a clearer picture of how the three models behave. Residual versus predictor plots (figures 1, 2, 3) indicate that for all three models, residuals are greater when there is less or no generation from wind. At the same time, a histogram of wind observations shows that most observations have low wind generation. Therefore, there is likely heteroscedasticity in the model (figure 4).

Table 2. Regression Output with Price, Volatility, and Congestion as Dependent Variables

VARIABLES	(1) LMP	(2) SD_DAILY_LMP	(3) MCC
% Wind	-0.574***	-0.326***	-0.441***
	(-6.734)	(-7.748)	(-5.696)
% Nuclear	-0.577***	-0.655***	-0.446***
	(-6.550)	(-15.15)	(-5.558)
% Hydro	-0.0224	-0.297***	-0.0271
	(-0.142)	(-3.814)	(-0.185)
% Gas	-0.114	-0.213***	-0.0450
	(-1.214)	(-4.683)	(-0.528)
% Coal	-0.173**	-0.269***	-0.0470
	(-2.042)	(-6.403)	(-0.613)
Total Gen	0.000432***	0.000207***	0.000405***
	(12.77)	(10.77)	(13.12)
%Wind * Peakhour	0.0446***	0.0142***	0.0492***
	(5.262)	(3.091)	(6.111)
Illinois.hub	0	0	0
Indiana.hub	2.261***	0.182***	1.186***
	(25.72)	(4.582)	(13.69)
Michigan.hub	3.283***	0.983***	1.613***
	(30.94)	(18.02)	(15.59)
Minn.hub	12.81***	4.104***	12.58***
	(15.35)	(8.963)	(16.51)
Constant	37.58***	36.50***	-2.705
	(4.422)	(8.830)	(-0.352)
Observations	100,692	100,664	100,692
R-squared	0.111	0.038	0.073
Number of datetime_sif	27,249	27,242	27,249
F	885.1	165.5	521.0
r2_overall	0.134	0.0437	0.00721
r2_between	0.141	0.0397	0.00686
r2_within	0.111	0.0379	0.0727
sigma_u	17.65	12.25	7.570
sigma_e	10.97	5.981	10.65
rho	0.722	0.808	0.336

Robust t-statistics in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Costs of different generation technologies vary significantly with context. The location of a generation project may change its bidding behavior such that merit order varies from a generalized hierarchy of technologies. Even given that, individual plants may bid out of merit order given market power. In order to account for this plant level behavior, a model of power plant level generation would be required. Future studies could be improved by accounting for the amount of energy contracted through power purchase agreements (PPAs) that count towards generation statistics but do not influence spot-market prices.

When interpreting the coefficients for generation mix, it is important to remember that it is not the value of *% Wind* alone that matters but the difference between that and the other coefficients. If wind generation increases by one percentage point, the total of all other generation sources must decrease by one percent. If wind offsets coal, the models tell us to expect LMP to decrease by $0.401/MWh, SD_LMP to decrease by $0.057/MWh, and MCC to decrease by $0.394/MWh.

VII. Conclusion

a. Summary

This paper develops a model for predicting price, volatility, and transmission constraints in MISO's real time locational marginal price market to examine the effect of wind power on these market behaviors. Fixed effects regressions conclude that between 2013 and 2016 in the Midwest, wind generation reduced price and increased transmission constraints when replacing coal generation. However, wind is predicted to reduce volatility, which is inconsistent with existing empirical findings.

b. Limitations

This analysis has been limited by availability of data as well as model complexity. Being constrained to publically available data, I only use market level data for a four-year period. With data on individual power plant costs and generation, a more dynamic model of plant level merit-order generation could be estimated so as to determine what form of generation is displacing what and at what price. Additionally, an analysis similar to the study by Davis & Hausman (2014) could determine which plants are exercising market power and generating out of merit-order. Another weakness of the data emerges from associating price nodes with planning regions. It is possible that generation in the northern region also affects prices in the central region and vice versa, this leakage weakens the analysis results. Furthermore, if data extend back before the current growth of wind a clearer control-and-treatment time period might be defined for time series analysis.

c. Future Research

As wind energy continues to expand it may eventually become one of the primary sources for generating electricity the U.S. At the point of market saturation, wind may have a very different effect on market behavior than when it is a small portion of generation. Denmark, for example, has at times generated 100% of its electricity from wind.[4] Future research might create projections of how wind's growth will alter power markets in order to better plan for these changes. In addition, improved models might incorporate power purchase contracts, day ahead contracts, and energy trading that influence market behavior exogenous of physical generation. Plant level data would allow modelling of a dynamic generation schedule to better predict prices

[4] https://www.theguardian.com/environment/2015/jul/10/denmark-wind-windfarm-power-exceed-electricity-demand

once a certain load threshold is passed. Additionally, external shocks to the market such as extreme weather events, national economic performance, and policy changes, all influence the power market. Thus, accounting for them may improve future studies.

References

Bakke, G. A. (2016). *The grid: the fraying wires between Americans and our energy future* (Paperback edition). New York, NY: Bloomsbury.

Borenstein, S., Bushnell, J. B., & Wolak, F. A. (2002). Measuring Market Inefficiencies in California's Restructured Wholesale Electricity Market. *The American Economic Review*, *92*(5), 1376–1405.

Borenstein, S., & Holland, S. (2005). On the Efficiency of Competitive Electricity Markets with Time-Invariant Retail Prices. *The RAND Journal of Economics*, *36*(3), 469–493.

Bushnell, J. B., Mansur, E. T., & Saravia, C. (2008). Vertical Arrangements, Market Structure, and Competition: An Analysis of Restructured US Electricity Markets. *The American Economic Review*, *98*(1), 237–266.

Cicala, S. (2017). *Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation* (Working Paper No. 23053). National Bureau of Economic Research.

Davis, L., & Hausman, C. (2014). *The Value of Transmission in Electricity Markets: Evidence from a Nuclear Power Plant Closure* (Working Paper No. 20186). National Bureau of Economic Research.

Edenhofer, O., Hirth, L., Knopf, B., Pahle, M., Schloemer, S., Schmid, E., & Ueckerdt, F. (2013). On the economics of renewable energy sources. *Energy Economics*, *40*, S12–S23.

Fabrizio, K. R., Rose, N. L., & Wolfram, C. D. (2007). Do Markets Reduce Costs? Assessing the Impact of Regulatory Restructuring on US Electric Generation Efficiency. *The American Economic Review*, *97*(4), 1250–1277.

Green, R., & Vasilakos, N. (2010). Market behaviour with large amounts of intermittent generation. *Energy Policy*, *38*(7), 3211–3220.

Heal, G. (2009). *The Economics of Renewable Energy* (Working Paper No. 15081). National Bureau of Economic Research.

Ito, K., & Reguant, M. (2016). Sequential Markets, Market Power, and Arbitrage. *American Economic Review*, *106*(7), 1921–1957.

Ketterer, J. C. (2014). The impact of wind power generation on the electricity price in Germany. *Energy Economics*, *44*, 270–280.

Knittel, C. R., & Roberts, M. R. (2005). An empirical examination of restructured electricity prices. *Energy Economics*, *27*(5), 791–817.

Milstein, I., & Tishler, A. (2011). Intermittently renewable energy, optimal capacity mix and prices in a deregulated electricity market. *Energy Policy*, *39*(7), 3922–3927.

MISO. (2017a). Historical Generation Mix. Midcontinent Independent System Operator. Available at: www.misoenergy.org/RTDisplays/MKTRPT/AllReportsList.html?rptName=Historical%20Generation%20Fuel%20Mix%20(xls)

MISO. (2017b). Historical LMP. Midcontinent Independent System Operator. Available at: https://www.misoenergy.org/RTDisplays/MKTRPT/AllReportsList.html?rptName=Historical%20Annual%20Real-Time%20LMPs%20(zip)

Ryan, N. (2017). *The Competitive Effects of Transmission Infrastructure in the Indian Electricity Market* (Working Paper No. 23106). National Bureau of Economic Research.

Sickles, R. (2008). energy economics. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed., pp. 842–848). Basingstoke: Nature Publishing Group.

Traber, T., & Kemfert, C. (2011). Gone with the wind? - Electricity market prices and incentives to invest in thermal power plants under increasing wind energy supply. *Energy Economics*, *33*(2), 249–256.

Wolak, F. A. (2014). Regulating Competition in Wholesale Electricity Supply. *NBER*, 195–289.

Wolak, F. A. (2015). Measuring the competitiveness benefits of a transmission investment policy: The case of the Alberta electricity market. *Energy Policy*, *85*, 426–444.

Wolak, F. A. (2016). *Level versus Variability Trade-offs in Wind and Solar Generation Investments: The Case of California* (Working Paper No. 22494). National Bureau of Economic Research.

Woo, C. K., Horowitz, I., Moore, J., & Pacheco, A. (2011). The impact of wind generation on the electricity spot-market price level and variance: The Texas experience. *Energy Policy*, *39*(7), 3939–3944.

Würzburg, K., Labandeira, X., & Linares, P. (2013). Renewable generation and electricity prices: Taking stock and new evidence for Germany and Austria. *Energy Economics*, *40*(Supplement 1), S159–S171.

Appendix

Figure 1. Residual versus Predictor Plot (Model 1, LMP)

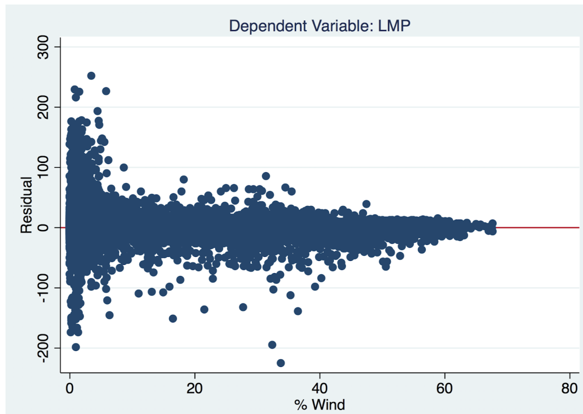


Figure 2. Residual versus Predictor Plot (Model 2, SD_Daily LMP)

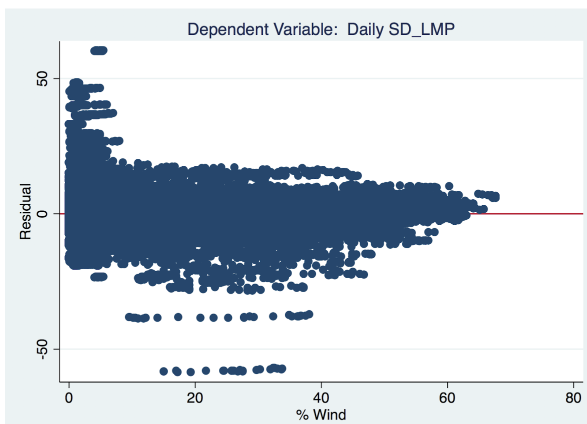


Figure 3. Residual versus Predictor Plot (Model 3, MCC)

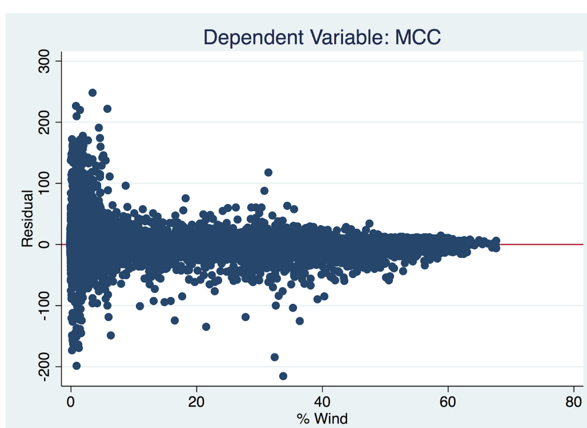


Figure 4. Normal distribution of %Wind observations

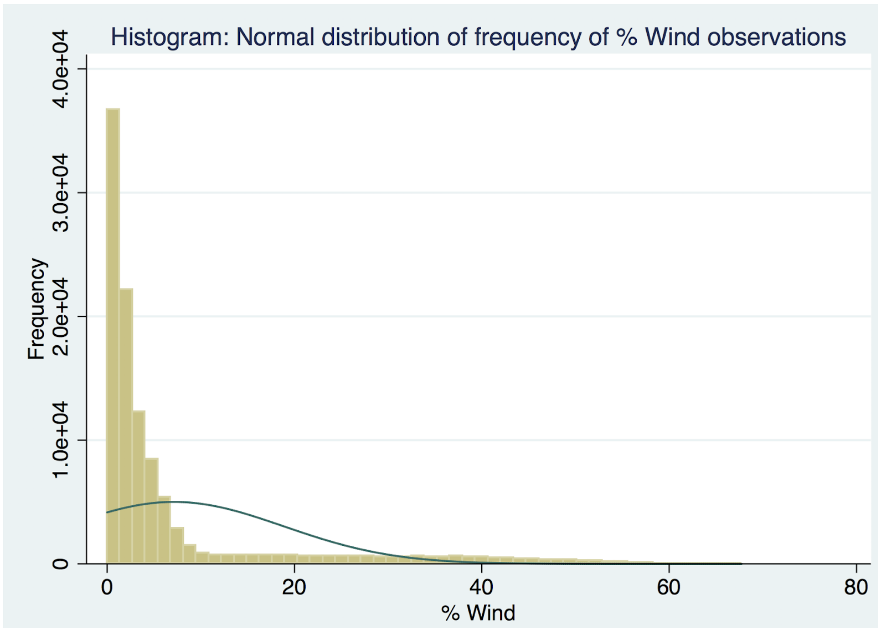


Figure 5. MISO Price Nodes Map

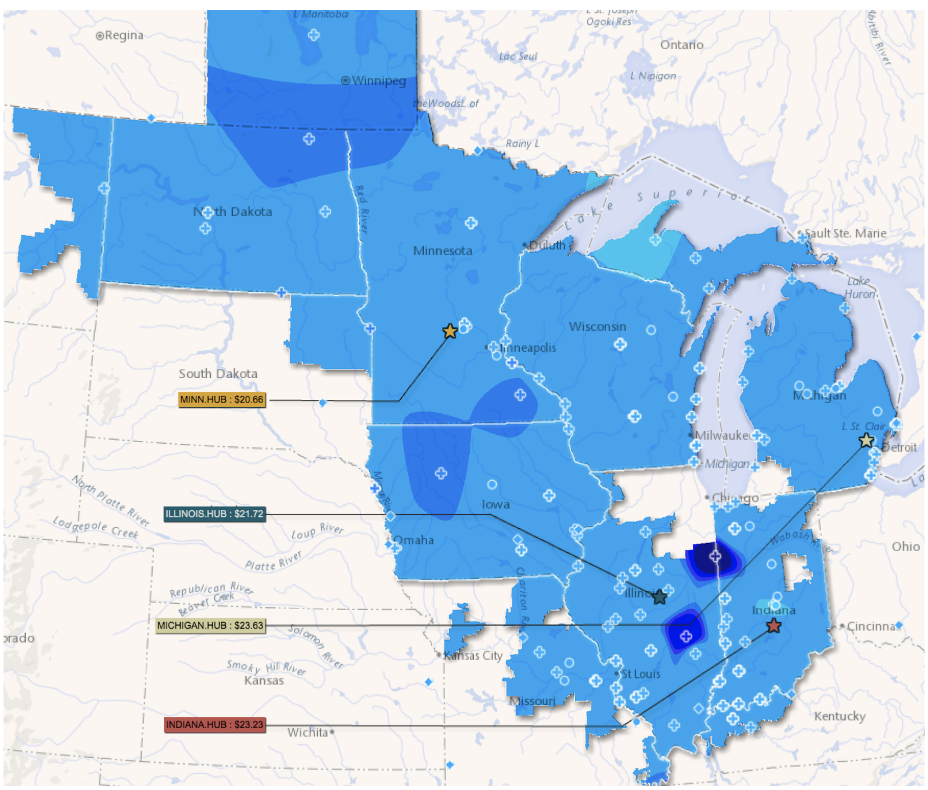


Figure 6. MISO Operations Regions Map

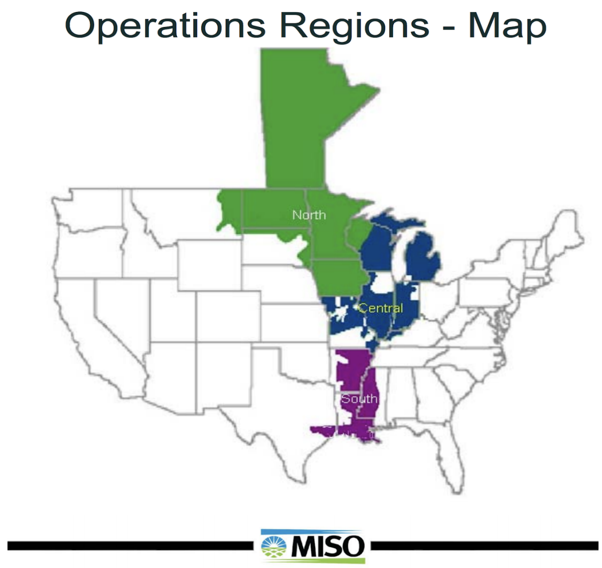


Figure 7. Average Market Behavior Metric by Price Node

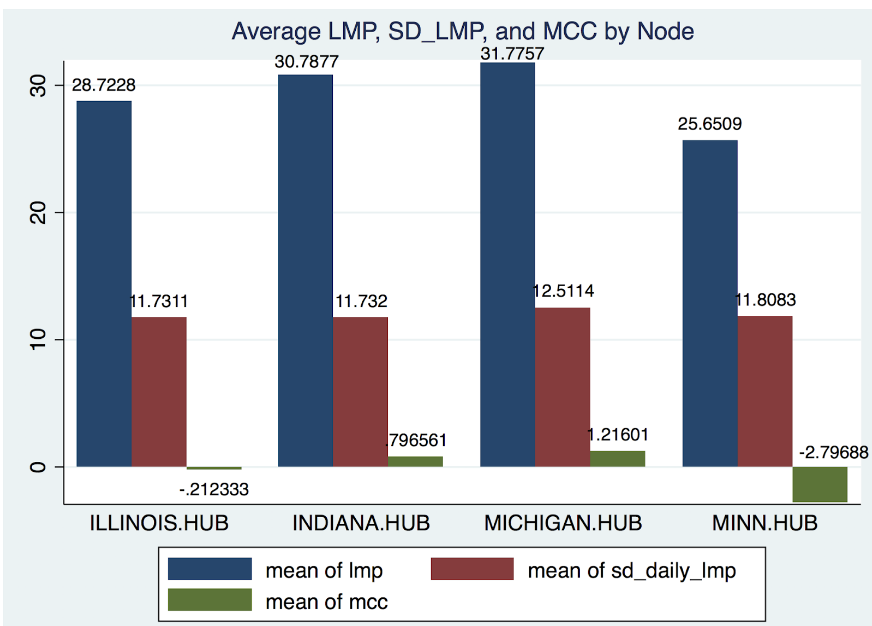
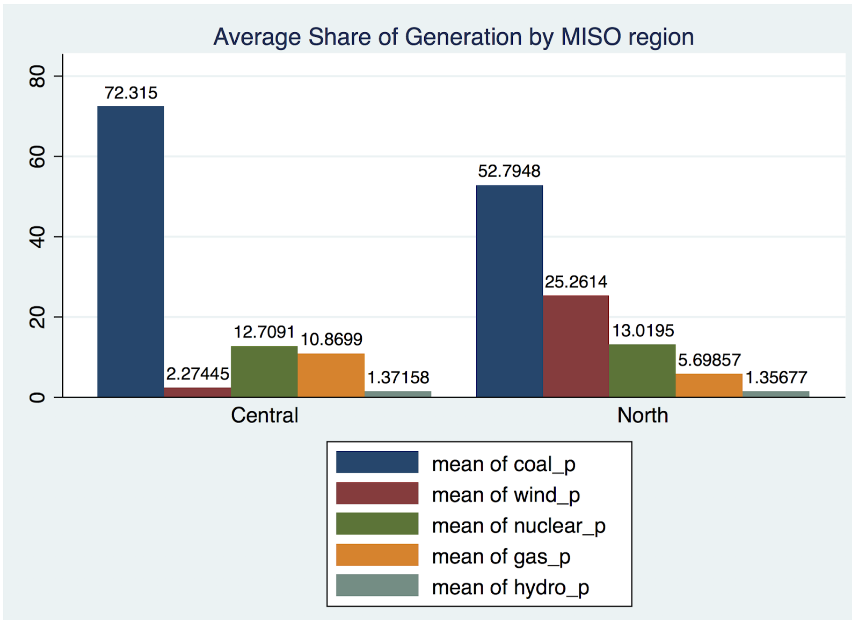


