Statistical analysis using Microsoft Excel

Microsoft Excel spreadsheets have become somewhat of a standard for data storage, at least for smaller data sets. This, along with the program often being packaged with new computers, naturally encourages its use for statistical analyses. This is unfortunate, since Excel is most decidedly **not** a statistical package.

Here's an example of how the numerical inaccuracies in Excel can get you into trouble. Consider the following data set:

```
Data Display

 Row            X                Y

   1  10000000001   1000000000.000
   2  10000000002   1000000000.000
   3  10000000003   1000000000.900
   4  10000000004   1000000001.100
   5  10000000005   1000000001.010
   6  10000000006   1000000000.990
   7  10000000007   1000000001.100
   8  10000000008   1000000000.999
   9  10000000009   1000000000.000
  10  10000000010   1000000000.001
```

Here is Minitab output for the regression:

```
Regression Analysis

The regression equation is
Y =9.71E+08 + 0.0029 X

Predictor         Coef        StDev           T         P
Constant     970667056    616256122        1.58     0.154
X              0.00293      0.06163        0.05     0.963

S = 0.5597      R-Sq = 0.0%      R-Sq(adj) = 0.0%

Analysis of Variance

Source             DF           SS           MS          F         P
Regression          1       0.0007       0.0007       0.00     0.963
```

```
Residual Error     8      2.5065      0.3133
Total              9      2.5072
```

Now, here are the values obtained when using the regression program available in the Analysis Toolpak of `Microsoft Excel 2002` (the same results came from earlier versions of `Excel`; I will say something about `Excel 2003` later):

```
SUMMARY OUTPUT

Regression Statistics
Multiple R 65535
R Square   -0.538274369
Adjusted R Square -0.730558665
Standard Error 0.694331016
Observations 10

ANOVA
              df         SS          MS              F       Signif F
Regression     1   -1.349562541  -1.349562541  -2.799367289    #NUM!
Residual       8    3.85676448    0.48209556
Total          9    2.507201939

                Coeff      Standard Error      t Stat    P-value
Intercept    2250000001           0             65535     #NUM!
X Variable 1   -0.125             0             65535     #NUM!
```

Each of the nine numbers given above is incorrect! The slope estimate has the wrong sign, the estimated standard errors of the coefficients are zero (making it impossible to construct t–statistics), and the values of R^2, F and the regression sum of squares are **negative**! It's obvious here that the output is garbage (even `Excel` seems to know this, as the `#NUM!`'s seem to imply), but what if the numbers that had come out weren't absurd — just wrong? Unless `Excel` does better at addressing these computational problems, it cannot be considered a serious candidate for use in statistical analysis.

What went wrong here? The summary statistics from `Excel` give us a clue:

```
              X                                 Y
Mean                  10000000006   Mean                 1000000001
Standard Error                  0   Standard Error       6.746192342
Median                10000000006   Median               1000000001
```

```
Mode                       #N/A   Mode                1000000000
Standard Deviation            0   Standard Deviation  21.33333333
Sample Variance               0   Sample Variance     455.1111111
```

Here are the corresponding values if the all `X` values are decreased by 10000000000, and all `Y` values are decreased by 1000000000. The standard deviations and sample variances should, of course, be identical in the two cases, but they are not (the values below are correct):

```
               X                                Y
Mean                        5.5   Mean                      0.61
Standard Error      0.957427108   Standard Error     0.166906561
Median                      5.5   Median                   0.945
Mode                       #N/A   Mode                         0
Standard Deviation  3.027650354   Standard Deviation 0.527804888
Sample Variance     9.166666667   Sample Variance       0.278578
```

Thus, simple descriptive statistics are not trustworthy either in situations where the standard deviation is small relative to the absolute level of the data.

Using `Excel` to analyze multiple regression data brings its own problems. Consider the following data set, provided by Gary Simon:

Y	X1	X2	X3	X4
5.88	1	1	1	1
2.56	6	1	1	1
11.11	1	1	1	1
0.79	6	1	1	1
0.00	6	1	1	1
0.00	0	1	1	1
15.60	8	1	1	1
3.70	4	1	1	1
8.49	3	1	1	1
51.20	6	1	1	1
14.20	7	1	1	1
7.14	5	1	1	1
4.20	7	1	1	1
6.15	4	1	1	1
10.46	6	1	1	1
0.00	8	1	1	1

