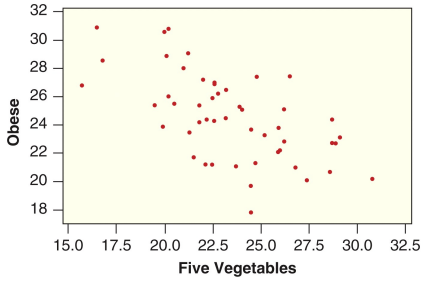
1. The graph below plots the percent of obese people in each of the 50 US states against the percent of people in the state that consume at least 5 servings of fruit or vegetables per day.



(a) (3 pts) Is the relationship described by the scatter plot for these two variables positive or negative? *Explain.*

The relation is negative: as the percentage of people who consume 5 servings of fruit/veggies ***increases****, the percentage of obese people* ***decreases****.*

(b) (3 pts) Which of $-1, -0.6, -0.02, 0.5, 0.98$ is most likely to be the correlation for the data above? *Explain briefly.*

Process of elimination: (i) The relation is negative so the correlation is negative, eliminating 0.5 *and* 0.98 *from contention; (ii) The relation is distinctly negative, eliminating* $-0.02 \approx 0$ *from contention; (iii) The relation is not purely linear, eliminating* -1*... The correlation is most likely to be* $r = -0.6$*.*

2. In 2005, the *Educational Testing Service* observed the average Math SAT score for each of the 50 states and the percentage of high school seniors in the state who took the test. The correlation between these two variables was -0.84.

(a) (3 pts) ***True*** or ***False***: test scores tend to be lower in states where a higher percentage of students take the test. *Explain your answer briefly.*

True*. The negative correlation means that there is a* ***negative relation*** *between percentage of students taking the test and the average score on the test — as the percentage goes up, the average score goes down.*

(b) (3 pts) In Connecticut the average Math SAT score in 2005 was 517, and in Iowa it was 608. Do the data show that schools in Iowa do a better job teaching math than schools in Connecticut, or is there another possible explanation? *Justify your answer.*

While it might be true, the given data do ***not*** *support the claim that Iowa schools do a better job of teaching math than Connecticut schools. The data* ***do*** *offer another* ***possible*** *explanation: perhaps the percentage of students taking the test in Connecticut is higher than the percentage of students taking the test in Iowa. Weaker students in Connecticut who take the test bring the average down, while in Iowa only the strongest students take the test.*

3. In a study of the stability of IQ scores, a large group of individuals is tested at age 18 and again at age 45. The following results are obtained.

age 18: average score ≈ 100, SD ≈ 15
age 45: average score ≈ 100, SD ≈ 18, $r \approx 0.8$

(a) (3 pts) Estimate the average score at age 45 of all the people who scored 120 at age 18.

The regression equation for predicting the (average) score at 45 *from score at* 18 *is*

$$\widehat{y}_j = \beta_0 + \beta_1 x_j$$

where $\widehat{y}_j$ *is the estimated average score at age* 45 *of all people who scored* x_j *at age* 18*, and*

$$\beta_1 = r \times \frac{SD_y}{SD_x} = 0.8 \times \frac{18}{15} = 0.96 \quad \textit{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x} = 100 - 0.96 \times 100 = 4.$$

This means that the predicted average score of all those that scored 120 *at age* 18 *is*

$$\widehat{y}(120) = 4 + 0.96 \times 120 = 119.2.$$

(b) (3 pts) Jane scored 115 on the test at age 18. What is her predicted score on the test at age 45? Please include an accurate margin of error — i.e., your answer should be in the form $A \pm B$. You may assume that the data is *homoscedastic*.

Assuming that the data is homoscedastic, we can use the standard error of regression (the root-mean-square error of regression) as the margin of error:

$$SER = \sqrt{1 - r^2} \times SD_y = \sqrt{1 - 0.64} \times 18 = 10.8.$$

Jane's predicted score at 45 *is then estimated to be*

$$\widehat{y}(115) \pm SER = (4 + 0.96 \times 115) \pm 10.8 = 114.4 \pm 10.8.$$

(c) (2 pts) The researchers observed that individuals who scored high at age 18 scored somewhat lower at age 45 on average, while individuals who scored low at age 18 improved their score at age 45 on average. To explain this phenomenon, they suggested that people who did well on the test at 18 were less concerned with their performance the second time around, while people who did not score high at 18, took the test at 45 more seriously.

Do you agree? Explain briefly.

No. *The observed phenomenon is due to the* ***regression effect****, not the attitudes of the test-takers at age* 45*.*