United States and Science Values 2006 - 2011

IS606 Final Project

Daniel Dittenhafer

December 7, 2015

Part 1 - Introduction:

How have the values of people from the United States changed over time with regard to science and technology and its positive/negative impact on the world? The World Values Survey includes a question where in the survey respondent is asked to characterize their view of science and the state of the world because of it (World Values Survey Association, 2014). Using the World Values Survey longitudinal data from 2006 and 2011, this project investigates changes in the views of United States respondents associated with science as well as correlations with educational level and/or the importance of religion.

Part 2 - Data:

The World Values Survey is an observational study, and this data project will be analyzing a subset of the observational study conducted by the World Values Survey Association. The data was collected and made available by the World Values Survey Association via their website. (World Values Survey Association, 2014).

The original data set in its entirety is available from the World Values Survey website: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp

Although the World Values Survey was conducted in the United States starting in 1995, the question regarding science and the world (005_203) was not introducted until 2006. As such, only the cases from 2006 and beyond, 3481 cases across 2006 and 2011, are considered in this study.

Scope of Inferencce Based on the data, collection methods and type of study, we review the scope of inference this data project can claim.

Generalizability: The population associated with the survey in this data project is the population of the United States between the ages of 18 and 85. *Is there bias, or is the data not representative of the age range?*

Causality: The data is derived from an observational survey and not an experiment. As such, no causality can be established from any conclusion.

Variables Included in Data Subset The following table lists the variables extracted from the original World Values Survey dataset which will be considered in this data project.

ID	Variable	Description
005_203	E234	The world is better off, or worse off, because of science and technology
010_023	S020	Year survey
010_028	S024	Country wave
010_004	S003	Country/region
014_003	X003	Age
014_030	X025	Highest educational level attained
001_006	A006	Important in life: Religion

Response Variable The response variable is the answer provided to the key question, "Is the world is better off, or worse off, because of science and technology?"

The answers are categorical in nature, but ordinal in their degree of support for the better/worse outcome. The distinct answer values and the description of each value are shown in the table below. Note that negative values are variations on missing data. These will be eliminated during the exploratory data phase.

Value	Description
1	A lot worse off
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A lot better off
-5	Missing; Unknown
-4	Not asked in survey
-3	Not applicable
-2	No answer
-1	Don´t know

Explanatory Variable(s) The explanatory variables considered for analysis are the "Highest educational level attained" value and the "Important in life: Religion" variable. Again, these are categorical variables, but ordinal in terms of level of education and degree of importance of religion. The values and descriptions for each variable follow:

Highest educational level attained

Value	Description
1	Inadequately completed elementary education
2	Completed (compulsory) elementary education
3	Incomplete secondary school: technical/vocational type/elementary education, basic vocational qual.
4	Complete secondary school: technical/vocational type/Secondary, intermediate vocational qualification
5	Incomplete secondary: university-preparatory type/Secondary, intermediate general qualification
6	Complete secondary: university-preparatory type/Full secondary, maturity level certificate
7	Some university without degree/Higher education - lower-level tertiary certificate
8	University with degree/Higher education - upper-level tertiary certificate
-5	Missing; Unknown
-4	Not asked in survey
-3	Not applicable; No formal education
-2	No answer
-1	Don´t know

Important in life: Religion

Value	Description
-5	Missing; Unknown
-4	Not asked in survey

Value	Description
-3	Not applicable
-2	No answer
-1	Don´t know
1	Very important
2	Rather important
3	Not very important
4	Not at all important

Part 3 - Exploratory data analysis:

Let us explore the variables within the data subset. First, some simple summary statistics:

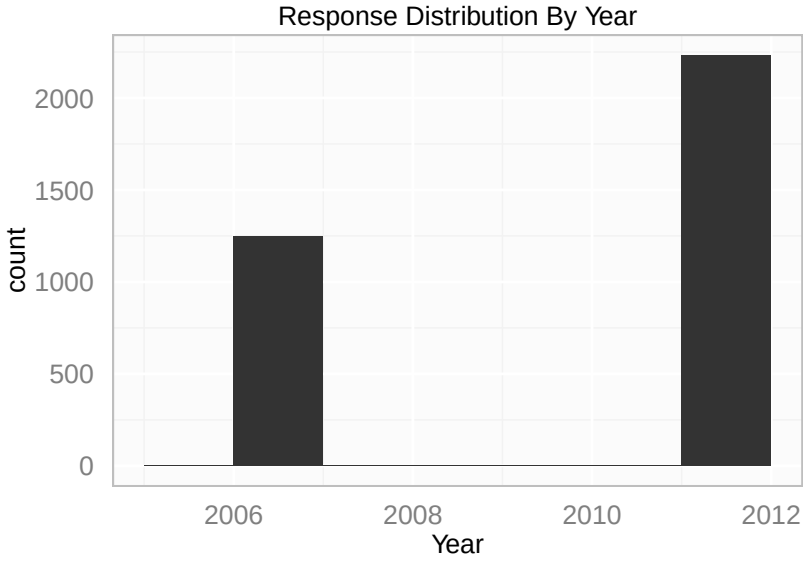
```
##   KeyQuestion          Year       CountryWave      Country
##  Min.   :-5.000   Min.   :1995   Min.   :8403   Min.   :840
##  1st Qu.:-4.000   1st Qu.:1999   1st Qu.:8404   1st Qu.:840
##  Median : 5.000   Median :2006   Median :8405   Median :840
##  Mean   : 2.236   Mean   :2004   Mean   :8405   Mean   :840
##  3rd Qu.: 8.000   3rd Qu.:2011   3rd Qu.:8406   3rd Qu.:840
##  Max.   :10.000   Max.   :2011   Max.   :8406   Max.   :840
##       Age        HighestEducation ReligionImportant CountryName
##  Min.   :-1.00   Min.   :-3.000   Min.   :-2.0      Length:6223
##  1st Qu.:33.00   1st Qu.: 4.000   1st Qu.: 1.0      Class :character
##  Median :46.00   Median : 6.000   Median : 1.0      Mode  :character
##  Mean   :47.15   Mean   : 5.771   Mean   : 1.8
##  3rd Qu.:61.00   3rd Qu.: 8.000   3rd Qu.: 2.0
##  Max.   :94.00   Max.   : 8.000   Max.   : 4.0
```

The following table shows the mean and standard deviation of the response variable for the various years for which the survey was conducted. Unfortunately, in 1995 and 1999 the question regarding science was not asked as evidenced by the mean of -4 ("Not asked in survey") and standard deviation of 0 (no variation). As a result, the focus will be on differences between 2006 and 2011.

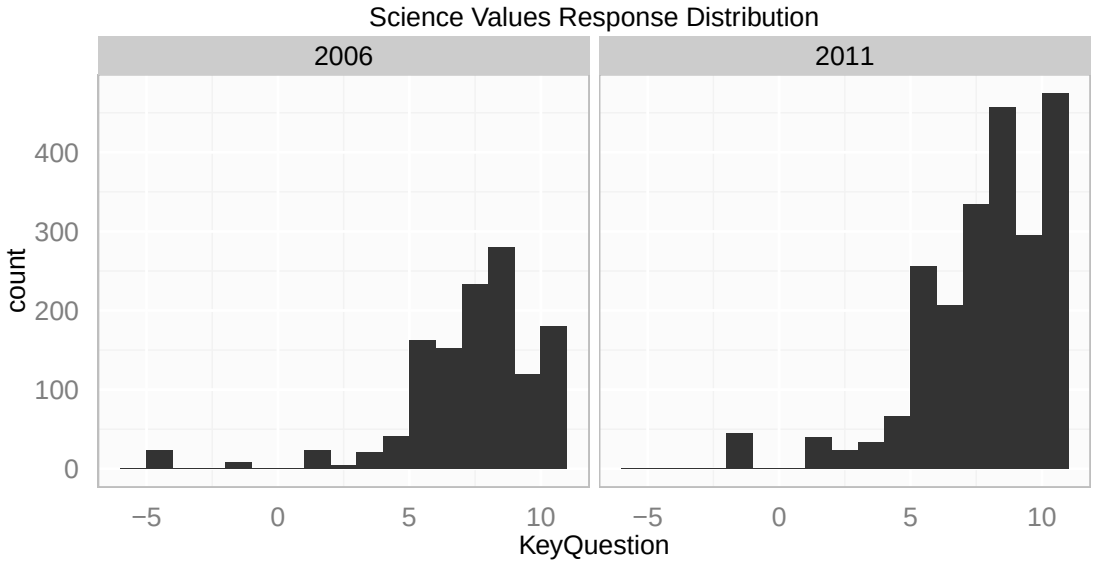
Year	Mean	Std Dev
1995	-4.000000	0.000000
1999	-4.000000	0.000000
2006	6.885508	2.671164
2011	7.294803	2.483114

