PGS.TS. BẢO HUY

PHÂN TÍCH THỐNG KÊ TRONG NGHIÊN CỨU THỰC NGHIỆM LÂM NGHIỆP – QUẢN LÝ TÀI NGUYÊN RỪNG – MÔI TRƯỜNG

Sử dụng các phần mềm Statgraphics, SPSS và Excel

Năm 2015

MỤC LỤC

LỜI NÓI ĐẦU

Tài liệu này được biên soạn phục vụ cho việc ứng dụng thống kê trong nghiên cứu lâm nghiệp, quản lý tài nguyên thiên nhiên cho nhà nghiên cứu, quản lý nghiên cứu. Mục đích là giúp cho thành viên tham gia phân tích, xử lý số liệu thống kê trên máy vi tính bằng các phần mềm thống kê để thực hiện các đề tài nghiên cứu cũng như ứng dụng vào thực tiễn.

Có rất nhiều phần mềm ứng dụng để xử lý thống kê như SPSS, Statgraphics Plus, Excel, R studio. Các phần mềm thống kê chuyên dụng và phổ biến trên thế giới là Statgrahics, SPSS, hoặc phần mềm mã nguồn mở R…. Đây là các phần mềm thống kê được ứng dụng rộng trong hầu hết các lĩnh vực nghiên cứu, phân tích dữ liệu của nhiều ngành khác nhau về xã hội, tự nhiên. Ứng dụng mạnh của các phần mềm này là phân tích hầu hết các chức năng thống kê cho nhiều lĩnh vực nghiên cứu, minh họa bằng đồ thị, biểu đồ. Ngoài ra Microsoft Excel được mọi người biết đến khi nói đến công cụ bảng tính, tính toán..., nhưng những chức năng chuyên sâu về ứng dụng thống kê cũng khá đầy đủ.

Tài liệu này sẽ không đi sâu vào lý thuyết xác suất thống kê, mà thiên về hướng ứng dụng đơn giản, dễ hiểu, kèm theo các ví dụ để người đọc có thể thực hành các chức năng xử lý, phân tích dữ liệu một cách nhanh chóng, thuận tiện trong hoạt động quản lý và nghiên cứu tập trung cho lâm nghiệp, quản lý tài nguyên rừng và môi trường. Đồng thời tài liệu này cũng không giới thiệu sử dụng từng phần mềm thống kê như SPSS, Statgraphics, … mà chỉ chọn lọc các chức năng thích hợp của chung cho từng nội dung nghiên cứu thực nghiệm trong phạm vi lâm nghiệp, sinh học, môi trường rừng.

1 TỔNG QUÁT VỀ CHỨC NĂNG XỬ LÝ THỐNG KÊ CỦA EXCEL, STATGRAPHICS VÀ SPSS

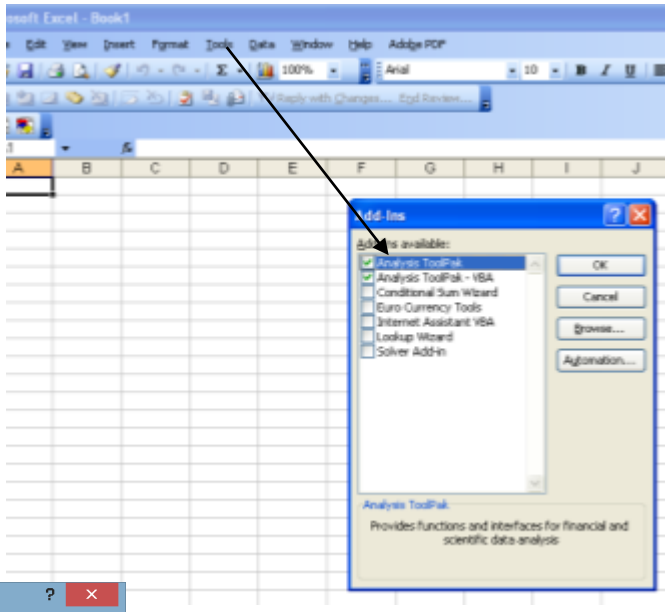
1.1 Tổng quát về phần xử lý thống kê trong Excel

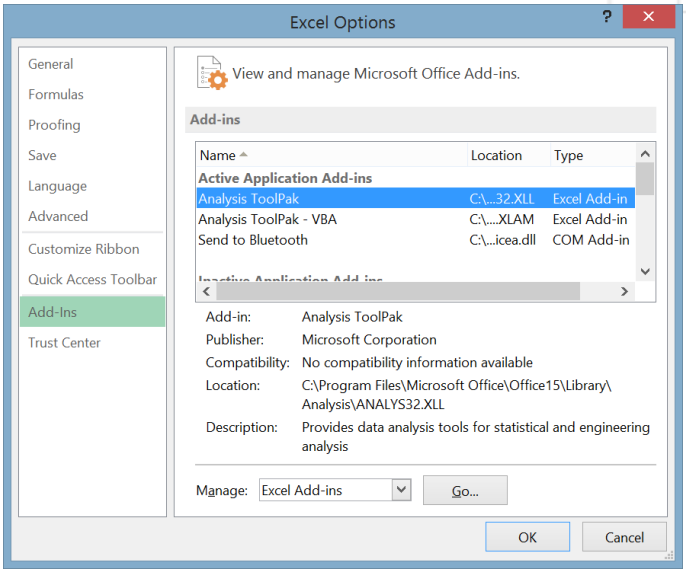
Excel thiết kế sẵn một số chương trình để xử lý số liệu và phân tích thống kê cơ bản ứng dụng trong nhiều lĩnh vực:

- Chức năng xử lý số liệu, tạo bảng tổng hợp dữ liệu: Sắp xếp, tính toán nhanh các bảng tổng hợp từ số liệu thô,...
- Chức năng của các hàm: Cung cấp hàng loạt các hàm về kỹ thuật, thống kê, kinh tế tài chính, hàm tra các chỉ tiêu thống kê như t, F, χ^2
- Chức năng Data Analysis: Dùng để phân tích thống kê như phân tích các đặc trưng mẫu, tiêu chuẩn t để so sánh sự sai khác, phân tích phương sai, ước lượng các tương quan hồi quy
- Phân tích mô hình tương quan hoặc hồi quy để dự báo các thay đổi theo thời gian ngay trên đề thị.

Lưu ý: Về việc cài đặt chương trình phân tích dữ liệu ***(Data Analysis)*** *trong Excel:*

- *Khi cài đặt phần mềm Excel phải thực hiện trong chế độ chọn lựa cài đặt, sau đó phải chọn mục: Add-Ins và Analysis Toolpak.*
- *Khi chạy Excel lần đầu cần mở chế độ phân tích dữ liệu bằng cách: Menu Tools/Add-Ins và chọn Analysis Toolpak-OK. (Đối với MS. Office 2003)*





Đối với MS. Office 2007 trở đi, tiến hành mở chế độ phân tích thống kê như sau: File/Option/Add-ins và chọn Analysis ToolPak – Go, sau đó kích chọn chức năng Analysis ToolPak trong hộp thoại - OK.

Trong thực tế quản lý xử lý dữ liệu, việc khai thác hết tiềm năng ứng dụng của Excel cũng mang lại hiệu quả tốt mà không nhất thiết phải tìm kiếm thêm một phần mềm chuyên dụng nào khác. Vấn đề đặt ra là xác định chiến lược ứng dụng và khai thác đúng và sâu các công cụ chức năng sẵn có ở một phần mềm phổ biến ở bất kỳ một vi tính cá nhân nào.

Một số hàm thông dụng trong thống kê:

- Tính tổng: =Sum(dãy đs).
- Tổng bình phương: =Sumq(dãy đs).
- Trung bình: =Average(dãy đs).
- Lấy giá trị tuyệt đối: =Abs(đs).
- Trị lớn nhất, nhỏ nhất: =Max(dãy đs), Min(dãy đs).
- Các hàm lượng giác: =Cos(đs), =Sin(đs), =tan(đs).
- Hàm mũ, log: =Exp(đs), =Ln(đs), =Log(đs).
- Căn bậc 2: =Sqrt(đs)..
- Sai tiêu chuẩn mẫu chưa hiệu đính: =Stdevp(dãy đs); đã hiệu đính =Stdev(dãy đs).
- Phương sai mẫu chưa hiệu đính: =Varp(dãy đs); đã hiệu đính =Var(dãy đs).
- Giai thừa: =Fact(n).
- Số Pi: =Pi().

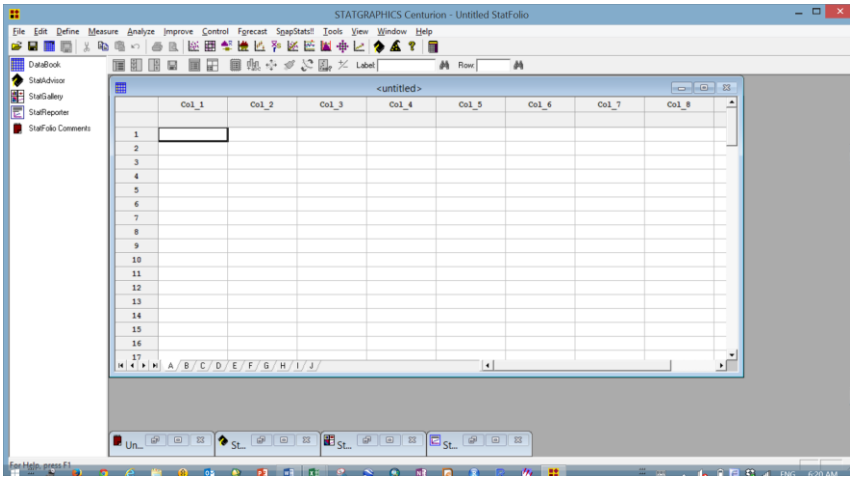
Tra các giá trị theo các tiêu chuẩn thống kê T, F, χ2:

- Chọn 1 ô lấy giá trị tra.
- Kích nút fx trên thanh công cụ chuẩn. Trong hộp thoại Function Category, chọn Statistical.
- Trong mục Function name, chọn 1 trong các hàm:
  Hàm Tinv: để tra T.
  Hàm Chiinv: để tra χ^2.
  Hàm Finv: để tra F.
  Bấm Next.
- Trong hộp thoại tiếp theo: Function Wizard chọn:
  1. Probability (fx): Gõ vào mức ý nghĩa α=0.05 ; 0.01 hay 0.001.
  2. Degrees Freedom (fx): Gõ vào bậc tự do. Đối với tiêu chuẩn F cần đưa vào 2 độ tự do.
  3. Finish.

1.2 Tổng quát về phần mềm xử lý thống kê Statgraphics Centuiron version 15.1.02

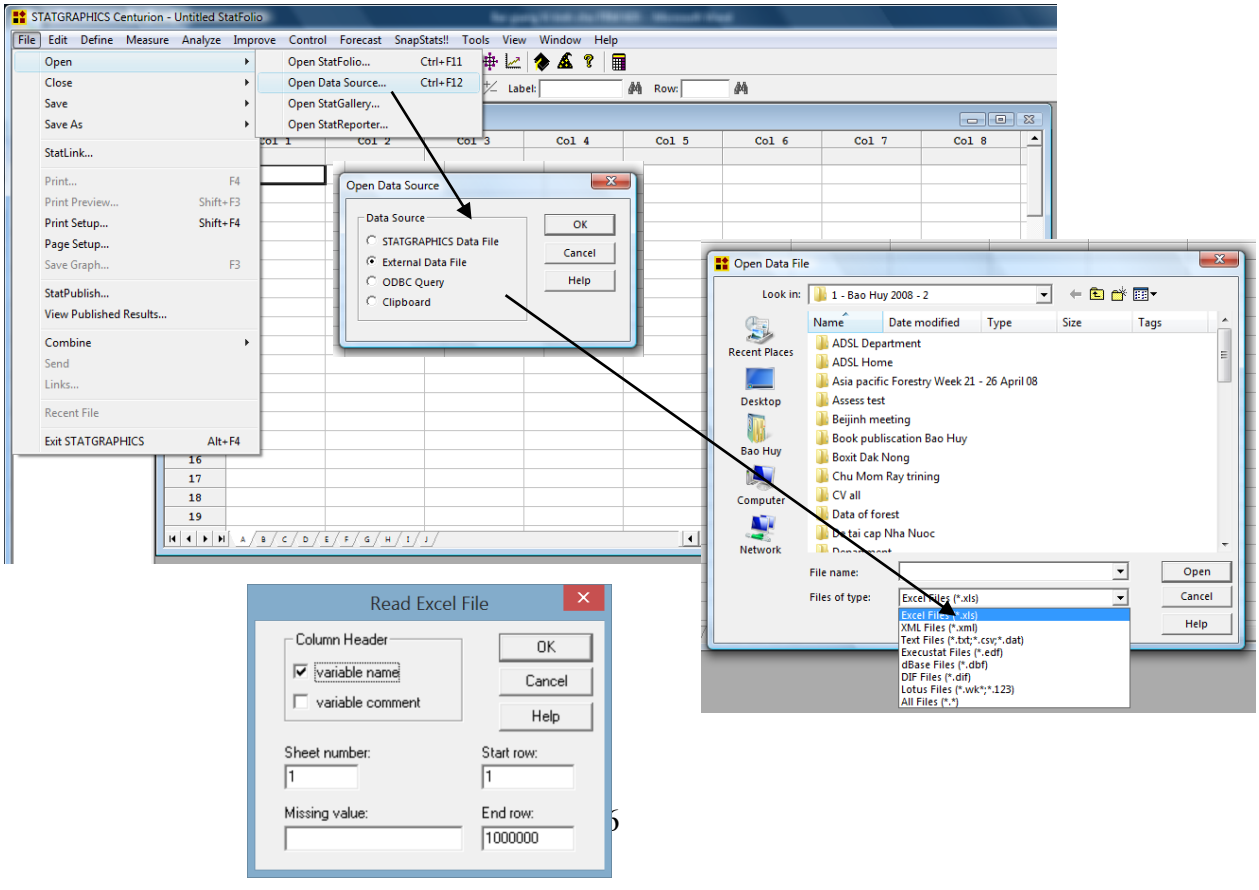
Đây là một phần mềm chuyên dụng trong xử lý thống kê, bao gồm các chức năng:

- Tạo lập cơ sở dữ liệu dưới dạng bảng tính
- Tính toán các đặc trưng mẫu, vẽ sơ đồ, đồ thị quan hệ
- So sánh hai hay nhiều mẫu bằng các tiêu chuẩn thống kê t, U, F và nhiều tiêu chuẩn phi tham số khác.
- Phân tích phương sai ANOVA.
- Kiểm tra tính chuẩn của dữ liệu và đổi biến số.
- Thiết lập các mô hình hồi quy tuyến tính hay phi tuyến tính từ một cho đến nhiều lớp, tổ hợp biến. Với cách xử lý đa dạng để chọn lựa được các biến ảnh hưởng đến một hậu qủa (biến phụ thuộc).



Giao tiếp trong Statgraphics Centurion, số liệu đầu vào có thể được nhập trực tiếp trong file bảng tính và cơ sở dữ liệu; song với các làm này đôi khi không thuận tiện trong các bước xử lý số liệu thô như đổi biến số, tính các biến trung gian, mã hóa biến số. Do đó thông thường nên tạo lập cơ sở dữ liệu trong bảng tính Excel để có thể sử dụng những chức năng bảng tính mạnh của nó trong xử lý dữ liệu thô, tạo lập cơ sở dữ liệu; sau đó sẽ nhập vào Statgraphics Centurion để tính toán, thiết lập mô hình, ... Cơ sở dữ liệu lập trong Excel cần lưu dưới dạng phiên bản của Excel 97 – 2003, vì nó chưa nhận được file Excel ở version từ 2010 - 2012.

Sau khi nhập dữ liệu trong Excel 97-2003, đóng file của Excel và mở nó trong Statgrahics Centurion như sau: File/Open/Open Data Source; chọn External Data File – OK. Trong hộp thoại mở file, chọn kiểu file Excel và chọn file cần mở đã tạo trước đó.
Có thể file excel có nhiều sheet, chọn số thứ tự sheet number và hàng bắt đầu tiêu đề của trường (Start row).



1.3 Tổng quát về phần mềm xử lý thống kê SPSS Statistics version 20

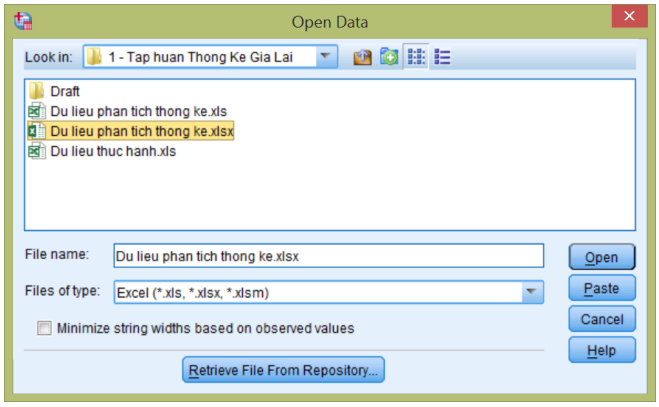
Đây là một phần mềm chuyên dụng trong xử lý thống kê, bao gồm các chức năng gần giống như Statgraphics, tuy nhiên có ưu nhược điểm khi so sánh với nhau:

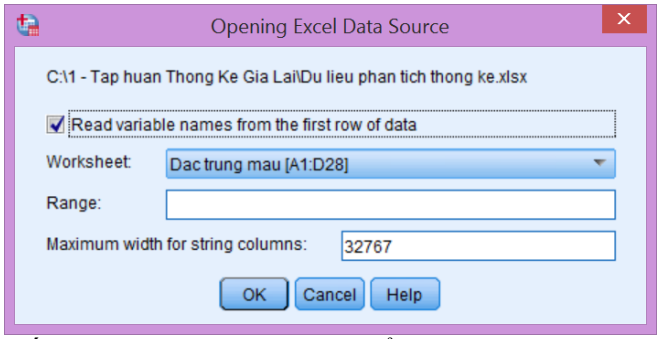
- Ưu điểm SPSS so với Stat:
    - Mã hóa biến số định tính
    - Có các chức năng phân tích so sánh phi tham số
    - Có chức năng lập mô hình hồi quy theo trong số Weight
- Nhược điểm SPSS so với Stat:
    - Không có tư vấn về kết quả phân tích thống kê
    - Không đổi biến số trực tiếp trong phân tích thống kê

Giao tiếp trong SPSS, số liệu đầu vào có thể được nhập trực tiếp trong file bảng tính và cơ sở dữ liệu; song với các làm này đôi khi không thuận tiện trong các bước xử lý số liệu thô như đổi biến số, tính các biến trung gian. Do đó thông thường nên tạo lập cơ sở dữ liệu trong bảng tính Excel để có thể sử dụng những chức năng bảng tính mạnh của nó trong xử lý dữ liệu thô, tạo lập cơ sở dữ liệu; sau đó sẽ nhập vào SPSS để tính toán, thiết lập mô hình, ...

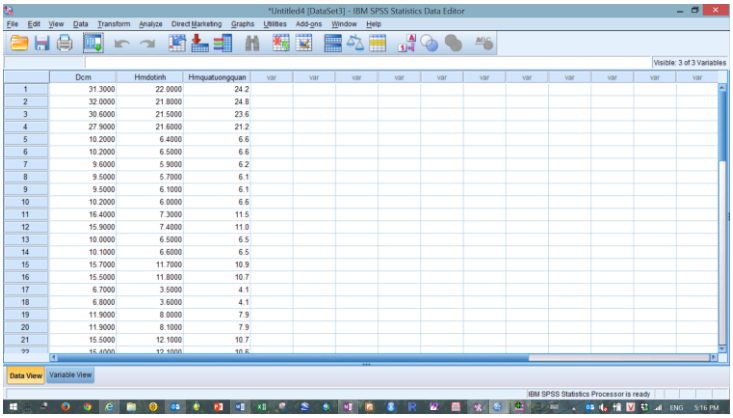
Sau khi nhập dữ liệu trong Excel, đóng file của Excel và mở nó trong SPSS như sau:

File/Open/Data. Trong hộp thoại mở file, chọn kiểu file Excel và chọn file cần mở đã tạo trước đó, và chọn row đầu tiên làm tên biến và Worksheet làm việc.





Kết quả dữ liệu đã được chuyển vào SPSS như sau



2 THỐNG KÊ MÔ TẢ MẪU VÀ KIỂM TRA LUẬT CHUẨN CỦA MẪU ĐỂ XỬ LÝ THỐNG KÊ

Để có những thông số đặc trưng về một đối tượng quan sát như sinh trưởng của một lô rừng, sự đa dạng loài của lô rừng, sự ảnh hưởng của cháy rừng đến mật độ, chất lượng tái sinh, biến động trữ lượng, mật độ của một lô rừng trồng, trạng thái rừng ..... cần tiến thành thu thập dữ liệu theo một nhân tố chủ đạo và sau đó ước lượng, tính toán các đặc trưng cơ bản. Đây là các thông tin cơ bản về một đối tượng quan sát, theo một chỉ tiêu, nhân tố quan tâm.

Các đặc trưng mẫu bao gồm tính các chỉ tiêu: Số trung bình, số trung vị, phương sai, sai tiêu chuẩn, độ lệch, độ nhọn của dãy số liệu quan sát, phạm vi biến động của nó với một mức sai số cho phép đặt trước và các biểu đồ phân bố

Ngoài ra đối với rút mẫu, cần quan tâm đến mẫu có đạt được phân bố chuẩn hay không. Việc này cần được làm rõ trong phân tích đặc trưng mẫu; đôi khi cũng cần xác định trước khi rút mẫu hoặc bố trí thí nghiệm

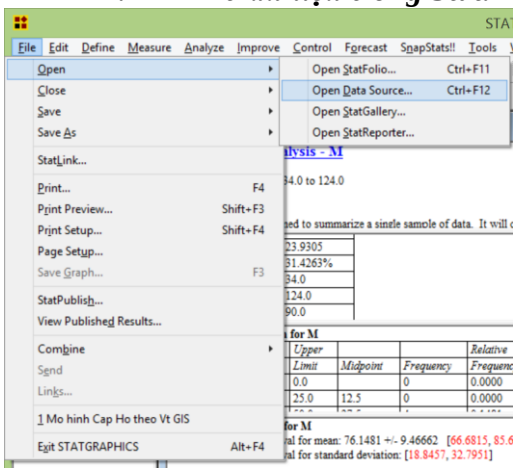
Ví dụ: Khảo sát trữ lượng rừng của một trạng thái; sử dụng ô mẫu dể đo tính trữ lượng m^3/ha (M); từ đây tính toán các đặc trưng cơ bản về trữ lượng rừng.

Các đặc trưng mẫu có thể tính trong Statgraphics theo các bước:

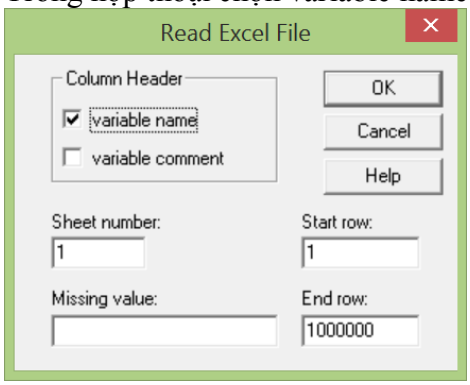
i. Nhập số liệu theo cột trong Exel:

Stt	D binh quan	H binh quan	M
1	15	17	34
2	16	18	34
3	17	19	45
4	21	23	45
5	21	23	56
6	22	24	56
7	23	25	56
8	21	23	56
9	22	24	67
10	21	23	67

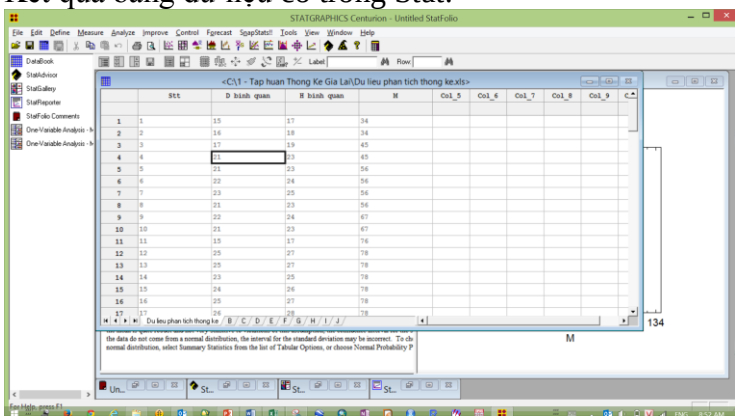
ii. Mở dữ liệu trong Stat: File/Open/Open Data Source/External data file



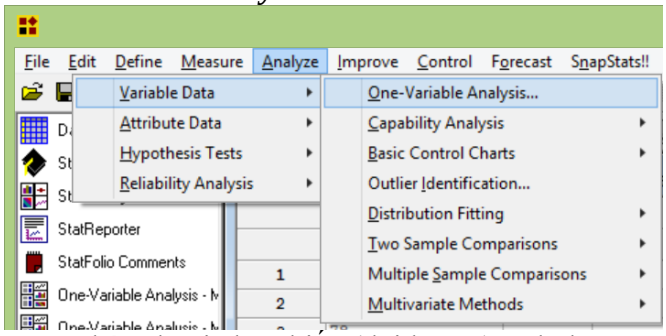
Trong hộp thoại chọn variable name và số thứ tự sheet của bảng tính làm việc



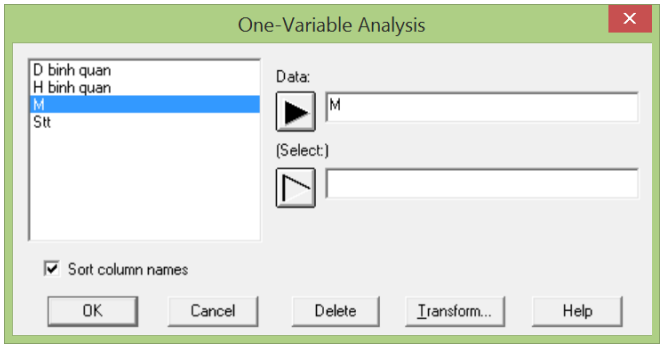
Kết quả bảng dữ liệu có trong Stat:



iii. Tính toán các đặc trưng mẫu trong Stat: Analyze/Variable Data/One-Variable Analysis

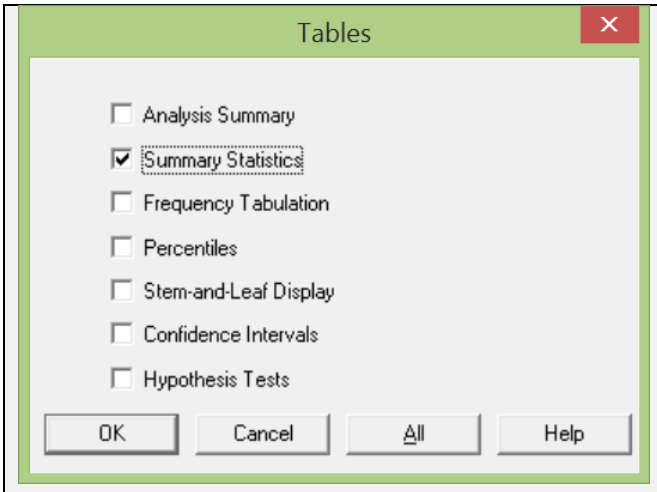


Trong hộp thoại chọn biến (đại lượng) tính đặc trưng mẫu ví dụ là M:



Từ đây có thể chọn ra kết quả mô tả mẫu trong hộp thoại sau

- ***Tóm tắt các chỉ tiêu thống kê mẫu (Summary Statistics):***



Summary Statistics for M

Count	27
Average	76.1481
Standard deviation	23.9305
Coeff. of variation	31.4263%
Minimum	34.0
Maximum	124.0
Range	90.0
Stnd. skewness	0.249982
Stnd. kurtosis	-0.415415

The StatAdvisor

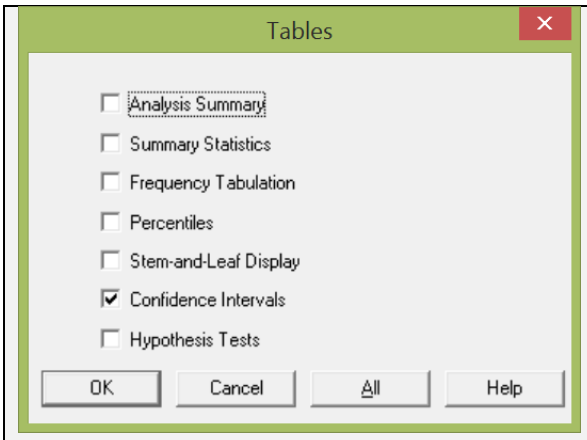
This table shows summary statistics for M. It includes measures of central tendency, measures of variability, and measures of shape. Of particular interest here are the standardized skewness and standardized kurtosis, which can be used to determine whether the sample comes from a normal distribution. Values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate any statistical test regarding the standard deviation. In this case, the standardized skewness value is within the range expected for data from a normal distribution. The standardized kurtosis value is within the range expected for data from a normal distribution.

Giải thích:

- Count (n): Dung lượng mẫu.
- Average (Xbq): Số trung bình.
- Standard deviation (S): Sai tiêu chuẩn mẫu.
- Coeff. of variation: Hệ số biến động CV% = S/X*100
- Minimum: Trị số quan sát bé nhất.
- Maximum: Trị số quan sát lớn nhất.
- Range: Trung vị của dãy quan sát

- Stnd. Kurtosis: Sai tiêu chuẩn của độ nhọn của phân bố nằm trong phạm vi ± 2, mẫu có phân bố chuẩn
- Stnd. Skewness: Sai tiêu chuẩn của độ lệch của phân bố nằm trong phạm vi ± 2, mẫu có phân bố chuẩn

iv. Biến động của giá trị trung bình và ước lượng với độ tin cậy cho trước: :Lựa chọn Confidence Intervals trong hộp thoại



Confidence Intervals for M

95.0% confidence interval for mean: 76.1481 +/- 9.46662 [66.6815, 85.6148]

95.0% confidence interval for standard deviation: [18.8457, 32.7951]

The StatAdvisor

This pane displays 95.0% confidence intervals for the mean and standard deviation of M. The classical interpretation of these intervals is that, in repeated sampling, these intervals will contain the true mean or standard deviation of the population from which the data come 95.0% of the time. In practical terms, we can state with 95.0% confidence that the true mean M is somewhere between 66.6815 and 85.6148, while the true standard deviation is somewhere between 18.8457 and 32.7951.