Figure 8. Average Generation Mix by MISO Region



Output Convergence in the United States: Evidence from State- Level Data

Chan Wang

Introduction to Econometrics

Abstract

The classical Solow model of economics development posits a negative relationship between GDP per capita and economics growth rate and therefore predicts output convergence. In this paper, I use Barro and Sala-i-Martin's (1990) specification to estimate the output convergence rate among 50 U.S states by 10 year subperiods from 1930 to 2015. The convergence rate significantly declines after 1980s but improves when I control for regional shocks, sectoral shocks and education attainment. I conclude that while the Solow model is still applicable to the current U.S economy, including these additional controls helps improve its significance and robustness.

I. Introduction

In his seminal paper, Solow (1956) ushered in a new field in neoclassical economics, income convergence. The principles behind his model are strikingly simple: steady state output is determined by capital per labor and exogenous factors – the savings rate, the capital depreciation rate, and the population growth rate. Additionally, capital per labor exhibits diminishing marginal returns. His model predicts that capital will flow from capital-intensive to capital-scarce economies, inducing output convergence in the long run. Research by Barro and Sala-i-Martin (1990) as well as Mankiw (1992) has empirically tested the convergence hypothesis. Barro and Sala-i-Martin (1990) studied convergence among US states. They included a regional dummy and a sector composition variable in their specifications to control for regional and sectoral shocks. Mankiw (1992) introduced human capital per labor into the Solow model, thereby improving its statistical fit and reducing the implausibly high capital share of income it implied.

This paper examines output convergence among the US states using Barro and Sala-i-Martin's (1990) specification for β-convergence. It contributes to the existing literature by testing the effectiveness of β-convergence models with data up to 2015, and by suggesting additional controls for Barro and Sala-i-Martin's (1990) specification. This paper is organized as the following. In section II, I discuss the classical Solow model, Barro and Sala-i-Martin's (1990) model, and past empirical research. In section III, I introduce the conceptual model. In section IV, I discuss the ideal data and measurement of concepts in the model. Section V describes my data and estimation equation. In section VI, I discuss my results. Section VII concludes.

II. Literature Review

a. Classical Solow Model

The Solow model begins by assuming a Cobb-Douglas production function with constant returns to scale and diminishing marginal returns to capital:

$$Y = F(K, L) = K^{\alpha} L^{1-\alpha} \text{ , where } 0 < \alpha < 1 \tag{1}$$

Since output per labor is $y = \frac{Y}{L}$, and capital per labor is $k = \frac{K}{L}$, dividing both side of (1) by L, we obtain:

$$y = k^{\alpha}, \text{ where } 0 < \alpha < 1 \tag{2}$$

The second assumption of Solow model is: income (Y) is saved and reinvested at the savings rate (s) and capital stock depreciates at rate δ. Both rates are exogenously given. Therefore, the change in capital (K') is given by:

$$K' = sY - \delta K \tag{3}$$

Solow then assumes that labor (L) is the entire population and grows at a fixed rate n:

$$L_n = L_0 e^{nt} \tag{4}$$

Taking log on both sides of the capital per labor equation $k = K/L$, we obtain:

$$\log k = logK - logL \tag{5}$$

Taking the derivative on both sides of (5) and substituting L for (4), we obtain:

$$\frac{k'}{k} = \frac{K'}{K} - \frac{L'}{L} = \frac{K'}{K} - n \tag{6}$$

Substituting K' for (3), we obtain:

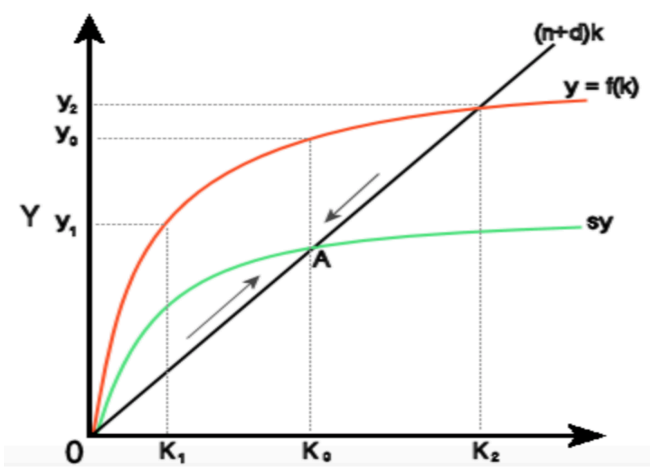
$$\frac{k'}{k} = \frac{sY - \delta K}{K} - n \tag{7}$$

Rearranging (7), we obtain:

$$k' = sy - (\delta + n)k \tag{8}$$

Therefore, the change in capital per labor (k') is a function of the savings rate (s), capital depreciation rate (δ) and population growth rate (n). If $k' = 0$, the inflow of new capital (sy) equals the outflow of capital $(\delta + n)k$. We arrive at the steady state capital per labor k_0 and output per labor y_0 according to Figure 1. If an economy starts below the steady state at k_1 and y_1 , $k' = sy - (\delta + n)k > 0$, capital and output per labor grow to approach k_0 and y_0. If an economy starts above the equilibrium at k_2 and y_2 , $k' = sy - (\delta + n)k < 0$, capital and output per labor decline to approach k_0 and y_0. Therefore, equation (8) predicts that holding all else constant (*s, δ* and *n*), poorer economies grow faster than richer economies and income converges.

Figure 1. Classical Solow Growth Model and the Steady State Output (y_0)



b. Barro and Sala-i-Martin's (1990) Specification

To model the inverse relationship between initial output and growth rate implied by the classical Solow model, Barro and Sala-i-Martin (1990) derived the following equation:

$$\frac{1}{T}\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right) = a - \frac{1-e^{-\beta t}}{T}\log(y_{i,t}) + (other\ variables) + u_{i,t,t+T} \qquad (9)$$

where $y_{i,t}$ is the output per labor of state *i* at time *t;* $y_{i,t+T}$ is the output per labor of state *i* at time *t+T*; and $u_{i,t,t+T}$ is the error term between time *t* and *t+T*. Therefore, *T* is the period that measures the convergence.

On the left side of equation (9), $\frac{1}{T}\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right)$ is the growth rate of output per labor between time *t* and *t+T*. On the right side of equation (9), *a* is the intercept; $\frac{1-e^{-\beta t}}{T}$ is the coefficient of $\log(y_{i,t})$, the initial output; and *ß* is the convergence coefficient.

c. Empirical Research Using Barro and Sala-i-Martin's (1990) Specification

Barro and Sala-i-Martin (1991) observed a robust convergence rate of 2% among the US states and European countries when controlling for regional and sectoral shocks.

Hart (2001) replicated Barro and Sala-i-Martin (1991) using updated data from the 2000 census and observed a robust convergence rate of 1.7% among the US states when controlling for regional and sectoral shocks. He attributed the low convergence rate in 1930s and 1980s to the great depression and stagflation, respectively.

Higgins (2006) tested convergence among the US states using county level data and 41 control variables. He observed a robust convergence rate of 1.6% when controlling for geographical region, state government size, education attainment and the size of the finance, insurance, real estate and entertainment sectors.

When using county level data from 1970 to 1998, Young (2008) found that convergence among the US states is significant but convergence among counties within each individual state is not always significant. Nevertheless, the convergence rates are always positive when significant.

III. Conceptual Model

In this paper, I use Barro and Sala-i-Martin's (1990) specification:

$$\frac{\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right)}{T} = a - [(1 - e^{-\beta t})/T]\log\left(y_{i,t}\right) + (other\ variables) + u_{i,t,t+T} \quad (9)$$

If we assume all states have the same steady state output, we do not need to control for the exogenous determinants of the steady state: the savings rate (s), the population growth rate (n) and the capital depreciation rate (δ). This assumption is tenable among the US states, where these exogenous factors are assumed to be identical under the same institutions, social norms, and culture. However, if we assume that each state has a different steady state output, we need to control for s, n and δ by using the *other variables* option on the right side of equation (9).

IV. Ideal Data and Concept Measurement

The output growth rate $\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right)/T$ and the initial output $\log\left(y_{i,t}\right)$ can be calculated from annual GDP per capita by state, or approximated with annual personal income per capita by state. The savings rate s is equal to the physical capital investment rate plus the human capital investment rate (Mankiw 1992). Unfortunately, neither are available by state as time series data. Besides, it is hard to define human capital investment, which includes education, job training and government programs among others. Therefore, past literature typically uses the percentage of population holding a bachelor's degree as a proxy for the human capital investment rate.

Capital depreciation δ is the rate at which the economic value of the capital stock decreases. Unfortunately, it is also not available by state as time series. Past literature typically fixes capital depreciation at 5% per year. Since it is assumed to be a constant, I do not control for capital depreciation in my specifications. The population growth rate can be calculated from the census in 10-year intervals or approximated with population estimates from the Bureau of Economic Analysis (BEA).

V. Data and Estimation

The equation I estimate is:

$$
\begin{aligned}
&Output\ Growth\ Rate \\
&\quad = \beta_0 + \beta_1 Intial\ Output + \beta_2 Regional\ Dummy + \beta_3 Sectoral\ Composition \\
&\quad + \beta_4 Population\ Growth\ Rate \\
&\quad + \beta_5 Percentage\ of\ Population\ with\ Bachelor's Degree + \mu
\end{aligned}
$$

My dependent variable is *Output Growth Rate*, calculated from BEA annual personal income per capita by state. My key independent variable is *Initial Output*, also calculated from BEA personal income by state.

Regional Dummy is a vector of four dummy variables that divides the states into four BEA regions: Northeast, Midwest, South and West. Each dummy variable has a value of one for states that belong to the region and zero otherwise. This variable controls for regional shocks.

Sectoral Composition is calculated by:

$$S_{it} = \sum_{j=1}^{9} \frac{W_{ijt} \log\left(\frac{y_{i,t+T}}{y_{it}}\right)}{T}$$

where W_{ijt} is the total income share of sector *j* in state *i* at time *t*. $\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right)/T$ is the national growth rate of sector *j* from time *t* to *t+T*. There are 9 sectors. Therefore, S_{it} is the output growth rate of state *i* at time *t* if each of its sectors grows at the national rate. *Sectoral Composition* controls for sectoral shocks. For example, if sector *j* suffers a shock between time *t* and *t+T*, $\log\left(\frac{y_{i,t+T}}{y_{i,t}}\right)/T$ will be low. States with high total income shares of sector *j* at time *t* will have a low $S_{i,t}$. All the inputs of *Sectoral Composition* are from BEA personal income by state and by industry. *Population Growth Rate* and *Percentage of Population with Bachelor's Degree* are obtained from the decennial census. They control for steady state output.

Table 1. Summary Statistics of Variables

Variable Name	Mean	Std. Deviation	Minimum	Maximum
Output Growth Rate	5.32%	0.38%	4.64%	6.11%
Initial Output	6.22	0.38	5.36	6.94
Percentage of Population with Bachelor's degree	15.3%	2.7%	9.9%	21.2%
Population Growth Rate	1.12%	0.83%	0.07%	4.05%
Sectoral Composition	0.062	0.002	0.058	0.066

Note: output growth rate, percentage of population with bachelor's degree, population growth rate and sectoral composition are aggregated over the full sample period 1930 – 2015. Initial output refers to 1930.

Table 2. Summary of Independent Variables by Availability, Source, Expected Sign and Economic Concept

Dependent Variable is Output Growth Rate					
Variable Name	**Availability**	**Source**	**Expected Sign**	**Justification of Expected Sign**	**Economic Concept**
Initial Output	1930 - 2010	BEA	Negative	Convergence predicts that the higher the initial output is, the lower the growth rate is	Initial Output
Percentage of Population with Bachelor's Degree	1940 - 2010	Census	Positive	The higher the human capital investment rate is, the higher the steady state output is and the higher the growth rate is	Human Capital Investment Rate
Population Growth Rate	1930 - 2010	Census	Negative	The higher the population growth rate is, the lower the steady state output is and the lower the growth rate is	Population Growth Rate
Regional Dummy	1930 - 2010	BEA	Indeterminate	The signs correspond to the directions of the regional shocks	Regional Shock
Sectoral Composition	1930 - 2010	BEA	Positive	The larger the sectoral composition variable is, the faster the state's sectors grow and the higher the growth rate is	Sectoral Shock

VI. Results

a. Assuming the Same Steady State Output for All States

If we assume all states have the same steady state output, we do not need to control for the exogenous determinants of the steady state. Therefore, I estimate three models. Model 1 is the basic model with initial output as the only independent variable. Model 2 is the basic model plus the regional dummy. Model 3 is the basic model plus the regional dummy and the sectoral composition variable. Each model is estimated in eleven periods: seven 10-year periods from 1930 to 2000, three 5-year periods from 2000 to 2015, and one full sample period from 1930 to 2015.

Each regression estimates $-(1-e^{-\beta t})/T$, the coefficient of initial output, from which I calculate the implied convergence rate *ß*. Intriguingly, the coefficient of initial output is inversely related to the convergence rate. If $-(1-e^{-\beta t})/T = 0$, then $\beta = 0$, implying that the initial output has no effect on the growth rate and thus no convergence. If $-(1-e^{-\beta t})/T > 0$ then $\beta > 0$, implying that states with higher initial output grow more slowly and thus output *converges*. . If $-(1-e^{-\beta t})/T > 0$ then $\beta > 0$ implying that states with higher initial outputs grow faster and thus output *diverges*.

Table 3. *Coefficients of Initial Output by periods*

	Model 1 Basic Equation				Model 2 Basic with Regional Dummy				Model 3 Basic with Regional Dummy and Sectoral Composition			
	Coef.	*T*	*ß*	R^2	*Coef.*	*T*	*ß*	R^2	*Coef.*	*T*	*ß*	R^2
1930-1940	-0.012 ***	-4.31	1.3%	0.27	-0.004	-1.25	0.4%	0.42	-0.007 ***	-2.18	0.7%	0.52
1940-1950	-0.041 ***	-11.21	5.2%	0.73	-0.040 ***	-13.02	5.2%	0.87	0.039 ***	-11.91	5.0%	0.87
1950-1960	-0.019 ***	-6.45	2.1%	0.46	-0.019 ***	-5.51	2.1%	0.56	-0.027 ***	-8.17	3.1%	0.70
1960-1970	-0.025 ***	-7.76	2.9%	0.56	-0.018 ***	-5.09	2.0%	0.67	-0.022 ***	-5.64	2.5%	0.70
1970-1980	-0.016 ***	-4.00	1.7%	0.24	-0.013 **	-2.79	1.3%	0.28	-0.015 **	-3.26	1.7%	0.31
1980-1990	-0.003	-0.34	0.3%	-0.02	-0.003	-0.57	0.3%	0.55	-0.007	-1.39	0.7%	0.67
1990-2000	0.003	1.02	-0.3%	0.00	0.002	0.58	-0.2%	-0.05	0.008	1.59	-0.8%	0.00
2000-2005	-0.022 **	-3.23	2.3%	0.17	-0.023 **	-2.99	2.5%	0.23	-0.023 ***	-4.00	2.4%	0.59
2005-2010	-0.007	-0.59	0.7%	-0.01	-0.022 *	-1.75	2.4%	0.16	-0.031 **	-2.38	3.3%	0.22
2010-2015	-0.005	-0.73	0.5%	-0.01	0.001	0.09	-0.1%	0.23	-0.002	-0.30	0.2%	0.29
Full Period	-0.009 ***	-18.37	0.9%	0.88	-0.009 ***	-14.36	0.9%	0.88	-0.009 ***	-10.44	0.9%	0.88
No. of Obs	48				48				48			

R-squared is Adjusted
Significance level: *p<0.1, **p<0.05, ***p<0.001

Note: Hawaii, Alaska and District of Columbia are excluded according to Barro and Sala-i-Martin (1990)

Table 3 shows that over the full sample period from 1930 to 2015, US states exhibit a robust and significant convergence rate of 0.9% across all three models, lower than the 2% consensus among earlier literature. The adjusted R-squared of the basic model is 0.88, indicating that the initial output in 1930 can explain most output growth rate variations in the subsequent 85 years. The fitted line in Figure 2 shows the strong negative linear relationship between the output growth rate and the initial output. The residual plot in Figure 3 shows that the basic model is homoscedastic.

