

Lesson 6: SLR: More inference

Nicky Wakim

2026-01-28

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 (expanding to test mult β 's)
- * 4. Calculate and interpret the coefficient of determination

→ parallel to
t-test
Lesson 4

So far in our regression example...

Lesson 3: SLR 1

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

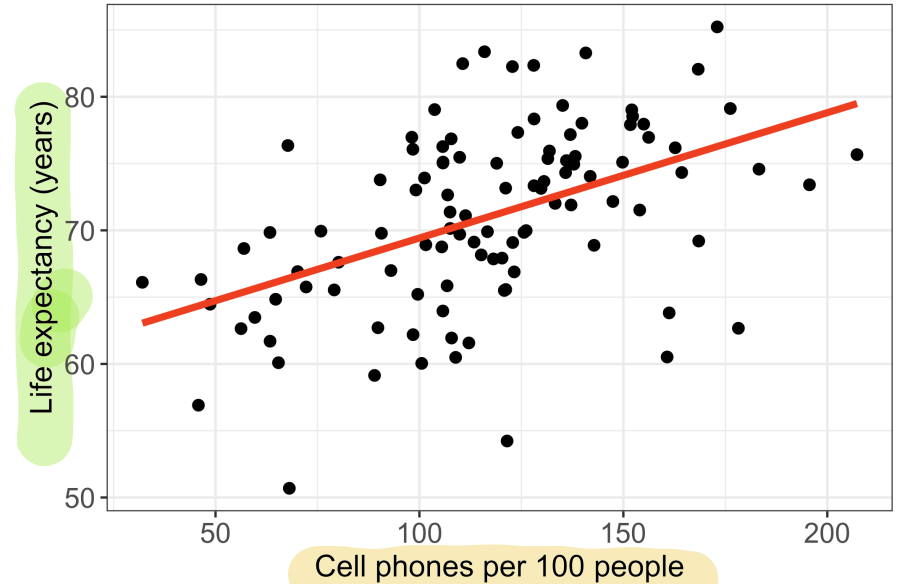
Lesson 4: SLR 2

- Estimate variance of the residuals
- Inference for slope & intercept: CI, p-value ✓
- Finding mean value of Y|X ✓

Lesson 5: Categorical Covariates

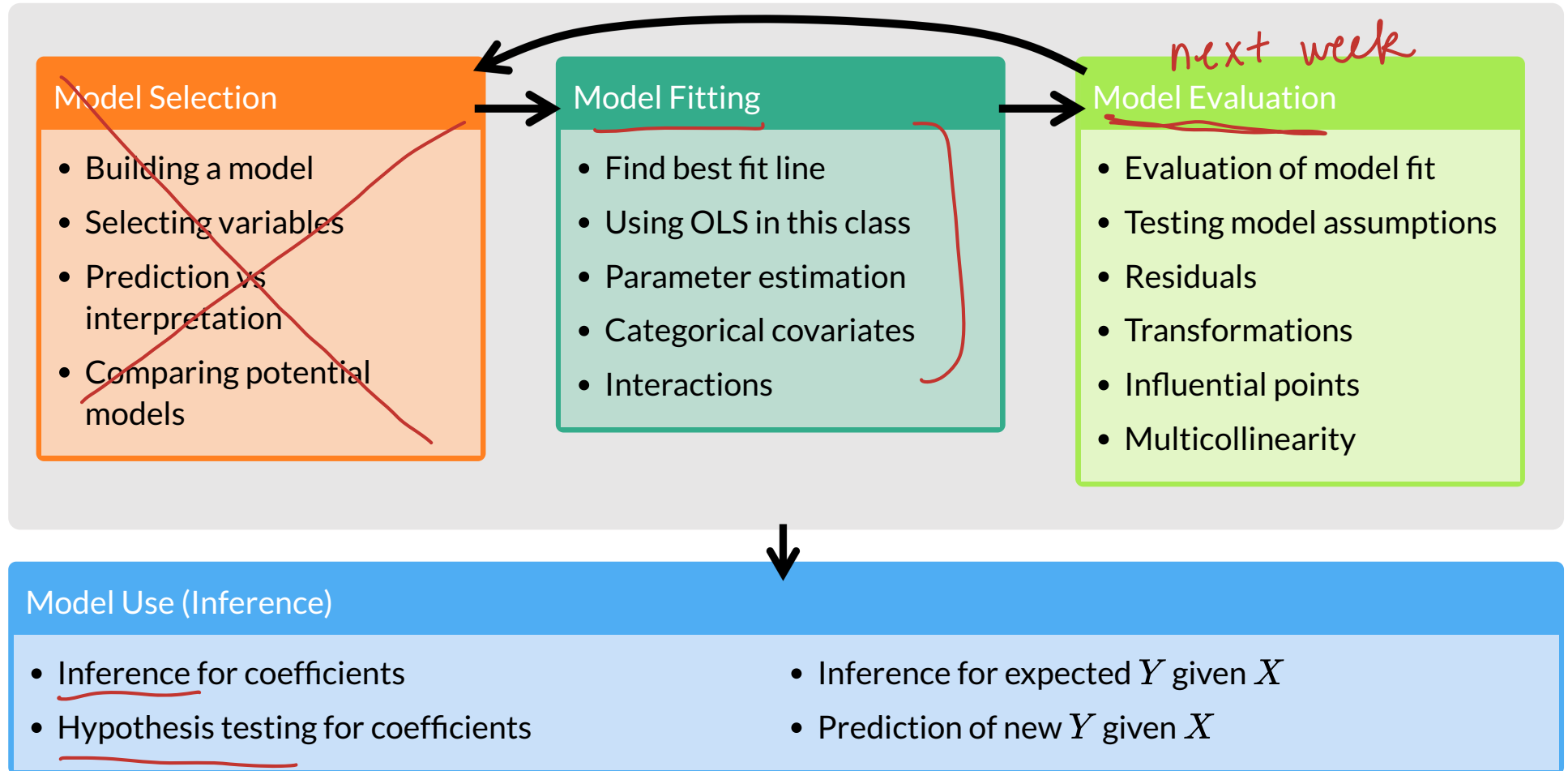
- Inference/Interpretation for different categories

