Examining the relationship between continuous variables

EDUC 641: Week XX

TBD

UNIVERSITY OF OREGON

Roadmap

Research is a partnership 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
- Articulate modern critiques of null-hypothesis significance testing framework
- Describe strategies to improve replicability and generalizability of quantitative research

Reminder of our motivating question

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

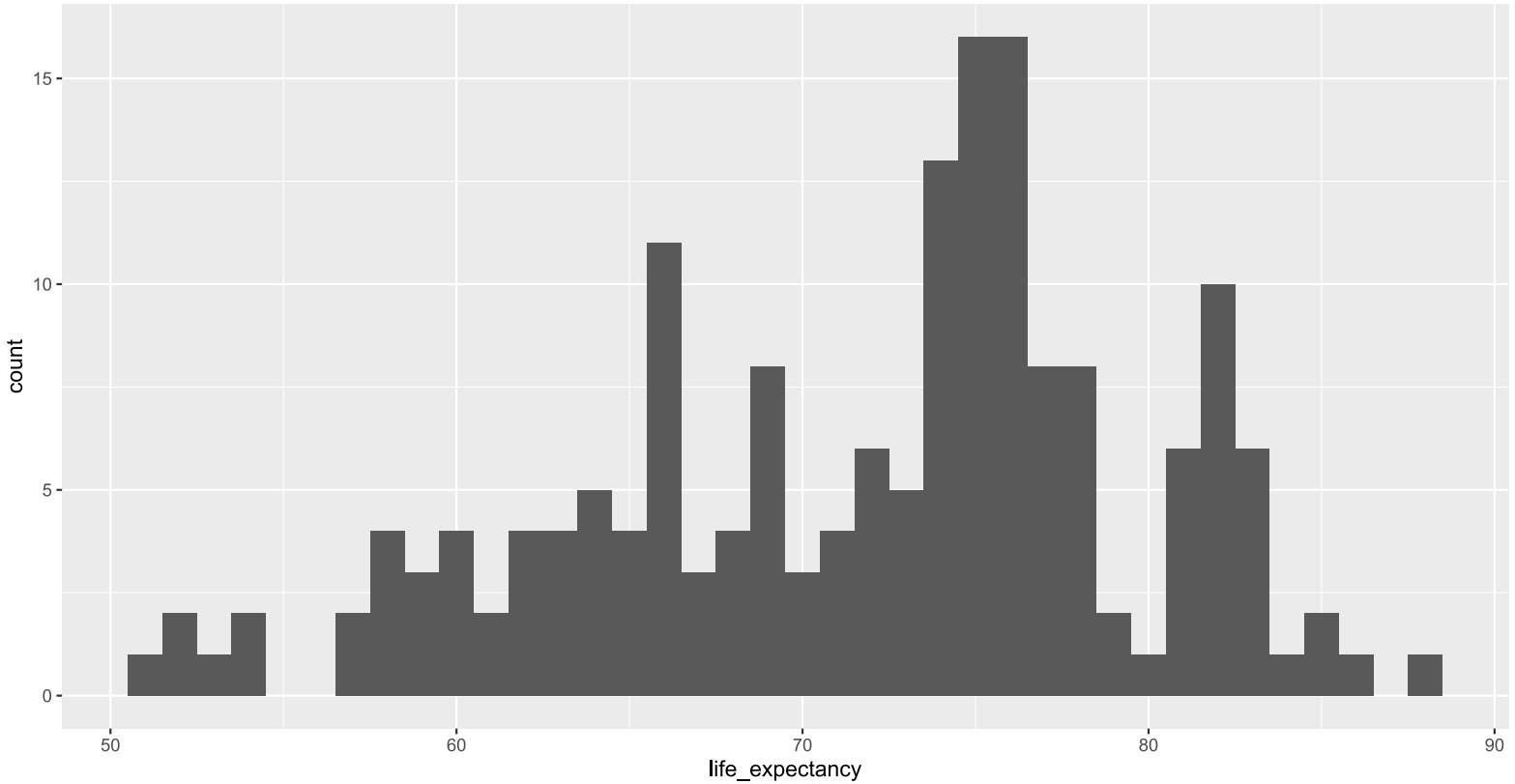
1. Death penalty data (in file called life_expectancy.csv)
2. Codebook describing the contents of said data
3. R script to conduct the data analytic tasks of the unit

Bivariate relationships between continuous variables

Recall life expectancy distr.

```
#>
#>   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 | 0000000000000000000000
#>   78 | 0000000000
#>   80 | 0000000
#>   82 | 00000000000000000
#>   84 | 000
#>   86 | 0
#>   88 | 0
```

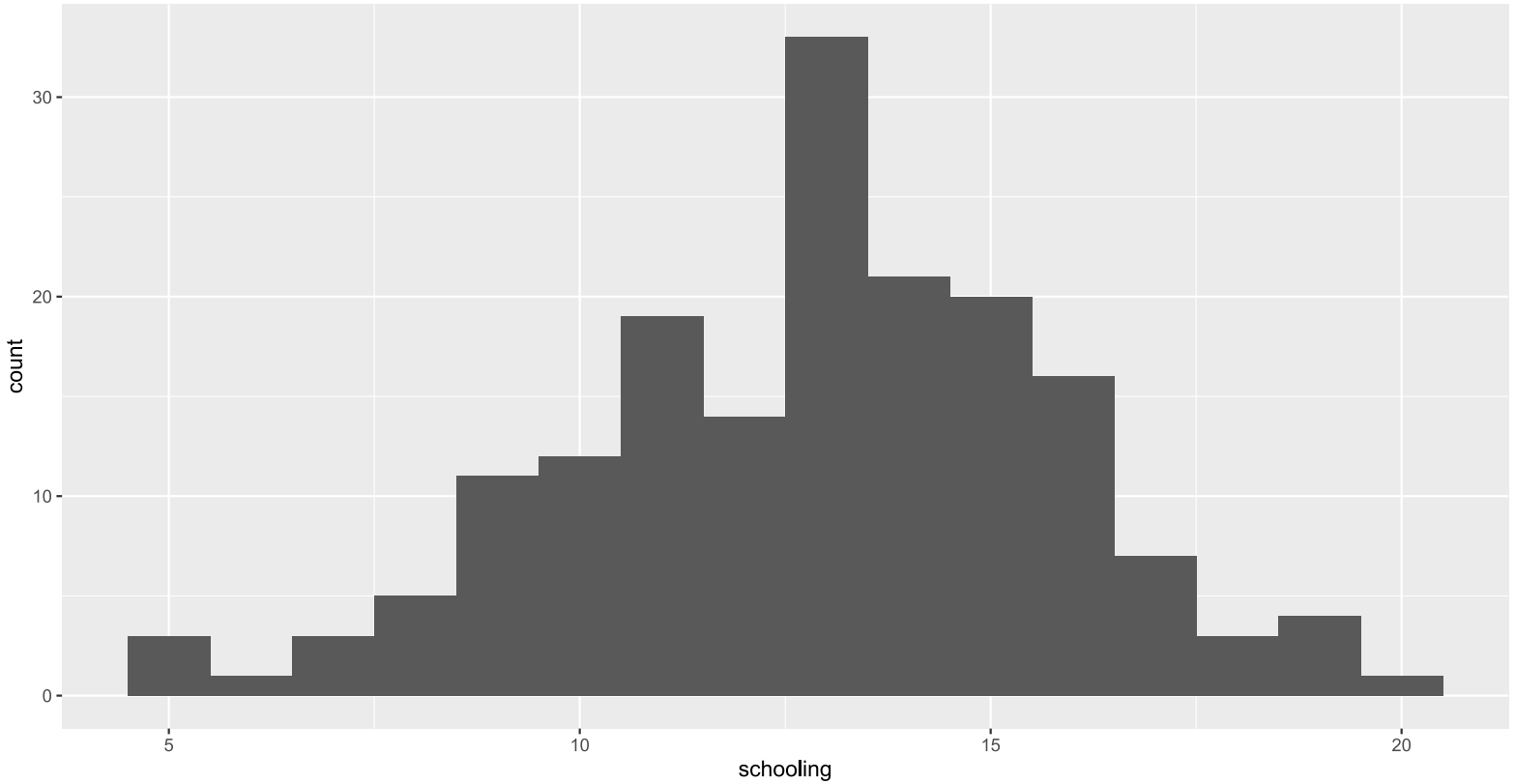
Another way



What about schooling?

```
#>
#>   The decimal point is at the |
#>
#>     4 | 9
#>     5 | 04
#>     6 | 3
#>     7 | 1237
#>     8 | 144589
#>     9 | 00111225569
#>    10 | 0001123334677788889
#>    11 | 111223444677779
#>    12 | 01123555666677777788999
#>    13 | 000111122333334445566789999
#>    14 | 00122233344556678889
#>    15 | 00001223333345668999
#>    16 | 0001333345566
#>    17 | 0123377
#>    18 | 16
#>    19 | 022
#>    20 | 4
```