Figure 2. Regression of Output Growth Rate from 1930 to 2015 Against Output in 1930

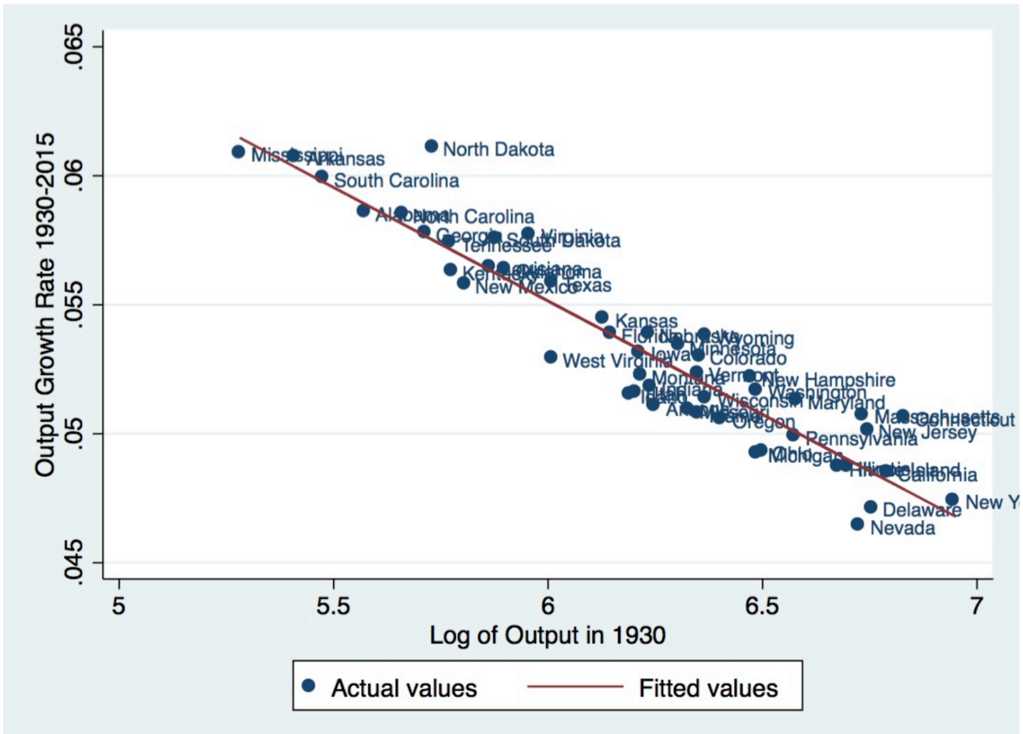
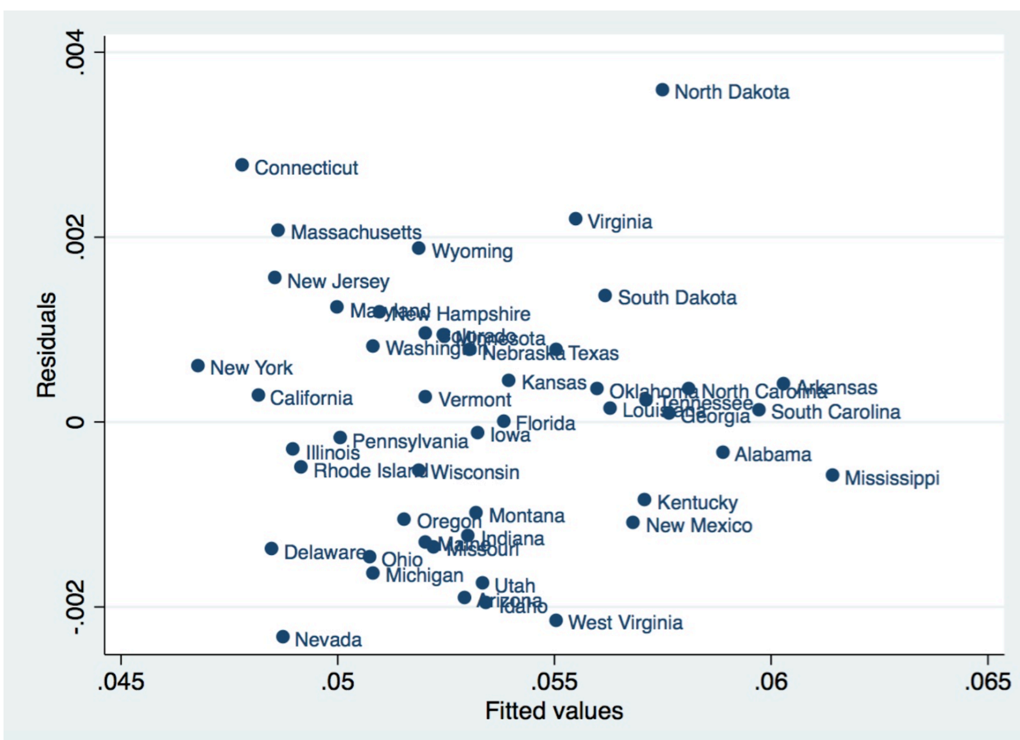


Figure 3. Residuals-versus-fitted (rvf) plot of the Regression in Figure 2



Unfortunately, output convergence in the 10-years and 5-year sub-periods is not consistent. From 1930 to 1980, β is nearly always positive and significant at 0.001 across all three models. In the 1980s, however, β suddenly drops from about 2% to about 0.5% and becomes insignificant across all three models. In the 1990s, β becomes negative, implying output *divergence*.

Figure 4. Fitted Line and Rvf plot of the 1940s and the 1990s (the Basic Model)

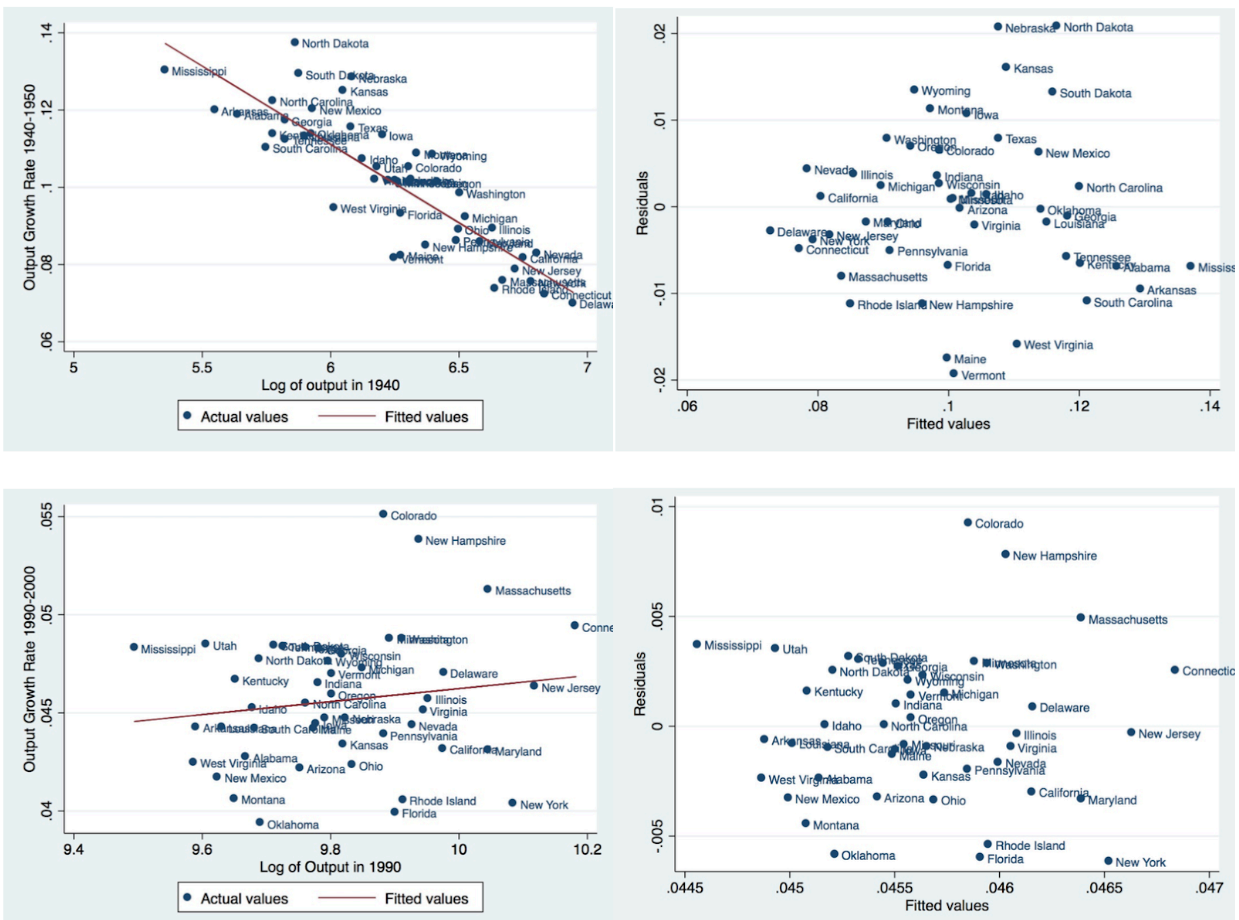


Figure 4 compares the 1940s (the most significant convergence) with the 1990s (divergence). The residual plots show that both sub-periods are homoscedastic. In fact, all the sub-periods from 1930 to 2015 are homoscedastic (Appendix). Furthermore, to examine whether weak convergence results from regional and sectoral shocks, we can compare the adjusted R-squared and T-statistics of Model 1 with those of Model 3. By adding the regional dummy and the sectoral composition variable, Model 3 improves the adjusted R-squared from -0.02 to 0.67 and the T-statistics from -0.34 to -1.39 in the 1980s. This indicates that regional and sectoral shocks can explain most output growth rate variations in the 1980s. The significance of *ß* improved after controlling for these shocks but is still insignificant. In the 1990s, Model 3 improves neither the adjusted R-squared nor the T-statistics. The output *divergence* becomes more significant when controlling for regional and sectoral shocks. These imply that during the 1980s and the 1990s, output convergence was disrupted by factors other than regional and sectoral shocks. Intensive technological evolution, the emerging information technology sector, and the expanding service sector might be responsible factors. These transformations correspond with BEA industrial sector reclassification in 2001, which added the information sector and split the service sector into 9 subsectors.

ß is positive and significant from 2000 to 2005 but becomes insignificant following the Great Recession. Fortunately, regional and sectoral shocks can explain the weak convergence from 2005 to 2010. When controlling for these shocks, *ß* rises from 0.7% to 3.3% and becomes significant. The adjusted R-squared improves from -0.01 to 0.22. However, the same is not true for the period 2010 to 2015. When controlling for regional and sectoral shocks, *ß* declines from 0.5% to 0.2% and becomes less significant.

Consequently, the full period convergence rate of 0.9% is lower than the 2% consensus among earlier literature, because the magnitude and significance of *ß* falls after 1980. Intensive technological evolution, uncontrolled by my models, might be responsible for the weak convergence since 1980.

Figure 5. Implied ß of Model 1 (the basic equation) and Model 3 (Controlled for Regional and Sectoral Shocks)

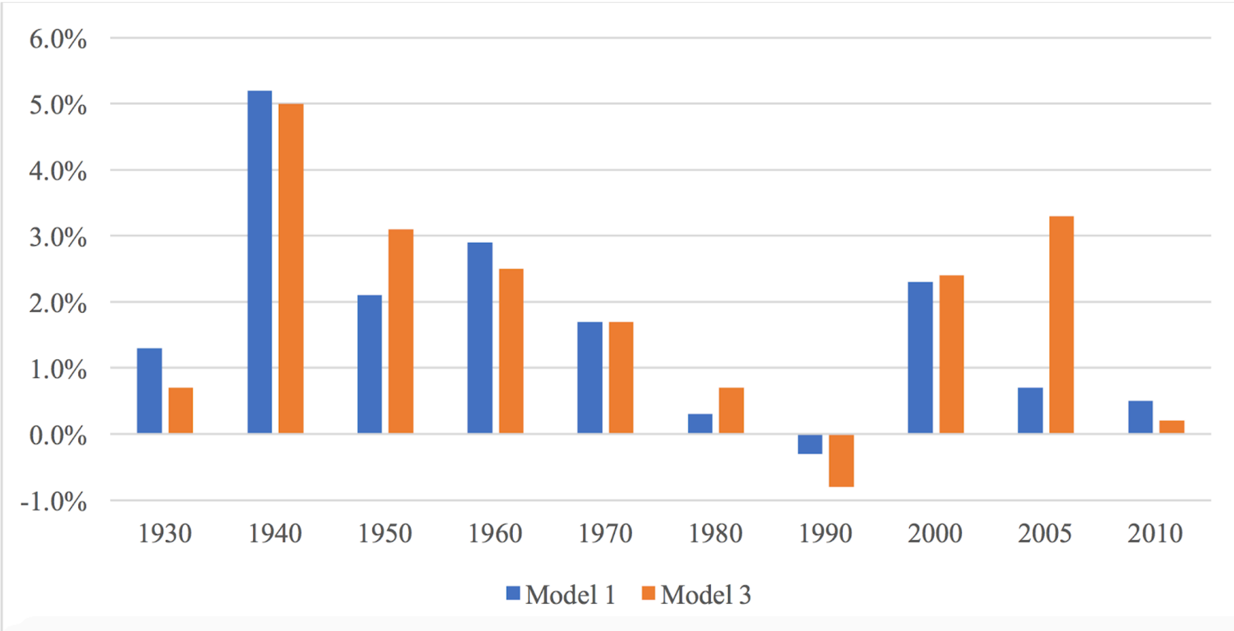


Figure 5 compares the implied *ß* of the basic equation and that of Model 3, which controls for regional and sectoral shocks. It shows that for the sub-periods after 1980, the implied *ß* becomes larger after controlling for these shocks. Besides, Table 3 shows that for the sub-periods after 1980, Model 3 improves the significance of *ß* and the adjusted R-squared 4 out of 5 times. Therefore, Model 3 is a good starting point for the next subsection, which relieves the assumption of identical steady state outputs.

b. Assuming Different Steady State Outputs

If we assume that each state has a different steady state output, we need to control for the exogenous determinants of the steady state. In this section, I control for the population growth rate and the percentage of population with bachelor's degree, a proxy for human capital investment rate. All three models in this section control for the regional and sectoral shocks.

They are identical to Model 3 except that Model 4 controls for educational attainment (bachelor's degree), Model 5 controls for population growth and Model 6 controls for both.

Table 4. Coefficients of Initial Output by Models and Estimation Periods

	Model 4 Education Attainment				**Model 5 Population Growth**				**Model 6 Education Attainment and Population Growth**			
	Coef.	*T*	*ß*	R^2	*Coef.*	*T*	*ß*	R^2	*Coef.*	*T*	*ß*	R^2
1930-1940					-0.00888	-2.31	0.8%	0.51				
1940-1950	-0.043 ***	-9.03	5.7%	0.87	-0.037 ***	-9.35	4.6%	0.87	-0.042 ***	-8.62	5.4%	0.88
1950-1960	-0.035 ***	-8.33	4.2%	0.74	-0.030 ***	-8.66	3.5%	0.73	-0.035 ***	-8.46	4.3%	0.75
1960-1970	-0.024 ***	-4.55	2.7%	0.69	-0.026 ***	-6.05	3.0%	0.71	-0.028 ***	-5.07	3.3%	0.71
1970-1980	-0.020 **	-3.14	2.2%	0.31	-0.015 **	-3.15	1.6%	0.29	-0.020 **	-3.07	2.2%	0.29
1980-1990	-0.014 **	-2.27	1.5%	0.69	-0.007	-1.37	0.8%	0.66	-0.014 **	-2.22	1.5%	0.68
1990-2000	0.000	-0.08	0.0%	0.11	0.009 *	1.90	-0.9%	0.09	-0.001	-0.13	0.1%	0.26
2000-2005	-0.025 **	-3.02	2.7%	0.58	-0.024 ***	-3.92	1.5%	0.58	-0.027 **	-3.01	2.9%	0.57
2005-2010	-0.040 **	-2.22	4.5%	0.21	-0.028 **	-2.16	3.1%	0.21	-0.039 **	-2.15	4.4%	0.21
2010-2015	-0.027 **	-2.65	2.9%	0.42	-0.009	-1.25	0.9%	0.42	-0.025 **	-2.57	2.7%	0.48
Full Period	-0.010 ***	-14.80	1.1%	0.93	-0.009 ***	-10.37	0.9%	0.88	-0.010 ***	-14.80	1.1%	0.93
No. of Obs	48				48				48			

R-squared is Adjusted
Significance level: *p<0.1, **p<0.05, ***p<0.001

Table 4 shows that over the full sample period from 1930 to 2015, US states exhibit a robust and significant convergence rate of about 1% across all three models. The adjusted R- squared for all three models is above 0.88. When controlling for educational attainment, Model 4 and Model 6 improve the T-Statistics from -10.44 (Model 3) to -14.80 and the adjusted R- squared from 0.88 (Model 3) to 0.93. *ß* rises from 0.9% to 1.1%. In contrast, when controlling for the population growth rate, Model 5 has the same adjusted R-squared as and a lower T- Statistics than Model 3. This evidence suggests that educational attainment *might be* necessary in the specification, whereas the population growth rate *might not be*.