Relationship between life expectancy and cell phones



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$
$$\widehat{LE} = 60.04 + 0.094 \cdot CP$$

Process of regression data analysis



Learning Objectives

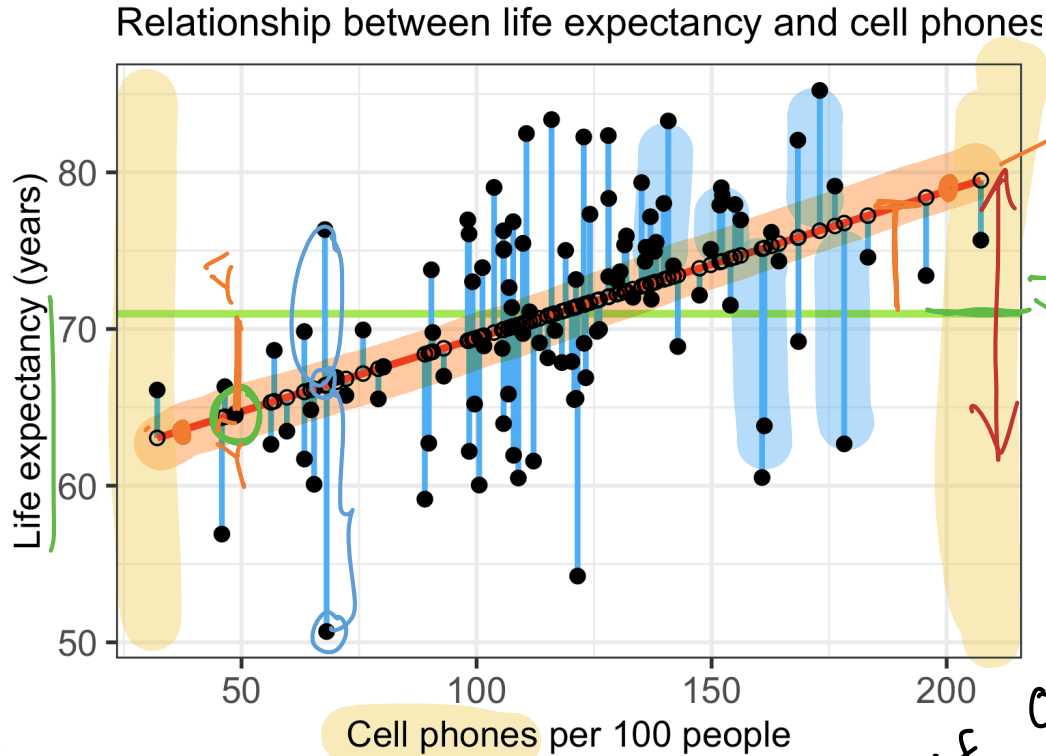
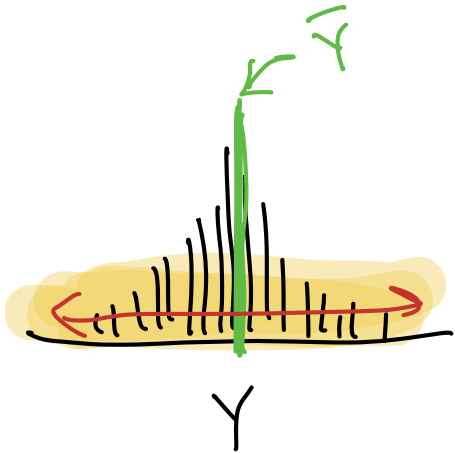
1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4.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



$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

if 0, cell phones
have given us NO
info abt life expectancy,
so not helpful
is model
(no assoc)

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

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

TOTAL = $\hat{\epsilon}$ UNEXPLAINED + EXPLAINED

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)
- $Y_i - \hat{Y}_i$ = the deviation of the observation Y 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)

