Literacy Rates from GapMinder

December 2017

Introduction

The literacy rate dynamic dataset comes from GapMinder http://www.gapminder.org/. GapMinder also has a plethora of variables from which students can choose (according to their own interests), but here we work with literacy rates measured at the country level. The analysis (and more importantly, the data scraping from GapMinder) could easily be extended for students interested in all sorts of political, social, environmental, or demographic data available. Understanding political and demographic trends across both time and location can provide very interesting insight into economic or political science questions. Alternatively the GapMinder data can be perused in a descriptive or graphical manner.

Below is code for pulling in a Google spreadsheet from GapMinder http://www.gapminder.org/. You can follow the instructions given in the following R Markdown file for downloading any Google spreadsheet (not just from Gap Minder), though adding authentication (for non public sheets) requires another step. See googlesheets for more information.

Data information & loading data

Three datasets are loaded. The datasets are the female literacy rate over time, the male literacy rate over time, and the overall literacy rate over time for dozens of countries going back to the mid-1970s. The R code is given in the Markdown file.

Looking at the data at this point is a good idea. One thing to notice is that there is a ton of missing data. That's expected (especially if you are used to looking at GapMinder data), because we wouldn't expect that every country has literacy data for each gender going back 40 years for every single year. Also, notice that the data aren't in tidy format (rows as observational units and columns of variables). After we wrangle the data again it will (a) look different and (b) be in tidy form.

```
glimpse(litALL)
```

```
## Observations: 260
## Variables: 38
## $ `Adult (15+) literacy rate (%). Total` <chr> "Afghanistan", "Albania...
## $ `1975`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1976`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1977`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1978`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1979`                                 <dbl> 18.15768, NA, NA, NA, N...
## $ `1980`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1981`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1982`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1983`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1984`                                 <dbl> NA, NA, NA, NA, NA, 95....
## $ `1985`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1986`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1987`                                 <dbl> NA, NA, 49.63088, NA, N...
## $ `1988`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1989`                                 <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1990`                                 <dbl> NA, NA, NA, NA, NA, NA,...
```

```
## $ `1991`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1992`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1993`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1994`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1995`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1996`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1997`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1998`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `1999`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2000`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2001`                          <dbl> NA, 98.71298, NA, NA, 6...
## $ `2002`                          <dbl> NA, NA, 69.87350, NA, N...
## $ `2003`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2004`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2005`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2006`                          <dbl> NA, NA, 72.64868, NA, N...
## $ `2007`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2008`                          <dbl> NA, 95.93864, NA, NA, N...
## $ `2009`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2010`                          <dbl> NA, NA, NA, NA, NA, NA,...
## $ `2011`                          <dbl> 39.00000, 96.84530, NA,...
```

Each of the original googlesheets comes as a spreadsheet with country as the row and year as the column. R imports the years as column names (which are difficult to deal with as numeric column headers), and we need to gather the data into a format such that "Year" is one of the variable names. At the end of the wrangling process, the variables will be: country, year, litRateF, litRateM, litRateALL, and continent.

```
litF = litF %>% select(country=starts_with("Adult"), everything()) %>%
        gather(year, litRateF, -country) %>%
        mutate( year = readr::parse_number(year)) %>%
        filter(!is.na(litRateF)) %>%
        mutate(litRateF = readr::parse_number(litRateF))

litM = litM %>% select(country=starts_with("Adult"), everything()) %>%
        gather(year, litRateM, -country) %>%
        mutate( year = readr::parse_number(year)) %>%
        filter(!is.na(litRateM)) %>%
        mutate(litRateM=readr::parse_number(litRateM))

litALL = litALL %>% select(country=starts_with("Adult"), everything()) %>%
        gather(year, litRateALL, -country) %>%
        mutate( year = readr::parse_number(year)) %>%
        filter(!is.na(litRateALL))

literacy = full_join(full_join(litF, litM, by=c("country", "year")),
                     litALL, by=c("country", "year"))

continentGM = gapminder %>% select(country, continent) %>% group_by(country) %>%
      summarize(continent = first(continent))

literacy = left_join(literacy, continentGM, by="country")
```

Now the data frame is in tidy format (rows are observational units, columns are variables), and the dataframe literacy has all of the information needed.

```
glimpse(literacy)
```

```
## Observations: 575
## Variables: 6
## $ country    <chr> "Burkina Faso", "Central African Rep.", "Kuwait", "...
## $ year       <dbl> 1975, 1975, 1975, 1975, 1975, 1975, 1976, 1976, 197...
## $ litRateF   <dbl> 3.182766, 8.399576, 48.015214, 45.098921, 38.124870...
## $ litRateM   <dbl> 14.52849, 29.59292, 68.02863, 77.50394, 58.38611, 9...
## $ litRateALL <dbl> 8.685145, 18.236172, 59.564392, 61.627683, 53.51487...
## $ continent  <fctr> Africa, NA, Asia, Europe, NA, Americas, Americas, ...
```

As mentioned before, we wouldn't really expect every country to have literacy data for every year. However, it is straightforward to tally how many observations per country and per year exist in the dataset, if desired.

