Simple Linear Regression Using Statgraphics

I Introduction

One of the features of Stagraphics that makes it useful for learning new statistical procedures are the YouTube videos and pdf files describing the implementation of the analyses in the program provided on Statgraphics home page. For instance, YouTubes of webinars covering the main topics of this course, regression and analysis of variance, can be viewed by going to http://www.statgraphics.com/webinars. Another resource is a user's manual that can be downloaded from http://www.statgraphics.com/Statistics-library.

A general introduction to the program, including how to input data, can be found as part of the webinar at http://blog.statgraphics.com/introducing-statgraphics-centurion-version-xvii. Inputting data appears near the beginning of the video.

Pdf files of common analyses that we'll be conducting in Statgraphics can be found at http://www.statgraphics.com/regression-analysis (for simple, multiple, and logistic regression) and http://www.statgraphics.com/analysis-of-variance (for one-way and two-way ANOVA). I've imported these files and the data used in the examples therein to a folder on our course Canvas page.

A. Importing Excel files into Statgraphics

Select the *Open Data File* button on the main tool bar (the third button from the left). If the file you want is a Statgraphics file then it will appear in the subsequent dialog box. If, however, the file is in Excel you must first select the *External Data File* radio button and use the *Input File Type* menu to change the file type to Excel. Upon selecting an Excel file, browse to a folder that contains the file you wish to import and select the file.

B. Models: Deterministic versus Probability

In the physical sciences one often encounters models of the form $v = -32t$, which describes the velocity v (in feet per second) of an object falling freely near the Earth's surface t seconds after being released. Such a model is called Deterministic because it allows us to predict the velocity *exactly* for different values of t.

Outside of the physical sciences, however, deterministic models are rare. Instead, we use Probability models, which take the form **Actual Value = Predicted Value + Error** (where the error term is considered random, i.e., uncertain and unknowable, and the *predicted value* in the equation is the value predicted by the *true* regression line). All models employed in this course are probability models. The first probability model we consider is the Simple Linear Regression model.

C. The Simple Linear Regression model

The model for Simple Linear Regression is given by $\boldsymbol{Y} = \beta_0 + \beta_1 X + \varepsilon$, where

- $\boldsymbol{Y}$ is the dependent variable, called the response variable
- X is the independent variable, called the predictor variable
- ε is the random error variable
- β_0 is the y-intercept of the line $y = \beta_0 + \beta_1 x$
- β_1 is the slope of the line $y = \beta_0 + \beta_1 x$

In the model above:

Y and ***X*** are assumed to be **correlated**, i.e., linearly related, and thus the model function takes the form of a line, $y = \beta_0 + \beta_1 x$. Although we will discuss ways to test the validity of this hypothesis later, a simple visual check can be performed by graphing a scatterplot of the *x* and *y* values in the sample and deciding if a line appears to fit the plot reasonably well. There is a button on the main toolbar for creating scatterplots.
In most applications, the **independent** variable ***X*** is just one of many variables affecting the value of the **dependent** variable ***Y***. For example, while we expect the size of a house (in square feet) to be correlated to the price at which it sells, we also expect the price to be influenced by other variables: the number of bedrooms, the age of the house, the size of the lot, the neighborhood, etc. Those variables affecting sales price which are not included in the model create variability in the sales price unaccounted for by differences in house size alone. The error variable $\boldsymbol{\varepsilon}$ represents the **random variation** in the sales price of a house due to all of the important variables missing from the model $\boldsymbol{Price} = \beta_0 + \beta_1 Sqft + \boldsymbol{\varepsilon}$.

In the model, ***Y*** and $\boldsymbol{\varepsilon}$ are both **random variables**, while *X* is considered **fixed**, i.e., known in advance. For example, several houses with the same **fixed** size of 1520 ft^2 will, nonetheless, have different sales prices for reasons discussed in the previous paragraph. For each of these houses, ***Y*** and $\boldsymbol{\varepsilon}$ represent the actual sales price and the difference between the actual price and the price predicted by the true regression line, respectively. They are both random because the model has excluded other variables important to determining sales price.

β_0 and β_1 in the model $\boldsymbol{Price} = \beta_0 + \beta_1 Sqft + \boldsymbol{\varepsilon}$ are called model **parameters**. They are unknown constants of the model, i.e., *numbers* rather than variables. Statgraphics estimates β_0 and β_1 using the data. The **sample statistics** $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate the model's parameters β_0 and β_1, respectively.

1. Model Assumptions

The Simple Linear Regression model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \varepsilon$ makes two different types of assumptions.

- The first of these, mentioned previously, postulates that ***Y*** and ***X*** are linearly related, i.e., that a line $y = \beta_0 + \beta_1 x$ appropriately models the relationship between the dependent and independent variables.
- The second *set* of assumptions involves the distribution of the error variable, $\boldsymbol{\varepsilon}$. Specifically:
  1. The random variable $\boldsymbol{\varepsilon}$ is assumed to be normally distributed, with mean $\mu_\varepsilon = 0$, and constant variance σ_{ε}^2. (Although the normality of the error variable, $\boldsymbol{\varepsilon}$, isn't an essential assumption, it *is* required if we wish to do inference on the model parameters β_0 and β_1.)
  2. The errors associated with different observations are assumed to be independent of each other.

The first assumption about the error variable makes construction of confidence intervals for the mean value of ***Y*,** for particular values of *X*, possible. It also allows us to conduct useful hypothesis tests. The constant variance part of the assumption states that the variation in the values of the dependent variable ***Y*** about the line $y = \beta_0 + \beta_1 x$ is the same for all values of *X*.

The second assumption about the error variable (that the errors are independent) is important in time-series regression, and will be addressed if we discuss the regression of time-series.

It is important to note that the assumptions made in Simple Linear Regression may not be justified by the data. *Using the results of a regression analysis when the assumptions are invalid may lead to serious errors!* Prior to reporting the results of a regression analysis, therefore, you must demonstrate that the assumptions underlying the analysis appear reasonable given the data upon which the analysis is based.

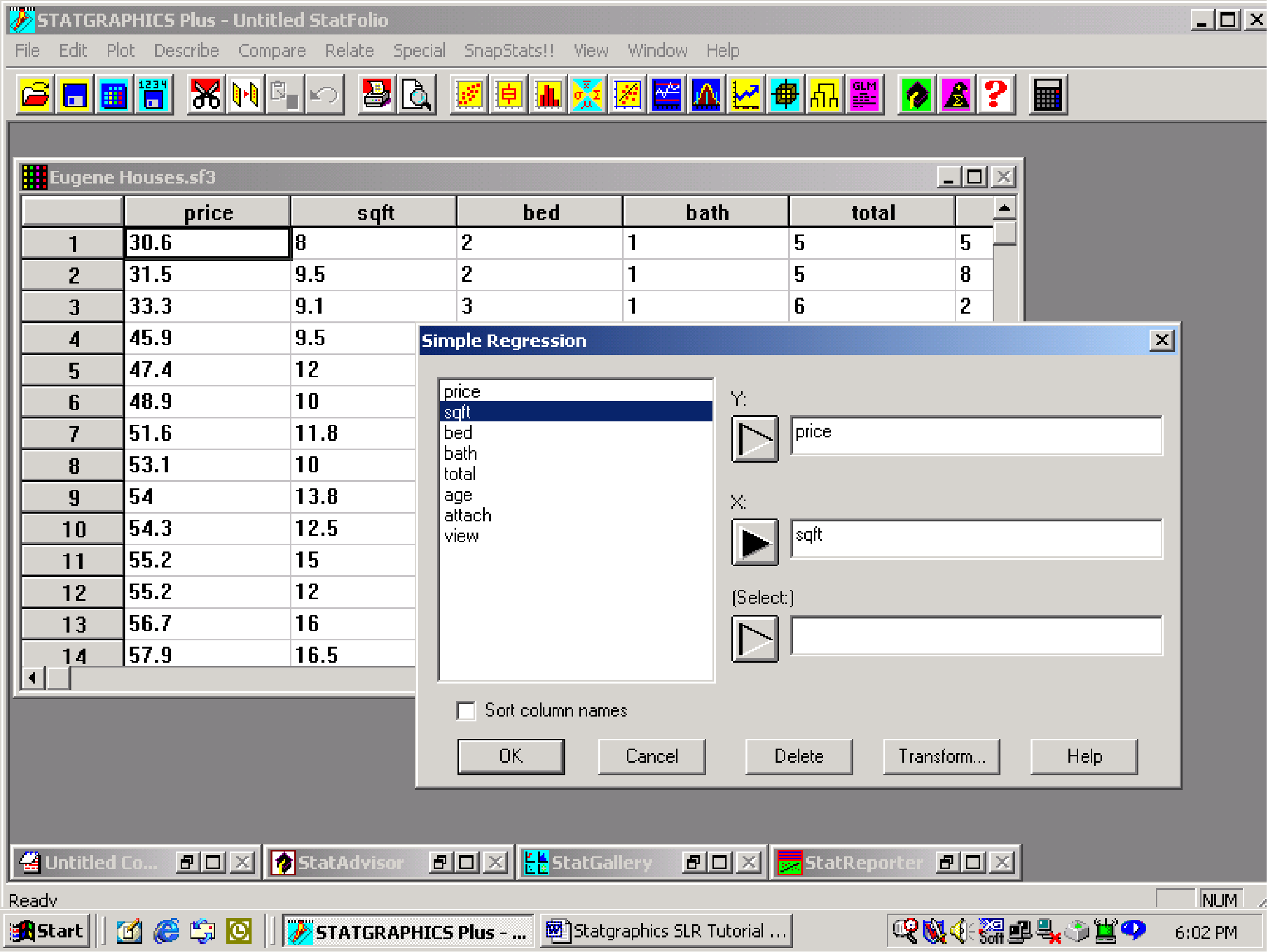
II The Analysis Window

Example 1: In the following discussion we'll use the file Example 1 - Eugene Houses consisting of 50 houses in Eugene, Oregon, sold in 1973. Below is a brief description of each of the variables.

- Price – sales price, in thousands of dollars
- Sqft – size of the house, in hundreds of square feet
- Bed – number of bedrooms
- Bath – number of bathrooms
- Total – total number of rooms
- Age – age of the house, in years
- *Attach – whether the house has an attached garage
- *View – whether the house has a nice view

* Note that Attach and View are **qualitative** (categorical) variables, while all other variables are **quantitative**. (We will postpone the discussion of the use of qualitative variables in regression until the notes for Multiple Linear Regression.)

To reach the analysis window for simple linear regression in Statgraphics, follow: *Relate > Simple Regression* and use the **input dialog box** to enter the dependent and independent variables. The example for "Eugene Houses", regressing price on sqft, appears below.



