RNR / ENTO 613 – Simple Linear Regression

Context

Regression analysis investigates the statistical relationship between 2 *continuous* variables: a *reponse* variable (Y) and an *explanatory* variable (X). The simple linear regression model specifies that the relationship between the *mean of a response* and the *explanatory* variable is a straight-line.

In general, linear regression is appropriate when the *roles* of response and explanatory variables are *distinguishable*. In both experimental and observational studies, the goal of the regression approach is to determine how *changes in the explanatory variable* affect the response.

Example: Describe the relationship between temperature and development time in a given insect species. Here the interest is in how *temperature* affects *growth rate*, not the reverse.

Example: Describe the relationship between leg and arm length in baboons; there is no clear dependence of one variable on the other. Here, the appropriate approach is *correlation analysis*, not regression.

Applications of linear regression technique

1) Estimation of *functional relationships* between variables, which is achieved by estimation of the *parameters* of the regression line (i.e. slope and intercept). Example: Relationship between time after slaughter and pH of meat (Sleuth, Chap. 7).

2) *Prediction* of the distribution of specific values of the explanatory or response variable (mean of Y at X, mean X at a single Y, value of a single Y at X, value of a single X at a single Y). Example: from the regression between meat pH and time after slaughter, it was predicted that at least 95% of steer carcasses would reach a pH of 6.0 between 2.9 and 5.1 hours after slaughter.

Overview of least squares regression approach

1- The least squares approach is used to estimate the value of the parameters of the regression line (i.e. slope and intercept)
2- The standard error of the statistics (estimated slope and intercept) is computed using:
   a) the variance of the explanatory variable
   b) the standard deviation of the response variable, computed from the residuals of the regression line
3- t-tests are used to assess possible values for the parameters (i.e., the slope and intercept): Confidence intervals are established for these parameters.
4- Confidence or prediction intervals may be used for prediction of a future response, or for guessing the X value that results in a given mean response.

Historical background

Resemblance between offspring and their parents is the fundamental concept in quantitative genetics (it is called the heritability). Sir Francis Galton, Charles Darwin's cousin, introduced a graphical representation of offspring-parent resemblance. He showed that a least squares line fitted to data of parent and offspring height was different from the line representing perfect inheritance (slope of 1). Offspring of tall parents regressed downward toward the offspring mean, while offspring of short parents regressed upward toward the mean. Galton thus called this best fit line a *regression*.

<<Fig. Galton>>

The regression phenomenon was first seen as a major problem for Darwin's theory of evolution, since continuous regression of traits toward the mean (following breeding) would oppose the effects of natural selection. The dilemma was solved by a rediscovery of Mendelian genetics. With Mendelian inheritance, the shape of the regression line remains constant across generations as long as selection does not change gene frequency. Large parents produce smaller offspring on average (with dominance), but large offspring are always regenerated in similar proportions in the population according to the rules of Mendelian inheritance. The regression (environmental) effect described in Sleuth (p. 193) would be a complementary explanation for the fact that extreme parents tend to produce offspring that are closer to the population mean.

On the origin of normal distributions

A central assumption in parametric statistics is that variables are normally distributed. The theory of quantitative genetics explains why such an assumption holds for many biological characters. Quantitative genetics assumes that most traits are controlled by many loci that contribute a small amount to the trait value, and environmental effects that act independently and also contribute small effects to the phenotype. If these effects combine *additively*, then the overall phenotype of an individual depends on the *sum of randomly selected small effects* that operate during development. The phenotype is thus a normally distributed random variable that centers on the average sum of those effects.

<<Fig 1.3 in Roff>>

Regression Approach and Terminology

Let Y denote the response variable, and X the explanatory variable. $\mu\{Y \mid X\}$ represents the regression of Y on X, which is read as "the <u>mean</u> of Y as a function of X".

The simple linear model is:

$$\mu\{Y \mid X\} = \beta_o + \beta_1 X$$

where β_o is the *intercept* of the line, and β_1 is the *slope*.

The slope is a *statistic* that estimates *the rate of change* in the mean response *per one-unit increase* in the explanatory variable, for any X value within the range of interest:

$$\beta_1 = \frac{\mu\{Y|X=b\} - \mu\{Y|X=a\}}{b-a}$$

$\{Y \mid X = x\}$ should be read as "the mean of Y when X = x".

The units for β_1 are the ratio of the units of the response variable to the units of the explanatory variable. In the insect growth example, this could be days / °C.