Table 5. Coefficients and T-Statistics of Educational Attainment in Model 4 and the Population Growth Rate in Model 5

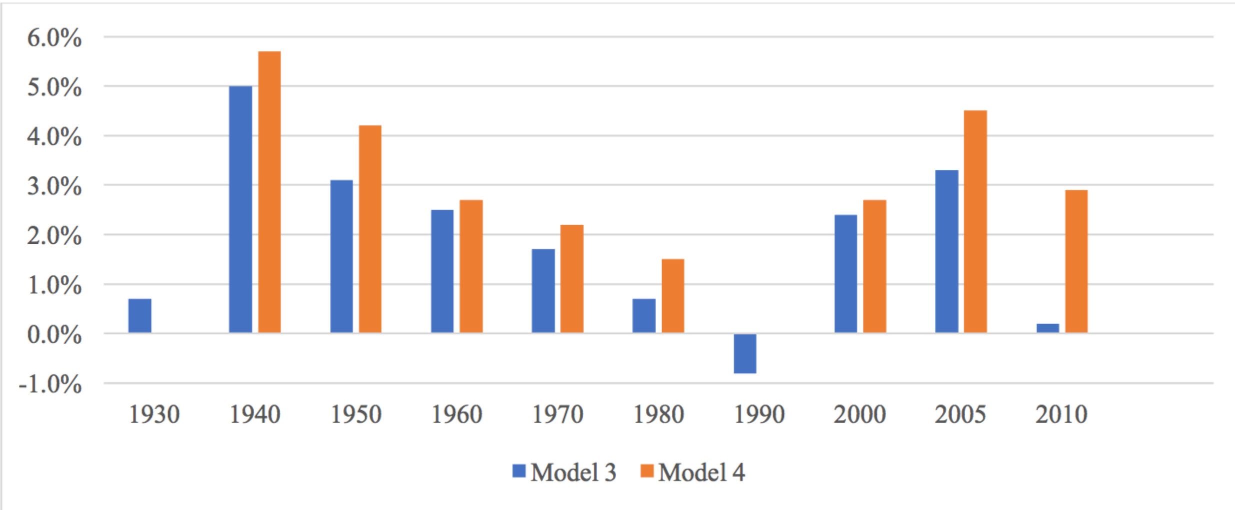
	Educational Attainment (Model 4)		Population Growth Rate (Model 5)	
	Coefficient	*T*	*Coefficient*	*T*
1930-1940	0.15	0.77		
1940-1950	-0.13	-1.03	0.0021	1.14
1950-1960	0.11 **	2.10	0.0022 **	2.73
1960-1970	0.16 *	1.88	0.0003	0.45
1970-1980	0.02	0.20	0.0005	1.10
1980-1990	0.02	0.17	0.0006 *	1.92
1990-2000	0.19 **	2.28	0.0005 **	2.51
2000-2005	0.08	0.53	0.0001	0.38
2005-2010	-0.27	-0.79	0.0004	0.76
2010-2015	0.62 **	3.27	0.0010 **	3.20
Full Period	-0.02	-0.59	0.0004 ***	5.69
No. of Obs	48		48	
Significance level: *p<0.1, **p<0.05, ***p<0.001				

To better determine whether educational attainment and the population growth rate are necessary in the specification, Table 5 summarizes the coefficients and t-statistics of educational attainment in Model 4 and population growth rate in Model 5. The coefficient of educational attainment is positive and significant over the full sample period and is positive in 10 out of 10 sub-periods. This aligns with our expectation that higher human capital investment rate implies higher steady state output and higher output growth rate. In contrast, the population growth rate coefficient is negative but insignificant over the full sample period, and is positive in 8 out of 10 sub-periods. This contradicts our expectation that higher population growth rate implies lower steady state output and lower output growth rate. This anomaly might result from interstate migration: people migrate from low-growth states to high-growth states seeking better opportunities, thereby driving up population growth in states with higher output growth rate. This evidence confirms that I should keep educational attainment and drop population growth rate in my specification. Therefore, I only use Model 4 for the rest of my discussion.

Compared to Model 1, 2 and 3, Model 4 has three major improvements. First, as discussed above, Model 4 improves the overall fit and the significance of β over the full sample period. Second, Model 4 has significant β in all sub-periods except the 1990s. Third, Model 4 observes a significant convergence rate in the 1980s and the 2010s, while Model 1, 2 and 3 all fail to do so. The improved significance of β is especially obvious in the 2010s. The t-statistic rises from -0.73, 0.09 and -0.30 to -2.65. β rises from 0.5%, -0.1% and 0.2% to 2.9%. The adjusted R- squared rises from -0.01, 0.23 and 0.29 to 0.42. Therefore, educational attainment is necessary in the specification as it makes β more significant, improves the overall fit, and has theoretically justified signs.

Figure 6 compares the implied β of Model 3 and Model 4, whose only difference is that Model 4 has educational attainment in its specification. Figure 6 shows that output convergence is more pronounced when Model 4 controls for educational attainment, especially in the 2010s. This indicates that human capital, especially investment in higher education (bachelor’s degree), has become more important in driving output growth recently.

Figure 6. Implied ß of Model 3 and Model 4



VII. Conclusion

This paper uses Barro and Sala-i-Martin's (1990) specification and estimates a robust and significant output convergence rate of 1% among US states over the full sample period 1930 – 2015. This is markedly lower than the 2% consensus among earlier literature, because convergence is inconsistent over the sub-periods and declines after 1980. When controlling for regional and sectoral shocks, Model 3 improves the overall fit and the significance of *ß*, but still fails to isolate a significant *ß* from 1980 to 2000, which I attribute to intensive technological evolution after 1980. When controlling for human capital investment rates, Model 4 isolates a significant *ß* in the 1980s and the 2010s whereas all other models fail to do so. Therefore, I conclude that educational attainment is necessary in Barro and Sala-i-Martin's (1990) specification if we assume different steady state outputs.

My research, however, is limited in three ways. First, the ten-years and five-years sub-periods are arbitrarily set up to accommodate for the availability of decennial census data. Different convergence rates might arise simply if the sub-periods are divided differently. Second, my model does not control for technological evolution, which I suspect is responsible for the weak convergence since 1980. Third, to ensure consistency, I use BEA's 9 industry classifications to calculate the sectoral composition variables throughout the entire period. However, this approach risks losing the nuances of sectoral shocks. For example, 9 subsectors are collapsed into one service sector, which now represents 80% of US GDP.

Therefore, future research can explore the effects of technological evolution on convergence by controlling for the employment share or total income share of technology sectors by state. Besides, since controlling for educational attainment helps isolate more pronounced convergence effects and improves the overall fit, future research can also include proxies for physical capital investment rate in the specification.

References

Barro, Robert J., & Sala-i-Martin, X. (1990). Economic Growth and Convergence across the United States. *NBER Working Paper Series*, no.3419

Barro, Robert J., & Sala-i-Martin, X. (1991). Convergence across States and Regions. *Brookings Papers Econ. Activity*, no.1, 107-182.

Hart, J. (2001). Government size and Convergence across the United States. *Macalester College Economics Honors Paper Series*

Higgins, Matthew J., Levy, Daniel., Young, Andrew T. (2006) Growth and Convergence across the United States: Evidence from County-Level Data. *The Review of Economics and Statistics*, no 88(4), 671-681.

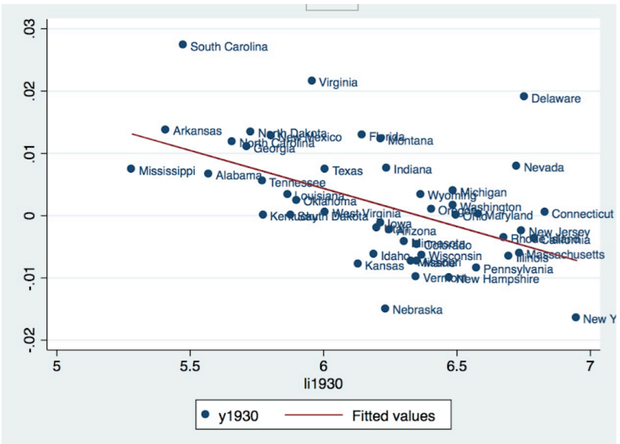
Mankiw, N.G., Romer, D., Weil, D.N. (1992) A contribution to the empirics of economic growth. *Quarterly Journal of Economics,* 107, 407–437.

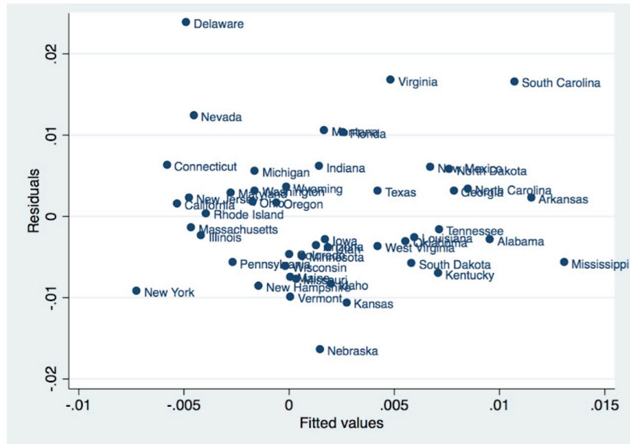
Solow, Robert M. (1956). A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics*, LXX, 65-94.

Young, Andrew T., Higgins, Matthew J., Levy, Daniel. (2008). Sigma Convergence versus Beta Convergence: Evidence from U.S. County-Level Data. *Journal of Money, Credit and Banking*, Vol. 40, No.5, 1083-1093.

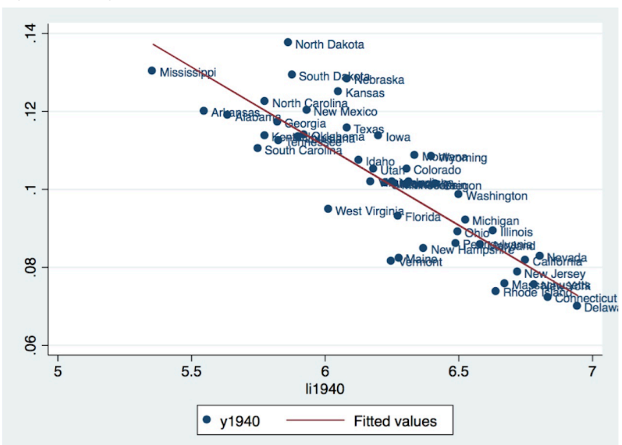
Appendix: Fitted Lines and Residual Plots for Each Sub-period (the Basic Model)

