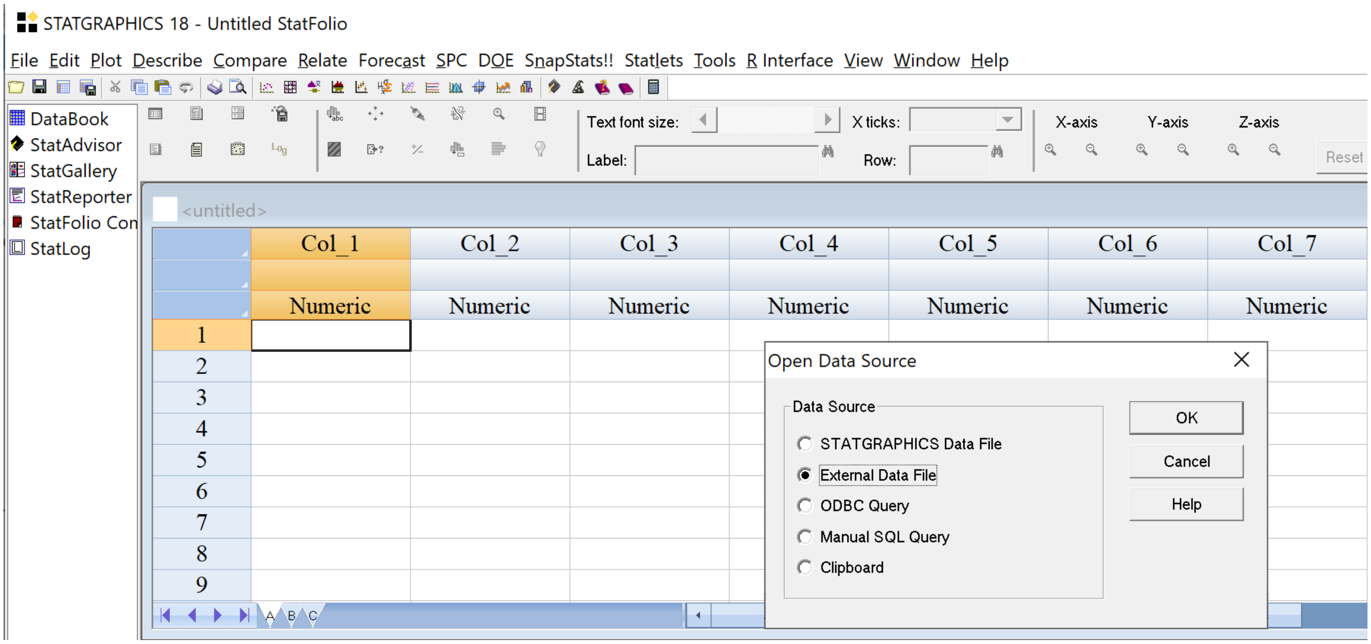
Lectures 9 & 10 - Regressing House Price on House Size

We are going to perform the first steps in exploring the relationship between the size of a house and its size in the dataset *House Price*, which can be found in the *Simple Regression* folder on Canvas. The data contains information on the sales price, house size, lot size, and number of bedrooms for 100 homes. Since it is reasonable to assume that the size of a house affects its price at time of sale, we will regress sales price (recorded in dollars) on house size (recorded in square feet).

We begin by loading the data into Statgraphics. Selecting the data input button (3rd from left) opens the *Open Data Source* tab as shown below. Select *External Data File* and click OK. (The default file format is Excel, which suits our purpose.) Then, browse to wherever you've stored the *House Price* file and select it.