Both intervals assume that the population from which the sample comes can be represented by a normal distribution. While the confidence interval for the mean is quite robust and not very sensitive to violations of this assumption, the confidence interval for the standard deviation is quite sensitive. If the data do not come from a normal distribution, the interval for the standard deviation may be incorrect. To check whether the data come from a normal distribution, select Summary Statistics from the list of Tabular Options, or choose Normal Probability Plot from the list of Graphical Options.

Giá trị Confidence Level (95%) cho phép ước lượng phạm vi biến động của số trung bình với độ tin cậy 95%:

$$P(Average - t.S/\sqrt{n} \le \mu \le Average + t.S/\sqrt{n}) = 0.95$$

trong đó $t.S/\sqrt{n}$ = Confidence Level (95%), S là Standard deviation, n = count (số mẫu)

Vì vậy giá trị biến động trung bình của tổng thể được ước lượng:

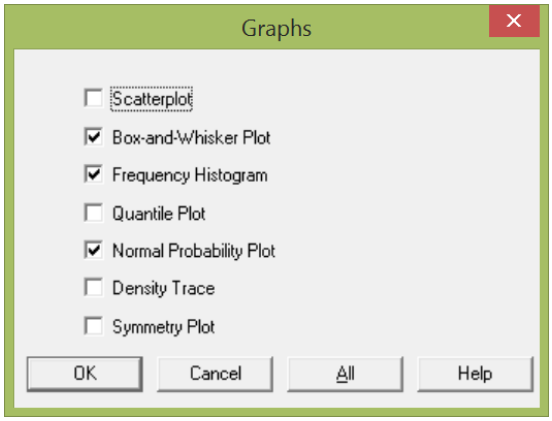
μ = Average ± Confidence Level (95%)

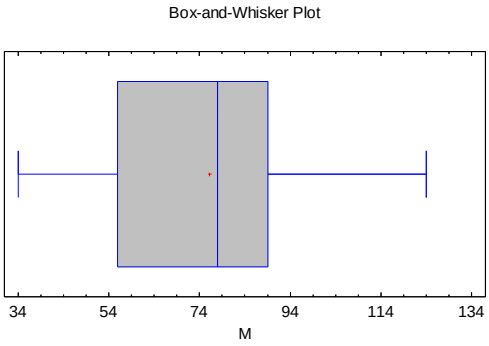
Tùy theo yêu cầu của cuộc điều tra đánh giá, thí nghiệm mà chọn mức độ tin cậy khác nhau: 90%, 95%, 99%.

Như vậy với độ tin cậy 95% thì M biến động trong khoảng: M = 76.1 ± 9.5 m^3

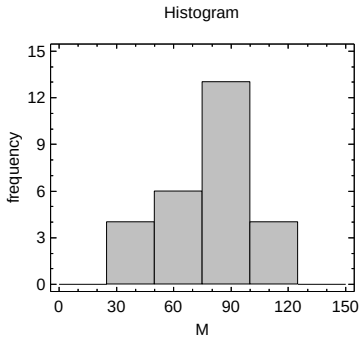
v. Các biểu đồ biểu diễn đặc trưng mẫu: Đặc trưng mẫu còn được biểu diễn dưới dạng biểu đồ. Có 3 loại biểu đồ cần quan tâm để minh họa:

- ✓ Sơ đồ hộp biến động giá trị bình quân (Box – and Whisker Plot)
- ✓ Frequency Histogram
- ✓ Normal Probability Plot

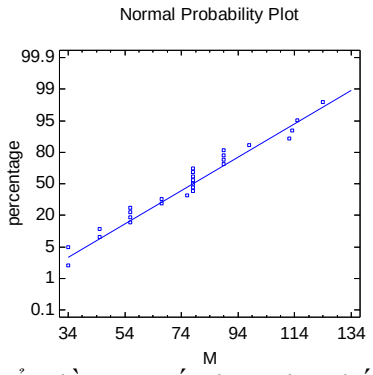




Biểu đồ hộp biến động giá trị bình quân



Phân bố tần số của giá trị quan sát



Biểu đồ xác xuất theo phân bố chuẩn của M

vi. Mẫu bảo đảm phân bố chuẩn hay không – Rút mẫu để đạt được phân bố chuẩn

Để kiểm tra mẫu chuẩn hay không, dựa vào 2 nhóm chỉ tiêu thống kê:

- ✓ Độ lệch và độ nhọn: Stnd. Kurtosis và Stnd. Skewness: nằm trong phạm vi ± 2, thì mẫu có phân bố chuẩn. Ngược lại thì mẫu chưa chuẩn
- ✓ Biểu đồ xác xuất theo phân bố chuẩn : Biểu đồ này chỉ ra mẫu chuẩn khi các giá trị quan sát nằm trên đường chéo xác xuất chuẩn.

Như vậy với kết quả ví dụ trên thì có thể tin mẫu này đạt phân bố chuẩn với phạm vi của sai tiêu chuẩn độ lệch và nhọn trong ± 2 và biểu đồ xác xuất khá bám sát đường chéo.

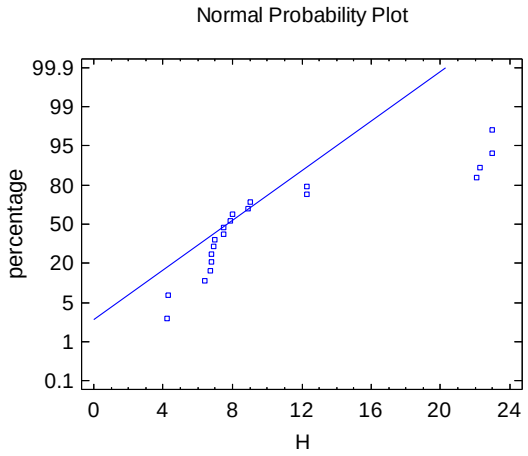
Một ví dụ khác là điều tra sinh trưởng chiều cao (H) cây Sao đen như bảng sau :

Stt	H
1	23.0
2	23.0
3	22.3
4	22.1
5	6.9
6	7.0
7	6.7
8	6.4
9	6.8
10	6.8
11	7.9
12	8.0
13	7.5
14	7.5
15	12.3
16	12.3
17	4.3
18	4.2
19	9.0
20	8.9

Kết quả tính đặc trưng mẫu và biểu đồ xác xuất cho thấy việc rút mẫu với 20 cây để đánh giá sinh trưởng chiều cao (H) keo là chưa có độ tin cậy, vì mẫu chưa đủ (chưa chuẩn). Với Stnd. Skewness = 2.34 > 2 và phân bố mẫu quan sát sai lệch quá lớn so với đường chéo chuẩn.

Summary Statistics for H

Count	20
Average	10.645
Standard deviation	6.44878
Coeff. of variation	60.5804%
Minimum	4.2
Maximum	23.0
Range	18.8
Stnd. skewness	2.34108
Stnd. kurtosis	0.0990205



Biểu đồ xác xuất theo phân bố chuẩn của H

Như vậy trong thực tế cần tiến hành :

- Trước nghiên cứu: Cần có chiến lược rút mẫu để bảo đảm chuẩn

Công thức tính số mẫu quan sát cần thiết (nct): Công thức này có thể áp dụng cho điều tra tự nhiên và xã hội

$$nct \geq t^2.CV\%^2/\Delta\%^2$$

Trong đó CV% (Coeff. of variation) là hệ số biến động: $CV\% = \frac{S}{Xbq}100$, với S là Standard deviation và Δ% là sai số tương đối cho trước ví dụ là 10%, Xbq là trung bình mẫu và t là giá trị hàm t theo độ tự do và độ tin cậy cho trước. Thường với độ tin cậy 95% thì t = 1.96; tuy nhiên tùy vào yêu cầu nghiên cứu có thể xác định độ tin cậy khác nhau; do đó t được xác định trong Excel theo hàm tinv (alpha, df), với df là độ tự do = n – 1 và alpha là % sai số ví dụ 5% = 0.05.
Như vậy để tính được mẫu bảo đảm chuẩn, trước hết phải rút mẫu thử, thường là > 30 mẫu để tính CV%.
Trong thực tế đối với nghiên cứu điều tra có thể áp dụng việc tính toán mẫu trước, tuy nhiên với nghiên cứu thực nghiệm như bố trí thí nghiệm cây trồng theo giống, xuất xứ, chúng ta chưa thể rút mẫu trước khi chưa thí nghiệm. Do vậy có thể áp dụng nguyên lý mẫu lớn để bố trí thí nghiệm, với mẫu > 30 thường có thể tiếp cận chuẩn.

- Trong xử lý số liệu : Nếu mẫu chưa chuẩn như ví dụ trên thì cần bổ sung cho đủ mẫu nct. Tuy nhiên nó chỉ áp dụng được đối với nghiên cứu khảo sát thông qua điều tra; còn với bố trí thí nghiệm trong phòng hoặc hiện trường thì không thể bổ sung.

  Trong ví dụ xác định H cây Sao đen với 20 cây đo tính đã không chuẩn, vì vậy cần bổ sung để mẫu đạt chuẩn như sau :

  Số mẫu cần có nct :

$$nct \geq \frac{t^2.CV\%^2}{\Delta\%^2}$$

Với t có độ tin cậy 95%: t = tiniv(0.05, 19) = 2.09. CV% = 60.5804%. Ví dụ sai số tương đối $\Delta\% = 10\%$.
Vậy

$$nct \geq \frac{2.09^2.60.58\%^2}{10\%^2} = 160\ cây$$

Như vậy nghiên cứu chỉ mới đo tính được 20 cây, vậy số mẫu cần bổ sung để đạt chuẩn là 160 – 20 = 140 cây.

3 SO SÁNH 1 – 2 MẪU QUAN SÁT BẰNG TIÊU CHUẨN T

Kiểm tra mẫu bằng tiêu chuẩn t dựa vào giả thiết phân phối chuẩn của mẫu quan sát. Có các loại kiểm tra t: kiểm tra t một mẫu (one-sample t-test), t cho hai mẫu (two-sample t-test) và t kiểm tra cho hai mẫu bắt cặp (Paired samples). Kiểm tra t một mẫu để đánh giá số trung bình của một mẫu có phải thật sự sai khác với một gía trị cho trước nào đó hay không?. Kiểm tra t hai mẫu là để so sánh hai mẫu xem có cùng một luật phân phối, hay cụ thể hơn là hai mẫu có thật sự có cùng trị số trung bình hay không? Hay nói khác đi có sự sai khác giữa hai mẫu quan sát hay không? Kiểm tra hai mẫu được chia ra là mẫu độc lập hay có bắt cặp.

3.1 So sánh một mẫu với một giá trị cho trước – Kiểm tra T một mẫu

Trong mô tả quan sát một mẫu, người ta có thể có yêu cầu đánh giá giá trị trung bình của mẫu với một giá trị cho trước, ví dụ từ đo đếm chiều cao của cây tái sinh trong rừng khộp, so sánh với một giá trị cho trước về chiều cao mong đợi để cây rừng vượt qua được lửa rừng, xem thật sự chiều cao tái sinh của lô rừng đó đã đạt yêu cầu hay chưa?
Có thể có nhiều ví dụ cho việc áp dụng tiêu chuẩn thống kê này như là so sánh bình quân chỉ số ô nhiễm nồng độ CO_2 trong không khí với tiêu chuẩn an toàn; so sánh chỉ tiêu hóa chất có trong thực phẩm với nồng độ/hàm lượng cho phép, …

Để giải quyết vấn đề này, sử dụng kiểm định t một mẫu với điều kiện mẫu có phân bố chuẩn.
Theo lí thuyết thống kê công thức t kiểm tra một mẫu với một giá trị cho trước:

$$t = \frac{Xbq - \mu}{\frac{S}{\sqrt{n}}}$$

Trong đó, Xbq là giá trị trung bình của mẫu, μ là trung bình theo giả thuyết, S là sai tiêu chuẩn và n là số lượng mẫu quan sát.

- Nếu giá trị tuyệt đối |t| tính cao hơn giá trị t lí thuyết ở mức sai có ý nghĩa, thường là 5% thì có thể kết luận có sự khác biệt có ý nghĩa thống kê giữa trung bình mẫu với giá trị cho trước đó. Và trong trường hợp này nếu t tính <0 thì có nghĩa trung bình của mẫu nhỏ thua có ý nghĩa so với trung bình lý thuyết, ngược lại nếu t tính > 0 thì trung bình của mẫu lớn hơn có ý nghĩa so với trung bình lý thuyết. Đồng thời để đơn giản, kết quả tính toán mức xác suất sai (thường là 5%) gọi là P hay significance alpha (Sig.), nếu Sig. < 0.05 thì kết luận có sự sai khác giữa trung bình mẫu với giá trị cho trước và t < 0 thì mẫu có bình quân bé hơn lý thuyết và ngược lai t > 0 thì lớn hơn lý thuyết.
- Nếu |t| tính ≤ t(0.05, df) thì có thể kết luận ở mức sai 5% trung bình mẫu quan sát xấp xỉ với trung bình lý thuyết. Hoặc Sig. > 0.05

Trong đó t lý thuyết được tính theo hàm =tinv(0.05, df), với độ tự do df = n-1.

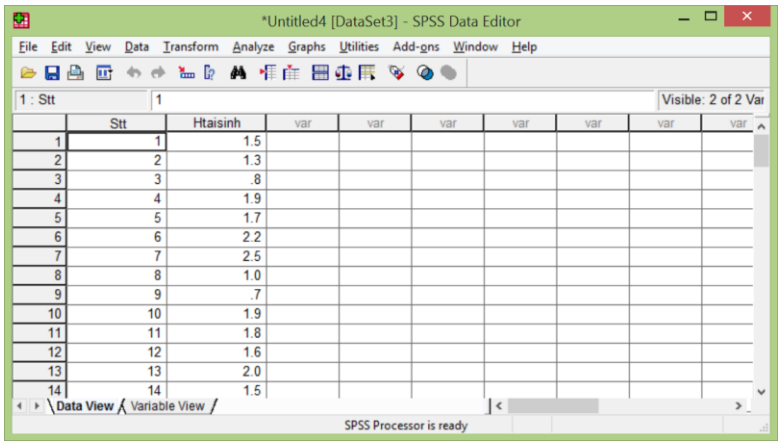
Ví dụ: Người ta rút mẫu do tính chiều cao (H) cây tái sinh trong rừng Khộp và kiểm tra xem trung bình H của cây tái sinh có lớn hơn 2m hay không; vì nếu đúng thì đây là cây tái sinh có triển vọng thành cây gỗ, vượt qua được lửa rừng.
Việc đánh giá được tiến hành như sau:

- Nhập số liệu đo H cây tái sinh trong Excel:

Số liệu đo cao cây tái sinh rừng khộp trong Excel

Stt	Chiều cao cây tái sinh (m)
1	1.5
2	1.3
3	0.8
4	1.9
5	1.7
6	2.2
7	2.5
8	1.0
9	0.7
10	1.9
11	1.8
.....	
58	1.6
59	2.0
60	1.9
61	1.7

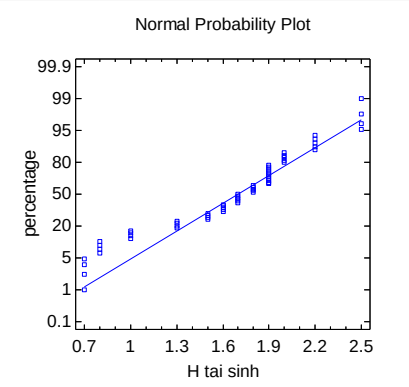
- So sánh H bình quân tái sinh với giá trị lý thuyết cho trước, ví dụ là 2m trong SPSS như sau:
  - ✓ Nhập dữ liệu vào SPSS để phân tích:



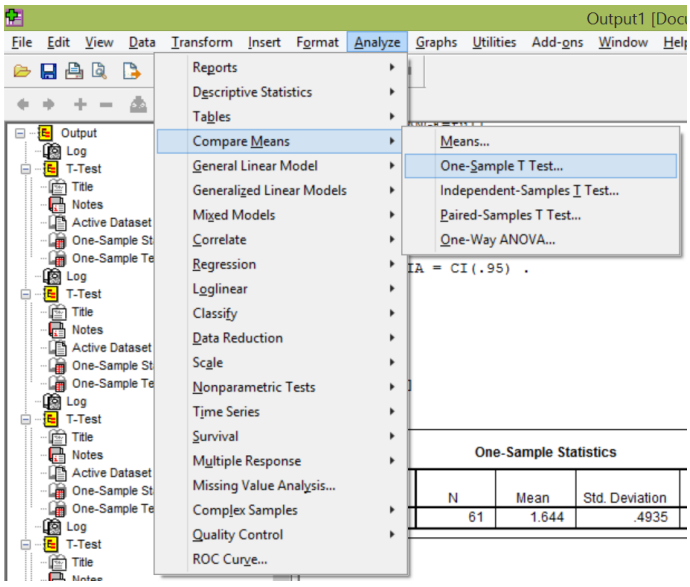
  - ✓ Kiểm tra phân bố chuẩn của mẫu (tiến hành như đã trình bày phần trên trong Statgraphics) và kết quả cho thấy việc rút mẫu đã bảo đảm chuẩn, không cần thu thập số liệu bổ sung

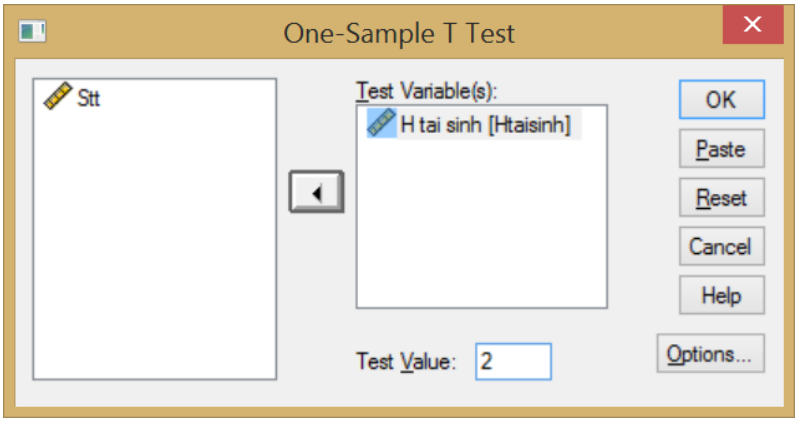
Summary Statistics for H tai sinh

Count	61
Average	1.64426
Standard deviation	0.493465
Coeff. of variation	30.0114%
Minimum	0.7
Maximum	2.5
Range	1.8
Stnd. skewness	-1.47523
Stnd. kurtosis	-0.71729



✓ Kiểm tra sai khác trung bình mẫu với giá trị cho trước (So sánh 1 mẫu) trong SPSS: Analyze/Compare Means/One-Sample T test. Trong hộp thoại chọn biến kiểm tra và giá trị so sánh: Test Value, trong ví dụ này là 2 (m)





Kết quả như sau:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
H tai sinh	61	1.644	.4935	.0632

One-Sample Test

	Test Value = 2					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
H tai sinh	-5.630	60	.000	-.3557	-.482	-.229

Bình quân chiều cao (H) cây tái sinh là 1.64m. Giá trị t = - 5.63 và Sig. = 0.000 < 0.05. Có nghĩa là có sai khác rõ rệt giữa bình quân mẫu quan sát với giá trị lý thuyết so sánh và t<0, do vậy kết luận rằng H bình quân tái sinh <2m rõ rệt và như vậy chưa đạt tái sinh triển vọng, chưa thoát được lửa rừng.

3.2 So sánh sự sai khác giữa trung bình 2 mẫu quan sát độc lập – Kiểm tra T 2 mẫu độc lập

Trong các nghiên cứu, thí nghiệm thường người ta cần so sánh kết quả của 2 mẫu hoặc 2 công thức độc lập, ví dụ: Bón phân hay không bón, che bóng hay không che, sinh trưởng, tái sinh của cây rừng nơi được chăm sóc và nơi không, sinh trưởng cây rừng nơi cháy và không cháy.....Việc kiểm tra thống kê được tiến hành theo 2 mẫu trên cơ sở so sánh 2 số trung bình bằng các tiêu chuẩn t.

Công thức tính giá trị kiểm tra t:

$$t = \frac{X1 - X2}{\sqrt{\frac{(n1-1)S1^2 + (n2-1)S2^2}{n1+n2-2}\left(\frac{1}{n1}+\frac{1}{n2}\right)}}$$

Với: X_1, X_2: Trung bình của mẫu 1 và 2.
S_1^2, S_2^2: Phương sai mẫu 1 và 2.
n_1, n_2: dung lượng 2 mẫu 1 và 2.

Nếu |t| tính lớn hơn t lý thuyết với Sig. α=0.05 và độ tự do K=n_1+n_2-2 thì bác bỏ giả thuyết Ho, có nghĩa trung bình 2 mẫu sai khác có ý nghĩa.

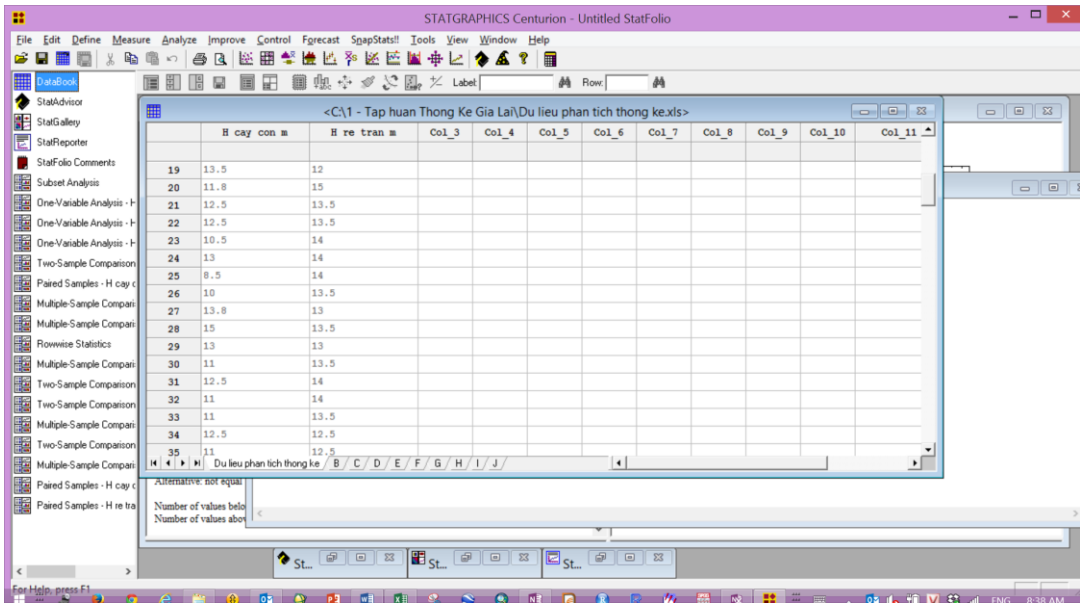
Khi sử dụng tiêu chuẩn t để so sánh 2 mẫu độc lập, cần kiểm tra 2 điều kiện:

- Hai mẫu có phân bố chuẩn.
- Sai tiêu chuẩn hoặc phương sai của hai mẫu phải bằng nhau

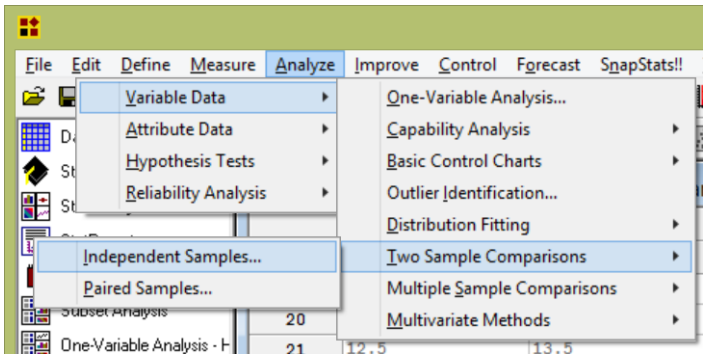
Ví dụ: Kiểm tra sinh trưởng chiều cao H của 2 phương pháp trồng thông 3 lá *Pinus kesiya* bằng cây con và rễ trần tại trạm thực nghiệm của Viện Nghiên cứu Lâm sinh ở Lang Hanh-Lâm Đồng: Mỗi công thức được rút mẫu độc lập theo ô tiêu chuẩn 1000m^2, đo đếm chiều cao:

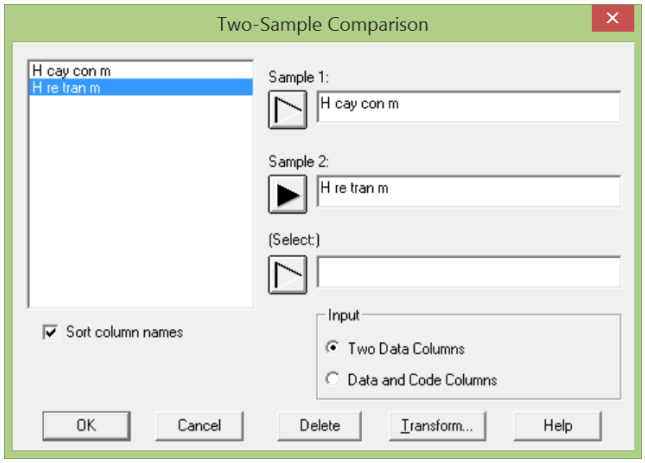
Sử dụng Statgraphics để kiểm tra thống kê bằng tiêu chuẩn t trong trường hợp 2 mẫu độc lập:

✔ Nhập số liệu vào Stat từ file Excel

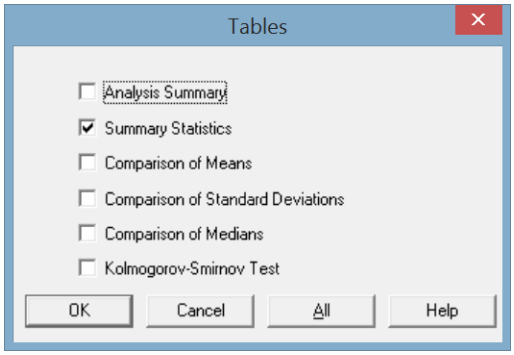


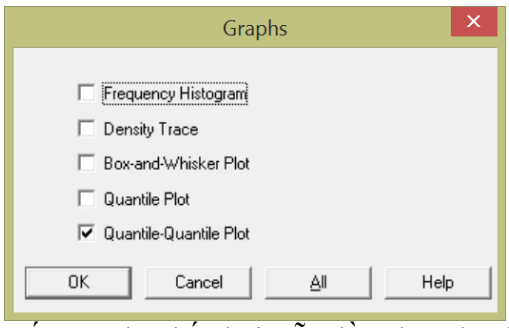
✔ Sử dụng so sánh t 2 mẫu độc lập: Analyze/Variable Data/Two Sample Comparisions/Independent Samples. Trong hộp thoại đưa biến từng mẫu vào





✔ Kiểm tra phân bố chuẩn của 2 mẫu: Mở hộp thoại phân tích thống kê và chọn Summary Statistics và hộp thoại biểu đồ chọn Quantile-Quantile Plot

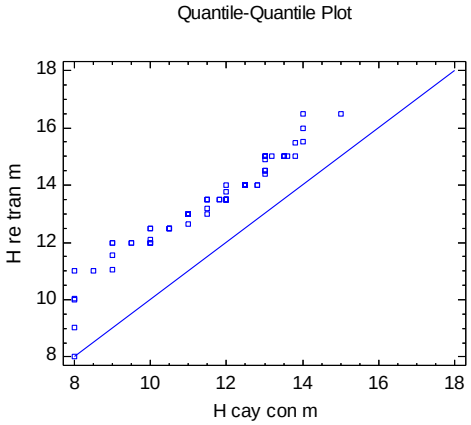




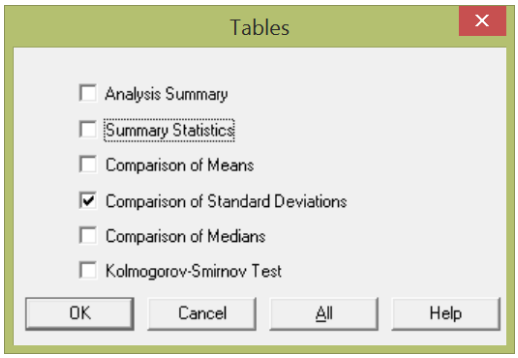
Kết quả cho thấy hai mẫu đều chưa đạt được phân bố chuẩn với Stnd. Skewness và Stnd. Kurtosis nằm ngoài phạm vi ± 2 và phân bố giá trị quan sát của hai mẫu không nằm trên đường chéo của phân bố chuẩn. Tuy nhiên ở đây mẫu được thu thập khá lớn (>90 cây cho mỗi mẫu), do đó tạm thời chấp nhận giả thuyết phân bố chuẩn của 2 mẫu. Nhưng để đánh giá chính xác hơn trong trường hợp không thể thu thập số liệu bổ sung, thì tiêu chuẩn phi tham số có thể hỗ trợ để so sánh vì nó không đòi hỏi yêu cầu phân bố chuẩn.

Summary Statistics

	H cay con m	*H re tran m*
Count	92	93
Average	11.6043	13.4032
Standard deviation	1.59993	1.46565
Coeff. of variation	13.7873%	10.9351%
Minimum	8.0	8.0
Maximum	15.0	16.5
Range	7.0	8.5
Stnd. skewness	-2.23744	-3.38989
Stnd. kurtosis	-0.398833	3.8466



✔ Kiểm tra phương sai của 2 mẫu bằng tiêu chuẩn F: Sử dụng hộp thoại để kiểm tra: Comparision of Standard Deviations.



