Rescaling Standard Deviation and Covariance

Fan Wang

2020-05-02

Contents

Coefficient of Variation and Correlation

> Go to the **RMD**, **R**, **PDF**, or **HTML** version of this file. Go back to fan's REconTools Package, R Code Examples Repository (bookdown site), or Intro Stats with R Repository (bookdown site).

We have various tools at our disposal to summarize variables and the relationship between variables. Imagine that we have multiple toolboxes. This is the first one. There are two levels to this toolbox. Three Basic Tools:

1. (sample) Mean of X (or Y)
2. (sample) Standard Deviation of X (or Y)
3. (sample) Covariance of X and Y

Additionally, we have two tools that combine the tools from the first level:

1. Coefficient of Variation = (Standard Deviation)/(Mean)
2. Correlation = (Covariance of X and Y)/((Standard Deviation of X)*(Standard Deviation of Y))

The tools on the second level rescale the standard deviation and covariance statistics.

Education and Wage

The dataset, *EPIStateEduWage2017.csv*, can be downloaded here.

Two variables:

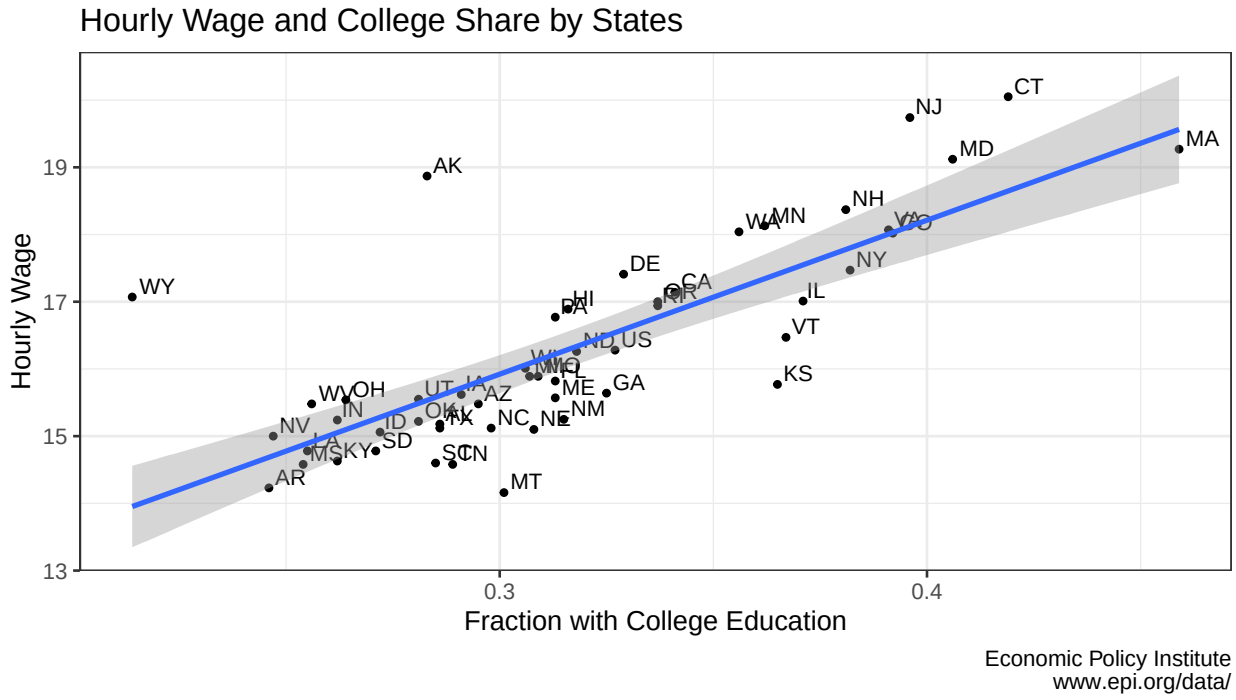
1. Fraction of individual with college degree in a state
   - this is in Fraction units, the minimum is 0.00, the maximum is 100 percent, which is 1.00
2. Average hourly salary in the state
   - this is in Dollar units

```
# Load in Data Tools
# For Reading/Loading Data
library(readr)
library(tibble)
library(dplyr)
library(ggplot2)
# Load in Data
df_wgedu <- read_csv('data/EPIStateEduWage2017.csv')
```

A Scatter Plot We can Visualize the Data with a Scatter Plot. There seems to be a positive relationship between the share of individuals in a state with a college education, and the average hourly salary in that state.

