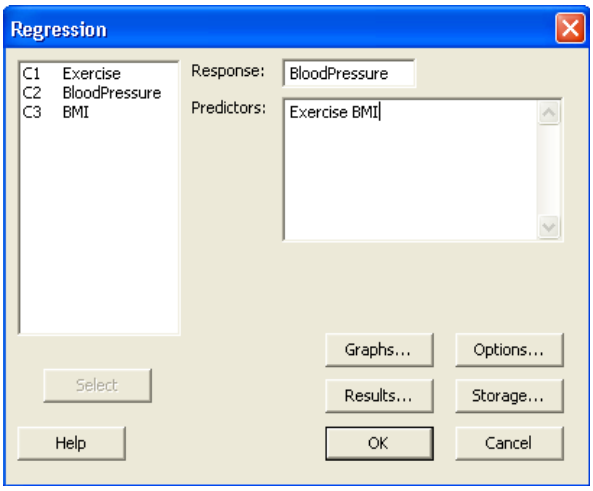
Suppose we are interested in how the exercise and body mass index affect the blood pressure. A random sample of 10 males 50 years of age is selected and their height, weight, number of hours of exercise and the blood pressure are measured. Body mass index is calculated by the following formula: $BMI\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

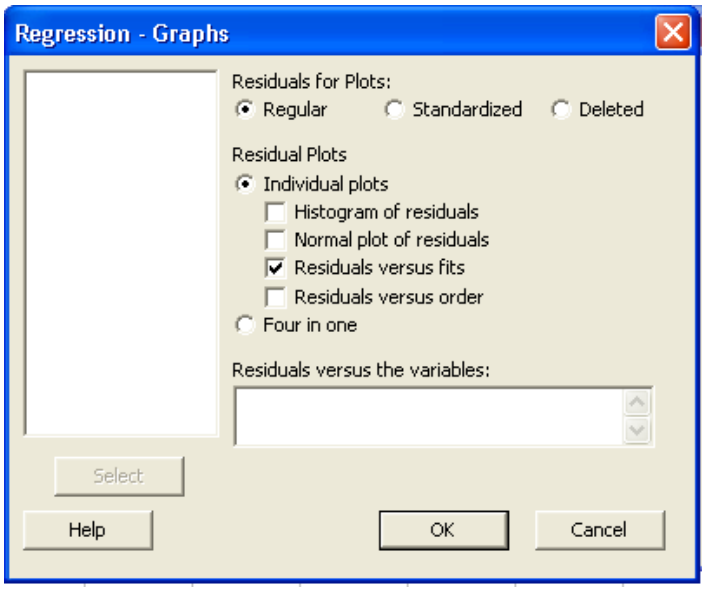
Worksheet 1 ***

	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					

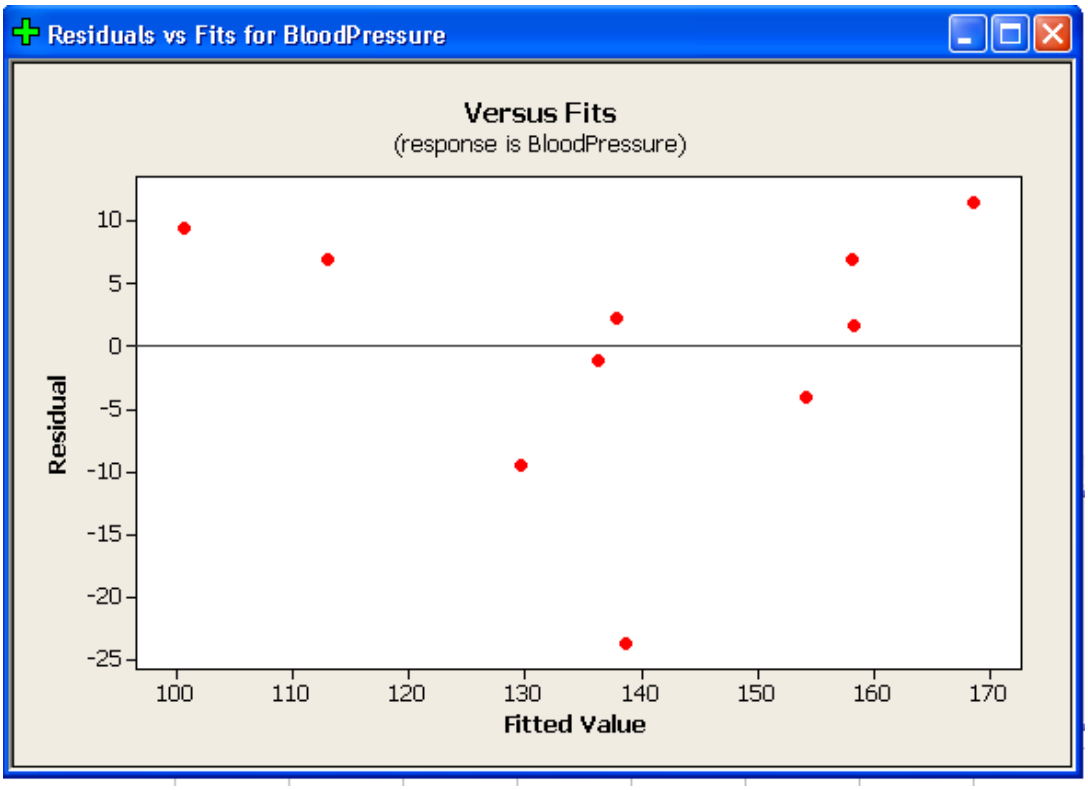


Residual plots are frequently used to identify violations to the assumptions of regression models. When all of the assumptions of the model hold, residuals should be randomly distributed about their mean (0) and there should be no systematic pattern. If there is a pattern or the residuals in the plot are not randomly distributed about their mean it does not mean the data is not linear, it means the data doesn't fit the model you are testing for (ex. cubic, log, linear, etc.).

In Minitab you can calculate and save as well as graph your residuals easily within the choices provided within the regression window. To save the residuals click **Stat-Regression-Regression…** and a window like that seen above will appear. Click the "Graphs…" button.