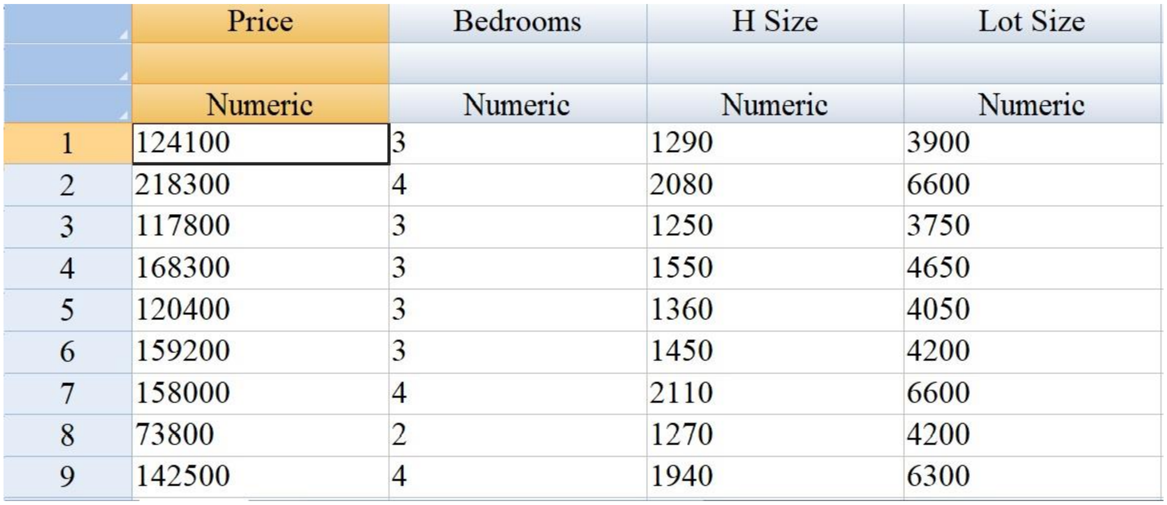


Open Data Source

Data Source

○ STATGRAPHICS Data File
● External Data File
○ ODBC Query
○ Manual SQL Query
○ Clipboard

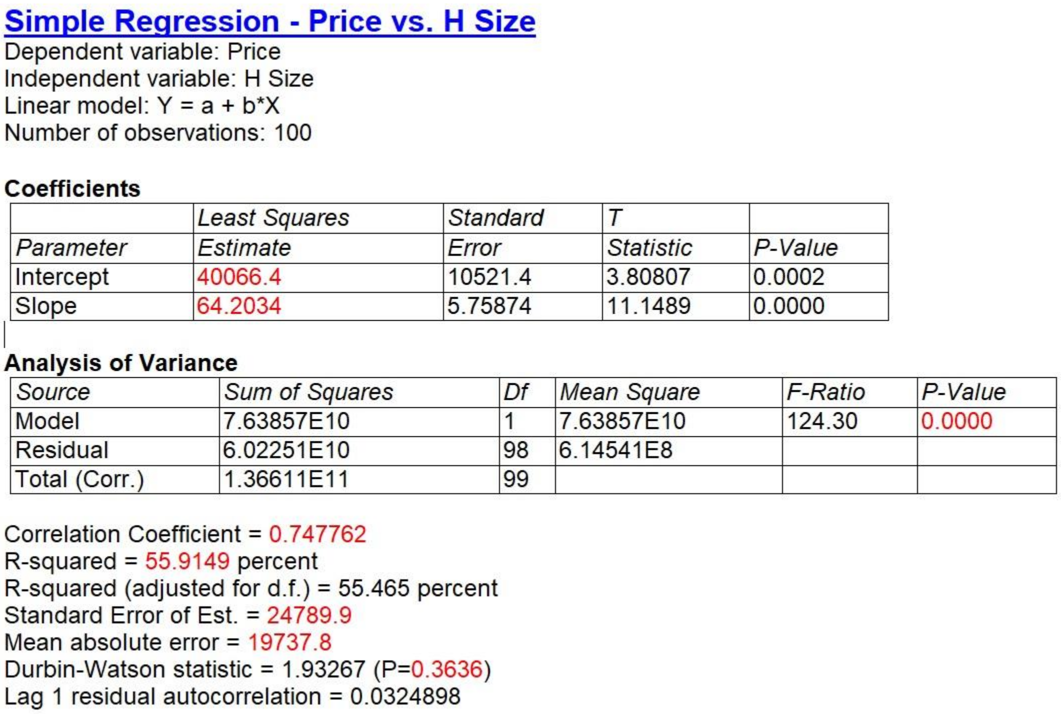
You should see the following spreadsheet (only the first 9 of 100 rows are displayed):



