Introduction to Handling Data

ECON20222 - Lecture 2

Ralf Becker and Martyn Andrews

January 2019

Aim for today

- Review hypothesis testing
- Review simple regresison anlysis
- Become more familiar with R

Preparing your workfile

We add the basic libraries needed for this week's work:

```
library(tidyverse)     # for almost all data handling tasks
library(readxl)        # to import Excel data
library(ggplot2)       # to produce nice graphiscs
library(stargazer)     # to produce nice results tables
```

New Dataset - Wellbeing

Doing Economics Project 8 deals with international wellbeing data.

Data are from the European Value Survey.

- A large catalogue of questions on Perceptions of Life, Politics and Society, Work, Religion etc
- 48 mainly European countries
- Four waves/years of data (1981, 1990, 1999 and 2008)
- 129,515 observations (people/respondents)

```
load("WBdata.Rdata")
```

This will load two objects into your environment

- `wb_data` - the actual data file
- `wb_data_Des` - a table which contains some description to each variable

To get to this dataset a significant amount of data handling and cleaning had to happen (see Project 8 in Doing Economics.)

Wellbeing Data

```
str(wb_data)  # prints some basic info on variables
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    129515 obs. of  19 variables:
##  $ S002EVS    : chr  "1981-1984" "1981-1984" "1981-1984" "1981-1984" ...
##  $ S003       : chr  "Belgium" "Belgium" "Belgium" "Belgium" ...
##  $ S006       : num  1001 1002 1003 1004 1005 ...
##  $ A009       : num  3 5 2 5 5 5 5 5 4 4 ...
##  $ A170       : num  9 9 3 9 9 9 9 10 8 10 ...
##  $ C036       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ C037       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ C038       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ C039       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ C041       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X001       : chr  "Male" "Male" "Male" "Female" ...
##  $ X003       : num  53 30 61 60 60 19 38 39 44 76 ...
##  $ X007       : chr  "Single/Never married" "Married" "Separated" "Married" ...
##  $ X011_01    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ X025A      : chr  NA NA NA NA ...
##  $ Education_1: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Education_2: chr  NA NA NA NA ...
##  $ X028       : chr  "Full time" "Full time" "Unemployed" "Housewife" ...
##  $ X047D      : num  NA NA NA NA NA NA NA NA NA NA ...
```

Data Description

```
wb_data_Des  # prints some basic info on variables
```

```
##          Names                  Labels
## 1      S002EVS               EVS-wave
## 2         S003          Country/region
## 3         S006       Respondent number
## 4         A009                  Health
## 5         A170       Life satisfaction
## 6         C036                 Work Q1
## 7         C037                 Work Q2
## 8         C038                 Work Q3
## 9         C039                 Work Q4
## 10        C041                 Work Q5
## 11        X001                     Sex
## 12        X003                     Age
## 13        X007          Marital status
## 14     X011_01      Number of children
## 15       X025A               Education
## 16 Education_1     Education category
## 17 Education_2  Education Description
## 18        X028              Employment
## 19       X047D Monthly household income
##                                                                                                  Description
## 1                                                                                                  EVS-wave
## 2                                                                                                Country/region
## 3                                                                                   Original respondent number
## 4                                                                  State of health (subjective), 1 = Very Poor, 5 = Very good
## 5                                                                                  Satisfaction with your life
## 6                                                  To develop talents you need to have a job, 1 = Strongly Agree, 5 = Strongly Disagree
## 7                                Humiliating to receive money without having to work for it, 1 = Strongly Agree, 5 = Strongly Disagree
## 8                                                People who don't work become lazy, 1 = Strongly Agree, 5 = Strongly Disagree
## 9                                                   Work is a duty towards society, 1 = Strongly Agree, 5 = Strongly Disagree
## 10                   Work should come first even if it means less spare time, 1 = Strongly Agree, 5 = Strongly Disagree
```

Data - Countries

Let's find out which countries are in the sample:

```
unique(wb_data$S003)    # unque finds all the different values in a variable
```

```
##  [1] "Belgium"            "Canada"             "Denmark"
##  [4] "France"             "Germany"            "Iceland"
##  [7] "Ireland"            "Italy"              "Malta"
## [10] "Netherlands"        "Norway"             "Spain"
## [13] "Sweden"             "Great Britain"      "United States"
## [16] "Northern Ireland"   "Austria"            "Bulgaria"
## [19] "Czech Republic"     "Estonia"            "Finland"
## [22] "Hungary"            "Latvia"             "Lithuania"
## [25] "Poland"             "Portugal"           "Romania"
## [28] "Slovakia"           "Slovenia"           "Croatia"
## [31] "Greece"             "Russian Federation" "Turkey"
## [34] "Albania"            "Armenia"            "Bosnia Herzegovina"
## [37] "Belarus"            "Cyprus"             "Northern Cyprus"
## [40] "Georgia"            "Luxembourg"         "Moldova"
## [43] "Montenegro"         "Serbia"             "Switzerland"
## [46] "Ukraine"            "Macedonia"          "Kosovo"
```

Data - Waves

Let's find out how many observations/respondents we have for each country (`S003`) in each wave (`S002EVS`).

Use piping technique of the `tidyverse`

```
table1 <- wb_data %>% group_by(S002EVS,S003) %>% # groups by Wave and Country
      summarise(n = n()) %>%                     # calculating no of obs
      spread(S002EVS,n) %>%                      # put Waves across columns
      print(n=4)
```

```
## # A tibble: 48 x 5
##   S003    `1981-1984` `1990-1993` `1999-2001` `2008-2010`
##   <chr>         <int>       <int>       <int>       <int>
## 1 Albania          NA          NA          NA        1200
## 2 Armenia          NA          NA          NA        1224
## 3 Austria          NA        1432          NA        1216
## 4 Belarus          NA          NA          NA        1237
## # ... with 44 more rows
```

For each country ($j = 1, ..., 48$) we have observations from potentially four years ($t = 1, ..., 4$). For each country-year (jt) we have ($i = 1, ..., n_{jt}$) observations, e.g. $n_{Austria,1990} = 1432$. Each observation can be identified/indexed by ijt.

