Benjamin Fenster
STAT-141 (MacPherson)
4 October 2003

Notes – Bivariate Statistics

I. Introduction
  a. So far we've only dealt with univariate statistics: one measurement for each EU.
  b. Could measure two things per EU (bivariate) or more (multivariate)
  c. We want to consider bivariate now.
  d. Consider X independent, Y dependent and use X to predict Y.
  e. If both are quantitative, need to see Chapters 5 and 13.
  f. If X is qualitative, Y quantitative, need Chapter 15 (analysis of variants, ANOVA)

II. Long-Volume Example
  a. Can plot data, draw line to represent it.
  b. Not a perfect predictor in this example.
  c. Want to capture the relationship between the two variables.

III. Notation
  a. Variance and Covariance
    i. $VAR(X) = s_x^2 = \sum(x - \bar{x}) / (n - 1) = S_{xx} / (n - 1)$
    ii. $VAR(Y) = s_y^2 = s_{yy} / (n - 1)$
    iii. $COV(x, y) = [\sum(x - \bar{x})(y - \bar{y})] / (n - 1) = s_{xy} / (n - 1)$
      1. Cross product sums of squares
      2. Variances are always ≥ 0.
      3. Covariance can be any number.
    iv. $s_x = 3.085$, $s_y = 0.501$ (standard deviation)
  b. Correlation Coefficient
    i. $r = COV(x, y) / (s_x * s_y)$
    ii. Scaled with s_x and s_y
    iii. Between –1 and +1
    iv. Lungs: 1.068 / (3.085 * 0.501) = +0.692
    v. General Formula: $\sum(x - \bar{x})(y - \bar{y}) / \sqrt{(\sum(x - \bar{x})^2 \sum(y - \bar{y})^2)} = \sqrt{(s_{xx} s_{xy})}$
    vi. Characteristics
      1. Population (parameter) correlation coefficient = ρ
      2. Sample (statistic) = r
      3. $-1 \leq r \leq 1$ always
      4. Magnitude measures *strength*, sign measures *direction*
      5. r = 0 means no correlation, +1 means perfect positive, -1 means perfect negative.
    vii. Lung Volumes
      1. $s_{xx} = (79.4 - 85.5)^2 + \ldots + (89.5 - 85.5)^2$
      2. $s_{yy} = \ldots$
      3. $s_{xy} = (79.4 - 85.5)(4.3 - 5.06) + \ldots + (89.5 - 85.5)(5.3 - 5.06)$
    viii. Caveats
      1. r = 0 doesn't imply that there is no relationship!
        a. Could have a more subtle relationship.
        b. Correlation coefficient measures *linear* correlation.
        c. The moral: Always plot the data!
      2. Large r does not imply causation
        a. Correlational fallacy: post hoc ergo proper hoc (after which, therefore because of which)
        b. Example
          i. x = 18 hole golf courses, y = number of divorces (1960 – 1990)
          ii. There's a definite positive correlation.
          iii. Causation? Perhaps merely the increasing population caused both!
        c. Lurking Variables: Some other variable may cause both x and y.

d. Establishing causation requires a carefully designed study, perhaps with control of lurking variables.
e. Partial Correlation
   i. $r_{xy*z} = (r_{xy} – r_{xz}r_{yz}) / \sqrt{((1 = r_{xz}^2)(1 – r_{yz}^2))}$
      1. Partial correlation for x & y, controlling for z.
      2. First order partial
   ii. $r_{xy*wz} = (r_{xy*w} – r_{xz*w} * r_{yz*w}) / \sqrt{((1 – r_{xz*w}^2)(1 – r_{yz*w}^2))}$
      1. Controlling for both w and z.
      2. Second order partial.

c. Spearman's Correlation Coefficient
   i. So far we've been using Pearson's product moment correlation (r), the traditional method.
   ii. Spearman's denoted r_s
   iii. Non-parametric coefficient (more resistant)
   iv. The formula in the book (pg 145) is a little tedius. Look anyway.
   v. Uses rank
      1. Ordering, compensating for equal values.
      2. 10, 20, 20, 30 becomes 1, 2½, 2½, 4
   vi. Spearman's coefficient represents the correlation between the ranks.
   vii. Resistant to outliers.
   viii. Can identy non-linear relationships too.
   ix. Excellent when data is already ranked.

IV. Regression Problem
a. How do we get a best-fit line?
b. $\hat{y}$ = "the line"
c. Deviation = y_i - τ (vertical distance between the real point and the line)
d. τ = a + bx
e. Want to find the best values for a and b.
f. $SSResid = \sum d_i^2 = \sum (y_i – \hat{y})^2 = \sum(y_i – (a + bx_i))^2$
g. By applying some calculus to find the minimum, we get…
   i. $b = \sum(x – \bar{x})(y – \bar{y}) / \sum(x – \bar{x})^2$
   ii. $a = \bar{y} - b\bar{x}$
   iii. $b = r (s_y / s_x)$
   iv. Note that these solutions are unique
   v. Don't worry about how to get from the calculus to the formula.
h. Predicting y given x
   i. Putting $\bar{x}$ into the equation, you always get $\bar{y}$
   ii. Interpolation is okay (predicting y for some x that falls within the range of the data)
   iii. Extrapolation = Beware! Don't try to go too far from the range of the original data.
i. Equation Forms
   i. Slope-Intercept
      1. $\hat{y} = a + bx$
      2. Makes mathematical sense, but the value for a often makes no practical sense.
   ii. Point-Slope
      1. $\hat{y} = \bar{y} + b(x - \bar{y})$
      2. Centered Form, or "Empirical Regression Form"
      3. Makes practical sense.

