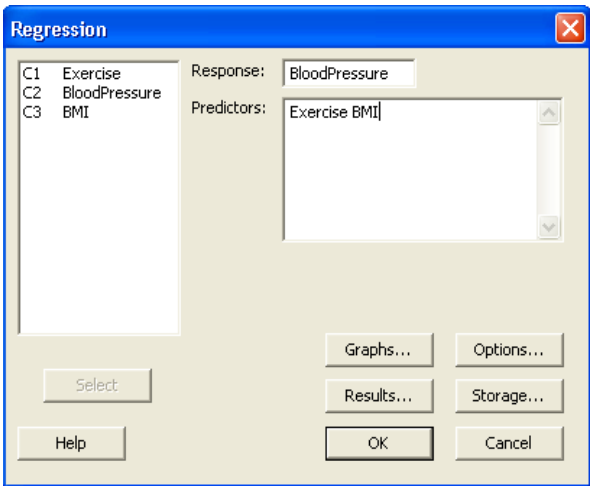
Suppose we are interested in how the exercise and body mass index affect the blood pressure. A random sample of 10 males 50 years of age is selected and their height, weight, number of hours of exercise and the blood pressure are measured. Body mass index is calculated by the following formula: $BMI\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

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					



Select **Stat-Regression-Regression…** from the pull-down menu.
Placing the variable we would like to predict, blood pressure, in the "Response:" and the variable we will use for prediction, exercise and body mass index in the "Predictors:" box. Click OK.
This generates the following Minitab output.

```
The regression equation is
BloodPressure = 74.5 - 2.84 Exercise + 2.71 BMI


Predictor     Coef  SE Coef      T      P
Constant     74.49    29.41   2.53  0.039
Exercise    -2.836    1.861  -1.52  0.171
BMI         2.7119   0.9144   2.97  0.021


S = 11.9087   R-Sq = 80.4%   R-Sq(adj) = 74.8%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        2  4079.8  2039.9  14.38  0.003
Residual Error    7   992.7   141.8
Total             9  5072.5
```

The interpretation of R^2 is same as before. We can see that 80.4% of the variation in Y is explained by the regression line. The fitted regression model found from the output is (Blood Pressure) = 74.5 - 2.84 * Exercise + 2.71 * BMI.

The next part of the output is the statistical analysis (ANOVA-analysis of variance) for the regression model. The ANOVA represents a hypothesis test with where the null hypothesis is

$H_o : \beta_i = 0$ for all i (In simple regression, i = 1)

