MODULE 12

LINEAR REGRESSION

OUTLINE

1. Linear regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

OUTLINE

1. Regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

SIMPLE LINEAR REGRESSION

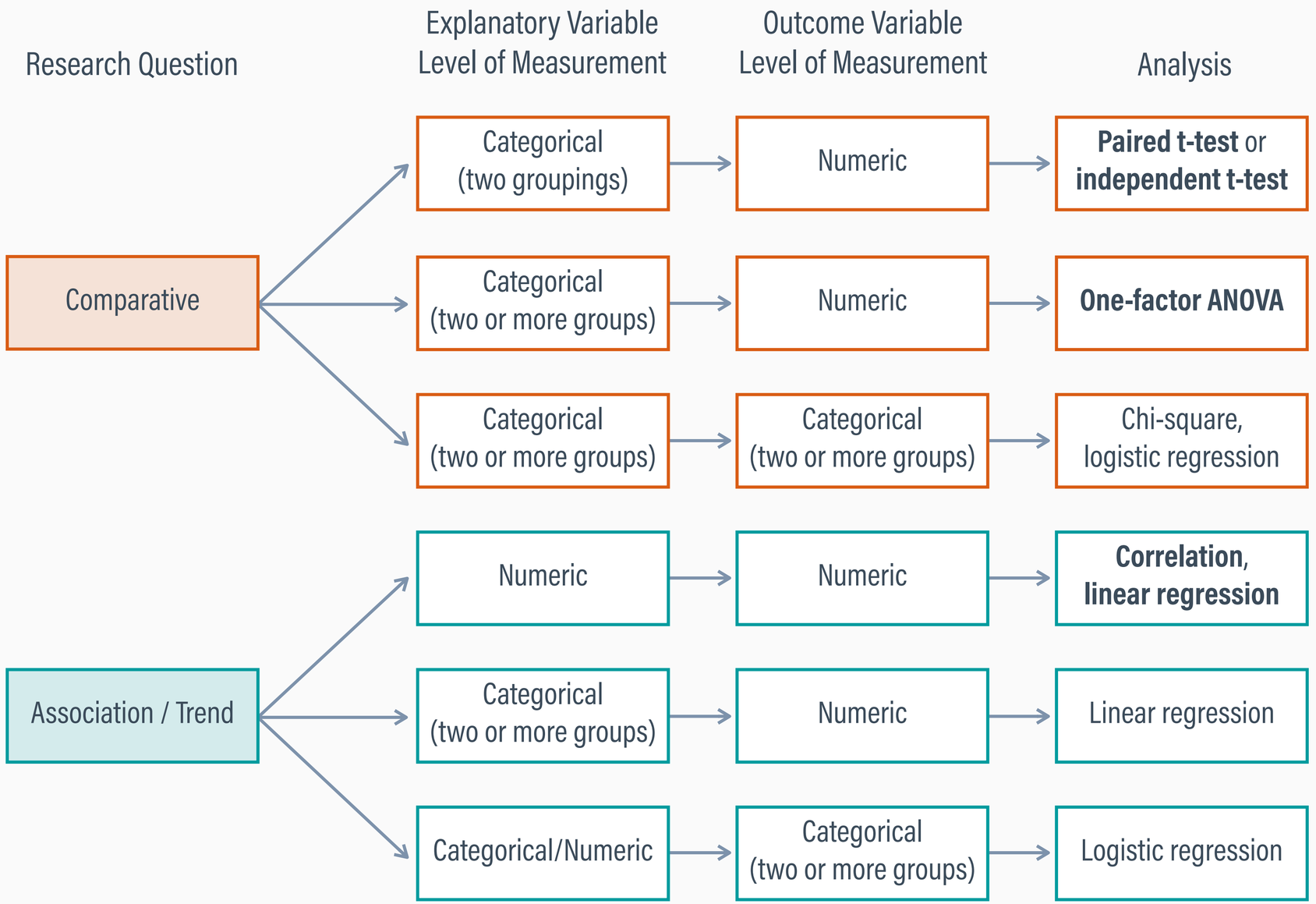
- Simple linear regression links an independent variable to a dependent variable via a linear equation
- An intercept and slope coefficient capture the direction and strength of the overall trend
- Applicable to association research questions involving a trend between two numeric variables

REGRESSION EXTENSIONS

- Regression is a very general family of procedures that can accommodate categorical and numeric variables
- With categorical independent variables and numeric dependent variables, regression is equivalent to ANOVA
- Regression is foundational to many advanced statistical modeling frameworks (e.g., structural equation models, multilevel models, mediation models)

STATISTICAL ORG CHART

Bold typeface = 250A topic



OUTLINE

1. Regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

LINEAR EQUATION REVIEWED

- Y is the dependent variable, and X is the predictor variable

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

- β_0 is the intercept, which is the predicted value of Y (the dependent variable) when X (predictor) equals 0
- β_1 is the slope, which is the expected change in Y (the dependent variable) for every one-point increase in X

LINEAR EQUATION APPLICATION

- Riding a Bird e-scooter costs $1 to activate and $0.20 per minute thereafter
- A linear equation describes the association between rental time and cost per trip

$$\text{cost} = \beta_0 + \beta_1(\text{minutes})$$

$$\text{cost} = 1.00 + 0.20(\text{minutes})$$

REGRESSION TERMINOLOGY

- The predictor variable is the “causal” variable on the right side of the equation
- The dependent variable on the left side of the equation changes in response to the predictor
- Rental time is the predictor because it influences cost, not vice versa

$$\text{cost} = \beta_0 + \beta_1(\text{minutes})$$

$$\text{dependent} = \beta_0 + \beta_1(\text{predictor})$$

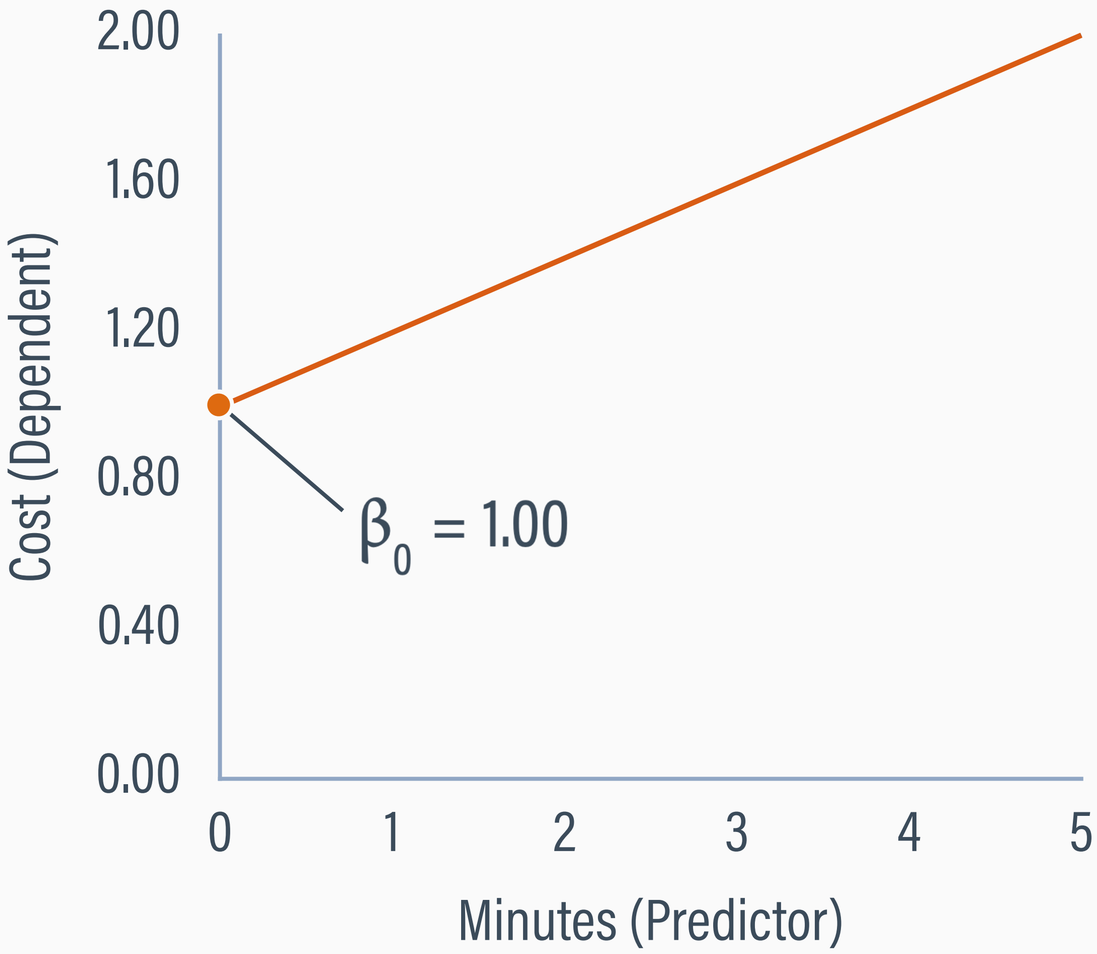
INTERCEPT: COST WHEN MINUTES = 0

- The intercept is the predicted value of the dependent variable (cost) when the predictor variable (minutes) equals zero
- Immediately cancelling the ride after activation (minutes = 0) will cost $1.00

$$\text{cost} = \beta_0 + \beta_1(\text{minutes})$$

$$\text{cost} = 1.00 + 0.20(\text{minutes})$$

$$\text{cost} = 1.00 + 0.20(0) = 1.00$$



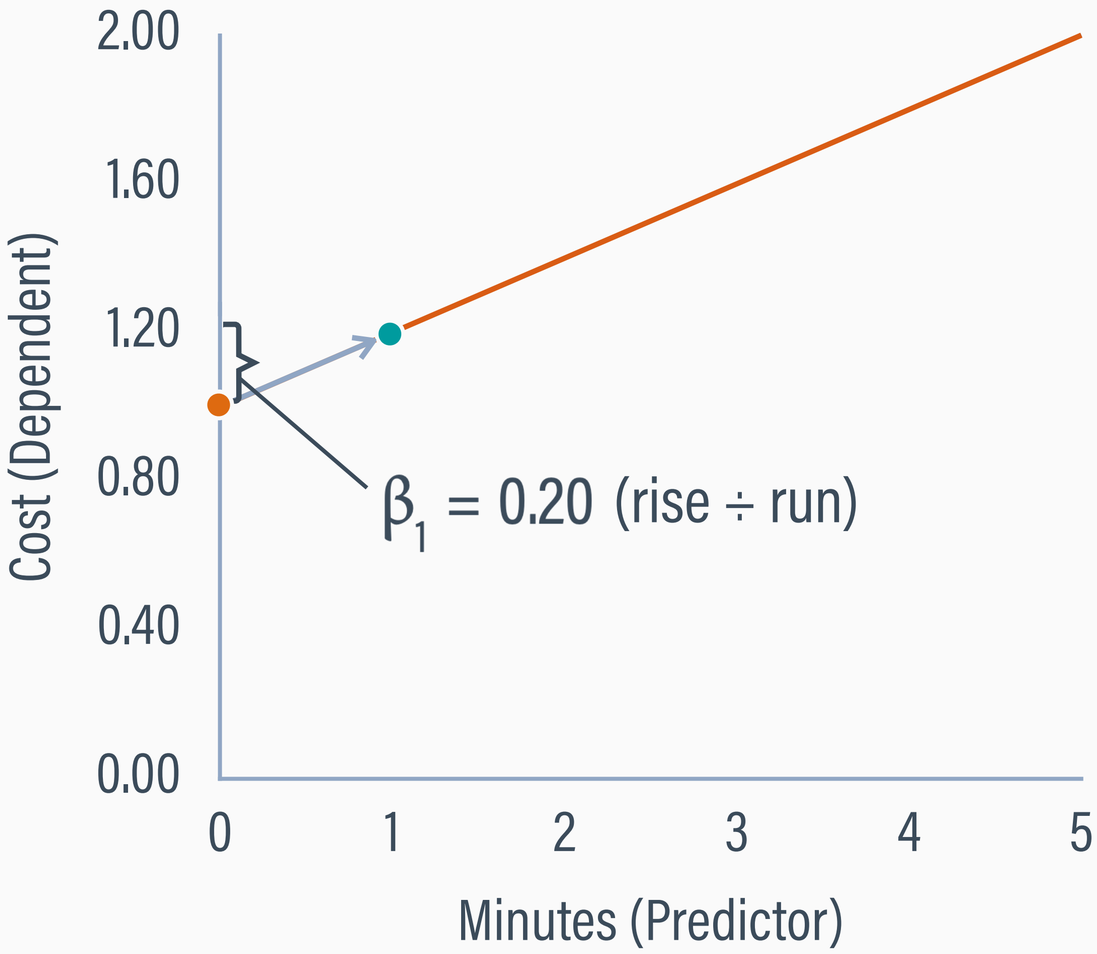
SLOPE: CHANGE IN COST PER MINUTE

- The slope is the expected change in the dependent variable (cost) when the predictor variable (minutes) increases by one point
- Every additional minute (one-point increase) adds $0.20 to the cost

$$\text{cost} = \beta_0 + \beta_1(\text{minutes})$$

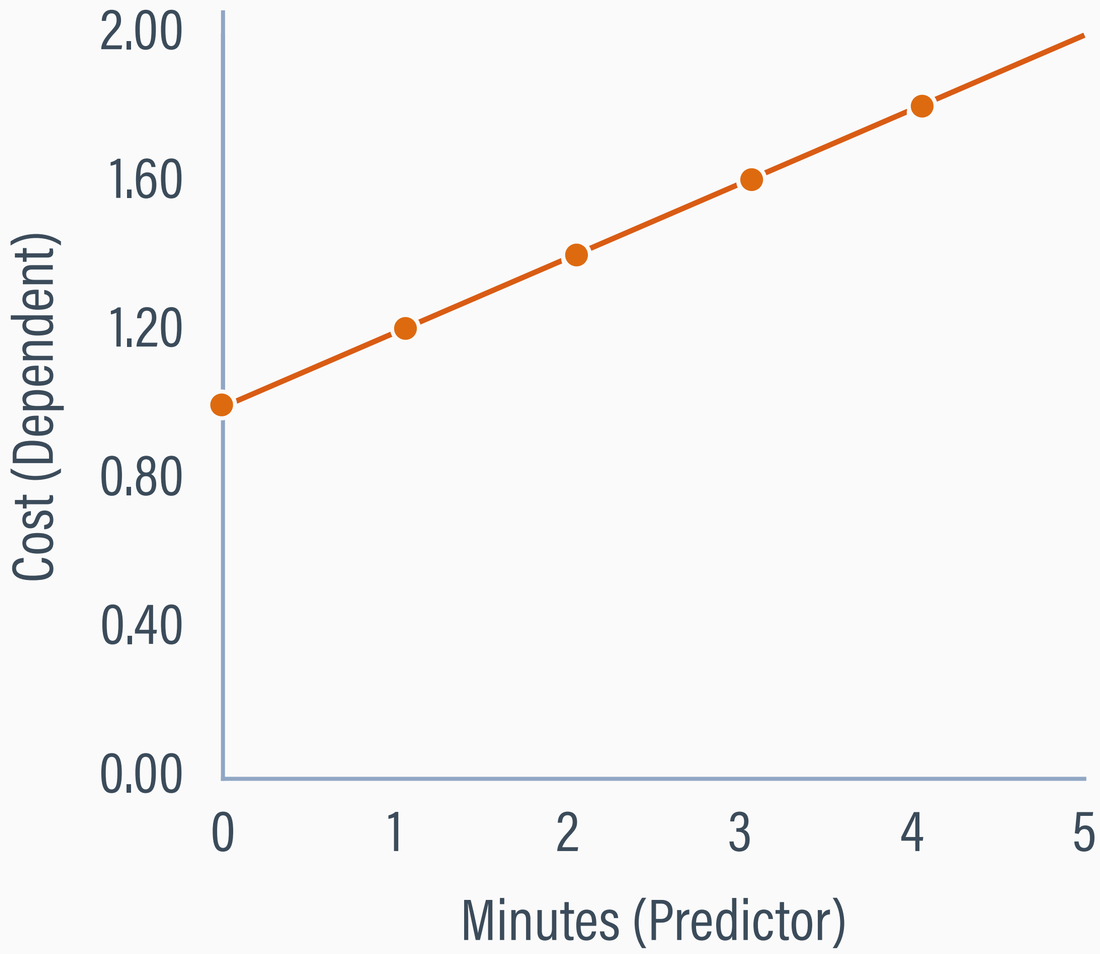
$$\text{cost} = 1.00 + 0.20(\text{minutes})$$

$$\text{cost} = 1.00 + 0.20(1) = 1.20$$



PERFECT PREDICTION

- The e-scooter application is atypical because all data points fall exactly on a straight line
- The linear equation perfectly predicts the cost for any value of minutes without error
- Real data may exhibit linear trends, but a straight line cannot perfectly describe data



OUTLINE

1. Regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

DAILY DIARY STUDY OF MOOD

This study delivered a daily diary to 61 HIV infected men who have sex with men (MSM) between 16 and 24 years old for 66 days to measure HIV-risk behaviors and other psychosocial variables. The study examined the association between daily life stressors and daily negative mood. The study also examined the person-level association between the average number of daily life stressors across the 66 days and the average level of negative mood across that period.

