Lesson 10: MLR: Using the F-test

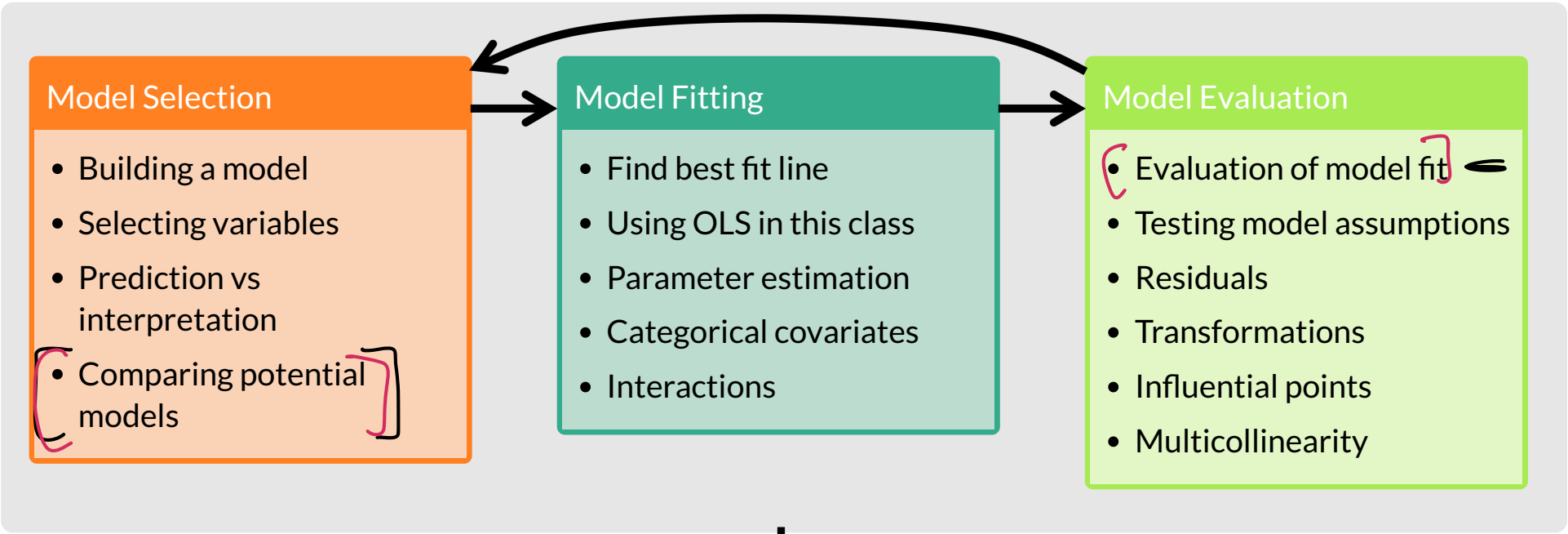
Nicky Wakim

2025-02-10

Learning Objectives

1. Understand the use of the general F-test and interpret what it measures.
2. Understand the context of the **Overall F-test**, conduct the needed hypothesis test, and interpret the results.
3. Understand the context of the **single covariate/coefficient F-test**, conduct the needed hypothesis test, and interpret the results.
4. Understand the context of the **group of covariates/coefficients F-test**, conduct the needed hypothesis test, and interpret the results.

Let's map that to our regression analysis process



Learning Objectives

1. Understand the use of the general F-test and interpret what it measures.
2. Understand the context of the **Overall F-test**, conduct the needed hypothesis test, and interpret the results.
3. Understand the context of the **single covariate/coefficient F-test**, conduct the needed hypothesis test, and interpret the results.
4. Understand the context of the **group of covariates/coefficients F-test**, conduct the needed hypothesis test, and interpret the results.

We must revisit our dear friend, the F-test!

https://www.writerswrite.co.za/foreshadowing/

Remember from Lesson 5: F-test vs. t-test for the population slope

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

SLR

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

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

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

- Note that the F-test does not support one-sided alternative tests, but the t-test does!

Remember from Lesson 5: 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!

We can extend this!!

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

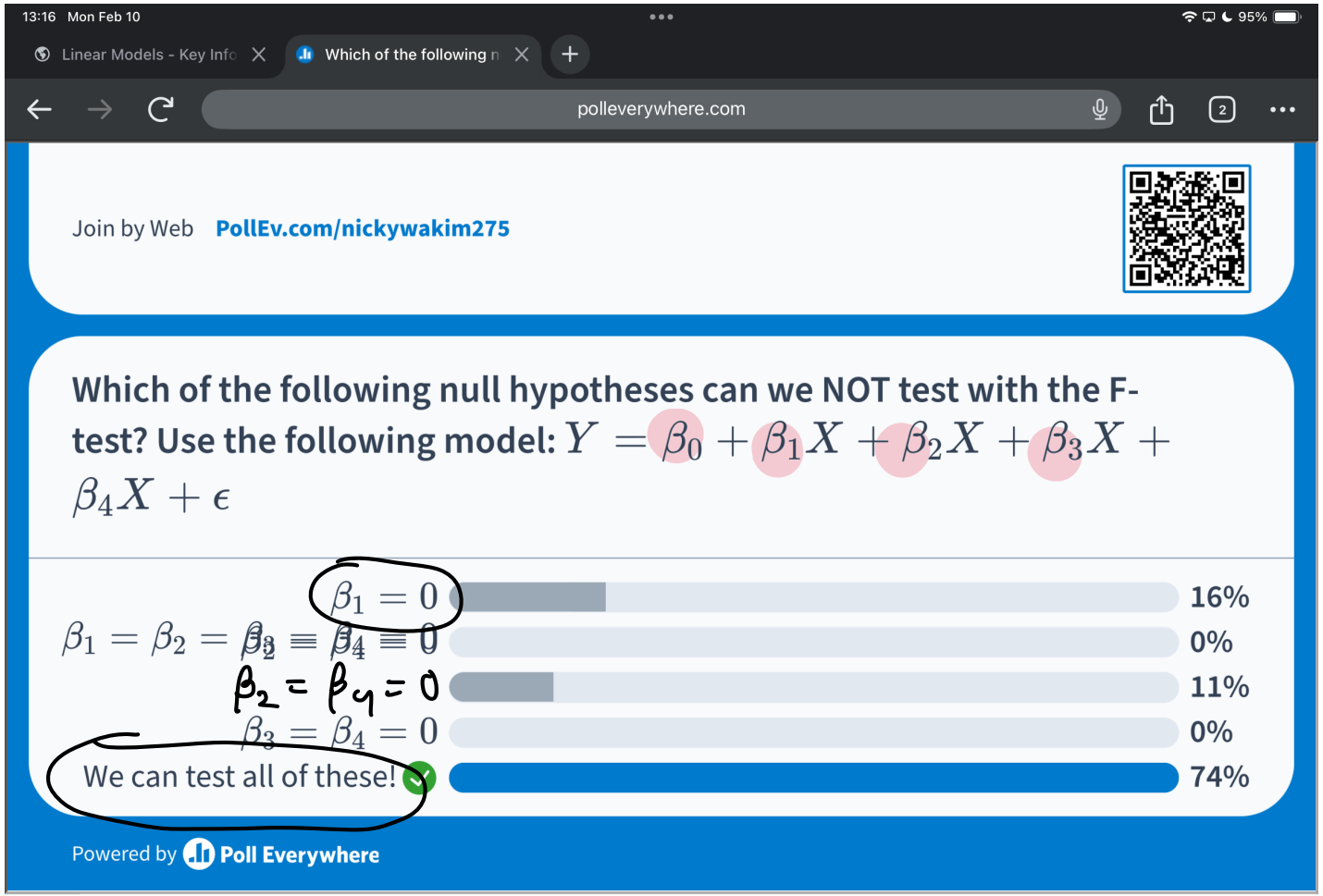
Null H_0	Alternative H_1
$\beta_1 = \beta_2 = 0$	$\beta_1 \neq 0$ and/or $\beta_2 \neq 0$