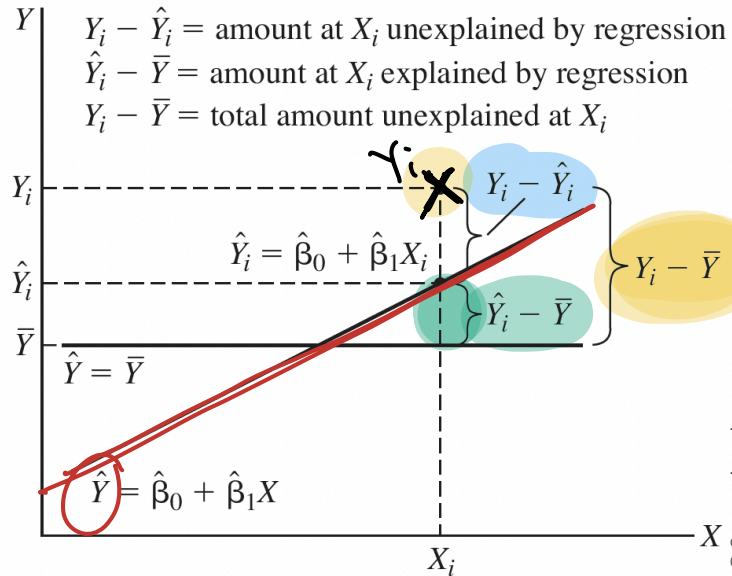
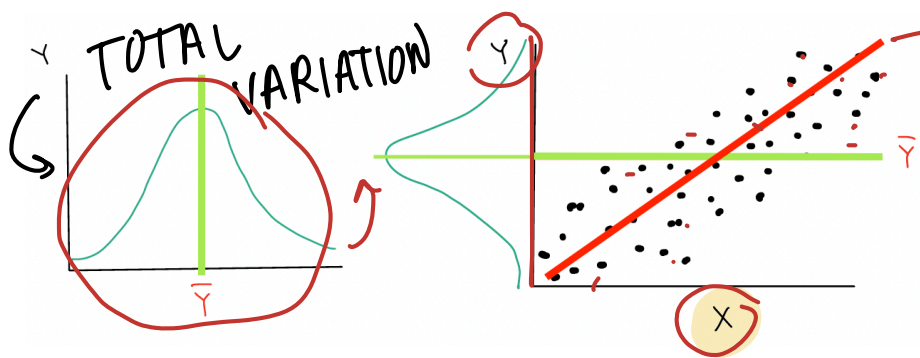


FIGURE 7.1 Variation explained and unexplained by straight-line regression

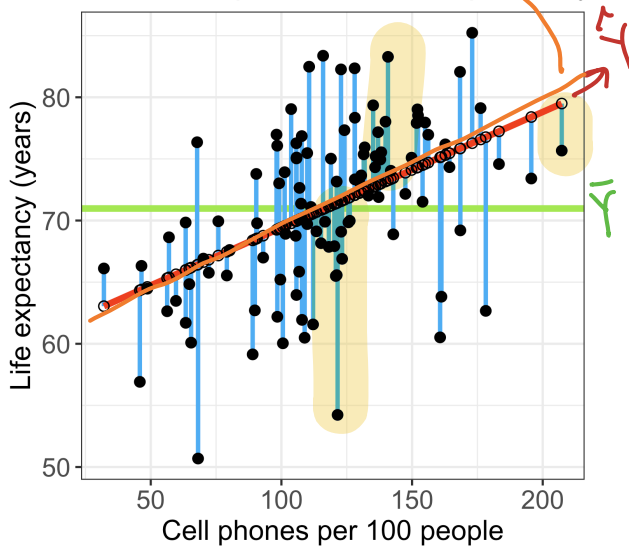
Another way of thinking about the different deviations



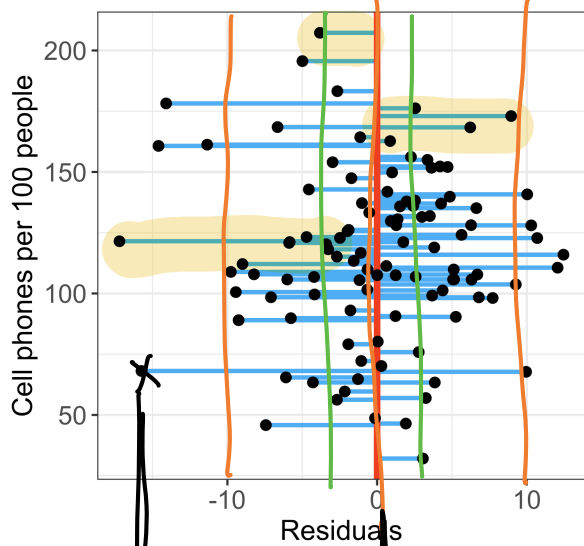
\hat{Y} is describing how Y changes when X changes



