Lesson 4: SLR Inference and Prediction

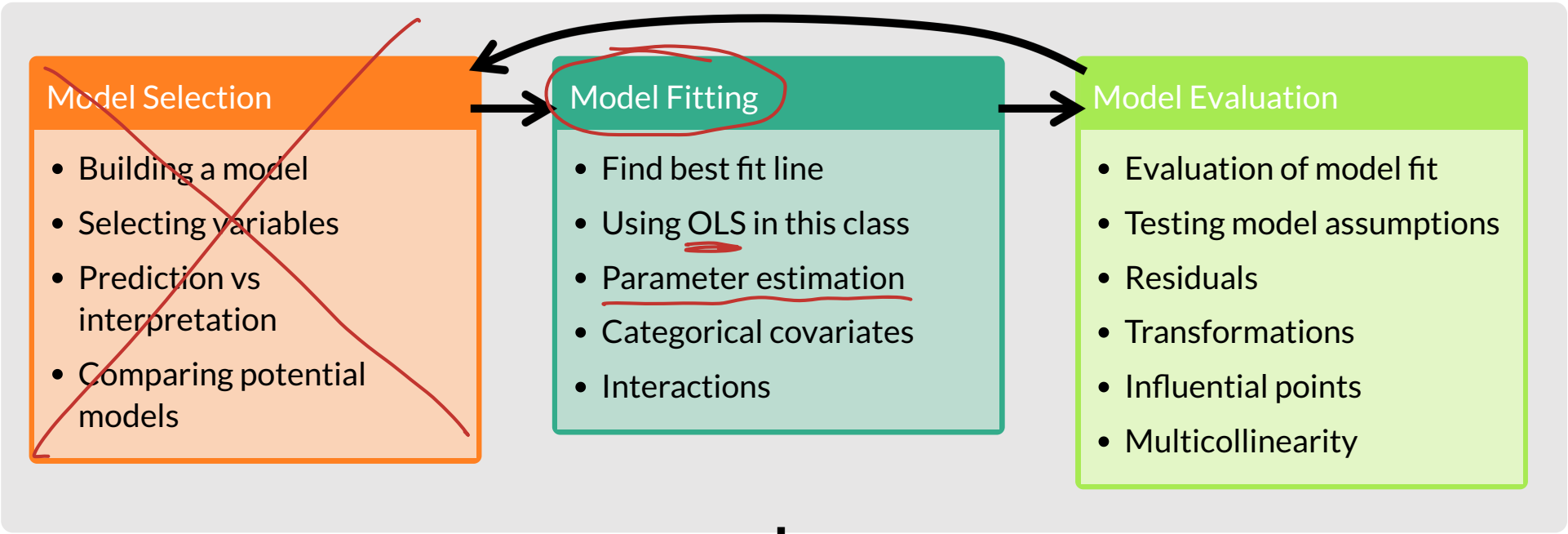
Nicky Wakim

2026-01-14

Learning Objectives

1. Estimate the variance of the residuals
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)
4. Calculate and report the estimate and confidence interval for the expected/mean response given X

Process of regression data analysis



Let's remind ourselves of the model that we fit last lesson

- We fit Gapminder data with cell phones as our independent variable and life expectancy as our dependent variable
- We used OLS to find the coefficient estimates of our best-fit line

```
1 model1 <- gapm %>%
2   lm(formula = life_exp ~ cell_phones_100)
```

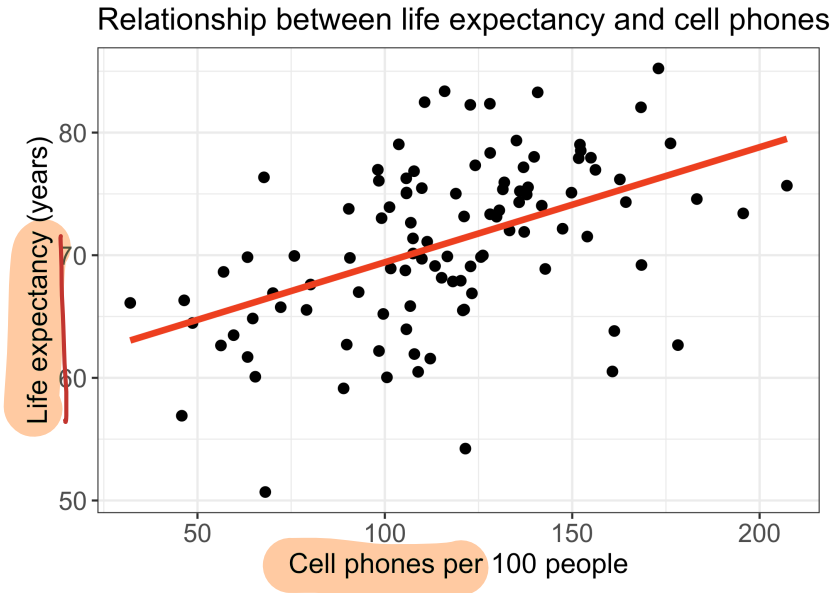
▶ Code

term	estimate	std.error	statistic	p.value
(Intercept)	60.04	2.06	29.21	0.00
cell_phones_100	0.09	0.02	5.55	0.00

