

# Two Topics in Mathematics and Procurement

Przemyslaw Stanislaw Stilger  
3432564

June 22, 2011

### **Abstract**

In the 1st part of the paper, I study a method of detecting anomalous bids in public procurement. I will then give a detailed treatment of principal component analysis. Next, a statistical method based on principal components to detect anomalous bids will be described. I show that under certain assumptions, this statistic follows the non-central  $\chi^2$  distribution. Finally, I give an example based on the data from a real tender to show how this statistic can detect anomalous bids. In the 2nd part, I investigate whether bidding performance of 6 companies in 5 tenders organized by 1 tendering entity is influenced by learning using random effects logistic regression.

### **Acknowledgments**

I am very much indebted to Prof. dr. ir. Erik J. Balder, dr. Alexander Gnedin, Prof Ian Jolliffe and Jan Siderius.

# Contents

<b>1</b>	<b>Principal Components Analysis and Detection of Anomalous Bids in Public Procurement</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Principal Component Analysis . . . . .	2
1.3	Statistic for Detecting Anomalous Bids . . . . .	8
1.4	Example . . . . .	10
1.5	Conclusion . . . . .	12
<b>2</b>	<b>Influence of Learning on the Bidding Performance</b>	<b>13</b>
2.1	Conclusion . . . . .	16
<b>A</b>	<b>Appendix</b>	<b>17</b>
A.1	Kolmogorov-Smirnov Test . . . . .	17
A.2	Anderson-Darling Test . . . . .	17
A.3	Shapiro-Wilk Normality Test . . . . .	18
A.4	Non-central $\chi^2$ Distribution . . . . .	19
A.5	Scatter Plot . . . . .	20
A.6	Non-Central $\chi^2$ Probability Plot . . . . .	21
A.7	Tenders Visualized . . . . .	22
A.8	Significance Tests . . . . .	24
A.9	<i>McFadden's Pseudo R<sup>2</sup></i> . . . . .	24

# 1 Principal Components Analysis and Detection of Anomalous Bids in Public Procurement

## 1.1 Introduction

In public procurement, contracts are awarded based on the lowest price or the Most Economically Advantageous Tender. As for anomalous bids, from a statistical point of view, these are the outliers. If the award criterion is the lowest price, an anomalous bid may be defined as a bid below certain threshold. For example, Spanish Ministry of Economy and Finance considers bids anomalous if at least 3 bids have been submitted that are lower than the average bid by at least 10%. If at least 2 bids have been submitted, an anomalous bid is the one that is lower than the other by at least 25%.

There are several reasons why anomalous bids occur[3]:

- Bidder underestimated the cost.
- Bidder is in desperate need of obtaining the contract and expects to renegotiate the contract.
- The bidder's financial conditions are in a poor state, forcing them to use the award of the contract as their last resort.
- Bidder is aiming at ousting other competitors, which is also known as predatory bidding.

If the award criteria is the Most Economically Advantageous Tender, it implies that other award criteria are taken into account in addition to the price. It is therefore assumed that bids are evaluated based on price and quality. This makes the anomalous bid detection problem multivariate. A naive approach to detect anomalous bids would be to define certain thresholds for both price and quality. However, a major problem in detecting multivariate outliers is that an observation that is not extreme on any of the original variables can still be an outlier. It is impossible to detect such outliers by only looking at the original variables. It is often an unusual combination of values of the original variables that makes an observation an outlier. Consider an example taken from[9]. Suppose that heights and weights are measured for children of various ages between 5 and 15 years. An observation with height and weight of 175 cm and 25 kg is not particularly extreme on either the original variables, as 175 cm is a plausible height for the older children and

25 kg is a plausible weight for the youngest children. However, the combination 175 cm and 25 kg is virtually impossible, and will be a clear outlier because it combines a large height with a small weight, thus violating the general pattern of a positive correlation between these 2 variables. Therefore, I apply a statistic based on principal components which is equivalent to the squared Mahalanobis distance to detect anomalous bids.

## 1.2 Principal Component Analysis

Assume that the data can be represented by  $n \times m$  matrix of real numbers, where  $n$  denotes the number of observations and  $m$  denotes the number of variables. Mathematically,

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix}.$$

Setting

$$\mathbf{x}_j = \begin{bmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{bmatrix},$$

we can write

$$\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_m].$$

Principal component analysis is a statistical technique that linearly transforms an original set of  $m$  variables into a set of  $k$  uncorrelated variables where  $k \leq m$ . It searches for  $k$  uncorrelated linear combinations of the original variables that capture the most of the information contained in the original variables. Principal components analysis has no underlying statistical model of the original variables and focuses on explaining the total variability of observations on the basis of the maximum variance property of the principal components.

Algebraically, the 1st principal component,  $\mathbf{y}_1$ , is a linear combination given by  $a_{1,1}\mathbf{x}_1 + \cdots + a_{1,m}\mathbf{x}_m$ , where  $\mathbf{a}_1' = [a_{1,1} \cdots a_{1,m}]$  is the weight vector. The weight vector is chosen such that the variance of  $\mathbf{y}_1$  is maximized given the constraint that the sum of the squared weights is equal to 1, i.e.  $\sum_{i=1}^m a_{1,i}^2 = 1$ . Hence the problem is to find such  $\mathbf{a}_1$  that  $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ , where  $\mathbf{S}$  denotes the sample covariance matrix, is maximized subject to the constraint  $\mathbf{a}_1'\mathbf{a}_1 = 1$ . This

constraint is necessary for a unique solution to exist. The problem of finding the optimal weights for the largest principal component involves multivariate calculus and theory of latent roots and latent vectors. It can be easily verified from the standard theory of differential calculus that the conditions for vector  $\mathbf{a}_1'$  to maximize  $\mathbf{a}_1' \mathbf{S} \mathbf{a}_1$  subject to the constraint  $\mathbf{a}_1' \mathbf{a}_1 = 1$  are precisely the same as those for the vector  $\mathbf{a}_1$  to maximize  $\mathbf{a}_1' \mathbf{S} \mathbf{a}_1 - \lambda_{1,1} (\mathbf{a}_1' \mathbf{a}_1 - 1)$ , where  $\lambda_{1,1}$  is a constant known as the Lagrange multiplier. Let

$$\begin{aligned} V_1 &= \mathbf{a}_1' \mathbf{S} \mathbf{a}_1 - \lambda_{1,1} (\mathbf{a}_1' \mathbf{a}_1 - 1) \\ &= \sum_{i=1}^m \sum_{j=1}^m a_{1,i} a_{1,j} s_{i,j} - \lambda_{1,1} \left( \sum_{i=1}^m a_{1,i}^2 - 1 \right). \end{aligned}$$

Then

$$\frac{\partial V_1}{\partial a_{1,k}} = 2 \sum_{j=1}^m s_{k,j} a_{1,j} - 2\lambda_{1,1} a_{1,k},$$

where  $k = 1, \dots, m$ .