Using dynamic data within a typical classroom

The introductory analysis considers gender differences in literacy rates and uses a linear model on the difference between female and male literacy rates across time. We show a graphical representation and discuss model assumptions including sampling and independence of residuals. The model indicates that the difference between male and female literacy rates is shrinking over time. However, we worry about the effects of other variables and encourage a more complete analysis. Indeed, there may be large biases in our model if important explanatory variables have been left out.

It is important to note that the data collected here (and on all of GapMinder) is observational. Causal mechanisms cannot be implied regardless of strength of correlation, and we recommend a converstation with students about the dangers of possible confounding variables that might explain any suggested causal realtionships.

```
ggplot(literacy, aes(x=litRateF, y=litRateM)) +
    geom_point(alpha=.75, aes(color=year)) + geom_abline(slope=1, intercept=0)
```

It might be interesting, however, to look at the relationship between literacy rates over time. To do that, we create a new variable which is the difference between male and female literacy rates.

```
literacy = literacy %>% mutate(diffLit = litRateM - litRateF)
summary(lm(diffLit~year, data=literacy))
```

```
##
## Call:
## lm(formula = diffLit ~ year, data = literacy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.125  -7.715  -2.804   7.229  30.072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 417.90345   80.11740   5.216 2.56e-07 ***
## year         -0.20419    0.04006  -5.097 4.70e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.826 on 568 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.04374,    Adjusted R-squared:  0.04206
```

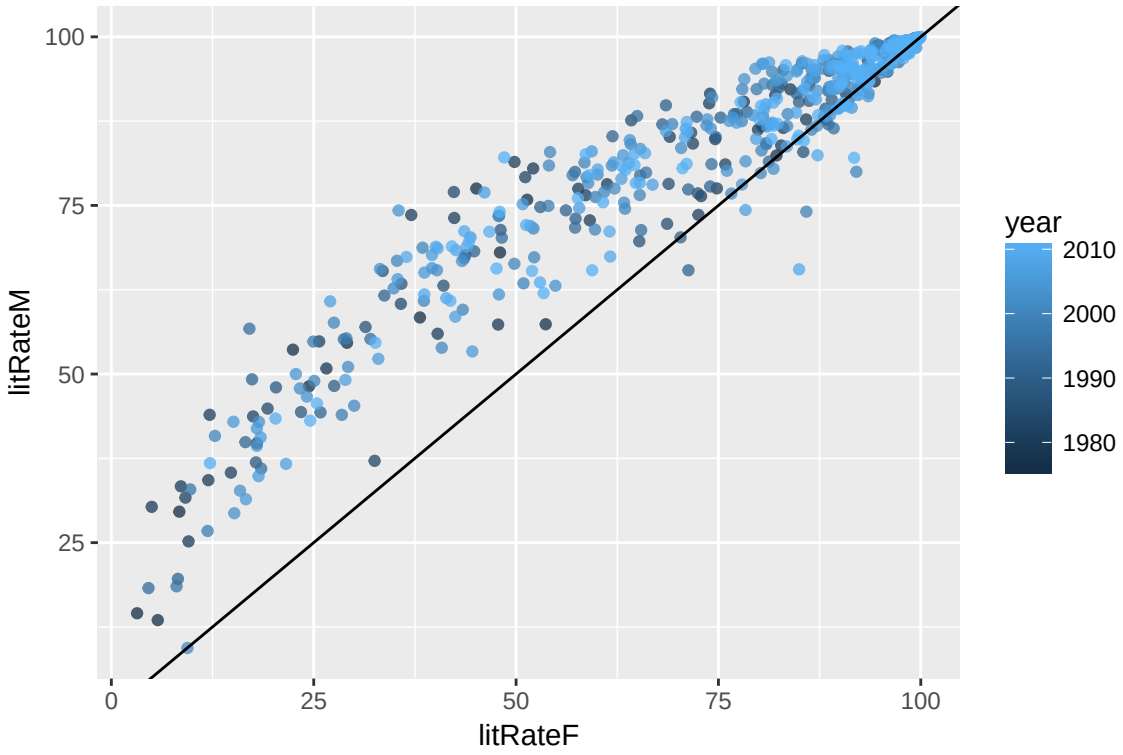

Figure 1: The plot shows that the higher the female literacy rate, the higher the male literacy rate. Additionally, across the board, the male literacy rate is higher than the female literacy rate (as referenced by the y=x line).

```
## F-statistic: 25.98 on 1 and 568 DF,  p-value: 4.698e-07
```

```
ggplot(literacy, aes(x=year, y=diffLit)) + geom_point(alpha=0.75, aes(color=continent)) +
    stat_smooth(method="lm", se=FALSE)
```

