# Productivity Breakthrough or Not: How Internet Industry Has Impacted College Premium in China

Yutong LI

June 25, 2024

**Abstract**

In China, the unusual simultaneous increase of college premium and college graduate supply during the 2010s coincides with the takeoff of Chinese Internet industry. Naturally, one may ask whether and how the latter can impact the former. Using CFPS household data of 2014, 2016, 2018, and 2020, I explore the effect of Internet industry on college premium over the 2010s with intersection-term regression. I propose and verify hypotheses of computer application growth, industry-specific productivity breakthrough, efficiency wage, and regional difference. I find significant evidence for productivity based effect in 2014 and 2018, and indirect evidence for effect in 2020 with unknown mechanism. Findings of this paper update understanding of college premium in China and provide a new explanation based on industry specific effect.

**Keywords:** college premium, Internet industry, Internet firms

# Introduction

In China, the 2010s witnesses the boom of Internet industry and the end of college enrolment expansion. Over the decade, Chinese youth all talk about graduating college and making a fortune in Internet industry. Seemingly, Internet firms have made college degrees more valuable, although there have been too many graduates at the time. The question is, how significant is the effect and how has it occurred? Using household data of 2014, 2016, 2018, and 2020, I explore the effect of Internet industry on college premium in China over the 2010s and discuss whether it is caused by productivity breakthroughs of the industry. I find significant evidence for productivity-based effect in 2014 and 2018, and indirect evidence for effect with unknown mechanism in 2020. Findings of this paper update understanding of college premium in China with most up-to-date data and provide a novel perspective of analysing how college premium can be impacted by an emerging industry. These results are of substantial relevance in an ever-changing world with rising inequality.

# Hypothesis

According to Katz and Murphy (1992), simultaneous increase of college premium and graduate supply in the entire labour market, which occurs in China during the 2010s (Hu & Bollinger, 2021), can only be supported by productivity growth. Therefore, highly likely the effect of Internet industry origins from productivity breakthrough. Nevertheless, it is also possible that non-productivity factors, like efficiency wage (Ikeuchi et al., 2024) or regional difference (Asadullah & Xiao, 2019), generate the effect from Internet industry, since the industry only constitutes a small part of labour market. Following these theoretical discussions, I propose four hypotheses in Table 1.

# Data and Variables

I draw data from China Family Panel Study (CFPS) 2014, 2016, 2018, and 2020. Led by Peking University, the project surveys national-wide household information, including income, education, employment, and personal characteristics, every two years since 2010. As I target on college premium, samples are further restricted to employed adults.

Table 2 summaries variables used. To measure changes in wage, I use natural logarithm of annual wage as dependent variable. After-tax data is used, as pre-tax data is unavailable in CFPS.

Main independent variables are constructed according to hypotheses. Edu captures the most widely defined notion of college premium, thus serving as a benchmark for various edu intersections proposed later. Prov assigns 1 to individuals working in Beijing, Shanghai, Jiangsu, Fujian, Tianjin, Zhejiang, and Guangdong, which are the most developed regions of China, thus capturing regional difference. Indu and *other_indu* are built on industrial dummies of CFPS, which are based on national classification standard (National Bureau of Statistics, 2017). Other_indu captures industries that traditionally tend to have higher labour payments, including finance, real estate, scientific research and technical service, and mining. Indu captures the concept of Internet industry. Non-zero values indicate that observed individuals work in information transmission, software and IT service, use computer for work, and have obtained at least junior secondary education. Last two requirements are made to

Table 1: The effect of Internet industry on college premium is due to ...

| | Hypothesis | Notes |
|---|---|---|
| H1 | Wide application of computer in the industry, which increases productivity compared to other industries | Hence, the effect of Internet industry is only a part of a larger trend of computer application widening led by college graduates. |
| H2 | A significant productivity breakthrough specific to the industry | This breakthrough is led by college graduates, and is strong enough to have a measurable impact on college premium. |
| H3 | Wide practice of efficiency wage in the industry | Although there has not been any significant productivity growth, Internet firms generally value college graduates highly and are willing to pay more to attract them. |
| H4 | Regional wage difference, as Internet firms mostly locate in developed areas which have higher wage level | Notably, this hypothesis does not rule out the possibility that factors intrinsic to Internet industry has contributed to college premium. One can argue that the high wages of Internet firms, while not directly observed, has contributed to the observed regional difference. |

define Internet industry more accurately, as national classification in CFPS also includes non-Internet firms like those in telecommunication.