Manuel Ramirez

Patrick Wilson

Ramirez, M, Wilson, P., Mitchell, R., Enders, C., & Woller, M. (in progress). Daily variability in depressed mood among gay and bisexual youth living with HIV. *Manuscript in preparation*.

KEY VARIABLES

Daily Life Stressors

Respondents were presented with list of stressful events (e.g., fights with family or friends, work stress, financial stress), and they checked how many they experienced each day.

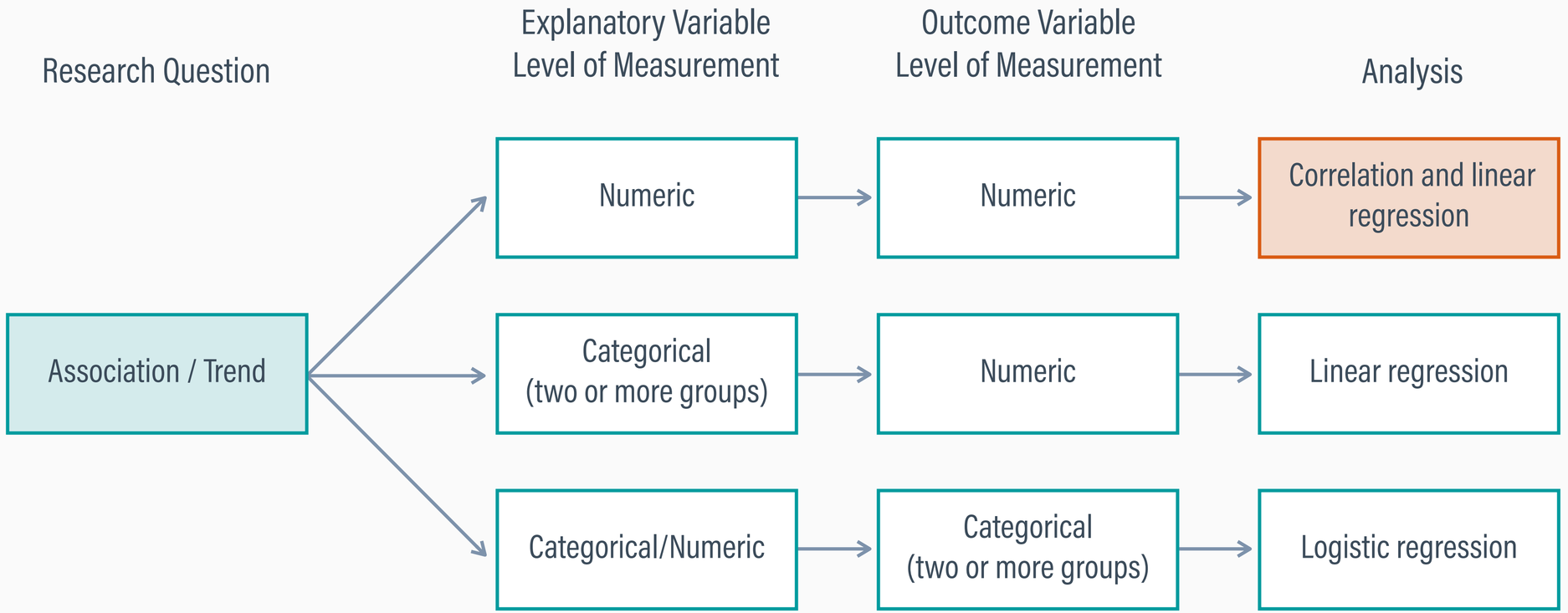
Depression

Depression was measured using the Profile of Mood States (POMS). The POMS Depression-Dejection scale is computed by summing responses to five questionnaire items, where higher ratings reflect more negative mood.

RESEARCH QUESTION

- Question: Is there an association or trend between one's number of life stressors and their average depressive mood?
- The explanatory (independent) variable, number of life stressors, is a numeric value derived from a checklist
- The outcome (dependent) variable, depressive mood, is a numeric scale derived from several questionnaire items

STATISTICAL ORG CHART



SIGNIFICANCE TESTING STEPS

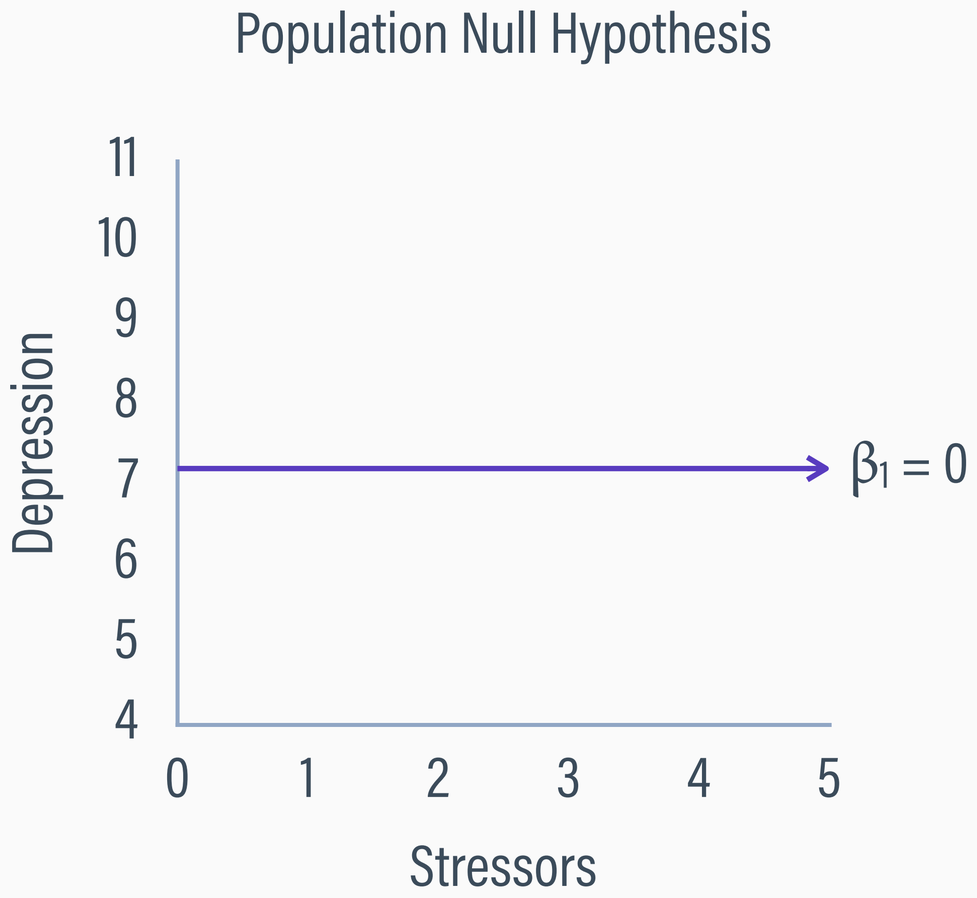
1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

NULL HYPOTHESIS

- Null hypothesis: In the population, there is no relation between stressors and depression

$$H_0: \beta_1 = 0$$

- The null is counter to expectations because researchers anticipate that changes in stressors could correspond with changes in depression



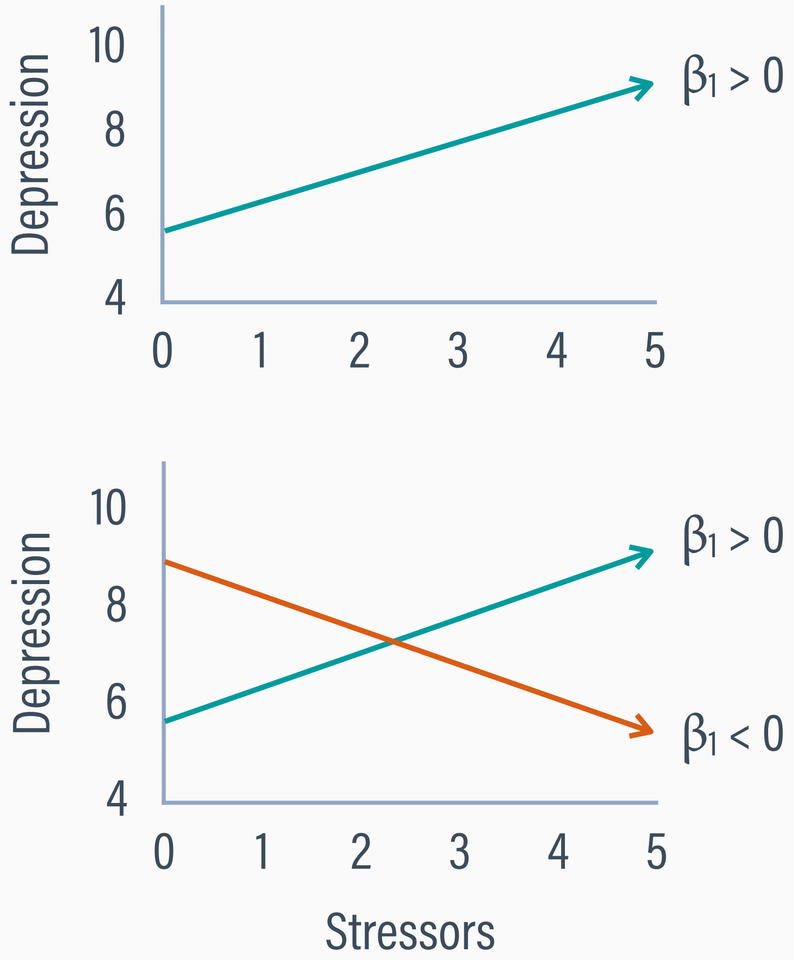
TWO POSSIBLE ALTERNATIVE HYPOTHESES

- One-tailed alternate: An increase in stressors could only be related to a corresponding increase in depression

$$H_A: \beta_1 > 0$$

- Two-tailed alternate: the relation between stressors and depression could be positive or negative

$$H_A: \beta_1 \neq 0 \ (\beta_1 > 0 \text{ or } \beta_1 < 0)$$

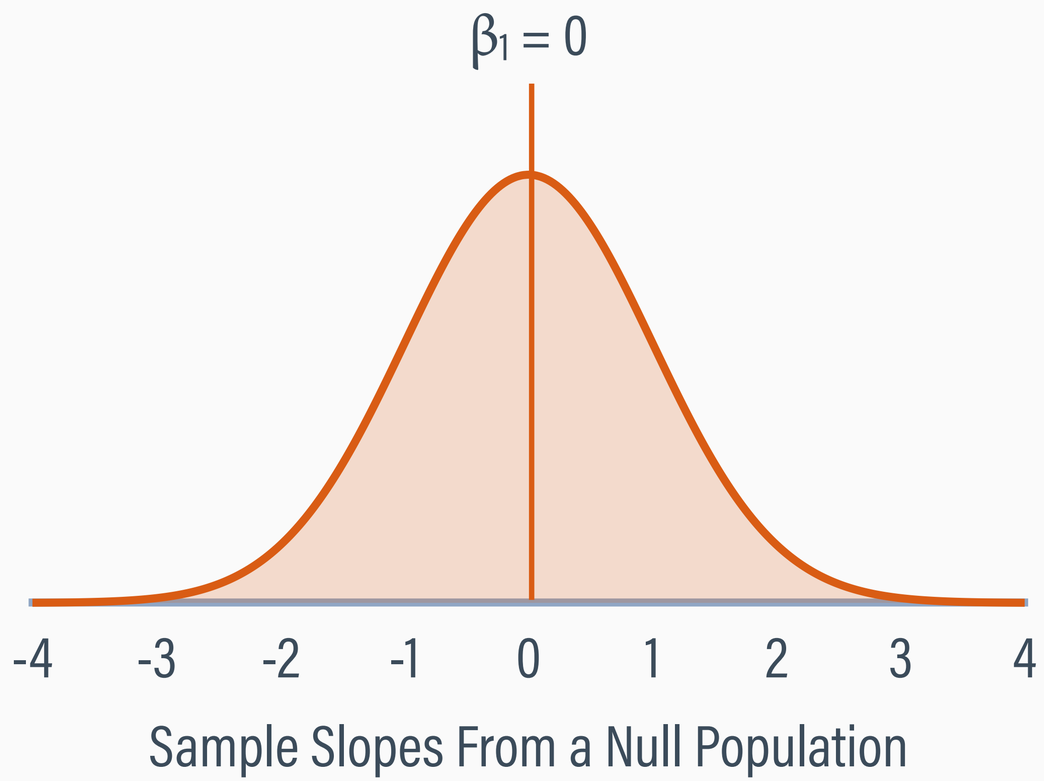


SIGNIFICANCE TESTING STEPS

1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

SAMPLING DISTRIBUTION IF THE NULL IS TRUE

- Like means and mean difference statistics, slopes vary across other hypothetical data sets that we could have worked with
- Samples from a null population would produce slopes that are normally distributed around a true slope of zero

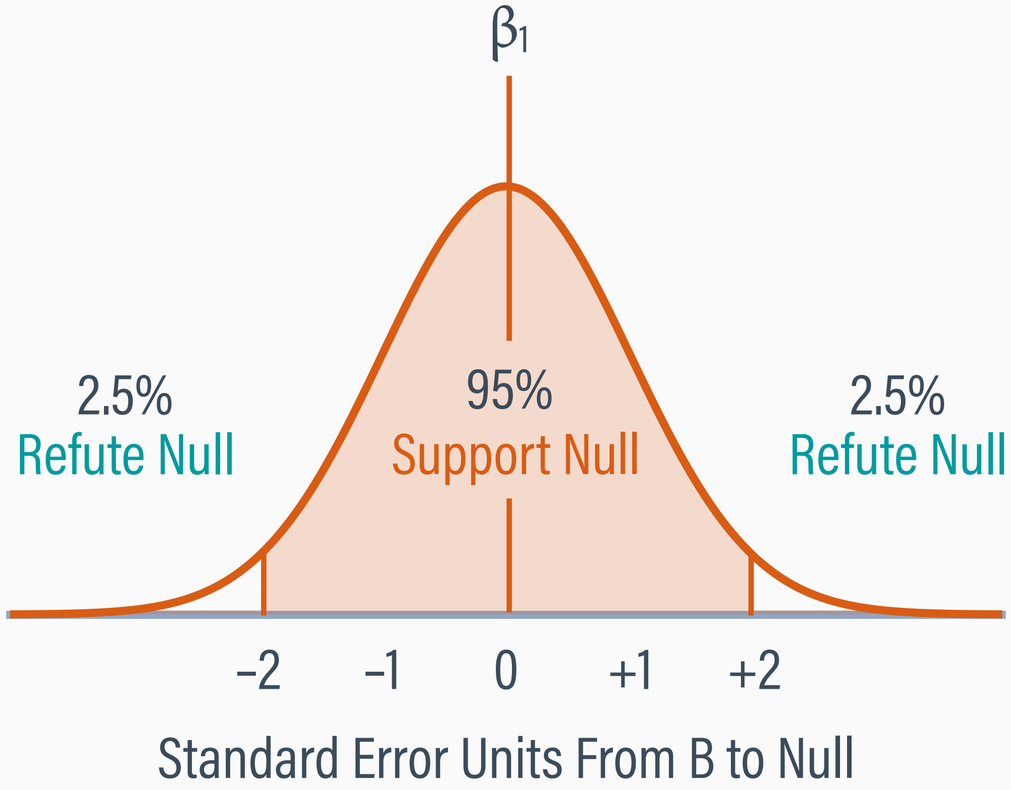


STANDARD OF EVIDENCE

- The data are the evidence that we use to conclude whether the null is plausible (“innocent”) or implausible (“guilty”)
- If the sample slope from our data (denoted B_1 with Roman letters) is very different from the null slope ($\beta_1 = 0$), then we conclude that the null hypothesis is implausible
- How big of a B_1 do we need to observe to refute the null?

