My short analysis

Vijay*

March 24, 2017

The results below are generated from an R script.

```
# First we load the packages
library(tidyr)
library(ggplot2)
library(Hmisc)
library(psych)

pers_exp <- USPersonalExpenditure

# Let's look at the data
pers_exp        #it looks like the row names are not actually in the matrix

##                       1940   1945  1950 1955  1960
## Food and Tobacco    22.200 44.500 59.60 73.2 86.80
## Household Operation 10.500 15.500 29.00 36.5 46.20
## Medical and Health   3.530  5.760  9.71 14.0 21.10
## Personal Care        1.040  1.980  2.45  3.4  5.40
## Private Education    0.341  0.974  1.80  2.6  3.64

                #let's fix that

expenditure_type <- rownames(pers_exp)
expenditure_type <- gsub( ' ' , '_' , expenditure_type) # Replace spaces with underscores
rownames(pers_exp) <- NULL
pers_exp <- as.data.frame(pers_exp)
pers_exp <- cbind(expenditure_type,pers_exp)

# Let's look at the data now
pers_exp        #better, but it doesn't look tidy to me

##      expenditure_type   1940   1945  1950 1955  1960
## 1    Food_and_Tobacco 22.200 44.500 59.60 73.2 86.80
## 2 Household_Operation 10.500 15.500 29.00 36.5 46.20
## 3  Medical_and_Health  3.530  5.760  9.71 14.0 21.10
## 4       Personal_Care  1.040  1.980  2.45  3.4  5.40
## 5   Private_Education  0.341  0.974  1.80  2.6  3.64

                #let's fix that


pers_exp <- gather( pers_exp , `1940` , `1945` , `1950` , `1955` , `1960` ,
```

*This report is automatically generated with the R package **knitr** (version 1.15.1).

```
        key = year, value =  expenditure)
pers_exp <- spread( pers_exp , key = expenditure_type , value = expenditure)

# Let's look at the data now
pers_exp        #better
```

```
##   year Food_and_Tobacco Household_Operation Medical_and_Health Personal_Care
## 1 1940             22.2                10.5               3.53          1.04
## 2 1945             44.5                15.5               5.76          1.98
## 3 1950             59.6                29.0               9.71          2.45
## 4 1955             73.2                36.5              14.00          3.40
## 5 1960             86.8                46.2              21.10          5.40
##   Private_Education
## 1             0.341
## 2             0.974
## 3             1.800
## 4             2.600
## 5             3.640
```

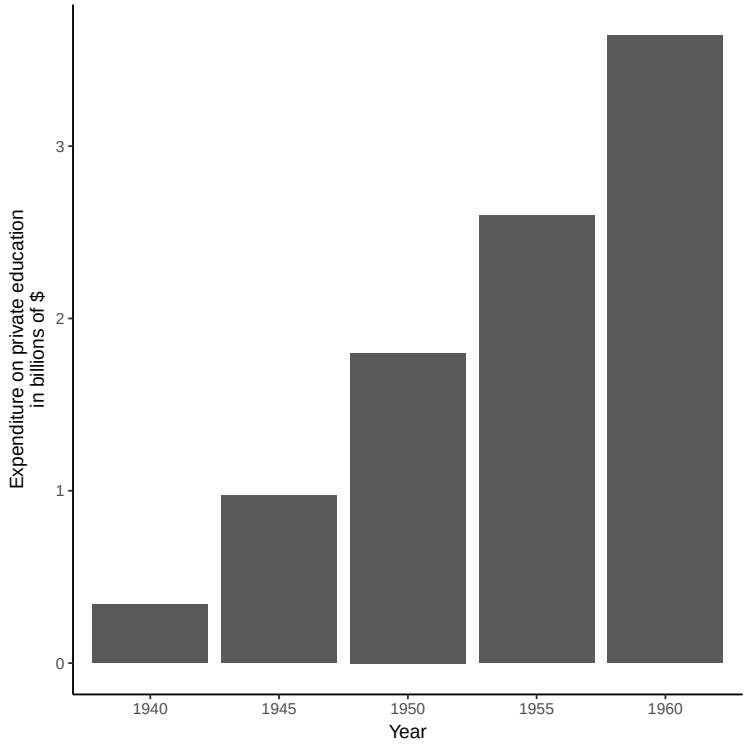
```
# We should now look at some descriptive stats
# Let's use the describe function from the psych package
psych::describe(pers_exp)
```

```
##                     vars n    mean    sd  median trimmed   mad     min     max range
## year*                  1 5 1950.00  7.91 1950.00 1950.00  7.41 1940.00 1960.00 20.00
## Food_and_Tobacco       2 5   57.26 25.12   59.60   57.26 22.39   22.20   86.80 64.60
## Household_Operation    3 5   27.54 14.71   29.00   27.54 20.02   10.50   46.20 35.70
## Medical_and_Health     4 5   10.82  7.00    9.71   10.82  6.36    3.53   21.10 17.57
## Personal_Care          5 5    2.85  1.66    2.45    2.85  1.41    1.04    5.40  4.36
## Private_Education      6 5    1.87  1.30    1.80    1.87  1.22    0.34    3.64  3.30
##                      skew kurtosis    se
## year*                0.00    -1.91  3.54
## Food_and_Tobacco    -0.19    -1.81 11.23
## Household_Operation  0.03    -2.01  6.58
## Medical_and_Health   0.35    -1.77  3.13
## Personal_Care        0.44    -1.58  0.74
## Private_Education    0.15    -1.88  0.58
```