$H_A : \beta_i \neq 0$ for at least 1 coefficient

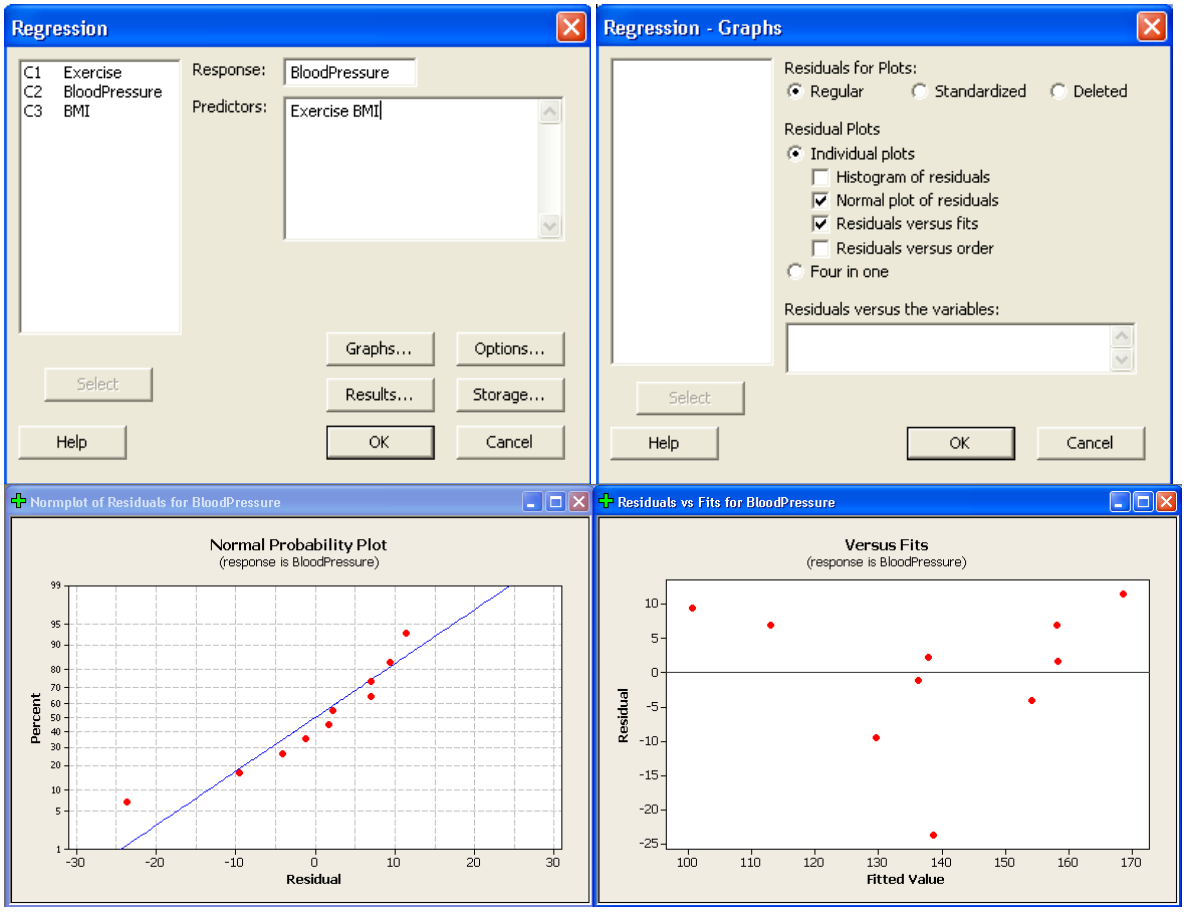
In this example, p-value for this overall test is .003 concluding at least one of independent variables is significantly meaningful to explain the blood pressure.

The individual t-test can also be performed.

$$H_0: \beta_1 = 0, H_A: \beta_1 \neq 0$$
$$H_0: \beta_2 = 0, H_A: \beta_2 \neq 0$$

In this example p-value is .171 and .021. Thus, β_1 is not significantly different from zero when body mass index is in the model, and β_2 is significantly different from zero when body mass index is in the model.

Model assumption checking and prediction interval can be done in the similar manner as the simple regression analysis. Normal probability plot and residual plot can be obtained by clicking the "Graphs" button in the "Regression" window, then checking the "Normal plot of residuals" and "Residuals versus fits" boxes. Click OK to exit the graphs window, click OK again to run the test.



Full and Reduced Models

Sometimes in multiple regression analysis, it is useful to test whether subsets of coefficients are equal to zero. To do this a partial F test will be considered. This answers the question, "Is the full model better than the reduced model at explaining variation in y?" The following hypotheses are considered:

$$H_0: \beta_{L+1} = \cdots = \beta_K = 0$$
$$H_A: at\ least\ one\ of\ the\ coefficients\ \beta_{L+1}, \ldots, \beta_K\ is\ not\ equal\ to\ zero$$

Where L represents the number of variables in the reduced model and K represents the number of variables in our full model. Rejecting H_0 means the full model is preferred over the reduced model, whereas not rejecting H_0 means the reduced model is preferred.The *Partial F* is used to test these hypotheses and is given by

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

Note that the denominator is the MS_E from the full model's output.

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)

In our example above we might consider comparing the full model with exercise and BMI predicting blood pressure to a reduced model of the BMI predicting blood pressure at the α=0.05 level. To calculate the partial F we will run the regression for the reduced model by clicking **Stat-Regression-Regression…** and putting BloodPressure in as the "Response:" variable and BMI in as the "Predictors:" variable and click "OK." The reduced model output reads:

```
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
```

