Tab 7:

One Variable Regression

Tab 7: One Variable Regression

PURPOSE:

Introduce Regression Analysis as an empirical model-building technique to model processes that have a continuous "Y" response.

OBJECTIVES:

- To determine when regression should be used and why.
- Understand the use of regression to model the relationship of one continuous "X" variable to a continuous response "Y".
- Apply regression in Minitab to fit a line to the data points so that the equation of the line can be used to predict "Y", given "X".
- To recognize the mathematical means of determining if the model is the best model for the data.
- Interpret and understand the graphical means of determining if the model is the "best fit" model for the data.

Regression Analysis

- Regression analysis is used to describe the relationship between a response variable and one or more predictors

- Minitab has several regression functions including linear, multiple, logistic and stepwise.

- ***Linear regression*** fits a simple model between the response variable and one predictor.
- ***Multiple regression*** fits a model between the response variable and two or more predictors.
- ***Logistic regression*** fits a model with discrete data.
- ***Stepwise regression*** is used to fit the best model from a pool of predictors.

Regression... a means of finding a relationship between the "Y" and "X"

What is it?

A mathematical means of describing a relationship between the "Y" and the "X"s - creating a "model" of the process.

Y = a + bx + error Where: a is the Y intercept
b is the slope of the line

Why use it?

- To find the potential Vital Few "X"s
- To predict / forecast the "Y"
- To optimize the "Y"
- To determine where to set the "X"s to optimize "Y"

When to use it?

- To screen passive data (historical or baseline data) for potential vital "X"s

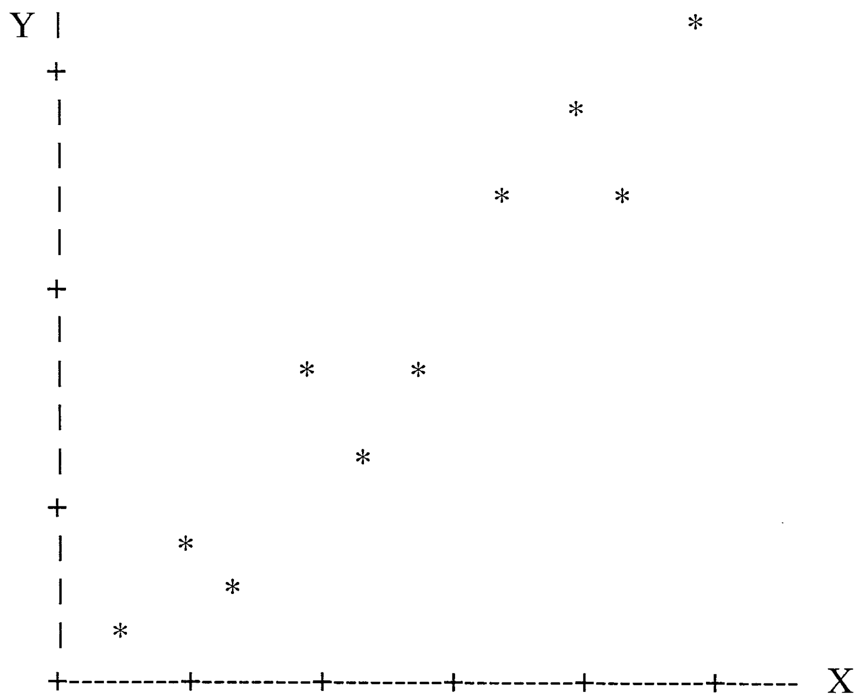
★ **Danger! Do not draw final conclusions using passive data. Follow up with a DOE (Design of Experiment)...**

- To analyze the results of a DOE (Design of Experiment)

Regression is a powerful tool that must be used carefully

One Variable Regression

We may be interested in the relationship between an independent variable (X) and a response variable (Y). A scatter plot of the relationship might be:



Suppose that the true relationship is: $\mathbf{Y_i = a + b * X_i + e_i}$

- linear relationship exists
- "a" (the constant) and "b" (the coefficient) will be fixed, but unknown, parameters
- "X"s are the independent variables
- "Y"s are the observed responses
- "e"s are errors. Usual assumptions on the errors are:
  - average is 0.0
  - uncorrelated
  - normal distribution
  - standard deviation of errors is the same for all levels of the "X" variable

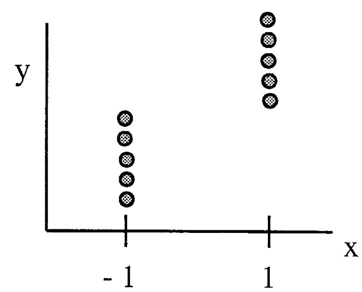
Some questions we might ask about the fitted equation include:

- What is the best way to **collect data** to estimate the equation?
- What are the **estimated** values of “a” and “b”?
- Is this the right **functional form** (a line)?
- Is the relationship **statistically significant** (not attributable to chance)?
- How large are the **errors**, “e”$_i$?

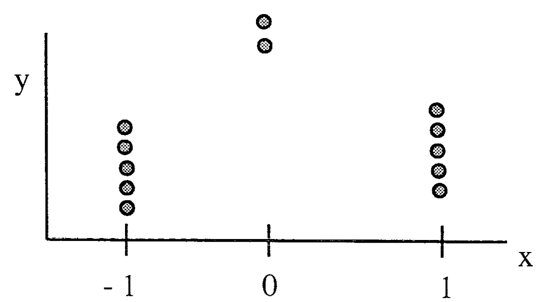
Collecting data

To give the **smallest variation** in the estimate of the slope, place one-half of the observations at the lowest limit of "X" and the other half at the highest limit, and use a wide range of the independent variable.

This is appropriate when the data are highly variable, the range for the independent variable is small, and the relationship is expected to be linear.



To determine the form of the relationship **(Is it a line? Or is it a curve?)**, use more than 2 levels of the independent variable. If the data are highly variable, then 3 levels are often used.



It is better to collect data in a **random order**, rather than starting with an "X" at the low value and then increasing - another variable may be changing over time that could affect the process.

One Variable Regression with Minitab

Example:

You are trying to optimize the performance of an paint cure oven. One theory says that blower fan velocity affects evaporation of solvent in the paint. You are trying to prove that such a relationship exists by analyzing the data below.