The blue line represents the linear relationship between year and difference in literacy rate. As we know, least squares is an optimization technique that does not require any assumptions about sampling or distribution of the data. However, the *inference* done on the slope statistics does require that we think about a null hypothesis and its relationship to the data. Certainly the data are not a representative sample of the entire population of countries over the last 40 years. Additionally, one might expect the residuals from one country to be correlated from year to year (certainly if the literacy rate is higher for women in one year, it will likely be higher for women in the following year).

Random sampling is not something we can change about the data. We can however, look carefully into the original spreadsheet and do our best to guage which countries have missing information over which time period. There may be additional information that will help the scope of conclusion for the inference.

Correlated errors is a problem that can be addressed. Indeed, adding in additional variables (country, continent) might remove the correlation structure all together. Teaching students to consider the multivariate structure of the data - even in an introductory class! - will go a long way toward them being able to make accurate assessments about future datasets.

We feel comfortable concluding that overall, the difference between male and female literacy rates is shrinking over time. However, we worry about the effects of other variables (e.g., the relationship might be different from continent to continent!) and encourage a more complete analysis.

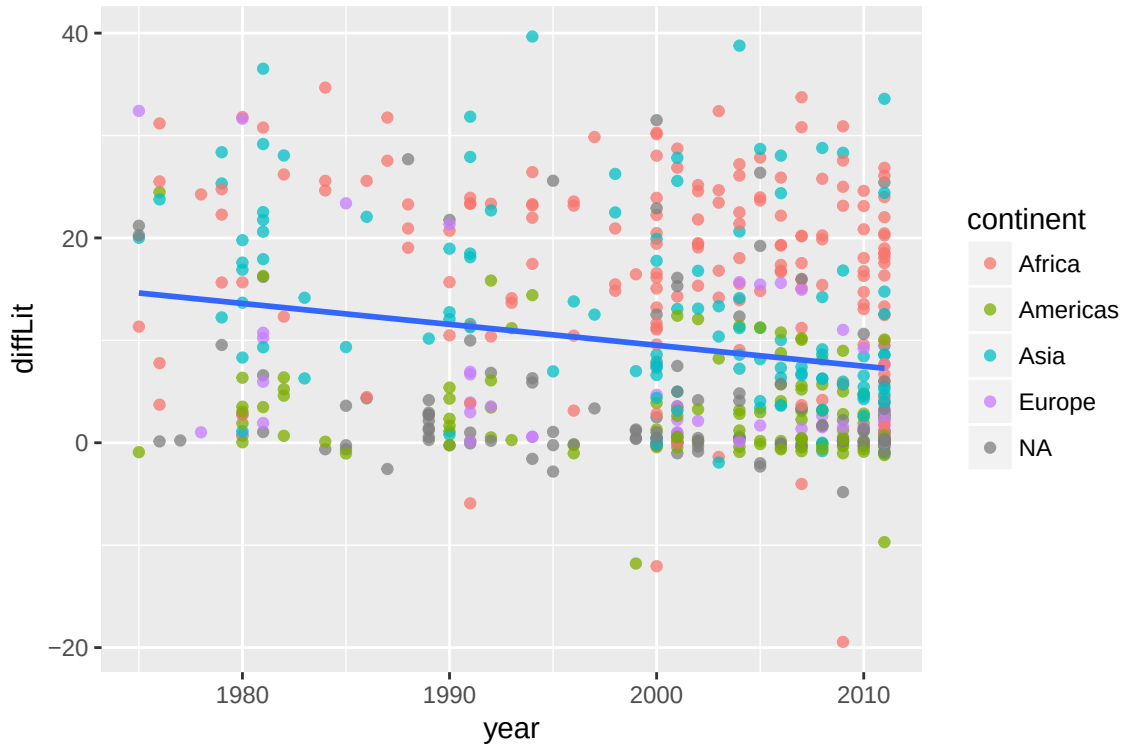

Figure 2: Difference in literacy rates vs year; blue line is the least squares regression line.