V. How Good is the Line
a. Assessing the fit.
b. Uses residuals.
   i. Points near the line are called "smooths"
   ii. Points farther away from the line are called "outliers"

c. Measures of deviation
   i. $y_i - \bar{y}$ (good for univariate data)
   ii. $y_I - \hat{y}$ (should be smaller on average than $y - \bar{y}$)
   iii. $\hat{y} - \bar{y}$ (how far is the line from the mean)
d. Fundamental Identity of Simple Linear Regression
   i. $\sum(y_i - \bar{y})^2 = \sum(\hat{y} - \bar{y})^2 + \sum(y_i - \hat{y})^2$
   ii. "Total Sum of Squares" = "Regression SS" + "Residual SS"
   iii. (Residual also called Error)
   iv. Doesn't make sense for a single point, but for the whole collection it's good.
   v. $SS_{TOTAL} = SS_{REGRESSION} + SS_{RESIDUAL}$
e. Degrees of Freedom
   i. Total (n – 1) = Regression(1) + Residual (n – 2)
   ii. (n – 1) = 1 + (n – 2)
   iii. How much information do we have for estimation purposes?
   iv. For perfect regression, $SS_{TOTAL} = SS_{REGRESSION}$
   v. For random data, $\hat{y}$ collapses to $\bar{y}$, $SS_{TOTAL} = SS_{RESIDUAL}$
f. Coefficient of Determination
   i. $R^2 = SS_{REGRESSION} / SS_{TOTAL} = 1 – SS_{RESIDUAL} / SS_{TOTAL}$
   ii. Scale is 0 to 1, often expressed as a percentage.
   iii. R^2 is just the correlation coefficient squared.
   iv. $R^2 = 0$ where $SS_{RESIDUAL} = SS_{TOTAL}$
   v. $R^2 = 1$ where $SS_{RESIDUAL} = 0$
   vi. See pages 141 & 170
      1. Called "Weak" where $0 \le R^2 < 25\%$
      2. Called "Moderate" where $25\% \le R^2 < 64\%$
      3. Called "Strong" where $64 \le R^2 \le 100\%$
   vii. This is a very big topic in statistics: it quantifies variability.
g. Standard Deviation of Regression
   i. $s_e^2 = \sum(y_i – \hat{y})^2 / (n – 2) = SS_{RESID} / (n – 2)$
   ii. s_e: Standard deviation of regression
   iii. How far does a typical observation deviate from the line?
   iv. This is another important measure of the adequacy of the model.
   v. Might want to use $s_e / \bar{y}$ to put the number in perspective.
h. Mean Square Residual
   i. s_e^2
   ii. Example

$SS_{RESIDUAL}$	100	100
$SS_{REGRESSION}$	100	900
SS_{TOTAL}	200	1000
R^2	50%	90%
s_e^2	10	10

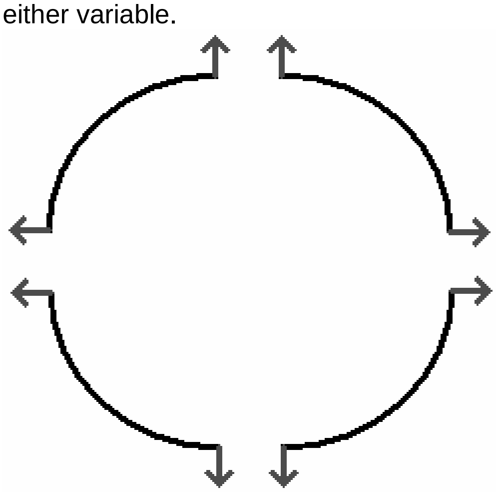
   iii. The moral: These numbers mean two different things!

VI. Residual Analysis
a. $d_i = r_i = y_i - \hat{y}$
b. Plot residuals against x.
c. Want to see values scattered randomly on the residual plot.
d. Standardized Residual
   i. Similar to Z-Score (not the same, but similar)
   ii. Want to see values between –2 and +2.
e. Seeing some pattern in the data suggests there's a non-linear relationship.
f. If variability increases (small residuals for small x's, larger residuals for big x's, for example), then may need to transform the data.
g. Interesting Points
   i. Outlier: Extreme Y value
   ii. Leverage Point: Extreme X value

iii. Influential Point: Both X and Y are extreme
  1. Has the ability to greatly influence the data. May create the illusion of linearity when it doesn't really exist, or change the slope.
  2. The Remedy: Remove the point and see what effect it has.

VII. Transforming Data

a. Sometimes data doesn't seem to be linear but has a linear relationship "hiding" in it somewhere.
b. We'd like to re-express (transform) non-linear data into some linear form.
c. x = original value, x' = transformed value.
d. Box-Cox Transforms
  i. $x' = x^p$ for some p
  ii. The Power Transform Ladder
    1. p = 2, square
    2. p = 1, no transform
    3. p = ½, square-root
    4. p = $^1/_3$, cube-root
    5. p = "0" (special case: x' = log(x))
    6. p = -1, x' = -1/x (use negative sign to preserve the order of the data)
e. Falling Body Example
  i. S vs. t looks non-linear
  ii. Could go up the ladder with respect to t
  iii. Could go down the ladder with respect to S.
  iv. Typically we'd look at S vs t^2
f. Use this graphic as a guide for which direction to go (up or down) with respect to either variable.



ERROR: undefinedfilename
OFFENDING COMMAND: </FONT>

STACK: