

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

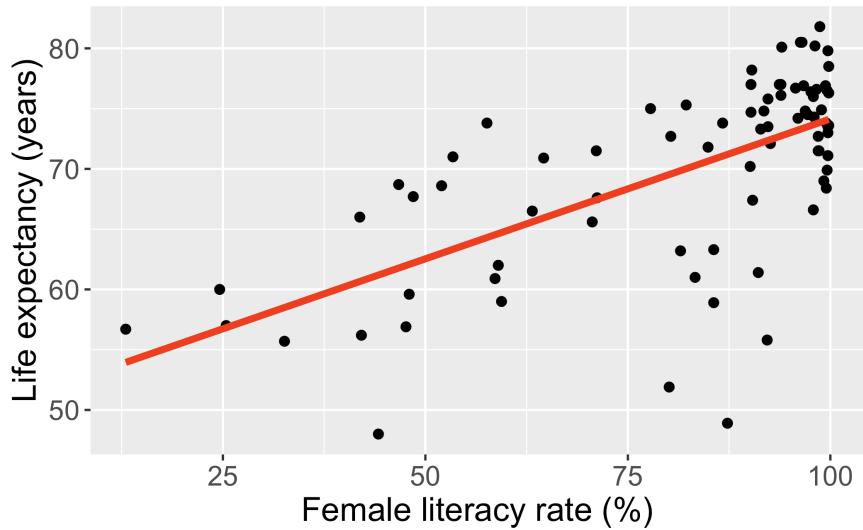
- Estimate variance of the residuals
- Inference for slope & intercept: CI, p-value
- ~~Confidence bands of regression line for mean value of $Y|X$~~

finding

Lesson 5: Categorical Covariates

- Inference for different categories

Relationship between life expectancy and the female literacy rate in 2011



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$
$$\widehat{\text{LE}} = 50.9 + 0.232 \cdot \text{FLR}$$

Process of regression data analysis

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

} ish

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates ✓
- Interactions

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals
- Transformations
- Influential points
- Multicollinearity

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

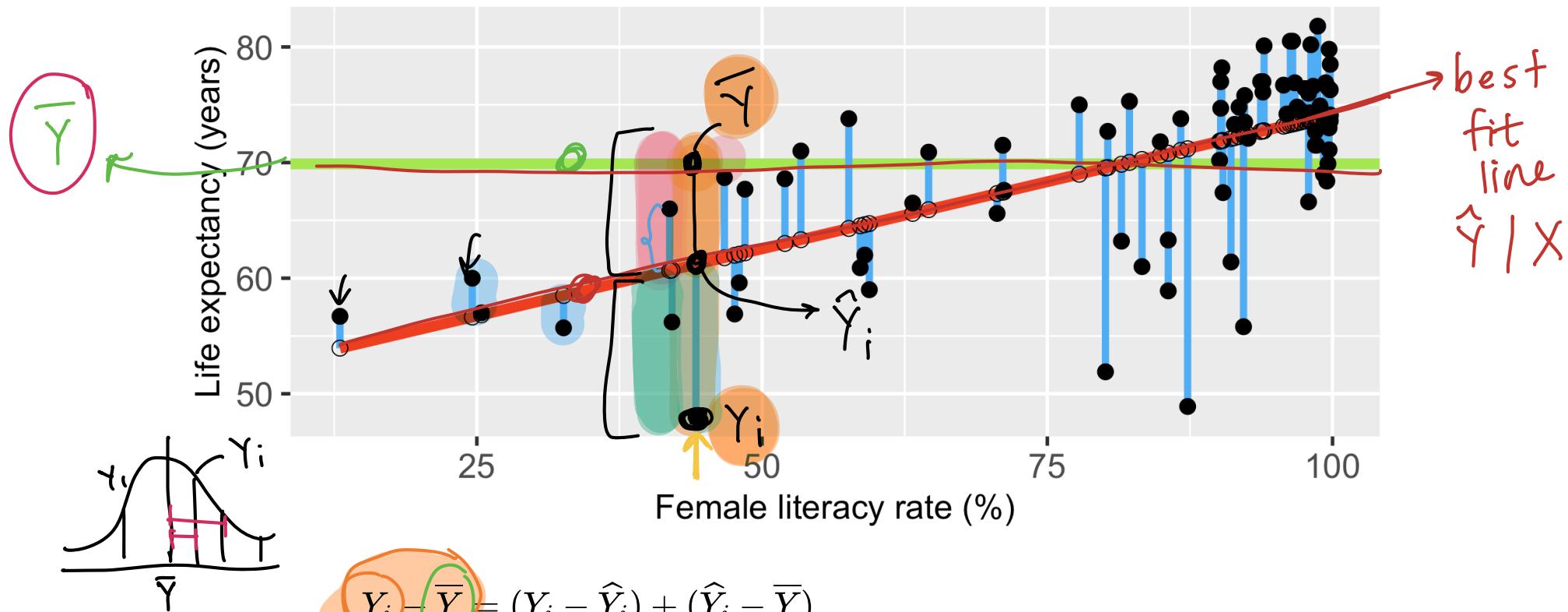
The F statistic in linear regression is essentially a proportion of the variance explained by the model vs. the variance not explained by the model