Data - Some graphical representation

Summarise data by country and wave.

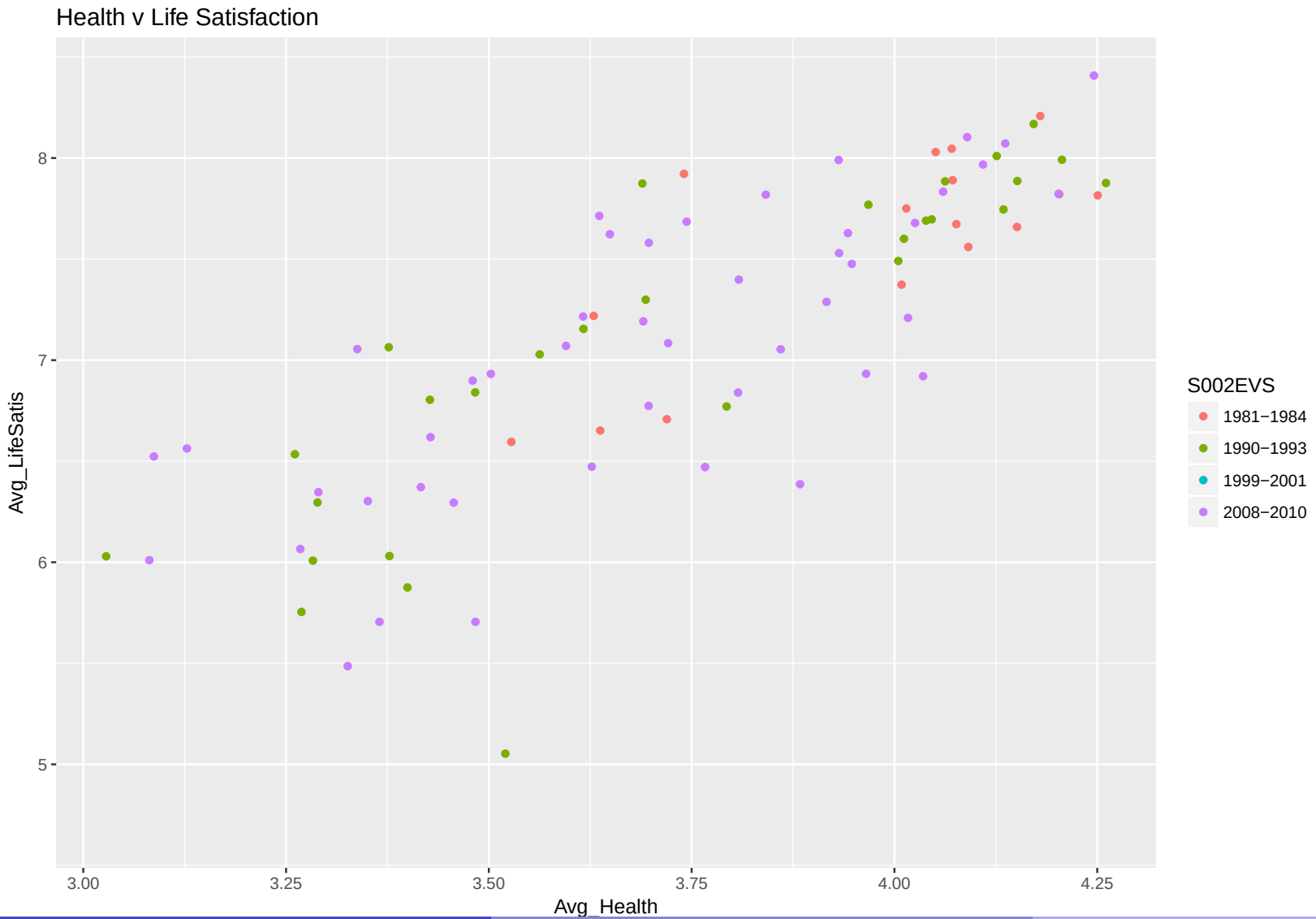
- A170: All things considered, how satisfied are you with your life as a whole these days? (1 Dissatisfied to 10 Satisfied)
- A009: All in all, how would you describe your state of health these days? Would you say it is ... 1 Very good to 5 Very poor

```
table2 <- wb_data %>% group_by(S002EVS,S003) %>% # groups by Wave and Country
            summarise(Avg_LifeSatis = mean(A170),Avg_Health = mean(A009))
head(table2,4)
```

```
## # A tibble: 4 x 4
## # Groups:   S002EVS [1]
##   S002EVS   S003    Avg_LifeSatis Avg_Health
##   <chr>     <chr>           <dbl>      <dbl>
## 1 1981-1984 Belgium          7.37       4.01
## 2 1981-1984 Canada           7.82       4.20
## 3 1981-1984 Denmark          8.21       4.18
## 4 1981-1984 France           6.71       3.72
```

Data - Some graphical representation

```
ggplot(table2,aes(Avg_Health,Avg_LifeSatis, colour=S002EVS)) +
  geom_point() +
  ggtitle("Health v Life Satisfaction")
```