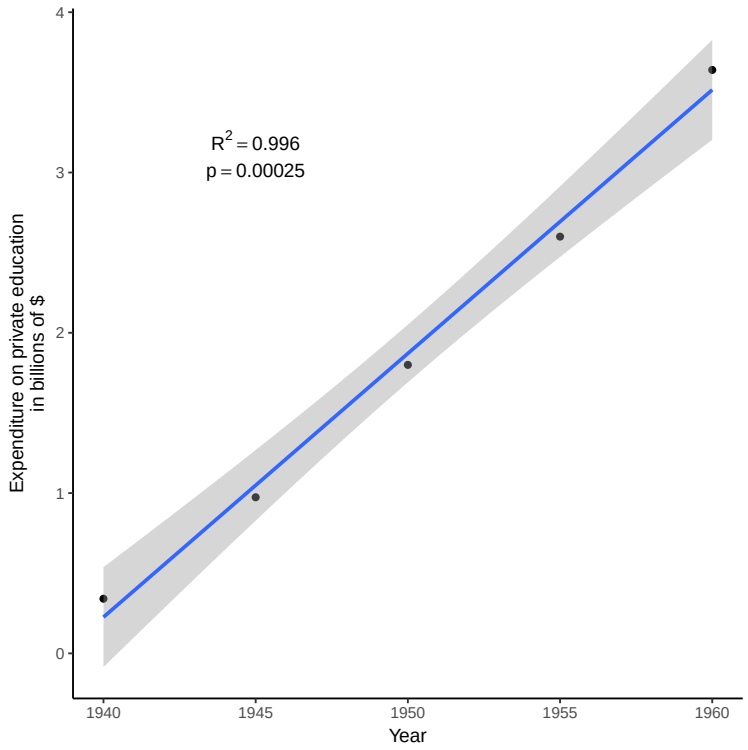
```
# Okay, that's nice but we really just want the Private Education info

psych::describe(pers_exp$Private_Education)
```

```
##    vars n mean  sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 5 1.87 1.3    1.8    1.87 1.22 0.34 3.64   3.3 0.15    -1.88 0.58
```