in a slightly different way…

Null model	Alternative* model
• $Y = \beta_0 + \epsilon$ • Smaller (reduced) 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

Poll Everywhere Question 1



F-test

- for single coeff
- for multiple coef
- restriction: two-sided sided test (aka $H_A: \neq$)

Building a very important toolkit: three types of tests

Overall test

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

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

Does the addition of one particular covariate (with a single coefficient) add significantly to the prediction of Y achieved by other covariates already present in the model?

Test for addition of group of variables' coefficient (covariate subset test)

Does the addition of some group of covariates (or one covariate with multiple coefficients) add significantly to the prediction of Y achieved by other covariates already present in the model?

Variation: Explained vs. Unexplained

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

$$SSY = SSR + SSE$$

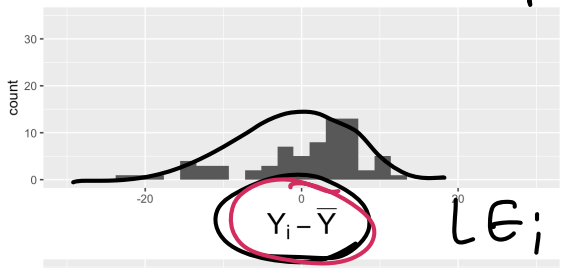
- $Y_i - \overline{Y}$ = the deviation of Y_i around the mean $\overline{Y}$
  - the **total** amount deviation
- $\widehat{Y}_i - \overline{Y}$ = the deviation of the fitted value $\widehat{Y}_i$ around the mean $\overline{Y}$
  - the amount deviation **explained** by the regression at $X_{i1}, \ldots, X_{ik}$
- $Y_i - \widehat{Y}_i$ = the deviation of the observation Y around the fitted regression line
  - the amount deviation **unexplained** by the regression at $X_{i1}, \ldots, X_{ik}$

Plot histogram of deviations for $LE = \beta_0 + \beta_1 FLR + \epsilon$

pop

- Code to make the below plots

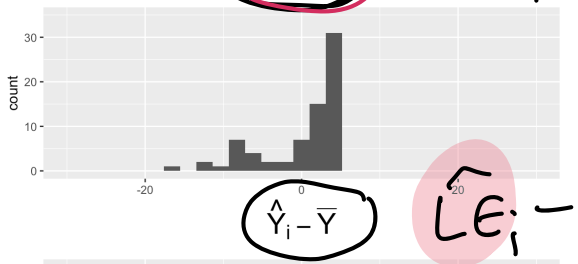
estimated model: $\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 FLR$



$LE_i - \overline{LE}$

$$SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 64.64$$

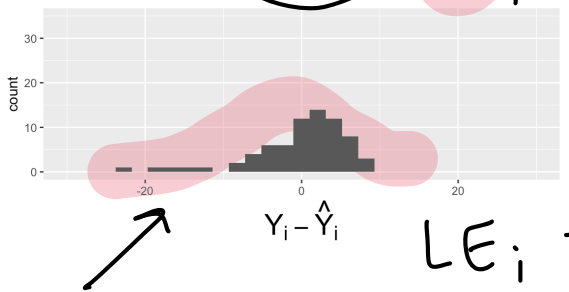
fixed



$\widehat{LE}_i - \overline{LE}$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 27.24$$

lower SSE, better model fit



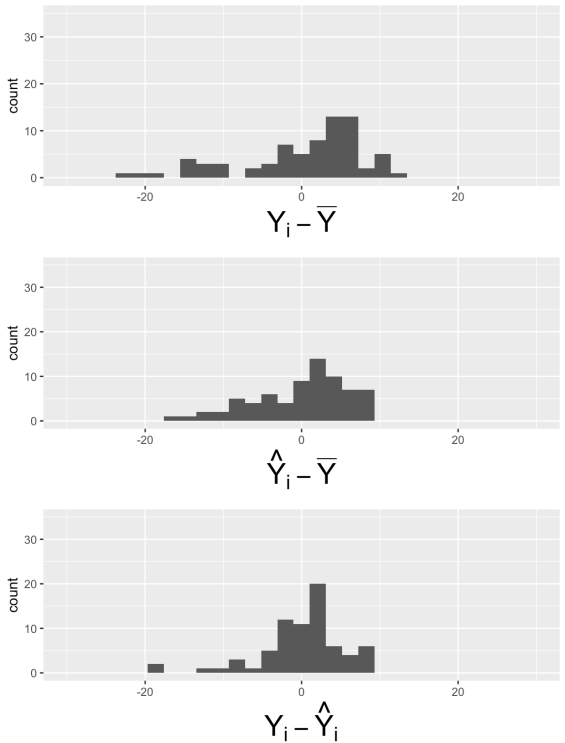
$LE_i - \widehat{LE}_i$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 37.39$$

variance of this is variance of LE NOT explained by model

Plot histogram of deviations for $LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \epsilon$