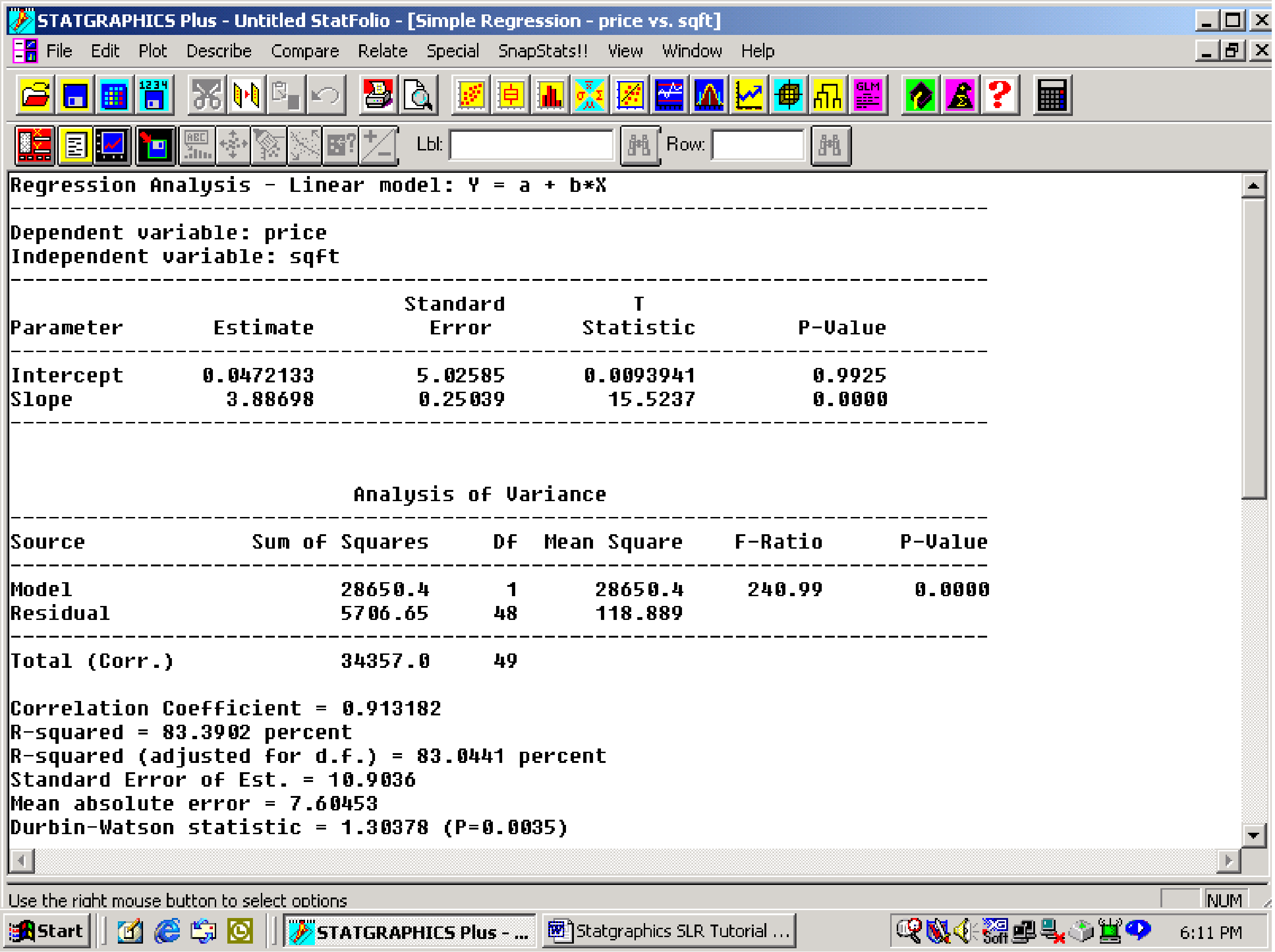
	price	sqft	bed	bath	total	
1	30.6	8	2	1	5	5
2	31.5	9.5	2	1	5	8
3	33.3	9.1	3	1	6	2
4	45.9	9.5				
5	47.4	12				
6	48.9	10				
7	51.6	11.8				
8	53.1	10				
9	54	13.8				
10	54.3	12.5				
11	55.2	15				
12	55.2	12				
13	56.7	16				
14	57.9	16.5				

Y: price

X: sqft

(Select:)

The **analysis summary** window, shown below, is the default *Tabular Options* (text) window. We next discuss the interpretation of some of the output appearing in the analysis window.



STATGRAPHICS Plus - Untitled StatFolio - [Simple Regression - price vs. sqft]
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help
Lbl:
Row:
Use the right mouse button to select options
NUM
Start
STATGRAPHICS Plus - ...
Statgraphics SLR Tutorial ...
6:11 PM

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: price
Independent variable: sqft
-----------------------------------------------------------------------------
                                Standard          T
Parameter       Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
Intercept      0.0472133        5.02585       0.0093941         0.9925
Slope            3.88698        0.25039         15.5237         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                         28650.4      1      28650.4     240.99       0.0000
Residual                      5706.65     48      118.889
-----------------------------------------------------------------------------
Total (Corr.)                 34357.0     49

Correlation Coefficient = 0.913182
R-squared = 83.3902 percent
R-squared (adjusted for d.f.) = 83.0441 percent
Standard Error of Est. = 10.9036
Mean absolute error = 7.60453
Durbin-Watson statistic = 1.30378 (P=0.0035)
```

A. *The Three Sums of Squares*

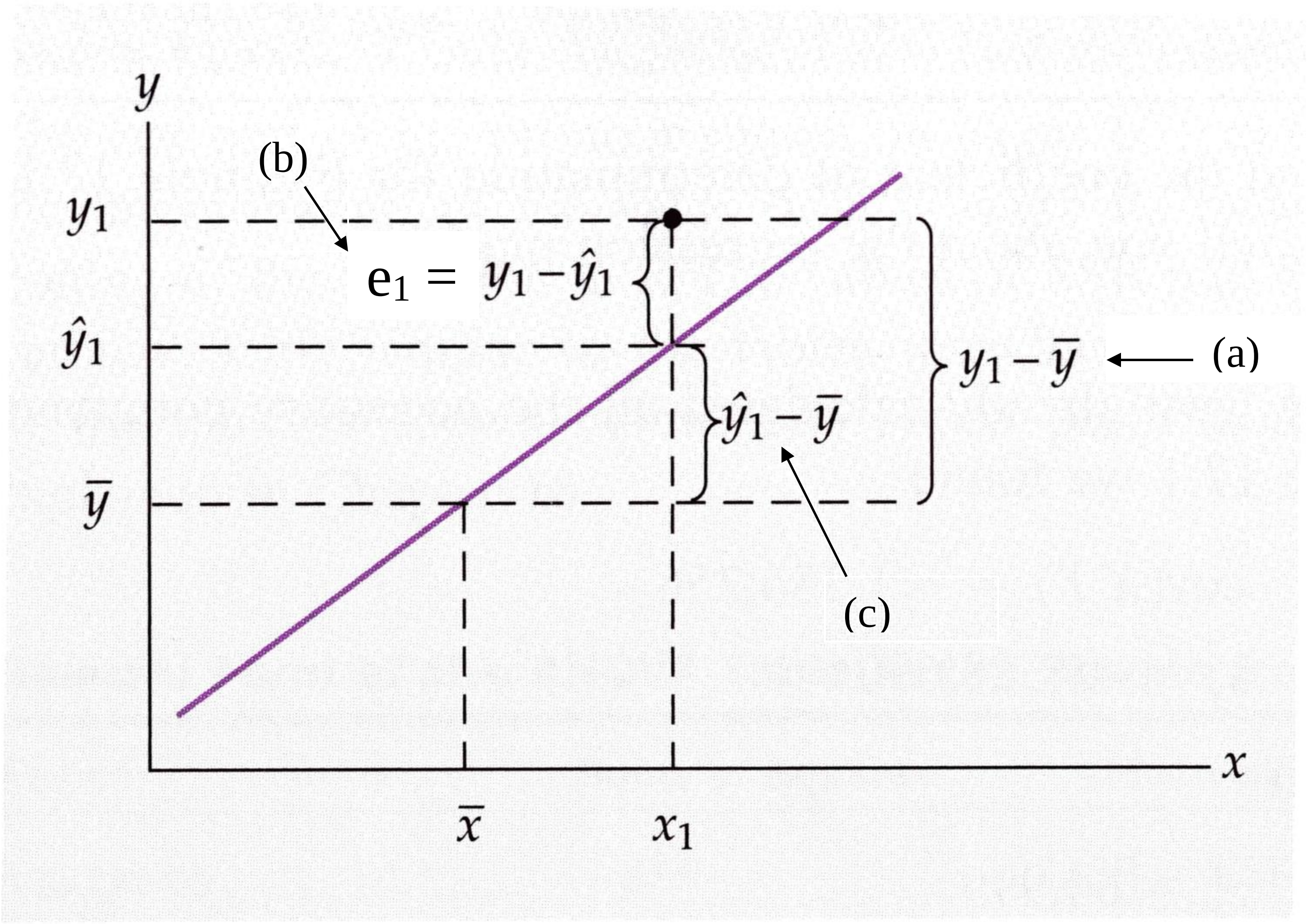
Let (x_i, y_i) represent the *x* and *y* values of the i^th^ observation. Define $\hat{y}_i$ to be the model's predicted value for *y* when $x = x_i$, i.e., $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

From the picture below, we derive the following three (vertical) differences for the ith observation:

(a) $y_i - \bar{y}$ = The deviation from the mean. (The deviation is *positive* if y_i is above the average for the sample.)

(b) $y_i - \hat{y}_i$ = The ith prediction error (or ith Residual, **e**$_i$ **)** for the regression line $\hat{y} = b_0 + b_1 x$

(c) $\hat{y}_i - \bar{y}$ = The difference between the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and the horizontal line $y = \bar{y}$ at $x = x_i$



From the picture, we note that part of the difference between y_i and $\bar{y}$ is *explained* by the difference between x_i and $\bar{x}$ (the *explained* part is given by the "rise" $\hat{y}_i - \bar{y}$ for the "run" $x_i - \bar{x}$). The *un*explained part of the difference between y_i and $\bar{y}$ is given by the ith residual e_i = $(y_i - \hat{y}_i)$. {You can verify, algebraically as well as visually, that the explained difference plus the unexplained difference equals the deviation from the mean: $(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) = y_i - \bar{y}$.} The goal in regression is to minimize the unexplained differences, i.e., the prediction errors e_i .

To find the equation of the line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ that minimizes the prediction errors (and to determine the effectiveness of ***X*** in explaining ***Y***), we might begin by examining the totals of the differences discussed in the previous paragraph over the *n* observations in the sample. However, since each of the three differences sum to zero for the sample, it is necessary to *square* the differences before summing them. This leads to the definition of the following three Sums of Squares:

Total Sum of Squares, $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$, is a measure of the total variation of *Y* about its mean for the sample

Note: $\frac{SST}{n-1} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$ is just the sample variance of the *y* values in the data.

Error Sum of Squares, $SSE = \sum_{i=1}^{n}(y_{\text{i}} - \hat{y}_{\text{i}})^2$**,** measures the variation of *Y* left unexplained by the variation in *X*, i.e., the variation of *Y* about the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Regression Sum of Squares, $SSR = \sum(\hat{y}_{\text{i}} - \bar{y})^2$, measures the variation of *Y* explained by the variation in *X*

Note: Statgraphics refers to *SSE* as the ***Residual*** Sum of Squares because it equals the sum of the squared residuals, and refers to *SSR* as the ***Model*** Sum of Squares because it relates to the regression model. In Statgraphics, the three sums of squares appear in the second column of the Analysis of Variance table in the *Analysis Summary* window as shown below.

```
                         Analysis of Variance
-----------------------------------------------------------------------------
Source          Sum of Squares     Df      Mean Square          F-Ratio      P-
-----------------------------------------------------------------------------
Model                SSR            1      MSR = SSR/1          F = MSR/MSE
Residual             SSE          n - 2    MSE = SSE/(n - 2)
-----------------------------------------------------------------------------
Total (Corr.)        SST          n - 1

*
*
* Note: The sample variance for Y, s^2 = SST/(n - 1), is an example of a Me
*       that is not computed by Statgraphics.
*
*
```

Example 1 (continued): The ANOVA Table for the Eugene data, regressing house price (the variable *Price*) on house size (the variable *Sqft*), appears below. Here *SSR* = 28,650; *SSE* = 5,706; and *SST* = 34,357 (all in units of $1,000 squared).

