NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF FINANCE AND MANAGEMENT SCIENCE

FOR 16 2008

ISSN: 1500-4066

SEPTEMBER 2008

Discussion paper

A regression surprise resolved

BY

JOSTEIN LILLESTØL AND **JONAS ANDERSSON**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS AND BUSINESS ADMINISTRATION

A regression surprise resolved

Jostein Lillestøl and Jonas Andersson*

Abstract

In this note we explore the following surprising fact: In regression with trend and seasonality, the prediction risk is constant for all seasons of a new cycle, despite the fact that it increases with time when the seasons are left out. Awareness of this may be useful to both the practicing statistician and to teachers of statistics. The challenge of resolving the issue may also be given to students of statistics as a research project.

KEY WORDS: Trend and seasonality; Prediction risk; Paradox.

1 The surprise

Regression analysis is in the nucleus of most elementary statistics courses, starting with the case of one explanatory variable. After the basics the students may learn about the error risks when using an estimated regression line for prediction. They may also learn that this risk increases as the value of the explanatory variable departs from the mean in the data set used for estimation. Some will learn this through the formula

$$SE(\hat{Y} - Y) = S \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{1}$$

which clearly expose this feature. Some may also learn that this extends to multiple regression, and even be presented with the corresponding formula in matrix terms. The students may also be told that regression can be used as framework for time series prediction in the case of trend and seasonality. The case of linear trend only becomes a special case of the formula above, where predictions for the future get gradually less reliable with the number of steps ahead. Seasonality is commonly taken into account by introducing a 0-1 indicators (dummies) for each season, and use all except one as variables in a multiple regression. As before we expect that the errors of prediction will increase with the the prediction horizon, but now we are faced with a surprise. Consider the following simple example with quarterly data observed for two years with a yearly season:

Time:	1	2	3	4	5	6	7	8
Obs Y:	3	2	4	6	4	3	7	9

Below is the standard regression estimation and predictions for each quarter of a third year, first by fitting a trend only and then with both trend and season, here represented by quarterly seasonal dummies Q1, Q2, Q3 and Q4, using the first quarter as basis.

```
The regression equation is Y = 1.43 + 0.738 t

Predictor     Coef  SE Coef     T      P
 Constant    1.429    1.297  1.10  0.313
 t          0.7381   0.2568  2.87  0.028

S = 1.66428   R-Sq = 57.9%   R-Sq(adj) = 50.9%

Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI           95% PI
      1   8.071   1.297  (4.898; 11.245)  (2.909; 13.234)
      2   8.810   1.530  (5.066; 12.554)  (3.278; 14.341)
      3   9.548   1.770  (5.217; 13.878)  (3.603; 15.492)
      4  10.286   2.014  (5.358; 15.214)  (3.893; 16.679)
```

*jostein.lillestol@nhh.no and jonas.andersson@nhh.no, Norwegian School of Economics and Business Adminstration, Department of Finance and Management Science, Helleveien 30, 5117 Bergen, Norway

```
Values of Predictors for New Observations

New Obs     t
      1   9.0
      2  10.0
      3  11.0
      4  12.0

The regression equation is

Y = 2.00 + 0.500 t - 1.50 Q2 + 1.00 Q3 + 2.50 Q4

Predictor     Coef  SE Coef      T      P

Constant    2.0000    0.7217   2.77   0.069
 t          0.5000    0.1443   3.46   0.041
 Q2        -1.5000    0.8292  -1.81   0.168
 Q3         1.0000    0.8660   1.15   0.332
 Q4         2.5000    0.9242   2.71   0.073

S = 0.816497   R-Sq = 94.9%   R-Sq(adj) = 88.2%

Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI           95% PI
      1   6.500   1.041  (3.188;  9.812)  (2.290; 10.710)
      2   5.500   1.041  (2.188;  8.812)  (1.290;  9.710)
      3   8.500   1.041  (5.188; 11.812)  (4.290; 12.710)
      4  10.500   1.041  (7.188; 13.812)  (6.290; 14.710)

Values of Predictors for New Observations

New Obs       t      Q2      Q3      Q4
      1     9.0    0.00    0.00    0.00
      2    10.0    1.00    0.00    0.00
      3    11.0    0.00    1.00    0.00
      4    12.0    0.00    0.00    1.00
```