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

total = unexplained
+ explained

Life expectancy vs. female literacy rate in 2011



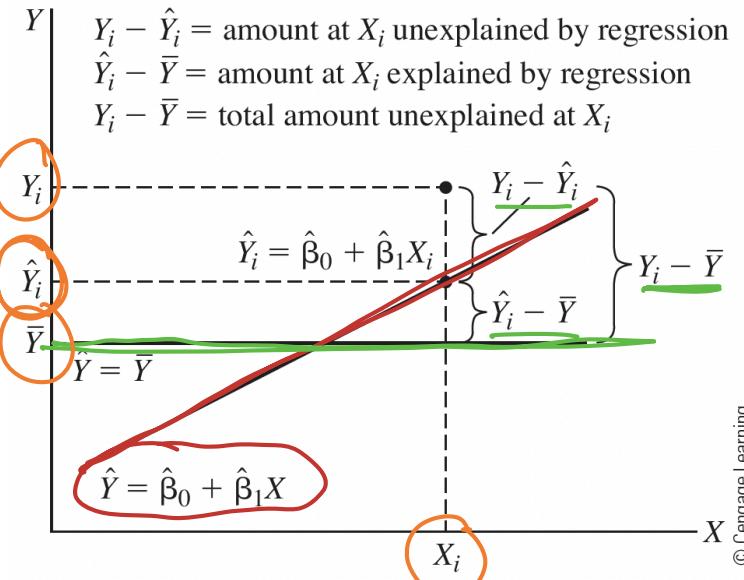
Total variation = Residual variation after regression + Variation explained by regression

