Lesson 10: MLR: Using the F-test

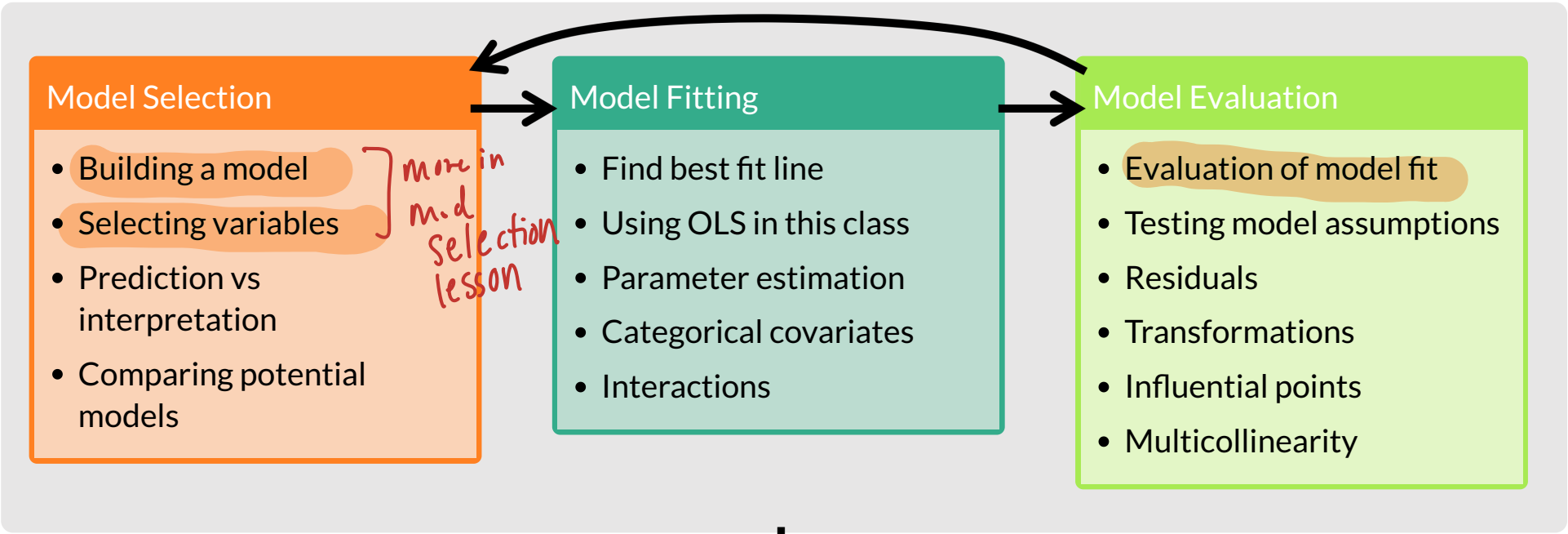
Nicky Wakim

2026-02-18

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



How do we estimate the model parameters?

- We need to estimate the population model coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$
- This can be done using the **ordinary least-squares method**
  - Find the $\widehat{\beta}$ values that **minimize** the sum of squares due to error (SSE)

residuals

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

$\hat{\epsilon}_i$ fitted residual for observation i

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

obs Y — fitted Y

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

$$SSE = \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \ldots + \widehat{\beta}_1 X_{ik}))^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \ldots - \widehat{\beta}_1 X_{ik})^2$$

will compare SSE's from diff models

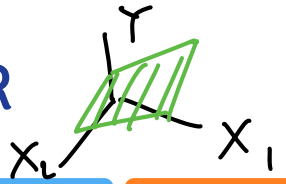
SLR

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

MLR

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots + \hat{\beta}_K X_K$$

LINE model assumptions in MLR



[L] **Linearity** of relationship between variables

The mean value of Y given any combination of $X_1, X_2, \ldots, X_k$ values, is a linear function of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$:

$$\mu_{Y|X_1,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

[I] **Independence** of the Y values

Observations $(X_1, X_2, \ldots, X_k, Y)$ are independent from one another

[N] **Normality** of the Y's given X (residuals)

Y has a normal distribution for any any combination of $X_1, X_2, \ldots, X_k$ values

- Thus, the residuals are normally distributed

[E] **Equality** of variance of the residuals (homoscedasticity)

The variance of Y is the same for any any combination of $X_1, X_2, \ldots, X_k$ values

$$\sigma^2_{Y|X_1,X_2,\ldots,X_k} = Var(Y|X_1, X_2, \ldots, X_k) = \sigma^2$$

residuals have constant variance

Summary of the LINE assumptions

- Equivalently, the **residuals** are independently and identically distributed (iid):
  - normal
  - with mean 0 and
  - constant variance σ^2

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Residuals are still $\hat{\epsilon}_i = Y_i - \widehat{Y}_i$ for each observation
  - It's just that $\widehat{Y}_i$ is now calculated with many covariates $(X_1, X_2, \ldots, X_k)$

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.

Remember from Lesson 6: 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!

Remember from Lesson 6: Planting a seed about the F-test

We can think about the hypothesis test for the slope…

Null H_0	Alternative H_1
$\beta_1 = 0$	$\beta_1 \neq 0$

in a slightly different way…

Null model ($\beta_1 = 0$)	Alternative model ($\beta_1 \neq 0$)
• $Y = \beta_0 + \epsilon$ • Smaller (reduced) model	• $Y = \beta_0 + \beta_1 X + \epsilon$ • 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!

Remember from Lesson 6: We can extend this!!