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

$$\widehat{\text{life expectancy}} = 50.9 + 0.232 \cdot \text{cell phones}$$

60.04 0.09



Fitted line is derived from the population SLR model

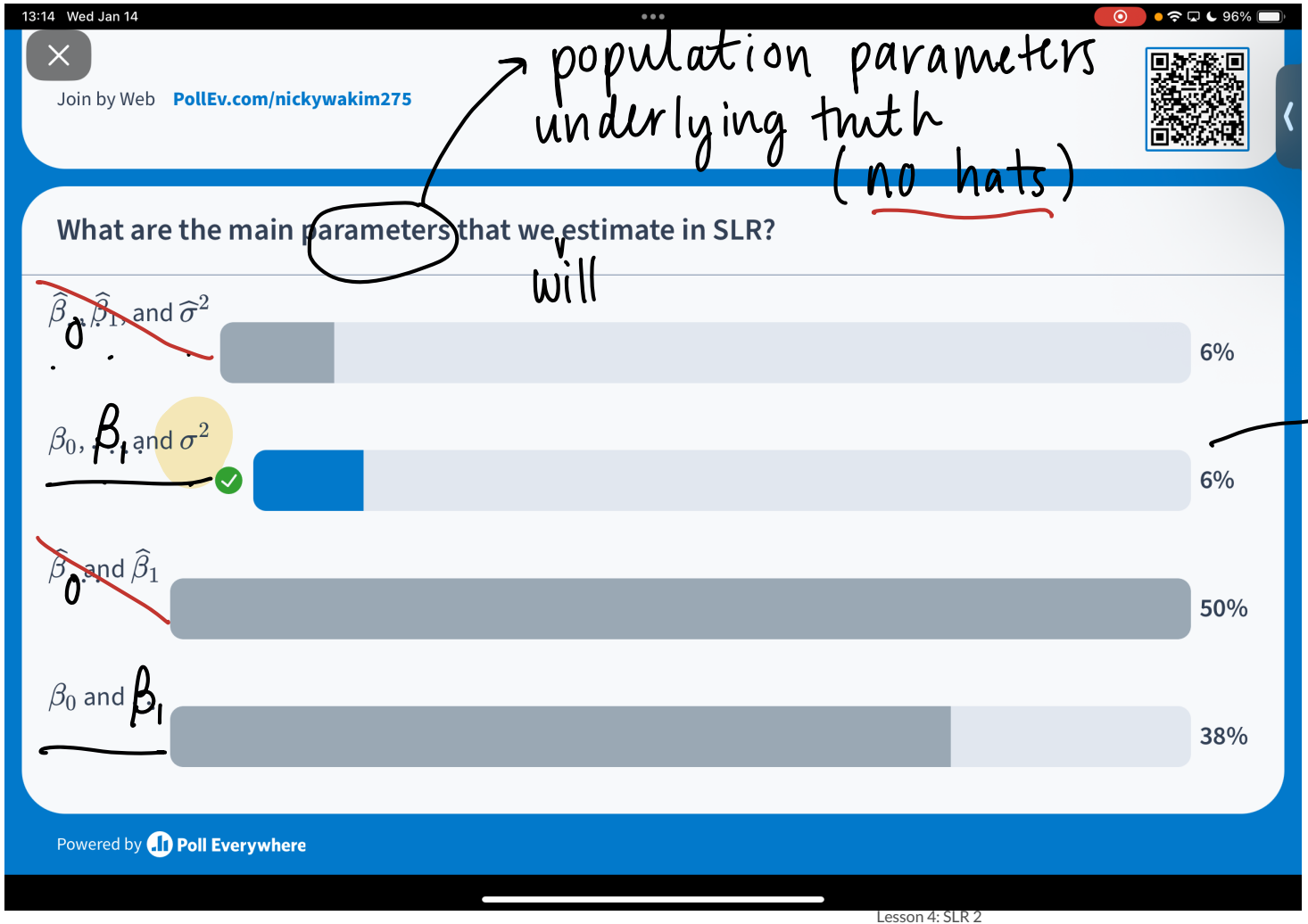
The (population) regression model is denoted by:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 and β_1 are **unknown** population parameters
- ϵ (epsilon) is the error about the line
  - It is assumed to be a random variable with a…
    - Normal distribution with mean 0 and constant variance σ^2
    - i.e. $\epsilon \sim N(0, \sigma^2)$

Poll Everywhere Question 1

hats: estimates (estimating a parameter value)



population parameters underlying truth (no hats)

will

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

true, underlying variance of residuals

Learning Objectives

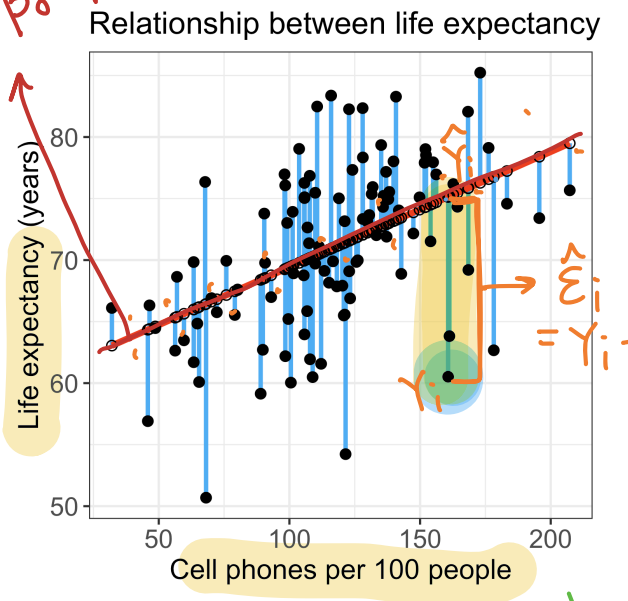
1. Estimate the variance of the residuals → σ^2 parameter we will calc $\hat{\sigma}^2$
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)
4. Calculate and report the estimate and confidence interval for the expected/mean response given X

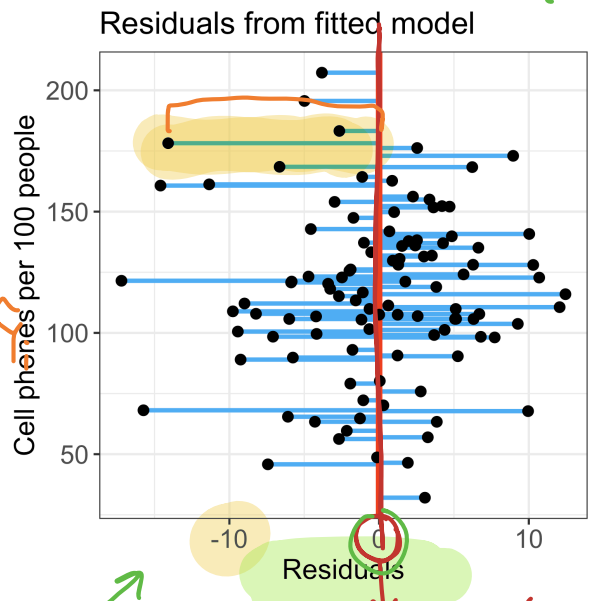
Residuals recap

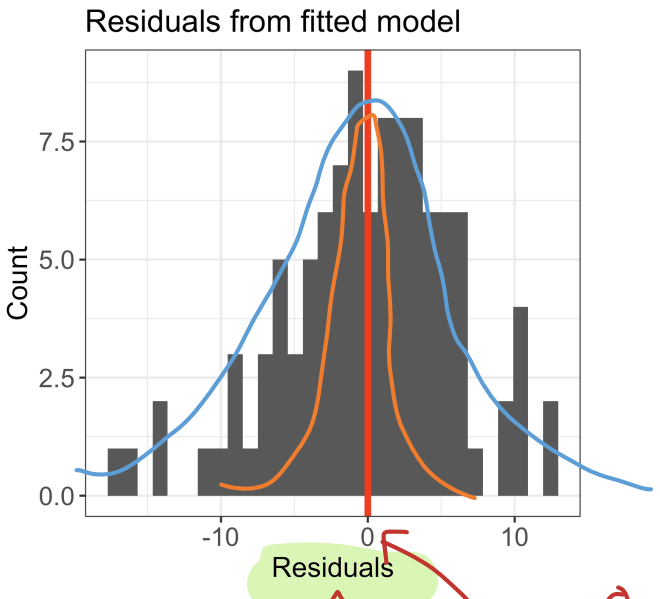
- Recall our population model residuals are distributed by $\epsilon \sim N(0, \sigma^2)$
- And our estimated residuals are $\hat{\epsilon} \sim N(0, \hat{\sigma}^2)$