Comparison of Standard Deviations

	H cay con m	*H re tran m*
Standard deviation	1.59993	1.46565
Variance	2.55976	2.14814
Df	91	92

Ratio of Variances = 1.19162

F-test to Compare Standard Deviations

Null hypothesis: sigma1 = sigma2

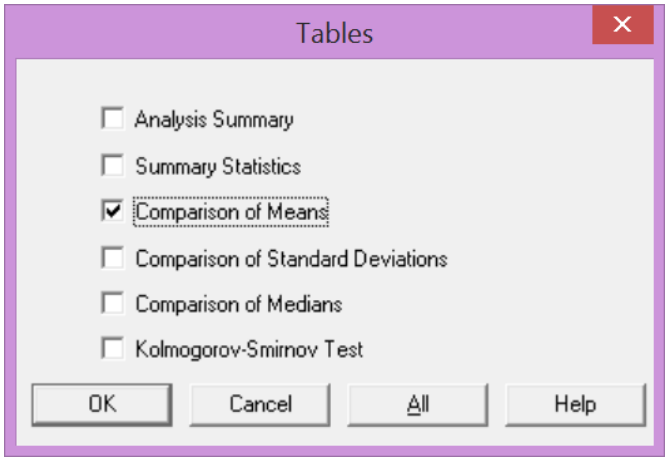
Alt. hypothesis: sigma1 NE sigma2

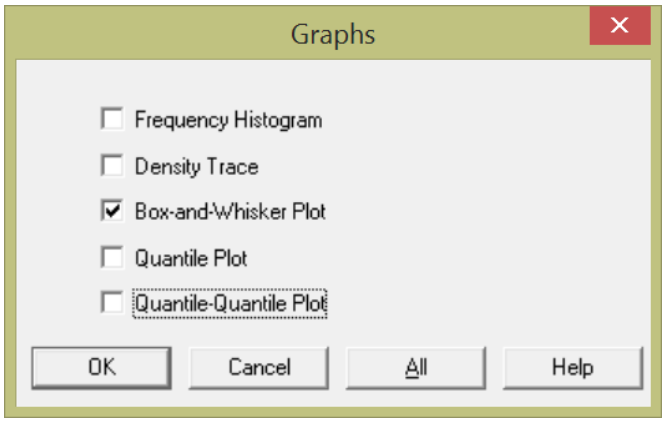
F = 1.19162 P-value = 0.403068

Do not reject the null hypothesis for alpha = 0.05.

Kết quả trên cho thấy P-value = 0.403 > 0.05, như vậy chấp nhận giả thuyết Ho (Null Hypothesis) là hai phương sai (sai tiêu chuẩn) của hai mẫu bằng nhau. Trong trường hợp ngược lại nếu P-value < 0,05 thì phương sai 2 mẫu không bằng nhau và không áp dụng tiêu chuẩn t để kiểm tra, như vậy hoặc bổ sung số liệu quan sát hoặc sử dụng tiêu chuẩn phi tham số không đòi hỏi luật chuẩn và phương sai bằng nhau (ở phần tiếp theo).

✔ So sánh 2 trung bình bằng tiêu chuẩn t: Sử dụng hộp thoại phân tích thống kê và chọn Comparision of Means và hộp thoại Graphs để có đồ thị so sánh biến động trung bình 2 mẫu





Comparison of Means

95.0% confidence interval for mean of H cay con m: 11.6043 +/- 0.331336 [11.273, 11.9357]

95.0% confidence interval for mean of H re tran m: 13.4032 +/- 0.301848 [13.1014, 13.7051]

95.0% confidence interval for the difference between the means

assuming equal variances: -1.79888 +/- 0.445016 [-2.24389, -1.35386]

t test to compare means

Null hypothesis: mean1 = mean2

Alt. hypothesis: mean1 NE mean2

assuming equal variances: t = -7.97547 P-value = 1.79536E-7

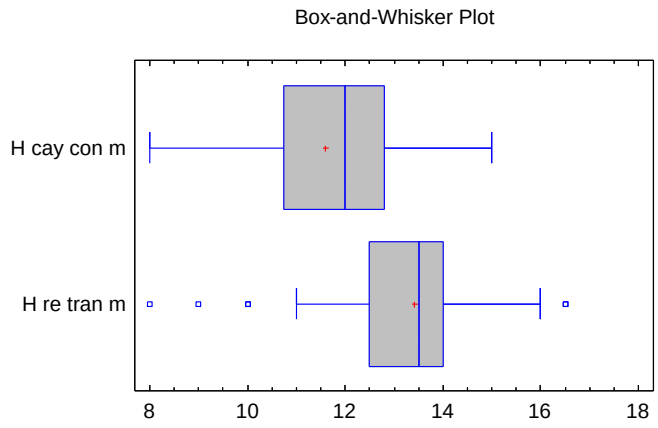
Reject the null hypothesis for alpha = 0.05.

The StatAdvisor

This option runs a t-test to compare the means of the two samples. It also constructs confidence intervals or bounds for each mean and for the difference between the means. Of particular interest is the confidence interval for the difference between the means, which extends from -2.24389 to -1.35386. Since the interval does not contain the value 0, there is a statistically significant difference between the means of the two samples at the 95.0% confidence level.

A t-test may also be used to test a specific hypothesis about the difference between the means of the populations from which the two samples come. In this case, the test has been constructed to determine whether the difference between the two means equals 0.0 versus the alternative hypothesis that the difference does not equal 0.0. Since the computed P-value is less than 0.05, we can reject the null hypothesis in favor of the alternative.

NOTE: these results assume that the variances of the two samples are equal. In this case, that assumption appears to be reasonable based on the results of an F-test to compare the standard deviations. You can see the results of that test by selecting Comparison of Standard Deviations from the Tabular Options menu.



Đồ thị biến động H bình quân của hai mẫu

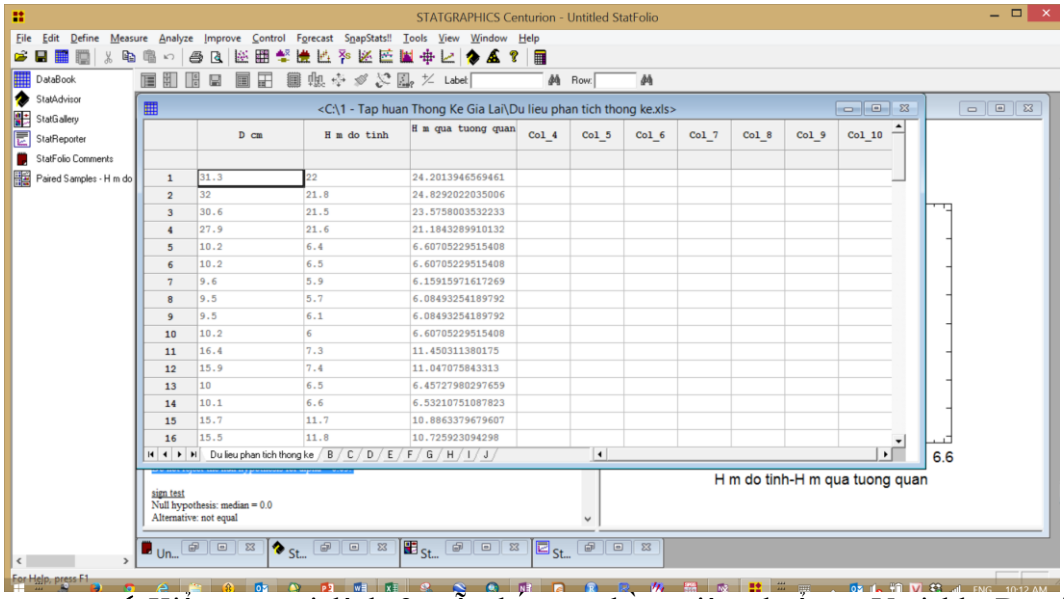
Kết quả trên cho thấy qua kiểm tra bằng tiêu chuẩn t có P-value = 1.79536E-7 < 0.05, có nghĩa là bác bỏ giải thuyết Ho (hai trung bình bằng nhau). Hay nói sinh trưởng của P. kesiya trồng bằng 2 phương pháp khác nhau sai dị rõ. Chiều cao bình quân cây trồng bằng rễ trần hơn hẳn trồng bằng cây con quan biểu đồ, do vậy phương pháp trồng thông 3 lá bằng rễ trần cần được ứng dụng trong thực tiễn.

3.3 So sánh sự sai khác giữa trung bình 2 mẫu quan sát bắt cặp – Kiểm tra T 2 mẫu bắt cặp

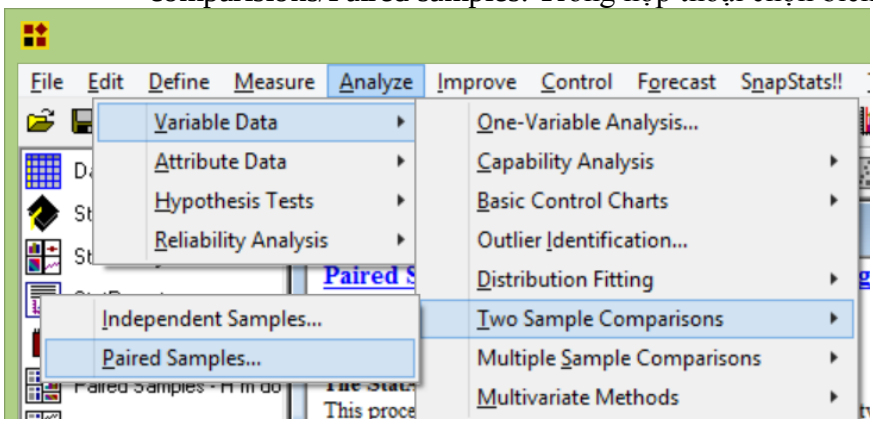
Trong các nghiên cứu, thí nghiệm thường người ta cần so sánh kết quả từ hai phương pháp khác nhau trên cùng một đối tượng. Ví dụ trên mỗi mẫu người tra dùng 2 phương pháp phân tích khác nhau và so sánh xem có sự khác biệt về kết quả hay không. Trường hợp này sử dụng so sánh bằng tiêu chuẩn t với 2 mẫu quan sát bắt cặp.
Điều kiện để áp dụng tiêu chuẩn t này là sai lệch giữa các cặp dữ liệu có phân bố chuẩn.

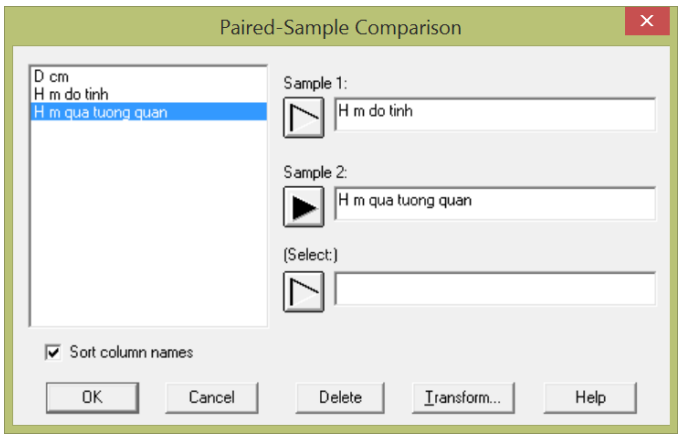
Ví dụ: Trong điều tra cây rừng, thường lập mô hình tương quan giữa chiều cao (H) theo đường kính (D) để từ đó giảm chi phí khi đo cao cây. Tuy nhiên để đánh giá độ tin cậy của mô hình tương quan, từ mỗi cây so sánh cặp dữ liệu gồm H đo cao trực tiếp và H ước tính qua mô hình tương quan. Đây là tường hợp so sánh 2 mẫu bắt cặp, tức là 2 giá trị trên một cây.
Sử dụng Statgraphics để so sánh bằng tiêu chuẩn t bắt cặp:

- ✔ Nhập dữ liệu từ Excel và Stat:

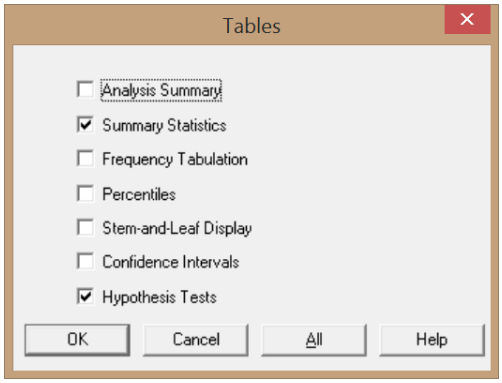


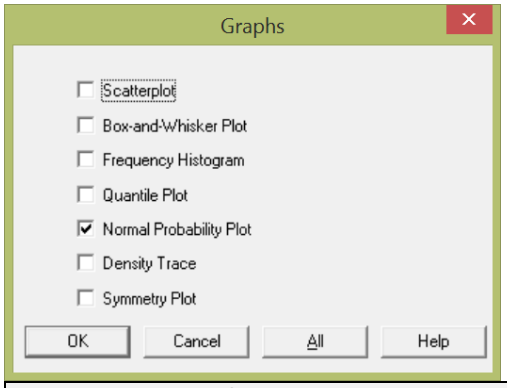
✔ Kiểm tra sai lệch 2 mẫu bắt cặp bằng tiêu chuẩn t: Variable Data/Two sample comparisions/Paired samples. Trong hộp thoại chọn biến so sánh cho từng mẫu.





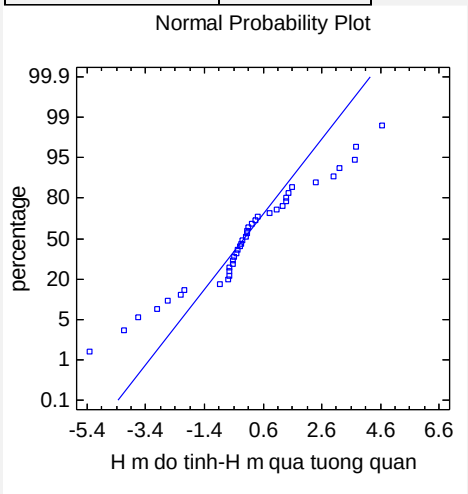
✔ Kiểm tra sai lệch giữa hai mẫu có chuẩn hay không: Trong hộp thoại Tables chọn Summary Statistics và trong Graphs chọn Normal Probability Plot





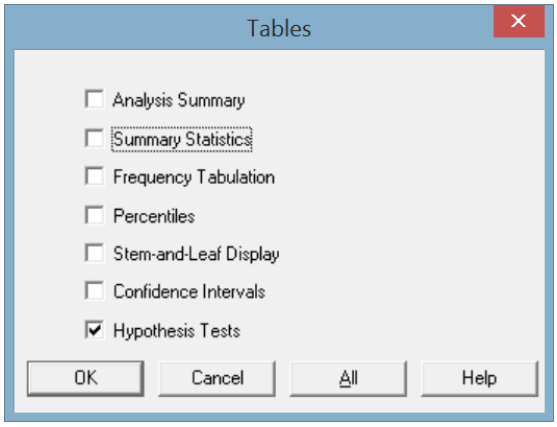
Summary Statistics for H m do tinh-H m qua tuong quan

Count	40
Average	0.0617335
Standard deviation	2.11221
Coeff. of variation	3421.49%
Minimum	-5.32
Maximum	4.66881
Range	9.98881
Stnd. skewness	-0.538061
Stnd. kurtosis	0.81107



Kết quả trên cho thấy sai lệch giữa các cặp quan sát H có phân bố chuẩn, với sai tiêu chuẩn độ lệch và nhọn nằm trong phạm vi ±2 và biểu đồ xác suất khá bám sát đường chéo chuẩn.

- ✓ Kiểm tra sự sai khác giữa các cặp quan sát trên cùng một mẫu: Trong hộp Table chọn Hypothesis



Hypothesis Tests for H m do tinh-H m qua tuong quan

Sample mean = 0.0617335

Sample median = -0.0459924

Sample standard deviation = 2.11221

t-test

Null hypothesis: mean = 0.0

Alternative: not equal

Computed t statistic = 0.184848

P-Value = 0.854306

Do not reject the null hypothesis for alpha = 0.05.

Ở đây dùng tiêu chuẩn t để kiểm tra sai lệch giữa H đo tính và H qua mô hình, với giả thuyết Ho (Null Hypothesis) là trung bình sai lệch giữa 2 mẫu = 0. Kết quả cho ra P-value = 0.854 > 0.05, có nghĩa không thể bác bỏ giải thuyết Ho, hay nói khác trung bình sai lệch là gần bằng 0, hay hai mẫu chưa có sự sai khác, hay H ước tính qua phương trình là bám sát với số liệu đo trực tiếp và có thể sử dụng phương trình vào thực tế.

4 TIÊU CHUẨN PHI THAM SỐ ĐỂ SO SÁNH NHIỀU MẪU QUAN SÁT ĐỘC LẬP HOẶC CÓ LIÊN HỆ

Trong sử dụng tiêu chuẩn t nói trên chỉ áp dụng để so sánh tối đa 2 mẫu độc lập hoặc bắt cặp và các mẫu đều phải thỏa mãn 2 điều kiện là phân bố chuẩn và phương sai bằng nhau. Trong thực tế số mẫu quan sát có thể >2 và các điều kiện rút mẫu không bảo đảm yêu cầu chuẩn và phương sai bằng nhau.Trường hợp như vậy, không thể áp dụng tiêu chuẩn t và như vậy các tiêu chuẩn thống kê phi tham số cần được sử dụng.

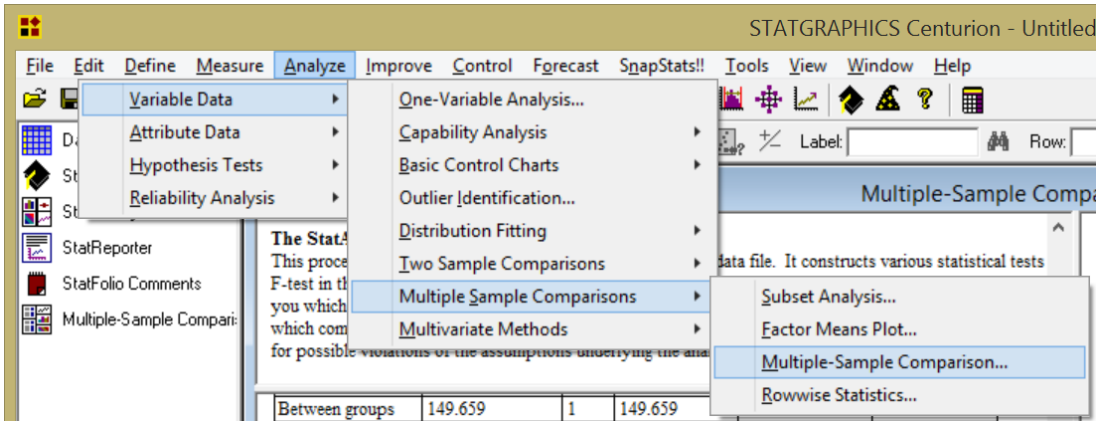
4.1 Tiêu chuẩn phi tham số kiểm tra các mẫu độc lập

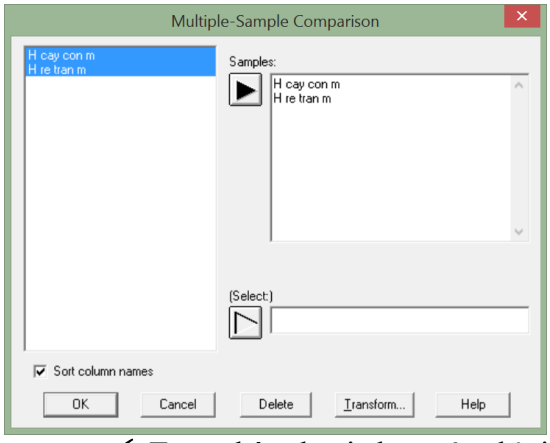
Tiêu chuẩn này chỉ đòi hỏi dãy số liệu quan sát độc lập của mỗi mẫu là liên tục. Đó là tiêu chuẩn phi tham số Kruskal Wallis và Friedman.

Tiêu chuẩn phi tham số Kruskal Wallis và Friedman là kiểm tra giả thuyết Ho trong đó dãy dữ liệu các mẫu được xem là đồng nhất. Dữ liệu quan sát của tất cả các mẫu kết hợp chung và được xếp hạng (thứ tự), từ đó tính trung bình thứ hạng (Median) cho từng mẫu và đem so sánh với nhau.

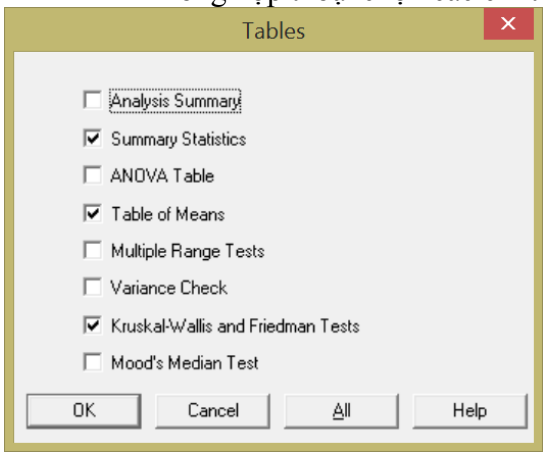
Ví dụ trong trường hợp so sánh hai mẫu độc lập theo hai phương pháp trồng cây thông 3 lá là cây con và rễ trần, với số liệu quan sát khá lớn (>90cây) nhưng cả hai mẫu đều chưa đạt chuẩn. Do đó nếu áp dụng t so sánh sẽ chưa đủ độ tin cậy. Trong trường hợp này nên sử dụng tiêu chuẩn phi tham số Kruskal Wallis và Friedman để so sánh vì nó loại trừ được yêu cầu chuẩn.
Sử dụng Stat để kiểm tra thống kê theo Kruskal Wallis và Friedman như sau:

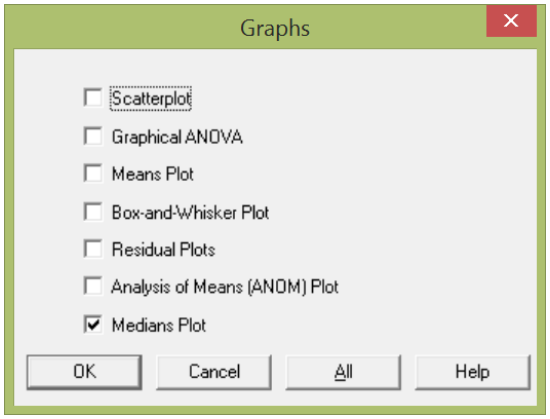
- ✓ So sánh nhiều mẫu bằng phương pháp phi tham số: Variable Data/Multiple- Sample Comparisions/Multiple-Sample Comnparirion. Trong hộp thoại chọn các mẫu so sánh.





- ✓ Trong hộp thoại chọn các chỉ tiêu mô tả thống kê và đánh giá như sau:





Summary Statistics

	Count	*Average*	*Standard deviation*	*Coeff. of variation*	*Minimum*	*Maximum*	*Range*	*Stnd. skewness*
H cay con m	92	11.6043	1.59993	13.7873%	8.0	15.0	7.0	-2.23744
H re tran m	93	13.4032	1.46565	10.9351%	8.0	16.5	8.5	-3.38989
Total	185	12.5086	1.77578	14.1964%	8.0	16.5	8.5	-3.0257

	Stnd. kurtosis
H cay con m	-0.398833
H re tran m	3.8466
Total	0.502338

Table of Means with 95.0 percent LSD intervals

			Stnd. error		
	Count	*Mean*	*(pooled s)*	*Lower limit*	*Upper limit*
H cay con m	92	11.6043	0.159919	11.3812	11.8275
H re tran m	93	13.4032	0.159057	13.1813	13.6251
Total	185	12.5086			

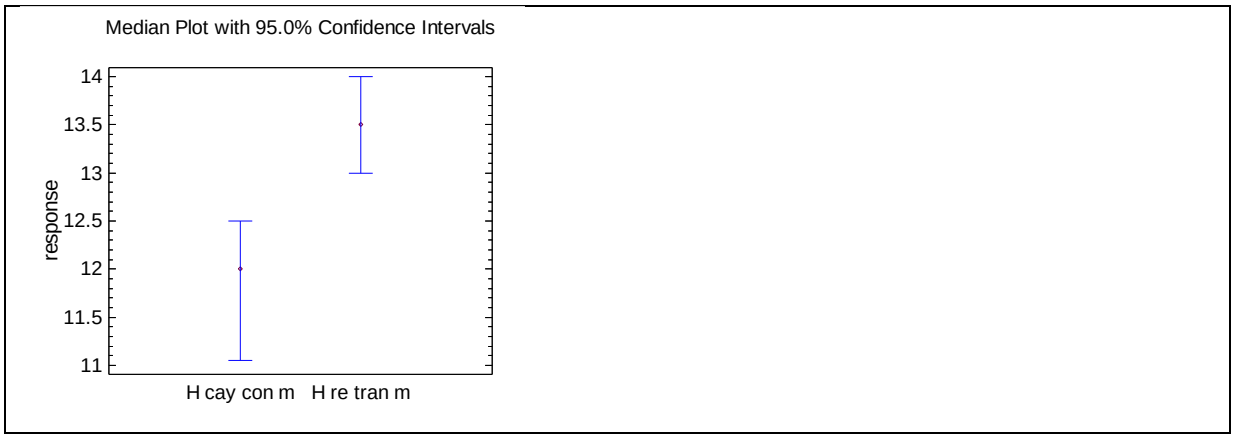
Kruskal-Wallis Test

	Sample Size	*Average Rank*
H cay con m	92	63.9402
H re tran m	93	121.747

Test statistic = 54.3713 P-Value = 0.0

The StatAdvisor

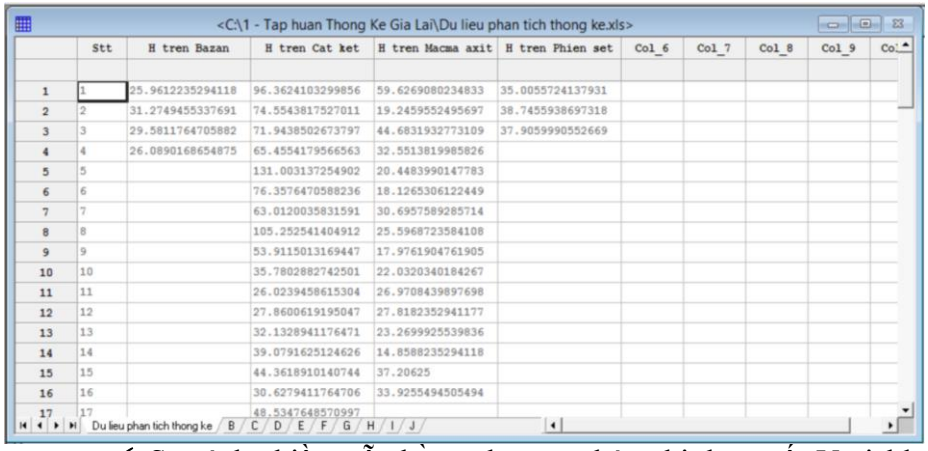
The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 2 columns is the same. The data from all the columns is first combined and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is less than 0.05, there is a statistically significant difference amongst the medians at the 95.0% confidence level. To determine which medians are significantly different from which others, select Box-and-Whisker Plot from the list of Graphical Options and select the median notch option.



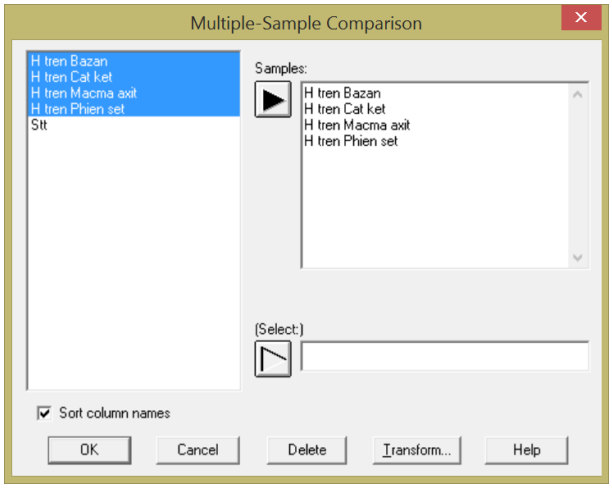
Kết quả trên cho thấy hai mẫu chưa đạt chuẩn, do đó nếu sử dụng tiêu chuẩn t sẽ chưa có độ tin cậy. Kết quả kiểm tra theo Kruskal-Wallis cho thấy P-value = 0.0 < 0.05, có nghĩa là dãy phân bố số liệu quan sát của hai mẫu trồng theo hai phương pháp khác nhau là có sự sai khác có ý nghĩa ở độ tin cậy 95%. Qua biểu đồ cho thấy trung bình vị trí xếp hạng của cây trồng bằng rễ trần cao hơn bằng cây con. Do vậy nên áp dụng phương pháp trồng bằng rễ trần.

Ví dụ khác cho việc kiểm tra trên hai mẫu độc lập theo tiêu chuẩn phi tham số. Dữ liệu là giá trị tăng trưởng chiều cao cây tếch (H, cm) khi trồng làm giàu rừng khộp trên 4 loại đá mẹ. Kiểm tra có hay không sự sai khác tăng trưởng tếch ở rừng khộp với 4 loại đá mẹ khác nhau (4 mẫu).

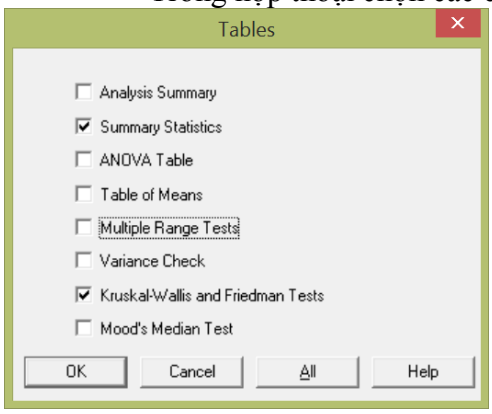
- ✔ Nhập dữ liệu từ Excel vào Stat, trong đó dãy sinh trưởng chiều cao H) của cây tếch được xếp theo từng loại đá mẹ khác nhau (4 loại ứng với 4 cột):

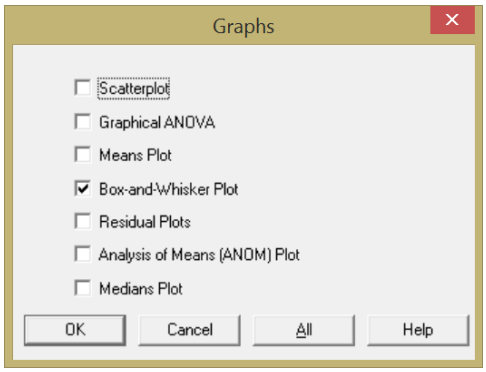


- ✔ So sánh nhiều mẫu bằng phương pháp phi tham số: Variable Data/Multiple- Sample Comparisions/Multiple-Sample Comnparirion. Trong hộp thoại chọn các mẫu so sánh.



