THE IMPACT OF PROGRAMMING LITERACY ON LOCAL WAGE GROWTH

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master's Degreee

by

Rongtao Duan

May 2024

ABSTRACT

This study examines the impact of programming skills on local wage growth, utilizing county-level data with Stack Overflow activity as a proxy for programming skills. Findings reveal that larger counties experience significant benefits from strong programming skills, evidenced by higher wage growth. However, this effect is less pronounced in regions with high income, advanced education levels, or concentrated IT industries. This suggests that the economic advantages of programming skills are enhanced by clustering. The results highlight how local characteristics critically influence the economic utilization of programming skills.

BIOGRAPHICAL SKETCH

Rongtao Duan is currently a second-year master's student at Cornell University's Dyson School of Applied Economics and Management. She earned his Bachelor's degree in Economics from Peking University. Rongtao's research interests are centered around the digital economy, focusing on how technological advancements influence economic structures and growth.

ACKNOWLEDGEMENTS

I extend my deepest gratitude to my advisor, Professor Chris Forman, for his invaluable assistance throughout my research journey. His guidance in data analysis, research methodology, and academic writing has been crucial in the development of this thesis.

I am also grateful to Professor Calum Turvey, a committee member who participated in my defense and provided insightful feedback and constructive suggestions that significantly enhanced this work.

Special thanks to Keith Jenkins, the tech consultant at Mann Library, whose expertise in data acquisition and assistance with ArcGIS were instrumental in my research. His support made a substantial difference in handling complex datasets and visualizations.

My heartfelt appreciation goes to each of these individuals for their contributions to my academic and personal growth during this project.

TABLE OF CONTENTS

1 INTRODUCTION

A substantial body of literature highlights that advancements in information technology (IT) have the potential to significantly boost productivity and economic growth both at the national level (Jorgenson, Ho, and Stiroh 2005) and in cities (Beaudry,Doms, and Lewis 2006). Simultaneously, extensive research underscores the crucial role of human capital—specifically skills—in enhancing individual wages and overall regional productivity (Barro, 1991; Barro and Lee, 2001; Becker, 1993; Lucas,1988; Romer, 1986). Programming literacy is increasingly recognized as a critical skill in the 21st-century workforce, enhancing individual employability and contributing to broader economic development through innovation and efficiency. Despite the well-documented impact of information technology (IT) on national and urban economic scales, and its profound role in boosting regional productivity and individual wages, there remains a significant gap in the literature concerning how IT-related skills directly influence regional economic growth. This study seeks to bridge this gap by analyzing county-level data from the United States from 2011 to 2017.

The 2010s have been characterized as a decade dominated by technological advancements and the rise of tech giants. At the start of 2010, the landscape of the world's most valuable public companies was quite diverse, with PetroChina leading at a valuation of $350 billion. Among the top ten, only Microsoft and Apple represented the tech industry, with valuations of $270 billion and $189 billion respectively. This period was primarily dominated by energy companies and financial institutions . However, by the end of the decade, the scenario had shifted dramatically. As of December 2019, the top five most valuable public companies globally were all U.S. tech companies (Divine 2019). This

shift underscores the decade's significant transformation, marking the 2010s as the era of tech dominance, wherein programming and IT-related skills surged in importance and impact on local and global economic scales.Therefore, our study focuses on the programming skills present in 2011 and their impact on local wage growth from 2011 to 2017 in the U.S. This timeframe was selected to analyze the period before the pandemic, which significantly influenced work practices by encouraging remote work and leading to the relocation of many programmers. This shift could affect the geographical distribution of programming skills since workplace interactions play a crucial role in promoting knowledge among workers(Allen 1997, Roche 2020, Roche et al. 2022) and physical proximity is a complement to online activity function when adopting the technologies(Brown,Roche 2023), which may in turn affect the distribution of local wage growth.

In this study, we will use the posting activity within Stack Overflow as a proxy for programming literacy. Stack Overflow plays a pivotal role in the software development industry as a comprehensive question-and-answer forum where developers converge to share knowledge, solve problems, and exchange code that can be directly applied to their own projects. Attracting over 100 million visitors each month, it serves not only as a public platform but also as a proprietary internal tool that major companies like Microsoft, Google, and Logitech employ to facilitate technical discussions among their engineering teams. As of November 2022, there were a total of 23 million questions asked throughout the year. This breaks down to an average of roughly 4 questions posted every minute (SignHouse 2024). Stack Overflow not only serves as community-driven hubs where individuals can seek and exchange knowledge but also act as vibrant networks of practice. These networks play a crucial role in the dis-

semination and enhancement of practice-related skills such as programming (Boudreau and Lakhani 2009).Given its high profile,active participation among programmers and its role of a skill development tool(Brown and Roche 2023), we can use the posting activity on Stack Overflow as an indicator of local programming skills. Our dataset contains a total of 1,453,770 posts originating from 2,311 counties across the U.S. in the year 2011.

Our study starts with an Ordinary Least Squares (OLS) regression to estimate the overall impact of programming literacy on local wage growth in the United States. We discover that the general impact of programming literacy on wage growth is economically modest, even after controlling for pre-sample demographic characteristics and hardware availability. However, we find that in more populous counties (had a population over 150,000 in 2011), the influence of programming skills is both larger and statistically significant.

To address potential endogeneity of posting activity, we employ two instrumental variables: the average slope of terrain within the county and the total fixed internet connections per capita. These instruments are chosen to provide exogenous variation in programming activity, thereby helping to isolate its true effect on local wage growth. Although the instruments are somewhat weak, the results align with our initial conclusions: programming skills primarily benefit larger urban areas. This consistency reinforces the notion that the advantages of programming literacy are most pronounced in regions with substantial populations.