To find the vector  $\mathbf{a}_1$  maximizing  $V_1$  we thus set  $\frac{\partial V_1}{\partial a_{1,k}} = 0$  for all  $k$  and solve the resulting set of simultaneous equations. Now

$$\frac{\partial V_1}{\partial a_{1,k}} = 0 \Rightarrow \sum_{j=1}^m s_{k,j} a_{1,j} = \lambda_{1,1} a_{1,k}.$$

$\sum_{j=1}^m s_{k,j} a_{1,j}$  is the  $k$ th element of  $\mathbf{S} \mathbf{a}_1$ , while  $\lambda_{1,1} a_{1,k}$  is the  $k$ th element of  $\lambda_{1,1} \mathbf{a}_1$ . Thus, when all equations are considered simultaneously, it follows that the maximizing value of  $\mathbf{a}_1$  must satisfy

$$\begin{aligned} \mathbf{S} \mathbf{a}_1 &= \lambda_{1,1} \mathbf{a}_1 \\ (\mathbf{S} - \lambda_{1,1} \mathbf{I}) \mathbf{a}_1 &= \mathbf{0}. \end{aligned}$$

This is a homogeneous set of  $m$  equations in  $m$  unknowns and for a non-trivial solution to exist it is required that

$$|\mathbf{S} - \lambda_{1,1} \mathbf{I}| = 0.$$

Hence  $\lambda_{1,1}$  is an eigenvalue of  $\mathbf{S}$  and the solution  $\mathbf{a}_1$  is its corresponding eigenvector. However, there are  $m$  eigenvalues of  $\mathbf{S}$ , so it is to be determined which one is the required one. Multiplying  $\mathbf{S} \mathbf{a}_1 = \lambda_{1,1} \mathbf{a}_1$  by  $\mathbf{a}_1'$ , we get  $\mathbf{a}_1' \mathbf{S} \mathbf{a}_1 = \lambda_{1,1} \mathbf{a}_1' \mathbf{a}_1$ . Since it must hold that  $\mathbf{a}_1' \mathbf{a}_1 = 1$ , we get  $\lambda_{1,1} = \mathbf{a}_1' \mathbf{S} \mathbf{a}_1$ .

Furthermore,  $\lambda_{1,1}$  is the sample variance of  $\mathbf{y}_1$ , which will be shown later. Since we are aiming at maximizing the sample variance, then  $\lambda_{1,1}$  must be chosen to be the largest eigenvalue of  $\mathbf{S}$ . It follows that the coefficients  $\mathbf{a}_1$  of the 1st principal component,  $\mathbf{y}_1$ , are given by the elements of the eigenvector  $\mathbf{a}_1$  that correspond to the largest eigenvalue  $\lambda_{1,1}$  of  $\mathbf{S}$ .

Now, consider the 2nd principal component. Algebraically, the 2nd principal component  $\mathbf{y}_2$  is a linear combination given by  $a_{2,1}\mathbf{x}_1 + \cdots + a_{2,m}\mathbf{x}_m$ , where  $\mathbf{a}_2' = [a_{2,1} \cdots a_{2,m}]$  is the weight vector. The weight vector is chosen such that the variance of  $\mathbf{y}_2$  is maximized given the constraints that the sum of the squared weights is equal to 1, i.e.  $\sum_{i=1}^m a_{2,i}^2 = 1$  and the 2nd principal component is uncorrelated with the 1st principal component, i.e.  $Cov(\mathbf{y}_1, \mathbf{y}_2) = 0$ . The former constraint  $\mathbf{a}_2'\mathbf{a}_2 = 1$  is necessary for a unique solution to exist. The latter constraint may be also expressed in a different way. A line defining the 2nd principal component must be orthogonal to the line defining the 1st principal component, i.e.  $\mathbf{a}_2'\mathbf{a}_1 = \mathbf{a}_1'\mathbf{a}_2 = 0$ . The variance of  $\mathbf{y}_2$  is clearly  $\mathbf{a}_2'\mathbf{S}\mathbf{a}_2$ , so maximization of this variance will again involve the Lagrange multipliers. With 2 constraints we need 2 such multipliers, say  $\lambda_{2,1}$  and  $\lambda_{2,2}$  and we thus require to maximize

$$\begin{aligned} V_2 &= \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 - \lambda_{2,1}(\mathbf{a}_2'\mathbf{a}_2 - 1) - \lambda_{2,2}(\mathbf{a}_2'\mathbf{a}_1) \\ &= \sum_{i=1}^m \sum_{j=1}^m a_{2,i}a_{2,j}s_{i,j} - \lambda_{2,1} \left( \sum_{i=1}^m a_{2,i}^2 - 1 \right) - \lambda_{2,2} \left( \sum_{i=1}^m a_{1,i}a_{2,i} \right). \end{aligned}$$

Thus

$$\frac{\partial V_2}{\partial a_{2,k}} = 2 \sum_{j=1}^m s_{k,j}a_{2,j} - 2\lambda_{2,1}a_{2,k} - \lambda_{2,2}a_{1,k},$$

where  $k = 1, \dots, m$ .

To find the vector  $\mathbf{a}_2$  maximizing  $V_2$  we thus set  $\frac{\partial V_2}{\partial a_{2,k}} = 0$  for all  $k$  and solve the resulting set of simultaneous equations. Now

$$\frac{\partial V_2}{\partial a_{2,k}} = 0 \Rightarrow \sum_{j=1}^m s_{k,j}a_{2,j} = \lambda_{2,1}a_{2,k} - \frac{1}{2}\lambda_{2,2}a_{1,k}.$$

$\sum_{j=1}^m s_{k,j}a_{2,j}$  is the  $k$ th element of  $\mathbf{S}\mathbf{a}_2$ , while  $\lambda_{2,1}a_{2,k}$  is the  $k$ th element of  $\lambda_{2,1}\mathbf{a}_2$  and  $\lambda_{2,2}a_{1,k}$  is the  $k$ th element of  $\lambda_{2,2}\mathbf{a}_1$ . Thus, when all equations are considered simultaneously, it follows that the maximizing value of  $\mathbf{a}_2$  must



satisfy

$$\begin{aligned}\mathbf{S}\mathbf{a}_2 &= \lambda_{2,1}\mathbf{a}_2 + \frac{1}{2}\lambda_{2,2}\mathbf{a}_1 \\ (\mathbf{S} - \lambda_{2,1}\mathbf{I})\mathbf{a}_2 &= \frac{1}{2}\lambda_{2,2}\mathbf{a}_1.\end{aligned}$$

