Lesson 6: SLR: More inference

Nicky Wakim

2025-01-27

Learning Objectives

1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4. Calculate and interpret the coefficient of determination

So far in our regression example...

Lesson 3: SLR 1

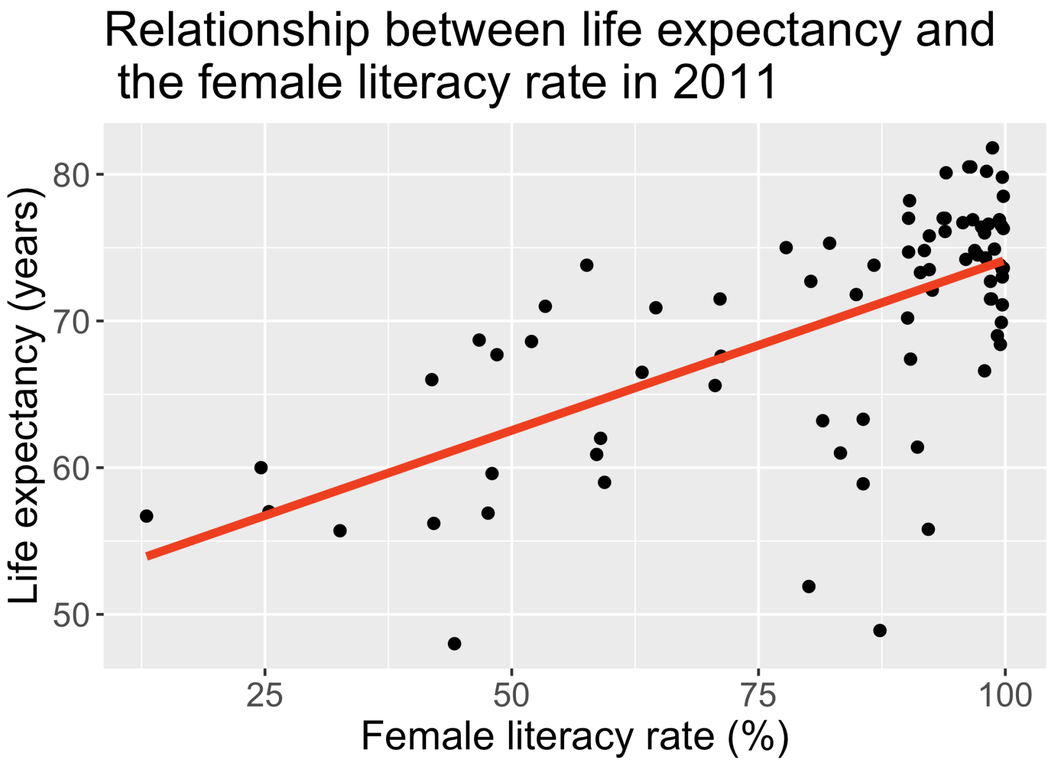
- Fit regression line
- Calculate slope & intercept
- Interpret slope & intercept

Lesson 4: SLR 2

- Estimate variance of the residuals
- Inference for slope & intercept: CI, p-value
- Confidence bands of regression line for mean value of Y|X

