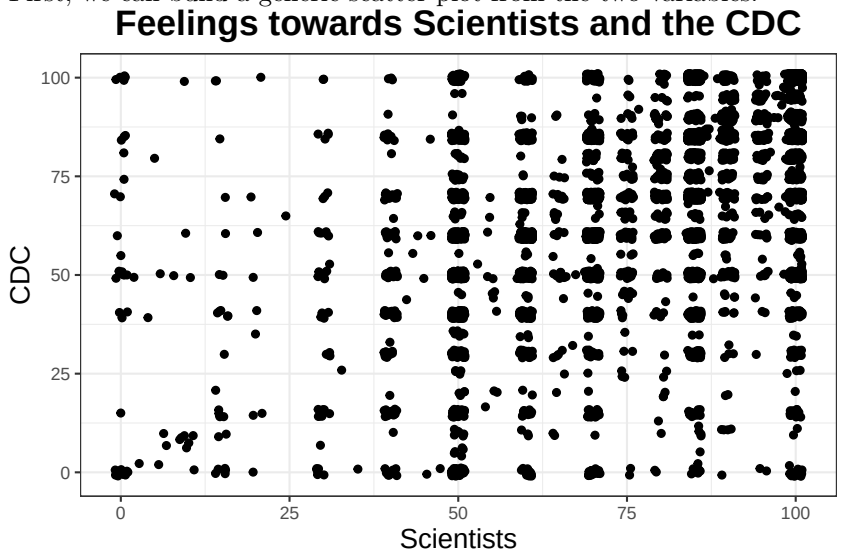
Week 7: Regression

Jennifer Lin

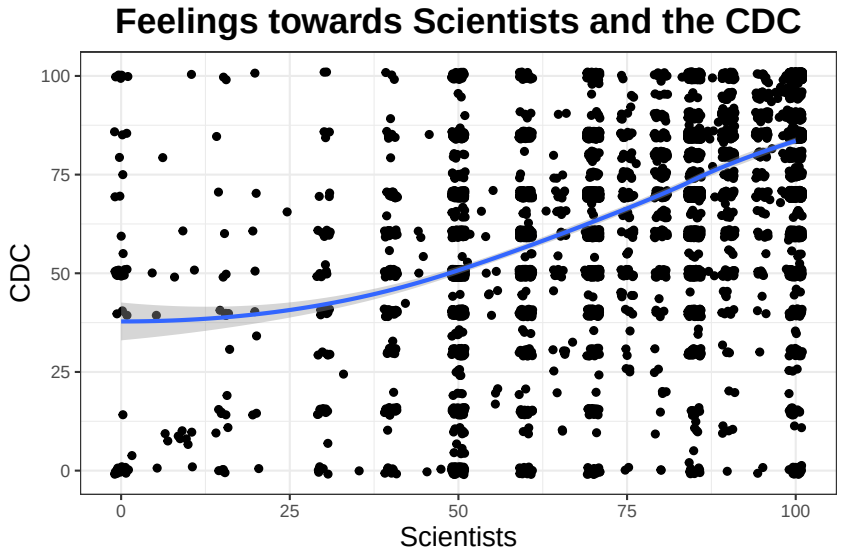
2021-11-03

For the exercises that we are going to go through today, we will use the American National Elections Studies (ANES). It is conducted every 4 years and measures a variety of political positions on various topics. Specifically, they include a series of feeling thermometers that measure how warmly people feel towards a particular subject from a scale of 0 to 100. Today, we will look at the feeling thermometers towards Scientists and the CDC.

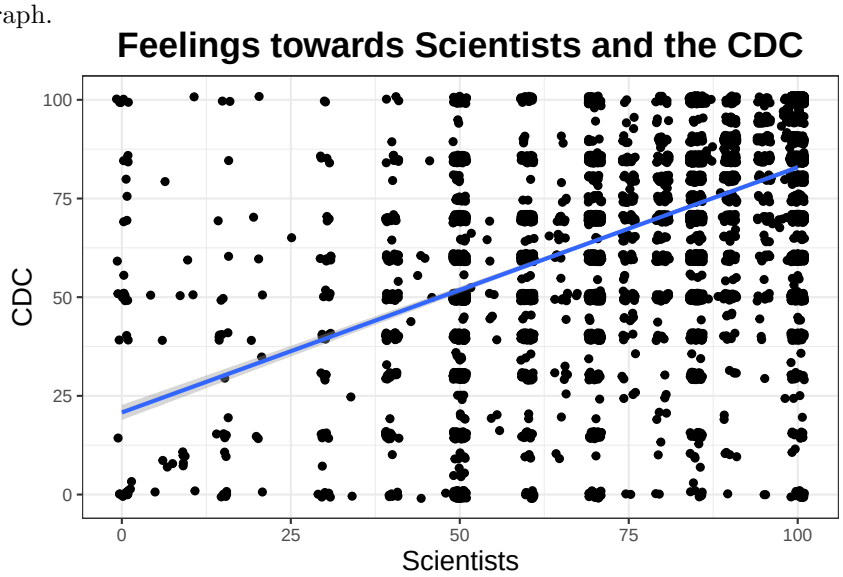
First, we can build a generic scatter plot from the two variables.



To find general trends, we can then fit what is called a "loess" line which finds the best curve based on all the points – it is not necessarily linear but it can give you a good idea of the variable relationships, so you can see if it is more linear or curvilinear.



Since the loess line looks more linear, we can go ahead and see what a standard Ordinary Least Squares Regression line would look on the graph.



To get a better understanding of the relationship of these variables, we can run a correlation. Using only the complete observations, we can see that the variable have a moderate, positive correlation.

```
cor(ANES$FT_Scientist, ANES$FT_CDC, use = "complete.obs")
```

```
## [1] 0.5235262
```

Regression

Using R

So if we want to predict a dependent variable from any set of independent variables, we can use Ordinary Least Squares Regressions (OLS), which is the Best Linear Unbiased Estimator (BLUE) of a relationship.

```
##
## Call:
## lm(formula = FT_CDC ~ FT_Scientist, data = ANES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.896 -11.846   2.104  14.322  79.205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.79532    0.97504   21.33   <2e-16 ***
## FT_Scientist  0.62101    0.01189   52.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.11 on 7230 degrees of freedom
##   (1048 observations deleted due to missingness)
## Multiple R-squared:  0.2741, Adjusted R-squared:  0.274
## F-statistic:  2730 on 1 and 7230 DF,  p-value: < 2.2e-16
```

From the results, the "Estimate" shows the predictor, "Std. Error" is the standard error, "t value" is the t-test value and "$Pr(> |t|)$" is the p-value associated with the particular variable. We can use the Estimate value next to the independent variable to interpret the results

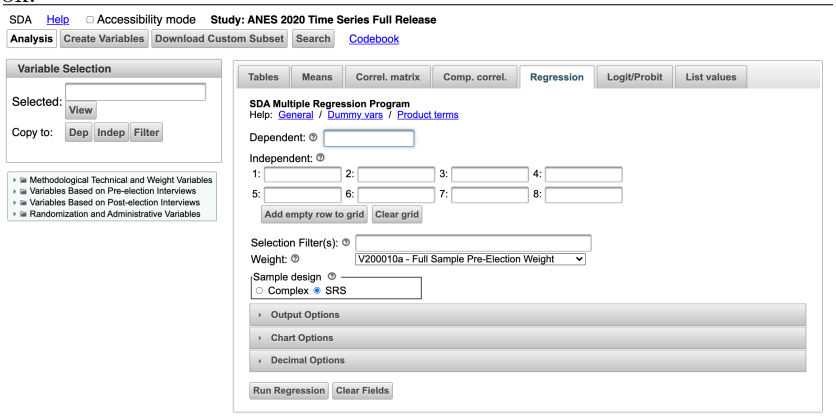
> *ON AVERAGE, for every one unit increase in Feelings towards Scientists, feelings towards the CDC increases by 0.62101*

