Use of Multiple Linear Regression to Model Differences in High-Tech Exports across 74 Countries

***12 October 2014***

***Submitted by:***

Nayef Ahmad

ID: 122 1167 053

***Submitted to:***

Dr. A. K. M. Atiqur Rahman

Course Instructor

Eco 372: Introduction to Econometrics

North South University

Use of Multiple Linear Regression to Model Differences in High-Tech Exports across 74 Countries

# 1 Introduction

The objective of this paper is to use the multiple linear regression framework to explore the relationship between two variables: high-technology exports (*htexp*) and international internet bandwidth (*mbps*) for a cross-section of 74 countries. For full definitions of these variables, please see Section 3. It is recognized that this paper does produce a full model to completely explain or predict the level of high-technology exports, but certain areas for future work are highlighted based on these preliminary results. The key finding of this paper is that the elasticity of high-tech exports with respect to international internet bandwidth is 1.1323.

# 2 Motivation for the research

There are several benefits that high-technology manufacturing bring to national economies. In 1998, the National Science Foundation highlighted the following as the most important of these advantages: [[1]](#footnote-1)

* High-tech firms are associated with innovation. Firms that innovate tend to gain market share, create new product markets, and use resources more productively
* High-tech firms are associated with high value-added production and success in foreign markets, which helps to support higher compensation to the workers they employ
* Industrial R&D performed by high-tech industries has other spillover effects. These effects benefit other commercial sectors by generating new products and processes that can often lead to productivity gains, business expansions, and the creation of high-wage jobs

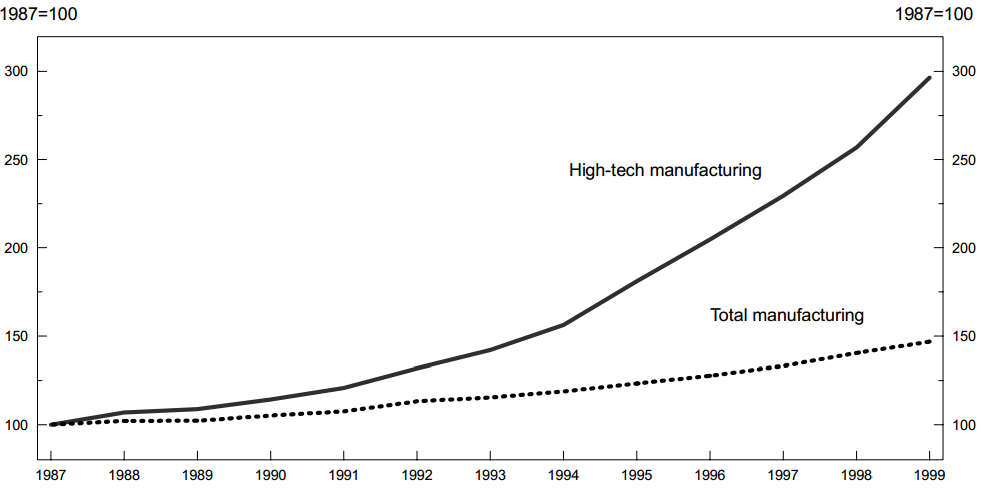


Figure 1: Index of output per hour in high-tech manufacturing and total manufacturing

(Source: U.S. Bureau of Labour Statistics)

Another study by the U.S. Bureau of Labor Statistics concentrated on the high-technology manufacturing sector in the U.S., and showed that productivity gains in high-technology manufacturing were rapidly exceeding those in the rest of the economy, as shown in the graph above.[[2]](#footnote-2)

Hence, it is clear that the important variables such as total amount of high-technology manufacturing, and of exports of high-technology manufactured goods, are worth exploring. In the present paper, this aspect of each nation’s economy is related with the total amount of international Internet bandwidth available to the country. Over the past few years, a separate body of literature has arisen that shows that broadband internet has multiple benefits for nations, through both direct and indirect channels. Readers interested in this literature may consult Katz (2012)[[3]](#footnote-3) for further details.

