Analysis of growth in internet users

Xintong Chen, Shahbaz Khan, Junkun Peng, Chen Ma, Yanyu Zhou

School of Mathematics and Statistics, University of Glasgow

Introduction

As a worldwide public computer network, there are more and more people in the whole world using the internet as their daily tools. This research aims to find the relationship between network usage rate and different factors from technological progress and technology adoption, such as mobile phone subscription, technology development like computer usage performance, and economic development in the whole world.

Data processing and analysis

Four related datasets about internet access and technology development are collected. Including percentage share of the population using the internet by years; mobile phone subscriptions, measured as the number per 100 people by years; computer efficiency by measuring transistors per microprocessor by years; mobile phone subscriptions vs GDP per capita by years. We fit the linear regression model about the transistors per microprocessor and year to predict the data missing in the dataset to fill the NAN.

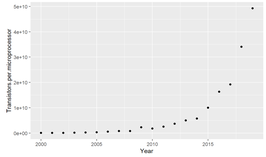


Figure 1. scatterplot

- We combined the four related datasets by years to discuss which factors might influence the share of the population with internet access. To select more suitable datasets, we choose the internet population's observations of more than zero and less than exp after 2007 (final observations count - 464)

- Relationship between the individuals using the internet and GDP per capita:
- By looking at the scatterplot, we can see a positive linear relationship. An increase in GDP per capita results in growth in individuals using the internet.

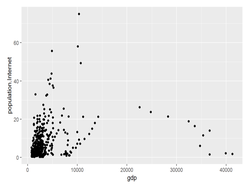


Figure 2. prediction scatterplot

Model

Let's consider pair wise correlations and relationships between with this set of explanatory variables and the response variable.

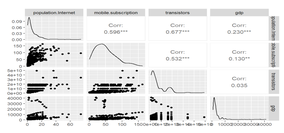


Figure 3. pairwise plot

- It is clear that the population with internet access is linearly related to all of the possible explanatory variables, and each explanatory variables are related to each other.

We use stepwise selection criterion rulers to determine which explanatory variables we need to include in the final regression model.

Model	AIC	BIC	R2	R2 adj
transistors	3057.6	3070.0	0.4587	0.4576
mobile transistors	2987.8	3004.4	0.5363	0.5343
mobile transistors gdp	2771.2	2791.6	0.5641	0.5610

Table 1. model choosing table.

Model 3, with all explanatory variables, has the highest R-square and R-square(adj), lowest AIC and BIC. So, the model with all three explanatory variables is selected for further analysis.
Firstly, consider the linear regression model:

$$population.Internet = \beta_0 + \beta_1 \times mobile.subscription + \beta_2 \times transistors + \beta_3 \times GDP$$

Then consider the log formation: the linear model with a natural transformation:

$$\log^{population.Internet} = \beta_0 + \beta_1 \times \log^{mobile.subscription} + \beta_2 \times \log^{transistors} + \beta_3 \times \log^{GDP}$$

Results and prediction

Compare the R square and adjusted R square of both two regression model:

Model	R2	R2 adj
linear	0.56	0.56
log-transform	0.75	0.75

Table 2. comparing table

- Since the R-square of log formation is greater than the normal linear model, which means the log formation model fits the data better than the normal one.

Therefore, the final model can be written as:

$$\log^{population.Internet} = -16.41 + 0.4 \times \log^{mobile.subscription} + 0.6 \times \log^{transistors} + 0.4 \times \log^{GDP}$$

Estimate	2.5%	97.5%
Intercept	-17.8203	-14.9964
lg.mobile.subscription	0.3098	0.4803
lg.transistors	0.5361	0.6656
lg.GDP	0.3232	0.4742

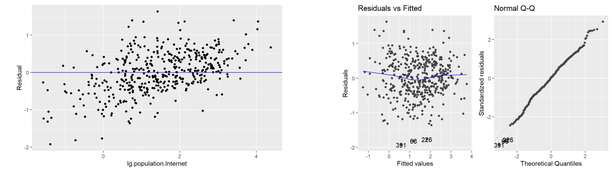
Table 3. Relevant confidence intervals table.

Also, we can see that the p-value is 2.2×10^{-16}, which is significantly less than 0.05.

When we use the final model to predict the future data about the population using the internet, the prediction data are not in the 95% confidence interval. Since the population using the internet might also be affected by other complex factors in the world. When the quantity of population using the internet is saturated, otherwise, the population using the internet around the world is constant even with the growth or decline of other three factors. The model we produced may not be suitable to predict future data.

Model Assumption

Then we check the assumptions of this log-formation regression model:



(a) Residual Independence (b) residuals vs fitted & Q-Q

Figure 4. assumptions checking plot

In Figure 4(a), it is clear that residuals are not independent. From residuals against fitted values plot in Figure 4(b), residuals seem randomly scattered around the zero line, so we can assume residuals have constant variance and mean zero. Also, we can see that residuals are normally distributed from the normal Q-Q plot.
Therefore, the final model with log transformation might be the optimal model for the selected datasets. We discuss the factors that affect the share of the population using the internet.

Conclusion

We can infer the positive relationships between the population using the internet and the other three explanatory variables from the model parameters. Moreover, there is a strong positive linear relationship between the transistors per microprocessor and the response variable among these three explanatory variables. There is a moderate relationship between the other two explanatory variables (mobile phone subscriptions and GDP) and the response variable.

Reference

https://ourworldindata.org/technology-adoptioninternet-access-technology
https://ourworldindata.org/technology-adoptionmobile-phone-adoption
https://ourworldindata.org/grapher/transistors-per-microprocessor
https://ourworldindata.org/grapher/mobile-phone-subscriptions-vs-gdp-per-capita