Data - Some graphical representation

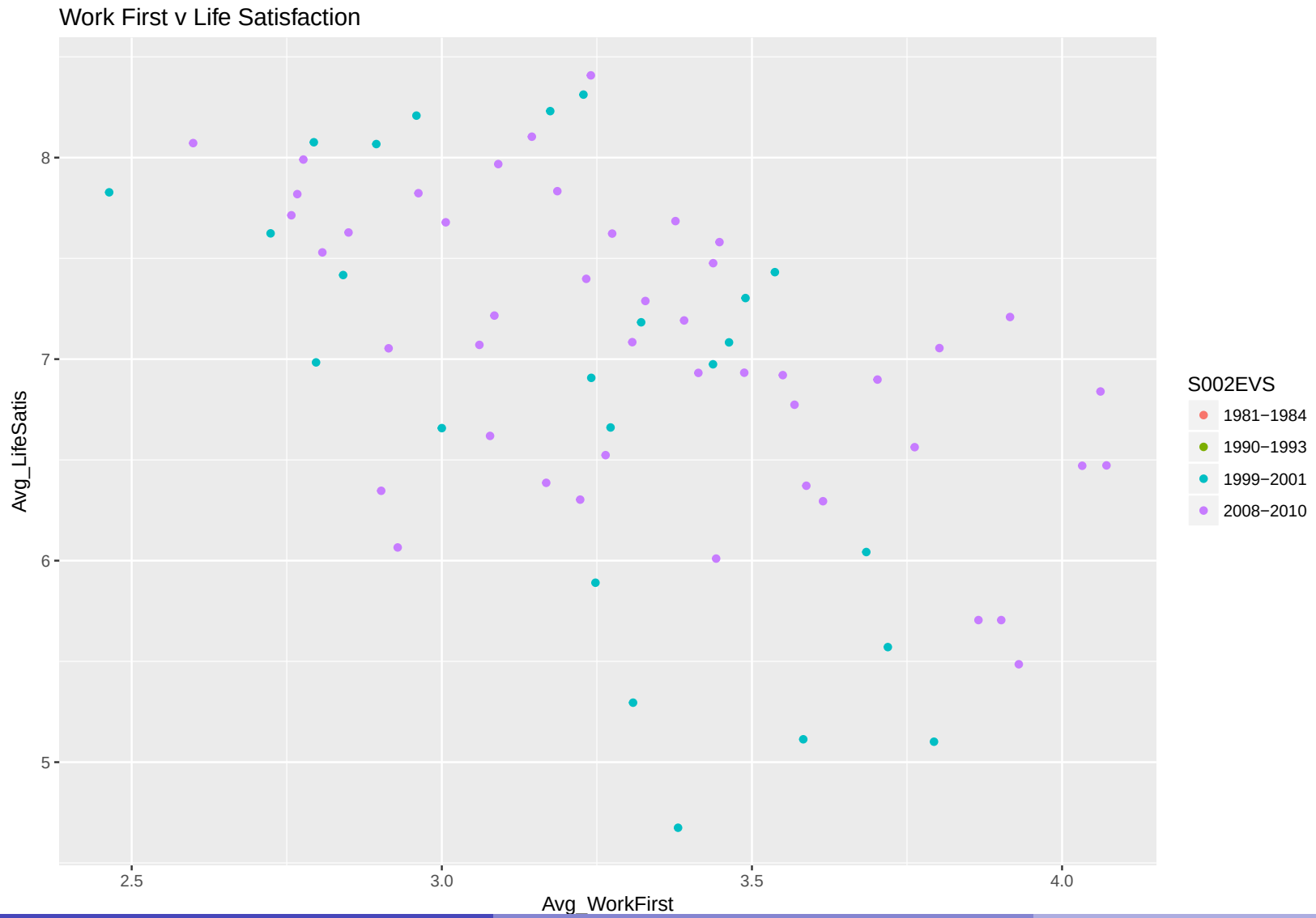
Summarise data by country and wave.

- A170: All things considered, how satisfied are you with your life as a whole these days? (1 Dissatisfied to 10 Satisfied)
- C041: Work should come first even if it means less spare time, 1 = Strongly Agree, 5 = Strongly Disagree

```
table2 <- wb_data %>% group_by(S002EVS,S003) %>%
            summarise(Avg_LifeSatis = mean(A170),Avg_WorkFirst = mean(C041))
```

Data - Some graphical representation

```
ggplot(table2,aes( x=Avg_WorkFirst, y=Avg_LifeSatis,colour=S002EVS)) +
  geom_point() +
  ggtitle("Work First v Life Satisfaction")
```



Data - Some graphical representation

There seems to be a negative relationship between Attitude to Work (WC) and Life Satisfaction (LS) (in countries with a "work-centric" ethic people were on average happier.)

Is there such a relationship inside countries as well?