Research questions often concern the change in the mean response associated with a specific change in the explanatory variable (for example from *a* to *b*). This is expressed as:

$$\mu\{Y|X=b\} - \{Y|X=a\} = \beta_1(b-a)$$

Regression model Assumptions

The Ideal Normal Simple Linear Regression Model assumes that the response variable (Y) is a *random* variable (i.e. a variable that we do not control). The explanatory variable (X) may be random or *fixed* (a variable that we control). The explanatory variable is *measured with small error* compared to Y.

Under these assumptions, the regression line is fitted with the *method of least squares*. We envision a series of subpopulations of responses (Y) at each level of X. All the subpopulations have equal standard deviation, and their means fall on a straight line.

Thus the major Assumptions of least square regression are:
1-Normality of subpopulations
2-Linearity of the subpopulation means
3-Equal SD of the subpopulations
4-Independence (within and among subpopulations)

<<Fig. 7.5 Sleuth>>

Note on measurement error of the X variable

Large measurement errors in the X-variables may cause estimation problems: the estimates of the slope of Y on X with a least squares method (explained below) in such case are *biased.* *<Fig Bias / Precision>* This is what is meant in the Big Bang study (p.177) : "Uncertainty due to errors in measuring velocities (i.e. the X's) is not included in the p-values and confidence coefficients. Such errors are potential source of pronounced bias".

In practice, the least square method is appropriate, as long as errors in the Xs is small compared to errors in the Ys.

The choice of Models for analyses of data when error measurement in X is not small is controversial. A general rule (see Sokal and Rohlf for more details):

1) Regression lines used mainly for purposes of *prediction* can be fitted with the least square method.
2) When one whish to determine *precisely* the slope of the association between 2 random variables (e.g., in morphometric work), other techniques such as *Major axis regression* should be used.

Example: Determine the relationship between weight (X) of individuals from a fish species and the number of eggs produced (Y). Here the interest is to compare the value of the slope among closely related species to assess evolutionary questions about allocation of resources to reproduction, not to predict reproductive output for a given fish size.

Weight (in g × 100)	Eggs (in thousands)
14	61
17	37
24	65
25	69
27	54
33	93
34	87
37	89
40	100
41	90
42	97

Least square estimation (assumes Xs measured with little error)*:*

Slope: 1.86 eggs × 10^3 / 100 g 95% CI (1.12, 2.62)
Intercept: 19.77

Reduced major axis (assumes 2 variables distributed according to a bivariate normal distribution)*:*

Slope: 2.12 95% CI (1.37, 2.87)
Intercept: 12.19

[With large error in the Xs, the slope of the reduced major axis regression is always greater than that of the least squares regression]

Least Squares regression Estimation

The method of least Squares is a general procedure to estimate parameters in a statistical model.

For simple linear regression, the goal is to find the *best-fitting* statistics, $\hat{\beta}_o$ and $\hat{\beta}_1$, for a data set comprising *n* pairs of observations (Y_i, X_i). These estimates then combine to provide an estimated mean function:

$$\mu\{Y|X\} = \hat{\beta}_o + \hat{\beta}_1 X \ .$$

This expression can be used to calculate a value of Y for any X_i. Y estimates the *mean of the distribution* of the response variable. The estimated mean is called a *fitted value* or *predicted value* and is represented by fit_i. The difference between the observed response and its estimated mean is the *residual*, denoted by res_i.

$$fit_i = \hat{\mu}\{Y|X_i\} = \hat{\beta}_o + \hat{\beta}_1 X_i \quad \text{and} \quad res_i = Y_i - fit_i$$

<<Fig 7.6 in Sleuth>>

A residual represents the distance between a *measured response* and its *fitted value.*

The distance between all responses and their fitted values is estimated by *the residual sum of squares*.

Least Squares Estimates

An estimate obtained with the method of least squares (LS) *minimizes the sum of the squared residuals:*

i.e. LS minimizes $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

For simple linear regression, the LS estimator of the slope is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

where the numerator is the sum of the *crossproducts* of X and Y, and the denominator is the *sum of squares* for X.