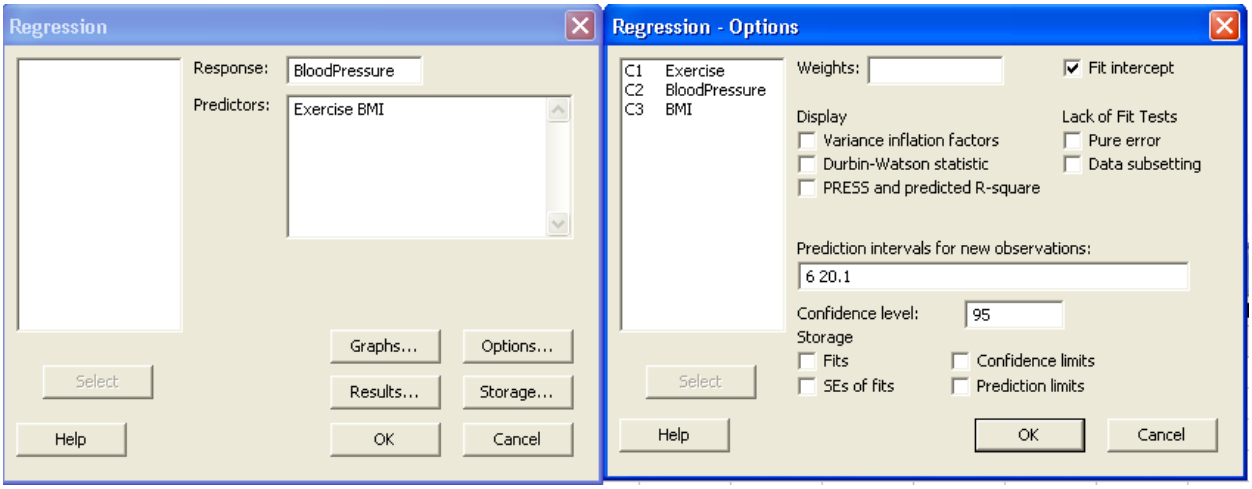
To calculate the partial F we will use the output for the full model found on page 1 and the reduced model output above.

$$F_{partial} = \frac{(SS_E(reduced) - SS_E(full))/(K-L)}{SS_E(full)/(n-K-1)} = \frac{(1321.9 - 992.7)/(2-1)}{(992.7)/(10-2-1)} = 2.32$$

Then use an F-table to look up the value for F(α; K-L, n-K-1) = F(0.05; 1, 7) = 5.59. According to our decision rule, 2.32 ≤ 5.59. This means we fail to reject H_0. This means that at the α=0.05 level we have found evidence that the reduced model is more efficient at explaining the games won.

Calculating Confidence Intervals and Prediction Intervals

Calculating CI and PI for multiple regressions are fairly similar to simple linear regressions. For multiple regressions you can create the intervals for your model based on the predictor variables. Consider the full model from earlier in this tutorial. We can predict the CI and PI for 6 hours of exercise and a BMI of 20.1 by entering the values in as seen below after clicking **Stat-Regression-Regression-Options…** to get to the window.



Then press "OK" and "OK" to run the regression analysis. The output will now include:

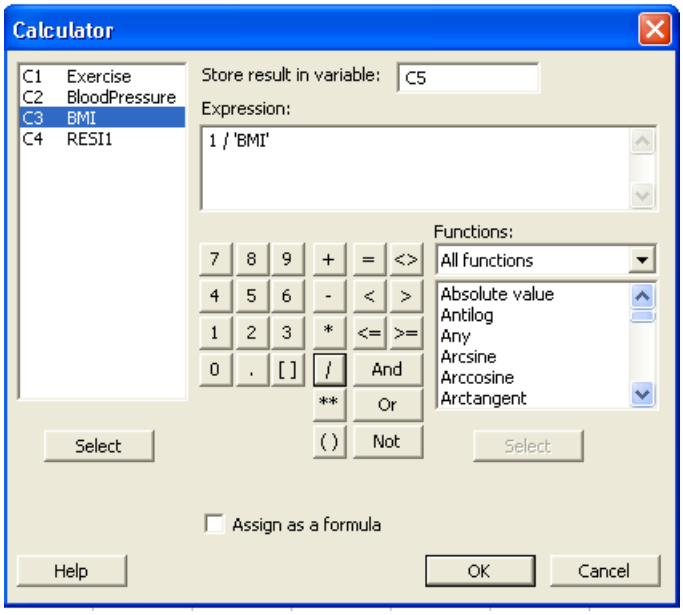
```
Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI            95% PI
      1  111.98    6.38  (96.88, 127.08)  (80.03, 143.93)


Values of Predictors for New Observations

New Obs  Exercise   BMI
      1      6.00  20.1
```

The 95% CI for this combination is (96.88, 127.08) and the 95% PI is (80.03, 143.93). The values entered can be seen at the bottom of the output to ensure each variable was correctly entered and not accidentally switched around or mistyped.

Transformation of Variables

It is not always obvious what to do when your model does not fit well. Transformations may be the easiest way to produce a better fit, especially when collecting more data is not feasible. Options to try include polynomial regression, inverse transformation, log transformation of the explanatory variable, log transformation of dependent and explanatory variable, and many more transformations. This tutorial will look at creating an inverse transformation for the model and storing this information in your Minitab sheet.

Click **Calc-Calculator…** and enter your information in the appropriate spaces in the window that pops up.



Choose a column without data stored in it to store your inverse transformation. In the expression blank enter the appropriate algebra and functions for your transformation. Then press "OK" and your transformation will then appear in your Minitab sheet. Use this variable instead of the original variable in your regression to see if the model becomes a better fit.

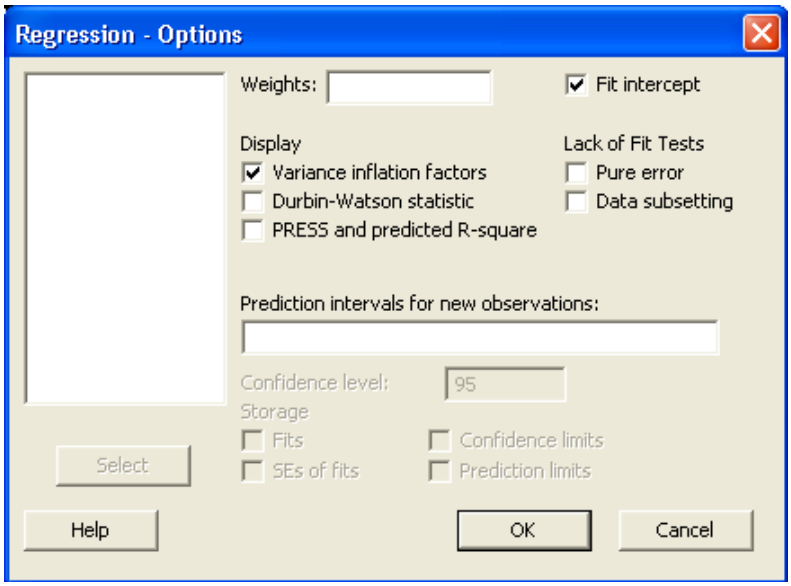
A Potential Problem with Multiple Regression

When explanatory variables are correlated with one another, the problem of *multicollinearity*, or near-linear dependence among regression variables, is said to exist. When a high degree of multicollinearity exists, the variance of the regression coefficients are infalted. This can lead to small *t*-values and unstable regression coefficients. Multicollinearity does not affect the ability to obtain a good fit to the regression (R^2) or the quality of forecasts or predictions from the regression.

The way to determine if this is a problem with your model is to look at the Variance Inflation Factors (VIFs). The equation for the VIF for the *j*th regression coefficent x_j can be written as $VIF_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of multiple determination obtained by performing the

regression of x_j on the remaining K-1 regressor variables. Any individual VIF larger than 10 should indiciate that multicollinearity is present.

To check for VIFs in Minitab click **Stat-Regression-Regression…** from the drop-down menu. Next click the Options button. Then check the "Variance inflation factors" box under Display, click OK. Then click OK again to run the test.



The data created in the output will look identical to the data collected before, except the table of coefficient will contain an additional column of information:

Predictor	Coef	SE Coef	T	P	VIF
Constant	74.49	29.41	2.53	0.039	
Exercise	-2.836	1.861	-1.52	0.171	1.701
BMI	2.7119	0.9144	2.97	0.021	1.701

Seeing the both our VIFs are below 10 we assume multicollinearity is not present in our model.

Correcting Multicollinearity

In order to correct for multicollinearity you need to remove the variables that are highly correlated with others. You can also try to add more data, this might break the pattern of multicollinearity.

There are drawbacks to these solutions. If you remove a variable you will obtain no information on the removed variable. So choosing which correlated variable to remove can be difficult. If you add more data the multicollinearity won't always disappear, and sometimes it is impossible to add more data (due to budget restraints or lack of data beyond what is known).

If a regression model is used strictly for forecasting, corrections may not be necessary.