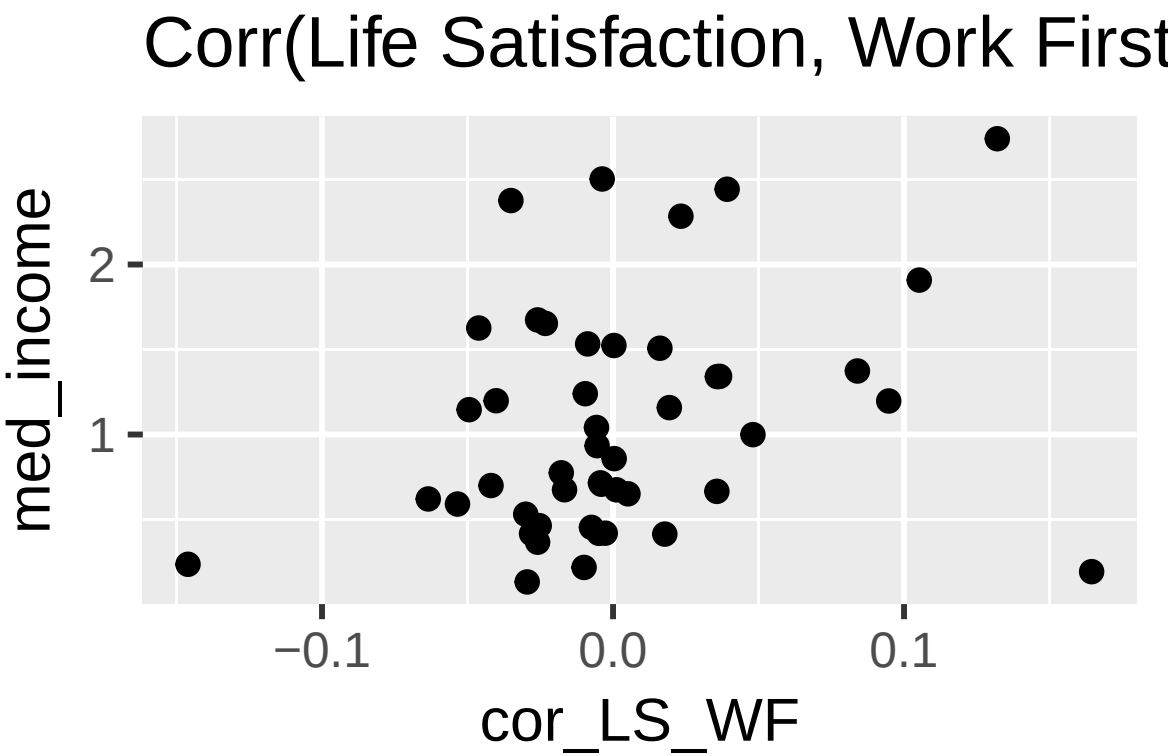
We will calculate correlations for each country-wave, e.g. Austria in 2008:

$$Corr_{Aut,2008}(LS_{i,Aut,2008}, WC_{i,Aut,2008})$$

```
table3 <- wb_data %>% filter(S002EVS == "2008-2010") %>%
            group_by(S003) %>% # groups by Country
            summarise(cor_LS_WF = cor(A170,C041,use = "pairwise.complete.obs"),
                      med_income = median(X047D)) %>%
            arrange(cor_LS_WF)
```

Data - Some graphical representation

```
ggplot(table3,aes( cor_LS_WF, med_income)) +
  geom_point() +
  ggtitle("Corr(Life Satisfaction, Work First) v Median Income")
```



Data on Maps

Geographical relationships are sometimes best illustrtaed with maps.

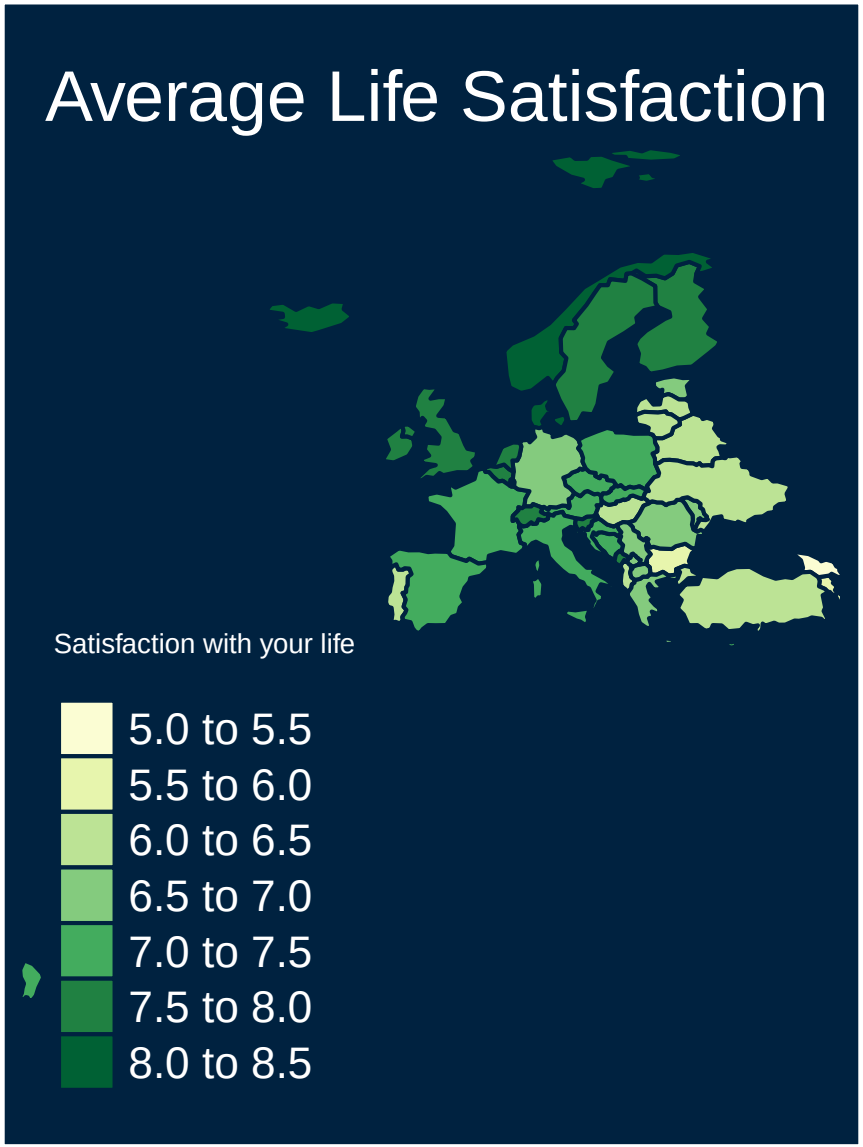
Sometimes these will reveal a pattern.

