® OBJECTIVE ASSESSMENT OF FORECASTING ASSIGNMENTS USING SOME FUNCTION OF PREDICTION ERRORS

CLARKE, Stephen R.
Swinburne University of Technology
Australia

One way of examining forecasting methods via assignments is to give each student a real or simulated set of data, with a requirement to forecast future values. However checking the accuracy of calculations for the host of possible methods can be onerous. One solution is to make part or the entire assessment dependent on the accuracy of the forecasts obtained. This mirrors real life, where forecasts are judged not by the method used but by how accurate the predictions turn out. This paper investigates how this might work with an actual example. Using simulated data from a model which incorporates trend, seasonality, Easter effect and randomness, we use a function of the mean square error of the forecasts to determine the final mark for a variety of methods. Results indicate that the students who have put in more work, and/or fitted the better models, would obtain the better marks.

INTRODUCTION

Before its course was abandoned, the Mathematics Department at Swinburne University put much effort into making graduates of its Mathematics and Computer Science course employable. The practical nature of the course was ensured by the use of a graduated set of project experiences - simple problems, individual and group workshops, extended case studies, group projects using staff as clients, a full year of Industry Based Learning (full time paid relevant employment), and group projects for an industrial client. Bailey and Weal (1993) give a good account of the structure of the first year of the course while Weal (1991) gives an overview of the practical components used during the entire course.

Clarke and Weal (1995) describe the use of workshops in the course. A Workshop refers to a small assignment involving the solution of an unstructured or semi-structured problem. They were used in the first year to wean students off the artificial problems usually used when teaching techniques, as well as to introduce them to the messier and less structured problems they are likely to meet in real life. In the first year of the course there were three workshops, each worth only four marks. One of the first workshops was basically a discriminant analysis problem (although we never referred to it as such to the students, who only had an elementary statistics background at this stage). The data consisted of the results of three different medical measurements on about 20 patients with a certain medical condition, along with similar data for another 20 patients with an alternative condition. The data on a further eight patients whose condition was unknown was given, and the students were asked to allocate each of these patients to the most likely condition. As the workshops were worth so little, it was inefficient to spend a lot of time in marking – our major purpose was to get students thinking about a different sort of problem. We solved this by simply giving half a mark to each correctly allocated patient.

This idea could be applied in a range of learning situations, but it appears particularly well suited to forecasting. One way of examining forecasting methods via assignments is to give each student a real set of data, with a requirement to forecast future values. The task is more realistic if students are given little direction as to method. They then have to decide whether to use naive methods, linear or non linear regression, or one of many exponential smoothing methods to account for possible trend or seasonality or other effect. However this can make marking onerous. Even if students are given the same set of data, checking the accuracy of calculations for the host of possible methods can be at best tedious or at worst impossible. One solution is to make part or the entire assessment dependant on the accuracy of the forecasts obtained. This mirrors real life, where forecasts are judged not by the method used but by how accurate the predictions turn out. The author has used this in a very simple example whereby students had to forecast future daily maximum temperatures. An alternative is to use data simulated from a known model which incorporates effects which students have studied, such as trend, seasonality and randomness. We investigate this approach in this paper.

THE DATA

The model given in Equation 1 is used within Excel to generate five years of fictitious monthly sales figures for a product with extra sales in the southern summer and at Easter.

Sales = 100*trend*monthly index*Easter index*random effect. (1)

Further randomness can be introduced by making any of these parameters random variables rather than constants. Here we make the monthly trend equal to an annual rate of increase of 10%, 11%, 12%, 13% or 14% with equal probability, and choose the other parameters as constants. Monthly indices of 1.17, 1.10, 1.00, 0.91, 0.85, 0.83, 0.85, 0.91, 1.00, 1.10, 1.17, 1.20 were used. Sales are increased by 15% if the month contains Easter (March in 2002, 2005, April in 2001, 2003, 2004, 2006). Finally the random effect is chosen as a uniform distributed fraction between 85% and 115%. Figure 1 gives a single example of the infinite number of data sets that can be generated by the model. It shows monthly sales for the period 2001 to 2005.

Clearly with careful setting up of the spreadsheet, parameters such as the average annual rate of increase, the maximum monthly seasonal index, the Easter index and the range or distribution of the random effect can easily be changed to produce a large number of different series. This allows for students to receive data sets with different structure or parameter values.

