

# MULTIPLE LINEAR REGRESSION IN MINITAB

This document shows a complicated Minitab multiple regression. It includes descriptions of the Minitab commands, and the Minitab output is heavily annotated.

Comments in { } are used to tell how the output was created. The comments will also cover some interpretations. Letters in square brackets, such as [a], identify endnotes which will give details of the calculations and explanations. The endnotes begin on page 9. Output from Minitab sometimes will be edited to reduce empty space or to improve page layout. This document was prepared with Minitab 14.

The data set used here can be found at the Web site [www.stern.nyu.edu/~gsimon/statdata](http://www.stern.nyu.edu/~gsimon/statdata); open the "Other Data Sets" folder M. The file name is SWISS.MTP, and it can be found on the Stern Web site as well. The data set concerns fertility rates in 47 Swiss cantons (provinces) in the year 1888. The dependent variable will be Fert, the fertility rate, and all the other variables will function as independent variables. The data are found in Data Analysis and Regression, by Mosteller and Tukey, pages 550-551.

This document was prepared by the Statistics Group of the I.O.M.S. Department. If you find this document to be helpful, we'd like to know! If you have comments that might improve this presentation, please let us know also. Please send e-mail to [gsimon@stern.nyu.edu](mailto:gsimon@stern.nyu.edu).

Revision date 14 NOV 2005

## GUIDE TO MINITAB REGRESSION


~~~~~

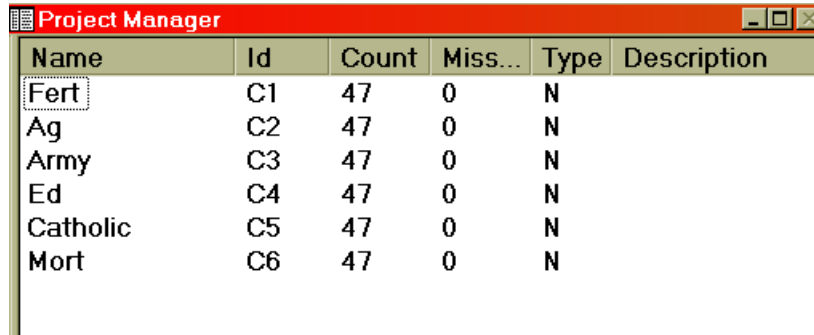
{Data was brought into the program through **File** ⇒ **Open Worksheet** ⇒. Minitab's default for **Files of type:** is (\*.mtw; \*.mpj), so you will want to change this to \*.mtp to obtain the file. On the Stern network, this file is in the folder X:\SOR\B011305\M, and the file name is SWISS.MTP. The listing below shows the data set, as copied directly from Minitab's data window.}

| Fert     | Ag    | Army | Ed   | Catholic | Mort  |
|----------|-------|------|------|----------|-------|
| 0.802[a] | 0.170 | 0.15 | 0.12 | 9.96     | 0.222 |
| 0.831    | 0.451 | 0.06 | 0.09 | 84.84    | 0.222 |
| 0.925    | 0.397 | 0.05 | 0.05 | 93.40    | 0.202 |
| 0.858    | 0.365 | 0.12 | 0.07 | 33.77    | 0.203 |
| 0.769    | 0.435 | 0.17 | 0.15 | 5.16     | 0.206 |
| 0.761    | 0.353 | 0.09 | 0.07 | 90.57    | 0.266 |
| 0.838    | 0.702 | 0.16 | 0.07 | 92.85    | 0.236 |
| 0.924    | 0.678 | 0.14 | 0.08 | 97.16    | 0.249 |
| 0.824    | 0.533 | 0.12 | 0.07 | 97.67    | 0.210 |
| 0.829    | 0.452 | 0.16 | 0.13 | 91.38    | 0.244 |
| 0.871    | 0.645 | 0.14 | 0.06 | 98.61    | 0.245 |
| 0.641    | 0.620 | 0.21 | 0.12 | 8.52     | 0.165 |
| 0.669    | 0.675 | 0.14 | 0.07 | 2.27     | 0.191 |
| 0.689    | 0.607 | 0.19 | 0.12 | 4.43     | 0.227 |
| 0.617    | 0.693 | 0.22 | 0.05 | 2.82     | 0.187 |
| 0.683    | 0.726 | 0.18 | 0.02 | 24.20    | 0.212 |
| 0.717    | 0.340 | 0.17 | 0.08 | 3.30     | 0.200 |
| 0.557    | 0.194 | 0.26 | 0.28 | 12.11    | 0.202 |
| 0.543    | 0.152 | 0.31 | 0.20 | 2.15     | 0.108 |
| 0.651    | 0.730 | 0.19 | 0.09 | 2.84     | 0.200 |
| 0.655    | 0.598 | 0.22 | 0.10 | 5.23     | 0.180 |
| 0.650    | 0.551 | 0.14 | 0.03 | 4.52     | 0.224 |
| 0.566    | 0.509 | 0.22 | 0.12 | 15.14    | 0.167 |
| 0.574    | 0.541 | 0.20 | 0.06 | 4.20     | 0.153 |
| 0.725    | 0.712 | 0.12 | 0.01 | 2.40     | 0.210 |
| 0.742    | 0.581 | 0.14 | 0.08 | 5.23     | 0.238 |
| 0.720    | 0.635 | 0.06 | 0.03 | 2.56     | 0.180 |
| 0.605    | 0.608 | 0.16 | 0.10 | 7.72     | 0.163 |
| 0.583    | 0.268 | 0.25 | 0.19 | 18.46    | 0.209 |
| 0.654    | 0.495 | 0.15 | 0.08 | 6.10     | 0.225 |
| 0.755    | 0.859 | 0.03 | 0.02 | 99.71    | 0.151 |
| 0.693    | 0.849 | 0.07 | 0.06 | 99.68    | 0.198 |
| 0.773    | 0.897 | 0.05 | 0.02 | 100.00   | 0.183 |
| 0.705    | 0.782 | 0.12 | 0.06 | 98.96    | 0.194 |
| 0.794    | 0.649 | 0.07 | 0.03 | 98.22    | 0.202 |
| 0.650    | 0.759 | 0.09 | 0.09 | 99.06    | 0.178 |
| 0.922    | 0.846 | 0.03 | 0.03 | 99.46    | 0.163 |
| 0.793    | 0.631 | 0.13 | 0.13 | 96.83    | 0.181 |
| 0.704    | 0.384 | 0.26 | 0.12 | 5.62     | 0.203 |
| 0.657    | 0.077 | 0.29 | 0.11 | 13.79    | 0.205 |
| 0.727    | 0.167 | 0.22 | 0.13 | 11.22    | 0.189 |
| 0.644    | 0.176 | 0.35 | 0.32 | 16.92    | 0.230 |
| 0.776    | 0.376 | 0.15 | 0.07 | 4.97     | 0.200 |
| 0.676    | 0.187 | 0.25 | 0.07 | 8.65     | 0.195 |
| 0.350    | 0.012 | 0.37 | 0.53 | 42.34    | 0.180 |
| 0.447    | 0.466 | 0.16 | 0.29 | 50.43    | 0.182 |
| 0.428    | 0.277 | 0.22 | 0.29 | 58.33    | 0.193 |

## GUIDE TO MINITAB REGRESSION

