## Fitting Equations

Idea:

- The variable of interest (dependent variable, $y_i$) is hard to measure.
- There are "easy to measure" variables (predictor/ independent) that are related to the variable of interest, labeled $x_{1i}$ , $x_{2i}$,.....$x_{mi}$

We measure the y and the x's for a sample and use this sample to fit a model.

Once the model is fitted, we can then just measure the $x$'s, and get an estimate of $y$ without measuring it

Types of Equations

Simple Linear Equation:

$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$

Multiple Linear Equation:

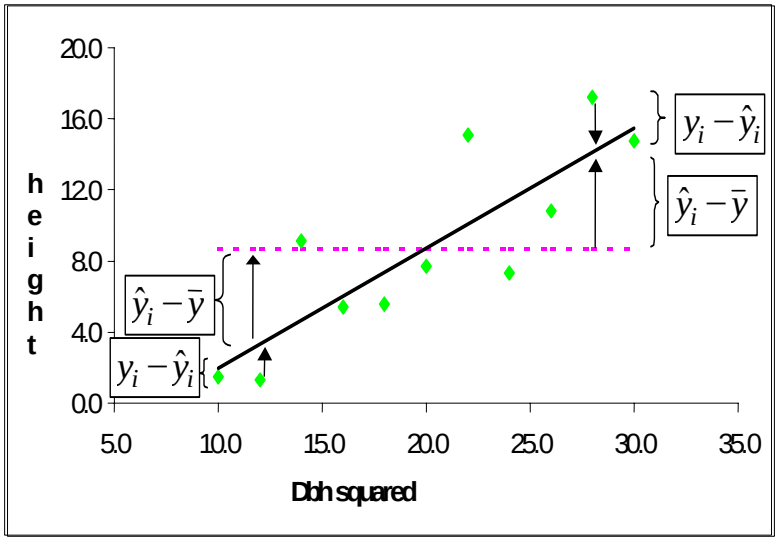
$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + \varepsilon_i$

Nonlinear Equation: takes many forms, for example:

$y_i = \beta_0 + \beta_1 x_{1i}^{\beta_2} x_{2i}^{\beta_3} + \varepsilon_i$



Example: Tree Height (m) – hard to measure; Dbh (diameter at 1.3 m above ground in cm) – easy to measure – use Dbh squared for a linear equation




$y_i - \bar{y}$ Difference between measured y and the mean of $y$

$y_i - \hat{y}_i$ Difference between measured $y$ and predicted $y$

$\hat{y}_i - \bar{y} = (y_i - \bar{y}) - (y_i - \hat{y}_i)$ Difference between predicted $y$ and mean of $y$

Objective:

Find estimates of $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ such that the sum of squared differences between measured $y_i$ and predicted $y_i$ (usually labeled as $\hat{y}_i$, values on the line or surface) is the smallest (*minimize* the sum of squared errors, called least squared error).

OR

Find estimates of $\beta_0, \beta_1, \beta_2 \ldots \beta_m$ such that the likelihood (probability) of getting these $y$ values is the largest (*maximize* the likelihood).

Finding the minimum of sum of squared errors is often easier. In some cases, they lead to the same estimates of parameters.



Least Squares Solution: Finding the Set of Coefficients that Minimizes the Sum of Squared Errors

To find the estimated coefficients that minimizes SSE for a particular set of sample data and a particular equation (form and variables):

1. Define the sum of squared errors (SSE) in terms of the measured minus the predicted $y$'s (the errors);
2. Take partial derivatives of the SSE equation with respect to each coefficient
3. Set these equal to zero (for the minimum) and solve for all of the equations (solve the set of equations using algebra or linear algebra).

## Simple Linear Regression

- There is only one x variable
- There will be two coefficients

The estimated intercept is found by:

$$b_0 = \bar{y} - b_1\bar{x}$$

And the estimated slope is found by:

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s^2{}_{xy}(n-1)}{s_x{}^2(n-1)} = \frac{SPxy}{SSx}$$

Where *SPxy* refers to the corrected sum of cross products for *x* and *y*; *SSx* refers to the corrected sum of squares for *x* [Class example]



*Properties of $b_0$ and $b_1$*

$b_0$ and $b_1$ are least squares estimates of $\beta_0$ and $\beta_1$. **Under assumptions** concerning the error term and sampling/ measurements, these are:

- Unbiased estimates; given many estimates of the slope and intercept for all possible samples, the average of the sample estimates will equal the true values
- The variability of these estimates from sample to sample can be estimated from the single sample; these estimated variances will be unbiased estimates of the true variances (and standard errors)
- The estimated intercept and slope will be the most precise (most efficient with the lowest variances) estimates possible (called "Best")
- These will also be the maximum likelihood estimates of the intercept and slope

Assumptions of SLR

Once coefficients are obtained, we must **check the assumptions of** SLR. Assumptions must be met to:

- obtain the desired characteristics
- assess goodness of fit (i.e., how well the regression line fits the sample data)
- test significance of the regression and other hypotheses
- calculate confidence intervals and test hypothesis for the true coefficients (population)
- calculate confidence intervals for mean predicted *y* value given a set of *x* value (i.e. for the predicted y given a particular value of the *x*)

Need good estimates (unbiased or at least consistent) of the standard errors of coefficients and a known probability distribution to test hypotheses and calculate confidence intervals.



*Checking the following assumptions using residual Plots*

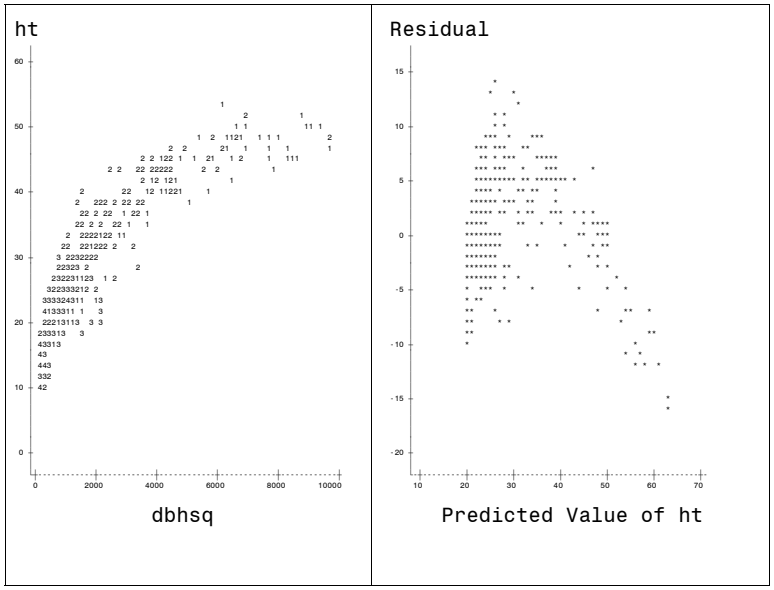
1. a linear relationship between the *y* and the *x*;
2. equal variance of errors across the range of the y variables; and
3. independence of errors (independent observations), not related in time or in space.

A residual plot shows the residual (i.e., $y_i - \hat{y}_i$) as the y-axis and the predicted value ($\hat{y}_i$) as the x-axis.

Residual plots can also indicate unusual points (outliers) that may be measurement errors, transcription errors, etc.

*Examples of Residual Plots Indicating Failures to Meet Assumptions:*

1. *The relationship between the x's and y is linear*. If not met, the residual plot and the plot of y vs. x will show a curved line: [CRITICAL ASSUMPTION!!]




Result: If this assumption is not met: the regression line does not fit the data well; biased estimates of coefficients and standard errors of the coefficients will occur



2. *The variance of the y values must be the same for every one of the x values.* If not met, the spread around the line will not be even.

Result: If this assumption is not met, the estimated coefficients (slopes and intercept) will be unbiased, but the estimates of the standard deviation of these coefficients will be biased.

∴ we cannot calculate CI nor test the significance of the x variable. However, estimates of the coefficients of the regression line and goodness of fit are still unbiased

3. *Each observation (i.e., $x_i$ and $y_i$) must be independent of all other observations.* In this case, we produce a different residual plot, where the residuals are on the y-axis as before, but the x-axis is the variable that is thought to produce the dependencies (e.g., time). If not met, this revised residual plot will show a trend, indicating the residuals are not independent.

Result: If this assumption is not met, the <u>estimated coefficients (slopes and intercept) will be unbiased</u>, but the <u>estimates of the standard deviation of these coefficients will be biased.</u>

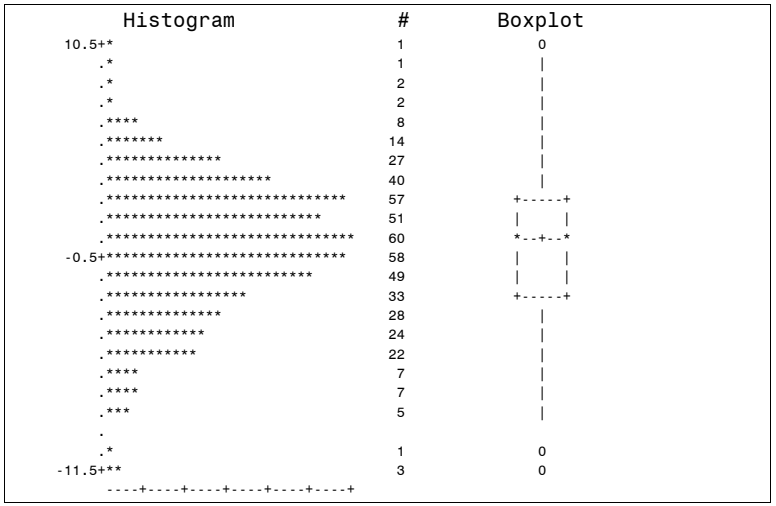
∴ we cannot calculate CI nor test the significance of the x variable. However, estimates of the coefficients of the regression line and goodness of fit are still unbiased



