Examining Relationships of Continuous Variables

EDUC 641: Unit 4 Part 3

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 (in file called EDUC641_13_code.R)

Our continuous relationship

A reminder of our relationship

```
biv <- ggplot(data = who, aes(x = schooling, y = life_expectancy)) +
        geom_point()
```
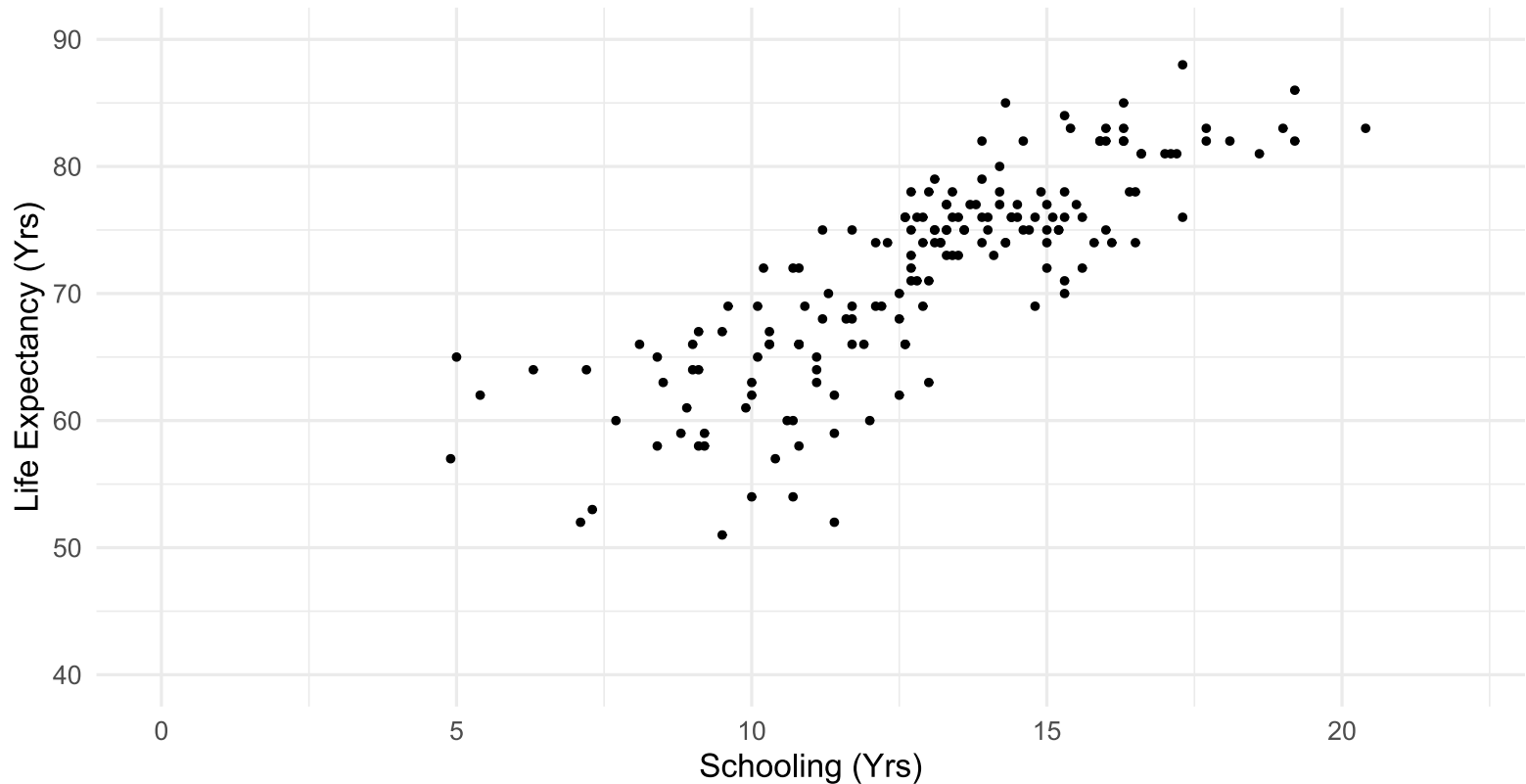
The results of our linear fit

```
#>
#> 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
```
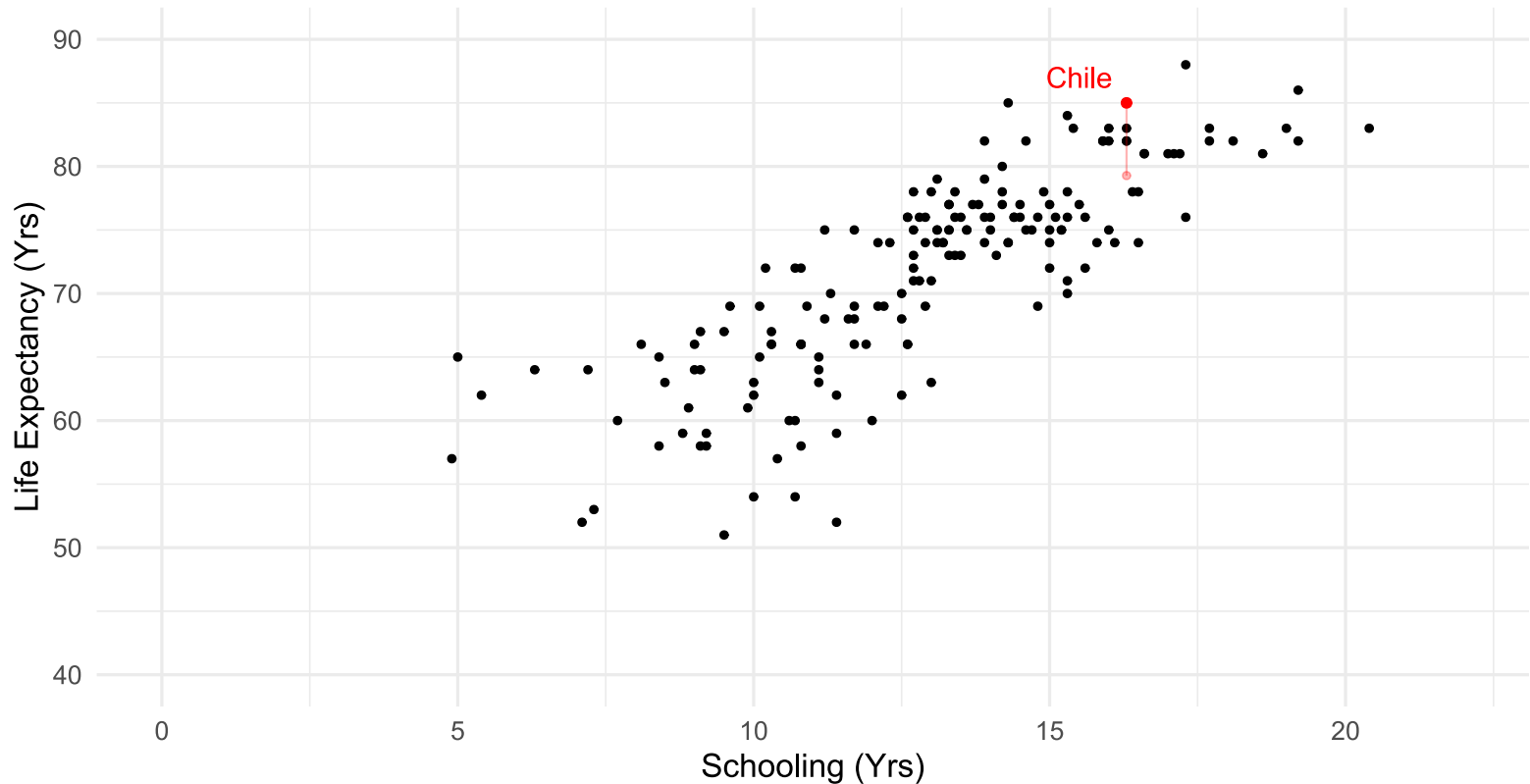
These coefficients tell you where the fitted trend line should be drawn:

$$[\text{Predicted value of } LIFE\_EXPECTANCY] = (42.85) + 2.23 * [\text{Observed value of } SCHOOLING]$$
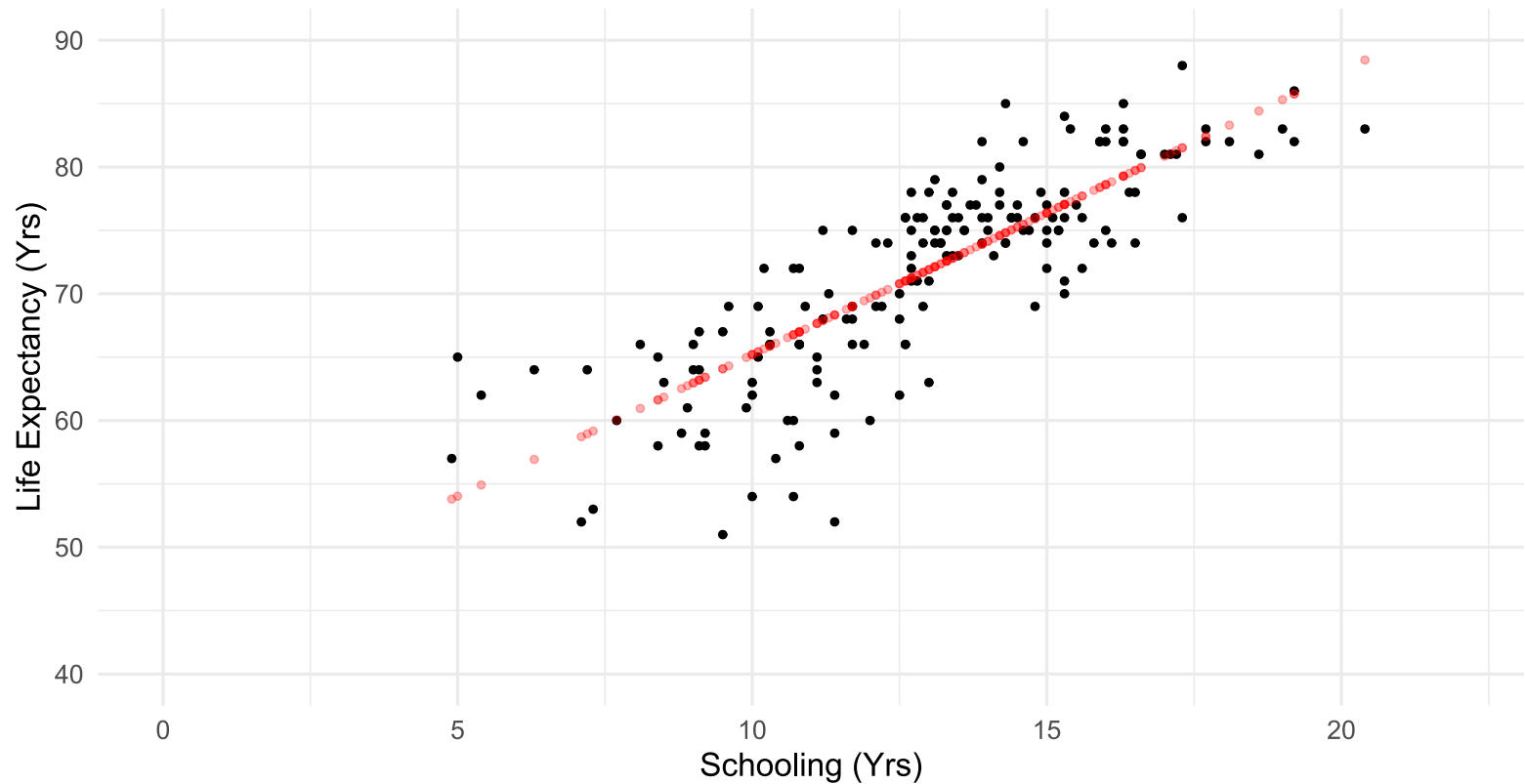
Fitted values

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

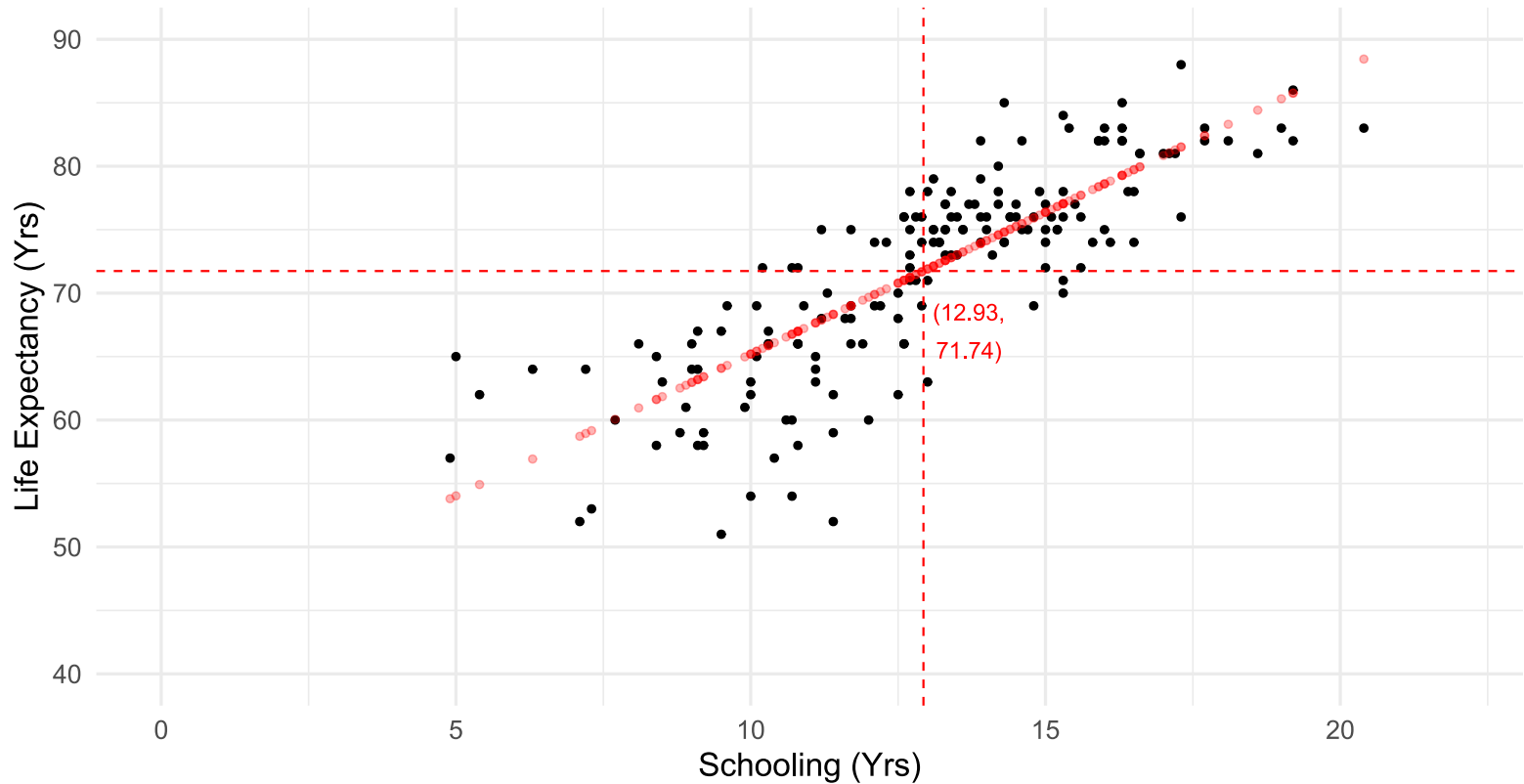Can do this for our old friend Chile ... and all others...