And differently again



Numerical univariate statistics

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

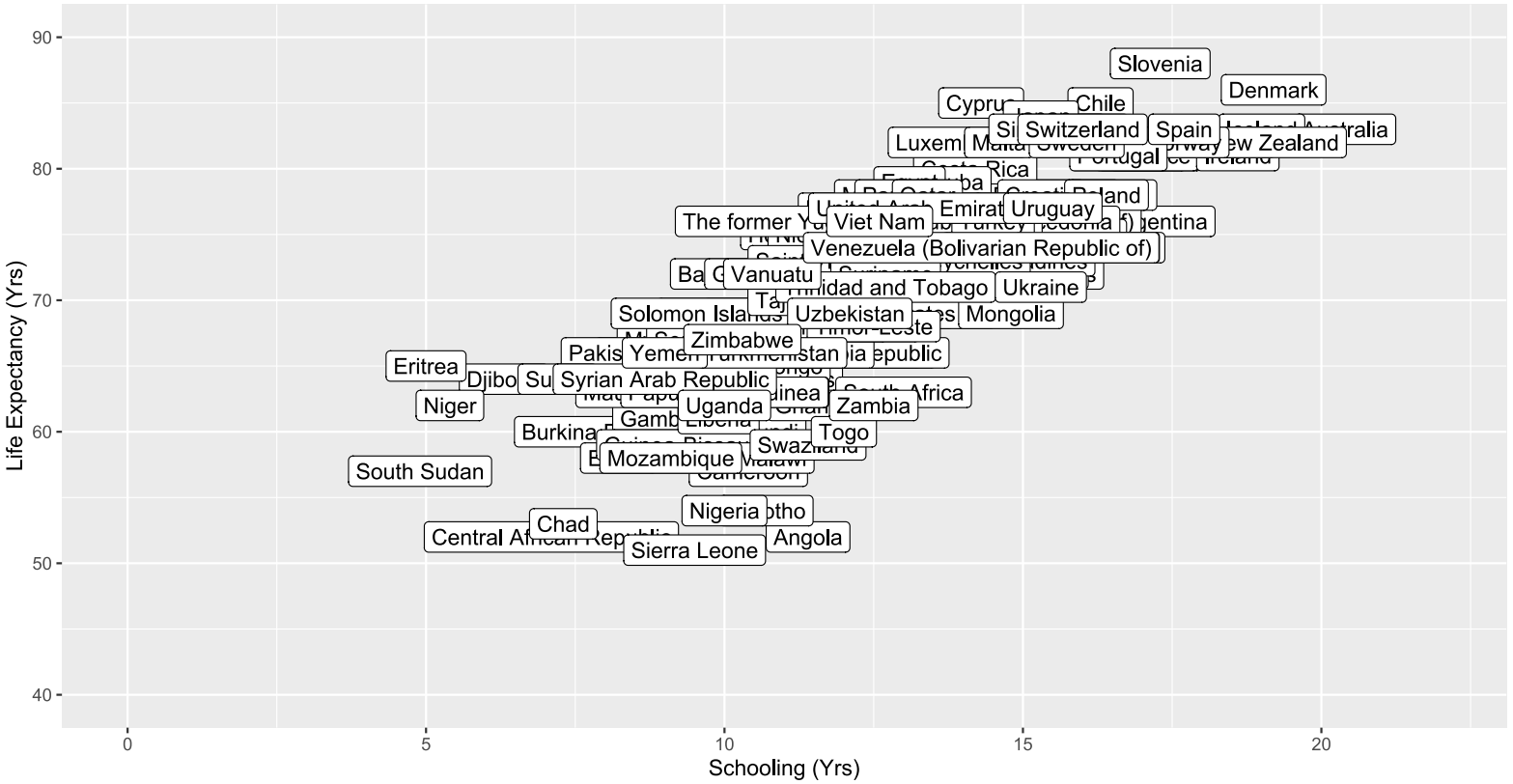
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   51.00   66.00   74.00   71.74   77.00   88.00
```

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

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

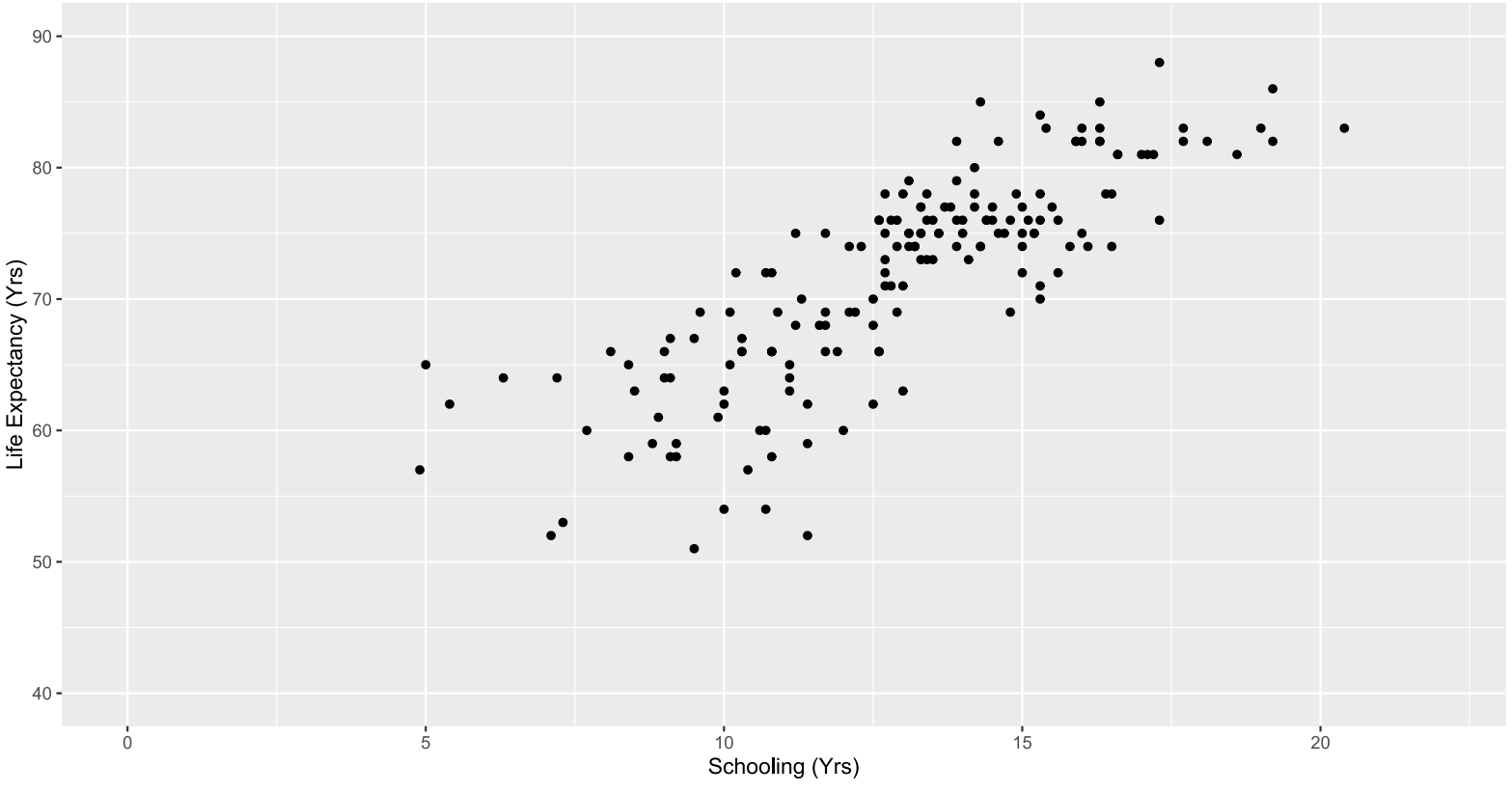
Can you interpret the univariate statistics and displays on this and the previous slides?

Visualizing the relationship



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

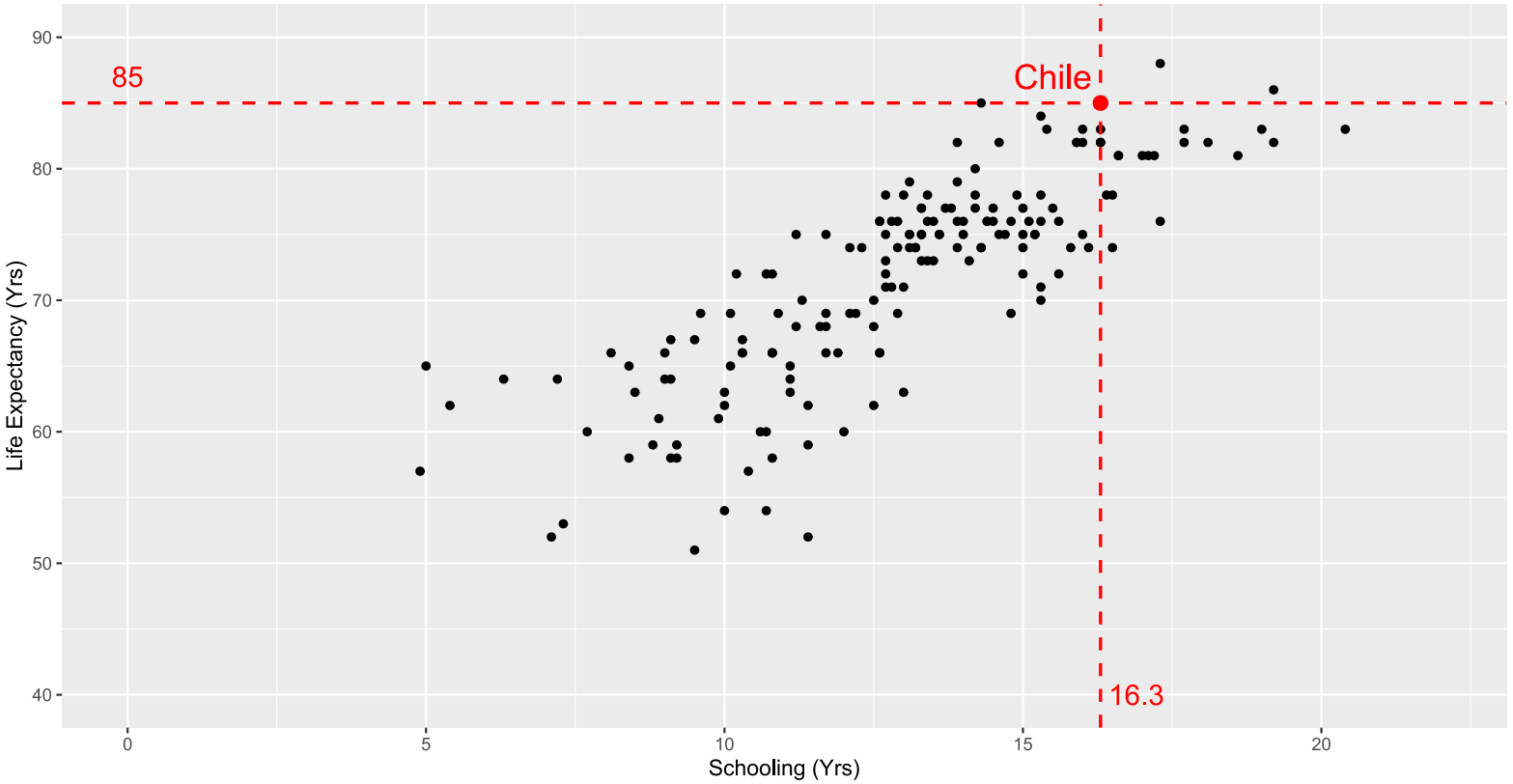
Visualizing the relationship