After eliminating the 1995 and 1999 data rows, 3481 cases remain. Of those, 35.88% are in 2006, and 64.12% are in 2011. The number of cases in the sample is less than 10% of the United States population in 2006 299,398,484 (United States Census Bureau, 2006). The same holds true for the 2011 population at 311,591,917 (United States Census Bureau, 2011).

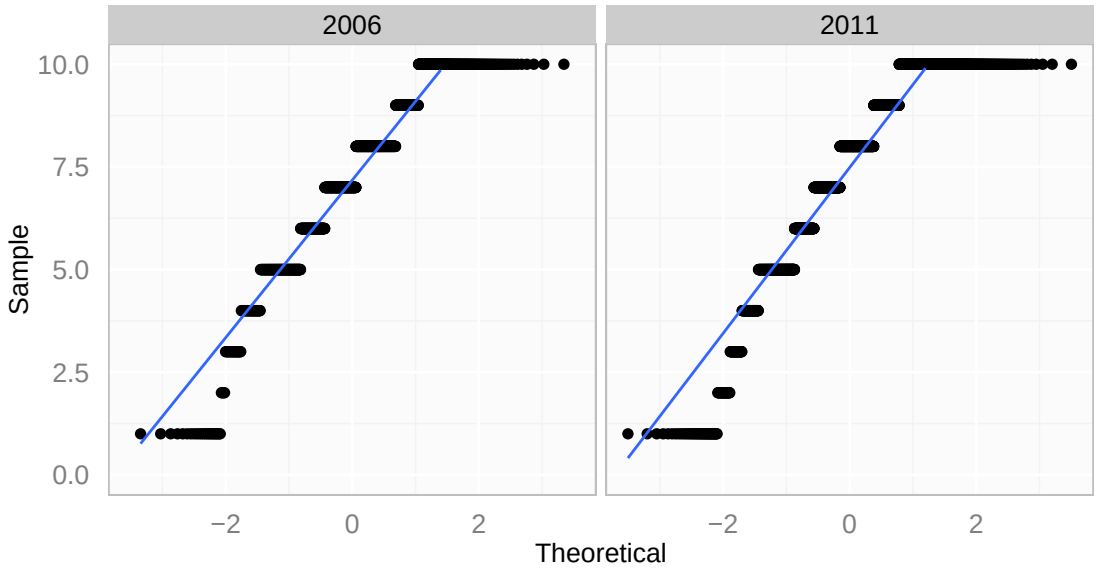
Year	Cases	Ratio
2006	1249	35.88049
2011	2232	64.11951
Total	3481	100.00000