*Normality Histogram or Plot*

A fourth assumption of the SLR is:

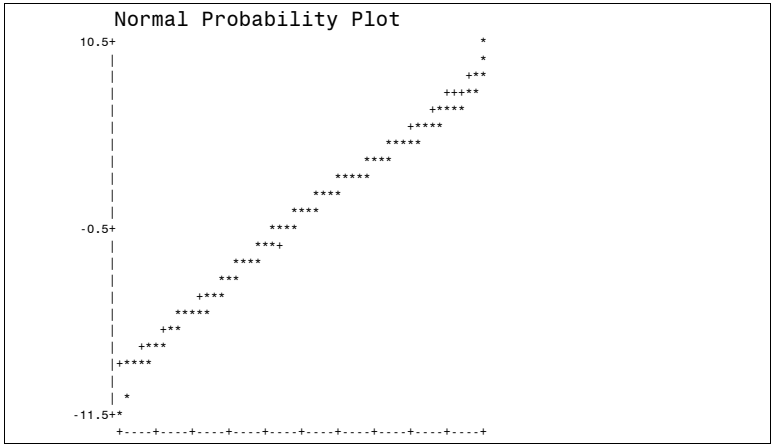
4. *The y values must be normally distributed for each of the x values.* A histogram of the errors, and/or a normality plot can be used to check this, as well as tests of normality




HO: residuals are normal H1: residuals are not normal

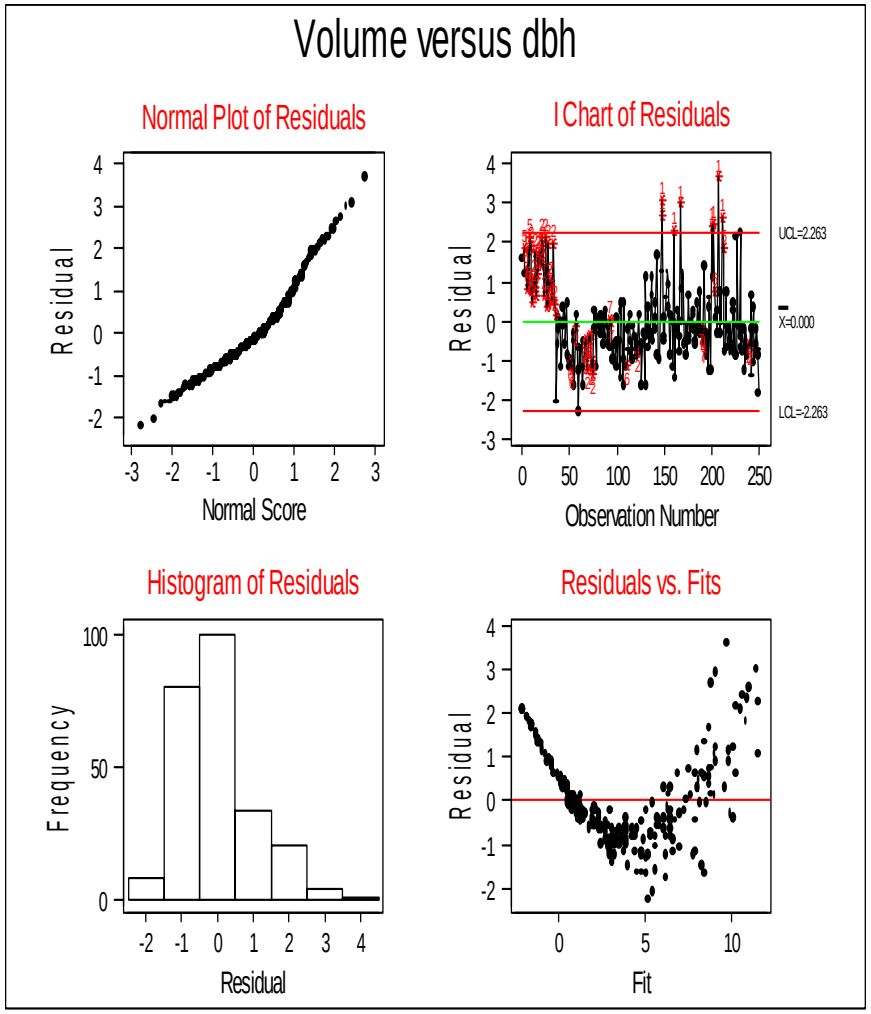
Tests for Normality

| Test | | --Statistic--- | -----p Value------ | |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.991021 | Pr < W | 0.0039 |
| Kolmogorov-Smirnov | D | 0.039181 | Pr > D | 0.0617 |
| Cramer-von Mises | W-Sq | 0.19362 | Pr > W-Sq | 0.0066 |
| Anderson-Darling | A-Sq | 1.193086 | Pr > A-Sq | <0.0050 |




Result: We cannot calculate CI nor test the significance of the x variable, since we do not know what probabilities to use. Also, estimated coefficients are no longer equal to the maximum likelihood solution.





*Measurements and Sampling Assumptions*

The remaining assumptions are based on the measurements and collection of the sampling data.

5. *The x values are measured without error (i.e., the x values are fixed).* This can only be known if the process of collecting the data is known. For example, if tree diameters are very precisely measured, there will be little error. If this assumption is not met, the estimated coefficients (slopes and intercept) and their variances will be biased, since the x values are varying.

6. *The y values are randomly selected for value of the x variables (i.e., for each x value, a list of all possible y values is made, and some are randomly selected).* Often, the observations will be gathered using systematic sampling (grid across the land area). This does not strictly meet this assumption. Also, more complex sampling design such as multistage sampling (sampling large units and sampling smaller units within the large units), this assumption is not met. If the equation is "correct", then this does not cause problems. If not, the estimated equation will be biased.



Transformations

*Common Transformations*

- Powers $x^3$, $x^{0.5}$, etc. for relationships that look nonlinear
- log10, loge also for relationships that look nonlinear, or when the variances of y are not equal around the line
- Sin-1 [arcsine] when the dependent variable is a proportion.
- Rank transformation: for non-normal data
  - Sort the y variable
  - Assign a rank to each variable from 1 to n
  - Transform the rank to normal (e.g., Blom Transformation)

  PROBLEM: loose some of the information in the original data
- Try to transform *x* first and leave $y_i$ = variable of interest; however, this is not always possible.

Use graphs to help choose transformations

Outliers: Unusual Points

Check for points that are quite different from the others on:

- Graph of *y* versus *x*
- Residual plot

**Do not delete** the point as it MAY BE VALID! Check:

- Is this a measurement error? E.g., a tree height of 100 m is very unlikely
- Is a transcription error? E.g. for adult person, a weight of 20 lbs was entered rather than 200 lbs.
- Is there something very unusual about this point? e.g., a bird has a short beak, because it was damaged.

Try to fix the observation. If it is very different than the others, or you know there is a measurement error that cannot be fixed, then **delete it and indicate this in your research report**.

On the residual plot, an outlier CAN occur if the model is not correct – may need a transformation of the variable(s), or an important variable is missing



Measures of Goodness of Fit

How well does the regression fit the sample data?

- For simple linear regression, a graph of the original data with the fitted line marked on the graph indicates how well the line fits the data [not possible with MLR]
- Two measures commonly used: coefficient of determination ($r^2$) and standard error of the estimate($SE_E$).

To calculate $r^2$ and $SE_E$, first, calculate the SSE (this is what was minimized):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

The sum of squared differences between the measured and estimated *y*'s.

Calculate the sum of squares for *y*:

$$SSy = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n = s_y^{\,2}(n-1)$$

The sum of squared difference between the measured y and the mean of y-measures. NOTE: In some texts, this is called the sum of squares total.

Calculate the sum of squares regression:

$$SSreg = \sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2 = b_1 SPxy = SSy - SSE$$

The sum of squared differences between the mean of y-measures and the predicted *y*'s from the fitted equation. Also, is the sum of squares for *y* – the sum of squared errors.

Then: $$r^2 = \frac{SSy - SSE}{SSy} = 1 - \frac{SSE}{SSy} = \frac{SSreg}{SSy}$$

- SSE, SSY are based on *y*'s used in the equation – will not be in original units if *y* was transformed
- $r^2$ = coefficient of determination; proportion of variance of *y*, accounted for by the regression using *x*
- Is the square of the correlation between *x* and *y*
- O (very poor – horizontal surface representing no relationship between y and x's) to 1 (perfect fit – surface passes through the data)



And: $$SE_E = \sqrt{\frac{SSE}{n-2}}$$

- SSE is based on *y*'s used in the equation – will not be in original units if *y* was transformed
- $SE_E$ - standard error of the estimate; in same units as y
- Under normality of the errors:
  - $\pm 1\ SE_E \cong 68\%$ of sample observations
  - $\pm 2\ SE_E \cong 95\%$ of sample observations
  - Want low SEE

y-variable was transformed: Can calculate estimates of these for the original y-variable unit, called $I^2$ (Fit Index) and estimated standard error of the estimate ($SE_E$'), in order to compare to $r^2$ and $SE_E$ of other equations where the *y* was not transformed.

$$I^2 = 1 - SSE/SSY$$

- where SSE, SSY are in original units. NOTE must "back-transform" the predicted *y*'s to calculate the SSE in original units.
- Does not have the same properties as $r^2$, however:
  - it can be less than 0
  - it is not the square of the correlation between the y (in original units) and the *x* used in the equation.

Estimated standard error of the estimate ($SE_E$') , when the dependent variable, *y*, has been transformed:

$$SE_E' = \sqrt{\frac{SSE(original\ units)}{n-2}}$$

- $SE_E$' - standard error of the estimate ; in same units as original units for the dependent variable
- want low $SE_E$' [Class example]



Estimated Variances, Confidence Intervals and Hypothesis Tests

*Testing Whether the Regression is Significant*

Does knowledge of x improve the estimate of the mean of y? Or is it a flat surface, which means we should just use the mean of y as an estimate of mean y for any x?