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*

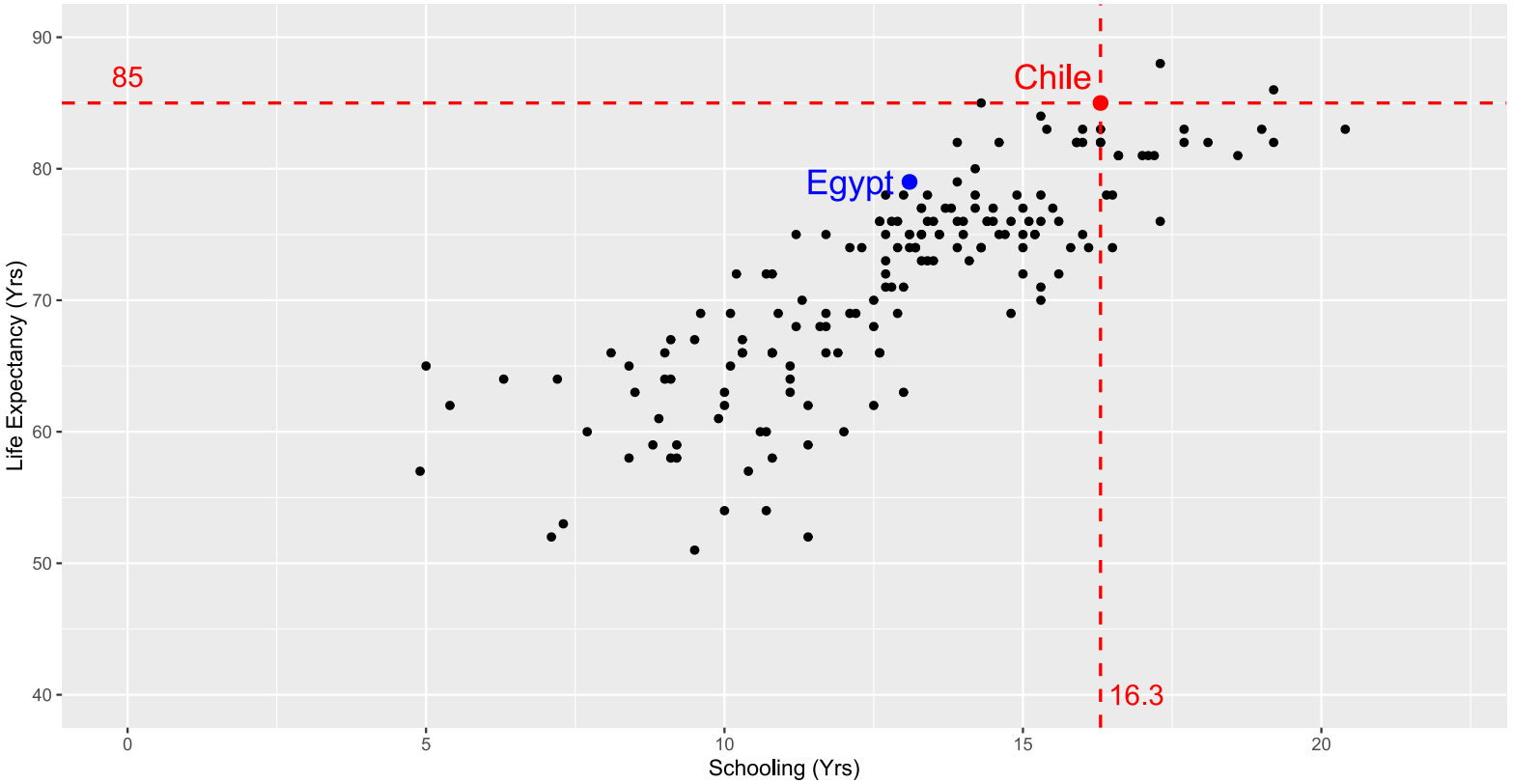
Can you interpret the bivariate display? What does it (and does it NOT) say about the relationship between schooling and life expectancy?

Visualizing the relationship



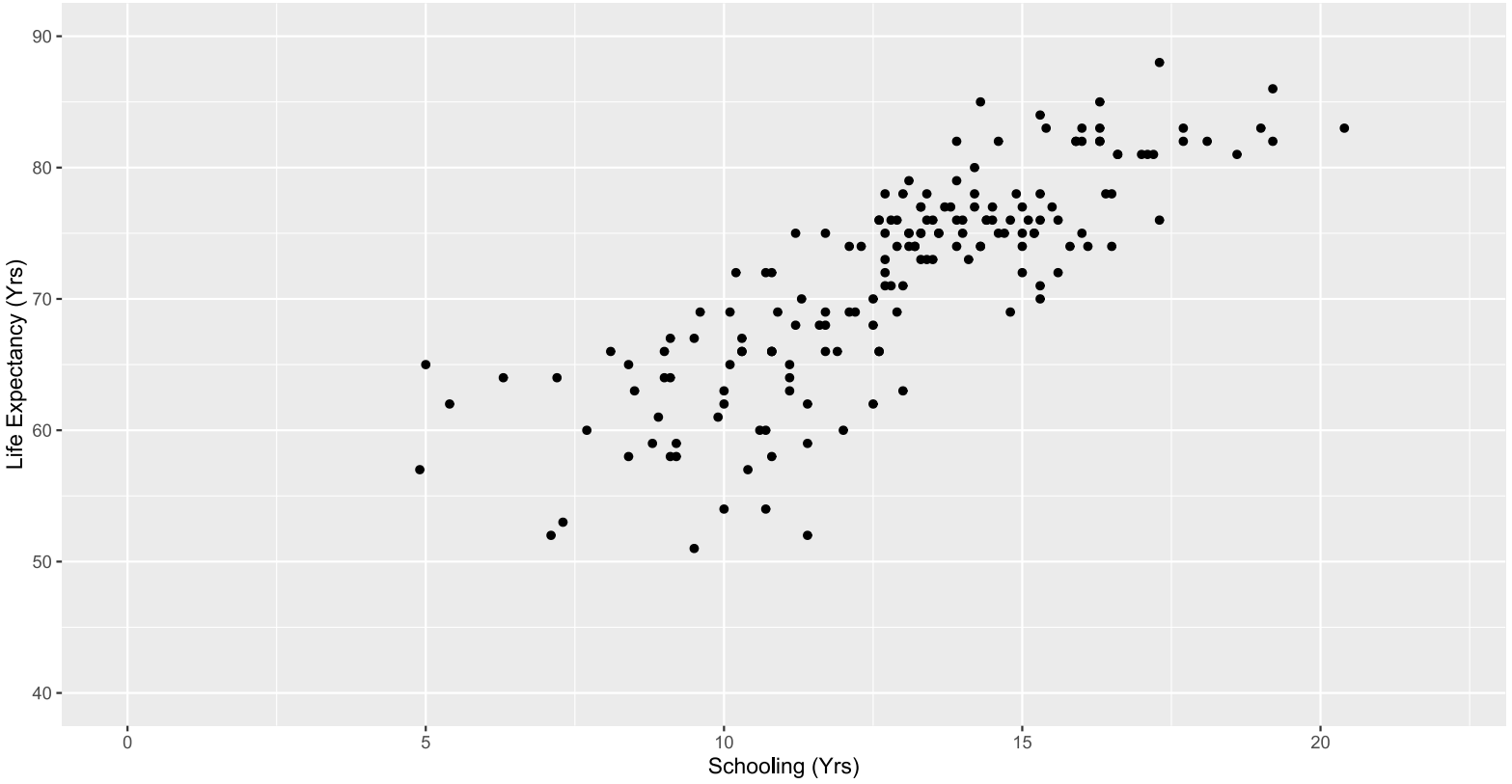
Can you interpret what this display says about the country of Chile?

You try...



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

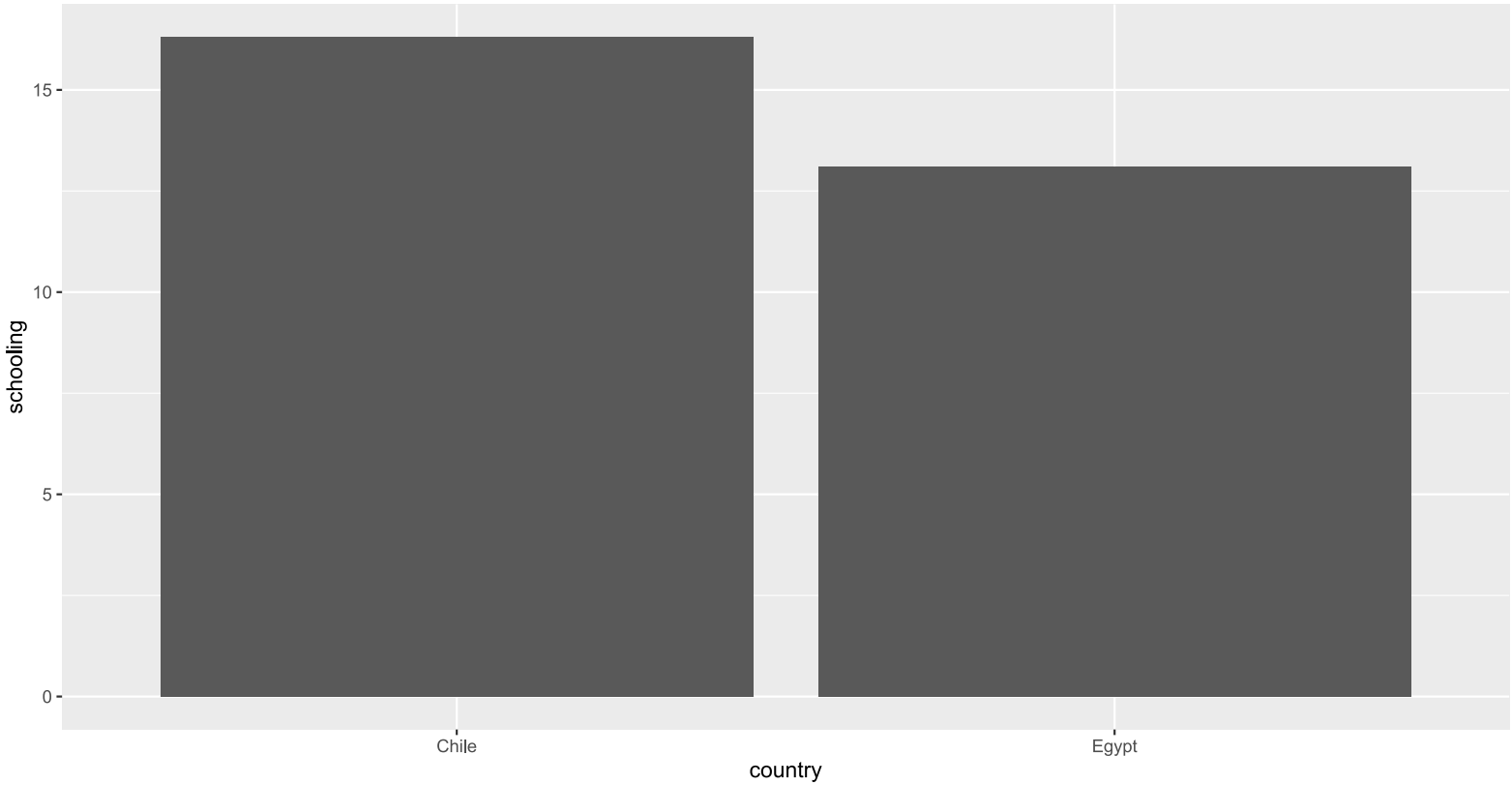
What about the relationship?



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

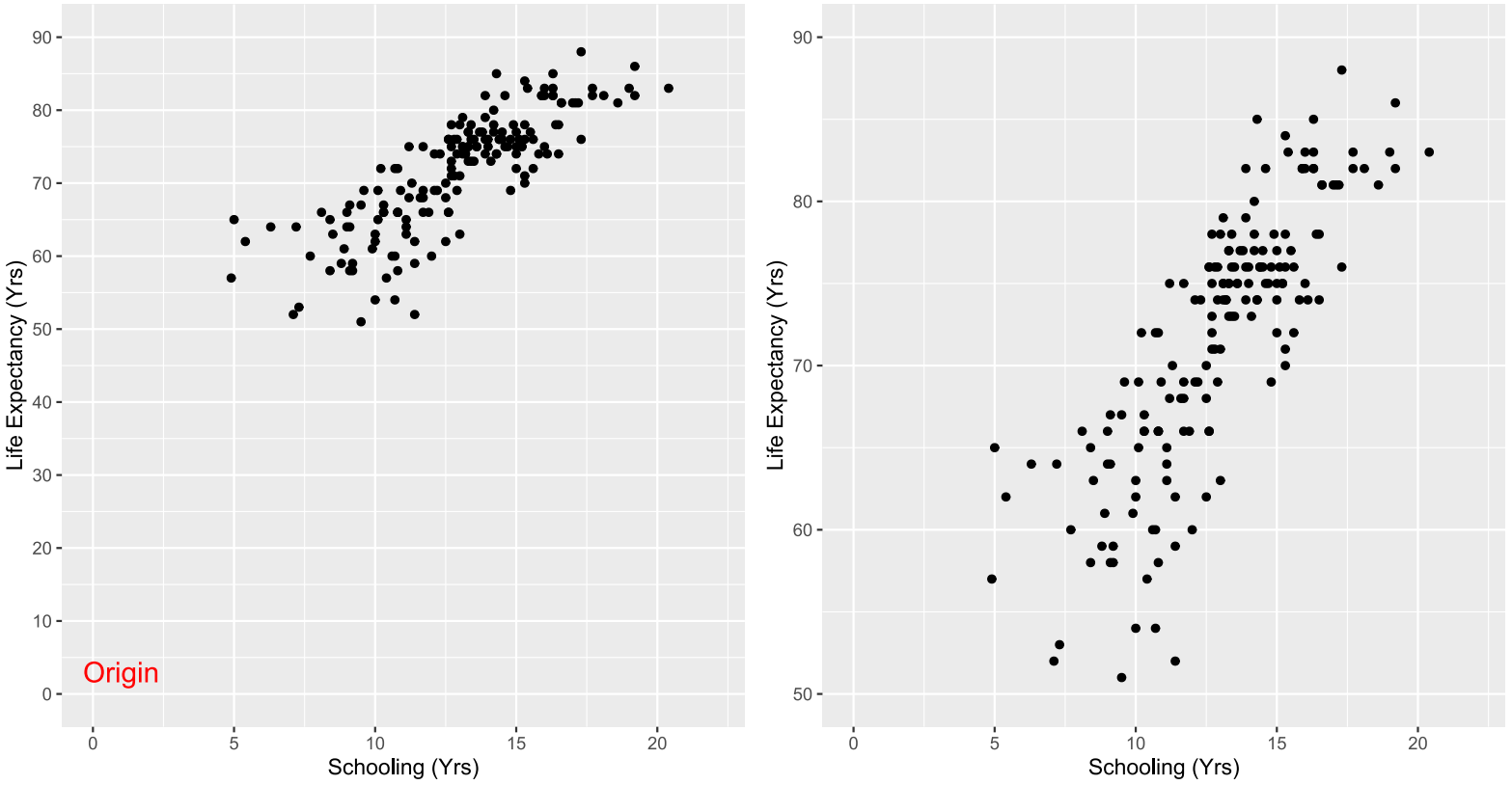
What kind of line, curve or other construction best summarizes the observed relationship between SCHOOLING and LIFE_EXPECTANCY?