Notably, while current requirements of indu ensure that all non-zero observations are from Internet firms, some individuals working in Internet industry are missed, since some Internet firms, such as Baidu, are not recorded as "information transmission, software and IT service" under national classification (Qichacha, 2024). This may explain why non-zero observations are few for indu. Consequently, indu is not an accurate estimator for the effect of Internet industry. Therefore, in H1 and H4, I do not separate Internet industry effect from comp and prov and I only consider intersections of edu with comp and prov, as in these cases indu cannot satisfactorily capture the separated effect. Moreover, this design cannot distinguish whether contribution from prov is intrinsic to Internet industry, so I only use H4 as indirect evidence.

Following Mincer (1974), I include gender, experience, squared experience, and marital status as control variables. Lang is another personal characteristic provided by CFPS. I also add a dummy on SOE, as difference in ownership can have considerable implication on wage level in China. No dummy on ethnicity is used, as over 91 per cent Chinese people are of Han ethnicity (National Bureau of Statistics, 2010). Due to data unavailability, no dummy on hukou is used. Fortunately, as most college graduates and Internet industry workers would have urban hukou, omitting hukou would not be a problem.

Table 2: Description of Variables

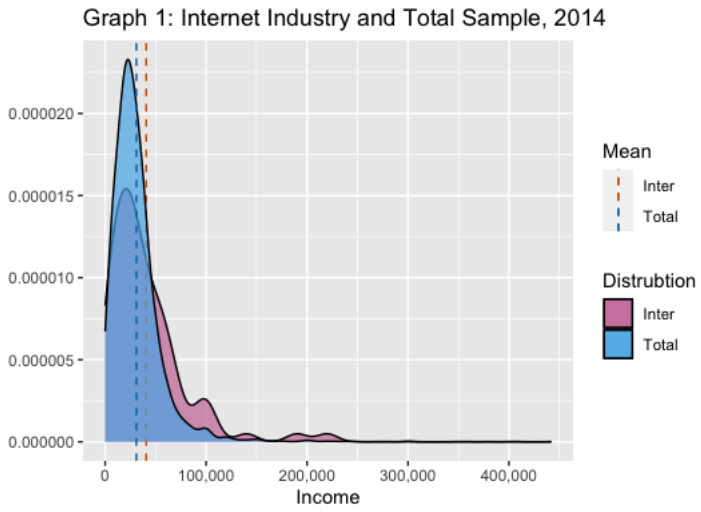
| | Symbol | Description |
|---|---|---|
| **Dependent Variable** | income | Natural logarithm of annual wage income after tax, unit in CNY (yuan). |
| **Main Independent Variable** | edu | Dummy variable for college education. 1 if individual has received college or above education (including college, bachelor's degree, master's degree, and doctor's degree), and 0 if not. |
| | comp | Dummy variable for computer usage. 1 if individual uses computer in work, and 0 if not. |
| | indu | Dummy variable for Internet industry. 1 if individual works in Internet industry, and 0 if not. |
| | other_indu | Dummy variable for traditional high-pay industries. 1 if individual works in such industries, and 0 if not. |
| | prov | Dummy variable for province. 1 if individual works in developed provinces of China, and 0 if not. |
| **Control Variables** | gender | Dummy variable for gender. 1 if individual is male, and 0 if individual is female. |
| | exp | Working experience of individuals. |
| | exp2 | Squared experience. |
| | marr | Dummy variable for marriage. 1 if individual has been married at least once, and 0 if individual has not. |
| | lang | Dummy variable for foreign language. 1 if individual uses at least one foreign language in work, and 0 if not. |
| | employer | Dummy variable for SOE. 1 if individual works in state-owned enterprises, governments, or other organisations affiliated to the state, and 0 if not. |