Restart Minitab (Don't save anything!), and
Enter the following data into C1 and C2:

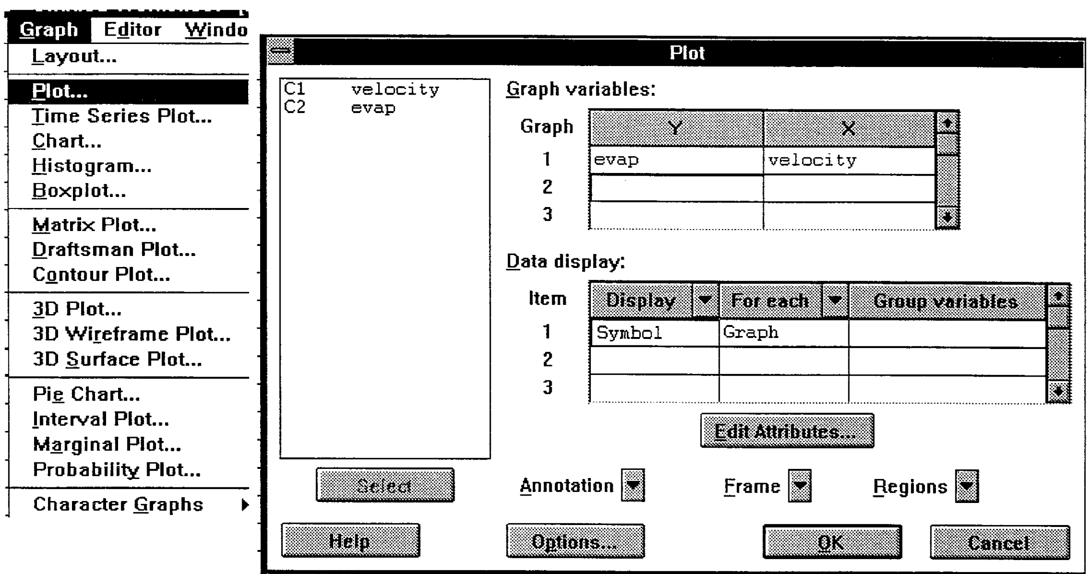
MINITAB - Unti

File Edit Manip Calc Stat Graph

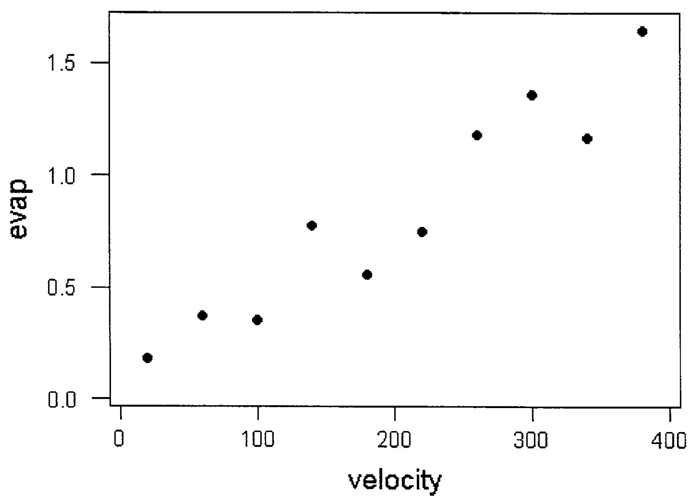
	C1	C2	C3
↓	velocity	evap	
1	20	0.18	
2	60	0.37	
3	100	0.35	
4	140	0.78	
5	180	0.56	
6	220	0.75	
7	260	1.18	
8	300	1.36	
9	340	1.17	
10	380	1.65	

1) Always Graph the Data First

Graph>Plot



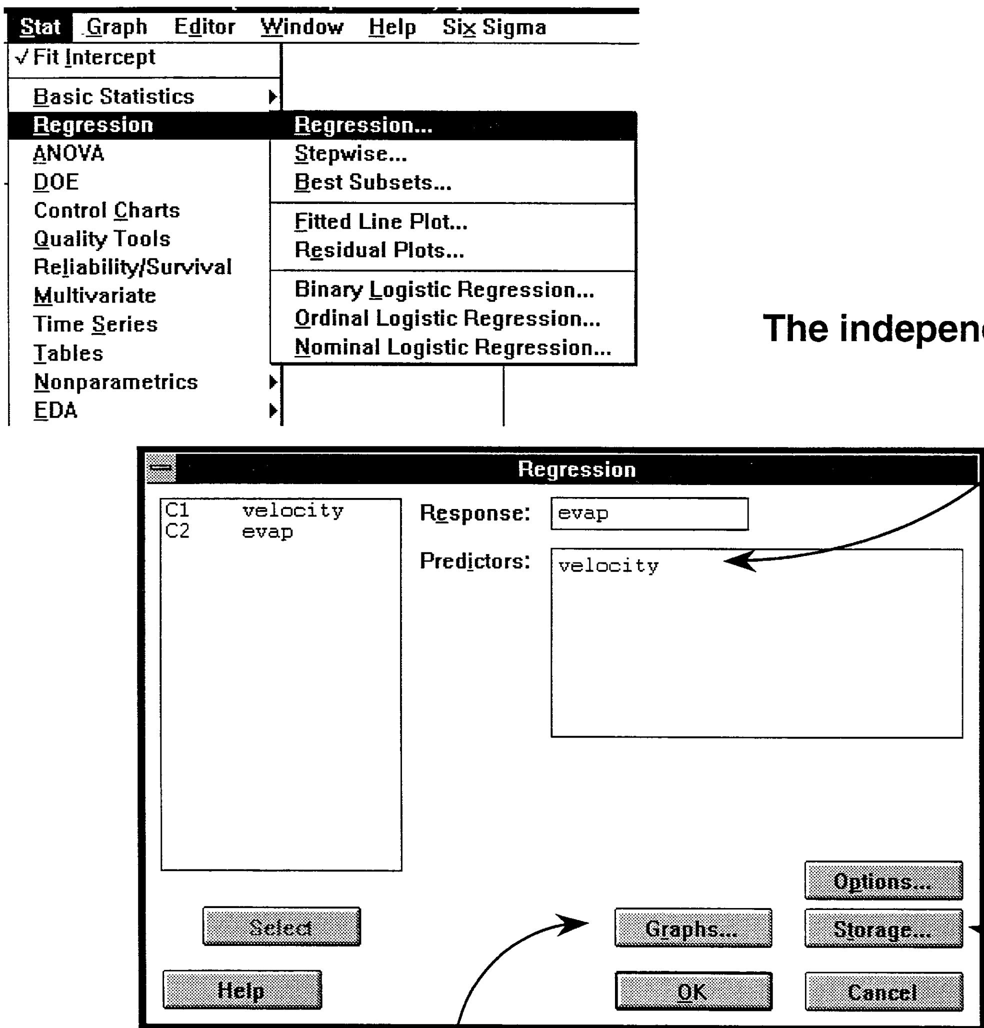
Click "OK" to run



Looks linear!!!

2) Run a Regression Analysis on the Data

Stat>Regression>Regression...



