A tenuous result: re-analysis of the link between internet-usage and young adults' driving-licence holding

Comments on 'Recent changes in the age composition of drivers in 15 countries'

(paper appeared in *Traffic Injury Prevention* in 2011)

Postprint of:

Le Vine, S., Latinopoulos, C., Polak, J. (2013) A tenuous result: re-analysis of the link between internet-usage and young adults' driving-licence holding. Comments on 'Recent changes in the age composition of drivers in 15 countries'. *Traffic Injury Prevention*. http://dx.doi.org/10.1080/15389588.2013.793583

Scott Le Vine[1]
Research associate, Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London, Exhibition Road, London, SW7 2AZ, UK
Telephone: +44 20 7594 6105
Fax: +44 20 7594 6102
Email : slevine@imperial.ac.uk

Charilaos Latinopoulos
Doctoral candidate, Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London, Exhibition Road, London, SW7 2AZ, UK
Telephone: +44 20 7594 6100
Fax: +44 20 7594 6102
Email : charilaos.latinopoulos10@imperial.ac.uk

John Polak
Director of Research, Department of Civil and Environmental Engineering
Chairman, Centre for Transport Studies
Imperial College London
Telephone: +44 20 7594 6089
Email: j.polak@imperial.ac.uk

[1] Corresponding author

1. Overview of Sivak and Schoettle (2011)

Sivak and Schoettle (2011) report a cross-national comparison of trends in driving-licencing rates amongst adults of different ages. The majority of the paper is devoted to presenting descriptive results. It is reported that in eight of the 15 countries studies, there has been a decrease in driving-licencing rates amongst young people (aged 20 – 24) and a corresponding increase amongst older adults, whilst in the other seven countries there has been an increase in licencing rates amongst adults of all ages. As crash rates vary strongly by age-of-driver, there are potentially large implications for future trends in road casualties.

The descriptive results are sound. The abstract states, however, that the primary implication of the study is based on a multivariate analysis: '*The results of the analysis are consistent with the hypothesis that access to virtual contact reduces the need for actual contact among young people*'.

This finding arises from the parameters estimated by linear regression with the various countries as the cases. Ten candidate country-level descriptors were used as independent variables. Stepwise linear regression with backward elimination, a data mining technique, was used to search amongst the 1,023 possible specifications. Four independent variables were retained in the preferred specification (Gross National Income at purchasing power parity, median age of the population, the percent of the population living in settlements of more than one million people, and the number of internet users per 100 people.)

Whilst the estimated regression parameters are not reported, what is reported is that the number of internet users per 100 [all age] population has a negative association with young adults' rate of licence-holding with a t-statistic of -3.33. This result forms the basis for what is reported as the main implication of the study, namely the negative association between internet usage and young adults' licence-holding.

This finding has been widely-cited within the scientific literature assessing the 'peak car' trend – the observation of a leveling-off or decline in per capita car driving mileage in many developed countries after decades of growth (Litman 2012, Puentes 2012, Delbosc and Currie 2013, Taylor et al. 2013). This result has also been covered by national general-readership media and referenced by the authors in public interviews (Weissman 2012, Alcindor 2012, Buccholz and Buccholz 2012).

It is argued in the remainder of the present paper that this key finding is questionable. Due to the interest it has generated it merits closer scrutiny to avoid misinterpretation by researchers and policymakers. It is relevant to note that other authors, using different analysis strategies, have reported more recent empirical findings that are different than Sivak and Schoettle's result (e.g. Delbosc and Currie 2013, Taylor et al. 2013). Delbosc and Currie (2013) carried out online-format discussion forums with a small sample of young adults, and report that within their sample '*Electronic communications were seen as a supplement to face-to-face contact, not a replacement for car travel*'. Taylor and colleagues (2013) employed linear regression techniques to analyse travel diary data from the 2009 National Household Travel

Survey (and predecessor surveys in 2001 and 1990). The authors report that their models '*suggest that daily web use is actually associated with increased PMT [person-miles of travel] across all age categories*'.

2. Weaknesses of the analysis

This section presents the issues that have been identified in the findings regarding licence-holding and internet usage.

First, no substantive argument is presented which would lead to the hypothesis that there is a negative link between internet usage and licence-acquisition by young adults. Nor is there any reference to the body of literature on the links between telecommunications and personal mobility; the hypothesis of a negative link is asserted without appeal to either a plausible argument or the extant literature. The literature is clear that the functional relationships between the two are complex and multifaceted, and that whether the net effect is substitution or complementarity is ambiguous (Salomon 1986, Mokhtarian 2002). Rather than a reasoned case to support testing the hypothesised negative link, an algorithmic variable-selection technique is instead employed to identify any statistical relationships that can be found in the data. Shmueli (2010) points out the distinction between *explanatory* and *predictive* modelling, and the appropriateness of algorithmic variable-selection methods for the latter but not the former. What is problematic is that such a method is employed despite the aim of the analysis being to explain the phenomenon under study (the rate of licence-holding amongst young adults).