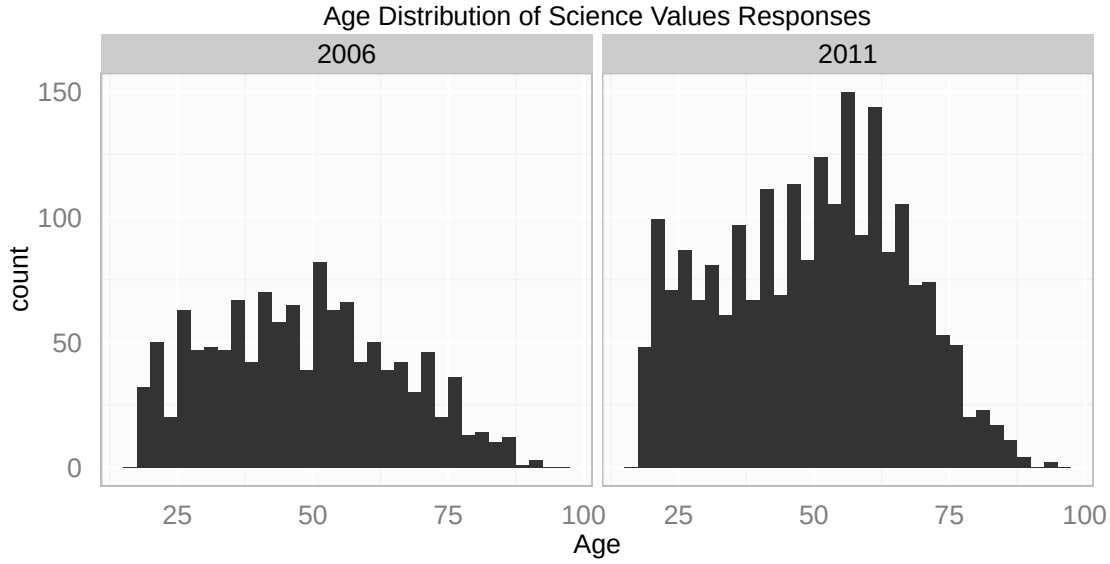
When viewing the data distribution of answer values for our `KeyQuestion` in each year, we see different distributions, as shown in the following Science Values Response Distribution histograms. Both years are skewed left, but 2006 appears a bit more normal where 2011 could be characterized as bi-modal at values 8 and 10. We see also that ~2.56% of 2006 answers and ~2.02% of 2011 answers are negative and constitute missing values. These data rows will be removed shortly.



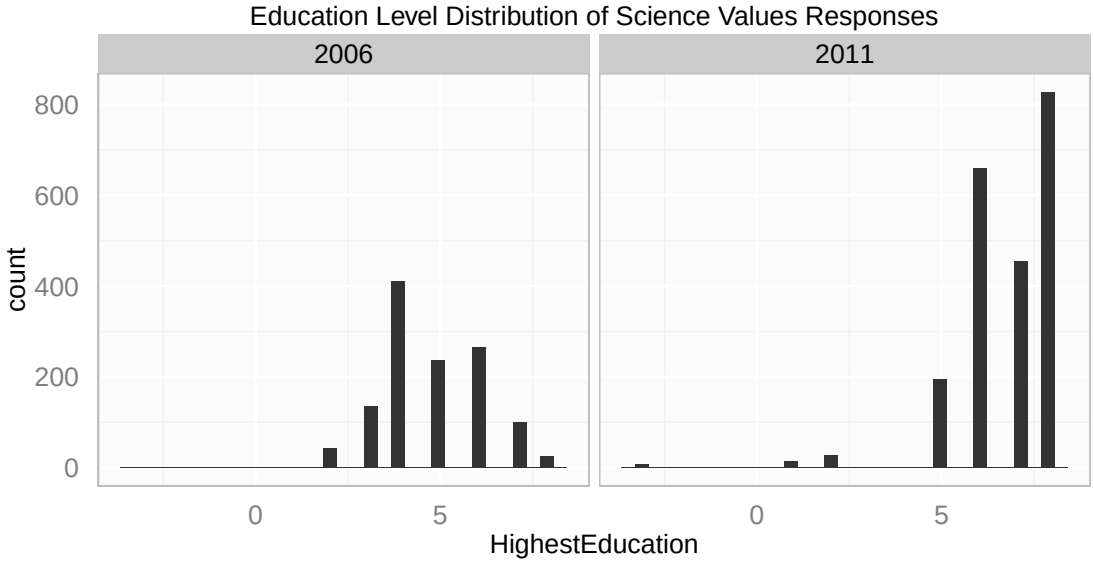
The missing/unknown response data rows have now been removed and we further explore the data. A Quartile-Quartile plot will be used to check the normalness of these distributions in more detail. As shown in the following charts, the distributions are not strictly normal, but for our purposes we will accept them as nearly normal. This allows us to proceed with the difference of two means analysis in Part 4 - Inference.