unexplained

More on the equation

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

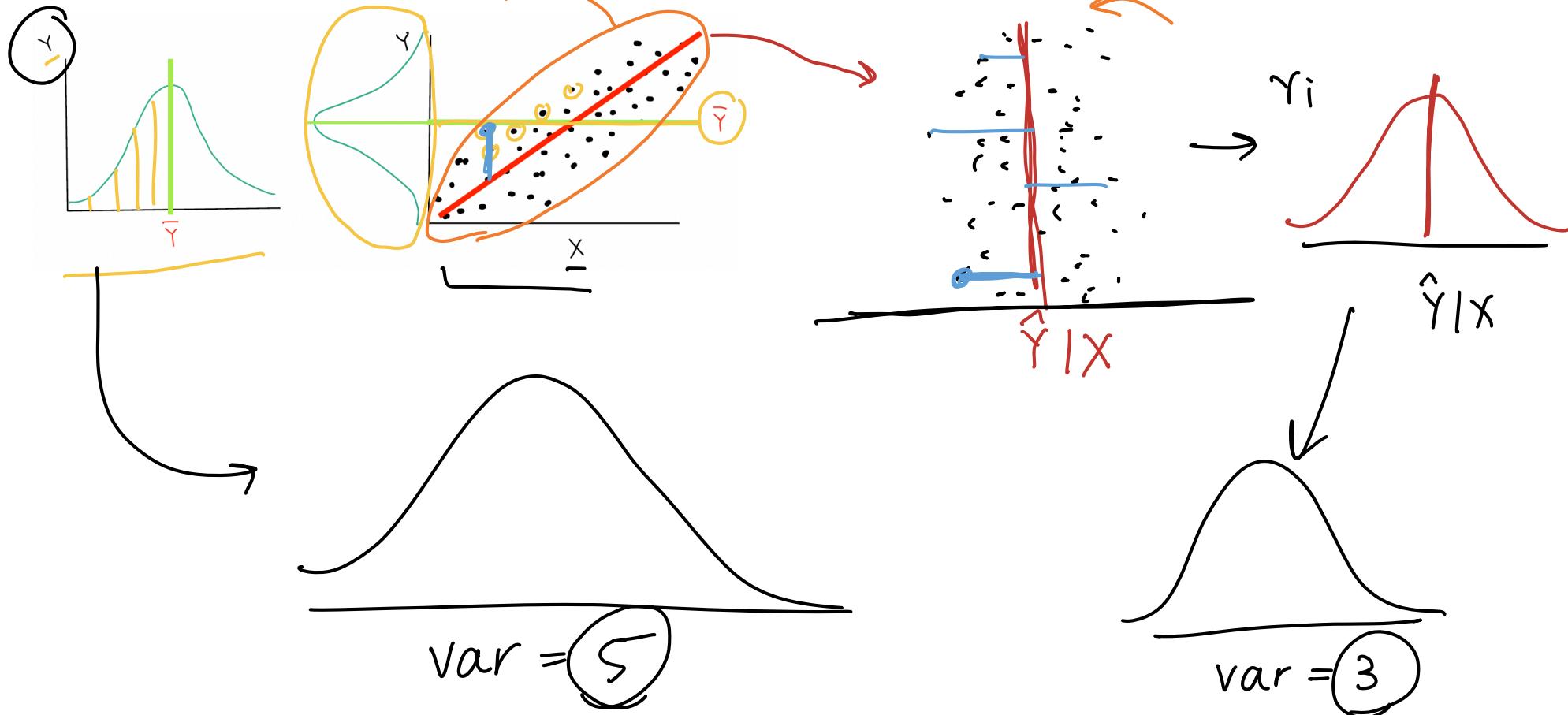
- $Y_i - \bar{Y}$ = the deviation of Y_i around the mean \bar{Y}
 - (the total amount deviation ~~unexplained~~).
- $Y_i - \hat{Y}_i$ = the deviation of the observation Y_i around the fitted regression line
 - (the amount deviation unexplained by the regression at X_i).
- $\hat{Y}_i - \bar{Y}$ = the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



