5. Multiple Linear Regression

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

Multiple Linear Regression Example

Objective: obtain an equation for estimating volume per ha from some of the easy to measure variables such as

- stems/ha
- basal area /ha (only need dbh on each tree)
- qdbh (need dbh on each tree and stems/ha)

VERY IMPORTANT assumption of MLR: The relationship between the x's and y is linear!

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

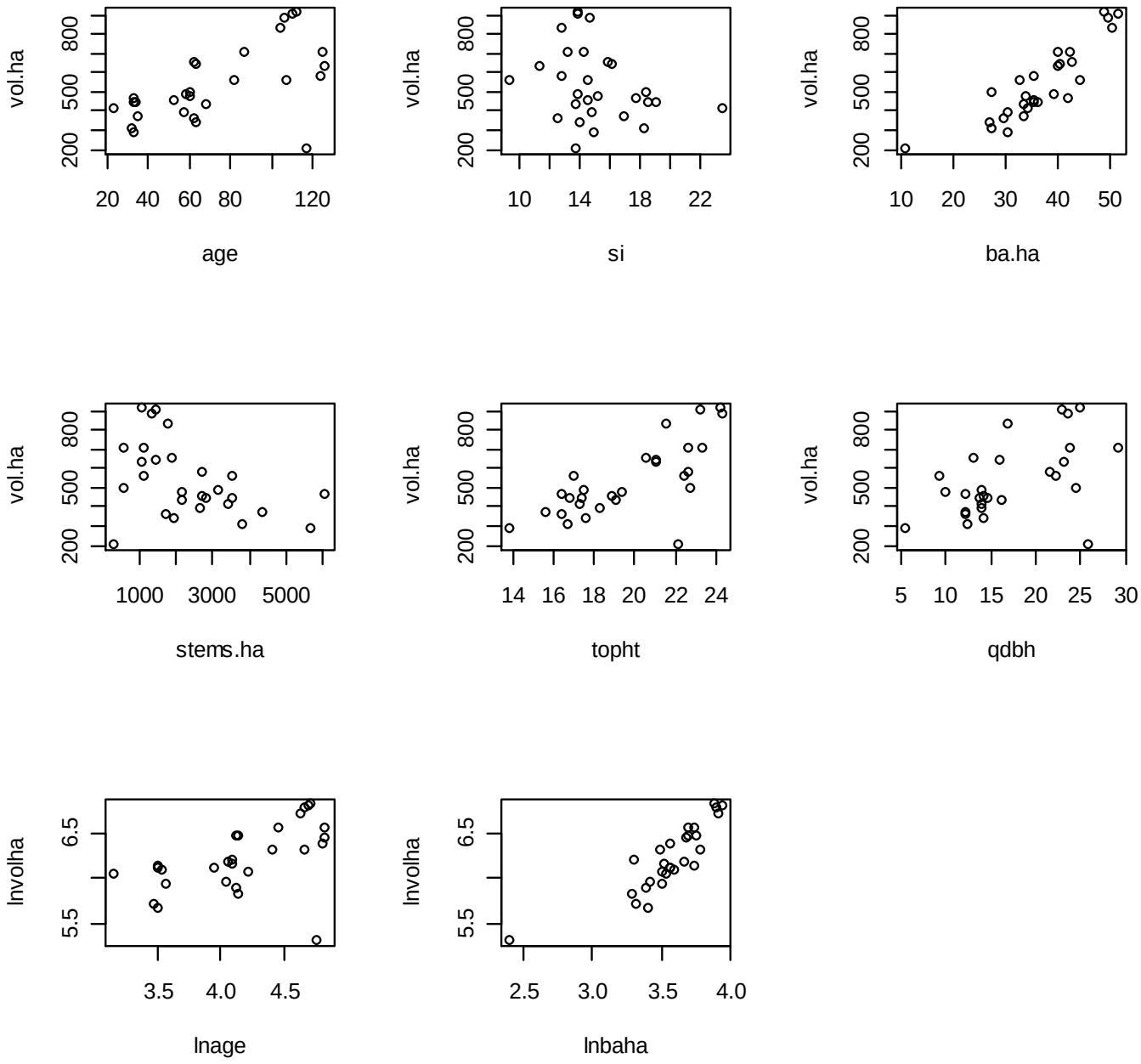
R code

```
> standdat<- read.table("../data/stand.txt",header=TRUE)
> # since standdat was created as a dataframe, the names can be attached for simpler commands
> attach(standdat)# this allows you to use the dataframe, treedat, with shorter names for
    variables
> names(standdat)
[1] "vol.ha"   "age"      "si"       "ba.ha"    "stems.ha" "topht"    "qdbh"
>

> lnvolha=log(vol.ha)
> lnage=log(age)
> lnbaha=log(ba.ha)
>
> detach(standdat)# This just detaches the dataframe, standdat, but it can be reattached
>
> standdat<-data.frame(standdat,lnvolha,lnage,lnbaha)
> rm(lnvolha,lnage,lnbaha)
>
> attach(standdat)
> names(standdat)
 [1] "vol.ha"   "age"      "si"       "ba.ha"    "stems.ha" "topht"
 [7] "qdbh"     "lnvolha"  "lnage"    "lnbaha"
>
```

Script 4_MLR.R

plots

```
> par(mfrow=c(3,3),cex=0.7)
> plot(vol.ha~age,data=standdat)
> plot(vol.ha~si,data=standdat)
> plot(vol.ha~ba.ha,data=standdat)
> plot(vol.ha~stems.ha,data=standdat)
> plot(vol.ha~topht,data=standdat)
> plot(vol.ha~qdbh,data=standdat)
> plot(lnvolha~lnage,data=standdat)
> plot(lnvolha~lnbaha,data=standdat)
> par(mfrow=c(1,1),cex=1)
>
```


vol.ha
200
500
800
20
40
60
80
120
age
vol.ha
200
500
800
10
14
18
22
si
vol.ha
200
500
800
10
20
30
40
50
ba.ha
vol.ha
200
500
800
1000
3000
5000
stems.ha
vol.ha
200
500
800
14
16
18
20
22
24
topht
vol.ha
200
500
800
5
10
15
20
25
30
qdbh
lnvolha
5.5
6.5
3.5
4.0
4.5
lnage
lnvolha
5.5
6.5
2.5
3.0
3.5
4.0
lnbaha

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

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

β_0 *is the y intercept parameter*

$\beta_1, \beta_2, \ldots, \beta_p$ are slope parameters

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

ε_i is the error term or the residual*;* is the variation in the dependent variable (the *y)* which is not accounted for by the independent variables (the *x's).*

Multiple Linear Regression (MLR)

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*.

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.

Least squares solution 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:

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

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

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

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

SP = sum of products between two variables, for example for *y with x1:*

$$SPx_1y = \sum_{i=1}^{n}(y_i - \bar{y})(x_{1i} - \bar{x}_1)$$

$$= \sum_{i=1}^{n} y_i x_{1i} - \frac{\left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = s^2{}_{x_1y}(n-1)$$

SS = sum of squares for one variable, for example for *x1:*

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

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!

```
> ##### model ######################################
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Coefficients:
(Intercept)        ba.ha     stems.ha         qdbh
 -198.17649     18.56615     -0.03124      7.54214
```

Properties of a least squares regression "surface":

1. Always passes through $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_m, \bar{y})$
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 hypotheses

For these tests 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.

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:

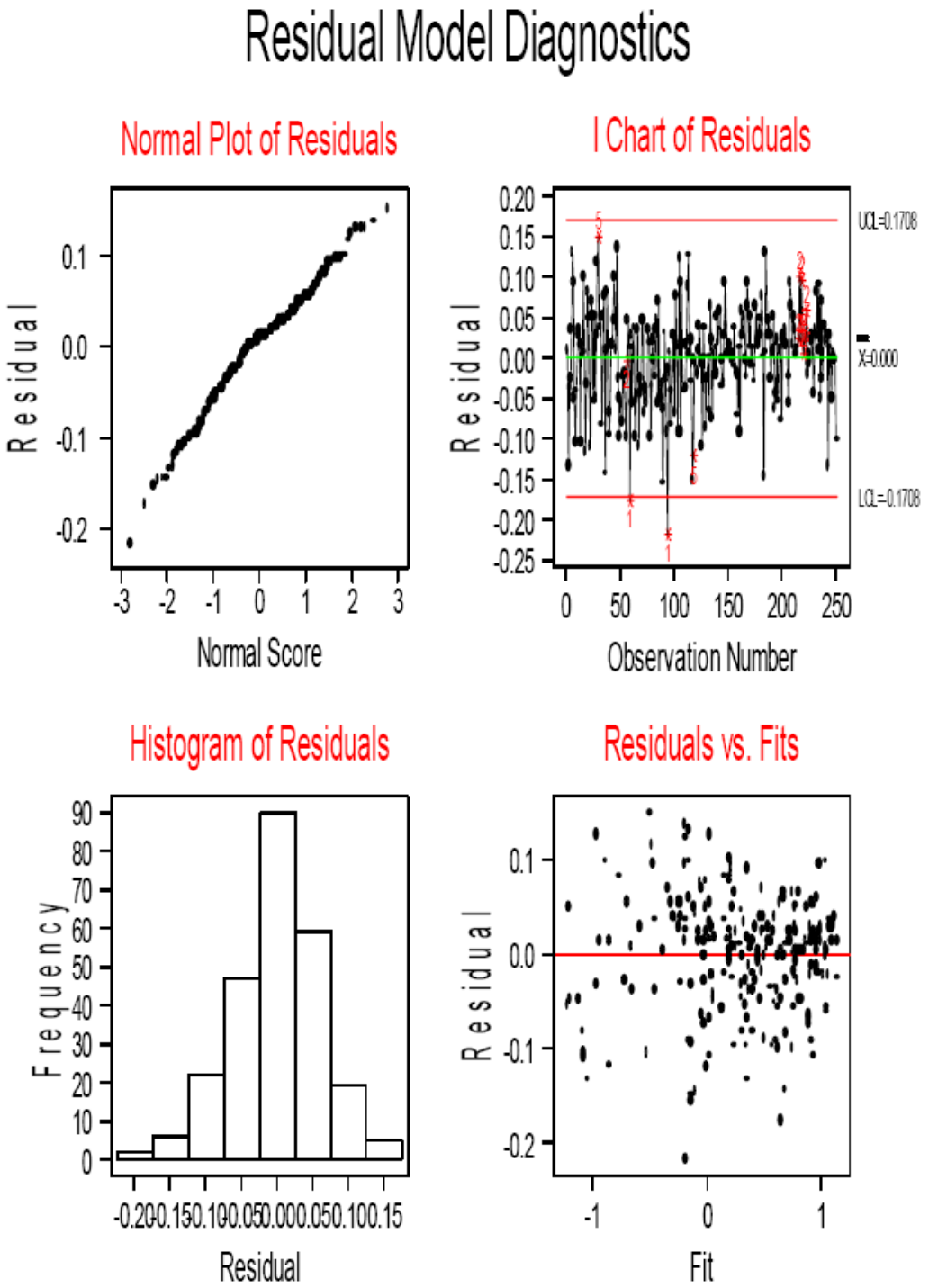
4. The y values must be normally distributed for each combination of x values.

