Autumn 2004 Business 41100-01
Applied Regression Analysis (Professor Hansen)
Problem Set #3: Data Collection and Analysis

Hatsuru Morita (Law School; Audit)
hmorita@law.uchicago.edu

Theme: Has the increase in human resource of the SESC improved its function?

1. Introduction

The SESC, Securities and Exchange Surveillance Commission, is the Japanese version of the SEC, Securities and Exchange Commission. In Japan, it was said that the department which was in charge of monitoring the securities market had too small resource to function effectively and that the weak body of the monitoring department had caused corporate scandals around 1990, after the collapse of the bubble economy, and "unfair" securities market which was considered to be a breeding place for fraudulent transactions such as insider tradings and market manipulations. So the Japanese government decided to establish the SESC in 1992 in order to achieve "free, fair, and transparent securities market". Because the SESC had only 84 staffs in its first year, public opinions, such as newspapers and TV media, insisted that the SESC had too few staffs to monitor the securities market effectively, compared to its US counterpart, the SEC. Facing these criticisms, the Japanese government decided to reinforce the SESC through increasing its staffs.

The purpose of this research is to analyze whether the above attempt of the Japanese government for this decade has been successful or not. Has the increase in number of staffs of the SESC improved its function?

2. The data

Relevant data are available at the SESC website[1]. These data are from 1992 to 2003 and are collected on annual-base, so they can be considered to be a random sample data. Although the data range is just twelve years, it is fair to say that the time is mature to evaluate the Japanese government's policy for this decade.

But the data is not free of flaw. The order of the dependent variable is not so large --- the maximum is 10 and the minimum is 1 --- that it is expected that the dependent variable fluctuates rather violently.

Let us turn to the detail of this analysis. The independent variable is the number of staffs of the SESC; the dependent variable is the number of filing of complaints per year. Although we can employ the number of inspections per year or the number of market surveillances per year as the dependent variable and these variables have much larger scale than the number of filing of complaints --- for example, the minimum and the maximum of the number of inspections is 80 and 118; as to surveillance, 203 and 392 ---, this research does not adopt such approaches.

The reason is that activities such as inspections or market surveillance are just 'process' of the SESC function, not its 'outcome'. The SESC could pretend to be working hard or effectively by making its process active, but it is not so easy for the SESC to camouflage the outcome. The number of filing of complaints is considered to be the less manipulable variable and more appropriate for this research.

The underlying data-generating-process is as follows: if the reinforcement of workforce of the SESC has been effective, the activity level, which is represented by the number of filing of complaints [the dependent variable: "Filing"], would increase as the number of staffs of the SESC [the independent variable: "Number"] increased.

A copy of the data set is saved in the attached floppy disk [SESC_new.MTP]. The descriptive statistics of the variables are as follows:

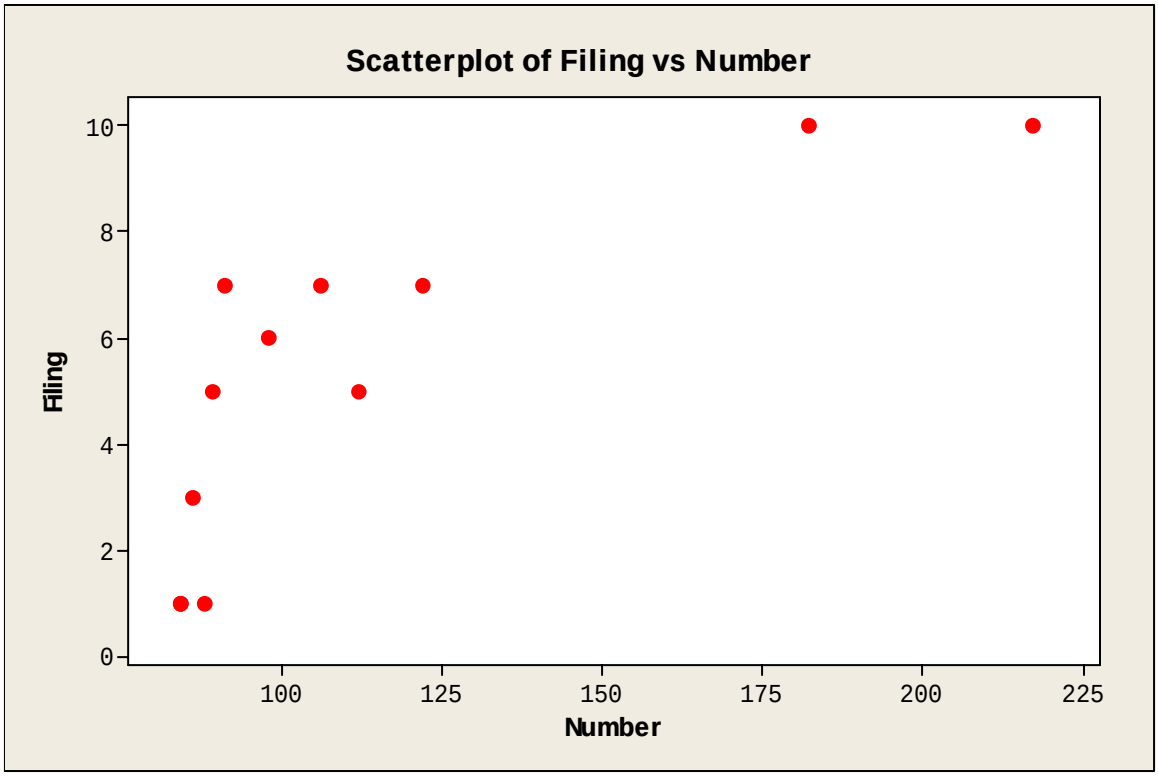
```
Variable  N   Mean  SE Mean  StDev  Minimum   Median   Maximum
```