Slopes can be positive, negative, or zero. They predict the change in Y that results from a *one unit* change in X. For the fish *fecundity* vs *weight* example, $\hat{\beta}_1$ = 1.86, indicating that a one unit change in X (weight × 10^2 g) results in an increase in fecundity by 1.86 units (eggs × 10^3). (or b_1 = 18.6 eggs/g)

Both $\overline{Y}$ and $\overline{X}$ always lie on the line fitted with least squares. Thus after estimating $\hat{\beta}_1$, we can solve for $\hat{\beta}_o$ by substituting the mean for Y and X into the formula for a line, $\overline{Y} = \beta_o + \beta_1 \overline{X}$: $\hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X}$

Sampling distribution the Least Squares estimates

The estimators $\hat{\beta}_o$ and $\hat{\beta}_1$ are *unbiased* when the X variable is measured with small error.

<<Fig Biais / Precision>>

The Student's t distribution $t = \dfrac{\hat{\theta} - \mu_{\hat{\theta}}}{SE_{\hat{\theta}}}$ is used to draw inferences about estimates of the slope and intercept. To do so, we need to estimate the standard deviations of $\hat{\beta}_o$ and $\hat{\beta}_1$:

$$SD(\hat{\beta}_1) = \sigma \sqrt{\frac{1}{(n-1)s_X^2}}$$

$$SD(\hat{\beta}_o) = \sigma \sqrt{\left(\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_X^2}\right)}$$

which depend on:

1. n, which is known
2. the sample variation in the explanatory variable, s_x^2, which is known
3. the population standard deviation of the response, σ, which is not known.

<< Fig. 7.7 Sleuth>>

σ is estimated using the residuals from the regression:

$$\hat{\sigma} = \sqrt{\frac{Sum\ of\ \ all\ squared\ residuals}{Degrees\ of\ \ freedom}}$$, where *df* for the residuals =

No. Observations – No. Parameters in the model for the means = n - 2.

Tests and Confidence intervals for Slope and intercept

The standard null hypotheses about simple linear regression parameters are:

Ho: $\beta_o = 0$ Ha: $\beta_o \neq 0$
Ho: $\beta_1 = 0$ Ha: $\beta_1 \neq 0$

The general form of the t-ratio, with n –2 *df* is:

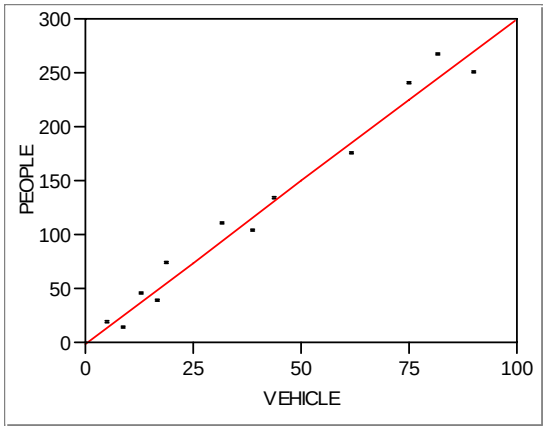
$$t = \frac{(parameter\ estimate - hypothesized\ value)}{SE\ of\ \ parameter\ estimate}$$

CI's are used to determine probable values of the regression parameters. They take the general form:

Parameter estimate $\pm$ $t_{n-2,\,(1-\alpha/2)}$ SE (***parameter estimate***)

An example:

Can we predict the number of visitors to a recreation area by knowing the number of cars in the parking lot? On 12 days, an employee monitors the number of cars in the parking lot and the number of users in the park. The goal is to assess whether it is reasonable to predict number of people in the park from knowledge on number of cars present, which is information that is easy to get.



PEOPLE = - 0.947634 + 3.0212969 VEHICLE

Summary of Fit

Rsquare	0.97669
RSquare Adj	0.974359
Root Mean Square Error	14.71194
Mean of Response	121.6667
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	90688.256	90688.3	418.9975
Error	10	2164.411	216.4	Prob > F
C. Total	11	92852.667		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.947634	7.342907	-0.13	0.8999
VEHICLE	3.0212969	0.1476	20.47	<.0001

The test of the null hypothesis of no association between car and visitor number (i.e. zero slope) is t = (3.021 – 0)/ 0.148 = 20.47 with n – 2 *df* (12 –2 =10) has a p-value smaller than 0.0001.

The CI's giving probable values for the regression parameters are:

$\hat{\beta}_1 \pm t_{n-2,(1-\alpha/2)} SE(\hat{\beta}_1)$ = 3.021 ± 2.228 (0.148) = 2.7, 3.3 people/ car. Each new car brings about 3 people, and the 95 % CI is 2.7, 3.3.