Here we assume all students receive the same data set, and are asked to forecast the next 12 months sales. We investigate the ramifications of marking on the closeness of the student’s forecasts to the values given by the model. In generating the extra sales to be forecast, some of the randomness can be removed to make the forecasting task easier for students. For example all the randomness could be removed and the expected values generated by the model used. Here we don’t go this far, and remove the variation in the annual trend, but leave the error term. Figure 1 also shows the model’s sales, with this alteration, for each month in 2006, and these are also given in the first two columns of Table 1.

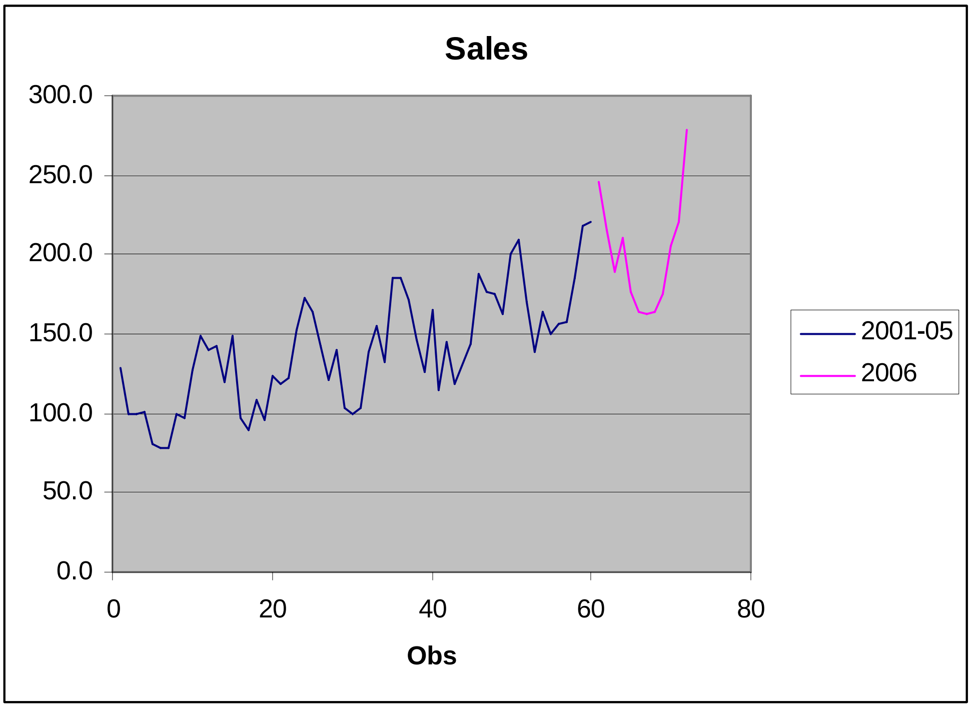

Figure 1: Simulated Sales data

ANALYSIS

The results of several techniques which students may try are shown in Table 1, ranging from simple naïve forecasts to more sophisticated techniques. The table shows the root mean square of the errors (RMSE) for the 60 fitted values, along with the 12 forecast values and their RMSE. The description of each method follows. This is not a comprehensive list of possible methods, merely a selection that students may try. We are interested in how the sophistication of the method is reflected in the closeness of the predicted and forecast values to their actual values.

Table 1

Forecasts and RMSE for each month of 2006, and RMSE of Fitted values for 2001-2005

Period	Sales	*Aver*	*ExpS*	*PrevV*	*OExpS*	*LReg*	*Season*	*Easter*	*AMReg*	*MMReg*
Jan 06	245.8	139.5	172.6	220.8	209.9	183.6	218.2	218.2	204.1	221.7
Feb 06	215.4	139.5	172.6	220.8	209.9	185.1	196.0	196.0	192.3	200.4
Mar 06	189.0	139.5	172.6	220.8	209.9	186.5	195.2	195.2	178.3	180.5
Apr 06	211.0	139.5	172.6	220.8	209.9	187.9	184.5	204.9	197.9	208.0
May 06	175.9	139.5	172.6	220.8	209.9	189.4	144.6	144.6	155.5	149.7
Jun 06	163.9	139.5	172.6	220.8	209.9	190.8	161.8	161.8	169.6	167.0
Jul 06	162.1	139.5	172.6	220.8	209.9	192.3	148.9	148.9	159.7	154.9
Aug 06	164.4	139.5	172.6	220.8	209.9	193.7	178.8	178.8	180.3	186.4
Sep 06	175.3	139.5	172.6	220.8	209.9	195.2	183.2	183.2	184.6	191.5
Oct 06	205.8	139.5	172.6	220.8	209.9	196.6	205.8	205.8	201.6	215.4
Nov 06	220.6	139.5	172.6	220.8	209.9	198.1	240.7	240.7	226.6	253.3
Dec 06	278.8	139.5	172.6	220.8	209.9	199.5	243.2	243.2	229.1	256.5
2006	RMSE	31.4	20.0	17.9	16.1	16.0	9.1	8.5	9.9	8.2
2001-5	RMSE	34.9	27.5	23.7	23.2	24.4	13.5	12.5	11.0	11.7