Multiplying  $(\mathbf{S} - \lambda_{2,1}\mathbf{I})\mathbf{a}_2 = \frac{1}{2}\lambda_{2,2}\mathbf{a}_1$  by  $\mathbf{a}_1'$ , and keeping in mind that  $\mathbf{a}_1'\mathbf{a}_1 = 1$ , while  $\mathbf{a}_1'\mathbf{a}_2 = 0$  yields  $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2 = \frac{1}{2}\lambda_{2,2}$ . However, multiplying  $\mathbf{S}\mathbf{a}_1 = \lambda_{1,1}\mathbf{a}_1$  by  $\mathbf{a}_2'$ , and keeping in mind that  $\mathbf{a}_2'\mathbf{a}_1 = 0$  yields  $\mathbf{a}_2'\mathbf{S}\mathbf{a}_1 = 0$ . Since  $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$  is a scalar and  $\mathbf{S}$  is a symmetric matrix, then  $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2 = \mathbf{a}_2'\mathbf{S}\mathbf{a}_1 = 0$ . Substituting this in  $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2 = \frac{1}{2}\lambda_{2,2}$  yields  $\lambda_{2,2} = 0$  and hence from  $(\mathbf{S} - \lambda_{2,1}\mathbf{I})\mathbf{a}_2 = \frac{1}{2}\lambda_{2,2}\mathbf{a}_1$  we get that  $\mathbf{a}_2$  also satisfies  $(\mathbf{S} - \lambda_{2,1}\mathbf{I})\mathbf{a}_2 = \mathbf{0}$ . Hence, we get again a homogeneous set of  $m$  equations in  $m$  unknowns and for a non-trivial solution to exist it is required that

$$|\mathbf{S} - \lambda_{2,1}\mathbf{I}| = 0.$$

Hence,  $\lambda_{2,1}$  is an eigenvalue of  $\mathbf{S}$  and the solution  $\mathbf{a}_2$  is its corresponding eigenvector. However, there are  $m$  eigenvalues of  $\mathbf{S}$ , so it is to be determined which one is the required one. Multiplying  $\mathbf{S}\mathbf{a}_2 = \lambda_{2,1}\mathbf{a}_2$  by  $\mathbf{a}_2'$ , we get  $\mathbf{a}_2'\mathbf{S}\mathbf{a}_2 = \lambda_{2,1}\mathbf{a}_2'\mathbf{a}_2$ . Since it must hold that  $\mathbf{a}_2'\mathbf{a}_2 = 1$ , we get  $\lambda_{2,1} = \mathbf{a}_2'\mathbf{S}\mathbf{a}_2$ . Furthermore,  $\lambda_{2,1}$  is the sample variance of  $\mathbf{y}_2$ , which will be shown later. Since we are aiming at maximizing the sample variance, then  $\lambda_{2,1}$  must be chosen to be the 2nd largest eigenvalue of  $\mathbf{S}$ . It follows that the coefficients  $\mathbf{a}_2$  of the 2nd principal component,  $\mathbf{y}_2$ , are given by the elements of the eigenvector  $\mathbf{a}_2$  that correspond to the 2nd largest eigenvalue  $\lambda_{2,1}$  of  $\mathbf{S}$ .

In case of the  $k$ th principal component, where  $k \leq m$ ,  $\mathbf{y}_k$  is of interest. It is a linear combination given by  $a_{k,1}\mathbf{x}_1 + \cdots + a_{k,m}\mathbf{x}_m$  which is orthogonal to all prior principal components and has maximum variance. Maximization of the variance subject to these constraints leads to the maximization of an expression with  $k$  Lagrange multipliers. Furthermore,  $\lambda_{k,1}$  is the sample variance of  $\mathbf{y}_k$ , which will be shown later. Using the same line of argument as for the 1st and the 2nd principal components, it follows the coefficients  $\mathbf{a}_k$  of the  $k$ th principal component,  $\mathbf{y}_k$ , are given by the elements of the eigenvector  $\mathbf{a}_k$  that correspond to the  $k$ th largest eigenvalue  $\lambda_{k,1}$  of  $\mathbf{S}$ .

For the 1st principal component we have maximized  $Var[y_1]$ . Note that

$$\begin{aligned} Var[\mathbf{y}_1] &= Var[a_{1,1}\mathbf{x}_1 + \cdots + a_{1,m}\mathbf{x}_m] \\ &= \mathbf{a}_1'Var[\mathbf{X}]\mathbf{a}_1 \\ &= \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 \\ &= \lambda_{1,1}. \end{aligned}$$

For the 2nd principal component we have maximized  $Var[y_2]$ . Note that

$$\begin{aligned} Var[\mathbf{y}_2] &= Var[a_{2,1}\mathbf{x}_1 + \cdots + a_{2,m}\mathbf{x}_m] \\ &= \mathbf{a}_2'Var[\mathbf{X}]\mathbf{a}_2 \\ &= \mathbf{a}_2'\mathbf{S}\mathbf{a}_2 \\ &= \lambda_{2,1}. \end{aligned}$$

For the  $k$ th principal component, where  $k \leq m$  we have maximized  $Var[y_k]$ . Note that

$$\begin{aligned} Var[\mathbf{y}_k] &= Var[a_{k,1}\mathbf{x}_1 + \cdots + a_{k,m}\mathbf{x}_m] \\ &= \mathbf{a}_k'Var[\mathbf{X}]\mathbf{a}_k \\ &= \mathbf{a}_k'\mathbf{S}\mathbf{a}_k \\ &= \lambda_{k,1}. \end{aligned}$$

Hence, the eigenvalues  $\lambda_{1,1}, \dots, \lambda_{m,1}$  are the sample variances of the principal components. Because the eigenvalues are variances of the principal components, we can speak about the proportion of variance explained by the first  $k$  principal components.

$$Proportion \ of \ variance = \frac{\lambda_{1,1}, \dots, \lambda_{k,1}}{\lambda_{1,1}, \dots, \lambda_{m,1}}.$$

Now, the total variance of the original set of  $m$  variables is simply the sum of the variances of each original variate  $\mathbf{x}_k$ , and this is the sum of the diagonal elements of the sample covariance matrix  $\mathbf{S}$ . This sum is mathematically denoted as  $tr(\mathbf{S})$ . Therefore, the expression of the proportion of variance can be rewritten as

$$Proportion \ of \ variance = \frac{\lambda_{1,1}, \dots, \lambda_{k,1}}{tr(\mathbf{S})}.$$

Principal component analysis can be considered as a rotation of the axes of the original coordinate system to the new coordinate system, such that the new axes coincide with directions of the greatest variability of observations.

The 1st principal component is a new coordinate axis which is oriented in the direction of greatest variability of observations. The 2nd principal component is another axis in the new coordinate system. It is oriented in the direction of the 2nd greatest variability of observations and is orthogonal to the 1st principal component. In general, the  $k$ th principal component is oriented in the direction of the  $k$ th greatest variability of observations and is orthogonal to all prior principal components.

### 1.3 Statistic for Detecting Anomalous Bids

The statistic which is the central point of this paper is given by

$$d_{1i}^2 = \sum_{j=m-q+1}^m \frac{y_{i,j}^2}{l_j^2},$$

where  $y_{i,j}$  is the value of the  $j$ th principal component for the  $i$ th observation and  $l_j^2$  is the variance of the  $j$ th principal component.

