Data exploration presentation week 2

Claire Le Barbenchon and Federico Ferrari

Welcome to Data Expeditioners Week 2

- Last time, we explored the data and learned how to do a Principal Component Analysis in order to create a poverty index based on household assets in 4 countries. This was your homework last week
- Last time we saw an example that used different elements of a turtle to get a measure of overall size. We can apply the same logic to our poverty data set.

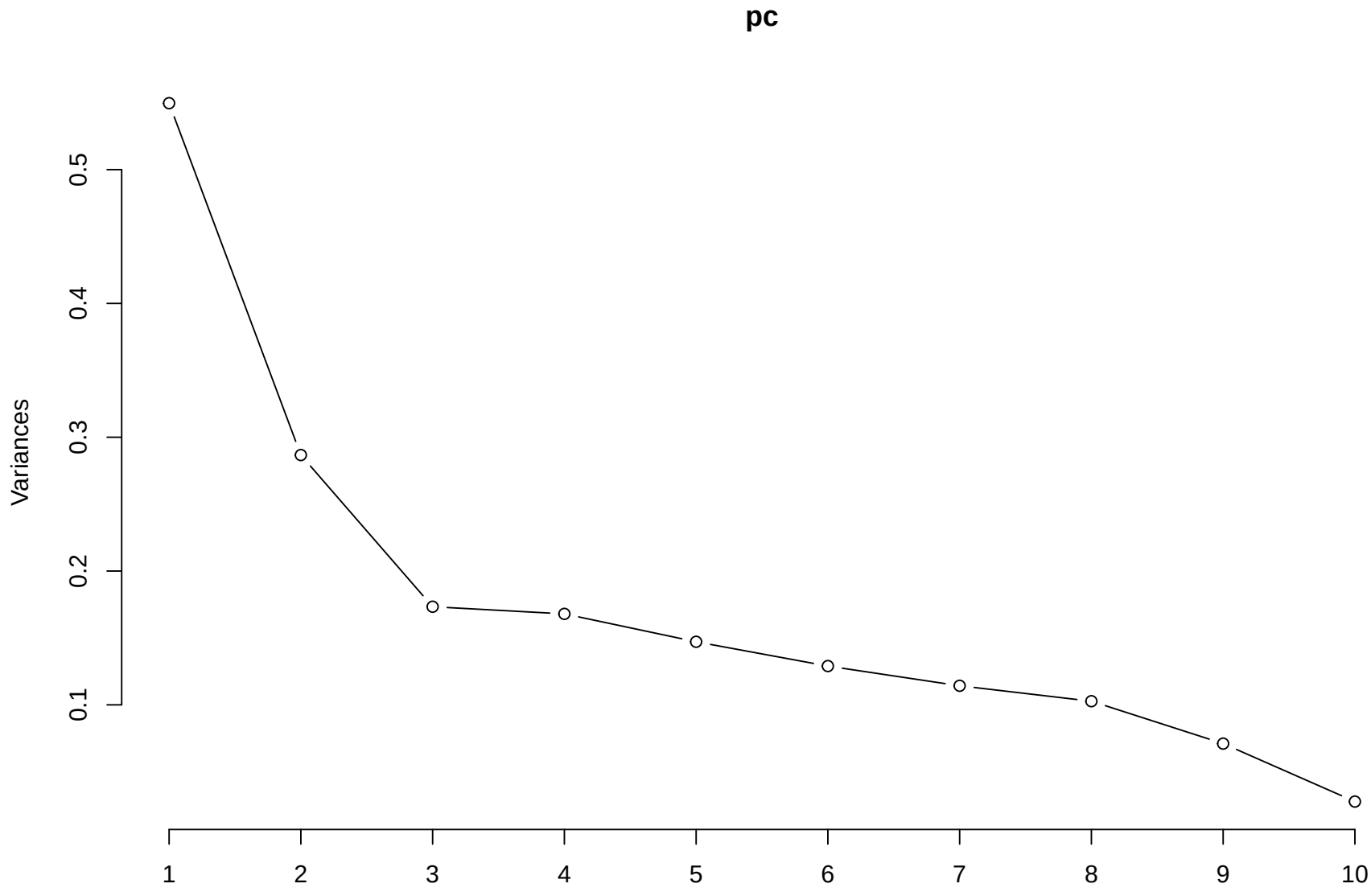
Principal Components Analysis: Review

```
household <- read_csv("data/household-survey.csv")
```

```
#Principal component analysis
assets = subset(household, select=c(stove,refrigerator,tv,bike,motorbike,
                        computer,car,video,stereo,sew))
#correlation pattern
C = cor(assets)
knitr::kable(round(C,1))
```