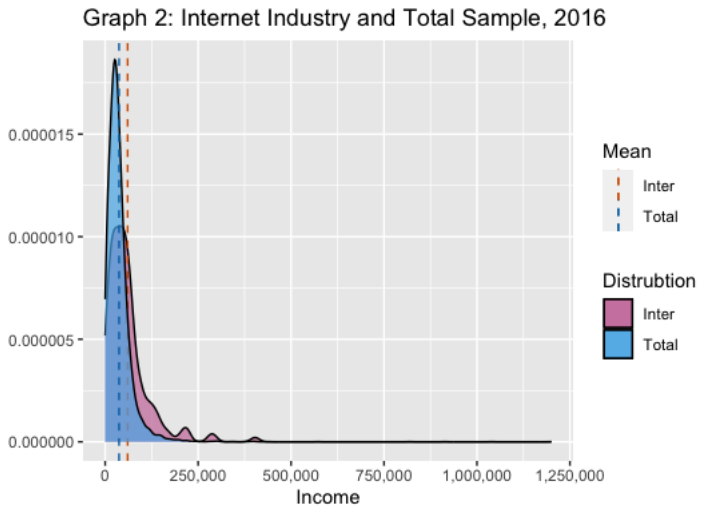
# Descriptive Statistics

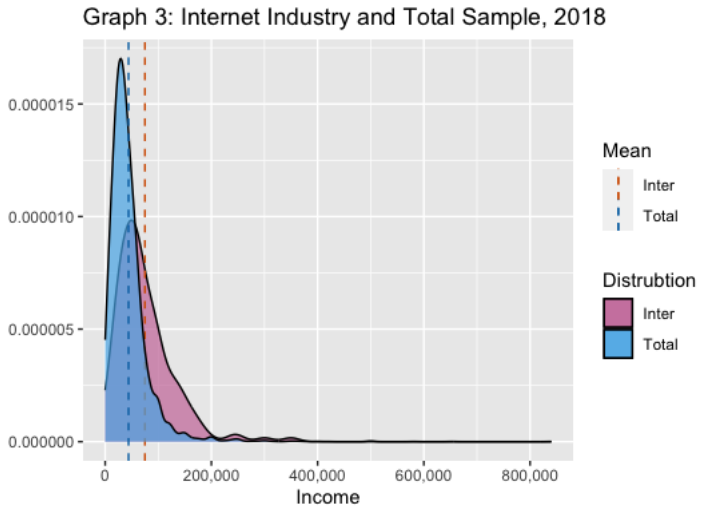
Table 3 compares mean of selected dummies and income, from the entire sample and Internet industry subsample. Internet workers are significantly more likely to obtain college above education, and on average earn much more than workers in other industries. Indeed, from Graph 1-4, compared to the total sample, Internet industry pays more at every income level. Moreover, Table 3 also suggests that Internet firms are more likely to locate in developed regions, aligning with H4.

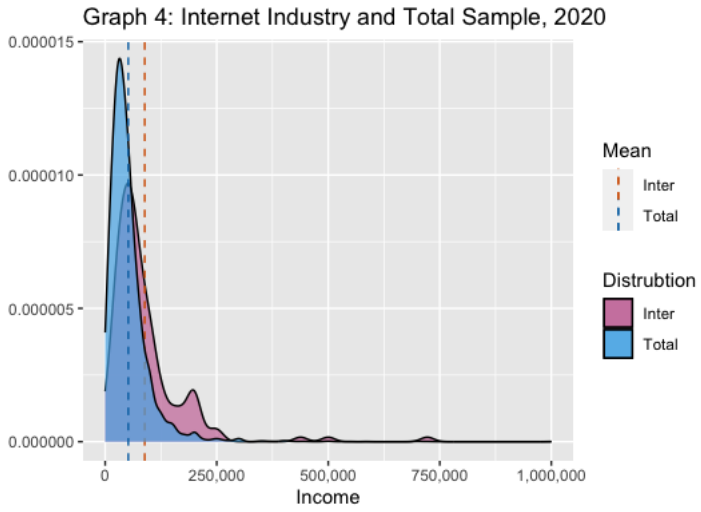
Table 3: Profile for workers in Internet industry