```
                             Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      28650.4      1      28650.4     240.99       0.0000
Residual                   5706.65     48      118.889
-----------------------------------------------------------------------------
Total (Corr.)              34357.0     49
```

B. The Least-Squares Criterion

The goal in simple linear regression is to determine the equation of the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ that *minimizes* the total unexplained variation in the observed values for ***Y***, and thus maximally "explains" the observed variation in ***Y***. However, the residuals, which represent the unexplained variation, sum to zero. Therefore, simple linear regression minimizes the sum of the *squared* residuals, *SSE*. This is called the Least-Squares Criterion and results in formulas for computing the *y* – intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ for the least-squares regression line. At this point there is no need to memorize formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$. It is enough to know that Statgraphics computes them and places them in the *Analysis Window* in the column labeled "Estimate."

Example 1 (continued): Below is the output for the regression of house price (in thousands) on square footage (in hundreds). The numbers in the "Estimate" column are $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
Regression Analysis - Linear model: Y = a + b*X
---------------------------------------------------------------------
Dependent variable: price
Independent variable: sqft
---------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
---------------------------------------------------------------------
Intercept      0.0472133        5.02585      0.0093941         0.9925
Slope            3.88698        0.25039        15.5237         0.0000
---------------------------------------------------------------------
```

The Mathematics of Least Squares

The quantity to be minimized is $SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$. In particular, we seek to fit the observed values y_i with a line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Replacing $\hat{y}_i$ in the equation for *SSE* with $\hat{\beta}_0 + \hat{\beta}_1 x_i$, we obtain $SSE = \sum_{i=1}^{n} \left(y_\mathrm{i} - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$.

The only free variables in the equation for *SSE* are the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$. From Calculus, the natural thing to do to minimize *SSE* is differentiate it with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set the derivatives to zero. (Note: since *SSE* is a function of the ***two*** independent variables $\hat{\beta}_0$ and $\hat{\beta}_1$, the derivatives are "partial" derivatives.)

$$\frac{\partial\, SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)$$

$$\frac{\partial\, SSE}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)$$

Setting the right-hand sides above equal to zero leads to the following system of two equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

$$\sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

Expanding the sums, we have:

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

Rearranging terms, we arrive at the following system of two equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i$$

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i$$

Because the system is *linear* in $\hat{\beta}_0$ and $\hat{\beta}_1$, it can be solved using the tools of linear algebra! We we'll postpone this until we introduce the matrix representation of the simple linear regression model later. For now, it's enough to know that a solution to the system is

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\left[\left(\sum x_i\right)\left(\sum y_i\right)\right]}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Note 1: To show that the two forms given above for $\hat{\beta}_1$ are equivalent, we use (for the numerator)

$$\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y} = \sum x_i y_i - n\bar{x}\bar{y} = \sum x_i y_i - \frac{1}{n}\left[\left(\sum x_i\right)\left(\sum y_i\right)\right]$$

A similar manipulation is used to show that $\sum\left(x_i - \bar{x}\right)^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$ in the denominator.

Note 2: $\hat{\beta}_1 = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}$ is the ratio of the sample covariance to the sample variance for *X*. It can be shown that the true slope β_1 is given by $\beta_1 = \frac{Cov\left(X, Y\right)}{\sigma_X^2}$

Discussion: Remember that $\hat{\beta}_0$ and $\hat{\beta}_1$ should be viewed as random variables until the data has been collected (it is somewhat unfortunate that the same symbols are used here for both the random intercept and slope, on the one hand, and their "observed" values computed from a particular sample). As random variables, $\hat{\beta}_0$ and $\hat{\beta}_1$ have probability distributions. We will investigate the distribution of $\hat{\beta}_1$ later.

C. Extrapolation

In algebra, the line *y* = *a* + *bx* is often assumed to continue forever. In Simple Linear Regression, assuming that the same linear relationship between ***X*** and ***Y*** continues for values outside the range of *X* observed is called **extrapolation** and should be avoided. For the Eugene data, the houses observed range in size from 800 ft^2 to 4,000 ft^2. Using the estimated regression line to predict the price of a 5,000 square foot house (in Eugene in 1973) would be inappropriate because it would involve extrapolating the regression line beyond the range of house sizes for which the linear relationship between price and size has been estimated.

D. Interpreting the Estimated Regression Coefficients

Example 1 (continued): The **sample statistics** $\hat{\beta}_0$ = 0.0472 and $\hat{\beta}_1$ = 3.887 estimate the model's *y*-intercept β_0 and slope β_1, respectively. In algebra, the *y*-intercept of a line is interpreted as the value of *y* when *x* = 0. In simple regression, however, it is not advisable to **extrapolate** the linear relationship between ***X*** and ***Y*** beyond the range of values contained in the data. Therefore, it is unwise to interpret $\hat{\beta}_0$ *unless x* = 0 is within the range of values for the independent variable actually observed. We will now interpret the estimated regression coefficients for the Eugene data.

- $\hat{\beta}_0$ **:** This would (naively) be interpreted as the estimated mean price (in thousands of dollars) of houses (in Eugene in 1973) with zero square feet, i.e., the mean price of the land. However, since all properties in the data involve houses, it is *not* appropriate to interpret $\hat{\beta}_0$.
- $\hat{\beta}_1$ **:** The estimated mean house price increases by $3,887 for each additional 100 ft^2 in the size of a house. Note that I have included the proper *units* in my interpretation. Note, also, that we are estimating the change in the **mean** house price associated with a 100 ft^2 increase in size. This reminds us that the estimated least-squares regression line is used to predict the *mean* value of ***Y*** for different values of *X*.

E. The Standard Error of the Estimate: S$_\varepsilon$

The Distribution of *Y*: In the simple linear regression model, $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$, only ***Y*** and ε are random variables, and the error ε is assumed to have mean 0. Thus, by the **linearity of expectation**, $\boldsymbol{E}\{\boldsymbol{Y} \mid X = x\} = \boldsymbol{E}\{\beta_0 + \beta_1 x + \boldsymbol{\varepsilon}\} = \beta_0 + \beta_1 x$. This states that the true regression line $y = \beta_0 + \beta_1 x$ is a line of means, specifically the conditional means for ***Y*** given *X*.

Also, because β_0, β_1, and *X* are fixed, the variance of ***Y*** derives from the variance of $\boldsymbol{\varepsilon}$, i.e., $\sigma_{\boldsymbol{Y}}^2 = \sigma^2$. So now we know the mean and variance of ***Y***, at least in theory. (Except for the small detail that we don't actually know the values of any of the parameters β_0, β_1, and σ^2, which is why we estimate them from the data.)

Having established the mean and variance of ***Y***, all that remains is to identify the family of distributions it comes from, i.e., its "shape." Here we make use of the assumption that the error is normal, $\boldsymbol{\varepsilon} \sim \boldsymbol{N}\left(0, \sigma^2\right)$. Because linear functions of normal variates are normal, and ***Y*** is a linear function of $\boldsymbol{\varepsilon}$ in the model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$, ***Y*** must be normally distributed.

Putting the previous three paragraphs together, we have $\boldsymbol{Y}_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$.

Having seen that the variance of ***Y*** is derived from σ^2, it will be seen later that other random variables, especially the estimator of the slope, $\hat{\beta}_1$, also have variances that are functions of the unknown parameter σ^2. So, it's time to estimate σ^2 !!!

Estimating the error variance, σ^2 : The model assumes that the variation in the *actual* values for ***Y*** about the TRUE regression line $y = \beta_0 + \beta_1 x$ is constant, i.e., the same for all values of the independent variable X. (Sadly, this is *not* true of the variation about the *estimated* regression line $\hat{\boldsymbol{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X$, but more on that shortly.) The variance of the error variable, σ^2, is a measure of this variation. $\boldsymbol{S}^2 = \dfrac{\sum_{i=1}^{n}\left(\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_i\right)^2}{n-2}$ is an unbiased estimator of σ^2, called the mean square error, abbreviated *MSE*. The estimated *MSE* for the Eugene house price example appears in the row containing the Error Sum of Squares, *SSE*, in the column labeled "Mean Square."

Although we will not derive the formula for the mean square error, $\boldsymbol{S}^2$, we can justify the degrees of freedom in the denominator as follows. We begin the problem of estimating model parameters with the *n* independent bits of information obtained from the sample. However, prior to estimating the error variance σ^2, we had to estimate the intercept β_0 and slope β_1 of the regression line $y = \beta_0 + \beta_1 x$ used to estimate the deviations $\boldsymbol{Y}_i - \hat{\boldsymbol{Y}}$ in the numerator of the formula for the *MSE*. In general, every time you estimate a parameter, you lose one degree of freedom going forward, and we've estimated the two parameters β_0 and β_1. Therefore, there are only *n* - 2 degrees of freedom (independent bits of information) still available for estimating σ^2.

Finally, the **standard error of the estimate,** $s=\sqrt{\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n-2}}$, estimates the standard deviation σ of the error variable ε. The estimated value of the **standard error of the estimate**, in units of thousands of dollars in the Eugene house price example, appears in the *Analysis Summary* window below the Analysis of Variance table.

Discussion: Statgraphics has no way of knowing what units the data is recorded in, so we must introduce the appropriate units for the independent and dependent variables when interpreting results or using the estimated regression line in prediction.

F. Are Y and X correlated? Testing β_1

If the slope, β_1, of the true regression line $y=\beta_0+\beta_1 x$ is *zero*, then the regression line is simply the horizontal line $y = \beta_0$, in which case the **expected value** of ***Y*** is the same for *all* values of *X*. This is just another way of saying that ***X*** and ***Y*** are *not* linearly related. Although the value of the true slope β_1 is unknown, inferences about β_1 can be drawn from the sample slope $\hat{\beta}_1$. A hypothesis test of the slope is used to determine if the evidence for a non-zero β_1 is strong enough to support the assumed linear dependence of ***Y*** on ***X***. For the test,

- H_0**:** $\beta_1 = 0$, i.e., *X* and *Y* are *not* linearly related

  H_A**:** $\beta_1 \neq 0$, i.e., the two variables *are* linearly related
- The Test Statistic is $\boldsymbol{t}=\frac{\hat{\beta}_1-\beta_1}{\boldsymbol{S}_{\hat{\beta}_1}}=\frac{\hat{\beta}_1}{\boldsymbol{S}_{\hat{\beta}_1}}$ (because $\beta_1 = 0$ in the null hypothesis), where $\boldsymbol{S}_{\hat{\beta}_1}$ is the sample standard deviation of $\hat{\beta}_1$ (called the **standard error of the slope**).

  Note: the test statistic has a ***t*** distribution with *n* – 2 degrees of freedom if the error variable is normally distributed with constant variance. For large samples, the assumption of constant variance is more important to the test than the requirement that the error be normally distributed.

Statgraphics reports the values of $\hat{\beta}_1$, $s_{\hat{\beta}_1}$, *t*, and the *P*-value for the test in the second, "Slope," row of the *Analysis Summary* window. The results for the Eugene example are shown below. The *P*-value of 0.0000 for the estimated slope allows us to reject the null hypothesis, H_0, and conclude that the data strongly suggests that ***X*** and ***Y*** are linearly related.

```
                                Standard          T
Parameter       Estimate          Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept      0.0472133         5.02585      0.0093941          0.9925
Slope            3.88698         0.25039        15.5237          0.0000
```

Discussion: Statgraphics also conducts a test of β_0, where the hypotheses are $\begin{array}{l}H_0:\beta_0=0\\ H_A:\beta_0\neq 0\end{array}$. A small *P*-value leads to rejection of the null hypothesis that the true regression line passes through the origin.

The Distribution of the Sample Slope, $\hat{\beta}_1$

The *t*-test of the slope $\hat{\beta}_1$, conducted above in StatGraphics, is valid if the distribution of $\hat{\beta}_1$ is normal. We now set out to derive the distribution of the sample slope $\hat{\beta}_1$. The distribution of $\hat{\beta}_1$ is based upon the following:

- $\hat{\beta}_1$ can be written as a **linear combination** of the $\boldsymbol{Y}_i$
- The $\boldsymbol{Y}_i$ are **independent** and have distribution $\boldsymbol{Y}_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$

To show that the sample slope $\hat{\beta}_1$ can be written as a linear combination of the observations on $\boldsymbol{Y}$, rewrite

$$\hat{\beta}_1 = \frac{\sum\left(X_i - \bar{X}\right)\left(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\sum \boldsymbol{Y}_i\left(X_i - \bar{X}\right) - \bar{\boldsymbol{Y}}\sum\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \frac{\sum \boldsymbol{Y}_i\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2}$$

, where we have used the fact that the deviations $X_i - \bar{X}$ sum to zero. We can now see that the sample slope $\hat{\beta}_1$ is linear in the $\boldsymbol{Y}_i$ by rewriting it as

$$\hat{\beta}_1 = \frac{\sum \boldsymbol{Y}_i\left(X_i - \bar{X}\right)}{\sum\left(X_i - \bar{X}\right)^2} = \sum \boldsymbol{Y}_i\left[\frac{\left(X_i - \bar{X}\right)}{\sum\left(X_j - \bar{X}\right)^2}\right] = \sum c_i \boldsymbol{Y}_i$$

, where the $c_i = \dfrac{\left(X_i - \bar{X}\right)}{\sum\left(X_j - \bar{X}\right)^2}$ are *constants* (because the X_i are treated as fixed), and are *functions only of the* X_i! Thus, $\hat{\beta}_1$ is a linear combination of the $\boldsymbol{Y}_i$ as claimed.

