Examining the Relationship Between Continuous Variables

EDUC 641: Unit 4 Part 2

David D. Liebowitz

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

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

Our continuous relationship

(and some data-cleaning)

Reading our data in

```r
who ← read.csv(here("data/life_expectancy.csv")) %>%
  # first making variable names take a common format
  janitor::clean_names() %>%
  # filtering to focus only on 2015
  filter(year == 2015) %>%
  # selecting only the variables we need
  select(country, status, schooling, life_expectancy) %>%
  # renaming one of the variables that is really misnamed
  rename(region = country) %>%
  # rounding life expectancy to nearest year
  mutate(life_expectancy = round(life_expectancy, digits = 0))
```

First data cleaning step:

Identify missingness

```
sum(is.na(who$life_expectancy))
```

```
#> [1] 0
```

```
sum(is.na(who$schooling))
```

```
#> [1] 10
```

```
### For the really ambitious...
sapply(who, function(x) sum(is.na(x)))
```

```
#>          region          status       schooling life_expectancy
#>               0               0              10               0
```

So some missingness...what do we do?

Listwise vs. pairwise deletion

- **Listwise**: any observations with any missingness (NA) for any of the variables to be used in our analysis are dropped. Analysis only conducted on observations that have complete data
- **Pairwise**: observations with missingness for some of the variables to be used in analysis are retained and included in sample when the particular analysis does rely on that variable, but are necessarily excluded in analyses that rely on the variable with missingness.

```
mean(who$life_expectancy, na.rm = T)
```

```
#> [1] 71.63934
```

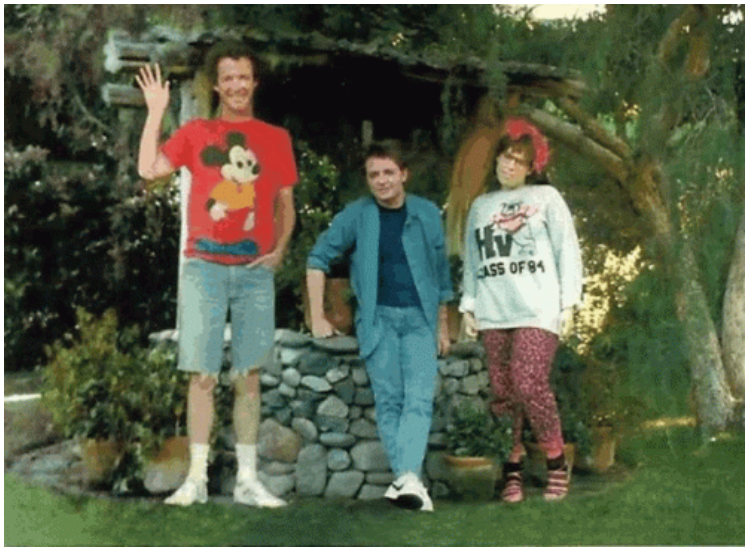
```
mean(who$schooling, na.rm = T)
```

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

How have we handled our missing data in estimating these univariate measures of central tendency?

The chainsaw approach

- Generally, we want to have a stable analytic sample so that differences across estimation strategies reflect differences in our models rather than sample differences
- However, simply dropping these observations may (severely) limit our desired external generalizability
- There are various imputation methods that you will explore more in EDUC 645
- With large data and a small amount of missingness, it generally doesn't matter what you do
- For now, we're going to employ **listwise** deletion