Lesson 5: Categorical Covariates

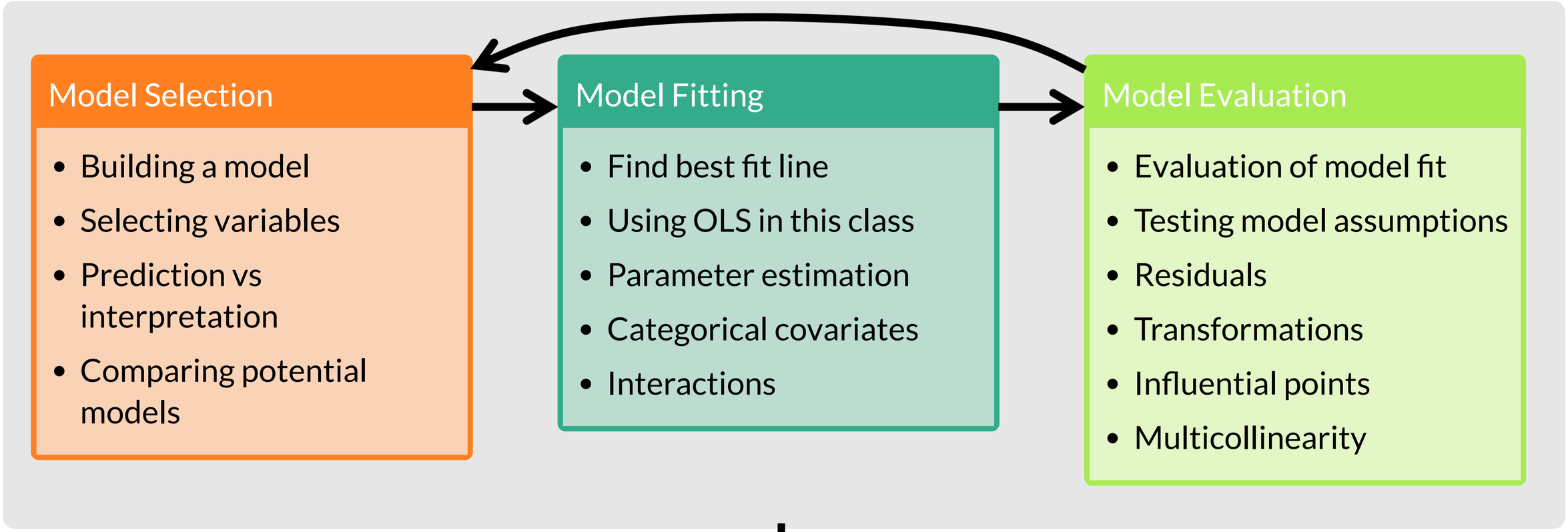
- Inference for different categories



$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot X$$

$$\widehat{\text{LE}} = 50.9 + 0.232 \cdot \text{FLR}$$

Process of regression data analysis



Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Learning Objectives

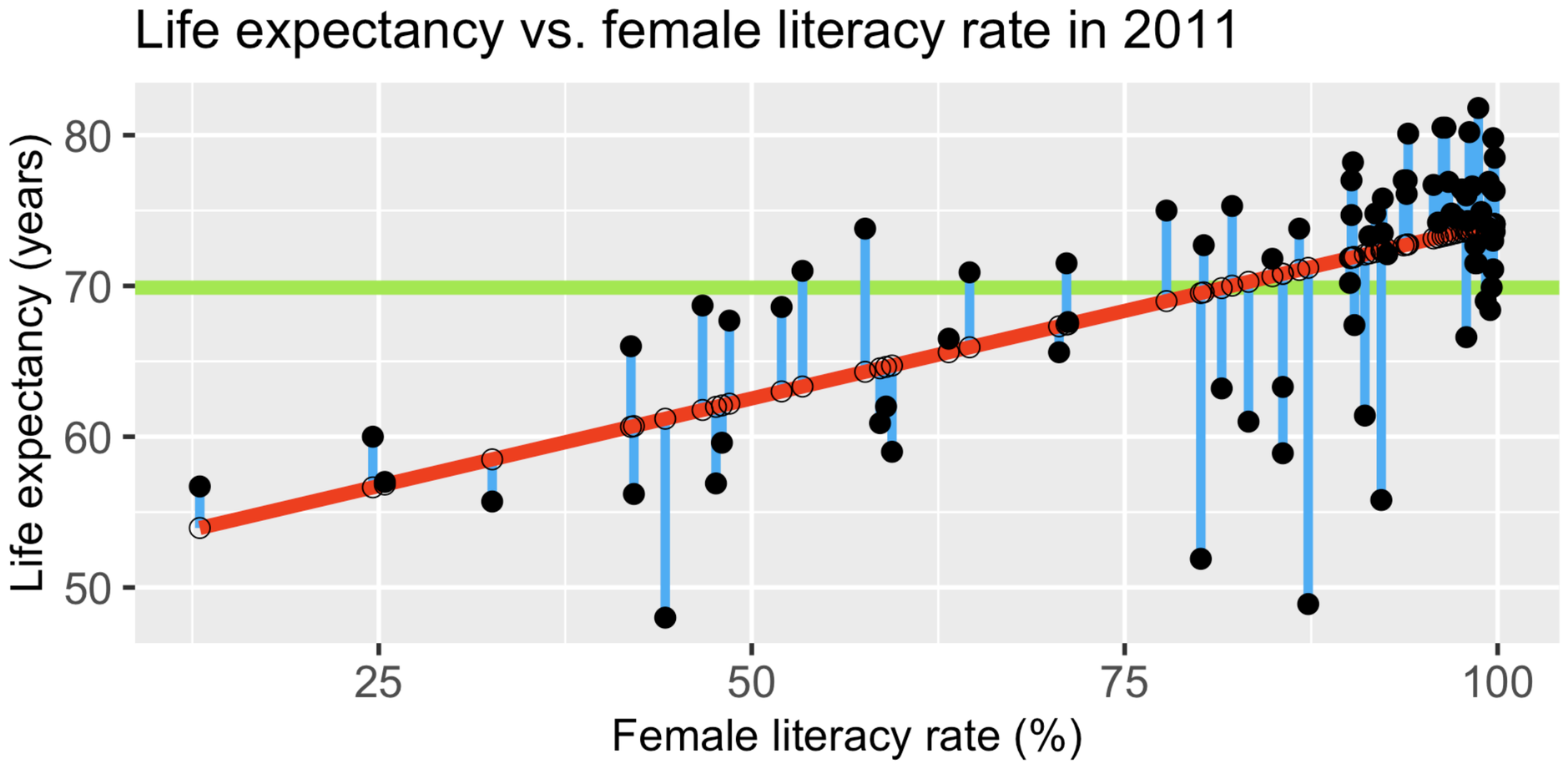
1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4. Calculate and interpret the coefficient of determination