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 the same problems as with SLR

Example:

- Linear relationship met
- equal variance met
- no evidence of trend with observation number (independence maybe met).
- normal distribution met

Logvol=f(dbh,logdbh)



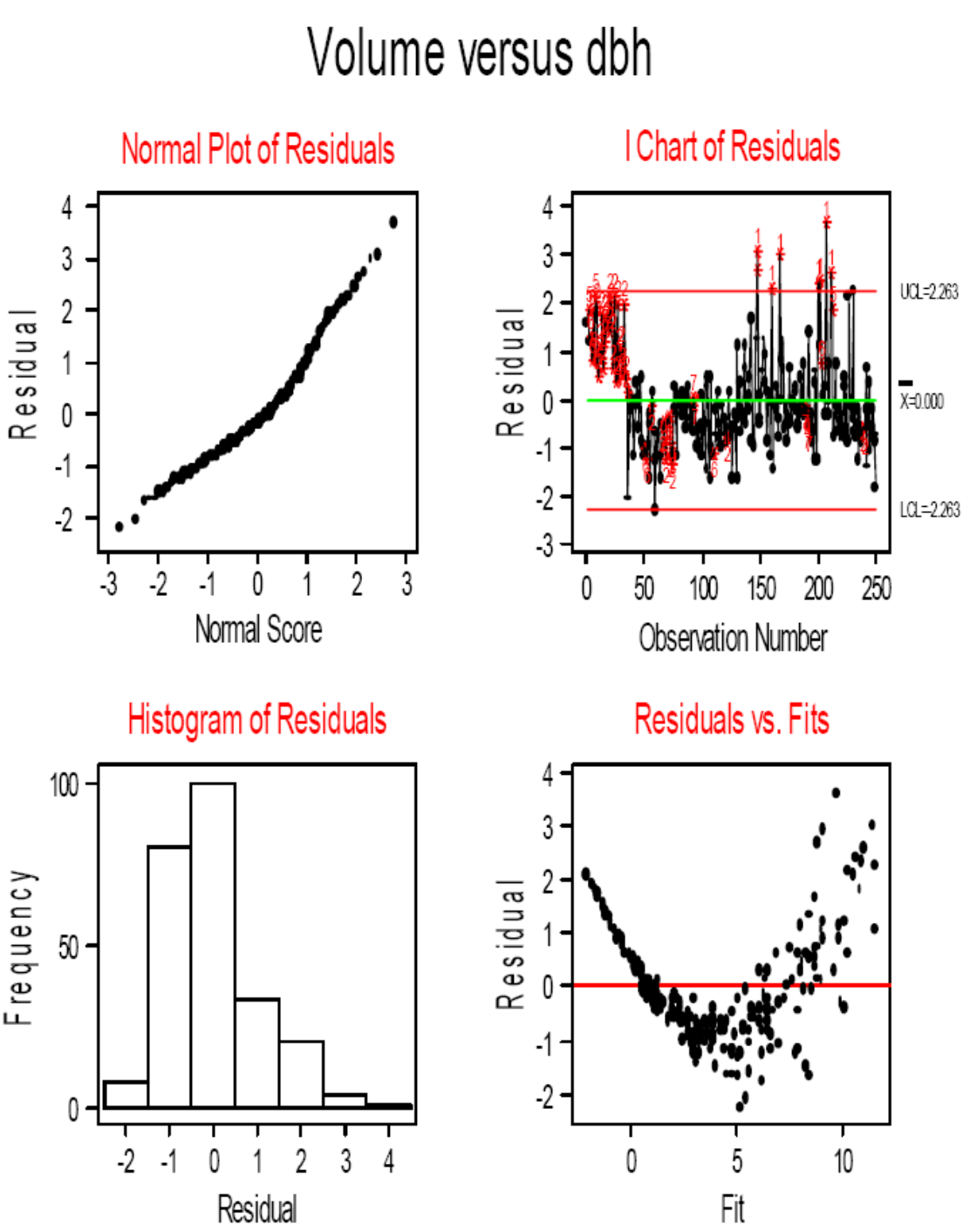
Example:

- Independence of observations not met

When we have no time relation in sequence of sampling (order in space or time), there is no reason to suspect relationship among our observations

- Linear relationship not met:

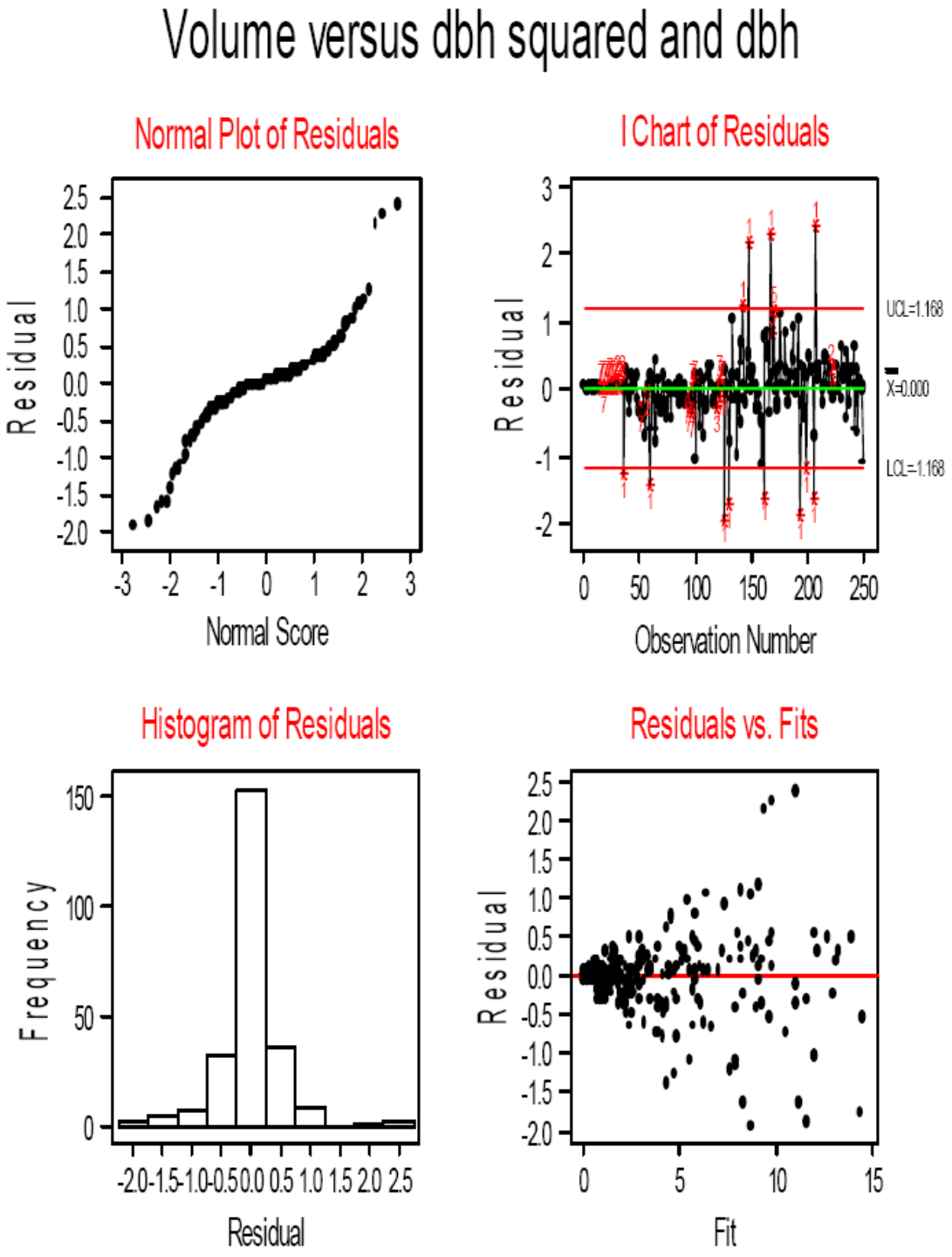
Pos. and neg. residuals are not balanced for all values of ŷ



Example:

- Variances are not equal:

The spread of the residuals is not the same for all ŷ



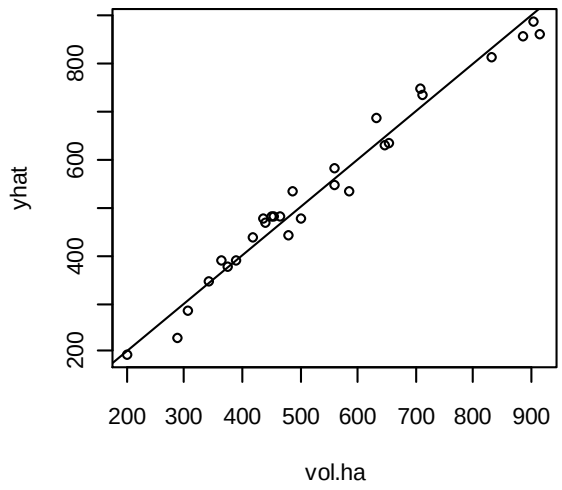
```
> yhat<-fitted(model.volha)
> resid<-resid(model.volha)
> par(mfrow=c(2,2),cex=0.7)
> plot(yhat~vol.ha)
> abline(a=0,b=1)    # plot a reference line where yhat equals vol.ha
>
> plot(resid~yhat)   # residual plot
>
> qqnorm(resid)      # normality plot
> qqline(resid,col=2)
>
> hist(resid, breaks =6 , density=10,col="green", border="black") # draws a histogram
> par(mfrow=c(1,1),cex=1)
> shapiro.test(resid)

        Shapiro-Wilk normality test

data:  resid
W = 0.9606, p-value = 0.36
```

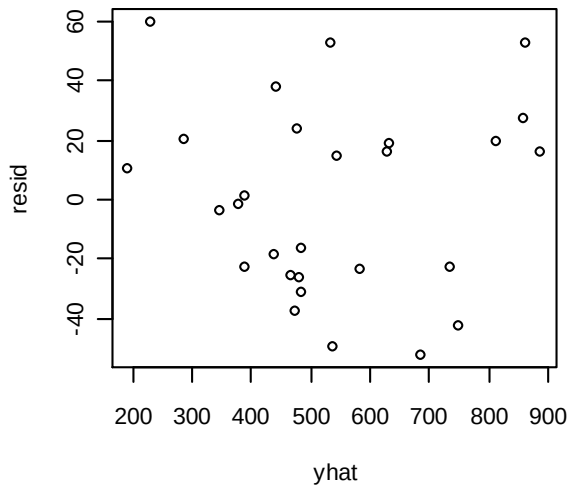
Normal distribution of the residuals:

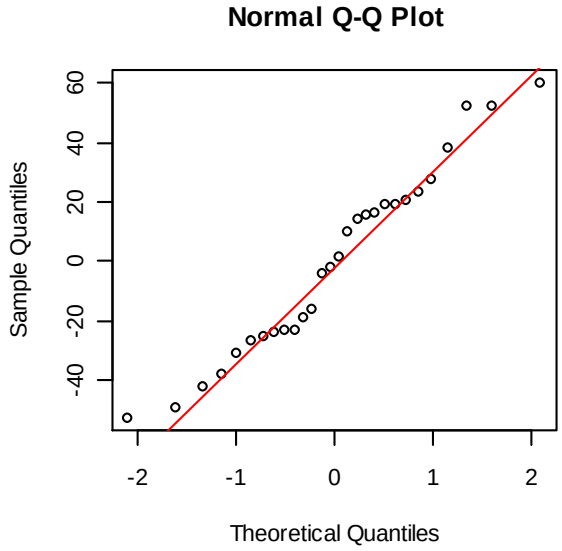
H0: Residuals are Normally distributed (this is what we hope)

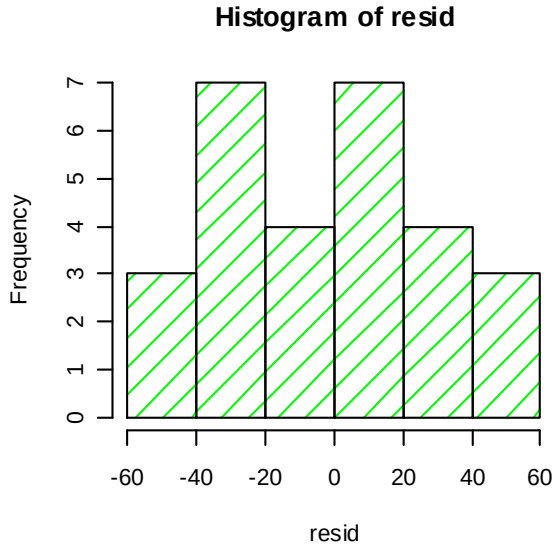
Ha: Residuals are not Normal

If P-value < α, then we reject the H0


yhat
200
400
600
800
200
300
400
500
600
700
800
900
vol.ha


resid
-40
-20
0
20
40
60
200
300
400
500
600
700
800
900
yhat


Normal Q-Q Plot
Sample Quantiles
-40
-20
0
20
40
60
-2
-1
0
1
2
Theoretical Quantiles


Histogram of resid
Frequency
0
1
2
3
4
5
6
7
-60
-40
-20
0
20
40
60
resid

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.

“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]

Careful with transformations!

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_2= 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. ∴ **volume (m^3) = 3.184. This represents the estimated average volume for trees with dbh=30 cm and height=28 m.**

```
> log10(3.184)
[1] 0.5029731
> 10^0.5029731
[1] 3.184
```

Note: This equation is originally a nonlinear equation:

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

Which was transformed to a linear equation using logarithms:

$\log10(vol) = log10(a) + b\ log10(dbh) + c\ log10(ht) + log10\varepsilon$

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.

```
> log10(3)
[1] 0.4771213
> 0.477-0.503
[1] -0.026
> 10^-0.026
[1] 0.9418896
> 3-3.184
[1] -0.184
```

Measures of Goodness of Fit

How well does the regression fit the sample data?

- For multiple linear regression, a graph of 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 SSE, SSy and SSreg:

- ***SSE*** (this is what was minimized):

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

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

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

To calculate R^2 and SE_E

- ***SSy***: the sum of squares for *y:*

$$SSy = \sum_{i=1}^{n} (y_i - \bar{y})^2 = s_y{}^2 (n-1)$$

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 (SSTO)**.

To calculate R^2 and SE_E

- ***SSreg***: the sum of squares regression:

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

The sum of squared differences between the mean of y-measures and the predicted *y's from the fitted equation.*

= the sum of squares for *y – the sum of squared* errors.

Remarkable property

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2$$

Or *SSy = SSreg + SSe*

Breakdown of Degrees of Freedom

- SSY (=SSTO)
  - 1 linear constraint due to the calculation and inclusion of the mean (equivalently: because sum must be 0)
    - n-1 degrees of freedom
- SSE
  - m + 1 linear constraints arising from the estimation of m + 1 parameters in the regression function
    - n-(m +1) = n-m-1 degrees of freedom
- SSR
  - All fitted values are calculated from the same regression function: m + 1 degrees of freedom in the regression parameters, one is lost due to linear constraint $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$
    - m degrees of freedom

Remarkable: n – 1 = (n – m – 1) + m

SE_E = Standard Error of the Estimate

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

= root of the mean square error (MSE)

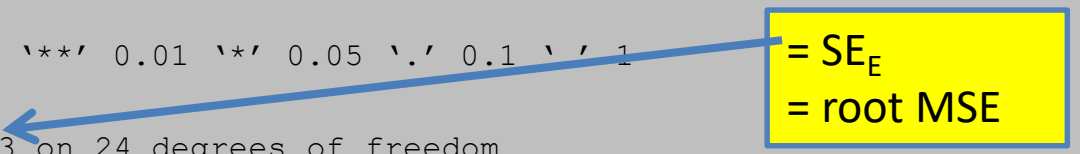
- SSE is based on *y's used in the equation – will not* be in original units if *y was transformed*
- n – m – 1 is the degrees of freedom of the error; is the number of observations minus the number of fitted coefficients
- SE_E - standard error of the estimate; in same units as y

SE_E = Standard Error of the Estimate

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

- 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 SE_E

```
> summary(model.volha)

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Residuals:
    Min      1Q  Median      3Q     Max
-52.334 -23.875  -0.104  19.642  59.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.982e+02  4.789e+01  -4.138 0.000372 ***
ba.ha        1.857e+01  7.564e-01  24.546  < 2e-16 ***
stems.ha    -3.124e-02  7.016e-03  -4.453 0.000167 ***
qdbh         7.542e+00  1.740e+00   4.335 0.000225 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.43 on 24 degrees of freedom
Multiple R-squared: 0.9727,     Adjusted R-squared: 0.9692
F-statistic: 284.6 on 3 and 24 DF,  p-value: < 2.2e-16
```



R^2

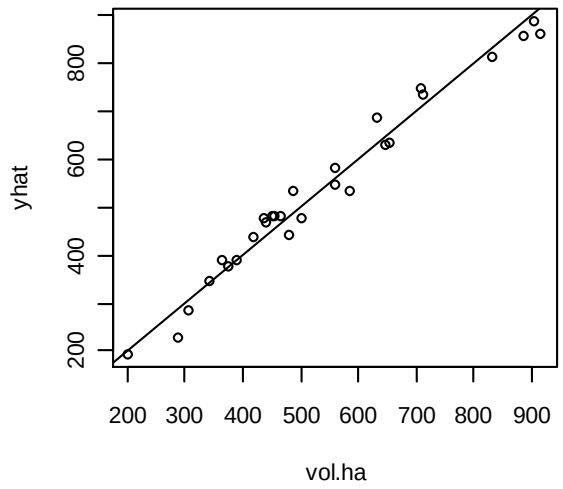
$$R^2 = \frac{SSy - SSE}{SSy} = 1 - \frac{SSE}{SSy} = \frac{SSreg}{SSy}$$