The first four methods are naïve methods that take little account of the structure of the data. The first uses the overall average (*Aver*), and because it doesn't pick up the general trend the forecasts are all far too low. The second uses exponential smoothing (*ExpS*) with a smoothing parameter of 0.15. This works up to a higher forecast, but is still 30 lower than the actual average of the 2006 sales, and again forecasts badly. Using the previous value (*PrevV*) as the forecast for the next period performs better for the fitted values, presumably because it picks up some of the seasonality as well as the trend. However this is lost when forecasting, as again the forecasts allow for neither seasonality nor trend. This method would also be highly variable, as clearly the forecasts depend on the single value of the last of the known sales. The fourth method avoids this last problem to some extent by using optimal exponential smoothing (*OExpS*) with the smoothing parameter (0.8) chosen to minimize the RMSE of the fitted values. This is easily accomplished using Solver in Excel. This shows only a slight improvement in the forecasts.

Any student who graphs the data as in Figure 1 will see the presence of trend and hopefully also seasonality. The next method fits a simple linear regression (*LReg*) to the raw data, giving Sales = 95.5 + 1.44 x Observation Number. The fitted RMSE shows no improvement over exponential smoothing, and the forecast errors are also about the same. Clearly to make reasonable improvement we need to allow for seasonality.

While there are more complicated methods of calculating seasonal indices, here we simply average the ratio of actual sales and the prediction given by the above linear regression for each month. This gives estimated monthly indices of 1.19, 1.06, 1.05, 0.98, 0.76, 0.85, 0.77, 0.92, 0.94, 1.05, 1.22, 1.22. These are all within 9 percentage points of the actual model values. The next method (*Season*) multiplies the linear regression forecasts by these values and gives a large reduction in both the fitted and forecast errors.

Top students may realize that sales might increase at Easter, particularly if the product is chosen carefully. Chocolate might be too obvious, and something associated with holidays such as duty free sales or demand for holiday accommodation might be more subtle. The difficulty is the Easter holiday moves around between March and April, and the next method (*Easter*) allows for this. Again taking the average ratio of Sales and the previous seasonal forecast for the 5 months containing Easter, we get an Easter index of 1.11 (compared with the model value of 1.15). This makes a slight improvement to the fitted values (as it must) and certainly improves the forecast for April 2006.

Clearly there are alternative and probably better methods of allowing for these effects. Comparing sales to 12 month moving average might be a better way to get seasonal indices, and this effect could be removed before trend is estimated. Easter and seasonal effects also interact. There are also Exponential smoothing methods that allow for trend and seasonality which could be tried. All these methods can be easily implemented using Excel.

Finally, we fit both an additive and multiplicative general linear model with a trend, month and Easter effect. While this could be done in Excel, here Proc GLM in SAS Version 9 was used. In the first case we use the raw sales figures to fit an additive model (*AMReg*). While this gives the best fit, it does not produce as good a forecast as the previous model. This is presumably because it fails to pick up the multiplicative nature of the model. While Figure 1 does not clearly show an increase in variance, most sales figures would be expected to grow exponentially rather than linearly, and it is reasonable to expect a multiplicative model. We fit this multiplicative model (*MMReg*) by using the previous method after taking logs of sales figures. This gives the best forecasts with a RMSE of only 8.2. Interestingly these are not all that much better than the student who uses some simple statistical theory and common sense.

Also of interest is the fact that the predicted RMSE is in each case less than the fitted RMSE. This is the reverse of the usual, where one would expect the fit to future values to be worse than the fit to the values from which the fitted model parameters have been derived. This is presumably because we have used the expected values of trend, rather than random ones. This reduces the variation by removing the variation due to the random error of the trend. It would be possible to use expected values with no error, or fully simulated values for the future sales, depending on the level of randomness with which teachers and students feel comfortable. Here the middle road was chosen.