It should be noted that when  $q = m$ , the statistic  $d_{2i}^2$  becomes

$$d_{1i}^2 = \sum_{j=1}^m \frac{y_{i,j}^2}{l_j^2},$$

which is simply the squared Mahalanobis distance between the  $i$ th observation and the sample mean. The Mahalanobis distance takes into account the covariance between the variables. Thanks to this, the problems of scale and correlation inherent in the Euclidean distance are no longer an issue. It can be seen that if the variances of the variables are equal to 1 and the variables are uncorrelated, then the Mahalanobis distance reduces to the Euclidean distance. The squared Mahalanobis distance is defined as

$$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}),$$

where  $\mathbf{S}$  denotes the sample covariance matrix. This follows because  $\mathbf{S} = \mathbf{A}\mathbf{L}^2\mathbf{A}'$ , where  $\mathbf{L}^2$  is the diagonal matrix whose  $j$ th diagonal element is  $l_j^2$  and  $\mathbf{A}$  is the matrix whose  $(i, j)$ -th element is  $a_{i,j}$ . Furthermore,

$$\begin{aligned} \mathbf{S}^{-1} &= \mathbf{A}\mathbf{L}^{-2}\mathbf{A}' \\ \mathbf{x}_i' &= \mathbf{y}_i'\mathbf{A}' \\ \bar{\mathbf{x}}' &= \bar{\mathbf{y}}'\mathbf{A}'. \end{aligned}$$

Thus,

$$\begin{aligned} D_i^2 &= (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \\ &= (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{A}'\mathbf{A}\mathbf{L}^{-2}\mathbf{A}'\mathbf{A} (\mathbf{y}_i - \bar{\mathbf{y}}) \\ &= (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{L}^{-2} (\mathbf{y}_i - \bar{\mathbf{y}}) \\ &= \sum_{j=1}^m \frac{y_{i,j}^2}{l_j^2}. \end{aligned}$$

Assuming that principal components are independent and normally distributed implies that  $d_{1i}^2$  follows the non-central  $\chi^2$  distribution with  $q$  degrees of freedom. An outline of the derivation of the non-central  $\chi^2$  distribution is to be found in Appendix A.4. An informal graphical technique, which may reveal outliers is a non-central  $\chi^2$  probability plot of the  $d_{1i}^2$ .

## 1.4 Example

From the data set consisting of 360 tenders, I chose 10 tenders with the highest number of submitted bids. Next, I find the principal components and I test both principal components for normality with the Anderson-Darling and Shapiro-Wilk tests. The Anderson-Darling test is described in the Appendix A.2 and Shapiro-Wilk test is described in the Appendix A.3. The results of normality tests are displayed in the table below.

n <sup>1</sup>	1st principal component		2nd principal component	
	Shapiro-Wilk	Anderson-Darling	Shapiro-Wilk	Anderson-Darling
29	0.1896	0.1573	0.1542	0.0575
38	0.1186	0.2891	0.1670	0.2470
33	0.7564	0.6081	0.5696	0.1722
31	0.5993	0.3604	0.2888	0.4753
30	0.7823	0.1778	0.0804	0.0319
21	0.4469	0.3375	0.1628	0.8888
17	0.4114	0.6950	0.1876	0.0512
38	0.3145	0.0937	0.0560	0.0128
16	0.2454	0.1106	0.5077	0.3100
18	0.8030	0.2942	0.3115	0.4020

Table 1: P-values

I set the significance level at 0.05. As regards the 1st principal component, I fail to reject the null hypothesis for all tenders. As regards the 2nd principal component, I fail to reject the null hypothesis for 8 tenders. Failing to reject the null hypothesis means that there is not enough evidence to prove that the data are not normally distributed. In practice, one says that it proves normality of the data. The other thing is that according to the results presented in the table above, the normality assumption doesn't seem implausible. There are 8 out of 10 tenders such that for both principal components I failed to reject the null hypothesis using Anderson-Darling and Shapiro-Wilk tests.

From a mathematical perspective it is very important how the quality is measured. Usually, this is done using weighted multi criteria analysis. Criteria are listed, prioritized and shown to suppliers. Each supplier can get a score between 0 and 1. Frequently, widely accepted industry criteria such as server uptime or delivery accuracy are utilized. However, the scale of measurement

---

<sup>1</sup>Number of bids that have been submitted in a given tender.

of different alternatives against these criteria is often different. Also, other less standard criteria are in use such as the buyer's score based on an evaluation of an implementation plan submitted by a supplier. Different criteria and different scales of measurement are 2 main reasons why quality comparisons from one tender to another tender are difficult if not impossible. What is quality and how it is measured has enormous influence on the outcome of the tender and thus on which bids are considered anomalous.

As I argued, normality is a reasonable assumption when it comes to detecting anomalous bids in public procurement. It holds for both principal components in 8 out of 10 tenders.

To show how the statistic  $d_{1i}^2$  works, I took 1 tender with 33 submitted bids. The results of Anderson-Darling and Shapiro-Wilk tests are shown in the 3rd row of Table 1. The scatter plot of the data is to be found in Appendix A.5

Setting  $q = m = 2$  and I calculate the statistic  $d_{1i}^2$ . I found the parameters of the non-central  $\chi^2$  distribution using the least squares, i.e. choosing such a non-centrality parameter  $\sum_{i=1}^m \left(\frac{\mu_i}{\sigma_i}\right)^2$  that the distance between the empirical cumulative distribution function and the theoretical cumulative distribution function is minimized.

Non-central  $\chi^2$  probability plot is to be found in Appendix A.6. Bids 1, 2, 15, 29, 32, 32 can be considered anomalous bids. Bid 1 is the 2nd highest price and the highest quality. Bid 2 is the 2nd highest quality. Bid 15 is the 2nd lowest price. Bid 29 is the lowest price. Bid 32 is the lowest quality. Finally, bid 33 is the highest price and the lowest quality. Yet, looking at the scatter of the data, one may wonder why bid 14 has not been considered anomalous.

## **1.5 Conclusion**

I have addressed a problem of detecting anomalous bids in public procurement. I believe that the described procedure may find its application in the area of anomalous bids detection by developing a general framework for determining whether a given bid is anomalous or not. Future research should include a study on a larger sample of tenders. It would be also interesting to see other methods of detecting anomalous bids in public procurement.



## 2 Influence of Learning on the Bidding Performance

I study bidding performance of 6 companies in 5 tenders organized by 1 tendering entity on a quarterly basis. The data set consists of 73 observations and the time span is 17 Apr 2009 - 21 Apr 2011. The data are displayed in Appendix A.7. The question here is whether bidding performance is influenced by learning.

I will use random effects logistic regression to investigate this question. The explained variable is a binary variable that takes value 1 if a company has been awarded the contract in a given tender and 0 otherwise. The explanatory variables are time, tender and the submitted price. Time is an important explanatory variable because if bidding performance was influenced by learning, the bidding performance would depend on time. Clearly, scarcity of data is a major disadvantage of the model. Unfortunately, data availability is a common problem in procurement. For example, I believe the model would be much better if I could include firm size and return on assets - which is the ratio of net income to total assets or operating ratio, which is the ratio of cost of goods sold plus operating expenses to net sales - as explanatory variables.

