Chapter 5: Summarizing Bivariate Data

Review Pack

Name ___________________________

1. What is it that the Pearson correlation coefficient quantifies?

2. If a scatter plot exhibits a strong positive relationship, what can be said about the value of the quantity, \[ \sum z_x z_y \]?

3. One of the properties of Pearson's \( r \) is: "The value of \( r \) does not depend on which of the two variables is labeled as \( x \)." In your own words, what does this mean?
4. The use of small aircraft with human observers is common in wildlife studies where the goal is to estimate the abundance of different species. Recently there has been interest in using unmanned aerial vehicles (UAV). The UAV, something about the size of a model airplane, would fly over the area of interest and take pictures to be analyzed by computers with imagery software when the UAV returns. The plot below is from a test run of the UAV over 10 areas in South Central Florida, using bird decoys to test the reliability of the process.

![Graph](image)

(a) The least squares best fit line is $Actual = -0.924 + 1.09(UAVCount)$. Plot this line on the graph above. Show any calculations in the space below.

(b) The least squares line is the line that minimizes the sum of the squared residuals. On the graph above pick 2 points and sketch the residuals associated with those points.
5. The data below were gathered on a random sample of 7 male black-footed albatrosses of known age. In an effort to monitor diseases of these animals, biologists would like to be able to estimate the age of animals that have died by flattening their gonads and measuring the resulting area.

**Gonad size vs. Age in Black-footed albatrosses**

<table>
<thead>
<tr>
<th>Gonad Size (sq mm)</th>
<th>Age (Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>1.42</td>
</tr>
<tr>
<td>60</td>
<td>4.75</td>
</tr>
<tr>
<td>20</td>
<td>0.67</td>
</tr>
<tr>
<td>96</td>
<td>23.64</td>
</tr>
<tr>
<td>24</td>
<td>0.52</td>
</tr>
<tr>
<td>27</td>
<td>2.35</td>
</tr>
<tr>
<td>27</td>
<td>1.4</td>
</tr>
</tbody>
</table>

a) What is the value of the correlation coefficient for these data?

\[ r = 0.928 \]

b) What is the equation of the least squares line describing the relationship between \( x = \text{Gonad Size} \) and \( y = \text{Age} \).

\[ y = -6.99 + 0.283x \]

c) If these albatrosses are representative of the population, what would you predict to be the age of a male albatross with a gonad size of 50 sq. mm? Show any calculations below.

\[ -6.99 + 0.283(50) = 7.15 \text{ years} \]

d) The largest albatross gonad size in the sample was 96 sq mm, with an age of 23.64 years. These animals are thought to live for up to 40 years. Would it be reasonable to use the equation from part (b) above to predict the age for a gonad size of 150 sq mm? Why or why not?

\[ \text{No, it would not be appropriate to use this equation to predict the age of this albatross. This would be an extreme extrapolation from the data that has been provided, which is an unreliable procedure.} \]
1. The *Des Moines Register* recently reported the ratings of high school sportsmanship as compiled by the Iowa High School Athletic Association. For each school the participants and coaches were rated by referees, where 1 = superior, and 5 = unsatisfactory. A regression analysis of the average scores given to football players and coaches is shown below.

\[
\text{FBParticipants} = 0.902 + 0.568 \text{ FBCoaches}
\]

**Summary of Fit**

- RSquare: 0.452
- RSquare Adj: 0.450
- s: 0.355

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>37.505</td>
<td>37.505</td>
<td>298.2723</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>362</td>
<td>45.518</td>
<td>0.126</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>363</td>
<td>83.022</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Interpret the value of the correlation between the ratings of coaches and participants.

\[ r = 0.467 \]

This indicates a moderately strong linear relationship between rated sportsmanship for spectators and participants.

b) Interpret the value of the coefficient of determination.

\[ r^2 = 0.467 \]

About 47% of the observed differences in sportsmanship among spectators can be explained by differences in participants’ sportsmanship.

c) Interpret the value of the standard deviation about the least squares line.

\[ s = 0.322 \]

This is a measure of how far a typical point will be above or below the least squares line.
2. As early as 3 years of age, children begin to show preferences for playing with members of their own sex, and report having more same-sex than opposite-sex friends. In a study of 3rd and 4th graders' views on 48 personality traits, children were asked to rate on a "5-point" scale:

-2 = "someone possessing that trait is probably a boy"
-1 = "someone possessing that trait might be a boy"
 0 = "can't tell"
 1 = "someone possessing that trait might be a girl"
 2 = "someone possessing that trait is probably a girl"

A plot of the data is presented below. A single point represents the (average girls' rating, average boys' rating) for a given trait.