✔ Trong hộp thoại chọn các chỉ tiêu mô tả thống kê và đánh giá như sau:





Summary Statistics

	Count	*Average*	*Standard deviation*	*Coeff. of variation*	*Minimum*	*Maximum*	*Range*
H tren Bazan	4	28.2266	2.63492	9.3349%	25.9612	31.2749	5.31372
H tren Cat ket	41	48.4679	23.4149	48.3101%	17.43	131.003	113.573
H tren Macma axit	16	28.4396	11.5254	40.5259%	14.8588	59.6269	44.7681
H tren Phien set	3	37.2191	1.96236	5.27246%	35.0056	38.7456	3.74002
Total	64	41.6684	21.6131	51.8693%	14.8588	131.003	116.144

	Stnd. skewness	*Stnd. kurtosis*
H tren Bazan	0.28028	-1.64562
H tren Cat ket	4.11222	4.25107
H tren Macma axit	2.36311	2.01465
H tren Phien set	-0.977387	
Total	5.99291	7.47169

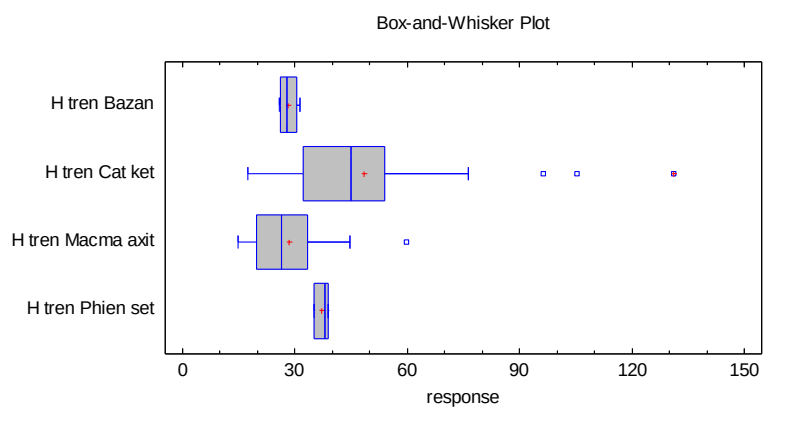
Kruskal-Wallis Test

	Sample Size	*Average Rank*
H tren Bazan	4	19.0
H tren Cat ket	41	39.0732
H tren Macma axit	16	18.625
H tren Phien set	3	34.6667

Test statistic = 16.1389 P-Value = 0.00106202

The StatAdvisor

The Kruskal-Wallis test tests the null hypothesis that the medians within each of the 4 columns is the same. The data from all the columns is first combined and ranked from smallest to largest. The average rank is then computed for the data in each column. Since the P-value is less than 0.05, there is a statistically significant difference amongst the medians at the 95.0% confidence level. To determine which medians are significantly different from which others, select Box-and-Whisker Plot from the list of Graphical Options and select the median notch option.



Kết quả kiểm tra theo Kruskal-Wallis cho thấy P-value = 0.000106 < 0.05, có nghĩa là dãy phân bố số liệu quan sát tăng trưởng chiều cao tếch trên 4 loại đá mẹ có sự sai khác có ý nghĩa ở độ tin cậy 95%. Qua biểu đồ cho thấy trung bình vị trí xếp hạng của cây tếch trên đá mẹ Cát kết là tốt nhất và kém nhất là trên Macma axit.

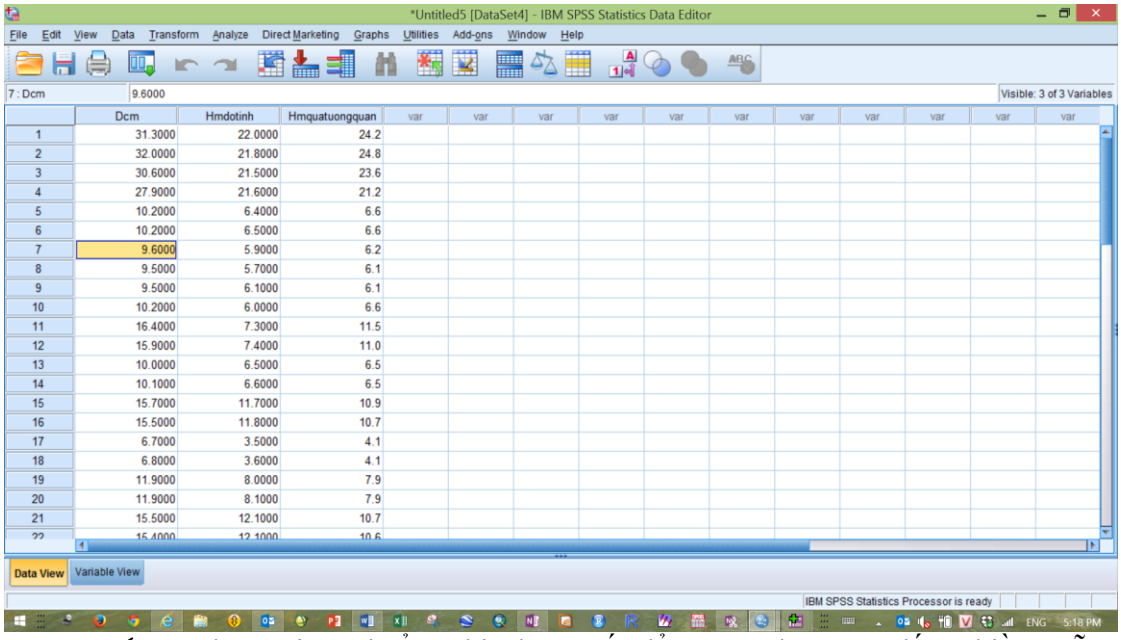
4.2 Tiêu chuẩn phi tham số kiểm tra các mẫu liên hệ

Trong trường hợp có hai hay nhiều hơn các mẫu có liên hệ với nhau; ngoài ra chưa đạt phân bố chuẩn, phương sai bằng nhau nên không thể sử dụng tiêu chuẩn t bắt cặp (với 2 mẫu); thì tiêu chuẩn phi tham số là thích hợp để so sánh.

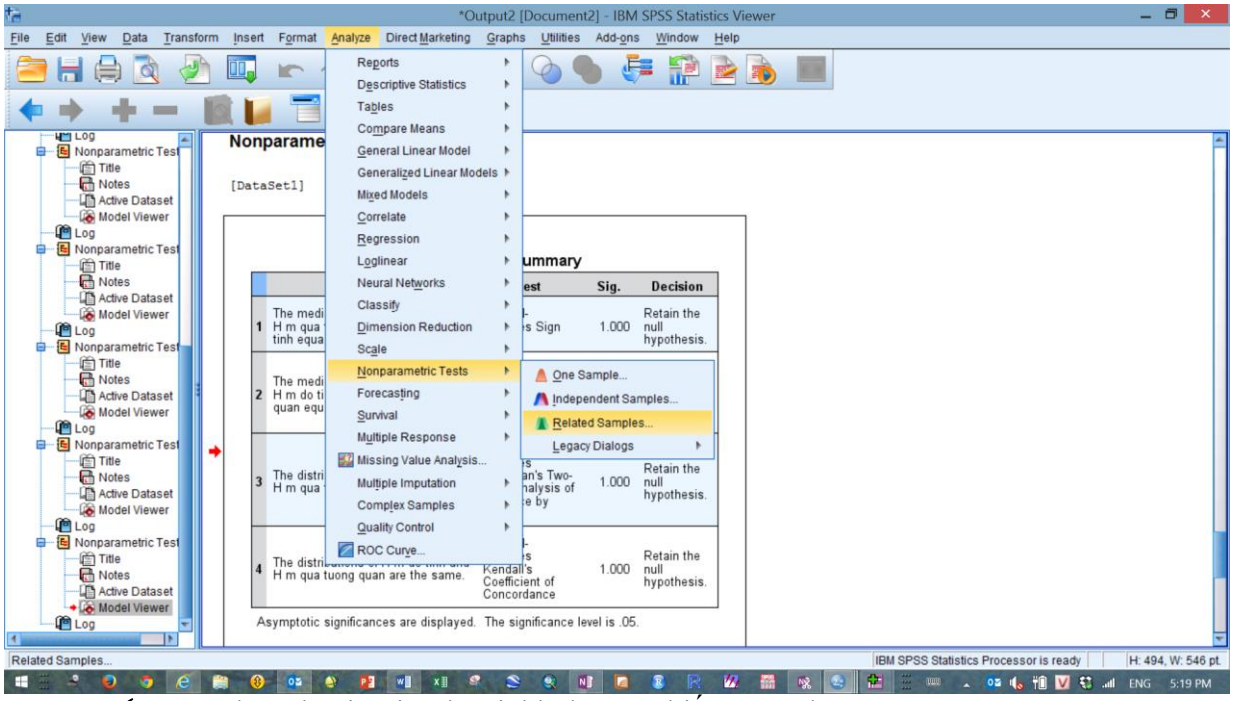
- Trường hợp hai mẫu liên hệ, có thể sử dụng tiêu chuẩn phi tham số Wilcoxon, trong đó chênh lệch giữa các cặp dữ liệu được xếp hạng và tính giá trị trung vị Median cho từng mẫu, sau đó so sánh sự sai khác
- Trường hợp có nhiều mẫu liên hệ thì tiêu chuẩn Kendall có thể được sử dụng để so sánh sự sai khác giữa các dãy phân bố dữ liệu của các mẫu

Sử dụng ví dụ chiều cao được đo trực tiếp và thông qua mô hình tương quan, sử dụng các tiêu chuẩn phi tham số Wilcoxon và Kendal để so sánh sự sai khác trong SPSS như sau:

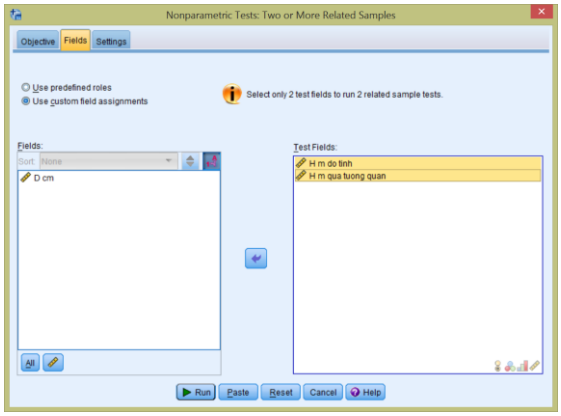
✔ Nhập dữ liệu trong SPSS từ Excel với từng cặp dữ liệu H theo từng cây:



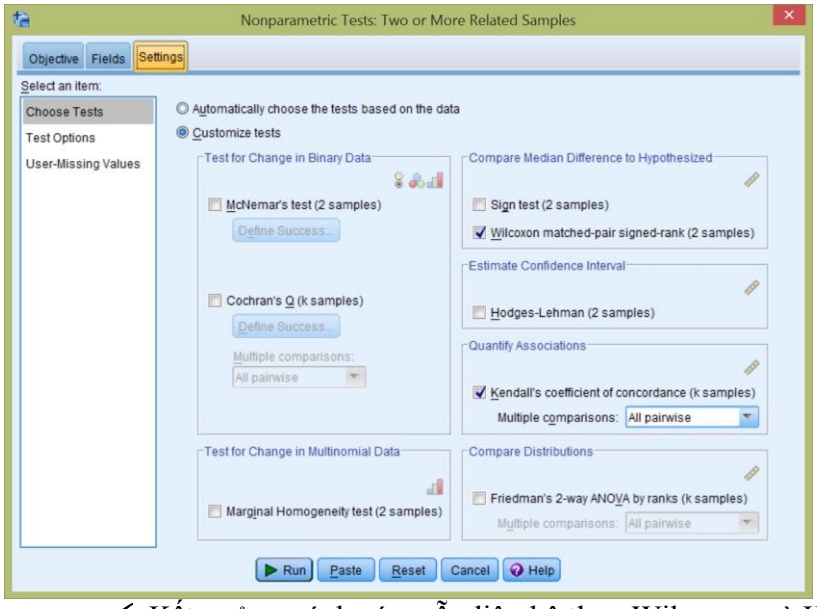
✔ Sử dụng tiêu chuẩn phi tham số để so sánh từ 2 đến nhiều mẫu liên hệ: Analyze/Nonparametric Test/Related Samples



✔ Trong hộp thoại với Tab: Field, đưa các biến so sánh vào



✔ Trong hộp thoại với Tab Setting/Choose Tests: Chọn Wilcoxon để so sánh hai mẫu theo Median và Kendall để so sánh dãy phân bố của 2 đến nhiều mẫu



✔ Kết quả so sánh các mẫu liên hệ theo Wilcoxon và Kendall như sau:

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between H m do tinh and H m qua tuong quan equals 0.	Related-Samples Wilcoxon Signed Rank Test	.936	Retain the null hypothesis.
2	The distributions of H m do tinh and H m qua tuong quan are the same.	Related-Samples Kendall's Coefficient of Concordance	1.000	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

- Kiểm tra theo Wilcoxon với 2 mẫu liên hệ theo Median cho thấy Sig. = 0.936 > 0.05, có nghĩa là chưa thể bác bỏ giả thuyết Ho với hai trung vị là bằng nhau. Như vậy việc xác đinh H qua đo trực tiếp và qua tương quan là chưa có sai khác. Có thể sử dụng tương quan để giảm chi phí đo đếm.
- Kiểm tra dãy phân bố của hai mẫu theo Kendall cho thấy Sig. = 1.0 > 0.05, như vậy Ho được chấp nhận, hay nói khác chưa có sự sai khác H giữa hai dãy số liệu đo H trực tiếp và qua phương trình. Có thể sử dụng phương trình để giảm chi phí điều tra.

Trong ví dụ trên, tiêu chuẩn Kendall được so sánh với 2 mẫu liên hệ; tuy nhiên tiêu chuẩn này được sử dụng tốt khi có trên 2 mẫu liên hệ.

5 PHÂN TÍCH PHƯƠNG SAI

Phân tích phương sai là một trong những phương pháp phân tích thống kê quan trọng, đặc biệt là trong các thí nghiệm giống, thí nghiệm các nhân tố tác động đến hiệu quả, chất lượng của cây trồng, vật nuôi, gieo uơm, kiểm nghiệm xuất xứ cây trồng. Chủ yếu đánh giá ảnh hưởng của các công thức, nhân tố đến kết quả thí nghiệm, làm cơ sở cho việc lựa chọn công thức, phương pháp tối ưu trong nông lâm nghiệp.

5.1. Phân tích phương sai 1 nhân tố với các thí nghiệm ngẫu nhiên hoàn toàn

Phân tích này có một nhân tố như xuất xứ cây trồng, mật độ trồng khác nhau, chế độ chăm sóc khác nhau, ...Có nghĩa trong đó có a công thức, mỗi công thức được lập lại m lần, số lần lặp của mỗi công thức có thể bằng hoặc không bằng nhau.

Trong trường hợp này có thể sử dụng chương trình phân tích phương sai một nhân tố để kiểm tra ảnh hưởng của các công thức đến kết quả thí nghiệm.

Cách bố trí thí nghiệm trên hiện trường để phân tích phương sai 1 nhân tố

Các công thức của 1 nhân tố	**Số lần lặp lại**			
	1	**2**	**3**	**m**
1	11	12	13	1m
2	21	22		
....	...	...	...	...
a	a1	a2		am

Ví dụ: Đánh giá kết quả khảo nghiệm xuất xứ Pinus caribeae tại Lang Hanh-Lâm Đồng.

Thí nghiệm 7 xuất xứ với 5 xuất xứ lặp lại 4 lần, còn 2 xuất xứ chỉ được lặp lại 2 lần vào năm 1991.

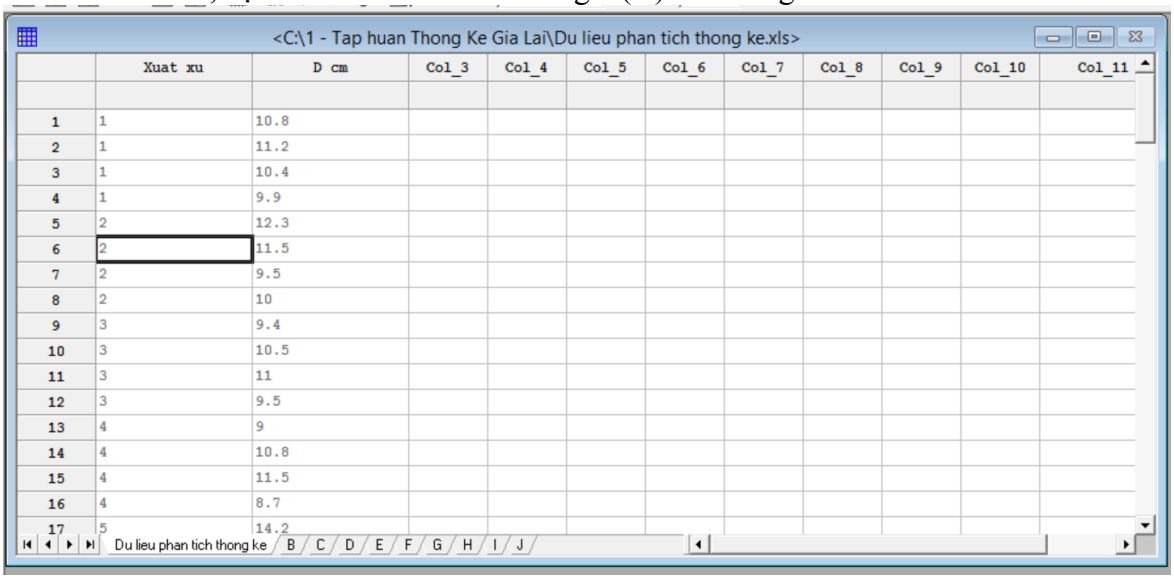
7 xuất xứ P.caribeae được trồng thực tế, được đánh số và lặp lại như sau:

1: Xuất xứ P.alamicamba (NIC) lặp lại 4 lần.
2: P.poptun (Guat) " 4 "
3: P.guanaja (Nonduras) " 4 "
4: P.linures (Nonduras) " 4 "
5: P.R482 (Australia) " 2 "
6: P.T473 (Australia) " 4 "
7: P.little asaco (Bahamas) 2 "

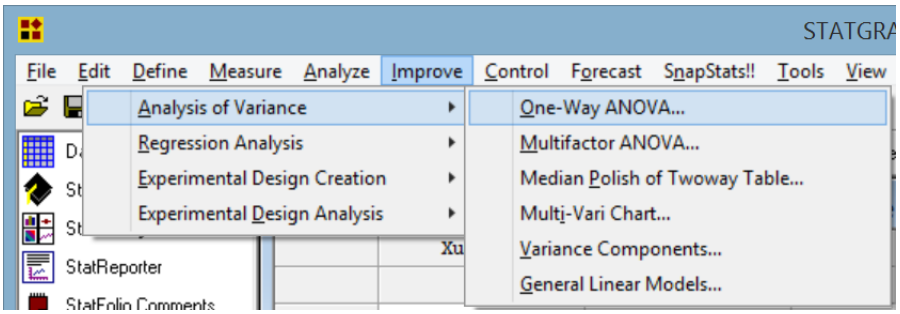
- Mỗi xuất xứ ứng với 1 lần lặp được trồng 25 cây, với cự ly 3x2m, tổng diện tích bố trí thí nghiệm là 1ha.
- Các điều kiện đất đai, vi khí hậu, địa hình, chăm sóc...đều được đồng nhất, nhân tố thay đổi để khảo sát chỉ còn lại là các xuất xứ khác nhau.
- Tại thời điểm điều tra (1996), cây trồng trong các ô thí nghiệm có tuổi là 5. Tiến hành đo đếm toàn diện các chỉ tiêu đường kính ngang ngực (D) , chiều cao (H), đường kính tán (D_t) , phẩm chất, tỉa cành, hình thân. Sử dụng 2 chỉ tiêu D và H để đánh giá sinh trưởng của các xuất xứ thử nghiệm.

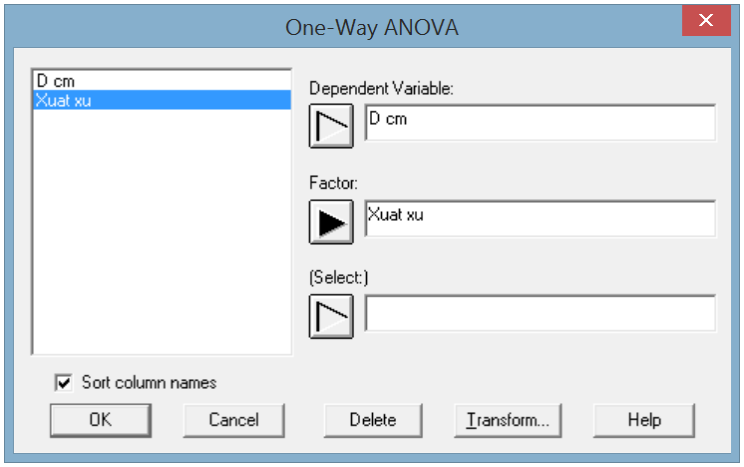
Dùng phân tích phương sai 1 nhân tố để kiểm tra sự sai khác sinh trưởng $D_{1.3}$ của 7 xuất xứ trong Statgraphics

✔ Nhập dữ liệu từ Excel vào Stat: Trong đó có hai cột: Cột nhân tố là xuất xứ khác nhau, cột thứ hai là chỉ tiêu đánh giá (D) theo từng nhân tố:

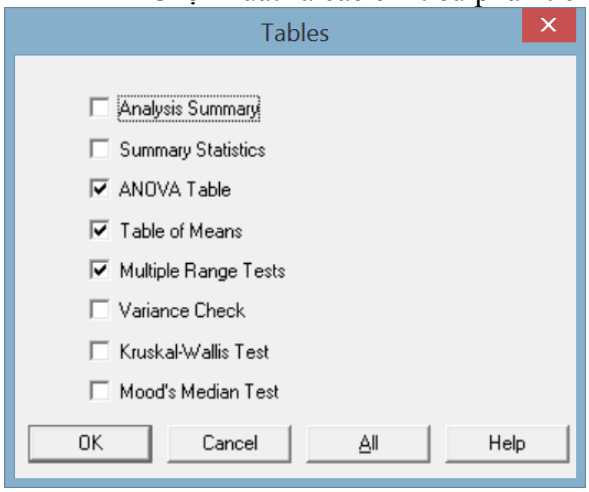


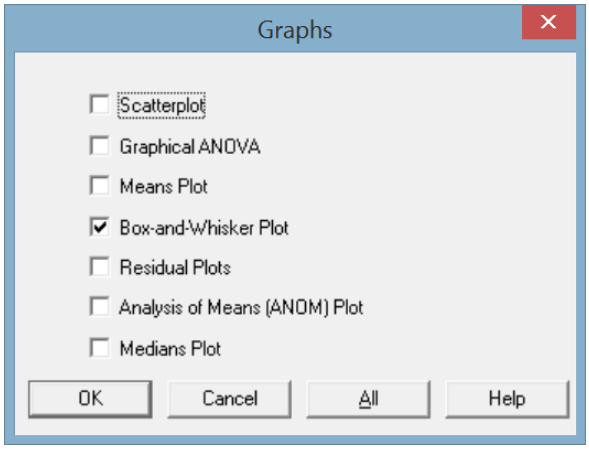
✔ Sử dụng phân tích ANOVA 1 nhân tố trong Stat: Improve/Analysis of Variance/One-Way ANOVA và xác định dữ liệu phụ thuộc vào nhân tố đánh giá





✓ Chọn xuất ra các chỉ tiêu phân tích, đánh giá và đồ thị như sau:





ANOVA Table for D cm by Xuat xu

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	37.0596	6	6.1766	5.22	0.0033
Within groups	20.1	17	1.18235		
Total (Corr.)	57.1596	23			

The StatAdvisor

The ANOVA table decomposes the variance of D cm into two components: a between-group component and a within-group component. The F-ratio, which in this case equals 5.22399, is a ratio of the between-group estimate to the within-group estimate. Since the P-value of the F-test is less than 0.05, there is a statistically significant difference between the

mean D cm from one level of Xuat xu to another at the 95.0% confidence level. To determine which means are significantly different from which others, select Multiple Range Tests from the list of Tabular Options.

Table of Means for D cm by Xuat xu with 95.0 percent LSD intervals

			Stnd. error		
Xuat xu	*Count*	*Mean*	*(pooled s)*	*Lower limit*	*Upper limit*
1	4	10.575	0.54368	9.7639	11.3861
2	4	10.825	0.54368	10.0139	11.6361
3	4	10.1	0.54368	9.2889	10.9111
4	4	10.0	0.54368	9.1889	10.8111
5	2	13.55	0.76888	12.4029	14.6971
6	4	12.0	0.54368	11.1889	12.8111
7	2	8.4	0.76888	7.25293	9.54707
Total	24	10.7458			

The StatAdvisor

This table shows the mean D cm for each level of Xuat xu. It also shows the standard error of each mean, which is a measure of its sampling variability. The standard error is formed by dividing the pooled standard deviation by the square root of the number of observations at each level. The table also displays an interval around each mean. The intervals currently displayed are based on Fisher's least significant difference (LSD) procedure. They are constructed in such a way that if two means are the same, their intervals will overlap 95.0% of the time. You can display the intervals graphically by selecting Means Plot from the list of Graphical Options. In the Multiple Range Tests, these intervals are used to determine which means are significantly different from which others.

Multiple Range Tests for D cm by Xuat xu

Method: 95.0 percent Duncan

Xuat xu	*Count*	*Mean*	*Homogeneous Groups*
7	2	8.4	X
4	4	10.0	XX
3	4	10.1	XX
1	4	10.575	X
2	4	10.825	X
6	4	12.0	XX
5	2	13.55	X

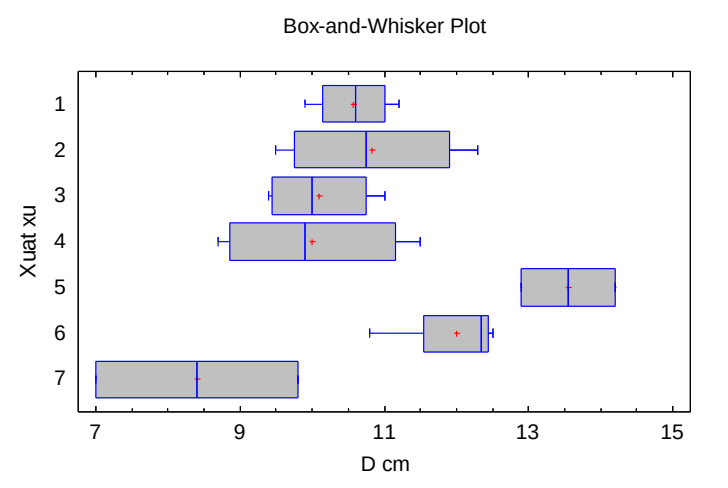
Contrast	*Sig.*	*Difference*
1 - 2		-0.25
1 - 3		0.475
1 - 4		0.575
1 - 5	<*	-2.975
1 - 6		-1.425
1 - 7	<*	2.175
2 - 3		0.725
2 - 4		0.825
2 - 5	<*	-2.725
2 - 6		-1.175
2 - 7	<*	2.425
3 - 4		0.1
3 - 5	<*	-3.45
3 - 6		-1.9
3 - 7		1.7
4 - 5	<*	-3.55
4 - 6		-2.0
4 - 7		1.6
5 - 6		1.55
5 - 7	<*	5.15
6 - 7	<*	3.6

* denotes a statistically significant difference.

The StatAdvisor

This table applies a multiple comparison procedure to determine which means are significantly different from which others. The bottom half of the output shows the estimated difference between each pair of means. An asterisk has been placed next

to 8 pairs, indicating that these pairs show statistically significant differences at the 95.0% confidence level. At the top of the page, 3 homogenous groups are identified using columns of X's. Within each column, the levels containing X's form a group of means within which there are no statistically significant differences. The method currently being used to discriminate among the means is Duncan's multiple comparison procedure. With this method, there is a 5.0% risk of calling one or more pairs significantly different when their actual difference equals 0.



- Bảng kết quả ANOVA nhận được P-value = 0.0033 < 0.05, dẫn đến bác bỏ giả thuyết Ho về sự bằng nhau của các trung bình mẫu; có nghĩa là trung bình D ở các xuất xứ là có sự sai khác ở mức 95%.
- Bảng Table of Means for D cm by Xuat xu with 95.0 percent LSD intervals: Cho kết quả trung bình D và biến động ở độ tin cậy 95% ở mỗi xuất xứ
- Bảng Multiple Range Tests for D cm by Xuat xu: Nếu chỉ dừng lại kiểm tra ANOVA thì mới cho biết là các xuất xứ có sai khác nhau về D, tuy nhiên chưa cho biết sự khác biệt giữa từng xuất xứ với nhau để lựa chọn xuất xứ tối ưu. Bảng này chỉ ra sự khác nhau giữa D bình quân của các xuất xứ ở mức 95% và xếp nhóm đồng nhất theo tiêu chuẩn Duncan. Kết quả cho thấy xuất xứ 5 tốt nhất và chưa có sự khác biệt với xuất 6 nhưng sai khác rõ với xuất xứ 2. Kết quả này cũng được minh họa trong biểu đồ. Như vậy trong nghiên cứu này, xuất xứ 5 là P.R482 (Australia) sẽ cho sinh trưởng tốt nhất và được lựa chọn.

5.2. Phân tích phương sai nhiều nhân tố

Trong các thí nghiệm người ta thường so sánh và phân tích tác động đồng thời nhiều nhân tố lên kết quả thí nghiệm như: năng suất, sinh khối...

5.2.1. Phân tích phương sai 2 nhân tố với 1 lần lặp lại: (Bố trí thí nghiệm theo khối ngẫu nhiên đầy đủ (Randomized Complete Blocks) (RCB):

Kiểu bố trí thí nghiệm RCB thường được sử dụng, nhân tố A chia làm a cấp và nhân tố B (Lần lặp) được chia b cấp (khối), tổ hợp 2 nhân tố chỉ có 1 lần lặp (1 ô thí nghiệm).

Bố trí thí nghiệm trên hiện trường

Lặp 1	Lặp 2	Lặp 3	Lặp 4
A1	Aa	A2	A2
A2	A4	A3	A3
A3	A3	Aa	A1
A4	A2	A4	A4

Aa	A1	A1	Aa

Nhân tố B được chia thành b khối, ở mỗi khối bố trí a công thức của nhân tố A một cách ngẫu nhiên.