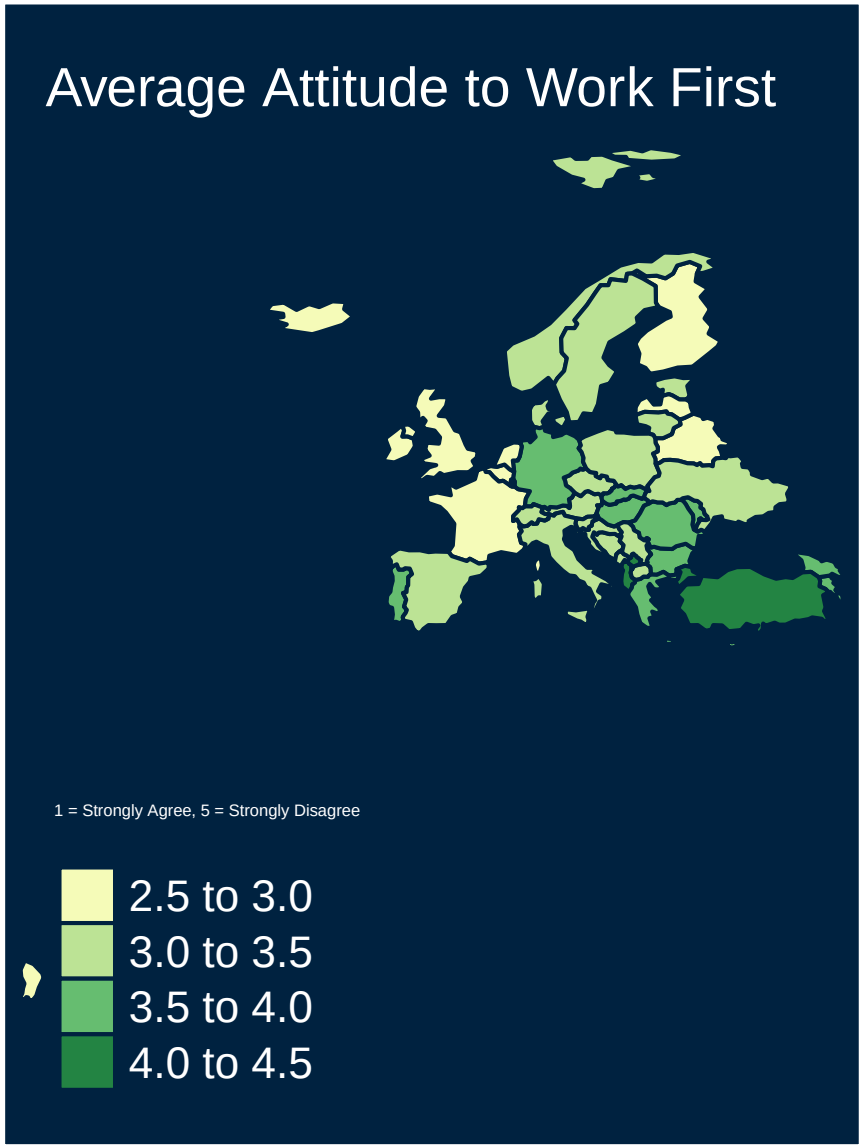
R can create great maps (but it requires a bit of setup - see the additional file on BB). You need the following

- A shape file for each country
- The statistics for each country
- a procedure to merge these bits of information in one data-frame (`merge`)

Let's look at average life satisfaction average attitude to the "work first" statement as these statistics vary by country.

Data on Maps


Average Life Satisfaction
Satisfaction with your life
5.0 to 5.5
5.5 to 6.0
6.0 to 6.5
6.5 to 7.0
7.0 to 7.5
7.5 to 8.0
8.0 to 8.5


Average Attitude to Work First
1 = Strongly Agree, 5 = Strongly Disagree
2.5 to 3.0
3.0 to 3.5
3.5 to 4.0
4.0 to 4.5

Hypothesis Testing - Introduction

Hypothesis testing is a core technique used in empirical analysis. Use sample data to infer something about the population mean (or correlation, or variance, etc). Hence *inference*.

It is crucial to understand that the particular sample we have is one of many different possible samples. Whatever conclusion we arrive at is not characterised by certainty.

Example

Are average life satisfaction in Germany and Britain the same? (for 2008-2010 wave)

$$\begin{aligned} H_0 &: \mu_{LS,Ger,2008} = \mu_{LS,GB,2008} \\ H_A &: \mu_{LS,Ger,2008} \neq \mu_{LS,GB,2008} \end{aligned}$$

When performing a test we need to calibrate some level of uncertainty. We typically fix the Probability with which we reject a correct null hypothesis (Type I error). This is also called the significance level.

Hypothesis Testing - Introduction

Depending on the type of hypothesis there will be a test statistic which will be used to come to a decision.

Assuming that H_0 is true this test statistic has a random distribution (frequently t, N, χ^2 or F). We can then use this distribution to evaluate how likely it would have been to get the type of sample we have if the null hypothesis was true ( ) or obtain .

Decision Rule 1: If that probability is smaller than our pre-specified significance level, then we H_0. If, however, that p-value is larger than our pre-specified significance level then we will H_0.

Decision Rule 2: If the absolute value of the test statistic is larger than the critical value (obtain from the Null distribution - see next slide), then we H_0. If, however, the absolute value of the test statistic is smaller than the critical value, then we will H_0.

Hypothesis Testing - Introduction

Example The test statistic

$$t = \frac{\bar{LS}_{Ger,2008} - \bar{LS}_{GB,2008}}{\sqrt{\frac{s_{LS,Ger,2008}}{n_{Ger,2008}} + \frac{s_{LS,GB,2008}}{n_{GB,2008}}}}$$

How is this test statistic, t, distributed (assuming H_0 is true)? **If**

1. The two samples are independent
2. The random variables $LS_{i,Ger,2008}$ and $LS_{i,GB,2008}$ are either normally distributed or we have sufficiently large samples
3. The variances in the two samples are identical

then $t \sim$

The above assumptions are crucial (and they differ from test to test). If they are not met then the resulting p-value (or critical values) are not correct.