[1] For the English version, http://www.fsa.go.jp/sesc/english/actions/actions.htm; for the Japanese version, http://www.fsa.go.jp/sesc/actions/actions.htm; the Japanese version provides more comprehensive data set. This research is based on the latter.

```
Number    12 113.3     12.3   42.7    84.0     94.5     217.0

Variable  N    Mean  SE Mean   StDev  Minimum   Median   Maximum
Filing    12  5.250    0.930   3.223    1.000    5.500    10.000
```

The following is the graph by plotting the data.



It looks like that there is a positive relationship between the dependent variable and the independent variable. However, the relationship does not seem to be a linear one, but rather a logarithmic one: as the number of staffs increases, the number of prosecution also increases, but the rate of increase steps down as the number of staffs increases.

3. Regression

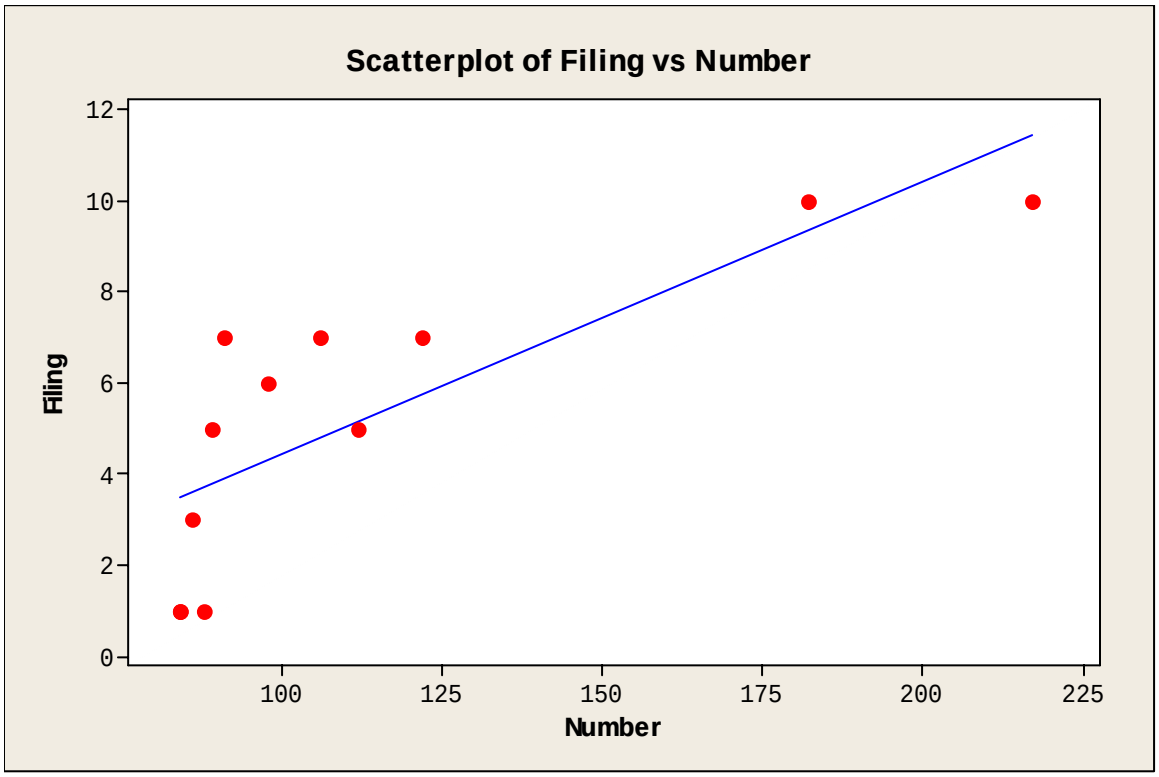
When we run least-squares regression on the data, we get the following output and graph.

```
The regression equation is
```

```
Filing = - 1.51 + 0.0597 Number

Predictor     Coef   SE Coef      T      P
Constant    -1.509     1.762  -0.86  0.412
Number     0.05968   0.01464   4.08  0.002
S = 2.07171   R-Sq = 62.4%

Analysis of Variance
Source          DF       SS      MS     F      P
Regression       1   71.330  71.330  16.62  0.002
Residual Error  10   42.920   4.292
Total           11  114.250
```



4. Discussion

4.1 Questions and hypothesis testing

First, we can consider whether Number affects Filing in a statistically significant

way.