~~~~~

{The item below is Minitab's Project Manager window. You can get this to appear by clicking on the  icon on the toolbar.}



Name	Id	Count	Miss...	Type	Description
Fert	C1	47	0	N	
Ag	C2	47	0	N	
Army	C3	47	0	N	
Ed	C4	47	0	N	
Catholic	C5	47	0	N	
Mort	C6	47	0	N	

[b]

{The following section gives basic statistical facts. It is obtained by **Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics ⇒**. All variables were requested. The request can be done by listing each variable by name (Fert Ag Army Ed Catholic Mort) or by listing the column numbers (C1-C6) or by clicking on the names in the variable listing.}

### Descriptive Statistics: Fert, Ag, Army, Ed, Catholic, Mort

Variable	[c][d]		[e]					[f]	
	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Fert	47	0	0.7014	0.0182	0.1249	0.3500	0.6440	0.7040	0.7930
Ag	47	0	0.5066	0.0331	0.2271	0.0120	0.3530	0.5410	0.6780
Army	47	0	0.1649	0.0116	0.0798	0.0300	0.1200	0.1600	0.2200
Ed	47	0	0.1098	0.0140	0.0962	0.0100	0.0600	0.0800	0.1200
Catholic	47	0	41.14	6.08	41.70	2.15	5.16	15.14	93.40
Mort	47	0	0.19943	0.00425	0.02913	0.10800	0.18100	0.20000	0.22200

Variable	Maximum
Fert	0.9250
Ag	0.8970
Army	0.3700
Ed	0.5300
Catholic	100.00
Mort	0.26600

{The next listing shows the correlations. It is obtained through **Stat ⇒ Basic Statistics ⇒ Correlation ⇒** and then listing all the variable names. For now, we have de-selected the feature **Display p-values.**}

