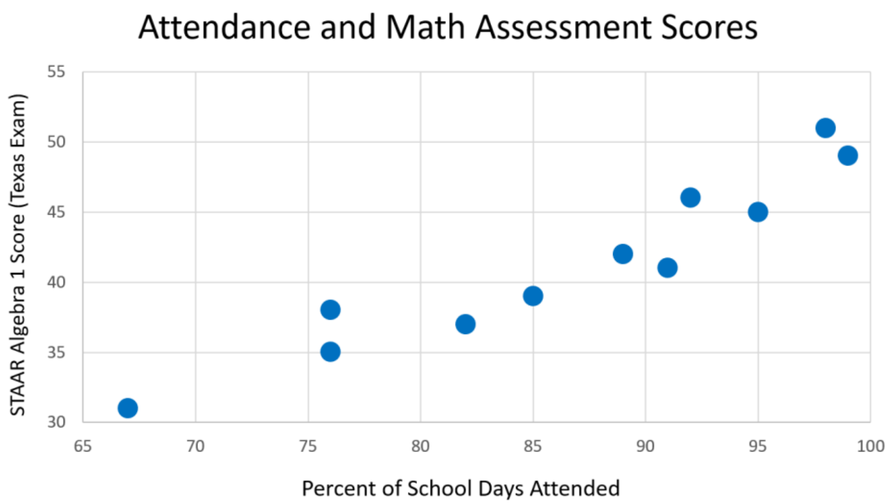
AP Statistics Handout Key: Lesson 3.2

Topics: least squares regression line, slope, y-intercept, predictions using LSRL, extrapolation

Lesson 3.2 Guided Notes

Closing the school achievement gap with attendance: Low-income students tend to have lower attendance rates and lower math test scores than their middle/upper income peers. Would raising their attendance close the achievement gap? To explore this possibility, a random sample was collected of students in Texas. For each student, data was collected on their attendance rate (percent of school days attended) and raw test scores on the Algebra 1 state exam.

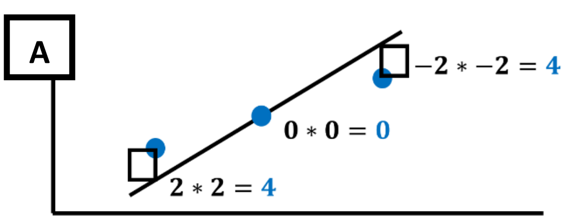
Percent Attendance (x)	Algebra 1 Raw Score (y)
95	45
89	42
67	31
98	51
99	49
76	38
92	46
91	41
76	35
85	39
82	37

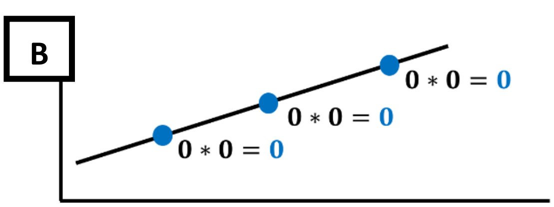


Least Squares Regression Line (LSRL)

a) For the above, what is the explanatory variable? What is the response variable? How can you tell?

Explanatory (x): Percent attendance. **Response (y):** Test score. We're investigating how test scores *respond* to variation in attendance.



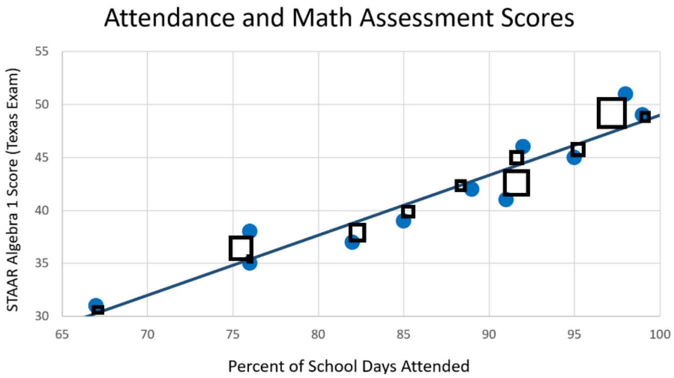


b) How can we use the information provided by the squared residuals to determine which model (A or B) better fits the data?