This paper is structured as follows: we start by describing the dataset, variables, and the empirical models employed. We then proceed with the OLS analysis, explore heterogeneity effects, and conduct an instrumental variable analy-

sis. Finally, we conclude with key insights derived from the results.

2 DATA

To investigate the impact of programming literacy on wage growth across various counties in the United States, we combined data from several comprehensive databases, encompassing demographic, economic, and educational metrics at the county level.

To quantify programming literacy, we utilize the posting volume in each county per population in 2011 on Stack Overflow as a proxy. The dataset for the posting activity variable is derived from the Stack Overflow PostsByLocation dataset. Utilizing posting-level data, we aggregate the posts from 2011 to the county level based on the latitude and longitude coordinates of the posters' IP addresses. This method allows us to compile the total number of posts made in each county during the year 2011, providing a comprehensive county-level view of posting activity.

To control for the impact of hardware in our analysis, we utilize the variable of average personal computers (PCs) per employee, drawn from the Computer Infrastructure (CI) Technology Market Intelligence database (hereinafter referred to as the CI database). This dataset provides detailed insights into establishment and firm-level technology usage, including metrics on the number of employees, PCs per employee, and internet application usage. Collected by Harte Hanks for technology market analysis, the CI database has also been widely used by economic researchers to study IT adoption in businesses. Our analysis specifically employs the dataset as of December 2011, offering a snapshot of the hardware availability at that time.

We sourced data on average weekly wages and total employment at the

county level from the Quarterly Census of Employment and Wages, a collaborative effort between the Bureau of Labor Statistics and State Employment Security Agencies. After aligning this data with our Stack Overflow activity data, we were left with observations from 2,311 counties. Out of the initial 3,108 counties, we excluded 745 due to the absence of posts in 2011. Predominantly, the counties omitted were those within the bottom quartile in terms of population size.

We aggregate county-level data from a multitude of sources to adjust for demographic variables potentially influencing the counties' capacity for growth and innovation. Specifically, from the American Community Survey (ACS) 5-year estimates, we obtain detailed demographic information, including population size, median income, the percentage of the population over 25 with a bachelor degree or higher, and the percentage of the population under the poverty line of 2011. Additionally, we consider population dynamics by evaluating the logarithmic difference between the populations in 2011 and 2017, offering insight into growth trends. The ACS, conducted by the U.S. Census Bureau, serves as a comprehensive source for understanding the demographic and socio-economic fabric of U.S. counties, pivotal for contextualizing our analysis within the broader framework of county characteristics.

To further control the counties' propensity for innovation, we employ two specific measures. Firstly, the count of patents granted from 2001 to 2010, sourced from the United States Patent and Trademark Office (USPTO), provides a direct metric of inventive activity. Secondly, we analyze the county's industrial composition with respect to Information Technology, measuring the fraction of firms in IT-using and IT-producing industries based on County Business

Table 2.1: Descriptive Statistics for Dependent Variables, IT Measures, and Instruments

VARIABLES	N	mean	sd	min	max
Number of Posts	2,311.00	629.07	2,240.01	1.00	42,117.00
Population 2011	2,311.00	130,667.51	363,033.84	690.00	9,873,700.00
Posts per Capita 2011	2,311.00	0.0031	0.0053	0.0000162	0.1288315
ln(Posts/Population 2011)	2,311.00	-6.45	1.25	-11.03	-2.05
PCs per employee	2,311.00	0.81	0.36	0.28	9.12
Average weekly wage 2011	2,311.00	692.61	152.54	407.00	1,950.00
Average weekly wage 2017	2,311.00	790.74	174.09	468.00	2,437.00
ln(wage17)-ln(wage11)	2,311.00	0.13	0.06	-0.46	0.54
Avearage slope	2,311.00	2.30	2.63	0.05	18.38
Total fixed connections(2011)	2,234.00	33.79	96.92	0.00	2,494.00
Total fixed connections per capita (2011)	2,234	0.0002129	0.0000643	0	0.0006034

Patterns data. These industries are identified by the classification utilized by Jorgenson et al.(2011). This combination of demographic data and innovation metrics provides a robust set of control for examining the interplay between software adoption and economic development at the county level.

Table 2.1 includes descriptive statistics on independent,dependent and instrumental variables. Table 2.2 includes a description of control variables.

Table 2.2: Data Description of Control Variables

Variable	Definition	Source	Mean
Total population 2011	Total population as of ACS(2011)	American Community Survey Data(2011)	130,667.51
Total population 2017	Total population as of ACS(2017)	American Community Survey Data(2017)	136,567.83
Median household income	Median county household income(2011)	American Community Survey Data(2011)	46,987.70
% population over 25 with bachelor degree or higher	% population over 25 with bachelor	American Community Survey Data(2011)	20.60
% below poverty line	% population below poverty line	American Community Survey Data(2011)	11.09
Percent population over 65	Percent of county population over 65	American Community Survey Data(2011)	15.07
Change in log total population between 2011 and 2017	log(population2017) – log(population2011)	American Community Survey Data(2011,2017)	0.01
ln(Patents)	Ln((Total number of patents from inventors located in county, 2001-2010)+1)	USPTO	3.58
Percentage of establishments of IT producing industries	Percentage of establishments of IT producing industries in county	County Business Pattern (2011)	0.01
Percentage of establishments of IT using industries	Percentage of establishments of IT using industries in county	County Business Pattern (2011)	0.61