## GUIDE TO MINITAB REGRESSION

~~~~~

### Correlations: Fert, Ag, Army, Ed, Catholic, Mort

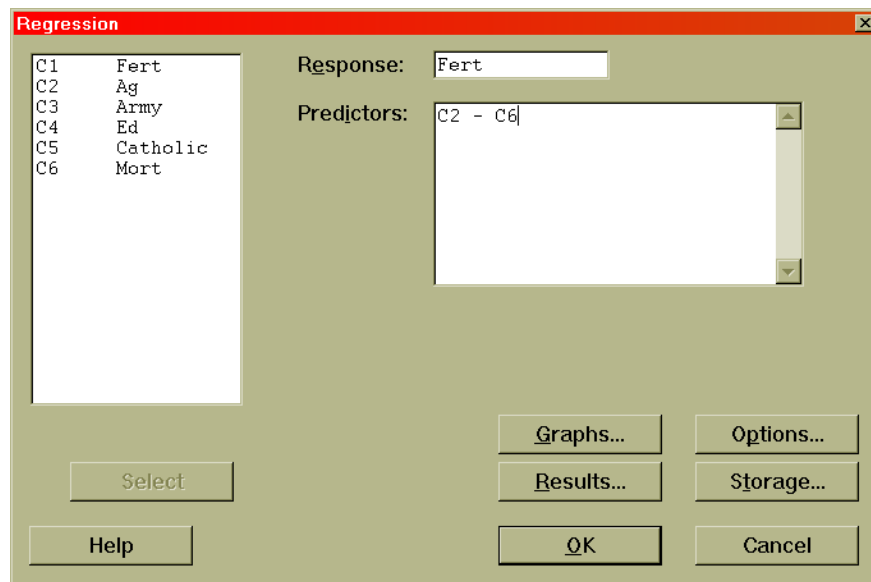
|          | Fert   | Ag         | Army   | Ed     | Catholic | Mort |
|----------|--------|------------|--------|--------|----------|------|
| Ag       | 0.353  |            |        |        |          |      |
| Army     | -0.646 | -0.687 [g] |        |        |          |      |
| Ed       | -0.664 | -0.640     | 0.698  |        |          |      |
| Catholic | 0.464  | 0.401      | -0.573 | -0.154 |          |      |
| Mort     | 0.417  | -0.061     | -0.114 | -0.099 | 0.175    |      |

Cell Contents: Pearson correlation

{The linear regression of dependent variable Fert on the independent variables can be started through

**Stat** ⇒ **Regression** ⇒ **Regression** ⇒

Set up the panel to look like this:



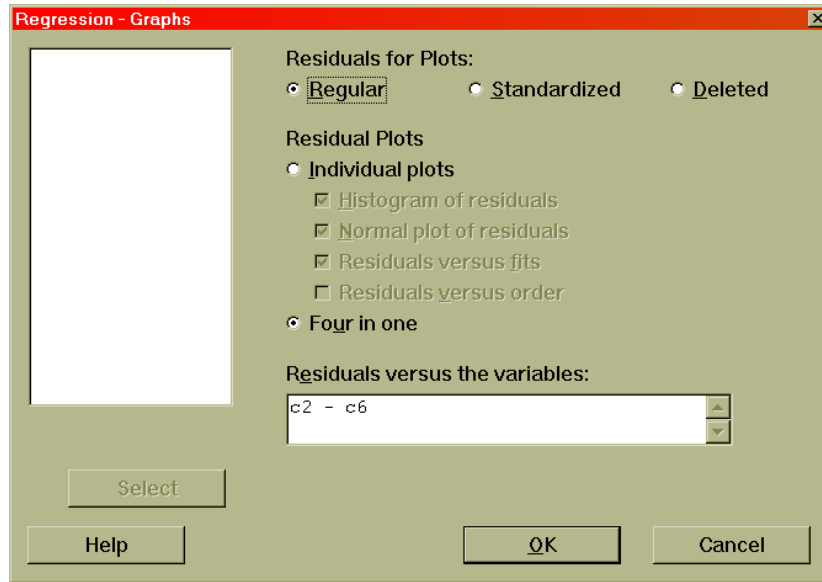
Observe that Fert was selected as the dependent variable (response) and all the others were used as independent variables (predictors). If you click **OK** you will see the basic regression results. For the sake of illustration, we'll show some additional features.

Click the **Options...** button and then select **Variance inflation factors**. The choice **Fit intercept** is the default and should already be selected; if it is not, please select it. The **Fit intercept** option should be de-selected only in extremely special situations.

We recommend that you routinely examine the variance inflation factors if strong collinearity is suspected. The Durbin-Watson statistic was not used here because the data are not time-sequenced.



Click the **Graphs...** button and select the indicated choices:



Examining the **Residuals versus fits** plot is now part of routine statistical practice. The other selections can show some interesting clues as well. Here we will use the **Four in one** option, as it shows the residual versus fitted plot, along with the other three as well. The **Residuals versus order** plot will not be useful, because the data are not time-ordered.

Some of the choices made here reflect features of this data set or particular desires of the analyst. Here the Regular form of the residuals was desired; other choices would be just as reasonable.

Click the **Storage...** button and select **Hi (leverages)**.

This provides a very thorough regression job. }

{ The model corresponding to this request is

$$\text{Fert}_i = \beta_0 + \beta_{AG} \text{Ag}_i + \beta_{Army} \text{Army}_i + \beta_{ED} \text{ED}_i + \beta_{CATH} \text{CATH}_i + \beta_{MORT} \text{MORT}_i + \varepsilon_i \quad \}$$

## GUIDE TO MINITAB REGRESSION

~~~~~

### Regression Analysis: Fert versus Ag, Army, Ed, Catholic, Mort

The regression equation is [h]  
 Fert = 0.669 - 0.172 Ag - 0.258 Army - 0.871 Ed  
 + 0.00104 Catholic + 1.08 Mort

Predictor	Coef	SE Coef	T	P	VIF
Constant [i]	0.6692 [j]	0.1071 [k]	6.25 [l]	0.000 [m]	[n]
Ag	-0.17211 [ø]	0.07030	-2.45	0.019	2.3 [p]
Army	-0.2580	0.2539	-1.02 [q]	0.315 [r]	3.7
Ed	-0.8709	0.1830	-4.76	0.000	2.8
Catholic	0.0010412	0.0003526	2.95	0.005	1.9
Mort	1.0770	0.3817	2.82	0.007	1.1

S = 0.0716537 [s]    R-Sq = 70.7% [t]    R-Sq(adj) = 67.1% [u]

#### Analysis of Variance [v]

Source	DF [w]	SS [aa]	MS [ee]	F [ii]	P [jj]
Regression	5 [x]	0.50729 [bb]	0.10146 [ff]	19.76	0.000
Residual Error	41 [y]	0.21050 [cc]	0.00513 [gg]		
Total	46 [z]	0.71780 [dd]	[hh]		

Source	DF	Seq SS [kk]
Ag	1	0.08948
Army	1	0.22104
Ed	1	0.08918
Catholic	1	0.06671
Mort	1	0.04088