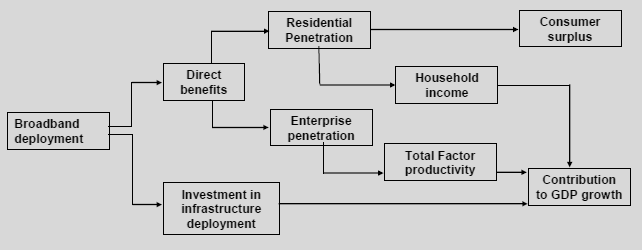


Figure 2: Channels through which broadband Internet benefits national economies (Source: Katz, 2012)

# 3 Key variables

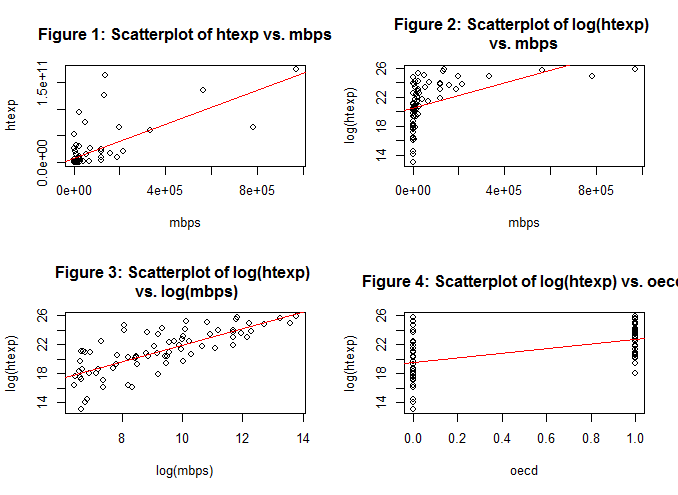
Three key variables were used in this analysis:

* *htexp*: the volume of high-technology exports in current U.S. dollars. The year 2004 was used for all 74 countries. This data was obtained from the World Bank’s online database, [[4]](#footnote-4) and is included in the appendix. The World Bank defined high-technology products as those with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery.
* *mbps:* thetotal volume of international internet bandwidth available to each country, in Mbps. Data form the year 2004 was used where available, or 2005 in a small number of cases. This data was obtained from the online database Nation Master.[[5]](#footnote-5)
* *oecd:* given large differences in high-tech exports across countries, it was hypothesized that the regression fit could be improved by introducing a dummy variable that records whether the country is a member of the OECD or not. The scatterplot in Figure 3, in particular, showed that this might be the case, as it seems to show that more than one line with the same slope, but different intercepts can be fitted to the data. The 34 OECD countries are among the most highly developed, and it is reasonable to hypothesize that their economies focus more on high-technology manufacturing, on average.

The regression analysis was performed using the R statistical computing environment. Samples of the raw output are given in the Appendix.

# 4 Results

Five regressions were done as part of this analysis, one using the variable *htexp*, and four using the transformed variable *log(htexp)*. Scatterplots for four of the five regressions are shown below.



It is clear that the scatterplot in Figure 2, of log(htexp) vs. mbps does not result in a good fit, and the results of this regression can safely be ignored, but they will be presented here for the sake of completeness.

The results of the five regressions conducted are summarized in the two tables below. In all cases, the notation “\*” represents significance at the 5% level, while “\*\*” and “\*\*\*” represent significance at the 1% and 0.1% levels.

The final model arrived at in this analysis is given below, from Model 3.

|  |  |
| --- | --- |
| **Table 1: Regression of htexp on mbps** | |
| ***Variable*** | ***Model 1*** |
| Intercept | 7.798e+09 \* |
| mbps | 1.605e+05 \*\*\* |
|  |  |
| Root MSE (72 df) | 2.727e+10 |
| R2 | 0.4811 |
| Adjusted R2 | 0.4739 |
| F-value ( 1 and 72 df) | 66.76 |
| p-value for F-statistic | 7.368e-12 |
| Number of observations (N) | 74 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 2: Regression of log(htexp)** | | | | |
| **Variable** | **Model 2** | **Model 3** | **Model 4** | **Model 5** |
| Intercept | 2.049e+01 \*\*\* | 10.6260 \*\*\* | 11.0207 \*\*\* | 19.6226 \*\*\* |
| mbps | 8.670e-06 \*\*\* |  |  |  |
| log(mbps) |  | 1.1323 \*\*\* | 1.0703 \*\*\* |  |
| oecd |  |  | 0.3844 | 3.1346 \*\*\* |
|  |  |  |  |  |
| Number of observations (N) | 74 | 74 | 74 | 74 |
| Root MSE (72 df) | 2.67 | 1.982 | 1.99 | 2.576 |
| R2 | 0.2202 | 0.5703 | 0.5727 | 0.2743 |
| Adjusted R2 | 0.2094 | 0.5643 | 0.5607 | 0.2642 |
| F-value ( 1 and 72 df) | 20.34 | 95.56 | 47.58 | 27.21 |
| p-value for F-statistic | 2.468e-05 | 7.64e-15 | 7.766e-14 | 1.681e-06 |