Ví dụ: Đánh giá kết quả khảo nghiệm 16 xuất xứ Pinus kesiya tại Lang Hanh-Lâm Đồng: 16 xuất xứ P.kesiya đã được trồng khảo nghiệm tại trạm thực nghiệm Lang Hanh năm 1991. Việc bố trí thí nghiệm đã được tiến hành theo khối ngẫu nhiên đầy đủ RCB (Randomized Complete Blocks), bao gồm 16 công thức chỉ thị 16 xuất xứ và được lặp lại ở 4 (khối)

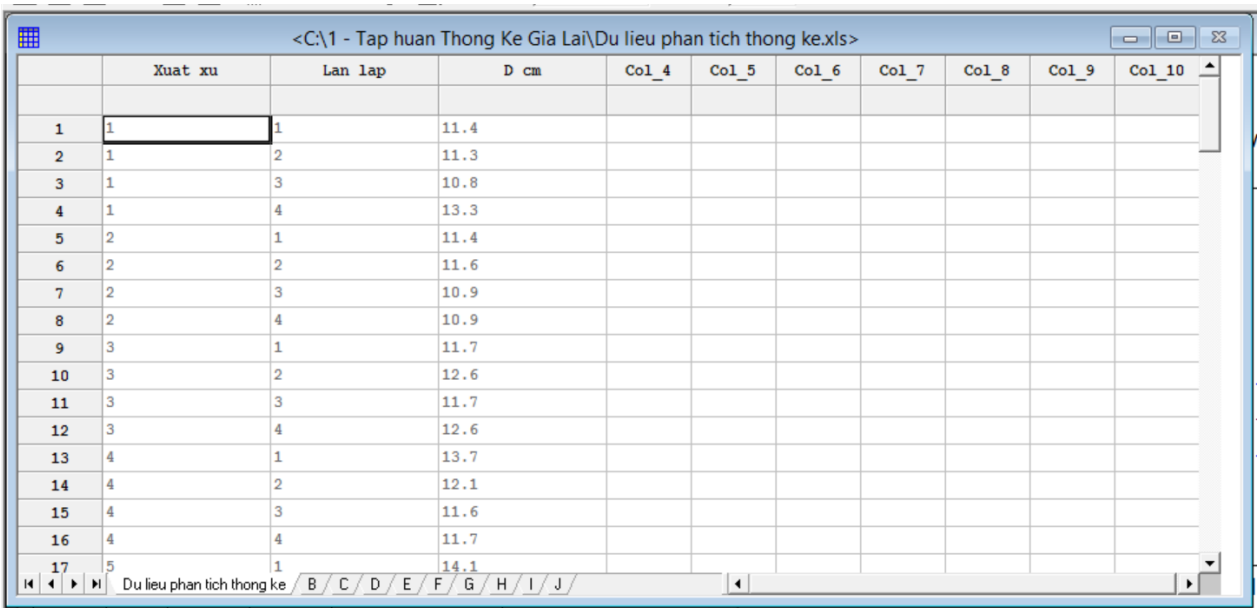
16 xuất xứ P.kesiya được đánh số như sau:

1: Xuất xứ Bengliet.
2: Faplac.
3: Xuân Thọ.
4: Thác Prenn.
5: Lang Hanh.
6: Nong Kiating.
7: Doisupthep.
8: Doiinthranon.
9: Phu Kradung.
10: Nam nouv.
11: Cotomines.
12: Simao.
13: Watchan.
14: Zo khua.
15: Aung ban.
16: Jingdury.

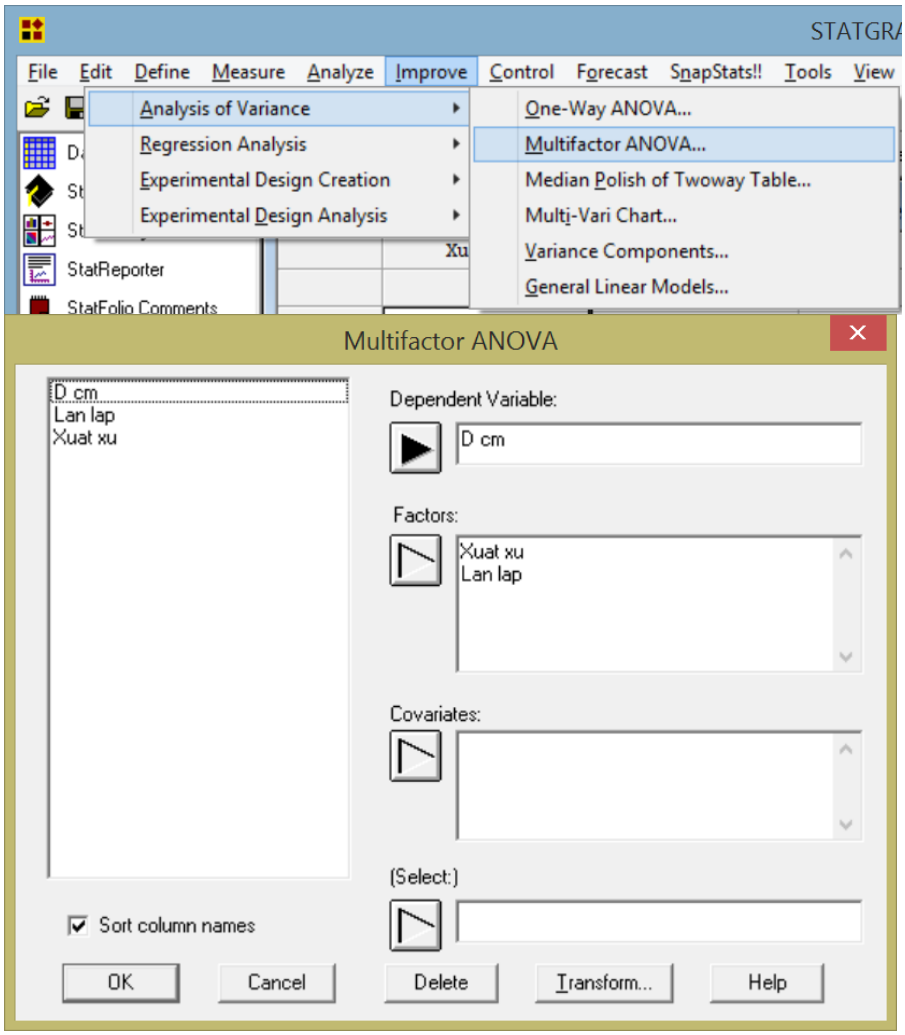
- Mỗi công thức ứng với 1 lần lặp được trồng 25 cây, với cự ly 3x2m, tổng diện tích bố trí thí nghiệm là 1,5ha.
- Các khí hậu, địa hình, chăm sóc...đều được đồng nhất, nhân tố thay đổi để khảo sát chỉ còn lại là các xuất xứ và cấp đất khác nhau.
- Tại thời điểm điều tra (1996), cây trồng trong các ô thí nghiệm có tuổi là 5. Tiến hành đo đếm toàn diện các chỉ tiêu D, H, D_t, phẩm chất, tỉa cành, hình thân. Sử dụng 2 chỉ tiêu D và H để đánh giá sinh trưởng của các xuất xứ thử nghiệm.

Dùng phân tích phương sai nhiều nhân tố để kiểm tra trong Stat:

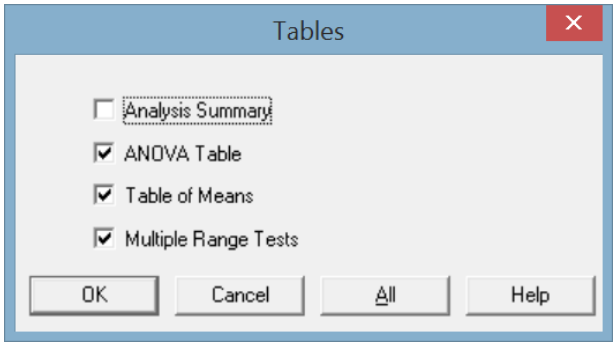
- ✓ Nhập dữ liệu từ Excel vào Stat: Gồm 3 trường dữ liệu: Nhân tố xuất xứ, lần lặp (khối) và trường dữ liệu quan sát ừng với từng xuất xứ là lần lặp lại.

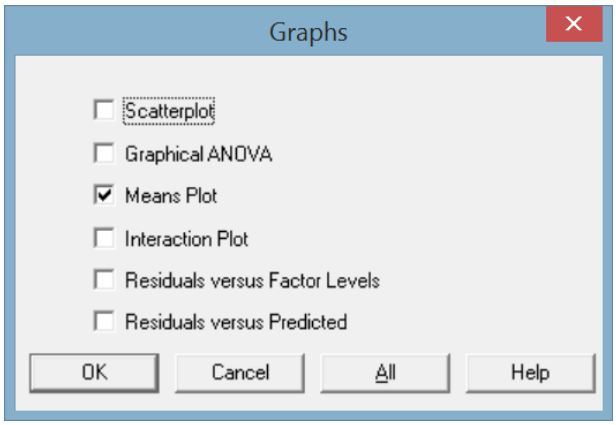


✔ Sử dụng phân tích ANOVA nhiều nhân tố trong Stat: Analysis of Variance/Multifactor ANOVA



✔ Xuất ra chỉ tiêu thống kê đánh giá và đồ thị so sánh:





Analysis of Variance for D cm - Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:Xuat xu	81.775	15	5.45167	7.72	0.0000
B:Lan lap	3.47625	3	1.15875	1.64	0.1931
RESIDUAL	31.7587	45	0.70575		
TOTAL (CORRECTED)	117.01	63			

All F-ratios are based on the residual mean square error.

Table of Least Squares Means for D cm with 95.0 Percent Confidence Intervals

Level	*Count*	*Mean*	*Stnd. Error*	*Lower Limit*	*Upper Limit*
GRAND MEAN	64	11.8875			
Xuat xu					
1	4	11.7	0.420045	10.854	12.546
2	4	11.2	0.420045	10.354	12.046
3	4	12.15	0.420045	11.304	12.996
4	4	12.275	0.420045	11.429	13.121
5	4	13.775	0.420045	12.929	14.621
6	4	12.1	0.420045	11.254	12.946
7	4	12.525	0.420045	11.679	13.371
8	4	13.9	0.420045	13.054	14.746
9	4	12.4	0.420045	11.554	13.246
10	4	11.75	0.420045	10.904	12.596
11	4	12.35	0.420045	11.504	13.196
12	4	11.55	0.420045	10.704	12.396
13	4	12.225	0.420045	11.379	13.071
14	4	9.35	0.420045	8.50398	10.196
15	4	9.975	0.420045	9.12898	10.821
16	4	10.975	0.420045	10.129	11.821
Lan lap					
1	16	12.2625	0.210022	11.8395	12.6855
2	16	11.8813	0.210022	11.4582	12.3043
3	16	11.7687	0.210022	11.3457	12.1918

4	16	11.6375	0.210022	11.2145	12.0605

Multiple Range Tests for D cm by Xuat xu

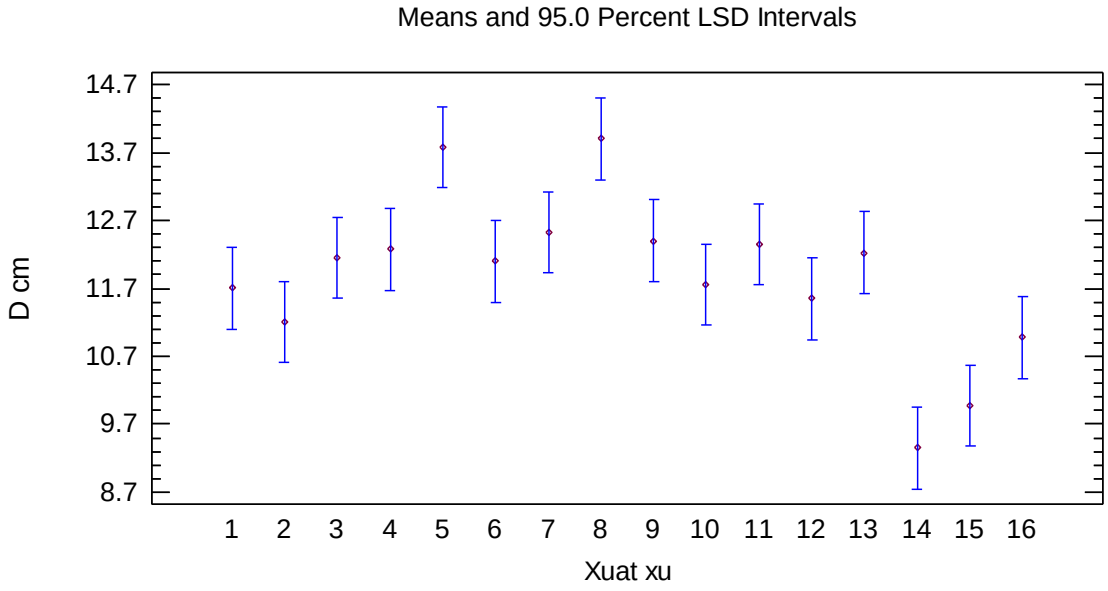
Method: 95.0 percent Duncan

Xuat xu	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
14	4	9.35	0.420045	X
15	4	9.975	0.420045	XX
16	4	10.975	0.420045	XX
2	4	11.2	0.420045	XXX
12	4	11.55	0.420045	XX
1	4	11.7	0.420045	XX
10	4	11.75	0.420045	XX
6	4	12.1	0.420045	XX
3	4	12.15	0.420045	XX
13	4	12.225	0.420045	XX
4	4	12.275	0.420045	XX
11	4	12.35	0.420045	XX
9	4	12.4	0.420045	X
7	4	12.525	0.420045	X
5	4	13.775	0.420045	X
8	4	13.9	0.420045	X

Multiple Range Tests for D cm by Lan lap

Method: 95.0 percent Duncan

Lan lap	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
4	16	11.6375	0.210022	X
3	16	11.7687	0.210022	X
2	16	11.8813	0.210022	X
1	16	12.2625	0.210022	X



Từ kết quả rút ra kết luận:

- Bảng ANOVA:
  - ✓ Giữa các xuất xứ có P-value = 0.000 < 0,05, bác bỏ Ho; có nghĩa là giữa các xuất xứ có sự sai khác về sinh trưởng D

- ✓ Giữa các lần lặp có P-value = 0.19 > 0.05, chấp nhận Ho; có nghĩa giữa các lần lặp chưa có sự sai khác, nói khác có sự đồng nhất điều kiện hoàn cảnh ở các lần lặp, bảo đảm khách quan

- Bảng Table of Least Squares Means for D cm with 95.0 Percent Confidence Intervals chỉ ra giá trị trung bình D và biến động ở mức tin cậy 95% cho từng xuất xứ.
- Bảng Multiple Range Tests for D cm by Xuat xu chỉ ra các nhóm xuất xứ đồng nhất và sai khác rõ rệt với các nhóm khác ở độ tin cậy 95% theo Duncan. Kết quả chỉ ra nhóm bao gồm xuất xứ Lang Hanh (5) và Doiinthranon (8) là các xuất xứ tốt nhất và sai khác rõ rệt với các nhóm xuất xứ khác. Kết quả biểu diễn rõ trên đồ thị
- Bảng Multiple Range Tests for D cm by Lan lap cũng chỉ ra sự đồng nhất ở 4 lần lặp.

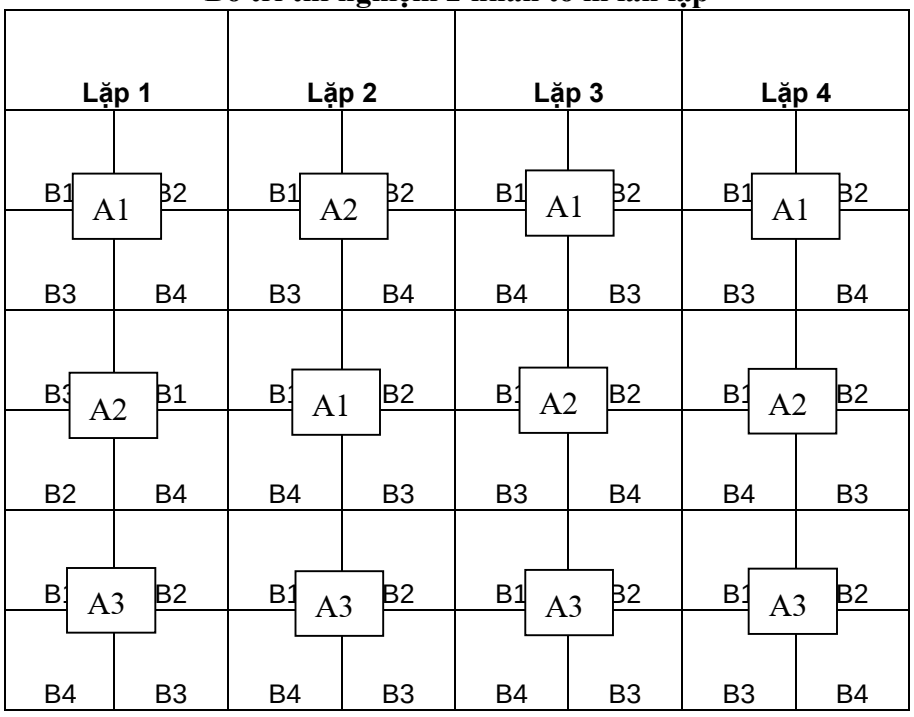
5.2.2. Phân tích phương sai 2 nhân tố m lần lặp

Trường hợp phân tích phương sau 2 nhân tố m lần lặp: Nhân tố A có a công thức và nhân tố B có b công thức; và này mỗi tổ hợp nhân tố A và B được lặp lại m lần một cách ngẫu nhiên. Lúc này ngoài việc đánh giá ảnh hưởng của từng nhân tố A, B ta còn phải tính ảnh hưởng qua lại của chúng đến kết quả thí nghiệm.

Ví dụ: Nghiên cứu ảnh hưởng của hai nhân tố thí nghiệm là mật độ và bón phân đến năng suất của Bông.

- Nhân tố A: Mật độ chia làm 3 cấp: A1, A2 và A3
- Nhân tố B: Phân bón được chia làm 4 mức: B1, B23, B3 và B4.
- Mỗi tổ hợp được thí nghiệm lập lại ngẫu nhiên 4 lần.

Bố trí thí nghiệm 2 nhân tố m lần lặp

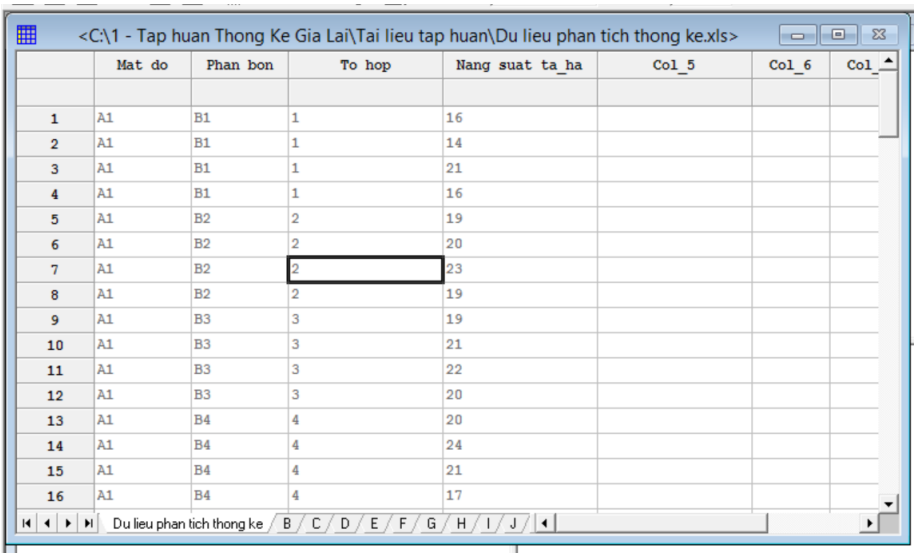


Bảng số liệu năng suất bông theo tổ hợp 2 nhân tố và lặp lại 4 lần ở một tổ hợp
(Đ/v: Tạ/ha)

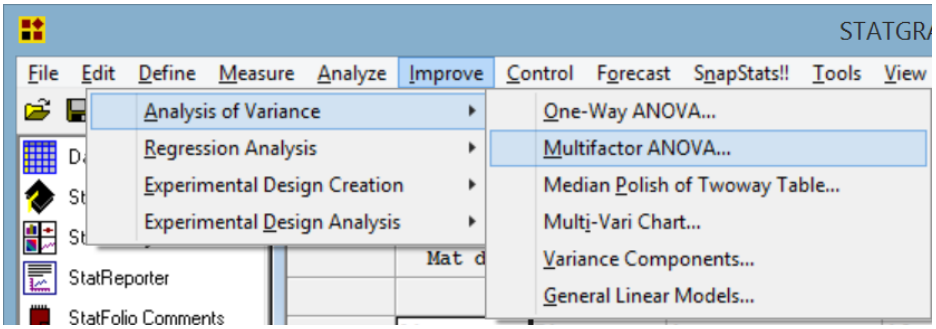
	A	B	C	D
1	**B\A**	A1	A2	A3
2	B1	16	17	18
3		14	15	18
4		21	17	19
5		16	19	17
6	B2	19	19	20
7		20	18	23
8		23	18	21
9		19	20	21
10	B3	19	21	22
11		21	21	18
12		22	22	21
13		20	23	21
14	B4	20	20	25
15		24	20	22
16		21	22	21
17		17	19	23

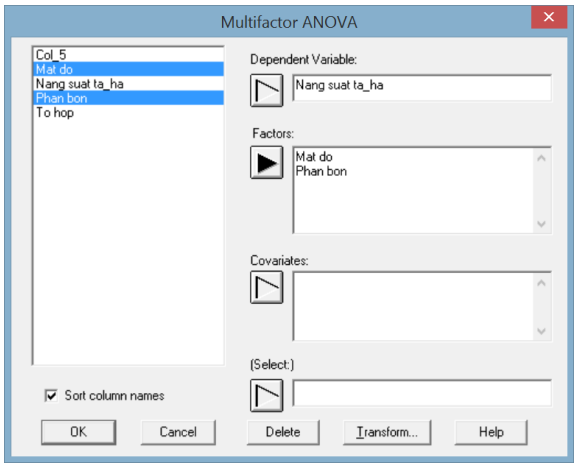
Sử dụng phân tích phương sai nhiều nhân tố m lần lặp trong Stat:

- ✓ Nhập dữ liệu từ Excel vào Stat: Gồm 4 trường: Nhân tố A (Mật độ), nhân tố B (Phân bón), Tổ hợp: Nhân tố A có 3 công thức * Nhân tố B có 4 công thức = 12 tổ hợp được đánh số lần lượt là 1, 2, 3, …. 12; trường cuối cùng là chỉ tiêu đánh giá (năng suất Bôn) tương ứng theo từng tổ hợp nhân tố.

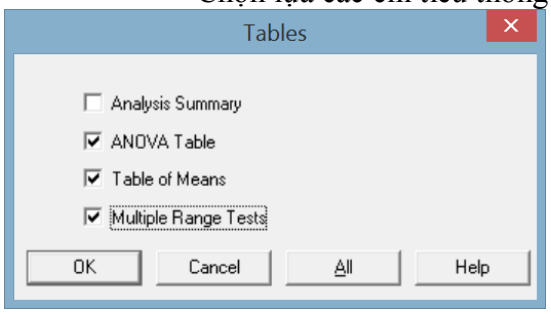


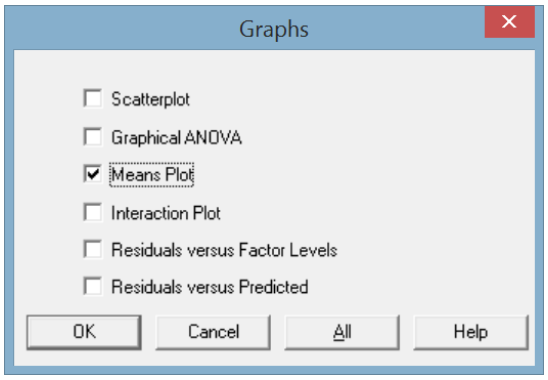
✓ Sử dụng chức năng phân tích ANVA đa biến; Improve/Analysis of Variance/Multifactor ANOVA và nhập biến số phụ thuộc, các nhân tố ảnh hưởng trong hộp thoại





✓ Chọn lựa các chỉ tiêu thống kê và đồ thị như sau:





✓ Kết quả như sau:

Analysis of Variance for Nang suat ta_ha - Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:Mat do	14.2917	2	7.14583	2.28	0.1145
B:Phan bon	116.229	3	38.7431	12.38	0.0000
RESIDUAL	131.458	42	3.12996		
TOTAL (CORRECTED)	261.979	47			

All F-ratios are based on the residual mean square error.

Table of Least Squares Means for Nang suat ta_ha with 95.0 Percent Confidence Intervals

			Stnd.	*Lower*	*Upper*
Level	*Count*	*Mean*	*Error*	*Limit*	*Limit*
GRAND MEAN	48	19.8542			
Mat do					
A1	16	19.5	0.442292	18.6074	20.3926
A2	16	19.4375	0.442292	18.5449	20.3301
A3	16	20.625	0.442292	19.7324	21.5176
Phan bon					
B1	12	17.25	0.510715	16.2193	18.2807
B2	12	20.0833	0.510715	19.0527	21.114
B3	12	20.9167	0.510715	19.886	21.9473
B4	12	21.1667	0.510715	20.136	22.1973

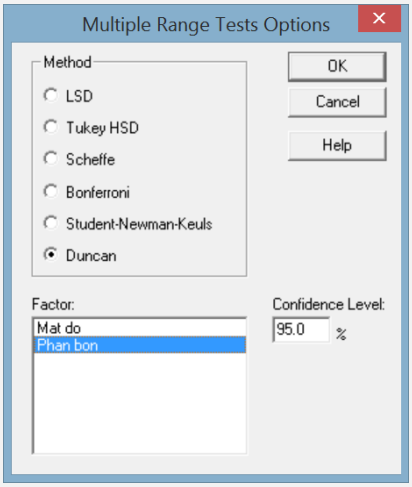
Multiple Range Tests for Nang suat ta_ha by Mat do

Method: 95.0 percent Duncan

Mat do	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
A2	16	19.4375	0.442292	X
A1	16	19.5	0.442292	X
A3	16	20.625	0.442292	X

Contrast	*Sig.*	*Difference*
A1 - A2		0.0625
A1 - A3		-1.125
A2 - A3		-1.1875

* denotes a statistically significant difference.



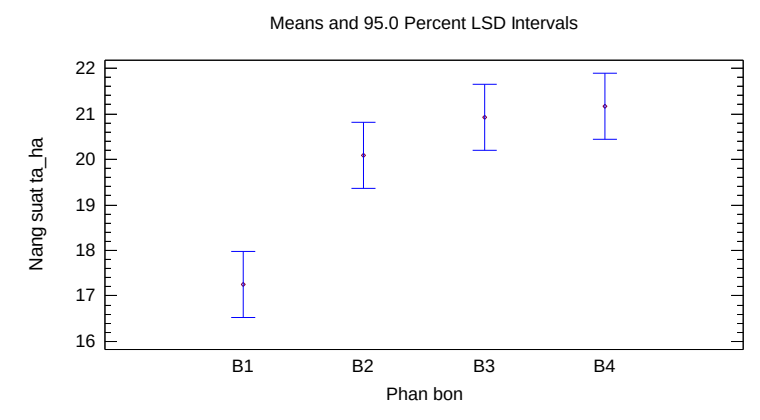
Multiple Range Tests for Nang suat ta_ha by Phan bon

Method: 95.0 percent Duncan

Phan bon	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
B1	12	17.25	0.510715	X
B2	12	20.0833	0.510715	X
B3	12	20.9167	0.510715	X
B4	12	21.1667	0.510715	X

Contrast	*Sig.*	*Difference*
B1 - B2	*	-2.83333
B1 - B3	*	-3.66667
B1 - B4	*	-3.91667
B2 - B3		-0.833333
B2 - B4		-1.08333
B3 - B4		-0.25

* denotes a statistically significant difference.

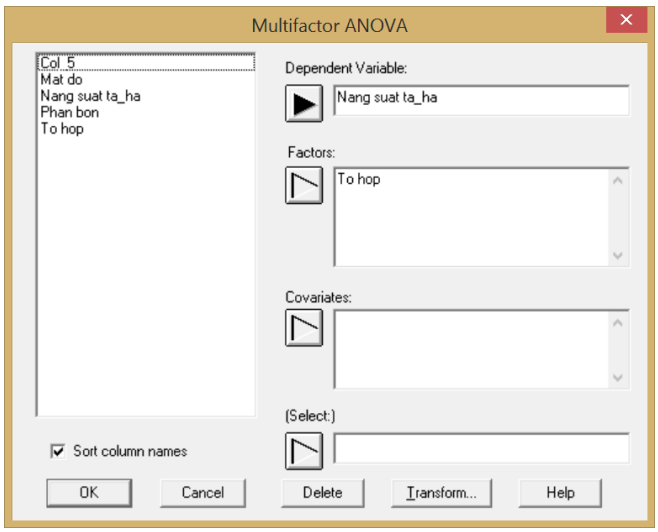


Kết quả cho thấy:

- Bảng ANOVA:
  - ✓ Đối với nhân tố mật độ, P-Value = 0.115 > 0.05; giả thuyết Ho được chấp nhận; có nghĩa là trung bình năng suất bông ở các mật độ là chưa có sự sai khác rõ
  - ✓ Đối với nhân tố phân bón, P-Value = 0.000 < 0.05; giả thuyết Ho bị bác bỏ; có nghĩa là trung bình năng suất bông ở các công thức phân bón là có sự sai khác có ý nghĩa ở 95%
- Bảng Table of Least Squares Means for Nang suat ta_ha with 95.0 Percent Confidence Intervals chỉ ra trung bình năng suất bông theo các công thức bón phân và mật độ và biến động ở 95%
- Bảng Multiple Range Tests for Nang suat ta_ha by Mat do chỉ ra tất cả công thức mật độ đều xuất phát từ một nhóm.
- Bảng Multiple Range Tests for Nang suat ta_ha by Phan bon qua phân tích DunCan cho thấy ảnh hưởng của phân bón thí nghiệm chia làm 2 nhóm: Nhóm năng suất thấp nhất ở công thức bón phân B1, và nhóm có nang suất bông cao nhất bao gồm các công thức B2-B3-B4 và chúng khác nhau có ý nghĩa ở mức tin cậy 95%. Đồ thị cũng chỉ ra sự phân nhóm năng suất theo công thức phân bón.

Trong trường hợp cả 2 nhân tố A và B đều có ảnh hưởng đến chỉ tiêu đánh giá, lúc này cần tìm ra tổ hợp 2 nhân tố tối ưu, tức là tổ hợp công thức mật và bón phân nào là tốt nhất và có sự sai khác với các tổ hợp khác. Trong ví dụ năng suất bông nói trên, có 12 tổ hợp công thức mật độ và bón phân đượng mã hóa từ 1, 2, … 12. Tiến hành so sánh 12 tổ hợp này này với nhau để tìm ra tổ hợp tối ưu.

Trong bảng ANOVA cho nhân tố là tổ hợp, biến phụ thuộc vẫn là năng suất bông:



Kết quả phân tích trong Stst cho ra như sau:

Analysis of Variance for Nang suat ta_ha - Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:To hop	151.729	11	13.7936	4.50	0.0003
RESIDUAL	110.25	36	3.0625		
TOTAL (CORRECTED)	261.979	47			

All F-ratios are based on the residual mean square error.