We see for the simple regression that the standard errors of fit (SE Fit) are increasing for each of the quarters of out of sample prediction, thus giving confidence intervals (CI) of increasing length, and subsequently also prediction intervals (PI) of increasing length. However, for the multiple regression the standard errors of fit (SE Fit) are all equal, thus giving confidence intervals (CI) of equal length, and subsequently also prediction intervals (PI) of equal length. This may come as a surprise, and some may call it a paradox. The first thought may be: Here is something wrong! It isn't! But how can it be explained intuitively and analytically

2 Formal analysis

In general suppose that the seasonal length is s, and that the number of observations n is a multiple of s, so that $n = k \cdot s$ Let X be the $n \times (s+1)$ regression matrix with columns representing successively the constant term, time, and the $s-1$ seasonal indicators. Furthermore let X_0 be the corresponding $s \times (s+1)$-matrix of predictor variables for the seasonal cycle to follow. The covariance matrix of the s fits is then proportional to

$$X_0(X'X)^{-1}X_0' \tag{2}$$

For the covariance matrix of the s prediction errors we have to add one to the diagonal elements. The challenge is now to show that all elements on the diagonal of this matrix are equal, leading to equal prediction variances for all seasons in the first cycle of seasons to come. Moreover, we may show that all elements off the diagonal are equal as well, leading to equal covariances between predictions. These results are due to the very specific structure of X_0 in conjunction with X.

For illustrative purposes take $s = 4$. The regression matrix for the standard representation of the problem, matching the computer output when $n = 8$:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & n & 0 & 0 & 1 \end{bmatrix} \tag{3}$$

and the predictor matrix is

$$X_0 = \begin{bmatrix} 1 & n+1 & 0 & 0 & 0 \\ 1 & n+2 & 1 & 0 & 0 \\ 1 & n+3 & 0 & 1 & 0 \\ 1 & n+4 & 0 & 0 & 1 \end{bmatrix} \tag{4}$$

It follows easily that

$$X'X = \begin{bmatrix} n & \frac{n(n+1)}{2} & \frac{n}{4} & \frac{n}{4} & \frac{n}{4} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} & \frac{n^2}{8} & \frac{n^2+2n}{8} & \frac{n^2+4n}{8} \\ \frac{n}{4} & \frac{n^2}{8} & \frac{n}{4} & 0 & 0 \\ \frac{n}{4} & \frac{n^2+2n}{8} & 0 & \frac{n}{4} & 0 \\ \frac{n}{4} & \frac{n^2+4n}{8} & 0 & 0 & \frac{n}{4} \end{bmatrix} \tag{5}$$

The inverse of this matrix is

$$(X'X)^{-1} = \frac{1}{n^3-16n} \begin{bmatrix} 7n^2-12n-52 & 12-6n & -4n^2+6n+52 & -4n^2+12n+40 & -4n^2+18n+28 \\ 12-6n & 12 & -12 & -24 & -36 \\ -4n^2+6n+52 & -12 & 8n^2-116 & 4n^2-40 & 4n^2-28 \\ -4n^2+12n+40 & -24 & 4n^2-40 & 8n^2-80 & 4n^2+8 \\ -4n^2+18n+28 & -36 & 4n^2-28 & 4n^2+8 & 8n^2-20 \end{bmatrix} \tag{6}$$