- 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*
- 0 (when SSy = SSE) (very poor – horizontal surface representing no relationship between y and x's) to 1 (perfect fit – surface passes through the data)

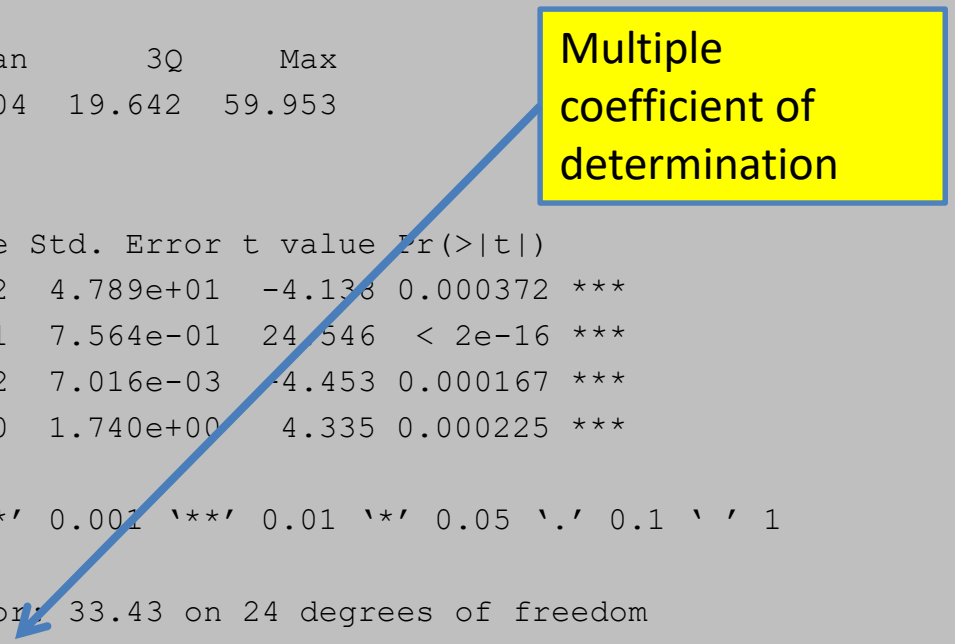
R^2

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

- It can be shown that the coefficient of multiple determination R^2 can be viewed as a coefficient of simple **determination r^2 between the responses y_i and the fitted values $\hat{y}_i$**



```
> summary(model.volha)

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Residuals:
    Min      1Q  Median      3Q     Max
-52.334 -23.875  -0.104  19.642  59.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.982e+02  4.789e+01  -4.138 0.000372 ***
ba.ha        1.857e+01  7.564e-01  24.546  < 2e-16 ***
stems.ha    -3.124e-02  7.016e-03  -4.453 0.000167 ***
qdbh         7.542e+00  1.740e+00   4.335 0.000225 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.43 on 24 degrees of freedom
Multiple R-squared: 0.9727,     Adjusted R-squared: 0.9692
F-statistic: 284.6 on 3 and 24 DF,  p-value: < 2.2e-16
```



y-variable was transformed

- Estimated standard error of the estimate (SE_E') *:*

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

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

y-variable was transformed

- Can calculate estimates of r^2 and SE_E for the original y-variable unit, in order to compare to R^2 and SE_E of other equations where the *y* was not transformed, similar to SLR

- Estimated r^2: **I^2 (Fit Index)**

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

  - 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

R-squared values can be too high!

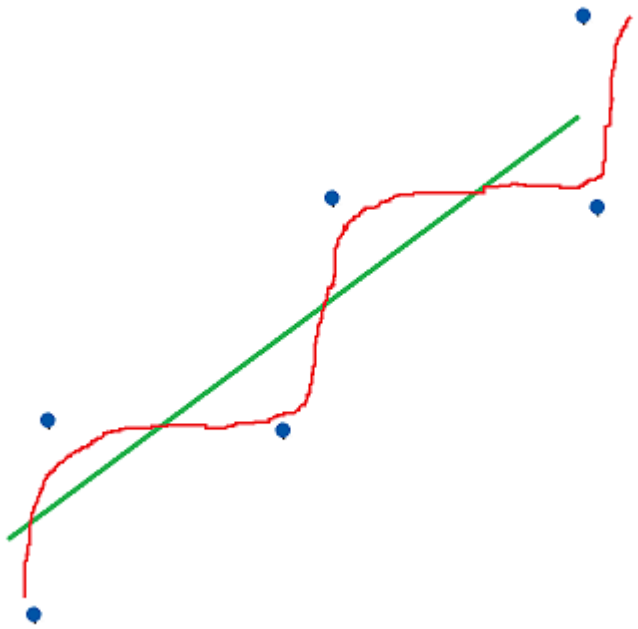
- **increases** every time you add an independent variable to the model.
- **never decreases**
- tempts you to add more.
- Inflated R^2 values actually are ***symptom* of overfit** !

Adjusted R-squared and **Predicted R-squared**

help you fight that impulse to add too many independent variables

Overfit regression models

- Adding **too many predictors** leads to **overfitting** regression model
- Regression coefficients then represent the **random error** in a **sample,** rather than the genuine relationships between the variables in the **population**.
- **Reduces its generalizability** of the model outside the original dataset.



- green line represents true relationship between the variables. The random error inherent in the data causes the data points to fall randomly around the green fit line. T
- red line represents an overfit model. This model is too complex, and it attempts to explain the random error present in the data. ➔ not generalizable to other samples of the population

Need a sample size that is large enough to handle the model complexity

Example: total sample size of 20

- 1-sample t-test to estimate population mean -> good estimate.
- 2-sample t-test to estimate the means of two populations (only ten observations to estimate each mean) -> estimates not so good.
- one-way ANOVA to estimate 3 or more means -> pretty bad estimates.

➔ As the number of observations per estimate decreases (20, 10, 6.7, etc.), the estimates become more erratic.

Need a sample size that is large enough to handle the model complexity

Overfitting a regression model is similar to the example above

= trying to estimate too many parameters from a sample with a fixed sample size.

= similar effect like having a small sample

Leads to **erratic estimates** AND **larger margins of error** (confidence intervals) for both the coefficients and predicted values ➔ **reduces model's precision**

The size of your sample restricts the number of terms that you can safely add to the model before you obtain erratic estimates.

Need a sample size that is large enough to handle the model complexity

Simulation studies indicate that you should have at least **10-15 observations for each term in a linear model**.

The number of terms in a model is the sum of all the independent variables + interactions (see later)

Overfit regression models

excessive number of independent variables

-> overly customized to fit the peculiarities and random noise in your sample rather than reflecting the entire population.

- Produces high R-squared values
- But decreases capability for precise predictions.

Adjusted R^2 value

SSE falls as *m (number of independent variables)* increases, so R^2 rises as more explanatory (independent or predictor) variables are added.

Adjusted R^2 value adjusts by dividing each sum of squares by its associated degrees of freedom

→A **penalty** is added **as you add x-variables** to the equation:

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

Adjusted R-squared

Increases only when the new term improves the model fit more than expected by chance alone.

Decreases when the term doesn't improve the model fit by a sufficient amount.

➔ Use Adjusted R-squared to **compare** the goodness-of-fit for regression models that contain **differing numbers of independent variables**.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

```
> summary(model.volha)

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Residuals:
    Min      1Q  Median      3Q     Max
-52.334 -23.875  -0.104  19.642  59.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.982e+02  4.789e+01  -4.138 0.000372 ***
ba.ha        1.857e+01  7.564e-01  24.546  < 2e-16 ***
stems.ha    -3.124e-02  7.016e-03  -4.453 0.000167 ***
qdbh         7.542e+00  1.740e+00   4.335 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.43 on 24 degrees of freedom
Multiple R-squared: 0.9727,     Adjusted R-squared: 0.9692
F-statistic: 284.6 on 3 and 24 DF,  p-value: < 2.2e-16
```

Multiple coefficient of determination

Adjusted R-squared

ratio Adjusted R-squared / R-Squared

Adjusted R-squared / R-Squared tells you the likely decrease in model fit when the model is applied to new data.

Ideally, adjusted R-squared should be very close to the R-squared for a good fit.

The higher the ratio Adjusted R2 / R2, the better.

