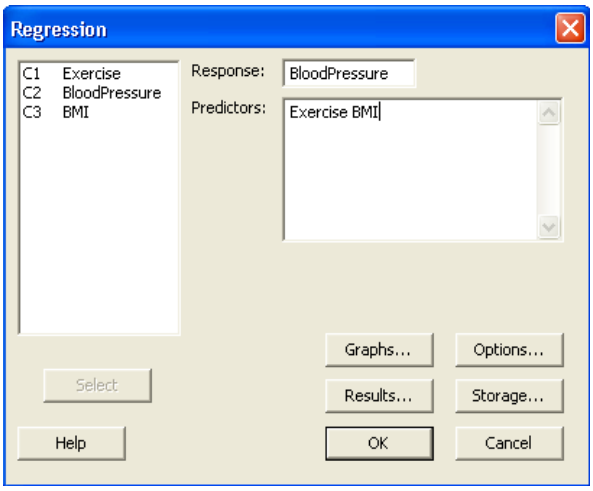
Suppose we are interested in how the exercise and body mass index affect the blood pressure. A random sample of 10 males 50 years of age is selected and their height, weight, number of hours of exercise and the blood pressure are measured. Body mass index is calculated by the following formula: $BMI\ (kg/m^2) = \frac{(Weight\ in\ pounds * 703)}{Height\ in\ Inches^2}$.

Worksheet 1 ***

	C1	C2	C3	C4	C5
	Exercise	BloodPressure	BMI		
1	4	120	18.4		
2	10	110	20.1		
3	2	120	22.4		
4	3	135	25.9		
5	3	140	26.5		
6	5	115	28.9		
7	1	150	30.4		
8	2	165	32.9		
9	2	160	33.0		
10	0	180	34.7		
11					
12					



The variables in our current model are all quantitative variables. We can include qualitative variables, which have no natural scale of measurement, as indicator variables. These variables are assigned a set of values that account for the effect the variable may have on the response.

Indicator variables, also known as dummy variables, usually take on the values of 0 and 1, to indicate whether an observation does (1) or does not (0) belong in a certain category. The general model with one explanatory variable and an indicator D:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D + \varepsilon$$

When an individual falls in a certain group and D=1, the model is

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

and when they do not fall in the group, D=0 and the model is

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

The following example will help to identify how to run and interpret the Minitab results for indicator variables.

Using the data set listed at the start of this tutorial, consider the variable employment status (where 1=employed and 0=not employed). The Minitab sheet would look like:

Worksheet 1 ***

	C1	C2	C3	C4	C5
	Exercise	BloodPressure	BMI	Employment	
1	4	120	18.4	0	
2	10	110	20.1	0	
3	2	120	22.4	1	
4	3	135	25.9	1	
5	3	140	26.5	0	
6	5	115	28.9	1	
7	1	150	30.4	1	
8	2	165	32.9	0	
9	2	160	33.0	0	
10	0	180	34.7	1	
11					
12					

To run the regression with the indicator variable click **Stat-Regression-Regression…** and select the "Response:" as you dependent variable and "Predictors" as your independent variables (including your indicator variable). Then click "OK" to run the regression. The output will look like

```
The regression equation is
BloodPressure = 85.7 - 3.83 Exercise + 2.65 BMI - 12.7 Employment


Predictor      Coef  SE Coef      T      P
Constant      85.72    26.42   3.24  0.018
Exercise     -3.830    1.717  -2.23  0.067
BMI          2.6495   0.7988   3.32  0.016
Employment  -12.701    7.110  -1.79  0.124


S = 10.3928   R-Sq = 87.2%   R-Sq(adj) = 80.8%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  4424.4  1474.8  13.65  0.004
Residual Error   6   648.1   108.0
Total            9  5072.5
```

The significance of the regression is determined in the same way. The null hypothesis (H_0: $\beta_1 = \beta_2 = \beta_3 = 0$) is significant in comparison to the alternative hypothesis (H_A: $at\ least\ one\ \beta_k \neq 0$) because p=0.004<0.05. This means at least one of the three variables is significant.

We can individually determine whether or not each variable is significant by looking at their T and P-values in the top part of the output. For example, the indicator variable employment is insignificant (p-value=0.124>0.05), exercise is insignificant (p-value=0.067>0.05), and BMI is significant (p-value=0.016<0.05) for the given model.

