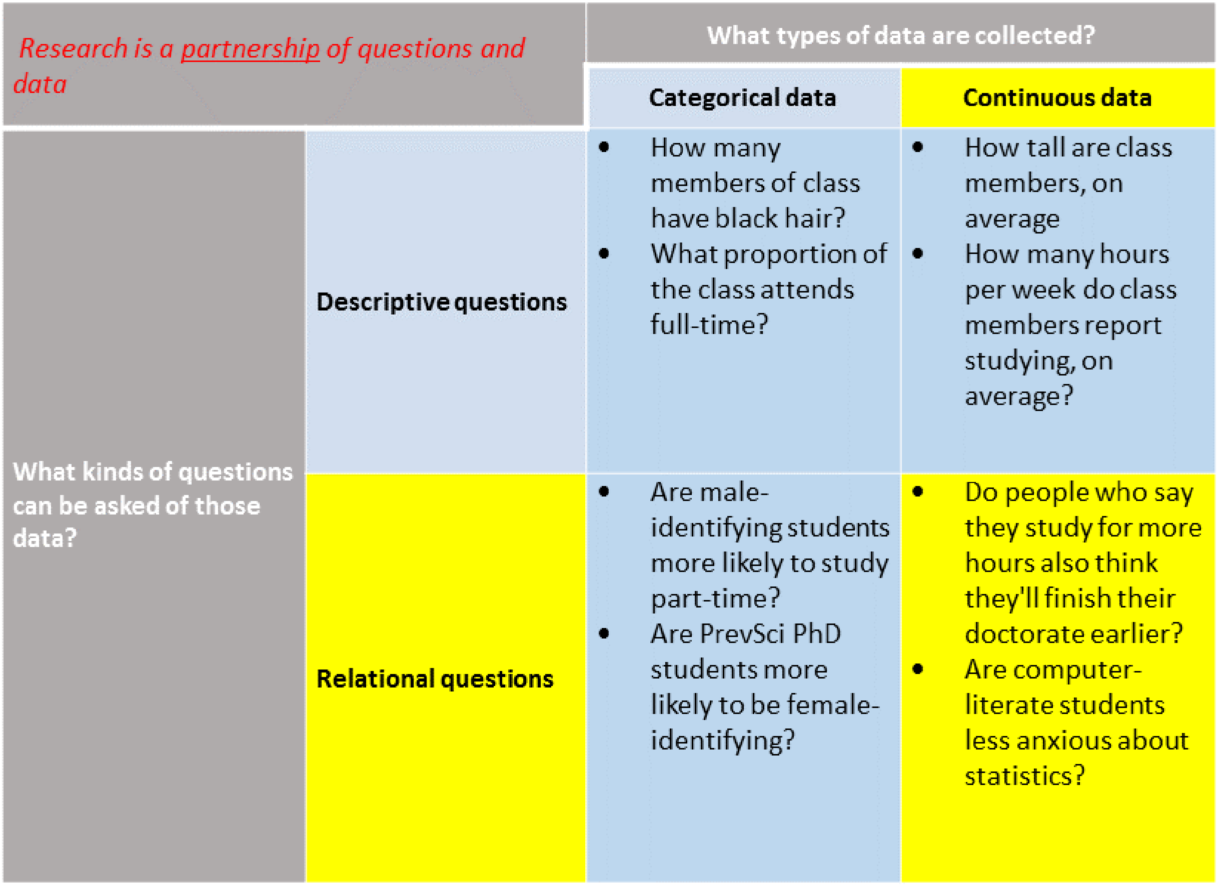
Relationships Between Continuous Variables

EDUC 641: Unit 4 Part 1

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap



Research is a <u>partnership</u> of questions and data		**What types of data are collected?**	
		Categorical data	**Continuous data**
What kinds of questions can be asked of those data?	**Descriptive questions**	- How many members of class have black hair? - What proportion of the class attends full-time?	- How tall are class members, on average - How many hours per week do class members report studying, on average?
	Relational questions	- Are male-identifying students more likely to study part-time? - Are PrevSci PhD students more likely to be female-identifying?	- Do people who say they study for more hours also think they'll finish their doctorate earlier? - Are computer-literate students less anxious about statistics?

Goals of the unit

- Describe relationships between quantitative data that are continuous
- Visualize and substantively describe the relationship between two continuous variables

- Describe and interpret a fitted bivariate regression line
- Describe and interpret components of a fitted bivariate linear regression model
- Visualize and substantively interpret residuals resulting from a bivariate regression model
- Conduct a statistical inference test of the slope and intercept of a bivariate regression model
- Write R scripts to conduct these analyses

Reminder of motivating question

We learned a lot about the distribution of life expectancy in countries, now we are turning to thinking about relationships between life expectancy and other variables. In particular:

Do individuals living in countries with more total years of attendance in school experience, on average, higher life expectancy?

In other words, we are asking whether the variables *SCHOOLING* and *LIFE_EXPECTANCY* are related.

Materials

1. Life expectancy data (in file called life_expectancy.csv)
2. Codebook describing the contents of said data
3. R script to conduct the data analytic tasks of the unit (EDUC641_13_code.R)

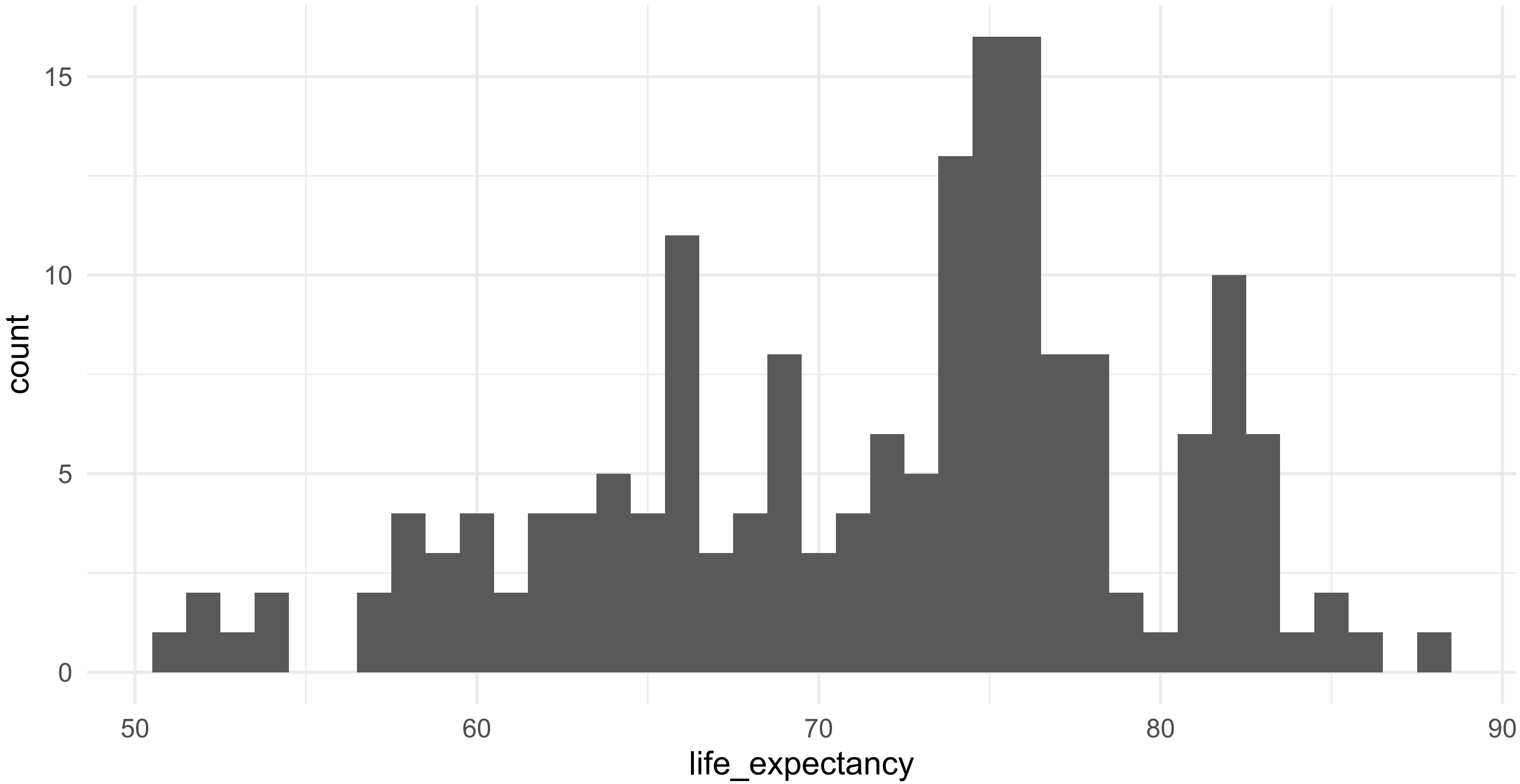
Bivariate relationships between continuous variables[1]

[1] We can also look at relationships between continuous and categorical variables with increasingly sophisticated--but functionally equivalent--methods, including two-sample t-tests, ANOVA, ANCOVA, regression, and more. We will examine these topics in EDUC 643.

Life expectancy distribution

```
#>
#>   The decimal point is at the |
#>
#>   50 | 0
#>   52 | 000
#>   54 | 00
#>   56 | 00
#>   58 | 0000000
#>   60 | 000000
#>   62 | 00000000
#>   64 | 000000000
#>   66 | 00000000000000
#>   68 | 000000000000
#>   70 | 0000000
#>   72 | 00000000000
#>   74 | 00000000000000000000000000000
#>   76 | 000000000000000000000000
#>   78 | 0000000000
#>   80 | 0000000
#>   82 | 0000000000000000
#>   84 | 000
#>   86 | 0
#>   88 | 0
```