	stove	refrigerator	tv	bike	motorbike	computer	car	video	stereo	sew
stove	1.0	0.3	0.4	0.0	0.1	0.2	0.2	0.2	0.0	0.1
refrigerator	0.3	1.0	0.5	-0.1	0.0	0.3	0.4	0.4	0.0	0.2
tv	0.4	0.5	1.0	0.0	0.1	0.3	0.2	0.4	0.0	0.1
bike	0.0	-0.1	0.0	1.0	0.1	0.1	0.0	0.1	0.2	0.0
motorbike	0.1	0.0	0.1	0.1	1.0	0.1	0.1	0.1	0.1	0.1
computer	0.2	0.3	0.3	0.1	0.1	1.0	0.4	0.3	0.1	0.1
car	0.2	0.4	0.2	0.0	0.1	0.4	1.0	0.3	0.0	0.1
video	0.2	0.4	0.4	0.1	0.1	0.3	0.3	1.0	0.3	0.1
stereo	0.0	0.0	0.0	0.2	0.1	0.1	0.0	0.3	1.0	0.0
sew	0.1	0.2	0.1	0.0	0.1	0.1	0.1	0.1	0.0	1.0

Principal Components Analysis: Screeplot



Homework Hints

- Remember to keep only the components that show the most variation, in our case this were the first two components
- Stoves has a higher PC coefficient since it summarizes the variation in the assets better than motorbikes. The descriptive statistics show that very few households have motorbike, therefore they do not provide us with useful information.

Principal Component Analysis: Results

	Pc1
stove	-0.32
refrigerator	-0.54
tv	-0.53
bike	-0.01
motorbike	-0.02
computer	-0.22
car	-0.29
video	-0.40
stereo	-0.09
sew	-0.13

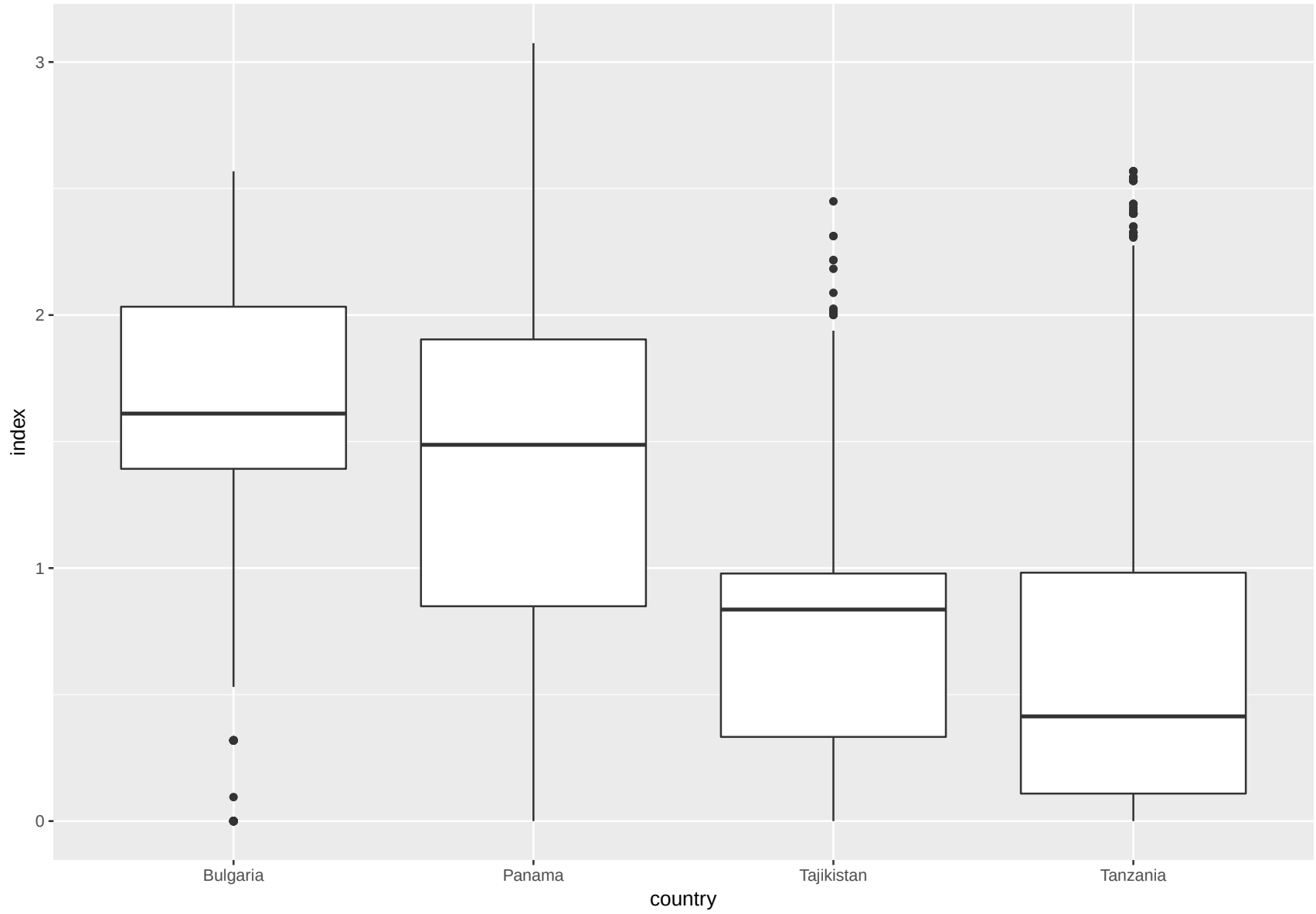
- We can interpret the first principal component as a weighted average of the assets, hence as an index of **poverty/wealth**, which we can use in further analysis.

