Homework 2

1) For the following 4 plots, state if the association is:
   (i) Direction: Positive, negative, or none
   (ii) Strength: Strongly correlated, or weakly correlated
   (iii) Shape: Linearly correlated, curved, other, or none

(1) Direction: None, Strength: Weak, Shape: None
(2) Direction: Positive, Strength: Strong, Shape: Linear
(3) Direction: Positive, Strength: Stronger Weak, Shape: Curved
(4) Direction: None, Strength: Strong, Shape: Linear

2) For the same plots in question 1, we calculated the following correlation coefficients: .9012, .5823, -.8345, -.0036. Match the values to the appropriate plot.

(Plot 1) -.0036
(Plot 2) .9012
(Plot 3) .5823
(Plot 4) -.8345
3) The National Insurance Crime Bureau reports that Honda Accords, Civics, and Toyota Camrys are stolen more often than Ford Tauruses, Pontiac Vibes, and Buick LeSabres. Is it reasonable to say that there's a correlation between the type of car you own and the risk that it will be stolen? Explain.

There may be an association, but not a correlation as the type of car is a categorical variable, not a quantitative variable.

4) A researcher studies elementary school children and finds there's a strong positive linear association between the height and reading scores.
   a) Does this mean that taller children are generally better readers? Explain.

Yes, as there is a strong positive association between the two variables, but this doesn't imply causation.

b) What might explain this strong correlation?

Age & grade level are lurking variables. Taller children are generally older, who are in higher grade levels, and are then better readers.

c) Suppose \( r = .9 \), what percent of the variability is explained by the researcher's model?

\[
R^2 = r^2 = .81
\]

81% of the variability is explained by the model.
5) A nutritionist analyzed several popular fast food burgers to see if there was a relationship between fat content and calorie content. The data is summarized below:

<table>
<thead>
<tr>
<th>X: Fat (g)</th>
<th>19</th>
<th>31</th>
<th>34</th>
<th>35</th>
<th>39</th>
<th>39</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y: Calories</td>
<td>410</td>
<td>580</td>
<td>590</td>
<td>570</td>
<td>640</td>
<td>680</td>
<td>660</td>
</tr>
</tbody>
</table>

a) Find the Standard deviation for X ($s_x$), and for Y ($s_y$).

$$s_x^2 = \frac{1}{n-1} \cdot \frac{\sum(X_i - \bar{X})^2}{n} = 60.905 \Rightarrow s_x = 7.80415$$

$$s_y^2 = \frac{1}{n-1} \cdot \frac{\sum(Y_i - \bar{Y})^2}{n} = 8066.6 \Rightarrow s_y = 89.81462$$

b) Find the covariance between X and Y ($s_{xy}$)

$$s_{xy} = \frac{1}{n-1} \cdot \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n} = 673.3$$

c) For the Linear Regression model: Calories = $a + b(fat)$, find $a$ and $b$, then state the model.

$$b = \frac{s_{xy}}{s_x^2} = \frac{673.3}{60.905} = 11.05555$$

$$a = \bar{y} - b \cdot \bar{x} = 590 - (39.29 \cdot 11.06) = 210.9539$$

$$\hat{y} = 210.95 + 11.06x$$

d) Find the correlation coefficient ($r$).

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{673.3}{7.8 \cdot 89.8} = .96063$$

e) Suppose there was a very bad error in the equipment, and it underestimated the calories. The true number of calories is twice as much than reported in the table. Will this change our correlation coefficient or our estimated slope and intercept? If so, find the new values, if not, explain or show why not.

$$b^{\prime} = \frac{s_{x'y'}}{(s_{x'}^2)} = 2 \cdot \frac{s_{xy}}{s_x^2} = 2 \cdot \frac{s_{xy}}{s_x^2} = 2 \cdot 6 = 22.0112$$

$$a^{\prime} = \bar{y} - b^{\prime} \cdot \bar{x} = 421.9077$$

$$r_{x'y'} = \frac{s_{x'y'}}{s_{x'} \cdot s_{y'}} = \frac{2 \cdot s_{xy}}{s_{x'} \cdot 2 \cdot s_{y'}} = \frac{s_{xy}}{s_{x'}^2} = r_{xy} \quad \text{No change}$$
6) For the following three residual plots, tell whether a linear model was appropriate for the data, and explain why if it is or isn't.

a) No, the residuals are not constant, or the variability is not consistent along x.

b) No, the residuals decrease as x increases, or the variability decreases as x increases.

c) Yes, there is no pattern in the residuals.
7) Fill in the missing pieces of information in the table. Each row is a separate case; the rows are not related to each other. Show work for each section below.

