Recap, Regression and Some R

Kathleen M. Saul

April 26, 2011

One Population, One Measurement, Assuming Normal Distributions

- Ho: mu_o = value
- Ha: mu_o > or < or not equal value
- Statistic to use, sigma not known:
  - t = (x bar – mu_o) / (s / sqrt(n))
  - With df = n – 1
- R command: t.test(dataset, mu=value)
  - Where mu is the Ho value, mu_o
  - Returns the t statistic, df, p-value
- If p-value < alpha, reject Ho.

- Example: Average tree diameter at breast height on the HJ Andrews site.

www.fsl.orst.edu/~bond/images/ws1%20tower.jpg

Two Independent Samples, One Measurement Each, Assuming Normality

- Ho: $mu_a - mu_b = 0$
- Ha: $mu_a - mu_b >$ or $<$ or not equal 0
- Are the variances equal or not equal?
  - First approach: If either sa >= 4 * sb or
    - sb >= 4 * sa, assume unequal variances.
  - Second approach: F test of variances
    - Based on Ho: $var_a - var_b = 0$
    - In R: var.test(a~b)
    - Returns an F statistic and p-value.
    - If the p-value < 0.05, reject Ho and assume unequal variances.

- For equal variances:
  - Calculate a t statistic using a pooled variance (sp^2) and df = ($n_a + n_b - 2$)
  - t = [(x_a bar – x_b bar) – ($mu_a - mu_b$)] / [sp^2 ($1/n_a + 1/n_b$)]
  - In R: t.test(a~b,var.equal=T)
    - Returns the Two Sample T test, t statistic, df, and p-value
    - If p-value < alpha, reject Ho.

- For unequal variances:
  - Calculate a t statistic using a df weighted by sample sizes
  - $t = [(x_a \text{ bar} – x_b \text{ bar}) – (mu_a – mu_b)] / [(s_a^2/n_a) + (s_b2/n_b)]$
  - In R: t.test(a~b)
  - Returns the Welch Two Sample T test, t statistic, df, and p-value
  - If p-value < alpha, reject Ho.

- Example: Tree diameter of a sample of trees adjacent to a stream versus tree diameter of a sample of trees along a forest service road.

journeyheretothere.com/new_page_1.htm

Matched Pairs

- For two samples linked by before/after, old/new measurements
- Focus on the difference D, where
  - $x_{1D} = x_{1a} – x_{1b}$ and
  - $mu_D = mu_a – mu_b$
- Ho: $mu_D = 0$
- Ha: mu_D > or < or not equal 0
- The test statistic is
  - $t_D = (x_D \text{ bar} – mu_D) / (s_D / \text{sqrt}(n_D))$
  - With df = $n_D - 1$

- In R:
  - t.test(a,b,paired=T)
  - Returns t statistic, df, and p-value
- If p-value < alpha, reject Ho.

- Example: Sediment levels in a river or lake before and after clear-cutting.

Two Populations of Very Non-normal Data

- The Wilcoxon test: Replaces the actual data values with ranks and finds the totals the ranks of positive and of negative differences.
  - For samples of n > 30, those totals will be normally distributed.
- Ho: Rank totals are the same (the population locations are the same)
- Ha: Rank totals are different (population locations are different)