SSE/ (*n*-2):

- Called the Mean squared error, as would be the average of the squared error if we divided by *n*.
- Instead, we divide by *n*-2. Why? The degrees of freedom are *n*-2; *n* observations with two statistics estimated from these, $b_0$ and $b_1$
- Under the assumptions of SLR, is an unbiased estimated of the true variance of the error terms (error variance)

SSR/1:

- Called the Mean Square Regression
- Degrees of Freedom=1: 1 *x*-variable
- Under the assumptions of SLR, this is an estimate the error variance PLUS a term of variance explained by the regression using *x*.

H0: Regression is not significant

H1: Regression is significant

Same as:

H0: $\beta_1 = 0$ [true slope is zero meaning no relationship with x]

H1: $\beta_1 \neq 0$ [slope is positive or negative, not zero]

This can be tested using an F-test, as it is the ratio of two variances, or with a t-test since we are only testing one coefficient (more on this later)

Using an F test statistic:

$$F = \frac{SSreg/1}{SSE/(n-2)} = \frac{MSreg}{MSE}$$

- Under H0, this follows an F distribution for a 1- $\alpha/2$ percentile with 1 and $n$-2 degrees of freedom.
- If the F for the fitted equation is larger than the F from the table, we reject H0 (not likely true). The regression is significant, in that the true slope is likely not equal to zero.



Information for the F-test is often shown as an Analysis of Variance Table:

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | *1* | *SSreg* | *MSreg*= *SSreg*/1 | *F*= *MSreg*/*MSE* | Prob F> $F_{(1,n-2,1-\alpha)}$ |
| Residual | *n*-2 | *SSE* | *MSE*= *SSE*/(*n*-2) | | |
| Total | *n*-1 | *SSy* | | | |

[Class example and explanation of the p-value]

*Estimated Standard Errors for the Slope and Intercept*

Under the assumptions, we can obtain an unbiased estimated of the standard errors for the slope and for the intercept [measure of how these would vary among different sample sets], using the one set of sample data.

$$s_{b_0} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{SSx}\right)} = \sqrt{\frac{MSE \times \sum_{i=1}^{n} x_i^{\,2}}{n \times SSx}}$$

$$s_{b_1} = \sqrt{\frac{MSE}{SSx}}$$

*Confidence Intervals for the True Slope and Intercept*

Under the assumptions, confidence intervals can be calculated as:

For $\beta_o$: $b_0 \pm t_{1-\alpha/2,n-2} \times s_{b_0}$

For $\beta_1$: $b_1 \pm t_{1-\alpha/2,n-2} \times s_{b_1}$

[class example]



*Hypothesis Tests for the True Slope and Intercept*

H0: $\beta_1 = c$ [true slope is equal to the constant, c]

H1: $\beta_1 \neq c$ [true slope differs from the constant c]

Test statistic:

$$t = \frac{b_1 - c}{s_{b_1}}$$

Under H0, this is distributed as a t value of $t_c = t_{n-2,\ 1-\alpha/2}$.

Reject $H_o$ if $|t| > t_c$.

- The procedure is similar for testing the true intercept for a particular value
- It is possible to do one-sided hypotheses also, where the alternative is that the true parameter (slope or intercept) is greater than (or less than) a specified constant c. MUST be careful with the $t_c$ as this is different.

[class example]

*Confidence Interval for the True Mean of y given a particular x value*

For the mean of all possible y-values given a particular value of x ($\mu_y|x_h$):

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$

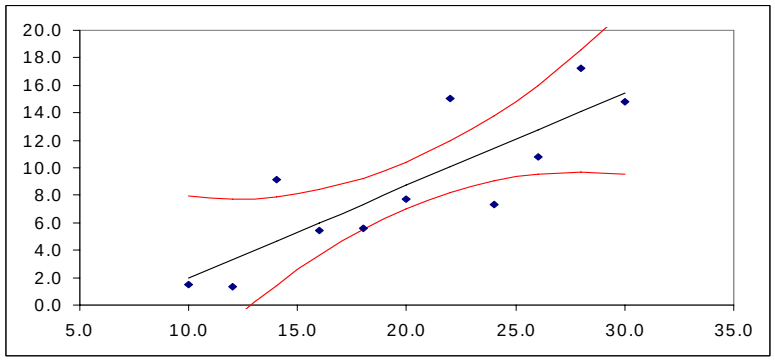
where

$$\hat{y} \mid x_h = b_0 + b_1 x_h$$

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

*Confidence Bands*

Plot of the confidence intervals for the mean of y for several x-values. Will appear as:






*Prediction Interval for 1 or more y-values given a particular x value*

For one possible new y-value given a particular value of x:

$$\hat{y}_{(new)} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}(new)|x_h}$$

Where

$$\hat{y}_{(new)} \mid x_h = b_0 + b_1 x_h$$

$$s_{\hat{y}(new)|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

For the average of *g* new possible y-values given a particular value of x:

$$\hat{y}_{(new)} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}(newg)|x_h}$$

where

$$\hat{y}_{(new)} \mid x_h = b_0 + b_1 x_h$$

$$s_{\hat{y}(newg)|x_h} = \sqrt{MSE\left(\frac{1}{g} + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$

[class example]

Selecting Among Alternative Models

*Process to Fit an Equation using Least Squares*

Steps:

1. Sample data are needed, on which the dependent variable and all explanatory (independent) variables are measured.
2. Make any transformations that are needed to meet the most critical assumption: The relationship between *y* and *x* is linear.

   Example: volume = $\beta_0 + \beta_1$ dbh$^2$ may be linear whereas volume versus dbh is not. Use $y_i$ = volume , $x_i$ = dbh$^2$.
3. Fit the equation to minimize the sum of squared error.
4. Check Assumptions. If not met, go back to Step 2.
5. If assumptions are met, then interpret the results.
   - Is the regression significant?
   - What is the $r^2$? What is the $SE_E$?
   - Plot the fitted equation over the plot of *y* versus *x*.



*For a number of models, select based on:*

1. Meeting assumptions: If an equation does not meet the assumption of a linear relationship, it is not a candidate model
2. Compare the fit statistics. Select higher $r^2$ (or $I^2$), and lower $SE_E$ (or $SE_E$')
3. Reject any models where the regression is not significant, since this model is no better than just using the mean of *y* as the predicted value.
4. Select a model that is biologically tractable. A simpler model is generally preferred, unless there are practical/biological reasons to select the more complex model
5. Consider the cost of using the model

[class example]

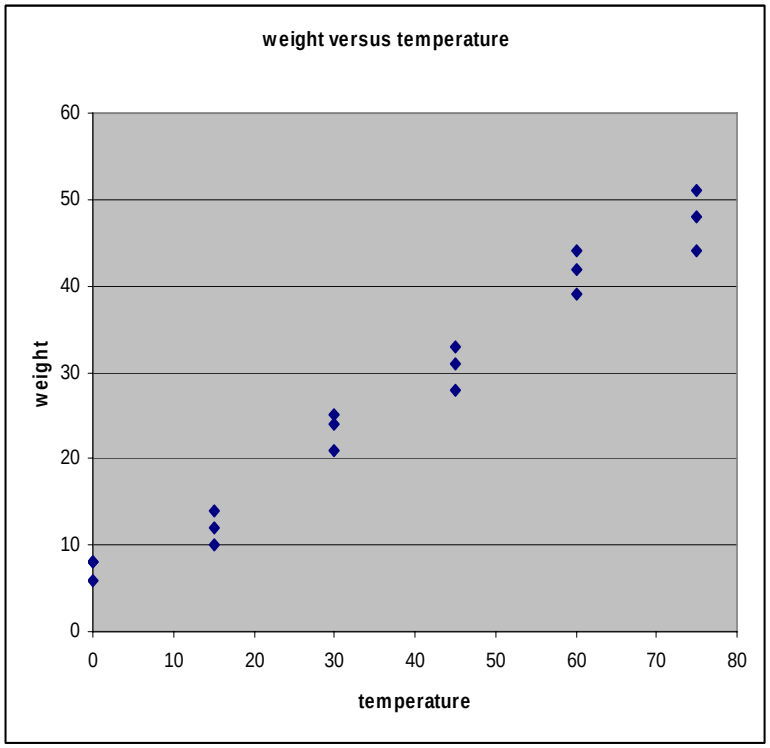
## Simple Linear Regression Example

| Temperature (x) | Weight (y) | Weight (y) | Weight (y) |
|---|---|---|---|
| 0 | 8 | 6 | 8 |
| 15 | 12 | 10 | 14 |
| 30 | 25 | 21 | 24 |
| 45 | 31 | 33 | 28 |
| 60 | 44 | 39 | 42 |
| 75 | 48 | 51 | 44 |

| Observation | temp | weight |
|---|---|---|
| 1 | 0 | 8 |
| 2 | 0 | 6 |
| 3 | 0 | 8 |
| 4 | 15 | 12 |
| 5 | 15 | 10 |
| 6 | 15 | 14 |
| 7 | 30 | 25 |
| 8 | 30 | 21 |

Et cetera…





| Obs. | temp | weight | x-diff | x-diff. sq. |
|---|---|---|---|---|
| 1 | 0 | 8 | -37.50 | 1406.25 |
| 2 | 0 | 6 | -37.50 | 1406.25 |
| 3 | 0 | 8 | -37.50 | 1406.25 |
| 4 | 15 | 12 | -22.50 | 506.25 |

Et cetera

| | | |
|---|---|---|
| mean | 37.5 | 27.11 |

SSX=11,812.5 SSY=3,911.8 SPXY=6,705.0

$$b_1 = \frac{SPxy}{SSx} \qquad b_0 = \bar{y} - b_1 \times \bar{x}$$

| | |
|---|---|
| b1: | 0.567619 |
| b0: | 5.825397 |

NOTE: calculate b1 first, since this is needed to calculate b0.