Getting to the F-test

The F statistic in linear regression is essentially a proportion of the variance explained by the model vs. the variance not explained by the model

1. Start with visual of explained vs. unexplained variation
2. Figure out the mathematical representations of this variation
3. Look at the ANOVA table to establish key values measuring our variance from our model
4. Build the F-test

Explained vs. Unexplained Variation



$$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y})$$

$$\text{Total unexplained variation} = \text{Residual variation after regression} + \text{Variation explained by regression}$$

More on the equation

$$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y})$$

- $Y_i - \overline{Y}$ = the deviation of Y_i around the mean $\overline{Y}$
  - (the **total** amount deviation unexplained at X_i).
- $Y_i - \widehat{Y}_i$ = the deviation of the observation Y around the fitted regression line
  - (the amount deviation **unexplained** by the regression at X_i).
- $\widehat{Y_i} - \overline{Y}$ = the deviation of the fitted value $\widehat{Y}_i$ around the mean $\overline{Y}$
  - (the amount deviation **explained** by the regression at X_i)

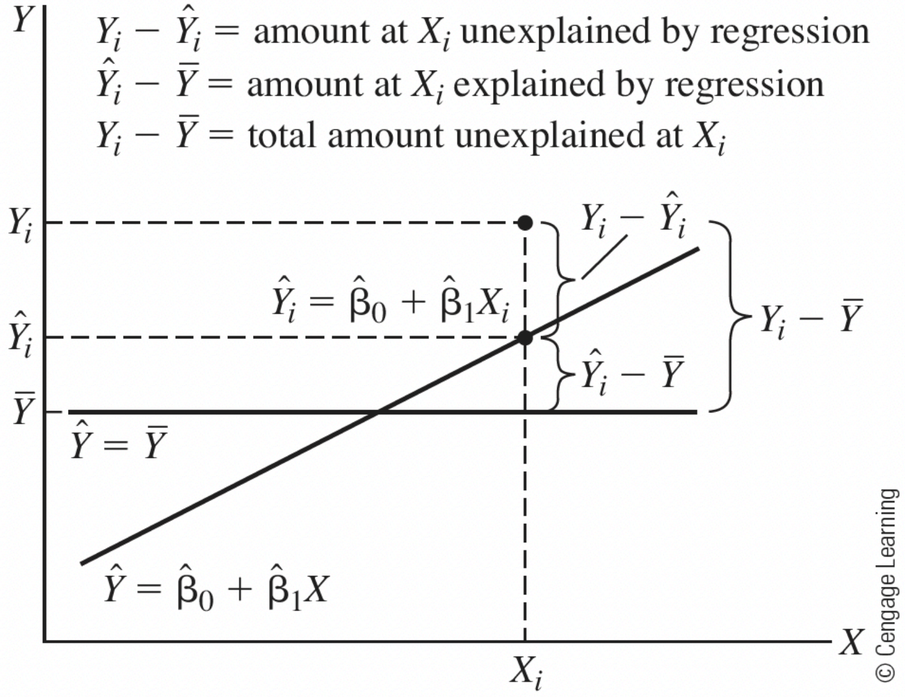
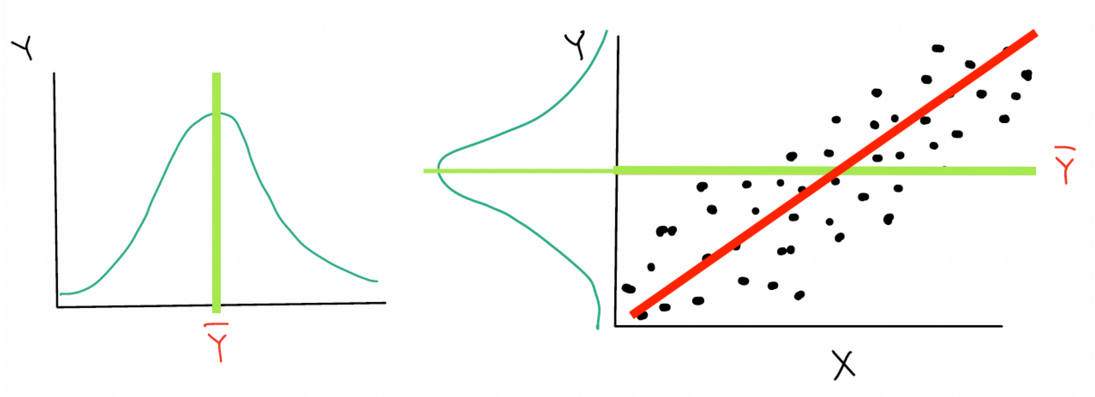