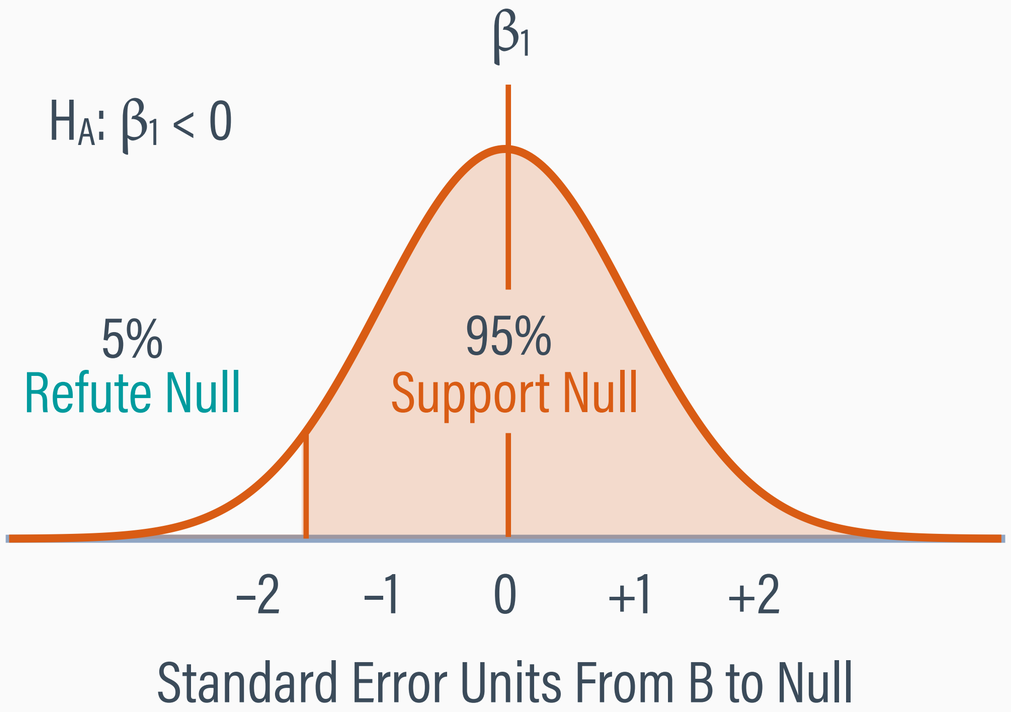
TWO-TAILED ALTERNATE HYPOTHESES

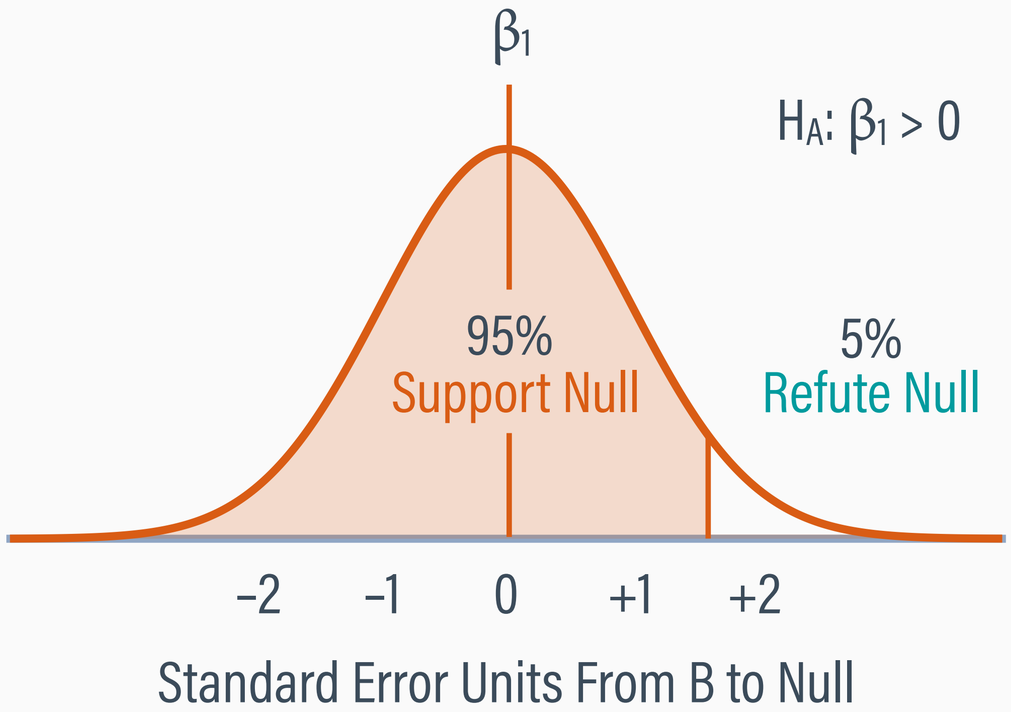
- By convention, we refute the null if the sample slope B_1 falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population ($p < .05$)
- The 5% rejection region (alpha level) is split in half to allow for the possibility that either an increase or a decrease provides evidence against H_0



ONE-TAILED ALTERNATE HYPOTHESES

- The 5% rejection region (alpha level) is placed in one tail, since only a positive (or only a negative) B_1 counts as evidence against H_0





SIGNIFICANCE TESTING STEPS

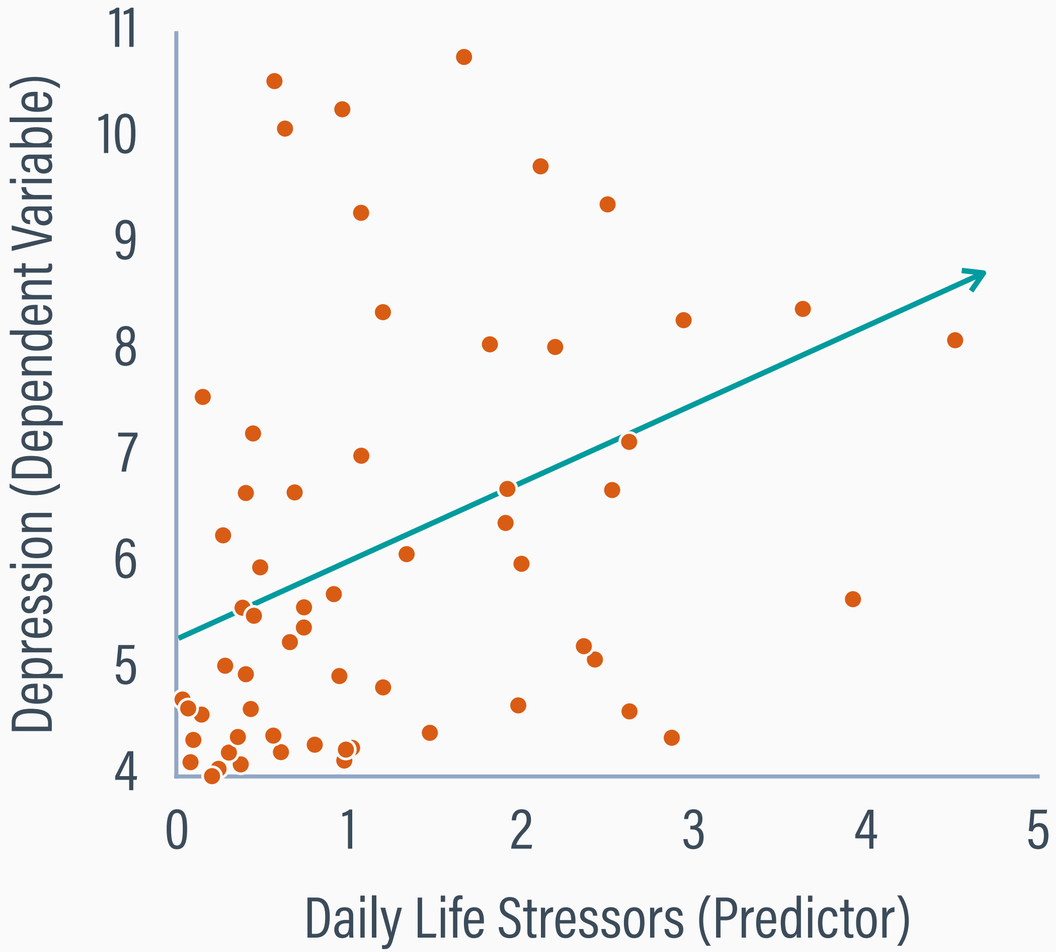
1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

LINEAR REGRESSION MODEL

- There is a positive trend: depression increases as daily stressors increase
- A straight line captures the overall positive trend, but the data do not fall on the line
- There is residual variation around the line

$$depress = \beta_0 + \beta_1(dailystress) + \varepsilon$$

depress = linear trend + residual



INTERCEPT AND SLOPE ESTIMATES

X = daily stressors

Y = depression

- The intercept and slope are calculated using means, standard deviations, and correlation

$$B_1 = r\left(\frac{s_Y}{s_X}\right) = \text{correlation} \times \frac{\text{dependent std. dev.}}{\text{predictor std. dev.}}$$

$$B_0 = \bar{Y} - B_1\bar{X} = \text{dependent mean} - (\text{slope} \times \text{predictor mean})$$

- Roman letters indicate that these are estimates from a sample rather than population statistics

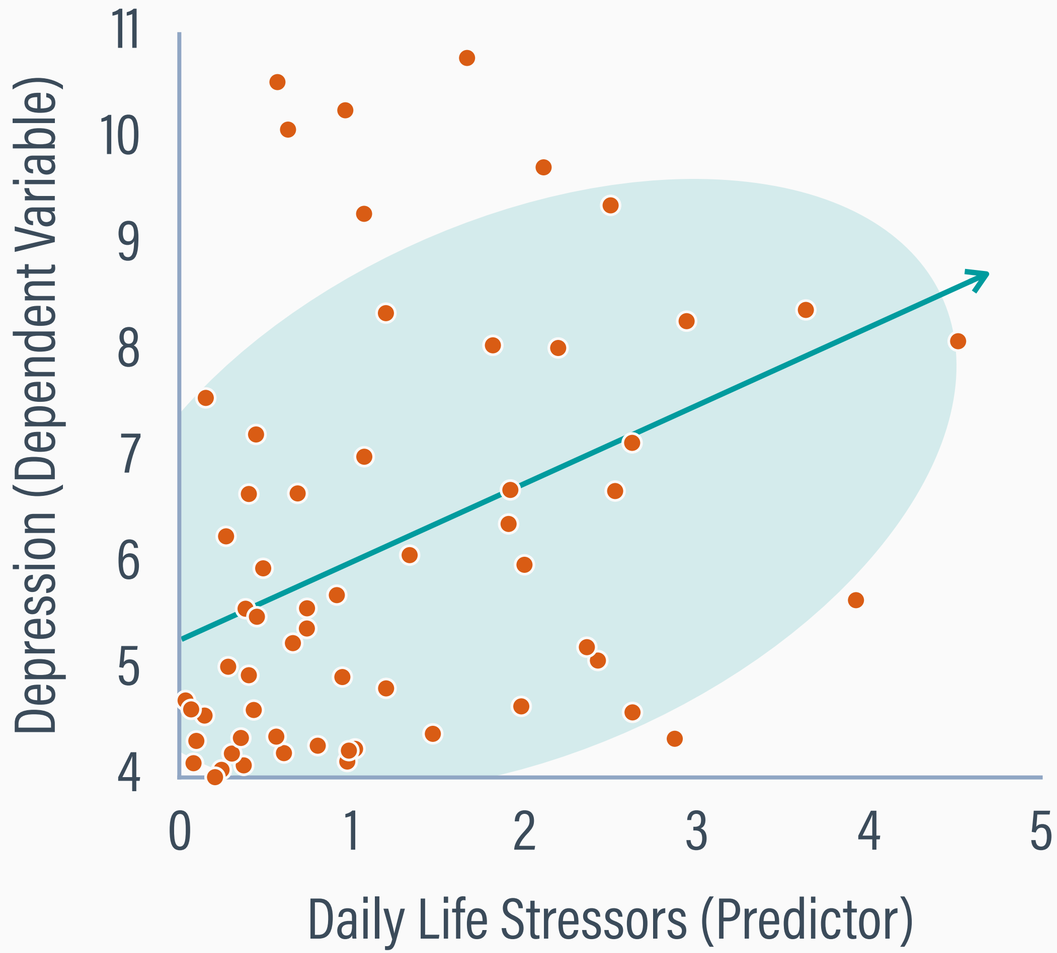
DIARY DATA EXAMPLE

- The slope is positive because the correlation is positive (r = +.37)

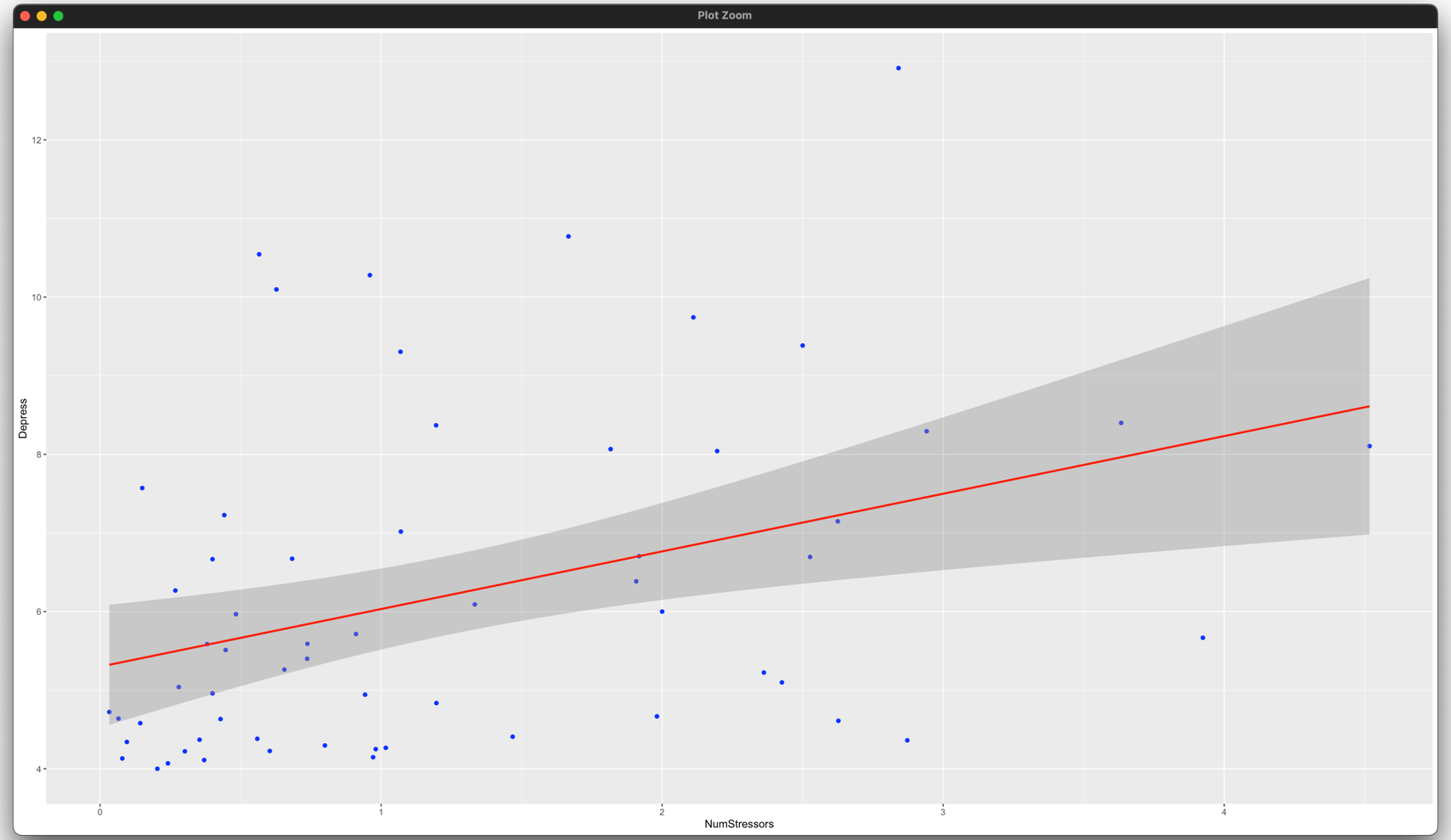
Variable	Mean	SD	r
Stressors	1.25	1.07	0.37
Depression	6.21	2.09	

$$B_1 = r\left(\frac{s_{depress}}{s_{stressors}}\right) = .37\left(\frac{2.09}{1.07}\right) = 0.73$$

$$B_0 = (\text{depression mean}) - B_1(\text{stressors mean})$$
$$= 6.21 - (0.73 \times 1.25) = 5.30$$