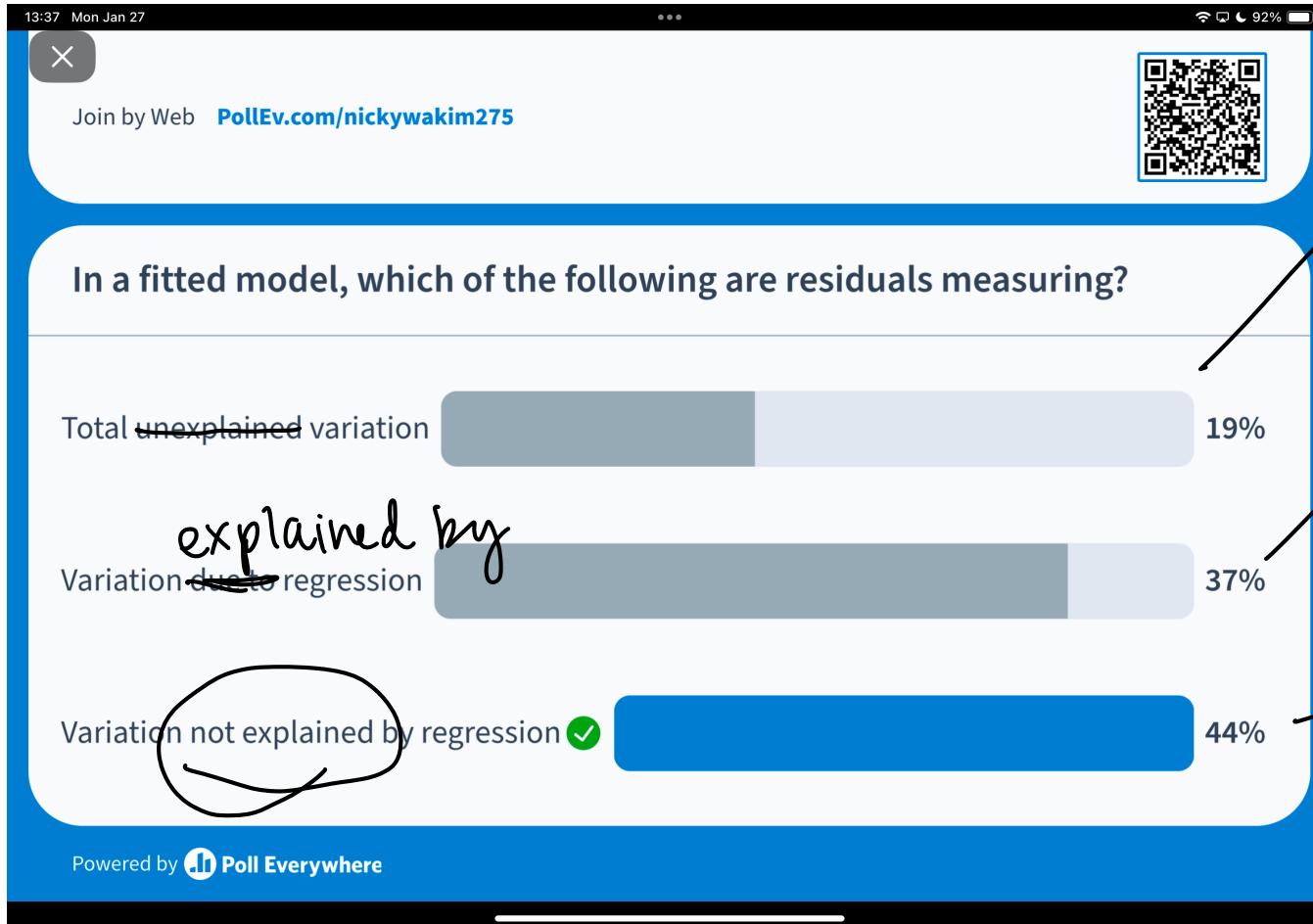
© Cengage Learning

FIGURE 7.1 Variation explained and unexplained by straight-line regression

Another way of thinking about the different deviations



Poll Everywhere Question 1



$$\begin{aligned} Y_i - \bar{Y} \\ \hat{Y}_i - \bar{Y} \\ L(\hat{\beta}_0 + \hat{\beta}_1 x) \\ \hat{\epsilon}_i = Y_i - \hat{Y}_i \end{aligned}$$

residuals

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

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Total unexplained variation = Variation due to regression + Residual variation after regression explained by

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

$$Y_i - \bar{Y} > 0 \\ < 0$$

$$SSY = SSR + SSE$$

SST

Total Sum of Squares = Sum of Squares due to Regression + Sum of Squares due to Error (residuals) explained by

ANOVA table:

Variation Source	df	sum of sq	mean sq	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	$n - 1$	$\frac{SSY}{SST}$	mean sq error		

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

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$
$$SSY = SSR + SSE$$

Total Sum of Squares = Sum of Squares due to Regression + 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

```
1 # Fit regression model:  
2 modell1 <- gapm %>% lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate)  
3  
4 anova(modell1)
```

Analysis of Variance Table

Response: LifeExpectancyYrs

model/regression 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

total

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
1 anova(modell1) %>% tidy() %>% gt()%>%  
2 tab_options(table.font.size = 40) %>%  
3 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

SSE MSE

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?

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

- It can be shown that

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

- Recall that σ^2 is the variance of the population residuals

- Thus if

- $\beta_1 = 0$, then $F \approx \frac{\hat{\sigma}^2}{\hat{\sigma}^2} = 1$

- $\beta_1 \neq 0$, then $F \approx \frac{\hat{\sigma}^2 + \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\hat{\sigma}^2} > 1$

- So the F statistic can also be used to test β_1

no association
b/w Y & X (cont),
aka $\beta_1 = 0$
 $\Rightarrow F = 1$

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$F > 1$: more variation explained by model

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:

basically only
for cont. x

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ \text{vs. } H_A : \beta_1 &\neq 0 \end{aligned}$$

