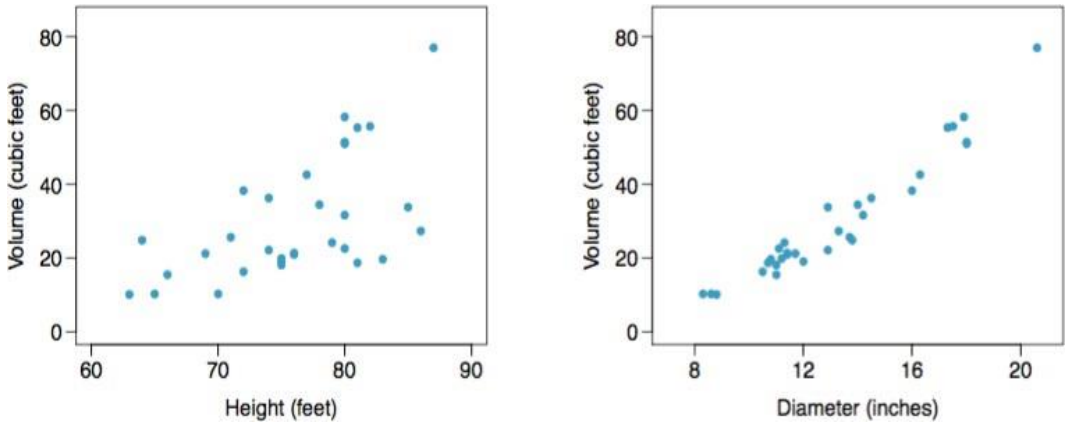
# Note # 11: Regression and Correlation

**Problem 1.** The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured at 4.5 feet above the ground.




**a.** Describe the relationship between the volume and height of these trees.

**b.** Describe the relationship between the volume and diameter of these trees.

**c.** Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

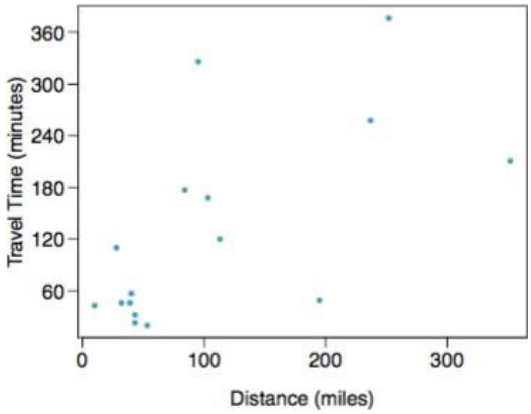
**Answer:**

**a.** Moderate, positive, linear relationship.
*Non-constant variance (as x increases, variability increases).

**b.** Strong, positive, linear / slightly curved relationship.
*Constant variance.

**c.** Diameter – stronger relationship.

**Problem 2.** The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays, the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).




**a.** Describe the relationship between distance and travel time.

**b.** How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers?

**c.** The correlation between travel time (in miles) and distance (in minutes) is $r = 0.636$ . What is the correlation between travel time (in kilometers) and distance (in hours)?

**Answer:**

**a.** Weak, positive, potentially linear.

**b.** It wouldn't change. Our units don't affect from, direction, or strength of a relationship.

**c.** $r = 0.636$ (units don't affect correlation).

**Problem 3.** What would be the correlation between the annual salaries of males and females at a company, if for a certain type of position men always made:

**a.** $5,000 more than women?

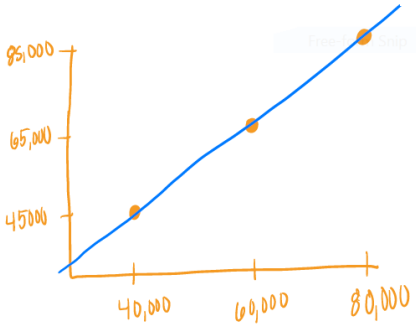
**b.** twice as much as women?