R OUTPUT



R OUTPUT

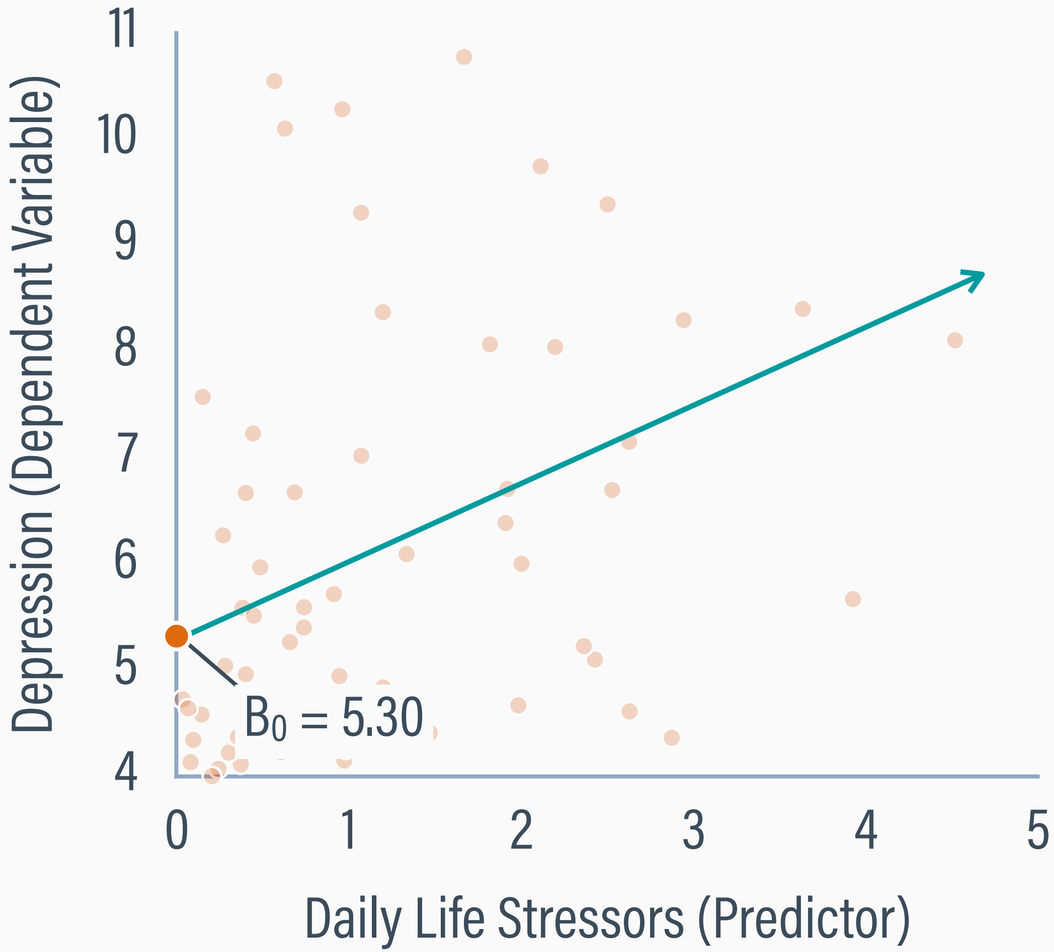
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0436 -1.3404 -0.4641  0.8722  5.5312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.2988     0.3877  13.666  < 2e-16 ***
NumStressors   0.7334     0.2371   3.093  0.00303 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.96 on 59 degrees of freedom
Multiple R-squared:  0.1395,  Adjusted R-squared:  0.1249
F-statistic: 9.565 on 1 and 59 DF,  p-value: 0.003028
```

INTERCEPT INTERPRETATION

- The intercept is the predicted value of the depression (the dependent variable) when daily stressors (the predictor) equals zero
- For a hypothetical person with no daily stressors, the predicted depression score is $B_0 = 5.30$

$$\text{depress} = B_0 + B_1(\text{stressors}) + \text{residual}$$

$$\text{depress} = 5.30 + 0.73(\text{stressors}) + \text{residual}$$



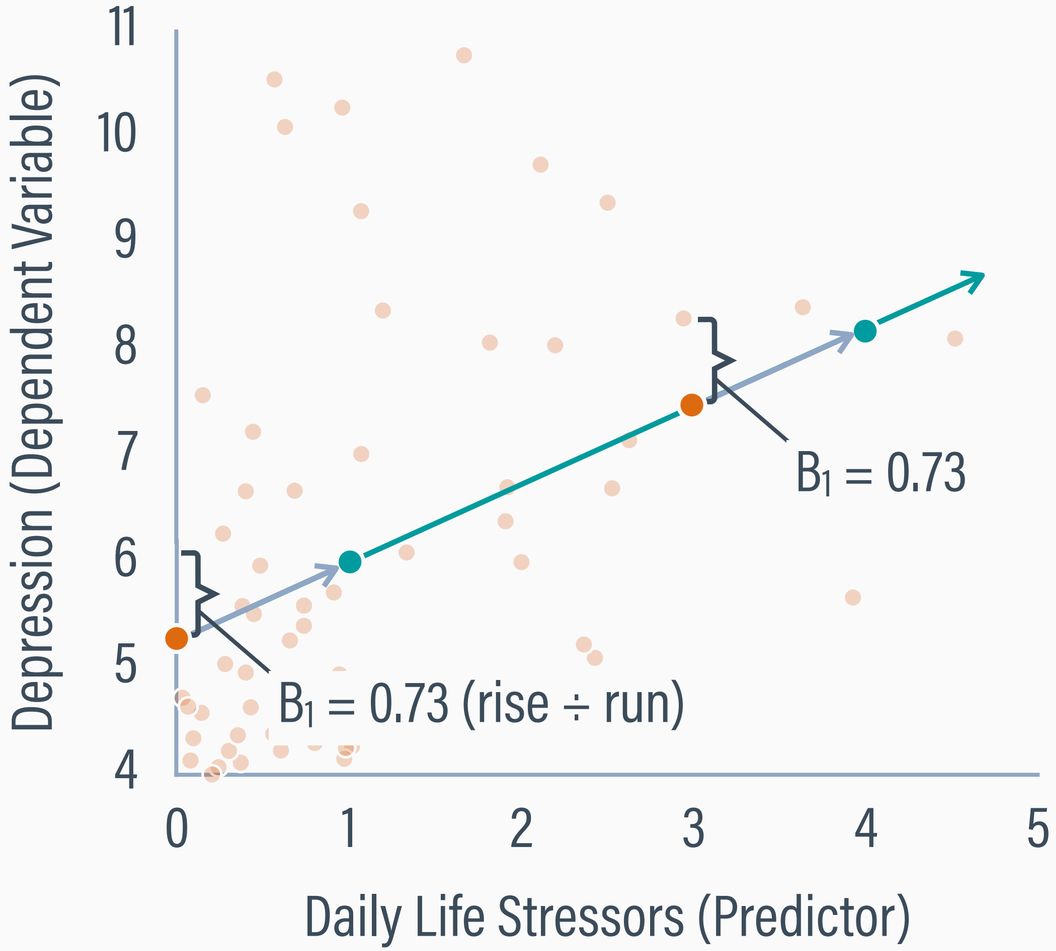
SLOPE INTERPRETATION

- The slope is the expected depression (dependent variable) difference for a 1-point increase in daily stressors (the predictor)

- For any two people who differ by one daily stressor, the person with the additional stressor is expected to be $B_1 = 0.73$ points higher on the depression scale

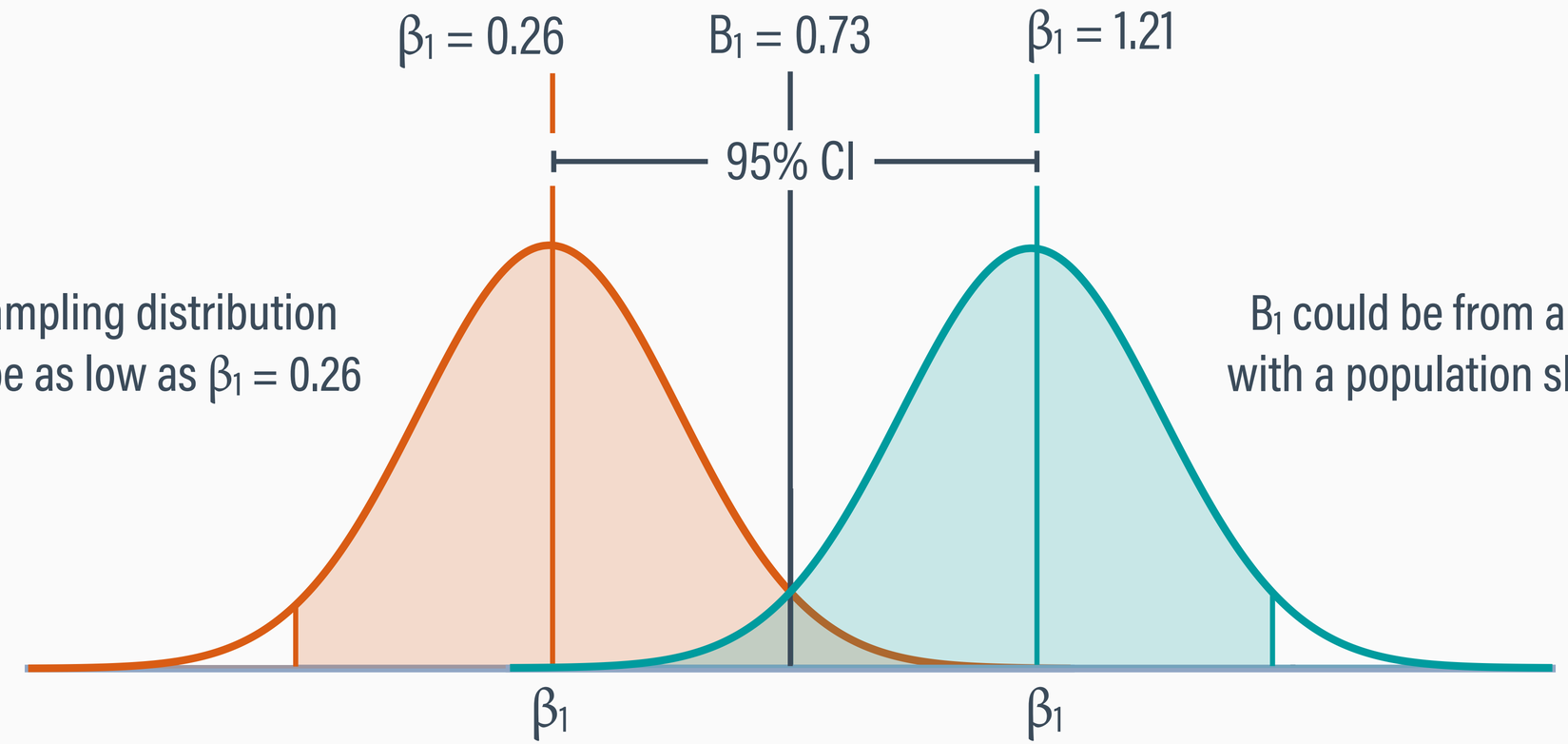
$$\text{depress} = B_0 + B_1(\text{stressors}) + \text{residual}$$

$$\text{depress} = 5.30 + 0.73(\text{stressors}) + \text{residual}$$



95% CONFIDENCE INTERVAL

- The 95% confidence interval gives the two most extreme values of the population slope that could have reasonably produced the sample B_1



R OUTPUT