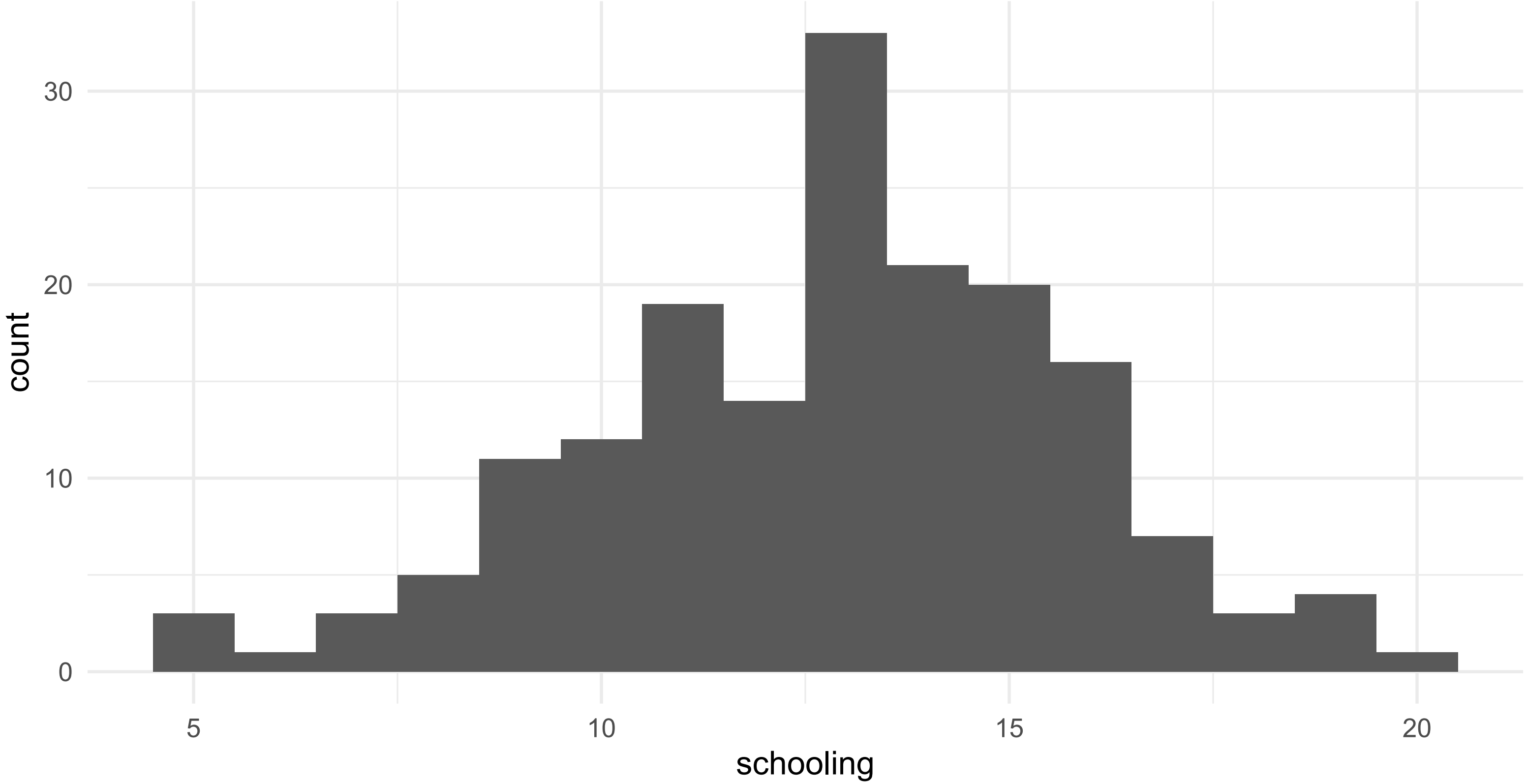
Another way



What about schooling?

```
#> 
#>   The decimal point is at the |
#> 
#>    4 | 9
#>    5 | 04
#>    6 | 3
#>    7 | 1237
#>    8 | 144589
#>    9 | 00111225569
#>   10 | 00011233346777888889
#>   11 | 111223444677779
#>   12 | 0112355566667777788999
#>   13 | 000111122333334445566789999
#>   14 | 0012223334455667889
#>   15 | 0000122333334566899
#>   16 | 0001333345566
#>   17 | 0123377
#>   18 | 16
#>   19 | 022
#>   20 | 4
```

And differently again



Numerical univariate statistics

```
summary(who$life_expectancy)
```

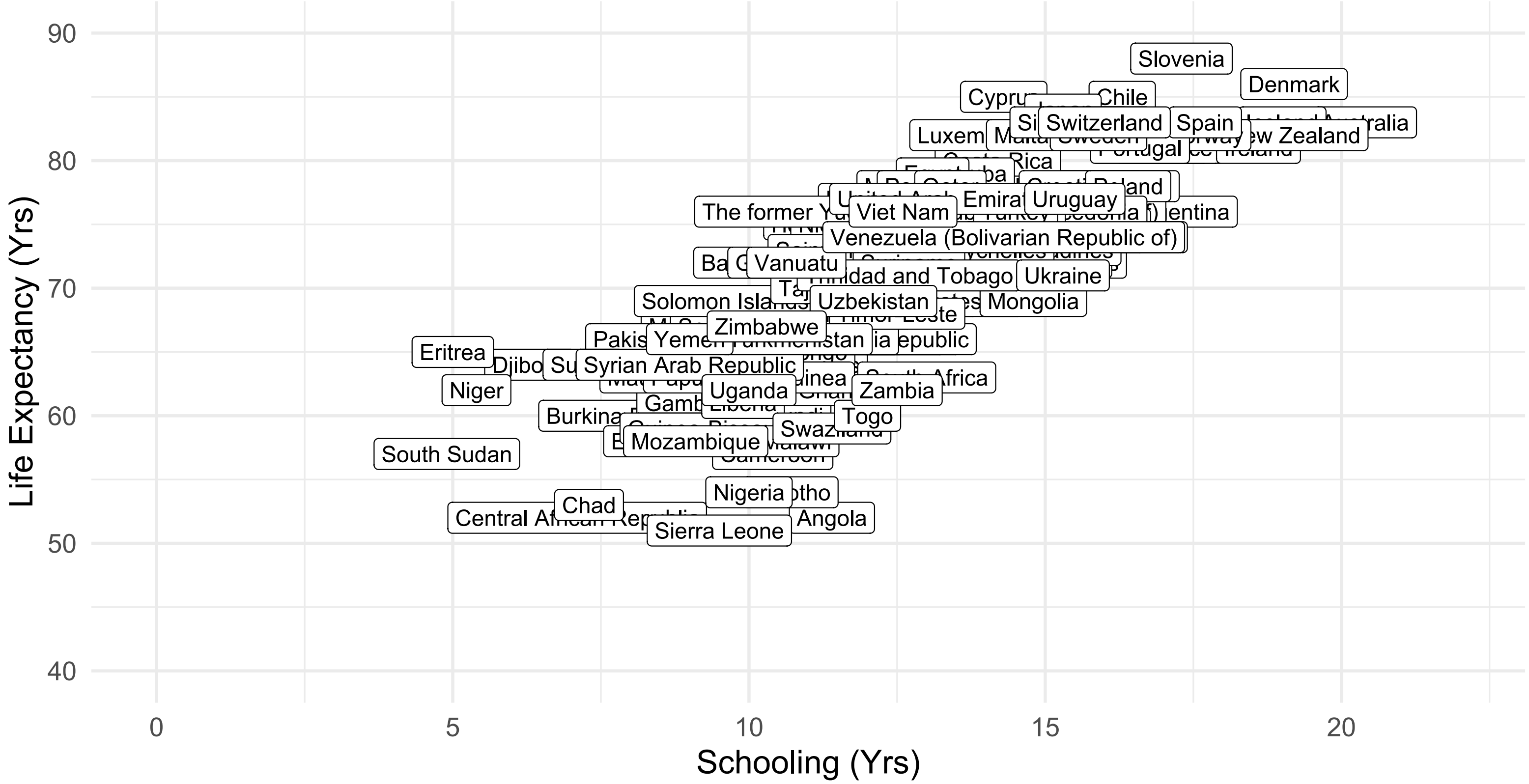
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   51.00   66.00   74.00   71.74   77.00   88.00
```

