Ch. 6 – Scatterplots, Association, and Correlation

So far, we've seen *univariate* data. This section, however, considers *bivariate* data and how two *numerical* variables are related. Methods of description are introduced here and formalized in Ch. 24 (or STAT 252).

Terminology:

x	y
Explanatory variable	Response variable
Independent variable	Dependent variable
Predictor variable	Predicted variable

Notation:

- bivariate sample of size n: $\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$
- sample means: $\bar{x}$, $\bar{y}$
- sample std dev.: s_x, s_y

Displaying relationships:

Def'n: An association exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

A scatterplot is a graphical display of two quantitative variables.

- x-variable goes on the x-axis, y-variable on the y-axis
- origin (0,0) may be included

Look for:
- form of relationship (any obvious pattern)
- strength of relationship (closeness of fitting to a line)
- direction of relationship (positive or negative association)
- any unusual observations or outliers

Ex6.1)

x	y
1	1
2	2
4	1
3	2

(graph of above data used to discuss scatterplot traits further)

Correlation:

Def'n: Pearson's Sample Correlation Coefficient r is given by

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

where z_{x_i} is the "standardized" observation for x_i and z_{y_i} is the "standardized" observation for y_i for $i = 1, \ldots, n$

Properties of r:

- A measure of the LINEAR relationship between two variables.
- $-1 \leq r \leq 1$
- The magnitude of r (or absolute value) measures the strength of the relationship:
  - If $r = \pm 1$, then the points follow a straight line.
  - If $r = 0$, then the pattern of scatter suggest no linear relationship.
- The sign of r indicates the nature of the relationship:
  - Positive association if $r > 0$,
  - Negative association if $r < 0$.
- Correlation treats x and y symmetrically.
- Center and scale invariance (unitless).
- We can have $r = 0$, even when the data reveal a strong nonlinear relationship.
  - Such as $y = x^2$
- Correlation does not imply causation (or vice versa).
- Since r depends on the mean and std. dev., it is sensitive to outliers.

Ch. 7/8 – Introduction to Simple Linear Regression

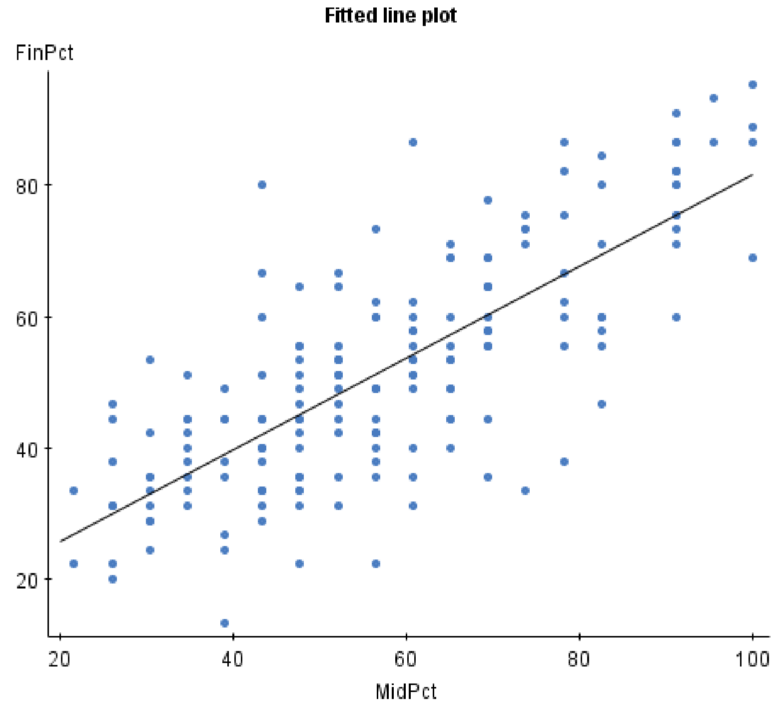
Ex7.1) Suppose you had 4 variables for the Oilers roster: height, weight, jersey, age

- which relationships might be valid?
- how can we describe the relationship between any pair?
- how do we use the description to make predictions?
- how do we quantify errors in estimates and predictions?

Def'n: The regression line predicts the value for the response variable y as a straight-line function of the value x, the explanatory variable.

Equation for the regression line: $\hat{y} = b_0 + b_1 x$

- b_0 is the intercept: the height of the line at $x = 0$.
- b_1 is the slope: the amount by which y changes when x increases by 1 unit.
- $\hat{y}$ ("y-hat") denotes the predicted value of y (or mean y for a given value of x).



What about a new student who gets a mark of 80.1%? No observation so can we estimate the final mark based on the pattern of the other observations? Try and fit a line through the data and use it as a model for final percentage given midterm percentage; then, use the line to estimate (or, interpolate) the final percentage for a student that gets 80.1% on the midterm.

Def'n: Regression analysis tells how to fit a line to the overall pattern. This equation, or "model", may estimate or predict other values of y given values of x. Simple linear regression refers specifically to fitting a straight line ("linear") and using only ONE explanatory variable ("simple").

Least squares estimation of b_0 and b_1:
Def'n: A residual is the difference between an observed value and its estimated value. Since $\hat{y}$ denotes the estimated value of y, then at some observed value of x, say x_i, the residual is defined as

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

The residual represents the vertical deviation of the point from the line. We want to choose (b_0, b_1) to minimize the sum of squared deviations (hence "least squares"):

$$\sum (y_i - b_0 - b_1 x_i)^2$$

Using calculus, the corresponding solution becomes

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = r\left(\frac{s_y}{s_x}\right) \qquad \text{and} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

Ex7.2) Choosing to predict final% from midterm% (both vars. are continuous)

x = midterm percentage, y = final percentage

$n = 180$, $\bar{x} = 57.923$, $\bar{y} = 52.123$, $s_x = 19.251$, $s_y = 17.588$, $r = 0.766$

a) Estimate and interpret slope and intercept.

Estimated equation for the regression line:

b) Estimate final percentage when midterm percentage is 80.1%.

c) Estimate final percentage when midterm percentage is 50%.

d) Estimate the *average* difference in final percentages for midterm% of 65% and 75%.

Assorted Topics on Simple Linear Regression:

- prediction and estimation:
  - Benefit: the model allows for prediction of y given values of x. This predicted value is also called the *fitted value*.
  - Benefit: estimating with values of x not contained in data but *within* the range of the observed values of x (or interpolation).
  - Caution: estimating values of y outside the range of the observed values of x (or extrapolation) is VERY dangerous.

- causation:
  - although x and y may be associated, this does NOT imply that x "causes" y.
    - → Association/correlation does not imply causation.
  - association may be due to a *lurking variable*.
  - causation is possible if a valid experiment design exists (see Ch. 9-11).

- re-expressing (or transforming) data:
  - if a scatterplot identifies a non-linear pattern, re-expressing the data can "straighten" the pattern. Common transformations are:
    - Square: x^2 → For left-skewed data.
    - Logarithm: $log(x)$ or $ln(x)$ → For right-skewed data.
    - Square-root: $\sqrt{x}$ → For counts.
    - Reciprocal: $\frac{1}{x}$ → For ratios of quantities (such as km/h).