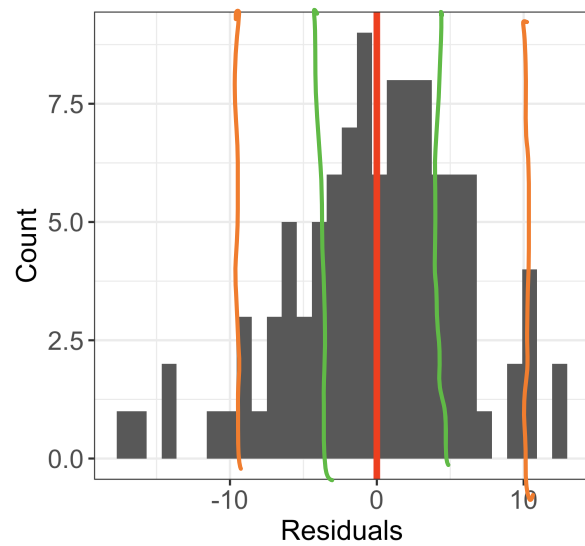
Relationship between life expectancy



Residuals from fitted model



Residuals from fitted model



$$\hat{\epsilon}_i = Y_i - \hat{Y}$$

Poll Everywhere Question 1

13:28 Wed Jan 28



Join by Web PollEv.com/nickywakim275



In a fitted model, which of the following are residuals measuring?

we want to minimize the sum of

squared residuals (errors)

Total variation



21%

Variation explained by regression



41%

sum of sq residuals become smaller as we explain more variation w/ model

Variation not explained by regression ✓



38%

residual / left over variation after we fit a model

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

deviation [

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

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

info from ALL obs →

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

$$\text{SSY} = \text{SSR} + \text{SSE errors}$$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

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

ANOVA table:

Variation Source	df	SS	MS normalized SS	test statistic	p-value
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	→ $\frac{(SSR/df1)}{(SSE/df2)}$
Error	$n - 2$	SSE	$MSE = \frac{SSE}{n-2}$		
Total	$n - 1$	SSY			

F test is SSR vs SSE

so we can compare SSR & SSE

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 explained by 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			

Handwritten notes: "add β_1 " with an arrow pointing to the df=1 cell; "coeff in model" with an arrow pointing to the df=n-2 cell.

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

want $F > 1$

Analysis of Variance (ANOVA) table in R

```

1 # Fit regression model:
2 model1 <- gam %>% lm(formula = life_exp ~ cell_phones_100)
3
4 anova(model1)

```

Analysis of Variance Table

Response: life_exp

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cell_phones_100	1	1094.1	1094.10	30.759	2.271e-07 ***
Residuals	103	3663.7	35.57		

Handwritten annotations: SSR points to Sum Sq, MSR points to Mean Sq, F points to F value, SSE points to Sum Sq of Residuals, MSE points to Mean Sq of Residuals.

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

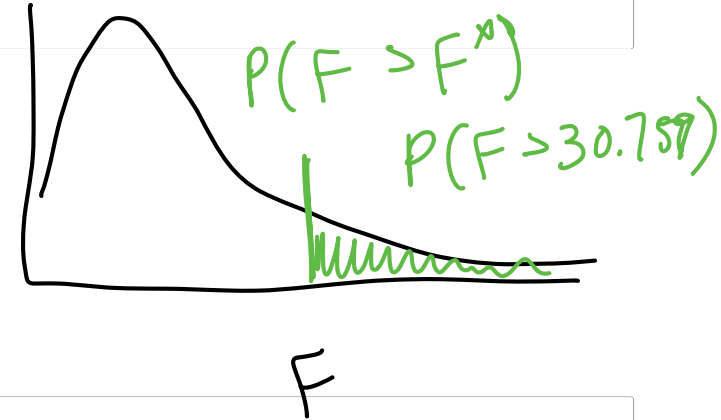
```

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

```

term	df	sumsq	meansq	statistic	p.value
cell_phones_100	1.000	1,094.102	1,094.102	30.759	0.000
Residuals	103.000	3,663.747	35.570	NA	NA

→



Learning Objectives

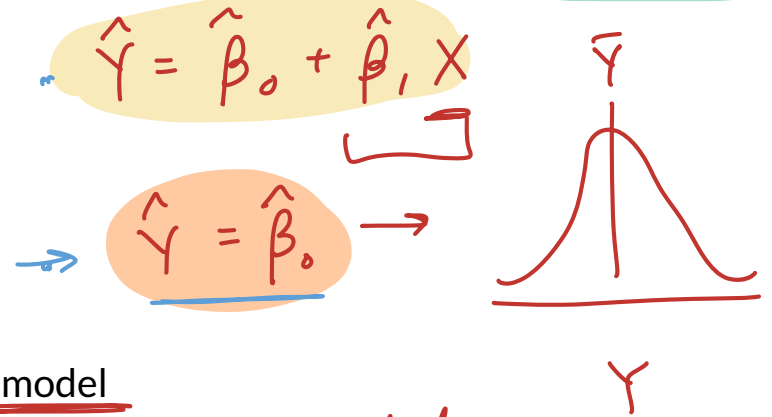
1. Identify different sources of variation in an Analysis of Variance (ANOVA) table
2. Using the F-test, determine if there is enough evidence that population slope β_1 is not 0
3. Using the F-test, determine if there is enough evidence for association between an outcome and a categorical variable
4.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

What is the F statistic testing?