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

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

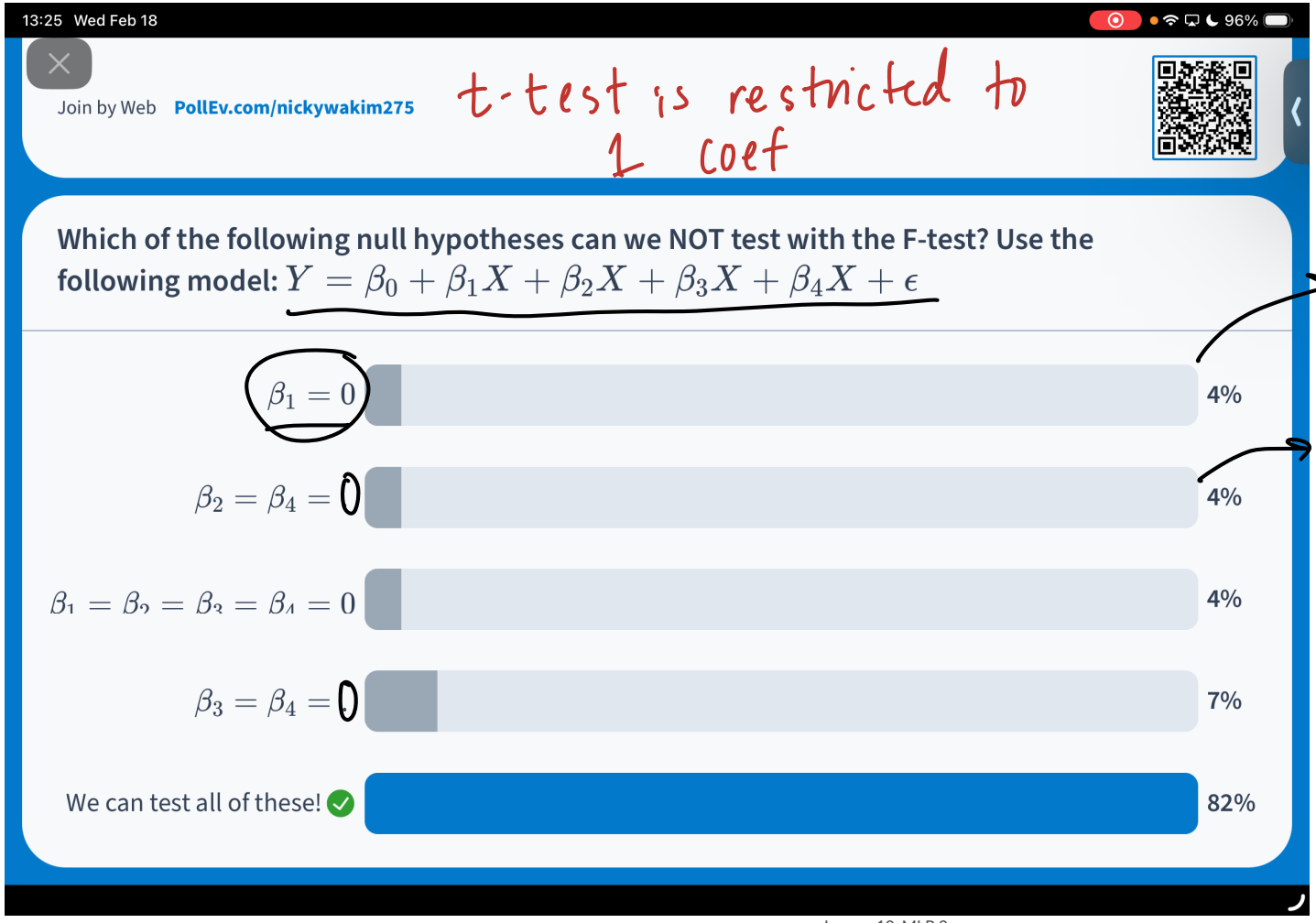
Null H_0	Alternative H_1
$\beta_1 = \beta_2 = 0$	$\beta_1 \neq 0$ and/or $\beta_2 \neq 0$

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



$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$

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

Variation: Explained vs. Unexplained

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

$$SSY = SSR + SSE$$

errors

R for regression

- $Y_i - \overline{Y}$ = 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 $X_{i1}, \ldots, X_{ik}$

Y_i obs Y

$\overline{Y}$ overall mean Y

$\widehat{Y}_i$ expected/fitted Y (or mean Y given X)

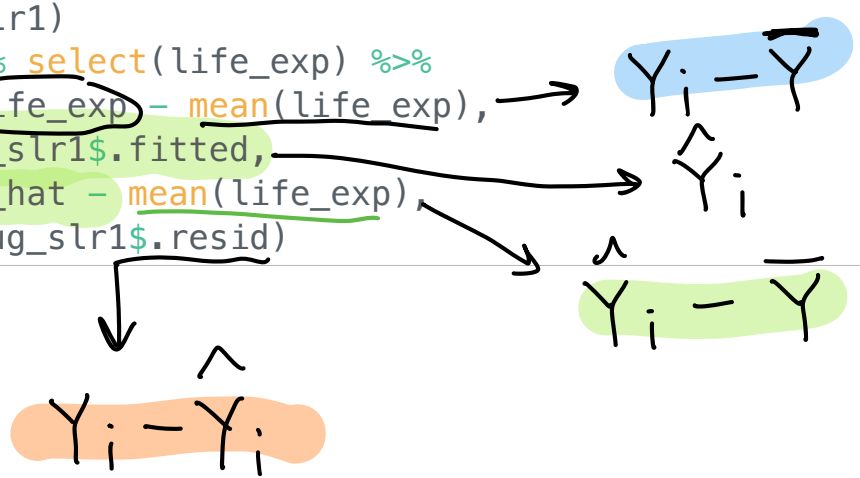
SLR: Another way to think of SSY, SSR, and SSE

- Let's create a data frame of each component within the SS's
  - Deviation in SSY: $Y_i - \overline{Y}$ → for every country
  - Deviation in SSR: $\widehat{Y}_i - \overline{Y}$
  - Deviation in SSE: $Y_i - \widehat{Y}_i$

each obs has each deviation

- Using our simple linear regression model as an example:

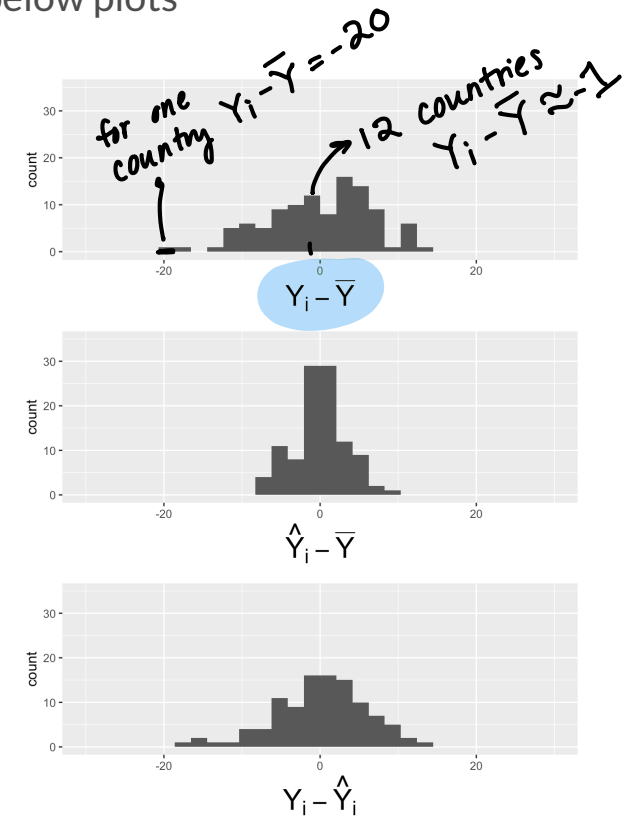
```
slr1 <- gapm %>%
  lm(formula = life_exp ~ cell_phones_100)
aug_slr1 = augment(slr1)
SS_dev_slr = gapm %>% select(life_exp) %>%
  mutate(SSY_dev = life_exp - mean(life_exp),
         y_hat = aug_slr1$.fitted,
         SSR_dev = y_hat - mean(life_exp),
         SSE_dev = aug_slr1$.resid)
```



SLR: Plot the components of each sum of squares → $\widehat{LE} = \hat{\beta}_0 + \hat{\beta} CP$

- Code to make the below plots

SSY for this dataset (and w/ LE as Y) will ALWAYS be 45.75



$$SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 45.75$$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 10.52$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 35.23$$

better model means sum moves from SSE to SSR (more explained)

$10.52 + 35.23 = 45.75$

MLR: Another way to think of SSY, SSR, and SSE

- Let's create a data frame of each component within the SS's
  - Deviation in SSY: $Y_i - \overline{Y}$
  - Deviation in SSR: $\widehat{Y}_i - \overline{Y}$
  - Deviation in SSE: $Y_i - \widehat{Y}_i$
- Using our simple linear regression model as an example:

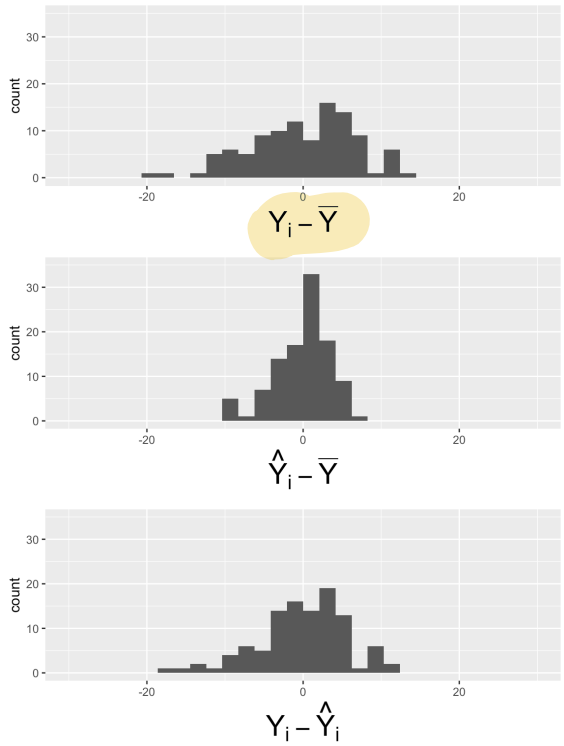
$$\widehat{LE} = \hat{\beta}_0 + \hat{\beta}_1 CP + \hat{\beta}_2 VR$$

```
1 mlr1 <- gapm %>%
2   lm(formula = life_exp ~ cell_phones_100 + vax_rate)
3 aug_mlr1 = augment(mlr1)
4 SS_df = gapm %>% select(life_exp) %>%
5   mutate(SSY_dev = life_exp - mean(life_exp),
6          y_hat = aug_mlr1$.fitted,
7          SSR_dev = y_hat - mean(life_exp),
8          SSE_dev = aug_mlr1$.resid)
```