The intercept is not different from zero (95 % CI is –17.3, 15.4), suggesting that the main access to the park is by car (an intercept greater than zero would be expected if users have access to the park by hiking, cycling, flying, etc…..).

Making a single prediction with the regression line

When making predictions based on a regression model, we assume that future values will behave similarly to the ones observed in the past. For example, making predictions about crowding in the park for the winter season based on a regression model estimated for the summer could be misleading.

We obtained the following equation for the regression line:

Y = -0.95 + 3.02 X

With that equation, we can use *interpolation* to make 2 types of predictions:

What is the expected *average number of people* in the park on days when we have 10 cars in the parking lot? We want to predict a probable value for a *subpopulation mean* with a *confidence interval.*

Our best guess is: $\mu \{Y \mid X = X_o\} = \mu \{Y \mid X = 10\} = -0.95 + 3.02\ (10) = 29.2$

What is the expected *number of people* in the park on days when we have 10 cars in the parking lot? Here we want to predict the *value of a single response* with a *prediction interval*.

Our best guess is: $\text{Pred}\ \{Y \mid X = X_o\} = \mu \{Y \mid X = 10\} = -0.95 + 3.02\ (10) = 29.2$

In both cases the best prediction is the same, but the *precision* of the estimate will vary. The standard error will be smaller for prediction of a mean response than for prediction of a single observation.

The Standard Error for the estimated mean response (i.e. $\overline{Y}$) is:

$$SE[\hat{\mu}\{Y|X\}] = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{(n-1)s_X^2}\right)},\ df = \text{n} - 2$$

The Standard Error for the estimated predicted value (i.e. $\mathbf{Y}$) is:

$$SE[\text{Pred}\{Y|X\}] = \sqrt{\hat{\sigma}^2 + SE\left[\hat{\mu}\{Y \mid X_o\}\right]^2}$$

$$= \hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{(n-1)s_X^2}\right)},\ df = \text{n-2}$$

The SE for an *estimated mean* only considers uncertainty about the position of the predicted subpopulation mean with respect to position of the true subpopulation mean.

The SE for a predicted value considers uncertainty about position of the *subpopulation mean* and in addition *of the future value* (which is $\hat{\sigma}^2$) with respect to the position of the true subpopulation mean. So:

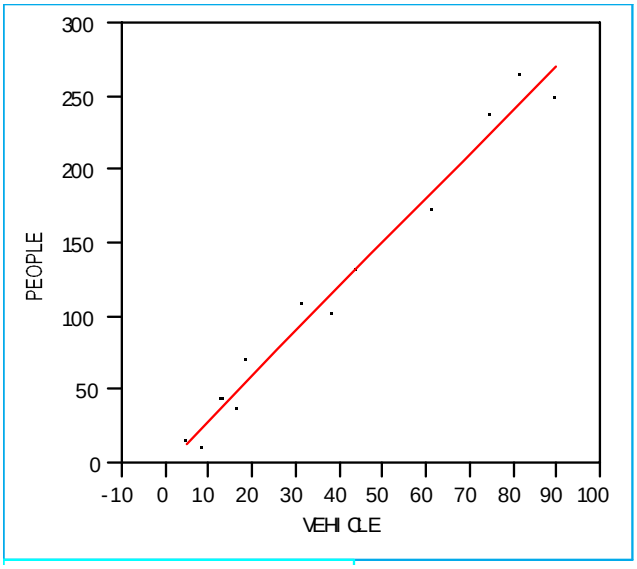
Prediction error = Sampling error (Y vs μ) + Estimation error ($\hat{\mu}$ vs μ)

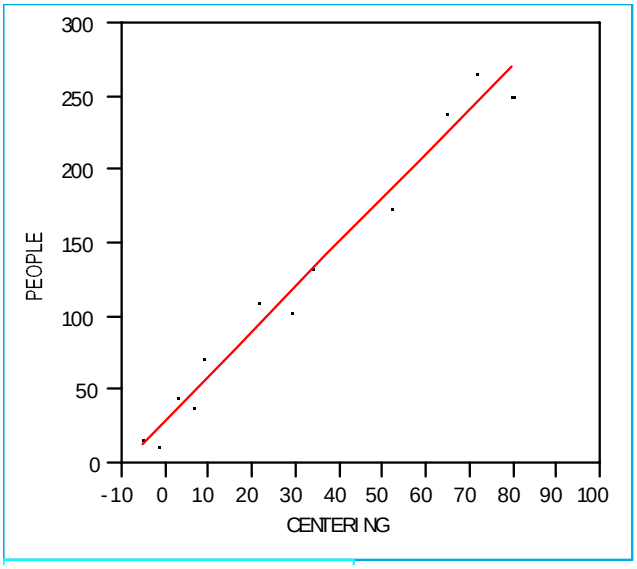
How do we estimate the above SE for a predicted subpopulation mean or a predicted individual value?

