Multiple Linear Regression

I Introduction

In simple linear regression, ε represented the combined effect upon the dependent variable of all the independent variables not included in the model. It seems only natural, therefore, that we might wish to add variables we think significant to the model with the goal of reducing the (unexplained) random variation in the dependent variable, i.e., producing a better fitting model.

II The Model

The model for multiple linear regression is given by
$\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \boldsymbol{\varepsilon}$, where

- k equals the number of independent variables in the model
- X_i is the i^{th} independent variable (out of k)
- $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$ are random variables
- $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters

III Assumptions

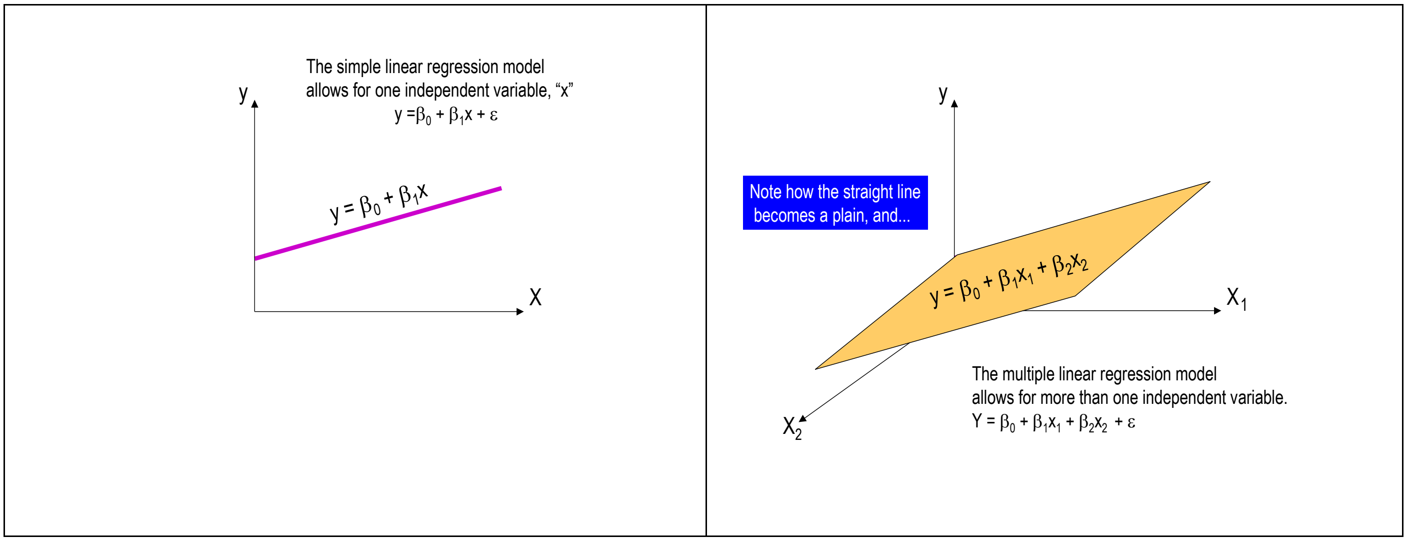
The Multiple Linear Regression model, $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \boldsymbol{\varepsilon}$, makes two different kinds of assumptions.

A. Linearity

- The first of these, mentioned previously, postulates that the dependent variable $\boldsymbol{Y}$ is linearly related to the *collection* of k independent variables *taken together*.

The **Simple Linear Regression** model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \boldsymbol{\varepsilon}$ hypothesizes that the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ can best be described as variation about a line. Similarly, the **Quadratic** (Polynomial) model $\boldsymbol{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \boldsymbol{\varepsilon}$ hypothesizes that the conditional distribution follows a parabola. The **Multiple Linear Regression** model $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \boldsymbol{\varepsilon}$ hypothesizes that the conditional distribution of the dependent variable $\boldsymbol{Y}$ given the independent variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can best be described as variation about a **plane**. The picture below right shows such a plane. This model may be suggested by experience, theoretical considerations, or exploratory data analysis. (For comparison, the picture below left is of a simple linear regression model.)

Note: Sets of variables, such as $\boldsymbol{Y}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, that are related through probability follow a Joint Probability Distribution. Then $\boldsymbol{Y}$, given fixed values of the independent variables, has a probability distribution conditioned on those fixed values. This conditional probability distribution has a mean, variance and may belong to a common family of distributions. In simple linear regression, if the assumptions of the model are valid, then the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X} = x$ is $N\left(\beta_0 + \beta_1 x, \sigma^2\right)$.


The simple linear regression model
allows for one independent variable, "x"
y =β0 + β1x + ε
y
y = β0 + β1x
X
Note how the straight line becomes a plain, and...
y
y = β0 + β1x1 + β2x2
X1
X2
The multiple linear regression model
allows for more than one independent variable.
Y = β0 + β1x1 + β2x2 + ε

The more general multiple regression model considered here, $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \boldsymbol{\varepsilon}$, is also thought of as defined by a plane, but for $k > 2$ we aren't able to picture the plane described by the model.

B. The Error Variable

- 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 $\boldsymbol{\mu_\varepsilon} = 0$, and constant variance $\sigma_{\boldsymbol{\varepsilon}}^2$. Note: They often drop the subscript and write simply σ^2, as in $\boldsymbol{\varepsilon} \sim N\left(0,\sigma^2\right)$.
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 $\boldsymbol{Y}$**,** for fixed values of the independent variables, 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 $\boldsymbol{Y}$ about the plane $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the same for all values of the independent variables observed.

Recall that the assumptions made in 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.

Example 1: Securicorp markets security equipment to airports. Management wishes to evaluate the effectiveness of a new program that offers performance bonuses to salespeople, while also taking into account the effect of advertising. The company currently markets in the West, Midwest, and South. The regions are further divided into smaller sales territories. The file SECURICORP contains data from last year for each sales territory for the following variables:

- ***Sales***: Sales, in thousands of dollars
- *Ad*: Advertising, in hundreds of dollars
- *Bonus*: Bonuses, in hundreds of dollars
- *Region*: The region to which the sales territory belongs

Selecting Relate > Multiple Factors > Multiple Regression from the menus, entering ***Sales*** as the dependent variable, and *Ad* and *Bonus* as the independent variables, produces the *Analysis Summary* window below.

```
Multiple Regression - Sales
Dependent variable: Sales
Independent variables:
     Ad
     Bonus
Number of observations: 25
```

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-515.073	190.759	-2.70013	0.0131
Ad	2.47216	0.275644	8.96869	0.0000
Bonus	1.85284	0.717485	2.5824	0.0170

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.06722E6	2	533609.	64.27	0.0000
Residual	182665.	22	8302.95		
Total (Corr.)	1.24988E6	24			

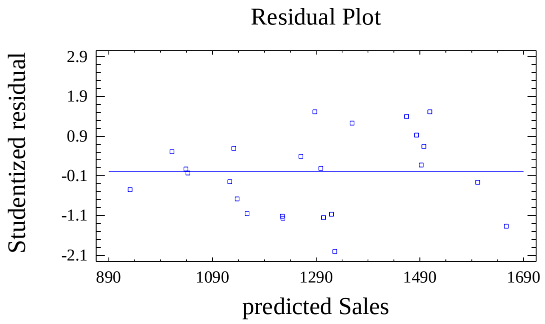
R-squared = 85.3854 percent
R-squared (adjusted for d.f.) = 84.0568 percent
Standard Error of Est. = 91.1205

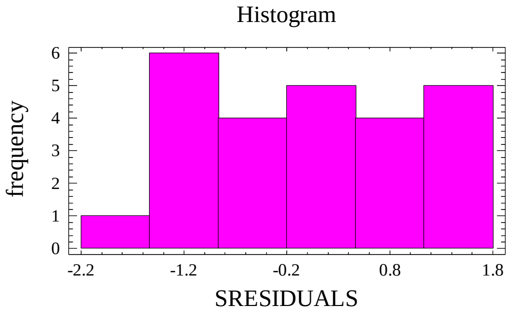
IV Checking the Error Variable Assumptions: Residual Analysis

The assumptions made about the distribution of the error variable can be checked by looking at the Plot of Residuals vs. Predicted ***Y*** and a histogram of the studentized residuals or a normal probability plot, as in simple regression. In addition, for time-series data the Plot of Residuals vs. Row Number (the row numbers represent the time periods in which the data was collected) should be free of any obvious patterns that would suggest that errors were correlated.

