Econometrics - Computer Workshops

Juergen Bracht (Ph.D. Economics, Pittsburgh, U.S.A.)

15 March 2009

Abstract

1 Computer Workshop 1 - Basics, Regression

1.1 Basics

Topics: Basics, Viewing and generating series, Histogram, Regression line, Scatter graph, commands.

1. Log in to WebCT at http://www.abdn.ac.uk/webct/. Choose the course from the list. Click on Computer workshops followed by DATA. Right-click on the file nysevolume.wf1 and save the file on your h drive.
2. On the PC desktop, go to the folders Arts & Social Science/Business School/Economics. To open the program, click on the EViews icon.
3. In the EViews menu, go to File/Open Workfile .... Choose the file nysevolume.wf1 on your h drive.

Remark: nysevolume.wf1 is a EViews workfile. A **workfile** is the basic EViews Document.

Remark: **Data**: Quarterly average trading volume on the New York Stock Exchange (NYSE). Over 400 observations taken across more than a century.

Remark: This things in EViews are called **objects**. EViews has "Views". One view is the **Spreadsheet view**. We will have a look.

1. View the series Volume: Right-click on the series object Volume in the workfile. Go to Open. Click on the button View in the series window. In the menu, specify Descriptive Statistics & Tests, then Histogram and Stats. Obtain figure 1.
2. Plot the time-series of Volume. In the series window Volume, hit the button View. Graph, Line & Symbol. Obtain figure 2.

Figure 1: Volume: Histogram and Stats

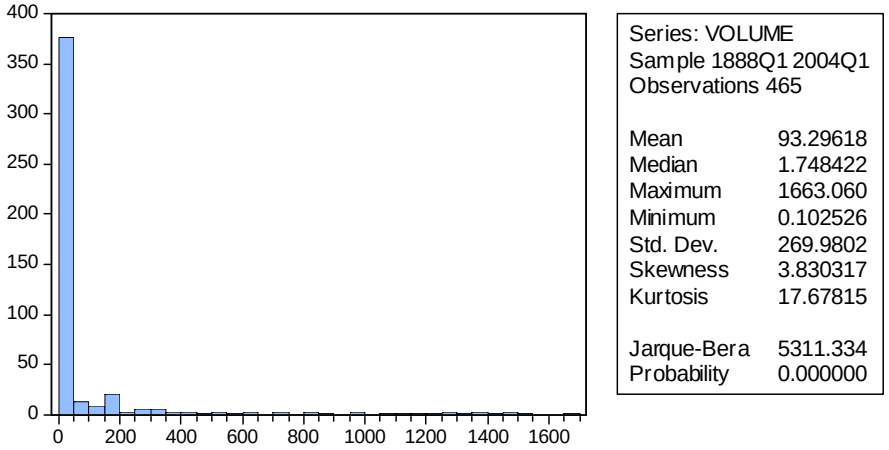


Figure 2: Volume: Time series, all observations

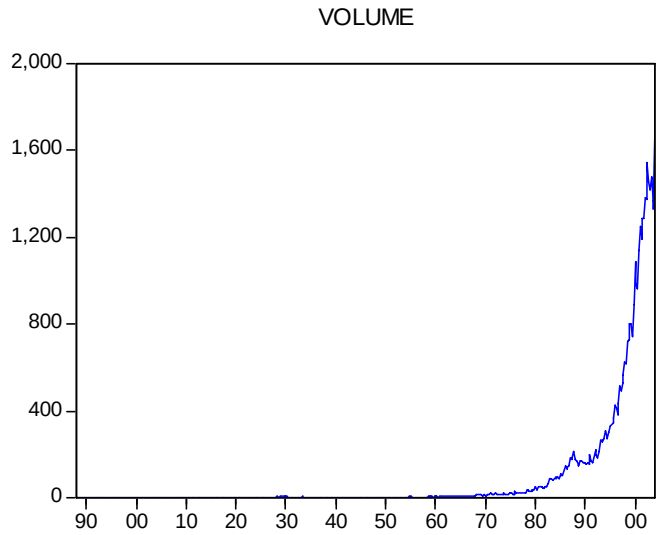
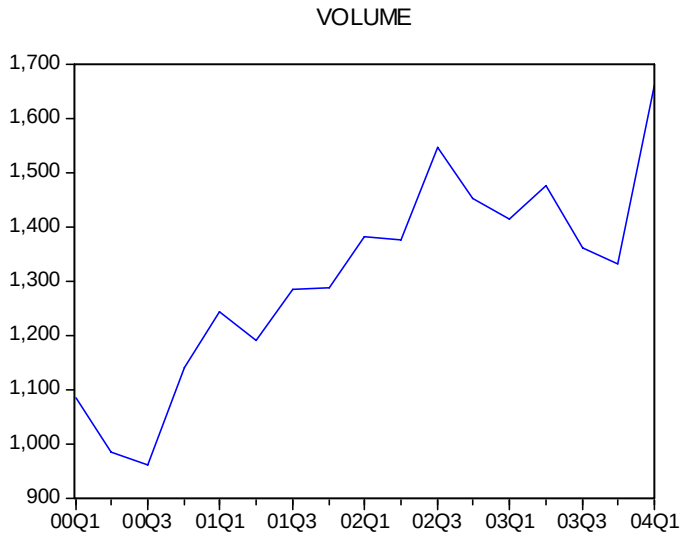