Age In the following charts we see the age distribution of respondents to the science question. Both years are somewhat normally distributed but with a right skew. Older people are less represented in the data particularlly over the age of 75. Obviously we have more responses in 2011 (which matches the Response by Year result seen earlier).



Education Level Now lets explore the education level of the science values respondents. Recall that the highest level in this survey is 8 - "University with degree/Higher education - upper-level tertiary certificate"", and -3 ="Not applicable; No formal education".



The distributions shown in the previous charts raise some questions. Did the education level distribution change that much in 5 years? What bias's might be introduced with these education levels? Unfortunately, these questions are out of scope for this data project.

Part 4 - Inference:

Our initial interest relates to differences between years for the response variable. Is there a statistically significant difference between 2006 and 2011 for the response to the science and the world question?

The following table lists the mean and standard deviation for our cleansed data set per our exploratory data analysis (Part 3).

Year	Mean	Std Dev	n
2006	7.178307	1.982304	1217
2011	7.486054	2.116065	2187

Our hypothesis test is set up as follows:

$H_0 : \mu_{2006} - \mu_{2011} = 0$

$H_a : \mu_{2006} - \mu_{2011} \neq 0$

$\alpha = 0.05$

The point estimate for the difference in means is -0.3077466.

$$\bar{x}_{2006} - \bar{x}_{2011} = -0.3077466$$

The standard error of the point estimate becomes:

$$SE_{\bar{x}_{2011} - \bar{x}_{2006}} = \sqrt{\frac{\sigma_{2011}^2}{n_{2011}} + \frac{\sigma_{2006}^2}{n_{2006}}} = 0.0726381$$

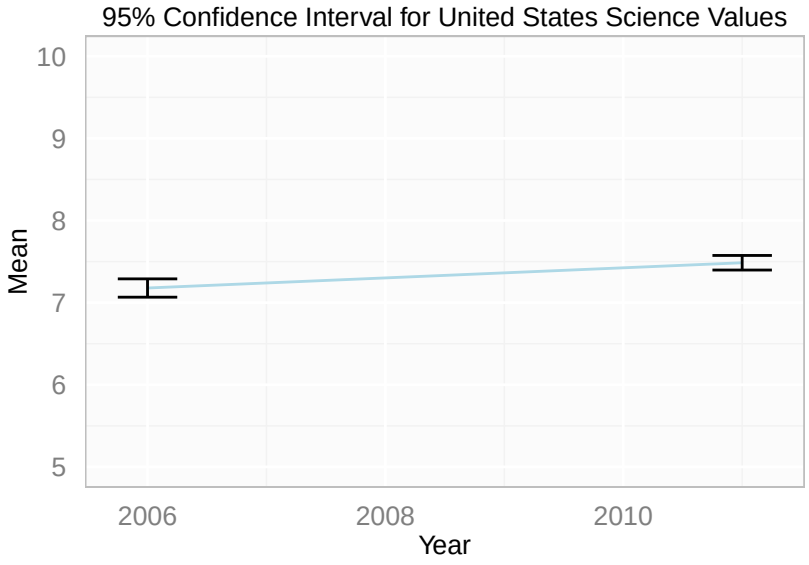
The T-score is computed as:

$$T = \frac{-0.3077466 - 0}{0.0726381} = -4.2367097$$

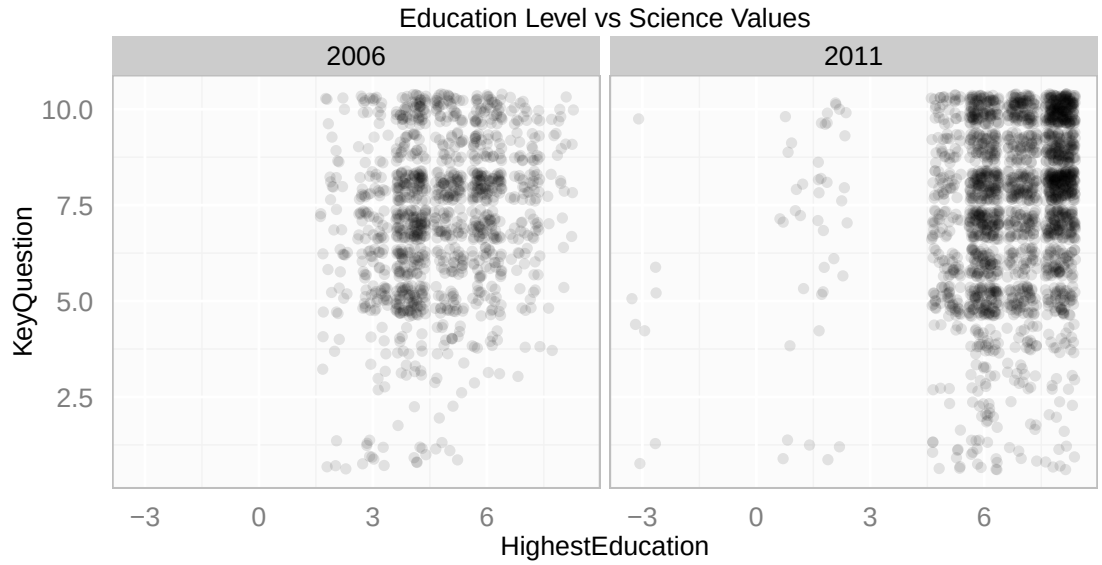
Using degrees of freedom based on the smaller of $n_{2006} - 1 = 1216$ vs $n_{2011} - 1 = 2186$: df=1216 we get a p-value $\approx 0 < 0.05$. Therefore, we reject the null hypothesis and conclude that the United States' mean view of the world being better off because of science and technology has increased from 2006 to 2011.

We can construct a 95% confidence interval around the response variable's means and visualize it to get sense of the change and the range of probable population values.