from which we get

$$X_0(X'X)^{-1}X_0' = \frac{1}{n^2-4n} \begin{bmatrix} 7n-4 & 3n+12 & 3n+12 & 3n+12 \\ 3n+12 & 7n-4 & 3n+12 & 3n+12 \\ 3n+12 & 3n+12 & 7n-4 & 3n+12 \\ 3n+12 & 3n+12 & 3n+12 & 7n-4 \end{bmatrix} \tag{7}$$

with equal diagonal elements and equal off-diagonal elements as claimed. In general we have

$$X'X = \begin{bmatrix} n & \frac{n(n+1)}{2} & \frac{n}{s} & \frac{n}{s} & \dots & \frac{n}{s} \\ \frac{n(n+1)}{2} & \frac{n(n+1)(2n+1)}{6} & \frac{n}{s}(2+\frac{n-s}{2}) & \frac{n}{s}(3+\frac{n-s}{2}) & \dots & \frac{n}{s}(s+\frac{n-s}{2}) \\ \frac{n}{s} & \frac{n}{s}(2+\frac{n-s}{2}) & \frac{n}{s} & 0 & \dots & 0 \\ \frac{n}{s} & \frac{n}{s}(3+\frac{n-s}{2}) & 0 & \frac{n}{s} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{n}{s} & \frac{n}{s}(s+\frac{n-s}{2}) & 0 & 0 & \dots & \frac{n}{s} \end{bmatrix} \tag{8}$$

Although this matrix has a patterned structure that can be utilized for obtaining an expression for the inverse, it turns out that this expression is fairly complicated. Nevertheless, by pre- and post-multiplication by X_0 each element simplifies "like magic" at the end. However, this requires painstaking book-keeping of terms, and provides no general insight to the issue. At this point it may be tempting to try a symbolic calculator, like the open source Maxima, to provide expressions in terms of n for various s, and conjecture the result, as we did. However, this gave no general insight either. There must be a better way!

3 Alternative formulation

Consider the alternative representation of the situation, where we instead of the constant term use all s seasonal indicators and represent time by s times the number of completed seasonal cycles. This variable is constant within each seasonal cycle and with k observed seasonal cycles, the last value is $(k-1)\cdot s$ and the following one corresponds to $k \cdot s = n$. In conjunction with the seasonal dummies we know for each observation exactly the time, and we essentially have the same time scale as before. As an illustration take the introductory case of $n = 8$ and $s = 4$ with just $k = 2$ seasonal cycles. We then have the regression matrix

$$X = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 4 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{9}$$

and the predictor matrix

$$X_0 = \begin{bmatrix} 8 & 1 & 0 & 0 & 0 \\ 8 & 0 & 1 & 0 & 0 \\ 8 & 0 & 0 & 1 & 0 \\ 8 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

Note that these matrices may be obtained by a nonsingular linear transformation of the columns of the earlier representation, taking $X_{new} = X_{old} \cdot C$, as follows:

$$C = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & 0 & 0 \\ -2 & -1 & 0 & 1 & 0 \\ -3 & -1 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

that is, the new column 1 is obtained from col2 by subtracting (col1 + col3 + 2 col4 + 3 col5) and new column 2 is obtained from col1 by subtracting (col3+col4 + col5). The columns 3, 4 and 5 are left unchanged.

With this equivalent formulation of the problem, the apparent paradox has found a transparent solution. For a given seasonal cycle, the seasons are exchangeable by design. What matters is just the first column, and this is constant for the new seasonal cycle as well. This also answers the question of what happens when prediction is extended beyond one seasonal cycle. Then the variances will increase to a new level, but still constant throughout this cycle.