| | income | | edu | | comp | | prov | | $N$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total | Inter | Total | Inter | Total | Inter | Total | Inter | Total | Inter |
| **2014** | 30971 | 40932 | 0.223 | 0.646 | 0.390 | 1.000 | 0.316 | 0.402 | 7394 | 81 |
| **2016** | 37387 | 59697 | 0.255 | 0.695 | 0.452 | 1.000 | 0.317 | 0.461 | 8949 | 164 |
| **2018** | 43988 | 74651 | 0.293 | 0.755 | 0.465 | 1.000 | 0.300 | 0.445 | 9024 | 154 |
| **2020** | 51832 | 88354 | 0.342 | 0.774 | 0.505 | 1.000 | 0.283 | 0.439 | 8118 | 163 |



# Regression Models

In my regression models, the effects of variables change with time. Hence, for every variable, four effects are estimated, corresponding to four samples drawn from different years. This allows me to track the evolution of Internet industry effect over time and omit time fixed effects.

$$income = \beta_0 + \beta_1 edu + \beta_2 comp + \beta_3 indu + \beta_4 other\_indu + \beta_5 prov + X^T\sigma \quad (1)$$

For each year's sample, as a benchmark, I first run a regression according to variables in Table 2 with no intersection term (Equation 1). This regression also checks the existence of college premium. Then, three regressions are run with intersection terms tailored towards proposed hypothesis.

## Productivity by Computer (H1)

$$income = \beta_0 + \beta_1 edu + \beta_2 comp + \beta_3 edu\_comp + \beta_4 prov + X^T\sigma \quad (2)$$

In Equation 2, edu_comp refers to the intersection of edu and comp, and X includes all control variables specified in Table 2. Prov is also used as a control. If college premium mainly comes from productivity gain of computer application led by college graduates, coefficient for edu_comp should be significant.

## Industry-Specific Breakthrough (H2) and Efficiency Wage (H3)

$$income = \beta_0 + \beta_1 edu + \beta_2 indu + \beta_3 other\_indu + \beta_4 edu\_indu + \beta_5 edu\_other\_indu + \beta_6 prov + X^T\sigma \quad (3)$$

In equation 3, edu_indu refers to the intersection of edu and indu, edu_other_indu refers to the intersection of edu and other_indu, prov is a control variable, and X includes all controls in Table 2. I assume industry-specific productivity breakthrough can only occur in Internet firms, and if efficiency wage towards college graduates does exist in Internet industry, its magnitude would not exceed that of traditional high-pay industries. Hence, if productivity breakthrough intrinsic to Internet industry has non-negligible impact on college premium, there should be significant coefficient for edu_indu and insignificant coefficient for edu_other_indu. If coefficients for both intersection terms are significant, then likely efficiency wage towards college graduates is how Internet industry has impacted college premium.

## Regional Difference (H4)

$$income = \beta_0 + \beta_1 edu + \beta_2 prov + \beta_3 edu\_prov + X^T\sigma \quad (4)$$

In Equation 4, edu_prov refers to the intersection of edu and prov, and X includes all controls in Table 2. If college premium is mainly contributed by regional difference, coefficient for edu_prov should be significant.

# Results and Discussions

Table 4 provides estimates from general regression (Equation 1). Edu always has significant impact on income, confirming existence of college premium. Estimates also demonstrate moderate increase in the size of premium, agreeing with Hu and Bollinger (2021).