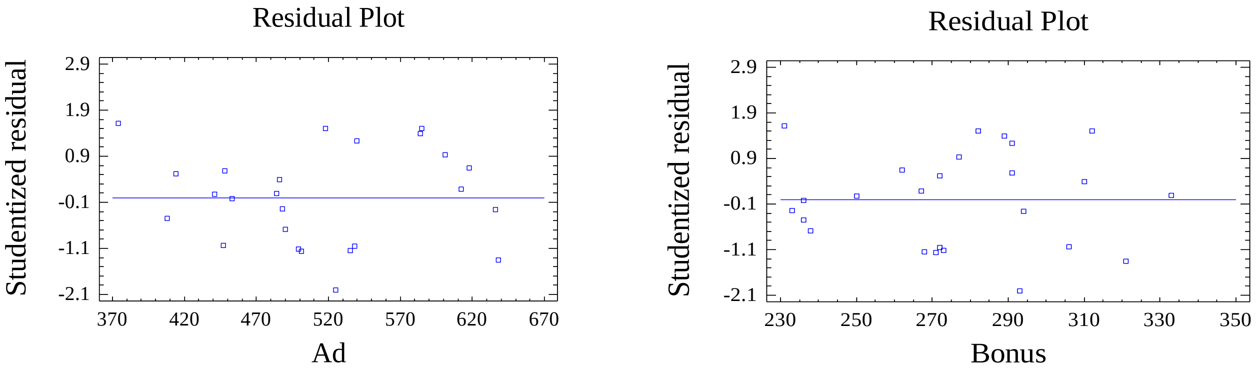
For Multiple Regression, it is also advisable to look at the *Plot of Residuals versus X* for each independent variable in the model. (From within the *Plot of Residuals versus X* window, use *Pane Options* to select independent variables for analysis.) As with other plots of residuals, we expect the residual plots for each independent variable to be random (free of obvious patterns) and to have a constant spread for all values of the independent variable.

Example 1 (continued): For Securicorp sales regressed on advertizing and bonuses, the studentized residuals plotted against the predicted sales, and the histogram of the studentized residuals appear below.

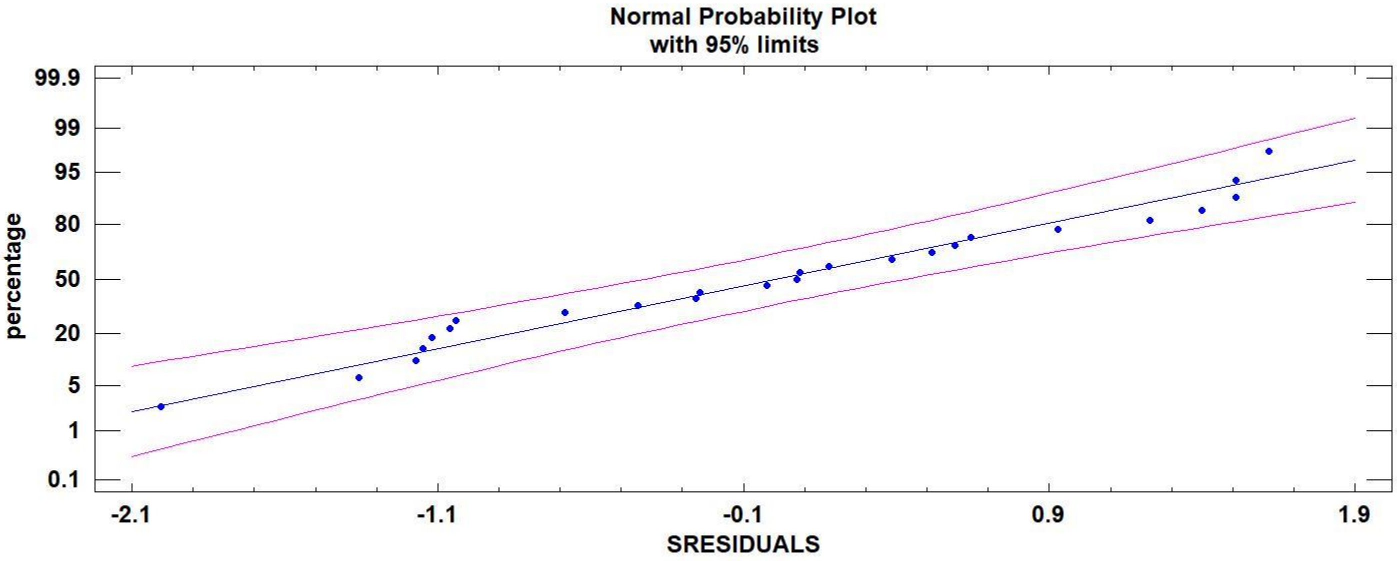

Residual Plot


Histogram

The *Plot of Residuals versus X* for advertising and bonus appear below. The plot for advertising is the default graph because the variable *Ad* appears as the first variable in our model. To view the plot for the variable *Bonus*, select *Pane Options* with the right mouse button and click on the variable Bonus.



Note that all of the residual plots appear plausibly random, and exhibit constant spread about the model (represented by the horizontal line in the plots). The histogram doesn't appear to be normal, but rather uniform (constant for most values on the studentized residual. Although we could play with the histogram by changing the number of classes and their boundaries to see if the pattern persists, I'll accept the model based on the plausible randomness of the residual plots, the results of the normal probability plot (shown below), and standardized skewness and kurtosis scores of -0.085 and -0.976 consistent with normally distributed errors.



V The Analysis of Variance (ANOVA) Table

As in simple regression, an ANOVA table is prominently featured in the output of the *Analysis Summary* window. Below is a description of the contents of the columns in the table.

A. Sums of Squares

The definition and interpretation of the sums of squares in multiple regression is similar to that in simple regression.

Total Sum of Squares, $SST = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$, is a measure of the total observed variation of $\boldsymbol{Y}$

Regression Sum of Squares, $SSR = \sum_{1}^{n} \left(\hat{y}_i - \bar{y} \right)^2$, measures the observed variation of ***Y*** **explained** by the model

Error Sum of Squares, $SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$, measures the observed variation of ***Y*** **unexplained** by the model

Note: Remarkably, we find that the equality *SST* = *SSR* + *SSE* always holds. Thus, the total variation in ***Y*** can be "decomposed" or "resolved" into the explained variation plus the unexplained variation for the model

B. Degrees of Freedom

The degrees of freedom *df* equal the amount of independent information available for the corresponding sum of squares. Starting with *n* degrees of freedom (one for each observation in the sample), we lose one degree of freedom for each parameter β in the model estimated by a sample statistic $\hat{\beta}$.

SST: *df* = *n* – 1, because the parameter $\mu_{\boldsymbol{Y}}$ is estimated by the sample statistic $\bar{\boldsymbol{Y}}$.

SSE: *df* = *n* – *k* – 1 = *n* – (*k* + 1), because the *k* + 1 model parameters, $\beta_0, \ldots, \beta_{\mathrm{k}}$, must be estimated by the sample statistics $\hat{\beta}_0, \ldots, \hat{\beta}_{\mathrm{k}}$.