An aside about the origin



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

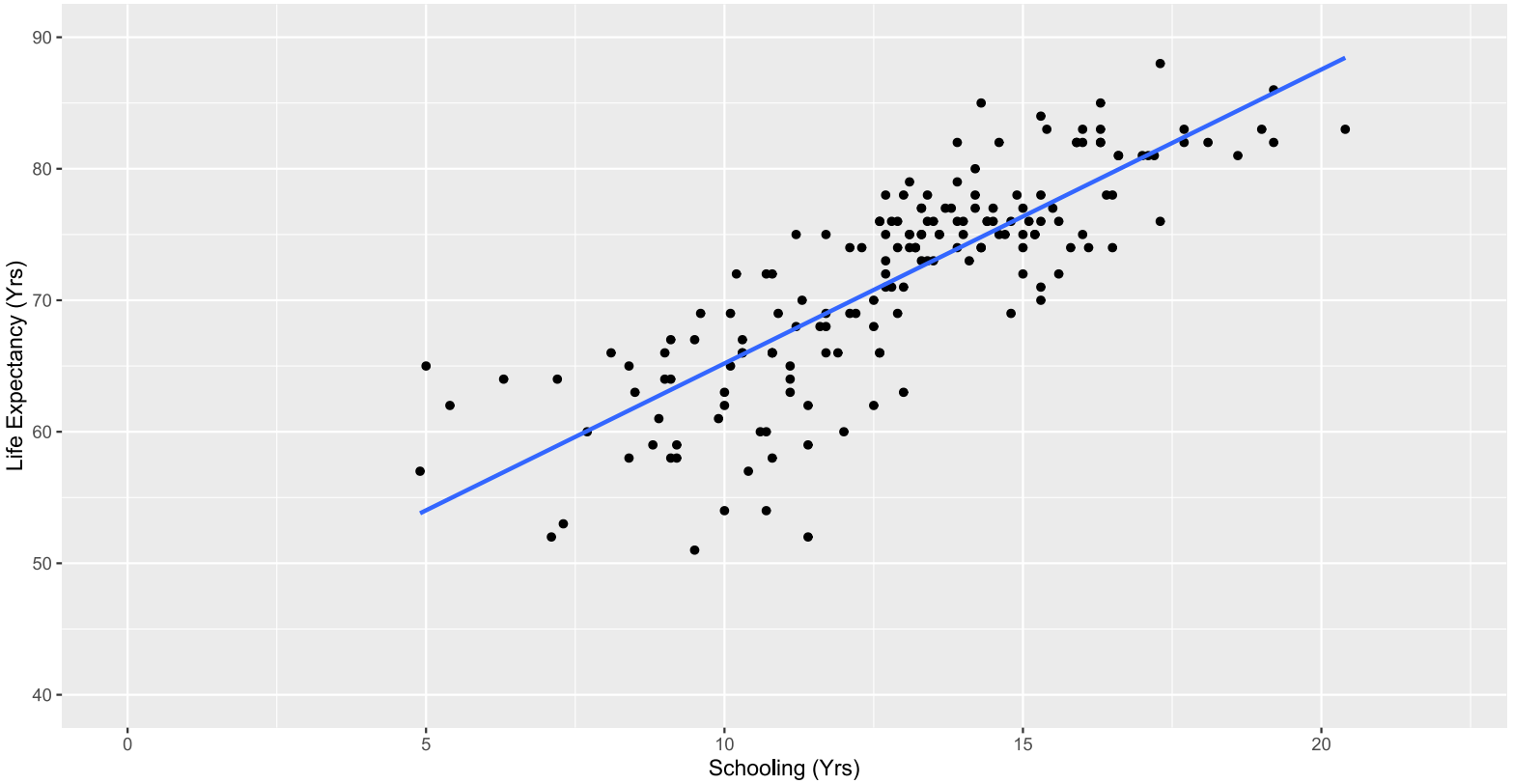
An aside about the origin



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...

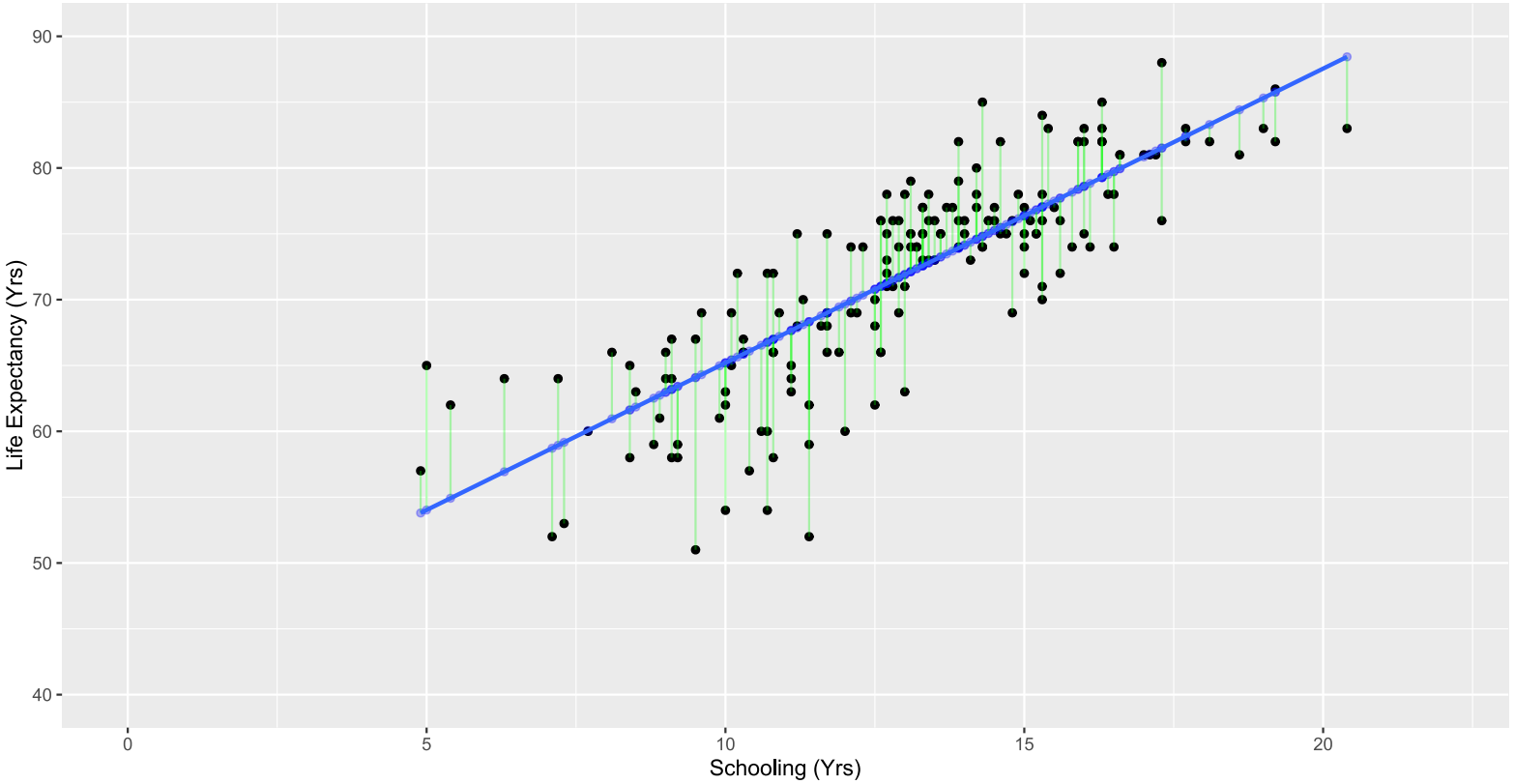
A gentle introduction to bivariate regression: Ordinary-Least Squares (OLS)-fitted regression lines

OLS-fitted regression line



The fitted regression line tells us the best prediction for the values of LIFE_EXPECTANCY.

Some intuition



Can think of the OLS-fitted regression line as a stick held in place by thumbtacks and elastic bands from each of the data points

A visualization