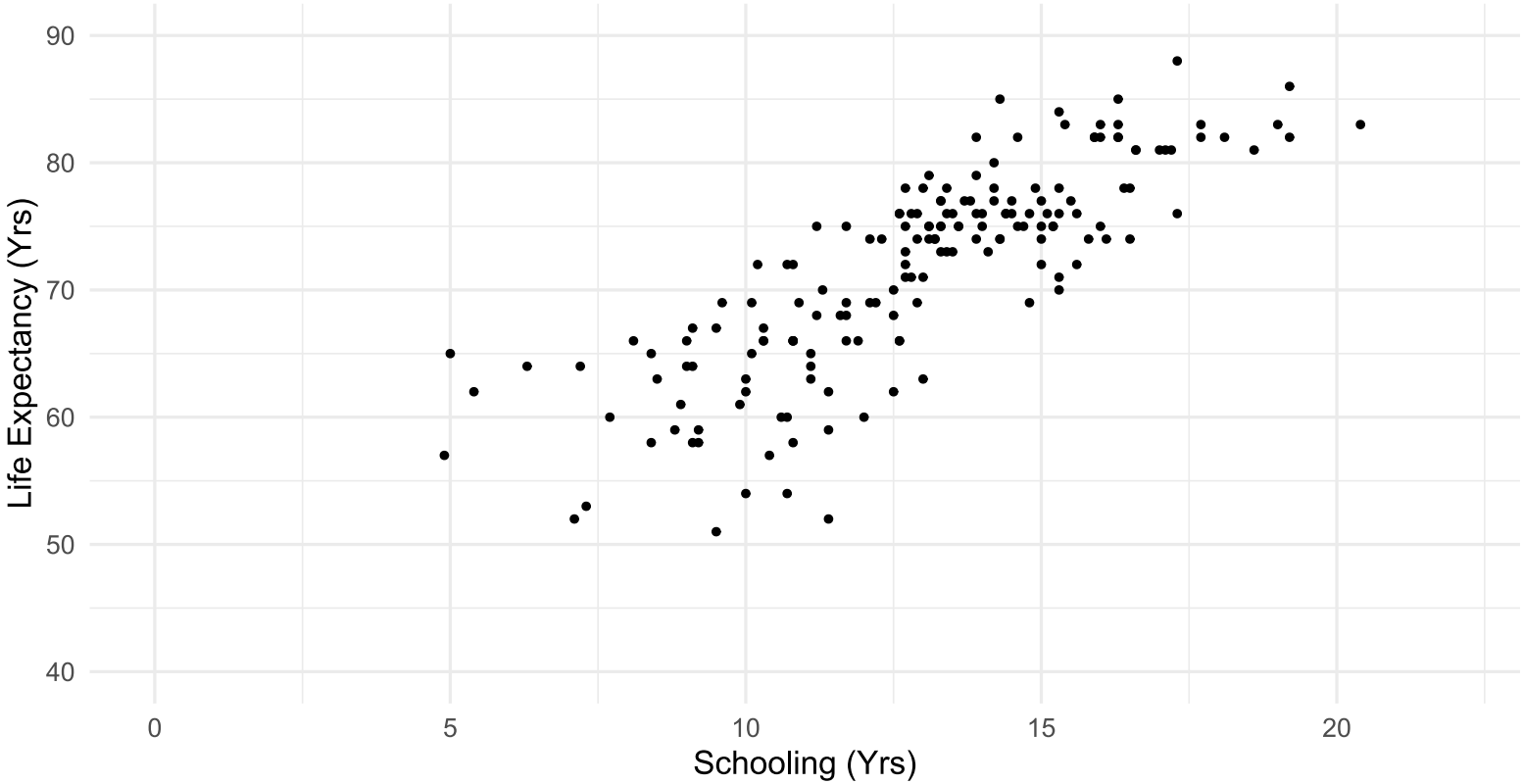


```
who ← filter(who, !is.na(schooling))
nrow(who)
```

```
#> [1] 173
```