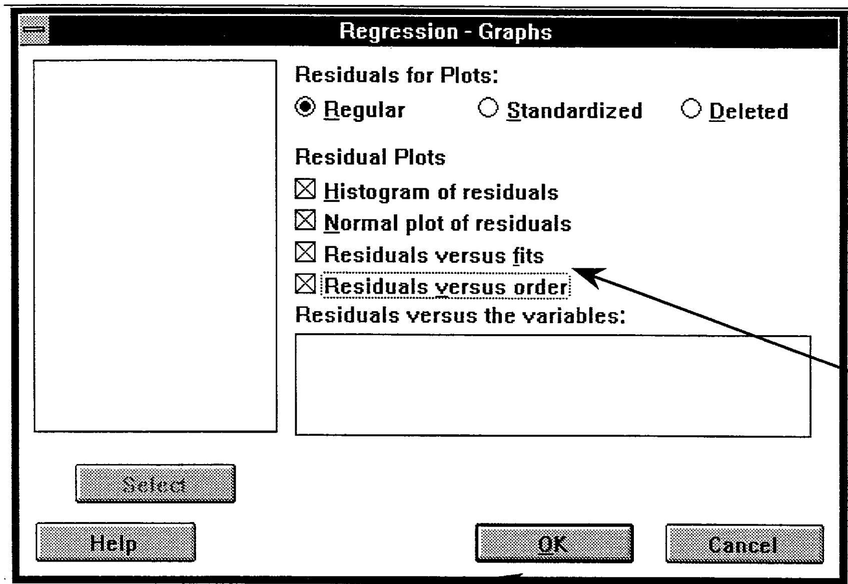
The independent variable

Click 'Graphs'

AND

Click 'Storage'

(see next page for the subdialog boxes)

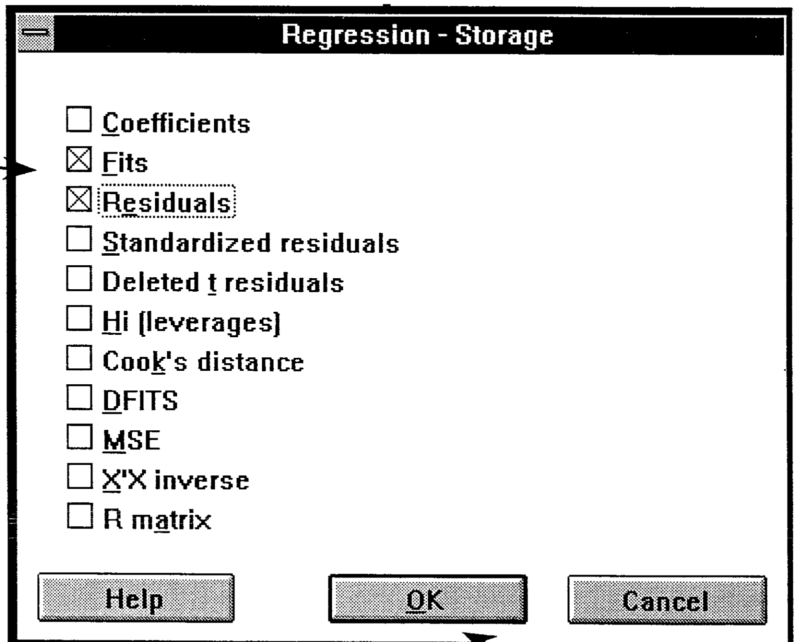


This dialog box is used to generate Residual (error) Plots - use these graphs to check the assumptions on the errors in your model

Click in the boxes to indicate which graphs you want to see.

Click 'OK', then click the 'Storage' button in the main dialog box

Click 'Fits' and 'Residuals' to store this information in the Data Window



Click 'OK' twice

The Data Window will have two new columns... "FITS1" and RESI1"

Type 'Ctrl-d' to return to the Data Window

MINITAB - Untitled Worksheet - [Data]

File Edit Manip Calc Stat Graph Editor Window Hel

	C1	C2	C3	C4	C5
↓	velocity	evap	FITS1	RESI1	
1	20	0.18	0.14582	0.034182	
2	60	0.37	0.29897	0.071030	
3	100	0.35	0.45212	-0.102121	
4	140	0.78	0.60527	0.174727	
5	180	0.56	0.75842	-0.198424	
6	220	0.75	0.91158	-0.161576	
7	260	1.18	1.06473	0.115273	
8	300	1.36	1.21788	0.142121	
9	340	1.17	1.37103	-0.201030	
10	380	1.65	1.52418	0.125818	
11					

FITS are the <u>predicted</u> values of "Y" calculated from the regression equation for each value of "X":

$C_3 = 0.069 + 0.00383\,C_1$ (this is the Regression equation found in the Session Window)

<u>or</u>

Predicted Response = 0.069 + 0.00383 (Velocity)

RESIDUALS are errors. The presence of residuals demonstrates that the model does not represent the data without mistakes. (Actual Response minus Predicted Response (Fits) for each point). Thus:

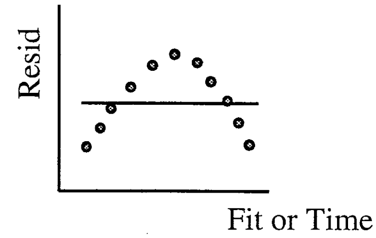
$$C_4 = C_2 - C_3$$

Residual Plots - A diagnostic tool to check the "goodness" of the regression model

- The average of the Residuals should always be 0.0
- The Residuals should be normally distributed
- The Residuals should be randomly distributed. A pattern in the Residuals may indicate that this model form is incorrect.

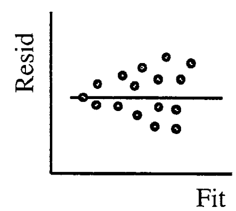
Examples of patterns are:

- curve (start low, increase, then decrease)

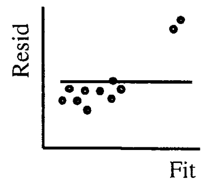


- trend over time of data collection

- unequal variation (usually larger variation for higher values)



- one or two extreme values



Some ways of improving poor fits:

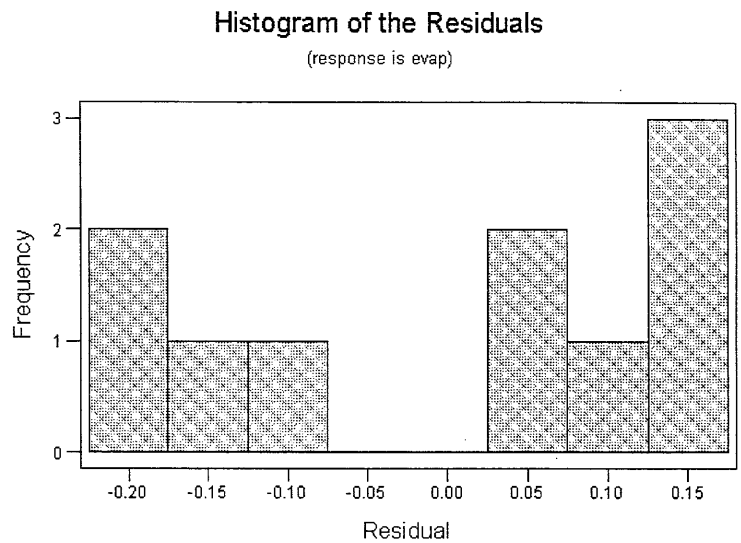
- Investigate interesting data. It may be incorrect, or it may be the most important information in your study.
- Fit a different equation (it may not be a linear relationship)
- Transform Y (log, square root, reciprocal, y^k . . .)
- Transform 'X' variables (log, square root, reciprocal . . .)