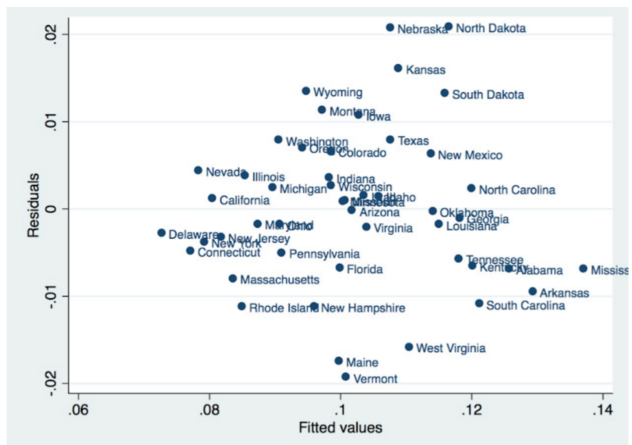
1930 – 1940



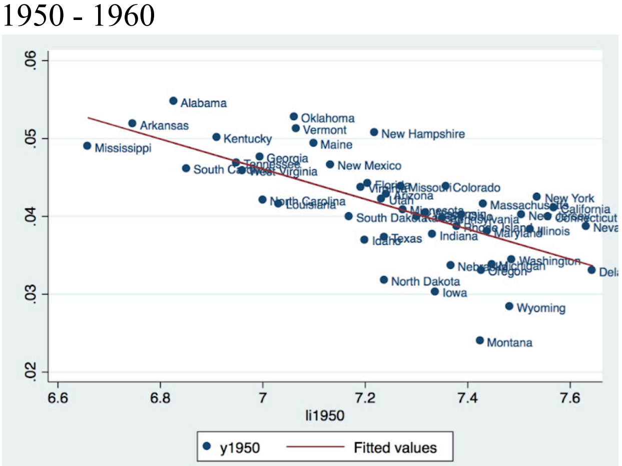


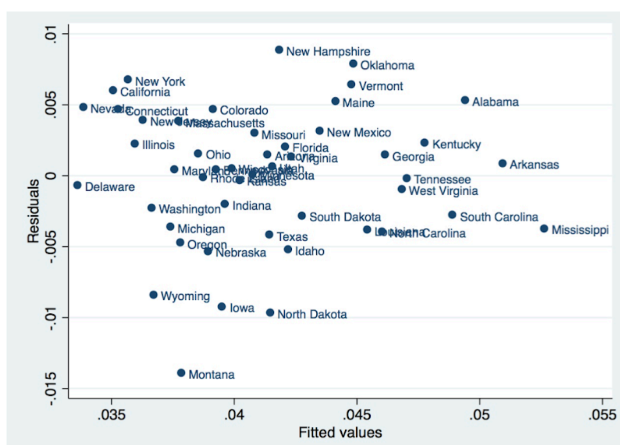
1940 - 1950



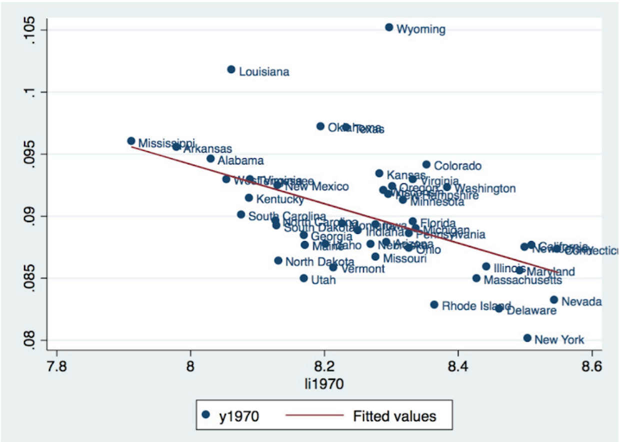


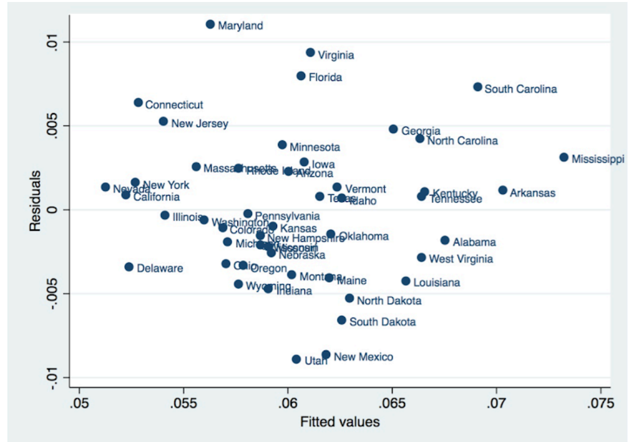
1950 - 1960



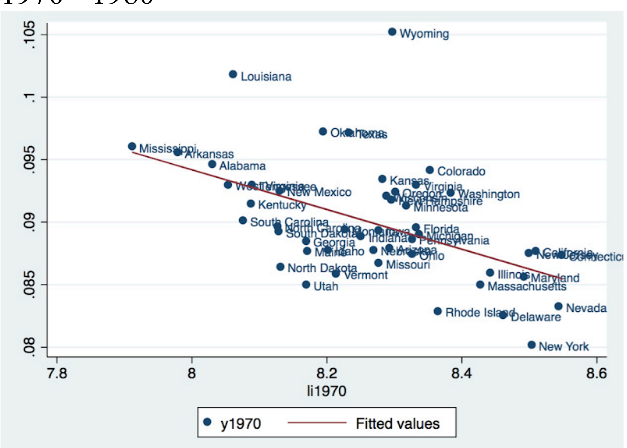


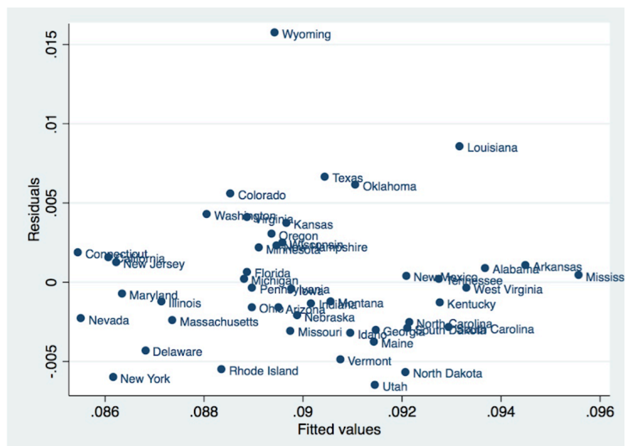
1960 - 1970



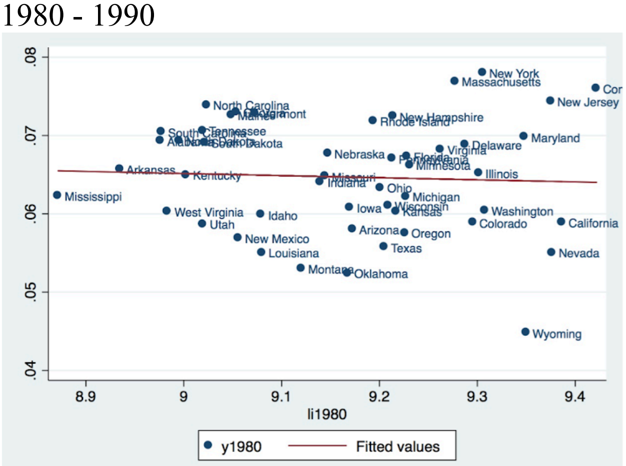


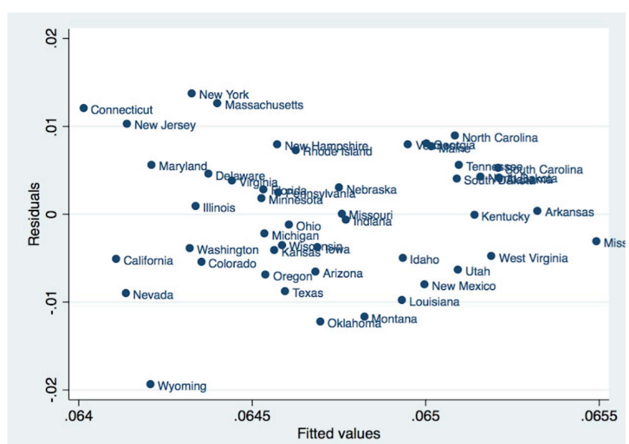
1970 - 1980



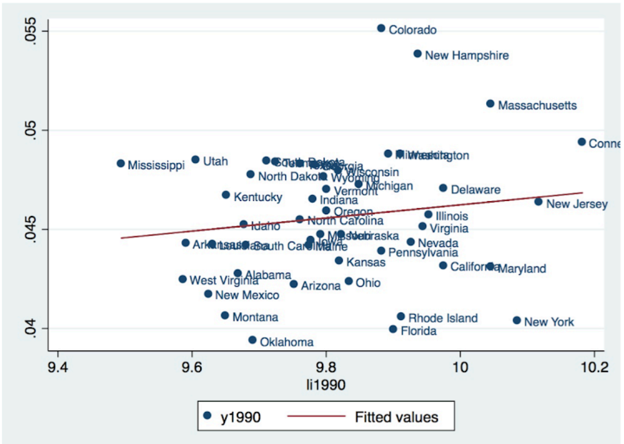


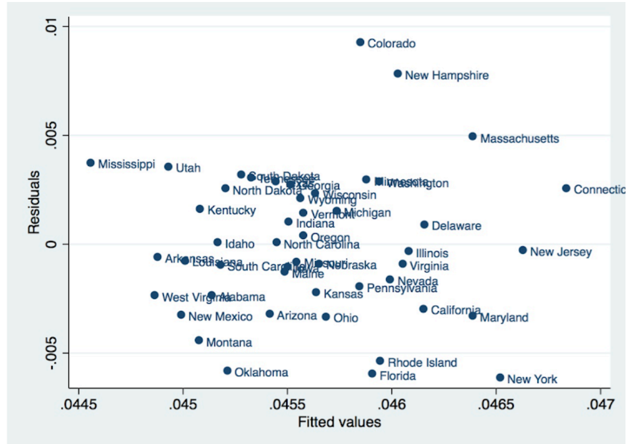
1980 - 1990



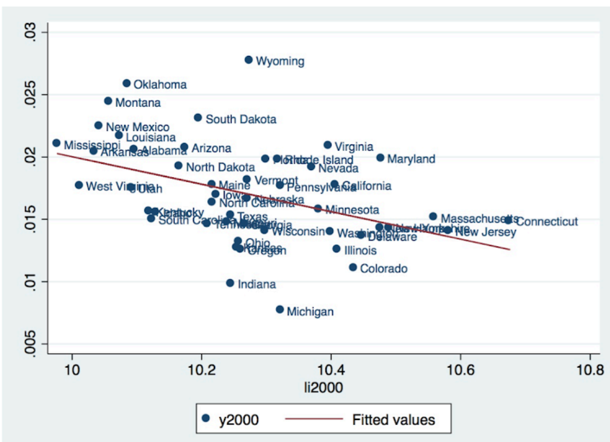


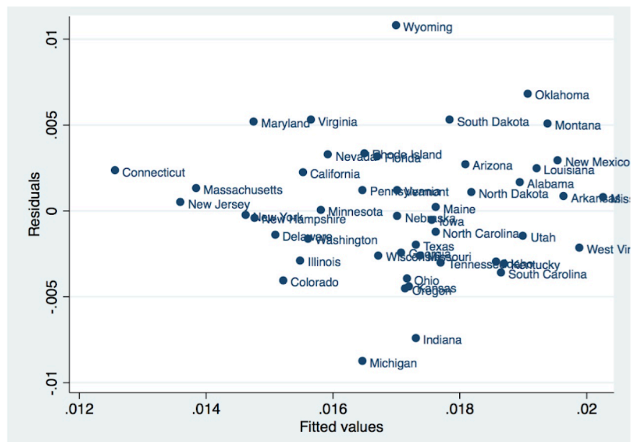
1990 - 2000



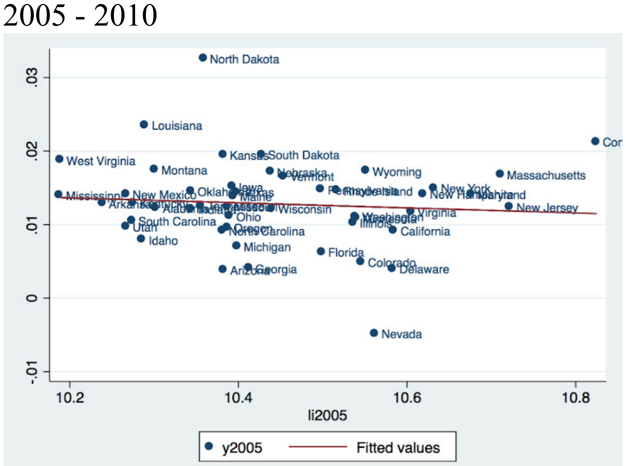


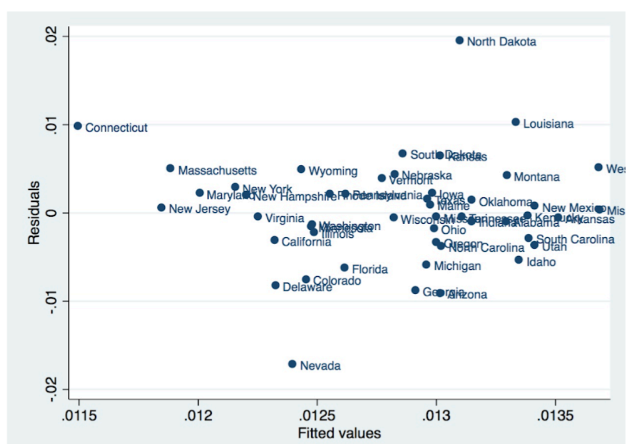
2000 - 2005



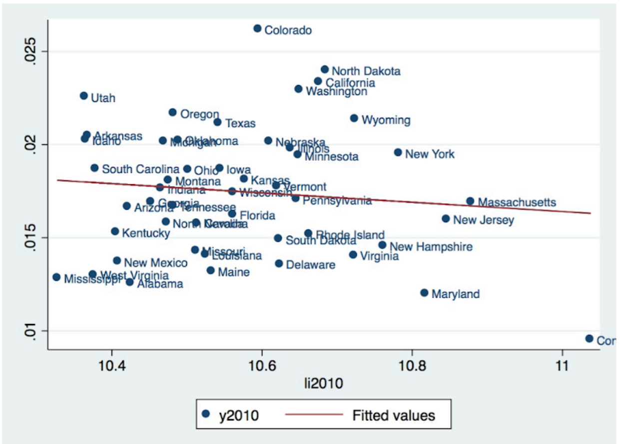


2005 - 2010

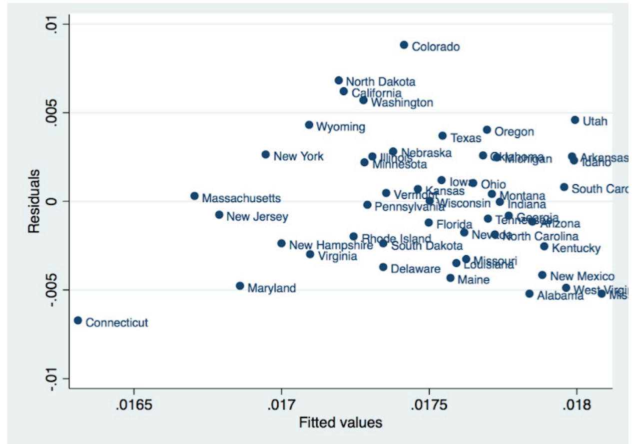




2010 - 2015



y2010
Fitted values
li2010
.01
.015
.02
.025
10.4
10.6
10.8
11
Colorado
North Dakota
California
Washington
Utah
Oregon
Texas
Wyoming
Arkansas
Idaho
Michigan
Oklahoma
Nebraska
Illinois
Minnesota
New York
South Carolina
Ohio
Iowa
Montana
Indiana
Kansas
Vermont
Wisconsin
Pennsylvania
Massachusetts
Arizona
Georgia
Tennessee
Florida
New Jersey
North Carolina
Nevada
Kentucky
Rhode Island
South Dakota
New Hampshire
Missouri
Louisiana
Virginia
New Mexico
Delaware
Maine
Mississippi
West Virginia
Alabama
Maryland



Residuals
Fitted values
-.01
-.005
0
.005
.01
.0165
.017
.0175
.018
Colorado
North Dakota
California
Washington
Wyoming
Utah
Texas
Oregon
New York
Nebraska
Illinois
Minnesota
Oklahoma
Michigan
Arkansas
Idaho
Iowa
Ohio
Kansas
Vermont
Montana
Massachusetts
Pennsylvania
Wisconsin
Indiana
South Carolina
New Jersey
Florida
Tennessee
Georgia
Arizona
Nevada
North Carolina
Rhode Island
New Hampshire
South Dakota
Kentucky
Virginia
Missouri
Louisiana
Delaware
Maine
New Mexico
Maryland
Alabama
West Virginia
Connecticut

A Game of Numbers: How *Moneyball* Affected Player Valuation and Competitive Balance in Major League Baseball

Matthew Yang

Introduction to Econometrics

Abstract

This paper investigates the effect that the publication of Michael Lewis's *Moneyball* had on the Major League Baseball (MLB) labor market. *Moneyball* details the analytical tools that the Oakland Athletic's used to strategically build a team that could compete with teams with more money. Using economic models, I measure the extent to which the MLB labor market was efficient prior to 2003, which was the central hypothesis in *Moneyball*. I find that there was indeed an inefficiency in the MLB labor market prior to 2003, but the market began to correct itself after the book was published. My results suggest that teams that were privy to "sabermetrics," the statistical analysis described in *Moneyball*, enjoyed a brief advantage over other teams in the league. However, as information on sabermetrics became more widespread, that advantage disappeared, and big-market teams were more successful relative to small-market teams.

I. Introduction

"When the numbers acquire the significance of language, they acquire the power to do all of the things language can do: to become fiction and drama and poetry" (Lewis, 2003). Bill James wrote these words about the power of statistics in baseball in his *Historical Baseball Abstract* (1985 edition). He continues, "Baseball statistics, unlike the statistics in any other area, have acquired the powers of language" (Lewis, 2003). Indeed, James saw baseball differently—not merely as a game of players who pitch, hit and field balls, but as a game of numbers. His novel ideas about the statistical analysis of baseball sparked the "sabermetrics"[5] revolution. Subsequently, economists and mathematicians began to use data and analytics not only to understand the game but also to change the way it is played.

Michael Lewis's *Moneyball: The Art of Winning an Unfair Game* (2003) is a vivid account of how the Oakland Athletics (A's) were able to use revolutionary statistical techniques to win on a meager budget. Although the A's did not win the championship in their 2002 season—the season that *Moneyball* chronicles—they showed how to allocate resources in an efficient way. Specifically, the A's exploited an inefficiency in the Major League Baseball (MLB) labor market. Teams and front offices had previously misjudged the productivity of players and had thus valued them incorrectly. Using sabermetrics, the A's front office discovered this inefficiency and then applied it to their advantage. The A's were able to find players who had been productive for their teams, yet undervalued in the labor market. But the A's ability to use sabermetrics to exploit an inefficient labor market was short-lived, as the market corrected after the publication of *Moneyball*.

---

[5] Sabermetrics, a term coined by James himself, was defined as "the search for objective knowledge about baseball." The term derives from the acronym SABR, which stands for the Society for American Baseball Research.

Using econometric techniques, this paper examines the central hypothesis in *Moneyball*: that a player's ability to get on base was undervalued in the MLB labor market. Although previous empirical papers have analyzed the *Moneyball* hypothesis, this paper extends the analysis to look at how the labor market has changed in the 15 years following the publication of *Moneyball*. As such, this paper examines the labor market inefficiency in the early 2000s, as well as the continuously changing labor market in the context of new baseball trends. I investigate whether there was truly an inefficiency in the labor market pre-*Moneyball*, and how the market responded post-*Moneyball.* Additionally, my paper seeks to determine how *Moneyball* changed the competitive balance in the MLB. This represents a departure from previous research and lends important insights into the role of sabermetrics in baseball. To answer these questions, I first construct a theoretical framework on which my analysis is based by reviewing the relevant economic literature. Next, I lay out my estimation strategy and discuss the data I am using. Lastly, I present and discuss my findings.

II. Theory and Literature Review

Market efficiency is a basic concept in economic theory. It is a concept that is applicable to every type of market, no matter the good or service. Fama (1970) discusses market theory in regard to capital markets. Fama explains that efficient markets "fully reflect" all available information. As a result, there are forces in efficient markets that always bring prices to an equilibrium, and thus such markets can never systematically generate abnormally high or low returns over long periods of time.

In a perfectly competitive labor market, a worker earns the exact amount that he or she contributes (in terms of revenue) to the firm. In other words, a worker's salary is equal to his or her marginal revenue product (MRP).[6] In most instances, measuring the contribution of an individual working in a large organization such as a firm is not an easy task. In baseball, however, economists have been able to do so as there exists a plethora of measures of player and team performance. As Hakes and Sauer (2006) argue, “A case can be made that more is known about pay and quantified performance in [the MLB] market than in any other labor market in the American economy”. It is this quantitative characteristic that makes the baseball market conducive to economic analysis. Many empirical papers have analyzed the labor market for MLB players. During the past decade, much of the literature on baseball economics has looked at the implications of Michael Lewis’s *Moneyball*.

In his best-selling book, Lewis wrote about the innovations by which Billy Beane, the general manager of the Oakland A’s, set the valuation of MLB players. Beane’s revolution and the economics debate that followed were actually based on previous academic work. Scully (1974) was the first to systematically calculate the MRPs of not just professional athletes, but of workers in any sector of the economy. A baseball player’s MRP is the amount of team revenue that is produced by the addition of the player to the team. Scully was able to estimate MRP by recognizing (1) that a team’s winning percentage is a function of player talent, and (2) that a team’s revenue is a function of its winning percentage. Thus, Scully established a functional relationship between player talent and team revenue. Scully concluded that the MLB’s now abolished reserve clause, which granted teams the property rights of players for an indefinite amount of time, gave monopsony power to baseball teams and resulted in player salaries that were below their MRPs.

[6] Marginal revenue product is a concept in economics that measures how much revenue an additional worker adds to a firm.

Being the first to develop a model to quantify MRPs, Scully was subject to a lot of criticism. Krautmann (1999) claims that Scully systematically overestimates MRP in professional baseball, arguing that Scully's estimates suggest that even the highest-paid players are significantly underpaid and exploited by teams. Krautmann proposes a new method to calculate MRP using information from free-agent contract negotiations, and finds that for players who are not bound by the reserve clause the MRPs are much closer to their actual salaries. Instead of estimating MRP using team revenues, Krautmann bases his method "on the intuitive notion that the intense bidding process that determines free agent salaries should align wages to marginal revenue products". In particular, the complex institutional features of the MLB labor market, especially free agency, affect the efficiency of the market. Thus, Krautmann concludes, it is only when a player becomes a free agent that we should see a close correspondence between salary and MRP. Both Scully and Krautmann offer valuable contributions to the existing literature. The methodological approach of this paper draws mainly from Scully because data on free agency is difficult to obtain.

Similar methodologies have been used to analyze the efficiency of labor markets in other professional sports. Roach (2017) examines which positions in the National Football League (NFL) are most valuable relative to their respective productivity. As with all professional sports teams, one of the main objectives of NFL teams is to sign talented players in order to maximize the team's winning percentage, subject to a limited budget. This constrained optimization problem can be solved by equating a player's marginal revenue product to his respective wage across all of a team's positions. Roach's paper examines this economic problem in the NFL, and finds that salary resources devoted to offensive players had a larger marginal impact on team productivity than did resources devoted to defensive players.

All professional sports teams face this optimization problem; it is the task of general managers and team front offices to solve it. *Moneyball* depicts how General Manager Billy Beane was able to use groundbreaking sabermetric techniques to solve the problem for the small-market Oakland A's.

Several analyses have evaluated the particular arguments put forth in *Moneyball.* Hakes and Sauer (2006) were the first to empirically test Lewis's central hypothesis: that the offensive skills of drawing walks and getting on base in ways other than hitting was undervalued in the MLB labor market before the publication of *Moneyball*. They employ data from 2000-2004 and find that in 2004, one year after *Moneyball* was published, there was a large appreciation in the value of the ability to get on base. Updating the work of Hakes and Sauer, Baumer and Zimbalist (2014) find that the market may have overcorrected, observing that the 2004 spike in value of on-base percentage decreased in 2005 and 2006. Both papers use the same methodology to test the efficiency of the MLB labor market.

In order to analyze the efficiency of the baseball labor market, both Hakes and Sauer (2006) and Baumer and Zimbalist (2014) first model player productivity to determine which offensive skills contribute the most to winning games. Next, they seek to identify the relationship between the same individual offensive statistics and salary levels. In an efficient labor market, the statistics that contribute the most to winning games will be valued highest and influence player salaries the most. This paper will use the same methodology and continue to look at how certain sabermetric statistics are valued in the marketplace for baseball players.

In the *Moneyball* debate, a key question, which previous literature has not addressed, remains unanswered: How did *Moneyball* affect the competitive balance in the MLB? The Oakland A's were able to compete with large-market teams because they had a competitive advantage: an innovative GM and front office that employed novel techniques to compile a championship-caliber team with very limited resources. After the publication of *Moneyball*, that advantage quickly disappeared with the dissemination of these techniques. As evident by the rise of young sabermetric-savvy GMs, these new analyses changed the competitive landscape of the MLB. Therefore this paper also seeks to address the question of how the *Moneyball* revolution tilted the MLB's competitive balance.

Zimbalist (2002) discusses the multitude of ways to measure competitiveness. Competitive balance can be measured within a single season or across many seasons. One measure of cross-season competitive balance is the concentration of championships among a league's teams over a period of time. Alternatively, it can be measured as how close teams are in the standings at the end of a given season. One common way to measure within-season competitive balance is to calculate the standard deviation of winning percentages for all teams in the league. Similarly, some use the ratio of the actual to the idealized standard deviation of winning percentages as a measure of competitive balance. The idealized standard deviation of winning percentage represents a perfectly balanced league and is defined as:

$$\sigma = \frac{.5}{\sqrt{N}}$$

where N is the number of games played in the season. In baseball, each team plays 162 games in a given season. I use this last method (the ratio of the standard deviation of winning percentages to σ) in my calculations.