From these, the residuals (errors) for the equation, and the sum of squared error (SSE) were calculated:

| Obs. | weight | y-pred | residual | residual sq. |
|---|---|---|---|---|
| 1 | 8 | 5.83 | 2.17 | 4.73 |
| 2 | 6 | 5.83 | 0.17 | 0.03 |
| 3 | 8 | 5.83 | 2.17 | 4.73 |
| 4 | 12 | 14.34 | -2.34 | 5.47 |

Et cetera

| | |
|---|---|
| SSE: | 105.89 |

And SSR=SSY-SSE=3805.89

**ANOVA**

| **Source** | **df** | **SS** | **MS** |
|---|---|---|---|
| **Model** | 1 | 3805.89 | 3805.89 |
| **Error** | 18-2=16 | 105.89 | 6.62 |
| **Total** | 18-1=17 | 3911.78 | |

F=575.06  with p=0.00 (very small)

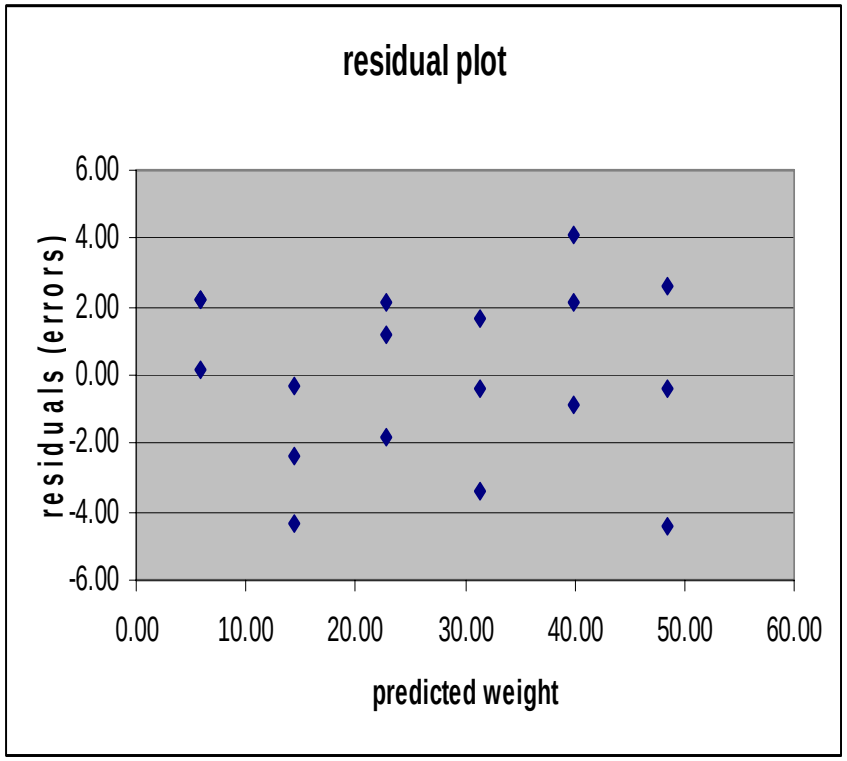
In excel use:  = fdist(x,df1,df2) to obtain a "p-value"

| | |
|---|---|
| **$r^2$:** | 0.97 |
| **Root MSE Or** | |
| $SE_E$ **:** | 2.57 |

**BUT:  Before interpreting the ANOVA table**, Are assumptions met?

**If assumptions were not met, we would have to make some transformations and start over again!**






Linear?

Equal variance?

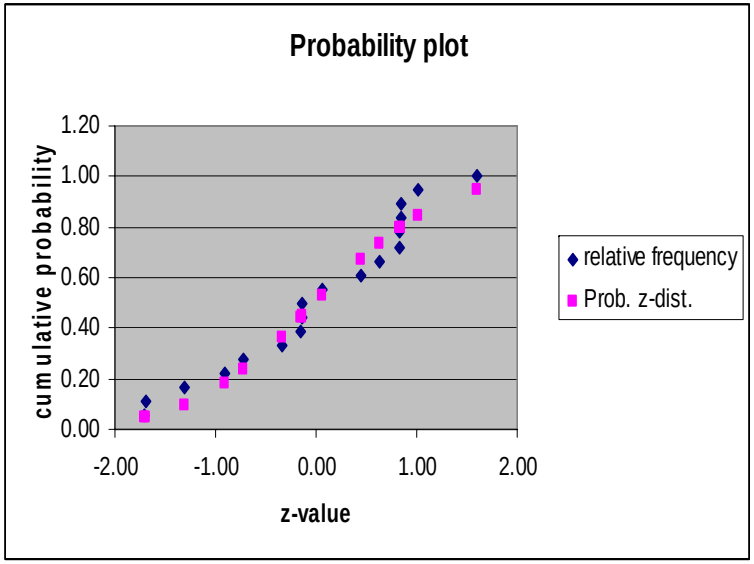
Independent observations? [need another plot – residuals versus time or space, that cause dependencies]

Normality plot:

| Obs. | sorted resids | Stand. resids | Rel. Freq. | Prob. z-dist. |
|---|---|---|---|---|
| 1 | -4.40 | -1.71 | 0.06 | 0.04 |
| 2 | -4.34 | -1.69 | 0.11 | 0.05 |
| 3 | -3.37 | -1.31 | 0.17 | 0.10 |
| 4 | -2.34 | -0.91 | 0.22 | 0.18 |
| 5 | -1.85 | -0.72 | 0.28 | 0.24 |
| 6 | -0.88 | -0.34 | 0.33 | 0.37 |
| 7 | -0.40 | -0.15 | 0.39 | 0.44 |
| 8 | -0.37 | -0.14 | 0.44 | 0.44 |
| 9 | -0.34 | -0.13 | 0.50 | 0.45 |

Etc.






Questions:

1. Are the assumptions of simple linear regression met? Evidence?

2. If so, interpret if this is a good equation based on goodness of it measures.

3. Is the regression significant?

For 95% confidence intervals for b0 and b1, would also need estimated standard errors:

$$s_{b_0} = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{SSx}\right)} = \sqrt{6.62 \times \left(\frac{1}{18} + \frac{37.5^2}{11812.50}\right)} = 1.075$$

$$s_{b_1} = \sqrt{\frac{MSE}{SSx}} = \sqrt{\frac{6.62}{11812.50}} = 0.0237$$

The t-value for 16 degrees of freedom and the 0.975 percentile is 2.12 (=tinv(0.05,16) in EXCEL)

$$b_0 \pm t_{1-\alpha/2,n-2} \times s_{b_0}$$

For $\beta_o$: $5.825 \pm 2.120 \times 1.075$

$$b_1 \pm t_{1-\alpha/2,n-2} \times s_{b_1}$$

For $\beta_1$: $0.568 \pm 2.120 \times 0.0237$



| | Est. Coeff | St. Error |
|---|---|---|
| **For b0:** | 5.825396825 | 1.074973559 |
| **For b1:** | 0.567619048 | 0.023670139 |

| **CI:** | **b0** | **b1** |
|---|---|---|
| t(0.975,16) | 2.12 | 2.12 |
| lower | 3.54645288 | 0.517438353 |
| upper | 8.104340771 | 0.617799742 |

Question: Could the real intercept be equal to 0?

Given a temperature of 22, what is the estimated average weight (predicted value) and a 95% confidence interval for this estimate?

$$\hat{y} \mid x_h = b_0 + b_1 x_h$$
$$\hat{y} \mid (x_h = 22) = 5.825 + 0.568 \times 22 = 18.313$$

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$
$$s_{\hat{y}|x_h} = \sqrt{6.62 \times \left(\frac{1}{18} + \frac{(22 - 37.5)^2}{11812.50}\right)} = 0.709$$

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$
$$18.313 - 2.12 \times 0.709 = 16.810$$
$$18.313 + 2.12 \times 0.709 = 19.816$$



Given a temperature of 22, what is the estimated weight for any new observation, and a 95% confidence interval for this estimate?

$$\hat{y} \mid x_h = b_0 + b_1 x_h$$
$$\hat{y} \mid (x_h = 22) = 5.825 + 0.568 \times 22 = 18.313$$

$$s_{\hat{y}|x_h} = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{SSx}\right)}$$
$$s_{\hat{y}|x_h} = \sqrt{6.62 \times \left(1 + \frac{1}{18} + \frac{(22 - 37.5)^2}{11812.50}\right)} = 2.669$$

$$\hat{y} \mid x_h \pm t_{n-2,1-\alpha/2} \times s_{\hat{y}|x_h}$$
$$18.313 - 2.12 \times 2.669 = 12.66$$
$$18.313 + 2.12 \times 2.669 = 23.97$$

## Multiple Linear Regression (MLR)

Population: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{mi} + \varepsilon_i$

Sample: $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{mi} + e_i$

$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_m x_{mi}$ $\quad e_i = y_i - \hat{y}_i$

$\beta_o$ is the y intercept parameter

$\beta_1, \beta_2, \beta_3, \ldots, \beta_m$ are slope parameters

$x_{1i}, x_{2i}, x_{3i} \ldots x_{mi}$ independent variables

$\varepsilon_i$ - is the error term or residual

- is the variation in the dependent variable (the *y*) which is not accounted for by the independent variables (the *x*'s).

For any fitted equation (we have the estimated parameters), we can get the *estimated average for the dependent variable,* for any set of *x*'s. This will be the "predicted" value for y, which is the estimated average of *y*, given the particular values for the *x* variables. NOTE: In text by Neter et al. $p=m+1$. This is not be confused with the p-value indicating significance in hypothesis tests.