	Price	Bedrooms	H Size	Lot Size
	Numeric	Numeric	Numeric	Numeric
1	124100	3	1290	3900
2	218300	4	2080	6600
3	117800	3	1250	3750
4	168300	3	1550	4650
5	120400	3	1360	4050
6	159200	3	1450	4200
7	158000	4	2110	6600
8	73800	2	1270	4200
9	142500	4	1940	6300

Next, either select *Relate* → *One Factor* → *Simple Regression*, or the *Simple Regression* button on the Main Toolbar. Let *Price* by the response (Y) and *H Size* be the predictor (X) variables and click OK. Click OK to move through the *Simple Regression Options* tab. Finally, select the Table and Graph options you want and click OK to see the results of a simple linear regression of *Price* on *Size*.

Looking first at the *Analysis Summary* window, it looks like the following:



Simple Regression - Price vs. H Size

Dependent variable: Price
Independent variable: H Size
Linear model: Y = a + b*X
Number of observations: 100

Coefficients

	Least Squares	Standard	T	
Parameter	Estimate	Error	Statistic	P-Value
Intercept	40066.4	10521.4	3.80807	0.0002
Slope	64.2034	5.75874	11.1489	0.0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	7.63857E10	1	7.63857E10	124.30	0.0000
Residual	6.02251E10	98	6.14541E8		
Total (Corr.)	1.36611E11	99			

Correction Coefficient = 0.747762
R-squared = 55.9149 percent
R-squared (adjusted for d.f.) = 55.465 percent
Standard Error of Est. = 24789.9
Mean absolute error = 19737.8
Durbin-Watson statistic = 1.93267 (P=0.3636)
Lag 1 residual autocorrelation = 0.0324898

We will analyze the output in this window in more detail in subsequent lectures, but the following stand out immediately.

- The equation of the estimated regression line is $\hat{y} = 40066 + 64.2x$
- The *P*-values for hypothesis tests of the intercept and slope of the model are both small, indicating that neither are likely to be zero. (We'll talk more about this in class.)
- In SLR, the *P*-value for the model, which appears in the row labeled *Model* in the *Analysis of Variance* window, equals the *P*-value for the slope (and the model *F* statistic equals the square of the *t* statistic for the slope). Both are small, indicating that regressing *Price* on *Size* is promising.
- *R*-squared, $R^2 = \dfrac{SSR}{SST} = \dfrac{SST - SSE}{SST}$, is almost 56 percent, which implies that house size is able to explain about 56% of the variation in house price observed in the sample.