```
                   2.5 %   97.5 %
(Intercept)  4.5229278 6.074675
NumStressors 0.2588819 1.207889
```

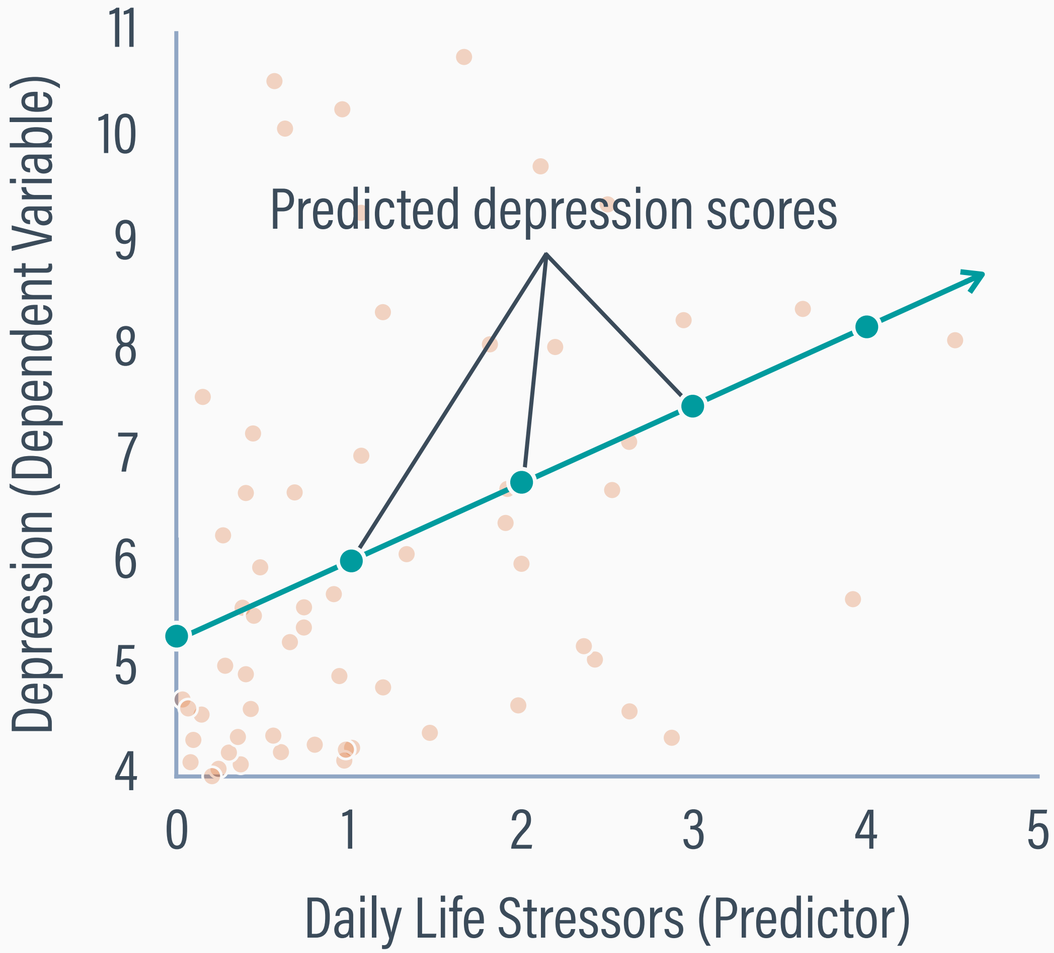
The study produced a sample slope and 95% confidence interval of $B_1 = 0.73$ and $CI_{95\%} = [0.26, 1.21]$. In small groups of two or three, discuss whether this sample of participants could have reasonably originated from a population where there is truly no association between daily stressors and depression ($\beta_1 = 0$).

PREDICTED VALUES

- Plugging daily stress scores into the linear equation (ignoring the residual) gives predicted depression scores on the line

$$\widehat{depress} = 5.30 + 0.73(stressors)$$

- Predicted points on the line are denoted with “hats” over the dependent variable’s name to differentiate them from the actual scores



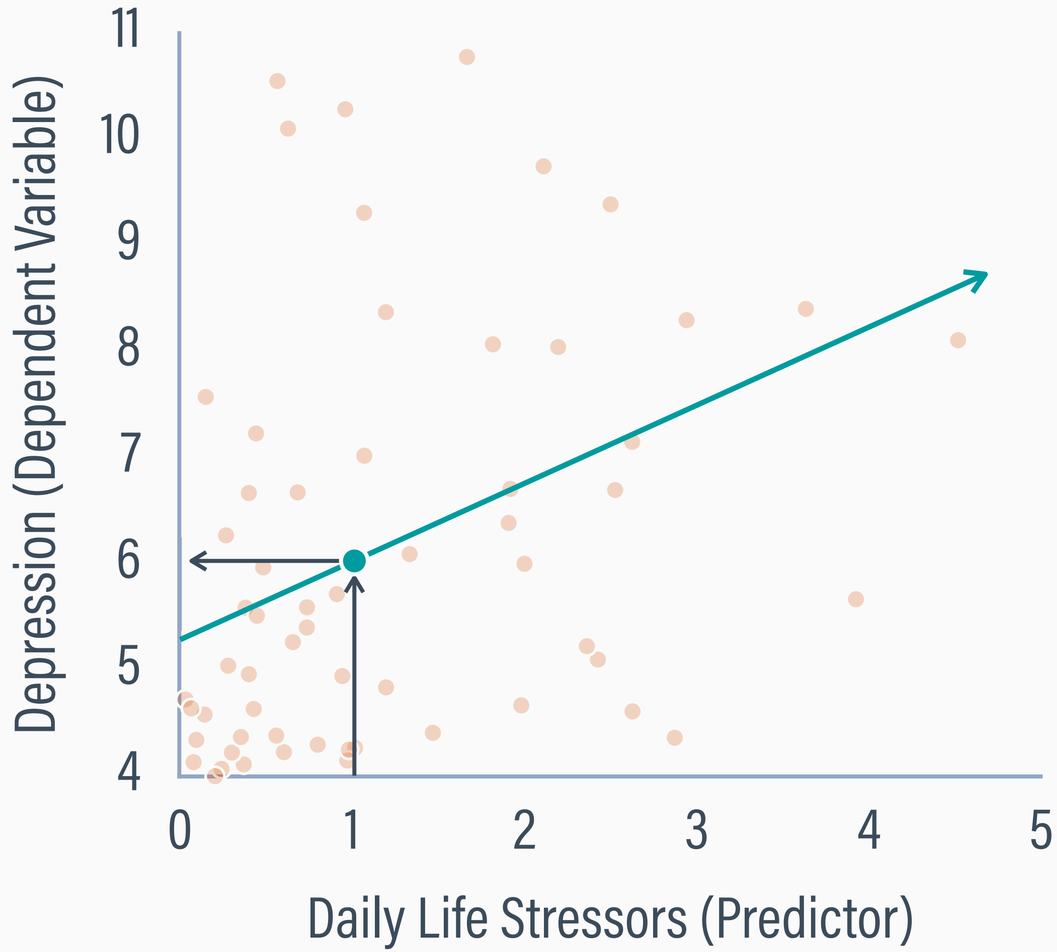
PREDICTED SCORE EXAMPLE

- Substituting the number of stressors gives the predicted depression score

$$\widehat{\text{depress}} = 5.30 + 0.73(\text{stressors})$$

$$\widehat{\text{depress}} = 5.30 + 0.73(1) = 6.03$$

- A predicted value is like a mean: the average depression for someone with a particular level of daily stressors (depression varies around this point)



RESIDUALS

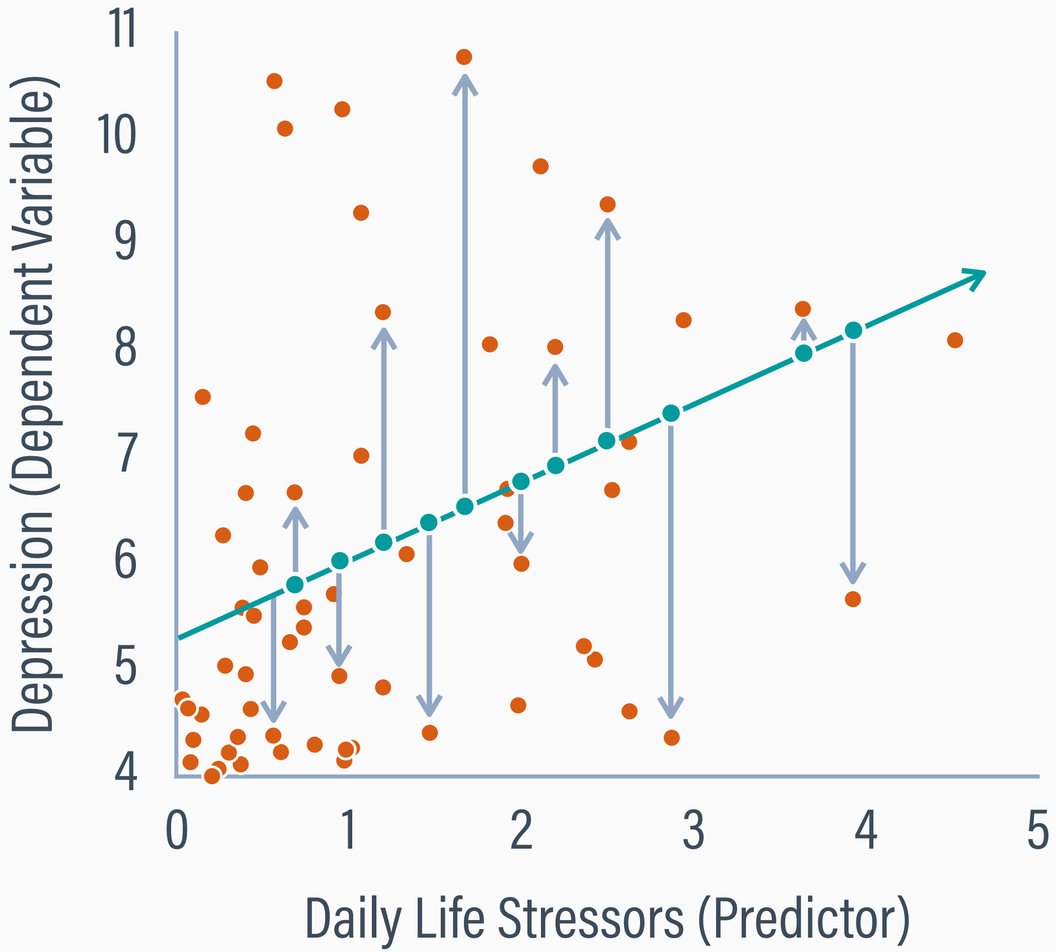
- Residuals are distances from the actual depression scores and the predicted depression values on the line

$$\text{depress} = B_0 + B_1(\text{stressors}) + \text{residual}$$

$$\text{depress} = \widehat{\text{depress}} + \text{residual}$$

$$\text{residual} = \text{depress} - \widehat{\text{depress}} = Y - \hat{Y}$$

- The standard deviation of the residuals, $s_{residual} = 1.96$, is the average distance to the line (the average prediction error)



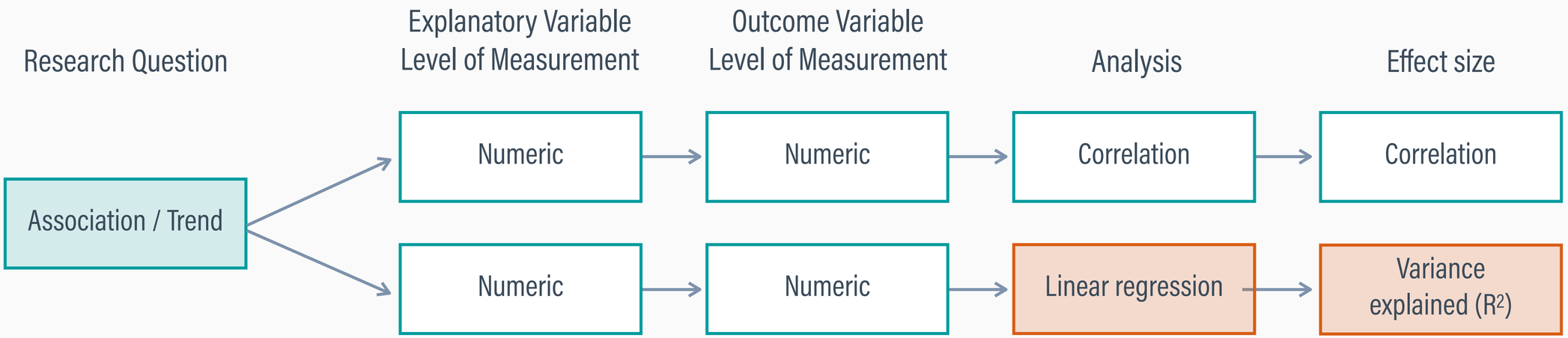
LEAST SQUARES ESTIMATES

- B_0 and B_1 are called the line of best fit or least squares estimates because they minimize the sum of the (squared) residuals or distances to the line
- No other line comes closer to the data points
- No other values of B_0 and B_1 could give a smaller residual standard deviation (prediction errors are minimized)

EFFECT SIZE

- Effect size refers to a family of measures that quantify the magnitude of an effect, independent of sample size
- The goal is to express how big an effect is, not just whether it likely exists at the population level
- Like ANOVA, regression models use the proportion variance explained (R^2) effect size

STATISTICAL ORG CHART

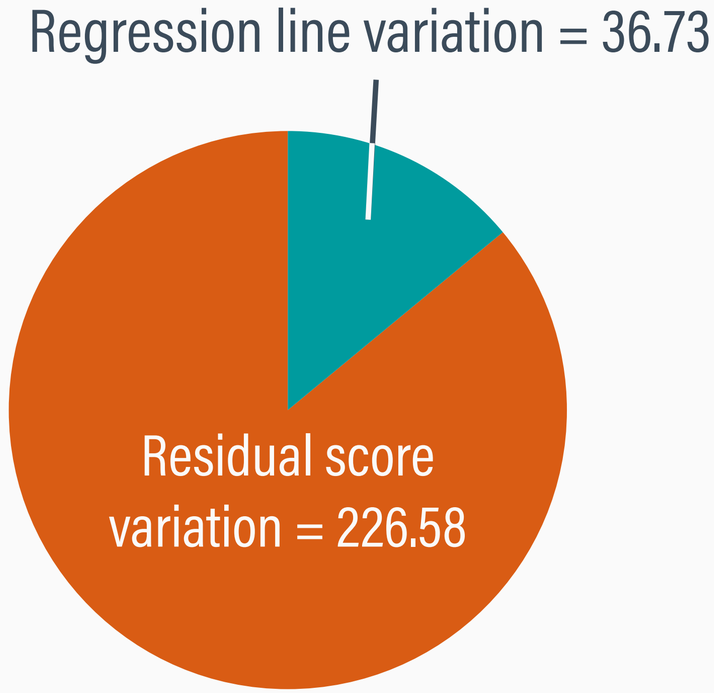


R-SQUARE EFFECT SIZE

- Like ANOVA, regression partitions variation into two sources: the effect of the predictor (regression line) and residual (leftover)

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{36.73}{263.31} = .14$$

- Depression differences attributable to changes in stressors comprise about 14% of the total depression variation pie



Sum of squares regression =	**36.73**
Sum of squares residual =	**226.58**
Sum of squares total =	**263.31**

R OUTPUT

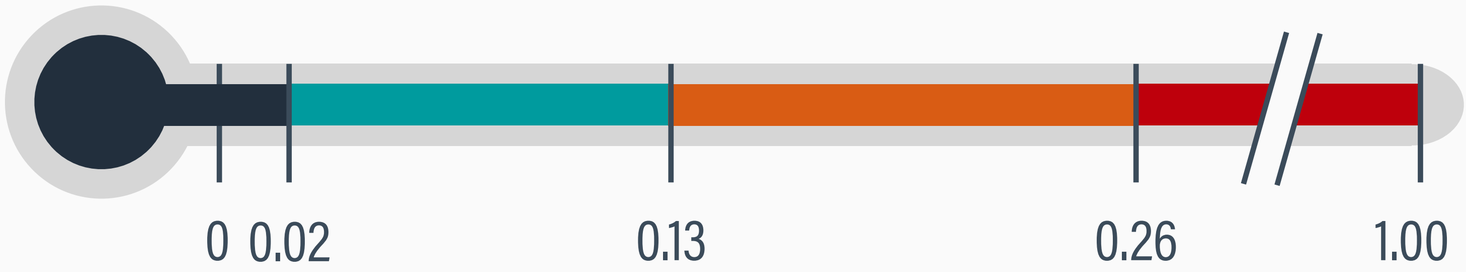
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0436 -1.3404 -0.4641  0.8722  5.5312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.2988     0.3877  13.666  < 2e-16 ***
NumStressors   0.7334     0.2371   3.093  0.00303 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.96 on 59 degrees of freedom
Multiple R-squared:  0.1395,  Adjusted R-squared:  0.1249
F-statistic: 9.565 on 1 and 59 DF,  p-value: 0.003028
```

R-SQUARE EFFECT SIZE GUIDELINES

Small = .02 to .13 (2% to 13%)

Moderate = .13 to .26 (13% to 26%)

Large = greater than .26 (26% to 100%)

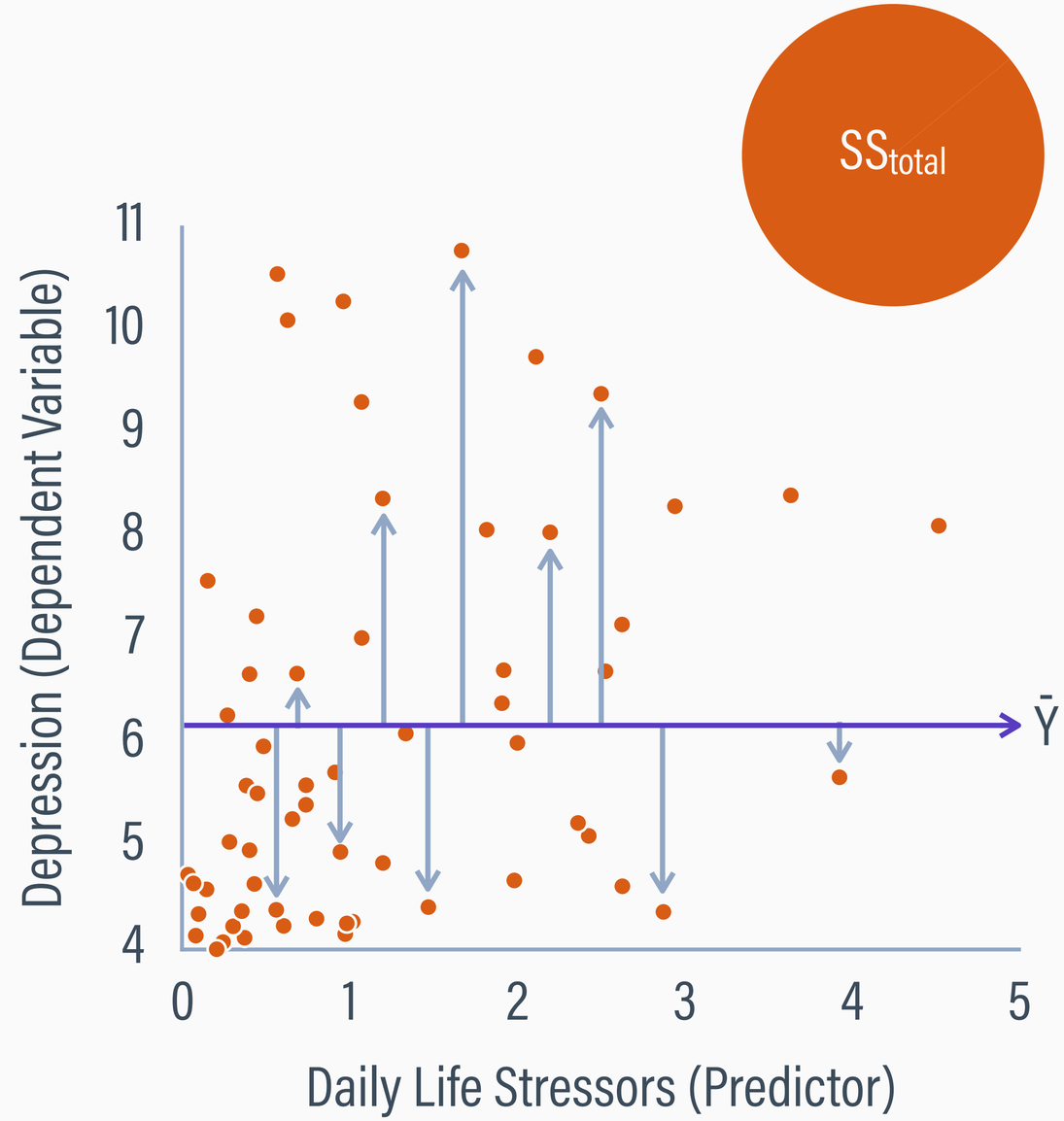


SUM OF SQUARES TOTAL

- The total sum of squares is the sum of squared distances from the scores to the mean (same as ANOVA)

$$SS_{total} = \sum(Y - \bar{Y})^2 = \sum(\text{score} - \text{grand mean})^2$$

- The sum of squares expresses the total amount of depression variability in the data as a lump sum (the total area of the pie)