III. Conceptual Model and Ideal Data

This paper's strategy for modeling labor-market efficiency directly follows from the methods employed by Hakes and Sauer's seminal 2006 paper. I develop two main models: (1) player productivity and (2) labor-market valuation.

The first model seeks to identify the statistics that contribute the most to a team's winning percentage. The objective of sabermetrics is to evaluate baseball players using rigorous quantitative analysis. Traditional (pre-sabermetrics) measures of a player's offensive contribution are statistics such as batting average and runs batted in (RBI). These statistics do not take into account a few very important aspects of offensive productivity. Specifically, the ability to draw walks (and generally get on base without getting a hit) is ignored by traditional baseball statisticians. After all, winning baseball games is about scoring more runs than the other team. The more runners a team gets on base (through any means), the higher is the probability of scoring a run. Moreover, one of the major findings of sabermetrics and the central premise of *Moneyball* is that the ability to get on base greatly increases the likelihood of winning games. The traditional batting-average statistic also does not take into account differences in the types of hits. A home run and a single are weighted the same in this traditional statistic, but the former is obviously more productive than the latter.

To address these weaknesses in traditional statistics, sabermetricians developed two new metrics: on-base percentage (OBP) and slugging percentage (SLG). OBP is a measure of how often a batter reaches base per plate appearance. A batter can get on base in several ways: making a hit, drawing a walk, getting hit by a pitch, benefiting from a defensive error, or grounding into a fielder's choice. These ways, except for reaching base on errors and fielder's choices, are included in OBP. SLG measures a batter's power by weighting extra-base hits proportionally. These metrics are explicitly defined as follows:

$$OBP = \frac{Hits + Walks + Hit\ By\ Pitch}{At\ Bats + Walks + Hit\ By\ Pitch + Sacrifice\ Flies}$$

$$SLG = \frac{(1 * Singles) + (2 * Doubles) + (3 * Triples) + (4 * Home\ Runs)}{At\ Bats}$$

My first model of player productivity tests the linear relationship between a team's OBP and SLG (independent variables) and the team's winning percentage (the dependent variable). After identifying how OBP and SLG relate to team winning percentage, I then create a second model for labor-market valuation, which relates these new offensive metrics to player salary.

The ideal data for measuring player productivity and labor market valuation would include detailed information on individual player contracts, as well as player and team batting statistics. The complex institutional features of the market for baseball players make modeling labor-market efficiency difficult. As Baumer and Zimbalist (2014) note, there are three categories of players in the MLB labor market. First, a player may be a free agent, meaning that his current contract is expiring and he can enter the open market to sell his services to the highest bidder. Second, baseball players who are still under contract but who have more experience are eligible to enter arbitration with their current teams. Third, rookies and other players fairly new to the league cannot enter either free agency or arbitration. Free agents and arbitration-eligible players tend to receive more accurate compensation for their skills than do players who are pre-arbitration eligible because they are being assessed their fair market value. Rookies receive low "rookie contracts" during their first few years, so they may receive less than their value to a team. For example, Los Angeles Angels star outfielder Mike Trout was arguably the most valuable player in the MLB during his 2012 season, but he earned only $482,500 because he wasn't eligible for arbitration. Both Hakes and Sauer (2006) and Baumer and Zimbalist (2014) are able to obtain data on each of these three categories of players, thus enabling them to better explain variation in player salaries. Due to a lack of access to detailed player contract information, this paper solely uses data on offensive player statistics to model salary.

IV. Actual Data and Model Specification

The team and player data was obtained from the Lahman baseball database, which includes statistics from 1871 to 2016. For the player productivity model, I use data from the 2000 to 2016 seasons because team data on sacrifice flies, a statistic that is included in the calculation of OBP, was not available before 2000. For the valuation model, I use data from 1995 to 2016. Using two decades of data allows me to see how the valuations of OBP and SLG change over different time periods and how the publication of *Moneyball* affected the valuations.

There are two sides of a baseball game: offense and defense. Successful ball clubs are able to score runs and limit the number of runs allowed. To account for this fundamental principle in the productivity model, I run a regression on the difference between the offensive statistics of a team and those of their opponent. This regression allows us to capture more of the variation in team winning percentage because a baseball team wins a game based on their performance relative to their opponent's performance. All regressions are performed using an ordinary least squares (OLS) estimator.

$$Team\ Winning\ Percentage = \beta_0 + \beta_1 \cdot (OnBase - OppOnBase) + \beta_2 \cdot (Slugging - OppSlugging)$$

A player's salary is determined by his expected contribution to the team. The best predictor of how a player will perform in a given season is his performance in the previous season. Thus the model takes one-year lagged values of OBP, SLG and Plate (which is the number of a player's plate appearances). Players who play more and accumulate more at-bats would be expected to be paid more relative to players who are mainly substitutes and do not play or bat regularly. I use the dummy variables "Catcher" and "Infielder" to account for defensive contributions to team winning percentage. These two positions are critically important for team defenses. Since OBP, SLG and Plate measure only offensive performance, these dummy variables isolate players of these key defensive positions. We would expect catchers and infielders to earn more relative to other position players of equal offensive skills. I also generate an indicator variable for the team that a player belongs to. Some big-market teams, like the New York Yankees, enjoy the benefit of especially large payrolls and thus pay more than most other teams for the same player.

$$\ln{(Salary)_t} = \beta_0 + \beta_1 OBP_{t-1} + \beta_2 SLG_{t-1} + \beta_3\ Plate_{t-1} + \beta_4\ Catcher + \beta_5\ Infielder + \gamma\ Team$$

V. Results

a. Labor Market Efficiency

Table 1: Dependent Variable is Team Winning Percentage

	Model		
	1	2	3
On-Base	2.657***		1.742***
	(36.32)		(19.52)
Slugging		1.816***	0.924***
		(37.92)	(14.25)
Constant	0.457***	0.411***	0.427***
	(226.63)	(148.06)	(157.05)
N	510	788	510
R^2	0.722	0.647	0.801

Notes: Coefficients were estimated using ordinary least squares. T-statistics are in parentheses. Asterisks represent significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 1 shows the results from the player-productivity model. I ran three separate regressions: one with just OBP, one with just SLG, and one with both. All variables are statistically significant at the .01-level in all three regressions. As evident in the first column of Table 1, OBP explains over 70 percent of the variance in team winning percentage. Even though winning games is a function of many variables, a team's ability to get on base is a large determinant of outcomes. When I introduce SLG into the model, the model is able to explain even more of the variation. The third column of Table 1 contains important information about the relative contribution to team winning percentage of OBP and SLG. The coefficient on OBP (1.742) is roughly twice that of SLG (.924). This 2:1 ratio means that a one-unit increase in OBP contributes twice as much to winning than does a one-unit increase in SLG. Thus an efficient labor market would value these statistics accordingly.

Table 2: Dependent Variable is Team Winning Percentage

	2000-2002	**2003-2007**	**2008-2012**	**2013-2016**
On-Base	2.052***	1.770***	1.620***	1.530***
	(11.36)	(12.65)	(8.23)	(7.27)
Slugging	0.790***	0.916***	1.067***	0.932***
	(6.31)	(8.72)	(7.72)	(5.79)
Constant	0.427***	0.424***	0.422***	0.434***
	(80.41)	(88.51)	(75.42)	(70.89)
N	90	150	150	120
R^2	0.873	0.821	0.780	0.747

Notes: Coefficients were estimated using ordinary least squares. T-statistics are in parentheses. Asterisks represent significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 2 shows this same model estimated over different time periods. Using data up until 2012, Baumer and Zimbalist (2014) find a slight decrease in the coefficient of OBP over time, similar to the trend seen in Table 2. They argue that the decrease in the coefficient of OBP is not significant and do not attribute this decrease to any particular factors. I find the coefficient on OBP to continue to steadily decline from 2013 to 2016. This continuing decline may be evidence that teams are changing their strategies in order to adapt to the new pitching-dominant era of baseball. That is, since fast-throwing pitchers give teams fewer opportunities to score, power hitters and the extra-base hits they produce may be more beneficial to teams relative to crafty players who find less powerful ways to get on base.

Regardless of the reason for the decrease in the OBP coefficient, it is important to note the relative proportions of the coefficients on OBP and SLG. In the 2000-2002 period, the ratio of OBP to SLG is roughly 3:1. From 2003-2007, it is roughly 2:1. After 2007, the ratio is about 3:2. We will see in the next model how the labor market valuations of these statistics changed over time.

Table 3: Dependent Variable is Natural Log of Player Salaries (USD)

	1995-1997	**1998-2002**	**2003-2007**	**2008-2016**
On-Base	3.669***	3.538***	4.912***	3.544***
	(3.22)	(4.08)	(4.65)	(4.06)
Slugging	2.668***	2.490***	2.491***	2.004***
	(4.92)	(5.93)	(4.66)	(4.41)
Plate	0.004***	0.003***	0.002***	0.002***
	(15.12)	(18.49)	(12.67)	(13.58)
Catcher	0.152	0.197**	0.0185	-0.00482
	(1.22)	(2.12)	(0.18)	(-0.05)
Infielder	0.0914	-0.101*	-0.272***	0.0295
	(1.23)	(-1.81)	(-4.33)	(0.58)
Constant	9.914***	10.361***	11.146***	11.588 ***
	(21.56)	(39.11)	(31.54)	(45.87)
N	755	1,391	1,305	2,229
R^2	0.477	0.449	0.361	0.253

Notes: Coefficients were estimated using ordinary least squares. T-statistics are in parentheses. Asterisks represent significance levels: *** p<0.01, ** p<0.05, * p<0.1.

Table 3 shows the regression with player salaries as the dependent variable. I find that the labor market was, indeed, inefficient in the periods leading up to *Moneyball.* As evident from the player-productivity model, OBP should be valued three times as much as SLG prior to 2003. However, the relative valuation of OBP to SLG from 1998 to 2002 was roughly 3.5 to 2.5. In contrast, from 2003 to 2007, the ratio of the productivity of OBP to SLG (2:1) matches with the relative valuations. During this period, which immediately follows the publication of *Moneyball,* the labor market corrected and was no longer inefficient. This correction is evident in the increased valuation of OBP from 3.5 in the 1998-2002 period to 4.9 in the 2003-2007 period, a 40 percent increase. In the following period (2008-2016), the valuation of OBP decreases, and the ratio of the valuations becomes just over 3:2. Baumer and Zimbalist (2014) also detect this trend, and cite the decreasing value of the coefficient on OBP as evidence that the market overcorrected. This overcorrection may well have been the case because many teams started copying what the A's did in 2002, and looked for players with high OBP.

In addition to a market correction, other factors may explain a decrease in the valuation of OBP. A widespread complaint after *Moneyball* was that teams were adopting a "small-ball" approach, which many fans view as a less exciting style of play. Players can generate team revenue through ways other than strict offensive performance. For example, a superstar player, both talented and charismatic, draws fans to the ballpark and thus increases ticket revenue. Likewise, a slugger who hits a lot of home runs, which is arguably the most exciting play in baseball, increases attendance and the sale of merchandise such as player jerseys. Thus, fan preferences and aesthetics may explain the relative decrease in the valuation of OBP to SLG in the 2008-2016 period.

The coefficients on the catcher and infielder dummy variables are all small in absolute value. In addition, the dummy variables are statistically significant in only some time periods. These two aspects suggest that a player's offensive performance dictates salary much more than defensive performance.

Figure 1

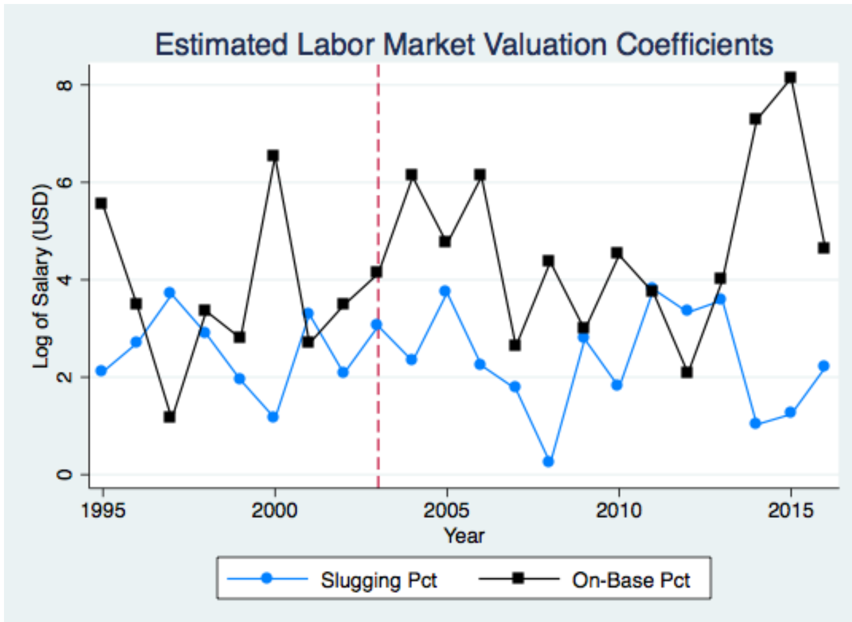


Figure 1 plots the value of the coefficients from Table 3 over time. Although there is a lot of variation across the period, there is nevertheless a clear spike in the value of OBP in 2004, the year following the publication of *Moneyball*. Although this spike is consistent with previous research, the value of the coefficients in Figure 1 continues to fluctuate a lot after 2003. This fluctuation was not expected. Hakes and Sauer (2006) and Baumer and Zimbalist (2014) do not find nearly as much variation in the value of these coefficients over time. They find more convincing evidence that OBP was undervalued pre-2003 and that the market adjusted post-2003. This difference in results can probably be attributed to the lack of information on arbitration-eligible players and free agents in my dataset, or perhaps to my observation about a possible stylistic backlash against the small-ball revolution. Nonetheless, the results show that in the period immediately after 2003 there was a sharp rise in the valuation of OBP, confirming the inefficiency in the MLB labor market.

b. Competitive Balance

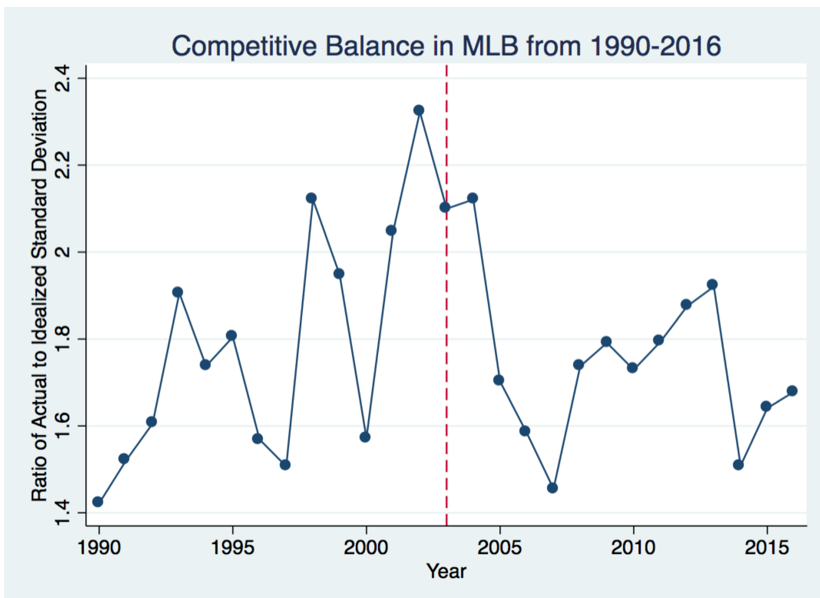
At its core, the concept of competitive balance is concerned with maintaining parity within sports leagues. The amount of revenue a team makes plays a large role in how competitive it is in a league. One of the main factors influencing team revenue is the specific geographic market in which the team is located. Teams located in large metropolitan areas earn more revenue, ceteris paribus, relative to teams located in smaller markets, simply due to the ability of large-market teams to attract more fans and thus sell more tickets, merchandise and TV advertising.

To address the advantages of large-market teams, professional sports leagues have introduced many institutional mechanisms to maintain competitiveness. For example, revenue-sharing is a common way to mitigate the inherent advantage large-market teams have over small-market teams. Additionally, reverse-order drafts give teams who perform poorly the opportunity to sign talented young players and perform better the next season. *Moneyball* spread innovative ideas throughout the league, helping teams to accurately measure player talent and productivity. So, it is important to ask the question of how the equality of information changed the competitive balance in the MLB.

Competitive balance in a sports league is crucial to the league's success for several reasons. For one, fans are simply more attracted to competitive duels. Maintaining a high level of competition in the league ensures that a fan's demand for the product stays high. It is important for teams to be given a reasonable chance to win each season. Otherwise, the sport becomes far less interesting for fans, players and teams alike.

Although competitive balance can be measured in a variety of ways, this paper focuses on the competitiveness of leagues within a given season. As such, I calculate the ratio between the standard deviation of win percentages among teams in a given year and the ideal standard deviation, which represents perfect parity. A ratio of 1 would represent perfect balance, since that would imply the actual standard deviation would be equal to the idealized standard deviation of winning percentages. Figure 2 shows this ratio from 1990 to 2016.

Figure 2



As shown, there is a sharp improvement in competitive balance following 2003. Even though the competitive balance in the league fluctuates a lot between 1990 and 2016, the results from Figure 2 support my hypothesis. As more and more teams adopted the sabermetric ideas and implemented them in their team strategy, the league became more competitive. After the release of *Moneyball,* many teams began to establish and expand a division of scouting purely devoted to advanced analytics, which was rarely seen before *Moneyball.* Although teams may have different strategies and interpret data differently, the overarching theme that rigorous statistical analysis was useful for scouting spread throughout the league. Thus, *Moneyball* provided all teams in the league access to techniques that could improve their performance. Yet as teams became more familiar with these methods, competitive balance declined, as the advantage of any one team to use sabermetrics diminished. This diminishing advantage may be why we see competitive balance begin to worsen in 2007, four years after the publication of *Moneyball.* As teams continued to develop scouting strategies and became competent with sabermetrics, large-market teams began to pull away from the pack.