Table of Least Squares Means for Nang suat ta_ha with 95.0 Percent Confidence Intervals

			Stnd.	*Lower*	*Upper*
Level	*Count*	*Mean*	*Error*	*Limit*	*Limit*
GRAND MEAN	48	19.8542			
To hop					
1	4	16.75	0.875	14.9754	18.5246
2	4	20.25	0.875	18.4754	22.0246
3	4	20.5	0.875	18.7254	22.2746
4	4	20.5	0.875	18.7254	22.2746
5	4	17.0	0.875	15.2254	18.7746
6	4	18.75	0.875	16.9754	20.5246
7	4	21.75	0.875	19.9754	23.5246
8	4	20.25	0.875	18.4754	22.0246
9	4	18.0	0.875	16.2254	19.7746
10	4	21.25	0.875	19.4754	23.0246
11	4	20.5	0.875	18.7254	22.2746
12	4	22.75	0.875	20.9754	24.5246

Multiple Range Tests for Nang suat ta_ha by To hop

Method: 95.0 percent Duncan

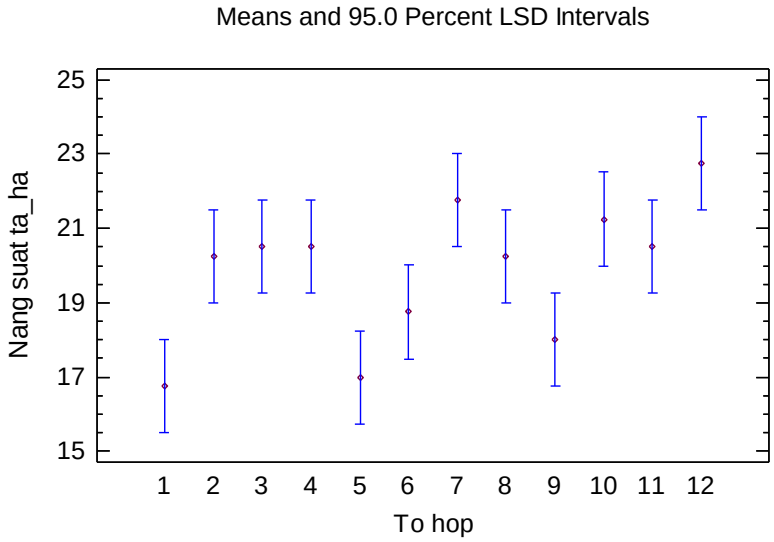
To hop	*Count*	*LS Mean*	*LS Sigma*	*Homogeneous Groups*
1	4	16.75	0.875	X
5	4	17.0	0.875	X
9	4	18.0	0.875	XX
6	4	18.75	0.875	XXX

8	4	20.25	0.875	XXX
2	4	20.25	0.875	XXX
3	4	20.5	0.875	XXX
11	4	20.5	0.875	XXX
4	4	20.5	0.875	XXX
10	4	21.25	0.875	XX
7	4	21.75	0.875	X
12	4	22.75	0.875	X

Contrast	*Sig.*	*Difference*
1 - 2	*	-3.5
1 - 3	*	-3.75
1 - 4	*	-3.75
1 - 5		-0.25
1 - 6		-2.0
1 - 7	*	-5.0
1 - 8	*	-3.5
1 - 9		-1.25
1 - 10	*	-4.5
1 - 11	*	-3.75
1 - 12	*	-6.0
2 - 3		-0.25
2 - 4		-0.25
2 - 5	*	3.25
2 - 6		1.5
2 - 7		-1.5
2 - 8		0.0
2 - 9		2.25
2 - 10		-1.0
2 - 11		-0.25
2 - 12		-2.5
3 - 4		0.0
3 - 5	*	3.5
3 - 6		1.75
3 - 7		-1.25
3 - 8		0.25
3 - 9		2.5
3 - 10		-0.75
3 - 11		0.0
3 - 12		-2.25
4 - 5	*	3.5
4 - 6		1.75
4 - 7		-1.25
4 - 8		0.25
4 - 9		2.5
4 - 10		-0.75
4 - 11		0.0
4 - 12		-2.25
5 - 6		-1.75
5 - 7	*	-4.75
5 - 8	*	-3.25
5 - 9		-1.0
5 - 10	*	-4.25
5 - 11	*	-3.5
5 - 12	*	-5.75
6 - 7	*	-3.0
6 - 8		-1.5
6 - 9		0.75
6 - 10		-2.5
6 - 11		-1.75
6 - 12	*	-4.0
7 - 8		1.5

7 - 9	*	3.75
7 - 10		0.5
7 - 11		1.25
7 - 12		-1.0
8 - 9		2.25
8 - 10		-1.0
8 - 11		-0.25
8 - 12		-2.5
9 - 10	*	-3.25
9 - 11		-2.5
9 - 12	*	-4.75
10 - 11		0.75
10 - 12		-1.5
11 - 12		-2.25

* denotes a statistically significant difference



Với kết quả phân tích sự khác biệt giữa 12 tổ hợp 2 nhân tố A và B cho thấy:

- Bảng Analysis of Variance for Nang suat ta_ha: P-value = 0.0003 < 0.05, có nghĩa là bác bổ giả thuyết Ho về sự bằng nhau giữa trung bình năng suất ở các tổ hợp 2 nhân tố. Hay nói khác có sự khác biệt có ý nghĩa về trung bình năng suất của 12 tổ hợp ở độ tin cậy 95%.
- Bảng Table of Least Squares Means for Nang suat ta_ha with 95.0 Percent Confidence Intervals: Chỉ ra trung bình và biến động của 12 tổ hợp ở mức tin cậy 95%.
- Bảng Multiple Range Tests for Nang suat ta_ha by To hop: Trắc nghiệm Duncan chỉ ra sự sắp xếp theo nhóm các tổ hợp đồng nhất và khác biệt nhau. Kết quả cho thấy tổ hợp 7 và 12 là cho năng suất cao nhất và chưa có sự sai khác với nhau, đồng thời chúng cũng cùng nhóm với các tổ hợp 6, 8, 2, 3, 11, 4 và 10 và có sự khác biệt rõ với các tổ hợp còn lại như 9, 5, 1. Trong trường hợp này có thể chọn tổ hợp công thức mật độ và bón phân 7 hoặc 12 làm tối ưu. Sự khác biệt năng suất ở 12 tổ hợp cũng được chỉ ra trên biểu đồ.

6. PHÂN TÍCH TƯƠNG QUAN - HỒI QUY

Trong thực tế người ta cần lập các mô hình tương quan hồi quy vì các mục đích:

- Để ước lượng một nhân tố khó đo đếm (gọi là biến phụ thuộc y) thông qua một hay nhiều biến dễ quan sát, đo đếm (gọi là biến độc lập x) và tất nhiên là phải có mối liên hệ giữa y và x. Từ đây có thể lập các biểu điều tra phục vụ cho việc giảm nhẹ các quan sát đo đếm một số nhân tố phức tạp
- Để nghiên cứu tác động, ảnh hưởng của một hoặc nhiều nhân tố đến một yếu tố cần quan tâm như sinh trưởng, sản lượng, chất lượng rừng, xói mòn đất, dòng chảy lưu vực. Trên cơ sở đó có giải pháp kỹ thuật thích hợp hoặc các biện pháp thích hợp.
- Để dự báo một nhân tố trong tương lai (gọi là biến dự báo y) với một số biến độc lập, đầu vào (gọi là là biến độc lập x)

Sử dụng chương trình Regression trong Statgraphics để thiết lập các mô hình tương quan/hồi quy tuyến tính từ một cho đến nhiều biến số độc lập. Trong tài liệu này này, các tham số được ước lượng bằng phương pháp bình phương tối thiểu có hay không có trọng số. Riêng các dạng phi tuyến khi ứng dụng chương trình này được đổi biến số để quy về dạng tuyến tính.

Tiêu chí thống kê lựa chọn mô hình tối ưu: Dạng hàm và biến số đầu vào

Quan hệ giữa đại lượng phụ thuộc y và biến số độc lập x trong sinh học, sinh thái môi trường rừng thường có kiểu dạng phức tạp, do vậy việc lựa chọn được mô hình tối ưu để mô tả tốt nhất mối quan hệ y = f(xi) trong thực tế cần dựa vào nhiều tiêu chí khác nhau.

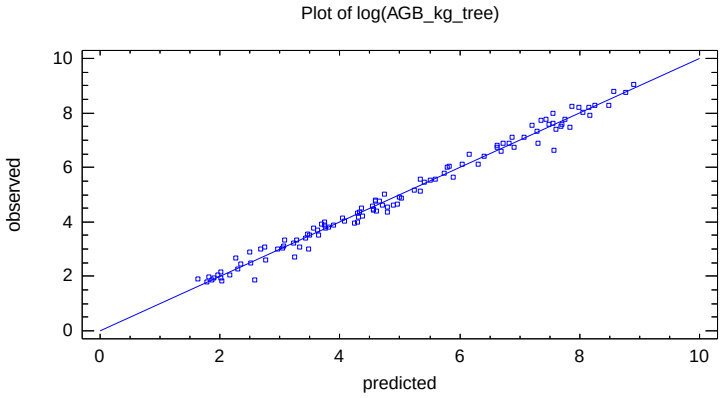
Phổ biến nhất dựa vào các chỉ tiêu thống kê:

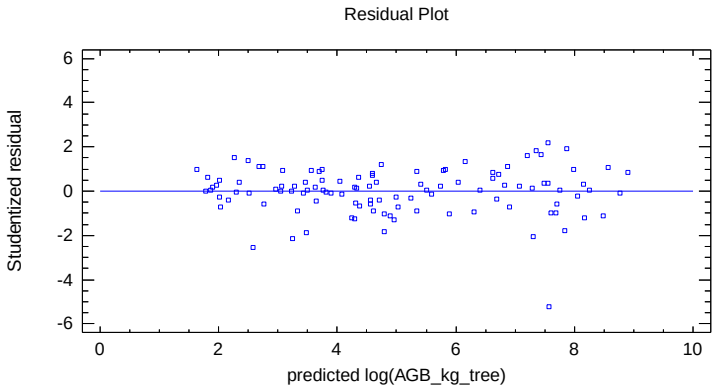
- *Hệ số xác định R^2:* Về tổng quát thì hàm tốt nhất khi R^2 đạt max và tồn tại ở mức sai $P < 0.05$. Tuy nhiên có trường hợp R^2 đạt max nhưng chưa phải là hàm phù hợp nhất, do vậy cần dựa thêm các chỉ tiêu thống kê khác.
- *Tồn tại của các tham số*: Nếu là hàm có từ 2 biến số độc lập trở lên, thì biến độc lập phải tồn tại qua kiểm tra theo tiêu chuẩn t ở mức $P < 0.05$.
- *MAE: Mean absolute error (Sai số tuyệt đối trung bình):* Giá trị MAE càng nhỏ thì mô hình càng tốt:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\mathrm{Yipre} - \mathrm{Yi}|$$

Với, Yipre: giá trị dự báo qua mô hình; Yi: giá trị quan sát; n = số mẫu.

- *Biểu đồ quan hệ giữa số liệu quan sát với ước tính qua mô hình và biểu đồ biến động phần dư (residual) ứng với các giá trị dự báo y của mô hình lựa chọn:* Mô hình tốt khi giá trị ước tính và quan sát bám sát nhau trên đường chéo và biến động residual tập trung quanh trục y = 0 và trong phạm vi giá trị -2 đến + 2 ứng với các giá trị dự báo y. Được minh họa ở 2 biểu đồ sau:





6.1. Mô hình một biến số

Hồi quy một biến số có nghĩa là có một biến số độc lập x được nghiên cứu ảnh hưởng đến biến phụ thuộc y, dạng quan hệ có thể là đường thẳng hoặc phi tuyến.

Ví dụ nghiên cứu thiết lập mô hình tối ưu ước tính tổng sinh khối cây rừng (AGB, kg) theo biến số đường kính (DBH, cm).

i) Mô hình tuyến tính một biến số:

Thiết lập trong Stat:

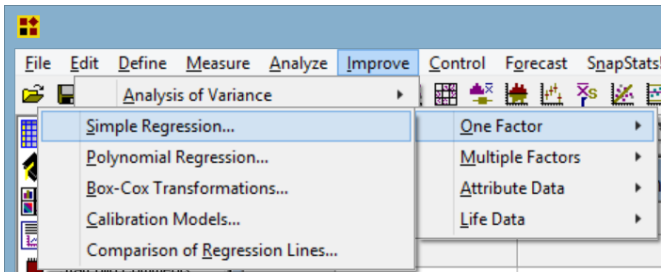
✓ Nhập dữ liệu đầu vào từ Excel:

<C:\1 - Tap huan Thong Ke Gia Lai\Du lieu phan tich thong ke.xls>

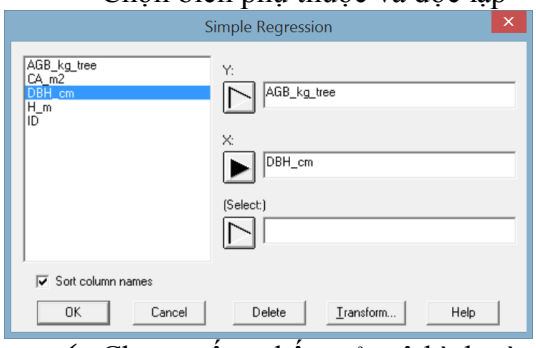
	ID	DBH_cm	CA_m2	H_m	AGB_kg_tree	Col_6	Col_7	Col_8
1	1	14.6	16.61902513749	12	67.6648920023576			
2	2	9.6	7.06858347057703	9.3	27.8868589281281			
3	3	12.1	9.0792027688745	11.5	47.8051988759639			
4	4	11.4	9.0792027688745	9.3	39.9128818615019			
5	5	13.6	13.854423602331	13.5	53.4249098217194			
6	6	6.5	7.06858347057703	6.2	14.6723394857834			
7	7	13.4	13.2025431267111	14	75.9996377584171			
8	8	9.3	5.7255526111674	11.6	21.8585100755737			
9	9	13.5	10.7521008569111	15.1	77.6830156380327			
10	10	12	13.2025431267111	13.9	57.1177193024927			
11	11	6.9	2.83528736986479	7.6	11.7745311547026			
12	12	11	6.15752160103599	11.5	33.9776045374154			
13	13	14.5	13.2025431267111	17.4	120.260913512052			
14	14	6.2	2.54469004940773	9.5	9.71040650556276			
15	15	10.6	9.62112750161874	12.9	54.262484712087			
16	16	5.6	1.32732289614169	12.2	7.73896626652478			
17	17	10.2	7.06858347057703	10.9	33.1638902926283			

Du lieu phan tich thong ke | B | C | D | E | F | G | H | I | J

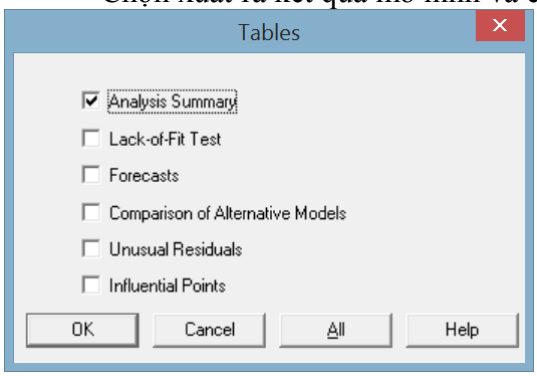
✓ Lập mô hình tuyến tính: AGB = a + b*DBH trong Stat: Improve/Regression Analysis/One Factor/Simple Regression

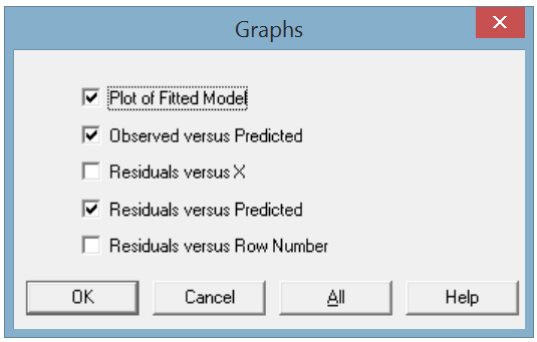


✓ Chọn biến phụ thuộc và độc lập



✓ Chọn xuất ra kết quả mô hình và các đồ thị





Simple Regression - AGB_kg_tree vs. DBH_cm

Dependent variable: AGB_kg_tree

Independent variable: DBH_cm

Linear model: Y = a + b*X

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*

Intercept	-794.609	101.323	-7.84232	0.0000
Slope	62.3168	3.04965	20.4341	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.90298E8	1	1.90298E8	417.55	0.0000
Residual	4.92205E7	108	455746.		
Total (Corr.)	2.39518E8	109			

Correlation Coefficient = 0.891348

R-squared = 79.4502 percent

R-squared (adjusted for d.f.) = 79.2599 percent

Standard Error of Est. = 675.089

Mean absolute error = 419.778

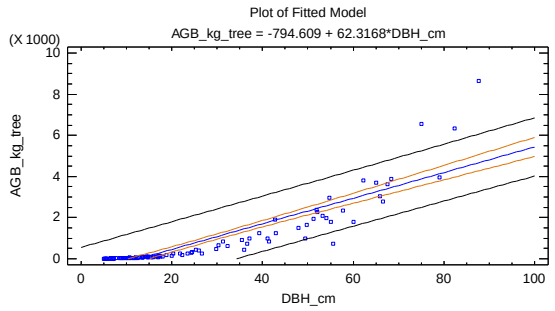
Durbin-Watson statistic = 1.39942 (P=0.0007)

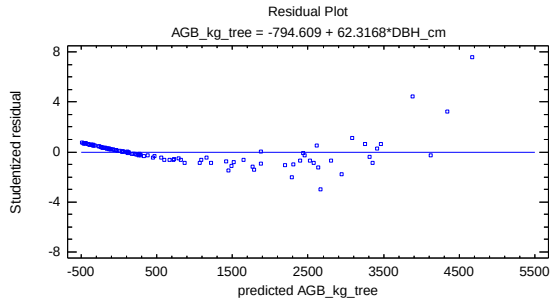
Lag 1 residual autocorrelation = 0.226734

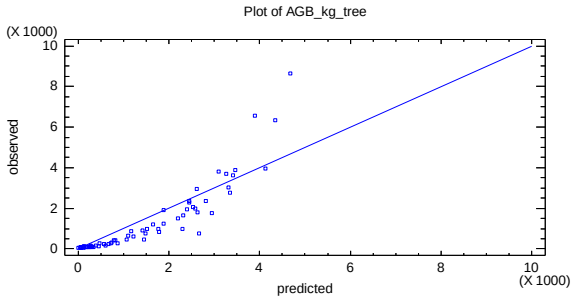
The StatAdvisor

The output shows the results of fitting a linear model to describe the relationship between AGB_kg_tree and DBH_cm. The equation of the fitted model is

AGB_kg_tree = -794.609 + 62.3168*DBH_cm







Kết quả cho thấy đối với mô hình tuyến tính:

- Hệ số R^2 cũng khá cao: R-squared (adjusted for d.f.) = 79.2599 percent và tồn tại với P < 0.05 (ANOVA)
- Tham số b (Slope) tồn tại ở mức P < 0.05
- MAE (Mean absolute error) = 419.778
- Biểu đồ biểu diễn quan hệ giữa quan sát (Observed) và dự báo (Predicted) nằm khá lệch đường chéo. Biểu đồ biến động phần dư Residuals không phân bố đều quanh giá trị dự báo.

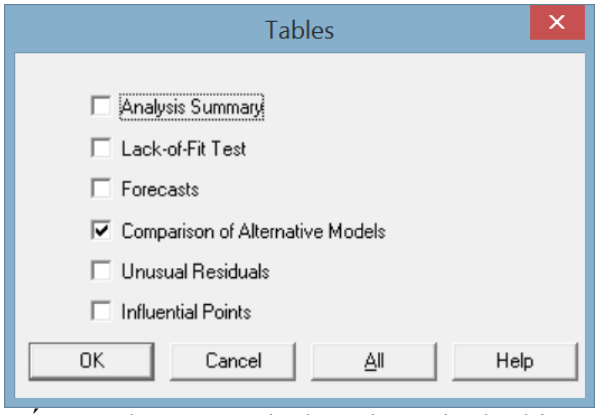
Như vậy có thể thấy mô hình quan hệ AGB = a + b*DBH là chưa phù hợp với dữ liệu quan sát

ii) Mô hình phi tuyến tính một biến số:

Trên cơ sở khảo sát trên cho thấy cần tìm mô hình phi tuyến để ước tính tốt hơn AGB theo DBH

Trong Statgraphics có công cụ hỗ trợ để phát hiện mô hình phi tuyến tốt nhất trên cơ sở R^2 cao nhất.

Trong hộp thoại chọn Comparision of Alternative Models



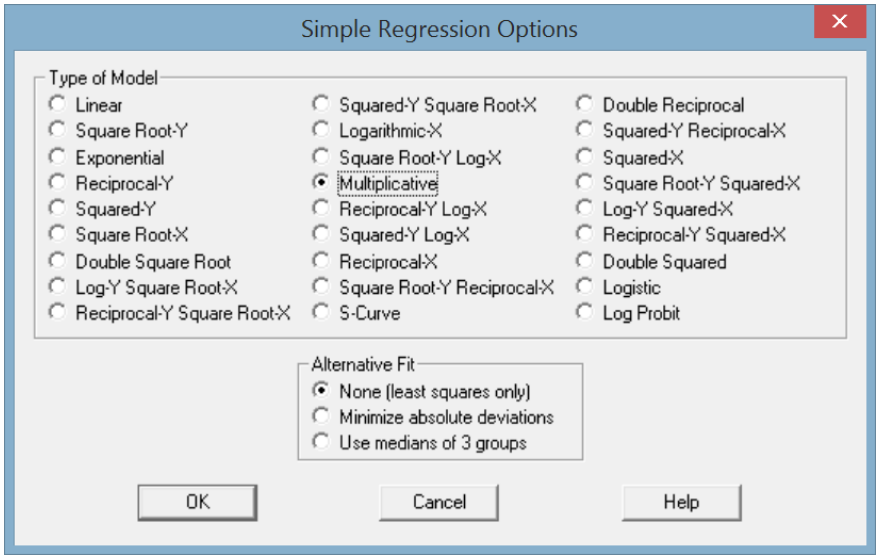
Kết quả cho ra một danh sách mô hình phi tuyến sắp xếp với R^2 cao nhất và thấp dần như sau

Comparison of Alternative Models

Model	*Correlation*	*R-Squared*
Multiplicative	0.9910	98.21%
Square root-Y	0.9801	96.05%
Logarithmic-Y square root-X	0.9760	95.27%
Square root-Y squared-X	0.9688	93.87%
Squared-X	0.9571	91.60%
Double square root	0.9560	91.38%
Exponential	0.9374	87.87%
S-curve model	-0.9259	85.73%
Double reciprocal	0.9057	82.02%
Square root-Y logarithmic-X	0.9033	81.60%
Linear	0.8913	79.45%
Logarithmic-Y squared-X	0.8341	69.57%
Square root-X	0.8294	68.79%
Double squared	0.7904	62.47%
Reciprocal-Y logarithmic-X	-0.7496	56.19%
Logarithmic-X	0.7462	55.69%
Square root-Y reciprocal-X	-0.7298	53.26%
Squared-Y	0.6630	43.96%
Squared-Y square root-X	0.5834	34.04%
Reciprocal-X	-0.5498	30.23%
Squared-Y logarithmic-X	0.4972	24.72%
Reciprocal-Y squared-X	-0.4133	17.08%
Squared-Y reciprocal-X	-0.3353	11.24%
Reciprocal-Y	<no fit>	
Reciprocal-Y square root-X	<no fit>	
Logistic	<no fit>	
Log probit	<no fit>	

Trong ví dụ này thì mô hình Multiplicative (Power – Mũ): AGB = a*DBH^b có R^2 cao nhất. Thiết lập mô hình theo dạng này.

Trong cửa sổ đồ thị, kích chuột phải và chọn Analysis Options để có bảng chọn mô hình tối ưu Multiplicative



Simple Regression - AGB_kg_tree vs. DBH_cm

Dependent variable: AGB_kg_tree

Independent variable: DBH_cm

Multiplicative model: Y = a*X^b

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	-2.2359	0.0972865	-22.9827	0.0000
Slope	2.47133	0.032121	76.9381	0.0000

NOTE: intercept = ln(a)

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	442.511	1	442.511	5919.46	0.0000
Residual	8.07356	108	0.0747552		
Total (Corr.)	450.584	109			

Correlation Coefficient = 0.991001

R-squared = 98.2082 percent

R-squared (adjusted for d.f.) = 98.1916 percent

Standard Error of Est. = 0.273414

Mean absolute error = 3.17096E6

Durbin-Watson statistic = 1.764 (P=62.4665)

Lag 1 residual autocorrelation = 56.1864

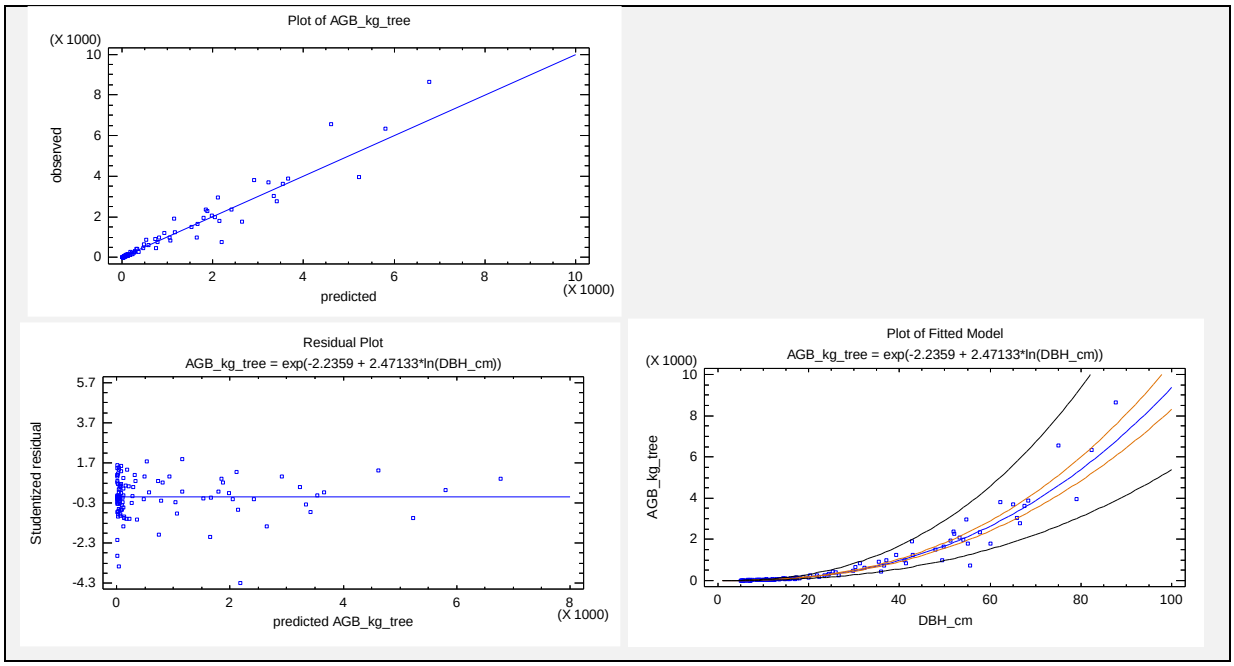
The StatAdvisor

The output shows the results of fitting a multiplicative model to describe the relationship between AGB_kg_tree and DBH_cm. The equation of the fitted model is

AGB_kg_tree = exp(-2.2359 + 2.47133*ln(DBH_cm))

or

ln(AGB_kg_tree) = -2.2359 + 2.47133*ln(DBH_cm)



Kết quả cho thấy mô hình phi tuyến mô tả tốt hơn tuyến tính với R^2 cao hơn và đồ thị quan sát và dự báo bám sát đường chéo, biến động phần dư phân bố khá đều quanh giá trị quan sát. Vì vậy mô hình này được lựa chọn.

6.2. Mô hình nhiều biến số

Trong thực tế biến phụ thuộc Y bị chi phối bởi nhiều biến số độc lập Xi. Ví dụ như trữ lượng rừng được đóng góp bởi nhiều nhân tố như mật độ, tiết diện ngang, chiều cao, cấp đất; năng suất cây trồng bị chi phối bởi các yếu tố phân bón, tưới nước, chăm sóc, …; sinh trưởng cây rừng phụ thuộc vào các yếu tố lập địa như loại đất, dinh dưỡng đất, lý tính đất, …

Tuy nhiên biến nào là chủ đạo thì chúng ta chưa biết, do vậy với phương pháp mô hình hóa với nhiều thử nghiệm khác nhau giúp chúng ta xác định được nhân tố ảnh hưởng quan trọng, trên cơ sở đó thiết lập mô hình dự báo theo các biến số ảnh hưởng.

Ví dụ nghiên cứu để xác định mô hình quan hệ tối ưu giữa sinh khối cây rừng (AGB) với 3 nhân tố đường kính (DBH), chiều cao (H) và diện tích tán lá (CA).

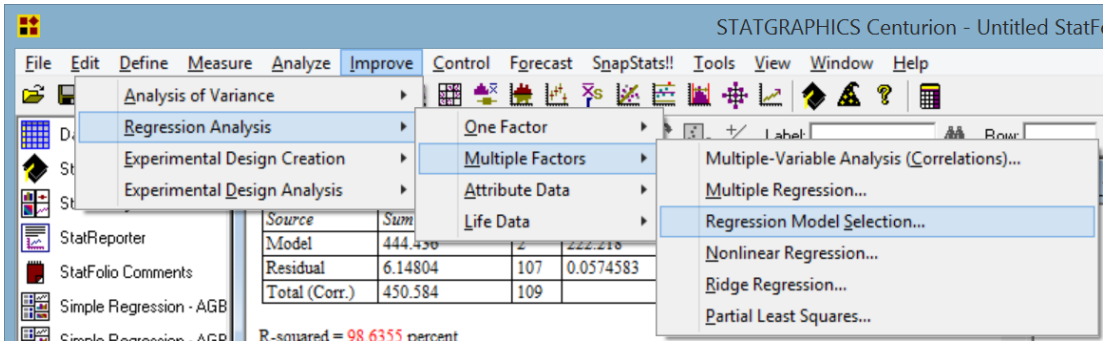
Thực hiện trong Stat như sau:

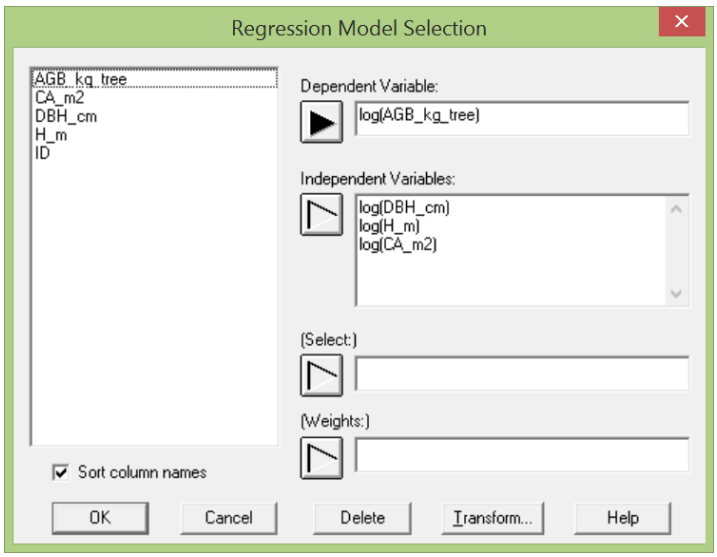
- ✓ Nhập dữ liệu đầu vào từ Excel sang Stat với các trường dữ liệu bao gồm biến phụ thuộc (AGB) và các biến độc lập (DBH, H, CA).