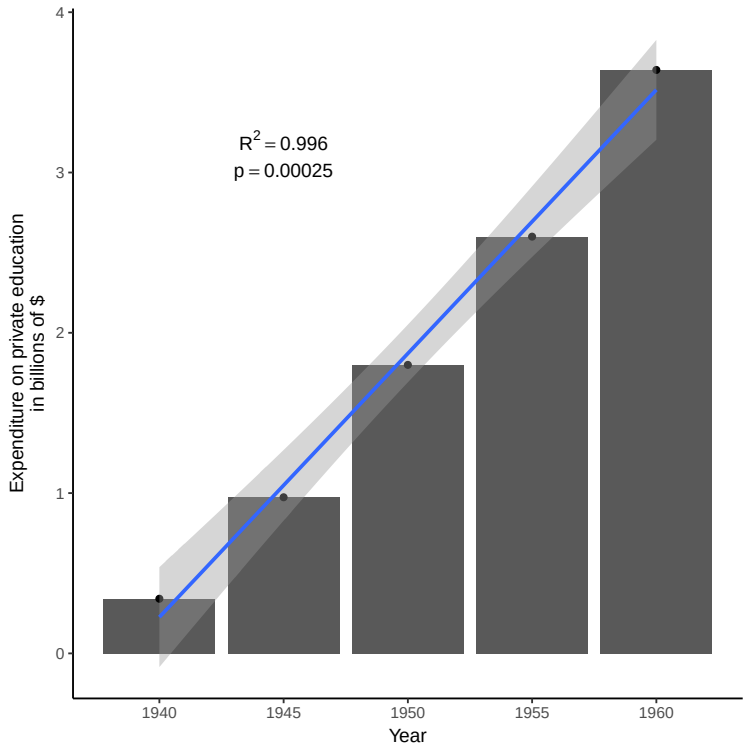
```
# This shows a difference between the minumum and maximum values
# with both the mean and median roughly in between.
# This would suggest a general change, but we cannot be sure of the direction.
# A plot might help with this

# Plot private education over time
ggplot( pers_exp , aes( year , Private_Education))+
        geom_bar(stat='identity')+
        labs( x = 'Year' , y = 'Expenditure on private education\nin billions of $') +
        theme_classic()
```




```
# This plot shows that expenditure on Private Education has
# increased between 1940 and 1960.

# Let's now see if the correlation between these values is significant

# First calculate correlation and p-value
cor_vals <- rcorr( pers_exp$year , pers_exp$Private_Education )

# Straight up scatter plot
ggplot( pers_exp , aes( x = as.numeric(year) , y = Private_Education)) +
  geom_point() +
  geom_smooth(method='lm') +
  annotate('text' , x = 1945 , y = 3.2 ,
    label = paste('R^2 == ' , round(cor_vals$r[1,2],3) , sep=''), parse = T) +
  annotate('text' , x = 1945 , y = 3 ,
    label = paste('p ==', round(cor_vals$P[1,2],5) , sep=''), parse = T) +
  labs(x = 'Year' , y = 'Expenditure on private education\nin billions of $') +
  theme_classic()
```




```
# Bar plot with scatter plot overlay
ggplot( pers_exp , aes( x = as.numeric(year) , y = Private_Education)) +
  geom_bar(stat='identity')+
  geom_point() +
  geom_smooth(method='lm') +
  annotate('text' , x = 1945 , y = 3.2 ,
    label = paste('R^2 == ' , round(cor_vals$r[1,2],3) , sep=''), parse = T) +
  annotate('text' , x = 1945 , y = 3 ,
    label = paste('p ==', round(cor_vals$P[1,2],5) , sep=''), parse = T) +
  labs(x = 'Year' , y = 'Expenditure on private education\nin billions of $') +
  theme_classic()
```



The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 3.3.2 (2016-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.2 LTS
##
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C               LC_TIME=en_GB.UTF-8
##  [4] LC_COLLATE=en_GB.UTF-8     LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C                  LC_ADDRESS=C
## [10] LC_TELEPHONE=C             LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] tidyr_0.6.1     knitr_1.15.1    Hmisc_4.0-1     Formula_1.2-1   survival_2.40-1
## [6] lattice_0.20-34 ggplot2_2.2.1   psych_1.6.12
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.10         RColorBrewer_1.1-2  plyr_1.8.4          highr_0.6
##  [5] tools_3.3.2          rpart_4.1-10        digest_0.6.12       base64_2.0
##  [9] evaluate_0.10        tibble_1.2          gtable_0.2.0        htmlTable_1.7
## [13] Matrix_1.2-7.1       DBI_0.6             parallel_3.3.2      gridExtra_2.2.1
## [17] stringr_1.2.0        dplyr_0.5.0         cluster_2.0.5       grid_3.3.2
## [21] nnet_7.3-12          data.table_1.10.0   R6_2.2.0            foreign_0.8-67
## [25] latticeExtra_0.6-28 magrittr_1.5         scales_0.4.1        htmltools_0.3.5
## [29] splines_3.3.2        assertthat_0.1      mnormt_1.5-5        colorspace_1.3-2
## [33] labeling_0.3         stringi_1.1.2       acepack_1.4.1       lazyeval_0.2.0
```

```
## [37] openssl_0.9.5       munsell_0.4.3
```

```
Sys.time()
```

```
## [1] "2017-03-24 12:55:31 GMT"
```