Predicted R-squared

calculated by systematically

1) removing each observation from the data set
2) estimating the regression equation
3) determining how well the model predicts the removed observation
4) repeating this for all data points in the dataset

In R: can calculate Predicted R-squared yourself using predicted residual sums of squares (PRESS)

https://rpubs.com/RatherBit/102428

```
predictive R-squared = [1 - (PRESS / sums of squares total)]
```

Predicted R-squared

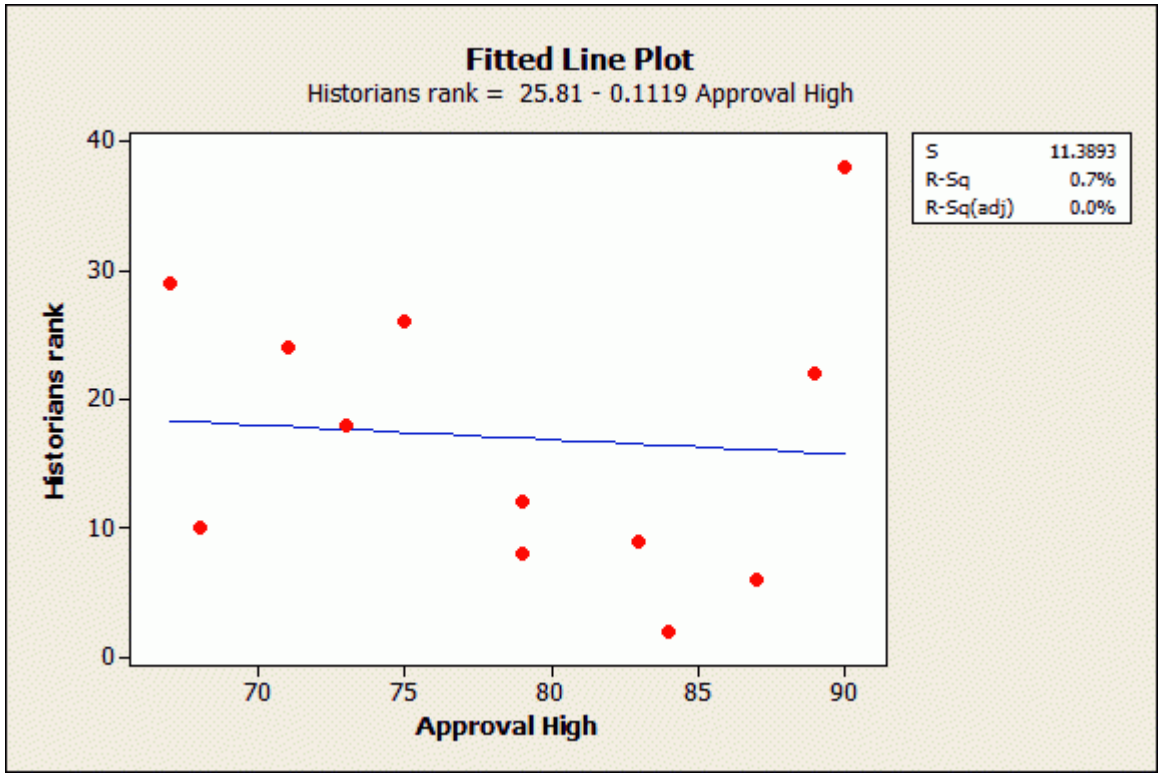
Predicted R-squared value **decreases** if the model fits random noise in the sample = **overfit** model *(because it is not possible to predict random noise!)*

If Predicted R-squared << R-squared

= warning sign that you are overfitting the model

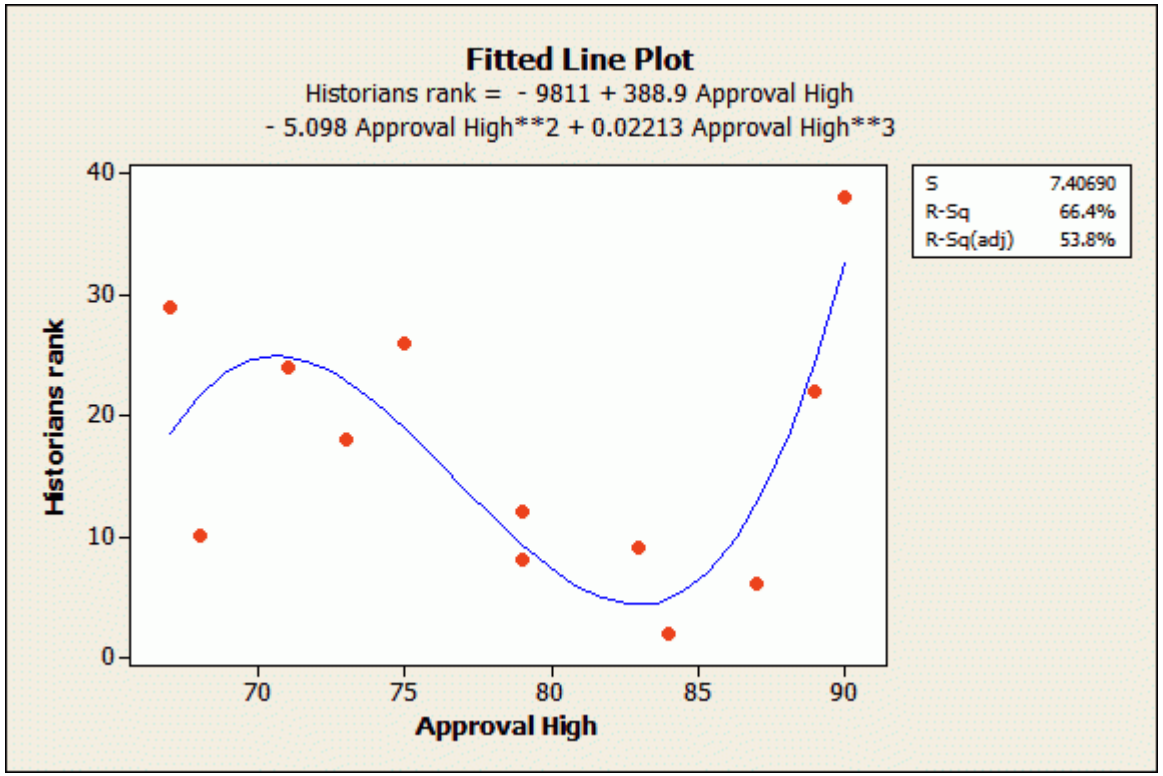
➔ Try reducing the number of terms.

Use Predicted R-squared to determine how well a regression model makes **predictions.**



No relationship between Rank by Historians and Highest Approval rating by population for U.S. presidents

R-squared 0.007



Chasing a high R-squared: we fit the model using a cubic term that provides an S-shape.
R-squared 0.66

But are we fitting real relationships or just playing "connect the dots"?

Regression Analysis: Historians rank versus Approval High

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	867.10	289.034	5.27	0.027
Approval High	1	438.35	438.347	7.99	0.022
Approval High*Approval High	1	460.23	460.225	8.39	0.020
Approval High*Approval High*Approval High	1	481.55	481.552	8.78	0.018
Error	8	438.90	54.862		
Lack-of-Fit	7	430.90	61.557	7.69	0.271
Pure Error	1	8.00	8.000		
Total	11	1306.00			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.40690	66.39%	53.79%	0.00%

Coefficients are statistically significant (p-values < 0.05)
R-squared and **adjusted R-squared** look great
Predicted R-squared << R-squared ➔ model is overfit!

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 y for any x?

Mean Square (MS)

= Sum of Squares divided by it's associated degrees of freedom

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

Called the **Mean squared error**, as would be the average of the squared error if we divided by *n.*

Instead, we divide by *n-m - 1. Why? The degrees of freedom* are *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)

- MSR = SSR/m:

Called the **Mean Square Regression**

Degrees of Freedom=m: m x-variables

Under the assumptions of MLR, this is an estimate of the error variance PLUS a term of variance explained by the regression using *x.*

Expected Mean Squares

(statistical theory provides these results:)

$$E\{MSE\} = \sigma^2$$

$$E\{MSR\} = \sigma^2 + SSreg$$

➔mean of sample distribution of MSE = σ^2

Independent of linear relationship X's~Y

Independent of $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_m = 0$ or not all slopes =0

➔mean of sample distribution of MSR = σ^2 <u>if</u> $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_m = 0$

Regression significant?

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

Analysis of Variance approach

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 testing several slope coefficients

Using an F test statistic: $$F = \frac{SSreg/m}{SSE/(n-m-1)} = \frac{MSreg}{MSE}$$

If F is 1 → no relationship
if F is bigger than 1 → there is a relationship

- 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 true slopes (the population slopes) are likely not equal to zero

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

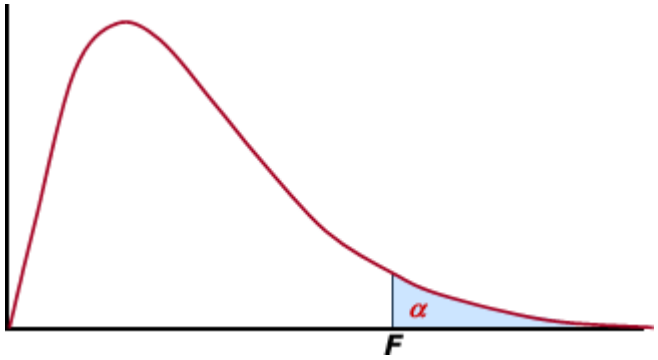
Source	df	SS	MS	F	p-value
Regression	m	$SSreg$	$MSreg = SSreg/\mathrm{m}$	$F = MSreg/MSE$	Prob F> $F_{(\mathrm{m},n-m-1,1-\alpha)}$
Error	n-m-1	SSE	$MSE = SSE/(n$ m-1)		
Total	n-1	SSy			

Is the regression significant?