	ID	DBH_cm	CA_m2	H_m	AGB_kg_tree	Col_6	Col_7	Col_8	Col_9	Col_10
1	1	14.6	16.61902513749	12	67.6648920023576					
2	2	9.6	7.06858347057703	9.3	27.8868589281281					
3	3	12.1	9.0792027688745	11.5	47.8051988759639					
4	4	11.4	9.0792027688745	9.3	39.9128818615019					
5	5	13.6	13.854423602331	13.5	53.4249098217194					
6	6	6.5	7.06858347057703	6.2	14.6723394857834					
7	7	13.4	13.2025431267111	14	75.9996377584171					
8	8	9.3	5.7255526111674	11.6	21.8585100755737					
9	9	13.5	10.7521008569111	15.1	77.6830156380327					
10	10	12	13.2025431267111	13.9	57.1177193024927					
11	11	6.9	2.83528736986479	7.6	11.7745311547026					
12	12	11	6.15752160103599	11.5	33.9776045374154					
13	13	14.5	13.2025431267111	17.4	120.260913512052					
14	14	6.2	2.54469004940773	9.5	9.71040650556276					
15	15	10.6	9.62112750161874	12.9	54.262484712087					
16	16	5.6	1.32732289614169	12.2	7.73896626652478					
17	17	10.2	7.06858347057703	10.9	33.1638902926283					
18	18	7.2	5.30929158456675	10.4	13.4721687588082					
19	19	10.2	3.80132711084365	12.1	35.3793155476641					
20	20	8.5	13.854423602331	10.3	25.1566537752885					
21	21	11.7	0.785398163397448	14.3	44.1975973249347					
22	22	7	2.54469004940773	9.4	11.9905395363346					
23	23	9.1	6.15752160103599	7.2	19.8389993651076					
24	24	6.8	6.15752160103599	8	17.9850902582745					

Du lieu phan tich thong ke / B / C / D / E / F / G / H / I / J

✓ Lựa chọn biến số ảnh hưởng: Improve/Regression Analysis/Multiple Factors/Regression Model Selection. Trong hộp thoại chọn biến phụ thuộc và các biến độc lập thăm dò, thông thường hàm Power mô phỏng tốt quan hệ phi tuyến, do đó nên lấy log các biến số phụ thuộc và độc lập.





Kết quả thăm dò tìm biến độc lập ảnh hưởng cho thấy theo tiêu chuẩn bé nhất Cp và R^2 cao nhất thì cả 3 biến số DBH, H và CA tham gia vào mô hình là tốt nhất (Cp gần bằng số biến số là 4 (3 biến số + sai số của mô hình), đồng thời và R^2 cao nhất)

Regression Model Selection - log(AGB_kg_tree)

Dependent variable: log(AGB_kg_tree)

Independent variables:

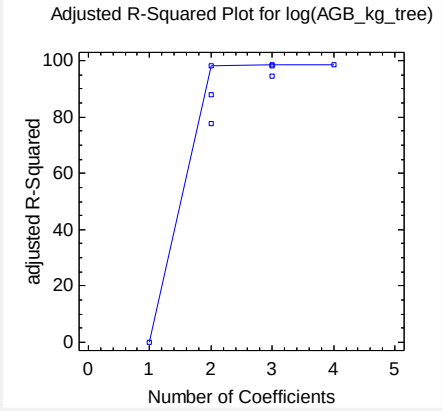
A=log(DBH_cm)

B=log(H_m)

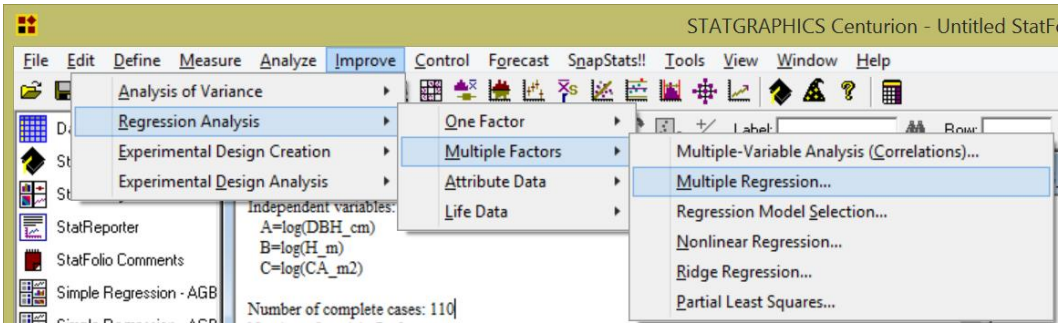
C=log(CA_m2)

Models with Smallest Cp

		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
0.0568979	98.6615	98.6236	4.0	ABC
0.0632183	98.4988	98.4707	14.8858	AB
0.0727907	98.2714	98.2391	32.8873	AC
0.0747552	98.2082	98.1916	35.8955	A
0.231951	94.4919	94.3889	332.197	BC
0.494177	88.1551	88.0455	832.015	B
0.923189	77.8722	77.6673	1646.34	C
4.1338	0.0	0.0	7811.17	

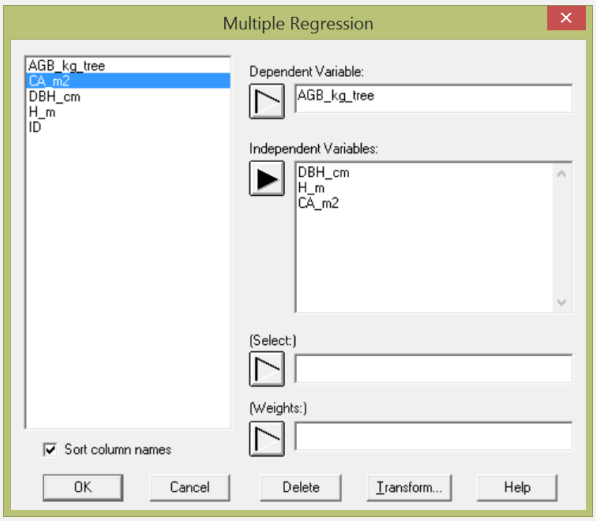


✓ Xây dựng mô hình đa biến số: Improve/Regression Analysis/Multiple Factors/Multiple Regression.



✓ Chọn mô hình (tuyến tính hay phi tuyến) và tổ hợp biến khác nhau trong hộp thoại. Mô hình được lựa chọn là mô hình có các chỉ tiêu tốt nhất về R^2 cao nhất, các tham số gắn biến số tồn tại ở mức P <0.05, MAE bé nhất, biến động residuals rải đều quanh giá trị dự báo trong phạm vi ±2. Sau đây là kết quả thử nghiệm các mô hình khác nhau.

Mô hình tuyến tính đa biến số:



Multiple Regression - AGB_kg_tree

Dependent variable: AGB_kg_tree

Independent variables:

DBH_cm

H_m

CA_m2

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-648.356	175.974	-3.68439	0.0004
DBH_cm	53.3999	9.5408	5.597	0.0000
H_m	-10.0609	18.9309	-0.531458	0.5962
CA_m2	10.5529	3.37177	3.12977	0.0023

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.95369E8	3	6.51231E7	156.36	0.0000
Residual	4.41488E7	106	416498.		
Total (Corr.)	2.39518E8	109			

R-squared = 81.5677 percent

R-squared (adjusted for d.f.) = 81.046 percent

Standard Error of Est. = 645.367

Mean absolute error = 383.513

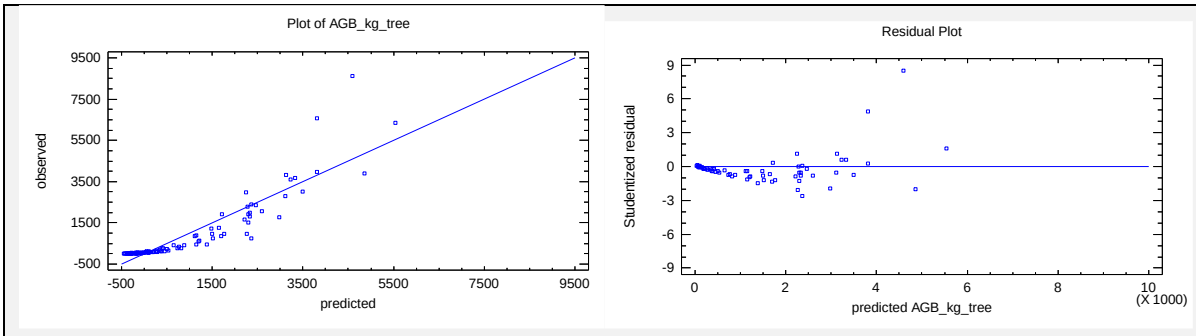
Durbin-Watson statistic = 1.32086 (P=0.0001)

Lag 1 residual autocorrelation = 0.2532

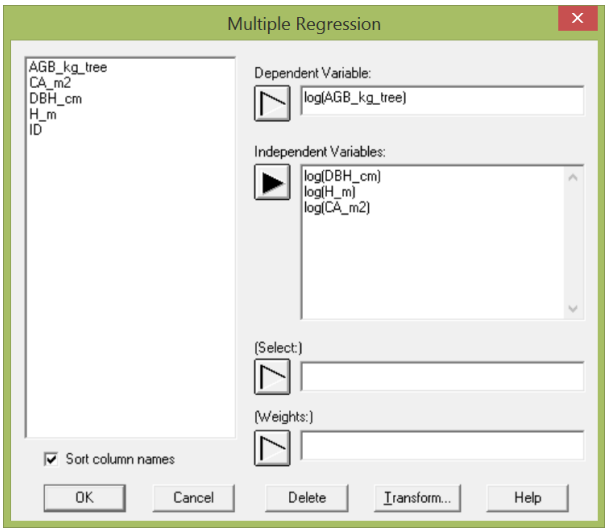
The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between AGB_kg_tree and 3 independent variables. The equation of the fitted model is

AGB_kg_tree = -648.356 + 53.3999*DBH_cm - 10.0609*H_m + 10.5529*CA_m2



Mô hình phi tuyến với đa biến số đơn



Multiple Regression - log(AGB_kg_tree)

Dependent variable: log(AGB_kg_tree)

Independent variables:

log(DBH_cm)

log(H_m)

log(CA_m2)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-2.85713	0.155287	-18.3991	0.0000
log(DBH_cm)	1.88169	0.103552	18.1713	0.0000
log(H_m)	0.696447	0.125314	5.55763	0.0000
log(CA_m2)	0.164251	0.0457565	3.58967	0.0005

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	444.553	3	148.184	2604.39	0.0000
Residual	6.03118	106	0.0568979		
Total (Corr.)	450.584	109			

R-squared = 98.6615 percent

R-squared (adjusted for d.f.) = 98.6236 percent

Standard Error of Est. = 0.238533

Mean absolute error = 0.174885

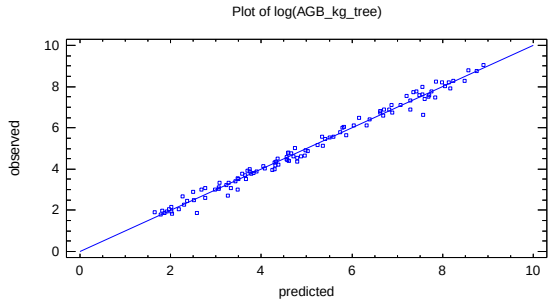
Durbin-Watson statistic = 1.94458 (P=0.3864)

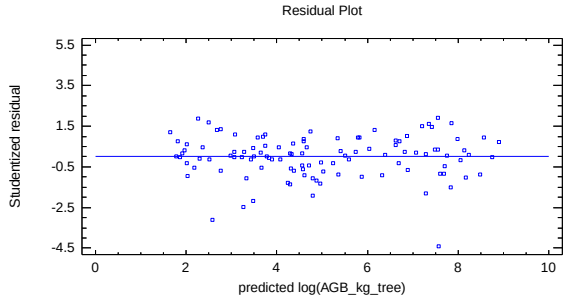
Lag 1 residual autocorrelation = 0.0214064

The StatAdvisor

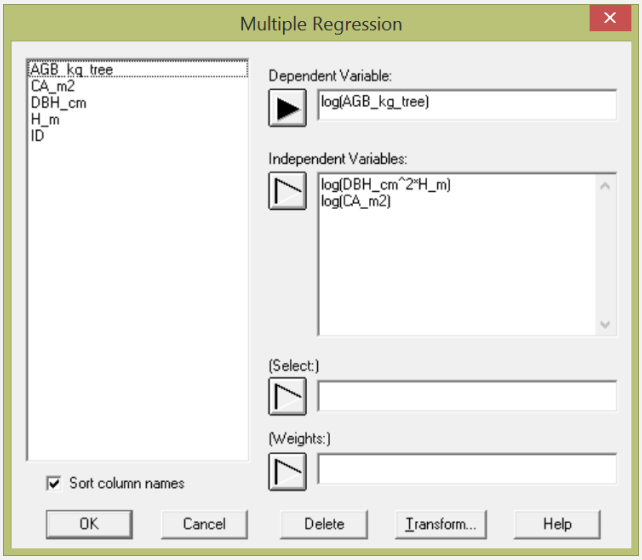
The output shows the results of fitting a multiple linear regression model to describe the relationship between log(AGB_kg_tree) and 3 independent variables. The equation of the fitted model is

log(AGB_kg_tree) = -2.85713 + 1.88169*log(DBH_cm) + 0.696447*log(H_m) + 0.164251*log(CA_m2)





Mô hình phi tuyến tổ hợp biến:



Multiple Regression - log(AGB_kg_tree)

Dependent variable: log(AGB_kg_tree)

Independent variables:

log(DBH_cm^2*H_m)

log(CA_m2)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-3.01731	0.108334	-27.8518	0.0000
log(DBH_cm^2*H_m)	0.873366	0.0216439	40.3515	0.0000
log(CA_m2)	0.190403	0.0421665	4.5155	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	444.436	2	222.218	3867.46	0.0000
Residual	6.14804	107	0.0574583		
Total (Corr.)	450.584	109			

R-squared = 98.6355 percent

R-squared (adjusted for d.f.) = 98.61 percent

Standard Error of Est. = 0.239705

Mean absolute error = 0.179352

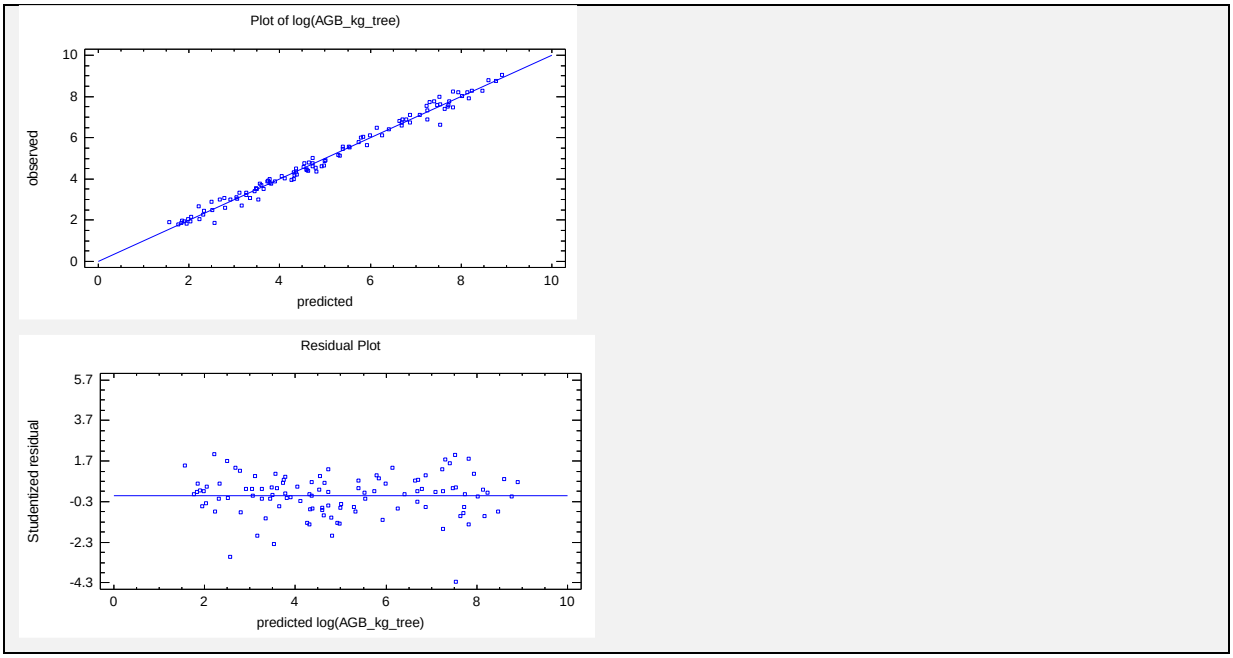
Durbin-Watson statistic = 1.88958 (P=0.2825)

Lag 1 residual autocorrelation = 0.0501669

The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between log(AGB_kg_tree) and 2 independent variables. The equation of the fitted model is

log(AGB_kg_tree) = -3.01731 + 0.873366*log(DBH_cm^2*H_m) + 0.190403*log(CA_m2)



Với kết quả thử nghiệm 3 loại mô hình trên cho thấy trong trường hợp này mô hình phi tuyến biến số đơn là tốt nhất với R^2 cao nhất, các tham số có P< 0.05, MAE bé nhất và biến động residuals rải đều quanh giá trị ước lượng.

Mô hình được lựa chọn là:

log(AGB_kg_tree) = -2.85713 + 1.88169*log(DBH_cm) + 0.696447*log(H_m) + 0.164251*log(CA_m2)

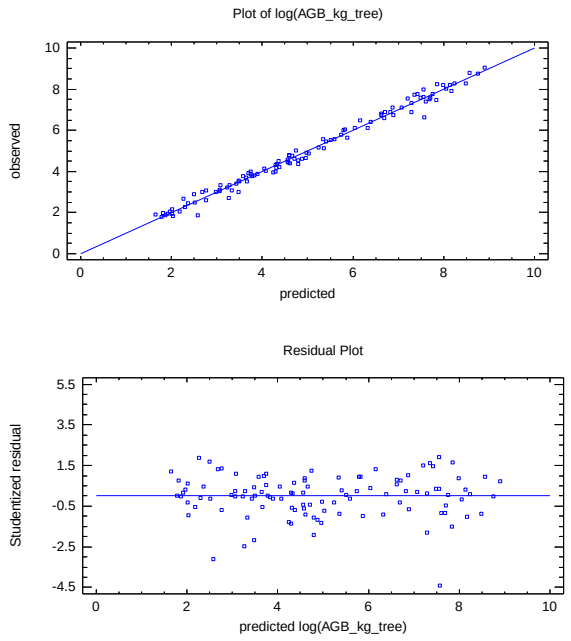
Với các chỉ tiêu thống kê:

R-squared (adjusted for d.f.) = 98.6236 %

Các tham số có P-value < 0.000

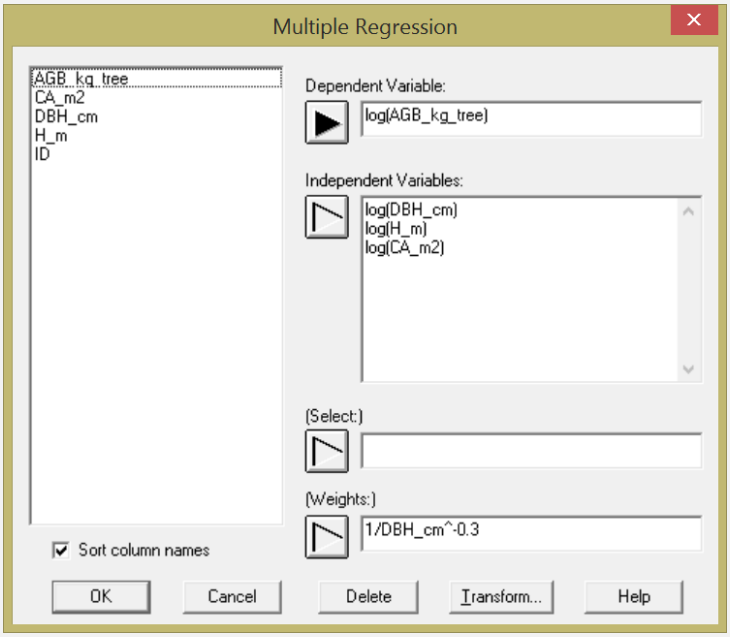
MAE = 0.174885

Biểu đồ biến động phần dư và biểu đồ quan hệ giữa quan sát với lý thuyết là tốt



Trong thực tế nghiên cứu lập mô hình, thường số liệu khó rải đều theo giá trị từ nhỏ đến lớn, ví dụ số liệu AGB theo cấp DBH thường tập trung ở cấp kính nhỏ. Vì vậy khi lập mô hình, sẽ có khả năng bị thiên lệch do số liệu tập trung ở một phạm vi nhất định. Để khắc phục điều này, trong lập mô hình đa biến, người ta sử dụng trọng số theo nhân tố độc lập chủ đạo.
Trọng số là một dạng hàm mũ: Weight = 1/X^c, trong đó X là biến số độc lập chủ đạo và c biến động từ -4 đến +4; thay đổi c ở bước nhảy khác nhau ví dụ là 0.1 để mô hình đạt tối ưu, trong đó lưu ý nhất chỉ tiêu biến động Residuals phân bố đều quanh trục ngang = 0 và trong phạm vi ±2.
Kết quả mô hình theo trọng số như sau: Sử dụng mô hình đã chọn trên là mô hình phi tuyến đa biến đơn, tiếp tục thử nghiệm trọng số để tìm mô hình tốt nhất

Mô hình có trọng số:



Multiple Regression - log(AGB_kg_tree)

Dependent variable: log(AGB_kg_tree)

Independent variables:

log(DBH_cm)

log(H_m)

log(CA_m2)

Weight variable: 1/DBH_cm^-0.3

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-2.87216	0.1687	-17.0253	0.0000
log(DBH_cm)	1.87475	0.107612	17.4214	0.0000
log(H_m)	0.701038	0.132705	5.28269	0.0000
log(CA_m2)	0.172687	0.0474493	3.6394	0.0004

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1119.6	3	373.201	2543.88	0.0000
Residual	15.5508	106	0.146706		
Total (Corr.)	1135.16	109			

R-squared = 98.6301 percent

R-squared (adjusted for d.f.) = 98.5913 percent

Standard Error of Est. = 0.383022

Mean absolute error = 0.178916

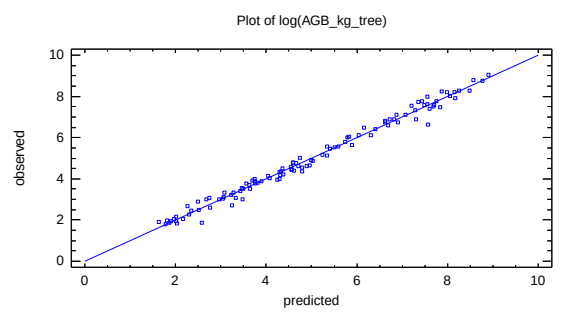
Durbin-Watson statistic = 1.94337 (P=0.3840)

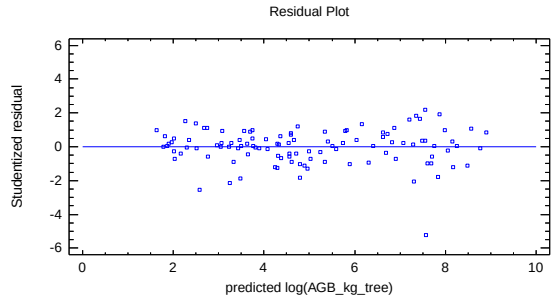
Lag 1 residual autocorrelation = 0.022304

The StatAdvisor

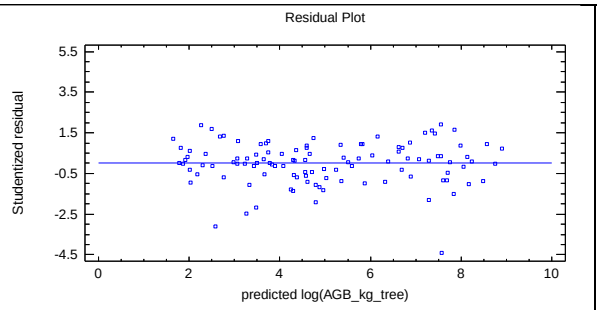
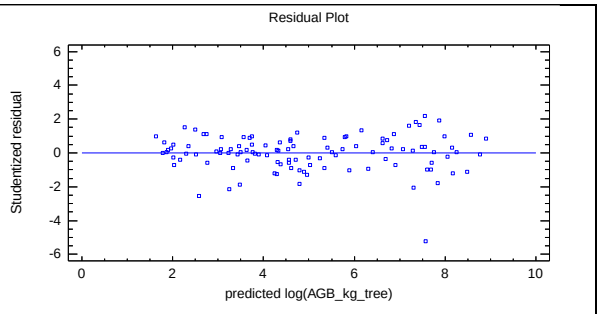
The output shows the results of fitting a multiple linear regression model to describe the relationship between log(AGB_kg_tree) and 3 independent variables. The equation of the fitted model is

log(AGB_kg_tree) = -2.87216 + 1.87475*log(DBH_cm) + 0.701038*log(H_m) + 0.172687*log(CA_m2)





Kết qủa mô hình có trọng số tuy có R^2 không cao hơn mô hình bình thường, tuy nhiên biến động Residuals được cải thiện rõ rệt, biến động quanh trục y = 0 và trong pham vi sai số ±2. Trong thiết lập mô hình hồi quy, hệ số xác định R^2 cao nhất chưa phải là mô hình tốt nhất, trong trường hợp này R^2 của mô hình có trọng số thấp hơn một ít, tuy nhiên biến động sai số được cân bằng và cải thiện tốt hơn. Do đó mô hình có trọng số được lựa chọn là tối ưu.

	
Mô hình không có trọng số	Mô hình có trọng số

Mô hình tối ưu có trọng số:

log(AGB_kg_tree) = -2.87216 + 1.87475*log(DBH_cm) + 0.701038*log(H_m) + 0.172687*log(CA_m2)

R-squared (adjusted for d.f.) = 98.5913%

Các tham số có P-value < 0.000
MAE = 0.178916
Biểu đồ biến động phần dư và biểu đồ quan hệ giữa quan sát với lý thuyết là tốt

7. PHÂN TÍCH PHÁT HIỆN CÁC NGUYÊN NHÂN ẢNH HƯỞNG ĐẾN VẤN ĐỀ

Trong thực thế chúng ta cần phát hiện các nhân tố chủ đạo ảnh hưởng đến một vấn đề, hậu quả. Ví dụ các nhân tố nào ảnh hưởng đến mức độ xung yếu của lưu vực, từ đây giúp cho việc quy hoạch lưu vực; hoặc tìm kiếm các nhân tố chủ đạo ảnh hưởng đến sinh trưởng sản lượng của một loài cây trồng, làm cơ sở quy hoạch, chọn vùng trồng thích hợp. Trong nghiên cứu liên quan đến xã hội thì cần xác định nhân tố ảnh hưởng đến quản lý tài nguyên thiên nhiên, nghèo đói ...
Mô hình hồi quy đa biến dạng tuyến tính hoặc phi tuyến hoặc tổ hợp biến sẽ là một công cụ mạnh giúp cho việc phát hiện các nhân tố ảnh hưởng rõ rệt cả về tự nhiên lẫn nhân tố xã hội.
Trong trường hợp nhiều biến số xi ảnh hưởng đến y không theo dạng tuyến tính mà có dạng quan hệ phi tuyến, trường hợp này cần đổi biến số để trở về dạng tuyến tính, hoặc lập mô hình tổ hợp biến. Trong Statgraphics, việc tính toán mô hình kiểu này rất đơn giản vì không cần tạo thêm các cột đổi biến số, biến số được đổi trực tiếp trong hộp thoại khi thiết lập mô hình.
Các bước tiến hành như sau:

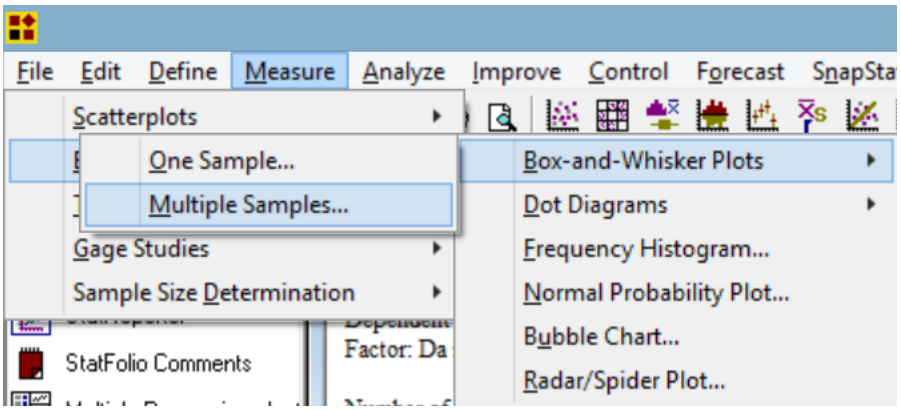
i) Thu thập dữ liệu về biến số phụ thuộc y và cùng với nó là các nhân tố xi dự kiến có ảnh hưởng (có thể định tính hay định lượng)
ii) Mã hóa các biến định tính
iii) Xác định biến số xi có ảnh hưởng đến y ở mức độ tin cậy 95% – Lập cây vấn đề nhân quả.
iv) Thử nghiệm các mô hình tuyến tính nhiều lớp hoặc được đổi biến số, khi cần thiết phải tổ hợp biến nếu các biến xi có quan hệ với nhau. Nên sử dụng trọng số Weight theo biến chủ đạo. Kiểm tra và lựa chọn mô hình tối ưu theo các tiêu chí thống kê: Hệ số xác định R^2 cao nhất với P <0.05; các tham số khác không với Pi <0.05, MAE bé nhất; và các đồ thị quan hệ giữa giá trị dự báo và thực tế và đồ thị giá trị phần dư Residuals nằm quanh trục y = 0 và biến động từ -2 và +2 ứng với giá trị dự báo trong´i độ tin cậy P = 95%.
v) Phân tích kết quả mô hình hồi quy đa biến để đánh giá chiều hướng tác động của các biến số đến biến phụ thuộc để đưa ra giải pháp.