Note also that in general the better the fit, the better the forecast. Thus a student who tries different methods and improves the fit, will generally be rewarded with better forecasts.

ASSESSMENT

The question now addressed is converting the closeness of the obtained solution to a mark. One could use a straight percentage of the best RMSE (as defined by that obtained by the lecturer, or by the best of the students) over that obtained by the student. This for instance would give 100* 8.2/31.4 = 26% for the simple average method, and 100*8.2/16.0 = 51% for the simple linear regression method. These marks might be considered too high if students had covered seasonality in lectures, but reasonable if the assignment was a preliminary one to introduce students to some forecasting concepts. Alternatively one could use the percentage of maximum reduction in RMSE. Thus the average gets zero, and the linear regression gets 100*(31.4-16.0) / (31.4-8.2) =66%.

One of the problems with these methods is that the law of diminishing returns means students are not rewarded for difficult small improvements once the main effects are accounted for. Using sums of squares of the error might go some way to correcting this. A better method is to create a table where a certain RSME corresponds to a particular mark, and use interpolation between these values. This allows the examiner great flexibility, and allows for a selection with easy interpretation of marks. Thus for example the table could be chosen so that a student who allows for a trend (which Table 1 shows should result in a RMSE of about 16) might generally obtain a pass (50%).. However to obtain a credit (65%) it would be necessary to allow for seasonality (RMSE of 9 from Table 1). This might result in a conversion as shown in Table 2.

Table 2
Possible conversion of RMSE to percentage mark.

RMSE	Mark
30	20%
20	40%
16	50%
9	65%
8	100%

Such a system makes marking virtually automatic with the use of a spreadsheet. Students submit their fitted values and forecasts electronically on a supplied spreadsheet. These are pasted into a master which then calculates the RMSE and mark.

ALTERNATIVE DATASETS

The question arises as to whether the results achieved here are just due to the particular simulated values – can a lecturer be sure that other values will yield the same results. For those methods that do not involve the use of Solver, Excel can be used to simulate several trials. This eliminates the two multiple regression methods and the *OExpS* method. Table 3 shows the above results along with the results obtained from 10 further simulations of the data. Clearly the order is generally maintained. On one occasion the *Easter* adjustment worsens the result of the *Season* method, and the *ExpS* method is as likely as not to better the *PrevV* method. But generally the table gives some confidence that the interpolation method above would yield reasonable results whatever data set was used.

However the table suggests that one needs to be careful in using separate data for each student. For example the RMSE for a student using *PrevV* who was lucky enough to get simulation 6 data would obtain a RMSE of 13.9, better than the students using *LReg* who received any data sets except Simulation 2, 3, 6 and 9. Such inconsistencies are few however. The *Aver* method never does better than *ExpS* or *PrevV*, which in turn never do better than the *Season* method. The inequities that do exist might be minimised by using larger growth percentages, Seasonal and Easter indices or smaller random errors. Alternatively using a spreadsheet such as that developed for this paper, it is possible to continually generate data sets and select for students only those for which the *Easter* method gave similar RMSEs.

Table 3
RMSE for different methods using 11 different data sets.

Simulation	*Aver*	*ExpS*	*PrevV*	*LReg*	*Season*	*Easter*
Above	31.4	20.0	17.9	16.1	9.1	8.5
1	31.8	21.7	16.7	17.0	10.3	9.5
2	27.3	17.4	21.3	12.5	7.1	7.7
3	28.0	16.4	23.6	13.5	7.6	7.6
4	35.1	22.7	17.5	18.1	12.7	12.3
5	30.2	18.7	26.7	14.6	9.6	9.4
6	29.9	18.1	13.9	11.8	10.3	10.2
7	29.2	17.7	24.0	15.4	11.0	10.8
8	34.8	23.2	19.2	18.2	11.5	11.1
9	28.5	15.9	20.1	13.2	7.0	6.8
10	34.1	22.2	22.2	18.0	9.9	9.6

FORECAST ERROR

For more advanced students, an alternative idea is to mark on the degree to which students forecasts match their confidence. Any forecast should really be accompanied by an estimate of error, or confidence interval. Students are then marked on a function that incorporates both their accuracy and their error. An accurate forecast with high confidence receives more reward than an accurate one with low confidence. An inaccurate forecast made with a high degree of confidence would receive less than one made with less confidence. Such a method has been used to mark AFL football tipping competitions. Many footy tipping competitions require the participant to nominate the winning team, others the margin. Dowe (1996) describes a probabilistic football tipping competition in which the person nominates the probability of each team winning, and the likelihood of the actual result [log(prob of winning team)] is used to determine a score. On the same site, the Gaussian competition involves the tipper nominating a winning margin and a standard deviation, and the reward is a constant plus the logarithm of the probability they assigned to the winning margin (see http://www.csse.monash.edu.au/~footy/about.shtml#gauss).

