Correlations in the Midwest Dataset

Jon Kinsey

Tue Dec 16 18:15:40 2014

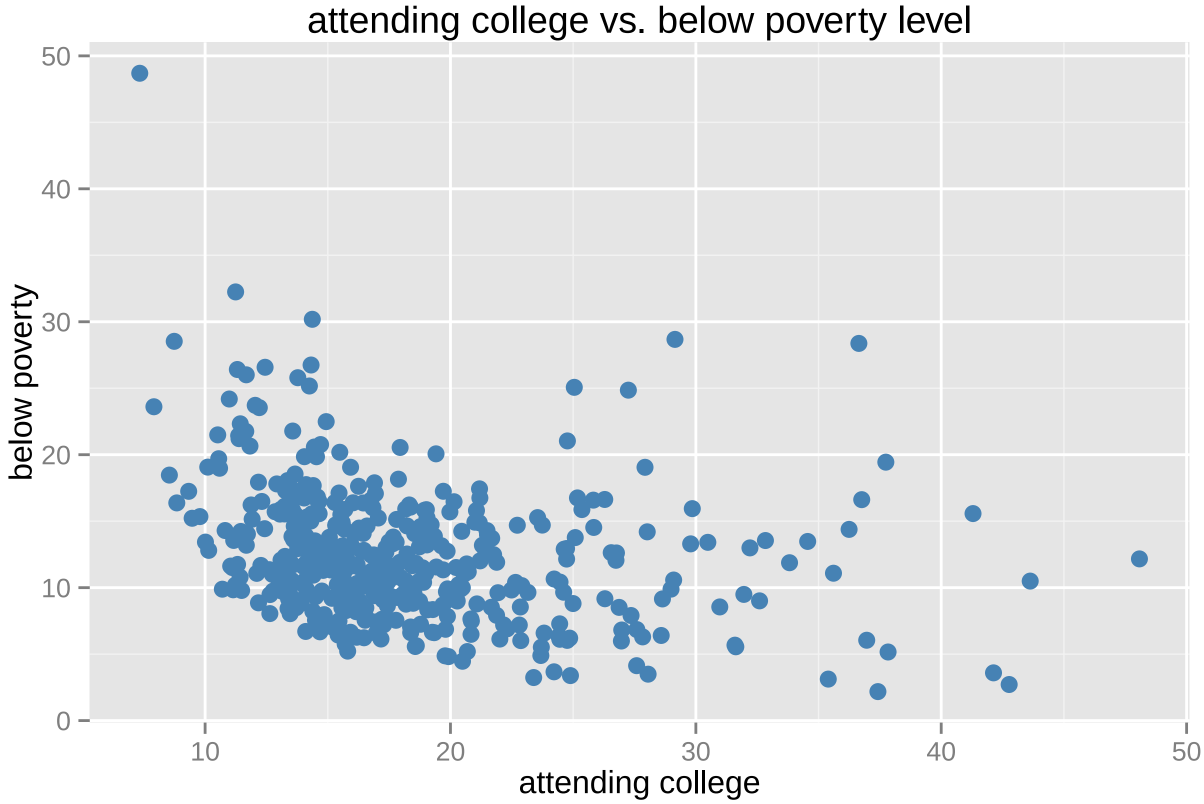
```
# Demographic information of midwest counties.
# A data frame with 437 rows and 28 variables
library(mosaic)
```

```
## Loading required package: car
## Loading required package: 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
##
## Loading required package: lattice
## Loading required package: ggplot2
##
## Attaching package: 'mosaic'
##
## The following objects are masked from 'package:dplyr':
##
##     do, tally
##
## The following object is masked from 'package:car':
##
##     logit
##
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cov, D, fivenum, IQR, median, prop.test, sd,
##     t.test, var
##
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(dplyr)
library(ggplot2)
# library(ggvis)
library(parallel)
#
data(midwest)
names(midwest)
```

```
##  [1] "PID"                  "county"               "state"
##  [4] "area"                 "poptotal"             "popdensity"
##  [7] "popwhite"             "popblack"             "popamerindian"
## [10] "popasian"             "popother"             "percwhite"
## [13] "percblack"            "percamerindan"        "percasian"
## [16] "percother"            "popadults"            "perchsd"
## [19] "percollege"           "percprof"             "poppovertyknown"
## [22] "percpovertyknown"     "percbelowpoverty"     "percchildbelowpovert"
## [25] "percadultpoverty"     "percelderlypoverty"   "inmetro"
## [28] "category"
```

```
ggplot(midwest,aes(x=percollege,y=percbelowpoverty)) +
  geom_point(size=3,color="steelblue") +
  xlab("attending college") + ylab("below poverty") +
  ggtitle('attending college vs. below poverty level')
```



```
# not working with knitr:
#ggvis(midwest,x=~percollege, y=~percbelowpoverty, fill := "steelblue", stroke:="steelblue", fillOpacit