For example:

Predicted log10(vol) = - 4.2 + 2.1 X log10(dbh) + 1.1 X log10(height)

where $b_o$= -4.2; $b_1$= 2.1 ; $b_1$= 1.1 estimated by finding the least squared error solution.

Using this equation for dbh =30 cm, height=28m, logten(dbh) =1.48, logten(height) =1.45; logten(vol) = 0.503. $\therefore$ **volume ($m^3$) = 3.184**. This represents the estimated average volume for trees with dbh=30 cm and height=28 m.

Note: This equation is originally a nonlinear equation:

$$vol = a \times dbh^b \times ht^c \varepsilon$$

Which was transformed to a linear equation using logarithms:

$$\log 10(vol) = \log 10(a) + b \log 10(dbh) + c \log 10(ht) + \log 10\varepsilon$$

And this was fitted using multiple linear regression

For the observations in the sample data used to fit the regression, we can also get an estimate of the error (we have measured volume).

If the measured volume for this tree was 3.000 **$m^3$,** or 0.477 in log10 units:

$$error = y_i - \hat{y}_i = 0.477 - 0.503 = -0.026$$

For the fitted equation using log10 units. In original units, the estimated error is 3.000-3.184= - 0.184

NOTE: This is not simply the antilog of -0.026.



Finding the Set of Coefficients that Minimizes the Sum of Squared Errors

- Same process as for SLR: Find the set of coefficients that results in the minimum SSE, just that there are more parameters, therefore more partial derivative equations and more equations
  - E.g., with 3 x-variables, there will be 4 coefficients (intercept plus 3 slopes) so four equations
- For linear models, there will be one unique mathematical solution.
- For nonlinear models, this is not possible and we must search to find a solution

Using the criterion of finding the maximum likelihood (probability) rather than the minimum SSE, we would need to search for a solution, even for linear models (covered in other courses, e.g., FRST 530).

Least Squares Method for MLR:

Find the set of estimated parameters (coefficients) that minimize sum of squared errors

$$\min(SSE) = \min(\sum_{i=1}^{n} e_i^2)$$

$$= \min\left( \sum_{i=1}^{n} \left( y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{mi}) \right)^2 \right)$$

Take partial derivatives with respect to each of the coefficients, set them equal to zero and solve.

For three x-variables we obtain:

$$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - b_3 \bar{x}_3$$

$$b_1 = \frac{SPx_1 y}{SSx_1} - b_2 \frac{SPx_1 x_2}{SSx_1} - b_3 \frac{SPx_1 x_3}{SSx_1}$$

$$b_2 = \frac{SPx_2 y}{SSx_2} - b_1 \frac{SPx_1 x_2}{SSx_2} - b_3 \frac{SPx_2 x_3}{SSx_2}$$

$$b_3 = \frac{SPx_3 y}{SSx_3} - b_1 \frac{SPx_1 x_3}{SSx_3} - b_2 \frac{SPx_2 x_3}{SSx_3}$$



Where SP= indicates sum of products between two variables, for example for $y$ with $x_1$:

$$SPx_1 y = \sum_{i=1}^{n} (y_i - \bar{y})(x_{1i} - \bar{x}_1)$$

$$= \sum_{i=1}^{n} y_i x_{1i} - \frac{\left( \sum_{i=1}^{n} x_{1i} \right) \left( \sum_{i=1}^{n} y_i \right)}{n} = s^2{}_{x_1 y}(n-1)$$

And SS indicates sums of squares for one variable, for example for $x_1$:

$$SSx_1 = \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 = \sum_{i=1}^{n} x_{1i}{}^2 - \frac{\left( \sum_{i=1}^{n} x_{1i} \right)^2}{n} = s^2{}_{x_1}(n-1)$$

Properties of a least squares regression "surface":

1. Always passes through $(\bar{x}_1, \bar{x}_2, \bar{x}_3, ..., \bar{x}_m, \bar{y})$
2. Sum of residuals is zero, i.e., $\Sigma e_i = 0$
3. SSE the least possible (least squares)
4. The slope for a particular x-variable is AFFECTED by correlation with other x-variables: CANNOT interpret the slope for a particular x-variable, UNLESS it has zero correlation with all other x-variables (or nearly zero if correlation is estimated from a sample).



Meeting Assumptions of MLR

Once coefficients are obtained, we must **check the assumptions of MLR** before we can:

- assess goodness of fit (i.e., how well the regression line fits the sample data)
- test significance of the regression
- calculate confidence intervals and test hypothesis

For these test to be valid, **assumptions of MLR concerning the observations and the errors (residuals) must be met.**

Residual Plots

Assumptions of:

1. The relationship between the x's and y is linear VERY IMPORTANT!
2. The variances of the y values must be the same for every combination of the x values.
3. Each observation (i.e., $x_i$'s and $y_i$) must be independent of all other observations.

can be visually checked by using **RESIDUAL PLOTS**

A residual plot shows the residual (i.e., $y_i - \hat{y}_i$) as the y-axis and the predicted value ($\hat{y}_i$) as the x-axis. For the indepence assumption, the x-axis is time or space that explains the dependence of the data.

THIS IS THE SAME as for SLR. Look for problems as with SLR. The effects of failing to meet a particular assumption are the same as for SLR

What is different? Since there are many x variables, it will be harder to decide what to do to fix any problems.



Normality Histogram or Plot

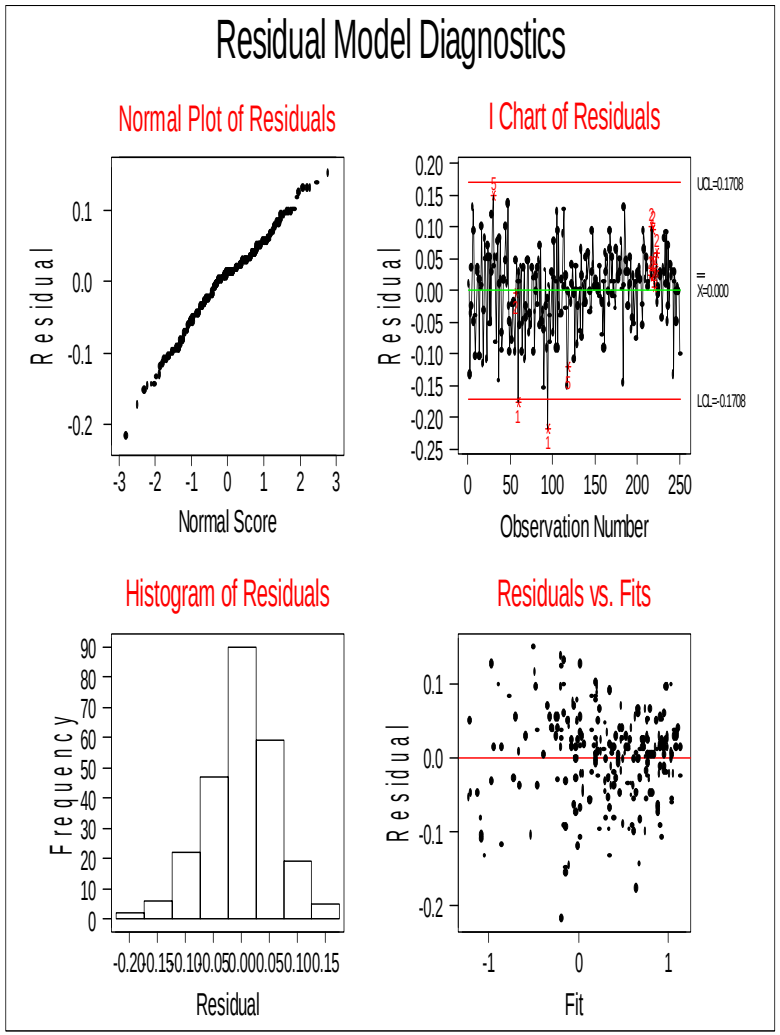
A fourth assumption of the MLR is:

4. The y values must be normally distributed for each combination of x values.

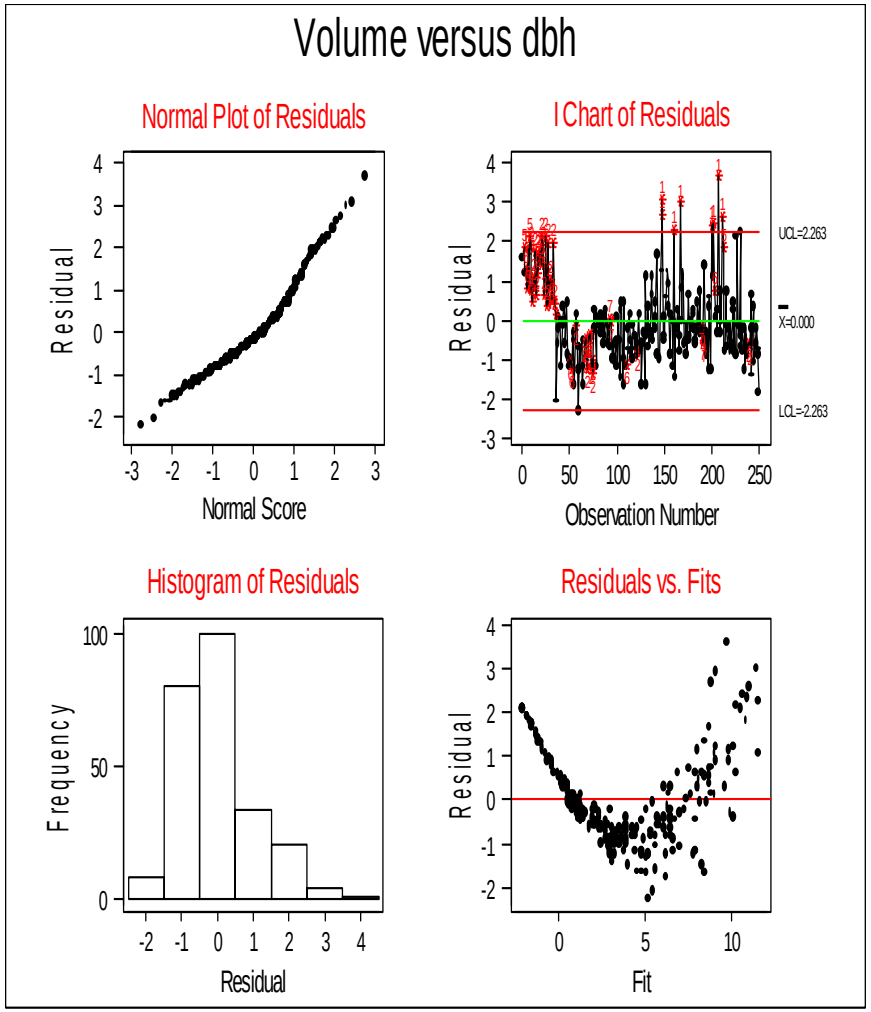
A histogram of the errors, and/or a normality plot can be used to check this, as well as tests of normality as with SLR. Failure to meet these assumptions will result in same problems as with SLR.