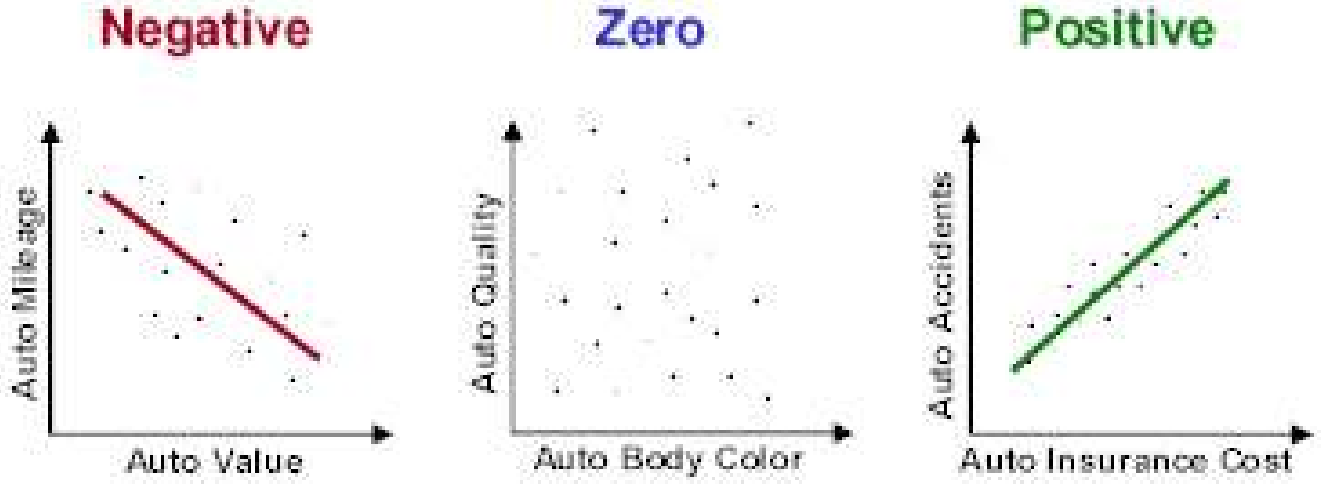
- Uses a z (or W) statistic, $z = T_t - E(t) / sigma_T$
  - With $E(T) = n (n +1) / 4$
  - and $sigma_T = sqrt [n (n+1) (n+2) / 24]$
- In R: wilcoxon.test(a~b)
- Returns the Wilcoxon Rank Sum Test, a W statistic and p-value.
- For p-value < alpha, reject Ho.

Correlation

- A measure of the direction and strength of the linear relationship between two quantitative variables.
- For normally distributed variables, we use Pearson's r to characterize the relationship: $-1 < r < +1$
- The closer to -1 or +1, the stronger the relationship. $r = 0$ indicates a weak linear relationship.

Correlation

Relationship Between Two Quantities
Such That When One Changes, the Other Does



www.russell.com/.../correlation_graphs.gif

- In R, for two vectors:
  - cor(vector1,vector2)
  - To ensure that R eliminates missing values, type cor(vector1,vector2,use="complete.obs")
  - Returns the r value indicating the relationship between the vectors.
- For a data frame containing many variables,
  - cor(data name,use="complete.obs")
  - Returns a matrix of r values

	Total monthly net income	Services as head of household's main economic activity	Number of income earners in household	Number of household members	Age of head of the household	Ratio of number of teenagers and youths to total number of individuals living in household	Ratio of male to total number of individuals living in household	Illiterate members in household	Maximum educational level attained by any household member	Zone	Region	Income level of household with a mobile telephone	Income level of household with a fixed telephone	Zone if there is any kind of telephone in the Coast region
Total monthly net income	1													
Services as head of household's main economic activity	0.1827	1												
Number of income earners in household	0.3385	0.0434	1											
Number of household members	0.1557	-0.0184	0.5068	1										
Age of head of the household	0.0596	-0.1264	0.1738	-0.0653	1									
Ratio of number of teenagers and youths to total number of individuals living in household	0.0591	0.0176	0.2084	0.1525	-0.2676	1								
Ratio of male to total number of individuals living in household	-0.0188	0.0007	-0.0194	-0.0589	-0.0599	0.1201	1							
Illiterate members in household	-0.1724	-0.1285	-0.0057	0.2739	0.024	-0.1581	-0.1014	1						
Maximum educational level attained by any household member	0.4833	0.3382	0.3289	0.1842	-0.0921	0.1976	-0.0007	-0.3247	1					
Zone	0.3289	0.2099	0.2176	-0.0057	-0.0139	0.0918	-0.0426	-0.2954	0.4737	1				
Region	0.2039	0.0364	0.1524	-0.0331	0.0713	-0.0078	-0.0323	-0.1385	0.1804	0.3029	1			
Income level of household with a mobile telephone	0.7238	0.1126	0.1208	0.0226	0.0198	0.0116	-0.0185	-0.086	0.2495	0.1482	0.13	1		
Income level of household with a fixed telephone	0.8105	0.1334	0.2233	0.0642	0.0954	0.015	-0.0466	-0.1538	0.3891	0.258	0.1657	0.6496	1	
Zone if there is any kind of telephone in the Coast region	0.394	0.1089	0.1907	0.0345	0.088	0.0179	-0.0464	-0.1847	0.3365	0.3365	0.4332	0.3315	0.5008	1