3 EMPIRICAL SPECIFICATION

We begin with a cross-sectional study, examining the growth in annual average weekly wages from 2011 to 2017. This log difference in wages between 2011 and 2017 serves as the dependent variable in our analysis. As the independent variable, we consider the logarithm of Stack Overflow post counts, adjusted per capita, to reflect the level of programming literacy across the population.

$$\log(Y_{i17}) - \log(Y_{i11}) = \alpha X_i + \beta \text{Posts}_i + \varepsilon_i \tag{3.1}$$

Here, Posts_i represents the logarithm of Stack Overflow posting frequency per capita in location i for the year 2011. This measurement involves taking the logarithm of posts normalized by local population, effectively adjusting for the impact of population on posting frequency and addressing the skewed distribution of normalized posts. This normalization ensures a more accurate reflection of programming activity relative to the size of the population in each area. ε_i is the error term which we assume to be i.i.d normal. Here, X_i includes a set of control variables that account for preexisting initial demographic conditions potentially influencing wage growth, including factors like median income , overall population, and educational attainment, as well as adjustments for shifts in population size to control for the potential changes in labor supply. The dependent variable in our study is the logarithmic difference in annual weekly wages between 2017 and 2011, capturing the percentage change in wages over the six-year period.

To test whether enhancements in programming literacy correlate with local wage increases, the focal point of interest is the coefficient β. We examine the hypothesis that β is positive, indicative of a positive relationship, against the

null hypothesis that β equals zero, suggesting no association on average.

Programming literacy does not thrive in isolation; rather, it relies on a vibrant environment comprising programmers, technology entrepreneurs, and tech product users who collectively contribute to creating, refining, and embedding programming into practical applications.(Wright,Nagle and Greenstein 2024, Forman, Goldfarb, and Greenstein 2012)

Hence, the quality of human capital becomes a pivotal element when assessing the impact of programming literacy on local economic landscapes. The existing body of literature, including models of skill-biased technical change (Autor, Katz, and Kearney 2006), suggests that a decrease in the cost of computing capital tends to increase the demand for skilled labor, consequently elevating the wages of skilled workers in comparison to their unskilled counterparts. Consequently, regions that boast a higher concentration of skilled professionals, particularly those in the tech sector, are more likely to experience augmented wage growth.

Further, studies have identified a location-specific bias in the effects of Internet adoption on local economies (Forman, Goldfarb, and Greenstein 2012). Enhanced internet technologies are correlated with more substantial wage growth in areas that are already prosperous, characterized by a well-educated populace, large metropolitan areas, and a density of IT-intensive industries.

To assess whether the impact of programming literacy is similarly location-biased, we examine several aspects of local economies, including income levels, educational attainment, population size, and the prevalence of IT-centric industries, as well as how these factors interact. We approach this analysis with the

following specification:

$$\log(Y_{i11}) - \log(Y_{i17}) = \alpha X_i + \beta \text{Posts}_i + \gamma_1(\text{Posts}_i \times \text{HighIncome}_i) + \varepsilon_i \tag{3.2}$$

$$\log(Y_{i11}) - \log(Y_{i17}) = \alpha X_i + \beta \text{Posts}_i + \gamma_2(\text{Posts}_i \times \text{HighEducation}_i) + \varepsilon_i \tag{3.3}$$

$$\log(Y_{i11}) - \log(Y_{i17}) = \alpha X_i + \beta \text{Posts}_i + \gamma_3(\text{Posts}_i \times \text{HighPopulation}_i) + \varepsilon_i \tag{3.4}$$

$$\log(Y_{i11}) - \log(Y_{i17}) = \alpha X_i + \beta \text{Posts}_i + \gamma_4(\text{Posts}_i \times \text{HighITIntensity}_i) + \varepsilon_i \tag{3.5}$$

$$\log(Y_{i11}) - \log(Y_{i17}) = \alpha X_i + \beta \text{Posts}_i + \gamma_5(\text{Posts}_i \times \text{HighAllFactors}_i) + \varepsilon_i \tag{3.6}$$

and

$$\begin{aligned} \log(Y_{i17}) - \log(Y_{i11}) = {} & \alpha X_i + \beta \text{Posts}_i \\ & + \gamma_1 \text{HighIncome}_i + \gamma_2 \text{HighEducation}_i \\ & + \gamma_3 \text{HighPopulation}_i + \gamma_4 \text{HighITIntensity}_i \\ & + \gamma_5 \text{HighAllFactors}_i \\ & + \phi_1(\text{Posts}_i \times \text{HighIncome}_i) \\ & + \phi_2(\text{Posts}_i \times \text{HighEducation}_i) \\ & + \phi_3(\text{Posts}_i \times \text{HighPopulation}_i) \\ & + \phi_4(\text{Posts}_i \times \text{HighITIntensity}_i) \\ & + \phi_5(\text{Posts}_i \times \text{HighAllFactors}_i) + \varepsilon_i \end{aligned} \tag{3.7}$$

The variable HighIncome_i is a dummy variable indicating whether the median income in location i in 2011 falls into the top quantile. Similarly, HighEducation_i is a dummy variable representing whether the proportion of the population aged 25 and over with at least a bachelor's degree in location i is in the top quantile. The HighPopulation_i variable is a dummy that indicates whether the population in location i exceeds 150,000. The HighITIntensity_i variable represents whether the ratio of IT-producing to IT-using establishments in

location i is in the top quantile. Finally, HighAllFactors_i is a composite dummy variable that indicates whether all four of these conditions are met in location i. The sets of γs and ϕs measure the difference in the relationship between wage growth and programming literacy across different locations. A divergence caused by programming skills is expected to produce $\gamma > 0$ ($\phi > 0$), indicating a positive impact on wage growth. Conversely, a relative convergence in wage growth rate between locations with and without pre-sample strength would be indicated by $\gamma < 0$ ($\phi < 0$), suggesting a negative impact. A finding of $\gamma = 0$ (or $\phi = 0$), where the null hypothesis cannot be rejected, suggests no significant heterogeneity effect.