VI. Conclusion

This paper finds evidence in support of Lewis's claim that the MLB labor market was inefficient during the A's successful run in the early 2000s. Following the publication of *Moneyball*, the valuation of a player's skill to get on base appreciated significantly, and may have even become overvalued. The market continued to adjust in the following years, which is consistent with the findings in previous research. In my model, the valuation of OBP and SLG in the MLB labor market varied greatly, and does not always align with the literature's prevailing theory. Limitations on player data could be a large reason for this discrepancy. Obtaining more data to more accurately model player compensation would most likely improve the explanatory power of my model, and enable me to make stronger claims in support of the central hypothesis that OBP was undervalued in the labor market. Alternatively, future research might seek to develop an econometric model to test my other observation: whether a stylistic rejection of "small ball" or a new pitching-dominant era led front offices to re-value power hitters at the expense of efficient hitters.

Additionally, I test the hypothesis that *Moneyball* improved competitive balance in the MLB, which is a departure from previous research. Although the competitive balance varies among seasons, I find a general trend towards improved competitiveness in the years immediately following 2003. In future research, it might be helpful to test this hypothesis in a more rigorous manner. Moreover, I am curious to see if the publication of *Moneyball* and the resulting improvement in both player valuation and league competitiveness also led to a stronger correlation between team payrolls and winning percentages. That is, can econometric analysis show that with a more efficient post-*Moneyball* labor market, MLB teams that spent more money on players actually earned a higher place in the standings?

References

Baumer, B. & Andrew Z. (2014). Quantifying Market Inefficiencies in the Baseball Players' Market. *Eastern Economic Journal,* 40(4): 488-498.

Bradbury, J. C. (2013). What is Right with Scully Estimates of a Player's Marginal Revenue Product. *Journal of Sports Economics,* 14(1): 87-96.

Butler, M. R. (1995). Competitive Balance in Major League Balance. *The American Economist,* 39(2): 46-52.

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance,* 25(2): 383-417.

Hakes, J. K. & Raymond D. S. (2006). An Economic Evaluation of the *Moneyball* Hypothesis. *Journal of Economic Perspectives*, 20(3):173-186.

Hakes, J. K. & R. D. Sauer. (2007). The *Moneyball* Anomaly and Payroll Efficiency: A Further Investigation. *International Journal of Sport Finance*, 2:177-189.

Krautmann, A. C. (1999). What's Wrong with Scully-Estimates of a Player's Marginal Revenue Product. *Western Economic Association International,* 37(2): 369-381.

Lahman, S. (2013). *Sean Lahman's Baseball Database*. Retrieved from http://www.seanlahman.com/baseball-archive/statistics/

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*. New York: WW Norton & Company.

Roach, M. A. (2017). Testing Labor Market Efficiency Across Position Groups in the NFL. *Journal of Sports Economics*, 1-29.

Scully, G. W. (1974). Pay and Performance in Major League Baseball. *The American Economic Review*, 64(6): 915-930.

Zimbalist, A. S. (2002). Competitive Balance in Sports Leagues. *Journal of Sports Economics*, 3(2): 111-121.

Overconfidence and Excess Entry in Entrepreneurship

Nana Adom Mills-Robertson, Qian Wang

Behavioral and Experimental Economics

Abstract

Entrepreneurs excessively enter the market and persist in their businesses despite high rates of failure and low rates of return. To investigate this phenomenon, we employ an interactive experiment to simulate market entry decisions. We have three major findings. First, overconfidence increases market entry. Second, overconfidence reduces market profit. Third, knowing more information about the market increases market profit by reducing overconfidence and excess entry.

I. Introduction

Frank Knight (1921) argue that entrepreneurship should not be simply conceived as investment under risk with a known distribution of returns because market entry would eliminate any profit if such information can be objectively learned. Rather, he argued that entrepreneurs face highly uncertain returns and are more skilled than others in perceiving profitable opportunities. Knight (1921) thereby ushered in a new field that studies what qualities enable entrepreneurs to earn above average market profits. Despite these arguments, empirical studies of entrepreneurial profits often find the opposite. Hall and Woodward (2010) find that 75 percent of all startups deliver zero exit value, while only a tiny fraction of entrepreneurs receive more than $100 million. This inflates the average exit value to $5.8 million. Given the highly skewed returns of startups, Hall and Woodward (2010) imputed that the expected utility of starting a business for individuals with normal risk aversion is negative. Nevertheless, over 500,000 individuals in the United States start firms with at least one employee every year according to Parker (2009). As a result, those who constantly generate losses persist in running businesses for a suboptimally long time (Hamilton, 2000).

Our paper aims to examine the relationship between oversaturation in the entrepreneur market and profits. This is achieved through an interactive experiment that simulates market entry decisions. We argue that entrepreneurs excessively enter the market because they overestimate their relative skills among their competitors and therefore inflate their expected profit. This phenomenon, which we'll refer to as overplacement, is well established in psychology literatures (Neil D. Weinstein, 1980; Shelly E. Taylor and J.D. Brown, 1988).

Our paper proceeds as the following. In section II, we summarize three major theories that explain excess entry in the entrepreneur market. In section III, we elaborate on our experimental design. In section IV, we describe the data collected. In section V, we introduce the specifications we use to test our hypothesis. In section VI, we discuss our regression results. In section VII, we present the implications of our study. In section VIII, we conclude with some final remarks.

II. Three Major Theories on Entrepreneur Excess Entry

Although our argument exclusively focuses on overplacement, a subcategory of overconfidence, it helps to contextualize our theory within the broader literature. Firstly we will discuss the three major entrepreneur motive theories: expected utility theory, overconfidence theory and non-pecuniary reward theory. Then, we will elaborate on overconfidence theories and one of it subcategories, overplacement, which is the focus of our paper.

a. Overview of three entrepreneur motive theories

Given the same distribution of returns, expected utility theories dictate that individuals who are more risk seeking have greater certainty equivalents of risky outcomes and thus are more likely to gamble. Given the extremely skewed distribution of startup exit values, Hall and Woodward (2010) calculate that for a risk-neutral individual, the certainty equivalent is $5.8 million. With mild risk aversion, the amount is only $0.6 million and with normal risk aversion, the certainty equivalent is slightly negative. Thus, Hall and Woodward (2010) conclude that risk seeking individuals perceive greater values in risky investments and are therefore more likely to become entrepreneurs. This argument is further corroborated by Astebro, Herz, Nanda and Weber (2014) who compare the earnings distribution of self-employed entrepreneurs and employees using income data from the Danish Labor Market Research database. The persistent higher income variance of the self-employed indicates that entrepreneurs are more risk tolerant than employees. Although theoretically tenable, the relationship between entrepreneurship and risk attitudes has yielded conflicting empirical results. (Parker, 2009, Ahn, 2010).

Overconfidence, an alternative entrepreneurial motive theory, argues that individuals enter the entrepreneur market because they evaluate the return distributions of their projects more favorably than is objectively the case. Surowiecki (2014) argues that individuals who become entrepreneurs are not more risk-seeking than everyone else. Rather, they are more confident about their profit-generating skills. Further, because entrepreneurs are “incurably optimistic” about their abilities, they always enter the market excessively, thereby driving down the market profit. Cooper, Woo and Dunkelberg (1988) substantiate the overconfidence theory with empirical evidences by asking 3,000 entrepreneurs the simple question: “What are the odds of your business succeeding?” 80% of the respondents answered greater than 70%; 33% predicted a 100% chance of success, while in reality 75% of startups fail within 3 years.

Non pecuniary utility theory denotes that entrepreneurs create a business not as a means to obtain monetary profits, but as an end in itself. One aspect of the non-pecuniary theory is the founder’s dilemma, whereby entrepreneurs maintain control over the business they created at the expense of reduced management efficiency and lower monetary profits according to Wasserman (2008). Additionally, Wasserman (2008) finds that entrepreneur founders persistently receive 20% less in cash compensation than non-founders who performed similar roles, implying that entrepreneurs derive non-pecuniary utilities from creating and controlling a business.

b. Overconfidence theories and overplacement

Overconfidence, generally speaking, happens when individuals rate their chances of success too favorably. However, it has 3 distinct psychological origins: overestimation, overplacement and overprecision (Moore and Healy, 2008). Overestimation refers to overrating one's actual ability, performance, level of control or chance of success. Overplacement refers to rating one's ability above others'. For example, people often rate themselves above the average, without noticing that others share the same belief. Overprecision refers to excessive certainty about the accuracy of one's beliefs.

Our paper exclusively focuses on overplacement, the type of overconfidence whereby people rate themselves more favorably than others within their peer group. When asked to assess how they perform among the general population on positive traits such as earnings prospects, longevity and driving ability, a majority of people rate themselves above the median, whereas, theoretically, only half of the population can achieve this (Svenson, 1981). The connection between entrepreneurship and overplacement is well documented in previous literatures. When asked two questions, "What are the odds of your business succeeding?" and "What are the odds of any business like yours succeeding?", entrepreneurs rated an 81% chance of success for themselves on average but only 59% for other businesses like their own (Cooper, Woo, Dunkelberg 1988). Camerer and Lovallo (1999) implement an experiment to test the relationship between overplacement and market entry. They find that participants enter the market more frequently when the payoff is based on their skills among their peers rather than pure luck.

III. Experimental Design

We model our experimental design after Camerer and Lovallo (1999). In this section, we first list the steps of our experimental design. Then, we highlight three key differences between our design and that of Camerer and Lovallo (1999) and justify our changes.

a. Assumptions

Our experimental design has four key assumptions. First, aggregate market profit is maximized when the number of entrants is equal to the market capacity. If the number of entrants is lower than the capacity, entrants can't capture the entire market. If the number of entrants exceeds the capacity, excess entrants generate losses and the aggregate market profit declines. Second, the more skilled an entrepreneur is relative to his or her peers, the larger profit he or she earns. Third, entrepreneurs receive market feedbacks and can adjust their decisions subsequently. Fourth, individuals don't know *objectively* how skilled they are among their peers, but rather base their market entry decisions on how they *subjectively* rank themselves among their peers. This assumption allows overconfidence to skew entry decisions.

The first two assumptions produce a payoff table shown below in Table 1. The first row denotes the market capacity and the first column denotes the skill rank of an individual. The first assumption is satisfied by setting the maximum market profit to $20 and the excess entry profit to negative $10. The second assumption is satisfied by assigning higher profits to higher ranked individuals.

Table 1. Payoff Table

Rank\Capacity	1	2	3	4	5	6	7	8
1	20	13	10	8	7	6	5	4
2	-10	7	7	6	5	5	4	4
3	-10	-10	3	4	4	4	4	3
4	-10	-10	-10	2	3	3	3	3
5	-10	-10	-10	-10	1	2	2	2
6	-10	-10	-10	-10	-10	1	1	2
7	-10	-10	-10	-10	-10	-10	1	1
8	-10	-10	-10	-10	-10	-10	-10	1

b. Steps of the experiment

Now we would elaborate on each step of our experiment:

1. After consenting, a group of 8-10 experiment participants each receive an anonymous ID, the experiment instruction (Appendix A) and the experiment answer sheet (Appendix B). The paper instruction is accompanied by a 5-minutes lecture instruction. Each participant has to finish a comprehension quiz before proceeding (Appendix C). The best performer of the experiment will receive $20 in cash. The incentive is communicated to all participants.
2. The experiment has two treatments: the random treatment and the skill treatment. In the random treatment, participants are randomly assigned a rank using a random sequence

generator. In the skill treatment, participants' scores in a 5-minute quiz determine their rank. In both treatments, the rank is *not* disclosed until the end of the experiment. This satisfies the fourth assumption that participants don't know their objective ranks among their peers.

3. The experiment has 16 rounds (8 for each treatment). Each round is exactly the same: First, the market capacity is announced. Second, each participant privately forecast how many people will enter the market. Each correct forecast is rewarded with $2 hypothetical earning. Incorrect forecast receives $0. Third, each participant privately chooses to enter or not enter the market. Payoffs for entrants are determined by Table 1.[7] Everyone else receives $0. Fourth, participants are asked to close their eyes and raise their hands if they entered the market. Finally, the researcher counts the number of entrants of the round and discloses the number to all participants.
4. The skill treatment is exactly the same as the random treatment, except that the ranks are determined by participants' score in the quiz rather than randomly assigned. For experiment 1-4, participants do the quiz after the skill treatment. For experiment 5-8, participants do the quiz before the skill treatment. The quiz is a 12 questions logic test.[8] (Appendix D)

Three points are worth noting: First, the rank of each participant remains the same throughout each treatment. Second, the only feedback from each round is the number of entrants.

---

[7] The rank in the payoff table is not the absolute rank assigned to the participants in random or skill treatments but relative to other entrants. For example, if only rank 3 and rank 6 enter the market, rank 3 corresponds to rank 1 and rank 6 corresponds to rank 2 in the payoff table.

[8] The quiz is composed of 12 questions chosen from the "Mensa Workout", a 30-questions problem set designed by Mensa International, the largest high IQ society in the world. The original test has a time limit of 30 minutes for 30 questions whereas we scaled up the difficulty by condensing it to 5 minutes for 12 questions, to differentiate the participants' performance.

Third, the only information participants can use to make entry decisions are: the market capacity, the number of entrants and the subjective self-assessment of their ranks.

c. Key changes to Camerer and Lovallo (1999)

Our experiment design differs from Camerer and Lovallo (1999) in three aspects:

First, we improve the incentive structure. Whereas Camerer and Lovallo (1999) award each participant the amount he/she earned in one randomly chosen round, we only award the participant with the highest cumulative earnings. Therefore, participants are incentivized to perform well throughout the game. Besides, since winning the reward requires outcompeting all other people in the group, participants are incentivized to assess their ranks more accurately among their peers.

Second, we measure the participants' general overconfidence. Whereas Camerer and Lovallo (1999) inform the participants about the type of quiz (trivia questions or logic puzzles) and provide them sample questions before the experiment, we intentionally did neither. Therefore, we measure the participants' general overconfidence rather than their overconfidence in trivia questions and logic puzzles.

Third, we added another variable to Camerer and Lovallo's (1999) design. That is, experiments 1-4 take the quiz *before* the skill treatment whereas experiments 5-8 take the quiz *after* the skill treatment. Note that in *both* arrangements the skill rank is *not* disclosed to the participants till the end of the experiment. By adding this variable, we test whether participants' knowledge about their absolute performance in the quiz *prior* to the skill treatment increases or decreases market entry. Or in other words, we test whether entrepreneurs' knowledge about their absolute performance in the market enhances or diminishes their overconfidence. We hypothesized that this can go both directions. An easy quiz might enhance overconfidence and increase market entry. A hard quiz might diminish overconfidence and decrease market entry.

IV. Data

Table 2. Description of Experiments

Experiment #	Sample Size (n)	Experiment Order
1	9	R-S-Q
2	8	R-S-Q
3	8	R-S-Q
4	8	R-S-Q
5	10	R-Q-S
6	8	R-Q-S
7	8	R-Q-S
8	7	R-Q-S

Note: In the experiment order column, R stands for random treatment, S stands for skill treatment and Q stands for quiz.

Table 2 summarizes the data we collected. We conducted 8 experiments in total. Sample sizes are listed in the second column. The third column denotes the treatment order. We always conducted the random treatment before the skill treatment and the quiz. In experiment 1-4, the participants did the quiz after the skill treatment. In experiment 5-8, the participants did the quiz before the skill treatment.

The total number of participants is 66, all of whom are Macalester College undergraduates. A quarter volunteered to participate and the rest three quarters were randomly invited by the researchers. Although subtle self-selection biases might exist, we believe that those who volunteered or accepted our invitation didn't exhibit significantly higher or lower degrees of overconfidence than other undergraduates. However, our research can't address the bias that college undergraduates might behave differently from the general population.

V. Specification

The key question of our study is whether overconfidence increases market entry. In other words, whether participants enter the market more frequently when the earnings are based on their skill rank rather than a random assigned rank. Here, the random treatment serves as a control group that measures individuals' entry decisions when the expected random rank is the group average. If overconfidence exists, individuals will expect their skill ranks to be higher than their random ranks and thus enter the market more frequently in the skill treatment. The specification that tests this hypothesis is:

$$\textbf{Entry decision}_{itj} = \beta_0 + \beta_1\, \textbf{capacity}_{it} + \beta_2\, \textbf{expected profit}_{itj} + \beta_3\, \textbf{skill}_{it} + \beta_4\, \textbf{QS}_i + \beta_5\, \textbf{skill}_{it} * \textbf{QS}_i + \gamma\, \textbf{round}_t + \varepsilon$$

Where $entry\ decision_{itj}$ is a dummy variable that equals to 1 if subject j enters the market in round t of experiment i and equals to 0 if the subject doesn't enter the market. $capacity_{it}$ is the market capacity in round t of experiment i. $expected\ profit_{itj}$ is subject j's expected profit for round t of experiment i, calculated from their forecast according to Table 1. For example, if $capacity_{it} = 2$ and $forecast_{itj} = 5$, then $expected\ profit_{itj} = 13 + 7 -10 -10 -10 = -10$. $skill_{it}$ is a dummy variable that equals1 if round t of experiment i is in the skill treatment and equals to 0 if it's in the random treatment. QS_i is a dummy variable that equals 1 if the quiz is before the skill treatment in experiment i and equals to 0 if the quiz is after the skill treatment. $skill_{it}*QS_i$ is the interaction term between $skill_{it}$ and QS_i. $round_t$ is a vector of dummy variables that equals 1 for round t and equals to 0 otherwise. It's included in the specification to control for round fixed effects. γ is a vector of coefficients for $round_t$. The table below lists the expected sign of each variable and their justifications:

Table 3. Expected Signs and Justifications for Variables

Dependent Variable: Entry Decision		
Variable Name	**Expected Sign**	**Justification**
capacity	positive	Higher market capacity allows more people to enter the market with a positive return and thus induces market entry.
expected profit	positive	The more profit people expect, the more likely they enter the market.
skill	positive	If overconfidence exists, people will enter the market more when the earnings are based on their skill ranks.
QS	indeterminate	Taking the quiz before the skill treatment either increases or decreases market entry depending on the difficulty of the quiz.
skill*QS	indeterminate	Same as above
round	N/A	Round is a vector of dummy variables that controls for round fixed effect, irrelevant to our theory.