I estimated the random effects logit model specified as  $y_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + u_i + \epsilon_{i,t}$  under the following assumptions:

1.  $P(y_{i,t} = 1|u_i) = F(\mathbf{x}'_{i,t}\boldsymbol{\beta} + u_i)$ , where  $F$  is the standard logistic distribution
2. error term,  $\epsilon_{i,t}$ , is independent and identically logistically distributed with mean 0 and variance  $\frac{\pi^2}{3}$
3. random effect,  $u_i$ , is independent and identically normally distributed with mean 0 and constant variance
4. no perfect linear relationships among the explanatory variables
5. random sample from the cross section

The results of the estimation are displayed below.

```

. xtlogit outcome tender price time, re
Fitting comparison model:
Iteration 0:  log likelihood = -50.428377
Iteration 1:  log likelihood = -49.816661
Iteration 2:  log likelihood = -49.816447
Iteration 3:  log likelihood = -49.816447
Fitting full model:
tau = 0.0    log likelihood = -49.816447
tau = 0.1    log likelihood = -48.894012
tau = 0.2    log likelihood = -48.388599
tau = 0.3    log likelihood = -48.113286
tau = 0.4    log likelihood = -48.006272
tau = 0.5    log likelihood = -48.071294
Iteration 0:  log likelihood = -47.994809
Iteration 1:  log likelihood = -45.75308
Iteration 2:  log likelihood = -45.354558
Iteration 3:  log likelihood = -45.273968
Iteration 4:  log likelihood = -45.273794
Iteration 5:  log likelihood = -45.273794
Random-effects logistic regression      Number of obs      =      73
Group variable: company                 Number of groups   =       6
Random effects u_i ~ Gaussian           Obs per group: min =       5
                                           avg =      12.2
                                           max =       29
                                           Wald chi2(3)      =       4.87
                                           Prob > chi2       =      0.1817
Log likelihood = -45.273794

```

outcome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
tender	.9684158	.4732938	2.05	0.041	.0407769 1.896055
price	-8.40e-06	4.67e-06	-1.80	0.072	-.0000175 7.56e-07
time	.0014505	.0306764	0.05	0.962	-.0586743 .0615752
_cons	-2.276475	1.794863	-1.27	0.205	-5.794342 1.241392
/lnsig2u	1.183552	.9420311			-.6627954 3.029899
sigma_u	1.807195	.8512169			.7179196 4.549191
rho	.4981761	.2355046			.1354458 .8628365

```

Likelihood-ratio test of rho=0: chibar2(01) =      9.09 Prob >= chibar2 = 0.001

```

I prefer logit to probit because for the former is possible to obtain a  $\sqrt{N}$ -consistent estimator of  $\beta$  without any assumptions about how  $u_i$  is related to  $\mathbf{x}_i$ . I decided to use the random effects instead of the fixed effects because I wanted to retain observations that lack time-series variation in the dependent variable. I decided to use the random effects logistic regression and not the ordinary logistic regression because in the likelihood ratio test of rho at the level of significance of 5%, I rejected the null hypothesis that rho is equal to 0. If rho was equal to 0, then there would be no variation in the  $u_i$  across companies, which would mean that there would be no need to control for company-specific effects. Variable tender is significant at the level of significance of 5%, which suggests that there is some tender specific

effect. Variable price is significant at the level of significance of 10% and it is insignificant at the level of significance of 5%. Variable time is insignificant at the level of significance of 5%. This means that there is no evidence of bidding performance being influenced by learning because it has been assumed that if bidding performance was influenced by learning, then the bidding performance would depend on time. The details of the significance tests are given in the Appendix A.8.

Exponentiating coefficients in the column Coef. gives the so called odds ratio. For example, if I assumed the level of significance of 10%, then an increase of 1 Euro in variable price would result in, *ceteris paribus*, a change of times  $\exp(-0.0000084) = 0.9999916$  in the odds ratio. In terms of percent change, one could say that, *ceteris paribus*, the odds for 1 Euro more expansive bid were 0.00084% lower than the odds for 1 Euro less expansive bid.

The Stata output presented below shows marginal effects.

```
. mfx compute, predict(pu0)
Marginal effects after xtlogit
      y = Pr(outcome=1 assuming u_i=0) (predict, pu0)
      = .30400531
```

variable	dy/dx	Std. Err.	z	P> z	[	95% C.I.	]	X
tender	.2049033	.10262	2.00	0.046	.003766	.406041		2.80822
price	-1.78e-06	.00000	-1.54	0.124	-4.0e-06	4.9e-07		154530
time	.0003069	.00648	0.05	0.962	-.012393	.013007		17.9589

For example, if I assumed the level of significance of 10%, then increasing the price of a given bid by 1 Euro would decrease, *ceteris paribus*, its probability of winning the tender by  $-0.000178\%$ . This relation is what one would have expected from an economic point of view.

To evaluate the goodness-of-fit of the model I use *McFadden's Pseudo R<sup>2</sup>*, which is described in Appendix A.9. It is equal to 0.0582513. Low value of *McFadden's Pseudo R<sup>2</sup>* can be explained by the scarcity of data.

## 2.1 Conclusion

I estimated the random effects logit model. It has been assumed that if bidding performance was influenced by learning, then the bidding performance would depend on time. The explained variable was a binary variable that takes value 1 if a company has been awarded the contract in a given tender and 0 otherwise. I used 3 explanatory variables. Variable tender was significant at the level of significance of 5%, which suggests that there is some tender specific effect. Variable price was significant at the level of significance of 10% and insignificant at the level of significance of 5%. I found that at the level of significance of 10% there is a negative relation between submitted price and probability of winning the tender. Variable time was insignificant at the level of significance of 5% which means that there is no evidence of bidding performance being influenced by learning. Since this study has been done using a data set containing a limited number of explanatory variables, further research should be done on a more extensive data set containing firm size, return on assets or operating ratio.

# A Appendix

## A.1 Kolmogorov-Smirnov Test

Suppose we wish to infer whether a sample comes from a certain specified distribution  $F_0(x)$ . This testing problem is well known as the simple goodness-of-fit problem. The null hypothesis is that the two samples come from a common distribution and the alternative hypothesis is that they don't. More formally,

$$\begin{aligned}H_0 : & \quad F(x) = F_0(x). \\H_1 : & \quad F(x) \neq F_0(x).\end{aligned}$$

The Kolmogorov-Smirnov test quantifies the distance between the empirical cumulative distribution, denoted by  $F(x)$  and the theoretical cumulative distribution, denoted by  $F_0(x)$ . The test statistic is given by

$$D = \sup |F(x) - F_0(x)|,$$

where  $\sup$  denotes the supremum of the set of distances. By the Glivenko-Cantelli theorem, if the sample comes from distribution  $F_0(x)$ , then  $D$  converges to 0 almost surely and hence in probability.

## A.2 Anderson-Darling Test

The Anderson-Darling test is an improvement upon the Kolmogorov-Smirnov test, which is only a historical curiosity and should never be used because of its poor power. The Anderson-Darling test is used to verify whether a sample comes from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov test that gives more attention to the tails. The Kolmogorov-Smirnov test is distribution free, meaning that the critical values do not depend on the specific distribution being tested, whereas the Anderson-Darling test makes use of the specific distribution in calculating the critical values. The null hypothesis is that the two samples come from a common distribution and the alternative hypothesis is that they don't. More formally,

$$\begin{aligned}H_0 : & \quad F(x) = F_0(x). \\H_1 : & \quad F(x) \neq F_0(x).\end{aligned}$$

The statistics for testing the null hypothesis versus the alternative hypothesis is  $A^2 = n^2 - S$ , where  $n$  is the sample size and

$$S = \sum_{k=1}^n \frac{2k-1}{n} [\ln F(Y_k) + \ln(1 - F(Y_{n+1-k}))].$$

Tabulated values and formulas have been published for the normal distribution.