www.idrc.ca/en/ev-114381-201-1-DO_TOPIC.html

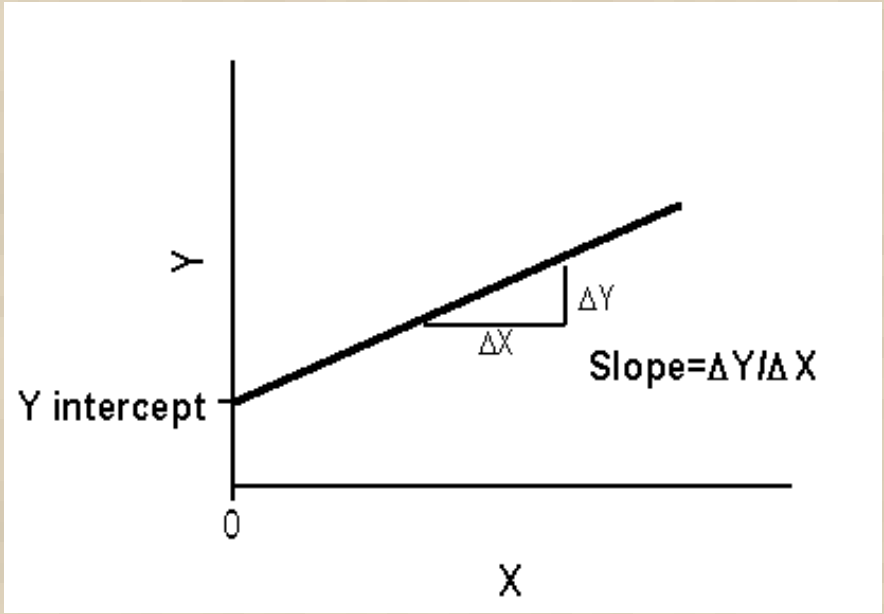
- For non-normal (skewed) data, the Spearman's rank coefficient (rho) uses ranks to compute correlation (similar to Wilcoxon)
  - Ho: a and b are independent (no statistical correlation) and rho = 0
  - Ha: a and b are not independent, rho not equal 0
- In R: cor.test(a,b,method="spearman")
  - Returns an S statistic, p-value, and rho value
  - If p-value < alpha, reject the null of no correlation
  - rho gives a measure of the strength and direction of the association between the variables: -1 < rho < +1

- Kendall's tau also applies to non-normal data but unlike the other measures, which indicate a proportion of variability accounted for, tau gives a probability based on ranking/ordering of data
  - Finding tau relies on looking at whether the pairs are condordant $_{+}$ + or discordant + $_{+}$
  - tau measures the strength of the relationship between paired variables (a_i, b_i)
  - Like r and rho, -1 < tau < +1
  - Hypotheses:
    - Ho: a and b are independent (no correlation) and tau = 0
    - Ha: a and b are not independent, tau not equal 0

- In R: cor.test(a,b,method="kendall")
  - Returns a z statistic, p-value, and tau
  - If p-value < alpha, reject the null of no correlation