```
summary(who$schooling)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    4.90   10.80   13.10   12.93   15.00   20.40
```

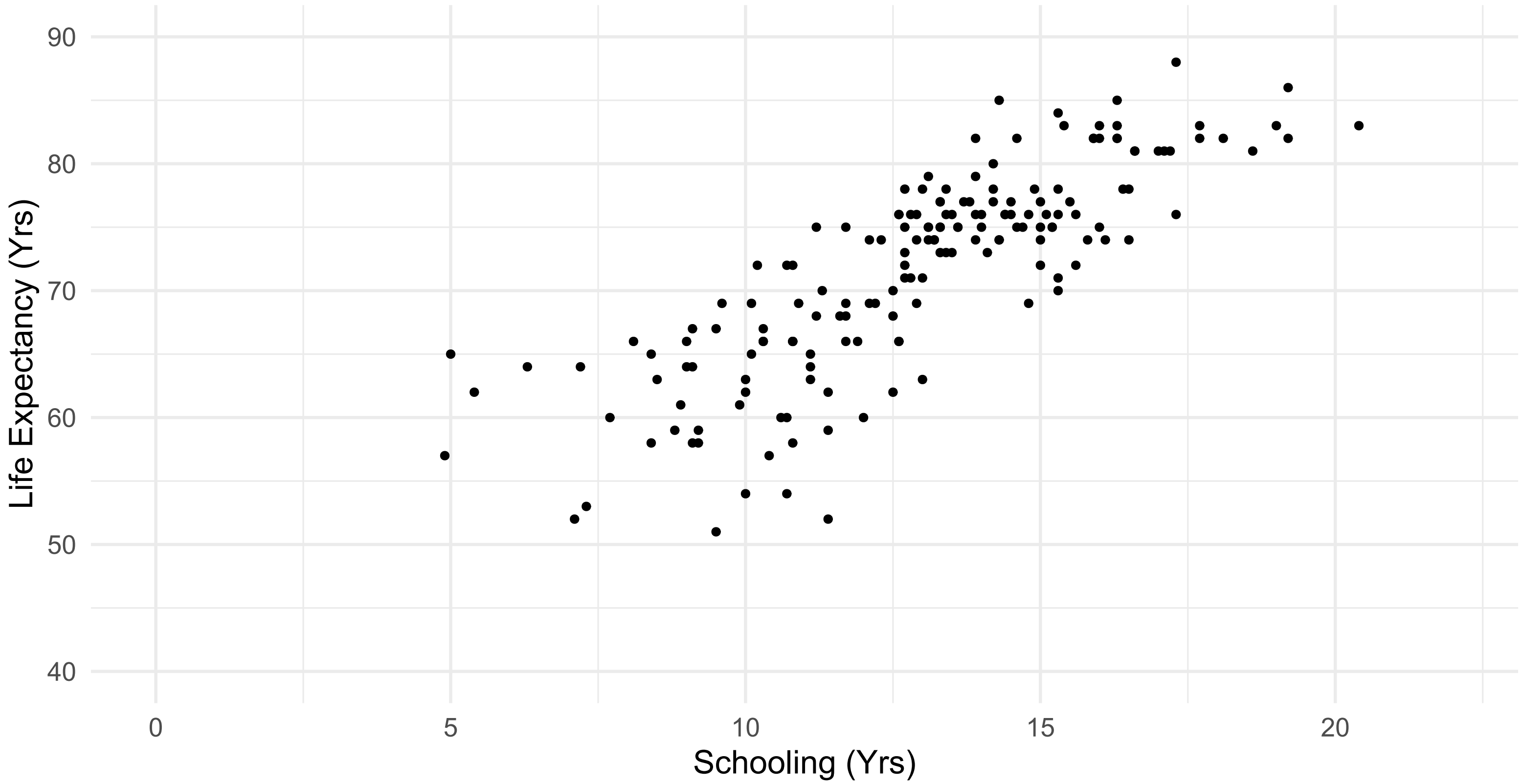
Can you interpret the univariate statistics and displays on this and the previous slides? Describe to folks at your table information about the measures of central tendency and the distributional shape of these two variables.

Visualizing the relationship



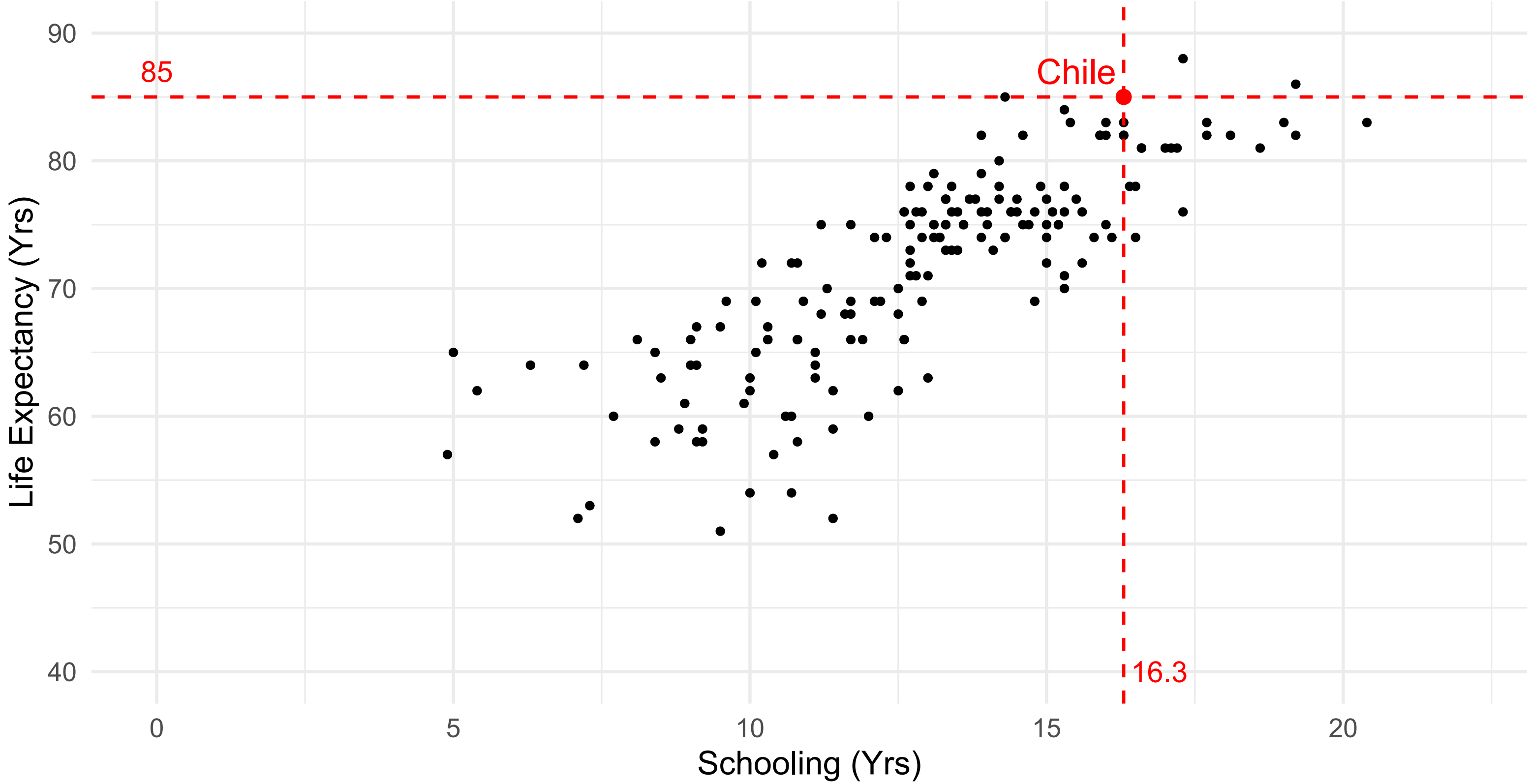
Probably easier to see if we have some symbolic way of representing our data…

Visualizing the relationship



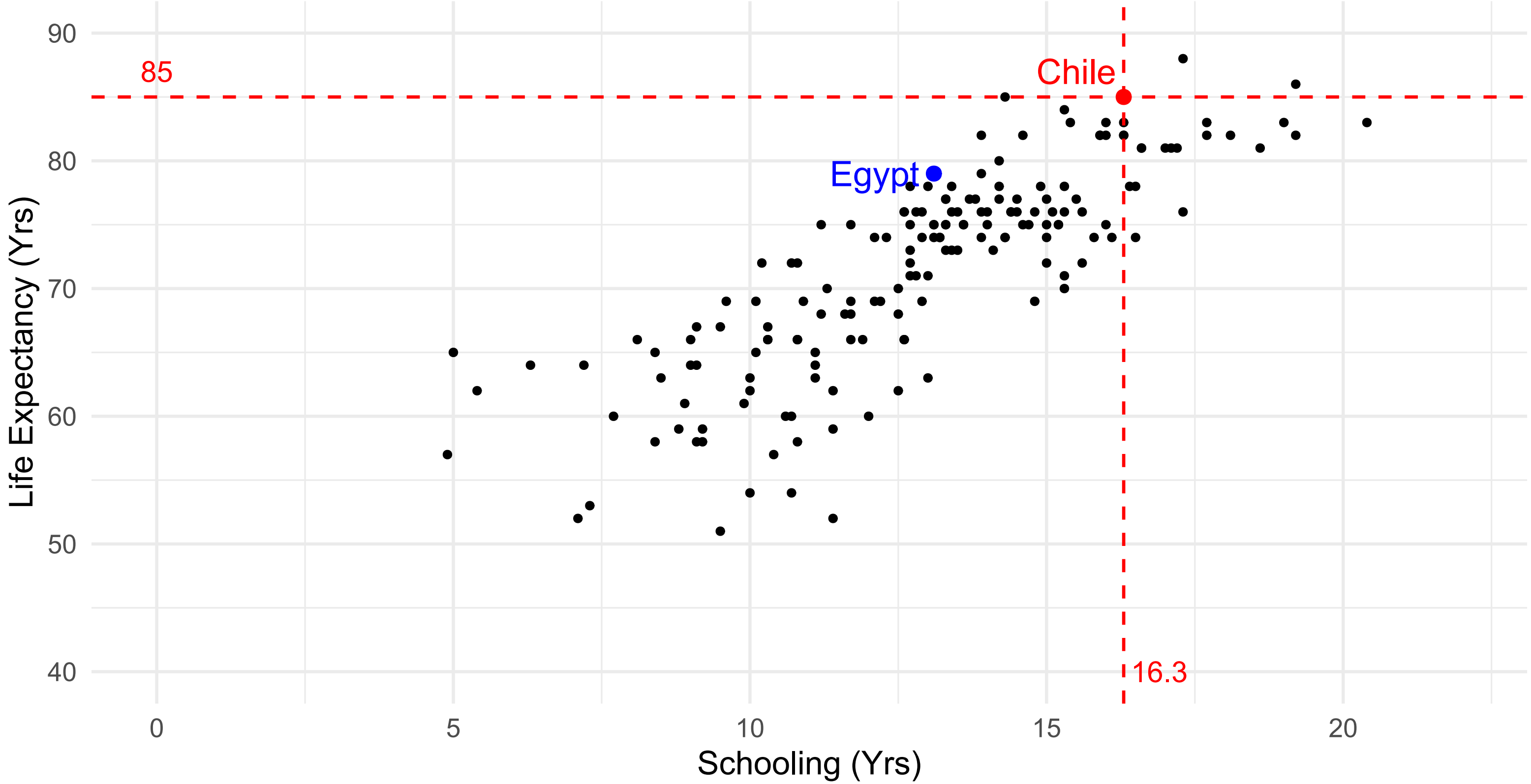
Horizontal axis (or *x*-axis) labels the value of the "predictor" *SCHOOLING*. Vertical axis (or *y*-axis) labels the value of the "outcome" *LIFE_EXPECTANCY*. *Can you interpret the bivariate display? What does it (and does it NOT) say about the relationship between schooling and life expectancy?*

Visualizing the relationship



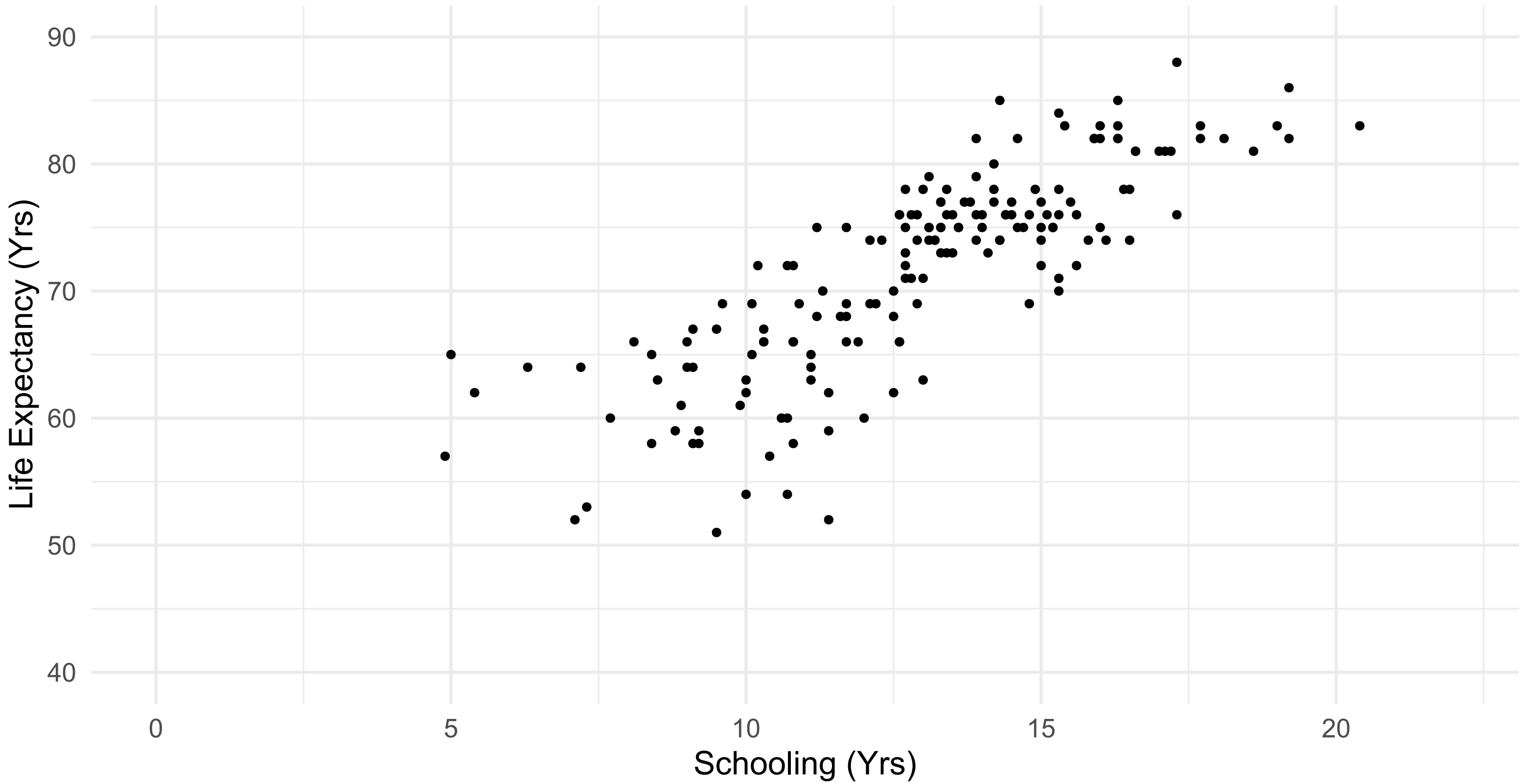
Can you interpret what this display says about the country of Chile?

You try...



Can you interpret what this display says about the country of Egypt?

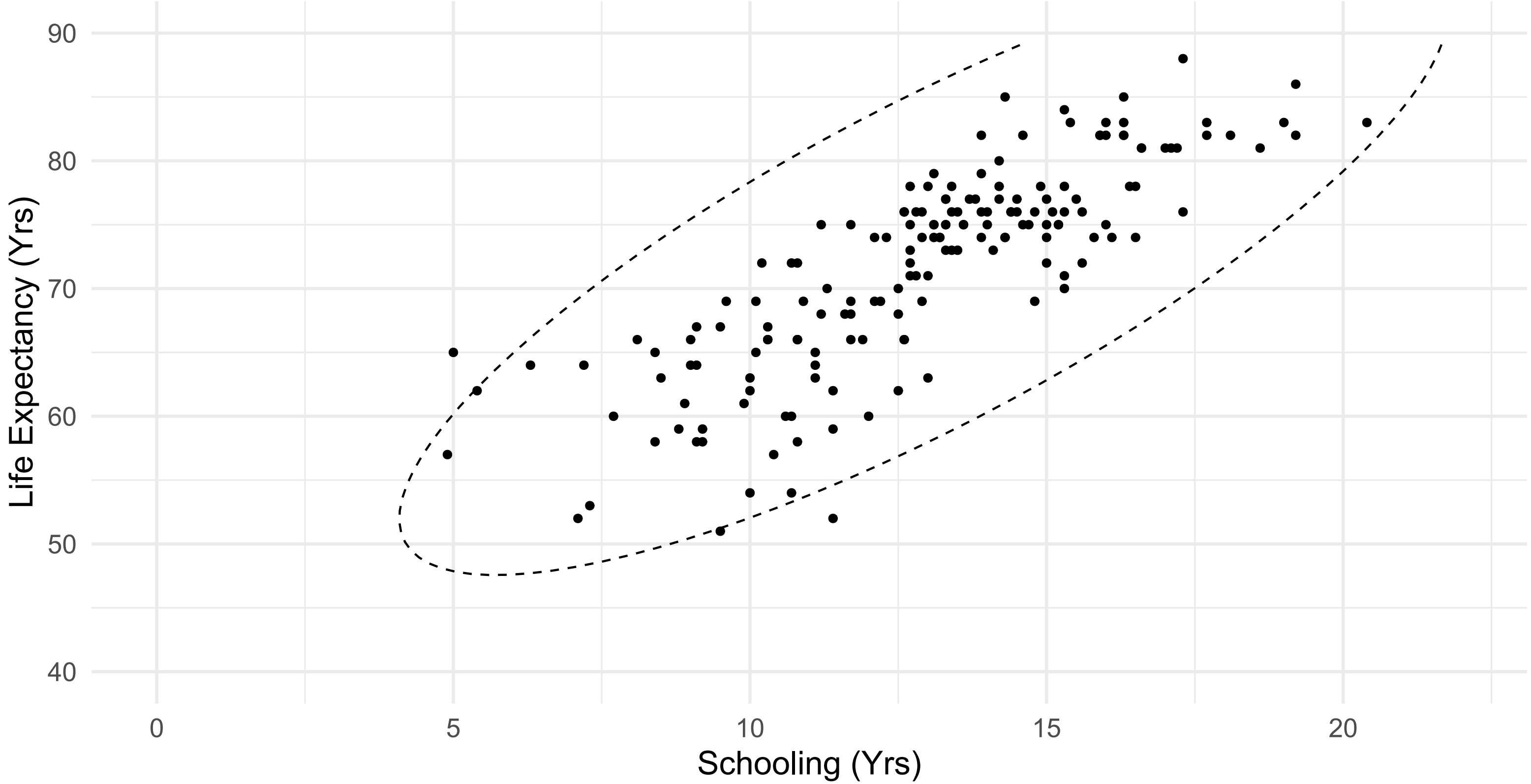
What about the relationship?



Is there a relationship between SCHOOLING and LIFE_EXPECTANCY? How do you know?

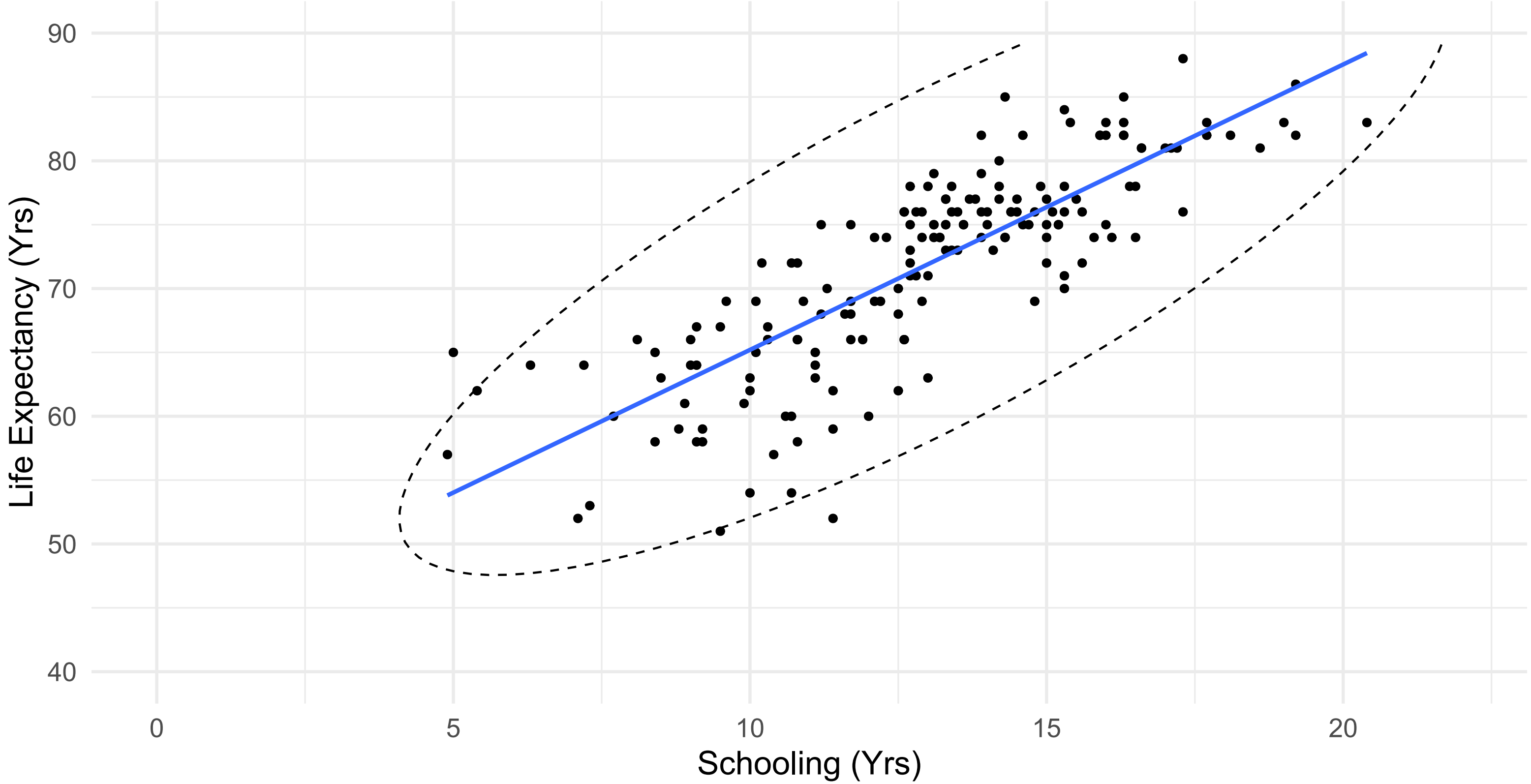
What kind of line, curve or other construction best summarizes the observed relationship between SCHOOLING and LIFE_EXPECTANCY?

What about the relationship?



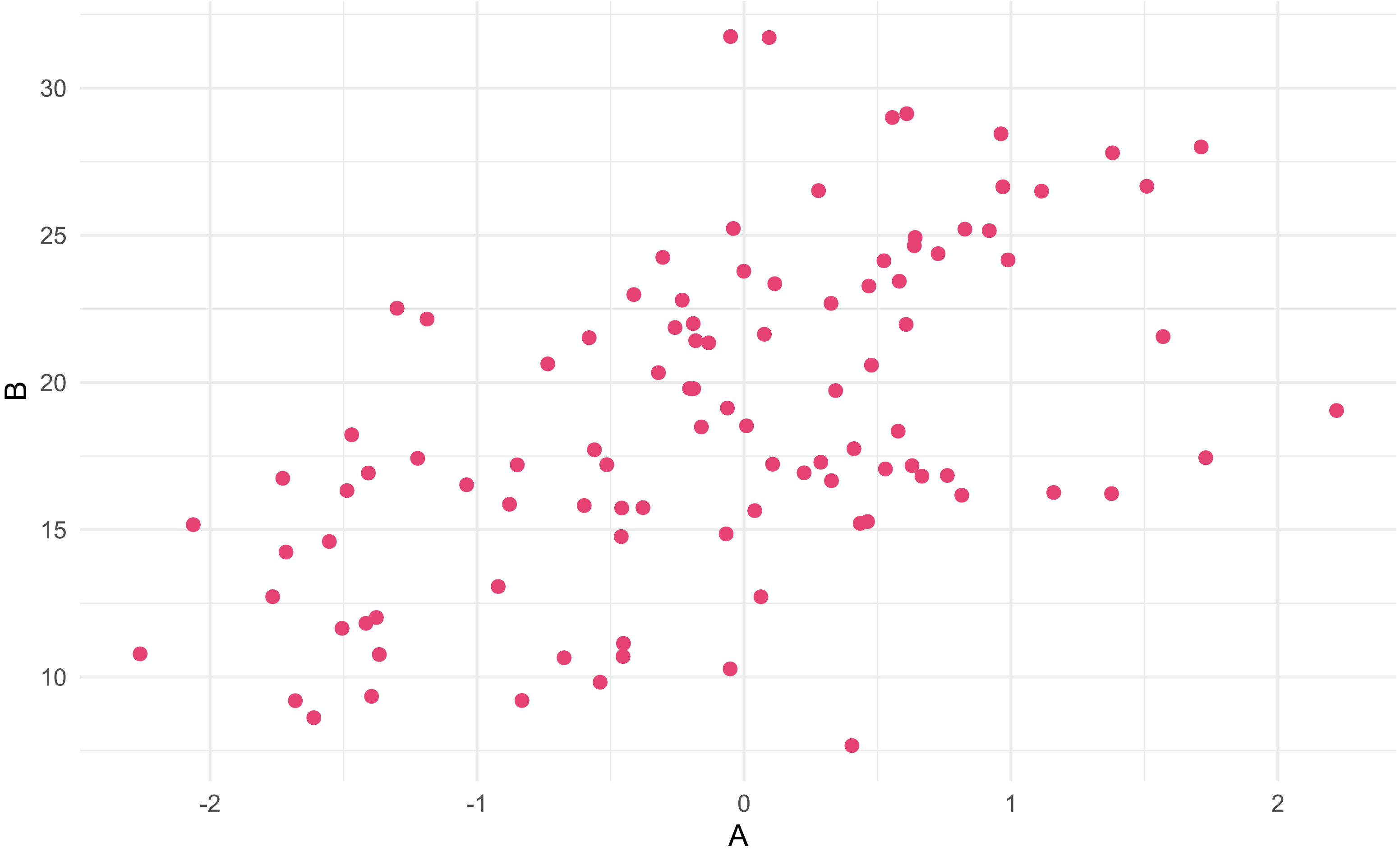
What kind of line, curve or other construction best summarizes the observed relationship between SCHOOLING and LIFE_EXPECTANCY?

What about the relationship?

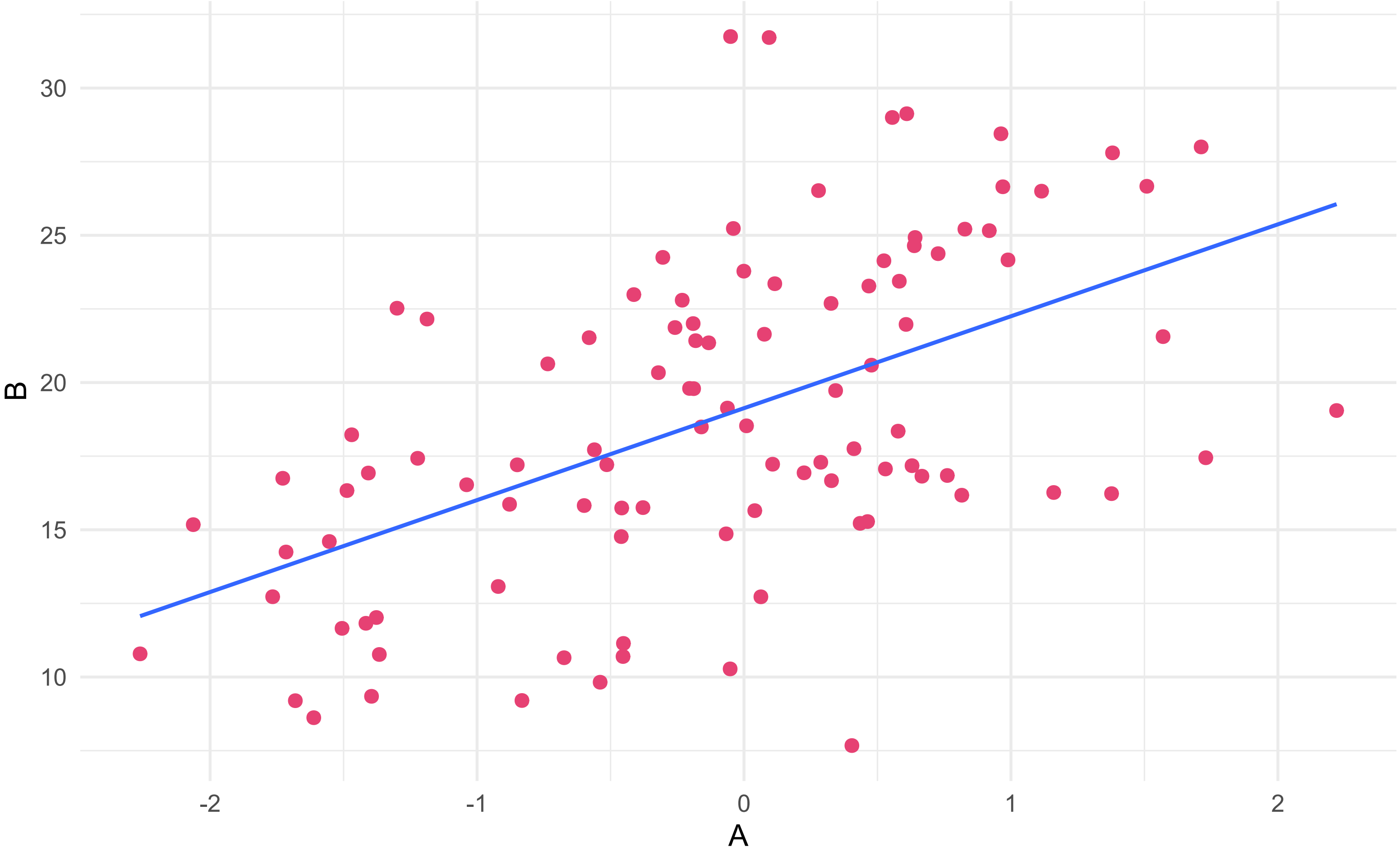


What kind of line, curve or other construction best summarizes the observed relationship between SCHOOLING and LIFE_EXPECTANCY?

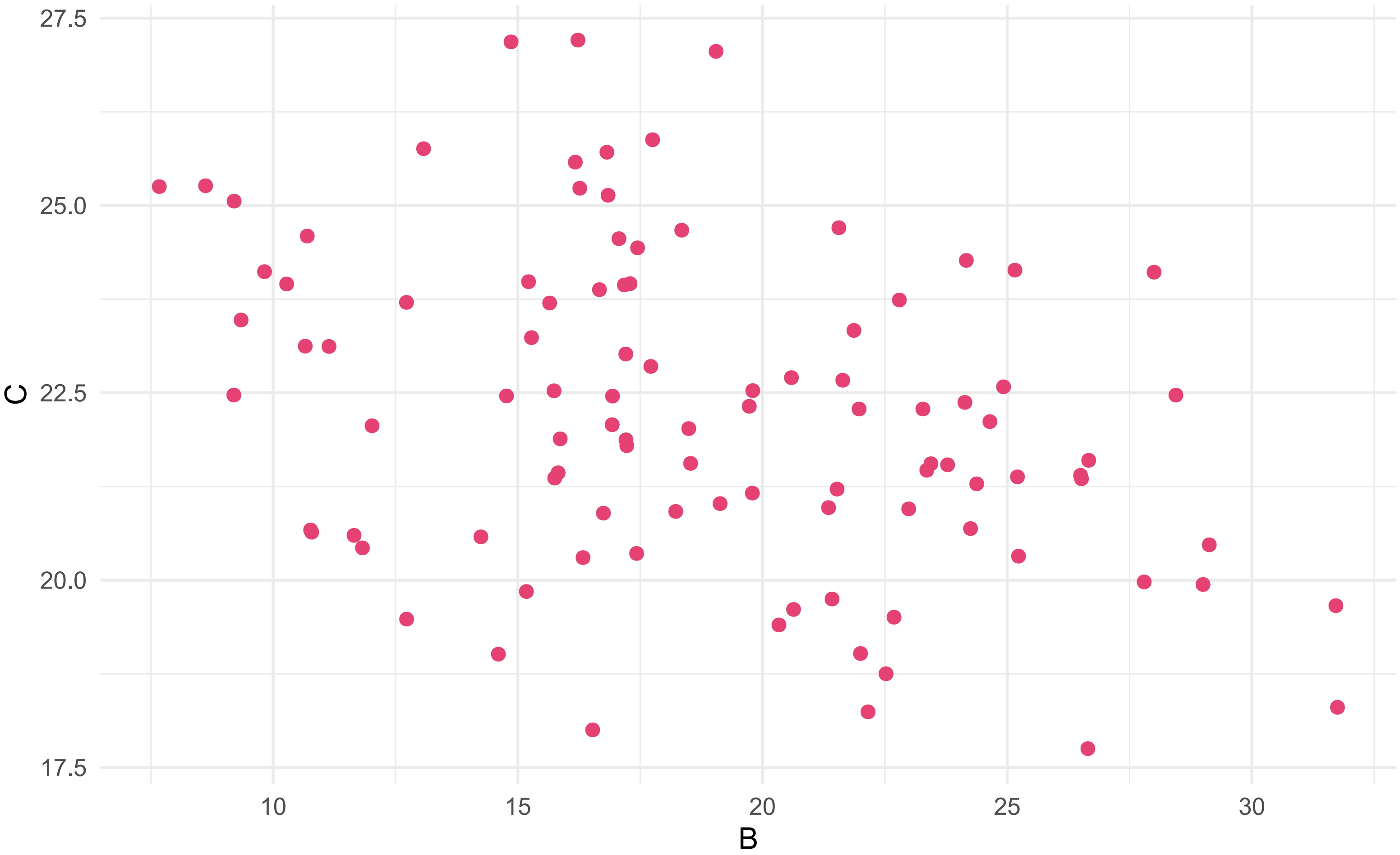
Pin the tail on the point cloud



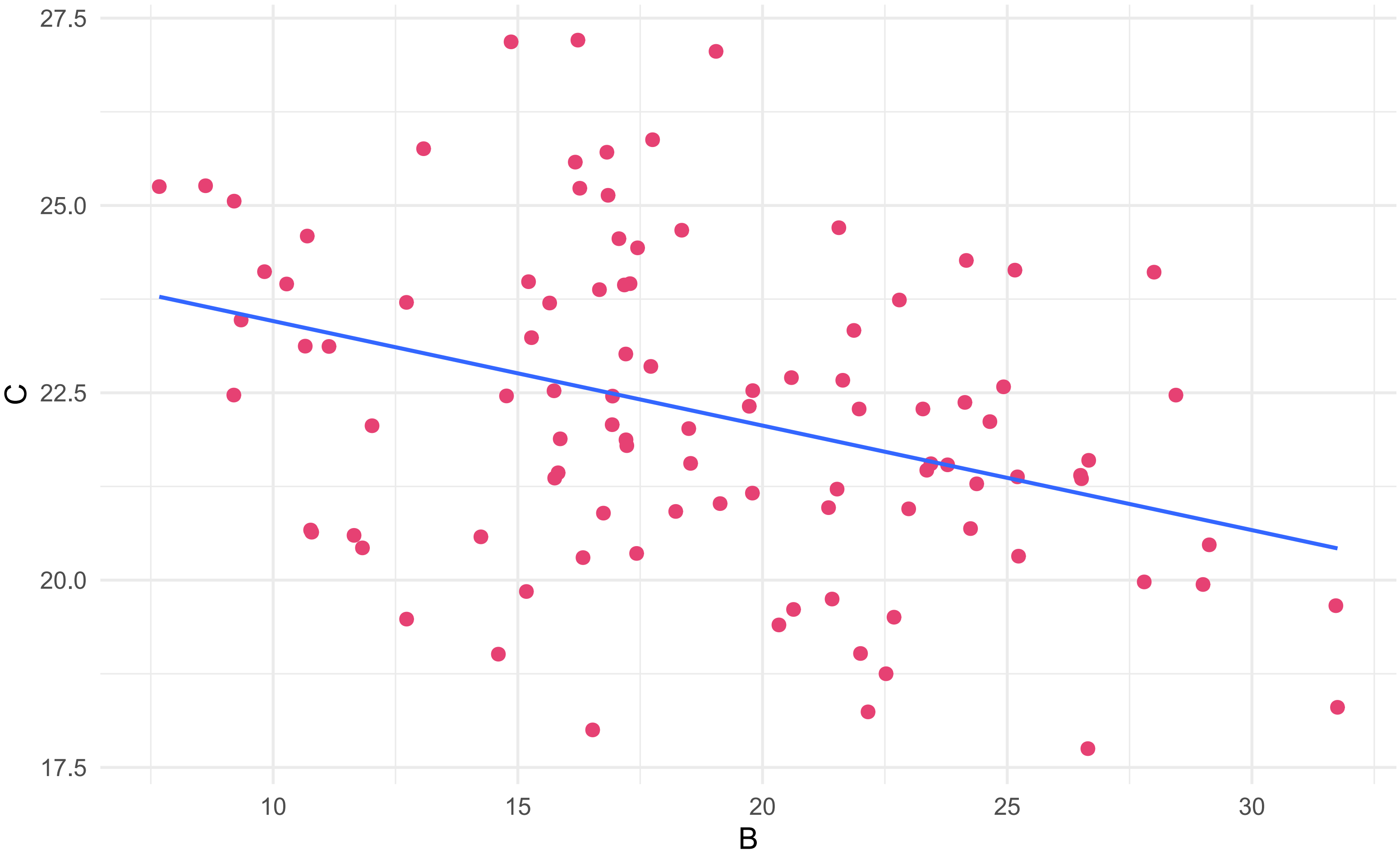
Pin the tail on the point cloud



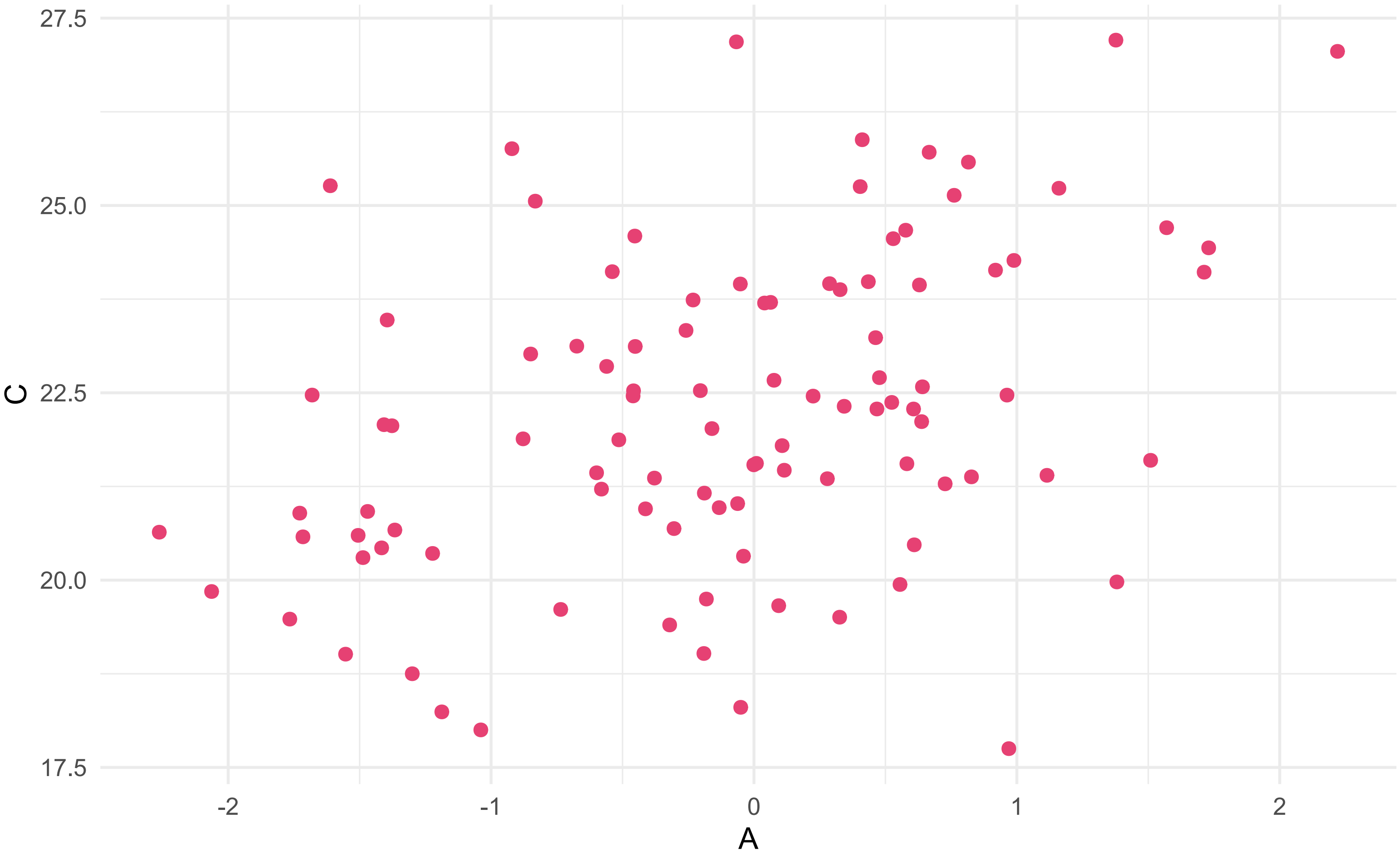
Pin the tail on the point cloud