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ $\quad \widehat{\text{life exp}} = \hat{\beta}_0 + \hat{\beta}_1 \text{cellphone}$ $\quad \hat{\epsilon}_i = Y_i - \hat{Y}_i$ (obs Y) (fitted Y)

want variance of this ($\hat{\sigma}^2$)







taking red line (aka best fit line) and making vertical @ $\hat{Y}_i$

$\epsilon_i = 0 \Rightarrow Y_i = \hat{Y}_i$

$\hat{\epsilon}$ $\quad \mu = 0$

$\widehat{\sigma}^2$: Needed ingredient for inference

- **We need the variance of the residuals $\widehat{\sigma}^2$ to perform inference on our coefficients**
- The *variance* of the errors (residuals) is estimated by $\widehat{\sigma}^2$

$$\widehat{\sigma}^2 = S^2_{y|x} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n-2}SSE = MSE$$

$\widehat{\sigma}^2$: I hope R can calculate that for me... (1/2)

- The *standard deviation* $\widehat{\sigma}$ is given in the R output as the `Residual standard error`
  - 4^{th} line from the bottom in the `summary()` output of the model:

```
1 summary(model1)
```

```
Call:
lm(formula = life_exp ~ cell_phones_100, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-17.211  -3.268   0.615   3.818  12.449

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     60.04051    2.05567  29.207  < 2e-16 ***
cell_phones_100  0.09384    0.01692   5.546 2.27e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.964 on 103 degrees of freedom
Multiple R-squared:   0.23, Adjusted R-squared:  0.2225
F-statistic: 30.76 on 1 and 103 DF,  p-value: 2.271e-07
```

$n-2 = 105-2 = 103$

$\widehat{\sigma} = 5.964$

$\widehat{\sigma}^2$: I hope R can calculate that for me... (2/2)

- It can!!

```
m1_sum = summary(model1)
m1_sum$sigma
```

5.964

```
[1] 5.964089
```

```
# number of observations (pairs of data) used to run the model
n = nobs(model1)
n
```

```
[1] 105
```

$\widehat{\sigma}^2$ to SSE

- Recall how we minimized the SSE to find our line of best fit
- SSE and $\widehat{\sigma}^2$ are closely related:

$$\widehat{\sigma}^2 = \frac{1}{n-2} SSE$$

$$6.142^2 = \frac{1}{105} SSE$$

$$SSE = 103 \cdot 6.142^2 = 2942.48$$

5.964

105

5.964

- 2942.48 is the smallest sums of squares of all possible regression lines through the data

Learning Objectives

1. Estimate the variance of the residuals
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)
4. Calculate and report the estimate and confidence interval for the expected/mean response given X

Do we trust our estimate $\widehat{\beta}_1$?

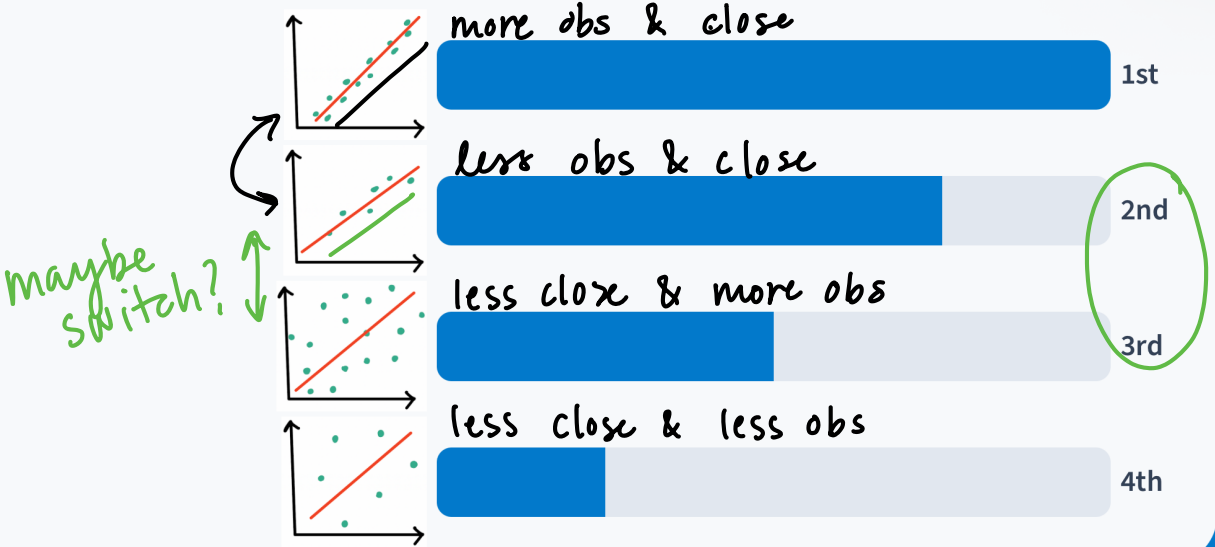
- So far, we have shown that we think the estimate is 0.094
- $\widehat{\beta}_1$ (coefficient estimate) uses our sample data to estimate the population parameter β_1
- Inference helps us figure out *mathematically* how much we trust our best-fit line
- Are we certain that the relationship between X and Y that we estimated reflects the true, underlying relationship?

Poll Everywhere Question 2

13:36 Wed Jan 14 86%

Join by Web PollEv.com/nickywakim275

Rank the following plots from most trustworthy β_1 estimation to least trustworthy estimation.



Powered by Poll Everywhere

# of obs & closeness to best fit line

Inference for the population slope: hypothesis test and CI

Population model

line + random "noise"

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$
σ^2 is the variance of the residuals

Sample best-fit (least-squares) line

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

Note: Some sources use b instead of $\widehat{\beta}$

We have two options for inference:

1. Conduct the **hypothesis test**

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

Note: R reports p-values for 2-sided tests

2. Construct a **95% confidence interval** for the **population slope** β_1

Learning Objectives

1. Estimate the variance of the residuals
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)
4. Calculate and report the estimate and confidence interval for the expected/mean response given X

Reference: Steps in a Hypothesis Test

1. Check the **assumptions**
   - What sampling distribution are you using? What assumptions are required for it?
2. Set the **level of significance** α
3. Specify the **null** (H_0) and **alternative** (H_A) **hypotheses**
   - In symbols and/or in words
   - Alternative: one- or two-sided?
4. Specify the test statistic and its **distribution under the null**
5. Calculate the **test statistic**.
6. Calculate the **p-value** based on the observed test statistic and its sampling distribution
7. Write a **conclusion** to the hypothesis test
   - Do we reject or fail to reject H_0?
   - Write a conclusion in the context of the problem

Steps for hypothesis test for population slope β_1 (using t-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:

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

4. Specify the test statistic and its **distribution under the null**
   - The test statistic is t, and follows a Student's t-distribution.
5. Calculate the **test statistic**.
   - The calculated **test statistic** for $\widehat{\beta}_1$ is

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\text{SE}_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{\text{SE}_{\widehat{\beta}_1}}$$

   when we assume $H_0 : \beta_1 = 0$ is true.
6. Calculate the **p-value**
   - We are generally calculating: $2 \cdot P(T > |t|)$
7. Write a **conclusion**
   - We (reject/fail to reject) the null hypothesis that the slope is 0 at the $100\alpha\%$ significiance level. There is (sufficient/insufficient) evidence that there is significant association between (Y) and (X) (p-value = $P(T > t)$).