```
#
# This is a Shiny web application. You can run the application by clicking
# the 'Run App' button above.
#
# Find out more about building applications with Shiny here:
#
#    http://shiny.rstudio.com/
#

library(shiny)
library(tidyverse)
library(MASS)
library(reactable)

# Define UI for application that draws a histogram
ui <- fluidPage(

    # Application title
    titlePanel("Sums of Squares Visualization"),

    # Sidebar with a slider input for number of bins
    sidebarLayout(
        sidebarPanel(
            numericInput("intercept",
                         "Intercept:",
                         value = 0),
```

Implementing in R

```
fit ← lm(life_expectancy ~ schooling, data=who)
summary(fit)
```

```
#>
#> Call:
#> lm(formula = life_expectancy ~ schooling, data = who)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -16.3270  -2.6565   0.1581   3.3095  10.9758
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  42.8501     1.5976   26.82   <2e-16 ***
#> schooling     2.2348     0.1206   18.53   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.606 on 171 degrees of freedom
#> Multiple R-squared:  0.6676,    Adjusted R-squared:  0.6657
#> F-statistic: 343.5 on 1 and 171 DF,  p-value: < 2.2e-16
```

The fitted equation

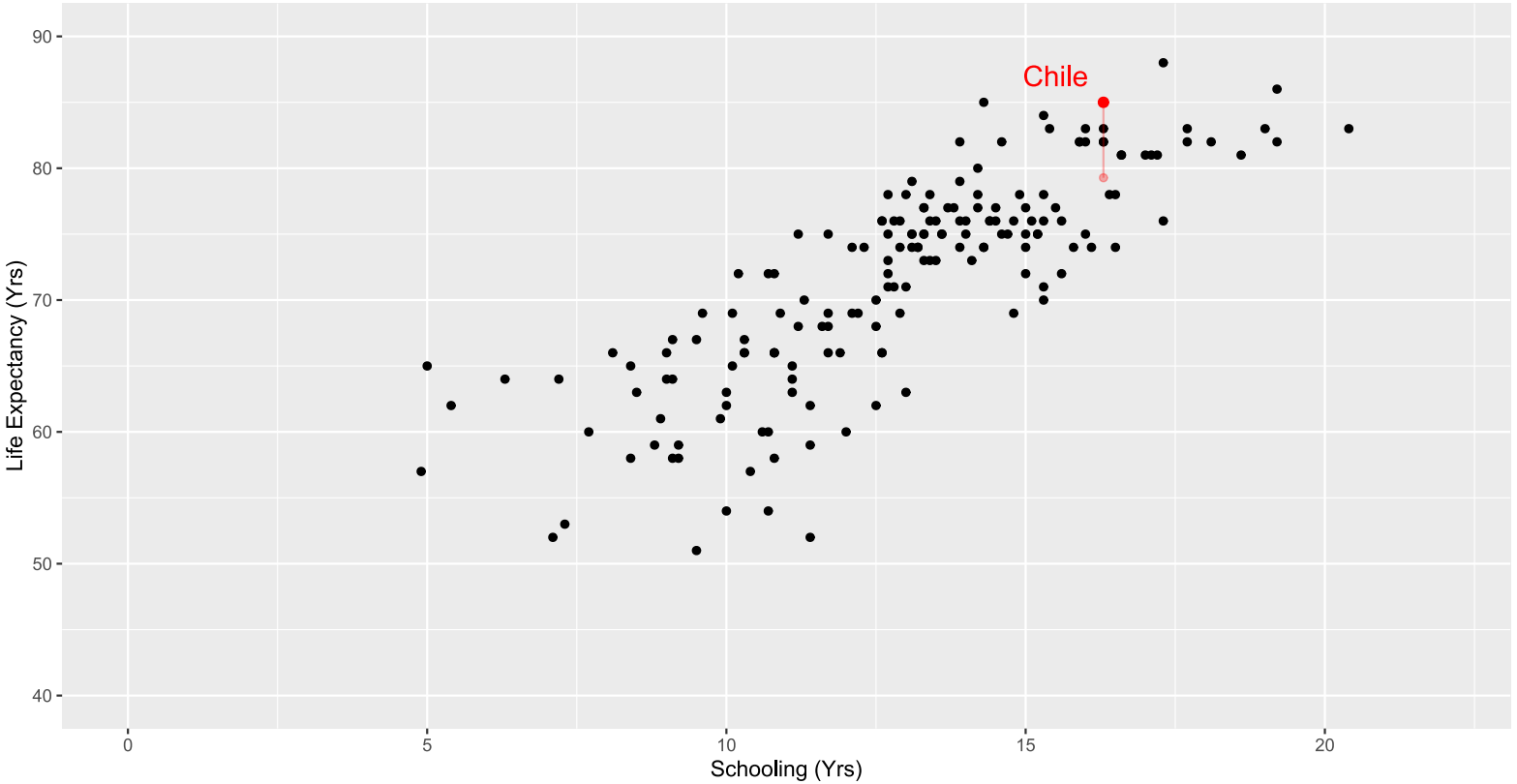
```
#>
#> Call:
#> lm(formula = life_expectancy ~ schooling, data = who)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -16.3270  -2.6565   0.1581   3.3095  10.9758
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  42.8501     1.5976   26.82   <2e-16 ***
#> schooling     2.2348     0.1206   18.53   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.606 on 171 degrees of freedom
#> Multiple R-squared:  0.6676,    Adjusted R-squared:  0.6657
#> F-statistic: 343.5 on 1 and 171 DF,  p-value: < 2.2e-16
```

These "coefficients" tell you where the fitted trend line should be drawn:

$$[\text{Predicted value of } LIFE_EXPECTANCY] = (42.85) + 2.23 * [\text{Observed value of } SCHOOLING]$$

Fitted values

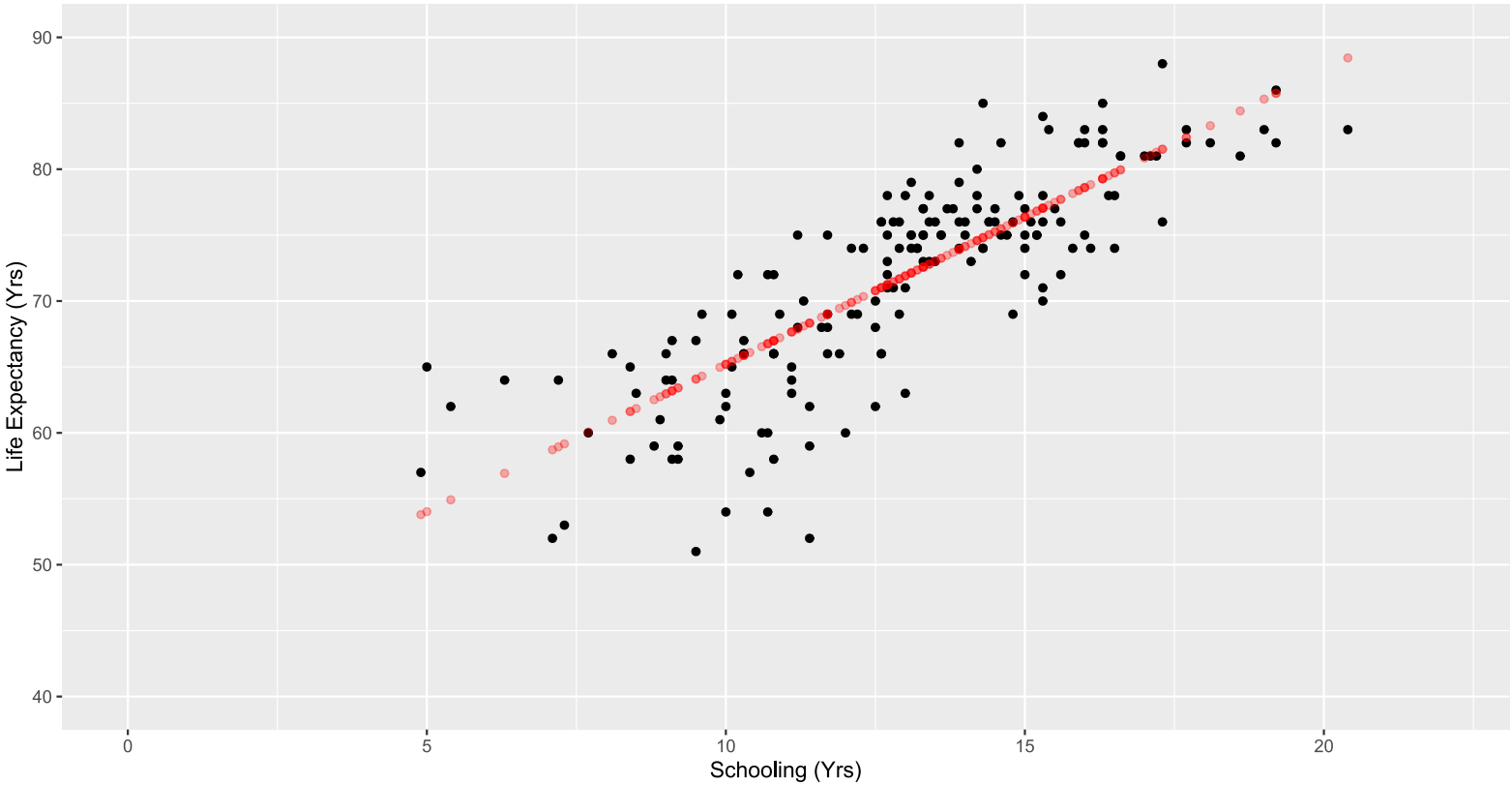
Can substitute values for the "predictor" ($SCHOOLING$) into the fitted equation to compute the *predicted* values of $LIFE\_EXPECTANCY$.



Can do this for our old friend Chile ... and all others...

Fitted values

So we can re-construct the line of best fit from the fitted values:



Fitted values

Note that the fitted line always goes through the average of the predictors

```
mean(who$schooling)
```

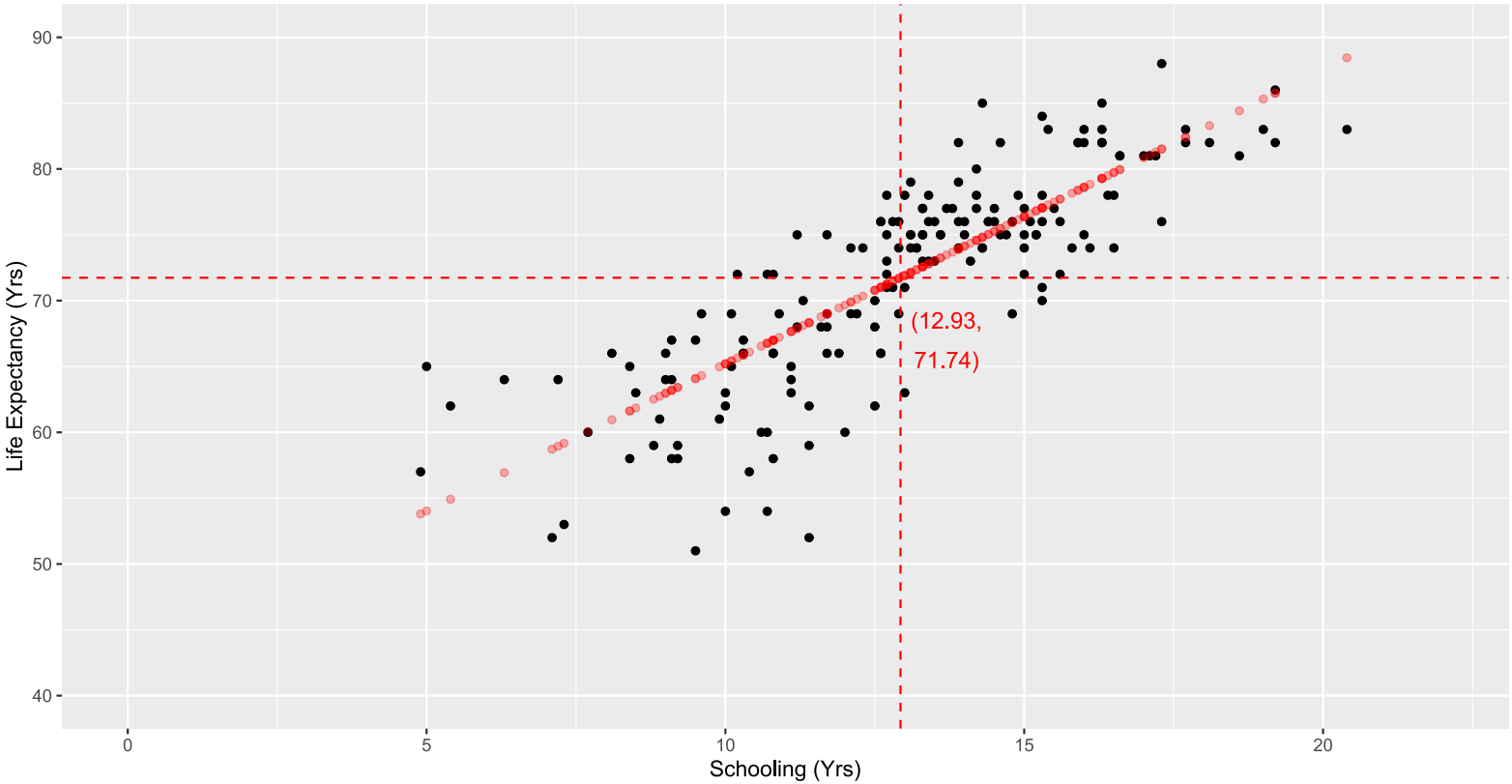
```
#> [1] 12.92717
```

```
mean(who$life_expectancy)
```

```
#> [1] 71.73988
```

Fitted values

Note that the fitted line always goes through the average of the predictors



The regression equation

Each term in the regression equation has a specific interpretation

$$\hat{LIFE_EXPECTANCY} = 42.85 + 2.23 * (SCHOOLING)$$

The regression equation

Each term in the regression equation has a specific interpretation:

$$\hat{LIFE_EXPECTANCY} = 42.85 + 2.23 * (SCHOOLING)$$

The predicted value of $\hat{LIFE\_EXPECTANCY}$ is based on the OLS regression fit. Its "hat" represents that it is a prediction.

The regression equation

Each term in the regression equation has a specific interpretation:

$$\hat{LIFE_EXPECTANCY} = 42.85 + 2.23 * (SCHOOLING)$$

42.85 represents the *estimated intercept*. It tells you the predicted value of $LIFE\_EXPECTANCY$ when $SCHOOLING$ is zero (0)

- *In this context, it doesn't make sense to interpret this. Why?*

The regression equation

Each term in the regression equation has a specific interpretation:

$$\hat{LIFE_EXPECTANCY} = 42.85 + 2.23 * (SCHOOLING)$$

2.23 represents the *estimated slope*. It summarizes the relationship between $LIFE\_EXPECTANCY$ and $SCHOOLING$. It tells you the difference in the predicted values of $LIFE\_EXPECTANCY$ *per unit difference* in $SCHOOLING$.

Slopes can be positive (as in this case) or negative. Here, we conclude that countries that experience one additional year of schooling have an average life expectancy of 2.23 more years.

The regression equation

Each term in the regression equation has a specific interpretation:

$$\hat{LIFE_EXPECTANCY} = 42.85 + 2.23 * (SCHOOLING)$$

$SCHOOLING$ represents the *actual values* of the predictor $SCHOOLING$.

Regression inference

As with our categorical and single-variable continuous data analysis, we can ask whether we might have observed a relationship between $LIFE\_EXPECTANCY$ and $SCHOOLING$ by an idiosyncratic accident of sampling.

Could we have gotten a slope value of 2.23 by sampling from a population in which there was **no relationship** between $LIFE\_EXPECTANCY$ and $SCHOOLING$?

- In other words, by sampling from a *null population* in which the slope was zero?

Regression inference

Could we have gotten a slope value of 2.23 by sampling from a population in which there was **no relationship** between $LIFE\_EXPECTANCY$ and $SCHOOLING$?