Due to the inherently endogenous nature of programming literacy, a potential concern is that unobservable local factors may be correlated with both programming activity and local wage growth. To address this issue, we first incorporate a variety of demographic controls that reflect the initial conditions of the county, thus mitigating the risk of omitted variable bias. We also include controls for population changes to account for variations in labor supply. Additionally, to more accurately gauge the impact of programming literacy, we control for local hardware capacity by including 'PCs per employee' as a specific measure. This control is crucial as it helps to adjust for the level of technological infrastructure available, which could significantly influence both the development of programming skills and the economic benefits derived from such skills. By integrating these controls, we aim to isolate the unique contribution of programming literacy to wage growth, ensuring that our findings robustly reflect its true effect while minimizing potential confounding factors.

Furthermore, we implement an instrumental variables analysis, employing

local average slope of terrain(Chan,Ghose and Seamans 2016,Kolko 2012) and local fixed connections as instruments for programming skills. Variations in these instruments serve as proxies for differences in internet speed and access, which directly influence posting activity on platforms like Stack Overflow. Engaging effectively on Stack Overflow requires users to search and research relevant issues, clearly describe problems, proofread submissions, and interact promptly with peers (Stack Overflow 2024). Poor internet connectivity can significantly increase the time and effort needed to post and resolve queries, potentially reducing the inclination to participate actively. However, these variations in internet speed and access are unlikely to be systematically correlated with unobservable factors that might affect local wage growth.This approach enhances the robustness of our conclusions by ensuring that our estimates of the impact of programming literacy are not confounded by unobservable variables.

4.1 Baseline Results

We initiate our investigation with a cross-sectional analysis to examine the overall relationship between programming skills and local wage growth. This is conducted using the Ordinary Least Squares (OLS) method outlined in Equation 3.1. To visualize this relationship, Figure A.1 presents a scatter plot where the x-axis features logged posts per capita in 2011, serving as a proxy for programming literacy, and the y-axis displays the log wage difference between 2011 and 2017.The figure suggests a modest positive relationship between the number of posts and wage growth, indicating a weak overall correlation.

In Table 4.1, we reports the baseline results at the county level. Column 1 details the correlation between posting activity and wage growth without the inclusion of control variables. Here, the coefficient is positive and statistically significant, yet the economic link is weak; a 1 percent increase in posts per capita is associated with a mere 0.003 percentage point increase in wage growth. Column 2 introduces estimates with a comprehensive set of demographic controls from 2011 and adjustments for changes in population between 2011 and 2017. In Column 3, we aim to isolate the effect of programming literacy on wage growth by controlling for hardware availability, specifically through the measure of PCs per employee within the county's establishments in 2011.While PCs per employee are positively correlated with wage growth, this relationship is not statistically significant. Given that previous literature has already demonstrated the heterogeneity of IT effects, which vary significantly across different contexts and notably provide greater benefits in top counties than in others (Forman,

Goldfarb, and Greenstein 2012), the overall subtle impact is not surprising.

Table 4.1: Baseline OLS regression

	(1)	(2)	(3)
VARIABLES	No controls	Full set of controls	Include measures of IT use
ln(post2011/population2011)	0.003***	0.002**	0.002**
	(0.001)	(0.001)	(0.001)
PCs per employee			0.002
			(0.004)
ln(population)		-0.009***	-0.009***
		(0.002)	(0.002)
ln(median income)		-0.026*	-0.026*
		(0.014)	(0.014)
Percentage of population over 25 with bachelor degree or higher		-0.000	-0.000
		(0.0002)	(0.0002)
ln(patent granted 2001-2010)		0.003**	0.003**
		(0.002)	(0.002)
Percentage of population under poverty line		-0.002***	-0.002***
		(0.001)	(0.001)
Percentage of population over 65		0.001	0.001
		(0.000)	(0.000)
log(population2017)-log(population2011)		0.218***	0.219***
		(0.032)	(0.032)
Percentage of establishments of IT producing industries		-0.288*	-0.299*
		(0.165)	(0.166)
Percentage of establishments of IT using industries		0.042	0.041
		(0.037)	(0.037)
Observations	2,311	2,311	2,311
R-squared	0.004	0.064	0.064

Dependent variable is change in logged annual weekly wage from 2011 to 2017. Heteroskedasticity-robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

4.2 In What Contexts was Programming Skills Related to Wage Growth

Given the significant role that human capital plays when technology exerts its economic influence, this section will explore the regional variations in how programming skills affect wage growth. We will investigate how factors such as income, population, education, and the structure of IT industries contribute to the effectiveness of programming skills in driving economic outcomes. The choice of factors is based on Forman, Goldfarb, and Greenstein(2012). This analysis aims to understand the contexts in which programming skills can be most beneficial and how they interact with other socio-economic variables.