SSR: *df* = *k*

Note: Paralleling the results from simple regression, observe that the **total** degrees of freedom *n* – 1 "decomposes" into the **explained** degrees of freedom *k* plus the **unexplained** degrees of freedom *n* - *k* – 1. We will discuss this in more detail when we cover the linear algebra behind linear statistical models.

C. Mean Squares

While a sum of squares measures the total variation, the ratio of a sum of squares to its degrees of freedom is a **variance**. The advantage of a variance is that it can be used to compare different data sets and different models (since it incorporates information about the size of the sample and the number of independent variables used in the model). These variances are called **mean squares**.

$\mathbf{S}_{\boldsymbol{Y}}^2 = \dfrac{SST}{n-1} = \dfrac{\sum_{\boldsymbol{i}=1}^{n} \left(y_{\boldsymbol{i}} - \bar{y} \right)^2}{n-1}$ = the (sample) variance of the *y*-values observed (see the notes "Review of Basic Statistical Concepts"). You probably never knew (or cared, perhaps) that the sample variance you computed in your introductory statistics course was an example of a mean square!

$MSE = \dfrac{SSE}{n-k-1} = \dfrac{\sum_{\boldsymbol{i}=1}^{n} \left(y_{\boldsymbol{i}} - \hat{y}_{\boldsymbol{i}} \right)^2}{n-k-1}$ = the (sample) variance of the dependent variable unexplained by the model.

The *M*ean *S*quare *E*rror is the sample estimate of the variance of the error variable, $\boldsymbol{\sigma}_{\varepsilon}^2$.

$MSR = \dfrac{SSR}{k} = \dfrac{\sum_{\boldsymbol{i}=1}^{n} \left(\hat{y}_{\boldsymbol{i}} - \bar{y} \right)^2}{k}$ = the variance of the dependent variable explained by the model, the *M*ean *S*quare for *R*egression.

D. The F – Ratio

An intuitive device for judging the effectiveness of the model in describing the relationship between the dependent variable ***Y*** and the independent variables in the model, taken together, is to compute the **ratio** of the *MSR* (the variance in ***Y*** explained by the model) to the *MSE* (the variance of ***Y*** about the model). The resulting ratio is called the *F-Ratio* (or *F* statistic when used to conduct hypothesis tests within a model):

$$F - Ratio = \frac{MSR}{MSE}$$

Properties of the *F* statistic:

- $F > 0$
- If *F* is "large" then the model explains much more of the variation in ***Y*** then it leaves unexplained, which is evidence that the model $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ may be appropriate, i.e., a large *F* supports the *linearity* assumption of the model.
- A "small" *F* for a model indicates that the model is inappropriate, while a small *F* for an independent variable (or set of independent variables) suggests the variable (or set of variables) may be removed without affecting the model's ability to predict.

E. P-Value

If the error variable $\boldsymbol{\varepsilon}$ is normally distributed with constant variance, then the *F – Ratio* for the model follows a probability distribution called the *F* distribution. The *P*-value in the ANOVA table is for a hypothesis test of the linearity assumption of the model, i.e., the assumption that the dependent variable is linearly related to the *set* of independent variables, *taken together*, $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. See the next section (Section **VI**) for a discussion of the test. (The F distribution has two parameters for degrees of freedom, df_1 and df_2. In the test conducted for the model, $df_1 = k$ and $df_2 = n - k - 1$.)

F. Summary

The ANOVA Table below summarizes the results in this section as they appear in Statgraphics, where *k* is the number of independent variables in the model.

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

*
*
* Note: The sample variance for Y, s^2 = SST/(n-1), is an example of a Mean Squa
*       is not computed by Statgraphics in the ANOVA Table.
*
*
```

Below is the ANOVA Table for the Securicorp example. (Note: The E6 in SSR and SST is scientific notation for 10^6. Thus, SSR = 1,067,220, and SST = 1,249,880.) Remember, also, that the sales figures appearing in the spreadsheet are in units of $1,000, and that these units are *squared* in the computation of the sums of squares!

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.06722E6	2	533609.	64.27	0.0000
Residual	182665.	22	8302.95		
Total (Corr.)	1.24988E6	24			

VI Testing the Assumption of Linearity

Is $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ an appropriate description of the relationship between the dependent and independent variables? To answer this question, we conduct a formal hypothesis test. For the test,

- H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, i.e., *none* of the independent variables are linearly related to the dependent variable.
- H_A: At least one β_i is not zero, i.e., *at least one* of the independent variables is linearly related to the dependent variable.
- Test Statistic: $F = \dfrac{MSR}{MSE}$ = the ratio of the explained and unexplained variance in the observed response.
- *P*-value: If the error variable satisfies the assumption made in section **III** then *F* follows an *F* distribution. Using the *F* distribution, Statgraphics computes the *P*-value for the test statistic. Note, however, that substantial deviations from the error variable assumptions can make the *P*-value unreliable. Since larger *F – Ratios* correspond to more convincing evidence that at least one of the independent variables is correlated to ***Y***, large values of *F* lead to small *P*-values and the rejection of H_0.

Example 1 (continued): Based upon the *P*-value of 0**.**0000 for the *F – Ratio* in the ANOVA Table below for Securicorp, we reject the hypothesis that neither *Ad* nor *Bonus* is linearly related to *Sales*.

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.06722E6	2	533609.	64.27	0.0000
Residual	182665.	22	8302.95		
Total (Corr.)	1.24988E6	24			

VII Testing the Importance of Individual Variables to the Model

Having established, via the *F***-**test, that the *k* independent variables, taken together, are correlated to ***Y***, we next ask which individual independent variables belong in the model. This involves conducting a *t*-test of the slope for each of the *k* independent variables in the model.

Statgraphics determines the utility of an independent variable by considering how much the variable *improves* the model if it is the *last one to enter*. The test statistic and *P*-value for the test, presented in the *Analysis Summary* window in the same row as the estimated slope for the variable, can be used to determine the importance of the variable in explaining the variation in ***Y*** *after accounting for the effects of the other variables in the model*. Thus, the *t*–test measures the *marginal* improvement the variable affords the model.

Because results of the individual *t*-tests depend upon the presence of the other variables in the model, each time you add or remove a variable from the model all of the test statistics and *P*-values will change.

Example 1 (continued): Based upon the *P*-values of 0.0000 for the variable *Ad* and 0.0170 for the variable *Bonus* in the ANOVA Table for Securicorp shown on page 4, both independent variables are linearly related to *Sales* in this model, and are therefore retained.

VIII Multicollinearity

Example 2: A real estate agent believes that the selling price of a house can be predicted using the number of bedrooms, the size of the house, and the size of the lot upon which the house sits. A random sample of 100 houses was drawn and the data recorded in the file HOUSE PRICE for the variables below.

- ***Price***: Price, in dollars
- *Bedrooms*: The number of bedrooms
- *H_Size*: House size, in square feet
- *Lot_Size*: Lot size, also in square feet

A. What is Multicollinearity?

Looking at the House Price data, we immediately note that if the variables *H_Size*, *Lot_Size*, and *Bedrooms* are all included in the model, their individual *P*-values are all high (see the output below). This might lead us to conclude that none of them are correlated to the price of a house, but the *P*-value for the model assures us that *at least one* of them is correlated to house price. (In fact, doing simple regressions of *Price* on the three independent variables, taken one at a time, leads to the conclusion that all of them are correlated to house price. You should verify these results.) These seemingly contradictory results are explained by the existence of Multicollinearity in the model.

In regression, we expect the independent variables to be correlated to the dependent variable. It often happens, however, that they are also correlated to each other. If these correlations are high then multicollinearity is said to exist.