A reminder of our relationship

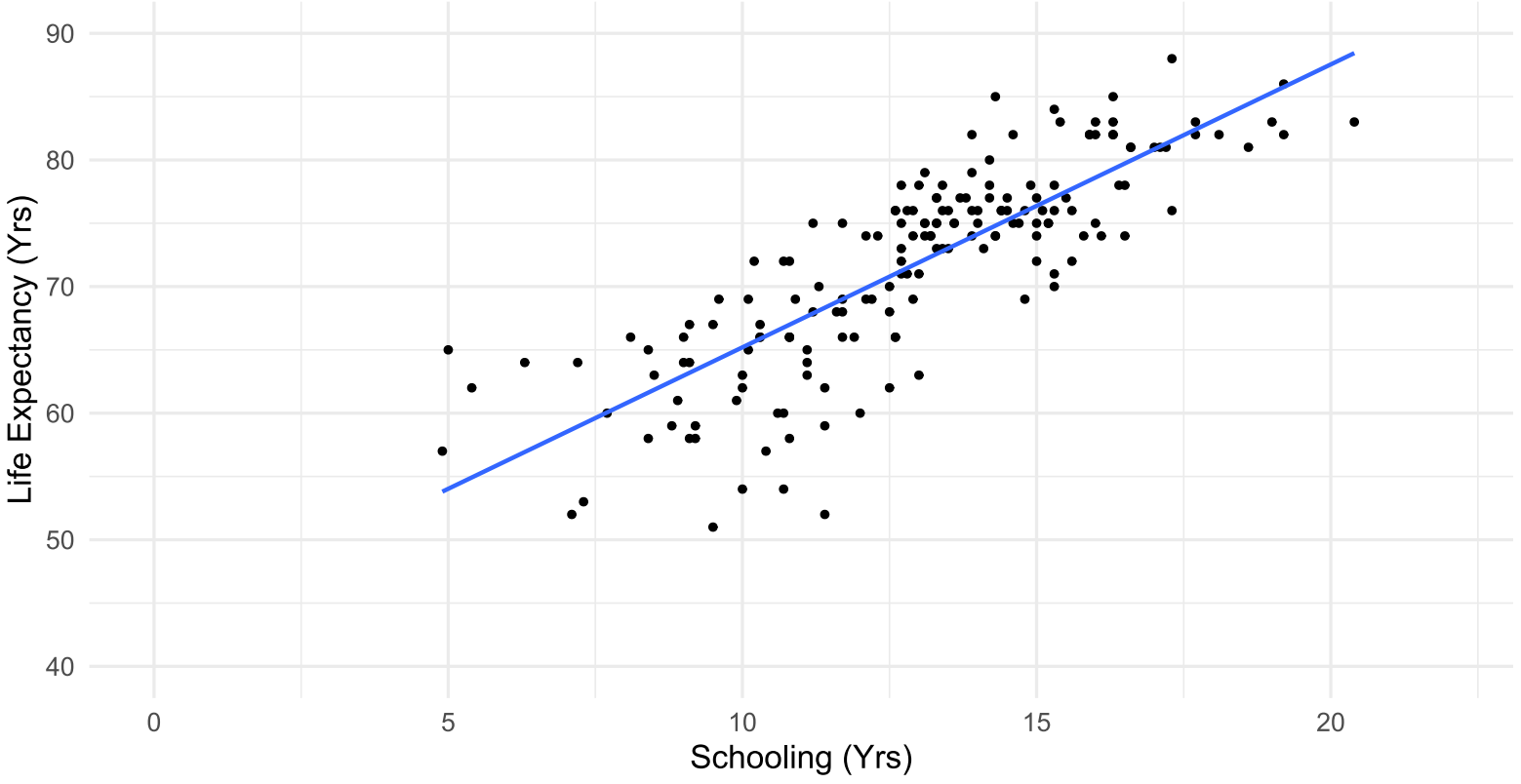
```
biv ← ggplot(data = who, aes(x = schooling, y = life_expectancy)) +
        geom_point()
```