### A.3 Shapiro-Wilk Normality Test

The Shapiro-Wilk test is another useful alternative to the Kolmogorov-Smirnov test and it is considered a better test than the Kolmogorov-Smirnov test, especially when the sample size is small. The Shapiro-Wilk test is relatively powerful when testing for departures from normality. It is based on the observed distance between symmetrically positioned observations. However, it requires a considerable amount of computational effort. The null hypothesis is that the sample is from a normal distribution with unknown mean and variance and the alternative hypothesis is that it doesn't. More formally,

$$\begin{aligned} H_0 : & \quad F(x) = N_{\mu,\sigma}(x) \\ H_1 : & \quad F(x) \neq N_{\mu,\sigma}(x). \end{aligned}$$

The Shapiro-Wilk statistic is given by

$$W = \frac{(\sum_{i=1}^n c_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

where  $x_{(i)}$  is the  $i$ th order statistic, i.e. the  $i$ th smallest number in the sample and  $\bar{x}$  is the sample mean.

The constants  $c_i$  are given by

$$[c_1, \dots, c_n] = \frac{\mathbf{b}'\mathbf{V}^{-1}}{(\mathbf{b}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{b})^{1/2}},$$

where  $\mathbf{b} = [b_1, \dots, b_n]'$  and  $b_1, \dots, b_n$  are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and  $\mathbf{V}$  is the covariance matrix of those order statistics.

## A.4 Non-central $\chi^2$ Distribution

Let  $X_i$  be  $m$  independent, normally distributed random variables with means  $\mu_i$  and variances  $\sigma_i^2$ . In brief, the distribution of  $Y = \sum_{i=1}^m \left(\frac{X_i}{\sigma_i}\right)^2$  may be derived as follows. For  $m = 1$  the cumulative distribution function of  $Y$  is

$$F_Y(y) = P(Y \leq y) = \dots = F_X(\sqrt{y}\sigma_1) - F_X(-\sqrt{y}\sigma_1).$$

The derivative of the cumulative distribution function of  $Y$  is

$$\frac{\sigma_1}{\sqrt{y}} P_X(\sqrt{y}\sigma_1) = \frac{1}{\sqrt{2\pi y}} \exp\left(\frac{-(\sqrt{y}\sigma_1 - \mu_1)^2}{2\sigma_1^2}\right).$$

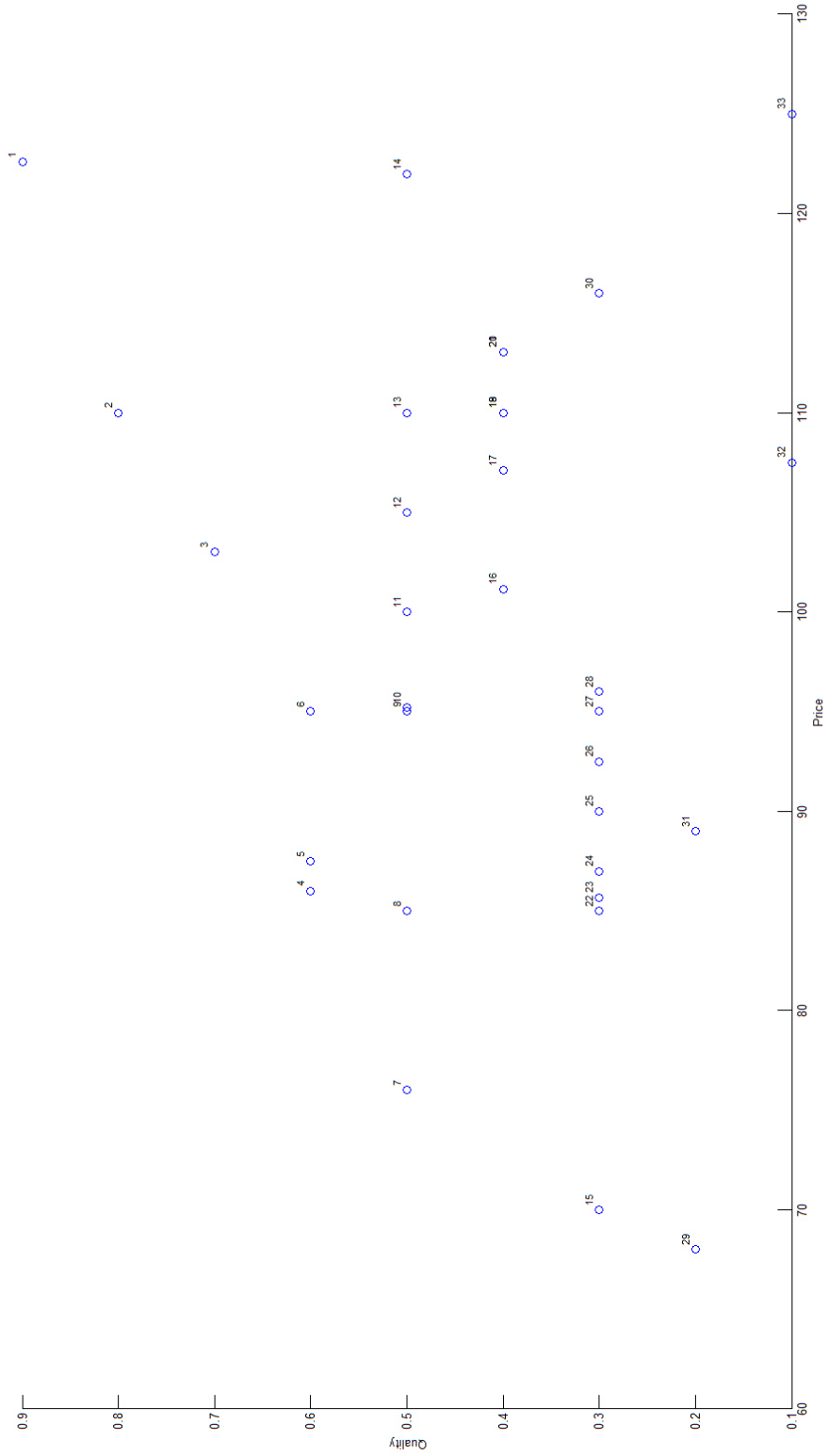
The characteristic function of  $Y$  is

$$\varphi_Y(t) = E[\exp(itY)] = \dots = \frac{1}{\sqrt{1-2it}} \exp\left(\frac{it\mu_1^2}{(1-2it)\sigma_1^2}\right).$$

For  $m \geq 2$  the characteristic function of the non-central chi-square distribution with  $m$  degrees of freedom is

