Applied Introduction to T-Tests, Correlation, and OLS regression

Alix Ziff, Gaea Morales, and Zachary Johnson

Summer 2020, Version 5: June 22

Introduction to Regression Analysis in R

In this module, we will be introducing basic commands and operations to conduct statistical T-Tests, calculating and analyzing correlations, and preliminary ordinary least squares (OLS) regressions.

We will be using the World Development Indicators dataset to analyze the relationship between HDI rank and Fertility Rates. Basically, we want to know if there is a correlation between a country's wealth and how many children citizens have.

We will begin by loading the necessary packages. Note that if you have not previously used the following, you will have to install them (using install.packages()). We must also set our working directory.

```
library(tidyverse)

setwd("/Volumes/GoogleDrive/My Drive/SPEC Summer 2020/Training Data Science/0 Training Data")
```

Data Preparation

We must first load and clean up the data to be analyzed. Because these variables are pulled from the World Development Indicators data, the years are currently variables (columns) with HDI rank and fertility rates as observations arranged per country in their respective dataframes.

Let's load our datasets.

```
hdi_index <- read.csv("hdi_index.csv")
glimpse(hdi_index)

wdi_fertility <- read.csv("wdi_fertility.csv")
glimpse(wdi_fertility)
```

Exercise 1:

Clean up our HDI data by renaming variables (separating and then removing the X in year variables, and only selecting our variables of interest), and pivoting (turning year variables into year observations).

```
hdi_index <- pivot_longer(hdi_index, X1990:X2018, names_to = "Year", values_to = "hdi_rank")

hdi_index <- separate(hdi_index, col = Year, into = c(NA, "Year"), sep = "X")
# splits off the X from the year observation name
```

```
hdi_index <- hdi_index %>%
  rename(Rank = HDI.Rank..2018.) %>%
  select(Country, Year, Rank) %>%
  mutate(Year = as.numeric(Year))
```

Do the same for the fertility data.

```
wdi_fertility <- wdi_fertility %>%
  select(Country = Country.Name,
         X1960:X2018)

wdi_fertility <- pivot_longer(wdi_fertility,
                              X1960:X2018, names_to = "Year",
                              values_to = "FertilityRate")

wdi_fertility <- separate(wdi_fertility,
                          col = Year, into = c(NA, "Year"),
                          sep = "X") # splits off the X from the year observation

wdi_fertility <- wdi_fertility %>%
  mutate(Year = as.numeric(Year)) %>%
  filter(Year >= 1990)
```

Now that we have two cleaned up datasets with our two variables of interest, we want to merge them into one dataframe.

```
frt_rank <- full_join(hdi_index, wdi_fertility, by = c("Country", "Year")) %>%
  arrange(Country, Year)
```

Helpful Hint: Notice that we lose observations when using inner_join because we are only including data with *both* HDI rank and fertility data for each country-year observation.

Regression Analysis

Our y (dependent or outcome variable) of interest is the Fertility Rate of our country sample. We want to know if HDI rank (our x, independent variable, explanatory variable) has a correlation with Fertility Rates.

If there is a relationship, we want to know how strong it is and whether it is positive or negative.The correlation coefficient tells us just that.

Calculate a correlation matrix for your variables. We will subset just our numeric variables: HDI rank and fertility rates.

```
smalldata <- frt_rank %>%
  select(Rank, FertilityRate)


cor(smalldata, use="complete.obs")
#cor() is the simplest way to calculate for the correlation coefficient
#matrix across all variables in the dataset.
#in this case, we are only working with two numeric variables.
```

We can interpret the correlation coefficient as having a weak and negative correlation.

We can also visualize the strength and direction of our correlation by creating a scatterplot.

```
ggplot(frt_rank, aes(x=Rank, y=FertilityRate)) + geom_jitter()
```

Notice that like our correlation coefficient, the scatterplot shows a loose (weak) grouping of observations with a slight downward slope.

##Ordinary Least Squares (OLS Regression)

While keeping our scatterplot that visualizes the relationship between fertility rate and HDI rank, we'll add our line of best fit. R will calculate this for us through the geom_smooth function.

```
ggplot(frt_rank, aes(y=FertilityRate, x=Rank))+
        geom_point(position = position_jitter(height = 1), alpha = 0.8 )+
        geom_smooth(method = "lm", formula = y ~ x, colour = "red", se = TRUE)
        #drawing the line "lm" through the data points
```

```
fit<-lm(FertilityRate~Rank, data=frt_rank,na.action = na.exclude)
summary(fit)
```

##Interpreting Regression Results

With this regression, we want to understand if there is a relationship between HDI rank and Fertility Rates. *Note: Here we are only highlighting a couple elements of our regression results to determine if there is relationship between our variables.*

residuals: the distance of each individual observation from the line of best fit *the median observation value–let's call it country A–is 0.056 units below the line of best fit*

estimate = regression coefficient: the expected change in the outcome variable for a one-unit increase in the explanatory variable *for simplified analysis we'll say that our estimate is -0.024 suggesting that for every one unit change in HDI rank we'll see a 0.024 unit increase in a country's fertility rate*

standard error: how accurate the mean of any sample from the population is compared to the true population mean *Our standard error is 0.00024–very low–suggesting that if we ran our model again with a different sample we'd get similar results*

t value: how similar the distribution of observations are between two variables *our t-value of -63.74 is statistically significant, suggesting that HDI rank and Fertility Rates are related to one another*

R^2: range between 0 and 1 suggesting how much variance of the DV can be explained by the IV *30% of the variance found in the response variable (DV) can be explained by the predictor variable (IV)*

p-value: indicates the strength of the evidence we have to determine if there is a correlatory relationship between our explanatory variable and our outcome variable. It is essentially the likelihood that our results occurred by random chance. *this very low p-value suggests that it is unlikely our results are due to random chance and it is more likely that our results are driven by the explanatory variable*

##Regression Analysis

Why do we expect HDI rank have a relationship with fertility rates?

Our intuition is that as populations experience economic development (accrue wealth), women may have fewer children (the fertility rate decreases) as they have more access to education and employment opportunity, as well as an adjustment of cultural expectations.

null hypothesis: *that fertility rates are not effected by an increase in HDI rank.* The relationship between our X and Y is no different than if we left it up to chance.

our hypothesis: As our X: HDI rank increases (populations experience economic development), our Y: fertility rates decrease.