1. SE for predicted subpopulation mean and computation of a *confidence interval*:

For the Park example, a computer centering trick can be used for estimating the SE for the *predicted mean response* when Xo =10. The trick consists in computing the regression of Y vs X* = X – Xo, which centers X at Xo (i.e. Xo becomes the intercept).

Remember that the mean Y predicted by the regression model when Xo = 10 was 29.2. This value (29.2) will correspond to the intercept of the "centered" regression (i.e. Y vs X * = X – 10). The Standard Error for the mean Y when X = 10 can therefore be obtained directly from the computer output if we use the centering trick...





RESULTS FOR THE CENTERED REGRESSION

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	29.265335	**6.197901**	4.72	0.0008
CENTERING	3.0212969	0.1476	20.47	<.0001

95 % CI for $\hat{\mu}\ \{Y \mid X=10\}$ = $\{Y \mid X=10\} \pm t_{10\,(0.975)}$ X SE [$\hat{\mu}\ \{Y \mid 10\}$]

= 29.2 ± 2.228 X 6.198

= 29.2 ± 13.8 = 15.4, 43.0

2. SE for a predicted value and computation of the corresponding *prediction interval*:

95 % prediction interval for Pred $\{Y \mid X = 10\}$

SE [Pred$\{Y \mid X=10\}$] = $\sqrt{\hat{\sigma}^2 + SE\left[\hat{\mu}\{Y \mid X_o\}\right]^2}$

$= \sqrt{\left(216.4 + [6.198]^{2}\right)}$ (216.4 is Error MS in ANOVA Table)

$= 15.96$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	90688.256	90688.3	418.9975
Error	10	2164.411	**216.4**	Prob > F
C. Total	11	92852.667		<.0001

$\{Y \mid X=10\} \pm t_{10\,(0.975)}$ X SE [Pred$\{Y \mid X=10\}$] $= 29.2 \pm 2.228$ X 15.96

$= 29.2 \pm 35.6$

$= -6.4, 64.8$

More about the estimated mean at some value of X

1- Precision in estimating $\mu\ \{Y \mid X\}$ is *not* constant for all values of X: *precision is greater near the sample average,* simply because the value $X_o - \overline{X}$ that enters in calculation of the SE gets larger as Xo gets farther from the mean X.

$$SE[\hat{\mu}\{Y|X\}] = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\left(X_o - \overline{X}\right)^2}{(n-1)s_X^2}\right)}$$

If the goal of an experimental study is to predict a response at a specified X, the values of the explanatory variable should be chosen to fall on each side of X.

2- There is *compound uncertainty* associated with predictions of more than one mean. To maintain the expected nominal rate (α) when more than one prediction is needed, a Bonferroni correction [the confidence level for k predictions is 100 (1 - α/k)] can be used when only few predictions are made.

Alternatively, the *Workman-Hotelling* procedure can be used to compute *confidence bands. In 95 % of the cases* when they are constructed, the confidence bands will include the correct predicted response corresponding to a value X_i within the range of the observed Xs.

To build confidence bands, the t-multiplier in the formula for confidence intervals is replaced with a Scheffé multiplier based on a F-percentile with 2 and (n –2) *df*.

For a 95 % CB, the lower and upper bounds at each X is given by:

$$\hat{\mu}\{Y|X\} \pm \sqrt{\left(2 \times F_{2,n-2}[.95]\right)} \times SE[\,\hat{\mu}\{Y|X\}]$$

3- Prediction bands delimiting the range of many predicted individual Y *values* can be obtained by using the appropriate SE in the above *Workman-Hotelling* method.