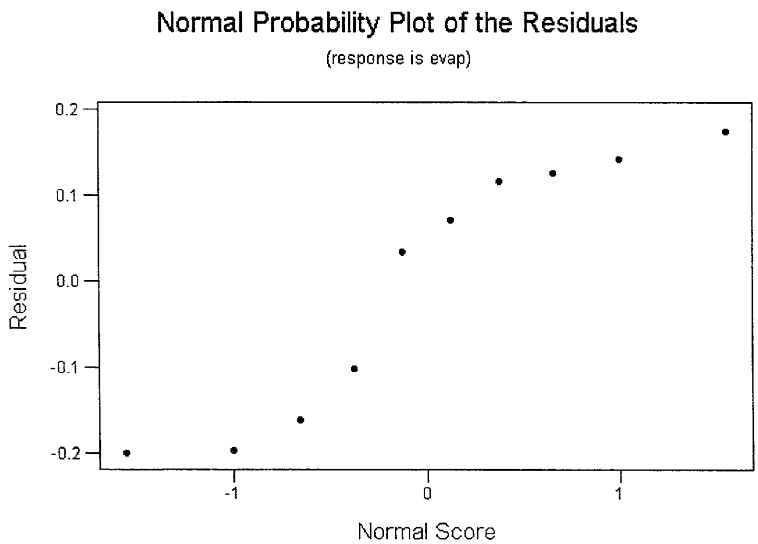
Check the Residuals:

Use “Crtl-Tab” to scroll through the windows until you find the Residual graphs

<u>Residuals should be normally distributed:</u>

It does not <u>look</u> like a bell curve...



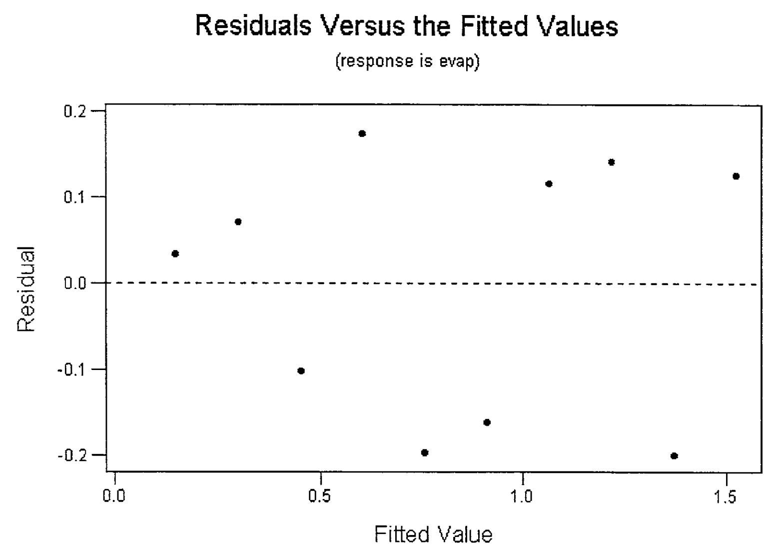


No p-value, but it looks like it could be non-normal. Check it using “Normality Test”

Stat>Basic Statistic> Normality Test
Variable: Resi1
p=0.092

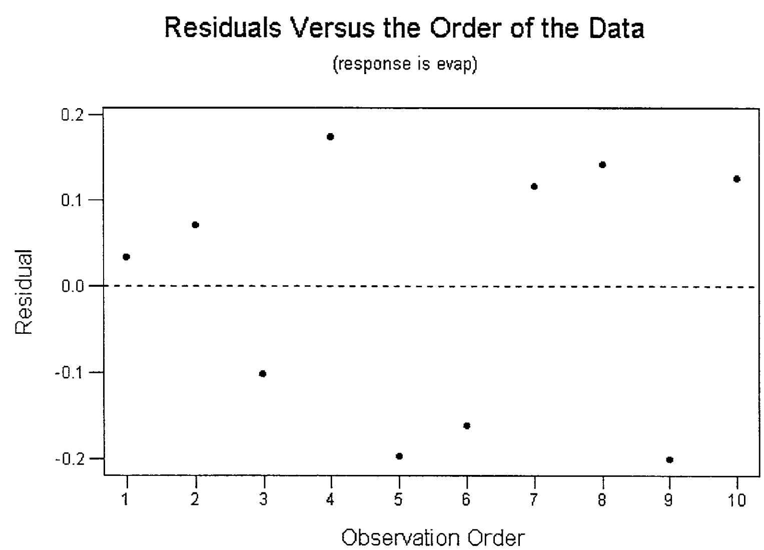
The assumption of normality must be checked

Residuals should be randomly distributed with an average of 0.0



Errors should be randomly distributed above and below the average value of 0.

These errors appear to be fairly randomly distributed.



This graph is interpreted the same as the one above, except the X-axis provides a picture of errors over time.

A good model will result in a random pattern of Residuals over time.

If a pattern is noticeable, the linear, one-variable model may not be the best fit for the data, or there could be more Vital "X"s

The Session Window contains the analysis results...
("Ctrl-M" to move to the Session window)

Regression Analysis

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef        StDev          T        P
Constant       0.0692       0.1010       0.69    0.512
velocity    0.0038288    0.0004378       8.75    0.000

S = 0.1591       R-Sq = 90.5%     R-Sq(adj) = 89.3%

Analysis of Variance

Source        DF         SS         MS         F        P
Regression     1     1.9351     1.9351     76.49    0.000
Error          8     0.2024     0.0253
Total          9     2.1375
```

Fail to reject H_o

Accept H_a

p-value of the <u>Constant</u>

H_o: The line passes through the origin (0,0)...
(0 velocity = 0 evaporation)

H_a: The line does <u>not</u> pass through the origin (0,0)...
(0 velocity $\neq$ 0 evaporation)

p-value of the "X" variable - <u>Velocity</u>

H_o: Slope = 0

H_a: Slope $\neq$ 0

or, another way of saying it:

H_o: The "X" is not significant

H_a: The "X" <u>is</u> significant

See Appendix for further descriptions of the Session Window output

Regression Analysis

The higher the R^2, the better fit of the model to the process

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor       Coef        StDev          T        P
Constant      0.0692       0.1010       0.69    0.512
velocity   0.0038288    0.0004378       8.75    0.000

S = 0.1591      R-Sq = 90.5%      R-Sq(adj) = 89.3%

Analysis of Variance

Source       DF        SS        MS        F        P
Regression    1    1.9351    1.9351    76.49    0.000
Error         8    0.2024    0.0253
Total         9    2.1375
```

For a good model, this number should be close to the same value as R^2

The smaller this value (the spread of the errors) is, the better the model

s: The standard deviation of the residuals (errors). Errors are observed values - expected values. In other words, the distance from the observed points to the fitted line described by the regression equation. (Should be small, for a good model)

$$s = MS_{(error)}^{1/2}$$

R-Sq: The percent of total variation "explained" by the fitted line. The variation accounted for by the "X"s. (Should be large, for a good model)

$$R\text{-}Sq = \frac{SS_{regression}}{SS_{total}} * 100$$

R-Sq(adj): Adjustment for an overfit condition (fitting too many variables into the equation) that incorporates the number of terms in the model compared to the number of observations.