```
Dependent variable: Price
-----------------------------------------------------------------------------
                                      Standard           T
Parameter              Estimate        Error         Statistic       P-Value
-----------------------------------------------------------------------------
CONSTANT                37717.6       14176.7         2.66053         0.0091
Bedrooms                2306.08       6994.19        0.329714         0.7423
H_Size                  74.2968       52.9786         1.40239         0.1640
Lot_Size               -4.36378        17.024       -0.256331         0.7982
-----------------------------------------------------------------------------
                        Analysis of Variance
-----------------------------------------------------------------------------
Source           Sum of Squares    Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model              7.65017E10       3   2.55006E10       40.73       0.0000
Residual            6.0109E10      96    6.26136E8
-----------------------------------------------------------------------------
Total (Corr.)      1.36611E11      99
```

R-squared = 55.9998 percent
R-squared (adjusted for d.f.) = 54.6248 percent
Standard Error of Est. = 25022.7

B. Diagnosing Multicollinearity

There are several ways to diagnose the existence of multicollinearity in a model. The following are the simplest indicators:

1. The *P*-values of important explanatory (independent) variables are high for the model. For example, the individual *P*-values for *Bedrooms*, *H_Size*, and *Lot_Size* are all high although we know that they are all correlated with House Price.
2. The algebraic sign (+/-) of one or more of the slopes is incorrect. For example, the regression coefficient for *Lot_Size* is negative, suggesting that *increasing* the size of the lot will tend, on average, to *decrease* the price of the house. A simple regression of *Price* on *Lot_Size*, however, confirms our suspicion that the two are *positively correlated*!

C. Problems Stemming from Multicollinearity

While the existence of multicollinearity doesn't violate any model assumptions, or make the model invalid, it does pose certain problems for the analyst:

1. Individual *t*-tests may prove unreliable, making it difficult to determine which variables in the model are correlated to ***Y***.
2. Because the estimated slopes may vary wildly from sample to sample (and even change algebraic sign), it may not be possible to interpret the slope of an independent variable as the *marginal* effect of a unit change in the variable upon the average value of ***Y*** (as described in the next section).

D. Remediation

The simplest way to remove multicollinearity is to remove one or more of the correlated variables. For example, for the House Price data removing the variables *Bedrooms* and *Lot_Size* produces a simpler (in fact, Simple Linear Regression) model without multicollinearity, as shown below.

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			

Correlation Coefficient = 0.747762
R-squared = 55.9149 percent
R-squared (adjusted for d.f.) = 55.465 percent
Standard Error of Est. = 24789.9

IX Interpreting the Regression Coefficients

The regression coefficients are interpreted essentially the same in multiple regression as they are in simple regression, *with one caveat*. The slope of an independent variable in multiple regression is usually interpreted as the *marginal* (or isolated) effect of a unit change in the variable upon the mean value of ***Y*** when "*the values of all of the other independent variables are held constant"*. Thus, as stated in the previous section, when multicollinearity is a problem it may *not* be possible to interpret all of the coefficients. This is because some of the independent variables are closely interrelated, making it impossible to change the value of one while holding the values of the others constant. (For those of you who've had Math 32, the interpretation of the coefficient of an independent variable in multiple regression is similar to the interpretation of a partial derivative taken with respect to one of the independent variables.)

Graphically, the coefficient β_i of the independent variable X_i is the slope of the plane $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ that we would encounter if we decided to walk in the direction of increasing values of X_i, i.e., parallel to the X_i axis. Specifically, if we move one unit in this direction, ***Y*** will change, on average, by β_i units. When multicollinearity is a problem, however, we may not be able to move around the plane in directions parallel to the axes of individual independent variables (we're encouraged to move in certain preferred directions on the plane). Thus, we are unable to "experience" the slope of X_i, and β_i can no longer be interpreted as the marginal effect of a unit change in the value of X_i upon the mean value of ***Y***.