Standard error of fitted slope $\widehat{\beta}_1$

$$\text{SE}_{\widehat{\beta}_1} = \frac{\widehat{\sigma}}{s_X\sqrt{n-1}}$$

$\text{SE}_{\widehat{\beta}_1}$ is a measure of **variability** of the estimate $\widehat{\beta}_1$

- $\widehat{\sigma}$ is the standard deviation of the residuals

$\widehat{\sigma} = \sqrt{5.964}$

- s_X is the sample standard deviation of the explanatory variable X
- n is the sample size, or the number of observations in the model

$n = 105$

Calculating standard error for $\widehat{\beta}_1$ (1/2)

$$SE_{\hat{\beta}_1} = \frac{\hat{\sigma}}{S_x\sqrt{n-1}}$$

- **Option 1:** Calculate using the formula

```
 1 glance(model1)
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic     p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.230         0.222  5.96      30.8 0.000000227     1  -335.  677.  685.
# ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
 1 # standard deviation of the residuals (Residual standard error in summary() output)
 2 (s_resid <- glance(model1)$sigma)
```

```
[1] 5.964089
```

```
 1 # standard deviation of x's
 2 (s_x <- sd(gapm$cell_phones_100, na.rm=T))
```

```
[1] 34.56469
```

```
 1 # number of pairs of complete observations
 2 (n <- nobs(model1))
```

```
[1] 105
```

$$\hat{\sigma}/(S_x\sqrt{n-1})$$

```
 1 (se_b1 <- s_resid/(s_x * sqrt(n-1))) # compare to SE in regression output
```

```
[1] 0.01691978
```

$SE_{\hat{\beta}_1}$

Calculating standard error for $\widehat{\beta}_1$ (2/2)

- **Option 2:** Use regression table

```
# recall model1_b1 is regression table restricted to b1 row
model1_b1 <-tidy(model1) %>% filter(term == "cell_phones_100")
model1_b1 %>% gt() %>%
  tab_options(table.font.size = 45) %>% fmt_number(decimals = 4)
```

term	estimate	std.error	statistic	p.value
cell_phones_100	0.0938	0.0169	5.5461	0.0000

$\hat{\beta}_1$ $SE_{\hat{\beta}_1}$ $t_{\hat{\beta}_1}$ p-val $2 \cdot P(T > t_{\hat{\beta}_1})$

Some important notes

- Today we are discussing the hypothesis test for a **single** coefficient

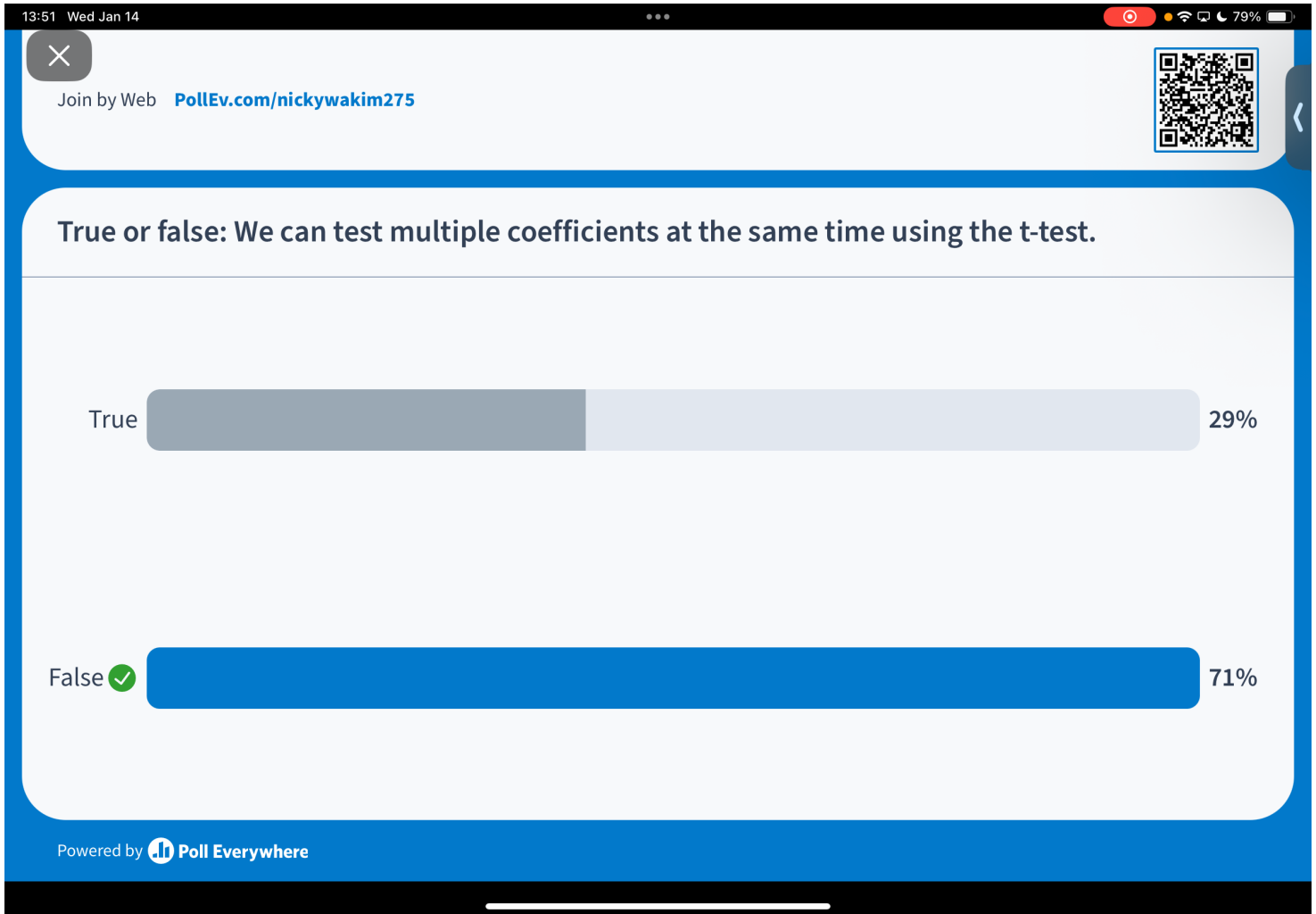
- The test statistic for a single coefficient follows a Student's t-distribution
  - It can also follow an F-distribution, but we will discuss this more with multiple linear regression and multi-level categorical covariates

- Single coefficient testing can be done on any coefficient, but it is most useful for continuous covariates or binary covariates
  - This is because testing the single coefficient will still tell us something about the overall relationship between the covariate and the outcome.
  - We will talk more about this with multiple linear regression and multi-level categorical covariates

↳ $H_0: \beta_0 = 0$?

(if covariate is cont or binary)

Poll Everywhere Question 3



test β_1 & β_0 at same time

H_0: $\beta_1 = 0$ & $\beta_0 = 0$

→ cannot use t-test!!

Life expectancy example: hypothesis test for population slope β_1 (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 slope is 0 or not:

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

4. Specify the test statistic and its **distribution under the null**
   - The test statistic is t, and follows a Student's t-distribution.