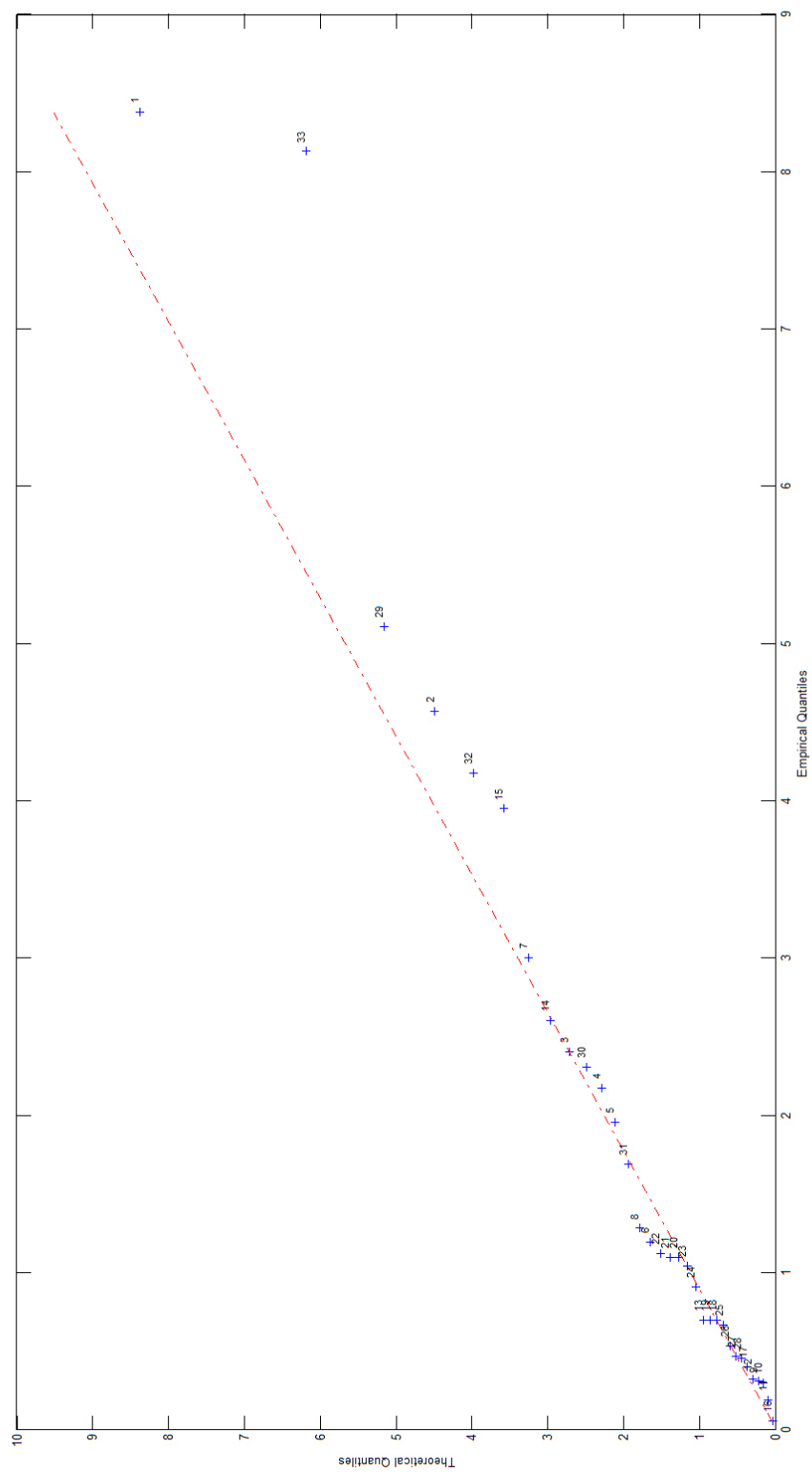
$$\varphi_Y(t) = E\left[\exp\left(\sum_{i=1}^m \left(\frac{X_i}{\sigma_i}\right)^2\right)\right] = \dots = (1-2it)^{-\frac{m}{2}} \prod_{i=1}^m \exp\left(\frac{it\mu_i^2}{(1-2it)\sigma_i^2}\right).$$