<table>
<thead>
<tr>
<th></th>
<th>$x$</th>
<th>$s_x$</th>
<th>$y$</th>
<th>$s_y$</th>
<th>$r$</th>
<th>$S_{xy}$</th>
<th>$y = a + bx$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>30</td>
<td>4</td>
<td>18</td>
<td>6</td>
<td>-.2</td>
<td>-4.8</td>
<td>$y = 2x - .3x$</td>
<td></td>
</tr>
<tr>
<td>b)</td>
<td>100</td>
<td>18</td>
<td>60</td>
<td>10</td>
<td>.9</td>
<td>162</td>
<td>$y = 10 + 4.5x$</td>
<td></td>
</tr>
<tr>
<td>c)</td>
<td>4</td>
<td>.8</td>
<td>50</td>
<td>15</td>
<td>-.8</td>
<td>9.6</td>
<td>$y = -10 + 15x$</td>
<td></td>
</tr>
<tr>
<td>d)</td>
<td>13</td>
<td>1.2</td>
<td>4</td>
<td>4</td>
<td>-.6</td>
<td>-2.88</td>
<td>$y = 30 - 2x$</td>
<td></td>
</tr>
</tbody>
</table>

a) $S_{xy} = r \cdot s_x \cdot s_y = -.2 \cdot 6 \cdot 4 = -4.8$

$b = \frac{S_{xy}}{S_x^2} = \frac{-4.8}{6^2} = -.3$

$a = \bar{y} - b \bar{x} = 18 - (-.3 \cdot 10) = 20$

b) $S_{xy} = .9 \cdot 10 \cdot 18 = 162$

$b = .9 \cdot \frac{10}{18} = .5$

$a = 60 - (.5 \cdot 100) = 10$

c) $\bar{y} = a + b \bar{x} \Rightarrow \frac{\bar{y} - a}{b} = \bar{x} = \frac{50 - (-10)}{15} = 4$

$b = r \cdot \frac{s_y}{s_x} \Rightarrow r = \frac{b \cdot s_x}{s_y} = \frac{.5 \cdot 6}{50} = .08$

$S_{xy} = .8 \cdot 15 - .8 = 9.6$

d) $\bar{x} = \frac{\bar{y} - a}{b} = \frac{41 - 30}{(-2)} = 13$

$s_x = r \cdot \frac{s_y}{b} = \frac{-6 \cdot 4}{(-2)} = 1.2$

$S_{xy} = -6 \cdot 1.2 \cdot 4 = -2.88$
8) Suppose at a certain college, administrators were interested in the incoming freshmen class. Suppose the average SAT scores of the freshmen were 1833, with standard deviation 123. After the first semester, the average GPA of the incoming freshmen was 2.66, with standard deviation .56. A scatterplot of the students SAT score and GPA shows it's reasonably linear, and r = .47.

a) Compute, and write down the regression line if we assume SAT is x.

\[ b = \frac{.47 \cdot .56}{123} = .00214 \quad \hat{y} = 2.66 - .00214 \cdot 1833 = -1.26262 \]

\[ y = -1.263 + .00214x \quad \text{or} \quad \hat{y} = -1.263 + .00214 \cdot \text{SAT} \]

b) Explain what the y-intercept means in this instance.

It's the predicted GPA if the SAT score is 0. Because a 0 SAT score is not possible, the y-intercept is only useful to define the best fit line, not predict.

c) Interpret what the slope here means.

The change in predicted GPA for every one point increase on the SAT.

d) Suppose an incoming freshman had a SAT score of 2100. Predict his GPA after one semester.

\[ \hat{y} = -1.263 + .00214 \cdot 2100 = 3.23 \]

e) Suppose you are in this group of freshmen. Would you rather have your data point be a positive or negative residual? Explain.

Positive, as it would mean I'm above average, or my GPA is larger than expected for my SAT score.
9) For the following 4 plots, there is a single unusual point. Answer the following 4 questions for each plot:

i) Does the unusual point have a high leverage, high residual, both, or neither?
ii) Is the point influential?
iii) If the point were removed, would the correlation become stronger, weaker, or have no effect?
iv) If the point were removed, would the slope of the regression line increase, decrease, or remain the same?

(a)
- (i) Both
- (ii) Yes
- (iii) Stronger
- (iv) Increase

(b)
- (i) High Leverage
- (ii) Yes
- (iii) Weaker or no effect
- (iv) Decrease

(c)
- (i) High Residual
- (ii) No
- (iii) Stronger
- (iv) Remain the same or Decrease

(d)
- (i) High Leverage
- (ii) No
- (iii) Weaker or No effect
- (iv) Remain the same