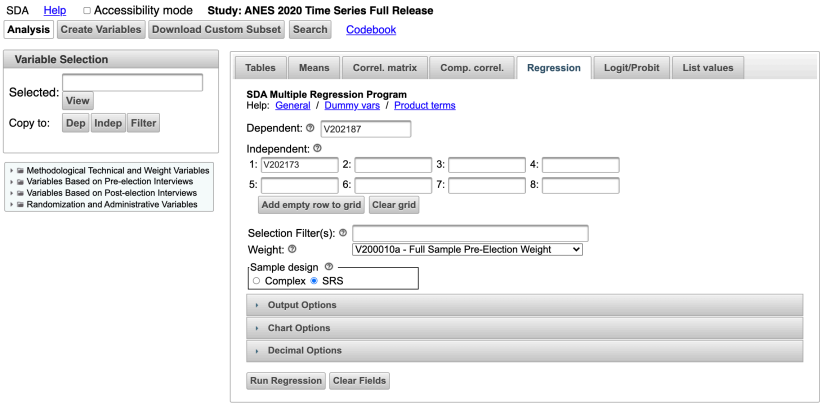
Why are we using the phrase "On Average"? We are making predictions based on a line that is drawn from the best fit (average) of the data. Therefore, the predictions only reflect average cases rather than extreme cases.

Using the Software

We can use a more user friendly tool to run regressions on the ANES data – `https://sda.berkeley.edu/sdaweb/analysis/?dataset=nes2020full`

Using the Regression tab, we con import dependent and independent variables using the variable names as they appear in the codebook.



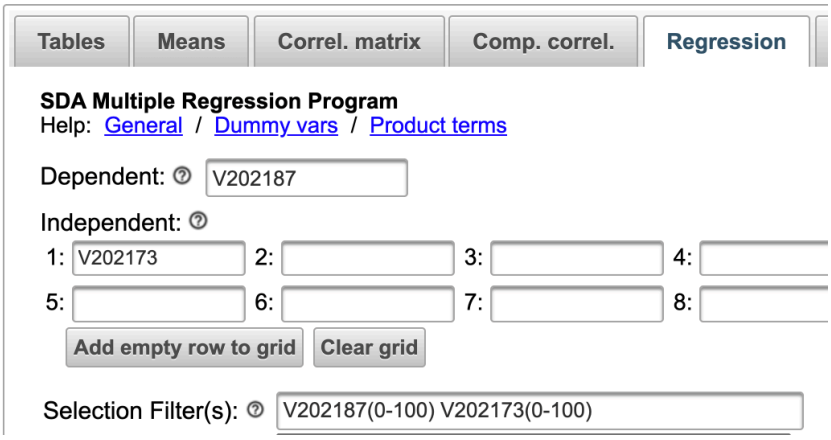


To get the variables, look at the codebook. Here, you can see information on the ways that the variables are coded and the question wordings. We can see if there are any values in the data that might be included but should be treated as "missing" data. For example, looking at the feeling thermometers, missing data is coded as negative values and 998/999 as don't know, which is not in the 0 - 100 range and can throw off the predictions.

V202173	POST: FEELING THERMOMETER: SCIENTISTS
Question	How would you rate: Scientists
Value Labels	-9. Refused -7. No post-election data, deleted due to incomplete interview -6. No post-election interview -5. Interview breakoff (sufficient partial IW) -4. Technical error 998. Don't know
Survey Question(s)	THERMGR_SCIENT
Randomization	Set 1: Randomize the order of THGRFUND, THGRFEM, THGRLIB, THGRLAB, THGRBIGB, THGRCONS, THGRSCT, THGRGAY, THGRCONG, THGRMUSL, THGRXTIAN, JEWS, POLICE, TRANS, SCIENT, BLM, JOURN
Interviewer Instruction	{PROBE FOR DON'T KNOW RESPONSE: when you say don't know, do you mean that you don't know who this is or do you have something else in mind? ENTER number 0-100 }

