MULTIPLE REGRESSION BASICS

Documents prepared for use in course B01.1305 and C22.0103,
New York University, Stern School of Business

Revision date: 27 MAR 2009

Cover photo: Praying mantis, 2003

INPUT TO A REGRESSION PROBLEM

Simple regression: $(x_1, Y_1), (x_1, Y_2), \dots , (x_n, Y_n)$

Multiple regression: $((x1)_1, (x2)_1, (x3)_1, \dots (xK)_1, Y_1),$
$((x1)_2, (x2)_2, (x3)_2, \dots (xK)_2, Y_2),$
$((x1)_3, (x2)_3, (x3)_3, \dots (xK)_3, Y_3),$
$\dots ,$
$((x1)_n, (x2)_n, (x3)_n, \dots (xK)_n, Y_n),$

The variable Y is designated as the "dependent variable." The only distinction between the two situations above is whether there is just one x predictor or many. The predictors are called "independent variables."

> There is a certain awkwardness about giving generic names for the independent variables in the multiple regression case. In this notation, $x1$ is the name of the first independent variable, and its values are $(x1)_1, (x1)_2, (x1)_3, \dots , (x1)_n$. In any application, this awkwardness disappears, as the independent variables will have application-based names such as *SALES*, *STAFF*, *RESERVE*, *BACKLOG*, and so on. Then *SALES* would be the first independent variable, and its values would be $SALES_1, SALES_2, SALES_3, \dots , SALES_n$.
>
> The listing for the multiple regression case suggests that the data are found in a spreadsheet. In application programs like Minitab, the variables can appear in any of the spreadsheet columns. The dependent variable and the independent variables may appear in any columns in any order. Microsoft's EXCEL requires that you identify the independent variables by blocking off a section of the spreadsheet; this means that the independent variables must appear in consecutive columns.

MINDLESS COMPUTATIONAL POINT OF VIEW

The output from a regression exercise is a "fitted regression model."

Simple regression: $\hat{Y} = b_0 + b_1 x$

Multiple regression: $\hat{Y} = b_0 + b_1(x1) + b_2(x2) + b_3(x3) + ... + b_K(xK)$

Many statistical summaries are also produced. These are R^2, standard error of estimate, t statistics for the b's, an F statistic for the whole regression, leverage values, path coefficients, and on and on and on and ...... This work is generally done by a computer program, and we'll give a separate document listing and explaining the output.

◩◩◩◩◩◩ INTRODUCTORY THOUGHTS ABOUT MULTIPLE REGRESSION ◩◩◩◩◩◩

WHY DO PEOPLE DO REGRESSIONS?

A cheap answer is that they want to explore the relationships among the variables.

A slightly better answer is that we would like to use the framework of the methodology to get a yes-or-no answer to this question: Is there a significant relationship between variable Y and one or more of the predictors? Be aware that the word *significant* has a very special jargon meaning.

An simple but honest answer pleads curiousity.

The most valuable (and correct) use of regression is in making predictions; see the next point. Only a small minority of regression exercises end up by making a prediction, however.

HOW DO WE USE REGRESSIONS TO MAKE PREDICTIONS?

The prediction situation is one in which we have new predictor variables but do not yet have the corresponding Y.

Simple regression: We have a new x value, call it x_{new} , and the predicted (or fitted) value for the corresponding Y value is $\hat{Y}_{\text{new}} = b_0 + b_1 x_{new}$.

Multiple regression: We have new predictors, call them $(x1)_{new}$, $(x2)_{new}$, $(x3)_{new}$, …, $(xK)_{new}$. The predicted (or fitted) value for the corresponding Y value is

$$\hat{Y}_{new} = b_0 + b_1 (x1)_{new} + b_2 (x2)_{new} + b_3 (x3)_{new} + ... + b_K (xK)_{new}$$

CAN I PERFORM REGRESSIONS WITHOUT ANY UNDERSTANDING OF THE UNDERLYING MODEL AND WHAT THE OUTPUT MEANS?

Yes, many people do. In fact, we’ll be able to come up with rote directions that will work in the great majority of cases. Of course, these rote directions will sometimes mislead you. And wisdom still works better than ignorance.

WHAT'S THE REGRESSION MODEL?

The model says that Y is a linear function of the predictors, plus statistical noise.

Simple regression: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Multiple regression: $Y_i = \beta_0 + \beta_1 (x1)_i + \beta_2 (x2)_i + \beta_3 (x3)_i + \ldots + \beta_K (xK)_i + \varepsilon_i$

The coefficients (the β's) are nonrandom but unknown quantities. The noise terms ε_1, ε_2, ε_3, …, ε_n are random and unobserved. Moreover, we assume that these ε's are statistically independent, each with mean 0 and (unknown) standard deviation σ.

The model is simple, except for the details about the ε's. We're just saying that each data point is obscured by noise of unknown magnitude. We assume that the noise terms are not out to deceive us by lining up in perverse ways, and this is accomplished by making the noise terms independent.

Sometimes we also assume that the noise terms are taken from normal populations, but this assumption is rarely crucial.

WHO GIVES ANYONE THE RIGHT TO MAKE A REGRESSION MODEL? DOES THIS MEAN THAT WE CAN JUST SAY SOMETHING AND IT AUTOMATICALLY IS CONSIDERED AS TRUE?

Good questions. Merely claiming that a model is correct does not make it correct. A model is a mathematical abstraction of reality. Models are selected on the basis of simplicity and credibility. The regression model used here has proved very effective. A careful user of regression will make a number of checks to determine if the regression model is believable. If the model is not believable, remedial action must be taken.

HOW CAN WE TELL IF A REGRESSION MODEL IS BELIEVABLE? AND WHAT'S THIS REMEDIAL ACTION STUFF?

Patience, please. It helps to examine some successful regression exercises before moving on to these questions.

THERE SEEMS TO BE SOME PARALLEL STRUCTURE INVOLVING THE MODEL AND THE FITTED MODEL.

It helps to see these things side-by-side.

Simple regression:

The model is $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

The fitted model is $\hat{Y} = b_0 + b_1 x$

Multiple regression:

The model is $Y_i = \beta_0 + \beta_1 (x1)_i + \beta_2 (x2)_i + \beta_3 (x3)_i + \ldots$
$+ \beta_K (xK)_i + \varepsilon_i$

The fitted model is $\hat{Y} = b_0 + b_1 (x1) + b_2 (x2) + b_3 (x3) + \ldots + b_K (xK)$

The Roman letters (the *b*’s) are estimates of the corresponding Greek letters (the β’s).