Table 4: General Regression Results

| | 2014 | 2016 | 2018 | 2020 |
|---|---|---|---|---|
| edu | $0.365^{***}$ | $0.357^{***}$ | $0.394^{***}$ | $0.419^{***}$ |
| | (.091) | (.041) | (.023) | (.025) |
| comp | $0.223^{***}$ | $0.098^{**}$ | $0.295^{***}$ | $0.321^{***}$ |
| | (.021) | (.032) | (.020) | (.021) |
| indu | $-1.382$ | 0.045 | $0.128^{*}$ | 0.083 |
| | (.091) | (.114) | (.064) | (.062) |
| other_indu | 0.116 | $0.164^{**}$ | $0.117^{***}$ | $0.203^{***}$ |
| | (.037) | (.058) | (.030) | (.032) |
| prov | $0.382^{***}$ | $0.510^{***}$ | $0.364^{***}$ | $0.375^{***}$ |
| | (.020) | (.032) | (.018) | (.019) |
| $N$ | 7394 | 8949 | 9024 | 8118 |

Note: Standard errors are presented in parenthesis.

$^{*}\ p < 0.1$; $^{**}\ p < 0.05$; $^{***}\ p < 0.01$

Table 5 summaries results from the regression by Equation 2. Comp is always significant, yet edu_comp only has significant impact in 2018, implying that while productivity growth created by computer use does exist, it does not heavily rely on college graduates in 2014, 2018, and 2020.

Table 5: Regression Results for Equation 2

| | 2014 | 2016 | 2018 | 2020 |
|---|---|---|---|---|
| edu | $0.315^{***}$ | $0.297^{***}$ | $0.288^{***}$ | $0.376^{***}$ |
| | (.051) | (.059) | (.044) | (.045) |
| comp | $0.211^{***}$ | $0.075^{*}$ | $0.271^{***}$ | $0.316^{***}$ |
| | (.025) | (.036) | (.022) | (.024) |
| edu_comp | 0.07 | 0.114 | $0.147^{**}$ | 0.064 |
| | (.051) | (.059) | (.048) | (.050) |
| $N$ | 7394 | 8949 | 9024 | 8118 |

Note: Standard errors are presented in parenthesis.

$^{*}\ p < 0.1$; $^{**}\ p < 0.05$; $^{***}\ p < 0.01$

Table 6 summaries regression by Equation 3. Other_edu is always significant, but edu_other_indu is never significant, suggesting that while traditional high-pay industries have practiced efficiency wage

in sampled years, they do not purposely target college graduates. Meanwhile, indu and edu_indu are both significant for 2014, implying effect on college premium caused by an industry-specific productivity breakthrough.

Table 6: Regression Results for Equation 3

| | 2014 | 2016 | 2018 | 2020 |
|---|---|---|---|---|
| edu | $0.354^{***}$ | $0.363^{***}$ | $0.395^{***}$ | $0.410^{***}$ |
| | (.029) | (.043) | (.024) | (.025) |
| indu | $-0.474^{**}$ | 0.282 | 0.103 | $-0.007$ |
| | (.153) | (.206) | (.127) | (.130) |
| edu_indu | $0.516^{**}$ | $-0.338$ | 0.033 | 0.120 |
| | (.189) | (.245) | (.146) | (.147) |
| other_indu | $0.115^{*}$ | $0.154^{*}$ | $0.127^{**}$ | $0.176^{***}$ |
| | (.045) | (.073) | (.040) | (.043) |
| edu_other_indu | 0.002 | 0.025 | $-0.022$ | 0.058 |
| | (.081) | (.121) | (.062) | (.063) |
| $N$ | 7394 | 8949 | 9024 | 8118 |

Note: Standard errors are presented in parenthesis.