The null hypothesis is that Number does not affect Filing, or the slope coefficient is zero. According to the above regression, the slope coefficient is 0.05968, which means that Filing increases by 0.05968 when we increase Number by 1. The t-value of 0.005 (1% level) and 10 degree of freedom (number of observance 12 minus 2) is 3.17. Because the t-statistic 4.08 is larger than the t-value 3.17, we can reject the null hypothesis at the 1% level.

We can reach the same conclusion by forming confidence interval. Because the estimated mean of coefficient is 0.0597 and the estimated standard error of the coefficient is 0.0146, we can form the 99% confidence interval for the slope coefficient as (0.0597-0.0146*3.17, 0.0597+0.0146*3.17), that is (0.0134, 0.106). Because the coefficient under the null hypothesis, zero, is out of this 99% confidence interval, we can reject the null hypothesis at 99% level.

We can also reach the same conclusion by observing the p-value 0.002, which is smaller than 0.005.

Another question regarding the regression is that of causation: Why is there a positive relationship between the number of filing of complaints and the number of staffs of the SESC? A simple answer is that the larger number of staffs of the SESC enhances the monitoring activity of the SESC and increases the detection rate of securities crimes.

Finally, we can also pose a question about effectiveness of increase in the number of staffs of the SESC. Forming the 95% confident level for the slope coefficient, (0.0271, 0.0923), we can predict that the filing will increase between 0.0271 and 0.0923 per year, plus the residual error, if the Japanese government increases one staff of the SESC. We cannot judge whether this effect is sufficient or not, but we can run a cost-benefit analysis on this prediction, if additional information, such as the cost of each staff, is provided.

4.2 Effectiveness of the SLR and additional factors

The above regression output tells us that the simple linear regression model is not so appropriate. Because the R-square is 0.624, Number can explain only 62.4% of Filing; other 37.6% cannot be explained by Number.

Probably this not-so-strong fit is caused by the above mentioned intuition: the model is not linear, but logarithmic. Also, we can imagine that some of the staffs of the

SESC do not engage in investigating or monitoring work but in back-office work, such as office administrating and human resources. Taking these factors into account, we can form another revised model.

Let us assume that half of the starting SESC staff in 1992, that is, 42 staffs are non-monitoring staffs. For ease of calculation, we assume that this number is constant as the total number of the SESC staffs increases. Then take logarithm of (Number - 42) and denote it as "LN42". The outcome of regression of Filing over LN42 is as follows.

```
The regression equation is
Filing = - 18.6 + 5.75 LN42

Predictor     Coef   SE Coef      T      P
Constant   -18.595     4.791   -3.88  0.003
LN42         5.751     1.149    5.01  0.001
S = 1.80489   R-Sq = 71.5%

Analysis of Variance
Source           DF       SS      MS      F     P
Regression        1   81.674  81.674  25.07 0.001
Residual Error   10   32.576   3.258
Total            11  114.250
```

This is also statistically significant at 99% level, but the difference we are focusing is that the R-square improves to 0.715, or about by 10% from the original linear model. This improvement may be interpreted as implicating that part of the members of the SESC does not directly affect activity level of the SESC or that marginal increase of staffs tends to decrease as the total number of staffs increase. Both possibilities are not favorable evaluation for the SESC.

Of course, we can think about other factors which will disturb the regression outcome. First, the number of securities crimes might be affected by the economic climate at the time. For example, when the economic climate turns downward, more people might be inclined to commit securities crime to earn more money. This is one explanation and the reverse causation is also possible: when the economic climate turns upward,

there are more chances to make money by committing securities crimes. Second, the number of securities crimes might decrease because the SESC was strengthened. Potential securities criminals might refrain from committing securities crimes because there is an increased possibility that the strengthened SESC would discover their crimes. Finally, the effect of increase in number of staffs of the SESC might affect activity level of the SESC not in the same year but in the succeeding years. If this conjecture was true, we might be able to improve our analysis by moving the dependent variable one or more years later.

4.3 [Un-]Usefulness of the regression result

As discussed above, the regression result could be used to evaluate the past Japanese government policy: whether its policy has been effective or not. In addition, we could utilize the result to evaluate whether we should further increase the number of staffs of the SESC in the future, referring prediction intervals from the result.

But we must be careful. When we extend the prediction out of the data range, the residual error becomes large and the prediction is not much reliable. For example, when we predict Filing at Number=300, MINITAB says:

```
Obs     Fit   SE Fit       95% CI           95% PI
  1  16.395    2.798  (10.160, 22.630)  (8.637, 24.153)XX
XX denotes a point that is an extreme outlier in the predictors.
```

This means that the regression might not be useful to form and evaluate the policy of the next year. Probably we could generalize this intuition: regression analysis could be useful for evaluating past performance of a specific policy, but might not be so useful for predicting future effectiveness of that policy, because future predictions often require predicting what we have not observed yet.