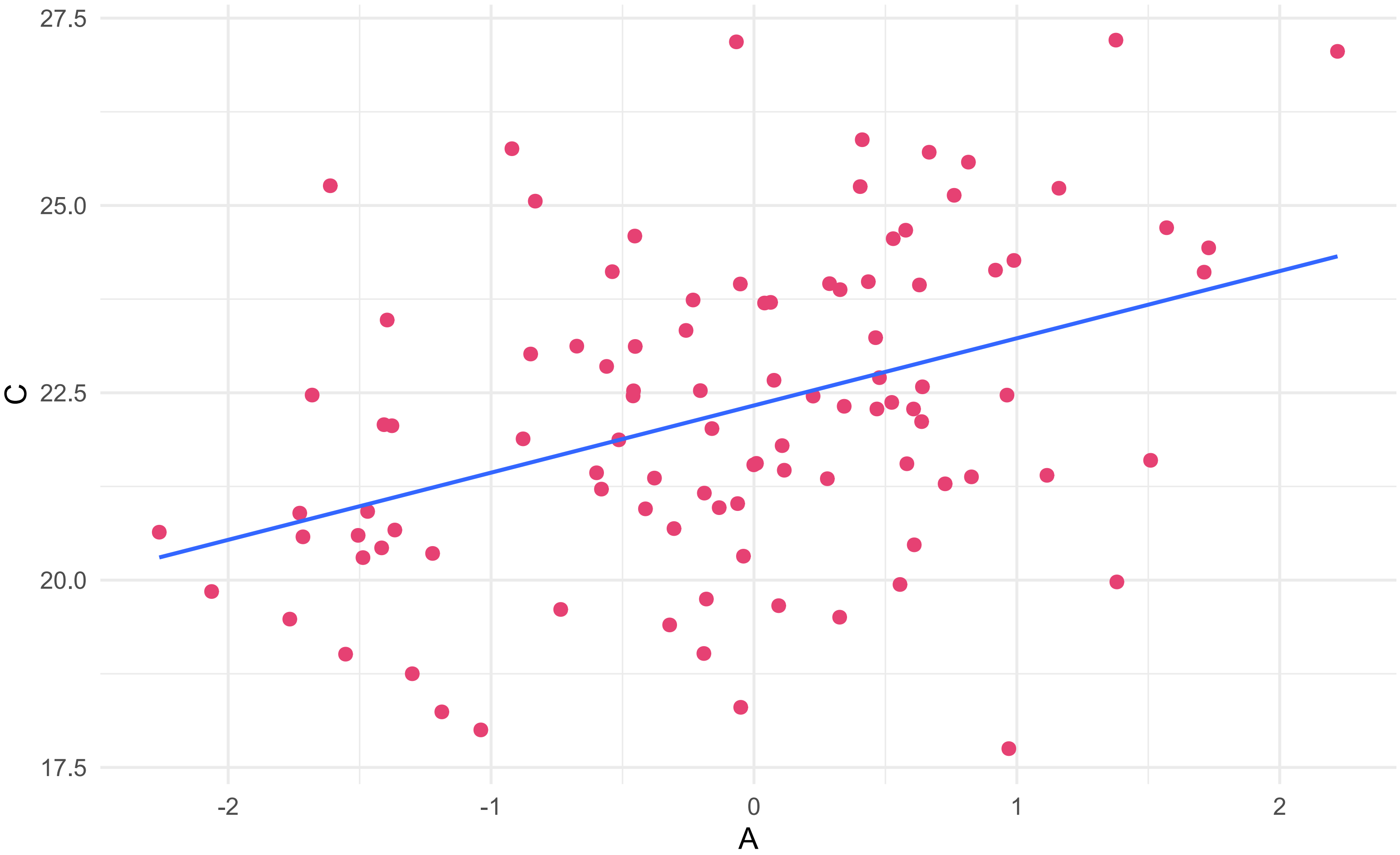
Pin the tail on the point cloud



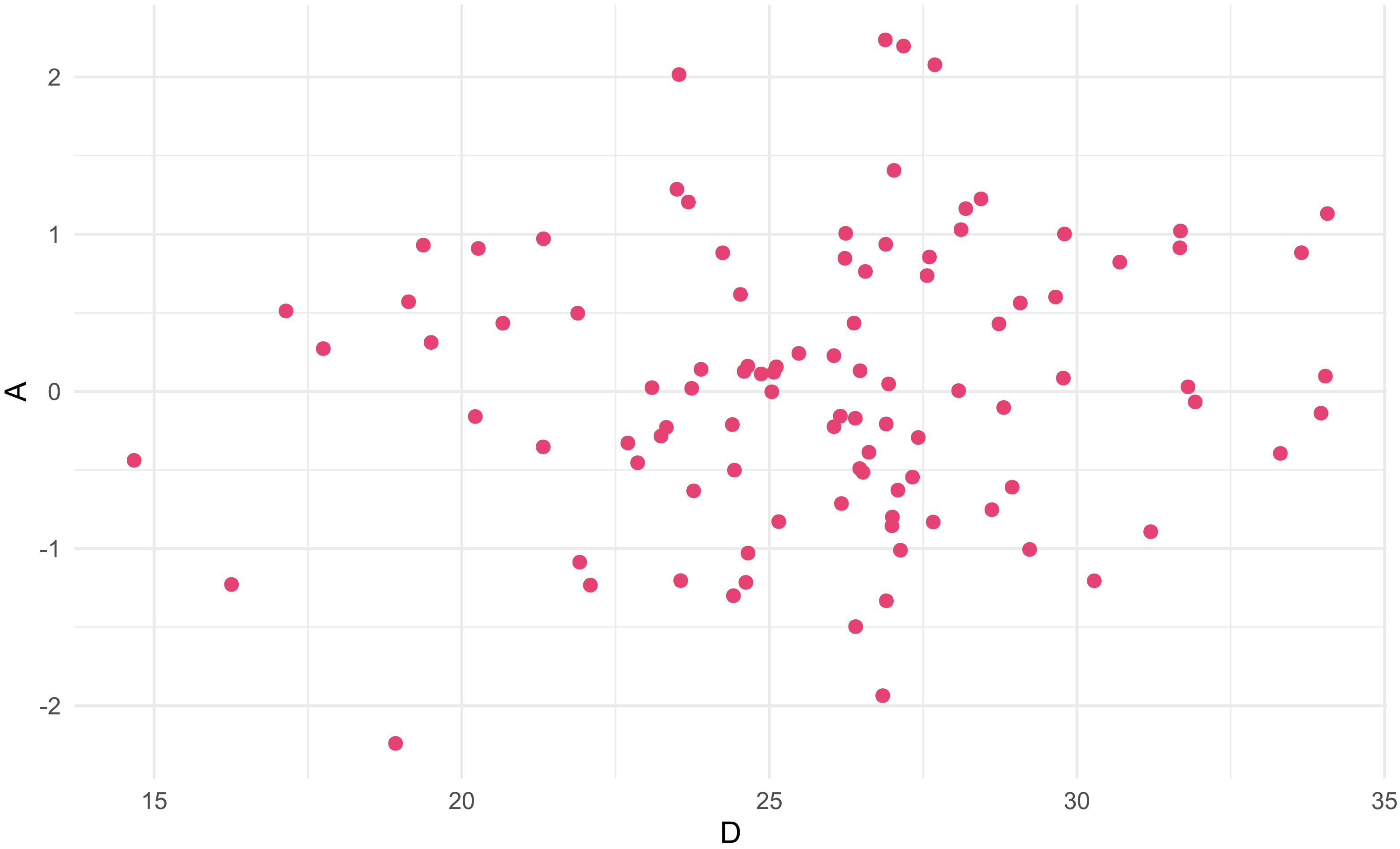
Pin the tail on the point cloud



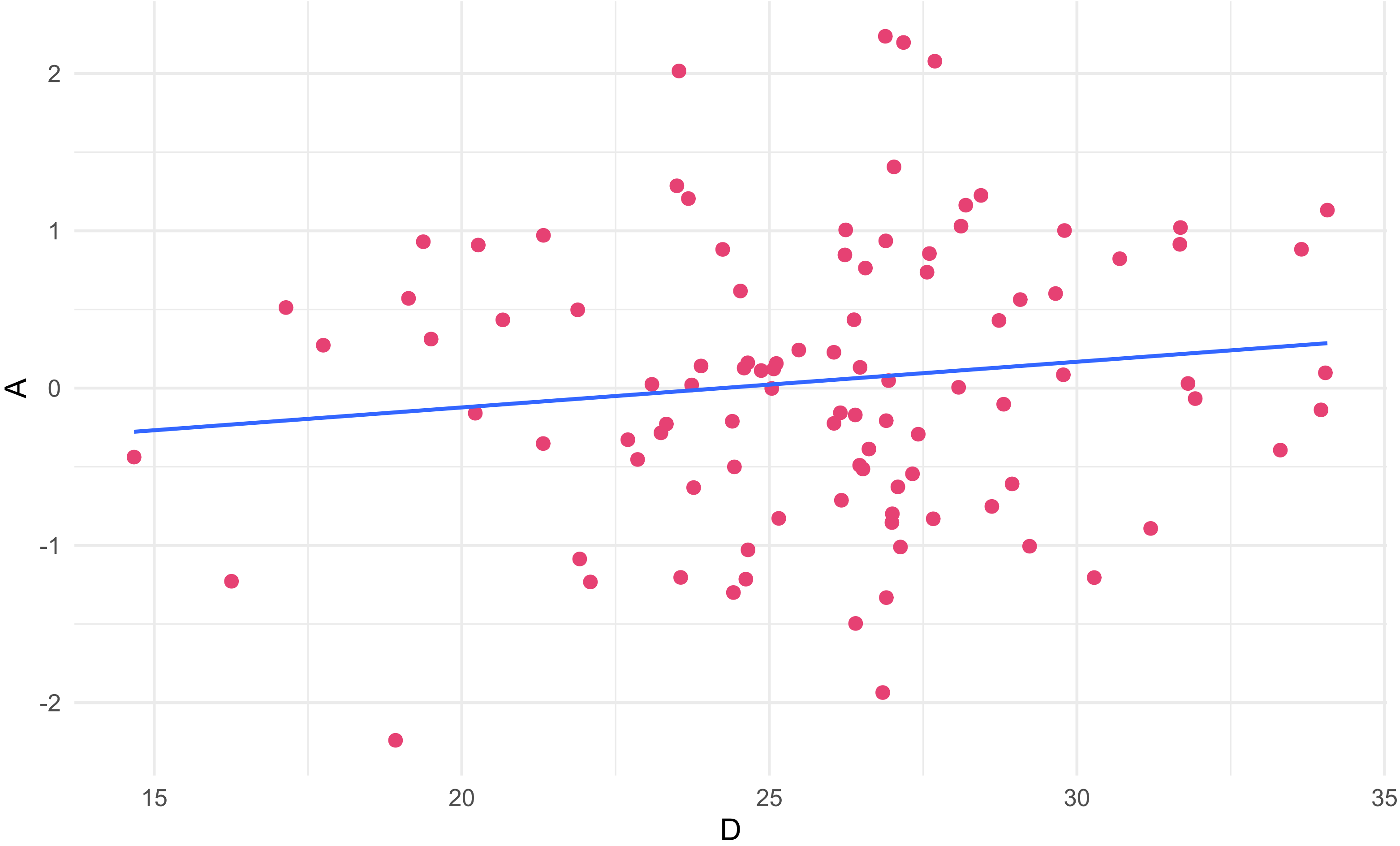
Pin the tail on the point cloud



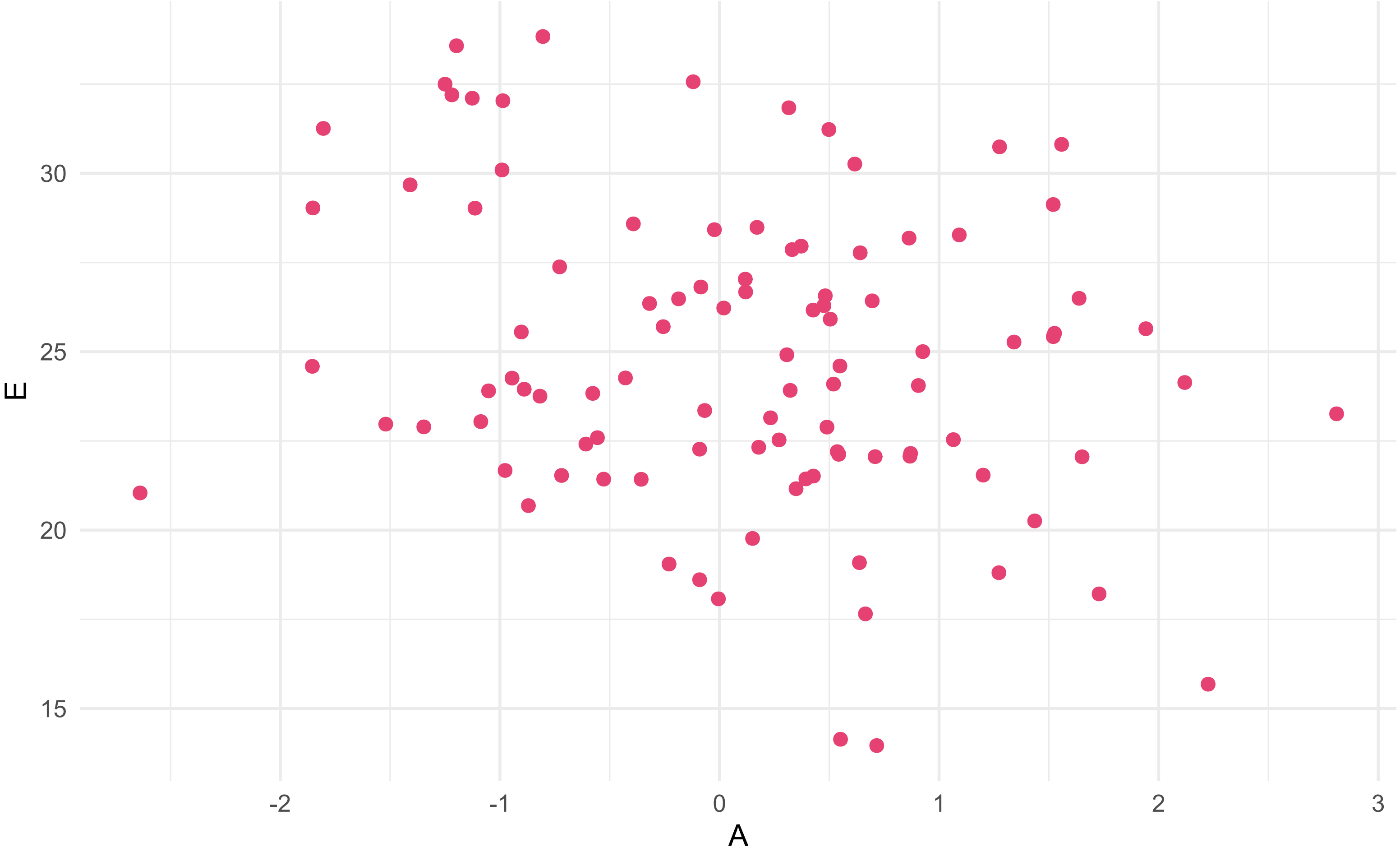
Pin the tail on the point cloud



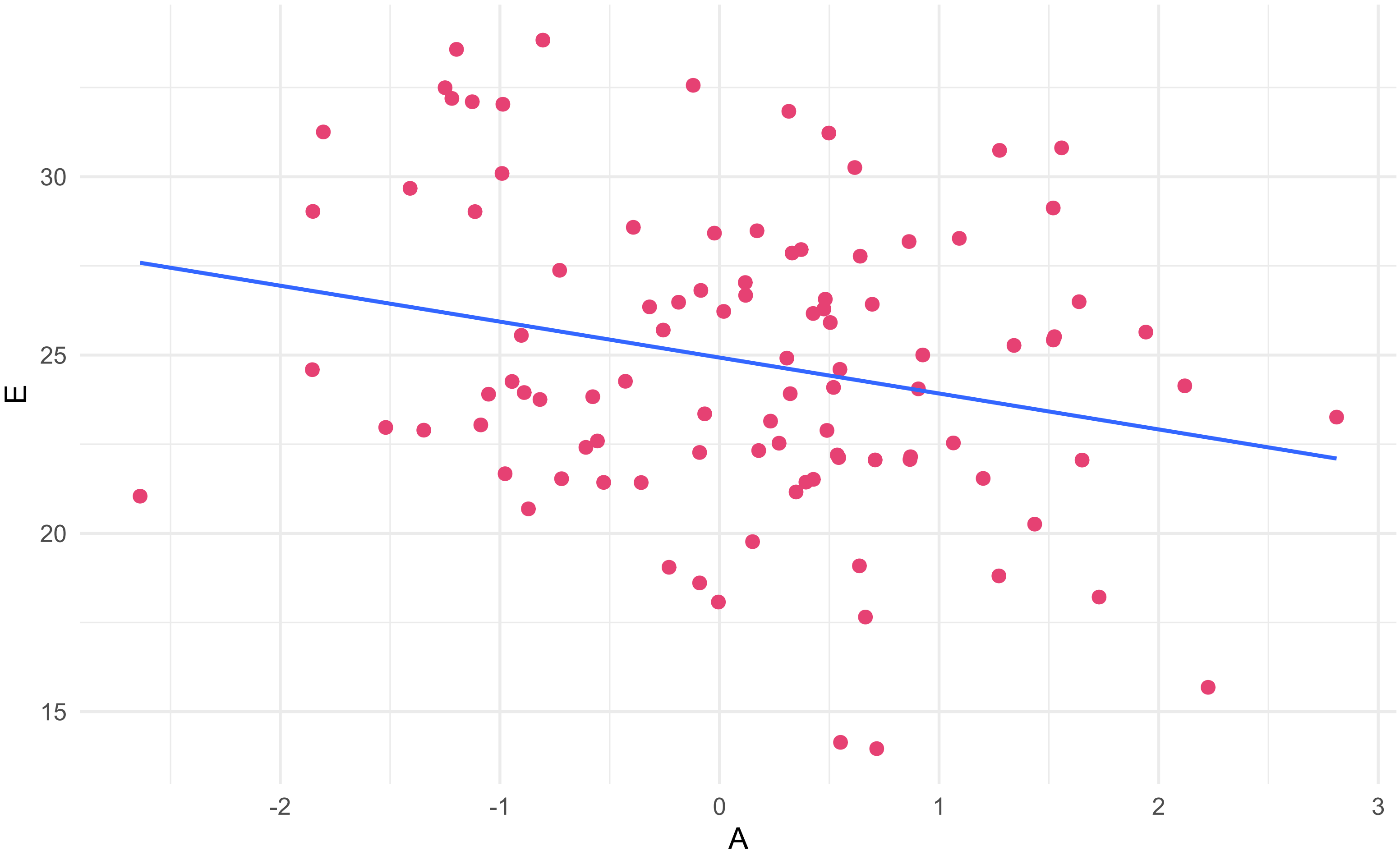
Pin the tail on the point cloud



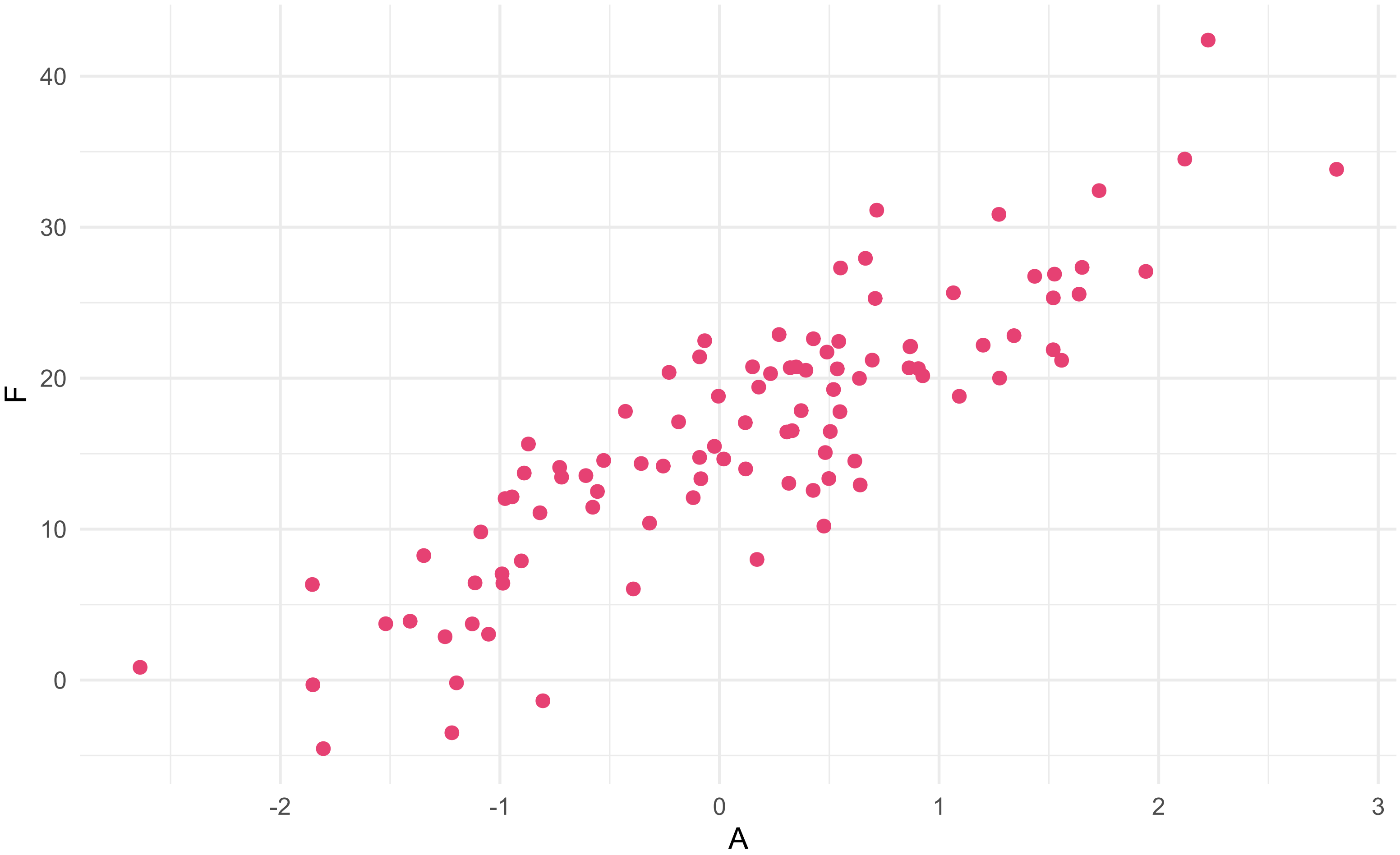
Pin the tail on the point cloud



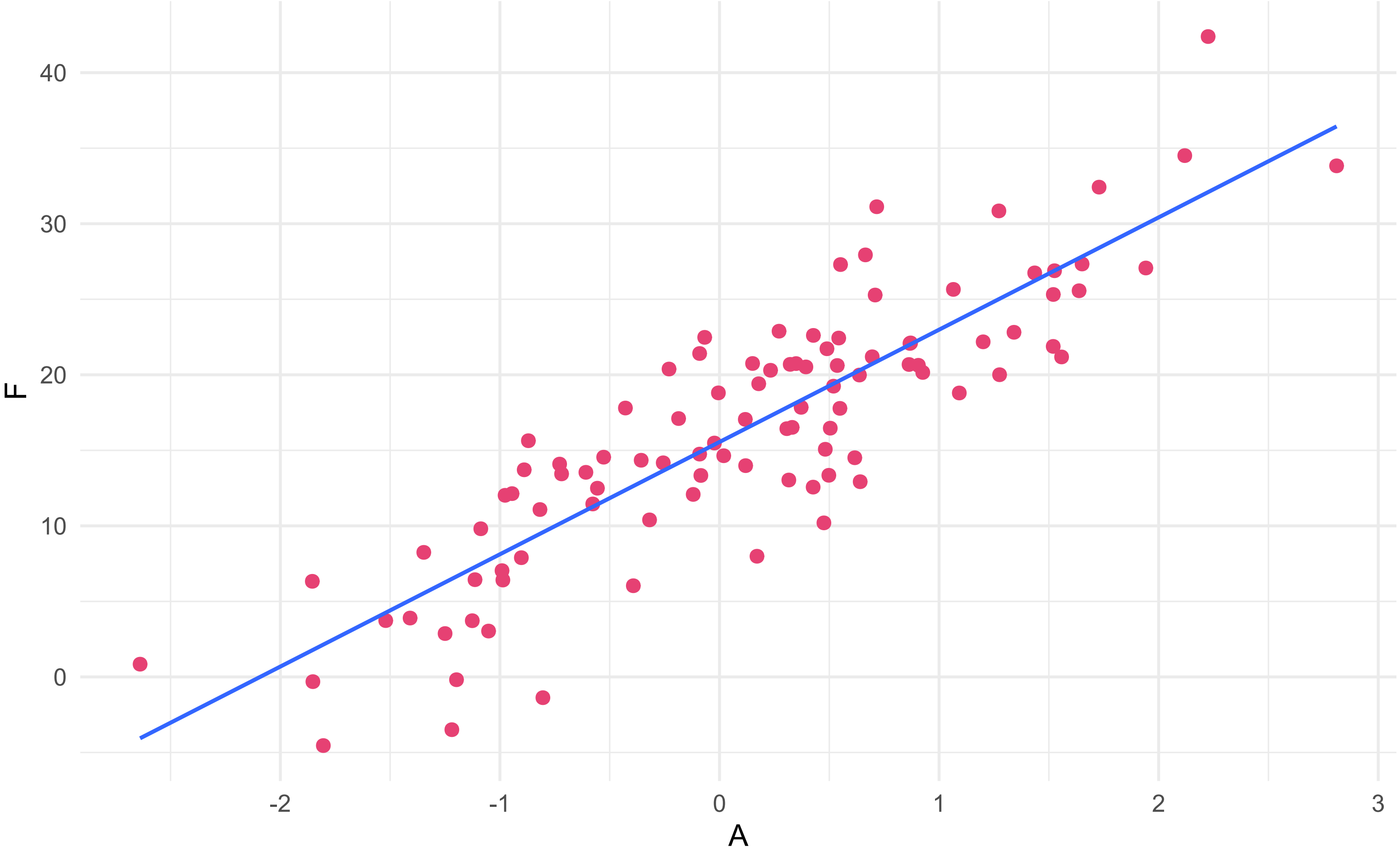
Pin the tail on the point cloud



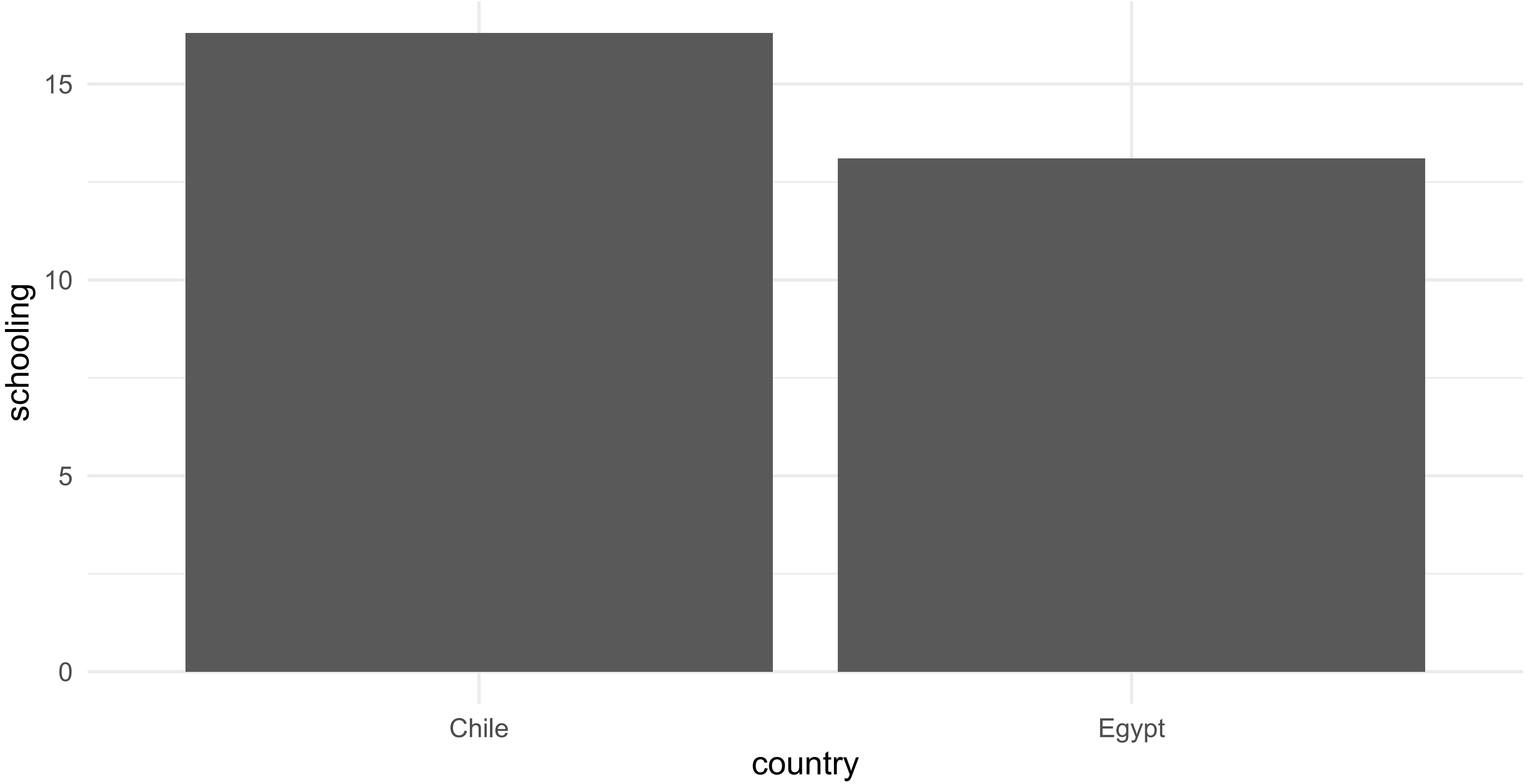
Pin the tail on the point cloud



Pin the tail on the point cloud

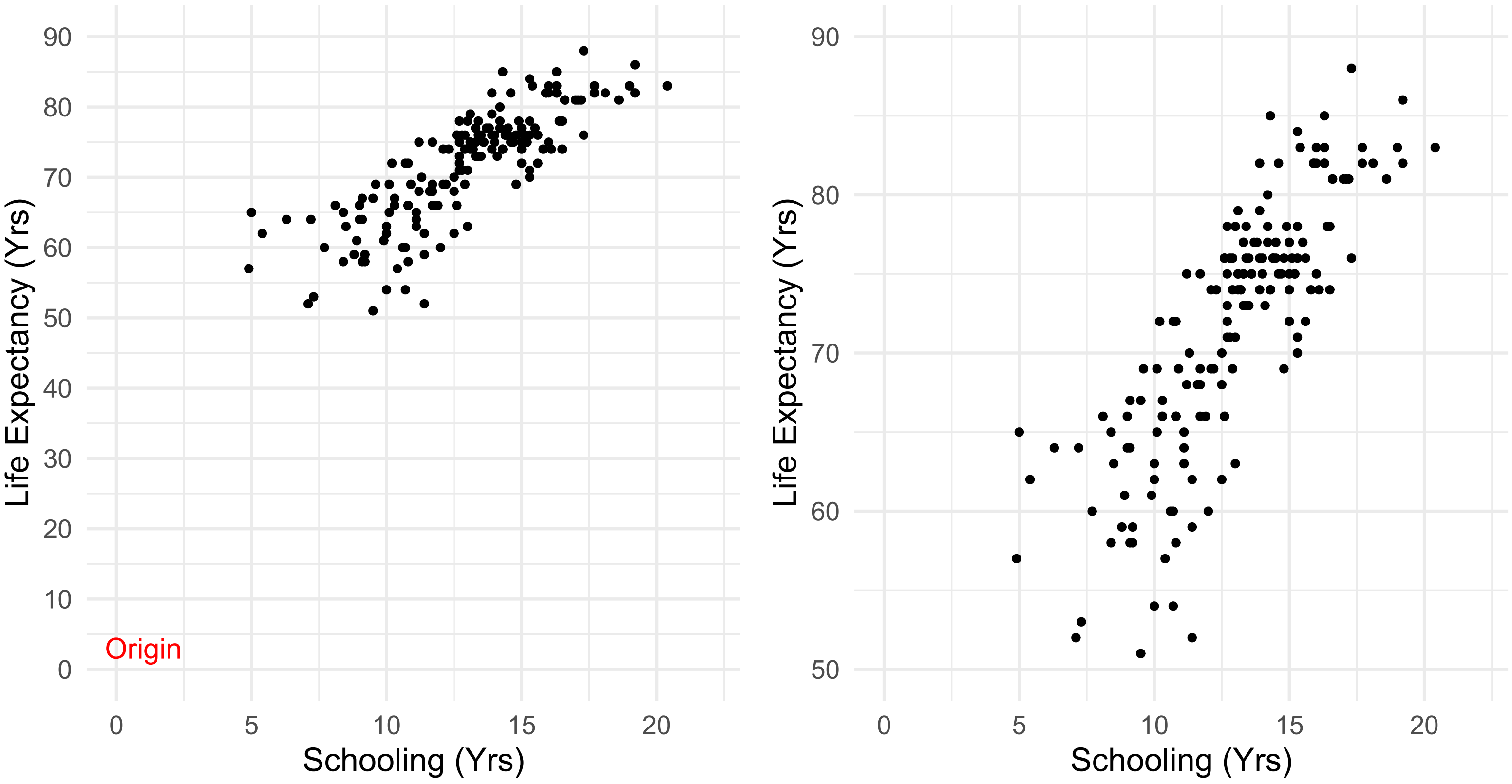


An aside about the origin



Figures that compare measures of central tendency across groups (e.g., bar charts) should generally start at zero (0) so as not to artificially inflate the differences between groups

An aside about the origin



Figures that describe relationships between two variables (e.g., scatter plots) might (or might not) include the origin (0, 0). The key concept these charts illustrate is the relationship. By adjusting the scale and range of each axis, we can make the relationship "look" different. But the strength and magnitude are the same. More to come in EDUC 643...

Synthesis and wrap-up

Goals of the unit

- Describe relationships between quantitative data that are continuous
- Visualize and substantively describe the relationship between two continuous variables

- Describe and interpret a fitted bivariate regression line
- Describe and interpret components of a fitted bivariate linear regression model
- Visualize and substantively interpret residuals resulting from a bivariate regression model
- Conduct a statistical inference test of the slope and intercept of a bivariate regression model
- Write R scripts to conduct these analyses

To Dos

Reading

- LSWR Chapter 10: Law of large numbers and CLT

Quiz

- QuiZ #4: Opens 3:45pm on Nov. 14, closes at 5pm Nov. 15

Assignment

- Assignment #3 Due Nov. 9, 11:59pm
- Assignment #4 Due Nov. 27, 11:59pm

Remember no class on Thursday (Nov. 9)!!