Model A: 4+0+4=8

Model B: 0+0+0=0

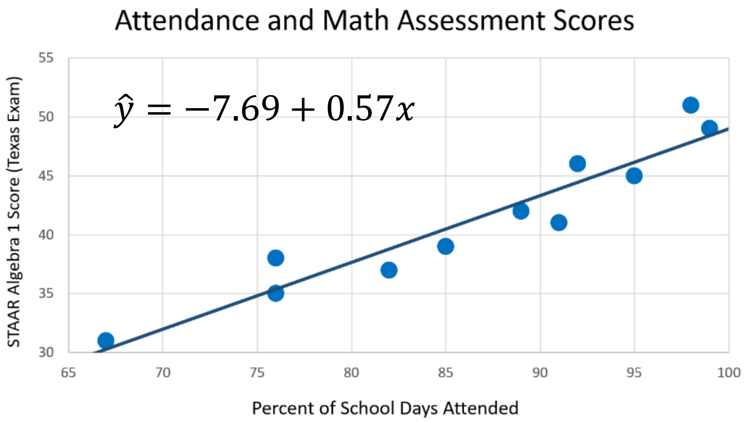
Model B is a better fit because it produces the lower sum of the residuals.

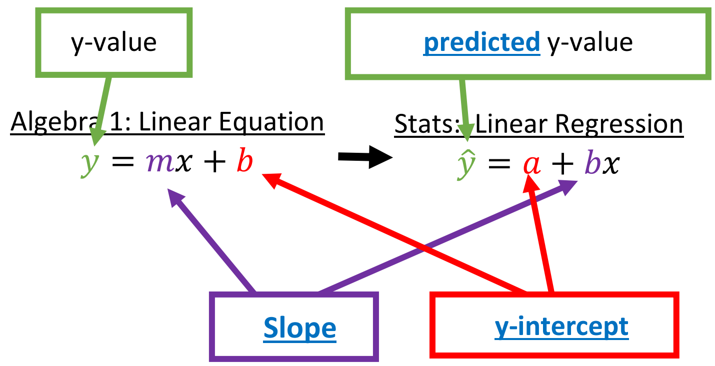


Least Squares Regression Line (LSRL): a linear model that minimizes the sum of the squared residuals between the data and the model.

- Also called "line of best fit"

Slope and y-intercept





$$\hat{y} = -7.69 + 0.57x$$

$\hat{y}$: Predicted test score
x: Percent attendance

1) Interpret the slope value:

Stem: For every 1 unit increase in explanatory variable, our model predicts an average **increase/decrease** of slope in response variable.

Your answer: For every 1 percentage point increase in attendance, our model predicts an average **increase** of 0.57 points in students' test scores.

2) Interpret the y-intercept:

Stem: When the explanatory variable is zero units, our model predicts that the response variable would be y-intercept.

Your answer: When attendance is zero percent, our model predicts that the students' test scores would be -7.689.

3) Is the y-intercept meaningful in this context? Explain…

This y-intercept value is not *statistically* meaningful since anyone with 0% attendance doesn't really go to the school or take the exam. It's also not possible for a student to receive a negative test score.

Predictions using the LSRL

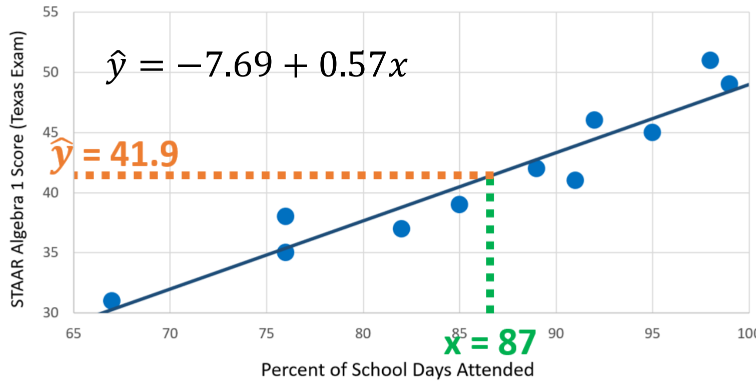
4) The superintendent of the district asks: "If a student meets our minimum attendance goal (87%), what would their predicted test score be?" Answer his question and show your work (including drawing your process on the scatterplot):