![Plot of data](image)

**Linear Fit**

\[ \text{MRating} = -0.765 + 0.714 \times \text{FRating} \]

**Summary of Fit**

- \( \text{RSquare} = 0.552 \)
- \( \text{RSquare Adj} = 0.529 \)
- \( s = 0.490 \)

**Analysis of Variance**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>5.63</td>
<td>5.63</td>
<td>23.45</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>4.56</td>
<td>0.24</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>20</td>
<td>10.20</td>
<td></td>
<td>0.0001</td>
</tr>
</tbody>
</table>

a) Circle the single point which represents the most influential observation. What aspect of this point makes it the most influential?
b) Suppose a personality trait similar to those used in the survey were given a 0.0 rating ("can't tell") by the girls. The predicted boys' average rating would be closest to which of the 5 categories described above?

\[ \text{Mashing} = -0.765 + 0.714(0) = -0.765. \] This is closest to the category "might be a girl".


c) The traits plotted above are those the researchers believe are "positive" traits, such as "mature," "honest," and "polite." The researchers thought that girls would rate these positive traits as characteristic of girls to a greater extent than boys would. What aspects of the plot and/or regression analysis presented above are consistent with this thinking?

If this is true, then the male rating numbers for each trait would tend to be lower than the female numbers. These would be below the graph of \( y = x \), shown on the graph (above). All but one or two of the points are consistent with this thinking.
3. A common statistical method for estimating a population size assumes each member of the population has an equal probability of being captured. To assess this assumption for crocodile populations, investigators repeatedly sampled sections of rivers in Australia. Crocodile lengths were measured in size classes. Crocs 0.0 – 0.3 meters in length are in size class 1, 0.3 – 0.6 meters in length are size class 2, etc. The normal maximum adult length is in a class size of 9 or 10. The investigators fit a quadratic function relating the probability of capture and the size class of captured crocodiles. The output from their analysis is shown below.

(a) What proportion of the variability in probability of capture is explained by the crocodile's size class?

\[ a) \ r^2 = .928 \]

(b) Some biologists speculate that as crocodiles grow they become more wary of humans, and are more difficult to detect in the wild. Support or refute this belief by appealing to the analysis above.

\[ b) \ This \ belief \ appears \ to \ be \ true, \ since \ the \ largest \ crocodiles \ appear \ to \ have \ the \ smallest \ probability \ of \ capture. \ This \ effect \ seems \ to \ be \ present \ for \ crocodiles \ in \ classes \ 6 \ and \ above \]
4. Assessing the goodness of fit of a regression line involves considering different information, and no single characteristic of data is sufficient for a good assessment. Consider the characteristics below. How does each contribute to an assessment of fit? That is, for each piece of information, what about it would indicate a "good" best-fit line?

a) The shape of the scatter plot

a) Points on the graph lined up in a pattern that is consistently increasing or decreasing, rather than curved.

b) The correlation coefficient

b) Values of $r$ that are close to -1 or 1.

c) The standard deviation of the residuals

\[ \text{c) Small value of the standard deviation of the residuals (close to zero).} \]

d) The coefficient of determination

d) Value of $r^2$ close to 1.
Paleontology, the study of forms of prehistoric life, can sometimes be aided by modern biology. The study of prehistoric birds depends on fossil information, which typically consists of imprints in stone of a prehistoric creature’s remains. To study the productivity of an ancient ecosystem it would be useful know the actual mass of the individual birds, but this information is not preserved in the fossil record. It seems reasonable that the biomechanics of birds operates much the same today as in the past. For example, relationship between the wing length and total weight of a bird should be very similar today to the relationship in the distant past. The wing lengths of ancient birds are readily obtainable from the fossil record, but the weight is not. Assuming similar biomechanical development for ancient birds and modern birds, a regression model expressing the relationship between wing length and total weight of a modern bird could be used to estimate the mass of similar prehistoric birds and thus gauge some aspects of the ancient ecosystem.