WHAT ARE THE FITTED VALUES?

In any regression, we can "predict" or retro-fit the *Y* values that we've already observed, in the spirit of the PREDICTIONS section above.

Simple regression:

The model is $Y_i = \alpha + \beta\, x_i + \varepsilon_i$

The fitted model is $\hat{Y} = a + bx$

The fitted value for point *i* is

$$\hat{Y}_i = a + bx_i$$

Multiple regression:

The model is
$$Y_i \quad = \beta_0 + \beta_1\,(x1)_i + \beta_2\,(x2)_i + \beta_3\,(x3)_i + \ldots + \beta_K\,(xK)_{\mathrm{i}} + \varepsilon_i$$

The fitted model is $\hat{Y} = b_0 + b_1\,(x1) + b_2\,(x2) + b_3\,(x3) + ... + b_K\,(xK)$

The fitted value for point *i* is

$$\hat{Y}_i = b_0 + b_1\,(x1)_i + b_2\,(x2)_i + b_3\,(x3)_i + ... + b_K\,(xK)_i$$

Indeed, one way to assess the success of the regression is the closeness of these fitted *Y* values, namely $\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \ldots, \hat{Y}_n$ to the actual observed *Y* values Y_1, Y_2, Y_3, …, Y_n.

THIS IS LOOKING COMPUTATIONALLY HOPELESS.

Indeed it is. These calculations should only be done by computer. Even a careful, well-intentioned person is going to make arithmetic errors if attempting this by a non-computer method. You should also be aware that computer programs seem to compete in using the latest innovations. Many of these innovations are passing fads, so don't feel too bad about not being up-to-the-minute on the latest changes.

The notation used here in the models is not universal. Here are some other possibilities.

Notation here	Other notation
Y_i	y_i
x_i	X_i
$\beta_0+\beta_1 x_i$	$\alpha+\beta\ x_i$
ε_i	e_i or r_i
$(x1)_i$, $(x2)_i$, $(x3)_i$, …, $(xK)_i$	x_{i1}, x_{i2}, x_{i3}, …, x_{iK}
b_j	$\hat{\beta}_j$

In many regression problems, the data points differ dramatically in gross size.

> EXAMPLE 1: In studying corporate accounting, the data base might involve firms ranging in size from 120 employees to 15,000 employees.
>
> EXAMPLE 2: In studying international quality of life indices, the data base might involve countries ranging in population from 0.8 million to 1,000 millions.

In Example 1, some of the variables might be highly dependent on the firm sizes. For example, the firm with 120 employees probably has low values for gross sales, assets, profits, and corporate debt.

In Example 2, some of the variables might be highly dependent on country sizes. For example, the county with population 0.8 million would have low values for GNP, imports, exports, savings, telephones, newspaper circulation, and doctors.

Regressions performed with such gross size variables tend to have very large R^2 values, but prove nothing. In Example 1, one would simply show that big firms have big profits. In Example 2, one would show that big countries have big GNPs. The explanation is excellent, but rather uninformative.

There are two common ways for dealing with the gross size issue: ratios and logarithms.

The ratio idea just puts the variables on a "per dollar" or "per person" basis.

For Example 1, suppose that you wanted to explain profits in terms of number of employees, sales, assets, corporate debt, and (numerically coded) bond rating. A regression of profits on the other variables would have a high R^2 but still be quite uninformative. A more interesting regression would create the dependent variable profits/assets and use as the independent variables employees/assets, sales/assets, debt/assets. The regression model is

$$\frac{PROFIT_i}{ASSETS_i} = \beta_0 + \beta_1 \frac{EMPLOYEES_i}{ASSETS_i} + \beta_2 \frac{SALES_i}{ASSETS_i} + \beta_3 \frac{DEBT_i}{ASSETS_i} + \beta_4 \, BOND_i + \varepsilon_i \qquad \text{(Model 1)}$$

Observe that BOND, the bond rating, is not a "gross size" variable; there is no need to scale it by dividing by ASSETS.

In Example 1, the scaling might be described in terms of quantities per $1,000,000 of ASSETS. It might also be reasonable to use SALES as the scaling variable, rather than ASSETS.

For Example 2, suppose that you wanted to explain number of doctors in terms of imports, exports, savings, telephones, newspaper circulation, and inflation rate. The populations give you the best scaling variable. The regression model is

$$\frac{DOCTORS_i}{POPN_i} = \beta_0 + \beta_1 \frac{IMPORTS_i}{POPN_i} + \beta_2 \frac{EXPORTS_i}{POPN_i} + \beta_3 \frac{SAVINGS_i}{POPN_i}$$

$$+ \beta_4 \frac{PHONES_i}{POPN_i} + \beta_5 \frac{PAPERS_i}{POPN_i} + \beta_6 INFLATE_i + \varepsilon_i \qquad \text{(Model 2)}$$

All the ratios used here could be described as “per capita” quantities. The inflation rate is not a “gross size” variable and need not be put on a per capita basis.

An alternate strategy is to take logarithms of all gross size variables. In Example 1, one might use the model

$$\log(PROFIT_i) = \gamma_0 + \gamma_1 \log(ASSETS_i) + \gamma_2 \log(EMPLOYEES_i) + \gamma_3 \log(SALES_i)$$

$$+ \ \gamma_4 \log(DEBT_i\,) \ + \ \gamma_5\, BOND_i \ + \ \varepsilon_i$$

Of course, the coefficients γ_0 through γ_5 are not simply related to β_0 through β_4 in the original form of the model. Unless the distribution of values of BOND is very unusual, one would not replace it with its logarithm.

Similarly, the logarithm version of model 2 is

$$\log(DOCTORS_i) = \gamma_0 + \gamma_1 \log(POPN_i) + \gamma_2 \log(IMPORTS_i) + \gamma_3 \log(EXPORTS_i)$$

$$+ \gamma_4 \log(SAVINGS_i) + \gamma_5 \log(PHONES_i) + \gamma_6 \log(PAPERS_i) + \gamma_7 INFLATE_i + \varepsilon_i$$

Since INFLATE is not a “gross size” variable, we are not immediately led to taking its logarithm. If this variable has other distributional defects, such as being highly skewed, then we might indeed want its logarithm.

Finally, it should be noted that one does not generally combine these methods. After all, since $\log\left(\frac{A}{B}\right) = \log(A) - \log(B)$ the logarithm makes the ratio a moot issue.