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

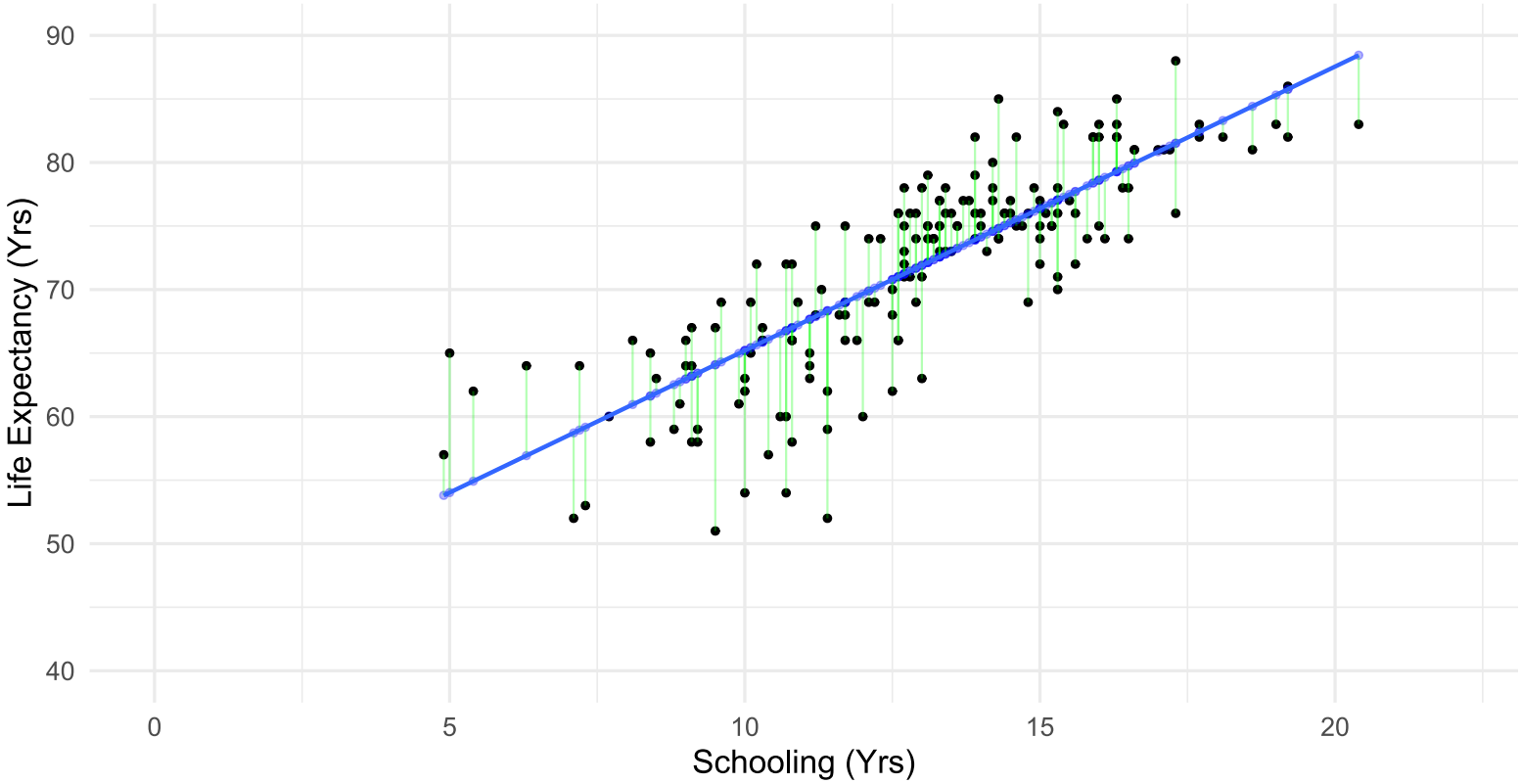
OLS-fitted regression line

```
biv + geom_smooth(method = lm, se = F)
```



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

Sums of Squares Visualization

Intercept:

0

Slope:

1

View Sums of Squares

- (•) Normal View
- ( ) View Residuals
- ( ) View Sums of Squares

Data Simulation

Mean of X:

5

Mean of Y:

5

Correlation

Error: An error has occurred. Check your logs or contact the app author for clarification.