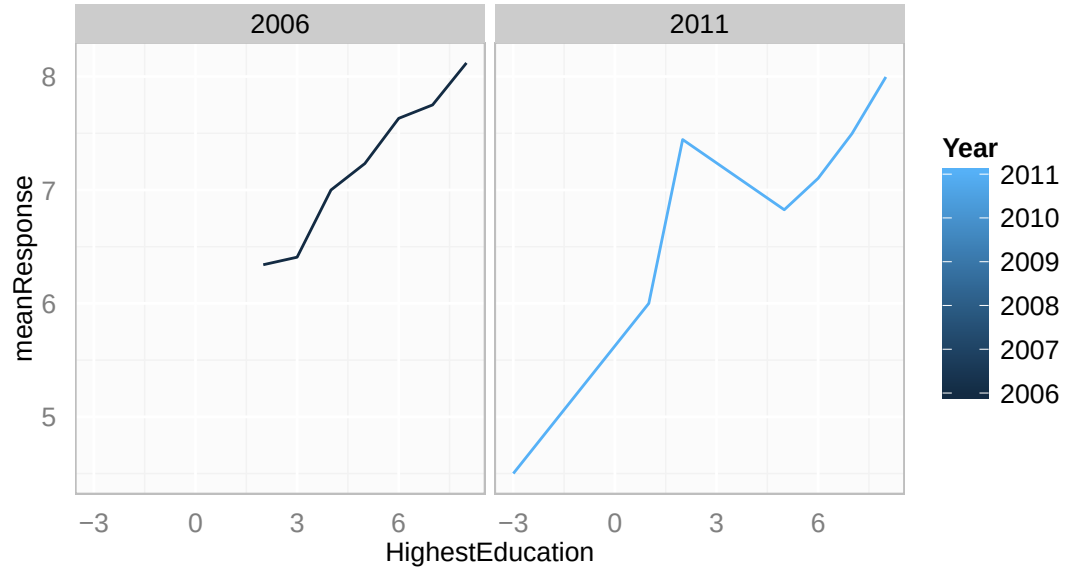
Year	Mean	Std Dev	n	LowerBound	UpperBound
2006	7.178307	1.982304	1217	7.066825	7.289790
2011	7.486054	2.116065	2187	7.397319	7.574789



Now lets focus on what factors might be related to the belief that science is making the world better off. We'll start out looking at education level through a simple scatter plot. As can be seen in both the combined and per year plots, there doesn't appear to be any dominate relationship.



When viewing the the response variable mean by educational level, it appears there is a positive relationship.



Our hypotheses for an analysis of variance (ANOVA) test are as follows:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8$, The mean science values response is the same across all education levels.

H_a : At least one mean is different.

```
aov_ed_sci <- aov(KeyQuestion ~ HighestEducation, data=WVS_US_0611_pos)
summary(aov_ed_sci)
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## HighestEducation    1    654   654.2   159.1 <2e-16 ***
## Residuals        3402  13986     4.1
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dfEdFactorStats
```

```
##    HighestEducation Year meanResponse sdResponse
## 1                -3 2011     4.500000   2.878492
## 2                 1 2011     6.000000   3.113247
## 3                 2 2006     6.340909   2.514545
## 4                 2 2011     7.444444   2.577019
## 5                 3 2006     6.407407   2.312511
## 6                 4 2006     7.000000   1.929236
## 7                 5 2006     7.233051   2.044382
## 8                 5 2011     6.825641   2.322094
## 9                 6 2006     7.631579   1.650494
## 10                6 2011     7.103030   2.160128
## 11                7 2006     7.750000   1.604130
## 12                7 2011     7.496703   2.108168
## 13                8 2006     8.120000   1.715615
## 14                8 2011     7.996377   1.837731
```

```
m_sci_ed <- lm(HighestEducation ~ KeyQuestion, WVS_US_0611_pos)
summary(m_sci_ed)
```

```
##
## Call:
## lm(formula = HighestEducation ~ KeyQuestion, data = WVS_US_0611_pos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4960 -0.9884  0.1039  1.5040  3.0269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.80385    0.10278   46.74   <2e-16 ***
## KeyQuestion  0.16922    0.01341   12.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.623 on 3402 degrees of freedom
## Multiple R-squared:  0.04468,    Adjusted R-squared:  0.0444
## F-statistic: 159.1 on 1 and 3402 DF,  p-value: < 2.2e-16
```

Part 5 - Conclusion:

References:

United States Census Bureau. Annual Population Estimates. 2011. URL: http://www.census.gov/popest/data/state/totals/201 EST2011-01.csv.

— Annual Population Estimates 2000 - 2006. 2006. URL: http://www.census.gov/popest/data/state/totals/2006/tables/NST-EST2006-01.csv.

World Values Survey Association. WORLD VALUES SURVEY 1981-2014 LONGITUDINAL AGGREGATE v.20150418. Aggregate File Producer: JDSystems. Madrid SPAIN, 2014. URL: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp.

Appendix (optional):

Remove this section if you don't have an appendix