While most states are along the trend line, we have some states, like WY, that are outliers. WY has a high hourly salary but low share with college education.

```
# Control Graph Size
options(repr.plot.width = 5, repr.plot.height = 5)
# Draw Scatter Plot
# 1. specify x and y
# 2. label each state
# 3. add in trend line
scatter <- ggplot(df_wgedu, aes(x=Share.College.Edu, y=Hourly.Salary)) +
      geom_point(size=1) +
      geom_text(aes(label=State), size=3, hjust=-.2, vjust=-.2) +
      geom_smooth(method=lm) +
      labs(title = 'Hourly Wage and College Share by States',
           x = 'Fraction with College Education',
           y = 'Hourly Wage',
           caption = 'Economic Policy Institute\n www.epi.org/data/') +
      theme_bw()
print(scatter)
```



Standard Deviations and Coefficient of Variation The two variables above are in different units. We first calculate the mean, standard deviation, and covariance. With just these, it is hard to compare the standard deviation of the two variables, which are on different scales.

The sample standard deviations for the two variables are: 0.051 and 1.51, in fraction and dollar units. Can we say the hourly salary has a larger standard deviation? But it is just a different scale. 1.51 is a large number, but that does not mean that variable has greater variation than the fraction with college education variable.

Converting the Statistics to Coefficient of Variations, now we have: 0.16 and 0.09. Because of the division, these are both in fraction units–standard deviations as a fraction of the mean. Now these are more comparable.

```r
# We can compute the three basic statistics
stats.msdv <- list(
              # Mean, SD and Var for the College Share variable
              Shr.Coll.Mean = mean(df_wgedu$Share.College.Edu),
              Shr.Coll.Std = sd(df_wgedu$Share.College.Edu),
              Shr.Coll.Var = var(df_wgedu$Share.College.Edu),

              # Mean, SD and Var for the Hourly Wage Variable
              Hr.Wage.Mean = mean(df_wgedu$Hourly.Salary),
              Hr.Wage.Std = sd(df_wgedu$Hourly.Salary),
              Hr.Wage.Var = var(df_wgedu$Hourly.Salary)
              )

# We can compute the three basic statistics
stats.coefvari <- list(
              # Coefficient of Variation
              Shr.Coll.Coef.Variation = (stats.msdv$Shr.Coll.Std)/(stats.msdv$Shr.Coll.Mean),
              Hr.Wage.Coef.Variation = (stats.msdv$Hr.Wage.Std)/(stats.msdv$Hr.Wage.Mean)
              )

# Let's Print the Statistics we Computed
as_tibble(stats.msdv)
as_tibble(stats.coefvari)
```

Covariance and Correlation For covariance, hard to tell whether it is large or small. To make comparisons possible, we calculate the coefficient of variations and correlation statistics.

The covariance we get is positive: 0.06, but is this actually large positive relationship? 0.06 seems like a small number.

Rescaling covariance to correlation, the correlation between the two variables is: 0.78. Since the correlation of two variable is below -1 and $+1$, we can now say actually the two variables are very positively related. A higher share of individuals with a college education is strongly positively correlated with a higher hourly salary.

```r
# We can compute the three basic statistics
states.covcor <- list(
              # Covariance between the two variables
              Shr.Wage.Cov = cov(df_wgedu$Hourly.Salary,
                                 df_wgedu$Share.College.Edu),
              # Correlation
              Shr.Wage.Cor = cor(df_wgedu$Hourly.Salary, df_wgedu$Share.College.Edu),
              Shr.Wage.Cor.Formula = (cov(df_wgedu$Hourly.Salary, df_wgedu$Share.College.Edu)
                                      /(stats.msdv$Shr.Coll.Std*stats.msdv$Hr.Wage.Std))
              )

# Let's Print the Statistics we Computed
as_tibble(states.covcor)
```