$$R\text{-}Sq(adj) = 1 - \frac{SS_{regression} / (n-p)}{SS_{total} / (n-1)}$$

where n = number of observations

p = total number of terms in the model

See Appendix for more definitions

Regression Analysis

The Regression terms (SS and MS) should be large relative to the Error terms (SS and MS)

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef       StDev          T        P
Constant       0.0692      0.1010       0.69    0.512
velocity    0.0038288   0.0004378       8.75    0.000

S = 0.1591      R-Sq = 90.5%      R-Sq(adj) = 89.3%

Analysis of Variance

Source        DF        SS        MS        F        P
Regression     1    1.9351    1.9351    76.49    0.000
Error          8    0.2024    0.0253
Total          9    2.1375
```

The Error term should be small relative to the Total

The p-value should be < 0.05 to demonstrate statistical significance (an equation with a "good" fit)

$SS_{regression}$: The **explained** variation in the "Y" response due to the presence of the "Xs" in the model. The sum of the squared difference between the predicted value for each run and the overall average response.

SS_{error}: The **unexplained** variation in the "Y"; the quality minimized by the regression line. The sum of the squared difference between each data point and the predicted value for that data point.

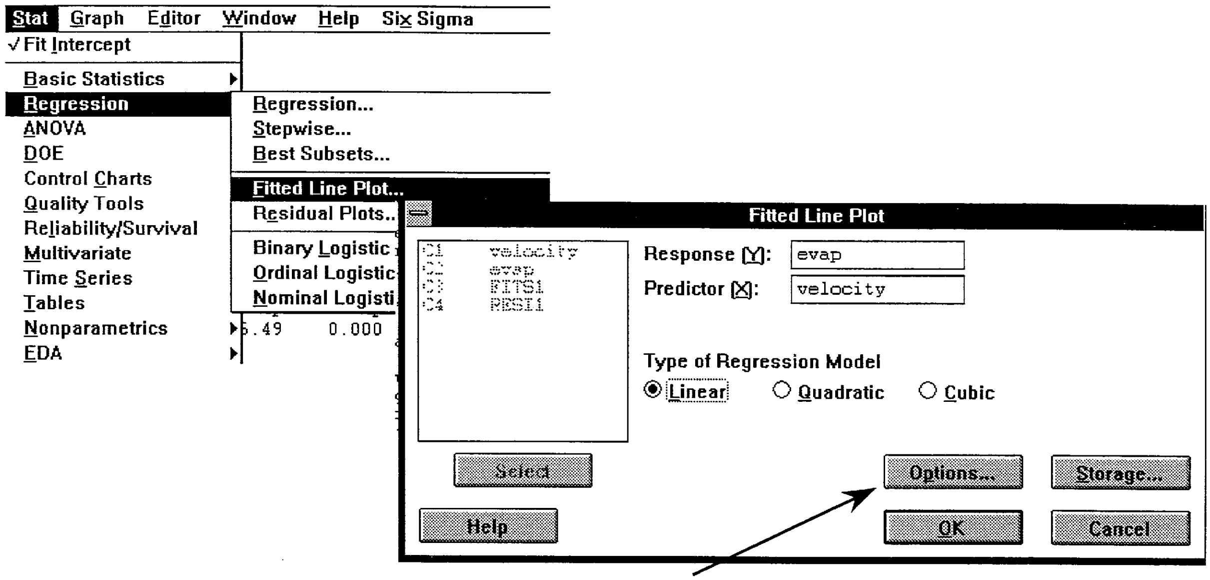
SS_{total}: The total variation of the "Y" around the average value.

Evaluate the model by looking at R-Sq, R-Sq(adj), s, and the p-values

See Appendix for more definitions

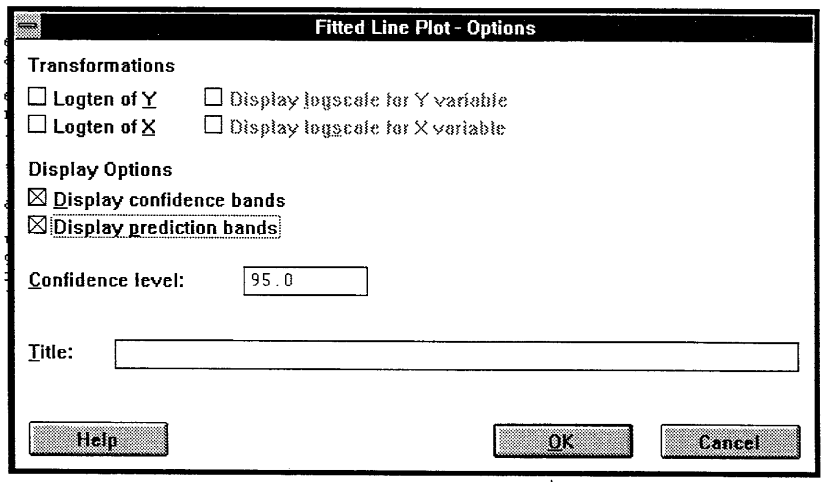
Regression Analysis can also be Graphical!

Stat>Regression>Fitted Line Plot



Click on "Options"

Click these Options to display more information in the graphical output



Click "OK" twice

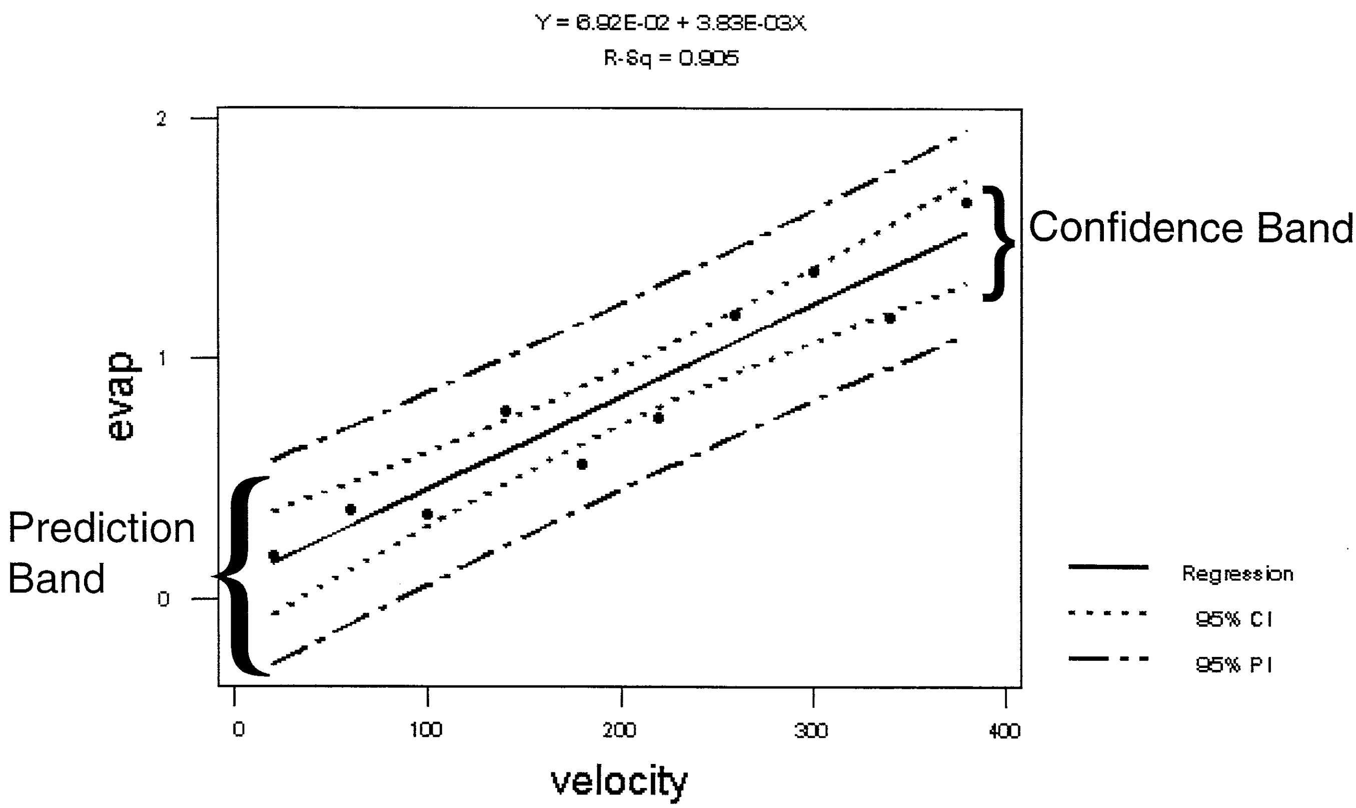
"Fitted Line Plot" provides:

- Regression Analysis in the Session Window
- A plot showing the Least Squares fit for the line *
- A plot showing Confidence Intervals (C.I.) and Prediction Intervals (P.I.)

*** See Appendix for Least Squares Method**

Confidence Intervals and Prediction Intervals

Regression Plot



C.I. = Confidence Interval (95% confidence that the <u>means</u> of all data will fall within this band)

P.I. = Prediction Interval (95% confidence that the <u>individual</u> data points will fall within this band)

Information in the Session Window is the same, as we generated earlier...

Regression Analysis

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef       StDev          T        P
Constant       0.0692      0.1010       0.69    0.512
velocity    0.0038288   0.0004378       8.75    0.000

S = 0.1591      R-Sq = 90.5%      R-Sq(adj) = 89.3%

Analysis of Variance

Source        DF        SS        MS        F        P
Regression     1    1.9351    1.9351    76.49    0.000
Error          8    0.2024    0.0253
Total          9    2.1375
```