Discussion: The two sums appearing in the quotient for $\hat{\beta}_1$ are evaluated separately. For this reason, the index in the denominator sum was changed from *i* to *j* prior to distributing the sum into the coefficient of $\boldsymbol{Y}_i$ to form c_i

An immediate consequence of the work above is that the sample slope $\hat{\beta}_1$ is normally distributed as a linear combination of the normally distributed $\boldsymbol{Y}_i$. Beyond that, however, it also allows us to determine the mean and standard error of $\hat{\beta}_1$. It makes a nice, and not too difficult, homework problem to fill in the missing steps below:

1. $\boxed{E\left(\hat{\beta}_1\right) = \beta_1}$, i.e., the sample slope $\hat{\beta}_1$ is an **unbiased estimator** of the true slope β_1. To show this, $E\left(\hat{\beta}_1\right) = E\left(\sum c_i \boldsymbol{Y}_i\right) = \sum c_i\, E\left(\boldsymbol{Y}_i\right) = \beta_1 \sum c_i X_i$, because β_0, β_1 are parameters and $\sum c_i = 0$. If we show that $\sum c_i X_i = 1$, we're done. $\sum c_i X_i = \sum c_i X_i - \bar{X}\sum c_i = \sum c_i X_i - \sum c\,\bar{X} = \sum c_i\left(X_i - \bar{X}\right) = 1$.
2. $\boxed{\sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}}$ because $\sigma^2_{\sum c_i \boldsymbol{Y}_i} = \sum c_i^2\, \sigma^2_{\boldsymbol{Y}_i}$, and is estimated by $\boldsymbol{S}^2_{\hat{\beta}_1} = \dfrac{\boldsymbol{S}^2}{\sum\left(X_i - \bar{X}\right)^2}$, where $\boldsymbol{S}^2$ is the mean square error, *MSE*, derived earlier.
3. The estimated standard error of $\hat{\beta}_1$ is $\boldsymbol{S}_{\hat{\beta}_1} = \dfrac{\boldsymbol{S}}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}}$.

Confidence Intervals for the true slope, β_1

Because $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}\right)$, we can construct a confidence interval for the true slope β_1 using a *t*-distribution with *n* - 2 degrees of freedom (see the previous discussion about the choice of degrees of freedom). Therefore, a $(1-\alpha)100\%$ confidence interval for β_1 is given by $\hat{\beta}_1 \pm t_{n-2,\alpha/2} \dfrac{\boldsymbol{S}}{\sqrt{\sum\left(X_i - \bar{X}\right)^2}}$, where $df = n-2$.

Confidence intervals for the true slope β_1 can be obtained from StatGraphics (or other software), so, in practice, there is no point in constructing them by hand.

The Distribution of the Sample Intercept, $\hat{\beta}_0$

$E\left(\hat{\beta}_0\right) = E\left(\bar{\boldsymbol{Y}} - \hat{\beta}_1 \bar{X}\right) = E\left(\bar{\boldsymbol{Y}}\right) - \bar{X}E\left(\hat{\beta}_1\right) = \left(\beta_0 + \bar{X}\beta_1\right) - \bar{X}E\left(\hat{\beta}_1\right) = \beta_0 + \bar{X}\beta_1 - \bar{X}\beta_1 = \beta_0$. So, $\boxed{E\left(\hat{\beta}_0\right) = \beta_0}$.

$\sigma^2_{\hat{\beta}_0} = \sigma^2_{\bar{Y} - \hat{\beta}_1\bar{X}} = \sigma^2_{\bar{Y}} + \bar{X}^2\sigma^2_{\hat{\beta}_1} = \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right)$. So, $\boxed{\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum\left(X_i - \bar{X}\right)^2}\right)\right)}$.

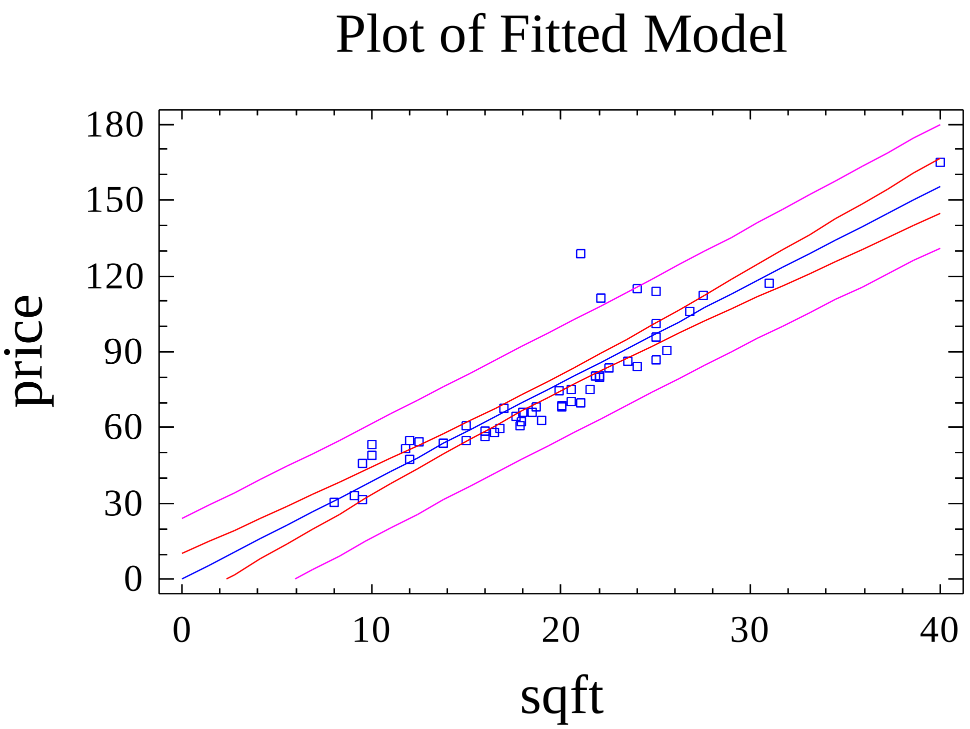
Confidence intervals for the true *y*-intercept β_0 can also be obtained in StatGraphics.

G. Measuring the strength of the correlation: R²

Having first determined, from the hypothesis test of β_1, that a statistically significant linear relationship exists between the dependent and independent variables, the *strength* of the linear relationship is measured by the **Coefficient of Determination**, $R^2 = \dfrac{SSR}{SST} = \dfrac{\text{the observed variation in } \boldsymbol{Y} \text{ explained by } X}{\text{the total observed variation in } \boldsymbol{Y}}$. From its definition, R^2 equals the *proportion* of the observed variation in ***Y*** explained by X, i.e., explained by the regression model. Thus $0 \le R^2 \le 1$, with $R^2 \cong 1$ if the line fits the data well, and $R^2 \cong 0$ if the line fits poorly. Finally, $R^2 \times 100\%$ equals the *percentage* of the observed variation in ***Y*** explained by *X*. Statgraphics displays R^2, as a percentage, beneath the *Analysis of Variance* table.

III The Plot of the Fitted Model

Statgraphics plots the scatterplot of the observations, along with the least-squares regression line and the prediction and confidence interval bands (see Section VI on estimation for a description of prediction and confidence intervals). The *Plot of the Fitted Model* for the Eugene example appears below.



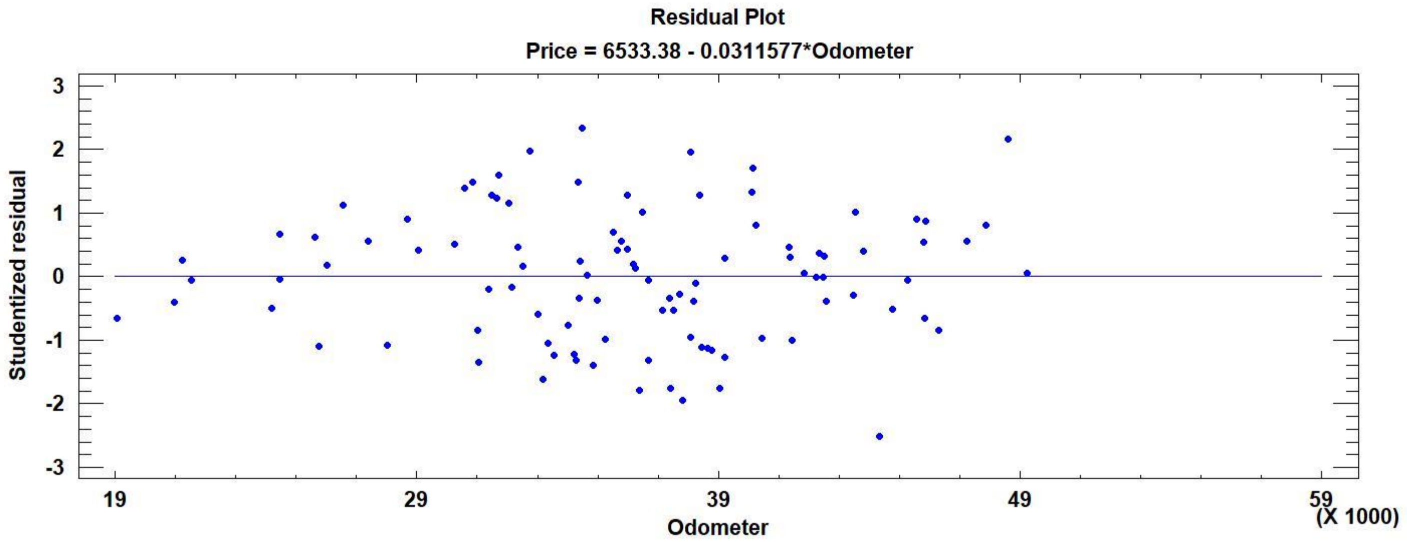
IV Checking the model assumptions: Residual Analysis

It is important to validate the model's assumptions about the error variable *prior* to testing the slope $\boldsymbol{\beta}_1$ or using the estimated regression line make predictions because both the hypothesis test of $\boldsymbol{\beta}_1$ and the interval (confidence and prediction) estimates in a regression analysis use the assumptions.

