11 Regression

- The Correlation Coefficient
- The Least-Squares Regression Line

The Correlation Coefficient

Introduction

A bivariate data set consists of \( n \) \( \ldots (x_1, y_1), \ldots (x_n, y_n) \).

A scatterplot is a \( \ldots \) of a bivariate data set.

e.g. Here are data for 13 sparrowhawk colonies relating the \% of adult sparrowhawks in a colony that return from the previous year and the number of new adults that join the colony:

\[
\begin{array}{llllllllllllll}
\text{%Returning adults} & 74 & 66 & 81 & 52 & 73 & 62 & 52 & 45 & 62 & 46 & 60 & 46 & 38 \\
\text{#New adults} & 5 & 6 & 8 & 11 & 12 & 15 & 16 & 17 & 18 & 18 & 19 & 20 & 20 \\
\end{array}
\]

The right-hand scatterplot, below, is from these data. It shows \( \ldots \)
The Correlation Coefficient

The correlation coefficient, $r$, measures the _______ and _______ of the linear relationship (if any) between $x$ and $y$:

$$ r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) $$

An Informal Explanation of $r$

- Start with a scatterplot.
- Shift reference point to $\bar{x}$ by subtracting $\bar{x}$ from each $x_i$ and $\bar{y}$ from each $y_i$.
- Rescale the $x$-axis by dividing each $x$ coordinate by $s_x$, and rescale the $y$-axis by dividing each $y$ coordinate by $s_y$.
  Now, $x$ coordinates, $\frac{x_i - \bar{x}}{s_x}$, have mean _______ and standard deviation _______. $y$ coordinates, $\frac{y_i - \bar{y}}{s_y}$, have the same mean and standard deviation.
- Analyze the sign of the $i^{th}$ term in the last sum above, $\left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$, by quadrant:

  ![Graph of 50 Random Points, standardized (r=0.09)](image1)
  ![Graph of 13 Sparrowhawk Colonies, standardized (r=-0.75)](image2)

  e.g. For the sparrowhawk data, $r = \frac{\text{value}}{\text{value}}$. For the random data, $r = \frac{\text{value}}{\text{value}}$. 
Properties of $r$

- $-1 \leq r \leq 1$, and

$$r = \pm 1 \implies \text{data are } \underline{\text{_________}}; \quad r \approx \pm 1 \implies \text{data are } \underline{\text{____________________}}$$

$$r \neq 0 \implies \text{some linear relationship: } x \text{ and } y \text{ are correlated}$$

$$r > 0 \implies \text{slope of line is } \underline{\text{_________}}$$

$$r < 0 \implies \text{slope of line is } \underline{\text{_________}}$$

$$r \approx 0 \implies \text{no linear relationship: } x \text{ and } y \text{ are } \underline{\text{____________________}}$$

- $r$ doesn’t distinguish between _____ and _____

- $r$ doesn’t depend on ________ or __________
Cautions

• \( r \) measures strength of a linear relationship; check scatterplot to avoid using \( r \) for a _______.

  e.g. The data \{ (-2, 4), (-1, 1), (0, 0), (1, 1), (2, 4) \} fit ________, but \( r = 0 \) because the data have no _______ relationship (draw).

  e.g. (from http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient)

  ![Correlation Coefficients Graph]

• \( r \) is not resistant to the influence of ________: don’t use it for a data set with ________

  e.g. Adding \( (0, 0) \) to the sparrowhawk data changes \( r \) to ________.

• Correlation does not imply causation:
  A _______ (or lurking) variable is one _______ under consideration that correlates with both the independent and dependent variables of interest.

  e.g.
  - Increasing ice cream sales are correlated with increasing ________ rates. Does ice cream cause ________? ______
    The confounding variable is ____________________________.
  - Sleeping with shoes on is correlated with ____________________________.
    Does sleeping with shoes on cause ________? ______
    The confounding variable is ____________________________.

If either the independent variable under study, or a _______ confounding variable, affects the dependent variable, then both will seem to by the (__________) criterion of correlation.

___ cartoon
The Least-Squares Regression Line

A line is one that describes how a dependent variable, \( y \), changes as an independent variable, \( x \), changes in a data set \((x_1, y_1), \ldots, (x_n, y_n)\). We use it to predict \( y \) for a given \( x \).

The least-squares regression line is the line that the data (according to a reasonable criterion).

Notation includes:

- \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \): an unknown true (model) regression line, where \( \beta_0 \) is the \( y \)-intercept, \( \beta_1 \) is the slope, and \( \epsilon_i \) is the \( i \)th random error

- \( y = \hat{\beta}_0 + \hat{\beta}_1 x \): estimated regression line, where
  - \( x \): ____________ variable
  - \( y \): dependent variable
  - \( \hat{\beta}_0 \): estimated \( y \)-intercept
  - \( \hat{\beta}_1 \): estimated ____________

- \((x_i, y_i)\): \( i \)th data point

- \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \): ____________ value of \( y \) given \( x = x_i \):

- \( e_i = y_i - \hat{y}_i \): residual, the difference between observed \( y_i \) and predicted \( \hat{y}_i \); estimates \( \epsilon_i \)

We predict \( y \) from \( x \), so minimize vertical error in the “least squares” sense by minimizing a “sum of squared errors”

\[
SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\]

(Alas, really it should be called a “sum of squared ____________.”) Ten lines of calculus gives:

For the data set \((x_1, y_1), \ldots, (x_n, y_n)\), the least-squares line is \( y = \hat{\beta}_0 + \hat{\beta}_1 x \), where

\[
\hat{\beta}_1 = \frac{s_y}{s_x} \text{ (slope)}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ (y-intercept)}
\]
e.g. Here again are data for 13 sparrowhawk colonies relating the % of adults in a colony that return from the previous year and the number of new adults that join the colony:

\[
\begin{align*}
    x &= \text{% Returning adults} \quad 74 \quad 66 \quad 81 \quad 52 \quad 73 \quad 62 \quad 52 \quad 45 \quad 62 \quad 46 \quad 60 \quad 46 \quad 38 \\
    y &= \text{# New adults} \quad 5 \quad 6 \quad 8 \quad 11 \quad 12 \quad 15 \quad 16 \quad 17 \quad 18 \quad 18 \quad 19 \quad 20 \quad 20
\end{align*}
\]

Use a calculator to find the least-squares line (recall slope \( \hat{\beta}_1 = \frac{s_{yw}}{s_x^2} \), y-intercept \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \)):

\[
\begin{align*}
    \bar{x} = \quad & \\
    \bar{y} = \quad & \\
    s_x = \quad & \\
    s_y = \quad & \\
    r = \quad & \\
\end{align*}
\]

\[
\Rightarrow \quad \hat{\beta}_1 = \quad & \\
\hat{\beta}_0 = \quad &
\]

So our model is \( y = \quad \)

Or we can do it more directly. (Figure out your \____________ labels.)

e.g. Predict the number of new adults in a colony to which 60% of last year’s adults return.

\( \hat{y} = \____________ \)

(Note that this is far from the data set value, (60, ____).)

**R code for correlation and regression**

```r
returning = c(74,66,81,52,73,62,52,45,62,46,60,46,38) 
new = c( 5,6,8,11,12,15,16,17,18,18,19,20,20) 
cor(x=returning, y=new) # cor() gives correlation
model = lm(new ~ returning) # lm() gives linear model
plot(returning, new, xlim=c(0, 85), ylim=c(0, 35)) # scatterplot
abline(model) # abline() adds line
```