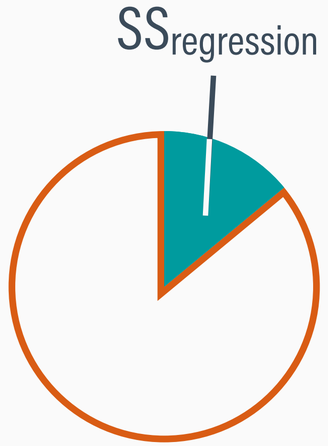
SUM OF SQUARES REGRESSION

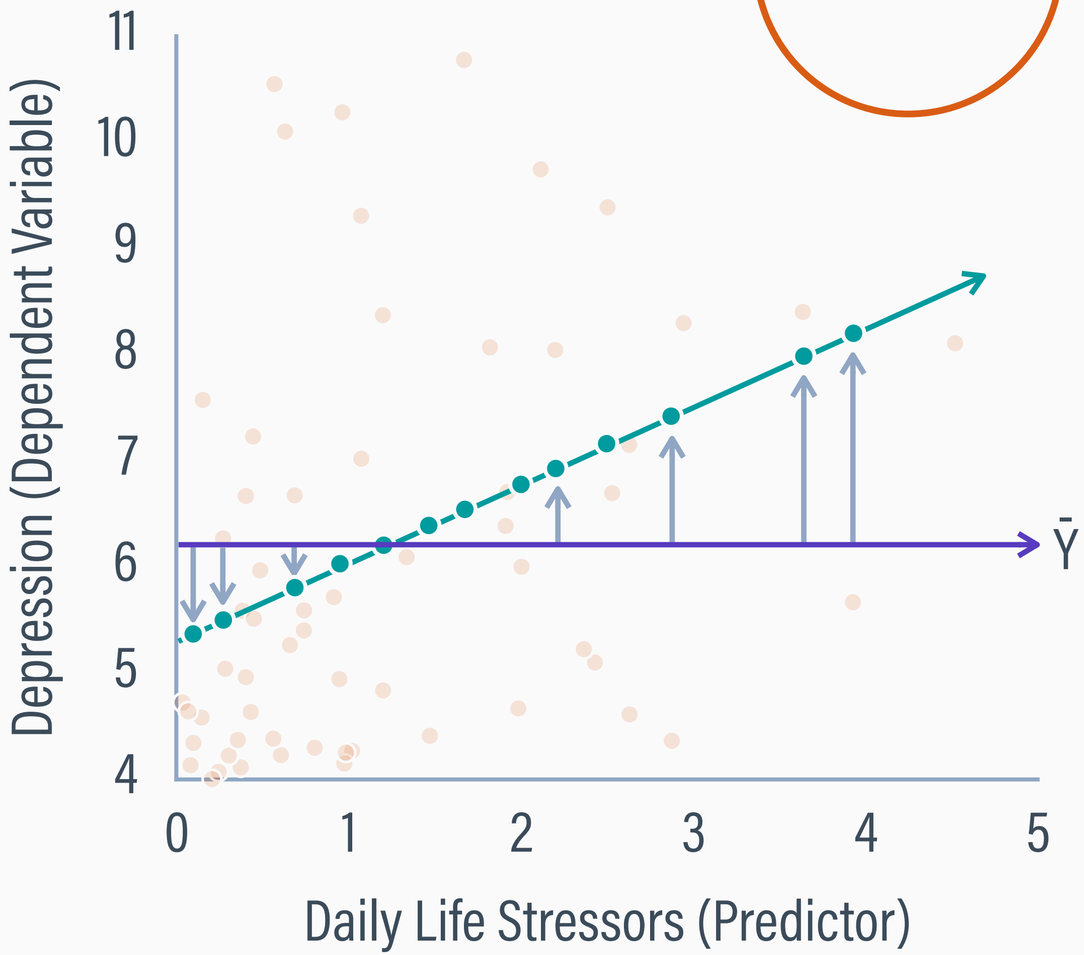
- In ANOVA, the effect of the independent variable was captured by a deviation between the group mean and grand mean

$$SS_{regression} = \sum(\hat{Y} - \bar{Y})^2 =$$

$$\sum(\text{predicted value on line} - \text{grand mean})^2$$

- Similarly, in regression, that effect is a deviation between a point on the regression line (similar to a mean) and the grand mean





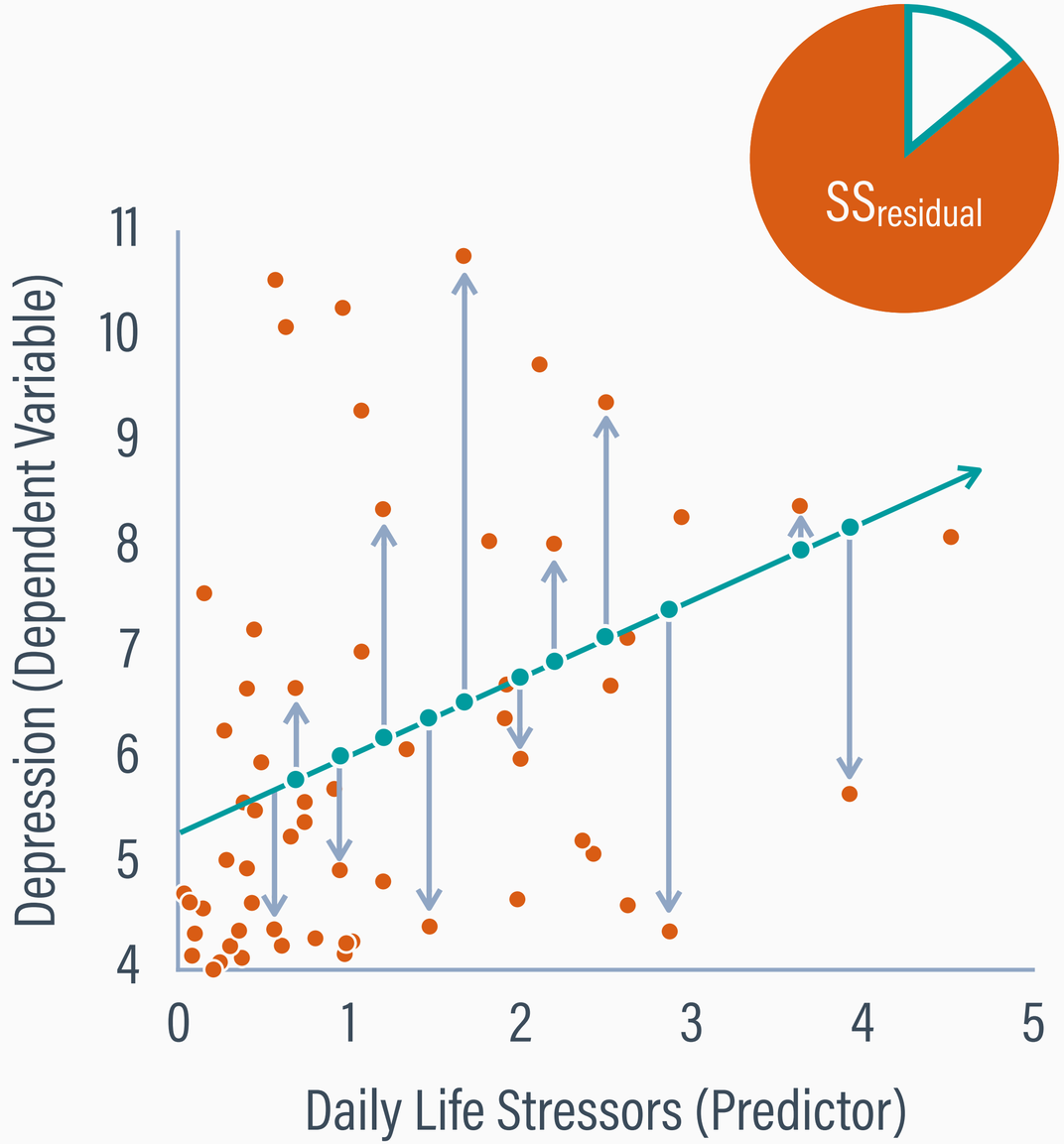
SUM OF SQUARES RESIDUAL

- In ANOVA, residual (leftover) variation was captured by a deviation between the scores and their group means

$$SS_{residual} = \sum (Y - \hat{Y})^2 =$$

$$\sum (\text{score} - \text{predicted value on line})^2$$

- Similarly, in regression, the residual is a deviation between the scores and points on the regression line (similar to means)



SIGNIFICANCE TESTING STEPS

1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

COMPARING DATA TO THE NULL

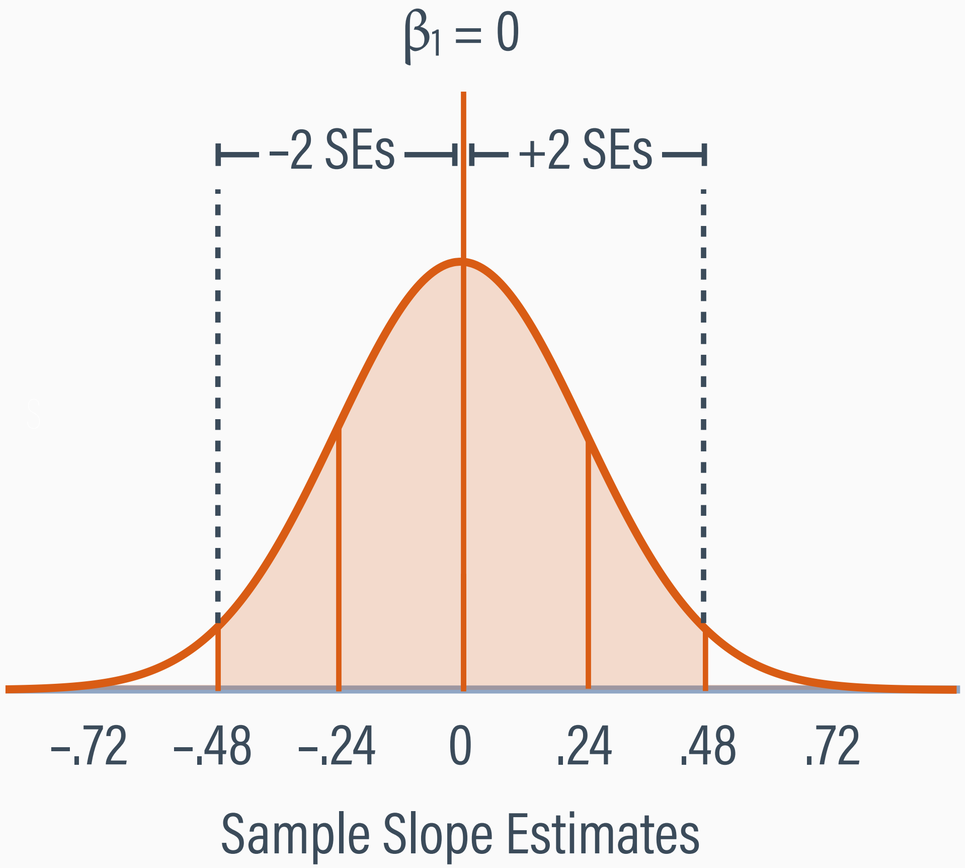
- Two ways to determine whether the sample B_1 is consistent (or inconsistent) with the null population slope
- The t-statistic gives a standardized distance between the sample slope and the null hypothesis slope (like a z-score)
- A p-value tells us how likely it is that hypothetical samples like our data would originate from the null population

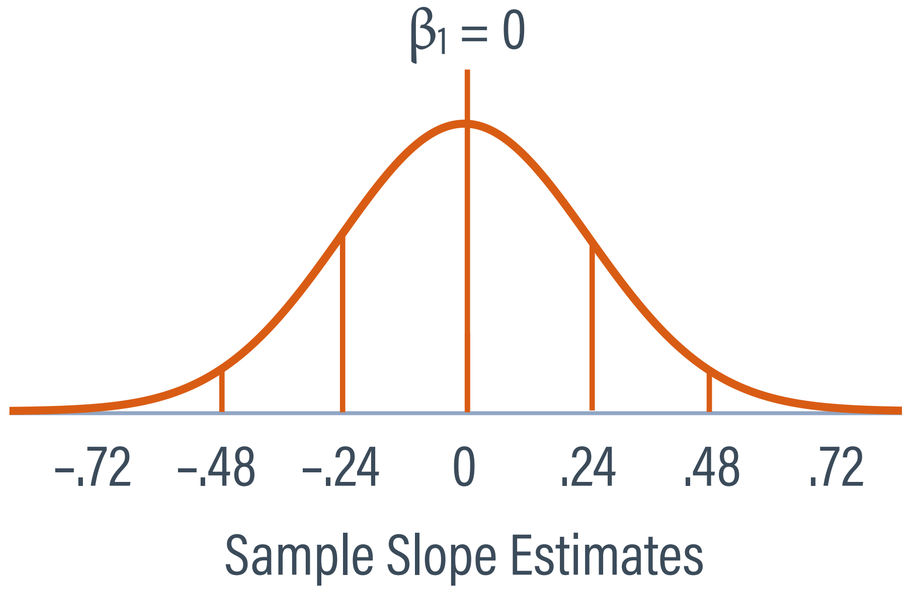
STANDARD ERROR OF A SLOPE

- Across many hypothetical samples from a null population, we would expect the data's slope to differ from $\beta_1 = 0$ by $\pm$ 0.24

$$s_{B_1} = \frac{\text{variance of residuals}}{\text{predictor sum of squares}} = 0.24$$

- The standard error allows us to construct the sampling distribution of B_1 under the assumption that the null is true





Consider the sampling distribution of sample slopes from a null population with $\beta_1 = 0$. The sample slope was $B_1 = 0.73$ ($s_{B1} = 0.24$). In small groups of two or three, discuss whether the data provide evidence for or against the null hypothesis.

t-STATISTIC

- The t-statistic quantifies the number of standard error units that separate the sample slope and null hypothesis population slope

$$t = \frac{B_1 - \beta_1}{s_{B_1}} = \frac{\text{distance from the null}}{\text{standard error (std. dev. of } B_1)}$$

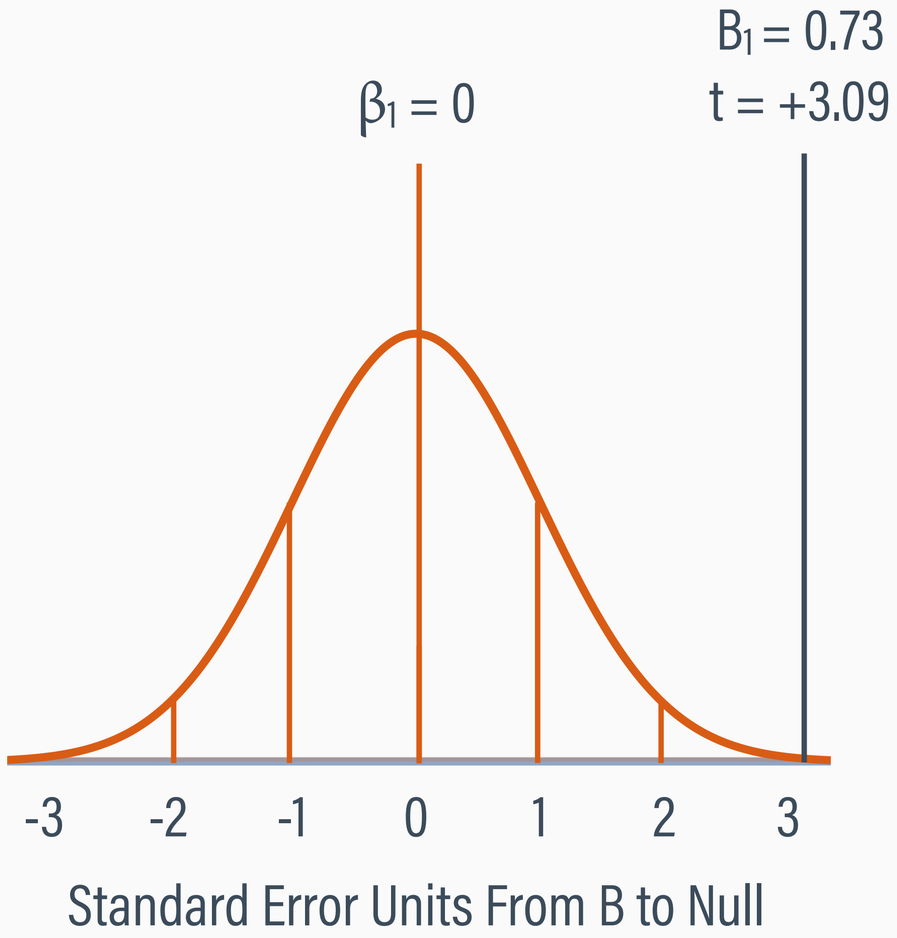
- The t-statistic is the same as a z-score (a standardized metric where distance is expressed in standard deviation units)