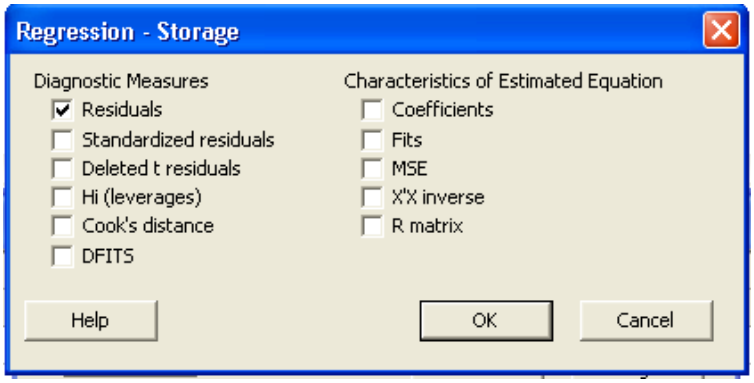
Select the appropriate graphs you would like Minitab to produce. Selecting "Four in one" will produce all four individual plots in one window for easy comparison. For the purpose of looking at residuals to determine if there is a pattern of it is randomly distributed around the mean we will look at the "Residuals versus fits" option. Once your options are selected click "OK" then click "OK" on the main regression window to run the regression. It will produce a graph that looks like:



This graph can then be interpreted to determine if it looks like the assumptions are met. In this case the residuals do look fairly randomly spread about the mean even though there appears to be an outlier around (138, -24).

Storing Residuals

Residuals can also be stored in Minitab. To do this click **Stat-Regression-Regression…** and then click "Storage" which will open the window that looks like:



Select “Residuals” under “Diagnostic Measures” and click “OK.” Then click “OK” to run the regression which will then store your residuals in an empty column by your data.