10.42	2	1	1	1
17.36	5	1	1	1
13.41	8	1	1	1
41.67	0	1	1	1
2.78	0	1	1	1
2.98	8	1	1	1
9.62	7	1	1	1
0.00	0	1	1	1
4.65	5	1	0	2
3.13	3	1	0	2
24.58	6	1	0	2
0.00	1	1	0	2
5.56	4	1	0	2
9.26	3	1	0	2
0.00	0	1	0	2
0.00	0	1	0	2
3.13	1	1	0	2
0.00	0	1	0	2
7.56	5	0	1	3
9.93	6	0	1	3
0.00	8	0	1	3
16.67	6	0	1	3
16.89	7	0	1	3
13.71	6	0	1	3
6.35	5	0	1	3
2.50	3	0	1	3
2.47	7	0	1	3
21.74	3	0	1	3
23.60	8	0	0	4
11.11	8	0	0	4
0.00	7	0	0	4
3.57	8	0	0	4
2.90	5	0	0	4
2.94	3	0	0	4
2.42	8	0	0	4
18.75	4	0	0	4
0.00	5	0	0	4
2.27	3	0	0	4

There is nothing apparently unusual about these data, and they are, in fact, from an actual clinical experiment. Here is output from `Excel 2002` (and earlier versions) for a regression of `Y` on `X1`, `X2`, `X3`, and `X4`:

```
SUMMARY OUTPUT

Regression Statistics
Multiple R          0.218341811
R Square            0.047673146
Adjusted R Square  -0.030067821
Standard Error     10.23964549
Observations       54

ANOVA
             df        SS           MS            F   Significance F
Regression    4  257.1897798   64.29744495 0.613230678 0.655111835
Residual     49  5137.666652  104.8503398
Total        53  5394.856431

              Coefficients   Standard Error     t Stat     P-value
Intercept      0.384972384   0                   65535      #NUM!
X Variable 1   0.386246607   0.570905635     0.6765507    0.501872378
X Variable 2   2.135547339   0                   65535      #NUM!
X Variable 3   4.659552583   0                   65535      #NUM!
X Variable 4   0.952380952   0                   65535      #NUM!
```

Obviously there's something strange going on here: the intercept and three of the four coefficients have standard error equal to zero, with undefined p–values (why Excel gives what would seem to be t–statistics equal to infinity as 65535 is a different matter!). One coefficient has more sensible–looking output. In any event, Excel does give a fitted regression with associated F–statistic and standard error of the estimate.

Unfortunately, this is all incorrect. There is **no** meaningful regression possible here, because the predictors are perfectly collinear (this was done inadvertently by the clinical researcher). That is, no regression model can be fit using all four predictors. Here is what happens if you try to use Minitab to fit the model:

```
Regression Analysis

* X4 is highly correlated with other X variables
* X4 has been removed from the equation

The regression equation is
Y = 4.19 + 0.386 X1 + 0.23 X2 + 3.71 X3

Predictor        Coef       StDev            T        P
```

```
Constant         4.194        3.975         1.06    0.296
X1              0.3862       0.5652         0.68    0.497
X2               0.231        3.159         0.07    0.942
X3               3.707        2.992         1.24    0.221

S = 10.14        R-Sq = 4.8%      R-Sq(adj) = 0.0%

Analysis of Variance

Source             DF           SS           MS          F         P
Regression          3        257.2         85.7       0.83     0.481
Residual Error     50       5137.7        102.8
Total              53       5394.9
```

Minitab correctly notes the perfect collinearity among the four predictors and drops one, allowing the regression to proceed. Which variable is dropped out depends on the order of the predictors given to Minitab, but all of the fitted models yield the same R^2, F, and standard error of the estimate (of these statistics, Excel only gets the R^2 right, since it mistakenly thinks that there are four predictors in the model, affecting the other calculations). This is another indication that the numerical methods used by these versions of Excel are hopelessly out of date, and cannot be trusted.

These problems have been known in the statistical community for many years, going back to the earliest versions of Excel, but new versions of Excel continued to be released without them being addressed. Finally, with the release of Excel 2003, the basic algorithmic instabilities in the regression function LINEST() were addressed, and the software yields correct answers for these regression examples (as well as for the univariate statistics example). Excel 2003 also recognizes the perfect collinearity in the previous example, and gives the slope coefficient for one variable as 0 with a standard error of 0 (although it still tries to calculate a t-test, resulting in $t = 65535$).