```
#>
#> Call:
#> lm(formula = life_expectancy ~ schooling, data = who)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -16.3270  -2.6565   0.1581   3.3095  10.9758
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  42.8501     1.5976   26.82   <2e-16 ***
#> schooling     2.2348     0.1206   18.53   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.606 on 171 degrees of freedom
#> Multiple R-squared:  0.6676,    Adjusted R-squared:  0.6657
#> F-statistic: 343.5 on 1 and 171 DF,  p-value: < 2.2e-16
```

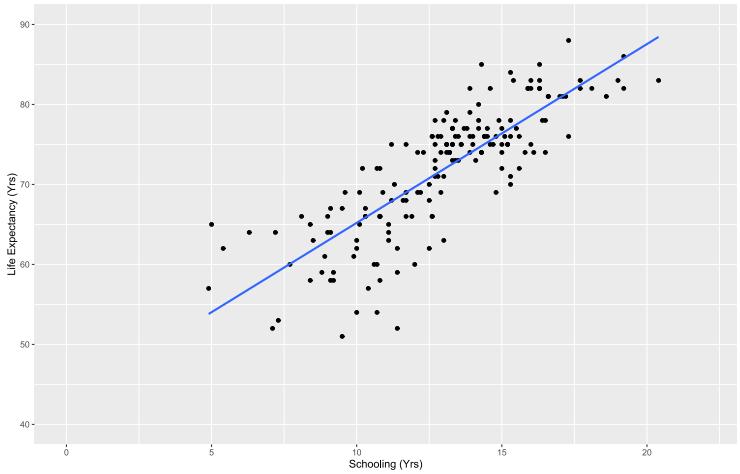
Regression inference

```
#>
#> Call:
#> lm(formula = life_expectancy ~ schooling, data = who)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -16.3270  -2.6565   0.1581   3.3095  10.9758
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  42.8501     1.5976   26.82   <2e-16 ***
#> schooling     2.2348     0.1206   18.53   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 4.606 on 171 degrees of freedom
#> Multiple R-squared:  0.6676,    Adjusted R-squared:  0.6657
#> F-statistic: 343.5 on 1 and 171 DF,  p-value: < 2.2e-16
```

Here, the *p*-value for the $\frac{LIFE\_EXPECTANCY}{SCHOOLING}$ regression slope is < 0.0001 (in fact, $< 2^{-16}$).

With an alpha-threshold of 0.05, 2^{-16} is definitely less than 0.05. Thus, we reject the null hypothesis that there is no relationship between $LIFE\_EXPECTANCY$ and

Writing it up

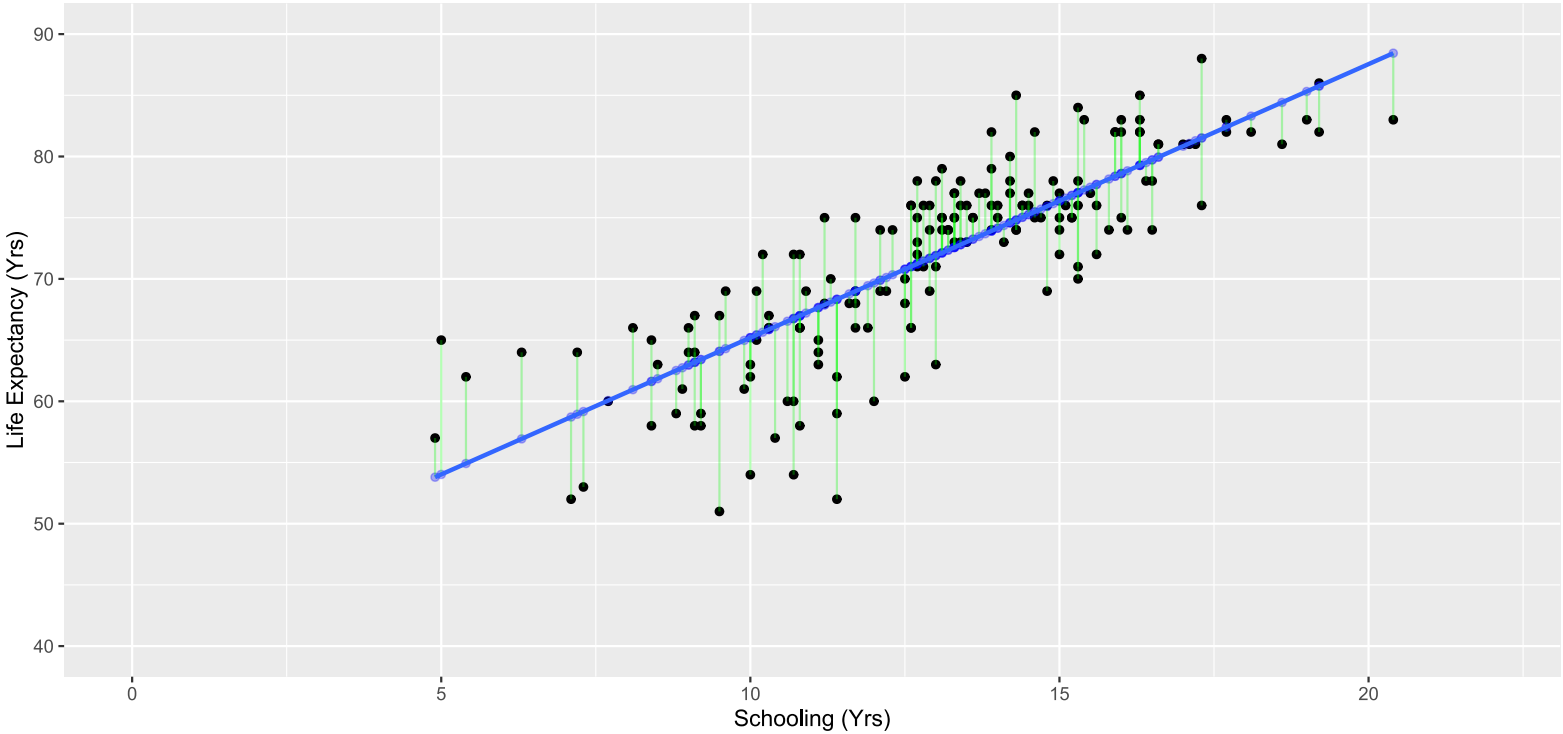


The story so far

In our investigation of country-level aggregate measures of schooling and life expectancy, we have found that the average years of schooling in a country is related to the average life expectancy. In particular, when we relate the country-level life expectancy (*LIFE_EXPECTANCY*) to the country-level mean years of schooling (*SCHOOLING*), we find that the trend-line estimated by ordinary-least-squares regression has a slope of 2.23 (p<0.0001). This suggests that two countries that differ in their average years of schooling attainment by 1 year will have, on average, a difference in life expectancy of 2.23 years. Of course, this relationship is far from causal...

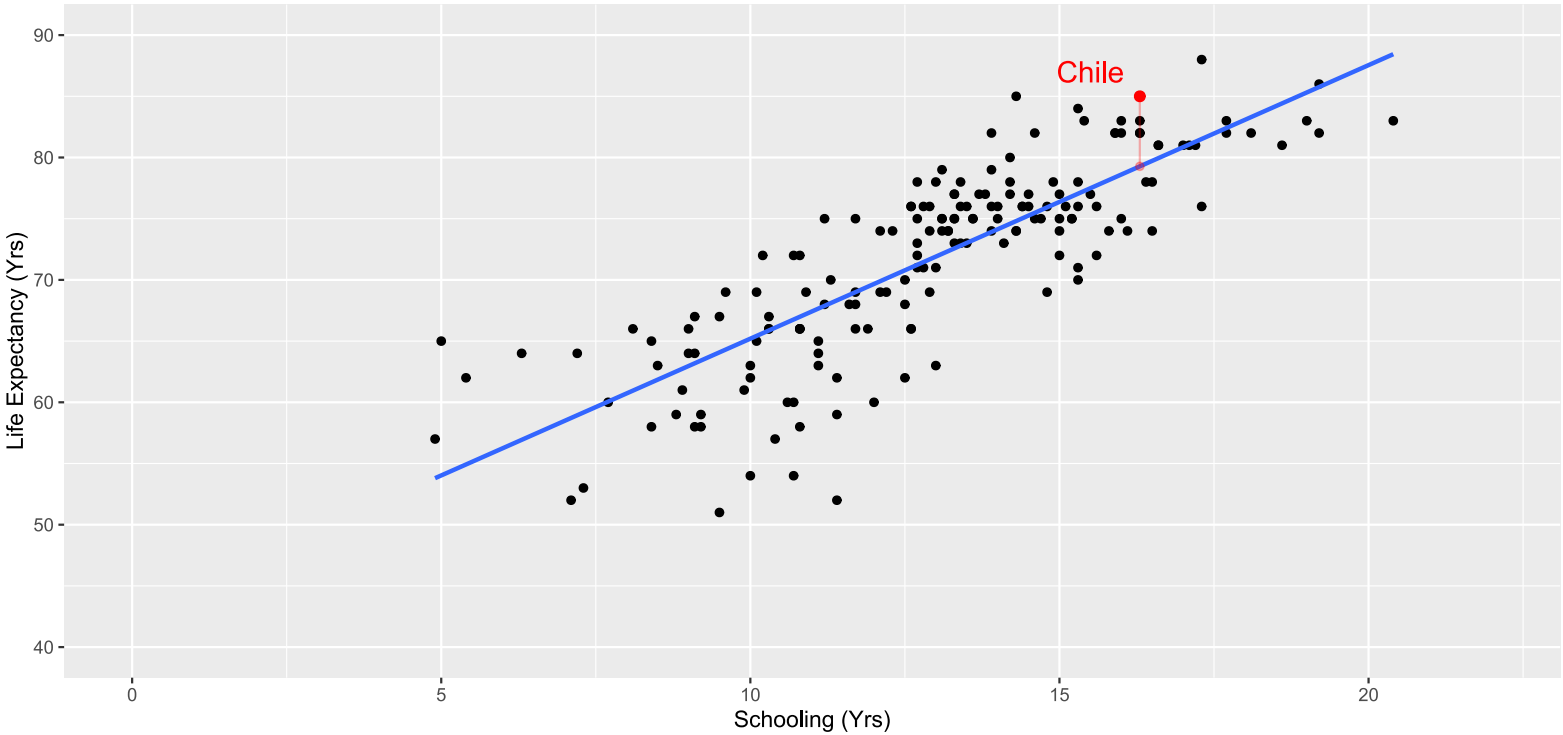
A gentle introduction to bivariate regression: Residual analysis

Residual analysis



Our fitted regression line contains the "predicted" values of *LIFE_EXPECTANCY* for each value of *SCHOOLING*. But almost all of the "actual" values of *LIFE_EXPECTANCY* lie off the actual line regression line.

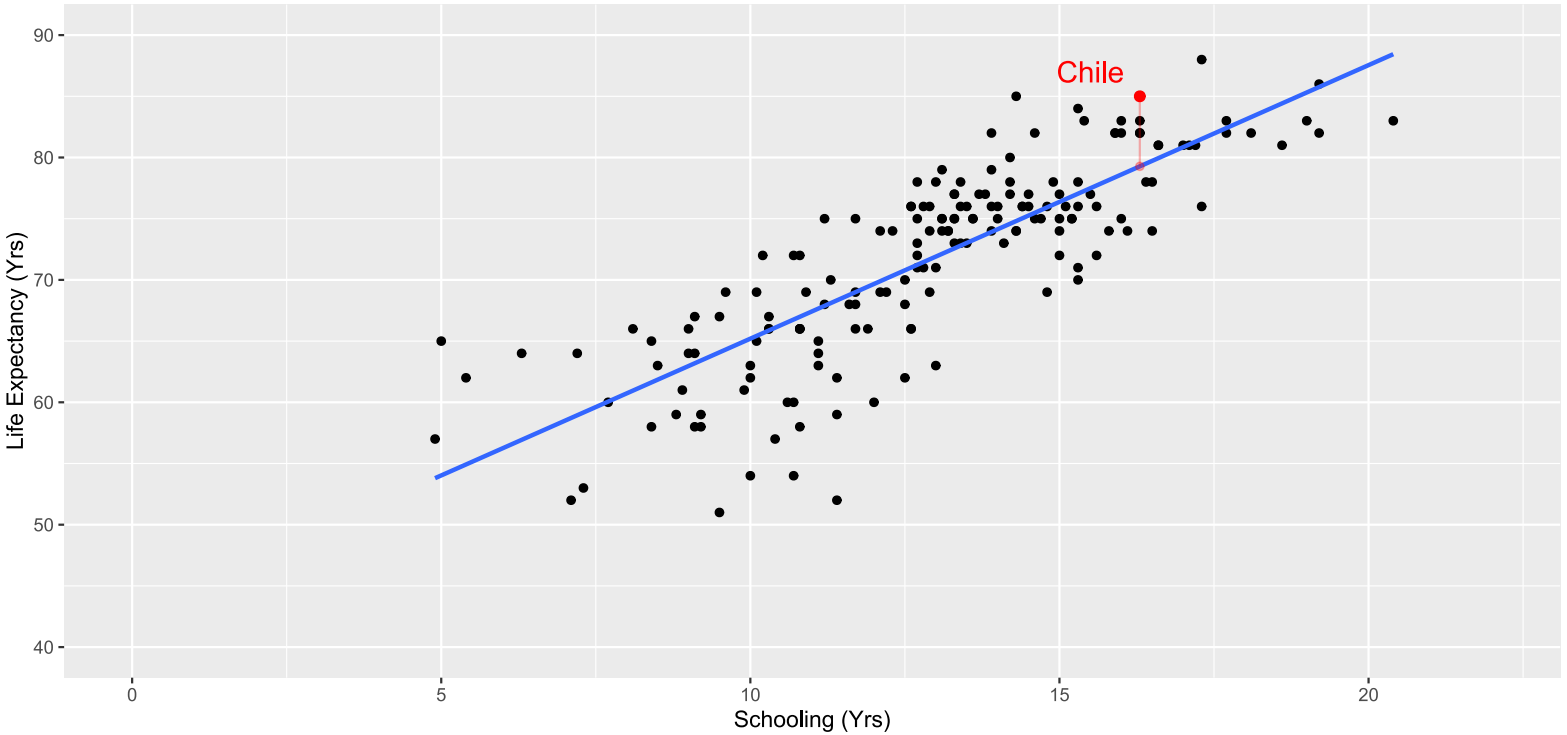
An example: Chile



Observed values for Chile: $LIFE\_EXPECTANCY = 85$; $SCHOOLING = 16.3$
Predicted value of *LIFE_EXPECTANCY* for Chile:

$$\begin{aligned} \hat{LIFE\_EXPECTANCY} &= 42.85 + 2.23 * (16.3) \\ &= 79.20 \end{aligned}$$

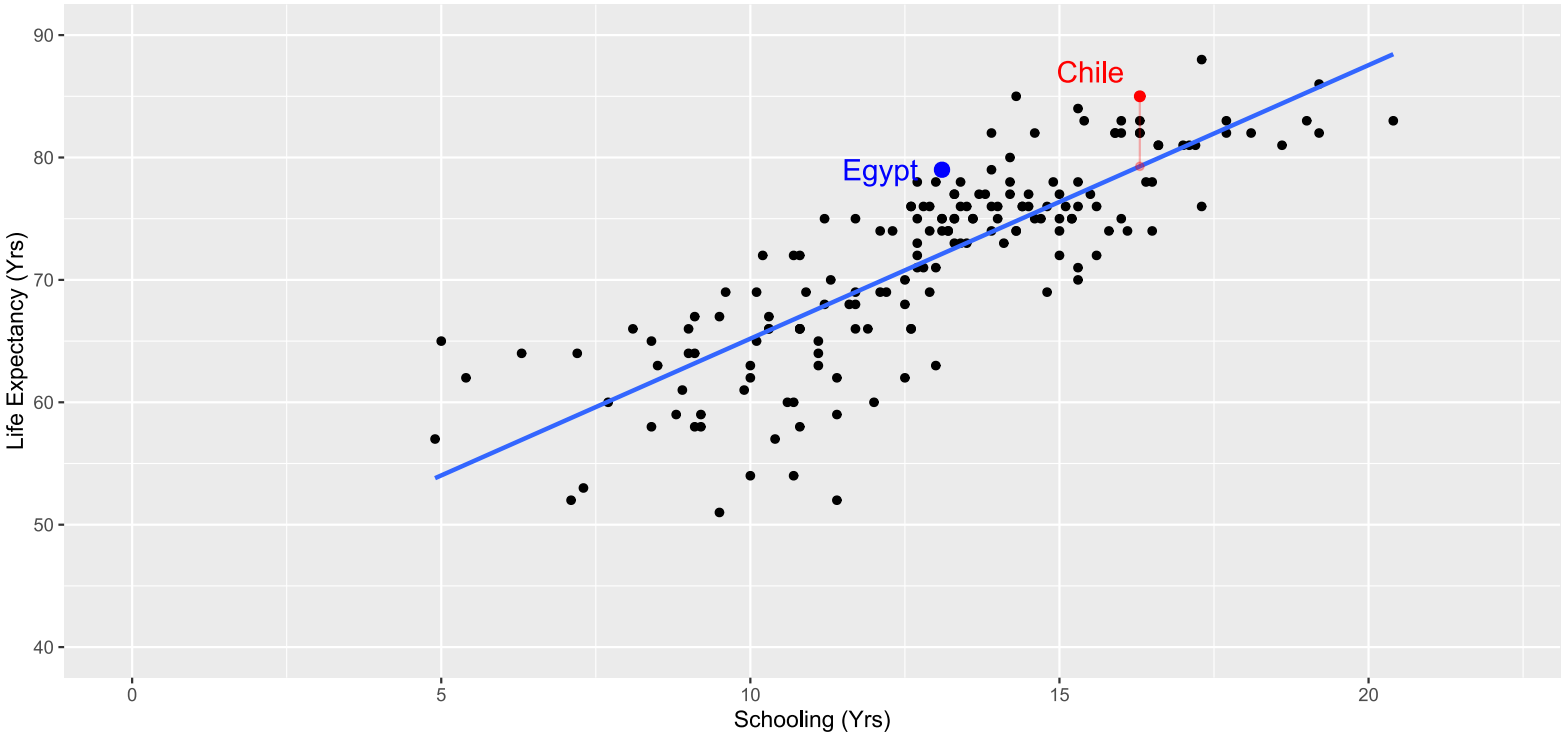
An example: Chile



$\hat{LIFE\_EXPECTANCY} = 79.20$
Actual life expectancy = 85

What can we say about the country of Chile's average life expectancy, relative to our prediction?

Now Egypt



Observed values for Egypt: $LIFE\_EXPECTANCY = 79$; $SCHOOLING = 13.1$

Can you calculate the predicted value of *LIFE_EXPECTANCY* for Egypt and compare it to the observed?

What is a "residual"?

The difference ("vertical distance") between the observed value of the outcome its predicted value is called the *residual*.

Residuals can be substantively and statistically useful:

- Represent individual deviations from average trend
- Tell us about values of the outcome after taking into account ("adjusting for") the predictor
    - In this case, tell us whether countries have better or worse life expectancies, given their average years of schooling

Residual analysis