Hypothesis Testing - Example 1

Let's create a sample statistic:

```
test_data_G <- wb_data %>%
  filter(S003 == "Germany") %>%      # pick German data
  filter(S002EVS == "2008-2010")     # pick latest wave
mean_G <- mean(test_data_G$A170)

test_data_GB <- wb_data %>%
  filter(S003 == "Great Britain") %>%  # pick British data
  filter(S002EVS == "2008-2010")       # pick latest wave
mean_GB <- mean(test_data_GB$A170)

sample_diff <- mean_G - mean_GB
sample_diff
```

```
## [1] -0.7559702
```

Is this different significant?

Hypothesis Testing - Example 1

Formulate a null hypothesis. Here that the difference in population means (`mu`) is equal to 0 using the `t.test` function. We deliver the `A170` series for oth countries to `t.test`.

```
t.test(test_data_G$A170,test_data_GB$A170, mu=0)  # testing that mu = 0
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_G$A170 and test_data_GB$A170
## t = -9.2244, df = 2198.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9166849 -0.5952556
## sample estimates:
## mean of x mean of y
##  6.773619  7.529589
```

The p-value is very small and hence

Hypothesis Testing - Example 2

What about the difference between Great Britain and Sweden though?

```
test_data_SW <- wb_data %>%
  filter(S003 == "Sweden") %>%  # pick British data
  filter(S002EVS == "2008-2010")        # pick latest wave

t.test(test_data_SW$A170,test_data_GB$A170, mu=0)  # testing that mu = 0
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_SW$A170 and test_data_GB$A170
## t = 1.5346, df = 1660.7, p-value = 0.1251
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04153322  0.34022371
## sample estimates:
## mean of x mean of y
##  7.678934  7.529589
```

The p-value is 0.1251 and hence

Hypothesis Testing - To reject or to not reject

When comparing between Germany and Britain the p-value was app. 0:

When comparing between Sweden and Britain the p-value was 0.1251:

- Conventional significance levels are 10%, 5%, 1% or 0.1%
- But what do they mean?

To illustrtate we add a random variable (**rvar**) to all observations. The value comes from the identical distribution for all observations, the standard normal ($N(0,1)$ or **rnorm** in R):

```
wb_data$rvar <- rnorm(nrow(wb_data))    # add random variable

test_data <- wb_data %>%
  filter(S002EVS == "2008-2010")        # pick latest wave

countries <- unique(test_data$S003)  # List of all countries
n_countries <- length(countries)    # Number of countries, 46
```

By construction we know that the true underlying mean is identical in all countries.

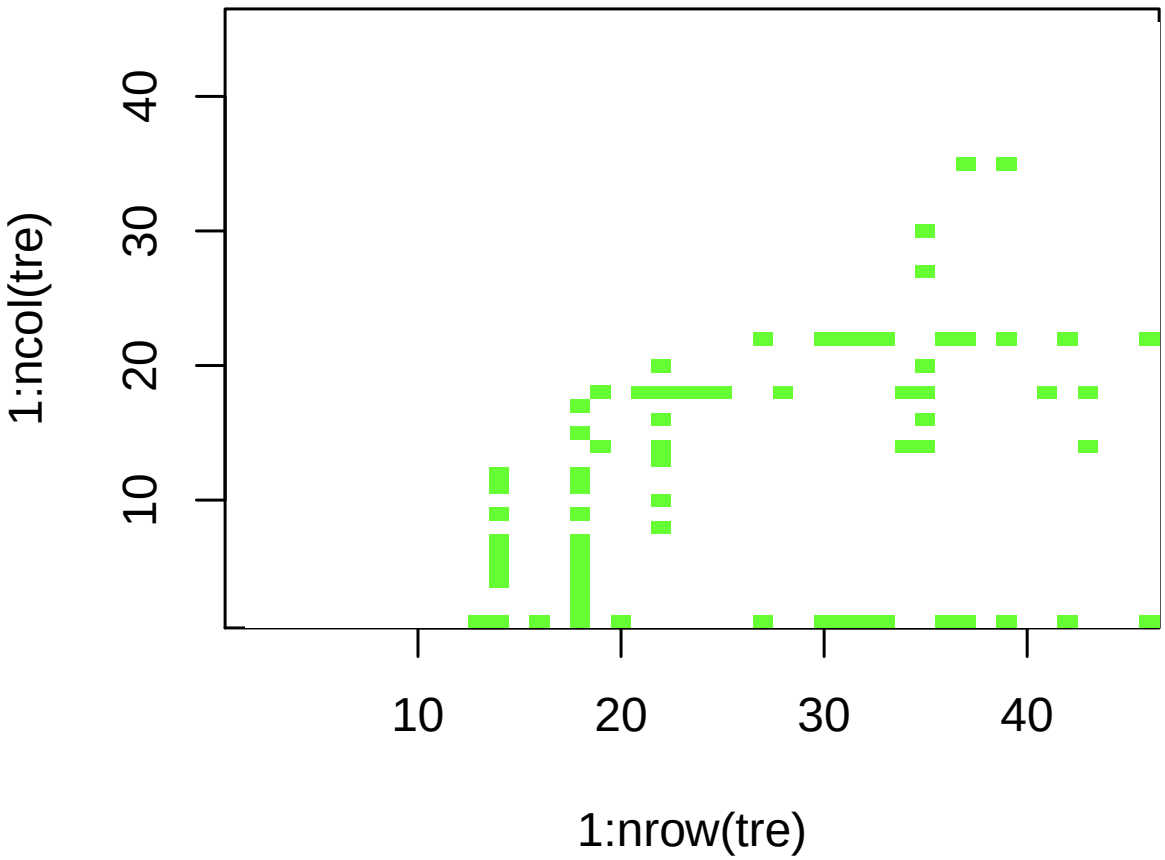
But what happens if we calculate sample means of **rvar** in all countries and then compare between countries?