t-STATISTIC EXAMPLE

- The t-statistic indicates that 3.09 standard error units separate the sample slope and null

$$t = \frac{B_1 - \beta_1}{s_{B_1}} = \frac{0.73 - 0}{0.24} = 3.09$$

- The positive sign reflects the slope's direction

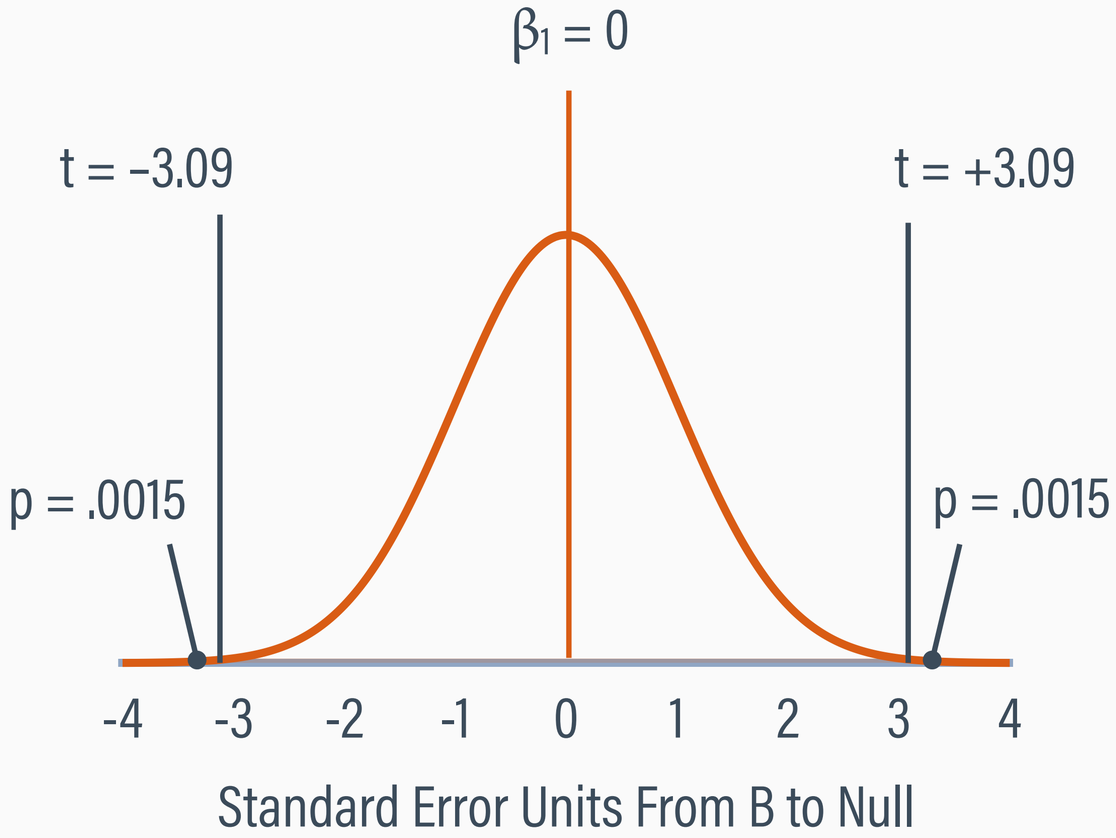


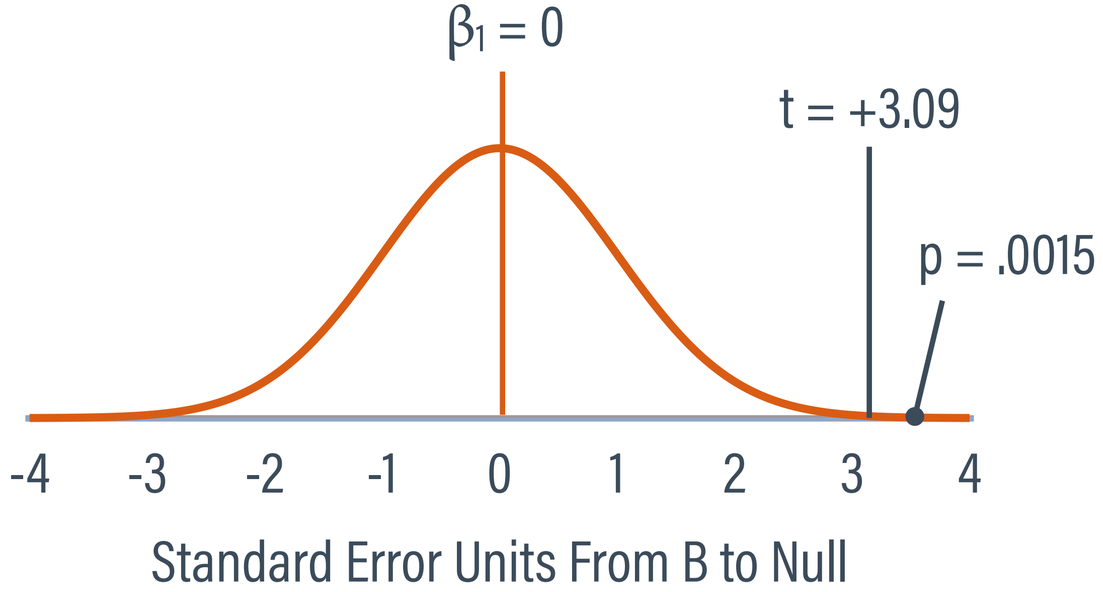
PROBABILITY VALUES (P-VALUES)

- A p-value is defined as proportion of hypothetical samples that have a t-statistic at least as large as the sample data
- Assuming the null is true, how likely is it to draw a sample with a B_1 value at least as large as the one from our data?
- Visually, probability is an area under the curve, obtained by applying calculus integrals to the t-distribution function

TWO-TAILED P-VALUE

- The p-value tells how likely it is to draw a sample slopes at least as extreme as ours from a null population with $\beta_1 = 0$
- The probability of drawing a sample from the null population with a t-statistic of at least ±3.09 is p = .003 (3 out of 1000)
- Only 3 out of every 1000 hypothetical samples from a null population would have t-statistics this large





Suppose the researchers had instead specified a one-tailed test where the predicted a positive association (i.e., increases in stressors could only lead to increases in depression). In small groups of two or three, discuss how the p-value would change with a one-tailed alternate hypothesis. Would your conclusion about significance change or stay the same?