## A.5 Scatter Plot



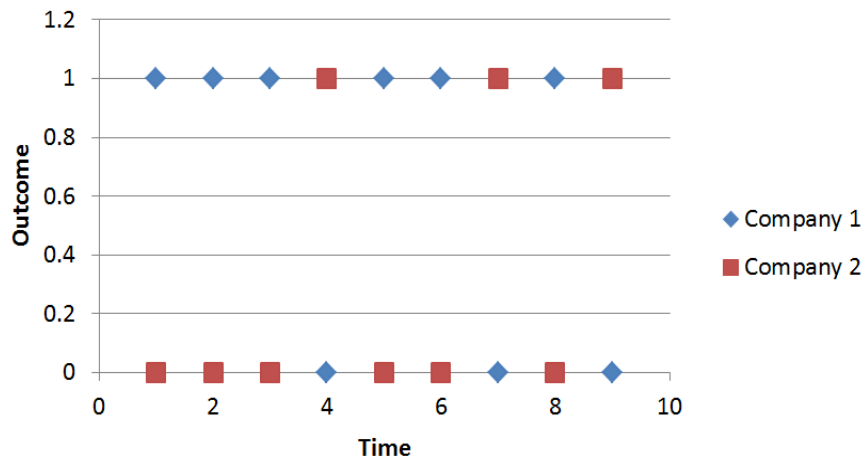


## A.6 Non-Central $\chi^2$ Probability Plot

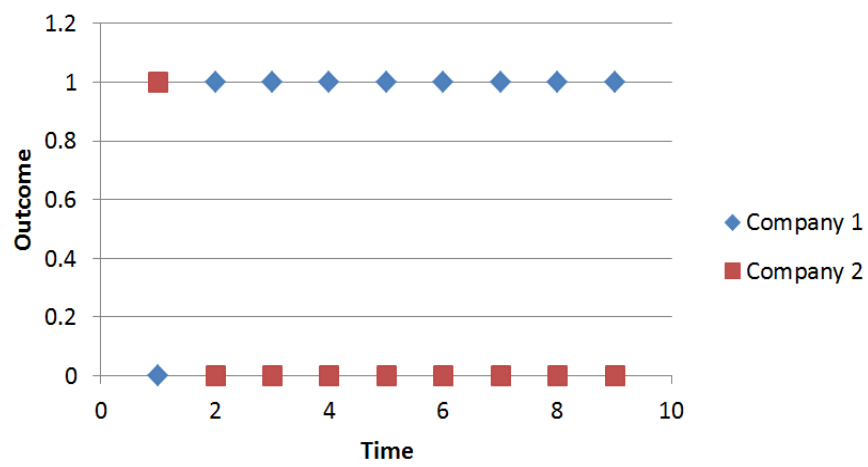


## A.7 Tenders Visualized

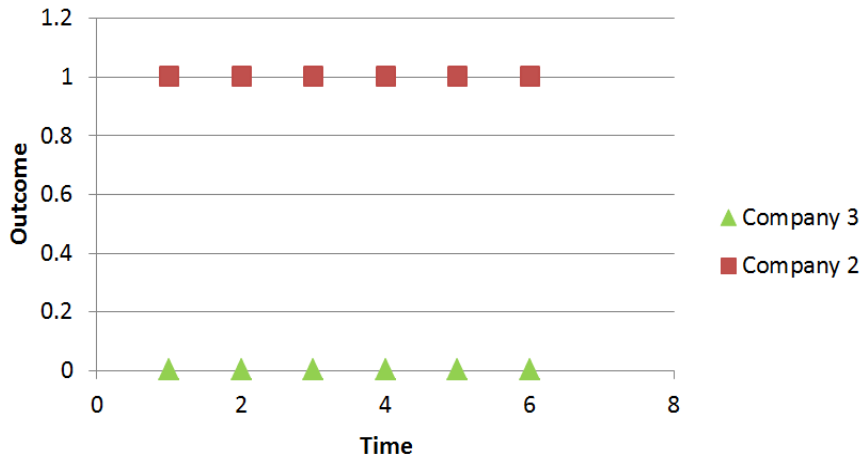
### Tender 1



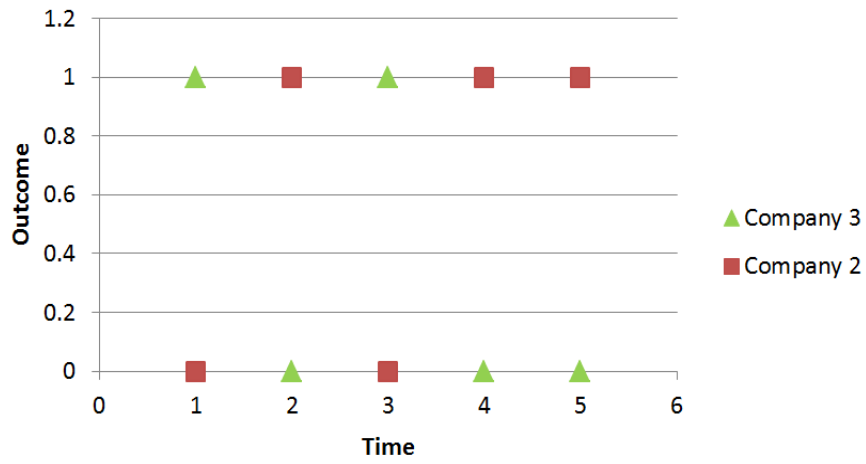
### Tender 2

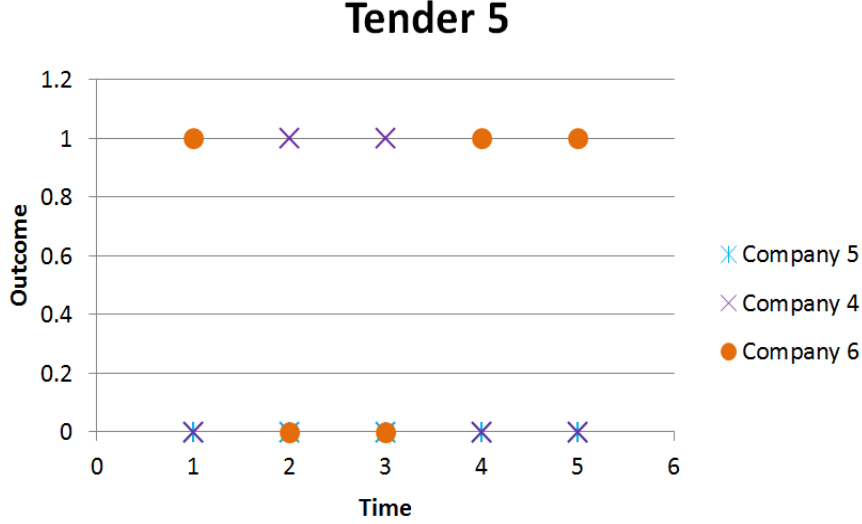


### Tender 3



### Tender 4





## A.8 Significance Tests

According to [13], little is known about the small sample properties of logistic regression coefficients, and therefore tests of significance for samples less than 100 prove risky. Hence, I use the squared ratio of the coefficient to the standard error to test whether a given variable is significant or not. This procedure is known as the Wald test. The critical value of  $\chi^2$  distribution with 1 degree of freedom is 3.8415 at the significance level of 5% and 2.7055 at the significance level of 10%. Variable tender is significant at the level of significance of 5%. Variable price is significant at the level of significance of 10% and insignificant at the level of significance of 5%. Variable time is insignificant at the level of significance of 5%. I don't test for joint significance with the likelihood ratio test because of insignificant variables that are present in the model.

## A.9 McFadden's Pseudo $R^2$

$$McFadden's\ Pseudo\ R^2 = 1 - \frac{LL_F}{LL_R}, \quad (1)$$

where  $LL_F$  is the log-likelihood of the full model and  $LL_R$  is the log-likelihood of the reduced model that contains only an intercept. Values between 0.2 and 0.4 are considered highly satisfactory.

The results of the estimation of the reduced model are displayed below.

```

. xtlogit outcome, re
Fitting comparison model:
Iteration 0:  log likelihood = -50.428377
Iteration 1:  log likelihood = -50.428377
Fitting full model:
tau = 0.0    log likelihood = -50.428377
tau = 0.1    log likelihood = -49.350874
tau = 0.2    log likelihood = -48.747869
tau = 0.3    log likelihood = -48.403518
tau = 0.4    log likelihood = -48.248936
tau = 0.5    log likelihood = -48.289507
Iteration 0:  log likelihood = -48.224087
Iteration 1:  log likelihood = -48.075015
Iteration 2:  log likelihood = -48.07418
Iteration 3:  log likelihood = -48.07418
Random-effects logistic regression
Group variable: company
Random effects u_i ~ Gaussian
Number of obs      =      73
Number of groups   =       6
Obs per group: min =       5
                  avg =     12.2
                  max =      29
Wald chi2(0)      =       .
Prob > chi2       =       .
Log likelihood    = -48.07418

```

outcome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	-.3747146	.5155439	-0.73	0.467	-1.385162	.6357329
/lnsig2u	-.0696916	1.051132			-2.129873	1.99049
sigma_u	.9657543	.5075678			.3447497	2.705387
rho	.2208811	.1808921			.0348671	.689898

```

Likelihood-ratio test of rho=0: chibar2(01) =      4.71 Prob >= chibar2 = 0.015

```

## References

- [1] Abdi H. and L.J. Williams (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2:433-459.
- [2] Armitage P. and T. Colton (eds) (2005) Encyclopedia of biostatistics. Wiley-Blackwell.
- [3] Conti P.L. and M. Naldi (2008) Detection of anomalous bids in procurement auctions. Decision Support Systems, 46(1):420-428.
- [4] Dunteman G.H. (1989) Principal components analysis. SAGE Publications.
- [5] Flury B. and H. Riedwyl (1988) Multivariate statistics: A practical approach. Chapman and Hall.
- [6] Gnanadesikan R. and J.R. Kettenring (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81-124.
- [7] Jackson E.J. (1991) A user's guide to principal components. John Wiley & Sons.
- [8] Johnson R.A. and D.W. Wichern (2002) Applied multivariate statistical analysis. Pearson Education.
- [9] Jolliffe I.T. (2002) Principal component analysis. Springer.
- [10] Krzanowski W.J. (1988) Principles of multivariate analysis: A user's perspective. OUP.
- [11] Mardia K.V., J.T. Kent and J.M. Bibby (1979) Multivariate analysis. Academic Press.
- [12] McNolty F. (1962) A contour-integral derivation of the non-central chi-square distribution. The Annals of Mathematical Statistics, 33(2):796-800.
- [13] Pampel F.C. (2000) Logistic regression: A primer. SAGE Publications.
- [14] Rencher A.C. (2002) Methods of multivariate analysis. John Wiley & Sons.
- [15] Sharma S. (1996) Applied multivariate techniques. John Wiley & Sons.

- [16] Stuart M. (1982) A geometric approach to principal components analysis. *The American Statistician* 36:365-367.
- [17] Wooldridge J.M. (2002) *Econometric analysis of cross section and panel data*. MIT Press.