Example 1 (continued): For Securicorp, the regression coefficients are interpreted below. For convenience, the Statgraphics' output found in the *Analysis Summary* window is shown again. Recall that *Sales* is in thousands of dollars, while *Ad* and *Bonus* are in hundreds of dollars.

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	-515.073	190.759	-2.70013	0.0131
Ad	2.47216	0.275644	8.96869	0.0000
Bonus	1.85284	0.717485	2.5824	0.0170

- $\hat{\beta}_0$ **:** $\hat{\beta}_0$ estimates the expected annual sales for a territory if $0.00 is spent on advertising and bonuses. Because these values are outside the range of values for *Ad* and *Bonus* observed, and upon which the estimated regression equation is based, the value of $\hat{\beta}_0$ has no practical interpretation. Put more concisely, an interpretation of $\hat{\beta}_0$ is not supported by the data. This will often, but not always, be the case in multiple regression. You should try to come up with a scenario, not involving the Securicorp example, where interpretation of the estimated intercept $\hat{\beta}_0$ would be appropriate.
- $\hat{\beta}_1$ **:** Expected (mean) sales increase by about $2,472 for each additional $100 spent on advertising, holding the amount of bonuses paid constant.
- $\hat{\beta}_2$ **:** Sales increase by $1,853, on average, for every $100 increase in bonuses, for a given amount spent on advertising

X The Standard Error of the Estimate

$S_\varepsilon = \sqrt{MSE}$ = the sample estimate of the standard deviation of the error variable, $\boldsymbol{\sigma_\varepsilon}$, which is a measure of the spread of the actual values of $\boldsymbol{Y}$ about the true plane $\boldsymbol{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. As such, the standard error should be reported in units appropriate to the dependent variable. For example, in the regression of *Sales* on *Ad* and *Bonus* for Securicorp the standard error of the estimate is $91,121. As in simple regression, Statgraphics displays the standard error below the ANOVA Table.

XI Preferred Measures of Fit

Although there are many statistics which can be used to measure the fit of a model to the data, such as $\boldsymbol{S_\varepsilon}$, the most commonly used statistics for this purpose are R^2 and R^2-*adjusted for degrees of freedom*.

- R^2 is defined as in simple regression and continues to have the same interpretation. The drawback to R^2 is that, because it can't decrease when new variables are added to a model, it is inappropriate for comparing models with different numbers of independent variables. For this reason, a statistic that included information about the number of independent variables (and that penalized models with lots of useless or redundant variables) was created. This new statistic is called R^2-*adjusted for degrees of freedom*, or simply R^2_{Adj} .
- $R^2_{Adj} = 1 - \dfrac{MSE}{s^2}$, where $s^2 = \dfrac{SST}{n-1}$ is the sample variance for $\boldsymbol{Y}$ (the "missing" Mean Square in the ANOVA Table). Because R^2_{Adj} includes information about the sample size and the number of independent variables, it is more appropriate than R^2 for comparing models with different numbers of independent variables. R^2_{Adj} appears directly below R^2 in Statgraphic's output.

There are other measures of model fit that are covered in more advanced courses. In Statgraphics, Follow *Relate* > *Multiple Factors* > *Regression Model Selection* to have the program use several different criteria to evaluate models you are considering, both at an initial stage when you are assembling prospective models and at the final stage to evaluate your final model. If you wish to know more (and who wouldn't), I have prepared a brief (but technical) introduction to two of the criteria displayed by Statgraphics, the Akaike Information Criterion and Mallows' C_p statistic, in the handout Model Selection Criteria Alternatives to R^2-adjusted, which can be found on the course Canvas page.

XII Dummy Variables

A. The Problem

Regression is designed for Quantitative variables, i.e., both the dependent and independent variables are quantitative. There are times, however, when we wish to include information about Qualitative variables in the model. For example, qualitative factors such as the presence of a fireplace, pool, or attached garage may have an effect upon the price of a house.

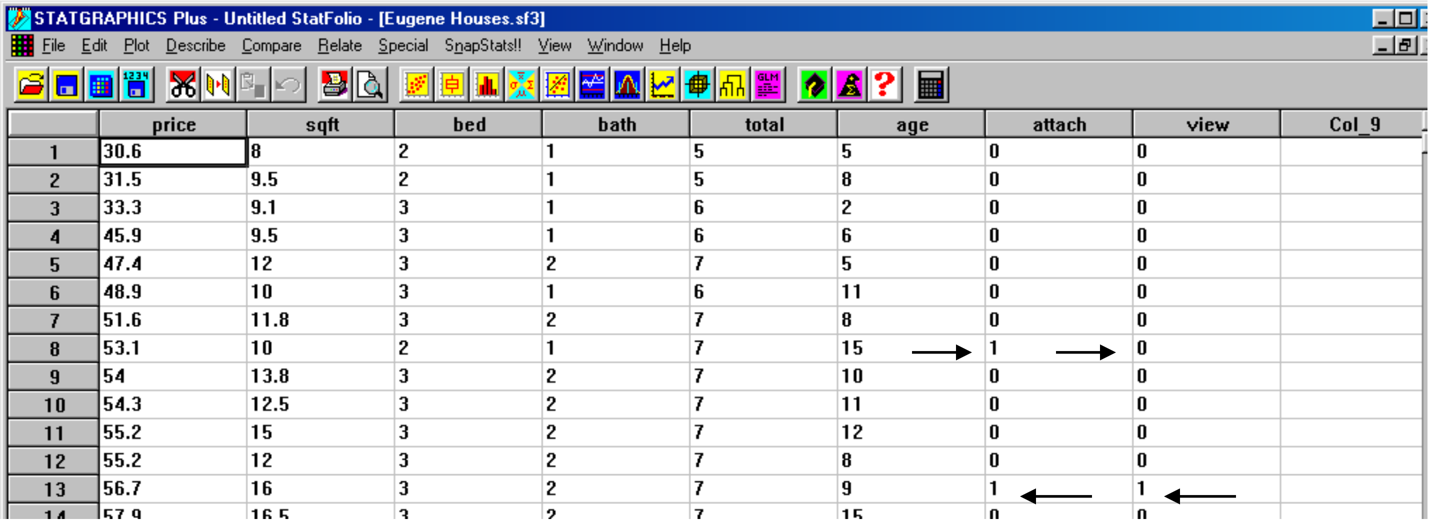
B. The Solution

The way to get around regression's restriction to quantitative variables is through the creation of Dummy Variables (also called indicator or Bernoulli variables).