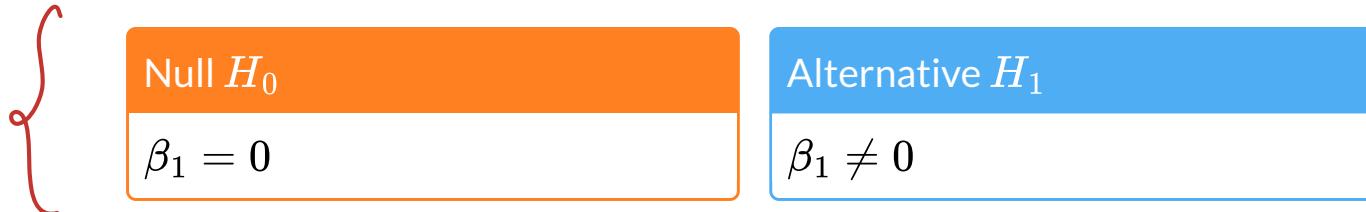
- Note that the F-test does not support one-sided alternative tests, but the t-test does!
 - F-test cannot handle alternatives like $\beta_1 > 0$ nor $\beta_2 < 0$

cannot use t-test

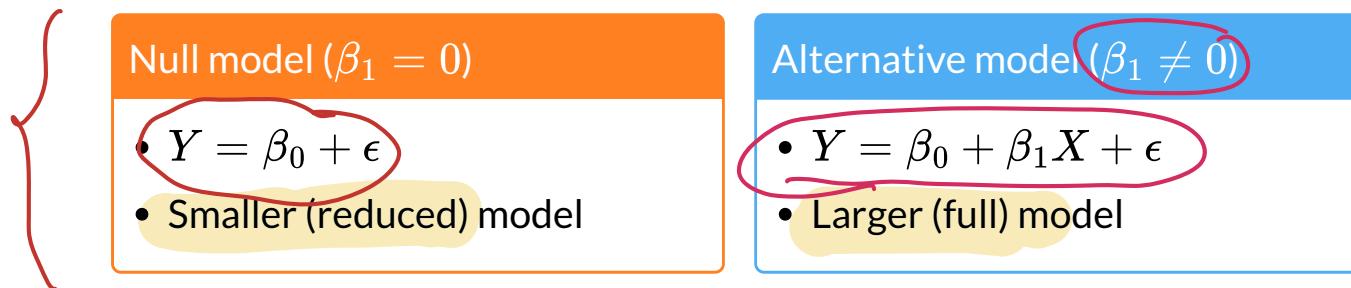
$$\left[\begin{array}{l} H_0: \beta_1 = 0, \beta_2 = 0 \\ H_A: \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \end{array} \right]$$

Planting a seed about the F-test

We can think about the hypothesis test for the slope...



in a slightly different way...



- 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

13:56 Mon Jan 27 ... 86% 

Join by Web PollEv.com/nickywakim275

Can I use the F-test to test the following null and alternative models?

Null: $Y = \beta_0 + \epsilon$ $\rightarrow \beta_1 = 0, \beta_2 = 0$

Alternative: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \rightarrow \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$

Yes!  92%

Yes and the t-test can do it, too!  4%

No!  4%

Powered by  Poll Everywhere

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

w/ an X
that is
contin-
uous

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:

$$H_0: \beta_1 = 0$$

vs. $H_A: \beta_1 \neq 0$

3. Specify the significance level.

Often we use $\alpha = 0.05$

R is # coef
testing

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

$$df = n - 2,$$

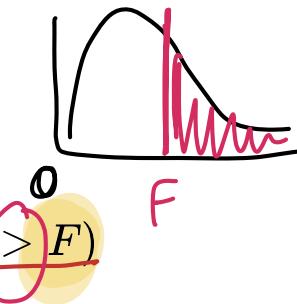
$$\hookrightarrow n - (R + 1)$$

of coef in model
- 1 (2-1)

5. Compute the value of the test statistic

The calculated test statistic for $\hat{\beta}_1$ 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 $100\alpha\%$ significance 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:

$$\left[H_0 : \beta_1 = 0 \text{ vs. } H_A : \beta_1 \neq 0 \right] \rightarrow Y = \beta_0 + \varepsilon \text{ (reduced / null)}$$
$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ (full / alt)}$$