Sums of Squares = 10.59

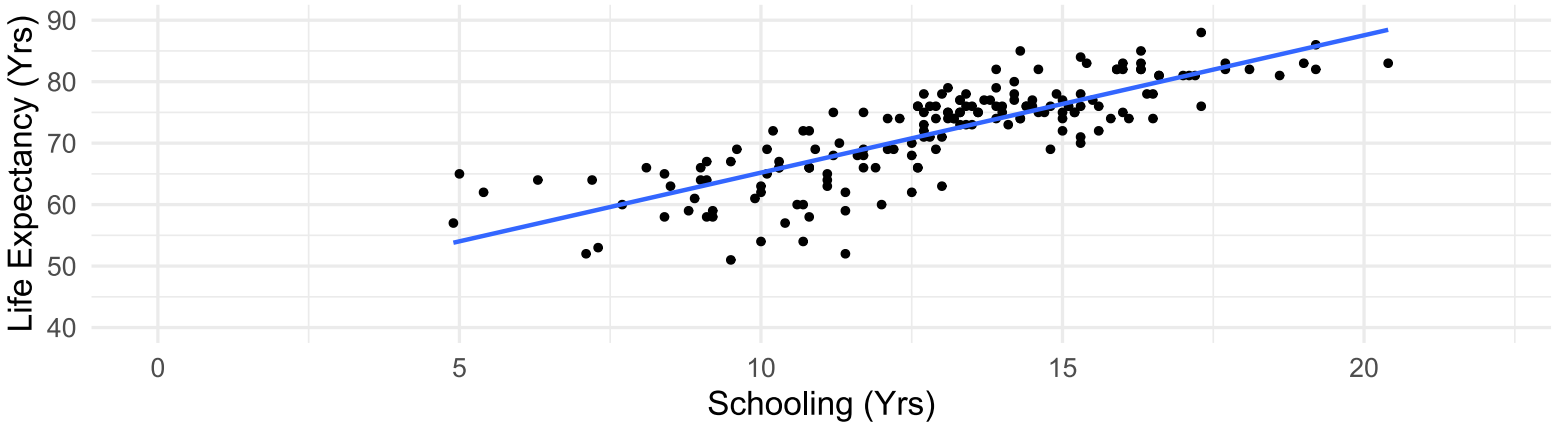
Pictures to equations

So, the Ordinary-Least Squares **line of best fit** minimizes the distance between it and all observations in the point cloud. Critically helpful to us: this line of best fit provides a two-number **summary of the relationship** between our two continuous variables.

As with any straight line, it can be characterized by a simple algebraic equation. Recall the **slope-intercept** form of a linear equation from 7th grade:

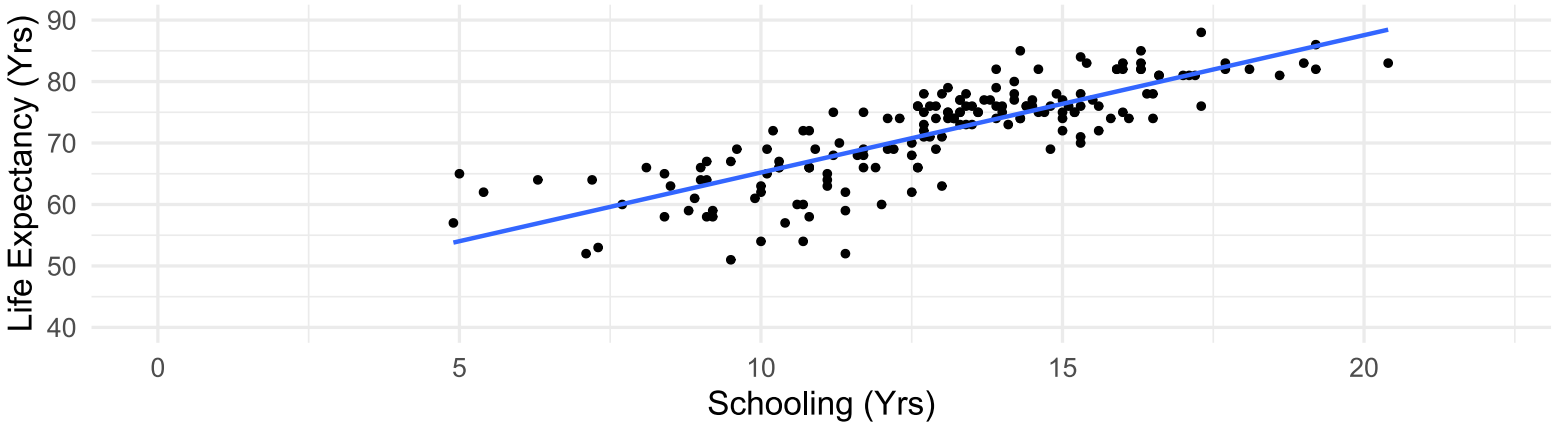
$$y = mx + b$$

What do each of these terms represent?



Pictures to equations

HOWEVER, we do not represent **lines of best fit** with equations in the **slope-intercept** form! **Why not?**



The slope-intercept form represents a deterministic relationship (*y* equals exactly *mx*+*b*). In statistics, we use the line of best fit to approximate the relationship. The line is straight ("smooth"), but there is a lot of variation ("roughness") around it, so we write this equation differently. We'll learn the formal way to represent this relationship in 643. For now, we'll use this slope-intercept form for convenience.