FIGURE 7.1 Variation explained and unexplained by straight-line regression

Another way of thinking about the different deviations



Poll Everywhere Question 1

How is this actually calculated for our fitted model? (1/2)

$$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y})$$

$$\text{Total unexplained variation} = \text{Variation due to regression} + \text{Residual variation after regression}$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

$$SSY = SSR + SSE$$

$$\text{Total Sum of Squares} = \text{Sum of Squares due to Regression} + \text{Sum of Squares due to Error (residuals)}$$

ANOVA table:

Variation Source	df	SS	MS	test statistic	p-value
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	
Error	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$		
Total	$n-1$	SSY			

How is this actually calculated for our fitted model? (2/2)

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2+\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2$$

$$SSY = SSR + SSE$$

$$\text{Total Sum of Squares} = \text{Sum of Squares due to Regression} + \text{Sum of Squares due to Error (residuals)}$$

ANOVA table:

Variation Source	df	SS	MS	test statistic	p-value
Regression	1	SSR	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$	
Error	$n-2$	SSE	$MSE=\frac{SSE}{n-2}$		
Total	$n-1$	SSY			

F-statistic: Proportion of variation that is explained by the model to variation not explained by the model

Analysis of Variance (ANOVA) table in R

```r
# Fit regression model:
model1 <- gapm %>% lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate)

anova(model1)
```

```
Analysis of Variance Table

Response: LifeExpectancyYrs
                    Df Sum Sq Mean Sq F value    Pr(>F)
FemaleLiteracyRate   1 2052.8 2052.81  54.414 1.501e-10 ***
Residuals           78 2942.6   37.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model1) %>% tidy() %>% gt() %>%
    tab_options(table.font.size = 40) %>%
    fmt_number(decimals = 3)
```

term	df	sumsq	meansq	statistic	p.value
FemaleLiteracyRate	1.000	2,052.812	2,052.812	54.414	0.000
Residuals	78.000	2,942.635	37.726	NA	NA

Learning Objectives

1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4. Calculate and interpret the coefficient of determination

What is the F statistic testing?

$$F = \frac{MSR}{MSE}$$

- It can be shown that

$$E(MSE) = \sigma^2 \text{ and } E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Recall that σ^2 is the variance of the population residuals
- Thus if
  - $\beta_1 = 0$, then $F \approx \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2} = 1$
  - $\beta_1 \neq 0$, then $F \approx \frac{\widehat{\sigma}^2 + \widehat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2}{\widehat{\sigma}^2} > 1$
- So the F statistic can also be used to test β_1

F-test vs. t-test for the population slope

The square of a t-distribution with $df = \nu$ is an F-distribution with $df = 1, \nu$

$$T_\nu^2 \sim F_{1,\nu}$$

- We can use either F-test or t-test to run the following hypothesis test:

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_A : \beta_1 \neq 0$$

- Note that the F-test does not support one-sided alternative tests, but the t-test does!
  - F-test cannot handle alternatives like $\beta_1 > 0$ nor $\beta_2 < 0$

Planting a seed about the F-test

We can think about the hypothesis test for the slope…

Null H_0

$\beta_1 = 0$

Alternative H_1

$\beta_1 \neq 0$

in a slightly different way…

Null model ($\beta_1 = 0$)

- $Y = \beta_0 + \epsilon$
- Smaller (reduced) model

Alternative model ($\beta_1 \neq 0$)

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Larger (full) model

- In multiple linear regression, we can start using this framework to test multiple coefficient parameters at once
  - Decide whether or not to reject the smaller reduced model in favor of the larger full model
  - Cannot do this with the t-test when we have multiple coefficients!

Poll Everywhere Question 2

F-test: general steps for hypothesis test for population slope β_1

1. For today's class, we are assuming that we have met the underlying assumptions

2. State the null hypothesis.

Often, we are curious if the coefficient is 0 or not:

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_A : \beta_1 \neq 0$$

3. Specify the significance level.

Often we use $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

The test statistic is F, and follows an F-distribution with numerator $df = 1$ and denominator $df = n - 2$.

5. Compute the value of the test statistic

The calculated **test statistic** for $\widehat{\beta}_1$ is

$$F = \frac{MSR}{MSE}$$

6. Calculate the p-value

We are generally calculating: $P(F_{1,n-2} > F)$

7. Write conclusion for hypothesis test

- Reject: $P(F_{1,n-2} > F) < \alpha$

We (reject/fail to reject) the null hypothesis that the slope is 0 at the $100\alpha\%$ significiance level. There is (sufficient/insufficient) evidence that there is significant association between (Y) and (X) (p-value = $P(F_{1,n-2} > F)$).