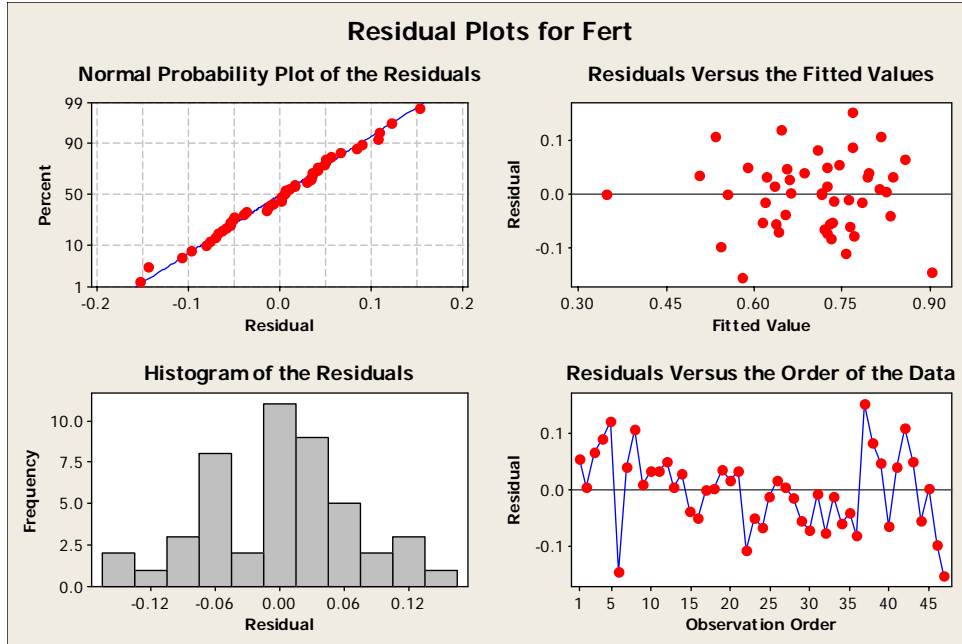
#### Unusual Observations [ll]

Obs	Ag	Fert	Fit	SE Fit	Residual	St Resid
6 [mm]	0.353	0.7610	0.9050 [nn]	0.0319 [øø]	-0.1440 [pp]	-2.24R [qq]
37	0.846	0.9220	0.7688	0.0270	0.1532	2.31R
45	0.012	0.3500	0.3480	0.0484	0.0020	0.04 X [rr]
47	0.277	0.4280	0.5807	0.0244	-0.1527	-2.27R

R denotes an observation with a large standardized residual  
 X denotes an observation whose X value gives it large influence.

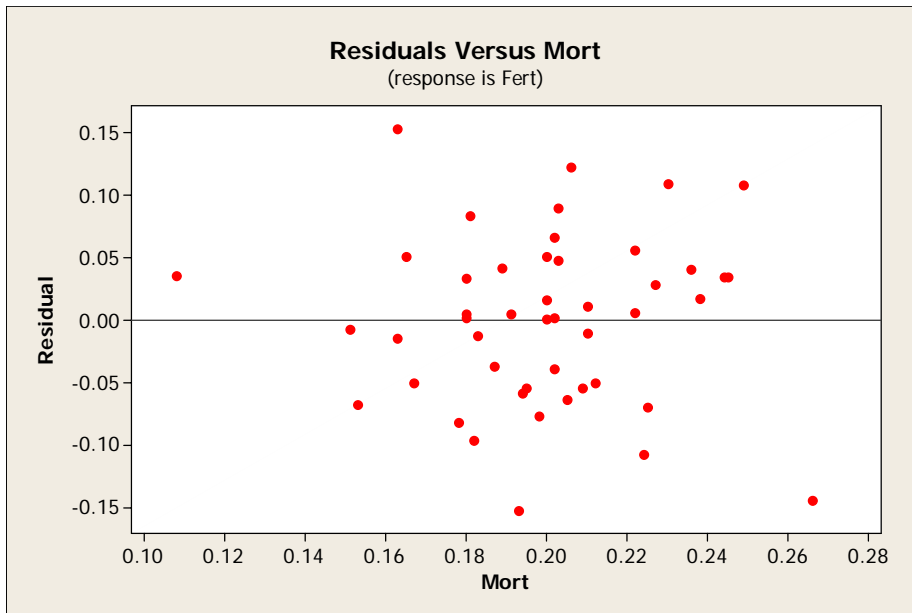
{Many graphs were requested in this run. The **Four in one** panel examines the behavior of the residuals because they provide clues as to the appropriateness of the assumptions made on the  $\epsilon_i$  terms in the model. The most important of these is the residuals versus fitted plot, the plot at the upper right on the next page. The normal probability plot and the histogram of the residuals are used to assess whether or not the noise terms are approximately normally distributed. Since the data points are not time-ordered, we will not use the plot of the residuals versus the order of the data.}

# GUIDE TO MINITAB REGRESSION



[ ss ]

{ Many users choose also to examine the plots of the residuals against each of the predictor variables. These were requested for this run, but this document will show only the plot of the residuals against the variable Mort. }



[ tt ]