We can, in fact, calculate by hand the slope and the y-intercept of the line of best fit, using each (*x*, *y*) pairing for each observation. However, as you can guess, this is much more straightforward to do using a statistical software package. Turn the page to observe the wonders of our first regression fit...!

Fitting a regression 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
```

Interpreting the results

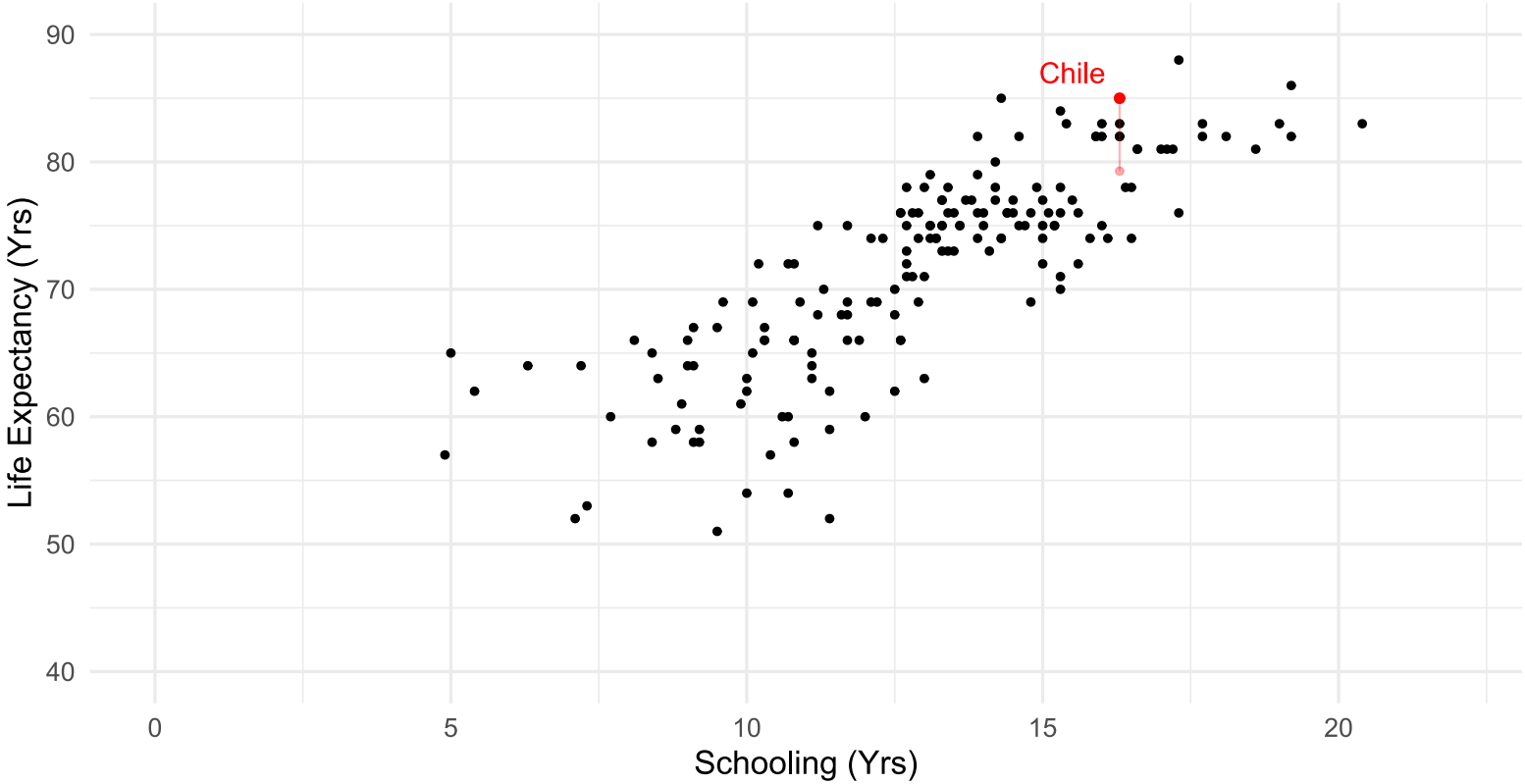
```
#>
#> Call:
#> lm(formula = life_expectancy ~ schooling, data = who)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -16.3270  -2.6565   0.1581   3.3095  10.9758
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  42.8501     1.5976   26.82   <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