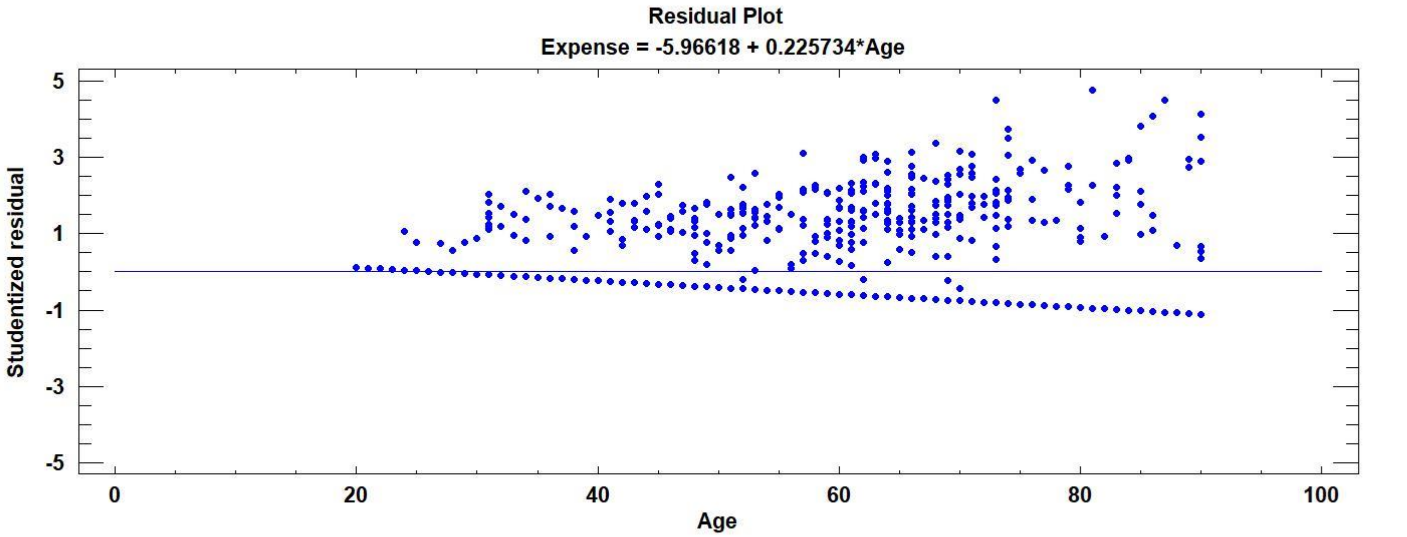
A. Constant Variance: the Plot of Residuals vs. Predicted Y

The assumption that $\boldsymbol{\varepsilon}$ has constant variance σ^2 can be checked visually by selecting the plot of *Residuals versus X* from the *Graphical Options* menu. If the spread of the residuals is roughly the same for all values of *x*, then the assumption is satisfied. If, however, there is a dramatic or systematic departure from constancy, then the assumption is violated and a remedial measure, such as a **transformation**, should be attempted before using the estimated regression equation. (Transformations are discussed later.)

Example 2: The file TAURUS contains the price (in dollars) and odometer reading (in miles) for 100 similarly equipped three-year-old Ford Tauruses sold at auction. Regressing price on odometer reading produces an acceptable *Residuals versus X* plot, shown below.



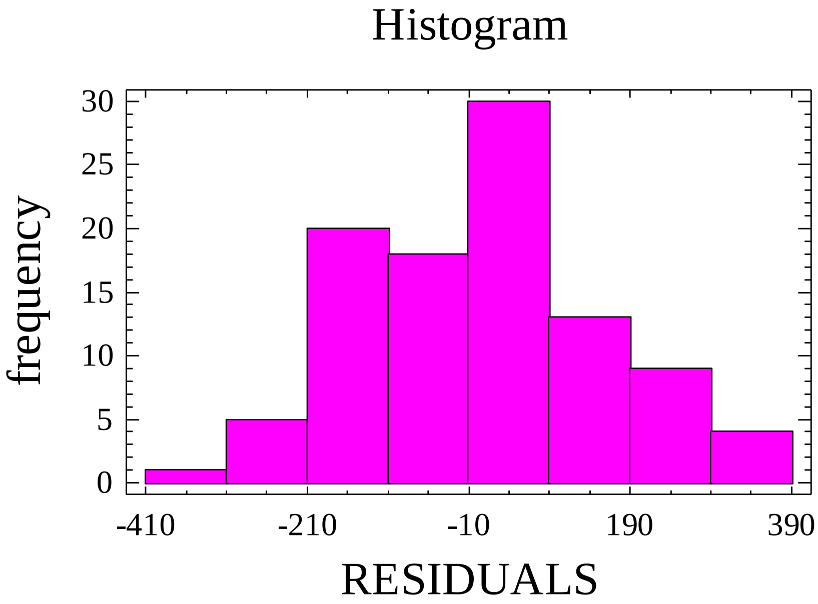
Example 3: The file CANADIAN HEALTH contains the age and mean-daily-health-expense for 1341 Canadians. The simple linear regression of mean-daily-health expense versus age produces an example of an unacceptable plot of *Residuals versus X*. Looking at the plot, you can see that as the Canadian's age increases (moving from left to right) the spread about the model also increases giving the graph a distinctive "fan" or "cone" shape. This is one of the most common forms that a violation of constant variance may take.



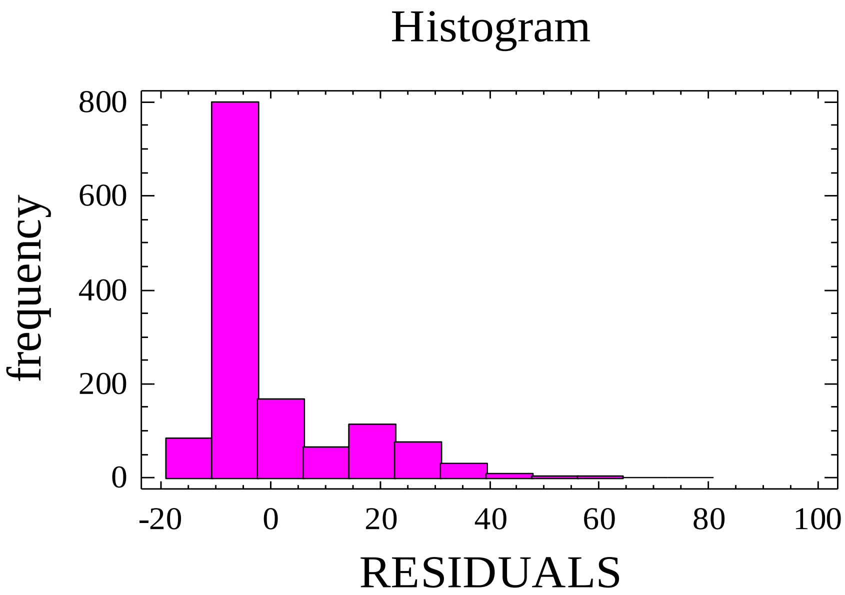
B. *Normality: Graphing a Histogram of the residuals*

The assumption that the error variable ε is normally distributed can be checked visually by graphing a histogram of the residuals. To create the histogram, first save the residuals using the *Save Results* button (fourth from the left in the region immediately to the right of the Navigation Toolbar). Then on Statgraphics' main toolbar select the *Statlets* > *Data Exploration* > *Interactive Histogram* option. If the histogram appears to be roughly bell shaped then the assumption is satisfied. If, however, it is strongly skewed then the assumption is violated and a remedial measure, such as a **transformation**, should be attempted before proceeding. (Selecting *Describe* > *Distributions* > *Distribution Fitting* (*Uncensored Data*), instead produces a histogram of the residuals with a normal curve superimposed; making the determination of normality easier.)

Example 2 (continued): The Taurus data below demonstrates an acceptably normal histogram of the residuals.



Example 3 (continued): The Canadian Health data below demonstrates an unacceptable histogram of the residuals (the histogram is strongly skewed to the right).



C. *Time-Series and Independent Errors: the Plot of Residuals vs. Row Number*

The assumption that the errors are independent of one another is often violated when regressing time-series data. Therefore, when modeling time-series data, you should look for patterns in the plot of *Residuals versus Row Number* selected from the *Graphical Options* menu. (Note: for time-series data the row number corresponds to the time period in which the values were collected.) A detailed discussion of this topic will be postponed until the notes on Multiple Linear Regression, where the regressing of time-series data may be considered if time permits.

V Influential Points and Outliers

A. Influence

An observation is said to be "influential" if the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ change markedly when the observation is removed and the least squares regression line is recalculated. This can be seen graphically by using the cursor to select the corresponding point on the *Plot of the Fitted Model* and then selecting the red and yellow **+/-** (include/exclude) button on the analysis toolbar.

Leverage is the *potential* an observation has to influence the slope of the least squares line. The further the *x* coordinate of the point is from $\bar{x}$ the more "leverage" the point has. It's useful to think of the line as a seesaw with the fulcrum at $\bar{x}$: the further the point is from the fulcrum, the more potential it has to "tilt" the line toward itself. The observations with the greatest leverage are listed in the *Influential Points* window under *Tabular Options*.

B. Outliers

1. Definition

An outlier is any point that does not seem to fit the overall pattern of the scatterplot. Any point that lies unusually far from the estimated regression line thus qualifies as an outlier. To determine potential outliers, Statgraphics computes the (deleted or externally) **Studentized Residual** t_i for every point on the scatterplot,

$$t_i = \frac{e_i - \mu_e}{s_{e_i}} = \frac{e_i}{s_i\sqrt{1 - \frac{1}{n} - \frac{\left(X_i - \bar{X}\right)^2}{\sum\left(X_j - \bar{X}\right)^2}}}$$ (because the mean of the residuals, μ_e, is always zero).

Discussion: In the equation for the *i*th studentized residual t_i , s_i is an estimate of σ_ε obtained from a model that excludes the *i*th observation. (e_i is still the ordinary residual based on the original model that included all *n* observations.) The purpose of excluding the *i*th observation prior to estimating σ_ε is to remove the effect it has on the model prior to calculating the number of standard errors (t_i) the observation lies from the *new* model. This prevents an outlier from influencing the model when calculating its studentized residual, rather like a judge recusing herself from a trial in which she has a personal interest, making it easier to identify the outlier. The studentized residual t_i has a *t*-distribution with $df = (n-1) - 2 = n - 3$ because *n* – 1 independent observations have been used to estimate the intercept and slope in a model that excluded the *i*th observation.

Statgraphics lists the row numbers of those observations whose studentized residuals have an absolute value greater than 2 in the *Unusual Residual* window under *Tables and Graphs*. These observations should be considered potential outliers.

Note: the standard error s_{e_i} used in studentizing residuals is *not* constant for all residuals, because the sample statistics $\hat{\beta}_0$ and $\hat{\beta}_1$ merely *estimate* the regression parameters β_0 and β_1 ! This has *absolutely nothing* to do with the assumption of constant variance of the errors in the model, but simply reflects the greater influence that some observations have upon the estimated parameters, especially the estimated slope $\hat{\beta}_1$.

Discussion: An outlier is any point on the scatterplot that is far removed from the bulk of the other points. Thus, a point with a large leverage value, because its *x*-coordinate is far from the average, may be an outlier even though it lies close to the regression line. Therefore, you should consider observations with either large deleted studentized residuals or large leverage to be potential outliers.

2. Sources

The most common sources of outlying observations are the following:

- A mismeasured or misreported value. For example, the value 39.4 is mistakenly entered as 394.
- The observation doesn't really belong to the population of interest. For example, a study of incomes in the software industry, using randomly selected employees, might include the income of one Bill Gates. However, as an owner of the company, as well as an employee, he may not be part of the population of interest in the study.
- The observation may represent a unique event that is not likely to be repeated. For instance, a study of retail sales in Sydney, Australia, might show a sharp spike in September 2000. A little research, however, would reveal that Sydney was hosting the Summer Olympics that month. The hosting of the Olympics is a rare event that is not likely to reoccur anytime soon.
- Finally, it is possible that the outlier is not the result of any of the above, in which case it may be considered a legitimate point for inclusion in the analysis. If this is the case then the observation has the potential to reveal important information about the dependent variable being studied, such as the nature of important independent variables not included in the regression model.

3. Remedies

If an outlier belongs to one of the first three categories above, then it may be appropriate to remove the observation from the data set before conducting the analysis. *However, one should never remove an observation from time-series data. We will discuss a technique for removing the* effect *of such an observation in time-series data when if cover time-series regression*.

VI Estimation

It is often of interest to be able to predict values of the dependent variable for given values of the independent variable. This may include the computation of point estimates, confidence interval estimates, or prediction interval estimates.

A. Point Estimates

The **predicted** or **fitted value** is obtained by substituting the required value of the independent variable into the estimated regression line. The result is a point that lies on the regression line with the specified *x*-value.