Table 4.2 reports the relationship between programming skills and local wage growth in different contexts. Column 1 reveals that there is no significant difference in the impact of programming skills on local wage growth between high-income counties (counties in the top quantile of median income as of 2011) and counties in other quantiles, as evidenced by the insignificant coefficient of the interaction term between the independent variable and the high-income counties dummy. Columns 2 and 4 demonstrate that the impact of programming skills on local wage growth does not vary significantly with differences in local education levels (classified by whether the proportion of the population over 25 with a bachelor's degree is in the top quantile as of 2011) or IT-intensity levels (classified by whether the proportion of IT-producing and IT-using establishments is in the top quantile as of 2011). In column 2, the analysis shows that programming skills are associated with local wage growth primarily in counties with large populations (counties with populations greater than 150,000). Of

the 2311 counties analyzed, 414 fall into this category.For the high-income, high-education, high IT-intensity, and "high all factors" counties, the breakdown is as follows: 699 for high-income, 717 for high-education, 630 for high IT-intensity, and 173 for counties that score high across all factors.

Column 5 shows that when we include all four measures related to a county's pre-sample strength in human capital, particularly within IT industries, only the dummy variable for high population counties and its interaction term with the independent variable appear to be significant. Given the potential overlap between the measures—where each measure encompasses roughly 700 counties (with high population comprising 414 counties), and 220 counties ranking in the top group for all factors—the results in Column 6 indicate that the correlation is not significant. This conclusion is drawn when considering all dummy variables and the interaction term of all high factors with the independent variable in the regression.Column 7 presents the results incorporating all dummy variables and interaction terms, confirming that local wage growth is primarily driven by programming skills in large counties with high populations.

These results indicate that in high population counties, the average posts per capita account for 13.8% of local wage growth, which translates to 1.84 percentage points out of a total of 13.02 percentage points. Conversely, in counties with populations less than 150,000 in 2011, average posts per capita contribute only 1.42%, or 0.18 percentage points out of an average of 13.34 percentage points[1].These 414 counties represent over 70% of the US population in 2011 across the entire sample. This results remain robust when using continu-

[1]These calculations are based on the margin estimates provided in Table 4.2, Column 5,using the average posts per capita for each group. The average posts per capita in high population counties is 0.0050723, while in other counties it is 0.0026951.

ous measures of income,education,population and IT-intensity.High All Factors gains statistical significant although the scale is very small even considering the scale of continuous measures of income and population size (with absolute value less than $10^{-}9$).[2] The findings indicate that urban areas with substantial labor markets are better positioned to capitalize on the benefits of programming skills. This aligns with insights from prior research that larger cities tend to offer greater rewards for analytical and social intelligence skills (Florida et al., 2012).

4.3 Justifying a Causal Interpretation

The relationship between programming skills and income can also be influenced by other omitted variables, such as the presence of higher education institutions. For example, areas with smaller populations but large numbers of students in higher education institutions might exhibit high levels of posting activity on platforms like Stack Overflow. However, these students often do not remain in the area after graduation, which means their presence does not necessarily translate into long-term local wage growth. Taking into account the potential for omitted variable bias, we incorporate an instrumental variables analysis to complement our previous analyses,such as the clustering of Higher Education institutes. The first instrumental variable selected for our analysis is the average slope of terrain , which is measured by average gradient of horizontal grid spacing of 7.5arc seconds (approximately 250 meters) within a county. Topographical variability not only affects the costs associated with building and operating network infrastructures but also influences the rate of broadband penetration and the quality of internet connectivity, which in turn could impact the

[2]See Table B.2

volume of posts on Stack Overflow, satisfying the relevance condition for an instrumental variable. Moreover, as a geographic characteristic, topographical variability is sufficiently exogenous and does not directly influence urban innovation, thereby fulfilling the exogeneity condition for an instrumental variable.[3] The second instrumental variable we employ is the total number of fixed Internet access connections with speeds over 200 kbps in at least one direction within each county (per capita) as of June 2010.[4] This measure reflects the infrastructure capacity for high-speed Internet, which is presumed to affect the use of online platforms such as Stack Overflow. This variable is unlikely to be correlated with local wage growth through other ways since the construction the Internet Access Service is unlikely to be related to omitted demographic variables.

Columns 1, 4, and 7 of Table 4.3 report the results of the instrumental variables analysis, with the average slope of terrain within the county and the total fixed internet connections per capita in 2010 serving as instruments.[5] The first stage regression reveals a positive, yet statistically insignificant, relationship between the average slope of terrain and the natural logarithm of posts per capita. The F-statistic for the instruments in the first stage is 1.77. Regarding the fixed internet connections instrument, the first stage demonstrates a positive and statistically significant coefficient, with an F-statistic of 8.68. When incorporating both instruments, only the coefficient for connections per capita remains significant, with an F-statistic of 5.7. The hypothesis of overidentification is rejected, as indicated by a p-value of 0.38. The coefficients for the second stage are not significant, which aligns with the results observed under the OLS model, indi-

[3]Source:USGS EROS Archive-Digital Elevation-Global 30 Arc-Second Elevation (GTOPO30)

[4]Source: Federal Communication Commision, Form 477 County Data on Internet Access Services

[5]The coefficient of fixed internet connections and its interaction terms are large because of the small scale of the variable itself, with a mean of 0.00021.

cating that the general relationship across all counties is not significant.

Columns 2, 5, and 7 report the heterogeneity analysis, focusing on "High all factors" counties and utilizing instrumental variables in the assessment.We interact each of our original instruments with an indicator for being located in one of the HighAllFactors counties. The F-statistics for both instruments are approximately 7 , and increase to 18.5 and 35.46 when the independent variable includes the interaction term with HighAllFactors. The coefficient of ln(posts per capita) interacting with HighAllFactors is positive and significant when using the average slope of terrain alone as the instrument, as well as when employing both the average slope of terrain and connections per capita together.