Principal Component Analysis: Results

	Pc1
stove	-0.32
refrigerator	-0.54
tv	-0.53
bike	-0.01
motorbike	-0.02
computer	-0.22
car	-0.29
video	-0.40
stereo	-0.09
sew	-0.13

- Why does the variable *motorbike* seem to be less important (in terms of contribution to total variance) than for example *refrigerator*?

Inequality in different Countries



Index

We use indeces in many different contexts, for example:

- The *2018 Global Multidimensional Poverty Index (MPI)*
- The *Environmental Sustainability Index (ESI)*
- The *Labor Market Conditions Index*

Applying our Index

- We used PCA to construct a wealth index based on assets, the higher the index value the higher the wealth level.

Applying our Index

- We used PCA to construct a wealth index based on assets, the higher the index value the higher the wealth level.
- Suppose we want to understand what sorts of factors contribute to wealth in countries in our dataset.

$$\widehat{\text{index}}_i = \beta_0 + \beta_1 \times \text{country}_i + \beta_2 \times \text{age}_i + \beta_3 \times \text{relhh}_i + \beta_4 \times \text{educ}_i + \beta_5 \times \text{sex}_i + \beta_6 \times \text{hosp12}_i$$

Factor vs Numeric

```
household %>%
   select(educ,relhh,age) %>%
   glimpse()
```

```
## Observations: 10,000
## Variables: 3
## $ educ  <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
## $ relhh <int> 3, 1, 2, 3, 2, 1, 1, 3, 1, 1, 2, 3, 3, 2,
## $ age   <int> 15, 47, 35, 21, 23, 41, 57, 14, 57, 33, 4(
```

- The variables *relhh*,*sex*,*marstat*,*wat source*,*hosp12* and *educ* are coded as **integers** but they are not actually numbers, and they represent different cathegories. $\rightarrow$ we need to code them as **factors**.

Your Turn: Factor vs Numeric

Trasform the variables from Integers to Factors

Hint: Use the `mutate_at()` function for changing multiple variables at once, together with `as.factor()`

Your Turn: Factor vs Numeric

Trasform the variables from Integers to Factors

Hint: Use the `mutate_at()` function for changing multiple variables at once, together with `as.factor()`

```
household = household %>%
  mutate_at(vars(educ,relhh,sex,marstat,
                 wat_source,hosp12),
            as.factor)
```

Your Turn: Regression Analysis

Run a regression with the following equation

$$\widehat{\text{index}}_i = \beta_0 + \beta_1 \times \text{country}_i + \beta_2 \times \text{age}_i + \beta_3 \times \text{relhh}_i + \beta_4 \times \text{educ}_i + \beta_5 \times \text{sex}_i + \beta_6 \times \text{hosp12}_i$$