## GUIDE TO MINITAB REGRESSION

~~~~~

{Finally, recall that we had requested the high leverage points through **Stat** ⇒ **Regression** ⇒ **Regression** ⇒ **Storage** ⇒ and then selecting **Hi (leverages)**. These will show up in a new column, called **HI1**, in the data window. This column can be used in plots, or it can simply be examined. What shows below is that column, copied out of the data window, and restacked to save space.}

| Case | HI1      |      | Case | HI1      | Case | HI1           |
|------|----------|------|------|----------|------|---------------|
| 1    | 0.156817 | [uu] | 17   | 0.068535 | 33   | 0.108342      |
| 2    | 0.122585 |      | 18   | 0.101750 | 34   | 0.098006      |
| 3    | 0.173683 |      | 19   | 0.351208 | 35   | 0.076759      |
| 4    | 0.079616 |      | 20   | 0.111375 | 36   | 0.091772      |
| 5    | 0.072190 |      | 21   | 0.074258 | 37   | 0.142462      |
| 6    | 0.198332 |      | 22   | 0.082771 | 38   | 0.081257      |
| 7    | 0.143082 |      | 23   | 0.064105 | 39   | 0.076831      |
| 8    | 0.141458 |      | 24   | 0.109214 | 40   | 0.226297      |
| 9    | 0.079940 |      | 25   | 0.100362 | 41   | 0.099816      |
| 10   | 0.106823 |      | 26   | 0.125696 | 42   | 0.205322      |
| 11   | 0.136769 |      | 27   | 0.180591 | 43   | 0.073667      |
| 12   | 0.083193 |      | 28   | 0.079051 | 44   | 0.172191      |
| 13   | 0.083926 |      | 29   | 0.053282 | 45   | 0.455836 [vv] |
| 14   | 0.109909 |      | 30   | 0.077062 | 46   | 0.210670      |
| 15   | 0.125512 |      | 31   | 0.173359 | 47   | 0.115954      |
| 16   | 0.106312 |      | 32   | 0.092047 |      |               |

{There is a commonly-used threshold of concern, as discussed in [uu]. Minitab will automatically mark points that exceed this threshold; see [ll] and [rr]. It is therefore *not* critical that the leverage, or **Hi**, values be computed.}

## GUIDE TO MINITAB REGRESSION

~~~~~

### ENDNOTES:

[a] This is the first line of the data listing. The line numbers (1 through 47) are not shown here, although they do appear in the Minitab data window. The numbers across this row indicate that this first canton had  $Fert_1 = 0.802$ ,  $Ag_1 = 0.17$ ,  $Army_1 = 0.15$ , and so on.

[b] Minitab's Project Manager window shows the variable names for the columns, and also some basic accounting. We see that each variable has 47 values, with none missing. Minitab data sets can also have Constants and Matrices, although this set has none. Descriptions are saved only with project (\*.MPJ) files.

[c] The symbol N refers to the sample size, *after* removing missing data. In this data set, there are 47 cantons and all information is complete. We have 47 pieces of information for each variable. In some data sets, the column of N-values might list several different numbers.

[d]  $N^*$  is the number of missing values. In this set of data, all variables are complete.

[e] This is the standard error of the mean. It's computed for each variable as  $\frac{SD}{\sqrt{N}}$ , where  $N$  is the number of non-missing values. Here you can confirm that for variable Fert,  $\frac{0.1249}{\sqrt{47}} \approx 0.0182$ .

[f] Minitab computes the quartiles Q1 and Q3 by an interpolation method. If the sample size is  $n$ , then Q1 is the observation at rank position  $(n+1)/4$ . If  $(n+1)/4$  is not an integer, then Q1 is obtained as a weighted average of the values at the surrounding integer positions. For instance, if  $(n+1)/4 = 6.75$ , then Q1 is  $\frac{3}{4}$  of the distance between the 6<sup>th</sup> and 7<sup>th</sup> values. The procedure for finding Q3 works from rank position  $3(n+1)/4$ .

[g] The value -0.687 is the correlation between Army and Ag. It is also the correlation between Ag and Army. Since correlations are symmetric, it is not necessary to print the entire correlation matrix. The correlation between a variable and itself is 1.000 (also not printed). Generally we like to see strong correlations (say above +0.9 or below -0.9) involving the dependent variable, here Fert. We prefer not to have strong correlations among the other variables.

[h] This is the estimated model or fitted equation. Some people like to place the "hat" on the dependent variable Fert as  $\hat{Fert}$  to denote estimation or fitting. Note that the numbers are repeated in the Coef column below. The letter  $b$  is used for estimated values; thus  $b_0 = 0.669$ ,  $b_{Ag} = -0.172$ , and so on.

[i] The term Constant refers to the inclusion of  $\beta_0$  in the regression model. This is sometimes called the *intercept*.