Even if the issue now has found a transparent solution, it is of interest to pursue the analytical details one step further. In general we now have

$$X'X = \begin{bmatrix} \frac{n(n-s1)(2n-s)}{6} & \frac{n(n-s)}{2s} & \frac{n(n-s}{2s} & \dots & \frac{n(n-s)}{2s} \\ \frac{n(n-s)}{2s} & \frac{n}{s} & 0 & \dots & 0 \\ \frac{n(n-s)}{2s} & 0 & \frac{n}{s} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{n(n-s)}{2s} & 0 & 0 & \dots & \frac{n}{s} \end{bmatrix} \tag{12}$$

We see that this matrix has a simpler structure than the one for the earlier representation. In fact, by cofactor expansion it is not difficult to show that

$$(X'X)^{-1} = \frac{1}{n(n+s)} \begin{bmatrix} \frac{12}{n-s} & -6 & -6 & \dots & -6 \\ -6 & s(n+s)+3(n-s) & 3(n-s) & \dots & 3(n-s) \\ -6 & 3(n-s) & s(n+s)+3(n-s) & \dots & 3(n-s) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -6 & 3(n-s) & 3(n-s) & \dots & s(n+s)+3(n-s) \end{bmatrix} \tag{13}$$

From this we get

$$X_0(X'X)^{-1}X_0' = \frac{1}{n(n-s)} \begin{bmatrix} s(n-s)+3(n+s) & 3(n+s) & 3(n+s) & \dots & 3(n+s) \\ 3(n+s) & s(n-s)+3(n+s) & 3(n+s) & \dots & 3(n+s) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 3(n+s) & 3(n+s) & 3(n+s) & \dots & s(n-s)+3(n+s) \end{bmatrix} \tag{14}$$

For $s = 4$ we see that this coincides with the result given for example of the preceding section. We note that the ratio between the covariance and the variance of the fits converges to $3/(3+s)$ when $n \to \infty$.

For the variance of the prediction error we have to add one to the diagonal elements, which then is simplified to

$$\frac{(n+s)(n-s+3)}{n(n-s)} \tag{15}$$

We note that the ratio between the covariance and the variance is $3/(n+3-s)$ which, of course, converges to zero when $n \to \infty$.

In terms of $k = n/s$ the variance may be written as

$$\frac{k+1}{k} \cdot (1 + \frac{3}{s(k-1)}) \tag{16}$$

We see that this decreases with s, despite the increase in the number of parameters to be estimated. We now get the following prediction interval for predicting each season of the coming seasonal cycle

$$\hat{Y} \pm t_{n-s-1,1-\alpha/2} \cdot S \sqrt{\frac{k+1}{k} \cdot (1 + \frac{3}{s(k-1)})} \tag{17}$$

where $t_{n-s-1,1-\alpha/2}$ is the $1-\alpha/2$ quantile of the t-distribution with $n-s-1$ degrees of freedom. In Table 1 we give the standard deviations as functions of s and k.

We could easily derive the formulas for prediction beyond one seasonal cycle as well. Since this is seldom requested, we omit this here. We have throughout assumed that the number of consecutive observations is a multiple of the seasonal length. If this is not the case the variances of prediction errors for a full seasonal cycle will be constant for the seasons up to the last season observed and then increase.

k	2	3	4	5	6	7	8	9
s= 4	1.620	1.354	1.250	1.194	1.158	1.134	1.116	1.102
s=12	1.369	1.225	1.164	1.129	1.107	1.091	1.079	1.070

Table 1: Standard deviations as functions of s and k

4 Further issues

The software specification of the reformulated model (II) will be as a non-intercept regression with $s+1$ explanatory variables, instead of a regression with intercept with s explanatory variables. For our example we get the following estimated equation.:

```
The regression equation is

Y = 0.500 ts + 2.50 Q1 + 1.50 Q2 + 4.50 Q3 + 6.50 Q4
```

Here ts denotes the new time variable, and it is worth noting that its regression coefficient 0.5 coincides with the one for the trend variable of the first formulation, which will be the case in general. Note also that the (initial) level is now represented by the regression coefficients of the seasonal dummies.

The regression coefficients for the seasonal dummies are often interpreted as seasonal factors or indices. In the current context the result will depend on the formulation chosen. For the data in our example the indices for the two formulations, using the first quarter as basis are given in Table 2.