R OUTPUT

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.0436 -1.3404 -0.4641  0.8722  5.5312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.2988     0.3877  13.666  < 2e-16 ***
NumStressors   0.7334     0.2371   3.093  0.00303 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.96 on 59 degrees of freedom
Multiple R-squared:  0.1395,  Adjusted R-squared:  0.1249
F-statistic: 9.565 on 1 and 59 DF,  p-value: 0.003028
```

SIGNIFICANCE TESTING STEPS

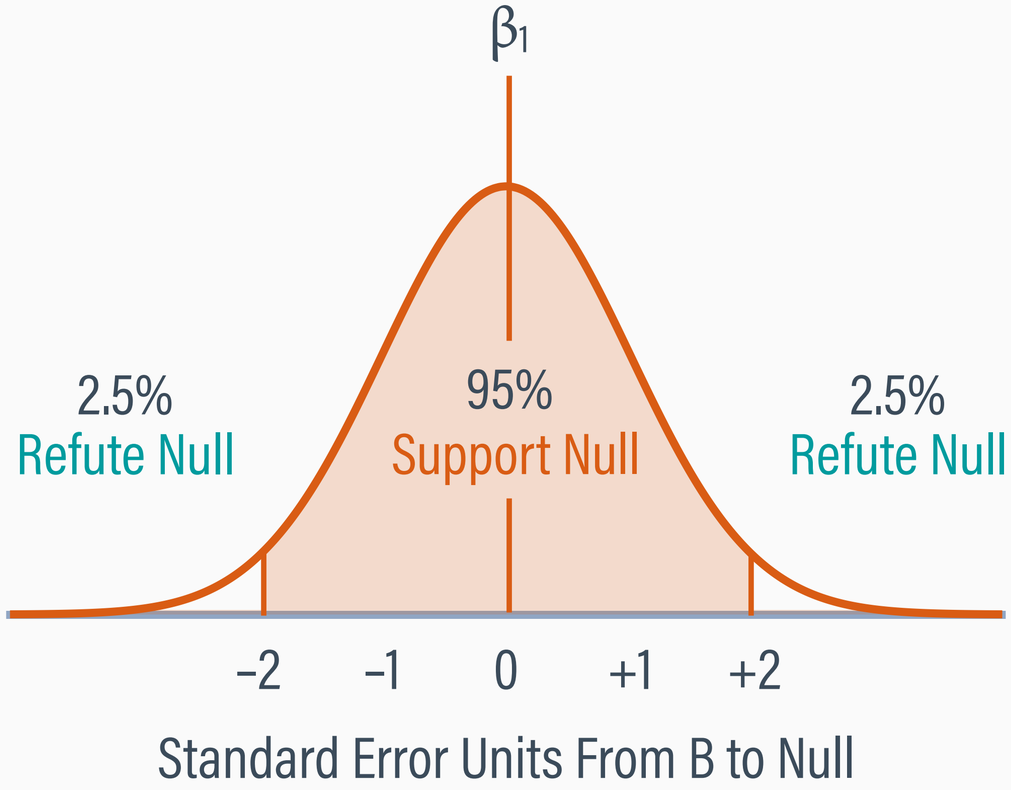
1. Specify hypotheses about population
2. Define standard of evidence
3. Design study and collect data
4. Compare data to null hypothesis
5. Evaluate hypotheses and draw conclusion

RESEARCH QUESTION REVISITED

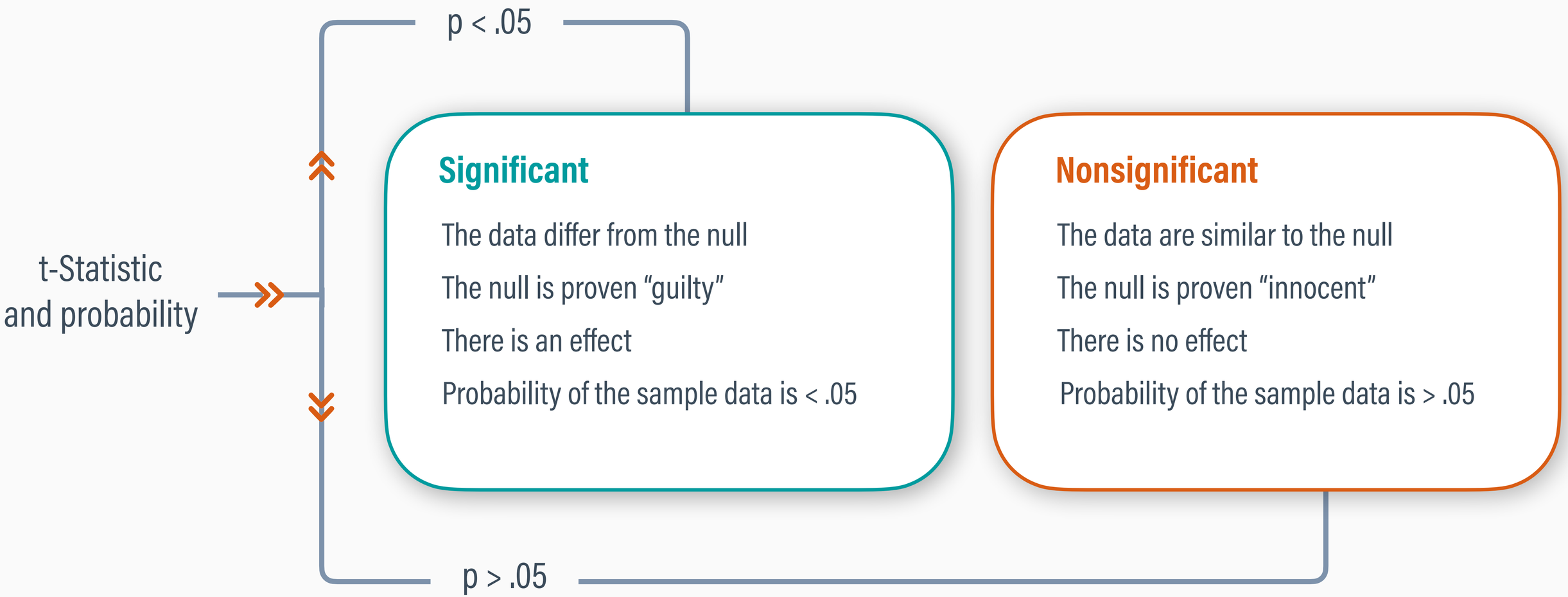
- Studies typically attempt to answer a handful of research questions involving associations between key variables
- Is there an association between daily life stressors and daily depressive symptoms?
- The null hypothesis states that there is no association between life stressors and depression (a flat slope)

5% SIGNIFICANCE CRITERION REVISITED

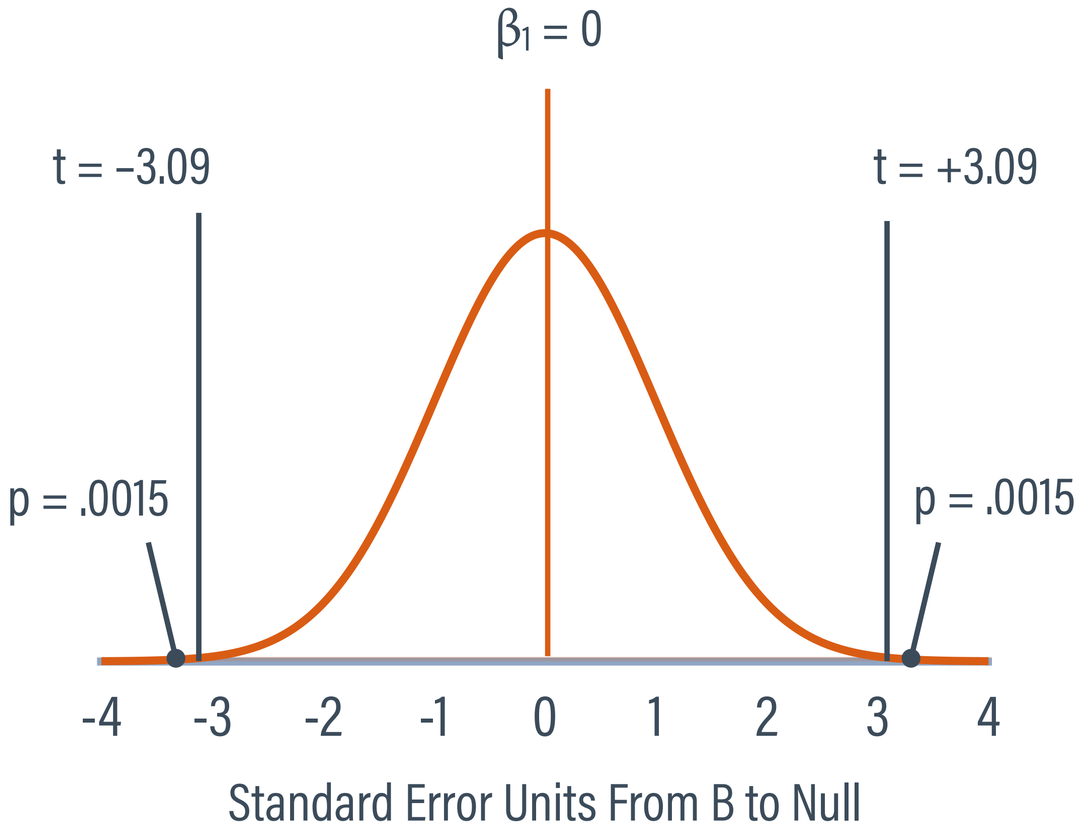
- By convention, we refute the null if the sample slope B_1 falls outside the middle 95% of the sampling distribution
- Such as sample has less than a 5% chance of originating from the null population
- We deem the null implausible because our data are unlikely to originate from that population



DECISION TREE



The two-tailed probability for the study is p = .003. In small groups of two or three, discuss your decision about the the null hypothesis. Translate your decision into a tangible statement about the association between daily stressors and depression.



CONCLUSION: TWO-TAILED ALTERNATE

- The p-value of .003 (3 out of 1000) would lead us to refute the null
- A sample slope as large as $B_1 = \pm 0.73$ (or a t-statistic at least ± 3.09) is very unlikely to have originated from a null population with $\beta_1 = 0$
- There is evidence that an increase in daily stressors is associated with a concurrent increase in depression

FALSE POSITIVES (TYPE I ERRORS)

- The 5% rejection region is an area of the distribution that contains outlier samples that are unlikely *but not impossible*
- When B_1 falls in the rejection region (evidence against the null), there is still a 5% chance it came from the null population
- We conclude there is an association, while acknowledging that there is a 5% chance of a false positive—incorrectly rejecting the null when it is actually true (a Type I error)

OUTLINE

1. Regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

STATISTICAL ASSUMPTIONS

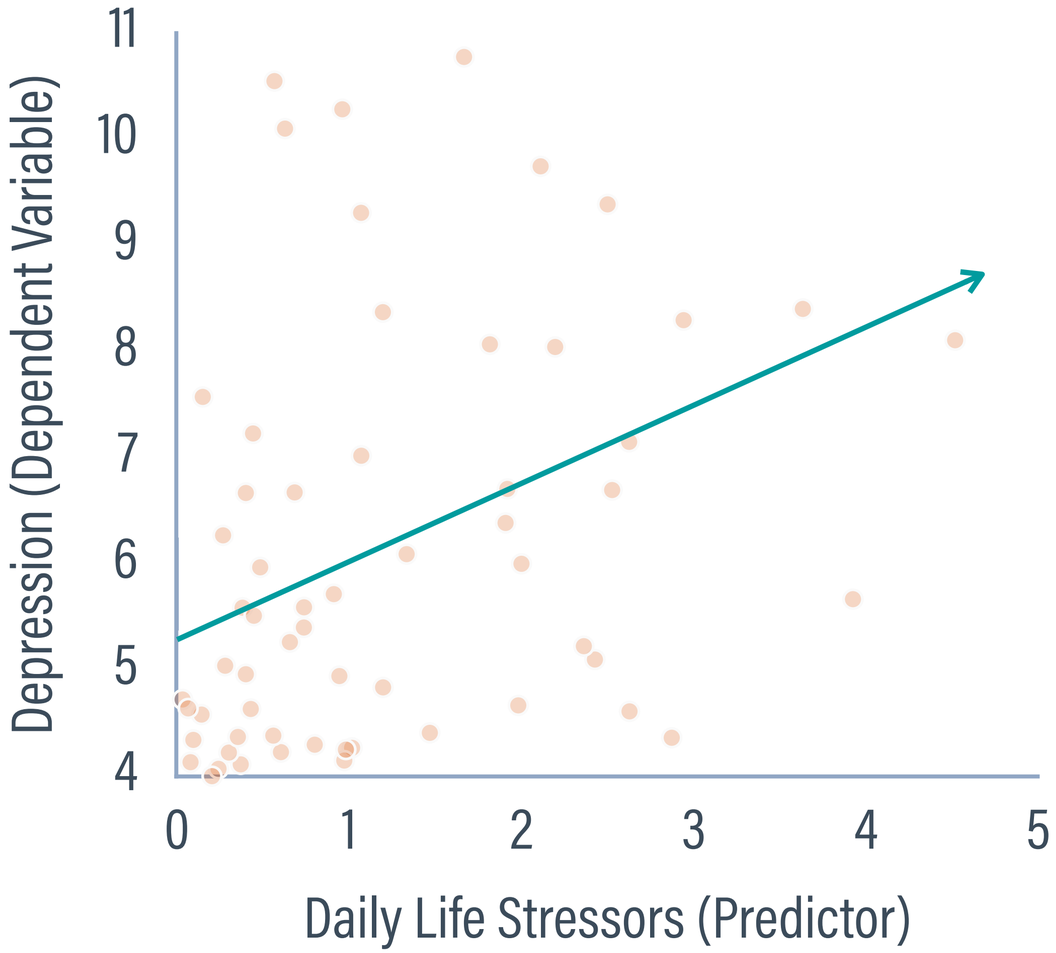
- The accuracy of t-tests (and other statistics) depends on certain conditions in the data being true (e.g., normality)
- Violations of assumptions can bias estimates, inflate or deflate standard errors, and distort significance tests
- Always check reasonableness of assumptions before drawing conclusions

REGRESSION ASSUMPTIONS

- Numeric (approximately continuous) dependent variable
- Residuals are approximately normal at every value on the line
- Independence of observations (no participant's score influences any other participant's score)
- Variation is constant across every value on the regression line (homoscedasticity, same as homogeneity of variance)

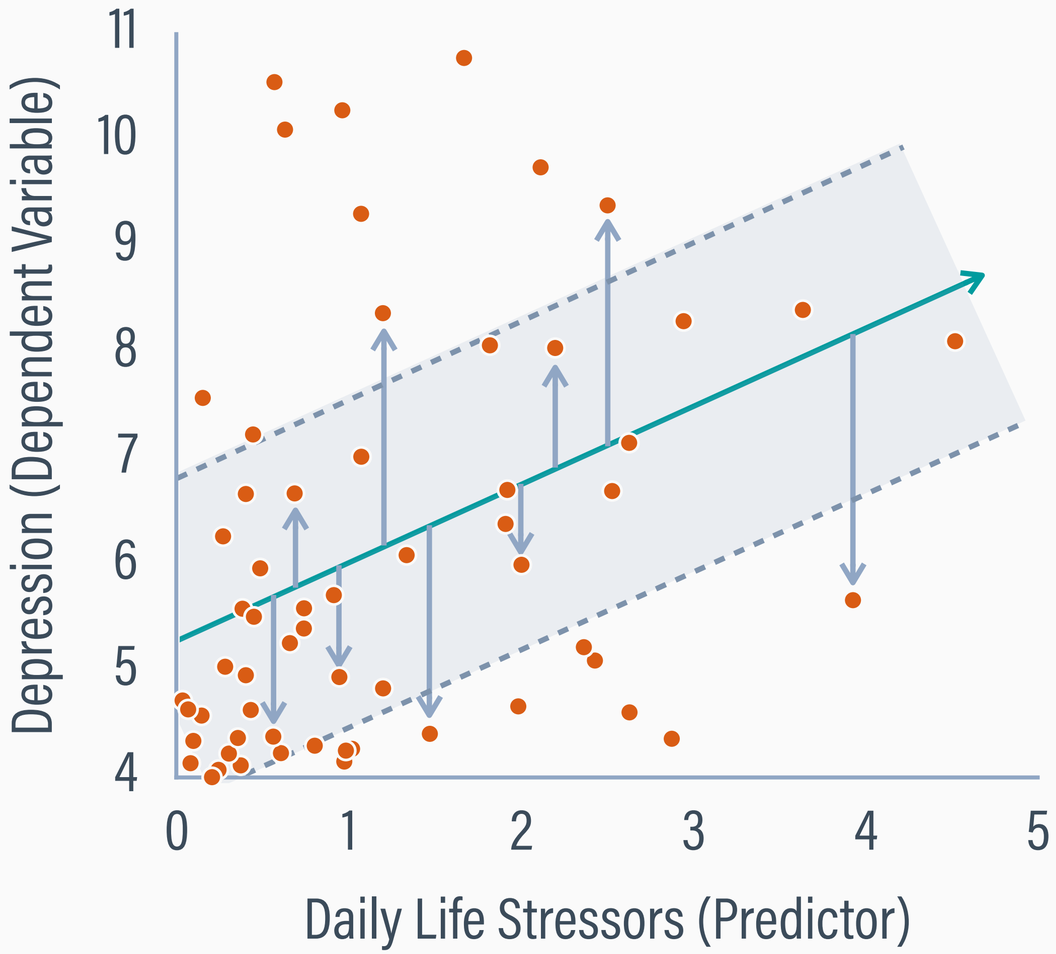
NORMALITY

- Linear regressions assumes that the residuals are normally distributed around the regression line
- In small samples, normality violations can artificially inflate or deflate standard errors, thus distorting significance tests
- Normality is less of a concern if the sample size is large enough (e.g., Ns > 40 to 50)

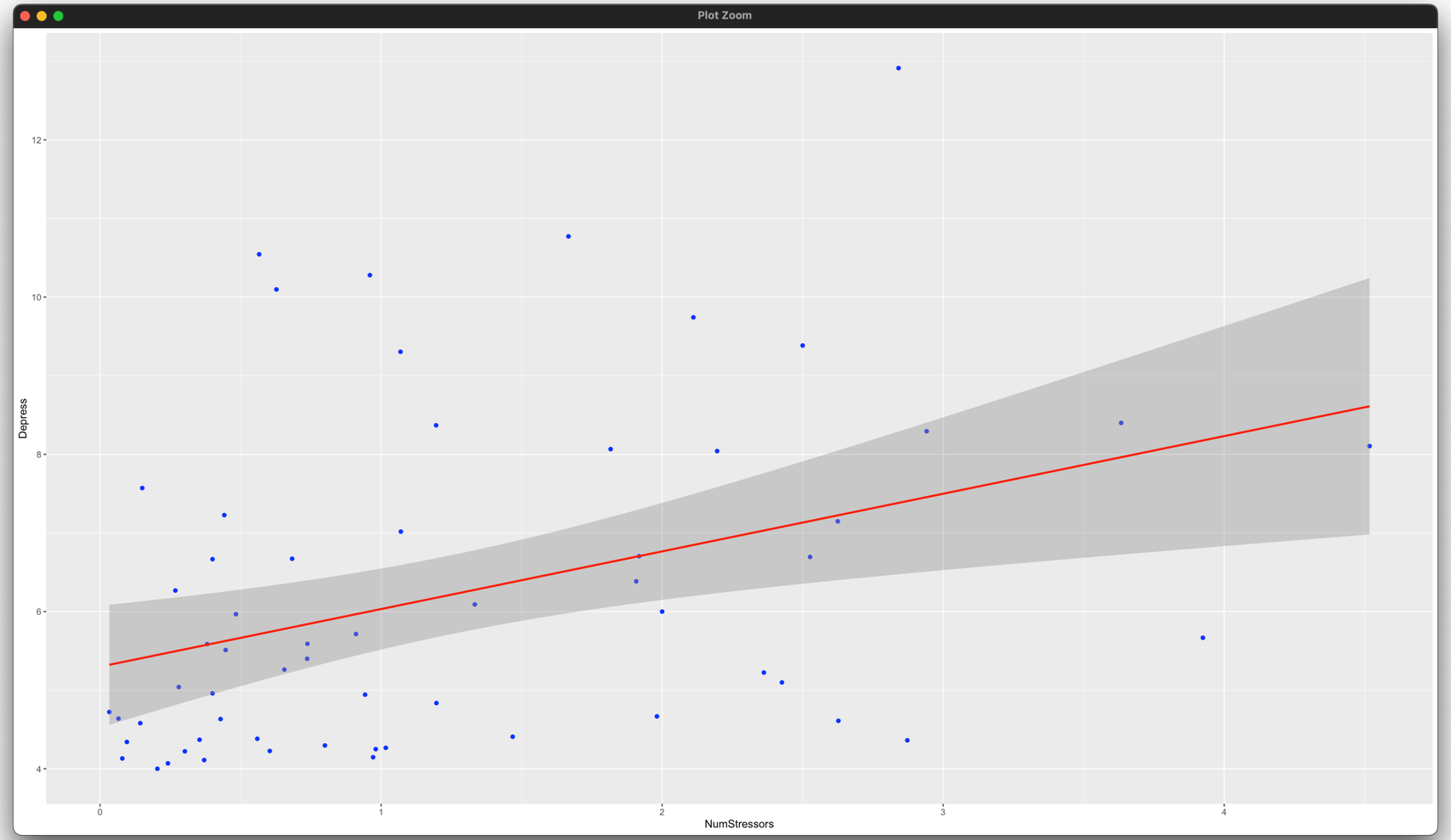


HOMOSCEDASTICITY

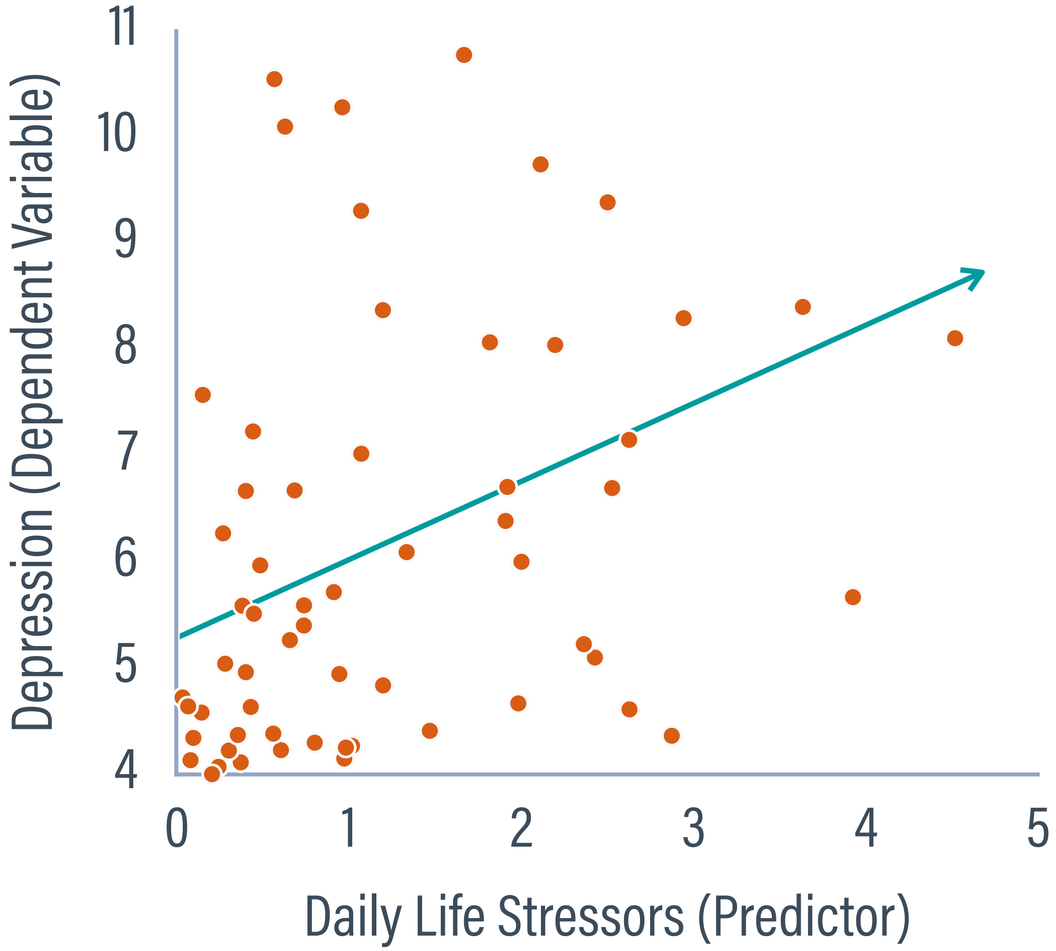
- Regression assumes that the spread of the scores around the regression line is the same across the entire line (called homoscedasticity)
- That is, residual distances don't change as the predictor variable (stressors) changes
- Non-constant variation inflates or deflates standard errors, thus distorting significance tests (large Ns do not mitigate the problem)
- Robust standard errors can address the issue



R OUTPUT



In small groups of two or three, examine the scatterplot and discuss whether residuals have constant variation. What features of the plot support your conclusion?



OUTLINE

1. Regression overview
2. Review of linear equations
3. Significance testing steps
4. Statistical assumptions
5. Study questions

STUDY QUESTIONS

For these study questions, consider the study we have discussed in class where participants with visual impairments are getting screened for ocular cancer. The researchers use linear regression to determine whether there is an association or trend between one's level of visual impairment and their optimism.

STUDY QUESTIONS (1)

1) What are the null and alternate hypotheses for the regression?

2) The intercept and slope coefficients from the data are $B_0 = 10$ and $B_1 = -0.25$. Provide an interpretation of the intercept and the slope.

3) The sampling distribution of B_1 under the null hypothesis plays a vital role in hypothesis testing with regression. Explain how the 5% significance criterion is applied to this distribution, and how it is used to decide whether to reject the null hypothesis.

STUDY QUESTIONS (2)

4) Describe the concept of a predicted value and residual in regression. In the context of this study, how would you go about computing them?

5) The researchers report the effect size as $R^2 = .03$. Provide an interpretation of the R^2 and describe its magnitude.

6) The probability value for the slope is .04 (4%). Provide an interpretation of the p-value. Note that I am not asking whether it is significant.

STUDY QUESTIONS (3)

7) Explain statistical assumptions. What are they and why do they matter? Which assumption(s) are most problematic for the accuracy of significance tests?

8) Regarding assumptions, consider the following statement: Homoscedasticity (constant residual variance) in regression is the same as homogeneity of variance in ANOVA, and normality of residuals in regression is the same as normality within each group in ANOVA. Explain whether this statement is correct or incorrect and why.