## GUIDE TO MINITAB REGRESSION

~~~~~

[j] This is the estimated value of  $\beta_0$  and is often called  $b_0$ .

[k] This is estimated standard deviation of the value in the Coef column; this is also called the *standard error*. In this instance, we believe that the estimated value, 0.6692, is good to within a standard error of 0.1071. We're about 95% confident that the true value of  $\beta_0$  is in the interval  $0.6692 \pm 2(0.1071)$ , which is the interval (0.4550 , 0.8834).

[l] The value of T, also called *Student's t*, is  $\frac{\text{Coef}}{\text{SE Coef}}$ ; here that arithmetic is

$\frac{0.6692}{0.1071} \approx 6.25$ . This is the number of estimated standard deviations that the estimate,

0.6692, is away from zero. The phrase "estimated standard deviations" refers to the distribution of the sample coefficient, and *not* to the standard deviation in the regression model. This T can be regarded as a test of  $H_0: \beta_0 = 0$  versus  $H_1: \beta_0 \neq 0$ . Here T is outside the interval (-2,+2) and we should certainly believe that  $\beta_0$  (the true-but-unknown population value) is different from zero. Some users believe that the intercept should not be subjected to a statistical test; indeed some software does not provide a T value or a P (see the next item) for the Constant line.

[m] The column P (for *p*-value) is the result of subjecting the data to a statistical test as to whether  $\beta_0 = 0$  or  $\beta_0 \neq 0$ . There is a precise technical definition, but crudely P is the smallest Type I error probability that you might make in deciding that  $\beta_0 \neq 0$ . Very small values of P suggest that  $\beta_0 \neq 0$ , while larger values indicate that you should maintain  $\beta_0 = 0$ . The typical cutoff between these actions is 0.05; thus  $P \leq 0.05$  causes you to decide that the population parameter, here  $\beta_0$ , is really different from zero. The *p*-value is never exactly zero, but it sometimes prints as 0.000. The *p*-value is directly related to T; values of T far outside the interval (-2,+2) lead to small P. When  $P \leq 0.05$ , we say that the estimated value is *statistically significant*.

As an important side note, many people believe that the Constant  $\beta_0$  should never be subjected to statistical tests. According to this point of view, we should not even ask whether  $\beta_0 = 0$  or  $\beta_0 \neq 0$ ; indeed we should not even list T and P in the Constant line. This side note applies *only* to the Constant.

[n] The VIF, variance inflation factor, comes as the result of a special request. The VIF does not apply to the Constant. See also item [p].

## GUIDE TO MINITAB REGRESSION

~~~~~

[ø] The value -0.1721 is  $b_{AG}$ , the estimated value for  $\beta_{AG}$ . In fact, the Coef column can be used to write the fitted equation [h]. Some people like to write the standard errors in parentheses under the estimated coefficients in the fitted equation in this fashion:

$$\begin{aligned} \hat{F}e_{rt} = & 0.669 & - & 0.172 & Ag & - & 0.258 & Army & - & 0.871 & Ed \\ & (0.107) & & (0.070) & & & (0.254) & & & (0.183) \\ + & 0.00104 & Catholic & + & 1.08 & Mort \\ & (0.00035) & & & (0.38) & & & & & & \end{aligned}$$

You may also see this kind of display using in parentheses the values of  $T$ , so indicate for your readers exactly what you are doing.

This is precisely the relationship that is used to determine the  $n = 47$  fitted values:

$$\begin{aligned} \hat{F}e_{rt_i} = & 0.669 & - & 0.172 & Ag_i & - & 0.258 & Army_i & - & 0.871 & Ed_i \\ + & 0.00104 & Catholic_i & + & 1.08 & Mort_i \end{aligned}$$

The differences between the observed and fitted values are the *residuals*. The  $i^{\text{th}}$  residual, usually denoted  $e_i$ , is  $Fert_i - \hat{F}e_{rt_i}$ .

[p] The VIF, variance inflation factor, measures how much of the standard error, SE Coef, can be accounted for by inter-relation of one independent variable with all of the other independent variables. The VIF can never be less than 1. If some VIF values are large (say 10 or more), then you have a *collinearity* problem. Plausible solutions to the collinearity are Stepwise Regression and Best Subsets Regression. A related concept is the *tolerance*; these are related through  $Tolerance = \frac{1}{VIF}$ .

[q] The value of  $T$  for the variable Army tests the hypothesis  $H_0: \beta_{Army} = 0$  versus  $H_1: \beta_{Army} \neq 0$ . This  $T$  is in the interval  $(-2,+2)$ , suggesting that  $H_0$  is correct, so you might consider repeating the problem without using Army in the model.

[r] The  $P$  for variable Army exceeds 0.05. This is consistent with the previous comment, and it suggests repeating the problem without using Army in the model. The comparison of  $P$  with 0.05 is a more precise standard than comparing  $|T|$  with 2.0.

[s] This is one of the most important numbers in the regression output. It's called *standard error of estimate* or *standard error of regression*. It is the estimate of  $\sigma$ , the standard deviation of the noise terms (the  $\epsilon_i$ 's). A common notation is  $s_e$ . In this data set, that value comes out to 0.07165. It is useful to compare this to 0.1249, which was the standard deviation of Fert on the **Descriptive Statistics** list. The original "noise" level in Fert was 0.1249 (without doing any regression); the "noise" left over after the regression was 0.07165. Item [s] is the square root of item [gg], the residual mean square, whose value is 0.00513; observe that  $\sqrt{0.00513} \approx 0.07162$ .