V202187	POST: FEELING THERMOMETER: CENTER FOR DISEASE CONTROL (CDC)
Question	How would you rate: The Centers for Disease Control (CDC)
Value Labels	-9. Refused -7. No post-election data, deleted due to incomplete interview -6. No post-election interview -5. Interview breakoff (sufficient partial IW) -4. Technical error 998. Don't know 999. Don't recognize
Survey Question(s)	THERMGR_CDC
Randomization	Set 2: Randomize the order of NATO, UN, NRA, SOCIAL, CAPITAL, FBI, ICE, METOO, RURAL, PLANPARENT, WHO, CDC
Interviewer Instruction	{PROBE FOR DON'T KNOW RESPONSE: when you say don't know, do you mean that you don't know who this is or do you have something else in mind? ENTER number 0-100 }

In the software, you can filter IN the observations so that the data do not reflect missing data. Simply do `VARNAME(start - end)` and separate the variables (if multiple) with a space or comma.



We can run the regression and look at the results. Notice how it matches my results from before!

SDA 4.1.3: Regression

ANES 2020 Time Series Full Release

Nov 03, 2021 (Wed 02:50 PM PDT)

Variables

Role	Name	Label	Range	MD	Dataset
Dependent	**V202187**	POST: Feeling thermometer: Center for Disease Control (CDC)	0-999		1
Independent	**V202173**	POST: Feeling thermometer: scientists	0-998		1
Filter	**V202187(0-100)**	POST: Feeling thermometer: Center for Disease Control (CDC)	0-999		1
Filter	**V202173(0-100)**	POST: Feeling thermometer: scientists	0-998		1

	Regression Coefficients				Test That Each Coefficient = 0	
	B	SE(B)	Beta	SE(Beta)	T-statistic	Probability
V202173	.621	.012	.524	.010	52.247	.000
Constant	20.795	.975			21.328	.000

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Effect of each variable:	Negative			Positive			

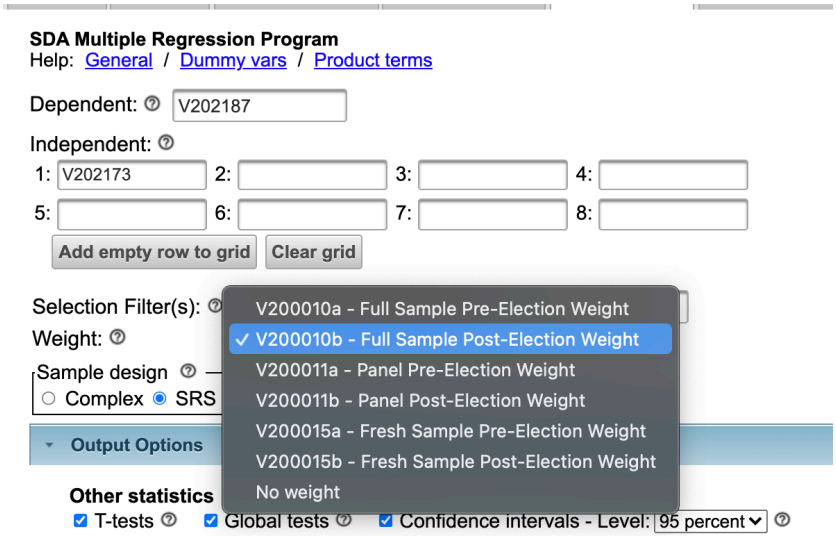
Multiple R = .524 R-Squared = .274 Adjusted R-Squared = .274 SE of Estimate (Root MSE) = 20.115

Global Tests for Groups of Variables

Group	Wald Chi-sq	df		Adjusted Wald F	P
		Numerator	Denominator		
All independent variables	2,729.771	1	7231	2,729.771	.000

In the preceding analyses, we did not weight the analyses, but for survey data, we should put weights on the data and in the analyses so we can account for biases in the sampling process.

From the dropdown menu, we can select the weight variable that we want, which, in the ANES, reflects the sample of data that we are interested in. Since the feeling thermometer variables are all in the post-election sample, we use the full-panel, post election weight.