Season	Q1	Q2	Q3	Q4
Index (I)	0	-1.5	1.0	2.5
Index (II)	0	- 1.0	2.0	4

Table 2: Seasonal indices (first quarter as basis

We see that the second formulation (II) "favors" the subsequent quarters by adding a 0.5 to each subsequent quarter in the cycle. This means that although we have the same trend scale, and same trend regression coefficient, the upward trend effect also shows up in the seasonal indices. This is of course something to be aware of when interpreting regression coefficients in the case of both trend and season.

The regression coefficient vector is given by

$$\hat{\beta} = (X'X)^{-1}X'y \tag{18}$$

Having the regression vector for formulations (II), the one for the standard formulation (I) may be obtaining by pre-multiplication of the matrix C.

From the latter formula we may expect simplified expressions in terms of n and s as well. However, both the first and the second formulation lead to tedious calculation where again things gtreatly simplifies at the end. A possibility is to modify the first column of X by subtracting its mean $(n-s)/2$, thus having elements in k blocks of size s ranging from $-(n-s)/2$ to $(n-s)/2$, and where the first column of X_0 is s elements equal to $(n+s)/2$. With this specification (III) all columns of X are orthogonal, and both $X'X$ and $(X'X)^{-1}$ become diagonal matrices. In fact

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{a} & 0 & 0 & \dots & 0 \\ 0 & \frac{s}{n} & 0 & \dots & 0 \\ 0 & 0 & \frac{s}{n} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{s}{n} \end{bmatrix} \tag{19}$$

where for $k = n/s$

$$a = \sum_{i=1}^{k}(s(2i-1)-n)^2 = \frac{1}{12}n(n+s)(n-s) = \frac{s^3}{12}k(k+1)(k-1) \tag{20}$$

Calculations show that

$$\hat{\beta} = \begin{bmatrix} \frac{1}{2a}\sum_{i=1}^{k}(s(2i-1)-n)\sum_{j=1}^{s} y_{ij} \\ \frac{1}{k}\sum_{i=1}^{k} y_{i1} \\ \frac{1}{k}\sum_{i=1}^{k} y_{i2} \\ \vdots \\ \frac{1}{k}\sum_{i=1}^{k} y_{is} \end{bmatrix} \tag{21}$$

where y_{ij} denotes the observation of the jth season in the ith seasonal cycle. Thus the seasonal regression coefficients for this formulation are simply their means.

```
The regression equation is

Y = 0.500 ts + 3.50 Q1 + 2.50 Q2 + 5.50 Q3 + 7.50 Q4
```

The transformation from this formulation (III) to formulation (II) is now transparent: The time-trend coefficient stays the same, while each seasonal coefficient is subtracted by mean of the time-trend variable (ts) times is regression coefficient, in our example $2 \times 0.5 = 1.0$.

We could of course have adopted the formulation (III) at the outset, and arrive at our formulas in Section 3 in a slightly less tedious way. However, we may not be willing to trade this technicality against a simpler and more natural specification. Anyway, the three specifications have provided some insight to the problem under study.

Another interesting question is how the prediction variance is behaving when we have a covariate to the trend and seasonal specification. Suppose we have no information about the future values of this covariate, and that they are replaced by their observed means in the prediction formula. If this is done for each season of the coming seasonal cycle, can we expect the variances of fits and prediction errors still to be equal? If not, does this provide a lower bound on the variances? The answer is no to both questions. The variances will now depend on the distribution of the covariate within and between each season. Specific seasons with observed average coinciding with the total average gives identical variances, despite different within spread. Seasons with observed mean at equal distance may give different variances, depending on the observed within spread. It may even happen that this is less than the one computed for a season with observed mean equal to the total mean.

5 Final comment

The issue presented in this note is about the observation of, at least to some of us, an unexpected regression result. Its resolution may involve several ingredients and skills, ranging from model formulation, elementary statistical theory and linear algebra, as well as creative use of software. As such it may be used as basis for a student project with elements of research.