MLR: Plot the components of each sum of squares

$\widehat{CE} = \hat{\beta}_0 + \hat{\beta}_1 CP + \hat{\beta}_2 VR$

- Code to make the below plots



$$SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 45.75$$

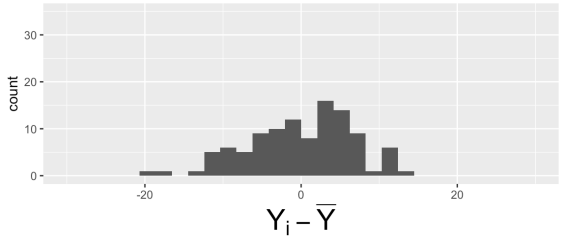
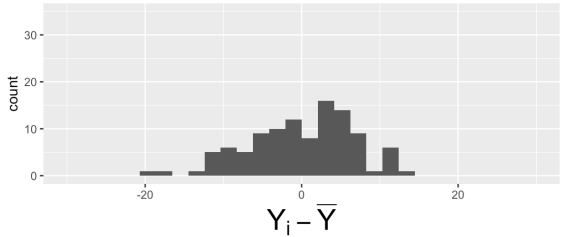
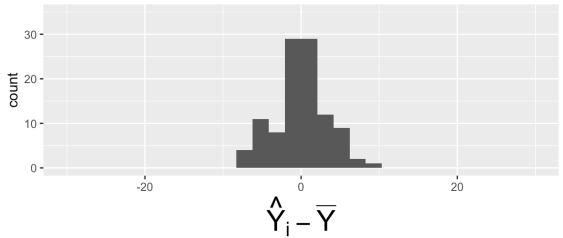
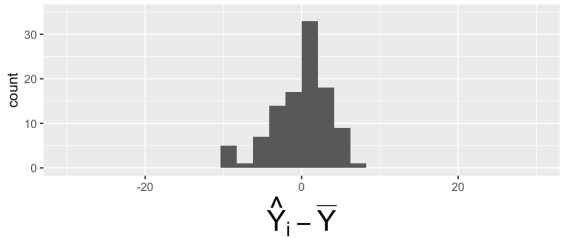
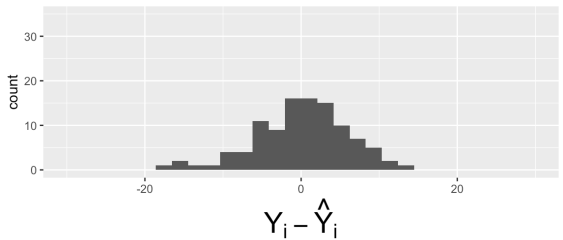
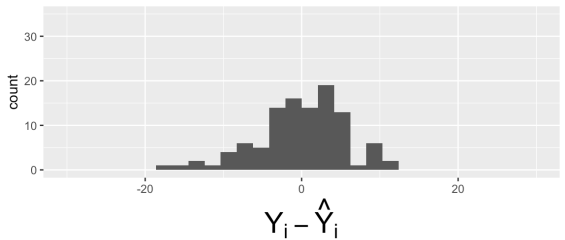
$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 12.28$$

↑ explaining more var in this model than SLR model

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 33.47$$

What did you notice in the plots?

$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \varepsilon$ alt: $\beta_2 \neq 0$

Simple Linear Regression (CP only, null: $\beta_2 = 0$)		Multiple Linear Regression (CP + VR)	
	$SSY = 45.75$		$SSY = 45.75$
	$SSR = 10.52$		$SSR = 12.28$
	$SSE = 35.23$		$SSE = 33.47$

- With F-test: we can determine if model fit is better by comparing the SSE's of different models

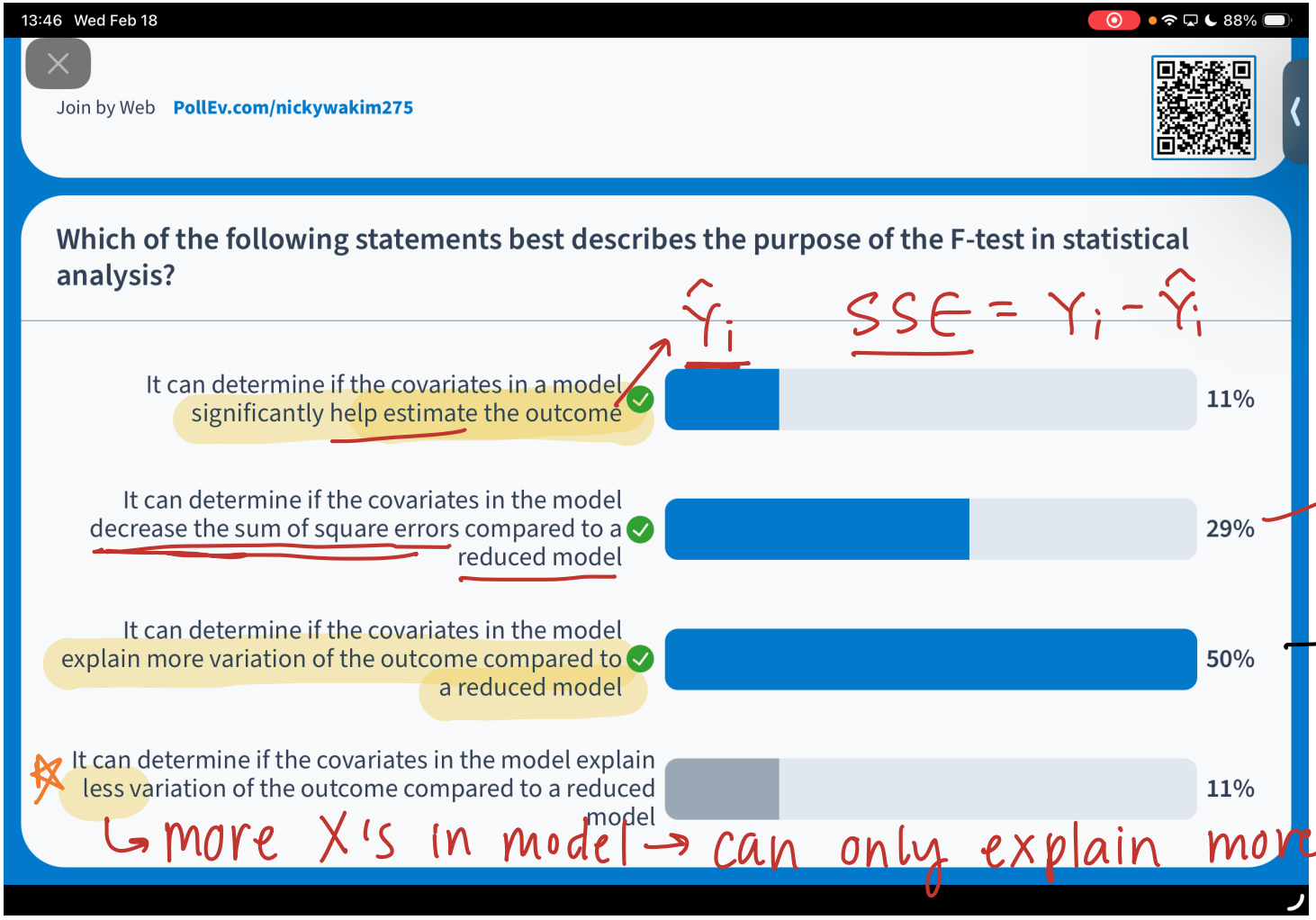
When running a F-test for linear models...