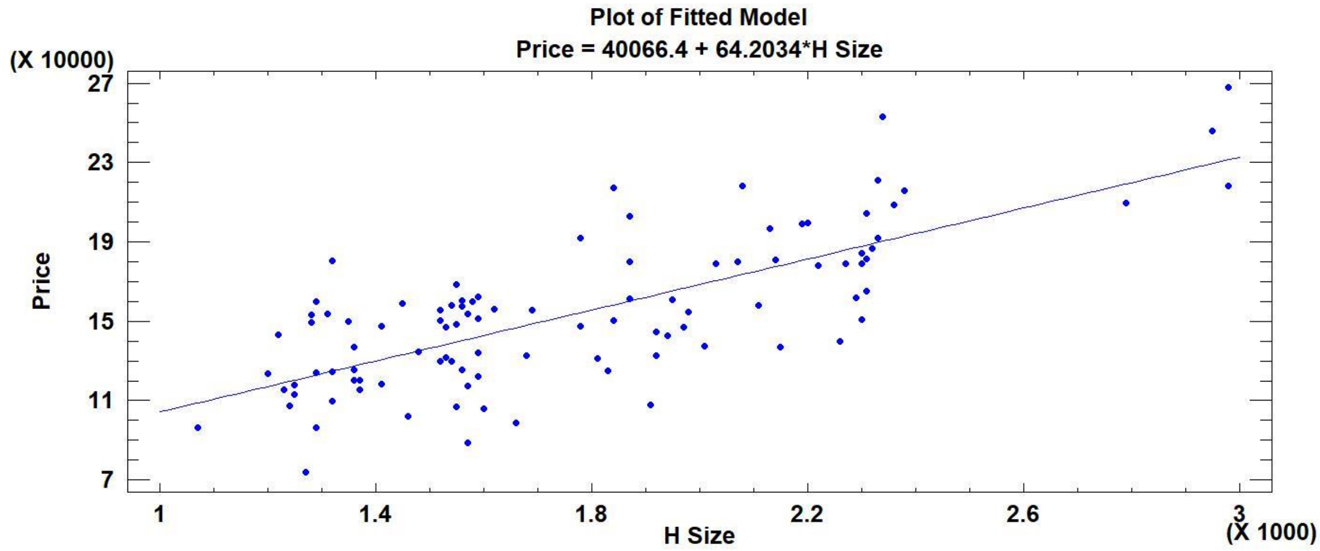
I'll interpret the slope of the estimated regression line for practice, but keep in mind that we wouldn't ordinarily interpret the slope until we had arrived at our final model. The slope is interpreted as:

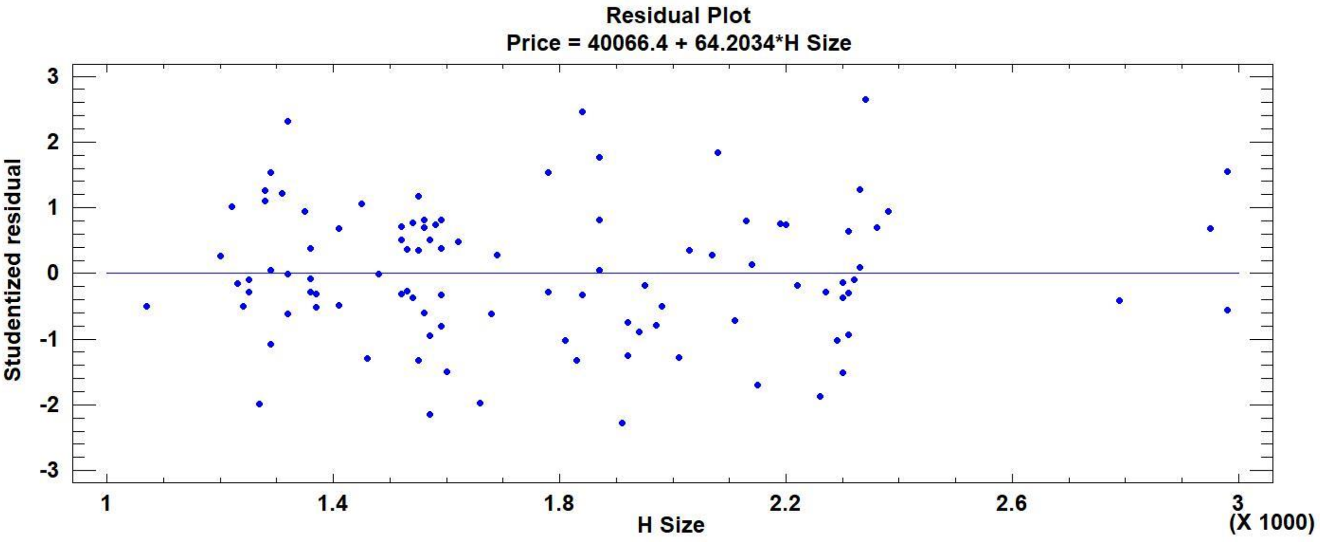
"The mean price of a house in the target population increases by about \$64 for each additional square foot."