In our case we might use the probability of the forecast being within 25 of the correct sales, assuming the normal distribution of errors with a standard deviation equal to the fitted RMSE. This is easily accomplished with the NORMAL functions in Excel. Table 4 shows the values obtained for each method for each month along with the total probability. Examination of the table clearly shows the strengths and weaknesses of each method. The average does poorly at the extremes and better in the middle. As expected, the final three methods all do better with the April forecast. Again some scaling of the total probability to produce marks seems feasible. The final row shows the Log likelihood – the sum of the logs of the probability. This shows a greater range than the sum of the probabilities, and clearly penalizes those forecasts that are way out. Thus because of the very poor Jan and Dec forecasts, the *AMReg* method gets a lower score than either the *Season* and the *Easter* method,

Table 4
Probability of actual value being within 25 of forecast.

	Aver	*ExpS*	*PrevV*	*OExpS*	*LReg*	*Season*	*Easter*	*AMReg*	*MMReg*
Jan	0.01	0.04	0.48	0.31	0.06	0.42	0.42	0.06	0.53
Feb	0.07	0.25	0.70	0.71	0.40	0.66	0.67	0.57	0.80
Mar	0.22	0.56	0.38	0.55	0.69	0.91	0.93	0.90	0.92
Apr	0.09	0.30	0.67	0.72	0.51	0.46	0.93	0.86	0.96
May	0.33	0.63	0.20	0.34	0.62	0.32	0.31	0.66	0.46
Jun	0.43	0.61	0.09	0.18	0.45	0.93	0.95	0.96	0.96
Jul	0.44	0.60	0.08	0.16	0.40	0.81	0.83	0.97	0.93
Aug	0.42	0.62	0.09	0.19	0.42	0.78	0.80	0.80	0.60
Sep	0.34	0.63	0.19	0.33	0.55	0.89	0.91	0.92	0.77
Oct	0.11	0.37	0.62	0.71	0.66	0.94	0.95	0.97	0.90
Nov	0.05	0.20	0.71	0.67	0.52	0.64	0.65	0.96	0.26
Dec	0.00	0.00	0.08	0.03	0.01	0.22	0.20	0.01	0.59
Total	2.53	4.81	4.28	4.90	5.30	7.97	8.55	8.64	8.69
LogL	-28.6	-17.9	-16.4	-14.1	-13.8	-6.0	-5.2	-8.8	-4.6

CONCLUSION

To reduce the burden of marking, teachers should consider using the accuracy of forecasts as the basis for marking, rather than a detailed examination of the method used. This mirrors real life, and is particularly suitable for assignments. The above analysis shows that in general increased sophistication of a fitted model produces higher accuracy, that the accuracy of the forecasts is highly correlated with the accuracy of the fitted model, and that the rank order of forecast accuracy of the different methods is reasonably independent of the data set used. This suggests that a marking scheme based on the RMSE of the forecasts is feasible. Such a scheme could be extended to include assessing confidence limits on the forecasts. In general this should give the students who have put in more work, and/or fitted the better models, the better marks. Such an assessment method has the advantage of not only needing little time to mark, but by being completely objective. Lecturers nervous about such a scheme could test the waters by allocating a portion of the assignment mark to be determined in such a manner.

REFERENCES

Bailey, T., & Weal, S. E. (1993). Introducing undergraduates to the spirit of OR whilst imparting substantive skills. *Journal of the Operational Research Society*, 44, 897-908.

Clarke, S. R. & Weal, S. (1995). Workshops: a valuable tool to teach problem solving. In D. L. Hoffman, (Ed.), *Network Conferencing, proceedings of ASOR: The 13th National Conference*, (pp. 57-74). Australian Society of Operations Research.

Dowe, D. L., Farr, G. E., Hurst, A. J. & Lentin, K. (1996). Information theoretic football tipping. In N. de Mestre (Ed.), *Third conference on: Mathematics & Computers in Sport conference*, (pp. 233-242). Bond University: Gold coast, Qld.

Weal, S. E. (1991). Practical OR for undergraduates in the Swinburne course. *Journal of the Operational Research Society*, 42, 1047-1059.