- We need to define a larger, full model (more parameters) → $LE = \beta_0 + \beta_1 CP + \beta_2 VR + \varepsilon$
- We need to define a smaller, reduced model (fewer parameters) → $LE = \beta_0 + \beta_1 CP + \varepsilon$
- 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

$$F = \frac{\frac{SSE_{red} - SSE_{full}}{df_{red} - df_{full}}}{\frac{SSE_{full}}{df_{full}}}$$

- $SSE(R) \geq SSE(F)$ → more covariates = explaining more or no additional variation
- 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 (does not inc explained variation very much)
- Take ratio of difference to the unexplained variation in the full model

Poll Everywhere Question 2



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

math def of F-stat

$SSE + SSR = SSY$

↓ ↑ explaining more variation

↳ more X's in model → can only explain more OR 0 additional variation

We will keep working with the MLR model from last class

New population model for example:

$$\text{LE} = \beta_0 + \beta_1 \text{CP} + \beta_2 \text{VR} + \epsilon$$

Cell phones

vax rate

```
1 mlr1 <- gapm %>% lm(formula = life_exp ~ cell_phones_100 + vax_rate)
2 tidy(mlr1, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt_number(decima
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	46.833	6.042	7.751	0.000	34.848	58.818
cell_phones_100	0.075	0.018	4.074	0.000	0.039	0.112
vax_rate	0.168	0.073	2.318	0.022	0.024	0.312

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{CP} + \widehat{\beta}_2 \text{VR}$$

$$\widehat{\text{LE}} = 46.833 + 0.075\ \text{CP} + 0.168\ \text{VR}$$

Building a very important toolkit: three types of tests

LOB 2

Overall test

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

LOB 3

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?

LOB 4

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?

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 covariates

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

Null H_0	Alternative H_1
$\beta_1 = \beta_2 = \ldots = \beta_k = 0$	At least one $\beta_j \neq 0$ (for $j = 1, 2, \ldots, k$)

Null / Smaller / Reduced model	Alternative / Larger / Full model
$Y = \beta_0 + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$

Overall F-test: general steps for hypothesis test

1. 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:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$\text{vs. } H_A : \text{At least one } \beta_j \neq 0, \text{for } j = 1, 2, \ldots, k$$

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~~ coefficients excluding β_0)