Hypothesis Testing - To reject or to not reject

We have 1035 hypothesis tests, all testing a correct. If a p-value is smaller than 10% we decide to reject H_0.

```
## tre
## FALSE  TRUE
##   965    70
```

Let's present a graphical representation of these results. Every green square representing a rejection of the null hypothesis.



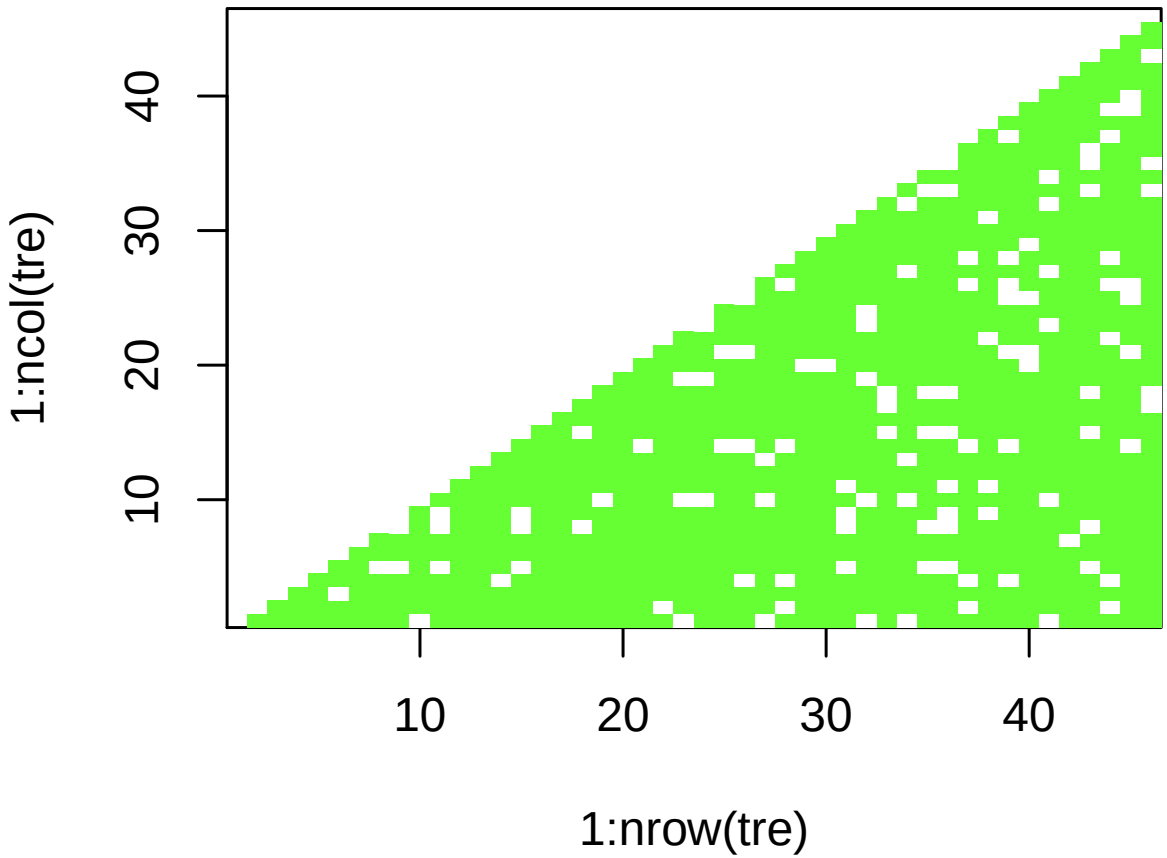
Is this what we would expect?

Hypothesis Testing - To reject or to not reject

Let's return to the Life Satisfaction data and repeat the above comparison between the average Life Satisfaction.

```
## tre
## FALSE  TRUE
##   120   915
```

Every green square representing a rejection of the null hypothesis.



Regression Analysis - Introduction

Tool on which most of the work in this unit is based

- Allows to quantify relationshis between 2 or more variables
- It can be used to implement hypothesis tests
- However it does not necessarily deliver causal relationships!

It is very easy to compute for everyone! Results will often have to be interpretated very carefully.

Your skill will be to interpret correctly!!!!

Regression Analysis - Example 1

Let's start by creating a new dataset which only contains the British data.

```
test_data <- wb_data %>%
  filter(S003 =="Great Britain") %>%  # pick British data
  filter(S002EVS == "2008-2010")          # pick latest wave
```

Now we run a regresison of the Life Satisfaction variable (`A170`) against a constant only.

$LifeSatis_i = \alpha + u_i$

```
mod1 <- lm(A170~1,data=test_data)
```

Regression Analysis - Example 1

We use the `stargazer` function to display regression results

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  A170
## ------------------------------------------------
## Constant                        7.530***
##                                  (0.063)
##
## ------------------------------------------------
## Observations                      997
## R2                               0.000
## Adjusted R2                      0.000
## Residual Std. Error        2.001 (df = 996)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The estimate for the constant, $\widehat{\alpha}$, is the sample mean.