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

The regression equation

Each term in the regression equation has a specific interpretation

$$LIFE\_EXP\hat{E}CTANCY = 42.85 + 2.23 * (SCHOOLING)$$

The regression equation

Each term in the regression equation has a specific interpretation:

$$LIFE\_EXP\hat{E}CTANCY = 42.85 + 2.23 * (SCHOOLING)$$

The predicted value of $LIFE\_EXP\hat{E}CTANCY$ 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:

$$LIFE\_EXP\hat{E}CTANCY = \textcolor{red}{42.85} + 2.23 * (SCHOOLING)$$

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:

$$LIFE\_EXP\hat{E}CTANCY = 42.85 + \textcolor{red}{2.23} * (SCHOOLING)$$

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 where children, on average, experience one additional year of schooling have an average life expectancy of 2.23 more years.

We do **NOT** say that increasing the average years that children attend school by one year increases average life expectancy in that country by 2.23 years. Why?

The regression equation

Each term in the regression equation has a specific interpretation:

$$LIFE\_EXP\hat{E}CTANCY = 42.85 + 2.23 * (SCHOOLING)$$

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

Regression inference

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 of the relationship between $LIFE\_EXPECTANCY$ and $SCHOOLING$ was zero?

Regression inference

What is the probability that we would have gotten a slope value of 2.23 (or a more extreme value) by sampling from a population in which there was **no relationship** between $LIFE\_EXPECTANCY$ and $SCHOOLING$?

```
...
#>
#> 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
...
```

As with our previous analysis, R provides us with a *p*–value which can help us to judge the likelihood that our results are driven by idiosyncrasies of sampling.

Regression inference

```
...
#>
#> 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 $SCHOOLING$, on average in the population.

Writing it up
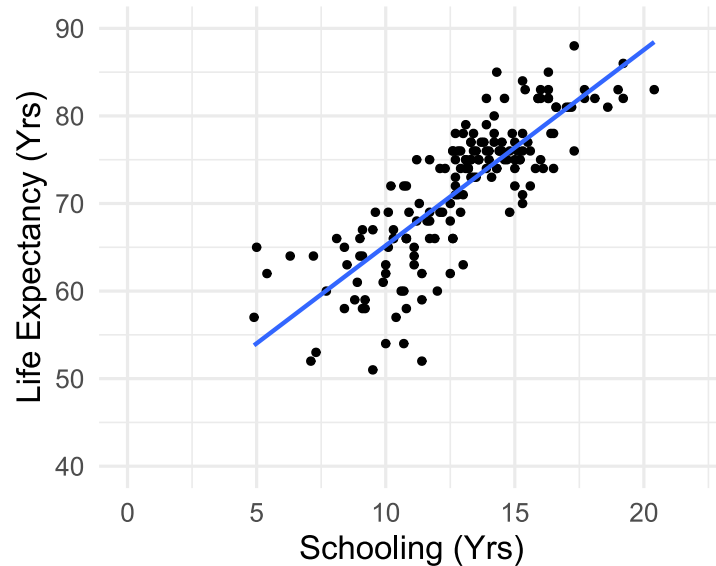
In our investigation of country-level aggregate measures of schooling and life expectancy, we have found that the average years of schooling in a country is related to the average life expectancy. In particular, when we relate the country-level life expectancy (*LIFE_EXPECTANCY*) to the country-level 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 implies 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...

Reporting results

Descriptive statistics

What do you want people to know about the nature of the variables in your data?

Descriptive statistics

What do you want people to know about the nature of the variables in your data?

Things people should probably know:

- Number of observations (N)
- Mean of continuous variables
- Measure of variance of continuous variables (probably *SD*)
- Count/proportion of values for categorical variables

Things people might need to know:

- Min/max values
- Median value
- IQR
- Missing %

Things people (probably) **don't** need to know: number of unique values, summary stats on ID variables, ?

(Re)producing beautiful results

A descriptive table (Table 1):

```
library(modelsummary)
datasummary_skim(who1)
```

	Unique	Missing Pct.	Mean	SD	Min	Median	Max
schooling	89	0	12.9	2.9	4.9	13.1	20.4
life_expectancy	35	0	71.7	8.0	51.0	74.0	88.0

(Re)producing beautiful results

A descriptive table (Table 1):

```
names(who1) <- c("Region", "Status", "Schooling (Yrs)",
                                "Life Expectancy (Yrs)")
datasummary_skim(who1,
            fun_numeric = list(Mean = Mean, SD = SD, Min = Min,
                                Median = Median, Max = Max))
```

	Mean	SD	Min	Median	Max
Schooling (Yrs)	12.9	2.9	4.9	13.1	20.4
Life Expectancy (Yrs)	71.7	8.0	51.0	74.0	88.0

(Re)producing beautiful results

A descriptive table (Table 1):

Saving it to a Word table:

```
datasummary_skim(who1,
                 fun_numeric = list(Mean = Mean, SD = SD, Min = Min,
                                    Median = Median, Max = Max),
                 output = "table.docx")
```

> Need to `install.packages("pandoc")` first

(Re)producing beautiful results

A descriptive table (Table 1):

Numeric variables by a categorical variable

```
# Tell R to cut by a given variable
datasummary_balance(`Schooling (Yrs)`+ `Life Expectancy (Yrs)` ~ Status,
                    data = who1)
```

	Developed (N=29)		Developing (N=144)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Schooling (Yrs)	16.5	1.6	12.2	2.5	−4.3	0.4
Life Expectancy (Yrs)	80.9	3.6	69.9	7.3	−11.0	0.9

Can you imagine when this might be an especially useful set of descriptive statistics to produce?

(Re)producing beautiful results

A descriptive table (Table 1):

Numeric variables by a categorical variable

```
datasummary_balance(`Schooling (Yrs)`+ `Life Expectancy (Yrs)` ~ Status,
                    dinm = F, # drop the diff-in-means
                    data = who1)
```

	Developed (N=29)		Developing (N=144)	
	Mean	Std. Dev.	Mean	Std. Dev.
Schooling (Yrs)	16.5	1.6	12.2	2.5
Life Expectancy (Yrs)	80.9	3.6	69.9	7.3

(Re)producing beautiful results

A regression output table (Table 2)

```
modelsummary(fit)
```

	(1)
(Intercept)	42.850
	(1.598)
schooling	2.235
	(0.121)
Num.Obs.	173
R2	0.668
R2 Adj.	0.666
AIC	1023.4
BIC	1032.8
Log.Lik.	−508.687
RMSE	4.58

(Re)producing beautiful results

Based on what you know so far, what do people need to know about your regression results?

(Re)producing beautiful results

Based on what you know so far, what do people need to know about your regression results?

People should know:

- Estimate of the intercept and coefficient(s)
- Uncertainty in estimates of the intercept and coefficient(s)
- Number of observations
- R^2 (we'll learn about this later)

Convention in most outlets to provide asterisks denoting conventional alpha thresholds (debatably helpful).

# (Re)producing beautiful results

A regression output table (Table 2)

```
modelsummary(fit, stars=T,
             gof_omit = "Adj.|AIC|BIC|RMSE|Log",
             coef_rename = c("schooling" = "Yrs. Schooling"))
```

	(1)
(Intercept)	42.850***
	(1.598)
Yrs. Schooling	2.235***
	(0.121)
Num.Obs.	173
R2	0.668
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

(Re)producing beautiful results

A regression output table (Table 2)

Saving it to a Word table:

```
modelsummary(fit,
             stars=T,
             gof_omit = "Adj.|AIC|BIC|RMSE|Log",
             coef_rename = c("schooling" = "Yrs. Schooling"),
             output = "table2.docx")
```
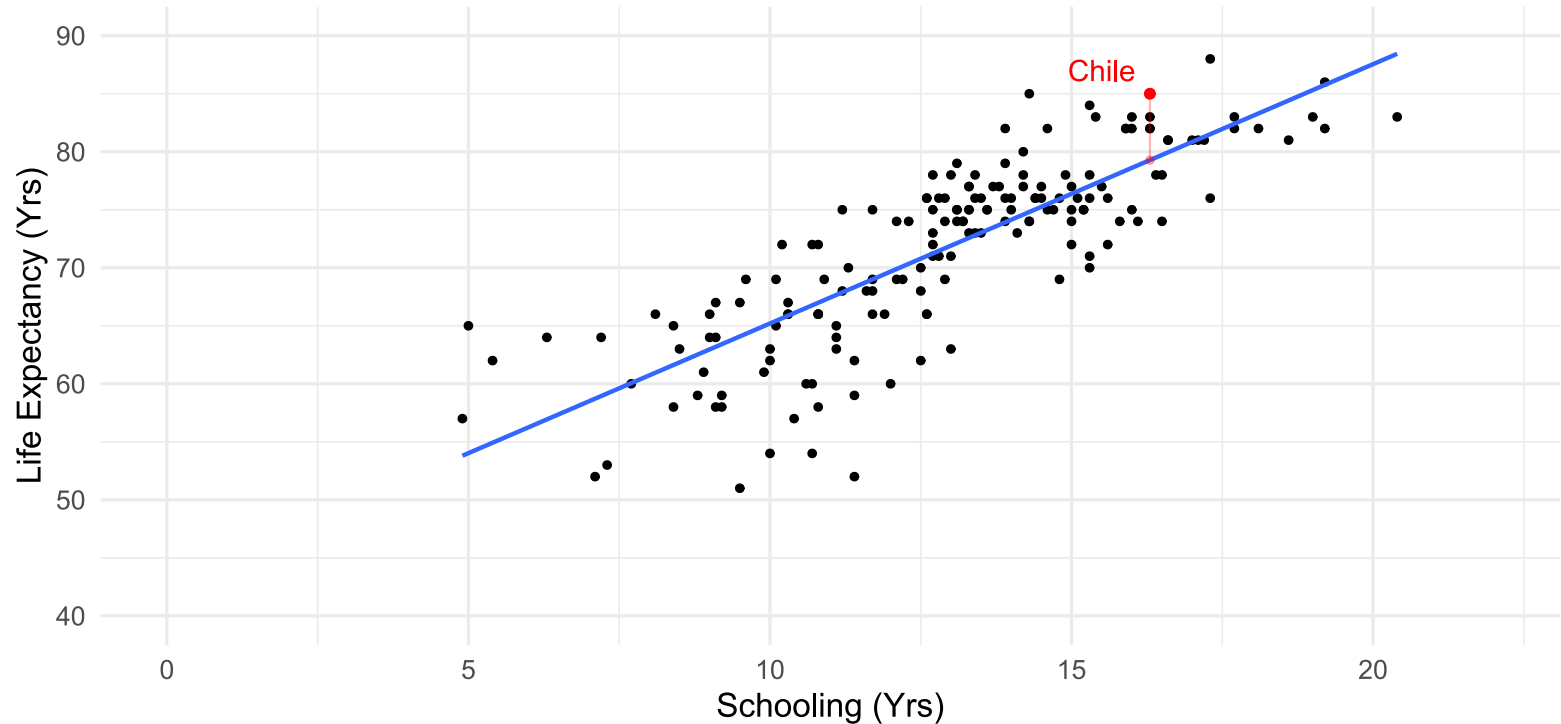
A gentle introduction to bivariate regression:

Residual analysis

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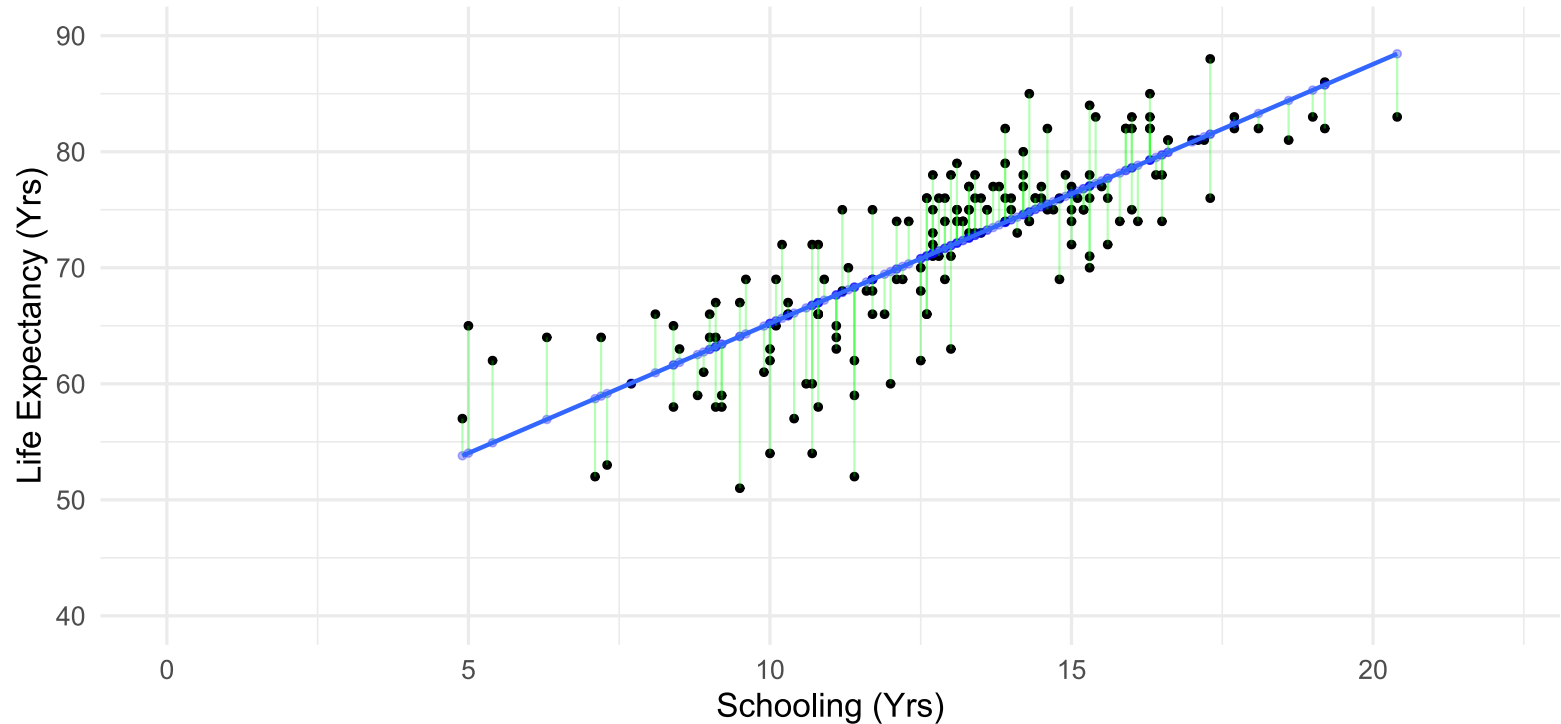
Our fitted regression line contains the "predicted" values of *LIFE_EXPECTANCY* for each value of *SCHOOLING*. But almost all of the "actual" values of *LIFE_EXPECTANCY* lie off the actual line regression line.

An example: Chile

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

$$LIFE\_EXP\hat{E}CTANCY = 42.85 + 2.23 * (16.3)$$
$$= 79.20$$

An example: Chile
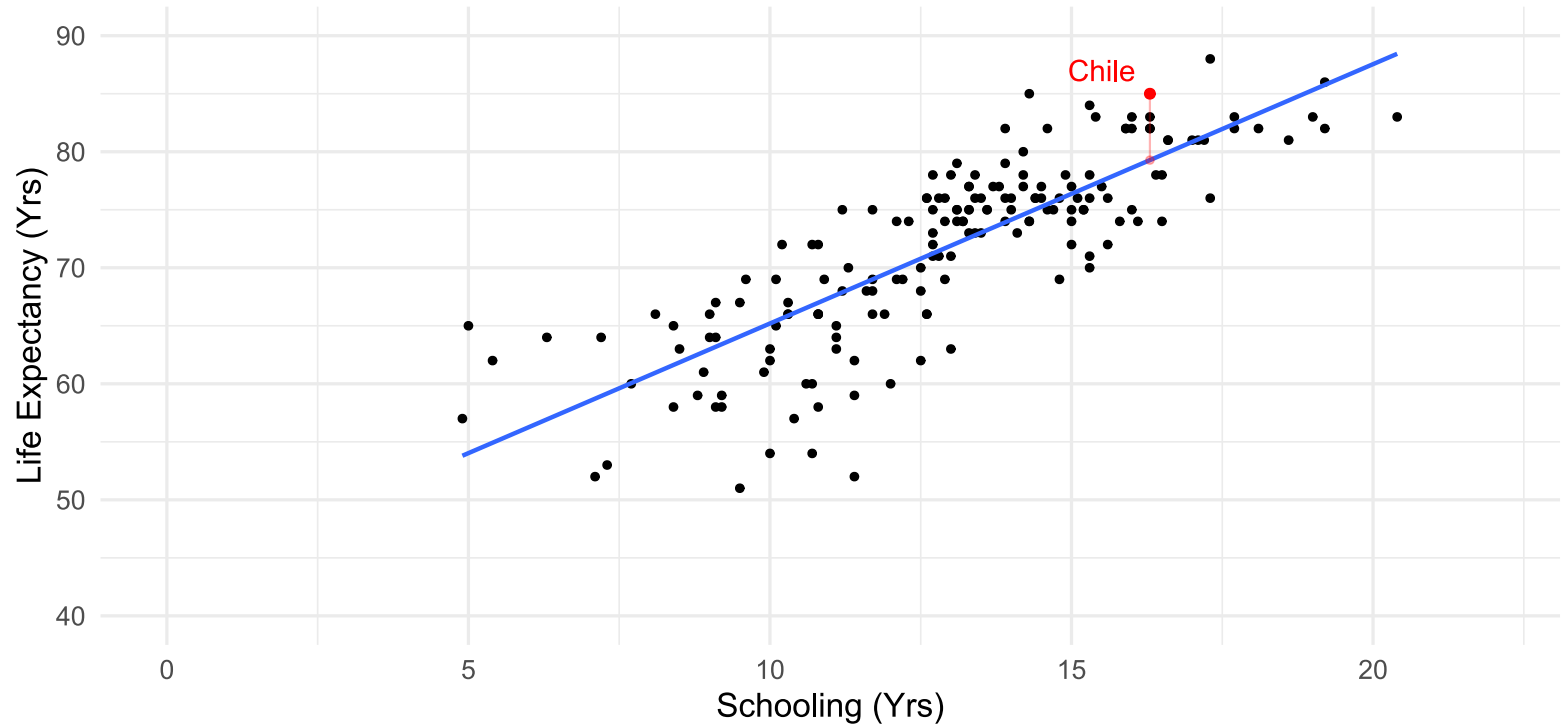
$$LIFE\_EXP\hat{E}CTANCY = 79.20$$
Actual life expectancy = 85

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

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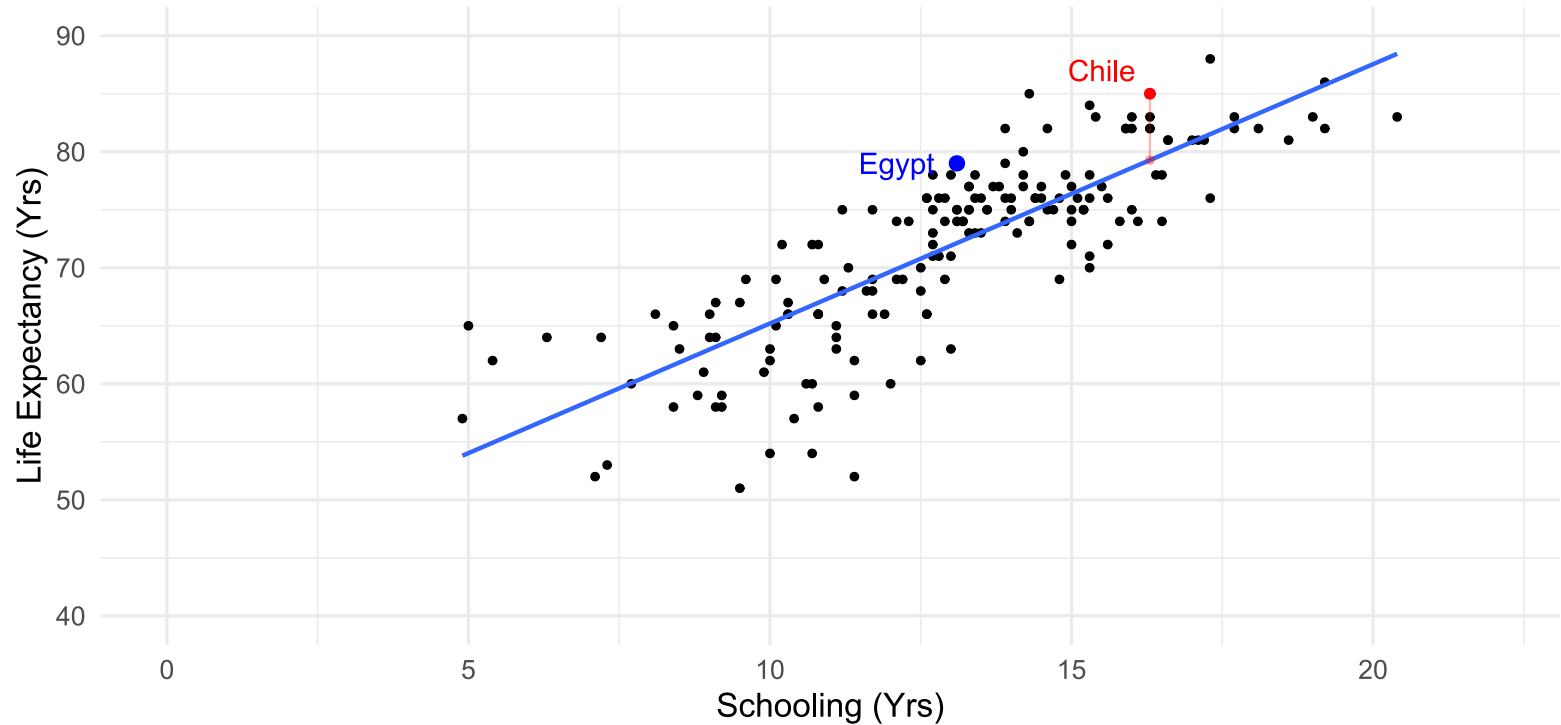
Observed values for Egypt: $LIFE\_EXPECTANCY = 79; SCHOOLING = 13.1$

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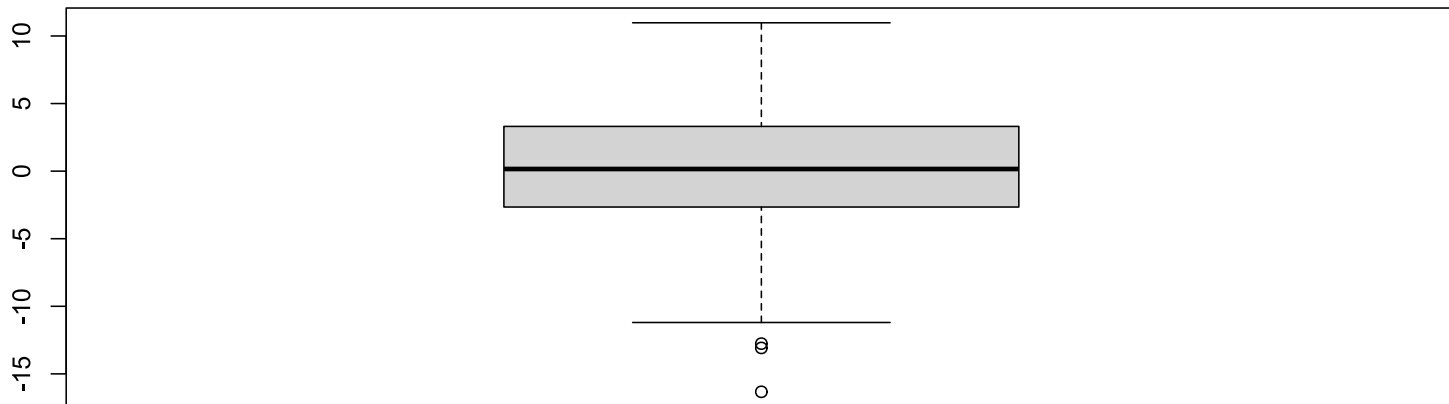
```
sd(who$resid)
```

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

- Standard deviation of the raw residuals can be quite useful in examining the quality of our fit. How?