B. Confidence Intervals (See separate notes for a more detailed discussion)

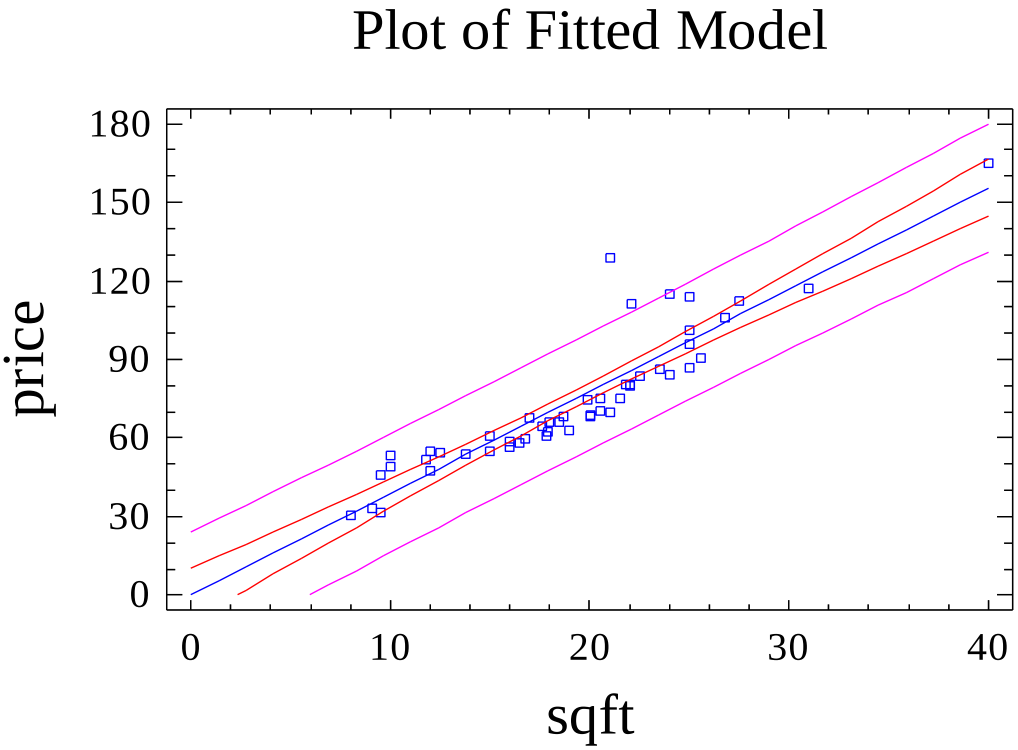
Confidence intervals play the same role in regression as they do in single variable statistics: they provide an interval with a specified likelihood (the coverage probability for the interval) of containing the true mean value of ***Y*** for the specified *x*-value. The width of the interval provides an indication of the accuracy of the estimate. A narrower interval indicates a more precise estimate than a wider interval bearing the same degree of confidence.

C. Prediction Intervals (See separate notes for a more detailed discussion)

While confidence intervals estimate the mean value of ***Y*** for a specified value of *x*, prediction intervals estimate an individual value of ***Y*** given the specified *x*-value. Remembering that there is more variability in individual values than in averaged values, it shouldn't surprise you to learn that prediction intervals are wider than confidence intervals. Both intervals, however, are centered about the point estimate for ***Y*** given the required value of *x*, i.e., both confidence and prediction intervals center on the estimated regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Confidence and Prediction "Bands"

In addition to constructing confidence and prediction intervals for $\boldsymbol{Y}$ given a specified x-value, we can construct curves, or "bands," which are functions of x. Note how the confidence band (orange curve) and prediction band (magenta curve) produced by StatGraphics below become wider as x differs more from $\bar{x}$ for the data. I'll give a heuristic reason for this in class, but it's because confidence and prediction intervals are more sensitive to errors in the estimation of the slope β_1 than to errors in the estimation of the intercept β_0. Far from the center of the scatterplot, i.e., far from $\bar{x}$, a small error in estimating the slope results in a greater error in estimating the conditional expectation $\mu_{Y|X=x} = \boldsymbol{E}\{\boldsymbol{Y} \mid X = x\}$. Thus, both types of intervals become wider to accommodate the additional uncertainty, causing the bands to bow out. **Note:** A $(1 - \alpha)$100% confidence band is designed to contain the true regression line $y = \beta_0 + \beta_1 x$ with probability $(1 - \alpha)$.



D. Using Statgraphics

To obtain predicted values, confidence intervals, and prediction intervals for a *particular* x-value select the *Forecasts* window under *Tables and Graphs*. Right click and select *Pane Options*. The *Forecast at X* dialog box lets you enter up to 10 values for the independent variable and/or change the level confidence from the default value of 95%.

Example 1 (continued): Forecasts for the price of a house (or the mean price of all houses) in Eugene with 1500 square feet, applicable to 1973, appear as follows:

```
Predicted Values
------------------------------------------------------------------------------
                                       95.00%                    95.00%
                 Predicted        Prediction Limits         Confidence Limits
            X            Y        Lower        Upper        Lower        Upper
------------------------------------------------------------------------------
         15.0      58.3519      36.1143      80.5894      54.6261      62.0776
------------------------------------------------------------------------------
```

Before using these forecasts, remember that all values are in the units for the variables *as they are presented in the data set,* i.e., X is in hundreds of square feet and Y is in thousands of dollars. Also, be careful not to report forecasted values that aren't possible in practice. Statgraphics isn't expected to know that a negative house price doesn't make sense as a lower limit in an interval estimate, but you are!

VII Transformations

While performing diagnostics for a particular regression model, you may discover serious violations of one or more of the error variable assumptions. It is natural to ask whether remedial measures can be taken that would allow us to use regression with more confidence. Here we explore a relatively simple remedy to the problems of non-constant variance and non-normality of the error variable. (Violations of the final assumption, row independent errors, are discussed in the notes on Multiple Linear Regression, where we consider time-series.)

Consider the original **specification** for the model, $\boldsymbol{Y} = \beta_0 + \beta_1 X$. This is the **Simple** Linear Regression model. This model may be inappropriate for either of two reasons: (1) a straight line may not provide the best model of the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and/or (2) the error variable assumptions may be violated for the model.

In case (1) the solution may involve specifying a **curvilinear** (curved) relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For instance, sales of a new product may increase over time, but at a decreasing rate, because of market saturation, the product's life cycle, etc. In this case, a polynomial model, such as $\boldsymbol{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$, or the logarithmic model $\boldsymbol{Y} = \beta_0 + \beta_1 \log(X)$ may be more appropriate. The latter is an example of a **transformation** of the **independent variable** $\boldsymbol{X}$, whereby X is replaced by $\log(X)$ in the model. If the **respecified** model fits the data better than the original model, we work with it instead. Although we will not discuss transformations of the independent variable X in detail in this course, we *will* consider polynomial models in the Multiple Linear Regression notes.

In case (2), where the assumptions about the error variable appear to be seriously violated, the solution may involve a **transformation** of the **dependent variable**, $\boldsymbol{Y}$. The transformation involves replacing $\boldsymbol{Y}$ in the model with some simple function of $\boldsymbol{Y}$. Although many transformations are possible, the most popular involve replacing $\boldsymbol{Y}$ with $\log(\boldsymbol{Y})$, $\boldsymbol{Y}^2$, $\sqrt{\boldsymbol{Y}}$, or $^1\!/_{\boldsymbol{Y}}$. Below I have provided the Statgraphics format for each of the four that must be entered into the **dependent** variable field of the input dialog box:

- Log: use LOG(variable) {**Note**: this is the *natural* logarithm, base *e*}
- Square: use variable^2
- Square-Root: use SQRT(variable)
- Reciprocal: use 1/variable

You can either type the appropriate transformation directly into the dependent variable field, or use the TRANSFORM button at the bottom of the input dialog box and use the built-in keypad and operators. Although it is not always clear which, if any, of the four transformations listed above will improve the model, some general guidelines are provide below.

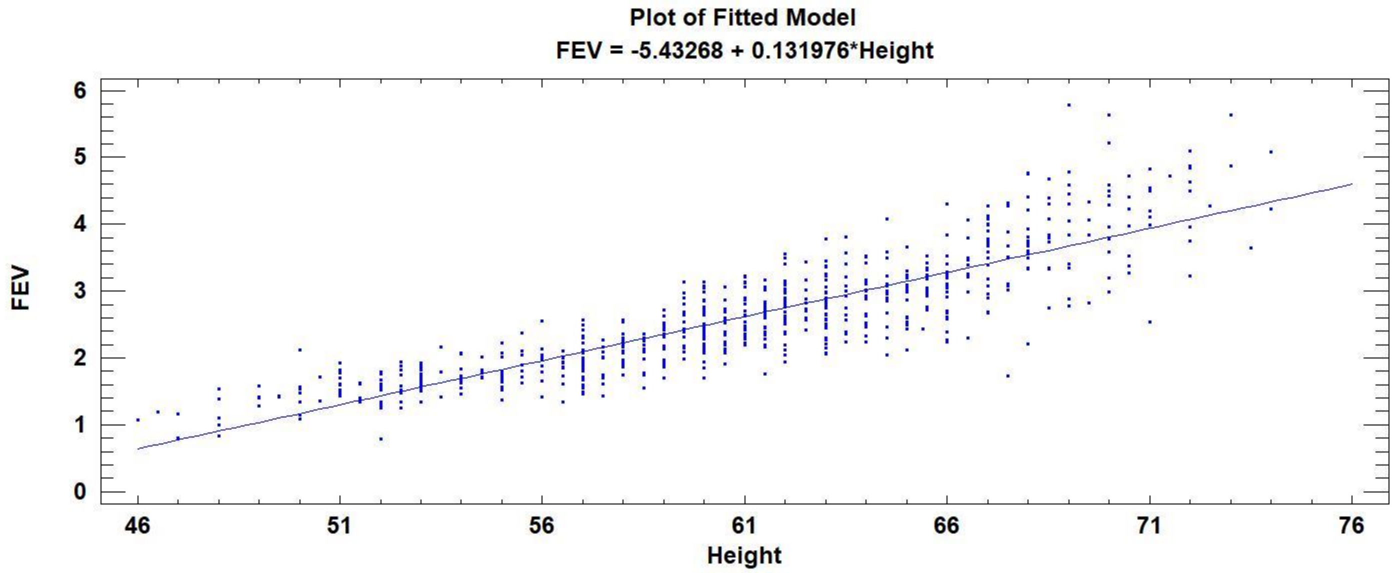
- Log: use LOG(variable) – may be useful, provided y > 0, when the variance increases as $\hat{y}$ increases or the distribution of ε is skewed to the right.
- Square: use variable^2 – may be useful when the variance is proportional to $\hat{y}$ or the distribution of ε is skewed to the left.
- Square-Root: use SQRT(variable) – may be useful, provided y > 0, when the variance is proportional to $\hat{y}$.
- Reciprocal: use 1/variable – sometimes useful when the variance increases significantly beyond some particular value of $\hat{y}$.