Example: Linear relationship met, equal variance, no evidence of trend with observation number (independence may be met). Also, normal distribution met.

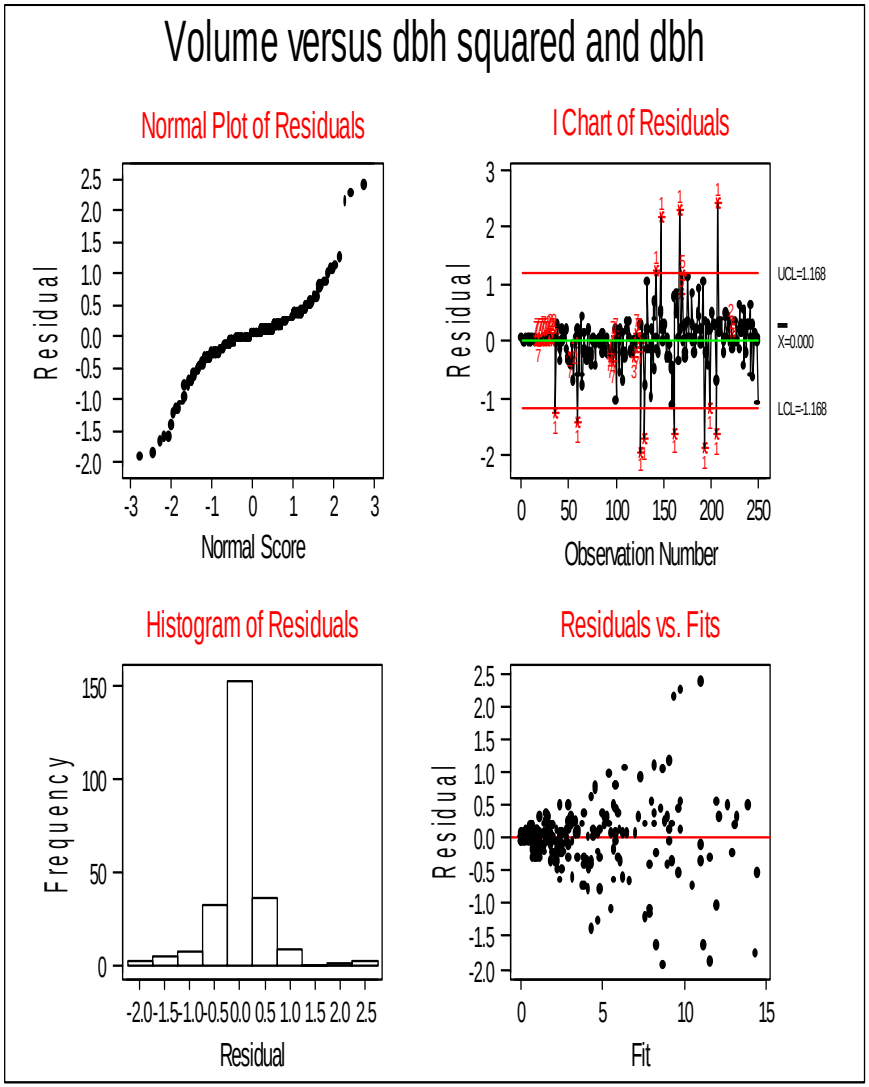
Logvol=f(dbh,logdbh)




Linear relationship assumption not met



Variances are not equal






*Measurements and Sampling Assumptions*

The remaining assumptions of MLR are based on the measurements and collection of the sampling data, as with SLR

5. The x values are measured without error (i.e., the x values are fixed).

6. The y values are randomly selected for each given set of the x variables (i.e., for each fixed set of x values, a list of all possible y values is made).

As with SLR, often observations will be gathered using simple random sampling or systematic sampling (grid across the land area). This does not strictly meet this assumption [much more difficult to meet with many x-variables!] <u>If the equation is "correct", then this does not cause problems.</u> If not, the estimated equation will be biased.

## Transformations

- Same as for SLR – except that there are more x variables; can also add variables e.g. use dbh and $dbh^2$ as x1 and x2.
- Try to transform *x's* first and leave y = variable of interest; not always possible.
- Use graphs to help choose transformations
- Will result in an "iterative" process:
  1. Fit the equation
  2. Check the assumptions [and check for outliers]
  3. Make any transformations based on the residual plot, and plots of *y* versus each *x*
  4. Also, check any very unusual points to see if these are measurement/transcription errors; ONLY remove the observation if there is a very good reason to do so
  5. Fit the equation again, and check the assumptions
  6. Continue until the assumptions are met [or nearly met]



## Measures of Goodness of Fit

How well does the regression fit the sample data?

- For multiple linear regression, a graph of the the predicted versus measured *y* values indicates how well the line fits the data
- Two measures commonly used: coefficient of multiple determination ($R^2$) and standard error of the estimate($SE_E$), similar to SLR

To calculate $R^2$ and $SE_E$, first, calculate the SSE (this is what was minimized, as with SLR):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + ...b_m x_{mi}))^2$$

The sum of squared differences between the measured and estimated *y*'s. This is the same as for SLR, but there are more slopes and more *x* (predictor) variables.

Calculate the sum of squares for *y*:

$$SSy = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i^{\,2}\right) \Big/ n = s_y^{\,2}(n-1)$$

The sum of squared difference between the measured *y* and the mean of *y*-measures.

Calculate the sum of squares regression:

$$SSreg = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2 = b_1 SPx_1 y + b_2 SPx_2 y + \ldots + b_3 SPx_3 y$$
$$= SSy - SSE$$

The sum of squared differences between the mean of y-measures and the predicted *y*'s from the fitted equation. Also, is the sum of squares for *y* – the sum of squared errors.



Then: $$R^2 = \frac{SSy - SSE}{SSy} = 1 - \frac{SSE}{SSy} = \frac{SSreg}{SSy}$$

- SSE, SSY are based on *y*'s used in the equation – will not be in original units if *y* was transformed
- $R^2$ = coefficient of multiple determination; proportion of variance of *y*, accounted for by the regression using *x's*
- O (very poor – horizontal surface representing no relationship between *y* and *x*'s) to 1 (perfect fit – surface passes through the data)
- SSE falls as *m* (number of independent variable) increases, so $R^2$ rises as more explanatory (independent or predictor) variables are added.

A similar measure is called the Adjusted $R^2$ value. A penalty is added as you add x-variables to the equation:

$$R_a^{\,2} = 1 - \left(\frac{n-1}{n-(m+1)}\right)\frac{SSE}{SSy}$$

And: $$SE_E = \sqrt{\frac{SSE}{n-m-1}}$$

- SSE is based on *y*'s used in the equation – will not be in original units if *y* was transformed
- *n*-*m*-1 is the degrees of freedom for the error; is the number of observations minus the number of fitted coefficients
- $SE_E$ - standard error of the estimate; in same units as *y*
- Under normality of the errors:
  - ±1 $SE_E$ ≅ 68% of sample observations
  - ±2 $SE_E$ ≅ 95% of sample observations
- Want low $SE_E$
- $SE_E$ falls as the number of predictor variables increases and SSE falls, but then rises, since *n*-*m* -1 is getting smaller



<u>y-variable was transformed</u>: Can calculate estimates of these for the original y-variable unit, $I^2$ (Fit Index) and estimated standard error of the estimate ($SE_E$'), in order to compare to $R^2$ and $SE_E$ of other equations where the *y* was not transformed, similar to SLR.

$I^2 = 1 - SSE/SSY$

- where SSE, SSY are in original units. NOTE must "back-transform" the predicted *y*'s to calculate the SSE in original units.
- Does not have the same properties as $R^2$, however it can be less than 0

Estimated standard error of the estimate ($SE_E$') , when the dependent variable, *y*, has been transformed:

$$SE_E' = \sqrt{\frac{SSE(original\ units)}{n-m-1}}$$

- SEE' - standard error of the estimate ; in same units as original units for the dependent variable
- want low SEE'

Estimated Variances, Confidence Intervals and Hypothesis Tests

*Testing Whether the Regression is Significant*

Does knowledge of *x*'s improve the estimate of the mean of *y*? Or is it a flat surface, which means we should just use the mean of *y* as an estimate of mean *y* for any set of *x* values?

SSE/ (*n*-*m*-1):

- Mean squared error.
  - The degrees of freedom are *n*-*m*-1 (same as *n*-(*m*+1)
  - *n* observations with (*m*+1) statistics estimated from these: $b_0$, $b_1$ , $b_2$ ,… $b_m$
- Under the assumptions of MLR, is an unbiased estimated of the true variance of the error terms (error variance)



SSR/*m*:

- Called the Mean Square Regression
- Degrees of Freedom=*m*: *m* *x*-variables
- Under the assumptions of MLR, this is an estimate the error variance PLUS a term of variance explained by the regression using *x's*.