Life expectancy example: hypothesis test for population slope β_1

- Steps 1-4 are setting up our hypothesis test: not much change from the general steps

1. For today's class, we are assuming that we have met the underlying assumptions 2. State the null hypothesis.

We are testing if the slope is 0 or not:

$$H_0 : \beta_1 = 0$$
$$\text{vs. } H_A : \beta_1 \neq 0$$

3. Specify the significance level.

Often we use $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

The test statistic is F, and follows an F-distribution with numerator $df = 1$ and denominator $df = n - 2 = 80 - 2$.

```
1 nobs(model1)
```

```
[1] 80
```

Life expectancy example: hypothesis test for population slope β_1 (2/4)

5. Compute the value of the test statistic

```
1  anova(model1) %>% tidy() %>% gt() %>%
2    tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
FemaleLiteracyRate	1	2052.812	2052.81234	54.4136	1.501286e-10
Residuals	78	2942.635	37.72609	NA	NA

- **Option 1:** Calculate the test statistic using the values in the ANOVA table

$$F = \frac{MSR}{MSE} = \frac{2052.81}{37.73} = 54.414$$

- **Option 2:** Get the test statistic value (F) from the ANOVA table

I tend to skip this step because I can do it all with step 6

Life expectancy example: hypothesis test for population slope β_1 (3/4)

6. Calculate the p-value

- As per Step 4, test statistic F can be modeled by a F-distribution with $df1 = 1$ and $df2 = n - 2$.
  - We had 80 countries' data, so $n = 80$
- **Option 1:** Use `pf()` and our calculated test statistic

```
# p-value is ALWAYS the right tail for F-test
pf(54.414, df1 = 1, df2 = 78, lower.tail = FALSE)
```

```
[1] 1.501104e-10
```

- **Option 2:** Use the ANOVA table

```
anova(model1) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
FemaleLiteracyRate	1	2052.812	2052.81234	54.4136	1.501286e-10
Residuals	78	2942.635	37.72609	NA	NA

Life expectancy example: hypothesis test for population slope β_1 (4/4)

7. Write conclusion for the hypothesis test

We reject the null hypothesis that the slope is 0 at the 5% significance level. There is sufficient evidence that there is significant association between female life expectancy and female literacy rates (p-value < 0.0001).

Did you notice anything about the p-value?

The p-value of the t-test and F-test are the same!!

- For the t-test:

```
1 tidy(model1) %>% gt() %>%
2   tab_options(table.font.size = 40)
```

term	estimate	std.error	statistic	p.value
(Intercept)	50.9278981	2.66040695	19.142898	3.325312e-31
FemaleLiteracyRate	0.2321951	0.03147744	7.376557	1.501286e-10