- The F statistic is testing whether the model with the predictor variable X explains significantly more variation in Y than a model with no predictors (intercept only model)

- In general, the F-test has 3 steps:

- Define the larger (full) model (more coefficients)
- Define the smaller (reduced) model (less coefficients)
- Decide whether to reject the reduced model in favor of the full model



- We “decide” by seeing how much more variation the full model explains compared to the reduced model

- If the full model explains significantly more variation, we reject the reduced model

- In simple linear regression, this is equivalent to testing whether the population slope β_1 is equal to 0

↳ is bigger model worth it? AKA does it explain enough variation

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:



$$H_0 : \beta_1 = 0$$

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 $\beta_2 < 0$

↳ use t-test if
one-sided
aka alt hyp
is $\beta_1 > 0$
or $\beta_1 < 0$

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

13:53 Wed Jan 28

84%



Join by Web PollEv.com/nickywakim275



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

Null: $Y = \beta_0 + \epsilon$

Alternative: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$



Yes!

78%

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

22%

No!

0%

t-test can...

null: $Y = \beta_0 + \epsilon$

alt: $Y = \beta_0 + \beta_2 X_2 + \epsilon$

null: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

alt: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

t-test can only handle each coefficient individually

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

1. Check the **assumptions** ✓
2. Set the **level of significance** ✓
 - Often we use $\alpha = 0.05$
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses** ✓
 - Often, we are curious if the coefficient is 0 or not:

$$H_0 : \beta_1 = 0$$

vs. $H_A : \beta_1 \neq 0$

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. Calculate 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 a **conclusion**

- 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. Check the **assumptions**: We have met the underlying assumptions (checked in our Model Evaluation step)
 2. Set the **level of significance**
 - Often we use $\alpha = 0.05$
 3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
 - Often, we are curious if the coefficient is 0 or not:

$$\left[\begin{array}{l} H_0 : \beta_1 = 0 \\ \text{vs. } H_A : \beta_1 \neq 0 \end{array} \right]$$

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 $df = n - 2 = \underline{105} - 2 = 103$

```
1 nobs(model1)
[1] 105
```

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

5. Calculate the **test statistic**

$$LE = \beta_0 + \beta_1 CP + \epsilon$$

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

term	df	sumsq	meansq	statistic	p.value
cell_phones_100	1	1094.102	1094.10190	30.75881	2.271176e-07
Residuals	103	3663.747	35.57035	NA	NA

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

$$F = \frac{MSR}{MSE} = \frac{1094.1019013}{35.5703546} = 30.759$$

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

$$F = 30.759$$

```
1 F_stat = anova_tab$statistic[1]
```

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 $df_1 = 1$ and $df_2 = n - 2$.
 - We had 105 countries' data, so $n = 105$
- **Option 1:** Use `pf()` and our calculated test statistic

```
1 # p-value is ALWAYS the right tail for F-test
2 n = nobs(model1)
3 pf(F_stat, df1 = 1, df2 = n-2, lower.tail = FALSE)
[1] 2.271176e-07
```

- **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
cell_phones_100	1	1094.102	1094.10190	30.75881	2.271176e-07
Residuals	103	3663.747	35.57035	NA	NA

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

7. Write a **conclusion**

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 life expectancy and cell phones per 100 people (p-value < 0.0001).

cell phones per 100 ppl explain
enough variation in life expectancy

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:

lm output

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

term	estimate	std.error	statistic	p.value
(Intercept)	60.04051297	2.05566959	29.207278	1.215444e-51
cell_phones_100	0.09383818	0.01691978	5.546063	2.271176e-07

- For the F-test:


ANOVA

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

term	df	sumsq	meansq	statistic	p.value
cell_phones_100	1	1094.102	1094.10190	30.75881	2.271176e-07
Residuals	103	3663.747	35.57035	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