Notice how the estimates change ever so slightly as you are accounting for biases in the sampling process

SDA 4.1.3: Regression

ANES 2020 Time Series Full Release

Nov 03, 2021 (Wed 03:08 PM PDT)

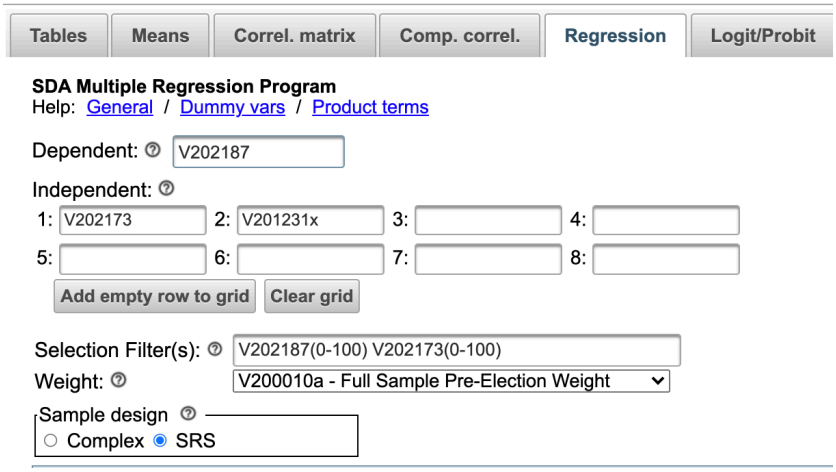
Variables		
Role	**Name**	**Label**
Dependent	**V202187**	POST: Feeling thermometer: Center for Disease Control (CDC)
Independent	**V202173**	POST: Feeling thermometer: scientists
Weight	**V200010b**	Full sample post-election weight
Filter	**V202187(0-100)**	POST: Feeling thermometer: Center for Disease Control (CDC)
Filter	**V202173(0-100)**	POST: Feeling thermometer: scientists

Regression Coefficients					Test That Each Coefficient = 0	
	B	**SE(B)**	**Beta**	**SE(Beta)**	**T-statistic**	**Probability**
V202173	.574	.012	.491	.010	47.872	.000
Constant	24.453	.972			25.150	.000

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Effect of each variable:	Negative			Positive			

Multiple R = .491 R-Squared = .241 Adjusted R-Squared = .241 SE of Estimate (Root MSE

Now, suppose, we want to add confounding variables, such as partisanship to the regression analysis.



Looking at the results, we now see estimates for all of the independent variables, but the one that that we would need to interpret are those associated with the main independent variable.

SDA 4.1.3: Regression

ANES 2020 Time Series Full Release

Nov 03, 2021 (Wed 03:00 PM PDT)

Variables		
Role	**Name**	**Label**
Dependent	**V202187**	POST: Feeling thermometer: Center for Disease Control (CDC)
Independent	**V202173**	POST: Feeling thermometer: scientists
Independent	**V201231x**	PRE: SUMMARY: Party ID
Weight	**V200010a**	Full sample pre-election weight
Filter	**V202187(0-100)**	POST: Feeling thermometer: Center for Disease Control (CDC)
Filter	**V202173(0-100)**	POST: Feeling thermometer: scientists

Regression Coefficients					Test That Each Coefficient = 0	
	B	**SE(B)**	**Beta**	**SE(Beta)**	**T-statistic**	**Probability**
V202173	.514	.013	.435	.011	40.543	.000
V201231x	-1.979	.114	-.185	.011	-17.296	.000
Constant	36.929	1.259			29.341	.000

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	T
Effect of each variable:	Negative			Positive			

ON AVERAGE, for every one unit increase in Feelings towards Scientists, feelings towards the CDC increases by 0.514 ($p < .000$)

Summary

Here is the template for interpreting regression results

> ***On average, for every one unit increase in [X], [Y] [increases/decreases] by [ESTIMATE]***

This is intuitive, if you think about the algebra of a linear model

$$[Y] = [ESTIMATE] * [X] + error$$

If [X] is 1, the [Y] increases by the value of the [ESTIMATE]