- For the F-test:

```
1 anova(model1) %>% tidy() %>% gt() %>%
2   tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
FemaleLiteracyRate	1	2052.812	2052.81234	54.4136	1.501286e-10
Residuals	78	2942.635	37.72609	NA	NA

This is true when we use the F-test for a single coefficient!

Learning Objectives

1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4. Calculate and interpret the coefficient of determination

Testing association between continuous outcome and categorical variable

- Before we used the F-test (or t-test) to determine association between two continuous variables
- We CANNOT use the t-test to determine association between a continuous outcome and a multi-level categorical variable
- We CAN use the F-test to do this!

- We can use the t-test or F-test for a categorical variable with only 2 levels

Poll Everywhere Question 3

Building a very important toolkit: three types of tests

Overall test (in a couple classes)

Does at least one of the covariates/predictors contribute significantly to the prediction of Y?

Test for addition of a single variable (covariate subset test)

Does the addition of one particular covariate add significantly to the prediction of Y achieved by other covariates already present in the model?

Test for addition of group of variables (covariate subset test) (in a couple classes)

Does the addition of some group of covariates add significantly to the prediction of Y achieved by other covariates already present in the model?

When running a F-test for linear models…

- We need to define a larger, full model (more parameters)
- We need to define a smaller, reduced model (fewer parameters)
- Use the F-statistic to decide whether or not we reject the smaller model
  - The F-statistic compares the SSE of each model to determine if the full model explains a significant amount of additional variance

$$F = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$$

- $SSE(R) \geq SSE(F)$
- Numerator measures difference in **unexplained** variation between the models
  - Big difference = added parameters greatly reduce the unexplained variation (increase explained variation)
  - Smaller difference = added parameters don't reduce the unexplained variation
- Take ratio of difference to the unexplained variation in the full model

We can extend our look at the F-test

We can create a hypothesis test for more than one coefficient at a time…

Null H_0

$\beta_1 = \beta_2 = 0$

Alternative H_1

$\beta_1 \neq 0$ and/or $\beta_2 \neq 0$

in a slightly different way…

Null model

- $Y = \beta_0 + \epsilon$
- Smaller (reduced) model

Alternative* model

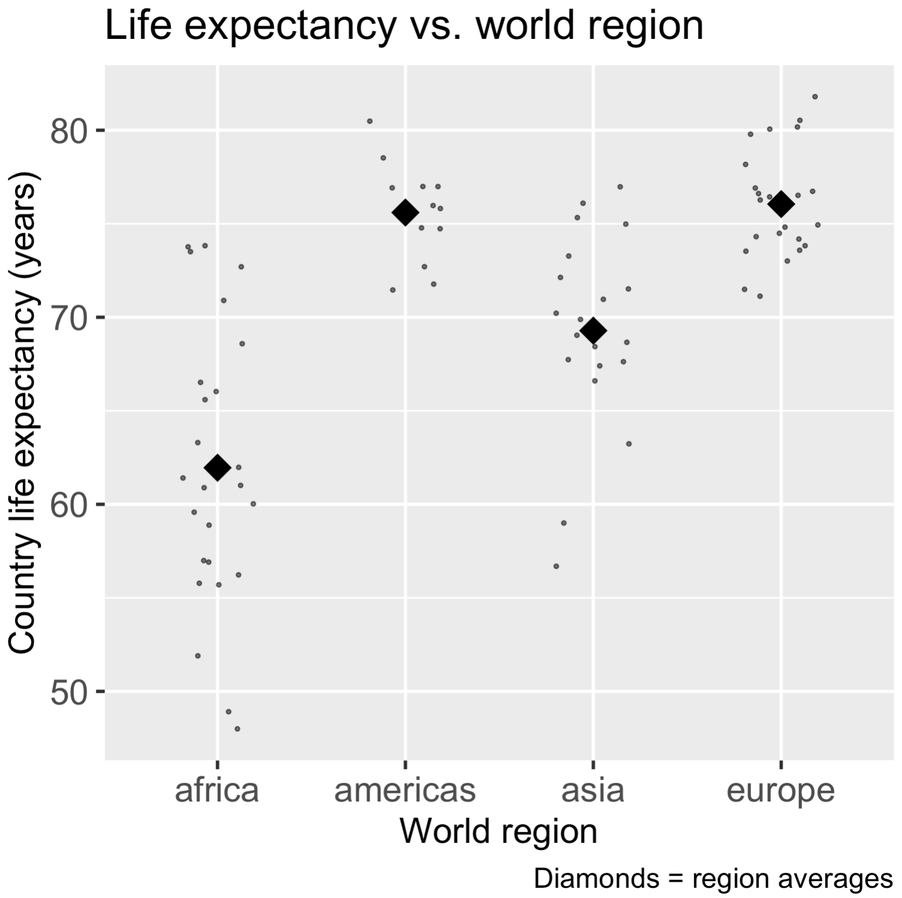
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- Larger (full) model

*This is **not quite** the alternative, but if we reject the null, then this is the model we move forward with

Let's say we want to test the association between life expectancy and world region

$$\begin{aligned}\widehat{\text{LE}} =& 61.96 + 13.64 \cdot I(\text{Americas}) + \\ & 7.33 \cdot I(\text{Asia}) + 14.1 \cdot I(\text{Europe}) \\ \widehat{\text{LE}} =& \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \\ & \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe})\end{aligned}$$

- We need to figure out if the model with world region explains significantly more variation than the model without world region!



F-test: general steps for hypothesis test for j-level categorical variable

Life expectancy example: hypothesis test for world region

- Steps 1-4 are setting up our hypothesis test: not much change from the general steps

1. For today's class, we are assuming that we have met the underlying assumptions 2. State the null hypothesis.

We are testing if the slope is 0 or not:

$$H_0 : \beta_1 = \beta_2 = \beta_3 0$$
$$\text{vs. } H_A : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0$$

3. Specify the significance level.

Often we use $\alpha = 0.05$

4. Specify the test statistic and its distribution under the null

The test statistic is F, and follows an F-distribution with numerator $df = j$ and denominator $df = n - (j + 1) = 80 - (3 + 1)$.

```
1 nobs(model1)
```

```
[1] 80
```

Life expectancy example: hypothesis test for world region (2/4)

5. Compute the value of the test statistic

```
1 model2 <- gapm %>% lm(formula = LifeExpectancyYrs ~ four_regions)
2 anova(model2) %>% tidy() %>% gt() %>% tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
four_regions	3	2845.643	948.5477	33.53311	6.514481e-14
Residuals	76	2149.804	28.2869	NA	NA

- **Option 1:** Calculate the test statistic using the values in the ANOVA table

$$F = \frac{MSR}{MSE} = \frac{948.5476696}{28.2869012} = 33.5331065$$

- **Option 2:** Get the test statistic value (F) from the ANOVA table

I tend to skip this step because I can do it all with step 6

Life expectancy example: hypothesis test for population slope β_1 (3/4)

6. Calculate the p-value

- As per Step 4, test statistic F can be modeled by a F-distribution with $df1 = 3$ and $df2 = n - 4$.
  - We had 80 countries' data, so $n = 80$
- **Option 1:** Use `pf()` and our calculated test statistic

```
# p-value is ALWAYS the right tail for F-test
pf(33.5331, df1 = 3, df2 = 76, lower.tail = FALSE)
```

```
[1] 6.514508e-14
```