Residual assumptions

For the *p*-values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**
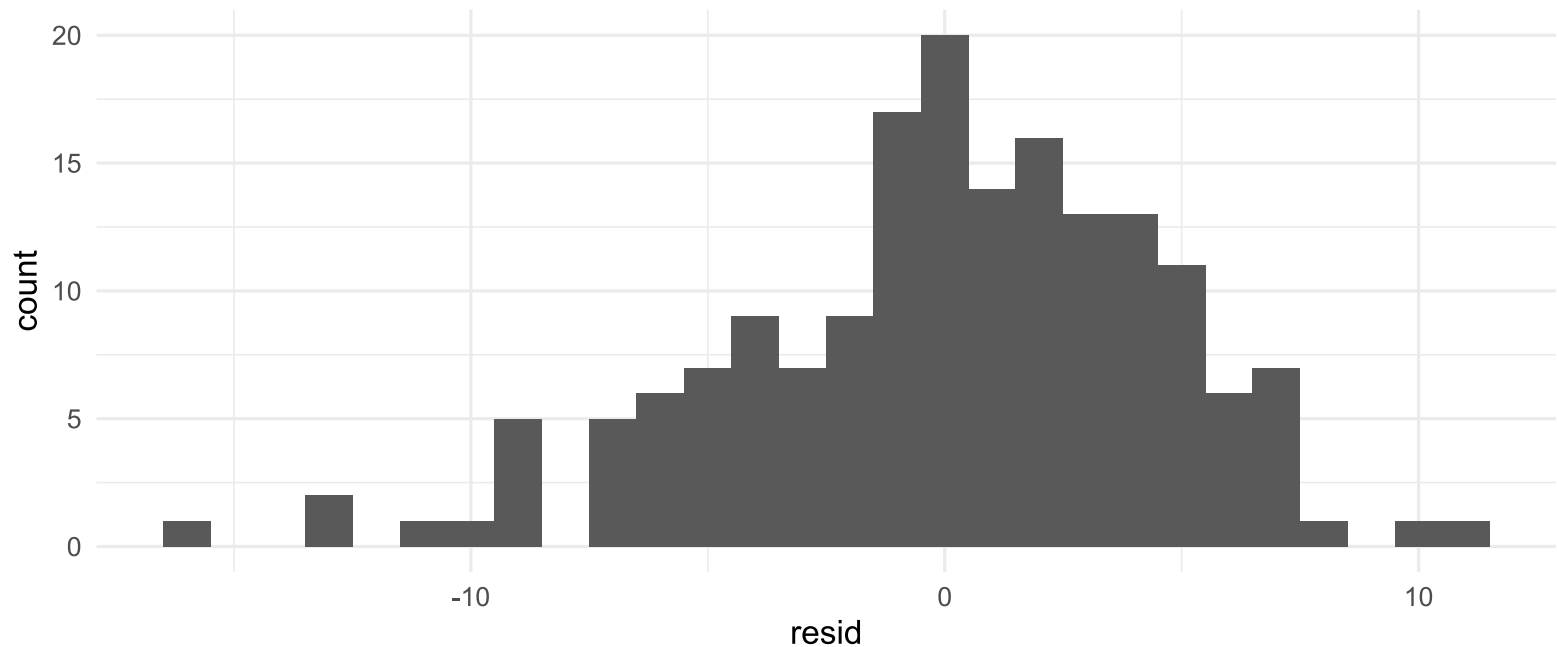
```
boxplot(resid(fit))
```


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

Residual assumptions

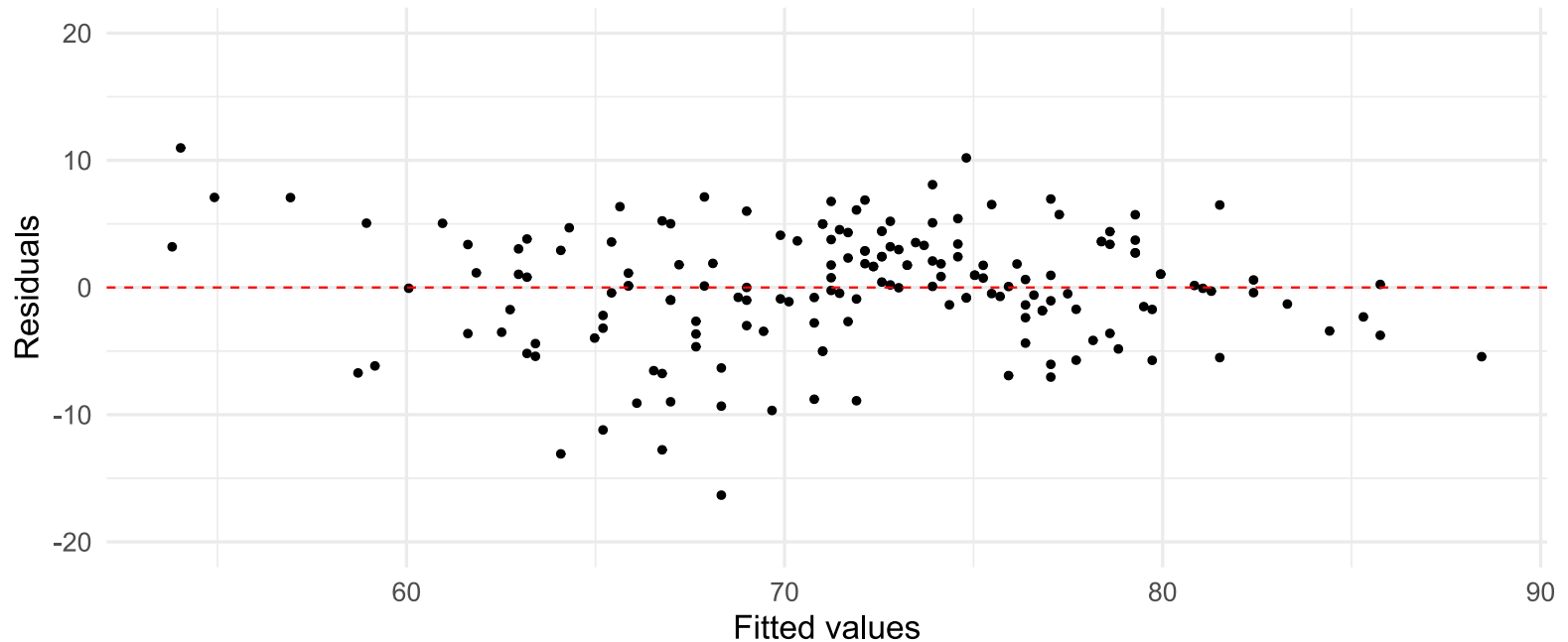For the *p*–values that we computed in the regression analysis to be correct, the residuals **must be normally distributed**[1]

Pretty good, pretty good... **Understanding check:** can you write out the code to create the above figure?

---

[1] We have solutions if they are not which we will learn about in EDUC 643.

Residual vs. fitted plot

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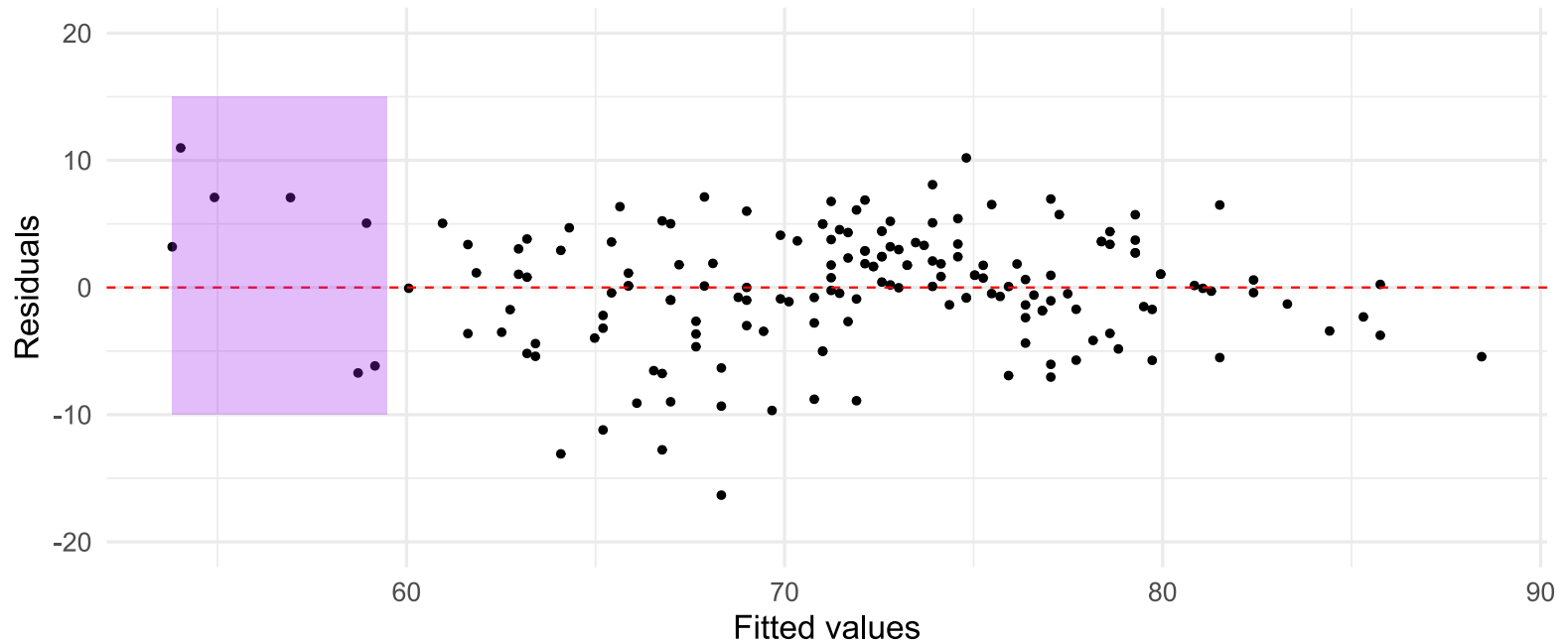

Key assumption checks for normality:

- 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.

Implementing residual v. fitted