#  layer_points() %>%
#  add_axis("x", title = "% attending college") %>%
#  scale_numeric("x", domain=c(0, 50), nice=FALSE, clamp=TRUE)  %>%
#  add_axis("y", title = "% below poverty") %>%
#  scale_numeric("y", domain=c(0, 50), nice=FALSE, clamp=TRUE)

# Lets look at the correlation between
# percbelowpoverty (% of population below poverty) and
# percollege (% of population who attended college)
```

```
## First let's calculate Pearson's r
# The Pearson product-moment correlation coefficient is a measure of the strength
# of the linear relationship between two variables. Pearson's r can range from -1 to 1.
# An r of -1 indicates a perfect negative linear relationship between variables,
# an r of 0 indicates no linear relationship between variables, and an r of 1
# indicates a perfect positive linear relationship between variables.
cor(percbelowpoverty, percollege, data=midwest)
```

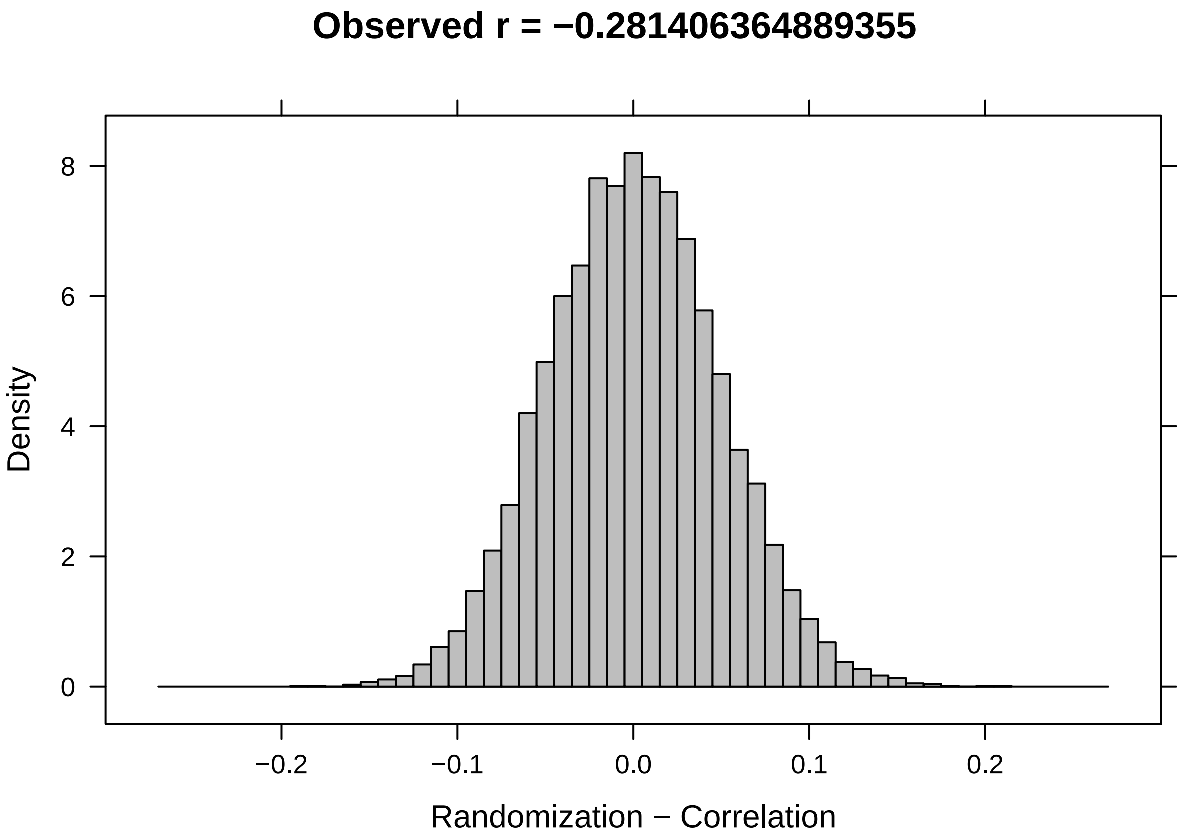
```
## [1] -0.2814064
```

```
## Now, let's run a test for that correlation
cor.test(midwest$percbelowpoverty, midwest$percollege, alternative="less")
```

```
##
##  Pearson's product-moment correlation
##
## data:  midwest$percbelowpoverty and midwest$percollege
## t = -6.1164, df = 435, p-value = 1.067e-09
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.2072089
## sample estimates:
##        cor
## -0.2814064
```

```
## We get a p-value of 1.067e-09

## Let's try a randomization-based test
test.stat <- cor(midwest$percbelowpoverty, midwest$percollege)
randomize <- do(10000) * cor(midwest$percbelowpoverty, shuffle(midwest$percollege))
histogram(~result, data = randomize, col="grey", xlim=c(-.3,.3),
     xlab = "Randomization - Correlation", width=.01,
     main=paste("Observed r =", cor(midwest$percbelowpoverty, midwest$percollege)))