Unfortunately, not all of Excel's problems were fixed in the latest version. Here is another data set:

```
 X1   X2

  1    1
  2    2
  3    3
  4    4
  5    5
  6    5
  7    4
  8    3
  9    2
 10    1
```

Let's say that these are paired data, and we are interested in whether the population mean for $X1$ is different from that of $X2$. `Minitab` output for a paired sample t–test is as follows:

```
Paired T-Test and Confidence Interval

Paired T for X1 - X2

                 N      Mean     StDev   SE Mean
X1              10     5.500     3.028     0.957
X2              10     3.000     1.491     0.471
Difference      10      2.50      3.37      1.07


95% CI for mean difference: (0.09, 4.91)
T-Test of mean difference = 0 (vs not = 0): T-Value = 2.34
                                            P-Value = 0.044
```

Here is output from `Excel`:

```
t-Test: Paired Two Sample for Means

                                   Variable 1     Variable 2
Mean                                      5.5              3
Variance                          9.166666667    2.222222222
Observations                               10             10
Pearson Correlation                         0
Hypothesized Mean Difference                0
df                                          9
t Stat                            2.342606428
P(T<=t) one-tail                  0.021916376
t Critical one-tail               1.833113856
P(T<=t) two-tail                  0.043832751
t Critical two-tail               2.262158887
```

The output is (basically) the same, of course, as it should be. Now, let's say that the data have a couple of more observations with missing data:

```
X1   X2

 1    1
 2    2
 3    3
 4    4
 5    5
 6    5
 7    4
 8    3
 9    2
10    1
     10
10
```

Obviously, these two additional observations don't provide any information about the difference between $X1$ and $X2$, so they shouldn't change the paired t–test. They don't change the `Minitab` output, but look at the `Excel` output:

```
t-Test: Paired Two Sample for Means

                           Variable 1     Variable 2
Mean                              5.909090909    3.636363636
Variance                          10.09090909    6.454545455
Observations                               11             11
Pearson Correlation                         0
Hypothesized Mean Difference                0
df                                         10
t Stat                            1.357813616
P(T<=t) one-tail                 0.1021848282
t Critical one-tail               1.812461505
P(T<=t) two-tail                  0.204369656
t Critical two-tail               2.228139238
```

I don't know what Excel has done here, but it's certainly not right! The statistics for each variable separately (means, variances) are correct, but irrelevant. Interestingly, the results were different (but still wrong) in Excel 97, so apparently a new error was introduced in the later versions of the software, which has still not been corrected. The same results are obtained if the observations with missing data are put in the first two rows, rather than the last two. These are **not** the results that are obtained if the two additional observations are collapsed into one (with no missing data), which are correct:

```
t-Test: Paired Two Sample for Means

                                    Variable 1      Variable 2
Mean                               5.909090909     3.636363636
Variance                           10.09090909     6.454545455
Observations                                11              11
Pearson Correlation                 0.35482964
Hypothesized Mean Difference                 0
df                                          10
t Stat                             2.291746243
P(T<=t) one-tail                   0.022440088
t Critical one-tail                1.812461505
P(T<=t) two-tail                   0.044880175
t Critical two-tail                2.228139238
```

Missing data can cause other problems in all versions of Excel. For example, if you try to perform a regression using variables with missing data (either in the predictors or target), you get the error message `Regression - LINEST() function returns error. Please check input ranges again.` This means that you would have to cut and paste the variables to new locations, omitting any rows with missing data yourself.

Other, less catastrophic, problems come from using (any version of) Excel to do statistical analysis. Excel requires you to put all predictors in a regression in contiguous columns, requiring repeated reorganizations of the data as different models are fit. Further, the software does not provide any record of what is done, making it virtually impossible to document or duplicate what was done. In addition, you might think that what Excel calls a "Normal probability plot" is a normal (qq) plot of the residuals, but you'd be wrong. In fact, the plot that comes out is a plot of the ordered target values $y_{(i)}$ versus $50(2i-1)/n$ (the ordered percentiles). That is, it is effectively a plot checking uniformity of the target variable (something of no interest in a regression context), and has nothing to do with normality at all! You should also know that if a column of an `Excel` spreadsheet is too narrow (so that pound signs replace an actual entry), and you copy and paste the column into Word, the pound signs are pasted over, not the actual entry (this cannot happen when pasting from `Minitab`, since it will automatically convert a number to scientific notation and automatically widen a text column to be as wide as is necessary).

To my way of thinking, at a bare minimum, to be considered even remotely useful *any* regression package must do the following (I've left out the obvious things, such as providing accurate least squares estimates, t-tests, F-tests, R^2 values, p-values, and so on):

(1) Provide an audit trail, so that changes in data and different analyses can be tracked and recorded. An alternative model (used by the packages `S-Plus`, `R`, and `SAS`, for example) is the construction and use of a command language, so that a data analyst can write scripts for this purpose.

(2) Allow for totally flexible choices of predictors from a spreadsheet/ worksheet (i.e., nonconsecutive columns).

(3) Have a large set of easy-to-use transformations.

(4) Easily produce **correct** residual plots.

(5) Easily produce columns of standard regression diagnostics (standardized residuals, leverage values, Cook's distances), and fitted values.