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

```
1 # p-value is ALWAYS the right tail for F-test
2 pf(54.414, df1 = 1, df2 = 78, lower.tail = FALSE)
[1] 1.501104e-10
```

pf(f, df1 = 1, df2 = 80 - 2,
num denom lower.
tail = FAIL)

- Option 2:** Use the ANOVA table

```
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

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

β_1

t-stat

- 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

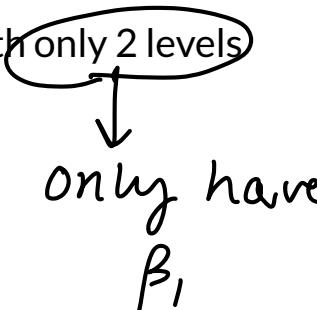
This is true when we use the F-test for a single coefficient!

F-stat

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 b/c multiple coeff!
one variable
but we're using
multiple
coeff
- 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

only have
 β_1

Poll Everywhere Question 3

14:21 Mon Jan 27

Join by Web PollEv.com/nickywakim275

Why can we use the t-test for binary categorical variables?

They only require one coefficient in the model  89%

They require multiple coefficients in the model 4%

You can use the t-test for any type of variable 7%

Powered by  Poll Everywhere



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) can have mult coeff

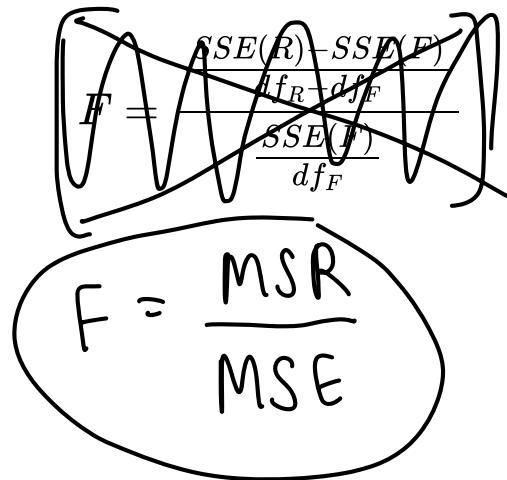
Does the addition of one particular covariate add significantly to the prediction of Y achieved by other covariates already present in the model?

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



- $SSE(R) \geq SSE(F)$
- 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

$$\beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$

in a slightly different way...

Null model

- $\underline{Y = \beta_0 + \epsilon}$

- Smaller (reduced) model



Alternative* model

- $\underline{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon}$

- 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

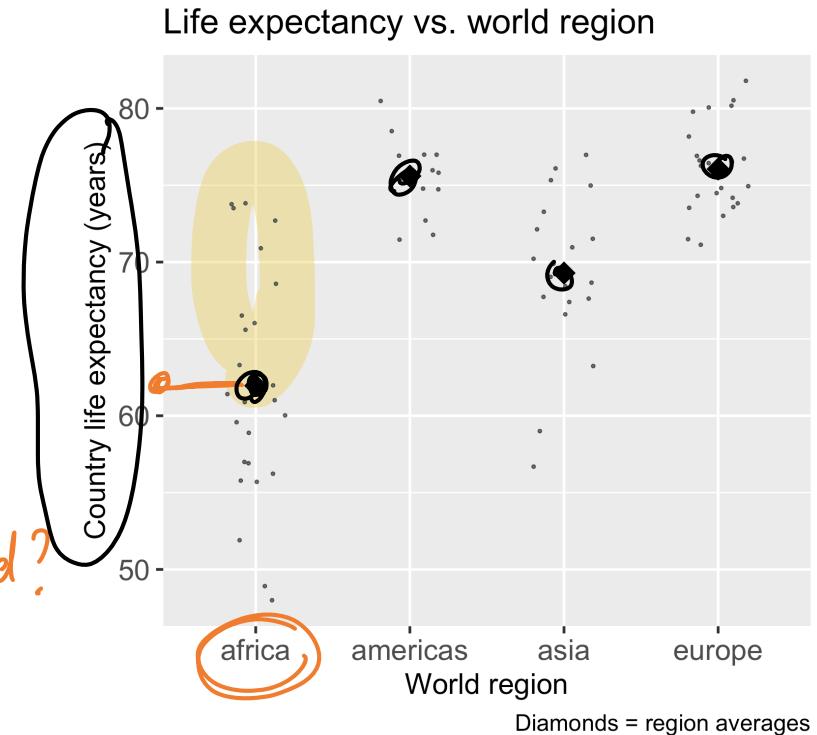
$$\widehat{LE} = 61.96 + 13.64 \cdot I(\text{Americas}) + 7.33 \cdot I(\text{Asia}) + 14.1 \cdot I(\text{Europe})$$

$$\left[\begin{array}{l} \widehat{LE} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot I(\text{Americas}) + \\ \widehat{\beta}_2 \cdot I(\text{Asia}) + \widehat{\beta}_3 \cdot I(\text{Europe}) \end{array} \right]$$

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

aka is WR & LE associated?

$$LE = \beta_0 + \beta_1 I(\text{Am}) + \beta_2 I(\text{Asia}) + \beta_3 I(\text{Eur})$$



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 ~~WR is 0 or not~~: WR is associated w/ LF

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

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

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 $df = n - (j + 1) = 80 - (3 + 1)$.

j = # coef testing
3