H0: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_m = 0$ regression is not significant

H1: not all slopes = 0 regression is significant

→ **F-test**



Two ways to look:

1) If F-value > critical F-value, then reject H0

F-value ? → summary(model.volha) → `F-statistic: 284.6`

$F_c = F_{m,\ n-m-1,\ 1-\alpha}$ α=0.05 → using the table (or R function)

m=3; n=28; n-m-1=24

```
> qf(0.95,3,24)
[1] 3.008787
```

2) OR if P-value < α, then reject H0

P-value? → summary(model.volha) → `p-value: < 2.2e-16`

```
> pf(284.6,3,24, lower.tail=FALSE)
[1] 6.942069e-19
```

```
> summary(model.volha)

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Residuals:
    Min      1Q  Median      3Q     Max
-52.334 -23.875  -0.104  19.642  59.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.982e+02  4.789e+01  -4.138 0.000372 ***
ba.ha        1.857e+01  7.564e-01  24.546  < 2e-16 ***
stems.ha    -3.124e-02  7.016e-03  -4.453 0.000167 ***
qdbh         7.542e+00  1.740e+00   4.335 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.43 on 24 degrees of freedom
Multiple R-squared: 0.9727,     Adjusted R-squared: 0.9692
F-statistic: 284.6 on 3 and 24 DF,  p-value: < 2.2e-16
```

DF of SSR

DF of SSE

```
> qf(0.95,3,24)
[1] 3.008787
```

```
> anova(model.volha)
Analysis of Variance Table

Response: vol.ha
          Df  Sum Sq Mean Sq F value    Pr(>F)
ba.ha      1  756843  756843 677.122 < 2.2e-16 ***
stems.ha   1  176537  176537 157.942 4.774e-12 ***
qdbh       1   21009   21009  18.796 0.0002251 ***
Residuals 24   26826    1118
---
Signif. codes:  0 '***' 0

> anovatable <- anova(model.volha)
> ssr <- sum(anovatable[1:3,2])
> ssr
[1] 954388.9
> dfr <- sum(anovatable[1:3,1])
> dfr
[1] 3
> msr <- ssr/dfr
> msr
[1] 318129.6
> mse <- anovatable[4,3]
> mse
[1] 1117.735
> fvalue <- msr/mse
> fvalue
[1] 284.6199
```

partial SSRegressions

MLR ANOVA table : R decomposes SSR in parts that sum up to SSR!

SSE for the full model (24 df)

Same result as in summary()

Estimated Standard Errors for the Slope and Intercept

Under the assumptions, we can obtain an unbiased estimate 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 β_o: $b_0 \pm t_{1-\alpha/2,n-m-1} \times s_{b_0}$

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

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]

Hypothesis Tests for one of the True Slopes or Intercept

Test statistic:

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

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

Reject H0 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.

III. Regression is significant – Now: which x-variables are significant, given the other variables in the equation?

→ **T-test**

```
> summary(model.volha)

Call:
lm(formula = vol.ha ~ ba.ha + stems.ha + qdbh)

Residuals:
    Min      1Q  Median      3Q     Max
-52.334 -23.875  -0.104  19.642  59.953

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.982e+02  4.789e+01  -4.138 0.000372 ***
ba.ha        1.857e+01  7.564e-01  24.546  < 2e-16 ***
stems.ha    -3.124e-02  7.016e-03  -4.453 0.000167 ***
qdbh         7.542e+00  1.740e+00   4.335 0.000225 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.43 on 24 degrees of freedom
Multiple R-squared: 0.9727,     Adjusted R-squared: 0.9692
F-statistic: 284.6 on 3 and 24 DF,  p-value: < 2.2e-16
```

$H0: \beta_0 = 0$

$Ha: \beta_0 \neq 0$

$$\text{t-statistic}: \frac{-(198.2e+02)-0}{47.89e+01}$$

if $|t| > t_{critical}$, we reject H0

this is the same as $Pv < \alpha$

We don't really care about the intercept: no x-variable is associated with it

Partial F-test

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]

H0: the dropped variables are not significant, the slopes are all zero

H1: not all the slopes of the dropped variables are zero

Partial F-test

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

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

Partial F-test

- 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 partial F is larger than the critical F-value from the table, we reject H0 (not likely true). The contribution of the dropped variables is significant, in that one or more of the true slopes are likely not equal to zero

Regression is significant – Now: which x-variables are significant , given the other variables in the equation?

→ partial F-test:

In R we can perform partial F-tests by fitting both the reduced and full models separately and thereafter comparing them using anova (reduced, full).

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha2<-lm(vol.ha~ba.ha+stems.ha)
> anova(model.volha2,model.volha)                # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     25 47835
2     24 26826  1     21009 18.796 0.0002251 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows the results of the partial F-test. Since F=18.796 (p-value=0.0002251) we can reject the null hypothesis ($\beta 3 = 0$) at the 5% level of significance. It appears that the variable qdbh contributes significant information to the vol/ha once the variables ba.ha and stems/ha have been taken into consideration.

Regression is significant – Now: which x-variables are significant , given the other variables in the equation?

→ partial F-test:

In R we can perform partial F-tests by fitting both the reduced and full models separately and thereafter comparing them using anova (reduced, full).

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha2<-lm(vol.ha~ba.ha+stems.ha)
> anova(model.volha2,model.volha)                    # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     25 47835
2     24 26826  1     21009 18.796 0.0002251 ***
---
Signif. codes:  0 '***' 0.001
```

Degrees of freedom of Sum of Squares Error (Residuals):
Reduced model: n-m-1 with m=2 → 25
Full model: n-m-1 with m=3 → 24

The output shows the results of the partial F-test. Since F=18.796 (p-value=0.0002251) we can reject the null hypothesis ($\beta 3 = 0$) at the 5% level of significance. It appears that the variable qdbh contributes significant information to the vol/ha once the variables ba.ha and stems/ha have been taken into consideration.

Regression is significant – Now: which x-variables are significant , given the other variables in the equation?

→ partial F-test:

In R we can perform partial F-tests by fitting both the reduced and full models separately and thereafter comparing them using anova (reduced, full).

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha2<-lm(vol.ha~ba.ha+stems.ha)
> anova(model.volha2,model.volha)                    # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh
  Res.Df   RSS Df Sum of Sq
1     25 47835
2     24 26826  1     21009 1
---
Signif. codes:  0 '***' 0.001
```

RSS: Residuals Sum of Squares
= SSE: Sum of Squares Error (Residuals):
-Reduced model
-Full model
→ MSE (full) = 26826/24 =1117.75

The output shows the results of the partial F-test. Since F=18.796 (p-value=0.0002251) we can reject the null hypothesis ($\beta 3 = 0$) at the 5% level of significance. It appears that the variable qdbh contributes significant information to the vol/ha once the variables ba.ha and stems/ha have been taken into consideration.

Regression is significant – Now: which x-variables are significant , given the other variables in the equation?

→ partial F-test:

In R we can perform partial F-tests by fitting both the reduced and full models separately and thereafter comparing them using anova (reduced, full).

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha2<-lm(vol.ha~ba.ha+stems
> anova(model.volha2,model.volha)
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     25 47835
2     24 26826  1     21009 18.796 0.0002251 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial SSregression for the omitted term:
= SSR(full)-SSR(reduced)
= SSE(reduced)-SSE(full)
= 47835-26826

Degrees of freedom of Partial SSregression
= number of omitted terms: 1

The output shows t (p-value=0.0002251) we can reject the null hypothesis (β3 = 0) at the 5% level of significance. It appears that the variable qdbh contributes significant information to the vol/ha once the variables ba.ha and stems/ha have been taken into consideration.

Regression is significant – Now: which x-variables are significant , given the other variables in the equation?

→ partial F-test:

In R we can perform partial F-tests by fitting both the reduced and full models separately and thereafter comparing them using anova (reduced, full).

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> model.volha2<-lm(vol.ha~ba.ha+stems.ha)
> anova(model.volha2,model.volha)                # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     25 47835
2     24 26826  1     21009 18.796 0.0002251 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F = [(SSR(full)-SSR(reduced))/r]/MSE(full)
F = [(SSE(reduced)-SSE(full))/r]/MSE(full)
F = [21009/1]/[26826/24]
F = 18.796

The output shows the results of the partial F-test. Since F=18.796 (p-value=0.0002251) we can reject the null hypothesis ($\beta 3 = 0$) at the 5% level of significance. It appears that the variable qdbh contributes significant information to the vol/ha once the variables ba.ha and stems/ha have been taken into consideration.

→ partial F-test:

In R, Partial F-values for single x-variables are given in the anova table of the full model!

```
> anova(model.volha)
Analysis of Variance Table

Response: vol.ha
          Df Sum Sq Mean Sq F value    Pr(>F)
ba.ha      1 756843  756843 677.122 < 2.2e-16 ***
stems.ha   1 176537  176537 157.942 4.774e-12 ***
qdbh       1  21009   21009  18.796 0.0002251 ***
Residuals 24  26826    1118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> anova(model.volha)
Analysis of Variance Table

Response: vol.ha
          Df  Sum Sq Mean Sq F value    Pr(>F)
ba.ha      1  756843  756843 677.122 < 2.2e-16 ***
stems.ha   1  176537  176537 157.942 4.774e-12 ***
qdbh       1   21009   21009  18.796 0.0002251 ***
Residuals 24   26826    1118
---
Signif. codes:  0 '***' 0

> anovatable <- anova(model.volha)
> ssr <- sum(anovatable[1:3,2])
> ssr
[1] 954388.9
> dfr <- sum(anovatable[1:3,1])
> dfr
[1] 3
> msr <- ssr/dfr
> msr
[1] 318129.6
> mse <- anovatable[4,3]
> mse
[1] 1117.735
> fvalue <- msr/mse
> fvalue
[1] 284.6199
```

partial SSRegressions

MLR ANOVA table : R decomposes SSR in parts that sum up to SSR!