Figure 3: Volume: Time series, selected observations



3. In the series window Volume, hit the button Sample. A **Dialog box** opens. The upper field, marked **Sample range pairs (or sample object to copy)**, indicates that all observations are being used. Replace "@all" with the beginning and ending dates you want. In this case, use "2000q1" for the first date, a space to separate the dates and the special code "@all" to pick up to the last date available in the workfile (i.e. type "2000q1 @last"). Hit the Button OK. Obtain figure 3.

Remark: Have a look at the **Line Graph**. The scaling on the horizontal axis has changed. Previously we could fit only one label for each decade. This close-up view gives a label every six months. For example, "03Q3" means year 2003, third quarter, which is to say, "July-September 2003". We can see lots of short-run up and down spikes. Note: When we changed the sample on the view, we have not changed the underlying data, just the portion of the data that we're viewing at the moment. The complete set of data remains intact, ready for us to use any time we'd like. The new sample applies to all our work until we change it again, not just this one graph. In the workfile, note the change in the **SAMPLE: line** of the workfile.

Our first graph (before we shortened the sample) presented a picture which looks a lot like exponential growth over time. A trick for dealing with exponential growth is to look at the logarithm of a variable, relying on the identity $y = e^{gt} \Leftrightarrow \log y = gt$. In order to look at the trend in the log instead of the level, we'll create a new variable Logvol which equals the log of Volume. This can be done with a **dialog**.

1. In the workfile, choose the **menu item** Quick/Generate series to bring up a dialog box. In the upper field, type "logvol=log(volume)". The workfile window has a new object, logvol. Double-click logvol and then scroll the window so that the beginning of 2000 is at the top. Starting in 2000 we see numbers, before 2000, only the letters "NA".

Remark: Two lessons. (1) EViews operates only on data in the current sample. (2) EViews marks data that are not available with the symbol NA. There are three ways to change the sample: (i) Click the

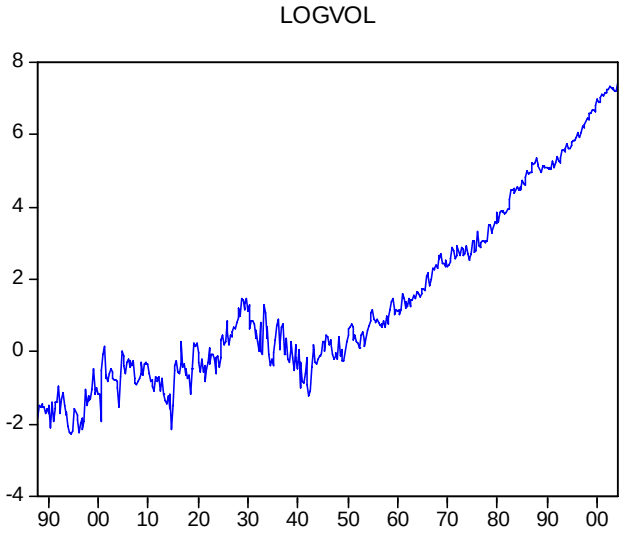


Figure 4:

menu selection QUICK/SAMPLE. (ii) Double-click on the sample line in the upper pane of the workfile window. (iii) Type a **command**.