```
fit ← lm(life_expectancy ~ schooling, data=who)

# predict asks for the predicted values
who$predict ← predict(fit)

# resid asks for the raw residual
who$resid ← residuals(fit)
```

We can now treat these residual and predicted values as new variables in our dataset and examine using all the other univariate and multivariate analysis tools we have.

Examining the residuals

```
summary(who$resid)
```

```
#>     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
#> -16.3270  -2.6565   0.1581   0.0000   3.3095  10.9758
```

- Sample mean of the residuals is *always* exactly zero
- We've done a very poor job of predicting life expectancy for some countries

```
sd(who$resid)
```

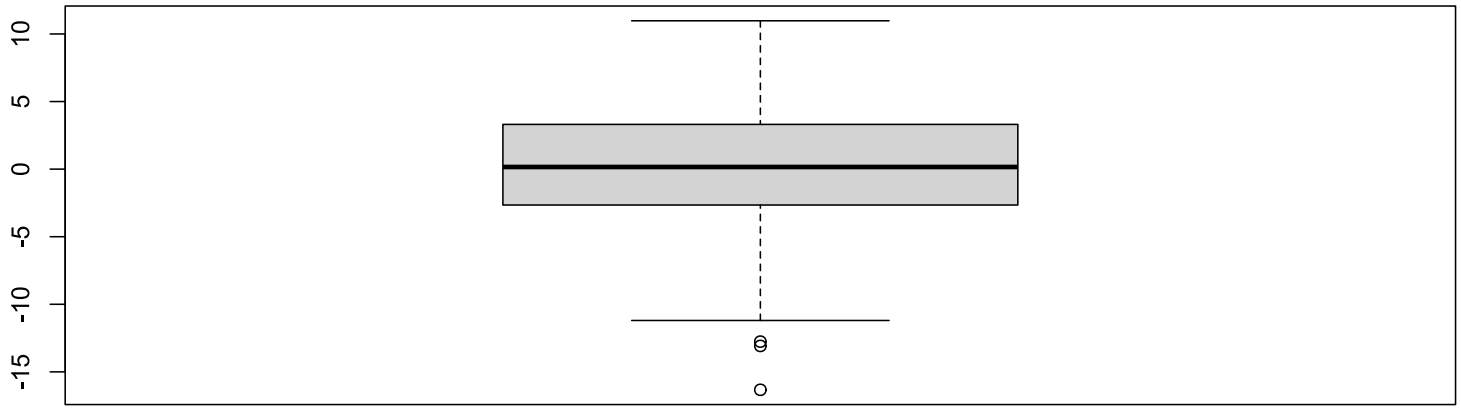
```
#> [1] 4.592143
```

- Standard deviation of the raw residuals can be quite useful in checking our assumptions
  - *What assumption?*

Residual assumptions

For the *p*-values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**