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

5.Calculate the **test statistic**

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

```
model1_b1 <-tidy(model1) %>%
  filter(term == "cell_phones_100")
model1_b1 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
cell_phones_100	0.094	0.017	5.546	0.000

```
(TestStat_b1 <- model1_b1$estimate /
    model1_b1$std.error)
```

```
[1] 5.546063
```

$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$$

Option 2: Get the test statistic value (t^*) from R

```
model1_b1 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
cell_phones_100	0.094	0.017	5.546	0.000

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

6.Calculate the **p-value**

$2 \cdot P(T > |t|)$

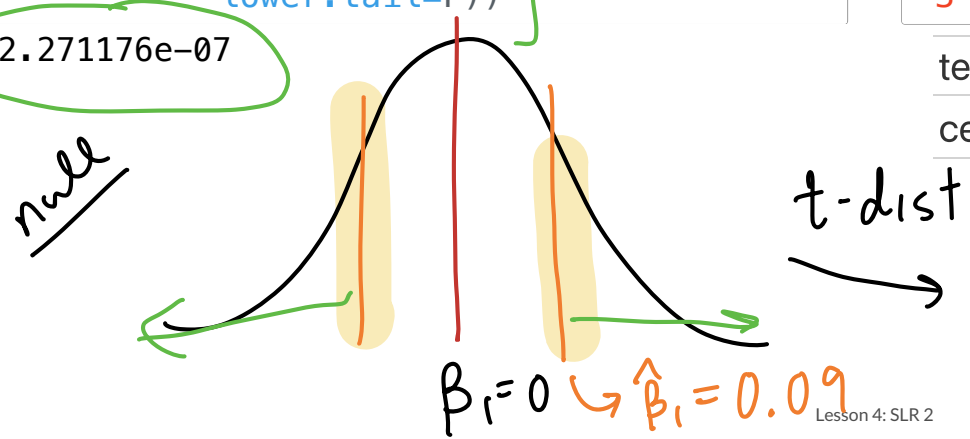
$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}$

- The p-value is the *probability of obtaining a test statistic* **just as extreme or more extreme** *than the* **observed** *test statistic assuming the null hypothesis* H_0 *is true*
- We know the probability distribution of the test statistic (the *null distribution*) assuming H_0 is true
- Statistical theory tells us that the test statistic t can be modeled by a t-distribution with $df = n - 2$.
  - We had 105 countries' data, so $n = 105$

Option 1: Use `pt()` and our calculated test statistic

```
(pv = 2*pt(TestStat_b1,
           df=n-2,
           lower.tail=F))
```

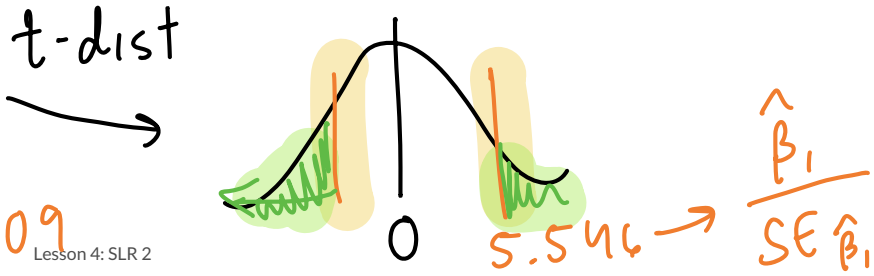
```
[1] 2.271176e-07
```



Option 2: Use the regression table output

```
model1_b1 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
cell_phones_100	0.094	0.017	5.546	0.000



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

7. Write a **conclusion**

p-val = 2.27×10^{-7} <<< 0.05

Reject null

We reject the null hypothesis that the slope is 0 at the 5% significance level. There is sufficient evidence that there is association between life expectancy and number of cell phones per 100 people (p-value < 0.0001).

↳ suff evidence that slope is NOT 0.

Note on hypothesis testing using R

- We can basically combine Step 5 and 6 if we are using the “Option 2” route

- In our assignments: if you use Option 2, Step 5and 6 become one
  - Unless I specifically ask for the test statistic!!

Life expectancy ex: hypothesis test for population intercept β_0 (1/4)

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 intercept is 0 or not:

$$H_0 : \beta_0 = 0$$
$$\text{vs. } H_A : \beta_0 \neq 0$$

4. Specify the test statistic and its **distribution under the null**
   - This is the same as the slope. The test statistic is t, and follows a Student's t-distribution.

Life expectancy ex: hypothesis test for population intercept β_0 (2/4)

5.Calculate the **test statistic**

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

```
model1_b0 <- tidy(model1) %>%
  filter(term == "(Intercept)")
model1_b0 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	60.041	2.056	29.207	0.000

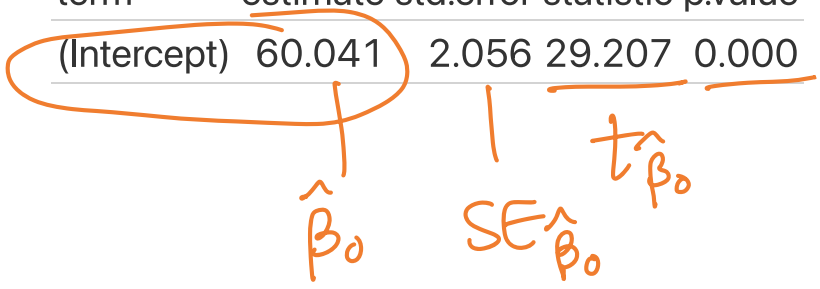
```
(TestStat_b0 <- model1_b0$estimate /
    model1_b0$std.error)
```

```
[1] 29.20728
```

Option 2: Get the test statistic value (t^*) from R

```
model1_b0 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	60.041	2.056	29.207	0.000



Life expectancy ex: hypothesis test for population intercept β_0 (3/4)

6.Calculate the **p-value**

Option 1: Use `pt()` and our calculated test statistic

```
(pv = 2*pt(TestStat_b0,
           df=n-2,
           lower.tail=F))
```

```
[1] 1.215444e-51
```

Option 2: Use the regression table output

```
model1_b0 %>% gt() %>%
  tab_options(table.font.size = 40) %>%
  fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	60.041	2.056	29.207	0.000

Life expectancy ex: hypothesis test for population intercept β_0 (4/4)

7. Write a **conclusion**

We reject the null hypothesis that the intercept is 0 at the 5% significance level. There is sufficient evidence that the intercept for the association between life expectancy and number of cell phones per 100 people is different from 0 (p-value < 0.0001).