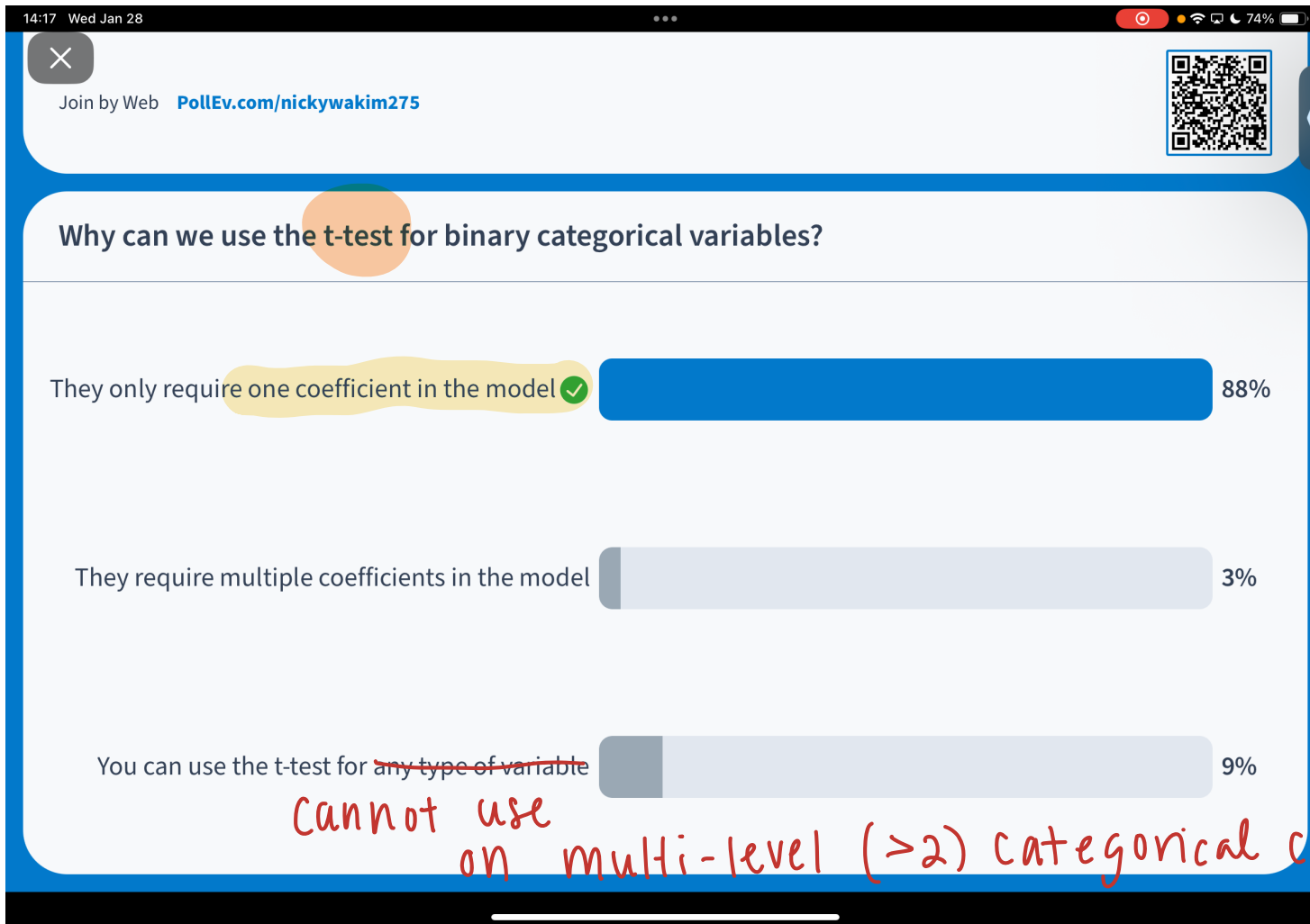
- Before we used the F-test (or t-test) to determine association between two continuous variables
- We CANNOT use the t-test to determine association between a continuous outcome and a multi-level categorical variable
- We CAN use the F-test to do this!

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

mult coeff =
F-test

one coeff to rep
2-level (binary)
cat vars.

Poll Everywhere Question 3



Binary cat
variable

Yes / No

1 / 0

$I(X = \text{Yes})$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 I(X = \text{Yes})$$

ref grp
 $X = \text{No}$

Building a very important toolkit: three types of tests

$$H_0: Y = \beta_0 + \varepsilon$$

$$H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$



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)

$$H_0: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

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)?

$$\hookrightarrow H_0: Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$H_1: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

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...

- ↳
- ① $\beta_1 \neq 0, \beta_2 = 0$
 - ② $\beta_1 = 0, \beta_2 \neq 0$
 - ③ $\beta_1 \neq 0, \beta_2 \neq 0$