A dummy variable for a characteristic indicates the presence or absence of the characteristic in the observation. For example, in the Eugene house data (see the notes for simple regression) the variables *Attach* and *View* indicate the presence or absence of an attached garage or a nice view, respectively, for each house observed. The variables are defined in the data as follows.

$$\textbf{Attach} = \begin{cases} \textbf{1} & \text{, if the house has an attached garage} \\ \textbf{0} & \text{, otherwise} \end{cases}$$

$$\textbf{View} = \begin{cases} \textbf{1} & \text{, if the house has a "nice" view} \\ \textbf{0} & \text{, otherwise} \end{cases}$$


STATGRAPHICS Plus - Untitled StatFolio - [Eugene Houses.sf3]
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help

	price	sqft	bed	bath	total	age	attach	view	Col_9
1	30.6	8	2	1	5	5	0	0	
2	31.5	9.5	2	1	5	8	0	0	
3	33.3	9.1	3	1	6	2	0	0	
4	45.9	9.5	3	1	6	6	0	0	
5	47.4	12	3	2	7	5	0	0	
6	48.9	10	3	1	6	11	0	0	
7	51.6	11.8	3	2	7	8	0	0	
8	53.1	10	2	1	7	15	1	0	
9	54	13.8	3	2	7	10	0	0	
10	54.3	12.5	3	2	7	11	0	0	
11	55.2	15	3	2	7	12	0	0	
12	55.2	12	3	2	7	8	0	0	
13	56.7	16	3	2	7	9	1	1	
14	57.9	16.5	3	2	7	15	0	0	

From looking at the spreadsheet above, we can tell that the eighth house in the sample had an attached garage but no view, while the thirteenth house had both.

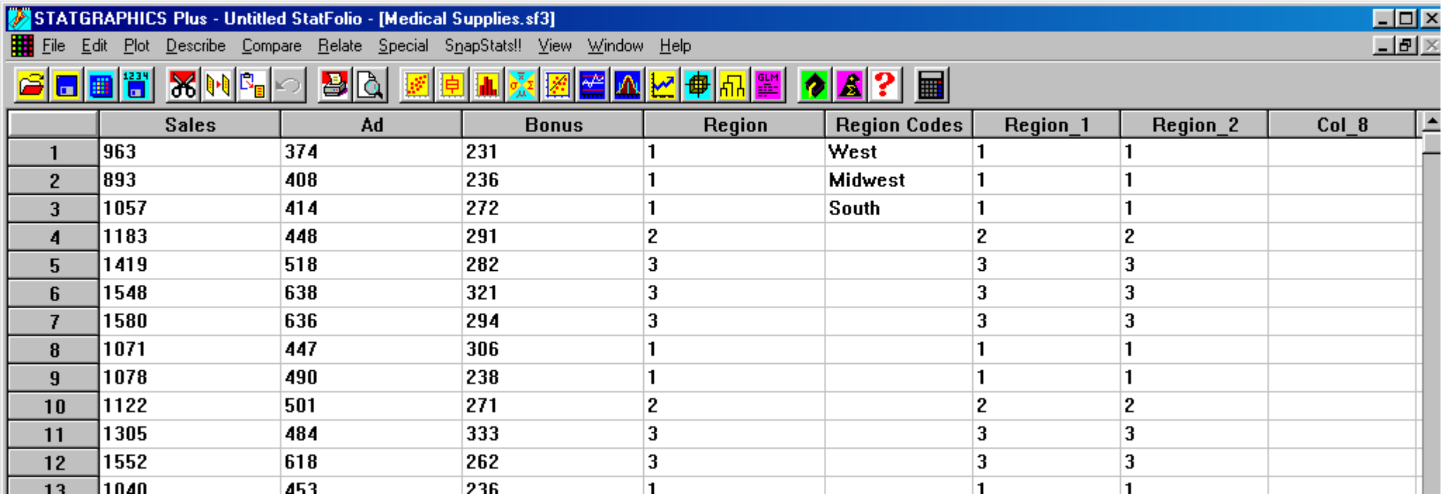
C. Qualitative Variables with More than Two Possible Outcomes

If a qualitative variable has more than two possible outcomes, for example a variable for the seasons may acquire the values Spring, Summer, Fall, or Winter, then a dummy is created for *all but one* of the outcomes. (It is important that you exclude one of the outcomes from the model. Statgraphics will get upset if you try to include them all!) Thus, we might create one dummy for Spring, another for Summer, and a third for Fall.

D. Creating Dummy Variables in Statgraphics

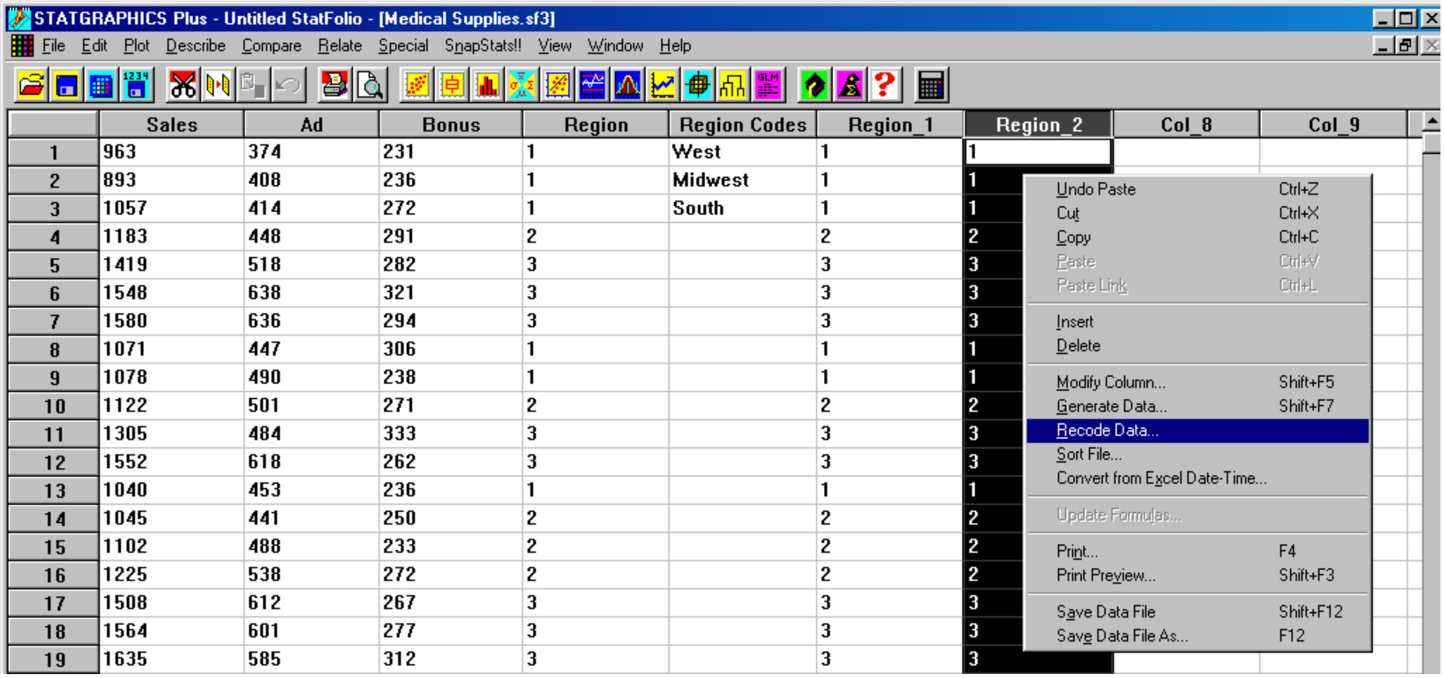
In the spreadsheet for Eugene houses the variables Attach and View were already represented as dummy (0 or 1) variables. Frequently, however, the values of a qualitative variable will appear as descriptions ("smoker" or "nonsmoker") or numerical codes ("1" for red, "2" for blue, "3" for green, etc.). To create dummy variables for a qualitative variable with *m* possible outcomes, begin by copying the variable and pasting it into *m* – 1 columns in the spreadsheet.

Example 1 (continued): Securicorp would also like to examine whether the marketing region to which a territory belongs affects expected sales, after taking into account the effect of advertising and bonuses. These regions have been coded in the data using 1 = West, 2 = Midwest, and 3 = South. (The column "Region Codes" provides the key to the codes.) Copying and pasting the Region variable twice (*m* = 3) we arrive at the view below.


STATGRAPHICS Plus - Untitled StatFolio - [Medical Supplies.sf3]
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help

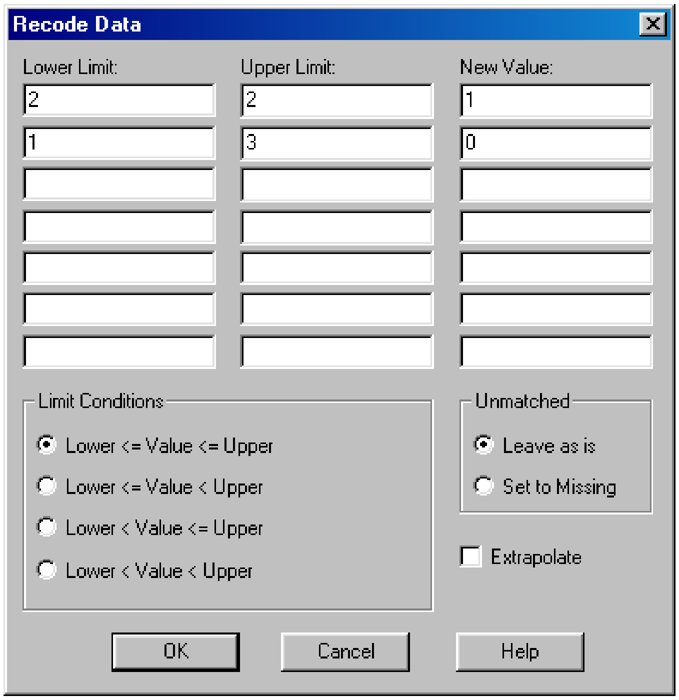
	Sales	Ad	Bonus	Region	Region Codes	Region_1	Region_2	Col_8
1	963	374	231	1	West	1	1	
2	893	408	236	1	Midwest	1	1	
3	1057	414	272	1	South	1	1	
4	1183	448	291	2		2	2	
5	1419	518	282	3		3	3	
6	1548	638	321	3		3	3	
7	1580	636	294	3		3	3	
8	1071	447	306	1		1	1	
9	1078	490	238	1		1	1	
10	1122	501	271	2		2	2	
11	1305	484	333	3		3	3	
12	1552	618	262	3		3	3	
13	1040	453	236	1		1	1	

To create a dummy variable for the Midwest, begin by selecting one of the pasted columns in the spreadsheet. Then use the right mouse button to access the **column menu** shown below and select *Recode Data*.


STATGRAPHICS Plus - Untitled StatFolio - [Medical Supplies.sf3]
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help
Undo Paste Ctrl+Z
Cut Ctrl+X
Copy Ctrl+C
Paste Ctrl+V
Paste Link Ctrl+L
Insert
Delete
Modify Column... Shift+F5
Generate Data... Shift+F7
Recode Data...
Sort File...
Convert from Excel Date-Time...
Update Formulas...
Print... F4
Print Preview... Shift+F3
Save Data File Shift+F12
Save Data File As... F12

	Sales	Ad	Bonus	Region	Region Codes	Region_1	Region_2	Col_8	Col_9
1	963	374	231	1	West	1	1		
2	893	408	236	1	Midwest	1	1		
3	1057	414	272	1	South	1	1		
4	1183	448	291	2		2	2		
5	1419	518	282	3		3	3		
6	1548	638	321	3		3	3		
7	1580	636	294	3		3	3		
8	1071	447	306	1		1	1		
9	1078	490	238	1		1	1		
10	1122	501	271	2		2	2		
11	1305	484	333	3		3	3		
12	1552	618	262	3		3	3		
13	1040	453	236	1		1	1		
14	1045	441	250	2		2	2		
15	1102	488	233	2		2	2		
16	1225	538	272	2		2	2		
17	1508	612	267	3		3	3		
18	1564	601	277	3		3	3		
19	1635	585	312	3		3	3		

This leads to the dialog box seen to the below. Using *Tab* on the keyboard to move around, we've instructed Statgraphics to turn all "2"s (midwestern territories) in the column into "1"s, while any other number (region) is set to zero.



Lower Limit:	Upper Limit:	New Value:
2	2	1
1	3	0

Limit Conditions

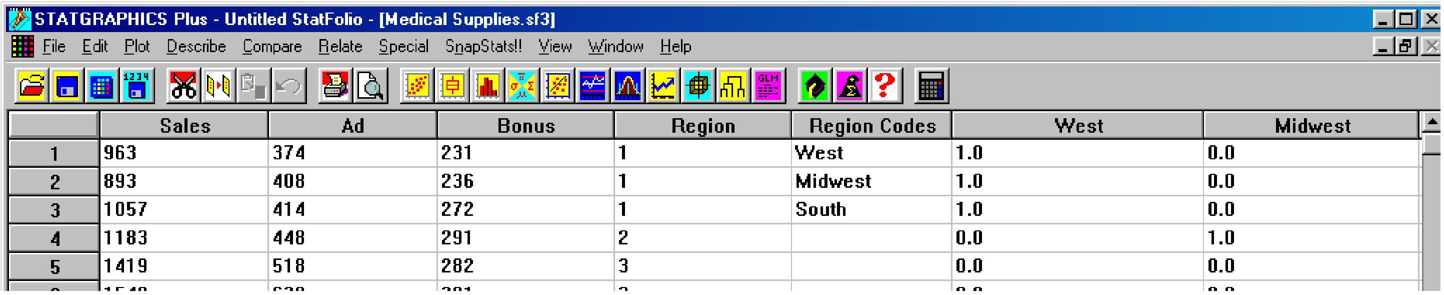
- ● Lower <= Value <= Upper
- ○ Lower <= Value < Upper
- ○ Lower < Value <= Upper
- ○ Lower < Value < Upper

Unmatched

- ● Leave as is
- ○ Set to Missing

□ Extrapolate

At this point, *Modify Column* may be selected from the column menu and used to rename the column *Midwest*. After similarly recoding the other pasted column for the western region, the spreadsheet will look like the one below.



	Sales	Ad	Bonus	Region	Region Codes	West	Midwest
1	963	374	231	1	West	1.0	0.0
2	893	408	236	1	Midwest	1.0	0.0
3	1057	414	272	1	South	1.0	0.0
4	1183	448	291	2		0.0	1.0
5	1419	518	282	3		0.0	0.0

Note that the third territory in the data, which belongs to the southern marketing region, appears with zeros in the columns for the western and midwestern regions, i.e., it belongs to neither *West* nor *Midwest*. Thus, only two dummy variables are required for the three regions! In general, $m - 1$ dummy variables are created for a qualitative variable with m possible outcomes.

We are now ready to include the region identification of a sales territory into our regression model for annual sales. The model now looks like **Sales** = $\beta_0 + \beta_1$Ad + β_2Bonus + β_3West + β_4Midwest + ε. The *Input Dialog* button (far left of the main toolbar) is used to rerun the regression analysis with the new set of independent variables. The *Analysis Window* for the new model appears below.

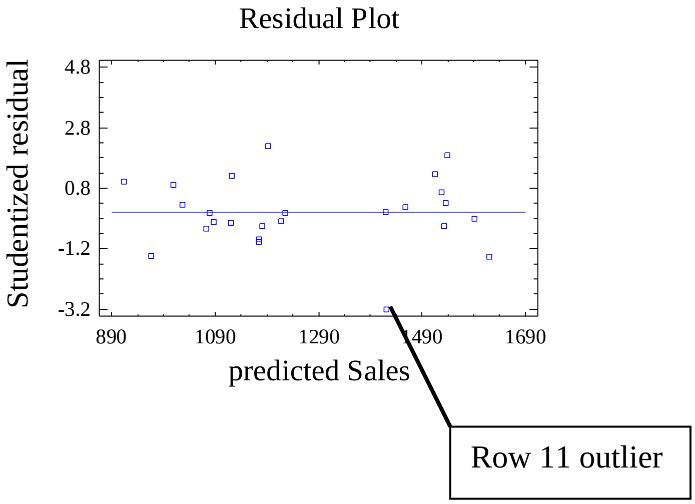
		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	439.193	206.222	2.1297	0.0458
Ad	1.36468	0.26179	5.21287	0.0000
Bonus	0.96759	0.480814	2.0124	0.0578
West	-258.877	48.4038	-5.34827	0.0000
Midwest	-210.456	37.4223	-5.62382	0.0000

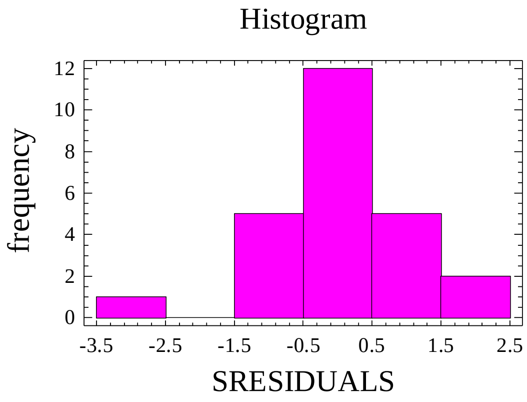
Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	1.18325E6	4	295812.	88.79	0.0000
Residual	66632.8	20	3331.64		
Total (Corr.)	1.24988E6	24			

R-squared = 94.6689 percent
R-squared (adjusted for d.f.) = 93.6027 percent
Standard Error of Est. = 57.7204

R^2_{Adj} has increased from 84.1% to 93.6% with the addition of the dummy variables for region. In addition, three of the four variables are significant at the 5% level of significance (including both dummies), while *Bonus* is significant at the 10% level. The *Plot of Residuals vs. Predicted* for Sales and the histogram of the studentized residuals for the new model are shown below. Note the outlier in row 11.





E. Interpreting Dummy Variables

For the Securicorp sales model with *Ad*, *Bonus*, *West*, and *Midwest* included, the estimated regression equation returned by Statgraphics is

$$\text{Sales} = 439.193 + 1.365{*}\text{Ad} + 0.968{*}\text{Bonus} - 258.877{*}\text{West} - 210.456{*}\text{Midwest}.$$

- For a southern territory, West = 0 and Midwest = 0, and the regression equation reduces to

$$\text{Sales} = 439.193 + 1.365{*}\text{Ad} + 0.968{*}\text{Bonus}.$$

- For a western territory, West = 1 and Midwest = 0, and the equation becomes

$$\text{Sales} = 439.193 + 1.365{*}\text{Ad} + 0.968{*}\text{Bonus} - 258.877.$$

Comparing these two equations, the model predicts that expected annual sales, for given expenditures on advertising and bonuses, will be $258,877 *less* for a western territory than for a southern territory. Thus, the estimated regression coefficient –258.877 for *West* is interpreted as the expected *difference* in sales between western and southern territories, in thousands of dollars, given identical advertisement and bonus expenditures. Similarly, the coefficient –210.456 for *Midwest* is interpreted as the *marginal difference* in mean annual sales between midwestern and southern territories. (How would we use the equation for *Sales* to compare the expected annual sales for western and midwestern territories?)

XIII Outliers and Leverage

There are many measures that may have been developed to flag possible outliers and points exerting significant influence on the fitting of the model.

Observations with large studentized residuals lie relatively far from the model and may be considered outliers. As with simple linear regression, Statgraphics computes deleted Studentized residuals (also called externally studentized residuals) to remove the point's influence on the model prior to computing its residual. Statgraphics then flags observations with absolute studentized residuals greater than 2 as possible outliers. These can be found in the *Unusual Residuals* table

Statgraphics computes leverages for observations. Leverage measures how far the *x*-coordinates of a point are from the center for the x-coordinates for the observations. An observation with a comparatively large leverage is an outlier in its *x*-coordinates. Points with high leverage have the potential to be influential, i.e., they may exert a larger than average influence on the regression coefficients (the *βetas*). Such points may not, however, exercise their potential to influence the model. Observations with leverage values greater than 3 times the average are reported in the *Influential Points* table.