Columns 3 and 6 explore heterogeneity by focusing on "High Population" counties. The F-statistics for both instruments are approximately 9, and they increase to more than 30 when the independent variable includes the interaction term with HighAllFactors. The coefficient of ln(posts per capita) interacting with HighPopulation is positive and significant when using the average slope of terrain alone as the instrument.

Hausman tests were conducted on the above regressions, and in all cases, the null hypothesis could not be rejected. In fact, the conclusions drawn from the instrumental variables analysis largely align with those from the OLS regression results. This consistency could be attributed to a couple of factors: firstly, the coefficients of the control variables did not undergo significant changes, and secondly, the instruments used may not be sufficiently strong.

Table 4.2: Relationship in different contexts

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(posts per capita)	0.002*	0.002	0.002	0.002	0.002	0.002*	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
High income county	0.003				-0.013	-0.005	-0.014
	(0.018)				(0.020)	(0.004)	(0.020)
ln(post per capita) and high income county	0.001				-0.001		-0.001
	(0.003)				(0.003)		(0.003)
High education county		0.012			-0.014	-0.003	-0.014
		(0.018)			(0.020)	(0.004)	(0.020)
ln(post per capita) and high education county		0.003			-0.002		-0.002
		(0.003)			(0.003)		(0.003)
High population county			0.097***		0.101***	0.004	0.101***
			(0.030)		(0.032)	(0.005)	(0.023)
ln(post per capita) and high population county			0.016***		0.017***		0.017***
			(0.005)		(0.005)		(0.004)
High IT-intensity county				0.026	0.018	-0.003	0.017
				(0.016)	(0.017)	(0.004)	(0.018)
ln(post per capita) and high IT-intensity county				0.004*	0.003		0.003
				(0.003)	(0.003)		(0.003)
High all factors county						0.087	-0.006
						(0.061)	(0.064)
ln(post per capita) and high all factors county						0.015	-0.002
						(0.011)	(0.012)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,311	2,311	2,311	2,311	2,311	2,311	2,311
R-squared	0.065	0.065	0.070	0.066	0.072	0.069	0.072

Notes:Dependent variable is change in logged annual weekly wage from 2011 to 2017. In addition to the controls in Table 4.1, regressions here include dummies for the high income ,high education level, high IT-intensity and,high population,high all factors counties and their interaction terms with the independent variable. Heteroskedasticity-robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 4.3: Instrumental Variables Analysis

	Average Slope			total fixed Internet access connections over 200 kbps in 2010 per capita			Both	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First Stage: DV is ln(posts per capita)								
Average Slope	0.011	0.012	0.006				0.012	0.007
	(0.009)	(0.009)	(0.009)				(0.008)	(0.009)
Fixed internet Connections				1,633.912***	1,605.983***	1,633.282***	1,584.149***	1,615.073***
				(554.735)	(555.938)	(555.253)	(558.726)	(560.243)
Average Slope and high all factors		-0.005						0.065***
		(0.016)						(0.020)
Fixed internet connections and high all factors					-1,002.927***			-1,453.752***
					(291.647)			(354.157)
Average Slope and high population			0.045***					
			(0.014)					
Fixed internet connections and high popualtion						11.853		
						(272.964)		
Partial R^2	0.001	0.002	0.002	0.005	0.008	0.011	0.006	0.0099
F- statistic	1.77	7.87	9.63	8.68	6.41	9.26	5.7	7.36
First Stage: DV is ln(posts per capita) and high all factors/population								
Average Slope		0.056***	0.113***					-0.002**
		(0.006)	(0.010)					(0.001)
Fixed internet Connections					40.473	715.561***		73.724
					(90.372)	(200.499)		(88.483)
Average Slope and high all factors		-0.782***						-0.571***
		(0.086)						(0.083)
Fixed internet connections and high all factors					-17,710.258***			-18,032.392***
					(213.681)			(303.231)
Average Slope and high population			-0.607***					
			(0.056)					
Fixed internet connections and high population						-20,102.954***		
						(226.034)		
Partial R^2		0.1	0.06		0.07	0.0393		0.12
F- statistic		18.5183	33.83		35.46	39.87		14.93
Second Stage: DV is log wage difference								
ln(posts per capita)	-0.100	-0.129	-0.129	-0.029	-0.025	-0.040	-0.036	-0.030
	(0.092)	(0.127)	(0.126)	(0.027)	(0.018)	(0.042)	(0.026)	(0.020)
ln(posts per capita) and high all factors		0.145*			0.020			0.074***
		(0.088)			(0.057)			(0.022)
ln(posts per capita) and high popoulation			0.143*			-0.044		
			(0.084)			(0.063)		
Overidentification test (p-value)							0.38	0.26
Hausman test (p-value)	0.99	1	0.99	0.97	0.99	1	0.93	0.88

Notes:Heteroskedasticity-robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

5 CONCLUSION

This research has explored the influence of programming skills on local wage growth. We initiated our analysis with baseline regressions, which revealed subtle correlations between posting activity and higher wage growth. Furthermore, we observed that in larger counties, strong programming skills are significantly associated with higher wage growth. However, this effect is not apparent in areas characterized by high income, high education levels, or high IT-intensity. These findings reinforce conclusions from previous literature, highlighting that analytical skills benefit from clustering, and affirming that a major driver of economic growth is the capacity to innovate and undertake new endeavors((Florida et al., 2012)). When compared to the effects of IT investment, where factors such as income, education, IT intensity, and population all play a role in local wage growth(Forman, Goldfarb, and Greenstein 2012), one possible explanation for the observed patterns is the presence of technological hubs. These hubs are predominantly located in larger counties that offer more programming-related job opportunities, potentially acting as a complement to online activities(Brown and Roche 2023). The insignificant results concerning IT intensity might be attributed to outdated and broad classifications of IT-related industries. These classifications often include traditional sectors that do not focus on the high-tech industries leading economic development in the 2010s. This misclassification could dilute the apparent impact of IT intensity on local economic growth.