multi level cat covariate → need multiple coefficients to rep in model

5. Calculate 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 a conclusion
   - Reject: $P(F_{k,n-k-1} > F) < \alpha$

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

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

does NOT necessarily mean ALL do

- 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

↳ similar, but not technically same, as none contribute significantly

Let's think about our MLR example for life expectancy

Our proposed population model

$$\text{LE} = \beta_0 + \beta_1 \text{CP} + \beta_2 \text{VR} + \epsilon$$

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{CP} + \widehat{\beta}_2 \text{VR}$$
$$\widehat{\text{LE}} = 46.833 + 0.075\ \text{CP} + 0.168\ \text{VR}$$

Our main question for the Overall F-test: Is the regression model containing cell phones and vaccination rate useful in estimating countries' life expectancy?

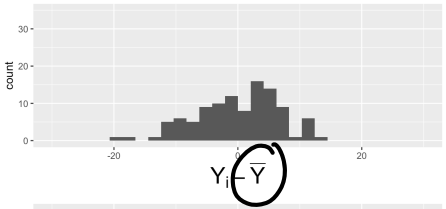
Null / Smaller / Reduced model	Alternative / Larger / Full model
$LE = \beta_0 + \epsilon$	$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \epsilon$
$\beta_1 = 0$ & $\beta_2 = 0$	$\beta_1 \neq 0$ and/or $\beta_2 \neq 0$

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

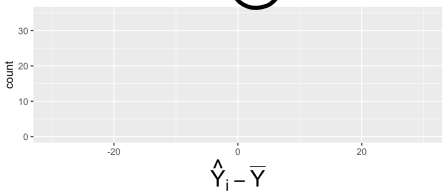
Reduced / null model

$$LE = \beta_0 + \epsilon$$

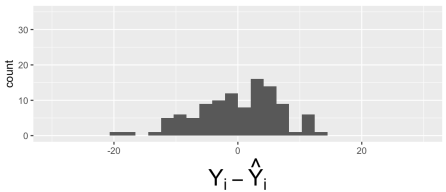
$\hat{\beta}_0 = \overline{LE}$



$SSY = 45.75$



$SSR = 0$

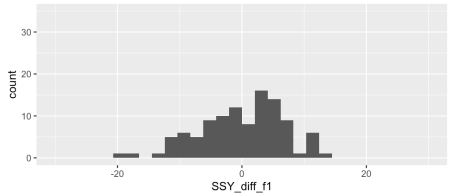


$SSE = 45.75$

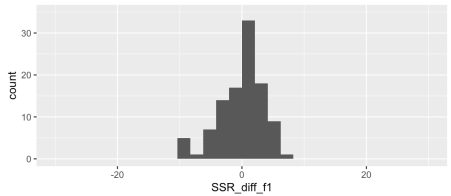
$SSE = SSY$

Full / Alternative model

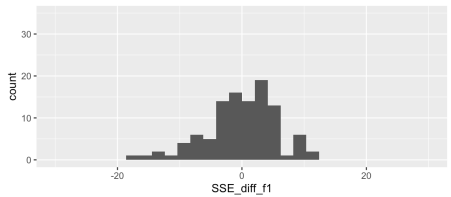
$$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \epsilon$$



$SSY = 45.75$



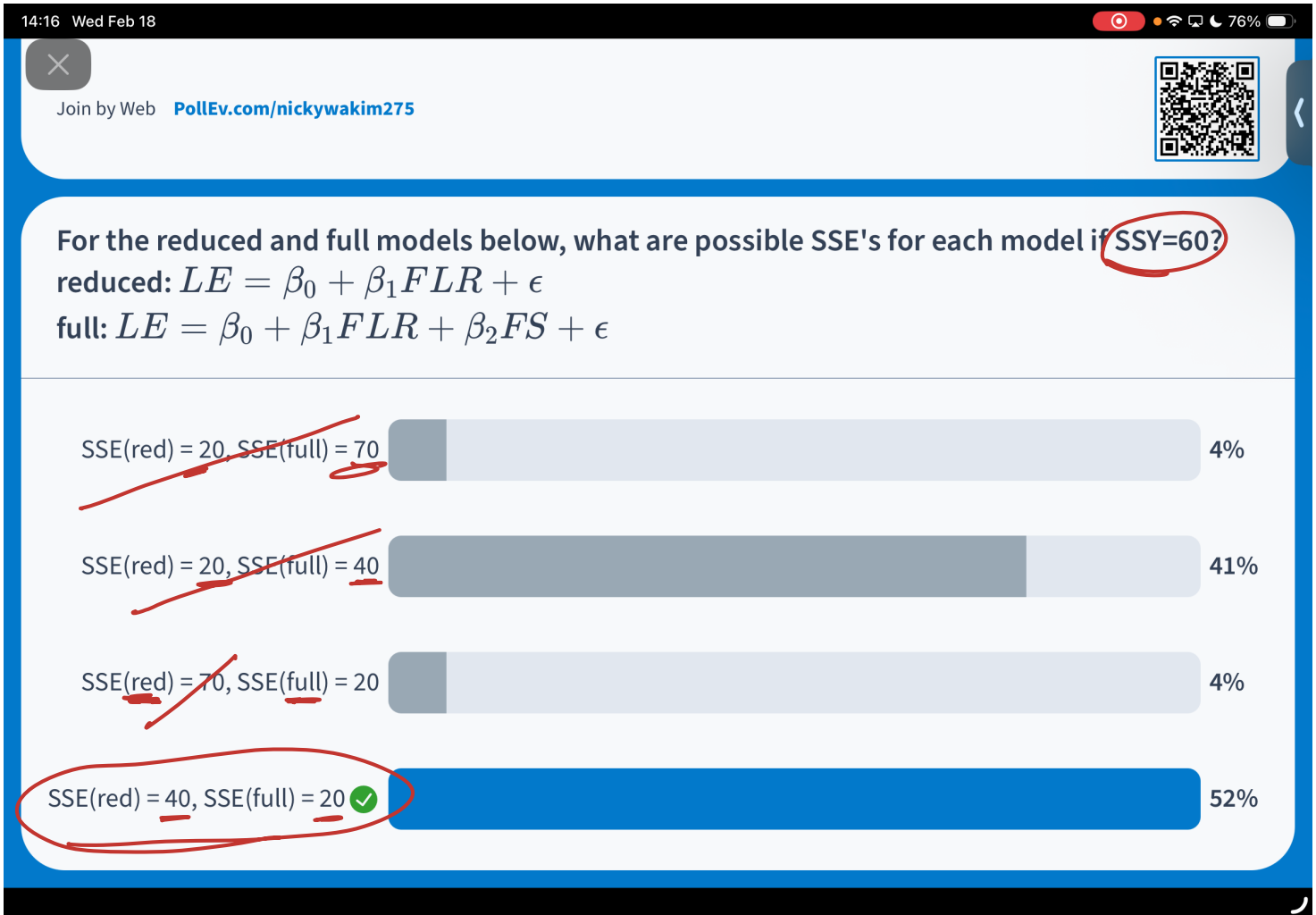
$SSR = 12.28$



$SSE = 33.47$

less unexplained

Poll Everywhere Question 3



$SSE_{red} \geq SSE_{full}$

full will always explain more (or no add) variation

$SSY = SSE + SSR$

$SSY = 60$

$SSE \leq 60$

$SSR \leq 60$

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

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**

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

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 $df = n - k - 1 = 105 - 2 - 1 = 102$. ($n$ = # obversation, k = # covariates)

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

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

The calculated **test statistic** is

$$F = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} = 18.72$$

OR use ANOVA table:

```
anova(mod_red1, mod_full1) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 35) %>% fmt_number(decimals = 3)
```

term	df.residual	rss	df	sumsq	statistic	p.value
life_exp ~ 1	104.000	4,757.848	NA	NA	NA	NA
life_exp ~ cell_phones_100 + vax_rate	102.000	3,480.371	2.000	1,277.478	18.720	0.000

$LE = \beta_0 + \varepsilon$

$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \varepsilon$

F

$P(F_{2,102} > 18.72)$

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

7. Write a **conclusion**

pval < 0.05

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that either countries' number of cell phones or vaccination rate (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

→ could be mult coef IF multi-level cat variable

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	Alternative H_1
$\beta_j = 0$	$\beta_j \neq 0$

if ML cat var: $\beta_j = 0$, $\beta_{j+1} = 0$, $\beta_{j+2} = 0$

Null / Smaller / Reduced model	Alternative / Larger / Full model
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_j X_j + \ldots + \beta_k X_k + \epsilon$

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

1. 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**

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

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

K > 1 for multikat

coef

k is typically 1 (cont or binary var)

5. Calculate 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 a **conclusion**

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

$$\text{LE} = \beta_0 + \beta_1 \text{CP} + \beta_2 \text{VR} + \epsilon$$

$LE = \beta_0 + \beta_1 CP +$
$\beta_2 I(FS = PF) +$
$\beta_3 I(FS = F)$
$+ \varepsilon$

Fitted multiple regression model:

$$\widehat{\text{LE}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{CP} + \widehat{\beta}_2 \text{VR}$$

$$\widehat{\text{LE}} = 33.595 + 0.157\ \text{CP} + 0.008\ \text{VR}$$

null: $\beta_2 = 0$ & $\beta_3 = 0$

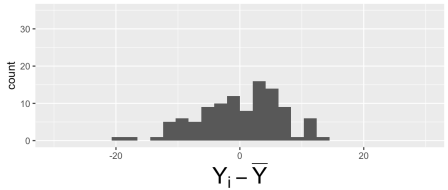
Our main question for the single covariate subset F-test: Does the regression model containing vaccination rate improve the estimation of countries' life expectancy, given cell phones per 100 people is already in the model?

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

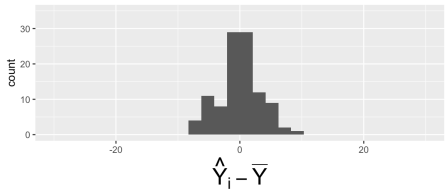
Comparing the SSY, SSR, and SSE for reduced and full model

Reduced / null model

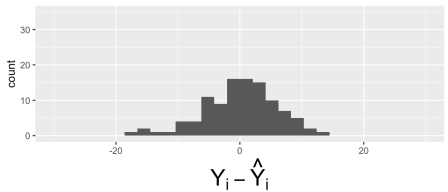
$$LE = \beta_0 + \beta_1 CP + \epsilon$$



$SSY = 45.75$



$SSR = 10.52$

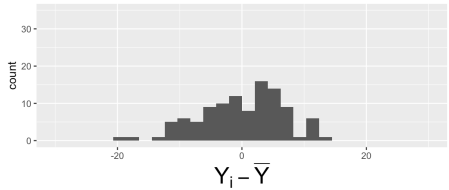


$SSE = 35.23$

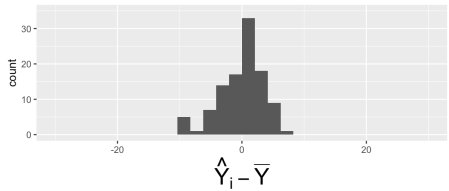
$df = 103$

Full / Alternative model

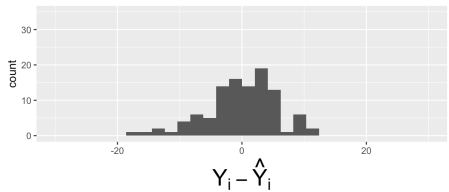
$$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \epsilon$$



$SSY = 45.75$



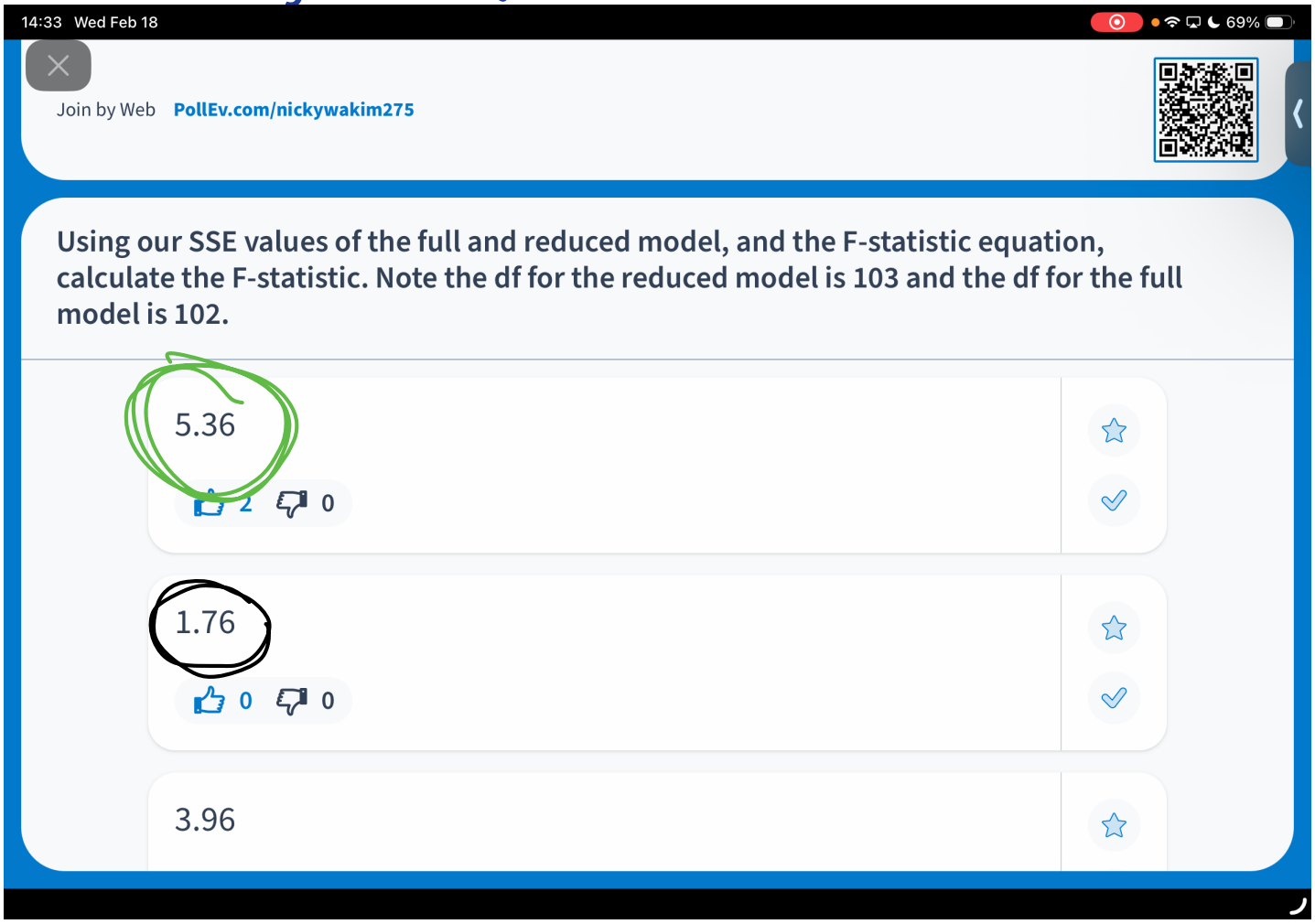
$SSR = 12.28$



$SSE = 33.47$

$df = 102$

Poll Everywhere Question 4



$$F = \frac{\frac{SSE_{red} - SSE_{full}}{df_{red} - df_{full}}}{\frac{SSE_{full}}{df_{full}}}$$

$$= \frac{\frac{35.23 - 33.47}{103 - 102}}{\frac{33.47}{102}}$$

$$= \frac{1.78}{0.328}$$

$$= 5.36$$

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

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**

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

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 $df = n - k - 1 = 105 - 1 - 1 = 103$. ($n$ = # obversation, k = # ~~covariates~~ Coef)

df red - df full
103 - 102

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

5. Calculate 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:

```
anova(mod_red2, mod_full2) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 35) %>% fmt_number(decimals = 3)
```

term	df.residual	rss	df	sumsq	statistic	p.value
life_exp ~ cell_phones_100	103.000	3,663.747	NA	NA	NA	NA
life_exp ~ cell_phones_100 + vax_rate	102.000	3,480.371	1.000	183.376	5.374	0.022

$P(F_{1,103} > 5.374) = 0.022$

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

7. Write a **conclusion**

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that countries' vaccination rate contributes significantly to the prediction of life expectancy, given that cell phones per 100 people 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

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?

- 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 $j = 2, 3$)

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. 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**

For example:

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

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~~ # Coef)

5. Calculate 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 a **conclusion**

We (reject/fail to reject) the null hypothesis at the $100\alpha\%$ significance level. There is (sufficient/insufficient) evidence that predictors/covariates $2, 3$ 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 percent access to basic sanitation (BS):

$$\text{LE} = \beta_0 + \beta_1 \text{CP} + \beta_2 \text{VR} + \beta_3 BS + \epsilon$$

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

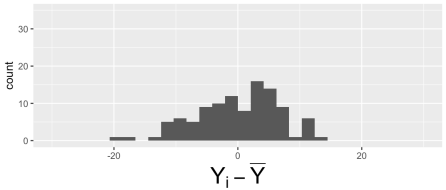
Our main question for the group covariate subset F-test: Does the regression model containing vaccination rate and basic sanitation percent improve the estimation of countries' life expectancy, given percent cell phones per 100 people is already in the model?

Null / Smaller / Reduced model	Alternative / Larger / Full model
$LE = \beta_0 + \beta_1 CP + \epsilon$	$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \beta_3 BS + \epsilon$

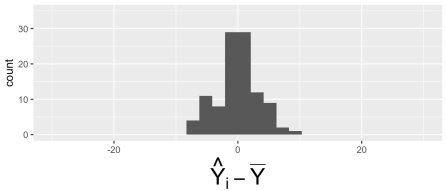
Comparing the SSY, SSR, and SSE for reduced and full model

Reduced / null model

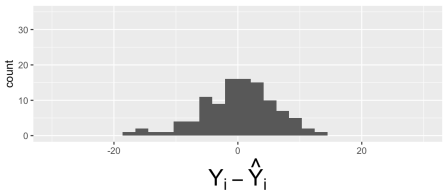
$$LE = \beta_0 + \beta_1 CP + \epsilon$$



$$SSY = 45.75$$



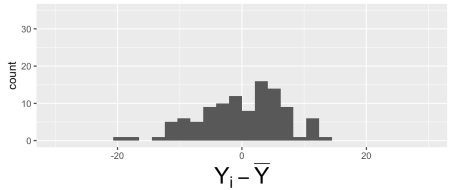
$$SSR = 10.52$$



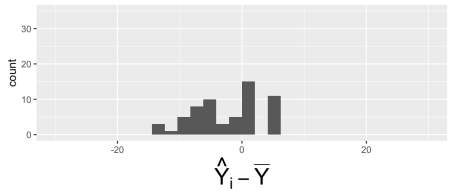
$$SSE = 35.23$$

Full / Alternative model

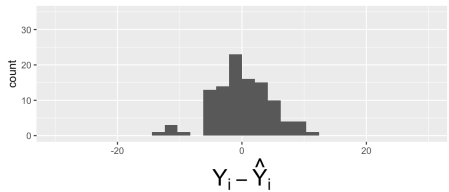
$$LE = \beta_0 + \beta_1 CP + \beta_2 VR + \beta_3 BS + \epsilon$$



$$SSY = 45.75$$



$$SSR = 24.47$$



$$SSE = 21.28$$

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

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**

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

testing 2 coef

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 $df = n - k - 1 = 105 - 2 - 1 = 102$. ($n$ = # obversation, k = # ~~covariates~~)

Coef

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

5. Calculate 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:

```
anova(mod_red3, mod_full3) %>% tidy() %>% gt() %>%
  tab_options(table.font.size = 35) %>% fmt_number(decimals = 3)
```

term	df.residual	rss	df	sumsq	statistic	p.value
life_exp ~ cell_phones_100	103.000	3,663.747	NA	NA	NA	NA
life_exp ~ cell_phones_100 + vax_rate + basic_sani	101.000	2,213.194	2.000	1,450.552	33.098	0.000

$P(F_{2,101} > 33.098) < 0.0001$

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

7. Write a **conclusion**

We reject the null hypothesis at the 5% significance level. There is sufficient evidence that countries' vaccination rate or basic sanitation (or both) contribute significantly to the prediction of life expectancy, given that cell phones per 100 people 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