Fail to reject H_o:

Accept H_a:

Conclusions:

- We have found a potential Vital "X" - Velocity (because $p < 0.05$) (Follow up with a DOE)
- The linear model appears to be a good fit, since no patterns were found in the residuals
- We could not prove Residuals were non-normal, which is consistent with our assumptions (p-value = .092, from Normality Test)
- The model should be acceptable for our purpose: predicting evaporation rates given a velocity (based upon: the small error term, $R^2 = 90.5\%$, p-value < 0.05)
- If the process is critical, more data should be taken. A regression model might then be developed with errors that are distributed closer to normal, and a higher R^2 value.

Key Concepts - Tab 7

- Regression can be used on passive data **with caution,** since it is not a controlled experiment.
- Always follow up with a DOE when drawing conclusions on passive data using regression.
- Regression is usually run on DOE results.
- Always graph the "Y" vs. "X" data before running regression - you need to see what the right model should be first.
- Look at p-values, s, R^2, R^2_{adj}, SS and MS to evaluate the model mathematically.
- Look at Residuals vs. Fits plots to focus on potential issues with your model. Use the Residuals graphs to diagnose "goodness of fit" graphically.
- Use Fitted Line Plot to create a graph of the regression line through the data and define both Confidence Intervals and Prediction Intervals for the model.

Class Exercise:

You believe that the amount of space our appliances occupy on the show room floor impacts the sales volume. You have gathered data on both sales volume and total floor space used over the last 12 months for a number of retail locations. Now you want to analyze the data to see if the amount of space does have a relationship to the annual sales volume.

Input the data below into Minitab

	($K)	(square feet)	
	C1	C2	C3
↓	Annual Sales	Floor Space	Location
1	280.0	180	1
2	217.0	120	2
3	221.5	60	3
4	295.0	180	4
5	285.0	120	5
6	173.0	60	6
7	336.0	180	7
8	212.5	60	8
9	206.0	60	9
10	290.5	120	10
11	312.5	180	11
12	263.5	120	12

Have fun applying what you have learned about one variable regression. Be prepared to explain your answer and the work that supports your conclusion.

Appendix

Regression Terminology

r:	The correlation coefficient (r) for multiple regression. The closer to +/ - 1, the better the fit of the model. '0' indicates no linear relationship.
R-Sq:	The correlation coefficient squared (R^2). A value of R^2 closer to 100% indicates that there is a possible relationship, and more variation is explained.
R-Sq (Adj):	Adjustment of R^2 for an overfit condition. (Takes into account the number of terms in the model).
Standard Error of the Estimate (s)	Expected deviation of data about the predictive "surface" $s = MS_{error}{}^{1/2}$
Mean Square of Regression ($MS_{regress}$)	"Between" estimate of variance for the overall model $MS_{regression} = SS_{regression} / DF_{regression}$ (DF = Degrees of Freedom)
Mean Square of the Residual (Error) (MS_{error})	"Within" estimate of variance. Best estimate of population variance. $MS_{error} = SS_{error} / DF_{regression}$
F-Ratio:	"F" statistic. A higher value indicates the model can detect a relationship between the factors and the response. $F = MS_{regression} / MS_{error}$
p-value:	Probability of an error if difference is claimed. p-value < 0.05 indicates a difference (significant) p-value >0.05 indicates that no conclusion of difference (significance) can be drawn. Probability that the model is not a "good" model. "Good" indicates that a relationship between factors and response has been found.

Regression Terminology (cont'd)

α and β are usually used to represent the population values.
"a" and "b" are estimates of the population values derived from the data.

Choose "a" and "b" to minimize the sum of the squared errors **"Least Squares":**

Minimize: $\Sigma (e_i^2) = \Sigma (Y_i - a - bX_i)^2$

Take partial derivatives with respect to "a" and "b," and set the derivatives equal to 0.0.