Dividing logarithms, as in log(*DOCTORS$_i$*)/log(*POPN$_i$*) is not likely to be useful.

One always has the option of doing a “weighted” regression. One can use one of the variables as a weight in doing the regression. The company assets might be used for Example 1 and the populations used for Example 2. The problem with this approach is that the solution will depend overwhelmingly on the large firms (or large countries).

☺☺☺☺☺☺☺☺☺☺ DATA CLEANING ☺☺☺☺☺☺☺☺☺☺

Data cleaning steps

We will describe the operations in terms of the computer program Minitab.

We will assume here that we are working with a spreadsheet. The columns of this spreadsheet will represent variables; each number in a column must be in the same units. The rows of the spreadsheet will represent data points.

As a preliminary step, check each column for basic integrity. Minitab distinguishes columns of two major varieties, ordinary data and text. (There are also minor varieties, including dates.) If a column is labeled C5-T, then Minitab has interpreted this column as text information.

> It sometimes happens that a column which is supposed to be numeric ends up as text. What should you do in such a case?
>
> > Scan the column to check for odd characters, such as N/A, DK, ?, unk; some people use markers like this to indicate missing or uncertain values. The Minitab missing numeric data code is the asterisk *, and this should be used to replace things like the above. The expression 2 1/2 was intended to represent 2.5 but Minitab can only interpret it as text; this repair is obvious.
> >
> > If you edit a text column so that all information is interpretable as numeric, Minitab will not instantly recognize the change. Use **Manipulate ⇒ Change Data Type ⇒ Text to Numeric**. If you do this to a column that still has text information, the corresponding entries will end up as *, the numeric missing data code.
>
> It sometimes happens that a column given as numeric really represents a nominal categorical variable and you would prefer to use the names. For example, a column might have used 1, 2, 3, 4 to represent single, married, widowed, and divorced. You would prefer the names. Use **Manipulate ⇒ Code ⇒ Numeric to Text**. You will be presented with a conversion table which allows you to do this.

The command **Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics** will give you the minimum and maximum of each column. The minimum and maximum values should make sense; unbelievable numbers for the minimum or the maximum could well be data coding errors. This same command will give you the number of missing values, noted as N^*. The count on missing values should make sense.

For many analyses you would prefer to deal with reasonably symmetric values. One of the cures for right-skewness is the taking of logarithms. Here are some general comments about this process:

Base *e* logarithms are usually preferred because of certain advantages in interpretation. It is still correct, however, to use base 10 logarithms.

Some variables are of the "gross size" variety. The minimum to maximum span runs over several orders of magnitudes.   For example, in a data set on countries of the world, the variable POPULATION will run from 10^5 to 10^9 with many countries at the low end of the scale. This variable should be replaced by its logarithm. In a data set on the Fortune 500 companies, the variable REVENUES will run over several orders of magnitude with most companies toward the low end of the scale. This variable should be replaced by its logarithm.

The command **Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics** will allow you to compare the mean and the standard deviation. If a variable which is always (or nearly always) positive has a standard deviation about as large as the mean, or even larger, is certainly positively skewed.

What should you do with data that are skewed but not necessarily of the "gross size" variety? This is a matter of judgment. Generally you prefer to keep variables in their original units.   If most of the other variables are to be transformed by logarithms, then maybe you want to transform this one as well.

If the skewed variable is going to be the dependent variable in a regression, then you will almost certainly want to take its logarithm.   (If you don't take the logarithms immediately, you may find expanding residuals on the residual versus fitted plot. Then you'll have take logarithms anyhow.)

If the variable to be transformed by logarithms as zero or negative values, then taking logarithms in Minitab will make trouble. (In releases 13 and earlier, the calculation will become a missing value with no warning. In release 14, the user will get a diagnostic message.) The technique is to pick a value c so that all values of $X + c$ are positive. Then consider $\log(X + c)$.

Logarithms will *not* cure left-skewed data. If X is such a variable and if M is a number larger than the biggest X, then you can consider $\log(M - X)$, provided you can make a sensible interpretation for this.

Logarithms should *not* be applied to binary variables. If a variable has only two values, then the logarithms will also have only two values.

Suppose that we regress Y on other variables, including J. The fitted model will be

$$\hat{Y} = b_0 + \ldots. + b_J\, J + \ldots\ldots$$

The interpretation of b_J is this:

> As J increases by 1, there is an associated increase in Y of b_J , while holding all other predictors fixed.

There's an important WARNING.

> WARNING: This interpretation should note that b_J is the "effect" of J on Y after adjusting for the presence of all other variables. (In particular, regressing Y on J without any other predictors could produce a very different value of b_J .) Also, this interpretation carries the disclaimer "while holding all other predictors fixed." Realistically, it may not be possible to change the value of J while leaving the other predictors unchanged.

Now… suppose that Y is really the base-e logarithm of Z, meaning $Y = \log Z$. What's the link between J and Z? The fitted model is

$$\log \hat{Z} = b_0 + \ldots\; b_J\, J + ..$$

Here the interpretation of b_J is this:

> As J increases by 1, there is an associated increase in $\log Z$ of b_J . This means that $\log Z$ changes to $\log Z + b_J$. By exponentiating, we find that $e^{\log Z} = Z$ changes to $e^{\log Z + b_J} = e^{\log Z}\, e^{b_J} = Z\;\, e^{b_J}$. Using the approximation that $e^t \approx 1 + t$ when t is near zero, we find that Z changes (approximately) to $Z(1+b_J)$. This is interpretable as a percent increase. We summarize thus: as J increases by 1, there is an associated proportional increase of b_J in Z.
>
> > If, for example, $b_J = 0.03$, then as J increases by 1, the associated increase in Z is 3%.

This next case is encountered only rarely.

Next suppose that Y is *not* the result of a transformation, but that $J = \log R$ is the base-e logarithm of variable R. What's the link between R and Y? Let's talk about increasing J by 0.01. (The reason why we consider an increase of 0.01 rather than an increase of 1 will be mentioned below.) Certainly we can say this:

The fitted model is $\hat{Y} = b_0 + \dots + b_J \log R + \dots$

As $J = \log R$ increases by 0.01, there is an associated increase in Y of 0.01 b_J . Saying that J increases by 0.01 is also saying that $\log R$ increases to $\log R + 0.01$. By exponentiating, we find that $e^{\log R} = R$ changes to $e^{\log R + 0.01} = e^{\log R}\, e^{0.01} = R\, e^{0.01} \approx R\,(1+0.01)$, which is a 1% increase in R.