Regression Analysis - Example 1

Testing $H_0 : \mu_{A170} = 0$ can be achieved by

```
##
##  One Sample t-test
##
## data:  test_data$A170
## t = 118.84, df = 996, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  7.405255 7.653922
## sample estimates:
## mean of x
##  7.529589
```

We can use the above regression to achieve the same:

$t - test = \widehat{\alpha}/se_{\widehat{\alpha}}$

Regression Analysis - Example 2

We now estimate a regression model which includes a constant and the household's monthly income (in 1,000 Euros) as an explanatory variable (Inc_i or variable `X047D` in our dataset).

$$LifeSatis_i = \alpha + \beta \; Inc_i + u_i$$

```
mod1 <- lm(A170~X047D,data=test_data)
```

How do we interprete the estimate of $\widehat{\beta}$?

Regression Analysis - Example 2

```
stargazer(mod1, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  A170
## ------------------------------------------------
## X047D                          0.184***
##                                 (0.039)
##
## Constant                       7.190***
##                                 (0.095)
##
## ------------------------------------------------
## Observations                     997
## R2                              0.022
## Adjusted R2                     0.021
## Residual Std. Error       1.980 (df = 995)
## F Statistic           22.302*** (df = 1; 995)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

As the income increases by one unit (increase of Euro 1,000) we should expect that Life Satisfaction increases by 0.184 units.

Regression Analysis - Example 2

Let's present a graphical representation.

```
ggplot(test_data, aes(x=X047D, y=A170)) +
    labs(x = "Income", y = "Life Satisfaction") +
    geom_jitter(width=0.2, size = 0.5) +    # Use jitter - try geom_point() instead
    geom_abline(intercept = mod1$coefficients[1],
                slope = mod1$coefficients[2], col = "blue")+
    ggtitle("Income v Life Satisfaction, Britain")
```

Regression Analysis - What does it actually do?

Two interpretations

1. Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that the residual sum of squares $\Sigma(LifeSatis_i - \widehat{\alpha} - \widehat{\beta}\ Inc_i)^2$. $\rightarrow$ Ordinary Least Squares (OLS)
2. Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that ensures that the residuals ($\widehat{u}_i = LifeSatis_i - \widehat{\alpha} - \widehat{\beta}\ Inc_i$) are with the explanatory variable(s) (here Inc_i).

In many ways 2) is the more insightful one.

Regression Analysis - What does it actually do?

$$LifeSatis = \alpha + \beta\ Inc + u$$

Assumptions

One of the regression assumptions is that the (unobserved) error terms u are uncorrelated with the explanatory variable(s), here Inc. Then we call Inc exogenous.

This implies that $Cov(Inc, u) = Corr(Inc, u) = 0$

In sample

$$LifeSatis_i = \quad \widehat{\alpha} + \widehat{\beta}\ Inc_i \quad + \widehat{u}$$

Where $\widehat{\alpha} + \widehat{\beta}\ Inc_i$ is the regression-line.

In sample $Corr(Inc_i, \widehat{u}_i) = 0$ (is ALWAYS TRUE BY CONSTRUCTION).

Regression Analysis - Underneath the hood?

$LifeSatis = \alpha + \beta\ Inc + u$

What happens if you call

`mod1 <- lm(A170~X047D,data=test_data)`?

You will recall the following from Year 1 stats:

$$\begin{aligned} \hat{\beta} &= \frac{\widehat{Cov}(LifeSatis, Inc)}{\widehat{Var}(Inc)} \\ \hat{\alpha} &= \overline{LifeSatis} - \hat{\beta}\ \overline{Inc} \end{aligned}$$

The software will then replace $\widehat{Cov}(LifeSatis, Inc)$ and $\widehat{Var}(Inc)$ with their sample estimates to obtain $\hat{\beta}$ and then use that and the two sample means to get $\hat{\alpha}$.

Regression Analysis - Underneath the hood?

Need to recognise that in a sample $\hat{\beta}$ and $\hat{\alpha}$ are really .

$$
\begin{aligned}
\hat{\beta} &= \frac{\widehat{Cov}(LifeSatis, Inc)}{\widehat{Var}(Inc)} \\
&= \frac{\widehat{Cov}(\alpha + \beta\ Inc + u, Inc)}{\widehat{Var}(Inc)} \\
&= \frac{\widehat{Cov}(\alpha, Inc) + \beta\widehat{Cov}(Inc, Inc) + \widehat{Cov}(u, Inc)}{\widehat{Var}(Inc)} \\
&= \beta\ \frac{\widehat{Var}(Inc)}{\widehat{Var}(Inc)} + \frac{\widehat{Cov}(u, Inc)}{\widehat{Var}(Inc)} = \beta + \frac{\widehat{Cov}(u, Inc)}{\widehat{Var}(Inc)}
\end{aligned}
$$

So $\hat{\beta}$ is a function of the random term u and hence is itself a random variable. Once $\widehat{Cov}(LifeSatis, Inc)$ and $\widehat{Var}(Inc)$ are replaced by sample estimates we get value which is draw from a

Regression Analysis - The Exogeneity Assumption

Why is **assuming** $Cov(Inc, u) = 0$ important when, in sample, we are guaranteed $Cov(Inc_i, \widehat{u}_i) = 0$?

If $Cov(Inc_i, u_i) = 0$ is **not true**, then

1. Estimating the model by OLS
2. The estimated coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ are

3. The regression model has no

As we cannot observe u_i, the assumption of exogeneity cannot be tested and we need to make an argument using economic understanding.

Regression Analysis - Outlook

$y = \alpha + \beta\ x + u$

Much of empirical econometric analysis is about making the exogeneity assumption ($Corr(x, u) = 0$) more plausible/as plausible as possible. But this begins with thinking why an explanatory variable x is endogenous.

1. Most models have more than one explanatory variable.
2. Including more relevant explanatory variables can make the exogeneity assumption more plausible.
3. But fundamentally, if $Cov(u, x) = 0$ is implausible we need to find another variable z for which $Cov(u, z) = 0$ is plausible.