The least squares line passes through $(\bar{X}, \bar{Y})$: $(Y_i - \bar{Y}) = b (X_i - \bar{X})$

$$\text{Slope is } b = \frac{\Sigma X_i Y_i - n \bar{X} \bar{Y}}{\Sigma X_i^2 - n \bar{X}^2} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

Calculating a Confidence Interval for the Coefficient (Slope)

(Refer to the example on page 7.11)

The regression equation from the Session Window is:

Evap = 0.069 + 0.00383 velocity

Estimate of the Slope

Regression Analysis

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef       StDev          T        P
Constant       0.0692      0.1010       0.69    0.512
velocity    0.0038288   0.0004378       8.75    0.000

S = 0.1591      R-Sq = 90.5%      R-Sq(adj) = 89.3%
```

0.00383 is the estimate, based upon the data, of the slope of the line. Since it's an estimate, we know that the actual value really falls within a range of plausible values - a Confidence Interval. The **Confidence Interval** for the slope can be calculated from the following equation:

Estimated value +/- ($t_{df, \alpha}$)(std. error of the estimate)

- The standard error of the slope estimate is found in the StDev column: 0.00044 (rounded up)
- The t-value is the tabled t-statistic using the degrees of freedom in the Error term of the model (8) and a confidence level of 0.025 (two-tailed test): t = 2.31

The 95% Confidence Interval for the Slope is:

0.00383 +/- 2.31(0.00044) ⟶ **(0.00281, 0.00485)**

Calculating a Confidence Interval for the Coefficient (Slope)

(Refer to the example on page 7.11)

The regression equation from the Session Window is:

Evap = 0.069 + 0.00383 velocity

Estimate of the Slope

Regression Analysis

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef       StDev          T        P
Constant       0.0692      0.1010       0.69    0.512
velocity    0.0038288   0.0004378       8.75    0.000

S = 0.1591       R-Sq = 90.5%      R-Sq(adj) = 89.3%
```

0.00383 is the estimate, based upon the data, of the slope of the line. Since it's an estimate, we know that the actual value really falls within a range of plausible values - a Confidence Interval. The **Confidence Interval** for the slope can be calculated from the following equation:

Estimated value +/- ($t_{df, \alpha}$)(std. error of the estimate)

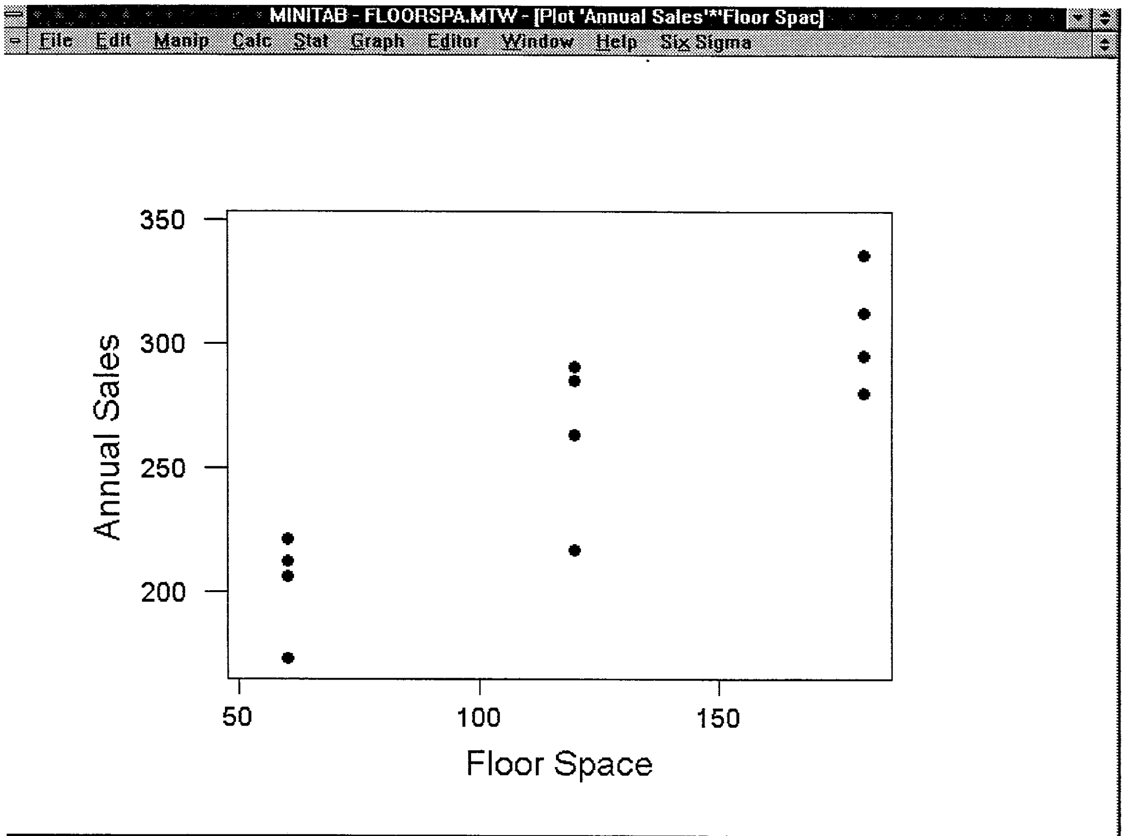
- The standard error of the slope estimate is found in the StDev column: 0.00044 (rounded up)
- The t-value is the tabled t-statistic using the degrees of freedom in the Error term of the model (8) and a confidence level of 0.025 (two-tailed test): t = 2.31

The 95% Confidence Interval for the Slope is:

0.00383 +/- 2.31(0.00044) ⟶ **(0.00281, 0.00485)**

Answer to Classroom Exercise

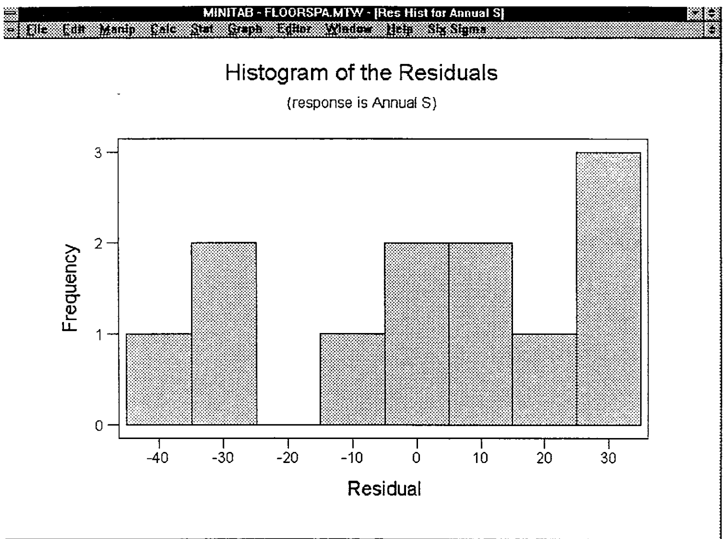
First graph the data...Graph>Plot



There appears to be a linear relationship between floor space and annual sales...