A FIGURES

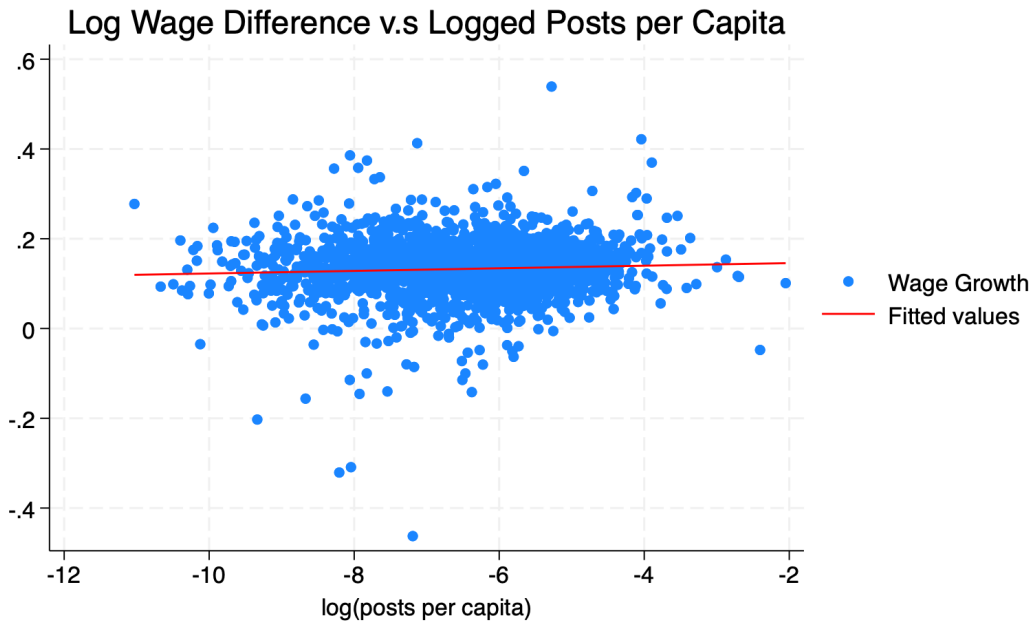

Figure A.1: Scatter Plot of Logged Posts per Capita and Wage Growth

B **TABLES**

Table B.1: Full Table of Table4.2

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(posts per capita)	0.002*	0.002	0.002	0.002	0.002	0.002*	0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
High income county	0.003				-0.013	-0.005	-0.014
	(0.018)				(0.020)	(0.004)	(0.020)
ln(post per capita) and high income county	0.001				-0.001		-0.001
	(0.003)				(0.003)		(0.003)
High education county		0.012			-0.014	-0.003	-0.014
		(0.018)			(0.020)	(0.004)	(0.020)
ln(post per capita) and high education county		0.003			-0.002		-0.002
		(0.003)			(0.003)		(0.003)
High population county			0.097***		0.101***	0.004	0.101***
			(0.030)		(0.032)	(0.005)	(0.023)
ln(post per capita) and high population county			0.016***		0.017***		0.017***
			(0.005)		(0.005)		(0.004)
High IT-intensity county				0.026	0.018	-0.003	0.017
				(0.016)	(0.017)	(0.004)	(0.018)
ln(post per capita) and high IT-intensity county				0.004*	0.003		0.003
				(0.003)	(0.003)		(0.003)
High all factors county						0.087	-0.006
						(0.061)	(0.064)
ln(post per capita) and high all factors county						0.015	-0.002
						(0.011)	(0.012)
PCs per employee	0.002	0.002	0.002	0.002	0.002	0.002	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
ln(population 2011)	-0.009***	-0.009***	-0.010***	-0.009***	-0.010***	-0.010***	-0.010***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ln(median income)	-0.017	-0.026*	-0.023	-0.025*	-0.014	-0.018	-0.014
	(0.015)	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)
Percentage of population over 25 with bachelor degree or higher	-0.000	0.000	-0.000	-0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Percentage of population under poverty line	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Percentage of population over 65	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
log(population2017)-log(population2011)	0.221***	0.221***	0.222***	0.220***	0.225***	0.225***	0.226***
	(0.032)	(0.032)	(0.032)	(0.032)	(0.033)	(0.033)	(0.033)
ln(patent granted 2001-2010)	0.003**	0.003**	0.003*	0.003**	0.003*	0.003*	0.003*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Percentage of establishments of IT producing industries	-0.327**	-0.341**	-0.456***	-0.341**	-0.481***	-0.449***	-0.487***
	(0.166)	(0.163)	(0.166)	(0.163)	(0.164)	(0.172)	(0.168)
Percentage of establishments of IT using industries	0.043	0.044	0.044	0.059	0.065	0.061	0.066
	(0.037)	(0.038)	(0.037)	(0.049)	(0.048)	(0.049)	(0.048)
Observations	2,311	2,311	2,311	2,311	2,311	2,311	2,311
R-squared	0.065	0.065	0.070	0.066	0.072	0.069	0.072

Notes:Dependent variable is change in logged annual weekly wage from 2011 to 2017. In addition to the controls in Table 4.1, regressions here include dummies for the high income ,high education level, high IT-intensity and,high population,high all factors counties and their interaction terms with the independent variable.