H0: Regression is not significant

H1: Regression is significant

Same as:

H0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_m = 0$ [all slopes are zero meaning no relationship with x's]

H1: not all slopes =0 [some or all slopes are not equal to zero]

If H0 is true, then the equation is:

$$y_i = \beta_0 + 0\, x_{1i} + 0\, x_{2i} + \ldots + 0\, x_{mi} + \varepsilon_i$$

$$y_i = \beta_0 + \varepsilon_i \qquad \hat{y}_i = \beta_0$$

Where the *x*-variables have no influence over y; they do not help to better estimate *y*.

As with SLR, we can use an F-test, as it is the ratio of two variances; unlike SLR we cannot use a t-test since we are only testing several slope coefficients.

Using an F test statistic:

$$F = \frac{SSreg/m}{SSE/(n-m-1)} = \frac{MSreg}{MSE}$$

- Under H0, this follows an F distribution for a 1- α percentile with *m* and *n*-*m*-1 degrees of freedom.
- If the F for the fitted equation is larger than the F from the table, we reject H0 (not likely true). The regression is significant, in that one or more of the the true slopes (the population slopes) are likely not equal to zero.

Information for the F-test in the Analysis of Variance Table:

| Source | df | SS | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | *m* | *SSreg* | *MSreg*= *SSreg*/m | *F*= *MSreg*/*MSE* | Prob F> $F_{(m,n-m-1,1-\alpha)}$ |
| Error | *n*-*m*-1 | *SSE* | *MSE*= *SSE*/(*n* *m*-1) | | |
| Total | *n*-1 | *SSy* | | | |



*Estimated Standard Errors for the Slope and Intercept*

Under the assumptions, we can obtain an unbiased estimated of the standard errors for the slope and for the intercept [measure of how these would vary among different sample sets], using the one set of sample data.

For multiple linear regression, these are more easily calculated using matrix algebra. If there are more than 2 x-variables, the calculations become difficult; we will rely on statistical packages to do these calculations.

*Confidence Intervals for the True Slope and Intercept*

Under the assumptions, confidence intervals can be calculated as:

For $\beta_o$: $b_0 \pm t_{1-\alpha/2,n-m-1} \times s_{b_0}$

For $\beta_j$: $b_j \pm t_{1-\alpha/2,n-m-1} \times s_{b_j}$ [ for any of the slopes]

[See example]

*Hypothesis Tests for one of the True Slopes or Intercept*

H0: $\beta_j = c$ [the parameter (true intercept or true slope is equal to the constant, c, given that the other x-variables are in the equation]

H1: $\beta_j \neq c$ [true intercept or slope differs from the constant c; given that the other x-variables are in the equation]

Test statistic:

$$t = \frac{b_j - c}{s_{b_j}}$$

Under H0, this is distributed as a t value of $t_c = t_{n\text{-}m\text{-}1,\ 1\text{-}\alpha/2}$.

Reject $H_o$ if $|t| > t_c$.

- It is possible to do one-sided hypotheses also, where the alternative is that the true parameter (slope or intercept) is greater than (or less than) a specified constant c. MUST be careful with the $t_c$ as this is different.



*The regression is significant, but which x-variables should we retain?*

With MLR, we are particularly interested in which x-variables to retain. We then test: Is variable $x_j$ significant given the other *x* variables? e.g. diameter, height - do we need both?

H0: $\beta_j = 0$, given other x-variables (i.e., variable not significant)

H1: $\beta_j \neq 0$, given other x-variables.

A t-test for that variable can be used to test this.

Another test, the partial F-test can be used to test one x-variable (as t-test) or to test a group of x-variables, given the other x-variables in the equation.

- Get regression analysis results for all x-variables [full model]
- Get regression analysis results for all but the x-variables to be tested [reduced model]

$$partial\ F = \frac{\left(SSreg(\ full) - SSreg(reduced)\right)/r}{SSE/(n-m-1)(\ full)}$$

OR

$$partial\ F = \frac{\left(SSE(reduced) - SSE(\ full)\right)/r}{SSE/(n-m-1)(\ full)}$$

$$= \frac{(SS\ \text{due to dropped variable(s)})/r}{MSE(\ full)}$$

Where *r* is the number of x-variables that were dropped (also equals: (1)the regression degrees of freedom for the full model minus the regression degrees of freedom for the reduced model, OR (2) the error degrees of freedom for the reduced model, minus the error degrees of freedom for the full model)



- Under H0, this follows an F distribution for a 1- α percentile with *r* and *n*-*m*-1 (full model) degrees of freedom.
- If the F for the fitted equation is larger than the F from the table, we reject H0 (not likely true). The regression is significant, in that the variable(s) that were dropped are significant (account for variance of the y-variable), given that the other x-variables are in the model.

*Confidence Interval for the True Mean of y given a particular set of x values*

For the mean of all possible y-values given a particular value set of x-values ($\mu_y|\mathbf{x}_h$):

$$\hat{y} \mid \mathbf{x}_h \pm t_{n-m-1,1-\alpha/2} \times s_{\hat{y}|\mathbf{x}_h}$$

where

$$\hat{y} \mid \mathbf{x}_h = b_0 + b_1 x_{1h} + b_2 x_{2h} + \cdots + b_m x_{mh}$$

$$s_{\hat{y}|\mathbf{x}_h} = \text{from statistical package output}$$

*Confidence Bands*

Plot of the confidence intervals for the mean of y for several sets x-values is not possible with MLR

*Prediction Interval for 1 or more y-values given a particular set of x values*

For one possible new y-value given a particular set of x values:

$$\hat{y}_{(new)} \mid \mathbf{x}_h \pm t_{n-m-1,1-\alpha/2} \times s_{\hat{y}(new)|\mathbf{x}_h}$$

Where

$$\hat{y} \mid \mathbf{x}_h = b_0 + b_1 x_{1h} + b_2 x_{2h} + \cdots + b_m x_{mh}$$

$$s_{\hat{y}(new)|\mathbf{x}_h} = \text{from statistical package output}$$

For the average of *g* new possible y-values given a particular value of x:

$$\hat{y}_{(new)} \mid \mathbf{x}_h \pm t_{n-m-1,1-\alpha/2} \times s_{\hat{y}(newg)|\mathbf{x}_h}$$

where

$$\hat{y} \mid \mathbf{x}_h = b_0 + b_1 x_{1h} + b_2 x_{2h} + \cdots + b_m x_{mh}$$

$$s_{\hat{y}(newg)|\mathbf{x}_h} = \text{from statistical package output}$$



Selecting and Comparing Alternative Models

*Process to Fit an Equation using Least Squares*

Steps (same as for SLR):

1. Sample data are needed, on which the dependent variable and all explanatory (independent) variables are measured.
2. Make any transformations that are needed to meet the most critical assumption: The relationship between *y* and *x*'s is linear.

Example: volume = $\beta_0$ + $\beta_1$ dbh +$\beta_2$ dbh$^2$ may be linear whereas volume versus dbh is not. Need both variables.

3. Fit the equation to minimize the sum of squared error.
4. Check Assumptions. If not met, go back to Step 2.
5. If assumptions are met, then check if the regression is significant. If it is not, then it is not a candidate model (need other x-variables). If yes, then go through further steps for MLR.

6. Are all variables needed? If there are x-variables that are not significant, given the other variables:

- drop the least significant one (highest p-value, or lowest absolute value of t)
- refit the regression and check assumptions.
- if assumptions are met, then repeat steps 5 and 6

continue until all variables in the regression are significant given the other x-variables also in the model

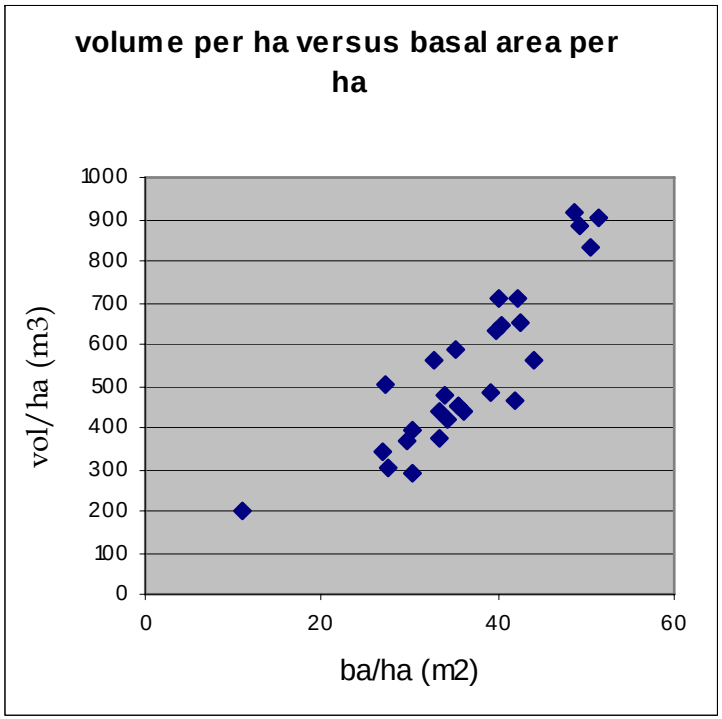


## Multiple Linear Regression Example

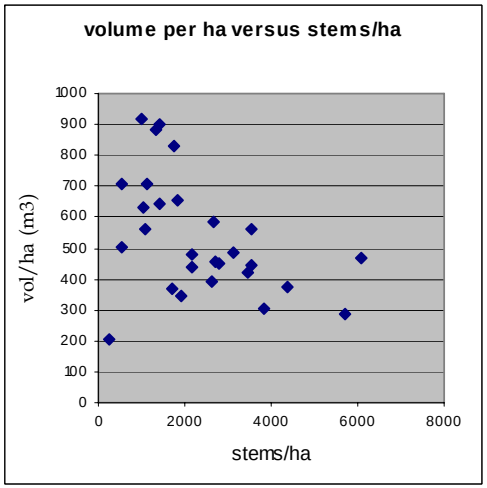
n=28 stands y=vol/ha (m3)