1. Follow the third way. The workfile window and the series window appear in the lower section of the **master EViews window**. The upper area is reserved for typing commands and is called the **command pane**. The command smpl is used to set the sample; the keyword @all signals EViews to use all available data in the current sample. Type the command "smpl @all" in the command pane.

Remark: You can see that the sample in the workfile window has changed back to 1888Q1 through 2004Q1.

1. Type "series logvol=log(volume)" in the command pane to generate the series logvol again; this time from the command line.

2. Double-click on the object logvol to check that we now have all the data.

3. Hit the button View, choose Graph, in the dialog Graphic Options choose a line graph. The line graph for logvol - the logarithm of our original variable volume - appears. Obtain figure 4.

Remark: Note that the line is closer to a straight line.

Remark: Conclude from looking at our logvol line graph that NYSE volume rises at a more or less constant percentage growth rate in the long run, with a lot of short-run fluctuation. Or perhaps the picture is better represented by a slow growth in the early years, a drop in volume during the depression (starting around 1929) and faster growth in the post-war era. If we have a variable t representing time (0, 1, 2,

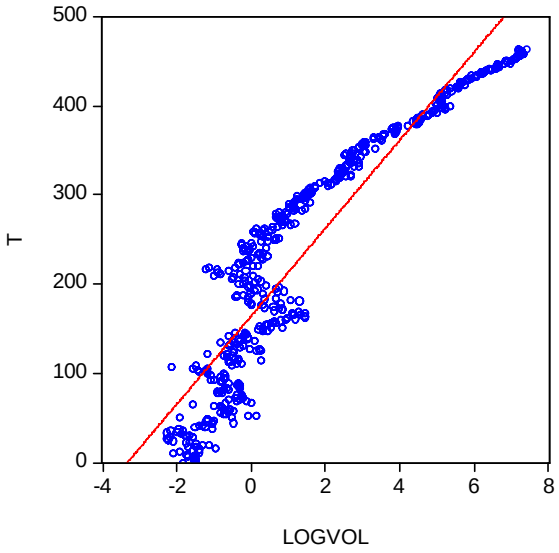


Figure 5:

3,...), then we can represent the idea of an upward trend with the algebraic model: "log(volume)=$\alpha+\beta t$" where the coefficient β gives the quarterly increase in LOGVOL. Remember that this type of model is a semi-elasticity model.

1. Create the variable t. In the command pane at the top of the EViews screen type "series t=@trend". Double-click on the object t.

Remark: @TREND is a function built into EViews for manipulating data. We want to think about how volume behaves over time, we want to look at the series t and logvol together. In EViews a collection of series dealt with together is called a **group**.

1. To create a group including t and logvol, first click on the object logvol. Now, while holding down the Ctrl-key, click on the object t. Then right-click highlighting Open, bringing up the content menu and choose **as Group**.

Remark: The group window shows the time and log volume, that is, the series t and logvol, together. Just as there are multiple ways to view the series, there are also a number of **group views**: a spreadsheet view; a scatter diagram.

1. In the group window, click on the View button and choose Graph .... In the Graph Options window that pops up, select the graph type Scatter on the left-hand side. To add a regression line, go back to the window Graph Options: Right-click somewhere in the graph and choose Options. Then select Regression Line from the Fit lines drop-down menu on the right-hand side. The default options for a regression line are fine, so hit the button OK to dismiss the dialog. Obtain figure 5.

Remark: Note that the straight line gives a good rough description of how log volume moves over time. We see that the intercept is roughly -2.5. When $t = 400$, logvol is around 4. The formula for the slope gives us an approximation for slope parameter β. $\widehat{\beta} = \frac{4-(-2.5)}{400-0} = 0.01625$.

1. Run a regression, type in the command pane: "ls logvol c t".

Remark: Here "ls" means least square. The first variable is the dependent variable. The second variable is the independent variable. "c" is a **special keyword** signaling EViews to estimate an intercept. The coefficient on the variable "c" is α, just as the coefficient on the variable t is β. Whether you use the menu or type a command, EViews pops up with regression results. The estimated intercept, $\widehat{\alpha}$, is -2.629649 and $\widehat{\beta} = 0.017278$.

1. Hit the Name button in the **equation window**. Spaces aren't allowed when naming an object in EViews. Note the **equation object** in workfile.