- **Option 2:** Use the ANOVA table

```
anova(model2) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
four_regions	3	2845.643	948.5477	33.53311	6.514481e-14
Residuals	76	2149.804	28.2869	NA	NA

Life expectancy example: hypothesis test for population slope β_1 (4/4)

7. Write conclusion for the hypothesis test

We reject the null hypothesis that all three coefficients are equal to 0 at the 5% significance level. There is sufficient evidence that there is association between female life expectancy and the country's world region (p-value < 0.0001).

Learning Objectives

1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable

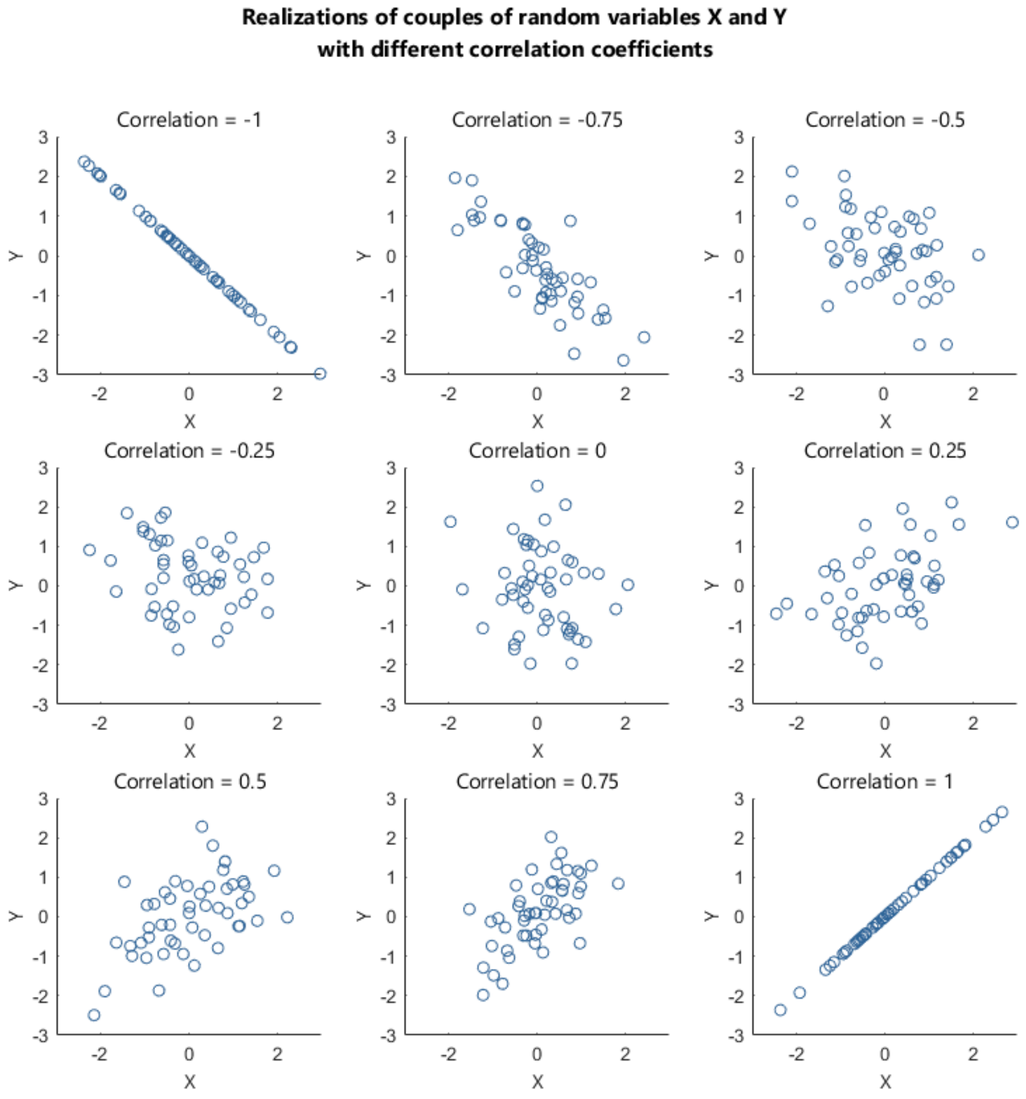
4. Calculate and interpret the coefficient of determination

Correlation coefficient from 511

Correlation coefficient r can tell us about the strength of a relationship **between two continuous variables**

- If $r = -1$, then there is a perfect negative linear relationship between X and Y
- If $r = 1$, then there is a perfect positive linear relationship between X and Y
- If $r = 0$, then there is no linear relationship between X and Y

Note: All other values of r tell us that the relationship between X and Y is not perfect. The closer r is to 0, the weaker the linear relationship.



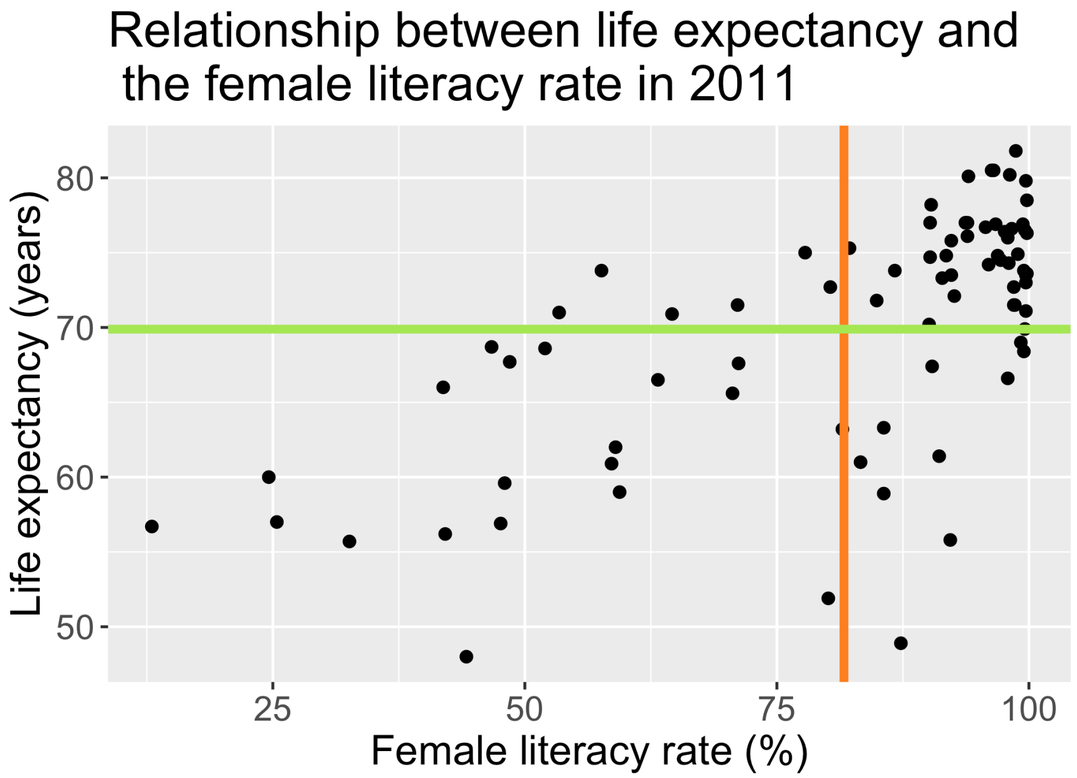
Correlation coefficient (r or R)

The (Pearson) correlation coefficient r of variables X and Y can be computed using the formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)^{1/2}}$$
$$= \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

we have the relationship

$$\widehat{\beta}_1 = r\frac{SSY}{SSX}, \quad \text{or}, \quad r = \widehat{\beta}_1 \frac{SSX}{SSY}$$



Coefficient of determination: R^2