Thinking outside the box

In the second analysis, we work with the additional variable `continent`. The trends observed in the first analysis hold up in the second analysis (i.e., the difference declines over time in each of the continents). However, there are additional considerations to be made, for example, the differences between the slopes across the continents (the analysis is available in the supplementary materials and not shown here). We suggest additional explorations into the independence of the residuals and more advanced spatio-temporal patterns of literacy.

```
summary(lm(diffLit~year*continent, data=literacy))
```

```
##
## Call:
## lm(formula = diffLit ~ year * continent, data = literacy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.806  -4.004  -1.074   4.596  27.544
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            260.63054  125.96058   2.069   0.0391 *
## year                    -0.12159    0.06298  -1.931   0.0542 .
## continentAmericas      -50.50739  192.78896  -0.262   0.7935
## continentAsia          421.00881  184.22529   2.285   0.0228 *
## continentEurope        329.59400  230.70597   1.429   0.1538
## year:continentAmericas   0.01798    0.09637   0.187   0.8521
```

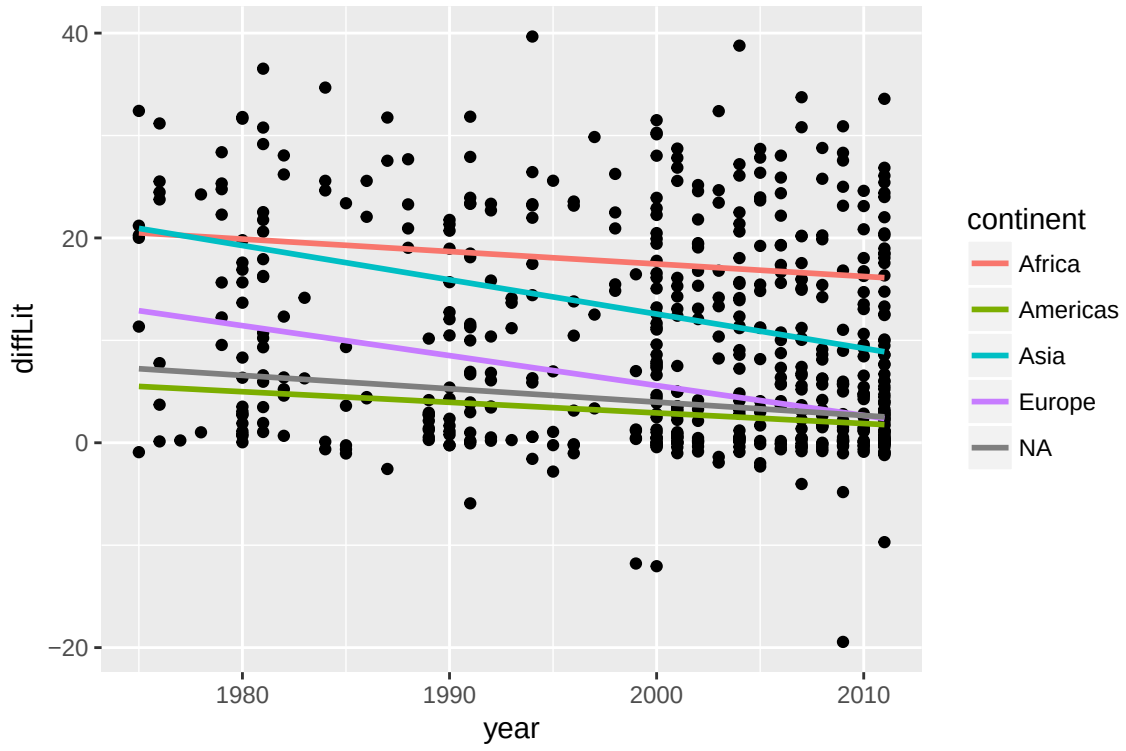

Figure 3: Least squares regression line between difference in literacy rate and year broken down by continent.

```
## year:continentAsia      -0.21295    0.09213  -2.311   0.0213 *
## year:continentEurope    -0.17073    0.11534  -1.480   0.1395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.885 on 439 degrees of freedom
##   (128 observations deleted due to missingness)
## Multiple R-squared:  0.4086, Adjusted R-squared:  0.3992
## F-statistic: 43.33 on 7 and 439 DF,  p-value: < 2.2e-16
```

```
ggplot(literacy, aes(x=year, y=diffLit)) + geom_point() +
  stat_smooth(method="lm", se=FALSE, aes(color=continent))
```

By looking at the data in a multivariate framework (with interaction), we see that the slope of the relationship over time is negative for all continents, however, the only continent which is significantly different from Africa (the baseline) is Asia.

Additional ideas for analysis:

- Further inspection of the residulas (e.g., looking for trends with respect to country)
- Adding a spatio-temporal analysis (e.g., a graphical representation of literacy rate on a world map)
- Considering other variables (e.g., GDP, infant mortality, etc.) as part of the analysis
- Confidence intervals for the difference in coefficients on slope or intercept to compare continents more specifically (e.g., Europe vs. Americas)