## GUIDE TO MINITAB REGRESSION

~~~~~

[t] This is the heavily-cited  $R^2$  ; it's generally given as a percent, here 70.7%, but it might also be given as decimal 0.707. The formal statement used is "The percent of the variation in Fert that is explained by the regression is 70.7%." Technically,  $R^2$  is the ratio of two sums of squares; it is the ratio of item [bb], the regression sum of squares, to item [dd], the total sum of squares. Observe that  $\frac{0.50729}{0.71780} \approx 0.7067 = 70.67\%$ . Large values of  $R^2$  are considered to be good.

[u] This is an adjustment made to  $R^2$  to account for the sample size and the number of independent variables and is given by  $R_{adj}^2 = 1 - \frac{n-1}{n-1-k}(1-R^2)$ . In this formula,  $n$  is the number of points (here 47) and  $k$  is the number of predictors used (here 5). Thus  $R^2$  has the neat interpretation in [t], and we have  $R_{adj}^2 = 1 - \left(\frac{s_\varepsilon}{s_{Fert}}\right)^2$ . Here  $s_\varepsilon$  is 0.07165 and  $s_{Fert} = 0.1249$  and given in [s]. Most users prefer  $R^2$  over  $R_{adj}^2$ .

[v] This is the analysis of variance table. The work is based on the algebraic identity

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

in which  $y_i$  denotes the value of the dependent variable for point  $i$ , and  $\hat{y}_i$  denotes the fitted value for point  $i$ . Since Fert is the dependent variable, we identify  $\hat{y}_i$  with  $\hat{Fert}_i$  as in [h] and [ø]. The three sums of squares in this equation are, respectively,  $SS_{total}$ ,  $SS_{regression}$ , and  $SS_{residual\ error}$ . These have other names or abbreviations. For instance

$SS_{total}$  is often written as  $SS_{tot}$ .

$SS_{regression}$  is often written as  $SS_{reg}$  and sometimes as  $SS_{fit}$  or  $SS_{model}$ .

$SS_{residual\ error}$  is often written as  $SS_{residual}$  or  $SS_{resid}$  or  $SS_{res}$  or  $SS_{error}$  or  $SS_{err}$ .

[w] The DF stands for degrees of freedom. This is an accounting of the dimensions of the problem, and the numbers in this column add up to the indicated total as  $5 + 41 = 46$ . See the next three notes.

[x] The Regression line in the analysis of variance table refers to the sum  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

which appeared in [v]. The degrees of freedom for this calculation is  $k$ , the number of independent variables. Here  $k = 5$ .

## GUIDE TO MINITAB REGRESSION

~~~~~

[y] The Residual Error line in the analysis of variance table refers to the sum  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$  which appeared in [v]. The degrees of freedom for this calculation is

$n - 1 - k$ , where  $n$  is the number of data points and  $k$  is the number of independent variables. Here  $n = 47$  and  $k = 5$ , so that  $47 - 1 - 5 = 41$  appears in this position.

[z] The Total line in the analysis of variance table refers to the sum  $\sum_{i=1}^n (y_i - \bar{y})^2$  which appears in [v]. The degrees of freedom for this calculation is  $n - 1$ , where  $n$  is the number of data points. Here  $n = 47$ , so that 46 appears in this position.

[aa] The SS stands for sum of squares. This column gives the numbers corresponding to the identity described in [v]. The values in this column add up to the indicated total as  $0.50729 + 0.21050 = 0.71780$ .

[bb] This is the sum  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  which appeared in [v]. This is  $SS_{\text{regression}}$ .

[cc] This is the sum  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$  which appeared in [v]. This is  $SS_{\text{residual error}}$ .

[dd] This is the sum  $\sum_{i=1}^n (y_i - \bar{y})^2$  which appears in [v]. This is  $SS_{\text{total}}$ , the total sum of squares. It involves only the dependent variable, so it says nothing about the regression. The regression is successful if the regression sum of squares is large relative to the residual error sum of squares. The  $F$  statistic, item [ii], is the appropriate measure of success.

[ee] The MS stands for Mean Squares. Each value is obtained by dividing the corresponding Sum Squares by its Degrees of Freedom.

[ff] This is  $MS_{\text{regression}} = 0.50729 \div 5 \approx 0.10146$ . In a successful regression, this is large relative to item [gg], the residual mean square.

[gg] This is  $MS_{\text{residual error}} = 0.21050 \div 41 \approx 0.00513$ .

[hh] There is no mean square entry in this position. The computation  $SS_{\text{total}} \div (n - 1)$  would nonetheless be useful, since it is the sample variance of the dependent variable.

[ii] This is the  $F$  statistic. It is computed as  $MS_{\text{regression}} \div MS_{\text{residual error}}$ , meaning  $0.10146 \div 0.00513 \approx 19.76$ . To test at significance level  $\alpha$ , this is to be compared to  $F_{k, n-1-k}^{\alpha}$ , the upper  $\alpha$  point from the  $F$  distribution with  $k$  and  $n - 1$  degrees of freedom, obtained from statistical tables or from Minitab. The  $F$  statistic is a formal test of the null hypothesis that all the independent variable coefficients are zero against the alternative that they are not. In our example, we would write