- Code to make the below plots



$$SSY = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 64.64$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = 36.39$$

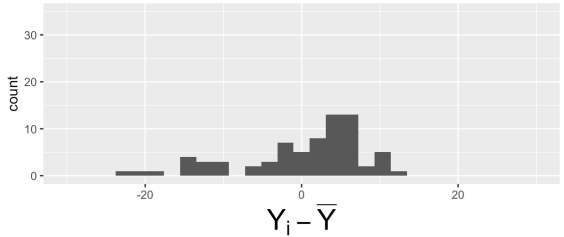
$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 28.25$$

What did you notice in the plots?

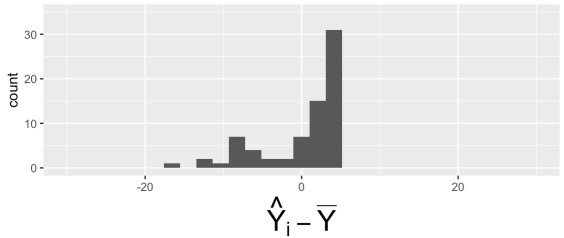
fixed

$SSY = SSR + SSE$

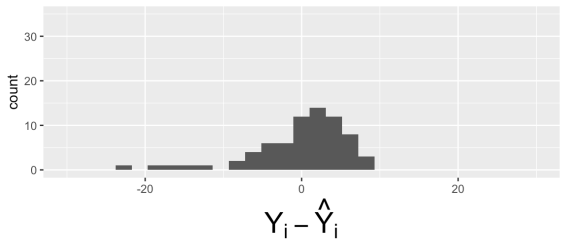
Simple Linear Regression



$SSY = 64.64$

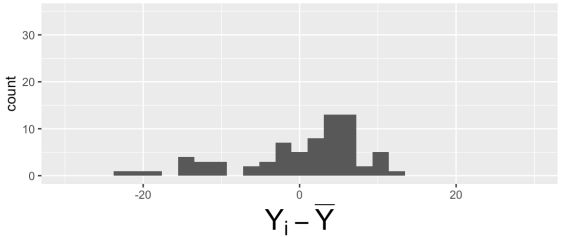


$SSR = 27.24$

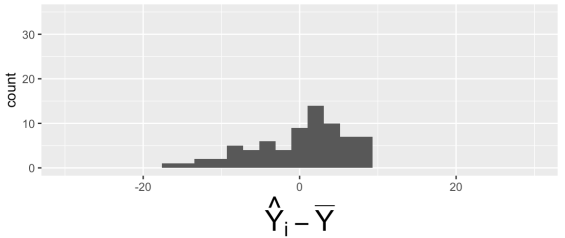


$SSE = 37.39$

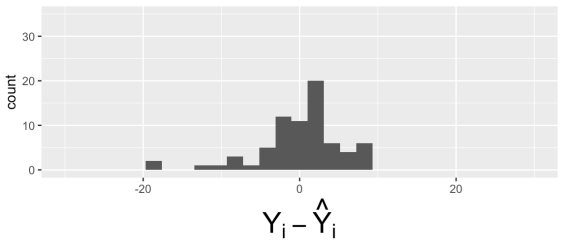
Multiple Linear Regression



$SSY = 64.64$



$SSR = 36.39$



$SSE = 28.25$

explains more variation of LE

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

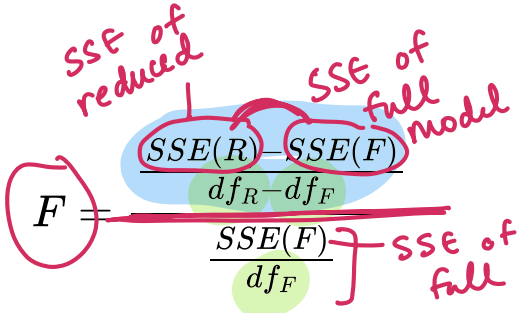
alt: $\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 FLR + \hat{\beta}_2 FS$