```
boxplot(resid(fit))
```

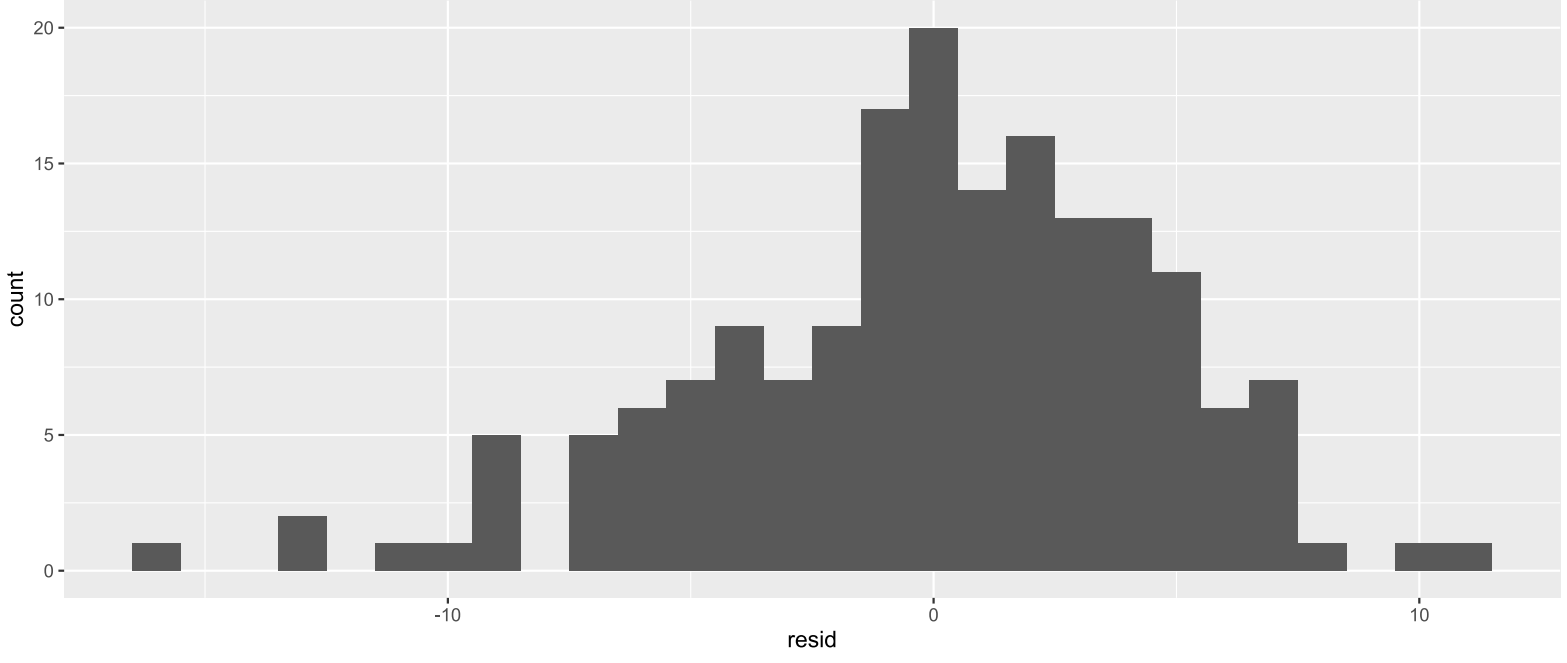


A few outliers, but we seem to be doing ok...

Residual assumptions

For the *p*-values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**

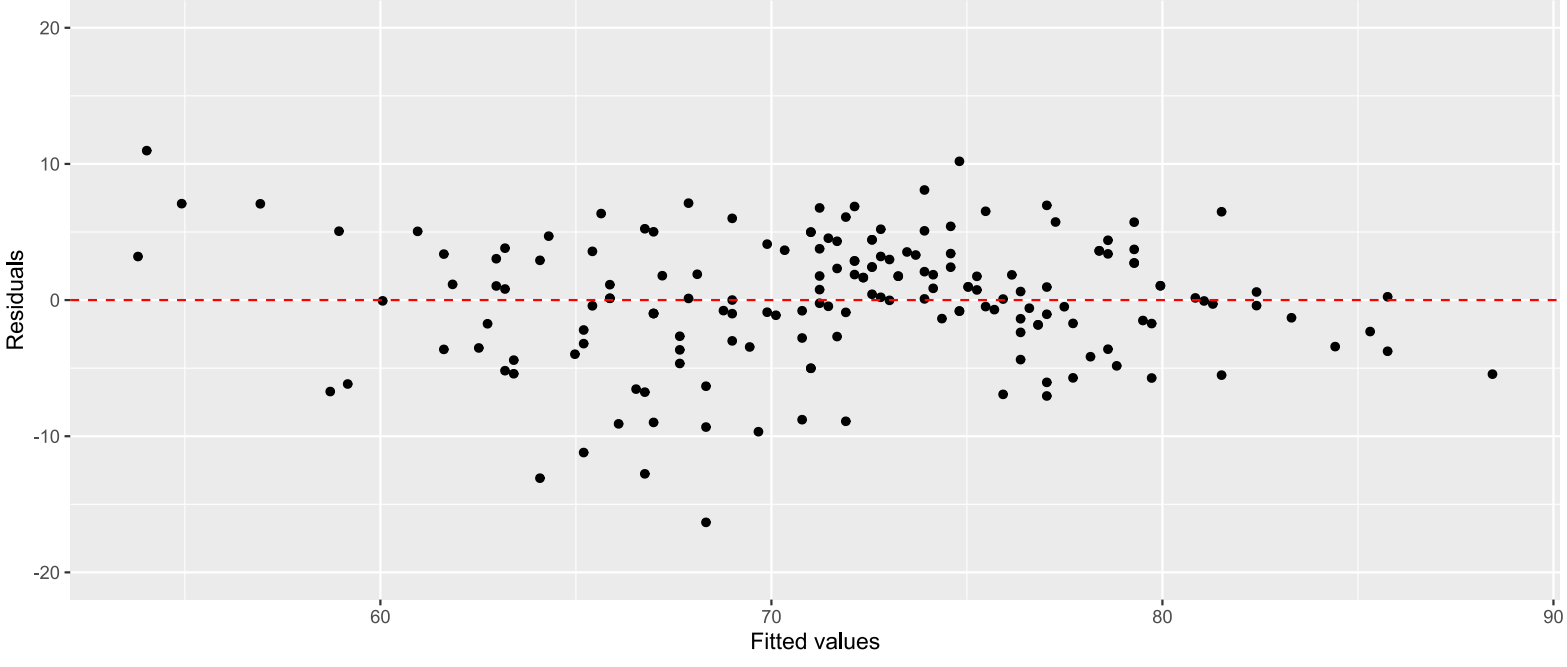
```
ggplot(who, aes(x = resid)) +
  geom_histogram(binwidth = 1)
```



Pretty good, pretty good...

Residual vs. fitted plot

For the *p*-values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**

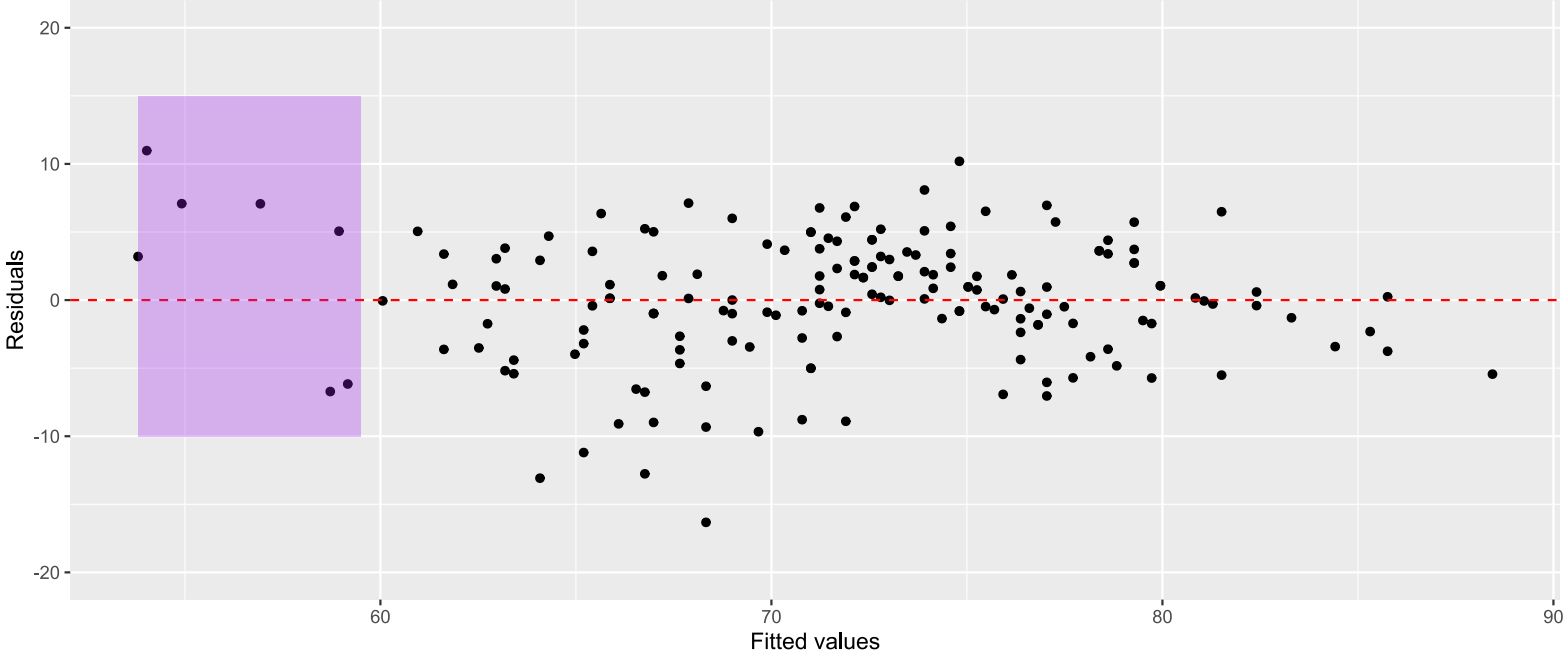


Key assumption checks for normality:

- The residuals "bounce randomly" around the 0 line.
- The residuals could be roughly contained within a rectangle around the 0 line.
- No one residual "stands out" from the basic random pattern of residuals.

Residual vs. fitted plot

For the *p*-values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**



Key assumption checks for normality:

- The residuals "bounce randomly" around the 0 line.
- The residuals could be roughly contained within a rectangle around the 0 line.
- No one residual "stands out" from the basic random pattern of residuals.

Writing it up

> In our investigation of country-level aggregate measures of schooling and life expectancy, we have found that the average years of schooling in a country is related to the average life expectancy. In particular, when we relate the country-level life expectancy (*LIFE_EXPECTANCY*) to the country-level mean years of schooling (*SCHOOLING*), we find that the trend-line estimated by ordinary-least-squares regression has a slope of 2.23 (*p*<0.0001). This suggests that two countries that differ in their average years of schooling attainment by 1 year will have, on average, a difference in life expectancy of 2.23 years. Of course, this relationship is far from causal...
>
> An analysis of the residuals from our fitted model suggests that our regression assumptions are reasonably well met and we have appropriately characterized the relationship between schooling and life expectancy. Despite the presence of a few outliers, our residuals are roughly symmetrically distributed around 0. Our fitted regression does seem to underpredict life expectancy for very low levels of schooling.

Key takeaways

- **Understand your data first**
  - Summarize and visualize each variable independently
  - Start with a visual representation of the relationship between your two variables
  - How you display the relationship will influence your perception of the relationship, but will not change the relationship
  - Try to describe what a particular observation in your visualized data represents
- **The regression model represents your hypothesis about the population**
  - When you fit a regression model, you are estimating *sample* values of *population* parameters that you will not directly observe
  - The goal of classical regression inference (just as with categorical relationships) is to understand how likely the observed data in your sample are in the presence of no relationship in the unobserved population
- **The regression model has a "smooth" and a "rough" component to it**
  - The "smooth" part is the portion of the relationship that your model explains
  - The "rough" part is the extent to which each observation (and the observations in aggregate) vary from the "smooth" part of your predictions
  - The "rough" parts (the residuals) provide important information on the extent to which our models satisfy their assumptions

More on all of this in EDUC 643

Synthesis and wrap-up