Worksheet 1 ***

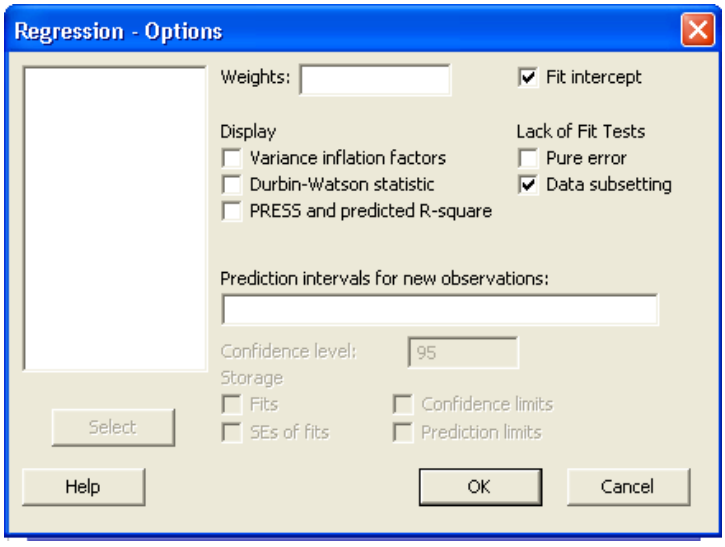
	C1	C2	C3	C4	C5
	Exercise	BloodPressure	BMI	RESI1	
1	4	120	18.4	6.9585	
2	10	110	20.1	9.3642	
3	2	120	22.4	-9.5609	
4	3	135	25.9	-1.2164	
5	3	140	26.5	2.1565	
6	5	115	28.9	-23.6800	
7	1	150	30.4	-4.0917	
8	2	165	32.9	6.9647	
9	2	160	33.0	1.6935	
10	0	180	34.7	11.4114	
11					
12					
13					

Testing Lack of Fit

If your data appears to not be randomly distributed or appears to have a systemic pattern you can test your data for a lack of fit for your model. The lack of fit test considers the following hypotheses:

H_0: The model is a good fit.
H_A: The model is not a good fit.

There are two options in Minitab, pure error which requires at least one replicate of an explanatory variable or data subsetting. To access these tests click **Stat-Regression-Regression…-Options…** and select the test you want. Click “OK” and then “OK” to run the test.



Your regression output will now include:

```
Lack of fit test
Possible interaction in variable Exercise  (P-Value = 0.031 )

Overall lack of fit test is significant at P = 0.031
```

We reject the null hypothesis since $p<0.05$. This means that the model is not a good fit. A better fit might be possible through variable transformation.

Influential Points

Influential points/observations are particular points that lie far from other data in a horizontal direction on the graph. Many of these points are considered outliers, which have large residual values. These points may have a significant impact on the slope of the regression line, creating a poor fit. Removing these variables or transforming the data may be able to remove or lessen the impact of this "influential" point. If influential points are removed, it is important to review the new graph to ensure that no new influential points are present.

Influential points can be determined more quantitatively by using DFITS and Cook's D statistics. Cook's D is determined to be a significant if a value differs significantly from the other Cook's D statistics, or using the general idea that a $D>4/n$ is an influential point. A point is considered an influential point if DFITS$>2/\sqrt{k/n}$. To access these tests click **Stat-Regression-Regression…-Storage** and check "Cook's distance" and "DFITS" and run the regression.

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6
	Exercise	BloodPressure	BMI	RESI1	COOK1	DFIT1
1	4	120	18.4	6.9585	0.19590	0.74520
2	10	110	20.1	9.3642	3.34617	3.79581
3	2	120	22.4	-9.5609	0.19896	-0.77394
4	3	135	25.9	-1.2164	0.00052	-0.03671
5	3	140	26.5	2.1565	0.00147	0.06174
6	5	115	28.9	-23.6800	0.53912	-2.31346
7	1	150	30.4	-4.0917	0.00969	-0.15951
8	2	165	32.9	6.9647	0.03995	0.33098
9	2	160	33.0	1.6935	0.00244	0.07943
10	0	180	34.7	11.4114	0.19621	0.78894
11						
12						
13						

Project...