Here's the conclusion: as R increases by 1%, there is an associated increase in Y of 0.01 b_J .

If, for example, $b_J = 25{,}400$, then a 1% increase in R is associated with an approximate increase in Y of 254.

We used an increase of 0.01 (rather than 1) to exploit the approximation $e^{0.01} \approx 1.01$.

Finally, suppose that both Y and J are obtained by taking logs. That is $Y = \log Z$ and $J = \log R$. What is the link between R and Z? Suppose we consider J increasing by 0.01; as in the previous note, this is approximately a 1% change in R.

As J increases by 0.01, there is an associated change from Y to $Y + 0.01\, b_J$. As $Y = \log Z$, we see that Z changes (approximately) to $Z(1+0.01\, b_J)$. Thus: as R increases by 1%, we find that there is an associated change in Z of 0.01 b_J , interpreted as a percent.

If, for example, $b_J = 1.26$, then a 1% increase in R is associated with an approximate increase of 1.26% in Z.

This document points out an interesting misunderstanding about multiple regression. There can be serious disagreement between

the regression coefficient b_H in the regression $\hat{Y} = b_0 + b_G\, G + b_H\, H$

and

the regression coefficient b_H in the regression $\hat{Y} = b_0 + b_H\, H$

While most people would not expect the values of b_H to be identical in these two regressions, it is somewhat shocking as to how far apart they can be.

Consider this very simple set of data with n = 20:

G	*H*	*Y*
73	7.3	3096
87	-6.0	3519
83	-3.7	3383
78	2.5	3261
82	-2.2	3360
80	0.7	3334
83	-2.9	3388
86	-6.2	3481
75	5.1	3120
82	-1.3	3378
80	0.8	3326
82	-2.4	3365
77	2.9	3215
81	-1.5	3306
79	1.1	3266
78	1.9	3229
76	3.5	3193
80	0.5	3315
80	-0.3	3280
81	-0.6	3335

Here is the regression of *Y* on (*G*, *H*) :

```
The regression equation is
Y = - 751 + 50.6 G + 20.5 H

Predictor       Coef       StDev          T        P
Constant      -751.2       515.9      -1.46    0.164
G             50.649       6.439       7.87    0.000
H             20.505       6.449       3.18    0.005

S = 13.63       R-Sq = 98.5%     R-Sq(adj) = 98.3%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    2      209106      104553     562.64    0.000
Error        17        3159         186
Total        19      212265
```

This shows a highly significant regression. The *F* statistic is enormous, and the individual *t* statistics are positive and significant.

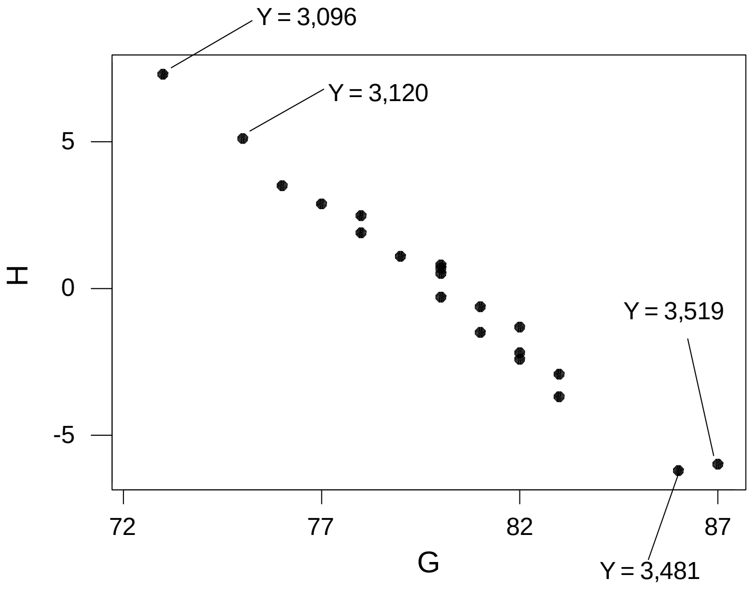
Now, suppose that you regressed *Y* on *H* only. You'd get the following:

```
The regression equation is
Y = 3306 - 29.7 H

Predictor       Coef       StDev          T        P
Constant     3306.31        6.38     518.17    0.000
H            -29.708       1.907     -15.58    0.000

S = 28.53       R-Sq = 93.1%     R-Sq(adj) = 92.7%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      197610      197610    242.71    0.000
Error        18       14655         814
Total        19      212265
```

This regression is also highly significant. However, it now happens that the relationship with *H* is significantly *negative*.

How could this possibly happen? It turns out that these data were strung out in the (*G*, *H*) plane with a negative relationship. The coefficient of *Y* on *G* was somewhat larger than the coefficient on *H*, so that when we look at *Y* and *H* alone we see a negative relationship.

The picture below shows the locations of the points in the (*G*, *H*) plane. The values of *Y* are shown at some extreme points, suggesting why the apparent relationship between *Y* and *H* appears to be negative.



The quantity S_{yy} = $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ measures variation in *Y*. Indeed we get s_y from this as s_y = $\sqrt{\frac{S_{yy}}{n-1}}$. We use the symbol $\hat{y}_i$ to denote the fitted value for point *i*.

One can show that $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ = $\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ + $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$. These sums have the names SS_{total}, $SS_{\text{regression}}$, and SS_{error} . They have other names or abbreviations. For instance

SS_{total} may be written as SS_{tot} .

$SS_{\text{regression}}$ may be written as SS_{reg} , SS_{fit} , or SS_{model} .

SS_{error} may be written as SS_{err} , SS_{residual} , SS_{resid} , or SS_{res} .

The degrees of freedom accounting is this:

SS_{total}	has *n* - 1 degrees of freedom
$SS_{\text{regression}}$	has *K* degrees of freedom (*K* is the number of independent variables)
SS_{error}	has *n* - 1 - *K* degrees of freedom

Here is how the quantities would be laid out in an analysis of variance table:

Source of Variation	Degrees of freedom	Sum of Squares	Mean Squares	F
Regression	*K*	$\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$	$\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{K}$	$\frac{MS_{\text{Regression}}}{MS_{\text{Error}}}$
Error	*n* - 1 - *K*	$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$	$\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-1-K}$	
Total	*n* - 1	$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$		

A measure of quality of the regression is the *F* statistic. Formally, this *F* statistic tests

H_0 : $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, …, $\beta_K = 0$ [Note that β_0 does not appear.]

versus

H_1 : at least one of β_1, β_2, β_3, …, β_K is not zero

Note that β_0 is not involved in this test.

Also, note that $s_\varepsilon = \sqrt{MS_{\text{Error}}}$ is the estimate of σ_ε. This has many names:

standard error of estimate
standard error of regression
estimated noise standard deviation
root mean square error (RMS error)
root mean square residual (RMS residual)

The measure called R^2 is computed as $\dfrac{SS_{\text{Regression}}}{SS_{\text{Total}}}$. This is often described as the "fraction of the variation in *Y* explained by the regression."

You can show, by the way, that

$$\frac{s_\varepsilon}{s_y} = \sqrt{\frac{n-1}{n-1-K}\left(1-R^2\right)}$$

The quantity $R^2_{adj} = 1 - \dfrac{n-1}{n-1-K}\left(1-R^2\right)$ is called the *adjusted R*-squared. This is supposed to adjust the value of R^2 to account for both the sample size and the number of predictors. With a little simple arithmetic,

$$R^2_{adj} = 1 - \left(\frac{s_\varepsilon}{s_y}\right)^2$$

This document considers the use of indicator variables, also called dummy variables, as predictors in multiple regression. Three situations will be covered.

EXAMPLE 1 gives a regression in which there are independent variables taking just two values. This is very easy.
EXAMPLE 2 gives a regression in which there is a discrete independent variable taking more than two values, but the values have a natural ordinal interpretation. This is also easy.
EXAMPLE 3 gives a regression with a discrete independent variable taking more than two values, and these values to not correspond to an ordering. This can get complicated.

EXAMPLE 1
Consider a regression in which the dependent variable SALARY is to be explained in terms of these predictors:

YEARS	years on the job
SKILLS	score on skills assessment (running from 0 to 40)
SUP	0 (not supervisor) or 1 (supervisor)
GENDER	0 (male) or 1 (female)

Suppose that the fitted regression turns out to be

$$\widehat{\text{SALARY}} = 16{,}000 + 1{,}680 \text{ YEARS}$$
$$+ 1{,}845 \text{ SKILLS} + 3{,}208 \text{ SUP} - 1{,}145 \text{ GENDER}$$

Suppose that all the coefficients are statistically significant, meaning that the *p*-values listed with their *t* statistics are all 0.05 or less. We have these very simple interpretations:

The value associated with each year on the job is $1,680 (holding all else fixed).
The value associated with each additional point on the skills assessment is $1,845 (holding all else fixed).
The value associated with being a supervisor is $3,208 (holding all else fixed).
The value associated with being female is -$1,145 (holding all else fixed).

The variables SUP and GENDER have conveniently been coded 0 and 1, and this makes the interpretation of the coefficients very easy. Variables that have only 0 and 1 as values are called *indicator* variables or *dummy* variables.

If the scores for such a variable are two other numbers, say 5 and 10, you might wish to recode them.

These might also be described as categorical variables with two levels.

In general, we will not offer interpretations on estimated coefficients that are not statistically significant.

EXAMPLE 2

Consider a regression in which the dependent variable HOURS (television viewing hours per week) is to be explained in terms of predictors

INCOME (in thousands of dollars)
JOB (hours per week spent at work)
FAM (number of people living in the household)
STRESS (self-reported level of stress, coded as
1 = none, 2 = low, 3 = some, 4 = considerable, 5 = extreme)

The variable STRESS is clearly categorical with five levels, and we are concerned about how it should be handled. The important feature here is that STRESS is an *ordinal* categorical variable, meaning that the (1, 2, 3, 4, 5) responses reflect the exact ordering of stress. Accordingly, you need not take any extra action on this variable; you can use it in the regression exactly as is.

If the fitted regression equation is

$$\widehat{\text{HOURS}} = -62.0 - 1.1 \text{ INCOME} - 0.1 \text{ JOB} + 2.4 \text{ FAM} - 0.2 \text{ STRESS}$$

then the interpretation of the coefficient on STRESS, assuming that this coefficient is statistically significant, is that each additional level of STRESS is associated with 0.2 hour (12 minutes) less time watching television.

It seems natural to encode STRESS with consecutive integers. These are some subtleties:

* If you replaced the codes (1, 2, 3, 4, 5) by (-2, -1, 0, 1, 2), the regression would produce exactly the same estimated coefficient -0.2. This replacement would alter the intercept however.

* If you replaced the codes (1, 2, 3, 4, 5) by (10, 20, 30, 40, 50), the regression coefficient would be produced as -0.02.

* If you do not like the equal-size spaces between the codes, you might replace (1, 2, 3, 4, 5) by (-3, -1, 0, 1, 3). The coefficient would now change from -0.2, and you’d have to rerun the regression to see what it would be.

EXAMPLE 3

We will consider next a data set on home prices with *n* = 370.

Variable	Interpretation	Average	Standard deviation
PRICE	Home price in dollars	154,422	14,883
STYLE	Home style, coded as 1 = split-level, 2 = ranch, 3 = colonial, 4 = Tudor	2.41	0.98
SIZE	Indoor area in square feet	2,007.5	320.9
BEDROOM	Number of bedrooms	3.29	0.61

The number of bedrooms is a small integer, and we can use it in the regression with no modification. The average and standard deviation are useful summaries for BEDROOM, but we might also be interested in a simple tally. The following was obtained in Minitab from **Stat** ⇒ **Tables** ⇒ **Tally Individual Variables**.

```
BEDROOM  Count
      2     21
      3    230
      4    109
      5     10
     N=    370
```

The variable STYLE is encoded as small integers, but the numbers function only as labels. Indeed, the information might have come to us as alphabetic names rather than these numbers. Note the inherent meaninglessness of the arithmetic

2 – 1 = ranch – split-level = 1 = 3 – 2 = colonial – ranch

From **Stat** ⇒ **Tables** ⇒ **Tally Individual Variables** for the variable STYLE we get this:

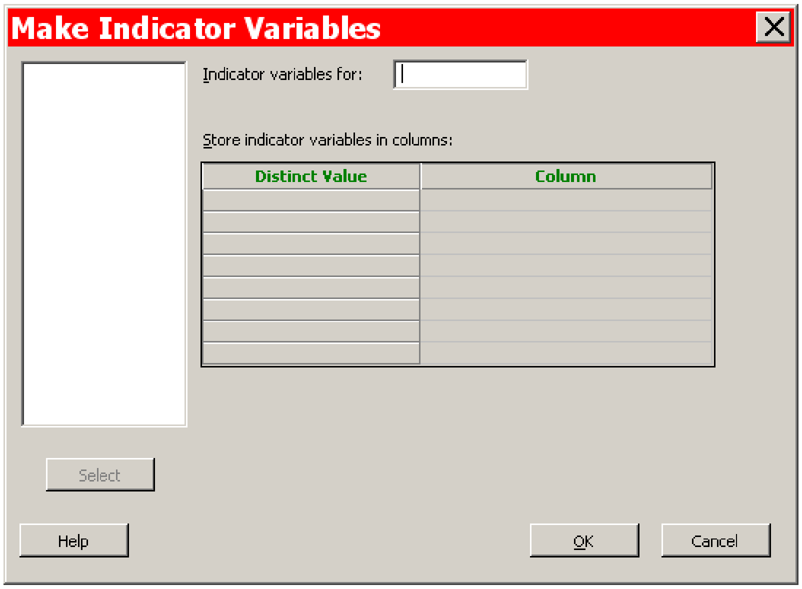
```
STYLE  Count
    1     85
    2     97
    3    141
    4     47
   N=    370
```

Since the numbers attached to STYLE do not mean anything, we cannot use this variable as presently structured.

> By the way, if you uncritically ran the regression of PRICE on (STYLE, SIZE, BEDROOMS) you'd get the fitted equation
>
> ```
> PRICE = 87443 + 5444 STYLE + 22.8 SIZE + 2436 BEDROOM
> ```
>
> and the coefficient on STYLE would be statistically significant. The interpretation would be that it's a $5,444 step up from split-level to ranch, also a $5,444 step up from ranch to colonial, and a $5,444 step up from ranch to Tudor. This is ridiculous.

If STYLE had only two values, we would be in the situation of EXAMPLE 1, and we could just use STYLE as an ordinary indicator (or dummy) variable. Here STYLE has four values, and we need a different method.

We will make a *set* of indicator variables for STYLE. In Minitab, do **Calc** ⇒ **Make Indicator Variables**. You will get this information panel:



Type STYLE in the location **Indicator variables for**, and the panel make a list in the **Distinct Value** column. The **Column** will then show new names STYLE_1 through STYLE_4. It would be convenient to overwrite these as SL, RANCH, COLONIAL, TUDOR.

Minitab will create four indicator (dummy) variable columns. In the column for SL, the value 1 will appear for any house that was a split-level, and the value 0 will appear for all other houses. In the column for RANCH, the value 1 will appear for any house that was a ranch, and the value 0 will appear for all other houses.

In each row of the data sheet, SL + RANCH + COLONIAL + TUDOR will be exactly 1. This just notes that each house is one, and only one, of the four styles.

The command **Calc** ⇒ **Make Indicator Variables** can be applied to a column of alphabetic information.

It seems natural now to run the regression of PRICE on (SL, RANCH, COLONIAL, TUDOR, SIZE, BEDROOM). Note that STYLE is not included.

If you do that, you'll get this message at the top of the Minitab run:

```
* TUDOR is highly correlated with other X variables
* TUDOR has been removed from the equation
```

This message happens because SL + RANCH + COLONIAL + TUDOR = 1 for every line of the data set. This creates total collinearity with the regression intercept, and the regression arithmetic is impossible. Minitab deals with this by removing the last-named variable involved. In this instance, TUDOR was named last and was eliminated.

Minitab then goes on to produce a useful regression run:

```
The regression equation is
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM - 21054 SL - 12504 RANCH
           - 12639 COLONIAL

Predictor        Coef     SE Coef          T        P
Constant       114696        4160      27.57    0.000
SIZE           21.832       1.993      10.96    0.000
BEDROOM          2682        1006       2.66    0.008
SL             -21054        1871     -11.26    0.000
RANCH          -12504        1821      -6.86    0.000
COLONIAL       -12639        1705      -7.41    0.000

S = 9882        R-Sq = 56.5%     R-Sq(adj) = 55.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5 46184185424  9236837085     94.58    0.000
Residual Error   364 35546964282    97656495
Total            369 81731149706
```

Parts of the output have been omitted.

The question now is the interpretation of the coefficients. For a split-level home, the indicators have values SL = 1, RANCH = 0, COLONIAL = 0. (Note that TUDOR has been omitted by Minitab). The fitted equation for a split-level home is then

```
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM - 21054           Split-Level
```

A ranch home has indicators SL = 0, RANCH = 1, COLONIAL = 0. This gives the fitted equation

```
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM - 12504           Ranch
```

Similarly, the fitted equation for colonial homes is

```
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM - 12639           Colonial
```

What about the Tudor homes? These have SL = 0, RANCH = 0, COLONIAL = 0, so that the fitted equation for these is

```
PRICE = 114696 + 21.8 SIZE + 2682 BEDROOM                   Tudor
```

The omitted indicator, here TUDOR, gives the base for interpreting the other estimated coefficients.

The suggestion is that a split-level home sells for 21,054 less than a Tudor home, holding all other variables fixed. A ranch sells for 12,504 less than a Tudor home, holding all other variables fixed. It follows that a ranch sells for 21, 054 - 12,504 = 8,550 more than a split-level, holding all other variables fixed.

If we had asked Minitab for the regression of PRICE on (SL, RANCH, TUDOR, SIZE, BEDROOM), we would have produced the following fitted equation:

```
PRICE = 102057 + 21.8 SIZE + 2682 BEDROOM - 8415 SL + 135 RANCH
               + 12639 TUDOR
```

This time the indicator for colonial was used as the baseline, and we see that the Tudor homes sell for 12,639 more than the colonial homes, holding all else fixed. Perfectly consistent.

The following display indicates exactly what happens as we change the baseline.

	Estimated coefficients			
Indicators used in the regression	SL	RANCH	COLONIAL	TUDOR
SL, RANCH, COLONIAL	-21,054	-12,504	-12,639	
SL, RANCH, TUDOR	-8,415	135		12,639
SL, COLONIAL, TUDOR	-8,550		-135	12,504
RANCH, COLONIAL, TUDOR		8,550	8,415	21,054

In all parts of this table, the other variables (SIZE, BEDROOM) were used as well.

All four lines of this table represent equivalent fits. All produce the same R^2, the same F statistic, and the same s_ε (S in Minitab). Moreover, the estimated coefficients on SIZE and BEDROOM will be the same in all four lines, as will the corresponding t statistics.