We won't interpret the intercept because it involves houses with zero square feet, which clearly isn't an interpretation suitable for the population of houses from which our sample was drawn.

All of the first four graphs available to us supply similar information in a simple linear regression (the fifth option, *Residuals versus Row Number*, is important in regressing time series, but not here). Below I've reproduced the output from the first and third graphical options: The *Plot of Fitted Model* and the *Residuals versus X*. The spread of the residuals about the regression line looks reasonably constant for all values of house size in the data, which supports the model assumption that the error variance is constant.

- The line in the *Plot of Fitted Model* graph is the estimated regression line $\hat{y} = 40066 + 64.2x$. (I've right-clicked and selected *Pane Options*, then removed the prediction and confidence intervals in order to be able to focus on the regression line.)
- The *Residuals versus X* graph is the result of removing the regression line from the *Plot of Fitted Model* graph. (I've right-clicked and selected *Pane Options* to display the residuals to facilitate the discussion.)





The Mean Square Error, *MSE*, which is the sample estimate of the variance of the error variable ε in the model, appears in the Mean Square column in the row for the residuals. Recall that $MSE = \frac{SSE}{df} = \frac{\sum e_i^2}{n-2}$.

The Standard Error, $S = \sqrt{MSE} = \$24,789.9$, is reported among the output below R^2 in the Analysis Window.

Next, we'll use the fitted model to estimate and predict the price of a 2,150 ft^2 house from this population. First, however, we'll need to open up a new window. Find the *Tables and Graphs* button in the toolbar and select *Forecast*s. Move to that table in the output and double-click to maximize and enter the window. Right-click, select *Pane Options*, and enter 2150 in one of the empty cells and hit *Enter*. You should see

Predicted Values

	Predicted	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
X	*Y*	*Pred. Limit*	*Pred. Limit*	*Conf. Limit*	*Conf. Limit*
1070.0	**108764.**	58670.5	158858.	99318.2	118210.
2980.0	**231393.**	180072.	282713.	216776.	246009.
2150.0	**178104.**	128479.	227729.	171584.	184624.

The predicted value of $178,104 is the estimated mean price for 2,150 ft^2 houses in this population, i.e., the fitted value on our estimated regression line for a house this size. The limits for a confidence interval for mean price, as well as limits for a prediction interval (recall Friday's lecture) are provided. The level of confidence can be changed by right-clicking and using *Pane Options*.

We may be interested in flagging outliers for further study (we might want to know whether they have something in common, for example, to see if we can use this information construct an improved model). Find the *Tables and Graphs* button again and select the tables for *Unusual Residuals* and *Influential Points*.

The table for unusual residuals flags rows (houses in the data) with studentized residuals with absolute values greater than two. Statgraphics uses a complicated formula to compute *deleted* studentized residuals (also called externally studentized residuals) for each observation. Although you are not expected to memorize the formula, I do cover it in the simple linear regression notes for completeness. The studentized residuals are a measure of how far the observations are from the fitted regression line. (Statgraphics also lists the residuals themselves, residual = observed *y* – predicted *y*, for the flagged rows.)