$\hat{y}$ = -7.69 + 0.57x

$\hat{y}$ = -7.69 + 0.57(**87**)

$\hat{y}$ = **41.9 points**

The predicted test score is about 41.9 points

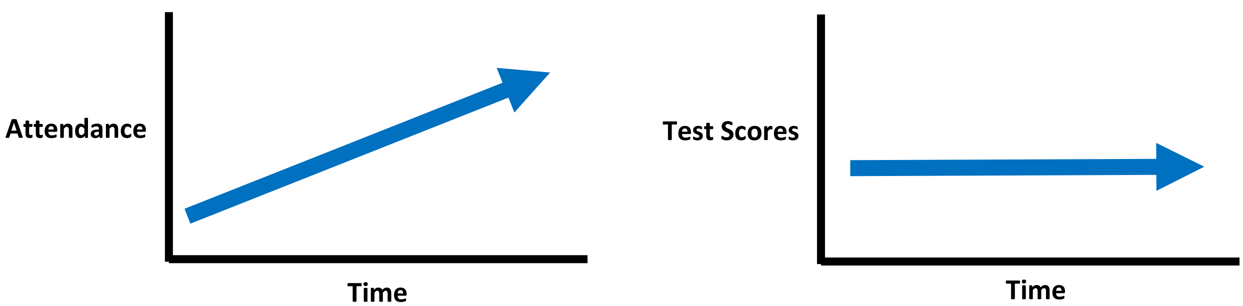


Recommended discussion norms:
skewthescript.org/discussion-norms

Lesson 3.2 Discussion

In the past several years, schools have piloted large-scale (and expensive) initiatives to improve student attendance.[1] These included call programs for chronically absent students, hiring attendance case managers and coordinators, and using Uber/Lyft for students with transportation issues.

The result:



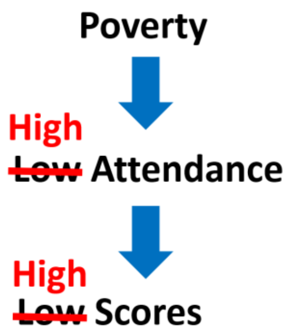
Discussion Question: Why didn't test scores grow when attendance rose?

Correlation ≠ causation

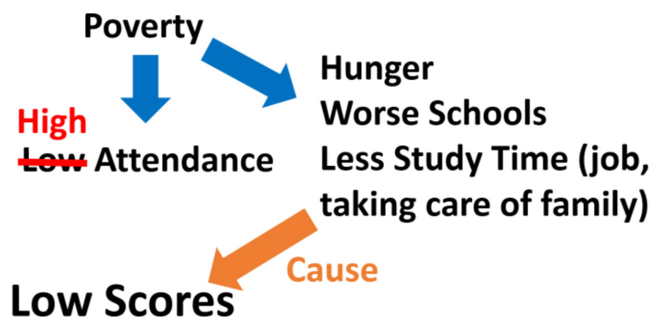
If you believe that raising attendance would raise test scores, you're making an assumption that missing class is the only reason students may be performing poorly on their tests. However, there may be other factors. For example, poverty may simultaneously cause low attendance (transportation issues, working, etc.) and cause students to suffer from hunger, to attend worse schools, or to have less study time because they're taking on a job or taking care of siblings. These latter variables may be causing the lower test scores, so changing attendance itself wouldn't actually move the needle on test scores. **See diagram on next page**.

[1] See attendance research: Pyne, Grodsky, et al., (2018). *What Happens When Children Miss School? Unpacking Elementary School Absences in MMSD*. Madison, WI: Madison Education Partnership.