Remark: EViews does not have an undo feature. Well, EViews does have an undo item in the usual place on the edit menu. It works when you're typing text. It does not undo changes to the workfile.

1.2 Regression

1. Hit the button Object/New object, pick Equation in the dialog box.

   Type LOG(VOLUME)=C(1)+C(2)*@TREND.

Remark: For the record, EViews' label the standard error of the regression.S.E. of the regression. The elements "Sum squared residuals", "Log likelihood", "Aikaike info criterion", "Schwarz info criterion" are used to make statistical comparison between two regressions. The sum of squared residuals is used in computing F tests, the log likelihood is used for computing likelihood ratio test, and the Akaike and Schwarz criteria are used in Bayesian model comparison. The mean dependent variable and the S.D. dependent variable report the sample mean and standard deviation for the left hand side variable. The standard deviation of the dependent variable is much larger than the standard error of the regression. This is good. F statistic and $Prob(F\ statistic)$ come as a pair and are used to test the hypotheses that none of the explanatory variables actually explain anything. Put more formally, the F statistic computes the standard F test of the joint hypothesis that all the coefficients, **except the intercept**, equal zero. $Prob(F\ statistic)$ displays the p value corresponding to the reported F statistic. In this example, there is essentially no chance at all that the coefficients of the right-hand-side variables all equal zero. This example has only one coefficient so the t statistic and the F statistic test exactly the same hypotheses. Not coincidentally, the reported p values are identical and the F is exactly the square of the t: $2672.9 = 51.70045^2$.

Dependent Variable: LOG(VOLUME)				
Method: Least Squares				
Sample: 1888Q1 2004Q1				
Included observations: 465				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.629649	0.089576	-29.35656	0.0000
@TREND	0.017278	0.000334	51.70045	0.0000
R-squared	0.852357	Mean dep var	1.378867	
Adjusted R-squared	0.852038	S.D. dependent var	2.514860	
S.E. of regression	0.967362	Akaike info criterion	2.775804	
Sum squared resid	433.2706	Schwarz criterion	2.793620	
Log likelihood	-643.3745	Hannan-Quinn criter.	2.782816	
F-statistic	2672.937	Durbin-Watson stat	0.095469	
Prob(F-statistic)	0.000000			

1.3 Multiple Regression

An addition with a multiple regression is that there are added right-hand-side variables and therefore added rows of coefficients, standard errors, etc..

1. Add two more variables time trend squared and lagged values of the dependent variable. Go to Quick, then Equation Estimation Dialog.

   Type log(volume)=c(1)+c(2)*@trend+c(3)*@trend^2+c(4)*log(volume(-1)).

   Here $@trend\hat{}2$ is the time trend squared and $log(volume(-1))$ is the one-period lagged Volume. EViews brings up the multiple regression output with $R^2 = 0.986826$.

Dependent Variable: LOG(VOLUME)				
Method: Least Squares				
Sample (adjusted): 1888Q2 2004Q1				
Included observations: 464 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.106396	0.045666	-2.329866	0.0202
@TREND	-0.000736	0.000417	-1.764606	0.0783
@TREND^2	6.63E-06	1.37E-06	4.829663	0.0000
LOG(VOLUME(-1))	0.868273	0.022910	37.89886	0.0000
R-squared	0.986826	Mean dep var	1.385802	
Adjusted R-squared	0.986740	S.D. dependent var	2.513120	
S.E. of regression	0.289391	Akaike info criterion	0.366505	
Sum squared resid	38.52359	Schwarz criterion	0.402193	
Log likelihood	-81.02909	Hannan-Quinn criter.	0.380553	
F-statistic	11485.70	Durbin-Watson stat	2.342018	
Prob(F-statistic)	0.000000			

1.4 (If time permits) Hypothesis Testing

1. In the equation window, click the View button and choose Coefficient Tests/Wald - Coefficient Restrictions to bring up a dialog. EViews names coefficients $c(1)$, $c(2)$, $c(3)$ etc numbering them in the order they appear in the regression. You specify a hypotheses as an equation restricting the values of the coefficients in the regression. To test that the coefficient on $log(volume(-1))$ equals zero, specify "$c(4) = 0$".

Remark: If a hypothesis involves multiple restrictions, you enter multiple coefficient equations separated by commas. EViews reports the F statistic rather than the t statistic because the F applies to both single and multiple restrictions. Here, t statistic is the square root of the F. That is, $\sqrt{1436.324} = 37.899$. The p value is 0.0000.