- Note: if we fail to reject H_0, then we *could* decide to remove the intercept from the model to force the regression line to go through the origin (0,0) if it makes sense to do so for the application
  - Not typical to remove the intercept

$$Y = \beta_0 + \beta_1 X$$

$$1$$

$$0$$

$$Y = \beta_1 X$$

Learning Objectives

1. Estimate the variance of the residuals
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)
4. Calculate and report the estimate and confidence interval for the expected/mean response given X

Inference for the population slope: hypothesis test and CI

Population model

line + random "noise"

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$
σ^2 is the variance of the residuals

Sample best-fit (least-squares) line

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

Note: Some sources use b instead of $\widehat{\beta}$

We have two options for inference:

1. Conduct the **hypothesis test**

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

Note: R reports p-values for 2-sided tests

2. Construct a **95% confidence interval** for the **population slope** β_1

Confidence interval for population slope β_1

Recall the general CI formula:

$$\widehat{\beta}_1 \pm t^*_{\alpha,n-2} \cdot SE_{\widehat{\beta}_1}$$

To construct the confidence interval, we need to:

- Set our α-level $\alpha = 0.05 \rightarrow$ 95% CI
- Find $\widehat{\beta}_1$
- Calculate the t^*_{n-2} $\rightarrow$ critical value for specific CI %
- Calculate $SE_{\widehat{\beta}_1}$

Calculate CI for population slope β_1 (1/2)

$$\widehat{\beta}_1 \pm t^* \cdot SE_{\widehat{\beta}_1}$$

where t^* is the t-distribution critical value with $df = n - 2$.

- **Option 1:** Calculate using each value

Save values needed for CI:

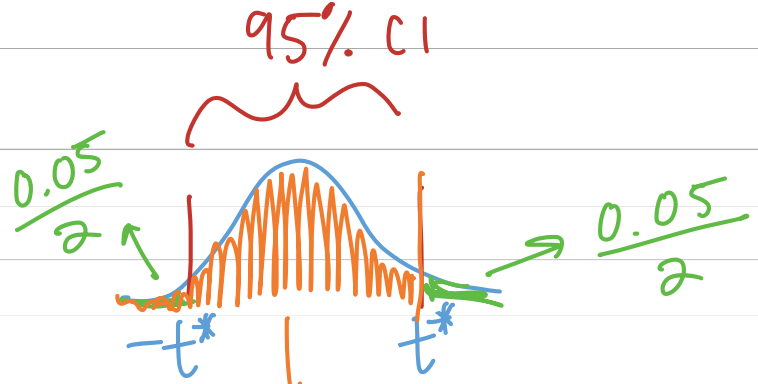
```
1 b1 <- model1_b1$estimate
2 SE_b1 <- model1_b1$std.error
```

```
1 nobs(model1) # sample size n
```

```
[1] 105
```

```
1 (tstar <- qt(.975, df = n - 2))
```

```
[1] 1.983264
```



Use formula to calculate each bound

```
1 (CI_LB <- b1 - tstar*SE_b1)
```

```
[1] 0.06028178
```

$$\hat{\beta}_1 - t^* \cdot SE_{\hat{\beta}_1}$$

```
1 (CI_UB <- b1 + tstar*SE_b1)
```

```
[1] 0.1273946
```

$$\hat{\beta}_1 + t^* SE_{\hat{\beta}_1}$$

Calculate CI for population slope β_1 (2/2)

$$\widehat{\beta}_1 \pm t^* \cdot SE_{\beta_1}$$

where t^* is the t-distribution critical value with $df = n - 2$.

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

```
1 tidy(model1, conf.int = T) %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	60.041	2.056	29.207	0.000	55.964	64.117
cell_phones_100	0.094	0.017	5.546	0.000	0.060	0.127

β_1 → cell_phones_100

$\hat{\beta}_1$ | $SE_{\hat{\beta}_1}$ | t (not t^*) | p-val | LCI | UCI

Interpreting the coefficient estimate of the population slope with CIs

- When we report our results to someone else, we don't usually show them our full hypothesis test
  - In an informal setting, someone may want to see it
- Typically, we report the estimate with the confidence interval
  - From the confidence interval, your audience can also deduce the results of a hypothesis test
- Once we found our CI, we often just write the interpretation of the coefficient estimate:

General statement for population slope inference

For every increase of 1 unit in the X-variable, there is an expected/average (pick one) increase of $\widehat{\beta}_1$ units in the Y-variable (95%: LB, UB).

- **In our example:** For every additional cell phone per 100 people, life expectancy increases, on average, 0.094 years (95% CI: 0.06, 0.127)

Usually three options for your interpretations

- **Option 1:** For every additional cell phone per 100 people, life expectancy increases, **on average**, 0.094 years (95% CI: 0.06, 0.127)

Y inc by avg of 0.094

- **Option 2:** For every additional cell phone per 100 people, **average** life expectancy increases 0.094 years (95% CI: 0.06, 0.127)

$\hat{Y}$

- **Option 3:** For every additional cell phone per 100 people, **expected** life expectancy increases 0.094 years (95% CI: 0.06, 0.127)

$\hat{E}(Y|X)$

Poll Everywhere Question 4

14:23 Wed Jan 14 64%

Join by Web **PollEv.com/nickywakim275**

Based off the following regression table, write a statement for the intercept's interpretation. Make sure to include the confidence interval.

When the X is 0, we expect the Y to be 50.9. (95% CI: 45.6, 56.2)

cellphones per 100 people

life expectancy

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	50.928	2.660	19.143	0.000	45.631	56.224
female_literacy_rate_2011	0.232	0.031	7.377	0.000	0.170	0.295

cellphone per 100

We are 95% confident that the average life expectancy intercept falls between 45.6 and 56.2 years

interpret of CI

Powered by Poll Everywhere

For reference: quick CI for β_0

- Calculate CI for population **intercept** β_0: $\widehat{\beta}_0 \pm t^* \cdot SE_{\widehat{\beta}_0}$

where t^* is the t-distribution critical value with $df = n - 2$

- Use the regression table

```
1 tidy(model1, conf.int = T) %>% gt() %>%
2   tab_options(table.font.size = 40) %>% fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	60.041	2.056	29.207	0.000	55.964	64.117
cell_phones_100	0.094	0.017	5.546	0.000	0.060	0.127

General statement for population intercept inference

The expected outcome for the Y-variable is $(\widehat{\beta}_0)$ when the X-variable is 0 (95% CI: LB, UB).

- **For example:** The average life expectancy is 60.04 years when the number of cell phones per ~~person~~ is 0 (95% CI: 55.96, 64.12).

Learning Objectives

1. Estimate the variance of the residuals
2. Using a hypothesis test, determine if there is enough evidence that population slope β_1 is not 0 (applies to β_0 as well)
3. Calculate and report the estimate and confidence interval for the population slope β_1 (applies to β_0 as well)

4. Calculate and report the estimate and confidence interval for the expected/mean response given X

Finding a mean response given a value of our independent variable

term	estimate	std.error	statistic	p.value
(Intercept)	60.041	2.056	29.207	0.000
cell_phones_100	0.094	0.017	5.546	0.000

fitted line:

$$\widehat{\text{life expectancy}} = 60.041 + 0.094 \cdot \text{cell phones}$$

- What is the expected/predicted life expectancy for any country with 60 cell phones per 100 people?

$$\widehat{\text{life expectancy}} = 60.041 + 0.094 \cdot 60 = 65.671$$

```
1 (y_60 <- tidy(model1)$estimate[1] + tidy(model1)$estimate[2]*60)
```

```
[1] 65.6708
```

- How do we interpret the expected value?
  - We sometimes call this "predicted" value, since we can technically use a number (of cell phones) that is not in our sample
- How variable is it?