Second, the analysis is missing important variables that are clearly relevant, such as changes in licence-acquisition testing regimes, prices of motor insurance, fuel, driving lessons, etc. This highlights the importance of, prior to quantitative analysis being undertaken, establishing reasoned hypotheses to explain as fully as possible the behaviour under study and then compiling the data that are required for rigorous testing. We focus here on the effect associated with internet-usage, but note that this argument applies to the other explanatory variables as well. It is not clear, for instance, why one would expect a positive link between young people's licencing and the percentage of people living in cities of over a million people, as reported by Sivak and Schoettle. Indeed one may reasonably expect the opposite, as car use levels tend to be lower in large cities than elsewhere.

Third, the use of data-mining techniques to select a preferred specification from a large set of candidate specifications runs a substantial risk of Type I errors – interpreting spurious correlation as a true effect. The significance levels estimated by stepwise regression are known to be biased low due to selection bias, thus the F-statistic reported by Sivak and Schoettle [$F(4, 10) = 10.20$, $p<0.01$] cannot be relied on as this statistic does not follow the required distribution. Rencher and Pun (1980) document, via simulations performed with small sample sizes and large sets of [random] candidate independent variables, that large expected r^2 values are obtained – which if obtained from a pre-specified linear regression would be highly-significant. For completeness we note that the F-statistic for the full

regression where all ten candidate variables are entered is not significant: $F(10,4) = 2.65$, $p=0.18$.

To place into context the apparent effect Sivak and Schoettle attribute to internet-usage, we performed exhaustive all-subsets regression with the dataset and candidate independent variables. It was found that of the 512 specifications that included the number-of-internet-users-per-100-population as an independent variable, this effect was estimated to be negatively-signed and statistically significant in only 51 (just under 10% of the regressions). This is shown graphically in Figure 1.

There is a body of statistical literature documenting the biases associated with stepwise regression methods, particularly with small sample sizes. Harrell (2001), for instance, writes: "*neither univariable screening nor stepwise variable selection in any way solve the problem of 'too many variables, too few subjects,' and they cause severe biases in the resulting multivariable model fits…the method [stepwise regression] yields standard errors of regression coefficient estimates that are biased low.*" Miller (1990) states: "*the application of standard [statistical] theory can be very misleading in such cases when the model has not been chosen* a priori*, but from the data. There is widespread awareness that considerable overfitting occurs*".

Harrell further provides guidance that the required sample size for reliable linear regression is typically a multiple of at least ten (preferably twenty) times the number of independent variables, when the independent variables are continuous (as in the case of Sivak & Schoettle 2011). As noted above in the analysis at issue there are fifteen observational units (countries) and ten candidate independent variables (even just the set of four independent variables that remain after the stepwise procedure, which is not the appropriate number of degrees-of-freedom for significance calculations, would indicate that a sample size is required that is comfortably larger than fifteen observations).

A conventional approach for assessing whether a regression model is 'overfitting' the estimation data is to remove part of the sample from the estimation dataset and to then assess the accuracy of predictions made for this 'holdout' sample. The small sample size precludes such a strategy in the present study, however. As an alternate strategy, we ran the stepwise regression technique from the original paper 15 times, in each instance using a sample size of 14 and systematically excluding a different country in each iteration. The output specification contained the internet-usage indicator in only 3 of the 15 iterations.

Fourth, even if it were to be accepted that a stepwise regression technique is appropriate, the analyst faces an arbitrary decision of whether to employ backward-elimination or forward-selection of explanatory variables. The former involves an initial iteration with all explanatory variables followed by incremental removal of insignificant effects, whilst in the latter case the explanatory variables are incrementally added into an initially-empty set of explanators. In Table 1 we show that the preferred specification using stepwise linear regression with forward-selection retains only one explanatory variable (the ratio of vehicles

per 1,000 people). It is cause for concern that the effect due to internet-usage is excluded from the preferred model form when this minor and arbitrary change in specification is made.

Fifth, the inadequate sample size is an artefact of having selected a weak experimental design strategy to study the issue at hand. Researchers are not limited to working with cross-sectional country-level data from a small set of countries. Appropriate microdata (where the individual person is the unit of analysis) are widely-available which can provide much more statistical information for this line of enquiry (e.g. datasets from travel surveys, omnibus social surveys, online-activity surveys, etc.). Beyond a strategy of appealing to microdata, another relevant dimension of variation is the passage of time. A country-level comparison would be more compelling if it were longitudinal – i.e. were it to provide evidence that the rate at which young adults acquire licences decreased concurrently with the increase in online-activity amongst young adults, after suitable correction for confounding effects. Alternate experimental design strategies, such as the two listed here, would admittedly be more resource-intensive, but this would need to be considered in light of the stronger evidence they would provide.