Assumption:

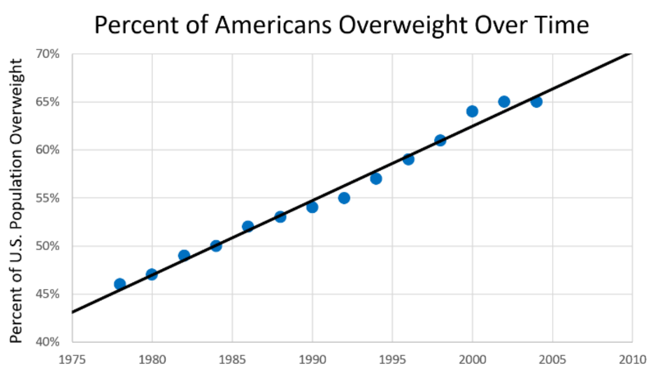


Reality:

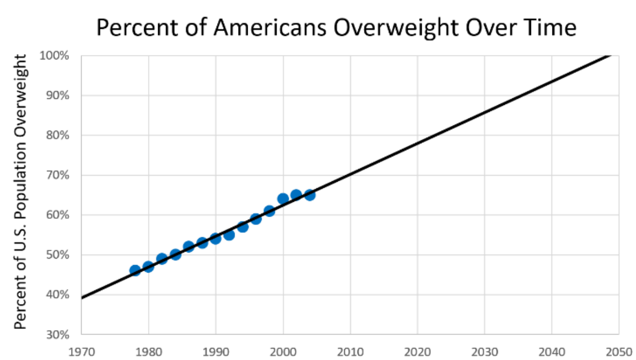


Lesson 3.2 Practice

1) Dr. Youfa Wang at University of North Carolina published a study[2] on obesity in America. Using linear regression, the study concluded that by 2048, if trends continue, **100% of Americans would be overweight**. Using the graphs[3] below, do you believe this conclusion is correct? Why or why not?



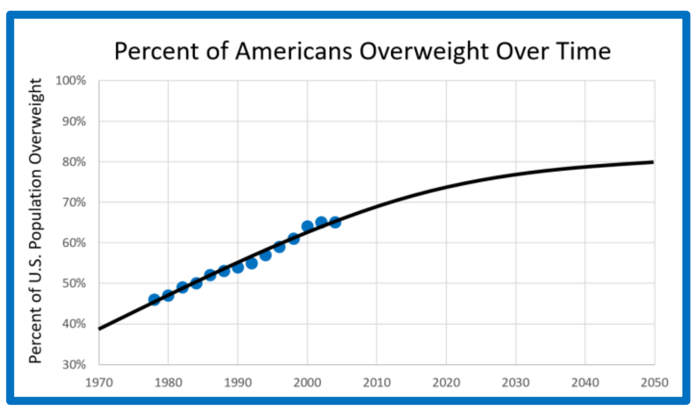
Data and Model



Model Projection

This conclusion is probably not correct because the model is making predictions outside the x-range of data. This is called **extrapolation.** It's not clear if the current trend will continue.

The following logarithmic model is more reasonable – it's improbable that *every single person* in the population would become overweight.



---

[2] Wang, Beydoun, et al., "Will all Americans become overweight or obese? estimating the progression and cost of the US obesity epidemic." *Obesity* (Silver Spring). 2008;16(10):2323-2330. doi:10.1038/oby.2008.351

[3] Graphs provided are representative approximations of analyses from the paper

2) Teachers: We recommend providing additional practice exercises from your AP Stats textbook or from prior AP Stats exams. The following textbook sections and AP exam questions are aligned to the content covered in this lesson.

- *The Practice of Statistics (AP Edition)*, 4th-6th editions: section 3.2
- *Stats: Modeling the World (AP Edition)*, 4th & 5th editions: chapter 7, 3rd edition: chapter 8
- *Statistics: Learning from Data (AP Edition)*, 2nd edition: sections 4.2-4.3
- *Advanced High School Statistics*, section 8.2
- AP Exam Free Response Questions (FRQs): 2017 Q1 (parts a&b), 2015 Q5

Handout Key by statistics student Greyson Zuniga