```



```
tally((~result <= test.stat), data=randomize, format="prop")
```

```
##
##   TRUE FALSE
##      0     1
```

```
# Let's try Spearman's rho, used to estimate a rank-based measure of association.
# The sign of the Spearman correlation indicates the direction of association
# between X (the independent variable) and Y (the dependent variable).
cor(midwest$percbelowpoverty, midwest$percollege, method="spearman")
```

```
## [1] -0.3334071
```

```
cor.test(midwest$percbelowpoverty, midwest$percollege, method="spearman", alternative="less")
```

```
##
##  Spearman's rank correlation rho
##
## data:  midwest$percbelowpoverty and midwest$percollege
## S = 18546140, p-value = 5.626e-13
## alternative hypothesis: true rho is less than 0
## sample estimates:
##        rho
## -0.3334071
```

```r
# Kendall's tau rank correlation. Like Spearman's rho a negative correlation
# indicates that when X is increasing, Y is decreasing. Values are in the range -1 <= tau <= 1
# Kendall’s Tau: usually smaller values than Spearman’s rho correlation.
#   Calculations based on concordant and discordant pairs. Insensitive to error.
#   P values are more accurate with smaller sample sizes.
# Spearman’s rho: usually have larger values than Kendall’s Tau.  Calculations based
#   on deviations.  Much more sensitive to error and discrepancies in data.
cor(midwest$percbelowpoverty, midwest$percollege, method="kendall")
```

```
## [1] -0.2334516
```

```r
cor.test(midwest$percbelowpoverty, midwest$percollege, method="kendall", alternative="less")
```

```
##
##  Kendall's rank correlation tau
##
## data:  midwest$percbelowpoverty and midwest$percollege
## z = -7.2911, p-value = 1.537e-13
## alternative hypothesis: true tau is less than 0
## sample estimates:
##        tau
## -0.2334516
```