

Lab 18

Millions Missing from Meters

In this lab we'll discover how statistical reasoning could be used to settle a particular legal dispute between the City of New York and a contractor. Statistics and probability have a long history of application to legal issues. In fact, some of the earliest applications of probability were to resolve legal issues.

About the Data

In May 1978, Brink's Inc. was awarded a contract to collect coins from some 70,000 parking meters in New York City for delivery to the City Department of Finance. Sometime later, the City became suspicious that not all of the money collected was being returned to the city. In April of 1978, five Brink's collectors were arrested and charged with grand larceny and were subsequently convicted. The city sued Brink's for negligent supervision of its employees, seeking to recover the amount stolen. Though the fact of theft had been established, a reasonable estimate of the amount stolen was needed for the lawsuit. This data was provided by the City's attorney. During some of the months in which the thefts occurred, the city had selected 47 parking meters near city hall and had emptied these itself. The other meters were emptied by various private contractors, including Brinks. The City could now compare the returns from their 47 meters with those from the

private contractors. The *priv* variable lists the amount of money collected from privately hired contractors each month. The *brinks* variable is an indicator variable indicating whether Brinks Inc. was the hired contractor for the month of collection. The *govt* variable lists the monthly returns from the 47 meters the city emptied.

Today's Lab

Your goal for this lab is to provide an estimate of the total amount stolen by Brinks, Inc. employees. Begin by answering the following question:

Question 1: Describe how you think you can use these data to estimate the total amount stolen.

Now, enter the data into *Stata*:

```
. use http://www.stat.ucla.edu/labs/datasets/metertheft.dta
```

Exploring and Examining the Data

The data you just loaded has the name label “Parking Meter Theft.” Using the **describe** command we can get a better idea about what is in the data set.

```
. describe
```

You should see that you have four variables with 47 observations. To see how these different variables relate to each other, we can create scatterplots to examine various variable pairs.

```
. scatter priv month
```

Question 2: Describe this graph. Do you see a trend? What amount is

collected by the private contractors in a “typical” month? What’s the least amount collected? What’s the most?

Graphs like this are sometimes called “time series” graphs because the points are displayed in the time-order in which they were collected. It can be easier to see this trend if you connect-the-dots:

```
. scatter priv month, c(1)
```

Question 3: Next examine the trend for the government contractors. Before you make this plot, predict what you think the plot will look like.

Question 4: How does your prediction compare with the actual plot? If they are different, describe how they are different and explain why your prediction was inaccurate. If your prediction is close, explain why you made that prediction.

```
. twoway (scatter priv month, c(1)) (scatter govt month, c(1)
yaxis(2)), yscale(axis(1) r(500000 2500000)) yscale(axis(2)
r(5000 20000)) indexgraph options@graph options!yscale@yscale
```

Stata hint: To make the previous graph more readable, add options. Append `ylabel(, angle(horizontal)) ylabel(, axis(2) angle(horizontal))` to the previous graph command to turn the *y*-axis labels on their side.

Question 5: Now consider the plot displaying the monthly amounts collected by the private contractors versus the monthly amounts collected by the City. Which graph shows the most fluctuation from month-to-month? Do the graphs look similar? How so? How do they differ? Why do you think this is?

To see how the Brinks contractors did with respect to the other private contractors, we can label the points on our plots.

```
. scatter priv month, mlabel(brinks)
```

Question 6: What does this graph tell us about Brinks that the other graph did not?

Notice that the amount collected from meters varies from month to month, and sometimes by quite a bit. This is troublesome because it helps hide the thievery. If the Brinks employees stole a little one month and it lowers the amount collected, it might just look like a typical “bad” month. This is why the city’s sample might be useful. If some months are “good,” then both the city and the private contractors should have higher than normal amounts collected. Likewise, in “bad” months, both should have lower than normal amounts.

Question 7: What do you think a graph of private amounts versus the government amounts will look like? Create this scatterplot. What does it tell you, if anything, about the relationship between the amount the government collected and the amounts the private contractors collected?

Sometimes a pattern can be difficult to detect “by eye.” If the relationship between two variables is linear, then the correlation coefficient is a useful summary of the strength of that relationship. *Stata* calculates the correlation coefficient using the **corr** command.

Question 8: Will the correlation coefficient between the amounts collected by private contractors and government contractors be positive, negative, or close to 0? Guess what value the correlation will be? Explain.

You can find the value by typing

```
. corr priv govt
```

Question 9: How close was your guess? If you were off, why do you think this is?

Another summary of the linear relationship is provided by a regression line (also called a least squares line or a best fit line.)

In *Stata*, we use the **regress** command to get the equation of the least squares regression line. With this set of data, we want to use the *govt* variable to predict the *priv* variable.

```
. regress priv govt
```

This command performs the regression and produces a table of output in the Results window.

