W203 Lab01 Broadband Exploration

Chandra Sekar, Bhuvnesh Sharma, Eugene Tang

Introduction

Research Question

With the recent rise of the Internet and its increasing importance in our lives, the availablility of access to the Internet has also become an increasingly large question and necessity in some societies. In this analysis, we look at broadband markets and in particular, three aspects of these markets and their relationships with each other:

- Price: how much does it cost to access the Internet
- Penetration: what fraction of customers have access to network service
- Speed: what rate can customers upload or download bits of data

We in particular consider this data in the context of open access policies. Much of the developed world has developed aggressive regulatory structures to compel network owners to increase penetration while there are some nations that do not. It is still an open debate on whether such policies are beneficial or harmful in price, penetration, and speed.

In this analysis we seek to tackle two main questions:

- Does a trade-off exist between network price, penetration, and speed?
- Is there evidence for beneifical effects of open access policies?

Dataset Setup (code)

Please see below sections to see why certain decisions were made in the preparation of the dataset.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
df_price = read.table("Price.csv", header = TRUE, sep = ",")
df_penetration = read.table("Penetration_Modified.csv", header = TRUE, sep = ",")
df_speed = read.table("Speed_Modified.csv", header = TRUE, sep = ",")
```

```r
# some dataset cleaning (see 'Data Quality Evaluation' for more details on why we
# did this)
drops_penetration <- c("X")  # extra column
df_penetration = df_penetration[, !(names(df_penetration) %in% drops_penetration)]
colnames(df_speed)[2] <- "Country.Code"

# convert numeric fields to the numeric type
convert_to_numeric = function(col) {
    return(as.numeric(sub("[\\$%,]", "", col)))
}

NON_ID_DATA_START <- 3  # columns 1 and 2 are the country and country code data
df_price[NON_ID_DATA_START:length(df_price)] = lapply(df_price[NON_ID_DATA_START:length(df_price)],
    convert_to_numeric)
df_penetration[NON_ID_DATA_START:length(df_penetration)] = lapply(df_penetration[NON_ID_DATA_START:leng
    convert_to_numeric)
df_speed[NON_ID_DATA_START:length(df_speed)] = lapply(df_speed[NON_ID_DATA_START:length(df_speed)],
    convert_to_numeric)
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```r
df_partial = full_join(df_penetration, df_price, by = c("Country", "Country.Code"))
df_full = full_join(df_partial, df_speed, by = c("Country", "Country.Code"))
```

Dataset Description

Our dataset comes in three csv files. One for price, penetration, and speed respectively. Each dataset contains observations on 30 countries, with one row for each country. Each dataset contains a variety of variables. Below we include tables of each variable, its type, and its description.

Each of the three datasets contains string columns to represent country and country code field. We use the country code field to join the datasets together since country code was unique across each row (though we could have used country as well). For conciseness, we exclude these two columns in the tables below. (TODO: check that country / country codes all match up)

Price Dataset

Our price dataset contained information on the cost to access different levels of Internet

Column	Interpretation

Data Quality Evaluation

Data Processing / Preparation

Univariate Analysis of Key Variables

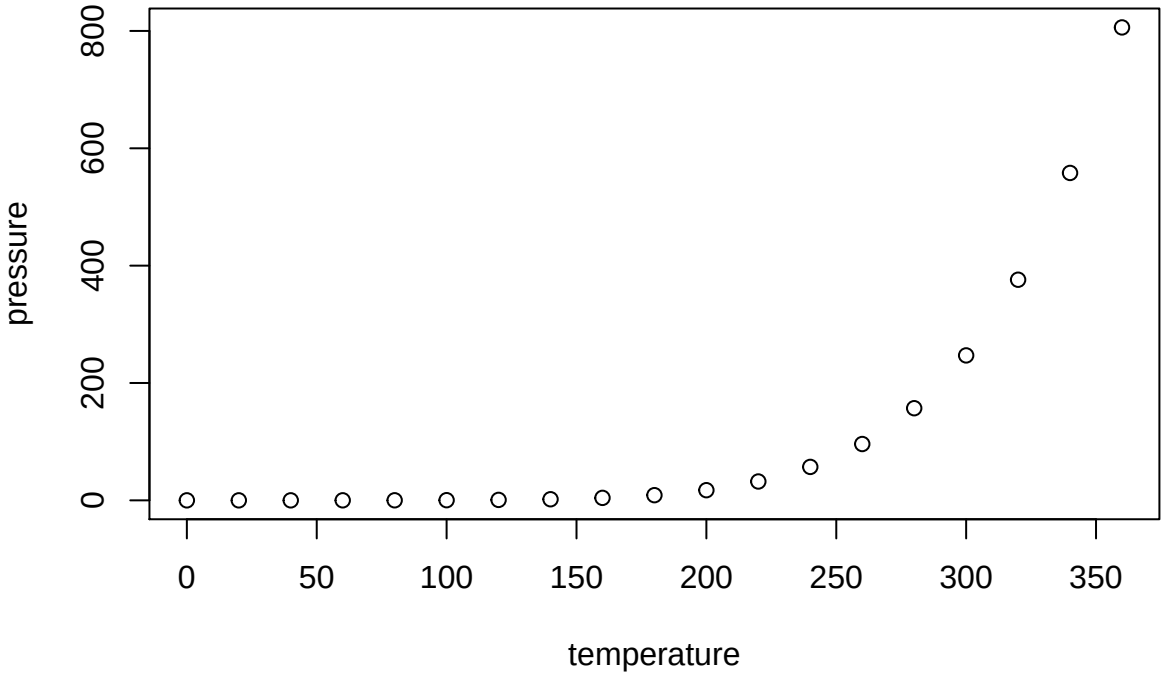
Analysis of Key Relationships

Analysis of Secondary Effects

Conclusion

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.