Your Turn: Regression Analysis

Run a regression with the following equation

$$\widehat{\text{index}}_i = \beta_0 + \beta_1 \times \text{country}_i + \beta_2 \times \text{age}_i + \beta_3 \times \text{relhh}_i + \beta_4 \times \text{educ}_i + \beta_5 \times \text{sex}_i + \beta_6 \times \text{hosp12}_i$$

```
model_poverty = lm(index ~ country + educ + sex + age +
                     hosp12 + relhh,
                   data = household)
```

Results from Regression

```
tidy(model_poverty)
```

```
## # A tibble: 14 x 5
##    term              estimate std.error statistic   p.value
##    <chr>                <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)         1.05     0.0368      28.6   8.59e-173
##  2 countryPanama      -0.379    0.0262     -14.5   5.39e- 47
##  3 countryTajikistan -0.854    0.0264     -32.4   1.43e-218
##  4 countryTanzania   -0.791    0.0263     -30.1   8.75e-191
##  5 educ1              0.281    0.0192      14.6   7.54e- 48
##  6 educ2              0.562    0.0194      28.9   6.67e-177
##  7 educ3              0.849    0.0261      32.6   4.99e-221
##  8 educ4              0.874    0.0418      20.9   5.72e- 95
##  9 sex1              -0.00788  0.0116      -0.681 4.96e-  1
## 10 age                0.00123  0.000392     3.13  1.73e-  3
## 11 hosp121            0.0224   0.0225       0.994 3.20e-  1
## 12 relhh2             0.0861   0.0152       5.68  1.41e-  8
## 13 relhh3             0.114    0.0185       6.17  7.25e- 10
## 14 relhh4             0.148    0.0203       7.29  3.40e- 13
```

Your Turn: Interpret the Results

Interpret the coefficients of education, age and sex

Your Turn: Interpret the Results

Interpret the coefficients of education, age and sex

- Increasing *education* level increases wealth index
- *Age* has little effect on one's wealth. This is because wealth is measured at the household level, such that children and their parents in the same house would have the same wealth.
- The same holds for *sex*

Interpret the Results

- Why do we estimate the coefficients for only three countries?

Interpret the Results

- Why do we estimate the coefficients for only three countries?
- This happens because we have the intercept in model $\rightarrow$ adding -1 in the regression equation will take the intercept out but include all the countries

```
model_poverty_noint = lm(index ~ -1 + country + educ +
                         sex + age +
                         hosp12 + relhh,
                       data = household)
```

Interpret the Results

- Model **with** intercept

```
##       (Intercept)     countryPanama countryTajikistan   countryTanzania
##       1.052129030      -0.379234117      -0.854280165      -0.791220993
##             educ1             educ2             educ3             educ4
##       0.280884435       0.562301939       0.848955158       0.873982749
##              sex1               age            hosp121             relhh2
##      -0.007884026       0.001227808       0.022417633       0.086099372
##            relhh3            relhh4
##       0.114168957       0.147660284
```

- Model **without** intercept

```
##   countryBulgaria     countryPanama countryTajikistan   countryTanzania
##       1.052129030       0.672894913       0.197848864       0.260908037
##             educ1             educ2             educ3             educ4
##       0.280884435       0.562301939       0.848955158       0.873982749
##              sex1               age            hosp121             relhh2
##      -0.007884026       0.001227808       0.022417633       0.086099372
##            relhh3            relhh4
##       0.114168957       0.147660284
```

Adding Interactions

- It is reasonable to assume that the effect of *education* on the outcome varies in different *countries*
- The *education* in Panama might have a different effect on *wealth* than *education* in Tajikistan.

Your Turn: Add the Interactions

Add the interaction between Country and Education

Hint: Use the symbol $*$ to make two variable interact

Your Turn: Add the Interactions

Add the interaction between Country and Education

Hint: Use the symbol * to make two variable interact

```
model_poverty_int = lm(index ~ -1 + country*educ +
                                sex + age +
                                hosp12 + relhh,
                       data = household)
```