If you are using a set of indicator variables, and if you go through a variable-selection process to remove variables, you must keep the indicator set intact. In the context of this problem, that means that any fitted model must use either

three out of the four indicators

or

none of the indicators

The indicators only make solid good sense when used together.

The regression of PRICE on (SIZE, BEDROOM, SL, RANCH, COLONIAL) which we saw above had significant *t* statistics on all independent variables. We would not be tempted to remove any of them. Moreover, a stepwise regression would select all the predictors.

The regression of PRICE on (SIZE, BEDROOM, SL, RANCH, TUDOR) produces this:

```
The regression equation is
PRICE = 102057 + 21.8 SIZE + 2682 BEDROOM - 8415 SL + 135 RANCH
         + 12639 TUDOR

Predictor        Coef     SE Coef          T        P
Constant       102057        3674      27.78    0.000
SIZE           21.832       1.993      10.96    0.000
BEDROOM          2682        1006       2.66    0.008
SL              -8415        1365      -6.16    0.000
RANCH             135        1309       0.10    0.918
TUDOR           12639        1705       7.41    0.000
```

This suggests that we might remove the indicator for RANCH. Indeed, stepwise regression selects all the variables except RANCH.

So what's the problem? If we removed RANCH, the other estimated coefficients would change, and we would no longer be able to assess correctly the differences between the home styles.

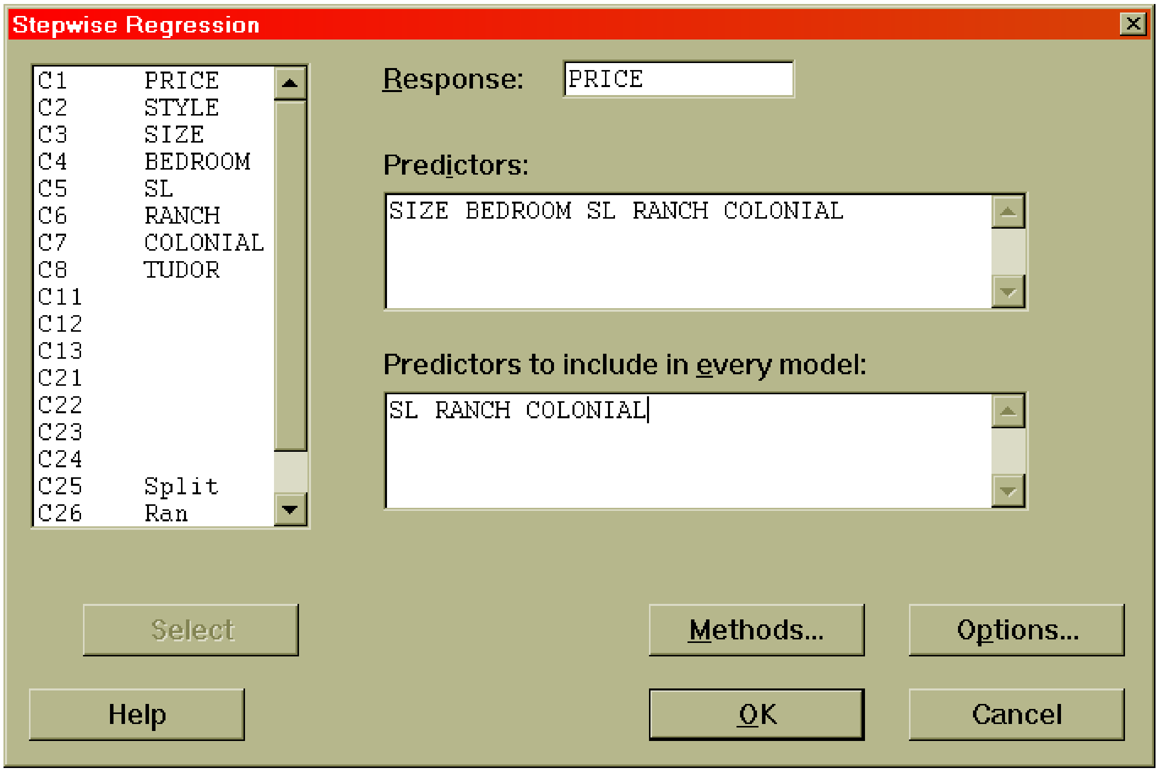
The advice, in generic form is this. If there are *K* indicators in a set, then a fitted model must use either

K – 1 of the indicators (leave out any one)

or

none of the indicators.

Specifying a model that has none of the indicators is easy. If you use a variable selection technique like stepwise regression or best subsets regression, you need a way to force the indicator set to stay together. Here is how you set that up for stepwise regression:



Response: PRICE

Predictors: SIZE BEDROOM SL RANCH COLONIAL

Predictors to include in every model: SL RANCH COLONIAL

Finally, we need an objective method to test whether an indicator variable set should be used at all. Let’s consider the context of our model, namely

$$\text{PRICE}_i = \beta_0 + \beta_{\text{SIZE}}\,\text{SIZE}_i + \beta_{\text{BEDROOM}}\,\text{BEDROOM}_i$$

$$+ \beta_{\text{SL}}\,\text{SL}_i + \beta_{\text{RANCH}}\,\text{RANCH}_i + \beta_{\text{COLONIAL}}\,\text{COLONIAL}_i + \varepsilon_i$$

The decision about whether or not to use the style indicators is really a test of the null hypothesis H_0: $\beta_{\text{SL}} = 0$, $\beta_{\text{RANCH}} = 0$, $\beta_{\text{COLONIAL}} = 0$.

There is a method for testing whether a *set* of coefficients is all zero. This method works for situations beyond what we are testing here. This requires the computation of this *F* statistic:

$$\frac{\left\{\left[\begin{array}{l}\text{Regression sum of squares}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{array}\right] - \left[\begin{array}{l}\text{Regression Sum of Squares}\\ \text{using SIZE, BEDROOM}\end{array}\right]\right\} \div \left\{\begin{array}{l}\text{Number of coefficients}\\ \text{being investigated}\end{array}\right\}}{\left[\begin{array}{l}\text{Residual Mean Square}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{array}\right]}$$

This is to be interpreted as an *F* statistic. We need to identify the two degrees of freedom numbers associated with *F*.

The numerator degrees of freedom is “Number of coefficients being investigated” in the calculation above.

The denominator degrees of freedom is the DF for residual in the regression on (SIZE, BEDROOM, SL, RANCH, COLONIAL).