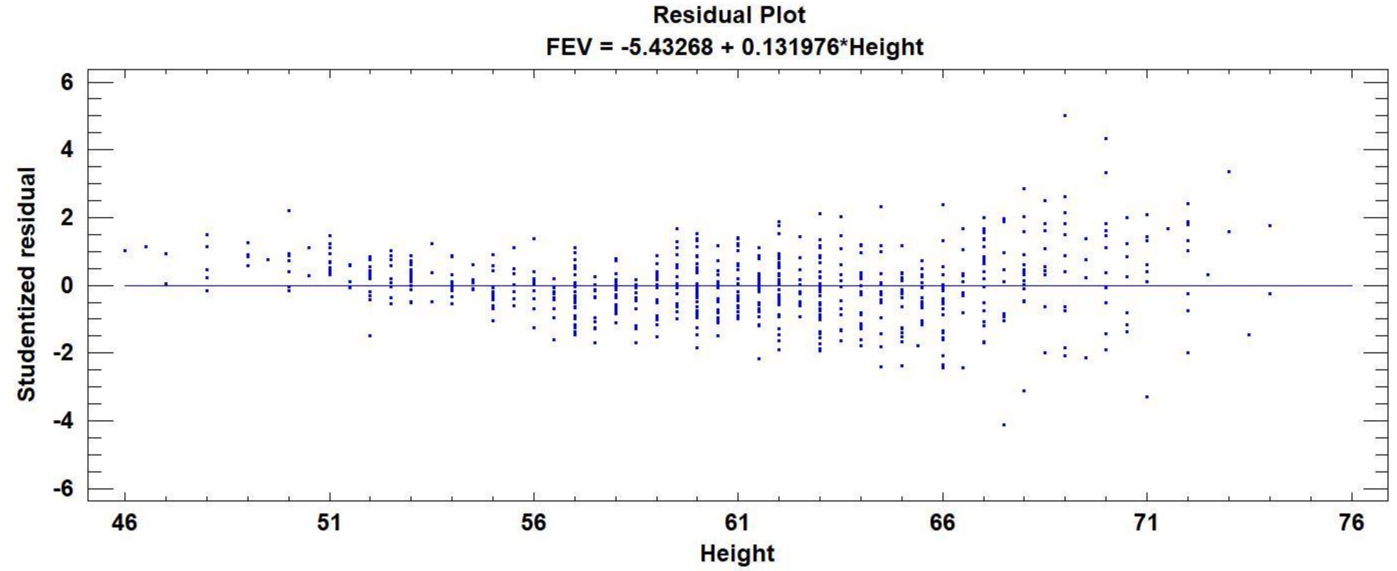
Example 4: The file FEV contains data collected from children for the following variables:

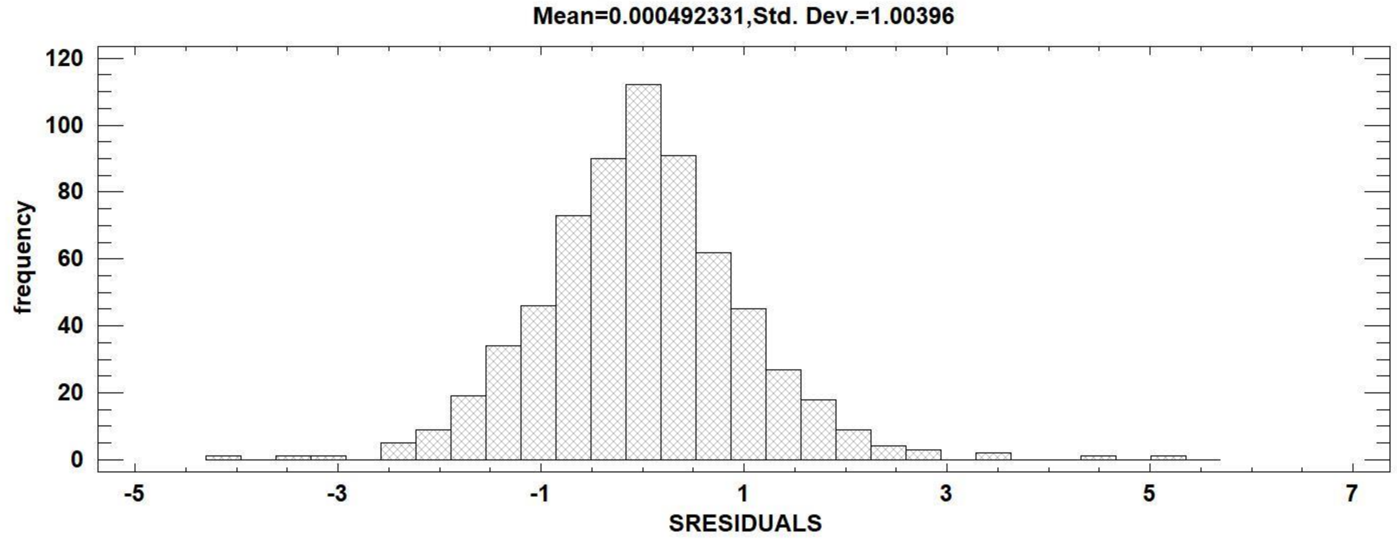
- ❑ FEV - Forced Expiratory Volume (in liters) is a measure of the child's lung capacity.
- ❑ Age - the child's age (in years)
- ❑ Height - the child's height (in inches)
- ❑ Sex – male (0) or female (1)
- ❑ Status – nonsmoker (0) or smoker (1)
- ❑ ID – not really a variable, the ID number identifies the child

Suppose that we wish to use a child's height to predict forced expiratory volume. (For instance, a child whose forced expiratory volume, as measured by appropriate instruments, is significantly below the predicted FEV for their height may qualify for a referral to a respiratory specialist.) Using FEV as the response variable, we obtain the following Statgraphics' output.

```
Dependent variable: FEV
Independent variable: Height
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept       -5.43268        0.18146       -29.9387         0.0000
Slope           0.131976     0.00295496        44.6624         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     369.986      1      369.986    1994.73       0.0000
Residual                  120.934    652     0.185482
-----------------------------------------------------------------------------
Total (Corr.)              490.92    653

Correlation Coefficient = 0.868135
R-squared = 75.3658 percent
R-squared (adjusted for d.f.) = 75.3281 percent
Standard Error of Est. = 0.430676
Mean absolute error = 0.323581
Durbin-Watson statistic = 1.60845 (P=0.0000)
Lag 1 residual autocorrelation = 0.195101
```

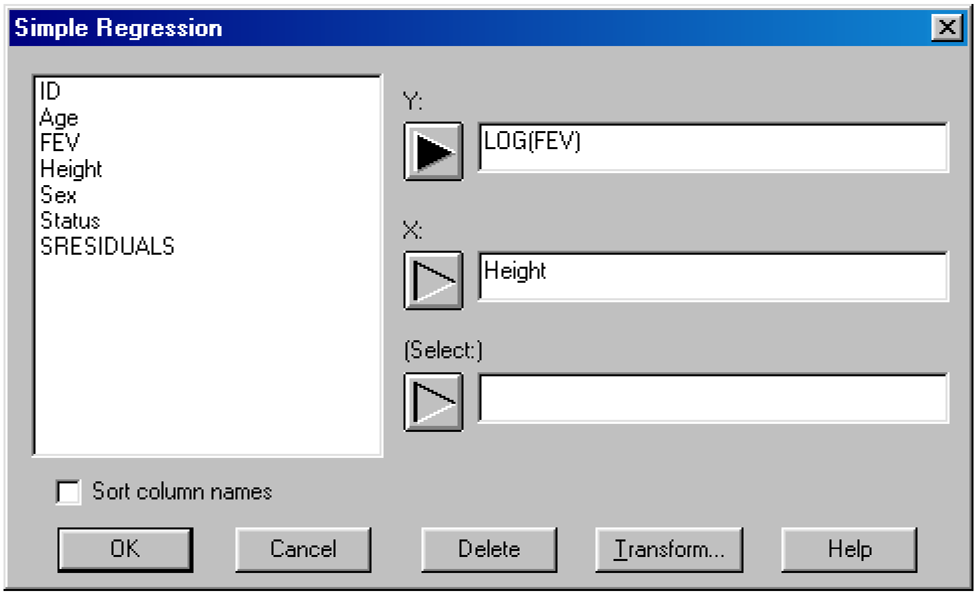






Although the histogram is not particularly skewed, there are several outliers with large studentized residuals. More dramatic evidence of problems, however, comes from the scatterplot (which displays curvature) and the residual plot (which shows variance increasing with increasing predicted FEV). Trying a logarithmic transformation on FEV (see the input dialog box below for details), we obtain the new model log(FEV) =

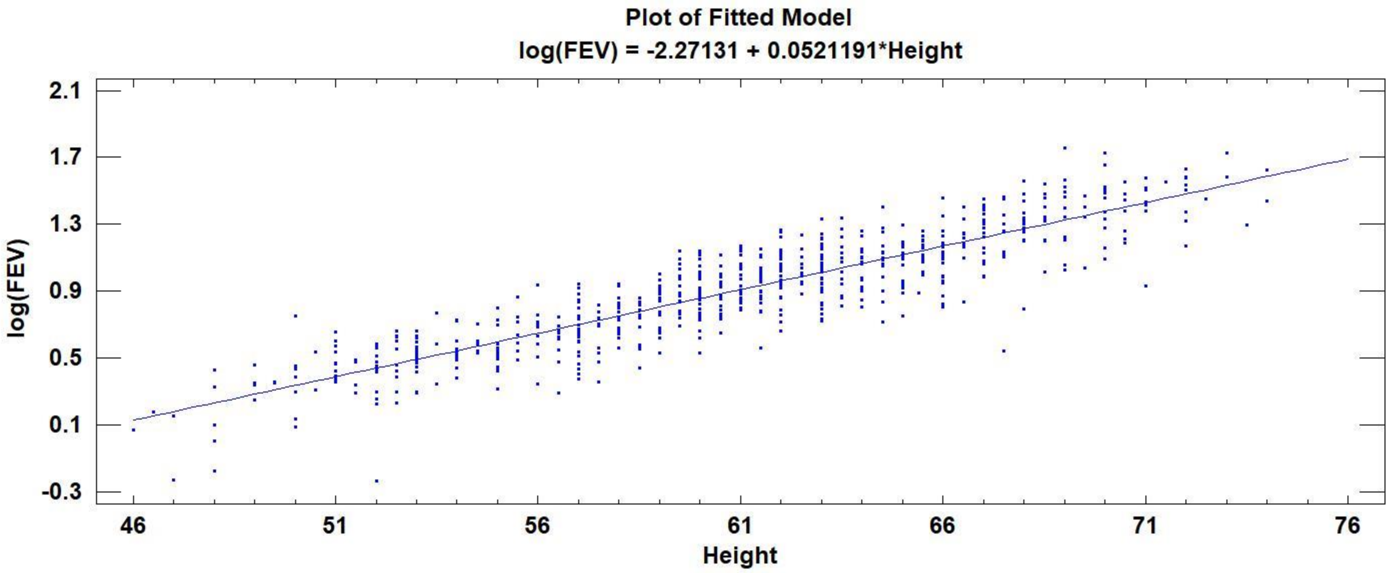
$\beta_0 + \beta_1 Height + \varepsilon$, which produces output more consistent with the regression assumptions (see output on the following pages).