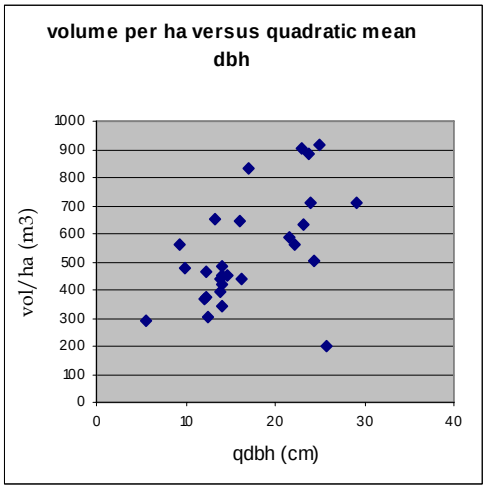
| volume/ha $m^3$ | Age years | Site Index | Basal area/ha $m^2$ | Stems /ha | Top height m | Qdbh cm |
|---|---|---|---|---|---|---|
| 559.3 | 82 | 14.6 | 32.8 | 1071 | 22.4 | 22.2 |
| 559 | 107 | 9.4 | 44.2 | 3528 | 17 | 9.3 |
| 831.9 | 104 | 12.8 | 50.5 | 1764 | 21.5 | 17 |
| 365.7 | 62 | 12.5 | 29.6 | 1728 | 16.4 | 12.1 |
| 454.3 | 52 | 14.6 | 35.4 | 2712 | 18.9 | 14.1 |
| 486 | 58 | 13.9 | 39.1 | 3144 | 17.5 | 14 |
| 441.6 | 34 | 18.5 | 36.2 | 3552 | 17.4 | 13.8 |
| 375.8 | 35 | 17 | 33.4 | 4368 | 15.6 | 12.2 |
| 451.4 | 33 | 19.1 | 35.4 | 2808 | 16.8 | 14.7 |
| 419.8 | 23 | 23.4 | 34.4 | 3444 | 17.3 | 14 |
| 467 | 33 | 17.7 | 42 | 6096 | 16.4 | 12.2 |
| 288.1 | 33 | 15 | 30.3 | 5712 | 13.8 | 5.6 |
| 306 | 32 | 18.2 | 27.4 | 3816 | 16.7 | 12.5 |
| 437.1 | 68 | 13.8 | 33.3 | 2160 | 19.1 | 16.2 |
| 633.2 | 126 | 11.4 | 39.9 | 1026 | 21 | 23.2 |
| 707.2 | 125 | 13.2 | 40.1 | 552 | 23.3 | 29.2 |
| 203 | 117 | 13.7 | 11 | 252 | 22.1 | 25.8 |
| 915.6 | 112 | 13.9 | 48.7 | 1017 | 24.2 | 25 |
| 903.5 | 110 | 13.9 | 51.5 | 1416 | 23.2 | 23 |
| 883.4 | 106 | 14.7 | 49.4 | 1341 | 24.3 | 23.7 |
| 586.5 | 124 | 12.8 | 35.2 | 2680 | 22.6 | 21.5 |
| 500.1 | 60 | 18.4 | 27.3 | 528 | 22.7 | 24.4 |
| 343.5 | 63 | 14 | 26.9 | 1935 | 17.6 | 14.1 |
| 478.6 | 60 | 15.2 | 34 | 2160 | 19.4 | 9.9 |
| 652.2 | 62 | 15.9 | 42.5 | 1843 | 20.5 | 13.2 |
| 644.7 | 63 | 16.2 | 40.4 | 1431 | 21 | 16.1 |
| 390.8 | 57 | 14.8 | 30.4 | 2616 | 18.3 | 13.9 |
| 709.8 | 87 | 14.3 | 42.3 | 1116 | 22.6 | 23.9 |

Objective: obtain an equation for estimating volume per ha from some of the easy to measure variables such as basal area /ha (only need dbh on each tree), qdbh (need dbh on each tree and stems/ha), and stems/ha





Then, we would need: SSY, $SSX_1$, $SSX_2$, $SSX_3$, $SPX_1Y$, $SPX_2Y$, $SPX_3Y$, $SPX_1X_2$, $SPX_1X_3$, $SPX_2X_3$, and insert these into the four equations and solve:

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$b_1 = \frac{SPx_1y}{SSx_1} - b_2\frac{SPx_1x_2}{SSx_1} - b_3\frac{SPx_1x_3}{SSx_1}$$

$$b_2 = \frac{SPx_2y}{SSx_2} - b_1\frac{SPx_1x_2}{SSx_2} - b_3\frac{SPx_2x_3}{SSx_2}$$

$$b_3 = \frac{SPx_3y}{SSx_3} - b_1\frac{SPx_1x_3}{SSx_3} - b_2\frac{SPx_2x_3}{SSx_3}$$

And then check assumptions, make any necessary transformations, and start over!



<u>Methods to aid in selecting predictor (x) variables</u>

Methods have been developed to help in choosing which x-variables to include in the equation. These include:

1. Forward: Bring in variables one at a time, until the remaining ones are no longer significant, given the others already in the equation. (in only)
2. Backward: Drop variables one at a time, until all remaining variables are significant, given the others still in the equation (out only)
3. Stepwise (in and out)

NOTE:

These tools just gives candidate models. You must check whether the assumptions are met and do a full assessment of the regression results

**Steps for Forward Stepwise, for example:**

To fit this "by hand", you would need to do the following steps:

1. Fit a simple linear regression for vol/ha with each of the explanatory (x) variables.
2. Of the equations that are significant (assumptions met?), select the one with the highest F-value.
3. Fit a MLR with vol/ha using the selected variable, plus each of the explanatory variables (2 x-variables in each equations). Check to see if the "new" variable is significant given the original variable (which may now be not significant, but forward stepwise does not drop variables). Of the ones that are significant (given the original variable is also in the equation), pick the one with the largest partial-F (for the new variable).
4. Repeat step 3, bringing in varables until i) there are no more variables or ii) the remaining variables are not significant given the other variables.



*For a number of models, select based on:*

1. Meeting assumptions: If an equation does not meet the assumption of a linear relationship, it is not a candidate model
2. Compare the fit statistics. Select higher $R^2$ (or $I^2$), and lower $SE_E$ (or $SE_E$')
3. Reject any models where the regression is not significant, since this model is no better than just using the mean of y as the predicted value.
4. Select a model that is biologically tractable. A simpler model is generally preferred, unless there are practical/biological reasons to select the more complex model
5. Consider the cost of using the model

Adding class variables as predictors

Want to add a class variable. Examples:

1. Add species to an equation to estimate tree height.
2. Add gender (male/female) to an equation to estimate weight of adult tailed frogs.
3. Add machine type to an equation that predicts lumber output.

How is this done?

- Use "dummy" or "indicator variables to represent the class variable

  e.g. have 3 species. Set up X1 and X2 as dummy variables:

| Species | X1 | X2 |
|---|---|---|
| Cedar | 1 | 0 |
| Hemlock | 0 | 1 |
| Douglas fir | 0 | 0 |

  Only need two dummy variables to represent the three species.



**The two dummy variables as a group represent the species.**

- Add the dummy variables to the equation – this will alter the intercept
- To alter the slopes, add an interaction between dummy variables and continuous variable(s)

  e.g. have 3 species, and a continuous variable, dbh

| Species | X1 | X2 | dbh | X4=X1 * dbh | X5=X2*dbh |
|---|---|---|---|---|---|
| Cedar | 1 | 0 | 10 | 10 | 0 |
| Hemlock | 0 | 1 | 22 | 0 | 22 |
| Douglas fir | 0 | 0 | 15 | 0 | 0 |

NOTE: There would be more than one line of data (sample) for each species.

- **The two dummy variables, and the interactions with the continuous variable as a group represent the species.**

How does this work?

$$y_i = b_0 \underbrace{+ b_1 x_{1i} + b_2 x_{2i}}_{\text{dummy variables}} + \underbrace{b_3 x_{3i}}_{\text{dbh}} \underbrace{+ b_4 x_{41i} + b_5 x_{5i}}_{i\text{interactions}} + e_i$$

For Cedar (CW):

$$y_i = (b_0 + b_1) + \underbrace{(b_3 + b_4) x_{3i}}_{\text{dbh}} + e_i$$

For Hemlock (HW):

$$y_i = (b_0 + b_2) + \underbrace{(b_3 + b_5) x_{3i}}_{\text{dbh}} + e_i$$

For Douglas fir (FD):

$$y_i = b_0 + \underbrace{b_3 x_{3i}}_{\text{dbh}} + e_i$$

Therefore: fit one equation using all data, but get different equations for different species. Also, can test for differences among species, using **a partial-F test.**



Other methods, than SLR and Multiple Linear Regression, when transformations do not work:

*Nonlinear least squares:* Least squares solution for nonlinear models; uses a search algorithm to find estimated coefficients; has good properties for large datasets; still assumes normality, equal variances, and independent observations

*Weighted least squares*: for unequal variances. Estimate the variances and use these in weighting the least squares fit of the regression; assumes normality and independent observations

*Generalized linear model*: used for distributions other than normal (e.g., binomial, Poisson, etc.), but with no correlation between observations; uses maximum likelihood

*Generalized least Squares and Mixed Models*: use maximum likelihood for fitting models with unequal variances, correlations over space, correlations over time, but normally distributed errors

*Generalized linear mixed models:* Allows for unequal variances, correlations over space and/or time, and non-normal distributions; uses maximum likelihood