# 5 Interpretation of results

## Model 1: htexp ~ c + mbps

The scatterplot in Figure 1 provides the initial indication that there may be a linear relationship between *htexp* and *mbps*; however, it is clear that some transformation of the variables is necessary to fit the model with a linear model. The correlation between *htexp* and *mbps* is relatively high, 0.694, again indicating that a linear model may be appropriate. Table 2 shows the results of the regression of *htexp* on *mbps*. At first glance, the model seems to fit quite well, with mbps significant at the 0.1% level, and the adjusted R2 at a moderate value of 0.4739. However, by studying Figure 5,[[6]](#footnote-6) the plot of residuals from the regression against the fitted values of *htexp*, it is clear that the residuals are not behaving stochastically. This means that a certain amount of information relevant to predicting *htexp* is being retained in the residuals, and is a violation of one of the assumptions of linear regression. For this reason, Model 1 was rejected in this analysis.

## Model 2: log(htexp) ~ c + mbps

As mentioned above, it was clear that a transformation of the variables would be necessary to appropriately fit a linear model to the data. The first transformation attempted was from *htexp* to *log(htexp).* The resulting scatterplot is given in Figure 2. It is clear that this regression model (Model 2) results in a poor fit. As expected, R2 (= 0.220) is significantly lower than for Model 1. Furthermore, as shown in Figures 6 and 7, the residual plots indicate very strongly that this is not a valid regression model.

## Model 3: log(htexp) ~ log(mbps)

This was the model that was finally judged to be most appropriate for the data. In this case, both initial variables were subjected to log transformations. The resulting scatter plot is Figure 3, which shows a very significant improvement over the previous two scatterplots. The correlation between *log(htexp)* and *log(mbps)* is high, at 0.755, further indicating that this is a good candidate for the most appropriate model. Adjusted R2 has increased over Model 1, from 0.4739 to 0.5643. Both the intercept and *log(mbps)* are significant at the 0.1% level. Furthermore, the residual plots (Figures 8 and 9) show very significant improvement over those for the earlier models. A histogram of the residuals (Figure 13) also shows that there is no obvious skew, which offers some evidence of normality of residuals (a key assumption for the regression model).

It must be noted, however, that the residual plots indicate that there is still room for improvement in the model, as there seems to be a detectable relationship between the residuals and fitted values in Figure 8. This is most likely due to missing variables. As stated in Section 1, developing a full model for predicting *htexp* is beyond the scope of this paper, but if further work is done using these results and adding more significant variables, it is anticipated that the residuals will become more well-behaved.

As noted above, the regression equation from Model 3 is:

Since this is a log-log equation, the interpretation of the coefficient of *log(mbps)* is that it is the elasticity of the *htexp* withrespect to mbps. In other words, the key finding of this paper is that the elasticity of high-tech exports with respect to international internet bandwidth is 1.1323. This means that a 1% increase in internet bandwidth will lead to a 1.1323% increase in high-tech exports. It is worth noting that this means the relationship is elastic (i.e. elasticity > 1).

## Model 4: log(htexp) ~ c + log(mbps) + oecd

The scatterplot in Figure 3 gives some indication that there may be more than one line that can be fit through the data, where different lines have the same slope but different intercepts. This would mean that, given a certain volume of international bandwidth, there may be some categorical variable that explains why one particular country has a higher volume of high-tech exports than another. In this analysis, it was hypothesized that these differences could be explained by introducing a categorical variable into Model 3, to indicate whether each country is a member of the OECD or not. Evidence for the suitability of the *oecd* variable is also available from Figure 4, and Model 5, which show that the regression of *log(htexp)* on *oecd* gives a low but not insignificant adjusted R2 of 0.2642.

However, the results of Model 4 show that it is a poorer fit than Model 3. The *oecd* variable is not significant at any reasonable level, and adjusted R2 decreased from 0.5643 in Model 3 to 0.5607. Furthermore, examination of the residual plots (Figures 10 and 11) reveals that they are almost identical to those for Model 3 (Figures 8 and 9), which also indicates that the inclusion of the *oecd* variable is not improving the fit of the model in any way. For these reasons, Model 4 was not judged to be the most suitable for fitting the data.