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

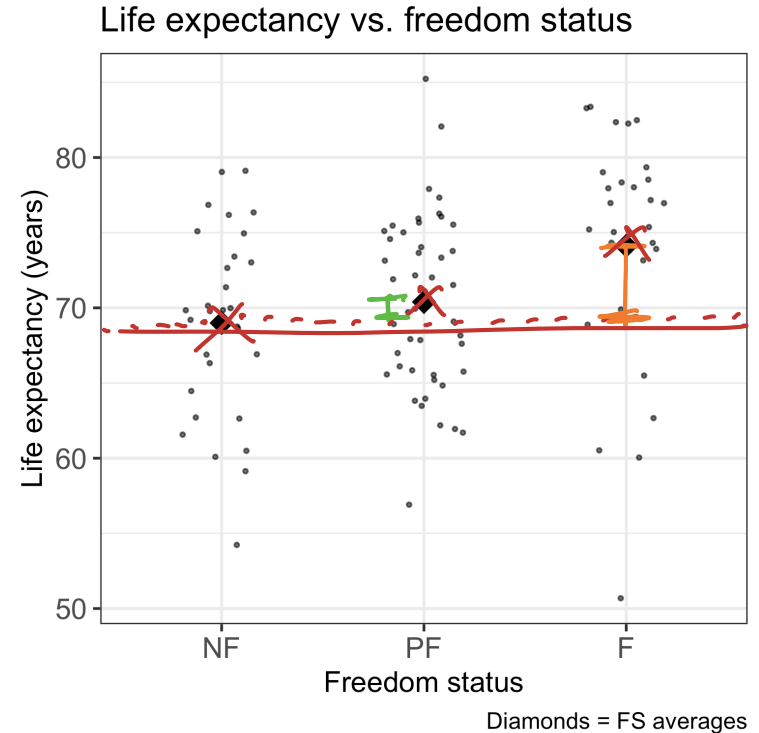
*This is not quite the alternative, but if we reject the null, then this is the model we move forward with

Let's say we want to test the association between life expectancy and freedom status

NF is ref

$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 \cdot I(\text{PF}) + \hat{\beta}_2 \cdot I(\text{F})$$
$$\widehat{LE} = 68.99 + 1.4 \cdot I(\text{PF}) + 5.14 \cdot I(\text{F})$$

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



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

testing
 k coeff-
icients

1. Check the **assumptions**
2. Set the **level of significance**
 - Often we use $\alpha = 0.05$
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
 - Often, we are curious if the coefficients are 0 or not:

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

vs. $H_A : \beta_1 \neq 0$ and/or $\beta_2 \neq 0 \dots$ and/or $\beta_k \neq 0$

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

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

variation explained by model

total coeff^m model ($\text{int} + \beta_1 + \beta_2 + \dots + \beta_k$)

5. Calculate the **test statistic**

The calculated test statistic for β s is

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

6. Calculate the **p-value**

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

7. Write a **conclusion**

- 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) ($p\text{-value} = P(F_{1, n-(k+1)} > F)$).

Life expectancy example: hypothesis test for freedom status (1/4)

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

✓ 1. Check the **assumptions**: We have met the underlying assumptions (checked in our Model Evaluation step)

2. Set the **level of significance**

- Often we use $\alpha = 0.05$

3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**

- We are testing if the FS is associated with life expectancy:

$$H_0: \beta_1 = \beta_2 = \underline{0}$$

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

$H_0: \bar{LE}$

$LE = \beta_0 + \varepsilon$

$H_1: LE = \beta_0 + \beta_1 I(PF) + \beta_2 I(F) + \varepsilon$

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

The test statistic is F , and follows an F-distribution with numerator $df = k$ and denominator $df = n - (k + 1) = 105 - (2 + 1)$

```
1 nobs(model1)
[1] 105
```

n residuals \downarrow total coeff in alt model $\beta_0, \beta_1, \beta_2 = 3 = k + 1$

Life expectancy example: hypothesis test for freedom status (2/4)

5. Calculate the test statistic

```
1 model2 <- gapm %>% lm(formula = life_exp ~ freedom_status)
2 anova_tab2 = anova(model2) %>% tidy()
3 anova_tab2 %>% gt() %>% tab_options(table.font.size = 40)
```

term	df	sumsq	meansq	statistic	p.value
freedom_status	2	415.9368	207.96841	4.885585	0.009414324
Residuals	102	4341.9116	42.56776	NA	NA

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

$$F = \frac{MSR}{MSE} = \frac{207.9684098}{42.5677608} = 4.886$$

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

$$F = 4.886$$

```
1 F_stat2 = anova_tab2$statistic[1]
```

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

Life expectancy example: hypothesis test for freedom status (3/4)

6. Calculate the p-value

- As per Step 4, test statistic F can be modeled by a F -distribution with $df_1 = 2$ and $df_2 = n - 3$.
 - We had 105 countries' data, so $n = 105$
- **Option 1:** Use `pf()` and our calculated test statistic

```
1 # p-value is ALWAYS the right tail for F-test
2 n = nobs(model1)
→ pf(F_stat2, df1 = 2, df2 = n-3, lower.tail = FALSE)
```

```
[1] 0.009414324
```

- **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
freedom_status	2	415.9368	207.96841	4.885585	0.009414324
Residuals	102	4341.9116	42.56776	NA	NA

Life expectancy example: hypothesis test for freedom status (4/4)

7. Write a **conclusion**

We reject the null hypothesis that both coefficients are equal to 0 at the 5% significance level. There is sufficient evidence that there is association between life expectancy and the country's freedom status (p-value = 0.009).