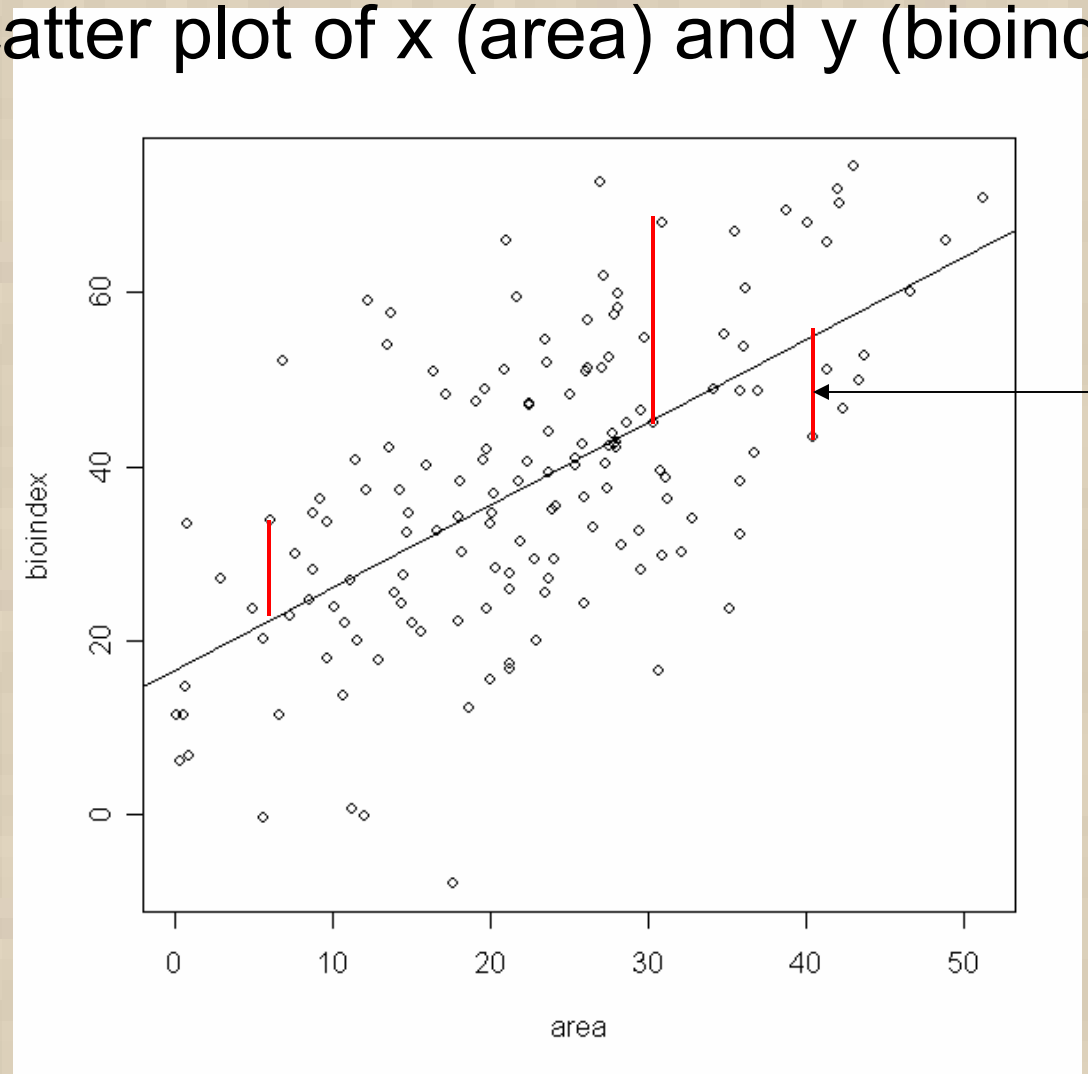
Simple Linear Regression

- Basic assumption: Causality or close relationship between variables x and y.
- Model: y = a + bx + error
  - With a = y intercept (x = 0) and b = slope



- The regression line describes how the response variable y changes as the explanatory variable x changes.
  - It can be used to predict the value of y for any given value of x.
- Least squares regression minimizes the sum of the squares of the vertical distances between the line and the plotted data points. That vertical distance is known as the “residual”.

A scatter plot of x (area) and y (bioindex):



- For regression, hypotheses relate to the coefficient of the x variable:
  - Ho: $B = 0$
    - Interpretation: There is no straight line relationship between x and y.
  - Ha: B not equal 0
    - There is a straight line relationship.
- In R:
  - lm(y~x)
  - Returns the intercept value and the x coefficients

- > lm(bioindex~area)
- Call: lm(formula = bioindex ~ area)
  - Coefficients:
  - (Intercept) area
  - 16.6471 0.9505
- Giving us bioindex = 16.6471 + 0.9505 * area

- > summary(lm(bioindex~area))

```
Call:
lm(formula = bioindex ~ area)
Residuals:
     Min       1Q   Median       3Q      Max
-41.1073  -9.0200  -0.3371   9.3601  30.8269
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.64706    2.40695   6.916 1.30e-10 ***
area         0.95050    0.09573   9.929  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 13 on 148 degrees of freedom
Multiple R-squared: 0.3998,     Adjusted R-squared: 0.3957
F-statistic: 98.58 on 1 and 148 DF,  p-value: < 2.2e-16
```

- r^2 gives the fraction of variation in the values of y explained by the regression of x on y, or the proportion of variation in y explained by the variation in x.
- Pr(>|t|) reflects the significance of the variable x in the model. The smaller the value, the more significant is x.
- p-value gives the p-value for the overall model. If p-value < alpha, we can reject our null of no linear reationship.

- To be continued . . .