If the indicator variable had been found significant it could be interpreted through a 95% confidence interval estimate for β_3 (our indicator variable). The equation for this confidence interval is $b_3 \pm t_{\frac{\alpha}{2},n-K-1} s_{b_3}$. In our example if we wanted a 95% CI,

-12.701±(2.447)(7.110)=(-30.099,4.697). So we are 95% confident that going from unemployed to employed changes your BMI by (-30.099, 4.697). [Note that this range does not make much sense, which is because this is not a significant factor in our model.]

If all the variables in our model were significant it would create two regression equations (because of our indicator variable's two options: 0 and 1). The general equation would be:

$$\hat{y} = 85.72 - 3.83x_1 + 2.6495x_2 - 12.701D$$

The two equations come from entering D=0 or D=1 into the general equation to produce:

Unemployed (D=0):

$$\hat{y} = 85.72 - 3.83x_1 + 2.6495x_2$$

Employed (D=1):

$$\hat{y} = 73.019 - 3.83x_1 + 2.6495x_2$$

The employed observation (person) has a lower blood pressure by about 12.701 mmHg than an unemployed observation (person).

Interaction Terms (Interaction of Variables with Indicator Variable)

If it is expected that the two models differ in both intercept and slope, it can be modeled with one equation by including an interaction variable. The model statement could look something like

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 x_1 D + \varepsilon.$$

To test whether there is any different between the two categories (in intercept or slope), a partial F test of the hypothesis H_0: $\beta_2 = \beta_3 = 0$ versus H_A: $at\ least\ one\ of\ the\ \beta_i \neq 0$ can be used. The partial F will be used where L represents the number of variables in the reduced model and K represents the number of variables in our full model.:

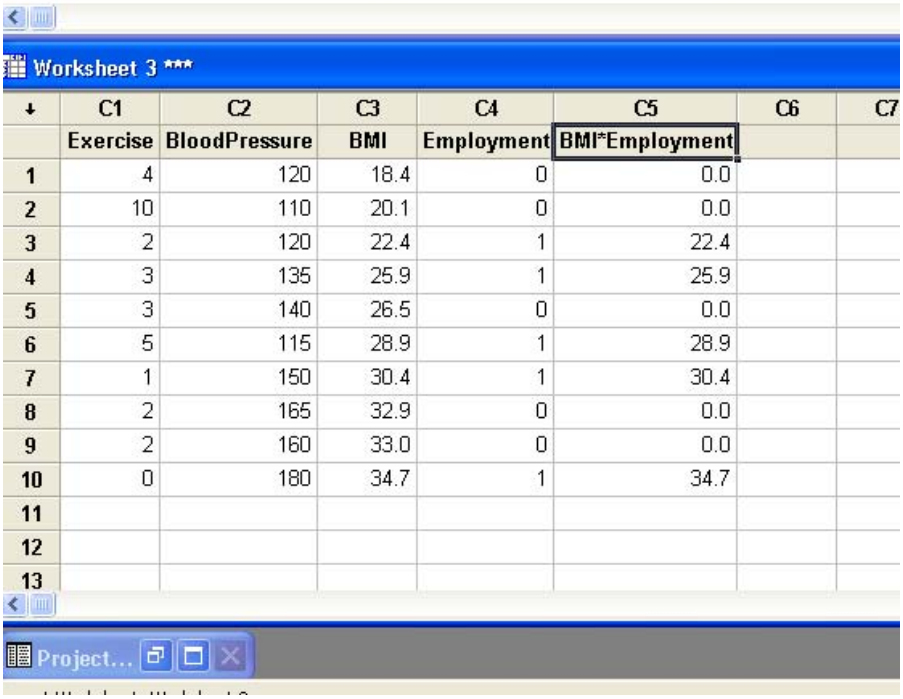
$$F_{partial} = \frac{(SS_E(reduced) - SS_E(full))/(K - L)}{SS_E(full)/(n - K - 1)}$$

The decision rule for the Partial F test is:

Reject H_0 if F > F(α; K-L, n-K-1)
Fail to reject H_0 if F ≤ F(α; K-L, n-K-1)

This requires a simple regression analysis on just the x_1 (reduced), and a multiple regression analysis to be run on $x_1, D,$ and $x_1 D$. The term $x_1 D$ can be built on our own by using the **Calc-Calculator…** function (see the Multiple Regression tutorial for information on how to do this step-by-step).

We will use our example on exercise, BMI, and employment in regards to blood pressure. For the sake of simplicity we will say we are interested in only looking at BMI, employment, and the interaction (BMI*Employment) between the two at the α=0.05 level. Then after adding the interaction variable the screen should look something like:

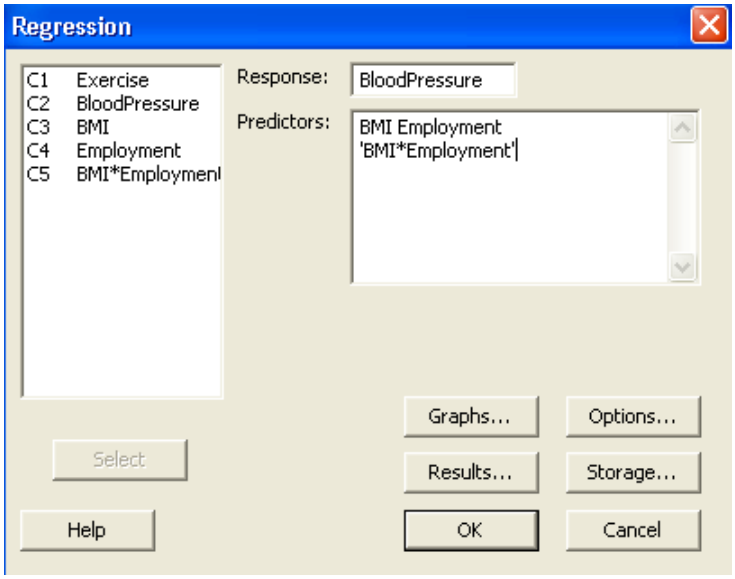


Worksheet 3 ***

	C1	C2	C3	C4	C5	C6	C7
	Exercise	BloodPressure	BMI	Employment	BMI*Employment		
1	4	120	18.4	0	0.0		
2	10	110	20.1	0	0.0		
3	2	120	22.4	1	22.4		
4	3	135	25.9	1	25.9		
5	3	140	26.5	0	0.0		
6	5	115	28.9	1	28.9		
7	1	150	30.4	1	30.4		
8	2	165	32.9	0	0.0		
9	2	160	33.0	0	0.0		
10	0	180	34.7	1	34.7		
11							
12							
13							

Project...

Next click **Stat-Regression-Regression…** and enter the information in so that Blood Pressure is the Response variable and the Predictors are BMI, Employment, and 'BMI*Employment'. Then click OK to run this regression as the full model. The window will look something like this screenshot: (Note that you can also put BMI|Employment in predictors which will give exactly same results.)



Part of the output will include an ANOVA table. For this example the ANOVA table looks like:

```
The regression equation is
BloodPressure = 50.1 + 3.39 BMI - 39.6 Employment + 1.15 BMI*Employment


Predictor         Coef  SE Coef      T      P
Constant         50.14    26.54   1.89  0.108
BMI            3.3941   0.9871   3.44  0.014
Employment      -39.61    49.75  -0.80  0.456
BMI*Employment   1.155    1.765   0.65  0.537


S = 13.5793   R-Sq = 78.2%   R-Sq(adj) = 67.3%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       3  3966.1  1322.0  7.17  0.021
Residual Error   6  1106.4   184.4
Total            9  5072.5
```

Next we will do the same steps, but we will only be testing the non-indicator variable of BMI (reduced model). The output for this looks like:

```
The regression equation is
BloodPressure = 41.0 + 3.61 BMI

Predictor    Coef  SE Coef     T      P
Constant    40.98    21.07  1.94  0.088
BMI        3.6060   0.7569  4.76  0.001

S = 12.8545   R-Sq = 73.9%   R-Sq(adj) = 70.7%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  3750.6  3750.6  22.70  0.001
Residual Error   8  1321.9   165.2
Total            9  5072.5
```

Now we will calculate our partial F:

$$F_{partial} = \frac{(SS_E(reduced) - SS_E(full))/(K-L)}{SS_E(full)/(n-K-1)} = \frac{(1321.9 - 1106.4)/(3-1)}{1106.4/(10-3-1)} = 0.584$$

The critical value is F(α; K-L, n-K-1) = F(0.05; 3-1, 10-3-1) = F(0.05; 2, 6) = 5.14.

Since 0.584<5.14, we fail to reject the null hypothesis. This means we cannot say that at least one of the coefficients does not equal 0.

Using the information from the full model we can still determine if the indicator or interaction varialbes were significant using the t-values and p-values similar to earlier in the tutorial (neither are significant—employment has p=0.456 and interaction has p=0.537). Thus the final model can be written as $\hat{y}$ = 41.0 +3.61 BMI (X1)