Heteroskedasticity-robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Table B.2: Analysis with continuous interaction terms

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ln(posts per capita)	0.00290	0.00124	-0.0176	0.0217	-0.00331	0.00406***	0.00220
	(0.00515)	(0.00300)	(0.0123)	(0.0175)	(0.0194)	(0.00149)	(0.0197)
ln(post per capita) x county-level median income	-1.14e-08				-1.37e-07		-1.51e-07
	(1.14e-07)				(1.54e-07)		(1.53e-07)
MedianIncome2011	-7.11e-07				-1.39e-06*	-1.15e-06***	-2.08e-06**
	(6.27e-07)				(8.36e-07)	(3.59e-07)	(8.66e-07)
Pop25over_bachelorhigher_pct		0.000168			0.000750	-0.000712**	0.000635
		(0.000842)			(0.000950)	(0.000357)	(0.000920)
edu_post		6.00e-05			0.000149		0.000259
		(0.000148)			(0.000169)		(0.000162)
ln(population)			0.00253		0.00674	-0.0108***	0.00467
			(0.00712)		(0.00753)	(0.00220)	(0.00734)
ln(posts per capita) x ln(population)			0.00190		0.00263**		0.00254**
			(0.00118)		(0.00126)		(0.00121)
IT using or producing establishments proportion				-0.203	-0.163	0.0118	-0.228
				(0.179)	(0.190)	(0.0376)	(0.192)
ln(posts per capita)x county level IT-intensity				-0.0315	-0.0295		-0.0358
				(0.0283)	(0.0297)		(0.0296)
ln(posts per capita)x income x education x lnpop x IT-intensity						-2.81e-10**	-3.51e-10**
						(1.35e-10)	(1.46e-10)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,311	2,311	2,311	2,311	2,311	2,311	2,311
R-squared	0.056	0.052	0.061	0.052	0.068	0.066	0.071

Notes:Dependent variable is change in logged annual weekly wage from 2011 to 2017.

Heteroskedasticity-robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

BIBLIOGRAPHY

[1] Dale W. Jorgenson, Mun S. Ho, and Kevin J. Stiroh. *Productivity, Volume 3: Information Technology and the American Growth Resurgence*. MIT Press, Cambridge, MA, 2005.

[2] Paul Beaudry, Mark Doms, and Ethan Lewis. Endogenous skill bias in technology adoption: City-level evidence from the it revolution. Working Paper 2006-24, Federal Reserve Bank of San Francisco, 2006.

[3] Robert J. Barro. Economic growth in a cross section of countries. *Quarterly Journal of Economics*, 106(2):407–443, 1991.

[4] Robert Barro and Jong-Wha Lee. International data on educational attainment: Updates and implications. *Oxford Economic Papers*, 53(3):541–563, 2001.

[5] Gary S. Becker. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. The University of Chicago Press, Chicago, 1993.

[6] Robert E. Lucas. On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42, 1988.

[7] Paul M. Romer. Increasing returns and long-run growth. *Journal of Political Economy*, 94(5):1002–1037, 1986.

[8] John Divine. Decade in review: 'big tech' gains enormous power, December 2019. Accessed: 2024.04.28.

[9] T.J. Allen. *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information Within the R&D Organization*. MIT Press, Cambridge, Massachusetts, 1997.

[10] M. P. Roche. Taking innovation to the streets: Microgeography, physical structure, and innovation. *Review of Economics and Statistics*, 102(5):912–928, 2020.

[11] M. Roche, A. Oettl, and C. Catalini. Co-working in close proximity: Knowledge spillovers and social interactions. Working paper, Harvard Business School, 2022.

[12] Daniel Jay Brown and Maria P. Roche. Learning to use: Stack overflow and technology adoption. Working Paper No. 24-001, Harvard Business School, July 2023.

[13] Signhouse. Stack overflow growth and usage statistics, 2024. Accessed: 2024.04.24.

[14] Kevin J. Boudreau and Karim R. Lakhani. How to manage outside innovation. *MIT Sloan Management Review*, 50:69–76, 2009.

[15] Dale W. Jorgenson, Mun S. Ho, and Jon D. Samuels. Information technology and u.s. productivity growth: evidence from a prototype industry production account. *Journal of Productivity Analysis*, 36(2):159–175, 2011.

[16] Nataliya Langburd Wright, Frank Nagle, and Shane Greenstein. Contributing to growth? the role of open source software for global startups. Working Paper No. 24-040, Harvard Business School, January 2024.

[17] Chris Forman, Avi Goldfarb, and Shane Greenstein. The internet and local wages: A puzzle. *American Economic Review*, 102(1):556–75, February 2012.

[18] Autor David, Lawrence F. Katz, and Melissa S. Kearney. The polarization of the U.S. labor market. *American Economic Review*, 96(2):189–194, 2006.

[19] Jason Chan, Anindya Ghose, and Robert Seamans. The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly*, 40(2):381–404, 2016.

[20] Jed Kolko. Broadband and local growth. *Journal of Urban Economics*, 71(1):100–113, 2012.

[21] Stack Overflow. How do i ask a good question. https://stackoverflow.com/help/how-to-ask, 2024. Accessed: 24.04.28.

[22] Richard Florida, Charlotta Mellander, Kevin Stolarick, and Adrienne Ross. Cities, skills and wages. *Journal of Economic Geography*, 12(2):355–377, 07 2011.