Next, run Regression to fit the model equation...
Don't forget to store the Residuals and create Residual Plots

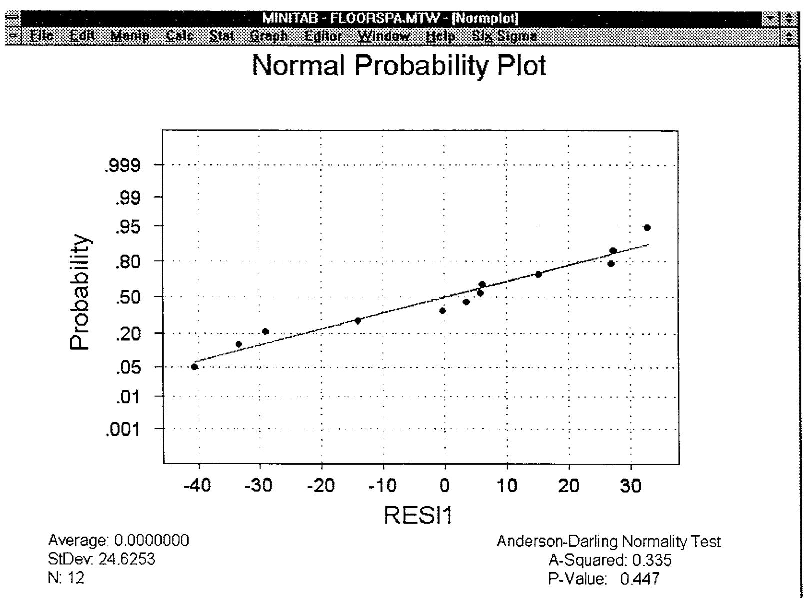
Analyze model by looking at Residual Plots



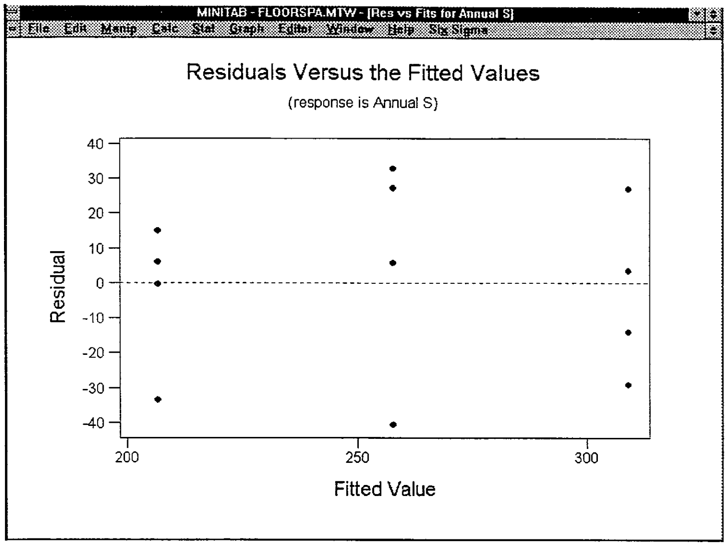
The Histogram does not look normal. Try to determine the reason for the shape of the distribution (mis-entered data, small number of data points, etc.).

Run a Normality Test on the Residuals (in column labeled 'RESI1').

We can't claim the Residuals are non-normal

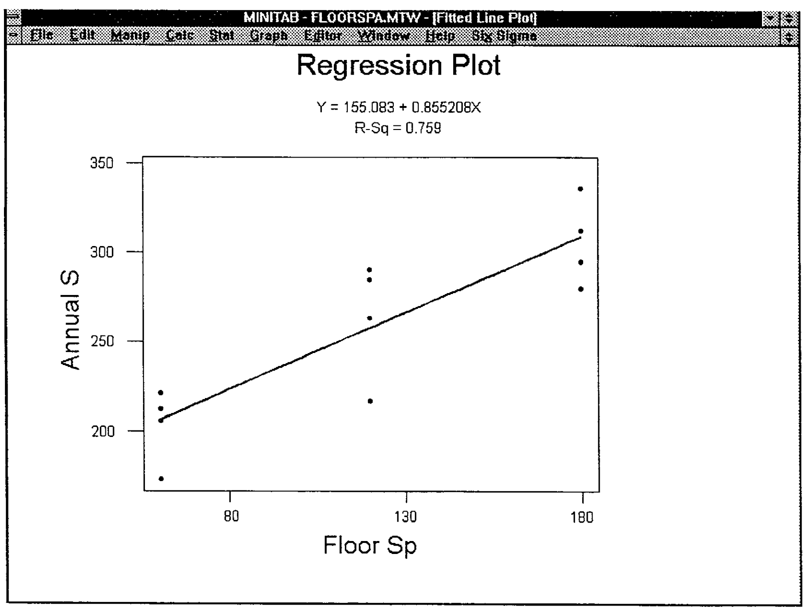


Let's look at the 'Residuals' vs 'Fits' plot



It looks like a pattern may exist. This may not be the "best fit" model for the data.

A line does not seem to fit the data well. Let's look at the Session Window



Next look at the Session Window

Regression Analysis

```
The regression equation is
Annual Sales = 155 + 0.855 Floor Space

Predictor        Coef       StDev          T        P
Constant       155.08       19.73       7.86    0.000
Floor Sp       0.8552      0.1522       5.62    0.000

S = 25.83       R-Sq = 75.9%     R-Sq(adj) = 73.5%

Analysis of Variance

Source       DF          SS          MS        F        P
Regression    1       21064       21064    31.58    0.000
Error        10        6670         667
Total        11       27734
```

This does not appear to be the "best fit" model even though the Regression p-value is <0.05 because:

- "S" is very high (the standard deviation of the error term)
- "R-Sq" is relatively low (probably good for this type of relationship, though!)
- "R-Sq(adj)" is also low (however, it is close to R-Sq, which is good)

Conclusion:

Amount of space does appear impact sales; however, the relationship may not be linear based on the Residual Plots. Also, there may be another Vital "X" that needs to be investigated and added to the equation.

Next Steps:

Find additional potential Vital "X"s or try to refit the data as a quadratic or cubic relationship...More about this in the next Tab...

Calculating a Confidence Interval for the Coefficient (Slope)

(Refer to the example on page 7.11)

The regression equation from the Session Window is:

Evap = 0.069 + 0.00383 velocity

Estimate of the Slope

Regression Analysis

```
The regression equation is
evap = 0.069 + 0.00383 velocity

Predictor        Coef        StDev          T         P
Constant       0.0692       0.1010       0.69     0.512
velocity    0.0038288    0.0004378       8.75     0.000

S = 0.1591        R-Sq = 90.5%       R-Sq(adj) = 89.3%
```

0.00383 is the estimate, based upon the data, of the slope of the line. Since it's an estimate, we know that the actual value really falls within a range of plausible values - a Confidence Interval. The **Confidence Interval** for the slope can be calculated from the following equation:

Estimated value +/- ($t_{df, \alpha}$)(std. error of the estimate)

- The standard error of the slope estimate is found in the StDev column: 0.00044 (rounded up)
- The t-value is the tabled t-statistic using the degrees of freedom in the Error term of the model (8) and a confidence level of 0.025 (two-tailed test): t = 2.31

The 95% Confidence Interval for the Slope is:

0.00383 +/- 2.31(0.00044) → **(0.00281, 0.00485)**