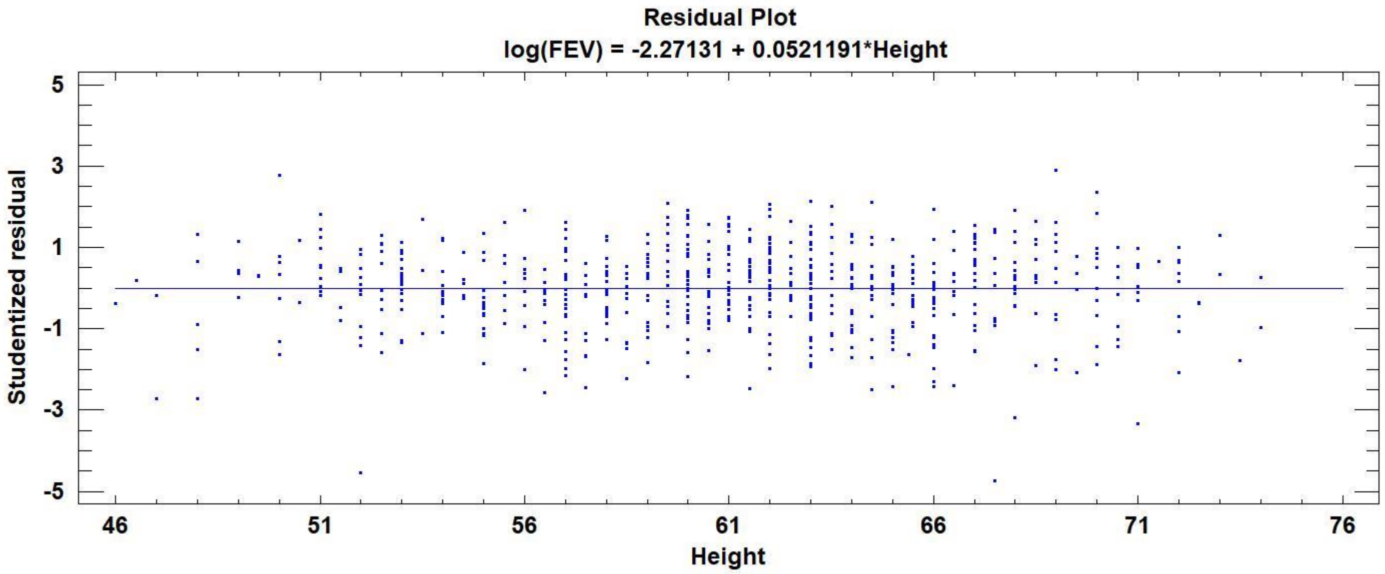


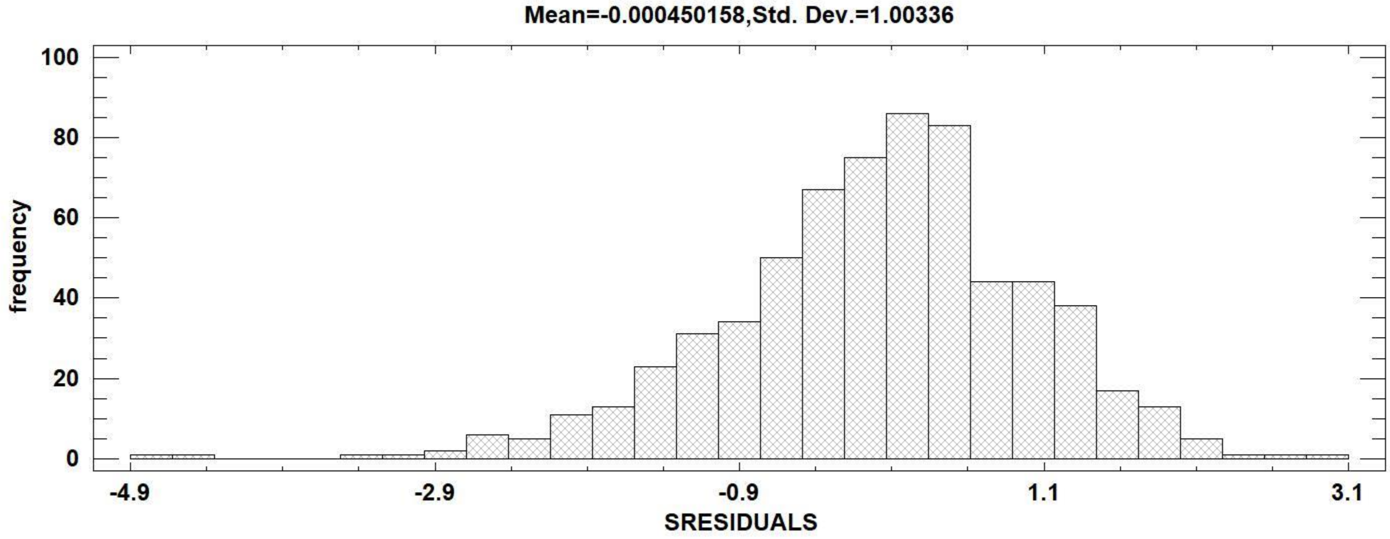
Y: LOG(FEV)

X: Height

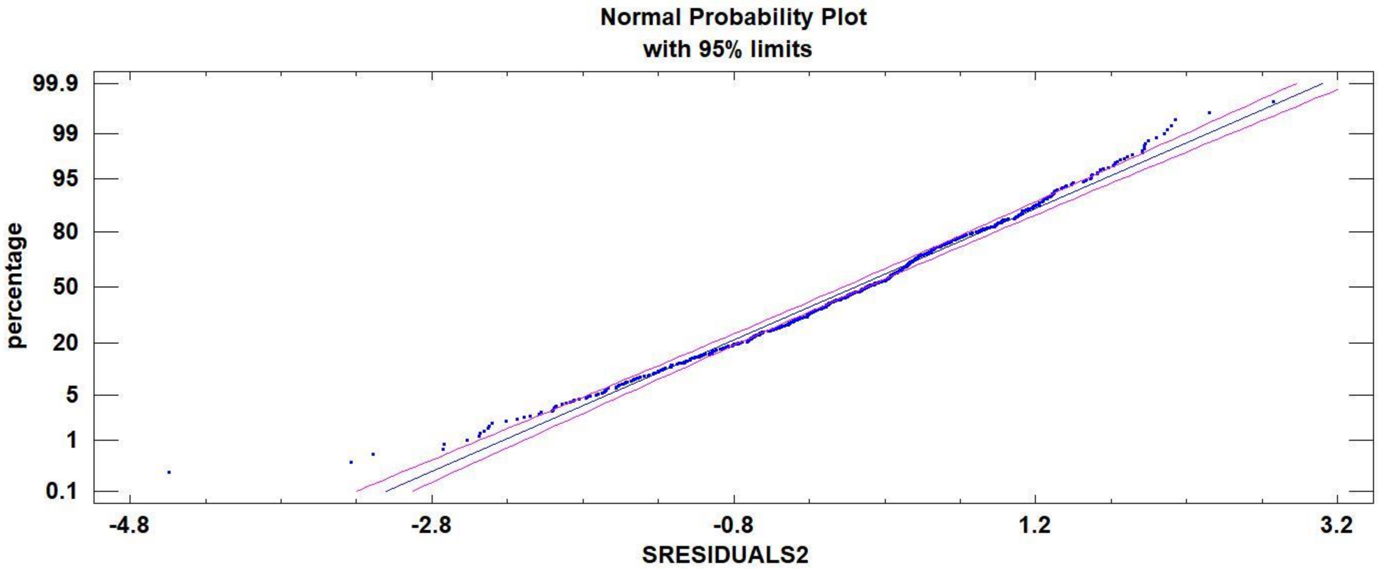
```
Dependent variable: LOG(FEV)
Independent variable: Height
-----------------------------------------------------------------------------
                                Standard          T
Parameter       Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
Intercept       -2.27131        0.063531      -35.7512         0.0000
Slope          0.0521191      0.00103456       50.3779         0.0000
-----------------------------------------------------------------------------


                           Analysis of Variance
-----------------------------------------------------------------------------
Source                   Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                          57.7021       1      57.7021    2537.94       0.0000
Residual                       14.8238     652    0.0227358
-----------------------------------------------------------------------------
Total (Corr.)                  72.5259     653

Correlation Coefficient = 0.891968
R-squared = 79.5607 percent
R-squared (adjusted for d.f.) = 79.5294 percent
Standard Error of Est. = 0.150784
Mean absolute error = 0.116218
Durbin-Watson statistic = 1.57644 (P=0.0000)
Lag 1 residual autocorrelation = 0.210846
```


Plot of Fitted Model
log(FEV) = -2.27131 + 0.0521191*Height
log(FEV)
2.1
1.7
1.3
0.9
0.5
0.1
-0.3
46
51
56
61
66
71
76
Height


Residual Plot
log(FEV) = -2.27131 + 0.0521191*Height
Studentized residual
5
3
1
-1
-3
-5
46
51
56
61
66
71
76
Height


Mean=-0.000450158,Std. Dev.=1.00336
frequency
100
80
60
40
20
0
-4.9
-2.9
-0.9
1.1
3.1
SRESIDUALS

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.



Discussion: The natural log transformation of the response variable FEV produce more nearly constant variance, a more linear fit, and a slightly higher R^2, but the normality assumption is still violated. The latter conclusion is suggested by the normal probability plot above (graphed after saving the studentized residuals for the new model), supported by the Summary Statistics that accompany the plot, and confirmed by StatAdvisor. The sample distribution of the residuals is left-skewed with substantial outliers in the left tail. For large samples the normality assumption may be relaxed if the model is otherwise satisfactory.

- **Note 1:** If a transformation of ***Y*** is employed, then all predictions pertain to the transformed variable. For instance, if we wish to predict the FEV for a child who is 60" tall we first use Statgraphics to predict log(FEV) for the child, and then raise the natural number *e* to that value (thereby "undoing" the log).

```
                                       95.00%                    95.00%
                  Predicted        Prediction Limits         Confidence Limits
            X             Y        Lower         Upper        Lower         Upper
----------------------------------------------------------------------------------
         60.0      0.855835     0.560068        1.1516     0.844048      0.867621
----------------------------------------------------------------------------------
```

In this example, the predicted FEV for the child is $e^{0.855835}$ = **2.353** liters.

- **Note 2:** Sometimes transformations of both *X* and ***Y*** may be necessary.

VIII Cautionary Note: Inferring Cause and Effect

“*Correlation does not imply causation.*”

There is a strong positive correlation between a child’s shoe size and the size of his or her vocabulary. Does this mean that you should buy Junior bigger shoes? Probably not. Both shoe size and vocabulary are correlated to the child’s age, and age is the determinative variable for both a child’s shoe size and vocabulary. (A variable that is not part of the analysis, but that drives both the Predictor and Response variables, is called a Lurking Variable (like the man behind the curtain in the *Wizard of Oz*).)

A more serious example is the correlation between cigarette smoking (either number of years or number of packs per day) and health variables such as life expectancy. The association between smoking cigarettes and negative health outcomes had been documented for decades, but the tobacco lobby successfully argued that there might be lurking variables driving both. Perhaps people in high stress occupations were more likely to take up smoking to relieve the stress, for example. Then it might be the stress, rather than smoking, that was reducing life expectancy. (**Note:** In this scenario, smoking is still viewed as a predictor of future negative health consequences, but we are unable to draw a causal inference from smoking to poor health.)

Usually, the only way to establish a **cause and effect** relationship between two variables is through a designed experiment, where levels of the independent variable (different dosages of a drug, for example) are randomly applied to subjects, and the corresponding value of the dependent variable for the subjects (life expectancy, perhaps) are recorded. By randomizing the treatments, the effects of lurking variables are reduced or eliminated (rather like shuffling a deck of cards to “randomize” it), making it possible to attribute any systematic changes in the dependent variable in response to different levels of the independent variable to cause and effect. In the case of tobacco, long-term laboratory experiments involving monkeys and chimpanzees, conducted in the 1950’s and 1960’s, were required to help establish beyond a reasonable doubt that tobacco products were a contributing factor to a number of adverse health outcomes.

Cause and effect can also be obscured by the presence of a Confounding Variable. The distinction between lurking and confounding variables is subtle, and confounding variables can only appear in a multivariate setting such as Multiple Regression and Two-Way ANOVA, so a discussion of confounding variables will be postponed, but both lurking and confounding variables can be controlled in a designed experiment.

IX Summary

The order in which the material has been presented in these notes is traditional. In a practical application, the residual analysis would be conducted earlier in the process. It is appropriate to investigate violations of the required conditions when the model is assessed and before using the regression equation to forecast. The following steps describe the entire process.

1. **Develop a model that has a theoretical basis. That is, for the dependent variable of interest find an independent variable that you believe is linearly related to it.**

2. **Gather data for the two variables. Ideally, conduct a controlled experiment that will allow you to control for lurking/confounding variables and/or establish causation. If that is not possible, collect observational data.**

3. **Begin a Simple Regression analysis and look at the *Plot of the Fitted Model* to see if the scatterplot supports the conclusion that the two variables are correlated (linearly related).**

4. **Determine the regression equation.**

5. **Save the residuals and check the required conditions.**

    - Is the error variable normal?
    - Is the variance constant?
    - Are the errors independent (applicable to time-series)?
    - Check the outliers and influential observations, and investigate them if necessary.

6. **Assess the model's fit.**

    - Compute the standard error of estimate s_ε.
    - Test the slope β_1 to determine whether X and Y are correlated.
    - Compute R^2 and interpret it.

7. **If the model fits the data, use it to predict a particular value of the dependent variable, or to estimate its mean.**