```
1 nobs(model1)
[1] 80
```

→ we also estimate β_0

$$\beta_1, \beta_2, \beta_3 + \frac{1}{\beta_0}$$

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

```
1 # p-value is ALWAYS the right tail for F-test
2 pf(33.5331, df1 = 3, df2 = 76, lower.tail = FALSE)
[1] 6.514508e-14
```

- Option 2:** Use the ANOVA table

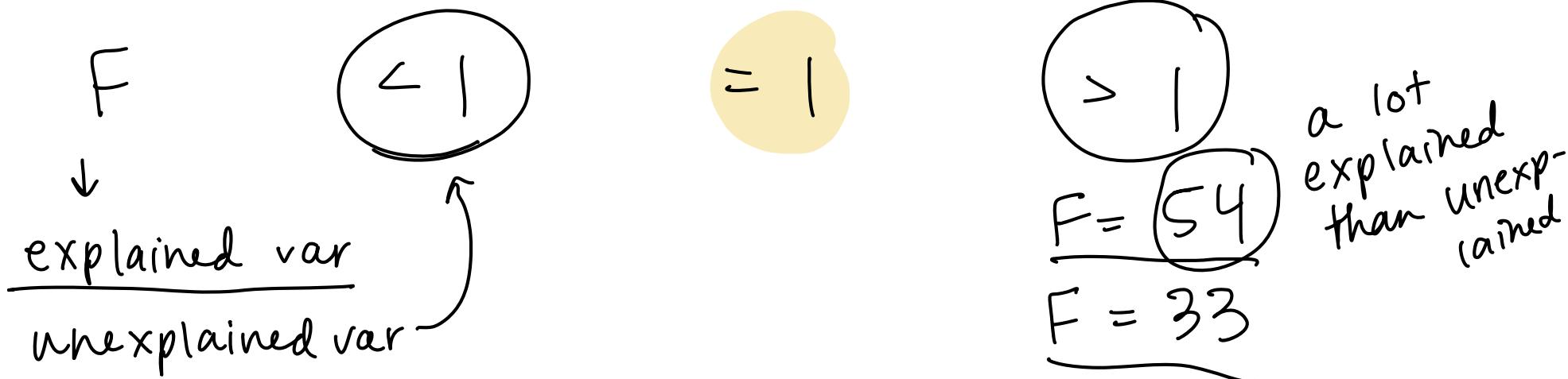
```
1 anova(model2) %>% tidy() %>% gt() %>%
2   tab_options(table.font.size = 40)
```

	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

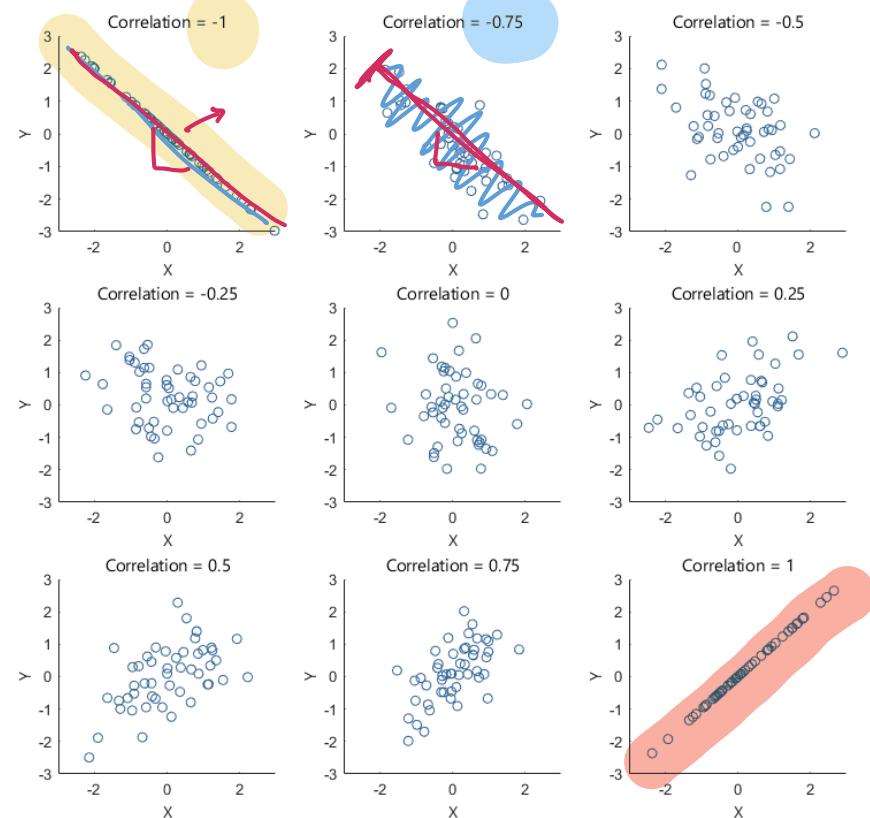
Correlation coefficient from 511 / 525/611

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.

Realizations of couples of random variables X and Y with different correlation coefficients



Correlation coefficient (r or R)

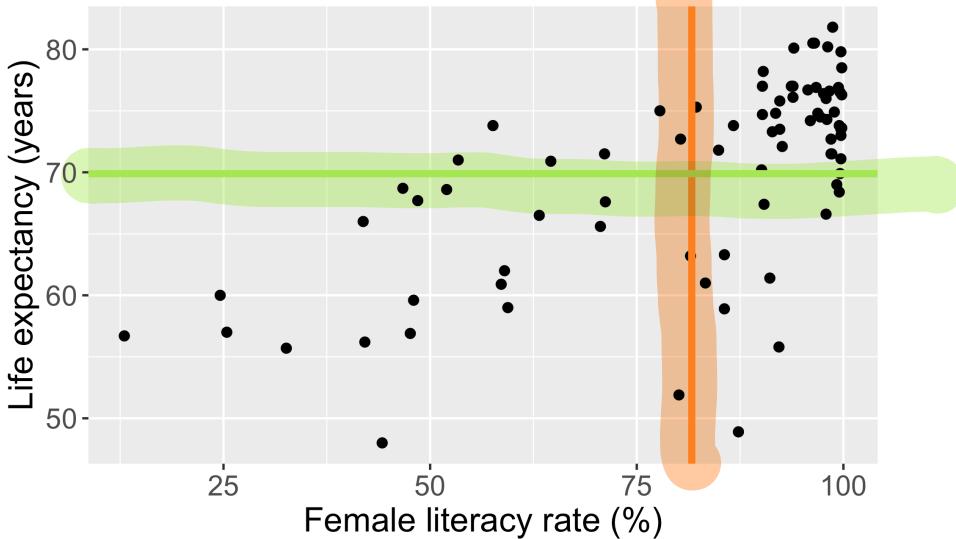
The (Pearson) correlation coefficient r of variables X and Y can be computed using the formula:

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

we have the relationship

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

Relationship between life expectancy and the female literacy rate in 2011



Coefficient of determination: R^2

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

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

variation explained
by model

total variation

- 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 \bar{Y} to model the Y values.

$$SSR = 0$$

0 variation explained by model

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

```
1 (r) = cor(x = gapm$LifeExpectancyYrs, y = gapm$FemaleLiteracyRate,  
2   use = "complete.obs"))  
[1] 0.6410434  
1 summary(model1) # for  $R^2$  value
```

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

1 \hat{r}^2
[1] 0.4109366

coef of determ.

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