Unusual Residuals

			Predicted		*Studentized*
Row	*X*	*Y*	*Y*	*Residual*	*Residual*
23	1840.0	217400.	158201.	59199.3	2.46
28	1910.0	107700.	162695.	-54994.9	-2.28
61	1570.0	88900.0	140866.	-51965.8	-2.15
76	2340.0	253200.	190302.	62897.6	2.65
96	1320.0	180500.	124815.	55685.1	2.32

The table for influential points flags the rows of observations that have the most potential to affect the equation of the fitted line, particularly the slope $\hat{\beta}_1$. In SLR, Statgraphics reports rows with leverage values greater than 3 times average. Leverage is discussed generally in the SLR Notes. The leverage of the *i*th observation is computed as $h_i = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum\left(X_j - \bar{X}\right)^2}$. If this looks vaguely familiar, we saw similar terms in Friday’s discussion of confidence and prediction intervals. The form of h_i gives greater leverage to points at the horizontal extremes of the scatterplot. There are other (potentially better) measures of influence, but Statgraphics doesn’t report them in simple linear regression.

Influential Points

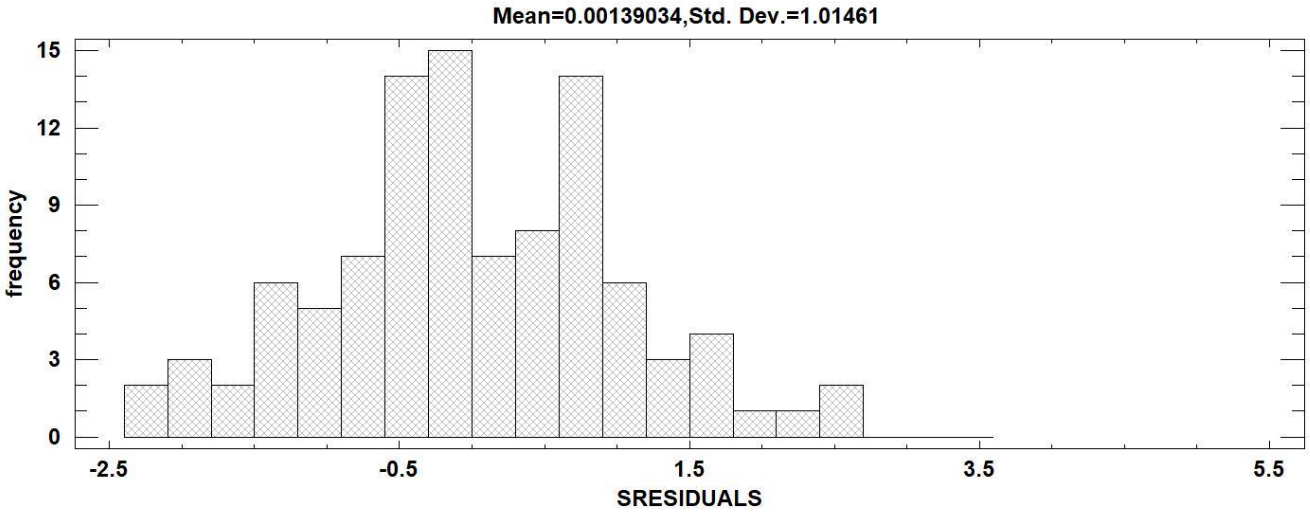
			Predicted	*Studentized*	
Row	*X*	*Y*	*Y*	*Residual*	*Leverage*
57	2980.0	267800.	231393.	1.55	0.0882791
58	2950.0	245700.	229467.	0.68	0.084428
68	2790.0	209400.	219194.	-0.41	0.0655294
97	2980.0	218100.	231393.	-0.56	0.0882791