<<Fig. 7.11 Sleuth>>

Calibration: or guessing the X that results in a given Y value

Interest is in guessing the X that produces a specific reponse. This is known as *calibration* or *inverse prediction.*

To get an inverse prediction, simply inverse the prediction relationship:

$$\Pr ed\{X \mid Y_o\} = \left(Y_o - \hat{\beta}_o\right) \Big/ \hat{\beta}_1$$

to estimate the X at which the *mean* of Y is Yo, the SE for calculation of the confidence interval is:

$$SE(\hat{X}) = \frac{SE(\hat{\mu}\{Y \mid \hat{X}\})}{\left|\hat{\beta}_1\right|}$$

<<Fig. 7.4 Sleuth>>

Correlation

Regression goal: determine functional relationship between a response variable Y and an explanatory variable X: the interest is in assessing how change in X affects Y. In regression, the roles of response and explanatory variables are easily distinguishable.

Correlation goal: determine the degree of *linear association* between 2 variables when there is no clear response variable.

The *sample correlation coefficient (r)* , also called the Pearson's product-moment correlation coefficient, is calculated as:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)/(n-1)}{s_x s_y}$$

where s_x and s_y are sample SD's.

r expresses the amount of linear association between two variables, and ranges between –1 and +1.

When r > 0, two variables are said to be *positively* correlated (or associated).
When r < 0, X and Y are *negatively* correlated
When r =0, X and Y are uncorrelated, i.e. do not show any linear association.

A hypothesis test involving r attempts to guess the value of ρ (rho, the population correlation coefficient), which is expressed as :

$$\rho = \frac{\sigma_{XY}}{\sqrt{\left(\sigma_X^2 \sigma_Y^2\right)}} = \frac{Covariance\,(X,Y)}{\sqrt{Variance(X)Variance(Y)}}$$

The test assume that samples for each variables have been drawn at random from a normal population, i.e that the 2 variables are distributed according to a bivariate normal distribution (Y normally distributed at each X; X normally distributed at each Y).

Correlation should only be used for statistical inferences when both the X and Y are random variables: never when X is fixed.

<<Fig Sokal and Rohlf>>

To test the null hypothesis that the correlation between two variables is zero (i.e. Ho: ρ = 0 vs. Ha: ρ ≠0), use a Student's t-test: $t = \frac{r}{se_r}$ which has n-2 *df*.

The standard error of r is: $se_r = \sqrt{\frac{1-r^2}{n-2}}$

However, for a null hypothesis that ρ ≠ 0, the distribution of

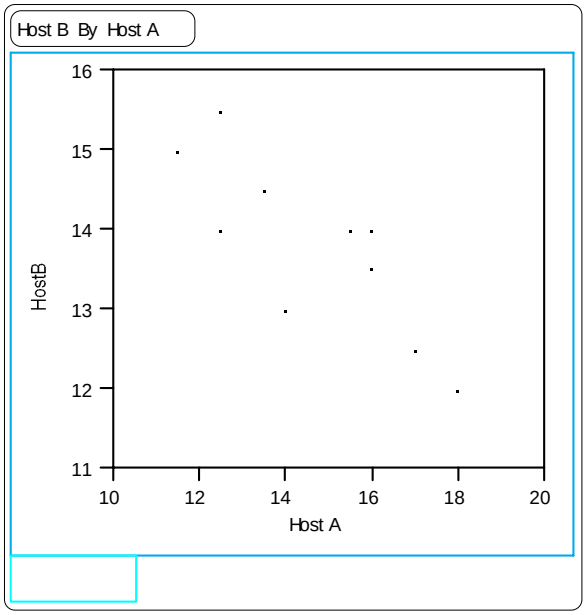
$$\frac{r-\rho_o}{\sqrt{\frac{1-r^2}{n-2}}}$$

is <u>not</u> a t distribution with n – 2 *df*.

When one is interested in testing whether ρ = *another value than zero*, a transformation of r (due to R.A Fisher) and the Z distribution must be used instead of the Students-t distribution (see Sokal and Rohlf for more details).

Example: Determine the correlation between average development time of aphid clones on two species of host-plant. (JMP: Analyze, Multivariate, Pairwise Correlations)

HostA	HostB
18	12
17	12.5
16	13.5
16	14
15.5	14
14	13
13.5	14.5
12.5	14
12.5	15.5
11.5	15



Correlations

Variable	Host A	Host B
Host A	1.0000	-0.8149
Host B	-0.8149	1.0000

Pairwise Correlations

Variable	by Variable	Correlation	Count	Signif Prob
Host B	Host A	-0.8149	10	0.0041