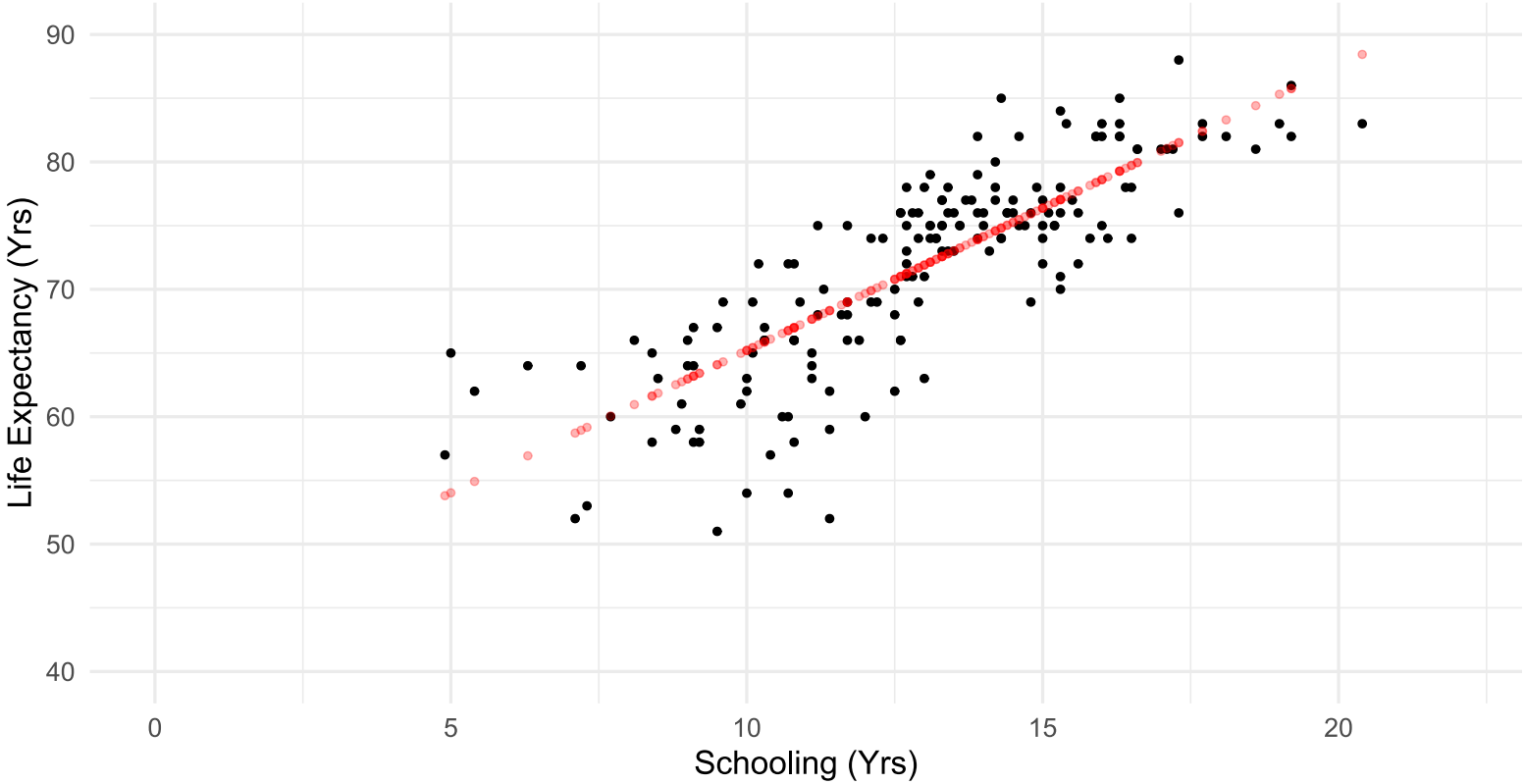
Can substitute values for the "predictor" ($SCHOOLING$) into the fitted equation to compute the *predicted* values of $LIFE\_EXPECTANCY$.