**c.** 25% less than women?

**Answer:**

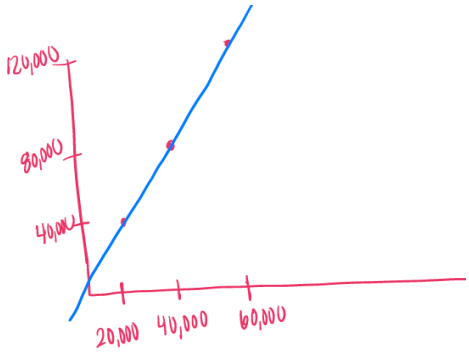
**a.** $Pay_M = Pay_W + 5000$

$r = 1$



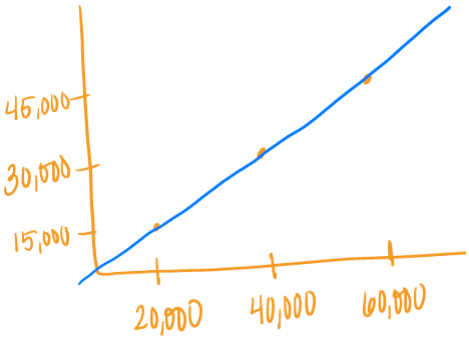

**b.** $Pay_M = 2(Pay_W)$

$r = 1$




**c.** $Pay_M = 0.75(Pay_W)$

$r = 1$



**Problem 4.** Determine if the following statements are true or false. Explain.

**a.** A correlation coefficient of -0.90 indicates a stronger linear relationship than a correlation of 0.5.

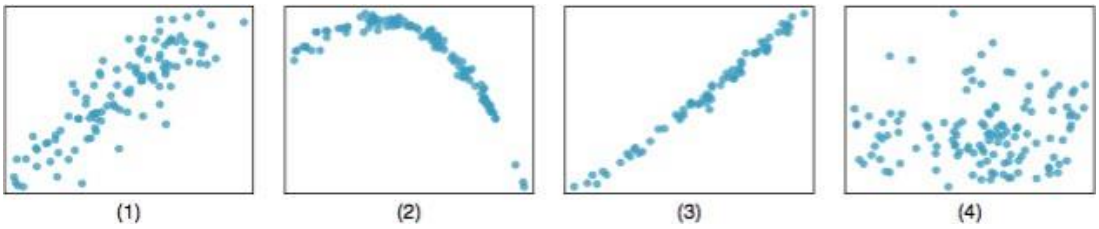
**b.** Correlation is a measure of the association between any two variables.

**Answer:**

**a.** True – strength is determined by the magnitude of $r = -0.90$ is further from 0 than 0.50.

**b.** False – correlation is a measure of the linear association between any two numerical variables.

**Problem 5.** Match each correlation to the corresponding scatterplot.




**a.** $r = -0.72$

**b.** $r = 0.07$

**c.** $r = 0.86$

**d.** $r = 0.99$

**Answer:**

**a.** $r = -0.72$. Scatterplot 2.

**b.** $r = 0.07$. Scatterplot 4.

**c.** $r = 0.86$. Scatterplot 1.

**d.** $r = 0.99$. Scatterplot 3.

**Problem 6.** In college freshman men, it appears as though there is a linear relationship between height (in inches) and weight (in pounds). In a sample of the population, we see that the average height is 68.4 inches, with a standard deviation of 4.0 inches. We see that the average weight is 141.6 pounds, with a standard deviation of 9.6 pounds. The correlation between height and weight in our sample is 0.73. We would like to create a linear model that can be used to predict a male college freshman's weight, given we know their height.

**a.** What is the formula for the linear model?
**b.** Interpret the slope.
**c.** Interpret the intercept.
**d.** What is r-squared?
**e.** Interpret r-squared.
**f.** James is a male college freshman who is 68 inches tall. What is his predicted weight?
**g.** James actually weighs 152 pounds. What is his residual?

**Answer:**

**a.** $b_1 = \left(\frac{s_y}{s_x}\right) \times r = \left(\frac{9.6}{4.0}\right) \times 0.73 = 1.752$ .

$b_0$: $\bar{y} = b_0 + b_1\bar{x} \rightarrow 141.6 = b_0 + (1.752) \times (68.4) \rightarrow 141.6 = b_0 + 119.8368$

$b_0 = 141.6 - 119.8368 \quad \rightarrow \quad b_0 = \mathbf{21.7632}$

$\therefore \hat{y} = \mathbf{21.7632 + 1.752\,x}$

**b.** Slope ($b_1$): For each additional inch in height, our predicted weight increases by 1.752 pounds.

**c.** Intercept ($b_0$): For someone who is zero inches tall, we predict their weight to be 21.7632 pounds (pretty meaningless).

**d.** $r^2 = (0.73)^2 = \mathbf{0.5329}$

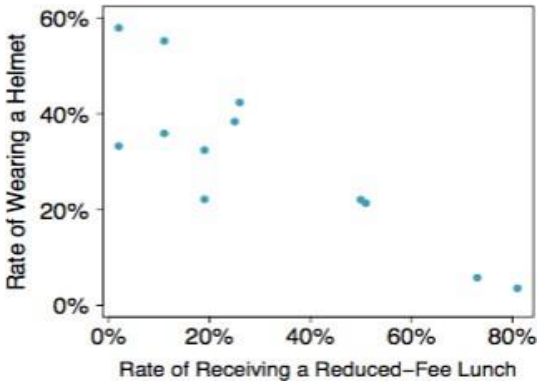
**e.** 53.29% of the variation in weight can be explained by our model.