SSE for the full model (24 df)

Same result as in summary()

→ partial F-test:

Sometimes we are interested in simultaneously testing whether a certain subset of the coefficients are equal to 0 (e.g. β3 = β4 = 0).

```
> model.volha3<-lm(vol.ha~ba.ha+stems.ha+qdbh+age+si)
> anova(model.volha,model.volha3)                # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha + qdbh
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh + age + si
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1     24 26826
2     22 23200  2    3625.6 1.719 0.2025
> model.volha4<-lm(vol.ha~ba.ha+stems.ha+qdbh+age+topht)
> anova(model.volha,model.volha4)                # partial F test to compare the two
    nested models
Analysis of Variance Table

Model 1: vol.ha ~ ba.ha + stems.ha + qdbh
Model 2: vol.ha ~ ba.ha + stems.ha + qdbh + age + topht
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     24 26826
2     22 11744  2     15082 14.127 0.0001132 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

β3 = β4 = 0, *if the other variables are included in the model*

Not all of β3 and β4 are 0, *if the other variables are included 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

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

For one possible new y-value given a particular value 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

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

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

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

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

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

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

Given stems/ha=300, qdbh=20 cm, and ba/ha=20 m2/ha, what is the estimated volume per ha? How would you get a CI for this estimate?

The function **predict()** can be used to make

- confidence intervals **for the mean response** : option interval="confidence".
- prediction intervals: option interval="prediction"

By default this makes 95% confidence and prediction intervals.

If you instead want to make a 99% confidence or prediction interval use the option level=0.99.

```
> model.volha<-lm(vol.ha~ba.ha+stems.ha+qdbh)
> predict(model.volha,data.frame(stems.ha=300, ba.ha=20,
    qdbh=20),interval="confidence")
       fit      lwr      upr
1 314.6161 279.4611 349.7711
> predict(model.volha,data.frame(stems.ha=300, ba.ha=20,
    qdbh=20),interval="prediction")
       fit      lwr      upr
1 314.6161 237.1754 392.0569
```

A 95% confidence interval of the mean is given by (279, 349)

A 95% prediction interval is given by (237,392). Note that this is quite a bit wider than the confidence interval, indicating the variation about the mean.

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; try a number of equations.
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, lowest F, 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

Variable selection

In many situations the set of predictor variables to be included is not predetermined

→ selecting predictor variables becomes part of the analysis.

Two main approaches towards variable selection:

1. all possible regressions approach

   considers all possible subsets of the pool of explanatory variables and finds the model that best fits the data according to some criteria (e.g. Adjusted R2, AIC and BIC). These criteria assign scores to each model and allow us to choose the model with the best score.

2. automatic methods (algorithm)

   useful when the number of explanatory variables is large and it is not feasible to fit all possible models.

Methods to aid in selecting predictor (x) variables

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

1. **R^2 (or Adjusted R^2).** The equation is fitted for a number of combinations of the x-variables to predict y. The ones with the highest R^2 are reported. CAUTION: You must check the assumptions of these fitted equations by fitting the equation with variables given. If assumptions are NOT met, these are NOT candidate models EVEN with a high R^2. ALSO, consider costs of measuring the x-variables, significance of the x-variables (given the other variables) etc. This only gives some ideas of models to try.

Methods to aid in selecting predictor (x) variables

2. Stepwise.

1) The most important variable is added to the model (highest partial F-value or absolute value of t; has lowest p-value).
2) Each of the other variables are added; the next most important variable is added to the model
3) Repeat Step 2
4) At any time, a variable already entered in, may become not significant. Drop it, and continue with Step 2.
5) Continue until all variables in the regression are significant, and the ones that are not in the equation are not significant, given the ones that are in the equation.

NOTES:

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

Methods to aid in selecting predictor (x) variables

3. Backwards Stepwise:

1) All x-variables are added to the model
2) Check to see if variables are not significant given the other variables in the equation (use partial F-test or t-test)
3) If all x-variables are significant given the other variables, stop. Otherwise, drop the variable with the lowest partial F-value (highest p-value)
4) Repeat step 2, until all variables in the equation are significant, given the other variables that are in the equation

NOTES:

- This again just gives candidate models. You must check whether the assumptions are met and do a full assessment of the regression results
- Unlike "stepwise", once a variable is dropped, it cannot come back in, even if it might be significant with a different set of x-variables than when it was dropped.

Methods to aid in selecting predictor (x) variables

4. **Forward Stepwise:** This is the same as Stepwise, EXCEPT, that once a x-variable is added to the model, it is not removed, even if it becomes non-significant at a particular step in the process.

NOTES:

- This again 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 variables until i) there are no more variables or ii) the remaining variables are not significant given the other variables.

R-code Stepwise

```
> null <- lm(vol.ha ~ 1,data = standdat)
> summary(null)

Call:
lm(formula = vol.ha ~ 1, data = standdat)

Residuals:
    Min      1Q  Median      3Q     Max
-332.54 -122.99  -53.24  111.04  380.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   535.54      36.03   14.87  1.6e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 190.6 on 27 degrees of freedom
```

Script 5_MLR_variableselection.R

R-code Stepwise

```
> null <- lm(vol.ha ~ 1,data = standdat)
> summary(null)

Call:
lm(formula = vol.ha ~ 1, data = standdat)

Residuals:
    Min      1Q  Median      3Q
-332.54 -122.99  -53.24  111.04  3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   535.54      36.03   14.87  1.6e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 190.6 on 27 degrees of freedom
```

Script 5_MLR_variableselection.R

```
> mean(vol.ha)
[1] 535.5393
```

```
> sem <- sd(vol.ha)/sqrt(length(vol.ha))
> sem
[1] 36.02642
```

```
> t.value <- (mean(vol.ha)-0)/sem
> t.value
[1] 14.86518
```

```
> sd(vol.ha-mean(vol.ha))
[1] 190.6339
```

```
> 2*pt(t.value,df=length(vol.ha)-1,lower.tail=FALSE)
[1] 1.601009e-14
```

```
> full <- lm(vol.ha ~ .,data = standdat)
> summary(full)
Call:
lm(formula = vol.ha ~ ., data = standdat)
Residuals:
    Min      1Q  Median      3Q     Max
-32.269 -15.239  -1.759  14.183  51.850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.523e+02  7.991e+01  -6.912  7.9e-07 ***
age          3.135e-01  4.665e-01   0.672 0.508860
si           2.778e+00  3.814e+00   0.728 0.474492
ba.ha        1.648e+01  6.747e-01  24.423  < 2e-16 ***
stems.ha    -1.443e-02  6.326e-03  -2.281 0.033114 *
topht        2.316e+01  4.991e+00   4.640 0.000141 ***
qdbh         2.808e-03  2.041e+00   0.001 0.998915
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 23.35 on 21 degrees of freedom
Multiple R-squared:  0.9883,    Adjusted R-squared:  0.985
F-statistic: 296.3 on 6 and 21 DF,  p-value: < 2.2e-16
```

R-code Stepwise

```
null <- lm(vol.ha ~ 1,data = standdat)
summary(null)
full <- lm(vol.ha ~ .,data = standdat)
summary(full)

step.model1 <- step(null, scope=list(lower=null, upper=full), direction="forward")
summary(step.model1)
step.model2 <- step(full,direction=c("backward"))
summary(step.model2)
step.model3 <- step(null, scope = list(upper=full), direction=c("both"))
summary(step.model3)
```

AIC represents a trade-off between the goodness of fit of the model and the complexity of the model.
AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that.

R-code R square: all subsets regression

```
> library(leaps)#Regression subset selection, including exhaustive
search.
> leaps<-regsubsets(vol.ha
~age+si+ba.ha+stems.ha+topht+qdbh,data=standdat,nbest=10)
> # view results
> summary(leaps)
Subset selection object
Call: regsubsets.formula(vol.ha ~ age + si + ba.ha + stems.ha + topht +
    qdbh, data = standdat, nbest = 10)
6 Variables  (and intercept)
         Forced in Forced out
age          FALSE      FALSE
si           FALSE      FALSE
ba.ha        FALSE      FALSE
stems.ha     FALSE      FALSE
topht        FALSE      FALSE
qdbh         FALSE      FALSE
10 subsets of each size up to 6
Selection Algorithm: exhaustive
```

R-code R square: all subsets regression