Mean response/prediction with regression line

Recall the population model:

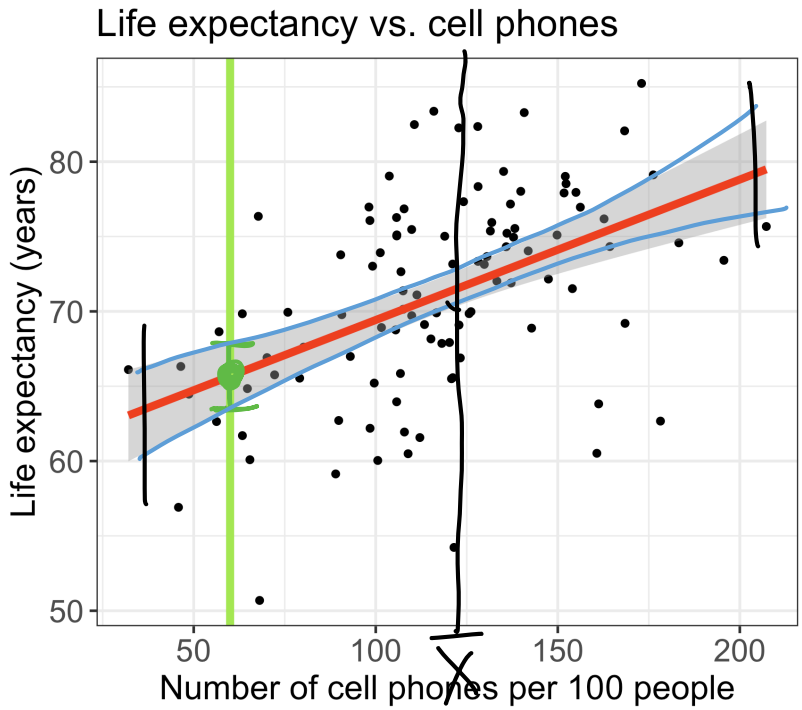
line + random "noise"

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon$$

with $\varepsilon \sim N(0, \sigma^2)$

- When we take the expected value, at a given value X^*, the average expected response at X^* is:

$$\widehat{E}[Y|X^*] = \widehat{\beta}_0 + \widehat{\beta}_1 X^*$$



- These are the points on the regression line
- The mean responses have variability, and we can calculate a CI for it, for every value of X^*

CI for population mean response ($\widehat{E}[Y|X^*]$ or $\mu_{Y|X^*}$)

$$\widehat{E}[Y|X^*] \pm t^*_{n-2} \, SE_{\widehat{E}[Y|X^*]}$$

critical

$$SE_{\widehat{E}[Y|X^*]} = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(X^* - \overline{X})^2}{(n-1)s_X^2}}$$

closer X^* is to $\overline{X}$, the smaller $(X^* - \overline{X})^2$

- $\widehat{E}[Y|X^*]$ is the predicted value at the specified point X^* of the explanatory variable
- $\widehat{\sigma}^2$ is the sd of the residuals
- n is the sample size, or the number of (complete) pairs of points
- $\overline{X}$ is the sample mean of the explanatory variable x
- s_X is the sample sd of the explanatory variable X

- Recall that t^*_{n-2} is calculated using `qt()` and depends on the confidence level $(1-\alpha)$

Example Option 1: CI for mean response $\mu_{Y|X^*}$

Find the 95% CI's for mean life expectancy for 60 cell phones per 100 people:

$$\widehat{E}[Y|X^*] \pm t^*_{n-2} \cdot SE_{\widehat{E}[Y|X^*]}$$

$$65.671 \pm 1.983 \cdot \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(X^* - \bar{x})^2}{(n-1)s_X^2}}$$

$$65.671 \pm 1.983 \cdot 5.964\sqrt{\frac{1}{105} + \frac{(60 - 116.523)^2}{(105-1)34.565^2}}$$

$$65.671 \pm 1.983 \cdot 1.12$$

$$65.671 \pm 2.22$$

$$(63.45, 67.891)$$

E(lifeexp | x* = 60)

Example Option 2: CI for mean response $\mu_{Y|X^*}$

Find the 95% CI's for mean life expectancy for 60 and 80 cell phones per 100 people, respectively.

- Use the base R `predict()` function
- Requires specification of a `newdata` "value"
  - The `newdata` value is X^*
  - This has to be in the format of a data frame though
  - with column name identical to the predictor variable in the model

```
1 newdata <- data.frame(cell_phones_100 = c(60, 80))
2 newdata
```

```
  cell_phones_100
1              60
2              80
```

should match variable name for X

```
1 predict(model1,
2         newdata=newdata,
3         interval="confidence")
```

E(Y|X*)

```
       fit      lwr      upr
1 65.67080 63.45044 67.89117
2 67.54757 65.86395 69.23118
```

lwr CI upr CI

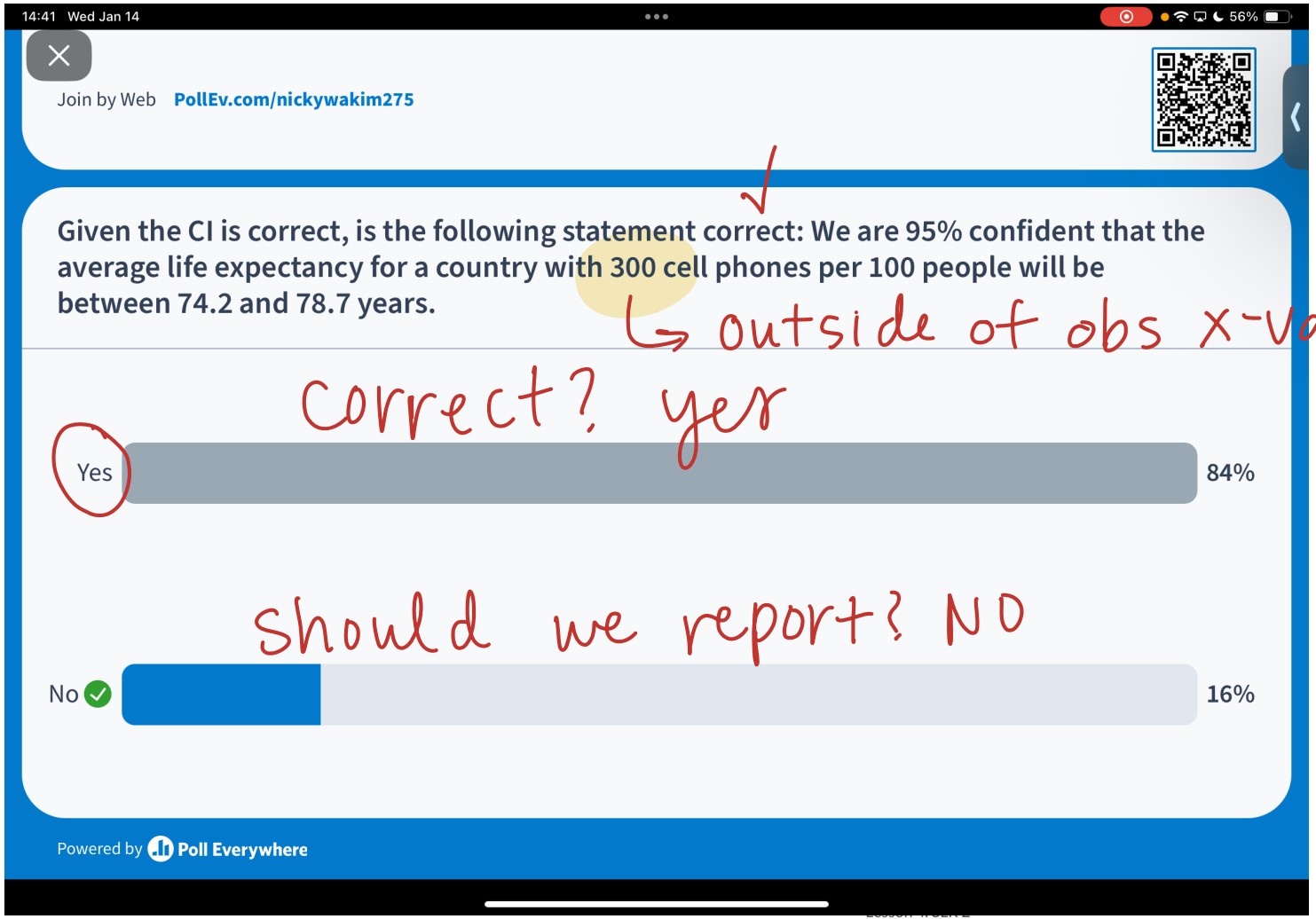
The avg life exp is 65.67 years for a country w/ 60 cellphones per 100 ppl (95% CI: 63.45, 67.89 yrs).