```r
ggplot(who, aes(x = predict, y = resid)) +
        geom_point() +
  geom_hline(yintercept = 0, color = "red", linetype="dashed") +
  ylab("Residuals") + xlab("Fitted values") +
  scale_y_continuous(limits=c(-20, 20)) +
  theme_minimal(base_size = 16)
```
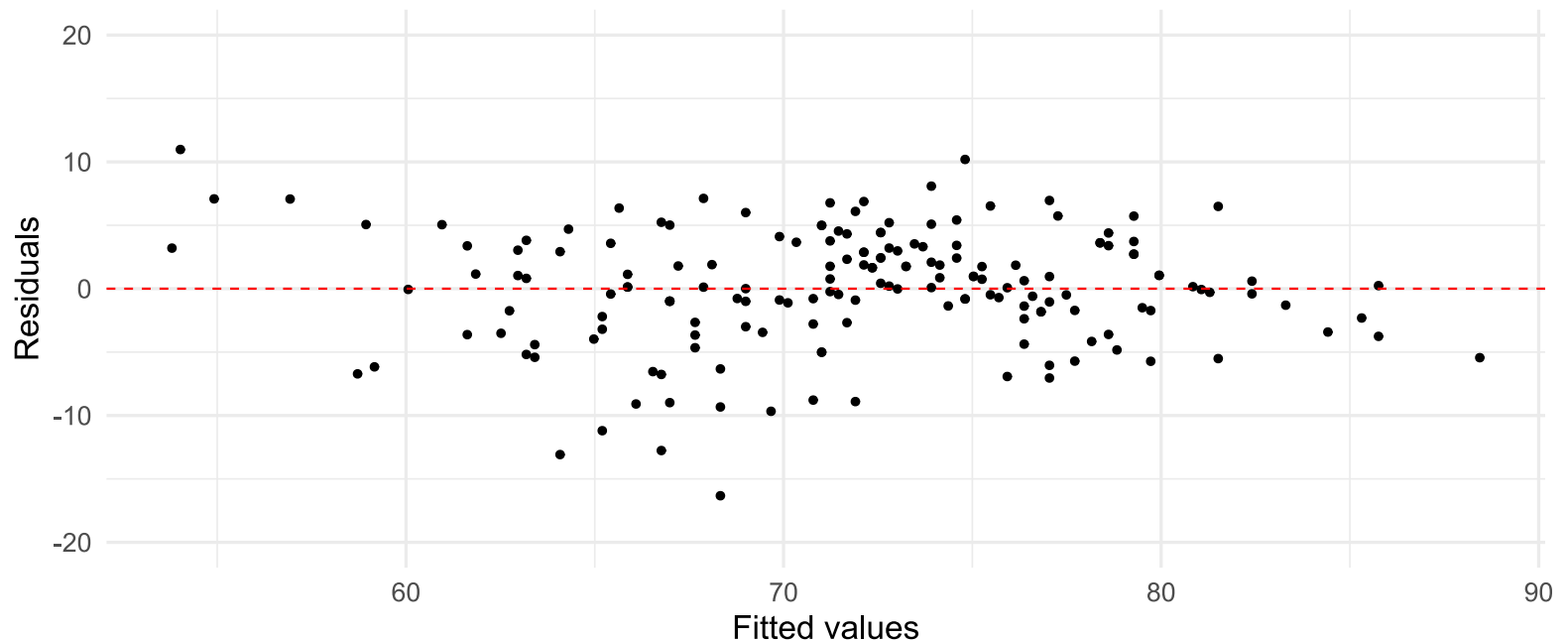
Writing it up

In our investigation of country-level measures of schooling and life expectancy, we found that the average years of schooling in a country is related to the average life expectancy. As we show in Table 2, when we relate the country-level life expectancy to the country-level mean years of 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.

Table 2. Estimates of relationship between life expectancy and schooling

	(1)
(Intercept)	42.850***
	(1.598)
Yrs. Schooling	2.235***
	(0.121)
Num.Obs.	173
R2	0.668
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

Writing it up II

> 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. As we note in Appendix Figure A1, our fitted regression does seem to underpredict life expectancy for very low levels of schooling.

Figure A1. Residuals vs. fitted plot of life expectancy and schooling
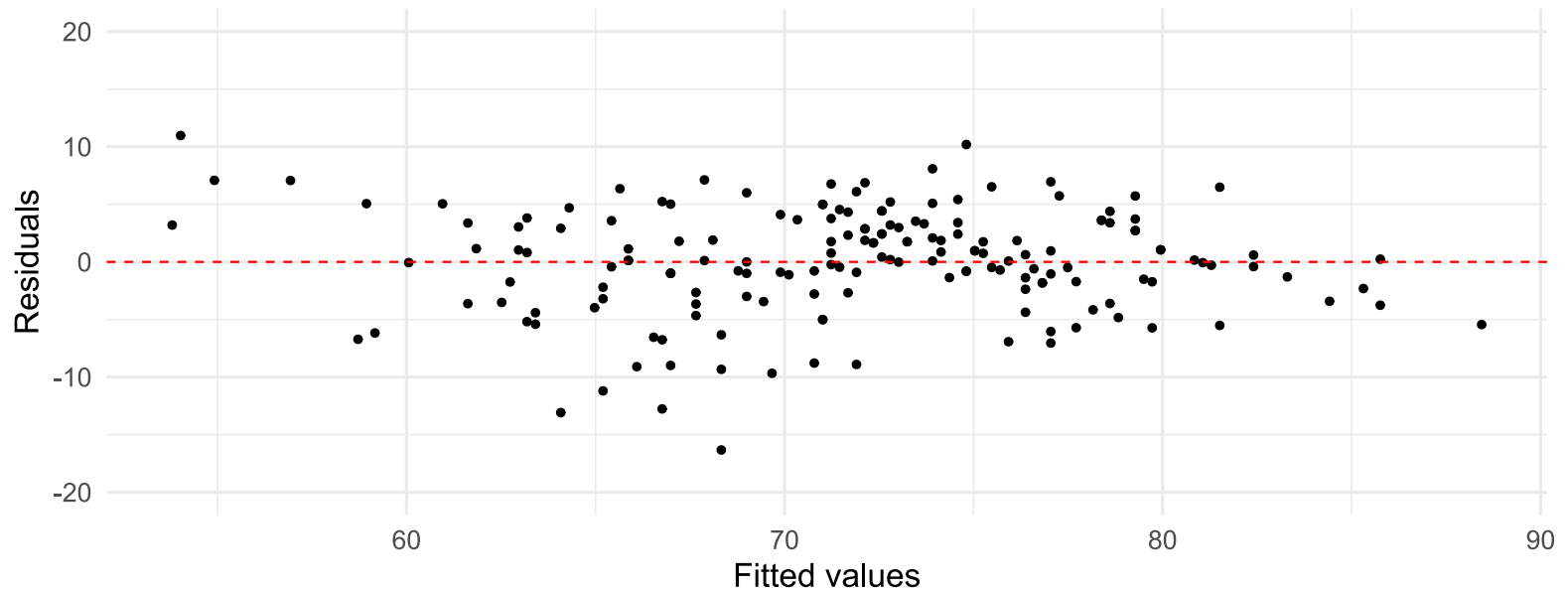
Key takeaways

- **Start with a RQ which you can answer in your data**
- **Understand your data first**
  - Summarize and visualize each variable independently
  - Start with a visual representation of the relationship between your 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

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 by **Nov. 21**

Assignments

- Quiz #5 due **November 27** at 5pm
- Assignment #4 due **December 2** at 11:59PM
- Final assignment due **December 11** at 4:59PM

No lab on 11/27 or 11/28! Lab replaces class on 11/26!