VI. Results

a. Does overconfidence increase market entry?

Since *entry decision*$_{itj}$ is a dummy dependent variable, we use the logit maximum likelihood estimator (MLE). Table 4 shows the coefficients and t-statistics of the 3 models we estimated. Significant coefficients are marked by asterisks.

Table 4. Logit Estimation of Entry Equation (Experiment 1-8)

Dependent Variable: Entry Decision (= 1)			
Variables	**Model 1**	**Model 2**	**Model 3**
capacity	0.471*** (7.95)	0.470*** (7.95)	0.467*** (7.95)
expected profit	-0.015** (-2.38)	-0.015** (-2.44)	-0.016*** (-2.66)
skill	0.594*** (2.99)	0.691*** (3.96)	0.350** (2.57)
QS	-0.195 (-1.03)		
skill*QS	-0.462* (-1.69)	-0.655*** (-3.30)	
observations	1,056	1,056	1,056
log-likelihood	-622.8	-623.3	-628.8
percent correct	69.2%	68.9%	68.8%
chi-square	147.5	146.5	135.4
p-value of X^2	0.0000	0.0000	0.0000

T-statistics are in parentheses. ***p<0.01, **p<0.05, *p<0.1

The coefficient of the skill treatment dummy (*skill*) is positive and significant at 0.01 across all three models, implying that people enter the market more frequently when the earnings are based on their skill ranks. Overconfidence *does* increase market entry.

On whether taking the quiz before the skill treatment *(QS=1)* increases or decreases market entry, the coefficients of the *QS* dummy and the *skill*QS* interaction dummy are negative but *not* significant in Model 1. The *QS* dummy, however, doesn't make theoretical sense: it is equal to one for *both* treatments, and thus models the effect of taking the quiz before the skill treatment on *both* treatments. The quiz, however, can't affect entry decisions in the random treatment, which is always before the quiz. Therefore, in Model 2, we dropped *QS* and *skill*QS* became negative and significant at 0.01 level, indicating that taking the quiz before the skill treatment decreases market entry. This implies that knowing more market information before making entry decisions dampens overconfidence.

To test whether *skill*QS* is necessary in our specification, we dropped it in Model 3. The coefficient of the skill treatment dummy (*skill*) dropped from 0.691 to 0.350 and lost its significance at the 0.01 level. Therefore, omitting *skill*QS* pushes the overconfidence-dampening effect to the skill treatment dummy, reducing both its coefficient and significance. This evidence suggests that *skill*QS* is a necessary variable and that taking the quiz before the skill treatment significantly reduces overconfidence.

Another noteworthy result is that the coefficient of the expected profit is negative and significant in all three models, contradicting our expectation that higher expected profit increases market entry. Camerer and Lovallo (1999) obtained the same results. Currently, we can't rigorously test why this happened. Camerer and Lovallo (1999) speculated that reverse causality might be the reason: "when subjects plan to enter, they also forecast a lot of entry, so the expected profit is lower when they enter".

b. Blind spot hypothesis: do people expect more profit in the skill treatment rather than being overconfident?

An alternative interpretation of more frequent market entry in the skill treatment is the blind spot hypothesis, which states that people enter the market not because they are overconfident about their skills, but because they expect more profit from the market. In other words, people are "blind", or misinformed about the market in their expectations.

To test this hypothesis, we calculated each individual's average expected profit in the random rounds and in the skill rounds, according to their forecast and the payoff table. If the blind spot hypothesis is true, we would observe higher expected profit in the skill rounds, where people enter the market more often.

Table 5. Average Difference in Expected Profits per Entrant per Round Between Random and Skill Treatments for Experiment 1 – 8

Measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
π_r-π_s	6.34	7.81	1.57	0.88	0.47	0.39	3.97	-1.85
# of subjects with π_r-π_s>0	8/9	7/8	6/8	3/8	4/10	6/8	5/8	3/7
percent	89%	88%	75%	38%	40%	75%	63%	43%
# of subjects with π_s<0	1/9	5/8	0/8	3/8	6/10	1/8	3/8	0/7
percent	11%	63%	0%	38%	60%	13%	38%	0%

In Table 5, we used three measures to compare the expected profit in random and skill rounds. In the first row, π_r-π_s denotes the average difference in expected profits per entrant per round between random and skill treatments. In 7 out of 8 experiments, π_r-π_s is positive, indicating that people on average expect more profits in the random rounds. This contradicts the blind spot hypothesis. In the second and third row, we calculated the number and percent of individuals who expect more profits in the random rounds. In 6 out of 8 experiments, the percentage is above 50%, contradicting the blind spot hypothesis. In the fourth and fifth row, we calculated the number and percent of individuals who on average expect negative profits in the skill rounds. 19 out of 66 participants expected negative profits in the skill rounds.

These three measures suggest that people expect *less* profit in the skill rounds but still choose to enter the market more frequently, thereby rejecting the blind spot hypothesis and substantiating the overconfidence effect. However, overconfidence is not strong enough such that most people enter the market more frequently with negative expected profits in the skill rounds.

c. Does overconfidence reduce market profit?

Another important phenomenon to address is whether overconfidence and greater entry in the skill rounds reduce market profit, defined as the sum of individual profits in each round. Greater entry doesn't necessarily reduce market profit. Recall that according to the payoff table, if entry is below market capacity, market profit increases as more people enter. If entry is above market capacity, market profit decreases as more people enter.

In table 6, we calculated the market profit of all 128 rounds (64 random, 64 skill) and aggregated the market profit of each experiment. In the random treatment, 41 out of 64 rounds (64%) have positive market profit but only 9 out of 64 rounds (14%) have negative market profit.

In contrast, in the skill treatment, only 24 out of 64 rounds (38%) have positive market profit but 21 out of 64 rounds (33%) have negative profit. The average market profit of the random rounds is $7.63 whereas that of the skill rounds is only $0.66. Considering the maximum market profit of $20, this difference of $6.97 is pretty large. Overconfidence and greater entry in the skill rounds reduce market profit.

Table 6--Market Profit by Round

Profit for random-rank treatment

Experiment	Sample Size	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	total
1	9	10	10	0	10	10	10	20	0	70
2	8	-10	10	10	10	-10	0	-10	-10	-10
3	8	20	10	10	0	19	20	20	20	119
4	8	0	20	20	10	-20	-10	10	-10	20
5	10	0	0	10	-10	0	0	0	19	19
6	8	10	13	0	10	10	0	0	20	63
7	8	20	-10	0	10	10	10	20	18	78
8	7	20	10	20	10	20	20	19	10	129

Profit for skill-rank treatment

Experiment	Sample Size	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8	total
1	9	0	-30	-10	0	0	-20	0	0	-60
2	8	18	-10	20	-20	0	10	-10	-10	-2
3	8	0	10	-20	13	0	-20	10	-10	-17
4	8	-20	0	10	-10	0	-10	0	-10	-40
5	10	-20	0	-10	0	10	-10	20	17	7
6	8	-20	10	20	-10	10	20	-20	10	20
7	8	0	0	10	0	10	0	-10	0	10
8	7	14	20	17	13	20	0	19	20	124

Alternatively, to formally test whether skill rounds has significantly lower profits than the random rounds, we conducted a *matched pairs test*. First, we paired up rounds with identical market capacity, location in the treatment and quiz/skill order, *only* differing in whether the round is in the random or skill treatment. Thus, we controlled for most differences between the two rounds in each pair. Second, we calculated the differences between the random round and the skill round in each pair. Third, we divided the mean of the differences by the standard error of the differences, to yield the t-statistic of the difference. The 64 pairs we matched yielded a skill round minus random round market profit difference t-statistic of -3.55, significant at 0.001 level. The matched pairs test suggests that overconfidence and excess entry reduce market profit in the skill rounds.

Furthermore, we tested whether taking the quiz before the skill treatment increases the market profit. We conducted another matched pairs test that matches rounds with identical treatment, market capacity and location in the treatment while only differing in the quiz/skill order. The 64 pairs we matched yielded a quiz/skill minus skill/quiz market profit difference t-statistic of 2.84, significant at 0.01 level. This suggests that knowing more market information before making entry decisions increases the market profit by dampening the overconfidence effect.

VII. Discussion

a. "Inside View" v.s. "Outside View"

We found that excess entry is caused by overconfidence, especially overplacement, whose psychological origin merits discussion. Kahneman and Lovallo (1993) argue that conflicts between the "inside view" and the "outside view" induce overplacement. Whereas the "inside view” induces people to subjectively perceive familiar subjects as unique and superior, the “outside view” enables people to objectively evaluate unfamiliar subjects by comparing it to similar subjects. In the context of startups, entrepreneurs are biased by the “inside view”. They attach too much uniqueness and superiority to the opportunities they identify and the businesses they create, thereby neglecting the outcomes of past similar businesses that would have enabled them to forecast their chance of success more accurately (Koellinger, Minniti, Schade 2007). In other words, the subjective story the “inside view” tells blinds the objective statistic the “outside view” observes (Camerer and Lovallo 1990).

Therefore, the “inside view” causes reference group neglect. It happens when people are overconfident about their skills without noticing that others share the same belief, thereby entering the market excessively with false expectations.

B. Overconfidence v.s. Underconfidence

We found that taking the quiz before the skill treatment reduces overconfidence and decreases market entry. However, this overconfidence-dampening effect might depend on the difficulty of the quiz. To differentiate performances, we deliberately designed a hard quiz. None of the 66 participants finished the quiz and the average score was 54%. Had we designed a easier quiz, we might have observed a weaker overconfidence dampening effect and even enhanced overconfidence.

This relationship between difficult tasks and underplacement is well documented: people routinely rate themselves below average on hard tasks such as juggling or unicycle riding (Moore and Healy, 2008). The greater the perceived difficulty of a task is, the bigger the effect of underplacement is.

This implies that the perception of entrepreneurship as an easy or hard task may lead people to underplace or overplace their skills. Therefore, future researches can dive into the relationship between entry rates and the perceived difficulty of starting a business in different industries. Entrepreneurs might overenter in reputedly difficult industries while underenter in reputedly easy industries, implying that market inefficiency can happen in both directions.

c. Is entrepreneur excess entry bad?

The high rate of failure, low rate of return and cognitive limitations such as overconfidence, “inside view” and reference group neglect all imply that entrepreneur excess entry is bad. However, is it necessarily detrimental to the society?

Entrepreneur excess entry harms society by driving down social returns. Unsuccessful ventures incur negative externalities by forcing stakeholders such as the government to absorb their failure. Peter Thiel, a prolific investor and entrepreneur who created PayPal, argues against entrepreneur excess entry in zero-sum industries (Thiel 2014). These industries are either declining or have a clear winner/incumbent that dominates. For example, excess entry into the newspaper industry is not beneficial due to its declining demand. Additionally, excess entry into the online search engine industry is a waste of resources because incumbents such as Google already dominates the market and possesses the best technology. However, entrepreneur excess entry and business failures accumulate valuable knowledge that improves future businesses. Besides, new entrants in a market improve efficiency by pressuring the incumbents to evolve and improve productivity (Koellinger, Minniti, Schade 2007).

VIII. Conclusion

To investigate the phenomenon that entrepreneurs excessively enter the market and persist in their businesses despite high rates of failure and low rates of return, we designed an experiment to simulate market entry decisions. We have three major findings. First, overconfidence, particularly overplacement, increases market entry. Participants enter the market more frequently when the earnings are based on their skill ranks rather than randomly assigned ranks, even if they expect less profit in the skill treatment. Second, overconfidence reduces market profit. According to the matched pairs test, participants earn significantly less in the skill rounds than in the random rounds. Third, knowing more information about the market increases market profit by reducing overconfidence and excess entry. The matched pairs test shows participants who took the quiz before the skill treatment enter the market less frequently and earn greater profits.

In the discussion section, we posited the "inside view" and reference group neglect as the origin of overplacement. We hypothesized that perceived difficulty of tasks can generate both overconfidence and underconfidence. Furthermore, despite all the negative evidences, entrepreneur excess entry is not necessarily detrimental to the society.

Besides investigating the rationales of entrepreneur excess entry, our paper has further policy implications. For example, the government can reduce irrational exuberance and excess entry by sharing the "outside view" of similar past ventures with potential entrepreneurs, thereby improving the quality and success rate of new ventures.

References

Ahn, T. (2010). Attitudes toward Risk and Self-employment of Young Workers. Labour Economics 17(2): 434–442.

Astebro, T., Holger H., Ramana N., & Roberto A. W. (2014). Seeking the Roots of Entrepreneurship: Insights from Behavioral Economics. *Journal of Economic Perspectives*, 28(3): 49-70.

Camerer, C. & Dan L. (1999). Overconfidence and Excess Entry: An Experimental Approach. *American Economic Review,* 89(1): 306–18

Cooper, Arnold C., Carolyn Y. Woo, and William C. Dunkelberg. (1988). Entrepreneurs' Perceived Chances for Success. *Journal of Business Venturing,* 3(2): 97–108.

Hamilton, Barton H. (2000). Does Entrepreneurship Pay? An Empirical Analysis of the Returns to Self-Employment. *Journal of Political Economy,* 108(3): 604–31.

Knight, Frank H. 1921. *Risk, Uncertainty, and Profit*. Houghton Mifflin, Boston

Koellinger, P., Maria M., and Christian S. (2007). I think I can, I think I can: Overconfidence and Entrepreneurial Behavior. *Journal of Economic Psychology,* 28(4): 502-527

Moore, Don A., and Paul J. Healy. (2008). The Trouble with Overconfidence. *Psychological Review,* 115(2): 502–17.

Parker, Simon C. (2009). *The Economics of Entrepreneurship*. Cambridge University Press

Surowiecki, J. (2014). Epic Fails of the Startup World. *New Yorker*, May 19. http:// www.newyorker.com/talk/financial/2014/05 /19/140519ta_talk_surowiecki.

Svenson, O. (1981). Are We All Less Risky and More Skillful Than Our Fellow Drivers? *Acta Psychologica* 47(2): 143- 48

Taylor, Shelley E., & Jonathan D. Brown. (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin,* 103(2) :193-210

Thiel, P. & Blake M. (2014). *Zero to One: Notes on Startups or How to Build the Future*. New York: Crown Business.

Wasserman, N. (2008) The Founder's Dilemma. *Harvard Business Review,* February, 103–109

Weinstein, D. N. (1980). Unrealistic Optimism About Future Life Events. *Journal of Personality and Social Psychology*, 39(5): 806-820

Appendix

(A) Instructions for Experiment

Participant Information
ID Number:
Sex:
Gender:
Class Year:
Major:
Ethnicity:

Instructions

There are 8 rounds in the first treatment. You would be randomly assigned a "rank" that's not disclosed to you till the end of the experiment.

At the beginning of each round, we will announce the market capacity "c’, the number of entrants that can enter the market with positive earnings.

Then we ask you to forecast the number of people who enter the market. We award each correct forecast with $2 (hypothetical earning). We do not penalize incorrect forecasts.

After that, you can decide whether to enter the market or not.

If the total number of entrants is less or equal to the market capacity "c", everyone would get a positive earning. Higher ranked entrants would earn more than lower ranked entrants, according to the payoff table.

If the total number of entrants is greater than the market capacity "c", the highest ranked "c" entrants would get a positive earning according to the payoff table, while entrants ranked below "c" would get a negative earning.

If you choose not to enter the market, your earning is $0 for the round.

After everyone made the entry decision, we would announce the total number of entrants for the round. Note that neither your randomly assigned rank nor your earning for the round would be announced till the end of the experiment.

The second treatment has another 8 rounds. Everything is the same except that your rank is based on your performance on a logic quiz which you will do at the end of the treatment.

After the experiment, individual with the highest hypothetical earning would obtain a $20 cash prize.

(B) Answer Sheet (Same Sheet for Skill Rank as well)

Random Rank Treatment Answer Sheet

Round 1
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 2
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 3
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 4
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 5
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 6
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 7
Market Capacity is __
I forecast __ people would enter the market.
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

Round 8
Market Capacity is __
I forecast __ people would enter the market
I choose to (enter/not enter) the market. (circle one)
Actual number of entrants is __

(C) Comprehension Quiz

1. Do you know your rank when making entry decisions?
   a. Yes
   b. No

2. Do you know the market capacity when making entry decisions?
   a. Yes
   b. No

3. If the market capacity is 2 and 3 people entered the market, how much rank 1 and rank 3 respectively earns according to the payoff table?
   a. $13, -$10
   b. $7, -$10
   c. $13, $7

(D) Skill Quiz

<u>ID Number:</u>
Make sure to circle the correct answer. Write in the answer where necessary. This quiz will take 5 mins.

(1) Which number comes next in this series of numbers?
2 3 5 7 11 13 ?

- o 14
- o 15
- o 16
- o 17
- o 18

(2) There are 1200 elephants in a herd. Some have pink and green stripes, some are all pink and some are all blue. One third are pure pink. Is it true that 400 elephants are definitely blue?
- o Yes
- o No

(3) What is the following word when it is unscrambled?
H C P R A A T E U

(4) If a circle is one, how many is an octagon?

- o 2
- o 4
- o 6
- o 8
- o 12

(5) Fill in the missing number:

0,1,1,2,3,5,8,13, ___ ,34,55

(6) Which letter comes next in this series of letters?

B A C B D C E D F __ ?

(7) How many four sided figures are in this diagram?
o 10
o 16
o 25
o 28

(8) What is the number that is one half of one quarter of one tenth of 400?

o 2
o 5
o 8
o 10
o 40

(9) If it were two hours later, it would be half as long until midnight as it would be if it were an hour later. What time is it now?

o 18:30
o 20:00
o 21:00
o 22:00
o 23:30

(10) Look at the drawing. The numbers alongside each column and row are the total of the values of the symbols within each column and row. What should replace the question mark?

o 23
o 25
o 28
o 30
o 32

(11) If two typists can type two pages in two minutes, how many typists will it take to type 18 pages in six minutes?
o 3

- 4
- 6
- 12
- 36

(12) Two men, starting at the same point, walk in opposite directions for 4 meters, turn left and walk another 3 meters. What is the distance between them?

- 2m
- 6m
- 10m
- 12.5m
- 14m