## Model 5: log(htexp) ~ c + oecd

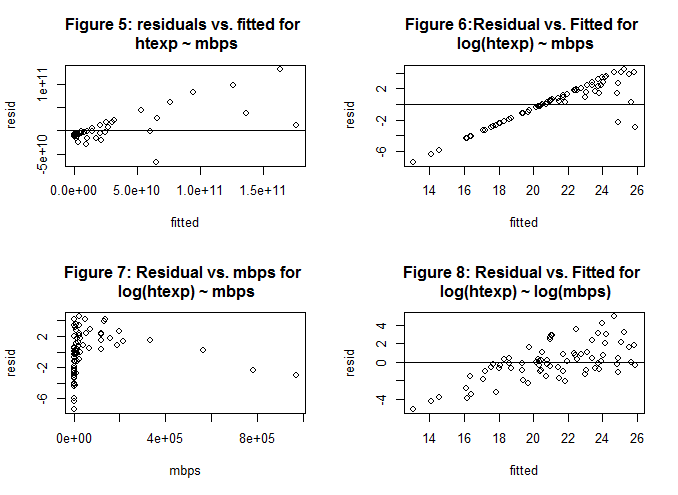
This model is purely used in assessing whether or not to include the variable *oecd* in Model 4. The variable was included in the end, because the adjusted R2 was low, but not insignificant.

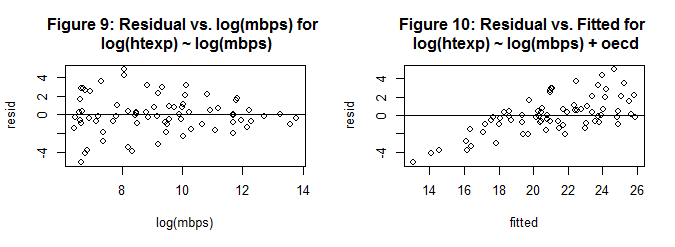
# 6 Conclusion

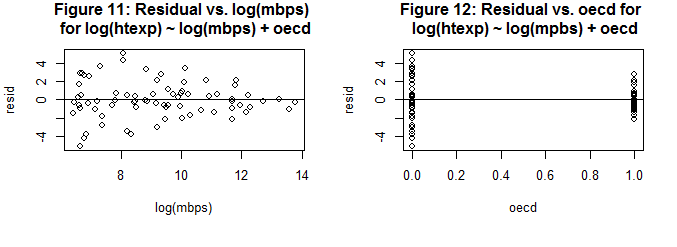
Thus, the conclusion of this analysis is the very preliminary result that the elasticity of high-tech exports with respect to the volume of international bandwidth across a cross-section of 74 countries is 1.1323. This is a qualified conclusion as there is clearly scope for adding more variables to the final model, to improve its fit to the data. This analysis also arrives at the preliminary result that classification of these 74 countries into the groups OECD and non-OECD is not significant in explaining/predicting differences in their volume of high-tech exports.

# Appendix

## Residual Plots

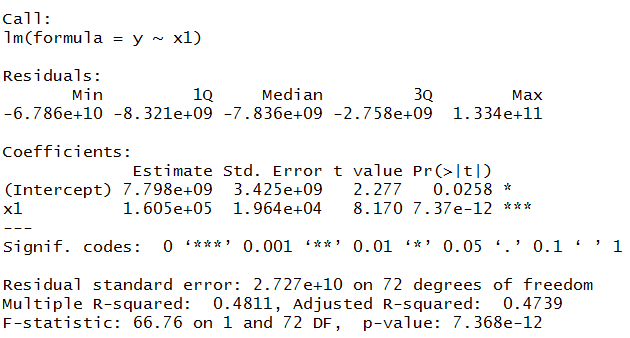






## C:\Users\Nayef\AppData\Local\Temp\ScreenClip.png

## Sample of raw output from R



1. National Science Foundation. http://www.nsf.gov/statistics/issuebrf/sib98319.htm [↑](#footnote-ref-1)
2. Bureau of Labor Statistics. http://www.bls.gov/opub/mlr/2002/03/art2full.pdf [↑](#footnote-ref-2)
3. Raoul Katz. 2012. *The Impact of Broadband on the Economy.* Available: http://www.itu.int/ITU-D/treg/broadband/ITU-BB-Reports\_Impact-of-Broadband-on-the-Economy.pdf [↑](#footnote-ref-3)
4. The World Bank. http://data.worldbank.org/indicator/TX.VAL.TECH.CD/countries [↑](#footnote-ref-4)
5. nationmaster.com. <http://www.nationmaster.com/country-info/stats/Media/Internet/International-Internet-bandwidth/Mbps> [↑](#footnote-ref-5)
6. Given in the Appendix [↑](#footnote-ref-6)