Statgraphics also reports values of an influence measure called DFITS. DFITS is a true measure of influence. For each point the procedure fits two models: a model based on all *n* observations and another model fitted after deleting the observation. It compares the residual for the point in both models. If the point is influential, the residual in the (second) deleted model will be significantly larger. Statgraphics flags observations with large absolute DFITS values in the *Influential Points* table.

You can save Studentized residuals, leverages, and DFITS values for all observations in the datasheet by clicking on the Save Results button and selecting them. Formulas for the statistics mentioned above are as follows:

Leverage value of the i^{th} observation: h_i . We will postpone further discussion of the computation of leverage scores until we cover the use of matrices in regression. It will turn out that leverage scores appear along the main diagonal of a matrix we will study in some detail.

Studentized Residual of the i^{th} observation: $t_i = e_i\sqrt{\dfrac{n-k-2}{(1-h_i)SSE - e_i^2}}$, where

- e_i is the ordinary residual for the i^{th} observation.
- h_i is the leverage value of the i^{th} observation.
- k is the number of independent variables in the model.
- SSE is the error sum of squares for the model fitted with all n observations.

DFITS value of the i^{th} observation: $\text{DFITS} = t_i\sqrt{\dfrac{h_i}{1-h_i}}$, where

- t_i is the Studentized residual for the i^{th} observation.
- h_i is the leverage value of the i^{th} observation.

Example 1 (continued): For the Securicorp sales model with *Ad*, *Bonus*, *West*, and *Midwest* included, selecting the *Unusual Residuals* table produces the following output, highlighting points in rows 11 and 22 as outliers far from the regression surface.

Unusual Residuals

		Predicted		*Studentized*
Row	*Y*	*Y*	*Residual*	*Residual*
11	1305.0	1421.9	-116.904	-3.20
22	1294.0	1191.92	102.077	2.20

Selecting the *Influential Points* table returns observations with high DFITS values. Note that Statgraphics has flagged rows 11 and 22 as both outliers and influential to the estimation of the regression coefficients.

Influential Points

		Mahalanobis	
Row	*Leverage*	*Distance*	*DFITS*
11	0.413671	15.2688	-2.68553
22	0.231615	5.97459	1.2097

Average leverage of single data point = 0.2

XIV Forecasting in Multiple Linear Regression Using Statgraphics

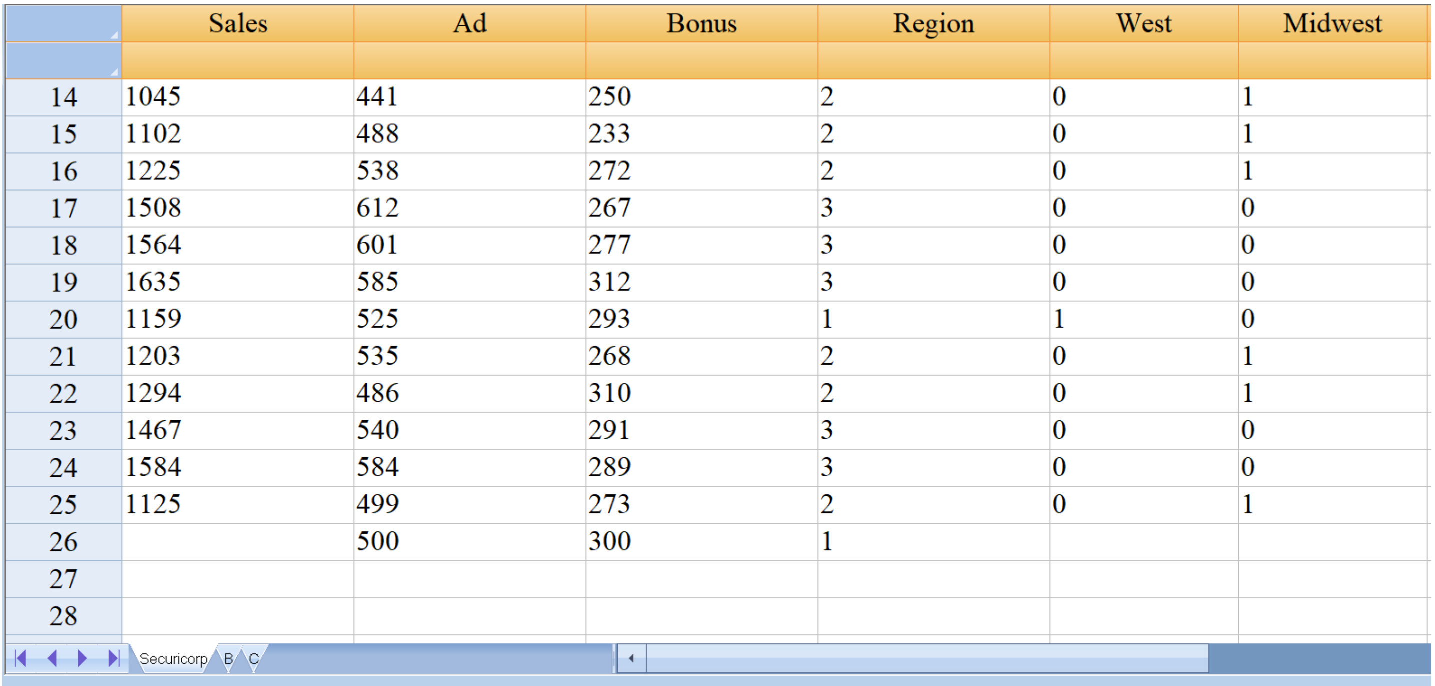
For the Securicorp example, the following represents the variables in a multiple regression model:

- ***Sales***: Sales, in thousands of dollars
- *Ad*: Advertising, in hundreds of dollars
- *Bonus*: Bonuses, in hundreds of dollars
- *West*: The Western sales territory dummy variable
- *Midwest*: the Midwestern sales territory dummy variable
- (Note: The Southern sales territory does not receive a dummy variable in this model)

The resulting model produced by StatGraphics is

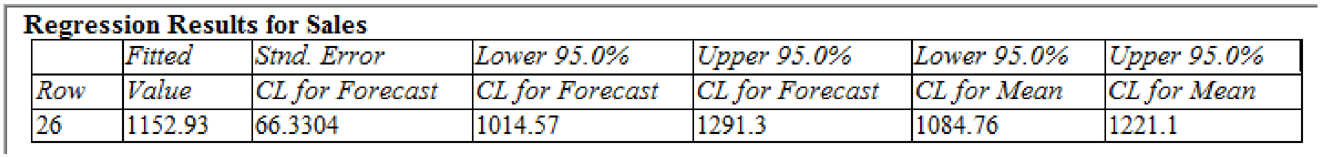
		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	439.193	206.222	2.1297	0.0458
Ad	1.36468	0.26179	5.21287	0.0000
Bonus	0.96759	0.480814	2.0124	0.0578
West	-258.877	48.4038	-5.34827	0.0000
Midwest	-210.456	37.4223	-5.62382	0.0000

To produce point estimates and prediction and confidence interval from the model in StatGraphics, check *Reports* in the *Tables and Graphs* menu. Now, suppose we wish to forecast sales in the Western region for $50,000 in advertising and $30,000 in bonuses. Place the appropriate values in the first empty row (row number 26) of the spreadsheet, as shown below.



	Sales	Ad	Bonus	Region	West	Midwest
14	1045	441	250	2	0	1
15	1102	488	233	2	0	1
16	1225	538	272	2	0	1
17	1508	612	267	3	0	0
18	1564	601	277	3	0	0
19	1635	585	312	3	0	0
20	1159	525	293	1	1	0
21	1203	535	268	2	0	1
22	1294	486	310	2	0	1
23	1467	540	291	3	0	0
24	1584	584	289	3	0	0
25	1125	499	273	2	0	1
26		500	300	1		
27						
28						

In the *Reports* window you will see the output below showing the point estimate (*Fitted Value*), Prediction Interval (*Lower and Upper CL for Forecast*), and Confidence Interval (*Lower and Upper CL for Mean*). The default level of 95% for the intervals can be changed by right-clicking in the *Reports* window and selecting *Pane Options*.



Regression Results for Sales

	Fitted	*Stnd. Error*	*Lower 95.0%*	*Upper 95.0%*	*Lower 95.0%*	*Upper 95.0%*
Row	*Value*	*CL for Forecast*	*CL for Forecast*	*CL for Forecast*	*CL for Mean*	*CL for Mean*
26	1152.93	66.3304	1014.57	1291.3	1084.76	1221.1