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)
```

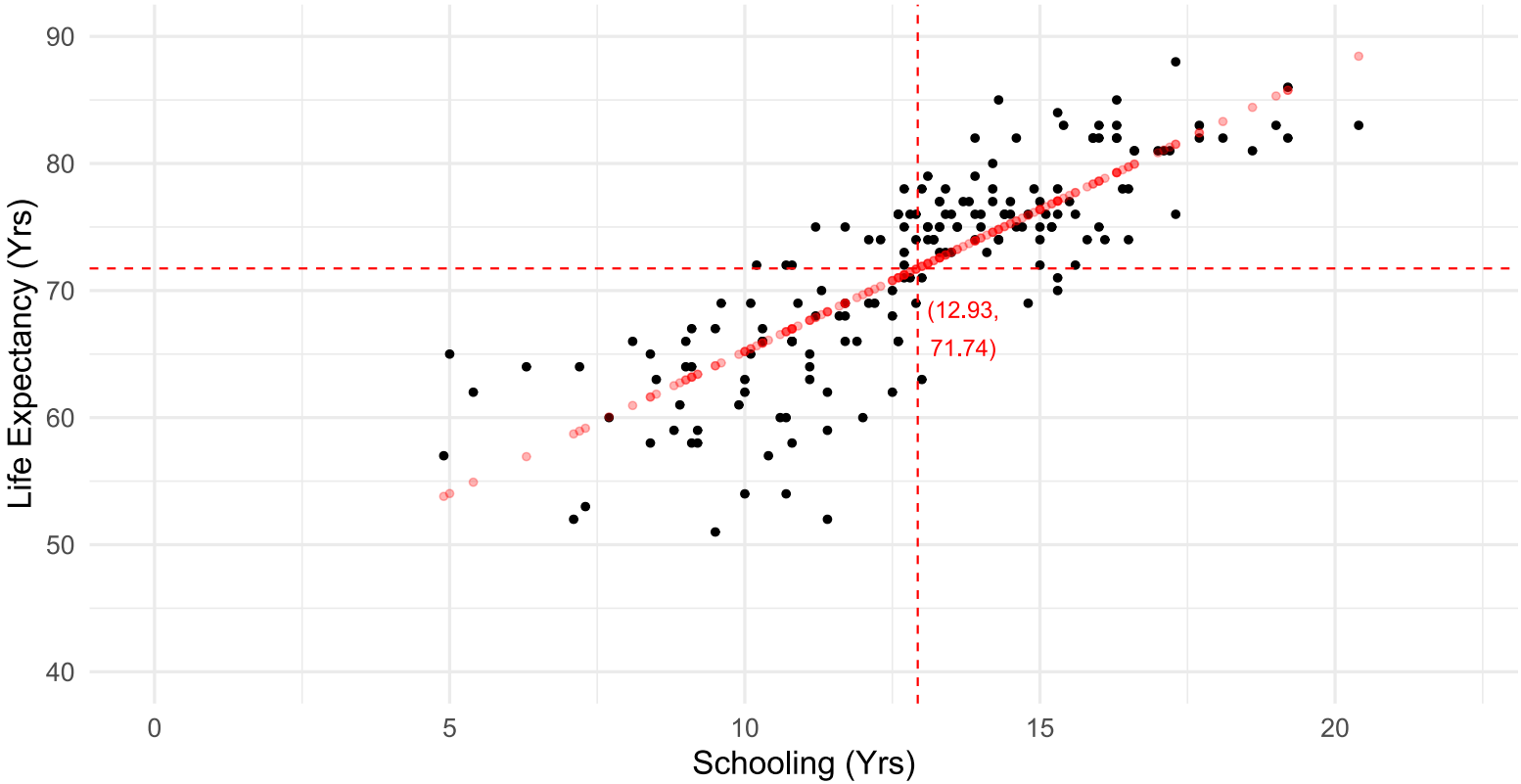
```
#> [1] 12.92717
```

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

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

Fitted values

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



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 15.1 and 15.2: bivariate regression

Assignments

Assignment #4 Due November 28 at 11:59PM