Ví dụ: Xác định các nhân tố sinh thái ảnh hưởng đến sinh trưởng cây tếch được trồng làm giàu rừng khộp.
Bước 1: Thu thập số liệu: Bố trí thí nghiệm trên nhiều tổ hợp sinh thái khác nhau của rừng khộp. Cây tếch ở các 64 ô thí nghiệm, sau khi trồng trên 3 năm được thu thập số liệu sinh trưởng, tăng trưởng tếch và các nhân tố sinh thái trên có ô thử nghiệm như đá mẹ, loại đất, tầng dày đất, đá nỏi, kết von, độ tàn che, mật độ cây rừng, ngập nước, …. vị trí, địa hình, độ dốc, ..
Bước 2: Mã hóa biến định tính: Các nhân tố định tính như đá mẹ, loại đất, …. Cần được mã hóa để tạo thành biến số định lượng.
Có hai phương án mã hóa:

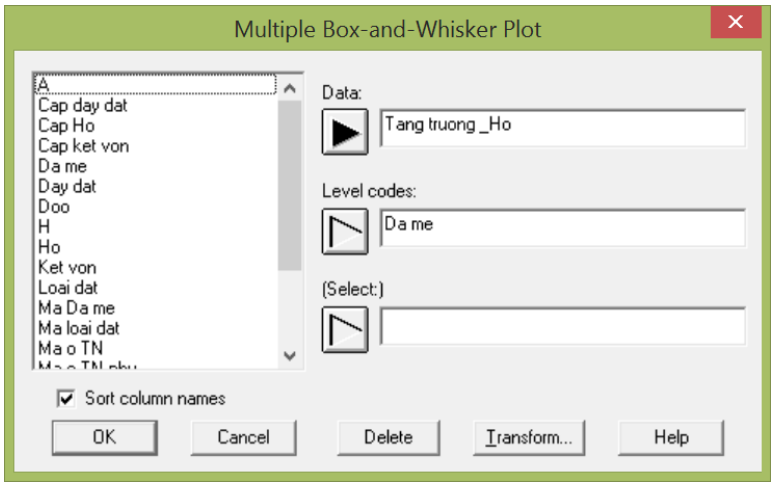
i. ***Mã hóa hệ thống:*** Các mức độ, cấp của của nhân tố được mã hóa hệ thống 1, 2, 3, ... Ví dụ mã hóa nhân tố vị trí địa hình: Bằng = 1; chân = 2; sườn = 3 và đỉnh = 4

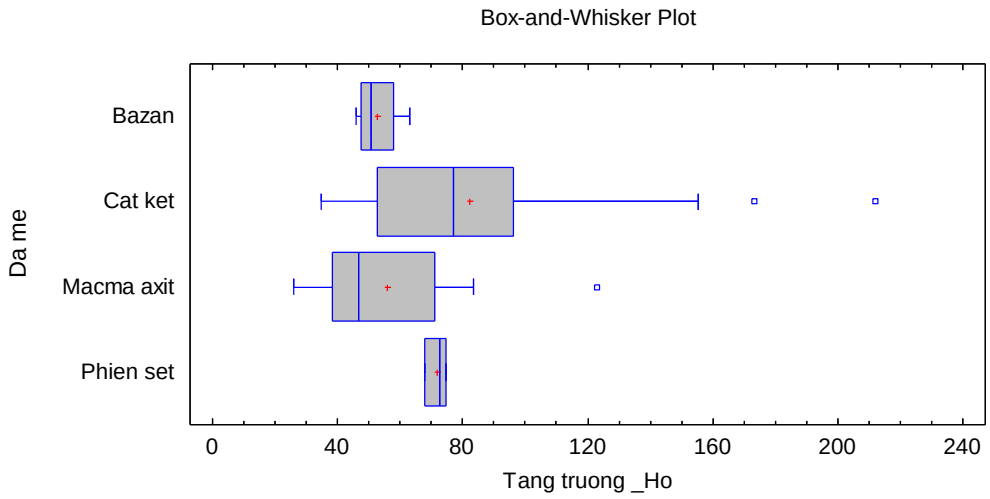
ii. ***Mã hóa theo chiều biến thiên:*** Các mức độ, cấp được mã hóa theo chiều biến thiên của nhân tố phụ thuộc. Sắp xếp nhân tố phụ thuộc theo một chiều nào đó (tăng hoặc giảm), sau đó các nhân tố được mã hóa theo cùng một vector như vậy.

Sử dụng chức năng vẽ biểu đồ biến động giá trị trung bình theo từng nhân tố trong Stat: Measure/Exploratory Plots/Box-and Whisker Plots/Multiple Samples:



Chọn biến dữ liệu quan sát và nhân tố khảo sát để mã hóa, ví dụ nhân tố là đá mẹ và biến số làn tang truong Ho.





Từ biểu đồ biến thiên dữ liệu quan sát theo sự thay đổi của nhân tố khảo sát, tiến hành mã hóa theo cùng chiều biến thiên với quan sát. Ví dụ trên, mã hóa các loại đá mẹ khác nhau theo chiều tăng của tăng trưởng tếch: Maxma axit = 1, Bazan = 2, Phien set = 3 và Cat ket = 4.

Cách thức mã hóa khác nhau sẽ dẫn đến việc lựa chọn mô hình hồi quy có mức độ phức tạp khác nhau

Hai phương án mã hóa biến định tính khác nhau sẽ dẫn đến việc chọn lựa mô hình hồi quy khác nhau

Kiểu dạng hàm mô phỏng	**Phương pháp mã hóa biến định tính**	
	Hệ thống **(Mã hóa đơn giản)**	Theo chiều biến thiên, vector của biến phụ thuộc **(Mã hóa phức tạp)**
Tuyến tính hoặc phi tuyến nhưng theo 1 chiều (tăng hoặc giảm) **(Xây dựng hàm đơn giản)**	Không thực hiện được hoặc sai quy luật	**Thực hiện được**
Phi tuyến dạng tăng giảm phức tạp, hoặc tổ hợp biến **(Xây dựng hàm phức tạp)**	**Thực hiện được**	Thực hiện được nhưng không cần thiết

Bước 3: Xác định các biến số xi có ảnh hưởng đến y – Cây vấn đề: Kết qủa phân tích này cũng chỉ ra được các biến số có quan hệ với nhau và ảnh hưởng đến y. Từ đây lập được cây vấn đề.

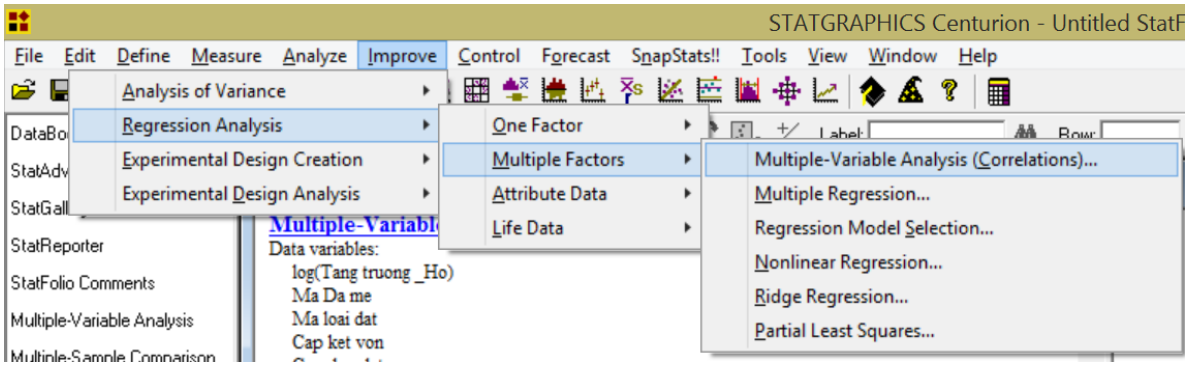
Nhập dữ liệu đã mã hóa trong Excel và chuyển vào Statgraphics.

<C:\1 - Tap huan Thong Ke Gia Lai\Du lieu phan tich thong ke.xls>

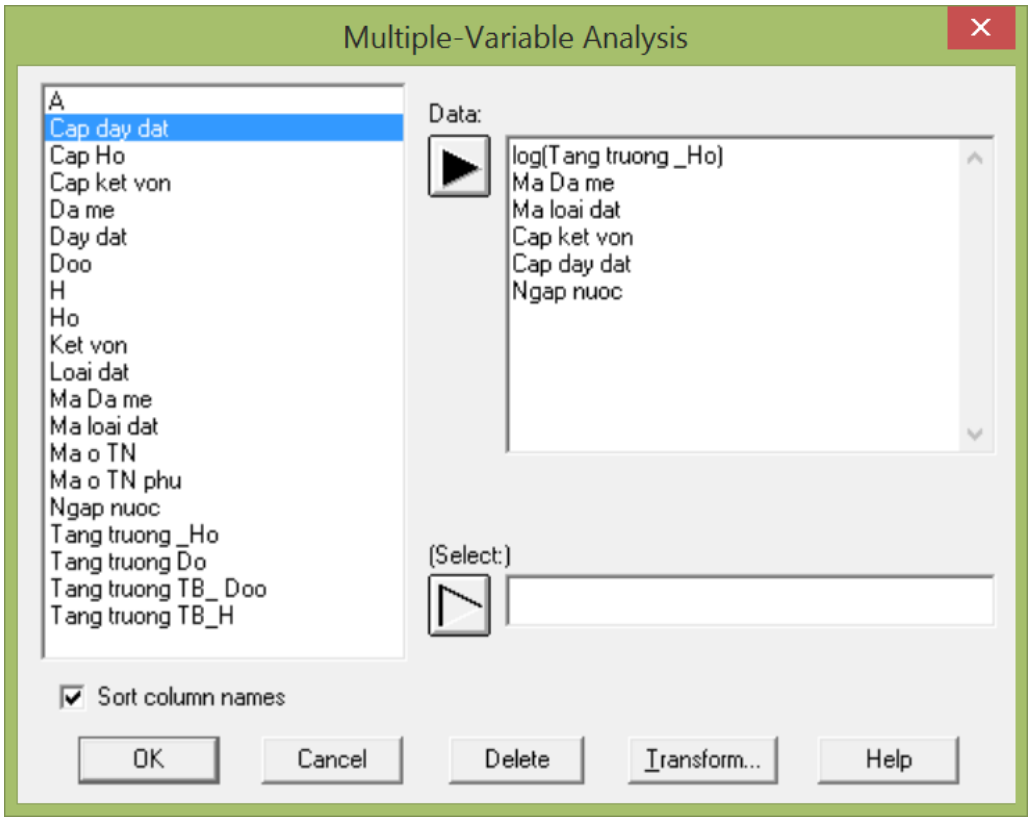
	Ma o TN	Ma o TN phu	Da me	Ma Da me	Loai dat	Ma loai dat	Da
1	BD1	BD1	Cat ket	3	Dat nau tang mong	3	40
2	BD2	BD2	Cat ket	3	Dat nau tang mong	3	40
3	BD3	BD3	Cat ket	3	Dat xam soi san nong	3	40
4	BD4	BD4.1	Cat ket	3	Dat xam soi san nong	3	40
5	BD4	BD4.2	Cat ket	3	Dat xam soi san nong	3	40
6	BD5	BD5	Cat ket	3	Dat nau tang mong	3	40
7	BD6	BD6.1	Cat ket	3	Dat nau tang mong	3	40
8	BD6	BD6.2	Cat ket	3	Dat nau tang mong	3	40
9	BN1	BN1.3	Cat ket	3	Dat den tang mong	3	60
10	BN1	BN1.2	Cat ket	3	Dat den tang mong	3	60
11	BN1	BN1.1	Cat ket	3	Dat den tang mong	3	60
12	BN1	BN1.4	Cat ket	3	Dat den tang mong	3	60
13	BN2	BN2.1	Cat ket	3	Dat den tang mong	3	40
14	BN2	BN2.2	Cat ket	3	Dat den tang mong	3	40
15	BN3	BN3.2	Cat ket	3	Dat xam tang rat mong	2	60
16	BN3	BN3.1	Cat ket	3	Dat den tang mong	3	20

Du lieu phan tich thong ke / B / C / D / E / F / G / H / I / J

Phân tích mối quan hệ giữa các biến số trong Stat: Improve/Regression Analysis/Mutiple Factors/Multiple-Variable Analysis



Trong hộp thoại đưa các biến y (tăng trưởng Ho) được lấy log để tạo ra biến liên tục và các biến sinh thái xi



Kết quả cho ra các chỉ tiêu thống kê của các biến y và xi; đồng thời trong bảng Correlations chỉ ra mức độ quan hệ giữa các biến, trong đó những biến có liên hệ với nhau được xác định với P-value < 0.05.

Summary Statistics

	log(Tang truong _Ho)	*Ma Da me*	*Ma loai dat*	*Cap ket von*	*Cap day dat*	*Ngap nuoc*
Count	64	64	64	64	64	64
Average	4.19609	2.39063	2.0625	1.35938	2.45313	1.82813
Standard deviation	0.434391	0.865882	0.774084	0.742522	0.73311	0.380254
Coeff. of variation	10.3523%	36.2199%	37.5314%	54.6223%	29.8847%	20.8002%
Minimum	3.25855	1.0	1.0	1.0	1.0	1.0
Maximum	5.35653	3.0	3.0	3.0	3.0	2.0
Range	2.09798	2.0	2.0	2.0	2.0	1.0
Stnd. skewness	1.18525	-2.81643	-0.35693	5.56615	-3.12134	-5.81832
Stnd. kurtosis	-0.122257	-1.81132	-2.13263	1.78137	-0.774089	1.97685

Correlations

	log(Tang truong _Ho)	Ma Da me	Ma loai dat	Cap ket von	Cap day dat	Ngap nuoc
log(Tang truong _Ho)		0.3785	0.5324	0.5051	-0.1016	-0.0545
		(64)	(64)	(64)	(64)	(64)
		0.0020	0.0000	0.0000	0.4245	0.6687
Ma Da me	0.3785		0.5787	0.2473	0.1168	-0.0821
	(64)		(64)	(64)	(64)	(64)
	0.0020		0.0000	0.0489	0.3579	0.5189
Ma loai dat	0.5324	0.5787		0.4850	-0.0507	0.0910
	(64)	(64)		(64)	(64)	(64)
	0.0000	0.0000		0.0000	0.6908	0.4745
Cap ket von	0.5051	0.2473	0.4850		-0.3331	0.2222
	(64)	(64)	(64)		(64)	(64)
	0.0000	0.0489	0.0000		0.0072	0.0776
Cap day dat	-0.1016	0.1168	-0.0507	-0.3331		-0.3425
	(64)	(64)	(64)	(64)		(64)
	0.4245	0.3579	0.6908	0.0072		0.0056
Ngap nuoc	-0.0545	-0.0821	0.0910	0.2222	-0.3425	
	(64)	(64)	(64)	(64)	(64)	
	0.6687	0.5189	0.4745	0.0776	0.0056	

Correlation

(Sample Size)

P-Value

The StatAdvisor

This table shows Pearson product moment correlations between each pair of variables. These correlation coefficients range between -1 and +1 and measure the strength of the linear relationship between the variables. Also shown in parentheses is the number of pairs of data values used to compute each coefficient. The third number in each location of the table is a P-value which tests the statistical significance of the estimated correlations. P-values below 0.05 indicate statistically significant non-zero correlations at the 95.0% confidence level. The following pairs of variables have P-values below 0.05:

log(Tang truong _Ho) and Ma Da me

log(Tang truong _Ho) and Ma loai dat

log(Tang truong _Ho) and Cap ket von

Ma Da me and Ma loai dat

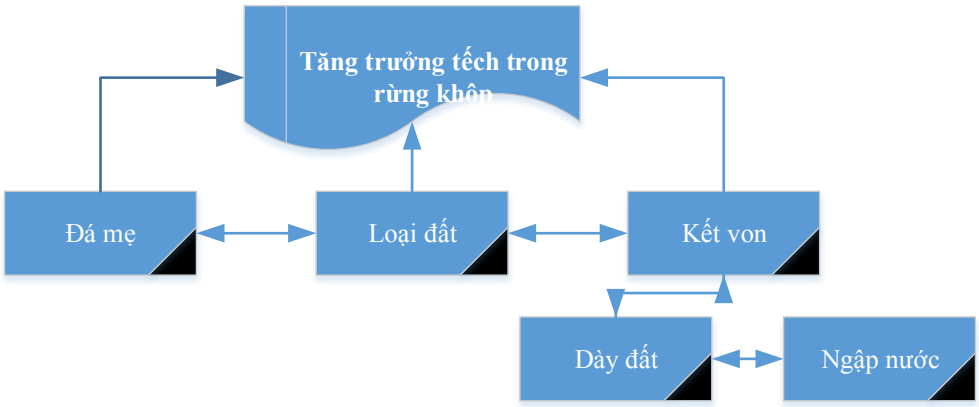
Ma Da me and Cap ket von

Ma loai dat and Cap ket von

Cap ket von and Cap day dat

Cap day dat and Ngap nuoc

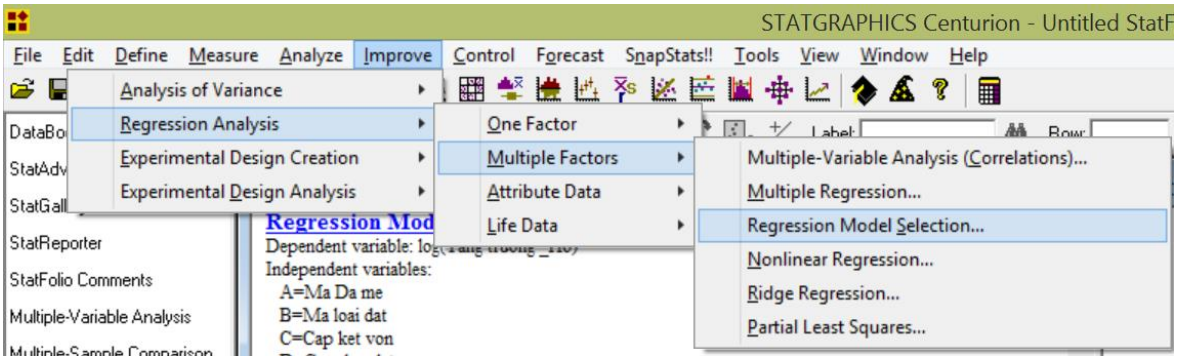
Kết quả trên cho thấy tăng trưởng Ho của tếch trong rừng khộp chịu ảnh hưởng trực tiếp của 3 nhân tố: Đá mẹ, loại đất và kết von; bị tác động gián tiếp bởi 3 nhân tố độ dày đất và mức độ ngập nước. Từ đây có thể vẽ ra cây nguyên nhân chi phối đến tăng trưởng tếch trong rừng khộp như sau:



Cây nhân tố ảnh hưởng đến tăng trưởng tếch làm giàu rừng khộp ở Đăk Lăk

Như vậy có 5 nhân tố xi ảnh hưởng trực tiếp hay gián tiếp đến tăng trưởng cây tếch ở các điều kiện lập địa khác nhau của rừng khộp. Tuy nhiên để tập trung lựa chọn nhân tố ảnh hưởng chính trong mô hình, tiến hành phân tích chọn biến trong Stat.

Sử dụng chức năng chọn biến số của Stat: Improve/Regression Analysis/Mutiple Factors/Rgression Model Selection:



Kết quả cho thấy có 3 biến số BCE (Ma loai dat, Cap ket von và Ngap nuoc) cho R^2 cao nhất và Cp tiến gần đến số biến số nhất. Vì vậy để đơn giản trong mô hình hồi quy, chỉ thiết lập với 3 biến số chủ đạo này.

Regression Model Selection - log(Tang truong _Ho)

Dependent variable: log(Tang truong _Ho)

Independent variables:

A=Ma Da me

B=Ma loai dat

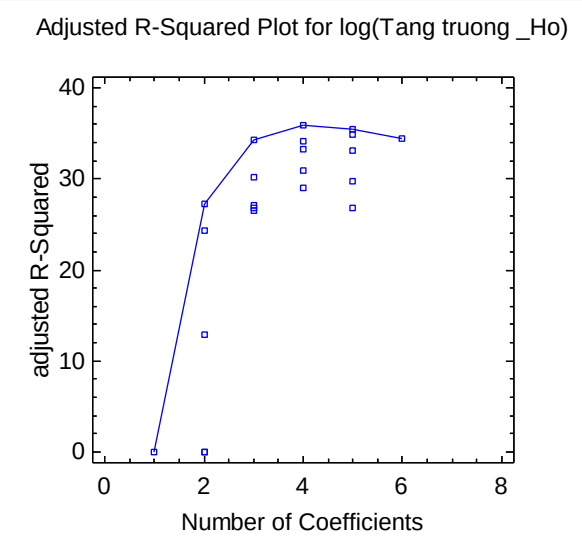
C=Cap ket von

D=Cap day dat

E=Ngap nuoc

Models with Smallest Cp

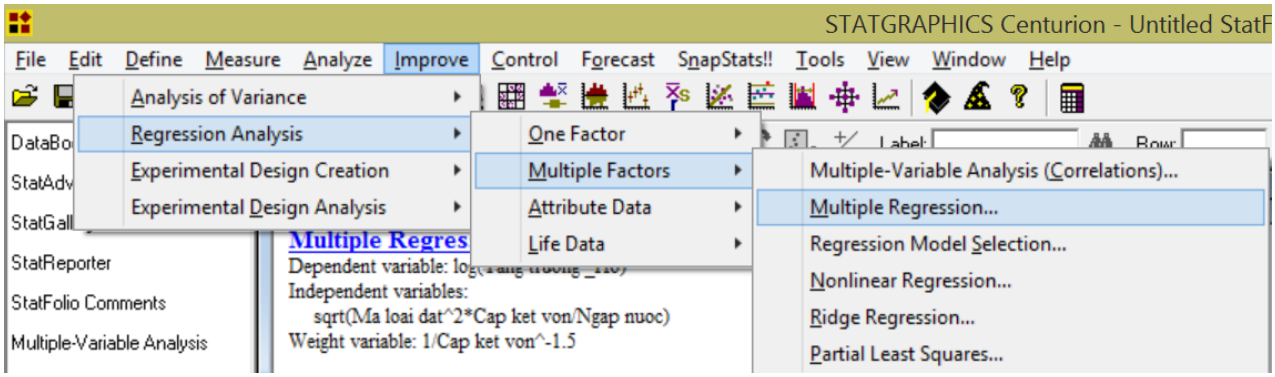
		Adjusted		*Included*
MSE	*R-Squared*	*R-Squared*	*Cp*	*Variables*
0.120811	39.0245	35.9757	2.63299	BCE
0.124112	36.3145	34.2264	3.23893	BC
0.121747	39.5763	35.4797	4.10244	ABCE
0.124213	37.3075	34.1728	4.28409	ABC
0.122759	39.0742	34.9437	4.58517	BCDE
0.126039	36.3861	33.2054	5.17005	BCD
0.123628	39.6828	34.483	6.0	ABCDE
0.1263	37.3166	33.0668	6.27534	ABCD
0.131813	32.3631	30.1455	7.03849	AC
0.130425	34.1725	30.8811	7.29862	ACE
0.137393	28.3441	27.1884	8.9031	B
0.133971	32.3826	29.0018	9.0197	ACD
0.132527	34.2261	29.7669	9.24703	ACDE
0.137561	29.4134	27.0991	9.87488	BE
0.138193	29.0891	26.7641	10.1867	AB
0.138558	28.9017	26.5706	10.3669	BD
0.142821	25.513	24.3116	11.6254	C
0.137974	31.5226	26.8801	11.8467	ABDE
0.164271	14.3259	12.9441	22.3827	A
0.188696	0.0	0.0	34.1583	
0.188696	1.5873	0.0	35.1664	D
0.188696	1.5873	0.0	35.8724	E

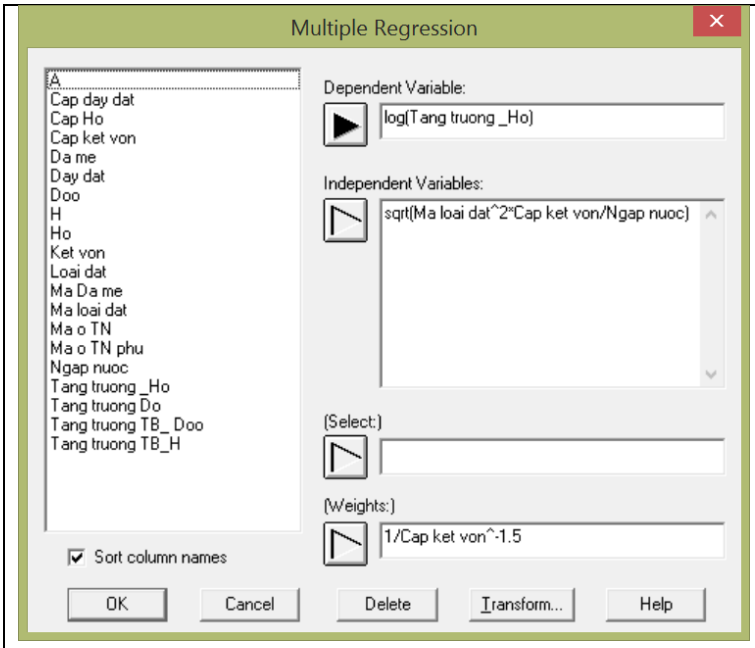


Bước 4: Thử nghiệm và lựa chọn mô hình tối ưu: Trên cơ sở xác đinh được 5 biến ảnh hưởng, trong đó có 3 biến độc lập ảnh hưởng chủ đạo là Ma loai dat, Cap ket von va Ngap nươc, tiến hành lập mô hình hồi quy có trọng số theo biến Cap ket von và lựa chọn mô hình tối ưu theo các chỉ tiêu thống kê. Kết quả như sau

Trong Stat thực hiện phân tích hồi quy đa biến: Improve/Regression Analysis/Mutiple Factors/Multiple Regression. Trong hộp thoại chọn biến số y và xi, đổi biến, tổ hợp biến và xác định trọng số Weight thích hợp. Trong đó Weight = 1/Cap ket von ^b, với b ± 4, tìm b để cho mô hình có các chỉ số thống kê tốt nhất và đồ thị biến động residuals phân bố quanh trục y = 0. Kết quả sau là một ví dụ với tổ hợp biến và giá trị trọng số tối ưu tìm được với b = -1.5.

Lưu ý khi tổ hợp biến cần quan tâm đến chiều hướng quan hệ của từng biến độc lập với biến phụ thuộc, nếu 2 biến độc lập có quan hệ + (cùng chiều với y) thì khi tổ hợp có thể tích lại với nhau; ngược lại 1 biến thuận và 1 biến nghịch với y thì chia nhau. Mối quan hệ thuận nghịch của y/xi được thể hiện trong bảng phân tích tương quan từng cặp biến số Correlations.





Multiple Regression - log(Tang truong _Ho)

Dependent variable: log(Tang truong _Ho)

Independent variables:

sqrt(Ma loai dat^2*Cap ket von/Ngap nuoc)

Weight variable: 1/Cap ket von^-1.5

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	3.64326	0.104536	34.8519	0.0000
sqrt(Ma loai dat^2*Cap ket von/Ngap nuoc)	0.298794	0.0384934	7.7622	0.0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	11.924	1	11.924	60.25	0.0000
Residual	12.27	62	0.197903		
Total (Corr.)	24.194	63			

R-squared = 49.285 percent

R-squared (adjusted for d.f.) = 48.467 percent

Standard Error of Est. = 0.444863

Mean absolute error = 0.266893

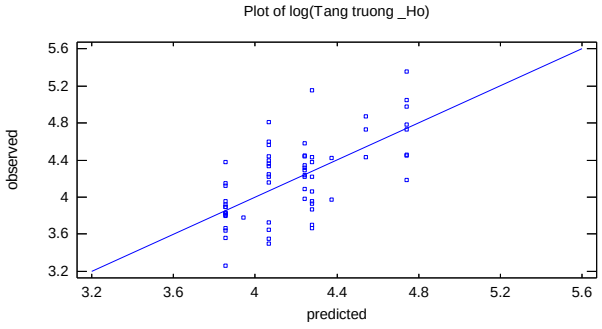
Durbin-Watson statistic = 1.69195 (P=0.0864)

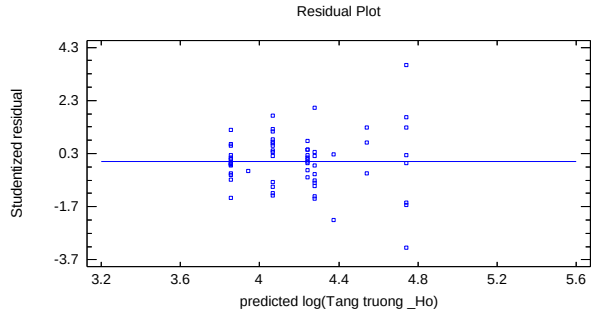
Lag 1 residual autocorrelation = 0.147319

The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between log(Tang truong _Ho) and 1 independent variables. The equation of the fitted model is

log(Tang truong _Ho) = 3.64326 + 0.298794*sqrt(Ma loai dat^2*Cap ket von/Ngap nuoc)





Bước 5: Phân tích kết quả ứng dụng của mô hình: Mô hình được thiết lập thể hiện quan hệ giữa 3 nhân tố loại đất, kết von và ngập nước đến tăng trưởng cây tếch trong rừng khộp. Thế các giá trị khác nhau của 3 biến này vào mô hình sẽ dự báo được tăng trưởng chiều cao tếch theo từng tổ hợp lập địa. Đây là cơ sở để: i) Dự báo năng suất theo tổ hợp 3 nhân tố ảnh hưởng; ii) Lập bản đồ thích nghi cho cây tếch trong rừng khộp theo 3 lớp dữ liệu của 3 nhân tố ảnh hưởng.