Sixth, there are specific weaknesses in the statistical analysis. It is inappropriate to use a proportion as a dependent variable in a linear regression, due to its bounding between zero and unity. When the percentage of young adults that hold a driving licence is transformed into log-odds (a standard monotonic transformation with the property of being unbounded from both above and below), the preferred specification identified by the same technique (stepwise regression with backward elimination) does not contain as an explanatory variable the internet-usage indicator (see Table 1). This is an improvement in model specification, and it is further cause for concern that the key finding is not robust to it.

3. Conclusions

The recent paper on which we comment has become an influential part of the scientific literature and contributed to the wider public discourse regarding the relationship between new telecommunication technologies and personal travel.

Of the various results in Sivak & Schoettle (2011), the authors draw attention to the negative association found between internet usage and young people's driving-licence holding; it is highlighted as the main implication of the study and has been covered by national media (USA Today, New York Times, etc.). We show that this statistical association is substantially more tenuous than presented in the original paper, a finding of note due to conflicting results emerging from other studies with different design strategies.

Establishing what has caused an increased share of young adults to delay (or forego completely) acquiring a driving licence is an issue with broad public-policy implications; further research to better understand the determinants of observed trends is urgently required.

References

Alcindor, Y. (2012) *Research shows that teens in no hurry to be behind the wheel*. Accessed 15 Feb. 2013 via: http://usatoday30.usatoday.com/NEWS/usaedition/2012-03-15-Putting-off-driving_ST_U.htm

Buccholz, T.G., and Buccholz, V. (2012) *The go-nowhere generation*. Accessed 15 Feb. 2013 via: http://www.nytimes.com/2012/03/11/opinion/sunday/the-go-nowhere-generation.html?_r=0

Delbosc, A., Currie, G. (2013) *Investigating attitudes towards cars among young people using online discussion forums*. Paper presented at 2013 Annual Meeting of the Transportation Research Board.

Harrell, F. E. (2001) *Regression modeling strategies with application to linear models, logistic regression, and survival analysis*.

Litman, T. (2012) *Current mobility trends: Implications for Sustainability*. In: *Keep moving, towards sustainable mobility*, edited by Bert van Wee.

Miller, A.J. (1990) *Subset selection in regression*. Chapman and Hall, London.

Mokhtarian, P. (2002) *Telecommunications and Travel: The Case for Complementarity*. Industrial Ecology, 6 (2) p.43-57.

Puentes, R. (2012) *Have Americans hit peak travel? A discussion of the changes in US driving habits*. OECD International Transport Forum Discussion Paper 2012-14. Prepared for the Roundtable on Long-Run Trends in Travel Demand, 29-30 Nov. 2012.

Rencher, A.C., Pun F.C. (1980) *Inflation of R^2 in best subset regression.* Technometrics, 22, p.49-53.

Salomon, I. (1986) *Telecommunications and travel relationships: a review*. Transportation Research Part A. 20(3) p.223-238.

Shmueli, G. (2010) *To explain or predict?* Statistical Science, 25 (3), p.289-310.

Sivak, M., Schoettle, B. (2011) *Recent changes in the age composition of drivers in 15 countries*. Traffic Injury Prevention, 13 (2), p.126-132.

Taylor, B.D., Ralph, K., Blumenberg, E., Smart, M. (2013) *Who knows about kids these days? Analyzing the determinants of youth and adult mobility between 1990 and 2009*. Paper presented at 2013 Annual Meeting of the Transportation Research Board.

Weissman, J. (2012) *The dramatic 30-year decline of young drivers (in 1 chart).* Accessed 15 Feb. 2013 via: http://www.theatlantic.com/business/archive/2012/07/the-dramatic-30-year-decline-of-young-drivers-in-1-chart/260126/

Tables and Figures

	Stepwise linear regression with backward-elimination, as reported in Sivak and Schoettle (2011)	**Linear regression with all candidate explanatory variables entered**	**Stepwise linear regression with forward-selection**	**Stepwise linear regression with backward elimination. Log-odds transformation applied to dependent variable**
	r^2=0.81	r^2=0.87	r^2=0.38	r^2=0.65
Constant	-31.7	-101	32.5	-3.18
GNI PPP per capita (USD)	1.16e-3 (<0.01)	4.86e-4 (0.50)	Excluded	Excluded
Vehicles per 1,000 population	Excluded	-0.0142 (0.76)	Excluded	Excluded
Cars per 1,000 population	Excluded	0.106 (0.32)	Excluded	0.00781 (<0.01)
Vehicles per km of road	Excluded	-0.0345 (0.75)	0.0676 (0.01)	Excluded
Percent unemployed	Excluded	-0.311 (0.74)	Excluded	Excluded
Percentage of population living in settlements over one million population	0.291 (0.02)	0.605 (0.10)	Excluded	0.0303 (<0.01)
Median age of population	2.76 (<0.01)	2.23 (0.15)	Excluded	Excluded
Average number of years of school attended	Excluded	3.86 (0.34)	Excluded	Excluded
Mobile phone subscriptions per 100 population	Excluded	0.106 (0.59)	Excluded	Excluded
Internet users per 100 population	-0.738 (<0.01)	-0.489 (0.32)	Excluded	Excluded