```
age si  ba.ha stems.ha topht qdbh
1  ( 1 ) " " " " " "*"  " "    " "  " "
1  ( 2 ) " " " " " " "  " "    "*"  " "
1  ( 3 ) "*" " " " " "  " "    " "  " "
1  ( 4 ) " " " " " " "  " "    " "  "*"
1  ( 5 ) " " " " " " "  "*"    " "  " "
1  ( 6 ) " " "*" " " "  " "    " "  " "
2  ( 1 ) " " " " " "*"  " "    "*"  " "
2  ( 2 ) " " " " " "*"  "*"    " "  " "
2  ( 3 ) " " " " " "*"  " "    " "  "*"
....
5  ( 1 ) "*" "*" "*" "*"  "*"    "*"  " "
5  ( 2 ) " " "*" "*" "*"  "*"    "*"  "*"
5  ( 3 ) "*" " " "*" "*"  "*"    "*"  "*"
5  ( 4 ) "*" "*" "*" "*"  " "    "*"  "*"
5  ( 5 ) "*" "*" "*" "*"  "*"    " "  "*"
5  ( 6 ) "*" "*" " " "*"  "*"    "*"  "*"
6  ( 1 ) "*" "*" "*" "*"  "*"    "*"  "*«
```

The best model with two predictors contains stems.ha and topht

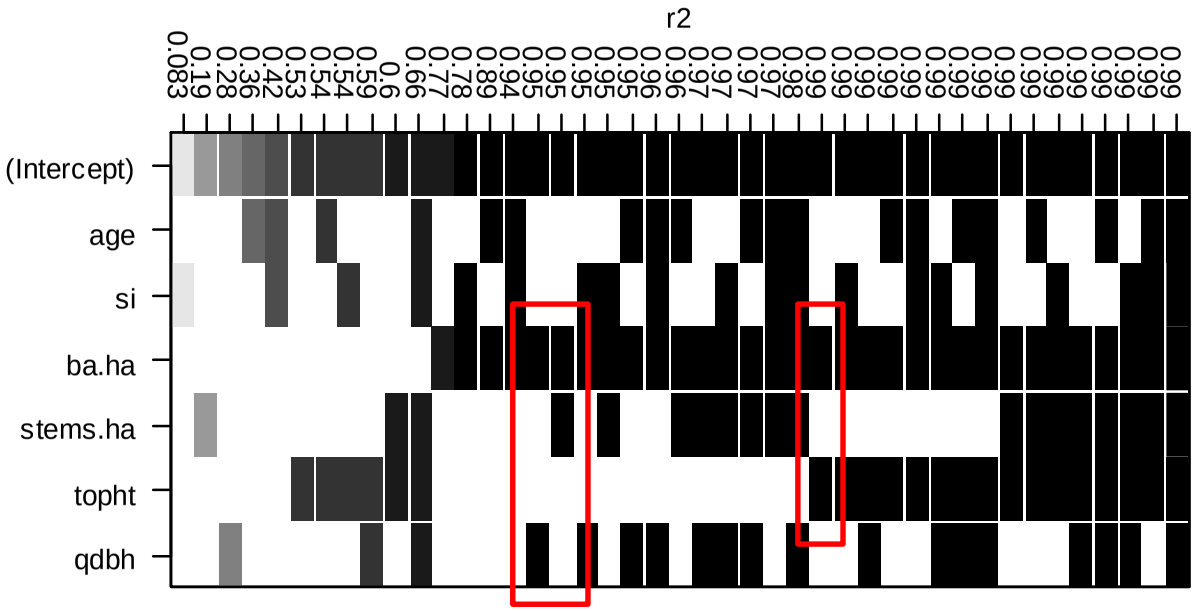
R-code R square: all subsets regression

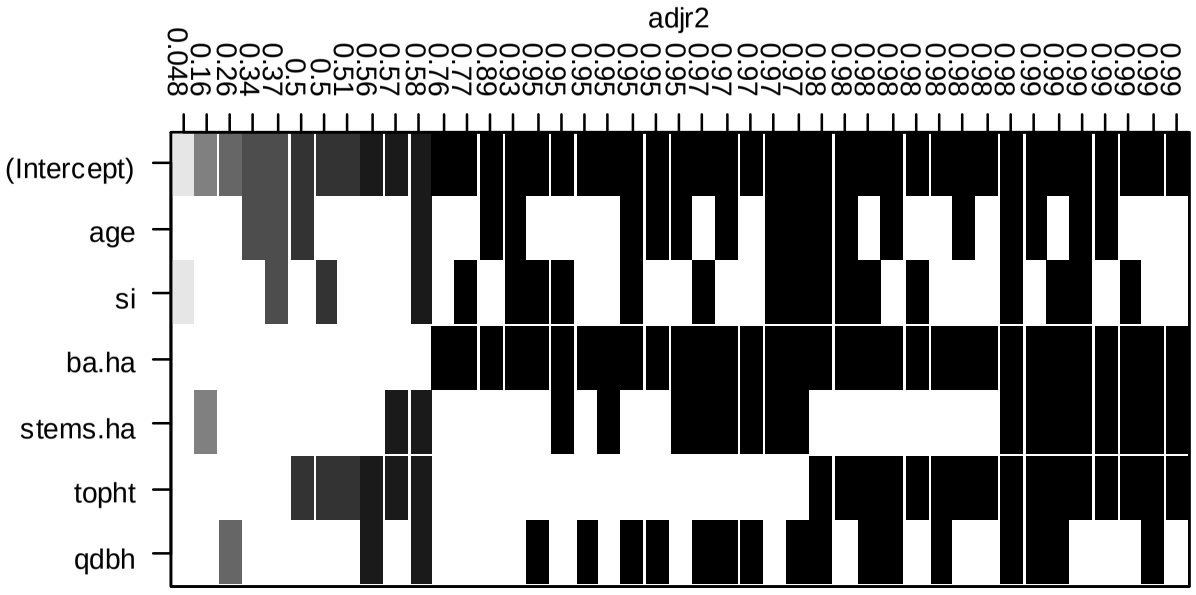
Plot a table of models showing variables in each model.

•This is particularly useful when there are more than ten models and the simple table produced by summary.regsubsets is too big to read.

•models are ordered by the selection statistic: "adjr2“or "r2"

```
> par(mfrow=c(1,2),cex=0.7)
> plot(leaps,scale="r2")
> plot(leaps,scale="adjr2")
> par(mfrow=c(1,1),cex=1)
```

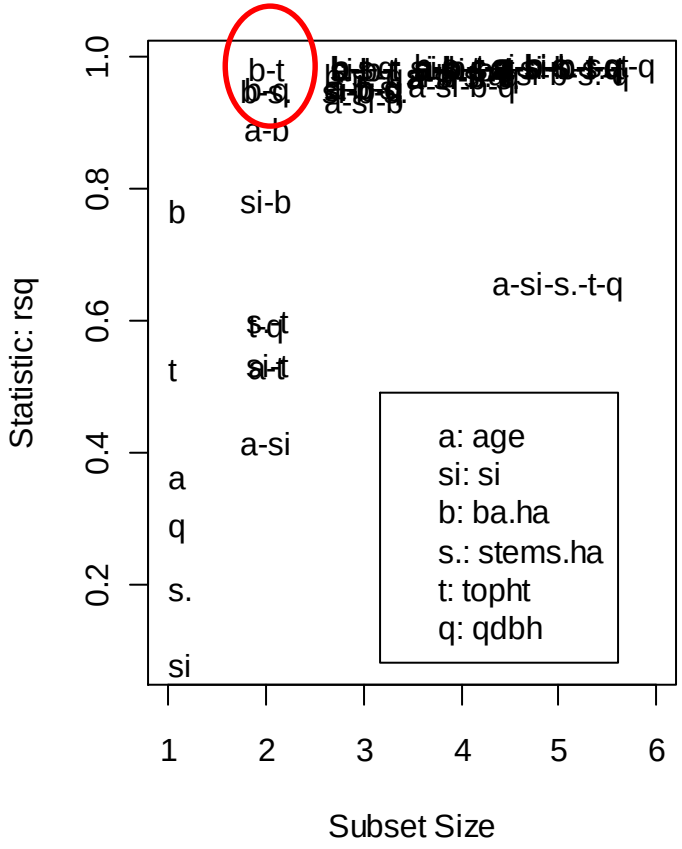

r2
0.083
0.19
0.28
0.36
0.42
0.53
0.54
0.54
0.56
0.6
0.69
0.77
0.78
0.86
0.84
0.95
0.95
0.95
0.95
0.96
0.96
0.97
0.97
0.97
0.97
0.98
0.98
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
(Intercept)
age
si
ba.ha
stems.ha
topht
qdbh

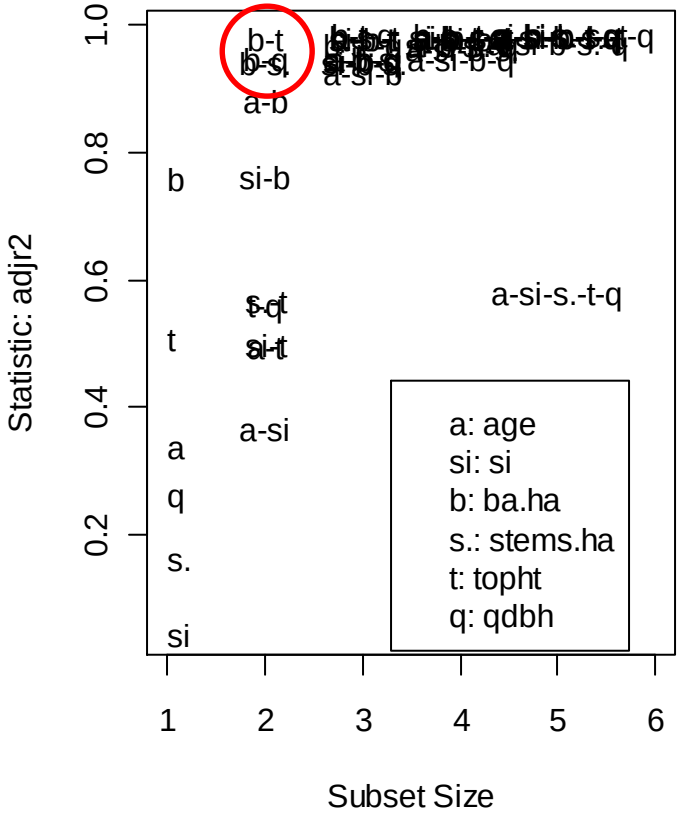

adjr2
0.048
0.16
0.26
0.34
0.37
0.5
0.5
0.51
0.56
0.57
0.58
0.76
0.77
0.86
0.93
0.95
0.95
0.95
0.95
0.95
0.96
0.97
0.97
0.97
0.97
0.97
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
0.99
(Intercept)
age
si
ba.ha
stems.ha
topht
qdbh

R-code R square: all subsets regression

Plot statistic by subset size

```
> library(car)
> par(mfrow=c(1,2))
> subsets(leaps, statistic="rsq")
> subsets(leaps, statistic="adjr2")
> par(mfrow=c(1,1))
```


Statistic: rsq
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
Subset Size
b-t
b-si
a-b
si-b
a-si-s.-t-q
t-q
si-t
a-si
b
t
a
q
s.
si
a: age
si: si
b: ba.ha
s.: stems.ha
t: topht
q: qdbh


Statistic: adjr2
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
Subset Size
b-t
b-si
a-b
si-b
a-si-s.-t-q
t-q
si-t
a-si
b
t
a
q
s.
si
a: age
si: si
b: ba.ha
s.: stems.ha
t: topht
q: qdbh

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