null: $\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 FLR$

Hyp test

$H_A: \beta_2 \neq 0$

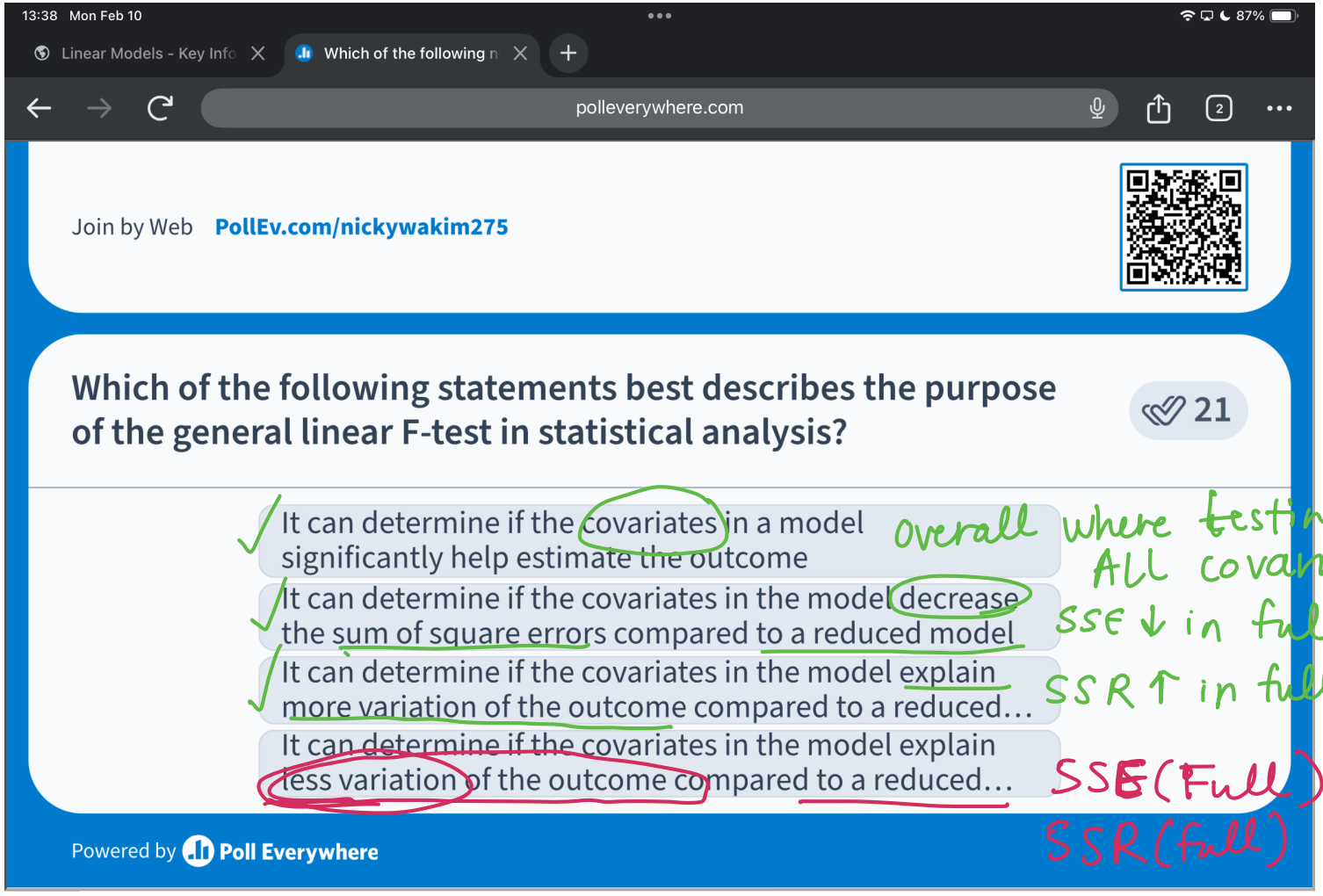
$H_0: \beta_2 = 0$



- $SSE(R) \geq SSE(F)$: b/c adding variable to model can only add info or nothing (will never take info away)
- Numerator measures difference in **unexplained** variation between the models
  - Big difference = added parameters greatly reduce the unexplained variation (increase explained variation) variable add A LOT of info
  - Smaller difference = added parameters don't reduce the unexplained variation variable adds little info
- Take ratio of difference to the unexplained variation in the full model

df: standardized by how many variables added → balance # added w/ variation explained

Poll Everywhere Question 2



overall where testing ALL covariates

SSE ↓ in full → compared to reduced

SSR ↑ in full

$SSE(Full) \leq SSE(Red)$

$SSR(full) \geq SSR(red)$

We will keep working with the MLR model from last class

New population model for example:

$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{FS} + \epsilon$$

```
1 # Fit regression model:
2 mr1 <- gapm_sub %>%
3   lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD)
4 tidy(mr1, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt_number
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	33.595	4.472	7.512	0.000	24.674	42.517
FemaleLiteracyRate	0.157	0.032	4.873	0.000	0.093	0.221
FoodSupplykcPPD	0.008	0.002	4.726	0.000	0.005	0.012

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{FLR} + \widehat{\beta}_2 \text{FS}$$

$$\widehat{\text{LE}} = 33.595 + 0.157\ \text{FLR} + 0.008\ \text{FS}$$

Learning Objectives

1. Understand the use of the general F-test and interpret what it measures.
2. Understand the context of the **Overall F-test**, conduct the needed hypothesis test, and interpret the results.
3. Understand the context of the **single covariate/coefficient F-test**, conduct the needed hypothesis test, and interpret the results.
4. Understand the context of the **group of covariates/coefficients F-test**, conduct the needed hypothesis test, and interpret the results.

Overall F-test

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

- For a general population MLR model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

up to k coefficients

We can create a hypothesis test for all the covariate coefficients...

Null H_0

$$\beta_1 = \beta_2 = \ldots = \beta_k = 0$$

Alternative H_1

At least one $\beta_j \neq 0$ (for $j = 1, 2, \ldots, k$)

Null / Smaller / Reduced model

$$Y = \beta_0 + \epsilon$$

intercept model

Alternative / Larger / Full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

Overall F-test: general steps for hypothesis test

1. Met underlying LINE assumptions EDA
2. State the null hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

vs. H_A : At least one $\beta_j \neq 0$, for $j = 1, 2, \ldots, k$

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 $df = k$ and denominator $df = n - k - 1$. (n = # obversation, k = # covariates)

5. Compute the value of the test statistic

The calculated **test statistic** is

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

6. Calculate the p-value

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

7. Write conclusion for hypothesis test

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

We (reject/fail to reject) the null hypothesis at the $100\alpha\%$ significance level. There is (sufficient/insufficient) evidence that at least one predictor's coefficient is not 0 (p-value = $P(F_{1,n-2} > F)$). explaining variation of outcome

Overall F-test: a word on the conclusion

- If H_0 is rejected, we conclude there is sufficient evidence that at least one predictor's coefficient is different from zero.
- Same as: at least one independent variable contributes significantly to the prediction of Y

- If H_0 is not rejected, we conclude there is insufficient evidence that at least one predictor's coefficient is different from zero.
- Same as: Not enough evidence that at least one independent variable contributes significantly to the prediction of Y

Let's think about our MLR example for life expectancy

Our proposed population model

$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{FS} + \epsilon$$

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{FLR} + \widehat{\beta}_2 \text{FS}$$
$$\widehat{\text{LE}} = 33.595 + 0.157\ \text{FLR} + 0.008\ \text{FS}$$

Our main question for the Overall F-test: Is the regression model containing female literacy rate and food supply useful in estimating countries' life expectancy?

Null / Smaller / Reduced model	Alternative / Larger / Full model
$LE = \beta_0 + \epsilon$	$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \epsilon$

Comparing the SSY, SSR, and SSE for reduced and full model

- Fit and get augmented values for reduced model:

```
mod_red1 = gapm_sub %>%
  lm(formula = LifeExpectancyYrs ~ 1)
aug_red1  = augment(mod_red1)
```

- Fit and get augmented values for full model:

```
mod_full1 = gapm_sub %>%
  lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD)
aug_full1  = augment(mod_full1)
```

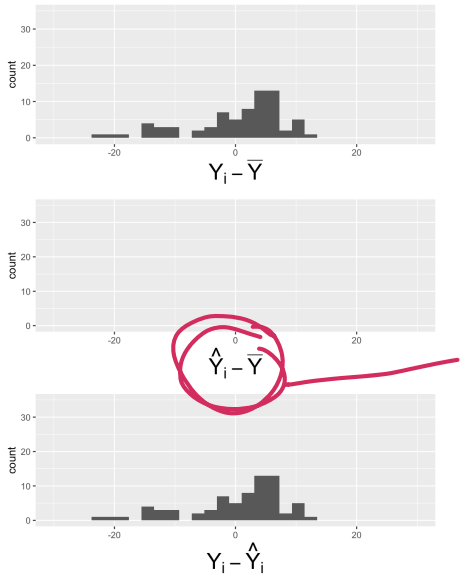
- Calculate the deviances for each model:

```
SS_df2 = gapm_sub %>% select(LifeExpectancyYrs) %>%
  mutate(SSY_diff_r1 = LifeExpectancyYrs - mean(LifeExpectancyYrs),
         SSR_diff_r1 = aug_red1$.fitted - mean(LifeExpectancyYrs),
         SSE_diff_r1 = aug_red1$.resid,
         SSY_diff_f1 = LifeExpectancyYrs - mean(LifeExpectancyYrs),
         SSR_diff_f1 = aug_full1$.fitted - mean(LifeExpectancyYrs),
         SSE_diff_f1 = aug_full1$.resid)
```

Comparing the SSY, SSR, and SSE for reduced and full model

Reduced / null model

$$LE = \beta_0 + \epsilon$$



$SSY = 64.64$

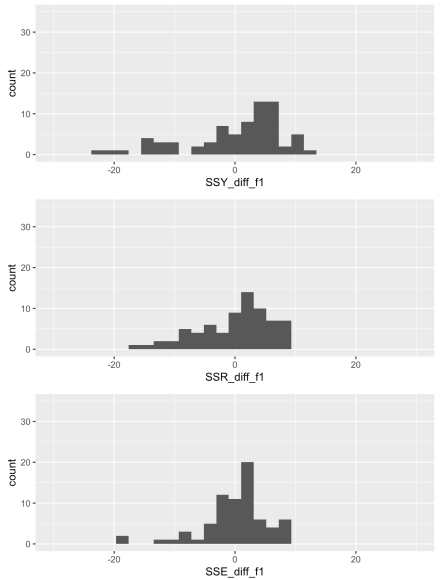
$\hat{\beta}_0 = \overline{LE} \Rightarrow \hat{LE} = \hat{\beta}_0 = \overline{LE}$

$SSR = 0$

$SSE = 64.64$

Full / Alternative model

$$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \epsilon$$



$SSY = 64.64$

$SSR = 36.39$

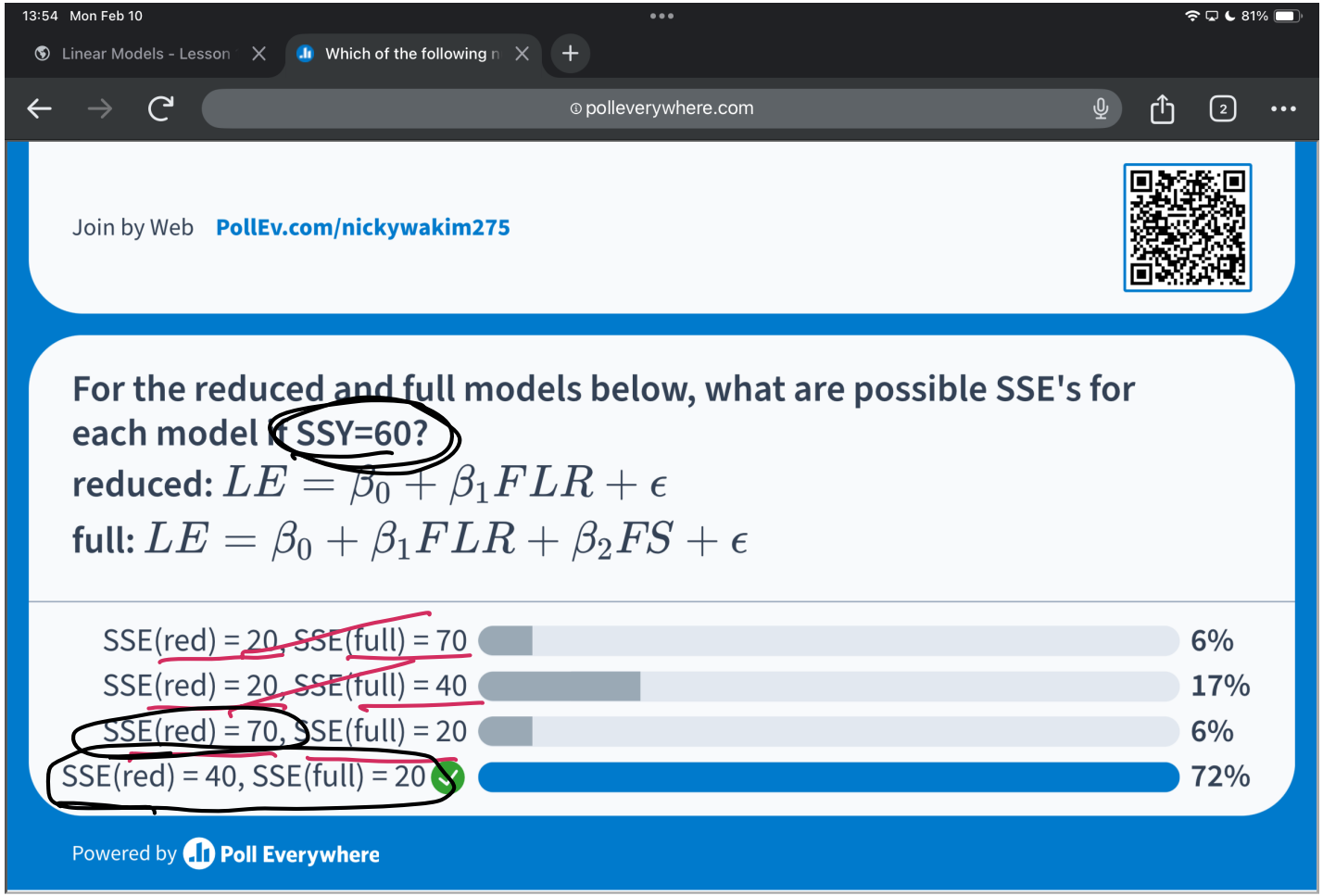
explained

$SSE = 28.25$

not explained

SSE(R) = 64.64 vs. SSE(full) = 28.25

Poll Everywhere Question 3



$SSE(red) \geq SSE(full)$

full has more variables, will explain more or same variance

$SSY = SSE + SSR$

$SSE \leq SSY$

$70 \not\leq 60$

So let's step through our hypothesis test (1/3)

1. Met underlying LINE assumptions ✓

2. State the null hypothesis

$$H_0 : \beta_1 = \beta_2 = 0$$
$$\text{vs. } H_A : \text{At least one } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

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 $df = k = 2$ and denominator $df = n - k - 1 = 72 - 2 - 1 = 69$. ($n$ = # obversation, k = # covariates)

# of coef testing

extra for intercept

So let's step through our hypothesis test (2/3)

5. Compute the value of the test statistic / 6. Calculate the p-value

The calculated **test statistic** is

$$F = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = 44.443$$

71 - 69 = 2

OR use ANOVA table:

tidying up output