Average leverage of single data point = 0.02

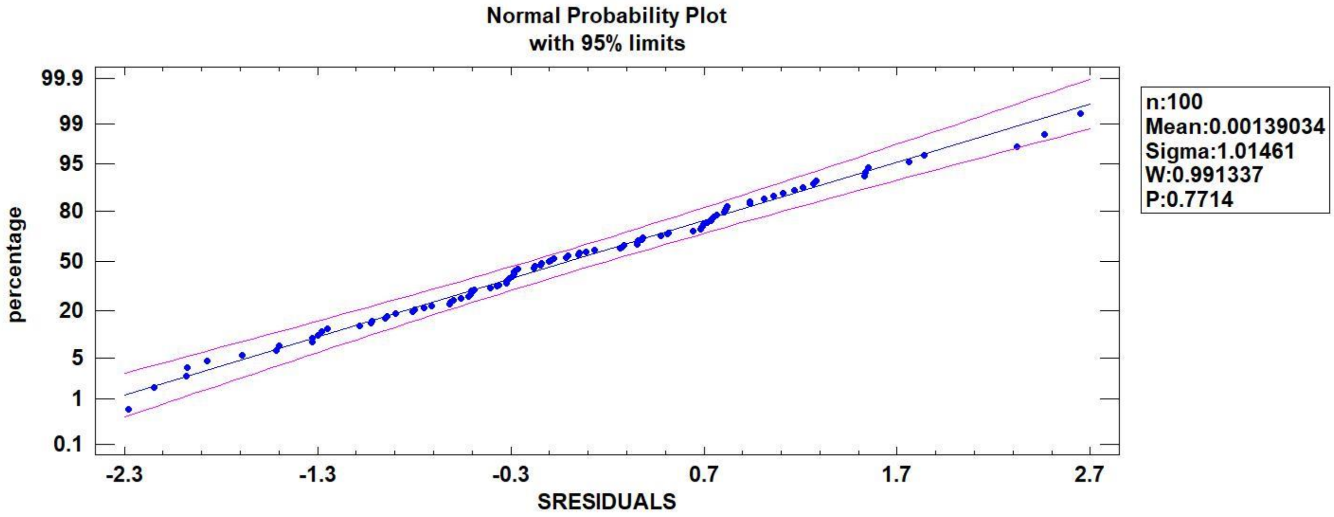
A visual way to judge an observation’s influence on the fitted regression line is to select the point on the *Plot of Fitted Model* graph and then click on the +/− button in the area below the *Main Toolbar*. This temporarily removes the point from consideration and refits the model. You will immediately see the new model and regression line. A point whose removal significantly affects the equation and graph of the fitted line is influential. (The row number of the point removed is displayed in the area above the graph.) Click on the +/− button again to return the point.

Thus far we’ve taken the normality of the error variable in the model for granted, but we can investigate whether this assumption appears reasonable as well. First, we’ll need to save the residuals (we don’t know the actual errors). To the right of the *Tables and Graphs* button you’ll find the *Save Results* button. Select it and save the Studentized Residuals to datasheet A. (Go to the datasheet to verify it’s been saved.)

A simple visual inspection can be performed on a histogram of the studentized residuals. Go to the topmost menu bar and select the *Statlets* menu. Choose *Data Exploration*→*Interactive Histogram* to view a histogram of the studentized residuals. The histogram below is close enough to bell-shaped to suggest that the normality assumption is reasonable.



Another way to evaluate whether it's plausible the errors come from a normal distribution is to have Statgraphics construct a normal probability plot of the residuals. (Either google normal probability plots or see my Stat 50 notes for a brief introduction.) Select *Plot* → *Exploratory Plots* → *Normal Probability Plot* and enter the studentized residuals as the data. Below is the graphical output from Statgraphics.



The points on the graph follow the line provided pretty well, which is indicative of the output we would expect for values drawn from a normal distribution. In addition, the StatAdvisor summary at the bottom of the *Summary Statistics* table states that the assumption that errors are normally distributed is supported by both the Skewness and Kurtosis scores in the table. (Skewness looks for asymmetry in the parent probability distribution (the normal is symmetric), while Kurtosis compares the probability concentrated in the tails of the parent distribution to that expected for a normal distribution.)

This concludes the basic introduction to fitting, validating, interpreting, and using a simple linear regression model Statgraphics. In the next set of lecture notes we'll consider some simple remedies that can be attempted if some of the assumptions, such as those for constant error and normal error are clearly violated