## GUIDE TO MINITAB REGRESSION

~~~~~

$$H_0: \beta_{Ag} = 0, \beta_{Army} = 0, \beta_{Ed} = 0, \beta_{Catholic} = 0, \beta_{Mort} = 0$$

$H_1$ : at least one of  $\beta_{Ag}$ ,  $\beta_{Army}$ ,  $\beta_{Ed}$ ,  $\beta_{Catholic}$ ,  $\beta_{Mort}$  is not zero

If the  $F$  statistic is larger than  $F_{k,n-1-k}^\alpha$ , then the null hypothesis is rejected. Otherwise, we accept  $H_0$  (or reserve judgment). In this instance, using significance level  $\alpha = 0.05$ , we find  $F_{k,n-1-k}^\alpha = F_{5,41}^{0.05} = 2.4434$  from a statistical table. Since  $19.76 > 2.4434$ , we would reject  $H_0$  at the 0.05 level of significance; we would describe this regression as *statistically significant*.

Here's how to use Minitab to get the values of  $F_{k,n-1-k}^\alpha$ .

**Calc**  $\Rightarrow$  **Probability Distributions**  $\Rightarrow$  **F**  $\Rightarrow$

Select **Inverse cumulative probability**, choose

**Numerator degrees of freedom:** 5

**Denominator degrees of freedom:** 41

**Input constant:** 0.95

Then click **OK**.

[jj] This is the  $p$ -value associated with the  $F$  statistic. This is the result of subjecting the data to a statistical test of  $H_0$  versus  $H_1$  in item [ii]. The  $p$ -value noted in item [m] is not cleanly related to this.

[kk] The Seq SS column is constructed by fitting the predictor variables *in the order given* and noting the change in  $SS_{\text{regression}}$ . Logically, this means here five regressions:

|                                      |                                                          |
|--------------------------------------|----------------------------------------------------------|
| Fert on Ag                           | has $SS_{\text{reg}} = 0.08948$                          |
| Fert on Ag, Army                     | has $SS_{\text{reg}} = 0.22104 + 0.08948 = 0.31052$      |
| Fert on Ag, Army, Ed<br>0.39970      | has $SS_{\text{reg}} = 0.08918 + 0.31052 =$<br>$0.39970$ |
| Fert on Ag, Army, Ed, Catholic       | has $SS_{\text{reg}} = 0.06671 + 0.39970 = 0.46641$      |
| Fert on Ag, Army, Ed, Catholic, Mort | has $SS_{\text{reg}} = 0.04088 + 0.46641 = 0.50729$      |

The final value 0.50729 is  $SS_{\text{reg}}$  for the whole regression, item [bb]. Naming the predictor variables in another order would produce different results. This arithmetic can only be interesting if the order of the variables is interesting. In this case, there is no reason to have any interest in these values.

## GUIDE TO MINITAB REGRESSION

~~~~~

[*ll*] Minitab will list for you data points which are “unusual observations” and are worthy (perhaps) of special attention. There are two concerns, unusual residuals and high influence.

Unfortunately, Minitab has too low a threshold of concern regarding the residuals, as it will list any standardized residual below -2 or above +2. Virtually every data set has points with this property, so that nothing “unusual” is involved. A more reasonable concern would be for residuals below -2.5 or above +2.5. Indeed, in large data sets, one might move the thresholds of concern to -3 and +3. The (-2, 2) thresholds cannot be reset, so you will have to live with the output.

Extreme residuals occur with points for which the regression model does not fit well. It is always worth examining these points. It is generally not appropriate to remove these points from the regression.

The geometric configuration of the predictor variables might indicate that some points unduly influence the regression results. These points are said to have high influence or high leverage. Whether such points should be removed from the regression is a difficult question. This section of the Minitab listing does not really provide enough guidance as to the degree of influence. You can get additional information on high influence points through **Storage** and then asking for **Hi (leverages)** when the regression is initiated. See also point [*uu*].

[*mm*] The unusual observations are identified by their case numbers (here 6, 37, 45, and 47), by their values on the first-named predictor variable, and by their values on the dependent variable.

[*nn*] The fitted values refer to the calculation suggested in item [*ø*], and here 0.9050 is the value of  $F\hat{e}t_6$ , the fitted value for point 6.

[*øø*] The SE Fit refers to the estimated standard deviation of the fitted value in the previous column. This is only marginally interesting.

[*pp*] This gives the actual residual; Here  $-0.1440 = F_{e6} - F\hat{e}t_6 = 0.7610 - 0.9050$ .

[*qq*] Since the actual residuals bear the units of the dependent variable, they are hard to appraise. Thus, we use the standardized residuals. These can be thought of as approximate z-scores, so that about 5% of them should be outside the interval (-2, 2).

[*rr*] This marks a large influence point. See item [*ll*].

## GUIDE TO MINITAB REGRESSION

~~~~~

[ss] The purpose of the residual versus fitted plot is to check for possible violations of regression assumptions, particularly non-homogeneous residuals. This pathology reveals itself through a pattern in which the spread of the residuals changes in moving from left to right on the plot. The residual versus fitted plot will sometimes reveal curvature as well. Large positive and large negative residuals will be seen on this plot. The detection and relieving of these pathologies are subtle processes which go beyond the content of this document. The appearance of this plot is not materially influenced by the particular choice made for standardizing the residuals; this was done here from the **Stat** ⇒ **Regression** ⇒ **Regression** ⇒ **Graphs** panel by choosing **Residuals for plots:** as **Regular**.

[tt] The purpose of plotting the residuals against the predictor variables (in turn) is to check for non-linearity. This particular plot shows no problems.

[uu] The leverage value for the first canton in the data set is 0.156817. This is computed as the (1, 1) entry of the matrix  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$  where  $\mathbf{X}$  is the  $n$ -by- $(k + 1)$  matrix whose  $n$  rows represent the  $n$  data points. The  $k + 1$  columns consist of the constant column (containing a 1 in each position) and one column for each of the  $k$  predictor variables. The dependent variable Fert is not involved in this arithmetic. The leverage value for the  $j^{\text{th}}$  canton is the (j, j) entry of this matrix. A commonly accepted cutoff marking off “high” leverage points is  $\frac{3(k+1)}{n}$ , which is here  $\frac{3(5+1)}{47} \approx 0.3830$ . Only the leverage value for canton 45 is larger than this; see [rr] and [vv].

[vv] The leverage value for the 45<sup>th</sup> canton is 0.455836. This is clearly a high leverage point. There is some cause for concern, because high leverage points can distort the estimation process. A quick look at the data set will show that this canton, the third from the bottom on the data list, has very unusual values for Ag, Army, and Ed. A thorough analysis of this data set would probably include another run in which canton 45 is deleted.