**f.** $\hat{y} = 21.7632 + 1.752\,x \rightarrow \hat{y} = 21.7632 + 1.752\,(68) = 21.7632 + 119.136$

$\therefore \hat{y} = \mathbf{140.8992}\ \boldsymbol{pounds}$

**g.** $Residual = y_i - \hat{y}_i = 152 - 140.8992 = 11.1008\ pounds.$

**Problem 7.** The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%.




**a.** If the r-squared for the least-squares regression line for the data is 72%, what is the correlation between the two variables?
**b.** What is the least-squares regression line?
**c.** Interpret the intercept of the least-squares regression line.
**d.** Interpret the slope of the least-squares regression line.
**e.** What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual.

**Answer:**

**a.** $r^2 = 0.72$

$r = \sqrt{0.72} = \pm 0.849$. Based on the scatter plot, $\boldsymbol{r = -0.849}$

**b.** $b_1 = \left(\frac{s_y}{s_x}\right) \times r = \left(\frac{16.9}{26.7}\right) \times (-0.849) = -0.537$ .

$b_0$: $\bar{y} - b_1\bar{x} \quad \rightarrow b_0 = 38.8 - (-0.537) \times (30.8) \rightarrow b_0 = 55.34$

$\therefore \boldsymbol{\hat{y} = 55.34 - 0.537\, x}$

**c.** For a neighborhood where 0% reduced-fee lunch, we predict 55.34% of bike riders to wear a helmet.

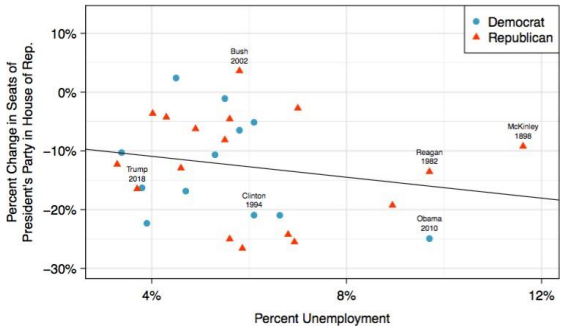
**d.** For each additional percentage point of children who receive reduced-fee lunch, we predict a decrease of 0.537% in the percent of bike riders who wear a helmet.

**e.** $\hat{y} = 55.34 - (0.537)x \rightarrow \hat{y} = 55.34 - (0.537)(40)$

$\therefore \boldsymbol{\hat{y} = 33.86}$

$Residual = y_i - \hat{y}_i = 40 - 33.86 = 6.14$

**Problem 8.** Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America's two-party system, one political theory suggests the higher the unemployment rate, the worse the President's party will do in the midterm elections. To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2018, with the exception of those elections during the Great Depression. We consider the percent change in the number of seats of the President's party (e.g. percent change in the number of seats for Republicans in 2018) against the unemployment rate. Below, you are given both a scatterplot of the data as well as the regression output. You want to test and see if there is a linear relationship at the 0.10 level.




| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -7.3644 | 5.1553 | -1.43 | 0.1646 |
| unemp | -0.8897 | 0.8350 | -1.07 | 0.2961 |
| | | | | $df = 27$ |

**a.** What kind of relationship do you notice?
**b.** The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?
**c.** What is the least-squares regression line?
**d.** What are the hypotheses?
**e.** What is the significance level?
**f.** What is the value of the test statistic?
**g.** What is the p-value?
**h.** What is the correct decision?
**i.** What is the appropriate conclusion/interpretation?
**j.** What is the 95% confidence interval for the slope parameter?

**Answer:**

**a.** Weak- moderate relationship – negative relationship.

**b.** These are high leverage points, potentially influential points.

**c.** $\hat{y} = -7.3644 - (0.8897)(x)$

**d.** $H_0: \beta_1 = 0 \quad vs. \quad H_A: \beta_1 \neq 0$

**e.** $\alpha = 0.10$

**f.** $TS = \frac{point\ estimate - null\ value}{std.error} = \frac{-0.8897 - 0}{0.8350} = \mathbf{-1.066}$

**g.** $TS = -1.066,\ df = 27 - Two\ sided$

**h.** $0.20 < p - value < 0.30, p - value > 0.20 > \alpha = 0.10,$ $\boldsymbol{Fail\ to\ reject\ H_0}$

**i.** The data does not provide statistically significant evidence that there is a linear relationship between unemployment rate and election results.