```
anova(mod_red1, mod_full1) %>% tidy() %>% gt() %>% tab_options(table.font.size = 35)
```

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ 1	71.000	4,589.119	NA	NA	NA	NA
LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD	69.000	2,005.556	2.000	2,583.563	44.443	0.000

red

full

So let's step through our hypothesis test (3/3)

7. Write conclusion for hypothesis test

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that either countries' female literacy rate or the food supply (or both) contributes significantly to the prediction of life expectancy (p-value < 0.001).

Learning Objectives

1. Understand the use of the general F-test and interpret what it measures.
2. Understand the context of the **Overall F-test**, conduct the needed hypothesis test, and interpret the results.
3. Understand the context of the **single covariate/coefficient F-test**, conduct the needed hypothesis test, and interpret the results.
4. Understand the context of the **group of covariates/coefficients F-test**, conduct the needed hypothesis test, and interpret the results.

Covariate subset test: Single variable

do w/ t-test

Does the addition of one particular covariate of interest (a numeric covariate with only one coefficient) add significantly to the prediction of Y achieved by other covariates already present in the model?

- For a general population MLR model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_j X_j + \ldots + \beta_k X_k + \epsilon$$

We can create a hypothesis test for a single j covariate coefficient (where j can be any value $1, 2, \ldots, k$)...

Null H_0

$\beta_j = 0$

Alternative H_1

$\beta_j \neq 0$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_j X_j$ vs $\hat{Y} = \hat{\beta}_0$

Null / Smaller / Reduced model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

β_j gone

Alternative / Larger / Full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_j X_j + \ldots + \beta_k X_k + \epsilon$$

Single covariate F-test: general steps for hypothesis test (reference)

1. Met underlying LINE assumptions

2. State the null hypothesis

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

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 $df = k$ and denominator $df = n - k - 1$. (n = # obversation, k = # covariates)

5. Compute the value of the test statistic

The calculated **test statistic** is

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

6. Calculate the p-value

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

7. Write conclusion for hypothesis test

We (reject/fail to reject) the null hypothesis at the $100\alpha\%$ significance level. There is (sufficient/insufficient) evidence that predictor/covariate j significantly improves the prediction of Y, given all the other covariates are in the model (p-value = $P(F_{1,n-2} > F)$).

Let's think about our MLR example for life expectancy

Our proposed population model

$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{FS} + \epsilon$$

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{FLR} + \widehat{\beta}_2 \text{FS}$$
$$\widehat{\text{LE}} = 33.595 + 0.157\ \text{FLR} + 0.008\ \text{FS}$$

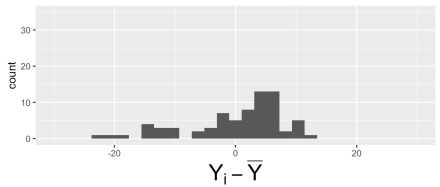
Our main question for the single covariate subset F-test: Is the regression model containing food supply improve the estimation of countries' life expectancy, given female literacy rate is already in the model?

Null / Smaller / Reduced model	Alternative / Larger / Full model
$LE = \beta_0 + \beta_1 FLR + \epsilon$	$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \epsilon$
$\beta_2 = 0$	$\beta_2 \neq 0$

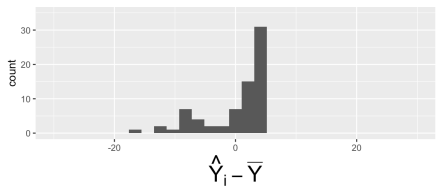
Comparing the SSY, SSR, and SSE for reduced and full model

Reduced / null model

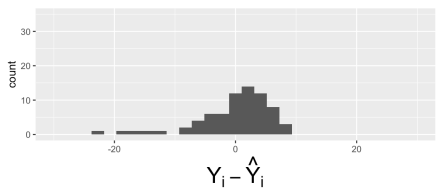
$$LE = \beta_0 + \beta_1 FLR + \epsilon$$



$SSY = 64.64$



$SSR = 27.24$

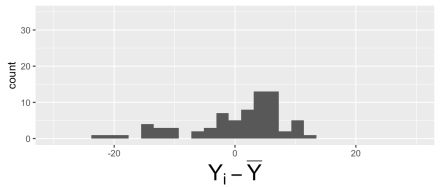


$SSE = 37.39$

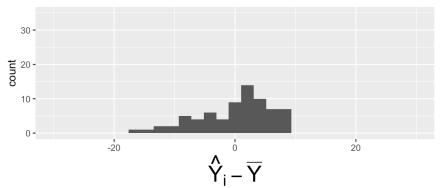
$df = 70$

Full / Alternative model

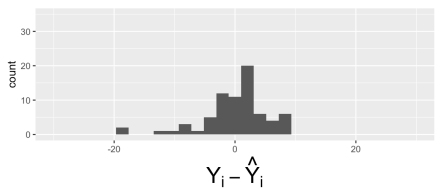
$$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \epsilon$$



$SSY = 64.64$



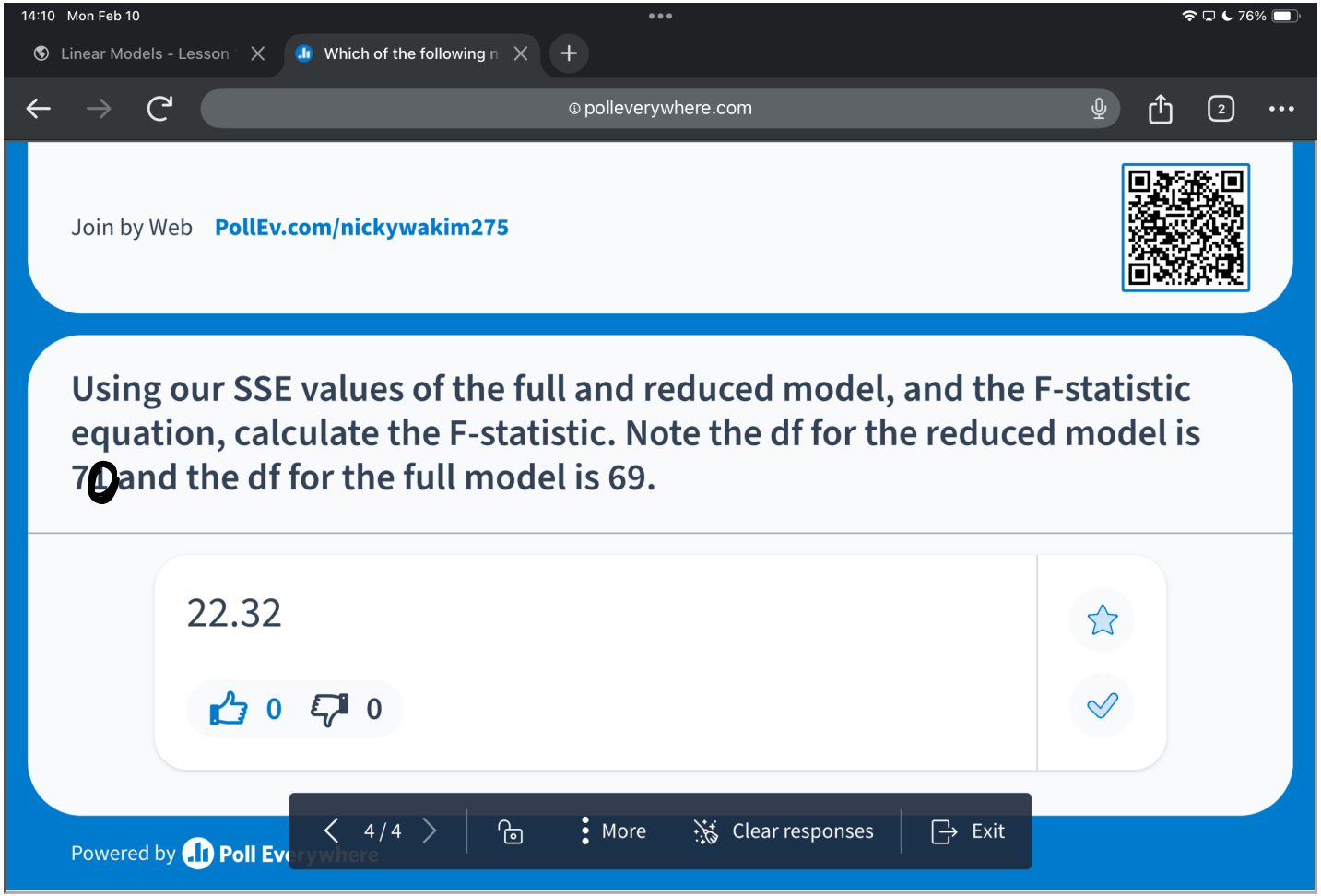
$SSR = 36.39$



$SSE = 28.25$

$df = 69$

Poll Everywhere Question 4



$$F = \frac{\frac{SSE(red) - SSE(f)}{df_R - df_F}}{\frac{SSE(ful)}{df_F}}$$

$$= \frac{\frac{37.39 - 28.25}{(70-69)}}{\frac{28.25}{69}}$$

$$\approx 22.33$$

So let's step through our hypothesis test (1/3)

1. Met underlying LINE assumptions

2. State the null hypothesis

$$H_0 : \beta_2 = 0$$
$$\text{vs. } H_A : \beta_2 \neq 0$$

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 $df = k = 1$ and denominator $df = n - k - 1 = 72 - 2 - 1 = 69$. ($n$ = # obversation, k = # covariates)

$\Delta df = \frac{df_R - df_F}{\text{diff in \# coeff}}$

So let's step through our hypothesis test (2/3)

5. Compute the value of the test statistic / 6. Calculate the p-value

The calculated **test statistic** is

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

ANOVA table:

```
1 anova(mod_red2, mod_full2) %>% tidy() %>% gt() %>% tab_options(table.font.size = 35
```

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FemaleLiteracyRate	70.000	2,654.875	NA	NA	NA	NA
LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD	69.000	2,005.556	1.000	649.319	22.339	0.000

So let's step through our hypothesis test (3/3)

7. Write conclusion for hypothesis test

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that countries' food supply contributes significantly to the prediction of life expectancy, given that female literacy rate is already in the model (p-value < 0.001).

Learning Objectives

1. Understand the use of the general F-test and interpret what it measures.
2. Understand the context of the **Overall F-test**, conduct the needed hypothesis test, and interpret the results.
3. Understand the context of the **single covariate/coefficient F-test**, conduct the needed hypothesis test, and interpret the results.
4. Understand the context of the **group of covariates/coefficients F-test**, conduct the needed hypothesis test, and interpret the results.

Covariate subset test: group of coefficients

1 multi level OR multiple covariates

Does the addition of some group of covariates of interest (or a multi-level categorical variable) add significantly to the prediction of Y obtained through other independent variables already present in the model?

including multi-level

- For a general population MLR model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

We can create a hypothesis test for a group of covariate coefficients (subset of many)... **For example...**

Null H_0	Alternative H_1
$\beta_1 = \beta_3 = 0$ (this can be any coefficients)	At least one $\beta_j \neq 0$ (for $j = 1, 3$)

$\beta_1 \neq 0$ and/or $\beta_3 \neq 0$

Null / Smaller / Reduced model	Alternative / Larger / Full model
$Y = \beta_0 + \beta_2 X_2 + \epsilon$	$Y = \beta_0 + \beta_1 X + \beta_2 X + \beta_3 X_3 + \epsilon$

Covariate subset F-test: general steps for hypothesis test (reference)

1. Met underlying LINE assumptions

2. State the null hypothesis

For example:

$$H_0 : \beta_1 = \beta_3 = 0$$
$$\text{vs. } H_A : \text{At least one } \beta_j \neq 0, \text{for } j = 1, 3$$

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 $df = k$ and denominator $df = n - k - 1$. (n = # obversation, k = # covariates)

5. Compute the value of the test statistic

The calculated **test statistic** is

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

6. Calculate the p-value

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

7. Write conclusion for hypothesis test

We (reject/fail to reject) the null hypothesis at the $100\alpha\%$ significance level. There is (sufficient/insufficient) evidence that predictors/covariates 1, 3 significantly improve the prediction of Y, given all the other covariates are in the model (p-value = $P(F_{1,n-2} > F)$).

We need to slightly alter our MLR example for life expectancy

Our proposed population model to include water source percent (WS):

$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{FS} + \beta_3 WS + \epsilon$$

- We don't have a fitted multiple regression model for this yet!

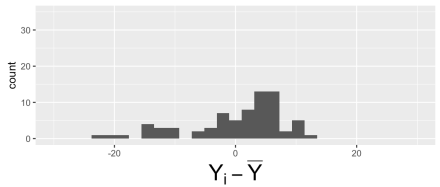
Our main question for the group covariate subset F-test: Is the regression model containing food supply and water source percent improve the estimation of countries' life expectancy, given percent female literacy rate is already in the model?

Null / Smaller / Reduced model	Alternative / Larger / Full model
$LE = \beta_0 + \beta_1 FLR + \epsilon$	$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \beta_3 WS + \epsilon$

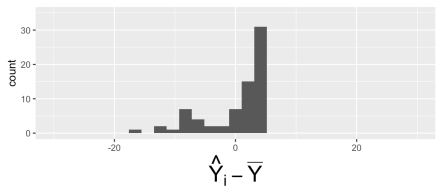
Comparing the SSY, SSR, and SSE for reduced and full model

Reduced / null model

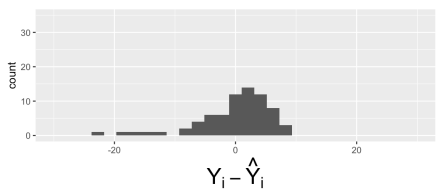
$$LE = \beta_0 + \beta_1 FLR + \epsilon$$



$SSY = 64.64$



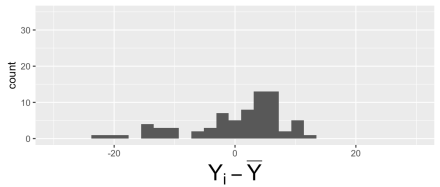
$SSR = 27.24$



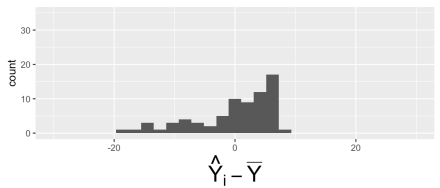
$SSE = 37.39$

Full / Alternative model

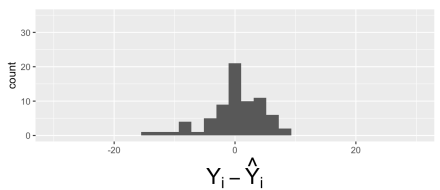
$$LE = \beta_0 + \beta_1 FLR + \beta_2 FS + \beta_3 WS + \epsilon$$



$SSY = 64.64$



$SSR = 43.26$



$SSE = 21.38$

So let's step through our hypothesis test (1/3)

1. Met underlying LINE assumptions

2. State the null hypothesis

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

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 $df = k = 2$ and denominator $df = n - k - 1 = 72 - 3 - 1 = 68$. ($n$ = # obversation, k = # covariates)

testing

in full: FLR, FS, WS → 3 conf.

So let's step through our hypothesis test (2/3)

5. Compute the value of the test statistic / 6. Calculate the p-value

The calculated **test statistic** is

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

ANOVA table:

```
1 anova(mod_red3, mod_full3) %>% tidy() %>% gt() %>% tab_options(table.font.size = 35
```

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FemaleLiteracyRate	70.000	2,654.875	NA	NA	NA	NA
LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD + WaterSourcePrct	68.000	1,517.916	2.000	1,136.959	25.467	0.000

So let's step through our hypothesis test (3/3)

7. Write conclusion for hypothesis test

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that countries' food supply or water source (or both) contribute significantly to the prediction of life expectancy, given that female literacy rate is already in the model (p-value < 0.001).

Other ways to word the hypothesis tests (reference)

- Single covariate subset F-test
  - H_0 : X^* does not significantly improve the prediction of Y, given that $X_1, X_2, \ldots, X_p$ are already in the model
  - H_A : X^* significantly improves the prediction of Y, given that $X_1, X_2, \ldots, X_p$ are already in the model
- Group covariate subset F-test
  - H_0 : The addition of the s variables $X_1^*, X_2^*, \ldots, X_s^*$ does not significantly improve the prediction of Y, given that $X_1, X_2, \ldots, X_q$ are already in the model
  - H_A : The addition of the s variables $X_1^*, X_2^*, \ldots, X_s^*$ significantly improves the prediction of Y, given that $X_1, X_2, \ldots, X_q$ are already in the model

null: $\beta_1 = 0$ alt: $\beta_1 \neq 0$

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