Question 10: Write the equation of the best fit line. Interpret the slope.

Stata allows us to superimpose the best-fit line onto the graph of private amounts versus the government amounts.

```
. twoway (scatter priv govt) (lfit priv govt)
```

Recall that one way to help us assess the fit of a regression line is to look at a residual plot. The residuals represent what is called the **error**: the difference between the actual observed value and the predicted value. If the regression line is a good summary of the scatter plot, then a plot of the residuals against the x -values should show no pattern and should be equally distributed around the line $y = 0$.

To calculate the residuals we first calculate the \hat{y} 's and then use the **generate** command to calculate the difference. Calculating the \hat{y} 's in *Stata* is easy. The **predict** command will compute the predicted values based on the most recent **regress** command input into *Stata*.

```
. predict predpriv  
. generate resid = priv - predpriv
```

We now can graph the residuals against the government amounts.

```
. scatter resid govt, yline(0)
```

Stata hint: The residuals can be plotted even easier using the built in *Stata* shortcut: `rvpplot govt, yline(0)`. This command does not require that you predict the \hat{y} 's or that you generate the residuals. It does all that for you.

Question 11: Do you think the regression line is a good summary of the relationship between the amount collected by the government and the amount collected by private contractors?

Looking closer at the residual plot, we can find a pattern that might otherwise be overlooked. Look at just the residuals for the months when Brinks collected money.

```
. scatter resid govt if brinks == 1, yline(0)
```

Question 12: What does this plot tell us about Brinks' performance? Does this change your opinion about whether or not the regression is a good summary?

Estimating the Amount Stolen

One approach to answering the original question about how much money the Brinks employees stole is to reason like this: we can use the honest contractors and compare them to the government to see what amount an honest contractor should collect whenever the government collects a certain amount. Then we go back to the months in which Brinks was the collector.

We see how much the city collected, and use this relationship to predict how much we should have gotten from Brinks.

To do a regression using only the honest contractors we add an **if** statement to our regress command.

```
. regress priv govt if brinks == 0
```

Now *Stata* has calculated the regression equation that best fits the points that corresponds to the times Brinks was not the collector.

```
. twoway (scatter priv govt) (lfit priv govt) if brinks == 0
```

The predict command will now use this equation to calculate for each observation the amount the private collectors were expected to collect based on the amount the City collected. Note that all values of *govt* are used, not just the values from the non-Brinks months that were used in fitting the model.

```
. predict predpriv2
```

The residuals tell us how much more or less the private contractors collected compared to the amount the regression equation predicts that they should have collected. Let's look at the non-Brinks or "honest" collectors first.

```
. generate resid2 = priv - predpriv2  
. scatter resid2 govt if brinks == 0, yline(0)
```

From the residual plot it is clear that sometimes the regression equation is over-predicting (i.e., the residual is negative) and sometimes the regression equation is under-predicting (the residual is positive). We can be assured though that when we sum up the residuals of the non-Brinks contractors, the total sums to 0. This is fact of least squares regression.

Stata doesn't let us simply add up all of the residual, but we can see the average of the residuals. Since the average is just the sum divided by the number of entries, if we multiply the average by the number of entries, we are left with the sum.

```
. summarize resid2 if brinks==0
```

As expected, the average of the residuals is 0 and thus the sum (0×22) of the residuals is also 0. (22 is the number of observations.)

We excluded the Brinks observations when calculating the regression model and furthermore, we know that some employees of Brinks were stealing money, thus we would expect that on average the regression equation will overestimate the amount Brinks was reported to have collected.

```
. scatter resid2 govt if brinks == 1, yline(0)
```

A look at the residuals for the Brinks observations verifies our hypothesis. We can use these residuals as a method to estimate the amount the Brinks' employees stole from the parking meters.

Question 13: Estimate the amount of money the Brinks' employees stole from the City's parking meters.

Assignment

For this in-class part of the lab you should hand in answers to all of the above questions, as well as the graphs you created.

On Your Own

Suppose you have been hired as a consultant by the city of New York to determine the amount of money that Brinks should pay. Write a report to the judge stating this amount and explaining how you reached your decision. Assume the judge has had a single introductory statistics course like this one. She will need to know any assumptions you made and will want to know whether or not you think they are true. Keep in mind that the employees were already found guilty. It is not your job to prove guilt, but to provide a statistically valid estimate of how much was stolen.

After reading your report, the judge asks this question: “Wouldn’t it be simpler and just as good to just find the average amount collected in the 22 months that Brinks was not the contractor, and then multiply this average by 24 to determine the amount that Brinks should have collected during the 24 months that they were the collector?” Write an explanation to the judge as to why you think this is or is not a good idea.