$^{*}\ p < 0.1$; $^{**}\ p < 0.05$; $^{***}\ p < 0.01$

Table 7 summaries regression by Equation 4. Prov and edu_prov are always significant, indicating that regional difference is constantly a major contributor to college premium, though its contribution is smaller in 2016 and 2018.

Table 7: Regression Results for Equation 4

| | 2014 | 2016 | 2018 | 2020 |
|---|---|---|---|---|
| edu | $0.305^{***}$ | $0.304^{***}$ | $0.376^{***}$ | $0.359^{***}$ |
| | (.031) | (.047) | (.026) | (.027) |
| prov | $0.335^{***}$ | $0.460^{***}$ | $0.340^{***}$ | $0.297^{***}$ |
| | (.023) | (.037) | (.021) | (.024) |
| edu_prov | $0.194^{***}$ | $0.178^{*}$ | $0.077^{*}$ | $0.217^{***}$ |
| | (.048) | (.073) | (.038) | (.040) |
| $N$ | 7394 | 8949 | 9024 | 8118 |

Note: Standard errors are presented in parenthesis.

$^{*}\ p < 0.1$; $^{**}\ p < 0.05$; $^{***}\ p < 0.01$

In summary, estimation results suggest that H4 always holds, H1 holds for 2018, H2 holds for 2014, and H3 never holds. Hence, I conclude that Internet industry has increased college premium in 2014 with an industry-specific productivity breakthrough and in 2018 with productivity growth from wider

computer application led by college graduates, and that efficiency wage is not the major channel how Internet industry has contributed to college premium. Notably, size of the effect on college premium is 0.516 in 2014, suggesting that ceteris paribus, college graduates earn 51.6 percentage more in Internet industry compared to any other industries. This is among the largest effects found in my regressions, which indicates that the effect of Internet industry is also economically significant. Additionally, since general regression reports weak significance for indu in 2018, possibly a breakthrough specific to Internet industry has also occurred during the time, only that edu_indu fails to capture the effect. Although there is insufficient evidence on the effect of Internet industry on college premium in 2016 and 2020, strong confirmation of H4 has provided indirect evidence for such effect in 2020.

(1499)

# Reference


Ao, X., 2009. *Introduction to endogeneity-related methods.*

Asadullah, M.N. and Xiao, S., 2019.
*Labor market returns to education and English language skills in the People's Republic of China: An update.* Asian Development Review, 36(1), pp.80-111.

Department for Environment Food & Rural Affairs (2023)
*National statistics: Total Factor Productivity of the United Kingdom agricultural industry in 2022* Available at https://www.gov.uk/government/statistics/total-factor-productivity-of-the-agricultural-industry

Ferré, J., 2009. '3.02 - Regression Diagnostics', in Brown, S.D. et al. (ed.) *Comprehensive Chemometrics.* Elsevier, pp. 33-89

Freedman, D.A., 2006. *On the so-called "Huber sandwich estimator" and "robust standard errors".* The American Statistician, 60(4), pp.299-302.

Guo, K. and Wang, Y., 2020. *College Wage Premium and Computer Use in China.* Available at SSRN 3613630.

Hu, C. and Bollinger, C., 2021.
*Effects of cohort size on college premium: Evidence from China's higher education expansion.* China Economic Review, 70, p.101700.

Ikeuchi, K., Fukao, K. and Perugini, C., 2024. *Which employers pay a higher college wage premium?.* International Journal of Manpower.

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

Mincer, J.A., 1974. Schooling and earnings. *In Schooling, experience, and earnings.* (pp. 41-63). NBER.

National Bureau of Statistics of China (2010) *Population Census of People's Republic of China.* Available at http://www.stats.gov.cn/tjsj/pcsj/

National Bureau of Statistics of China (2017) *Industrial classification for national economic activities.* Available at https://www.stats.gov.cn/sj/tjbz/gmjjhyfl/202302/P020230213400314380798.pdf

Qichacha (2024) *Baidu Inc.*, accessed 30 March 2024

Zhu, X., 2023. *China's Productivity Challenge.* In Presentation at The Arc of the Chinese Economy, the 2023 annual conference of the Center for the Study of Contemporary China, March.