The regression on (SIZE, BEDROOM, SL, RANCH, COLONIAL) had this analysis of variance table:

```
Analysis of Variance
Source            DF           SS          MS         F        P
Regression         5  46184185424  9236837085     94.58    0.000
Residual Error   364  35546964282    97656495
Total            369  81731149706
```

The regression sum of squares is 46,184,185,424. The residual mean square is 97,656,495. We note also that the degrees of freedom in the residual line is 364.

The regression on just (SIZE, BEDROOM) will have this analysis of variance table:

```
Analysis of Variance
Source            DF          SS          MS         F        P
Regression         2 33675069487 16837534743    128.59    0.000
Residual Error   367 48056080220   130942998
Total            369 81731149706
```

The regression sum of squares is 33,675,069,487.

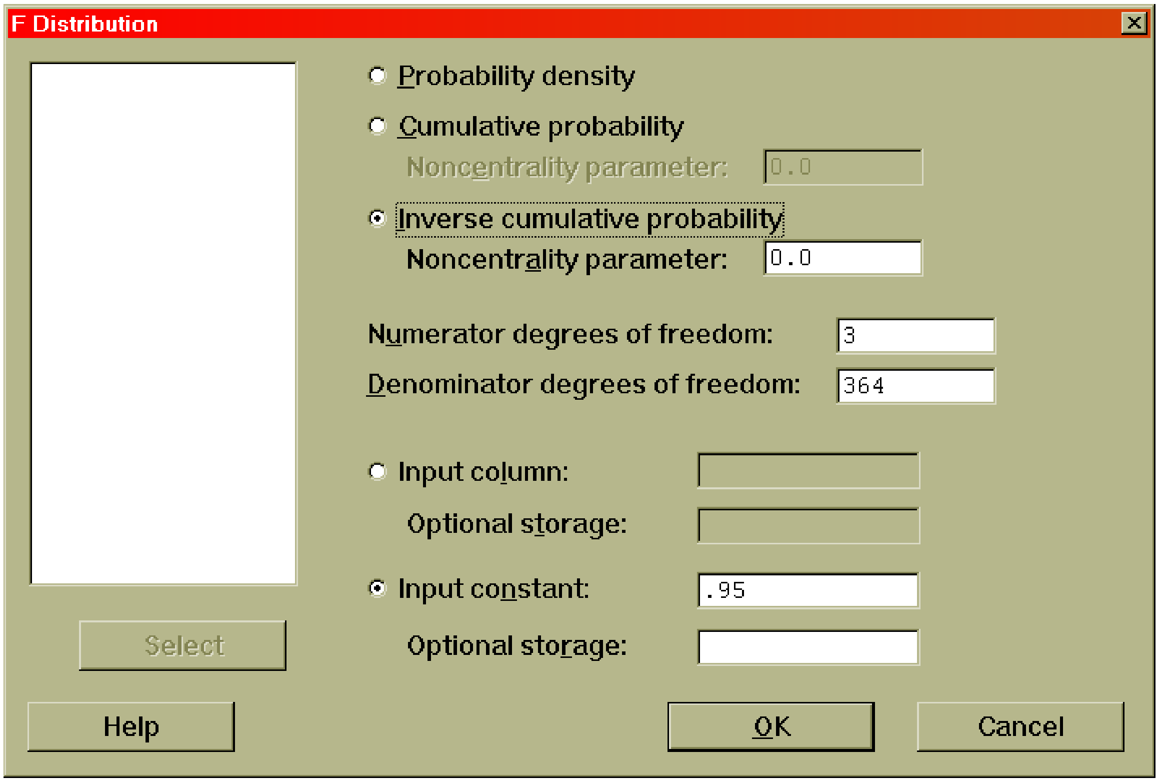
We'll note that three coefficients are under test. We now have enough information to assemble the test statistic:

$$\frac{\{46{,}184{,}185{,}424 \; - \; 33{,}675{,}069{,}487\} \; \div \; 3}{97{,}656{,}495} \; \approx \; 42.70$$

Minitab does not have a procedure for computing this number. The user needs to assemble it.

So what do we do with this number? The null hypothesis above should be rejected at the 0.05 level of significance if this exceeds $F_{3,\,364}^{0.05}$, the upper 5% point for the *F* distribution with (3, 364) degrees of freedom. It happens that $F_{3,\,364}^{0.05}$ = 2.6294. Since our computed statistic, 42.70 exceeds 2.6294, we would reject the null hypothesis that all the coefficients of the style indicators are zero. It appears that the style indicators are useful as predictors of home price.

You can find this cutoff point for the *F* distribution from Minitab. Just do **Calc** ⇒ **Probability Distributions** ⇒ **F**, and then fill in the resulting panel as follows:



○ Probability density

○ Cumulative probability

Noncentrality parameter: 0.0

● Inverse cumulative probability

Noncentrality parameter: 0.0

Numerator degrees of freedom: 3

Denominator degrees of freedom: 364

○ Input column:

Optional storage:

● Input constant: .95

Optional storage:

This particular *F* test had been defined through this statistic:

$$\frac{\left\{\begin{bmatrix}\text{Regression sum of squares}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{bmatrix} - \begin{bmatrix}\text{Regression Sum of Squares}\\ \text{using SIZE, BEDROOM}\end{bmatrix}\right\} \div \begin{Bmatrix}\text{Number of coefficients}\\ \text{being investigated}\end{Bmatrix}}{\begin{bmatrix}\text{Residual Mean Square}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{bmatrix}}$$

You will sometimes see this in the exactly equivalent form

$$\frac{\left\{\begin{bmatrix}\text{Residual sum of squares}\\ \text{using SIZE, BEDROOM}\end{bmatrix} - \begin{bmatrix}\text{Residual Sum of Squares}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{bmatrix}\right\} \div \begin{Bmatrix}\text{Number of coefficients}\\ \text{being investigated}\end{Bmatrix}}{\begin{bmatrix}\text{Residual Mean Square}\\ \text{using SIZE, BEDROOM,}\\ \text{SL, RANCH, COLONIAL}\end{bmatrix}}$$

This equivalent form lays out the arithmetic as

$$\frac{\{48,056,080,220 \;-\; 35,546,964,282\} \;\div\; 3}{97,656,495} \;\approx\; 42.70$$

This produces exactly the same number, as it must.

Original documents (not part of the formal handout)

introthoughts.doc
grossSize.doc
DataCleaning.doc
coefintr.doc
regpath.doc
anova.doc
indicator.doc (has a few things on variable selection)