Table 1: Results of linear regression analyses (p-values in parantheses)

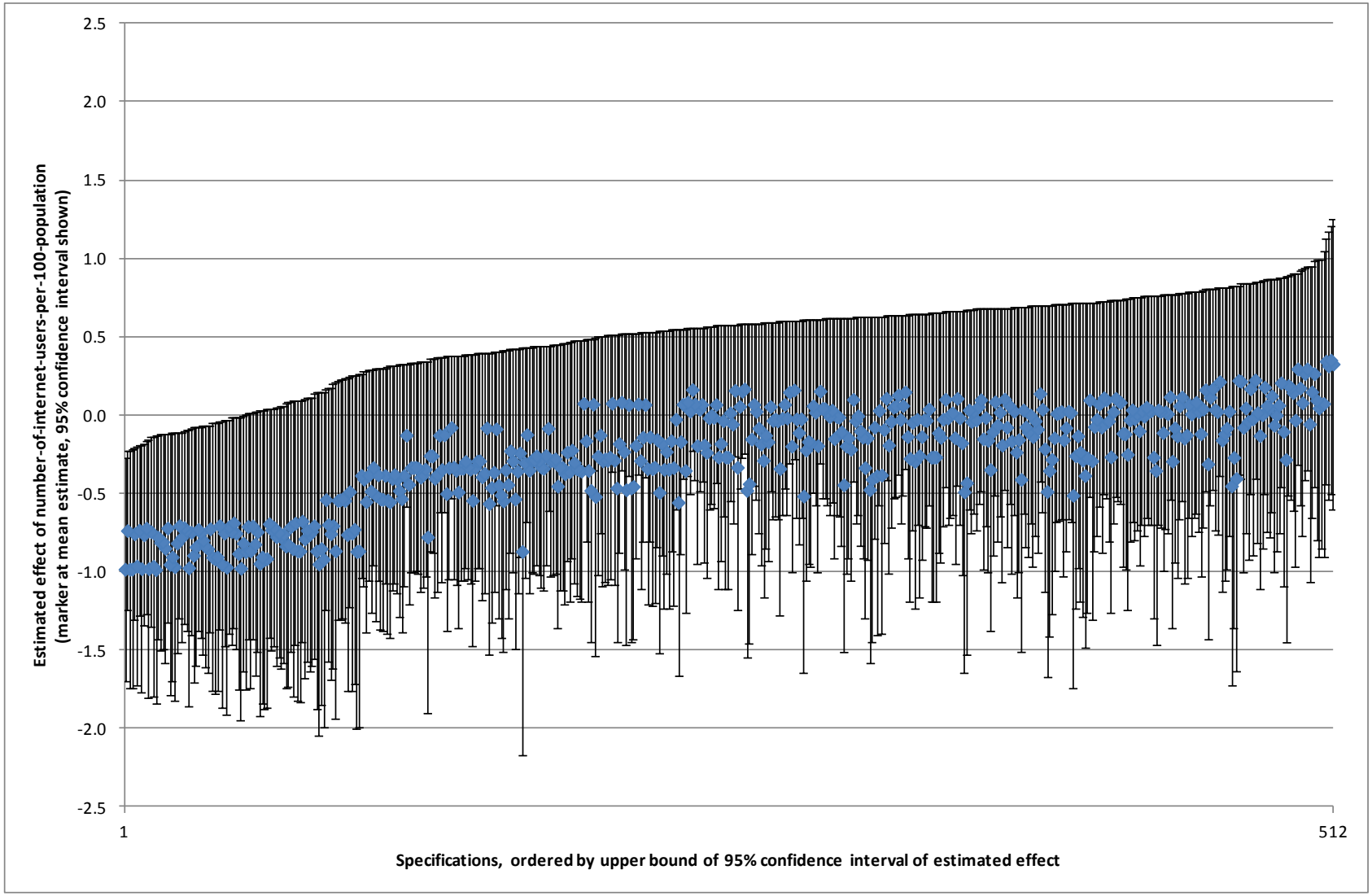

Figure 1: Estimated effects of number-of-internet-users-per-100-population from exhaustive all-subsets search of the 512 specifications that estimate this effect