Results from the model with Interactions

```
tidy(model_poverty_int)
```

```
## # A tibble: 26 x 5
##    term                      estimate std.error statistic   p.value
##    <chr>                        <dbl>     <dbl>     <dbl>     <dbl>
##  1 countryBulgaria             0.964    0.0449      21.5   5.30e-100
##  2 countryPanama               0.368    0.0452       8.15  4.23e- 16
##  3 countryTajikistan           0.530    0.0427      12.4   4.05e- 35
##  4 countryTanzania             0.173    0.0318       5.43  5.73e-  8
##  5 educ1                       0.286    0.0389       7.35  2.06e- 13
##  6 educ2                       0.601    0.0370      16.2   1.57e- 58
##  7 educ3                       0.773    0.0534      14.5   5.11e- 47
##  8 educ4                       0.847    0.0500      16.9   1.86e- 63
##  9 sex1                       -0.00168  0.0113      -0.148 8.82e-  1
## 10 age                         0.00252  0.000394     6.41  1.54e- 10
## 11 hosp121                     0.0297   0.0221       1.35  1.78e-  1
## 12 relhh2                      0.0949   0.0149       6.38  1.83e- 10
## 13 relhh3                      0.104    0.0182       5.73  1.02e-  8
## 14 relhh4                      0.142    0.0199       7.15  9.51e- 13
## 15 countryPanama:educ1         0.591    0.257        2.30  2.13e-  2
## 16 countryTajikistan:educ1    -0.263    0.0578      -4.56  5.15e-  6
## 17 countryTanzania:educ1      -0.0215   0.0480      -0.449 6.53e-  1
## 18 countryPanama:educ2         0.217    0.0542       4.01  6.19e-  5
## 19 countryTajikistan:educ2    -0.527    0.0564      -9.34  1.13e- 20
## 20 countryTanzania:educ2       0.376    0.0574       6.56  5.69e- 11
## 21 countryPanama:educ3         0.422    0.0699       6.04  1.64e-  9
## 22 countryTajikistan:educ3    -0.584    0.0784      -7.46  9.65e- 14
## 23 countryTanzania:educ3       0.567    0.100        5.67  1.46e-  8
## 24 countryPanama:educ4         0.519    0.208        2.50  1.25e-  2
## 25 countryTajikistan:educ4    -0.491    0.402       -1.22  2.21e-  1
## 26 countryTanzania:educ4       1.13     0.236        4.79  1.71e-  6
```

Results from the model with Interactions

- How do we calculate and interpret educ1 for Panama?

```
tidy(model_poverty_int) %>% select(term, estimate)
```

```
## # A tibble: 26 x 2
##    term                     estimate
##    <chr>                       <dbl>
##  1 countryBulgaria           0.964
##  2 countryPanama             0.368
##  3 countryTajikistan         0.530
##  4 countryTanzania           0.173
##  5 educ1                     0.286
##  6 educ2                     0.601
##  7 educ3                     0.773
##  8 educ4                     0.847
##  9 sex1                     -0.00168
## 10 age                       0.00252
## 11 hosp121                   0.0297
## 12 relhh2                    0.0949
## 13 relhh3                    0.104
## 14 relhh4                    0.142
## 15 countryPanama:educ1       0.591
## 16 countryTajikistan:educ1  -0.263
## 17 countryTanzania:educ1    -0.0215
## 18 countryPanama:educ2       0.217
## 19 countryTajikistan:educ2  -0.527
## 20 countryTanzania:educ2     0.376
## 21 countryPanama:educ3       0.422
## 22 countryTajikistan:educ3  -0.584
## 23 countryTanzania:educ3     0.567
## 24 countryPanama:educ4       0.519
## 25 countryTajikistan:educ4  -0.491
## 26 countryTanzania:educ4     1.13
```

Results from the model with Interactions

- Why do you think that the coefficients *countryPanama:educ1* is greater than *countryPanama:educ2,3,4* ?

Results from the model with Interactions

- Why do you think that the coefficients *countryPanama:educ1* is greater than *countryPanama:educ2,3,4* ?

- We need to consider the base levels as well!
- educ1 in Panama = educ1 + Panama + countryPanama:educ1

End

Thank You!

- Claire: cl426@duke.edu
- Federico: ff31@duke.edu

Bibliography

- The World Bank, Living Standards Measurement Study LSMS (2007). Bulgaria Multitopic Household Survey 2007 [BGR_2007_MTHS_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/2273/study-description
- The World Bank, Living Standards Measurement Study - Integrated Surveys on Agriculture (2010-2011). Tanzania - National Panel Survey 2010-2011, Wave 2 [TZA_2010_NPS-R2_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/1050
- The World Bank, Living Standards Measurement Study LSMS (2008). Panama - Encuesta de Niveles de Vida 2008 [PAN_2008_ENV_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/70
- Tajikistan Statistical Agency, Living Standards Measurement Study LSMS (2009). Tajikistan - Living Standards Survey 2009 [TJK_2009_TLSS_v01_M]. Retrieved from http://microdata.worldbank.org/index.php/catalog/73%5Bc1%5D