(6) Provide variance inflation factors.

(7) Provide confidence and prediction intervals for new observations.

(8) Allow for weights (i.e., weighted least squares).

(9) Handle categorical predictors, at least at a basic level.

Items 1 to 7 are fundamentally necessary even for regression at the level of this course, and if you ever need to perform a regression analysis after you leave the course, you almost immediately need items 8 and 9, so encouraging people to use software to do regression if those things aren't provided is to my mind not a good idea.

I've left other things out that are very useful (for example, the ability to easily create subsets of the data based on conditional statements, reasonably high quality graphics [histograms, scatter plots, side-by-side boxplots, etc.], correct accounting for missing values, built-in probability distributions for the F, t, and binomial, so tests for other hypotheses can be constructed, etc.), but you get the point. The solution to all of these problems is to perform statistical analyses using the appropriate tool — a good statistical package. Many such packages are available, usually with a Windows-type graphical user interface (such as `Minitab`), often costing $100 or less. A remarkably powerful package, `R`, is free!

(See www.r-project.org for information.) If you must do statistical calculations within Excel, there are add-on packages available that do not use the Excel algorithmic engine, but these can cost as much as many standalone packages, and you must be sure that you trust the designers to have carefully checked their code for problems.

Notes: The document "Using Excel for Statistical Data Analysis," by Eva Goldwater, provided some of the original information used here. The document is available on the World Wide Web at

www-unix.oit.umass.edu/∼evagold/excel.html

The United Kingdom Department of Industry's National Measurement System has produced a report on the inadequacies of the intrinsic mathematical and statistical functions in versions of Excel prior to Excel 2003. This 1999 report, written by H.R. Cook, M.G. Cox, M.P. Dainton, and P.M. Harris, is available on the World Wide Web at

www.npl.co.uk/ssfm/download/documents/cise27_99.pdf

Some published discussions of the use of Excel for statistical calculations are given below. The first reference describes other dangers in using Excel (including for purposes for which it is designed!), and gives a link to a document describing how a spreadsheet user can get started using R. References (6) and (11) discuss Excel 2003, noting remaining problems in its statistical distribution functions, random number generation, and nonlinear regression capabilities; reference (8) updates this to Excel 2007 and notes that similar problems still exist. The European Spreadsheet Risks Interest Group web site (www.eusprig.org) contains papers and news stories about potential problems and actual (sometimes multimillion dollar) errors that have occurred from inappropriate spreadsheet usage.

1. Burns, P. (2005), "Spreadsheet addiction," (www.burns-stat.com/pages/Tutor/spreadsheet_addiction.html).
2. Cryer, J. (2002), "Problems using Microsoft Excel for statistics," *Proceedings of the 2001 Joint Statistical Meetings* (www.cs.uiowa.edu/∼jcryer/JSMTalk2001.pdf).
3. Helsel, D.R. (2002), "Is it practical to use Excel for stats?," (http://www.practicalstats.com/Pages/excelstats.html).
4. Knüsel, L. (1998), "On the accuracy of statistical distributions in Microsoft Excel 97," *Computational Statistics and Data Analysis*, **26**, 375–377.
5. Knüsel, L. (2002), "On the reliability of Microsoft Excel XP for statistical purposes," *Computational Statistics and Data Analysis*, **39**, 109–110.

6. Knüsel, L. (2005), "On the accuracy of statistical distributions in Microsoft Excel 2003," *Computational Statistics and Data Analysis*, **48**, 445–449.
7. McCullough, B.D. (2002), "Does Microsoft fix errors in Excel?" *Proceedings of the 2001 Joint Statistical Meetings.*
8. McCullough, B.D. and Heiser, D.A. (2008), "On the accuracy of statistical procedures in Microsoft Excel 2007," *Computational Statistics and Data Analysis*, **52**, 4570–4578.
9. McCullough, B.D. and Wilson, B. (1999), "On the accuracy of statistical procedures in Microsoft Excel 97," *Computational Statistics and Data Analysis*, **31**, 27–37.
10. McCullough, B.D. and Wilson, B. (2002), "On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP," *Computational Statistics and Data Analysis*, **40**, 713–721.
11. McCullough, B.D. and Wilson, B. (2005), "On the accuracy of statistical procedures in Microsoft Excel 2003," *Computational Statistics and Data Analysis*, **49**, 1244–1252.
12. Pottel, H. (2001), "Statistical flaws in Excel," (`www.mis.coventry.ac.uk/~nhunt/pottel.pdf`).
13. Rotz, W., Falk, E., Wood, D., and Mulrow, J. (2002), "A comparison of random number generators used in business," *Proceedings of the 2001 Joint Statistical Meetings.*