# Appendix

## Data Source

The data are from China Family Panel Studies (CFPS), funded by 985 Program of Peking University and carried out by the Institute of Social Science Survey of Peking University.

Specifically, adult sample of 2014, 2016, 2018, and 2020 are used in the paper. As the 2020 CFPS data is published in June 2023, my paper would be among the first studies to use it.

## Technical Considerations in Regression Design

### Justification for Assumptions behind Equation 3

In the setup of Equation 3, occurrence of industry specific productivity breakthrough is limited to Internet industry only. This assumption is supported by the fact that Chinese economy of the 2010s does not feature significant overall productivity growth. According to Zhu (2023), between 2008 and 2023, average growth rate for is only 1% in China. In comparison, TFP growth rate in China is above 4% in average during 1978 and 2007 (Zhu, 2023), and is 3.4% between 2021 and 2022 for agricultural industry of UK (Department for Environment Food & Rural Affairs, 2023).

In the identification part of Equation 3, I state that if there are significant coefficients for both edu_indu and edu_other_indu, efficiency wage would be the channel through which Internet industry impacts college premium. Strictly speaking, with two coefficients being significant, it is possible that a productivity breakthrough of Internet industry has occurred simultaneously with wide practice of efficiency wage towards college graduates in other high-pay industries. Nevertheless, edu_other_indu is never significant in my regressions, so this possibility can be safely ignored. In fact, with the assumption that the level of efficiency wage is always higher in traditional high-pay industries, absence of significance results for edu_other_indu essentially suggests that efficiency wage towards college graduates is present neither in traditional high-pay industries nor in Internet industry.

### On Using Prov as a Control Variable

Although prov is designed as a main independent variable, I have used it as a control for both Equation 2 and Equation 3. This unusual practice is driven by the high level of regional difference in China, and the phenomenon that prov-controlled regressions often yield results significantly different from that of uncontrolled ones (Guo & Wang, 2020). Additionally, since I only intend to treat verification of H4 as indirect evidence, using it as a main independent variable elsewhere should not create much problem.

### On Using Standard Errors instead of Robust Standard Errors

According to Freedman (2006), robust standard errors do not differ from standard errors much when the regression is roughly accurate, but often suggest wrong conclusions when the underlying model is wrongly specified. Considering indu variable in my regressions can potentially create some misspecification and standard errors are statistically simpler to interpretate, I only report standard errors for all my regression results.

## Discussions about Regression Robustness

### On Multicollinearity

Since the number of observations for individual in Internet industry is not large (N ranges from 81 to 164), one may legitimately suspect that correlation among indu and other variables is nonnegligible. To address the issue, I compute variation inflation factor (VIF), which is a popular measure for correlation among variables, for all main independent variables across all regressions. As there are too many regressions and variables, I refrain from reporting details here. In short, results are satisfactory: VIF readings are mostly close to 1 and are always below 10, which suggests that multicollinearity is never a problem (Ferré, 2009).

### On Endogeneity

Possibly, main independent variables specified in my regressions are not independent from the error term. Intuitively, what major people have pursued in college may impact what industry they would later work in and what ability they would develop. For example, if an individual has studied computer in college, more likely he would be able to use computer and foreign language (namely English) in work, and want to choose Internet industry, as his college years have prepared him for these. Then, since I do not have a variable for individual's major in college, the above effect, would be included in the error term. This would make comp, lang, and indu variables dependent on the residual, hence creating endogeneity. Moreover, this endogeneity has no easy remedy. Without appropriate instrumental variables for my main regressors, I can neither avoid endogeneity with 2SLS methods nor check for endogeneity with Hausman test (Ao, 2009). The potential problem of endogeneity is thus left unattended to in my paper, and readers are hereby warned to treat the issue with caution.