freedom status explains enough
variation in life expectancy

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

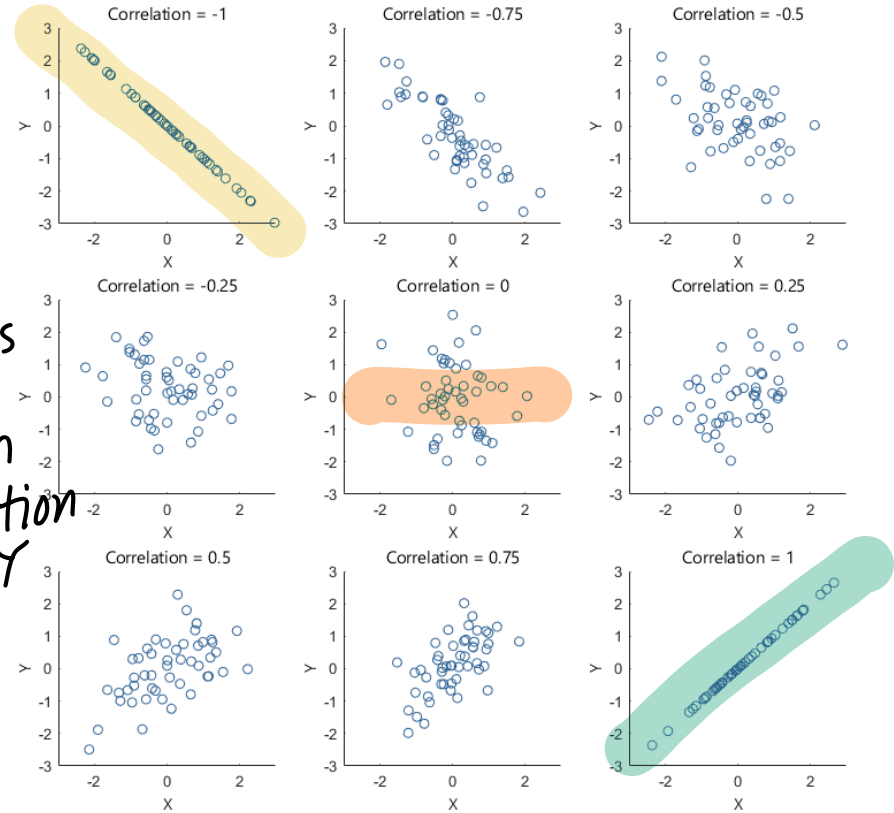
Correlation coefficient r can tell us about the strength of a relationship between two continuous variables

- 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 $Y \Rightarrow \beta_1 = 0$

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.

X does NOT explain variation in Y

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



Coefficient of determination: R^2

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

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

$$SSY = SSR + SSE$$

$$SSR = SSY - SSE$$

- R^2 is called the **coefficient of determination**.

Interpretation: The proportion of variation in the Y values explained by the regression model

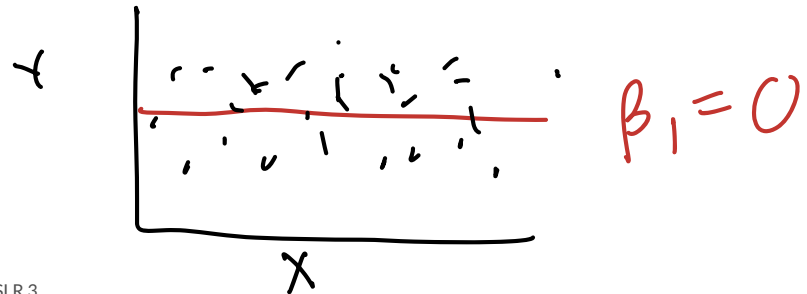
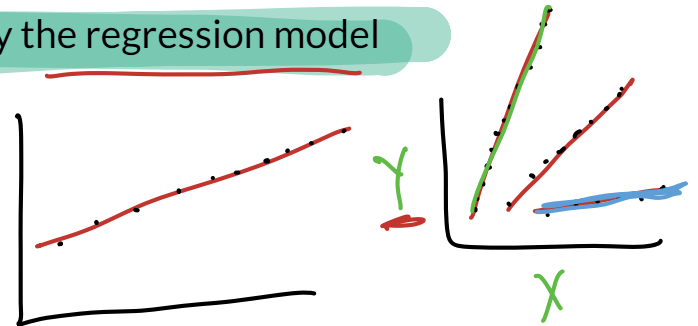
- R^2 measures the strength of the **linear** relationship between X and Y :

- $R^2 = 1$: 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 improve model fit over using \bar{Y} to model the Y values.



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

We can find the correlation coefficient and square it:

```
1 r = cor(x = gapm$life_exp,  
2       y = gapm$cell_phones_100,  
3       use = "complete.obs")  
4 r^2  
[1] 0.2299573
```

We can pull the coefficient of determination from the model summary:

```
1 model1_sum = summary(model1)  
2 r_squared = model1_sum$r_squared  
3 r_squared  
[1] 0.2299573
```

Interpretation

23% of the variation in countries' life expectancy is explained by the linear model with number of cell phones per 100 people 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