60 cellphone / 100 ppl

Interpretation

We are 95% confident that the **average** life expectancy for a country with a 60 cell phones per 100 people will be between 63.45 and 67.89 years.

Poll Everywhere Question 5

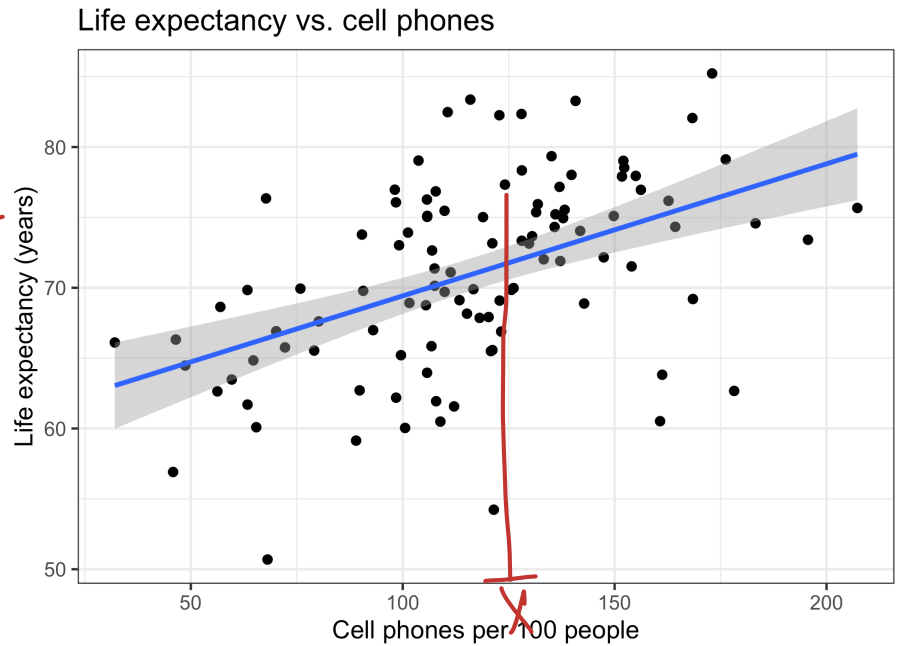


Confidence bands for mean response $\mu_{Y|X^*}$

- Often we plot the CI for many values of X, creating **confidence bands**
- The confidence bands are what ggplot creates when we set `se = TRUE` within `geom_smooth`
- Think about it: for what values of X are the confidence bands (intervals) narrowest?

```
gapm %>%
  ggplot(
    aes(
      x=cell_phones_100,
      y=life_exp
      )
    ) +
  geom_point() +
  geom_smooth(method = lm, se=TRUE) +
  labs(
    x = "Cell phones per 100 people",
    y = "Life expectancy (years)",
    title = "Life expectancy vs. cell phone
    )
```

gives us bands



Width of confidence bands for mean response $\mu_{Y|X^*}$

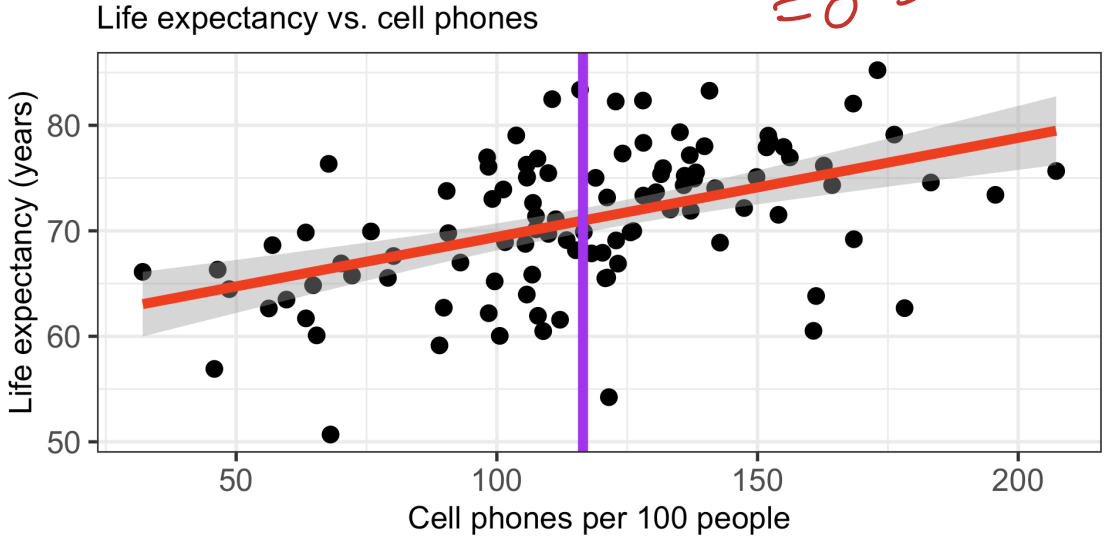
- For what values of X^* are the confidence bands (intervals) narrowest? widest?

$$\widehat{E}[Y|X^*] \pm t^*_{n-2} \cdot SE_{\widehat{E}[Y|X^*]}$$

$$\widehat{E}[Y|X^*] \pm t^*_{n-2} \cdot \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(X^* - \bar{x})^2}{(n-1)s_x^2}}$$

When $X^* = \bar{X}$

$=0$



Textbook readings

- Introduction to Regression Methods for Public Health Using R
  - 4.5 Interpreting p-values
  - 4.6 Predictions from the model
  - 4.7 Confidence intervals and prediction intervals
- A Progressive Introduction to Linear Models
  - Not really any good sections