Data is available for some modern birds of prey. Specifically, data on the mean wing length and mean total weight of species of hawk-like birds of prey is given below.

<table>
<thead>
<tr>
<th>Bird species</th>
<th>Wing length (cm)</th>
<th>Total weight (kilograms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gyps fulvus</td>
<td>69.8</td>
<td>7.27</td>
</tr>
<tr>
<td>Gypaetus barbatus grandis</td>
<td>71.7</td>
<td>5.39</td>
</tr>
<tr>
<td>Catharista atrata</td>
<td>50.2</td>
<td>1.70</td>
</tr>
<tr>
<td>Aguila chrysatus</td>
<td>68.2</td>
<td>3.71</td>
</tr>
<tr>
<td>Hieraeus fasciatus</td>
<td>56.0</td>
<td>2.06</td>
</tr>
<tr>
<td>Helotarsus ecaudatus</td>
<td>51.2</td>
<td>2.10</td>
</tr>
<tr>
<td>Geranoatus melanoleucus</td>
<td>51.5</td>
<td>2.12</td>
</tr>
<tr>
<td>Circatus gallicus</td>
<td>53.3</td>
<td>1.66</td>
</tr>
<tr>
<td>Buteo bueto</td>
<td>40.4</td>
<td>1.03</td>
</tr>
<tr>
<td>Pernis apivorus</td>
<td>45.1</td>
<td>0.62</td>
</tr>
<tr>
<td>Pandion haliaetus</td>
<td>49.6</td>
<td>1.11</td>
</tr>
<tr>
<td>Circus aeruginosos</td>
<td>41.3</td>
<td>0.68</td>
</tr>
<tr>
<td>Circus cyaneus (female)</td>
<td>37.4</td>
<td>0.472</td>
</tr>
<tr>
<td>Circus cyaneus (male)</td>
<td>33.9</td>
<td>0.331</td>
</tr>
<tr>
<td>Circus pygargus</td>
<td>35.9</td>
<td>0.237</td>
</tr>
<tr>
<td>Circus macrurus</td>
<td>35.7</td>
<td>0.386</td>
</tr>
<tr>
<td>Milvus milvus</td>
<td>50.7</td>
<td>0.927</td>
</tr>
</tbody>
</table>
Using these data, construct the least squares best-fit line for predicting total weight using wing length as a predictor.

1) What is the equation of the least-squares line?

\[ \hat{W} = -5.35 + 0.146L, \text{ where } \hat{W} \text{ is the wing weight and } L \text{ is the wing length.} \]

2) Graph the least-squares line on the scatter plot below.

3) Approximately what proportion of the variability in weight is explained by the wing length?

\[ r^2 = 0.805 \]

Biological theory suggests that the relationship between the weight of these animals and their wing length is exponential, i.e. \( W = \alpha 10^{\beta L} \), or \( W = \alpha e^{\beta L} \) where \( W \) is the wing weight and \( L \) is the wing length. Perform the appropriate transformation of variable(s) and fit an exponential model to the data.

4) What is the resulting best fit line using the transformed model?

\[ \log_{10}(\hat{W}) = -1.60 + 0.0339L \text{ or } \ln(\hat{W}) = -3.68 + 0.0780L. \]

5) For a wing length of the data point where \( L = 56.0 \) (\( Hieraeus fasciatus \)), what is the predicted bird weight? Show your work below.

\[
\begin{align*}
\text{base 10 logs:} & \\
\log_{10}(\hat{W}) &= -1.60 + 0.0339(56.0) \\
&= 0.301 \\
\hat{W} &= 10^{0.301} = 2.00 \text{ kg}
\end{align*}
\]

\[
\begin{align*}
\text{natural logs:} & \\
\ln(\hat{W}) &= -3.68 + 0.0780(56.0) \\
&= 0.692 \\
\hat{W} &= e^{0.692} = 2.00 \text{ kg}
\end{align*}
\]
6) The model $W = a + bL$ cannot be directly compared to a model with a response variable $\log W$ using the correlation or the standard error of the residuals, because the scales differ. How would you evaluate your transformed model in question (4) to see if it is an improvement over the linear model?

6. The pattern of residuals for the linear and exponential case allows us to evaluate the two models. As seen below, the clear pattern in the linear model residual plot vs. no pattern in the exponential model residual plot indicates that the exponential model fits the data better.