It can be shown that the square of the correlation coefficient r is equal to

$$R^2 = \frac{SSR}{SSY} = \frac{SSY - SSE}{SSY}$$

- R^2 is called the **coefficient of determination**.
- **Interpretation**: The proportion of variation in the Y values explained by the regression model
- R^2 measures the strength of the **linear** relationship between X and Y:
  - $R^2 = \pm 1$: Perfect relationship
    - Happens when $SSE = 0$, i.e. no error, all points on the line
  - $R^2 = 0$: No relationship
    - Happens when $SSY = SSE$, i.e. using the line doesn't not improve model fit over using $\overline{Y}$ to model the Y values.

Life expectancy example: correlation coeffiicent r and coefficient of determination R^2

```
(r = cor(x = gapm$LifeExpectancyYrs, y = gapm$FemaleLiteracyRate,
         use =  "complete.obs"))
```

```
[1] 0.6410434
```

```
summary(model1) # for R^2 value
```

```
r^2
```

```
[1] 0.4109366
```

```
Call:
lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-22.299  -2.670   1.145   4.114   9.498

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        50.92790    2.66041  19.143  < 2e-16 ***
FemaleLiteracyRate  0.23220    0.03148   7.377  1.5e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.142 on 78 degrees of freedom
Multiple R-squared:  0.4109,    Adjusted R-squared:  0.4034
F-statistic: 54.41 on 1 and 78 DF,  p-value: 1.501e-10
```

Interpretation

41.1% of the variation in countries' life expectancy is explained by the linear model with female literacy rate as the independent variable.

What does R^2 not measure?

- R^2 is not a measure of the magnitude of the slope of the regression line
  - Example: can have $R^2 = 1$ for many different slopes!!
- R^2 is not a measure of the appropriateness of the straight-line model
  - Example: figure