1. Test whether the coefficient on $log(volume(-1))$ equals one rather than zero. Enter "c(4)=1" to find the new test statistic.

Remark: So this hypothesis is also easily rejected. Warning: If you were to study the advanced topic called the "unit root problem" you would learn that standard theory doesn't apply in this test (although the issue is harmless for this particular set of data).

A good example of a hypothesis involving multiple restrictions is the hypothesis that there is no time trend, so the coefficients on both t and t^2 equal zero.

1. In the Wald Test dialog enter "c(2)=0, c(3)=0".

Remark: The hypothesis is rejected (EViews correctly reports 2 degrees of freedom for the test statistic).

1.5 Appendix

Source: EViews Illustrated

Data source: http://www.eviews.com/download/download.html

Econometrics - Computer Workshops

Juergen Bracht (Ph.D. Economics, Pittsburgh, U.S.A.)

15 March 2009

Abstract

1 Computer Workshop 2 - Model Specification

We will investigate empirically which factors affect chief executives officer (CEO) salaries.

1.1 Empirical analysis

We will work with the EViews workfile CEOSAL1. The set contain various measures of firm performance. The file contains information on 209 CEOs for the year 1990; these data are obtained from the magazine Business Week (5/6/91).

Exercise 1 ***(Definitions, Measurements and a Histogram)*** *Click on the series salary, roe, ros, sales in the workfile. How are the variables defined and how are they measured? Why might the independent variables roe, ros and sales "cause" salary?*

Exercise 2 *Plot a histogram of salary. Describe the distribution. Obtain figure 1.*

EViews hint (histogram): Right-click on the series object salary in the workfile. Choose Open. Alternatively, double-click on the object,. Click on the button View in the series window. In the menu, specify Descriptive Statistics & Tests, then Histogram and Stats.

Partial Solution: In this sample, the average annual salary is \$1281120 with the smallest and largest being \$223000 and \$14822000 respectively. The average return on equity, 88-90 avg, is 17.18% with the largest and smallest values being 56.30% and 0.50%, respectively.

Exercise 3 ***(Histogram and Scatter diagram)*** *Plot a histogram of roe. Describe the distribution. Plot a scatter diagram roe and salary. Is there a relationship between roe and salary? Obtain figure 2.*

EViews hint (scatter diagram): We want to look at the variables *roe* and *salary* together. In EViews a collection of series dealt with together is called a **group**. First, click on the object *salary*. Now, while holding down the Ctrl-key, click on the object *roe*. Right-click. Choose to open **as Group**. The group window shows *roe* and *salary*, that is, the series $SALARY$ and ROE, together. In the group

Figure 1: Salary: Histogram and Stats

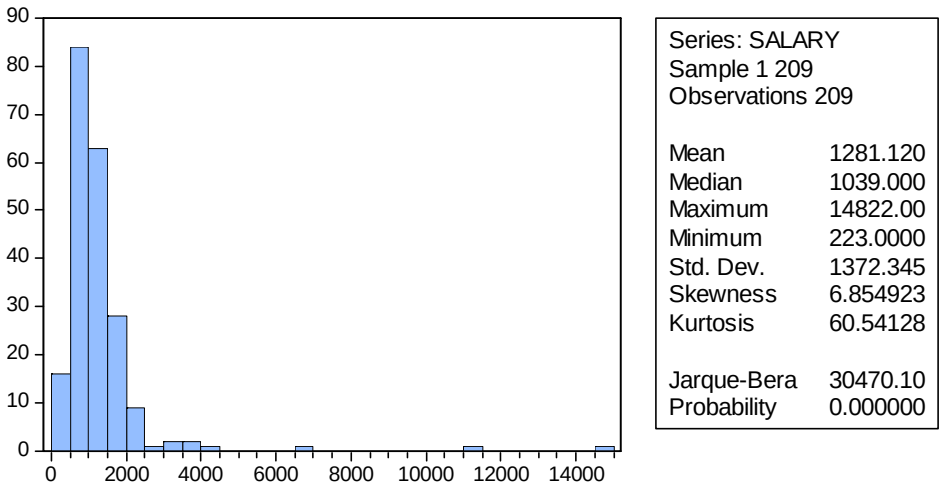
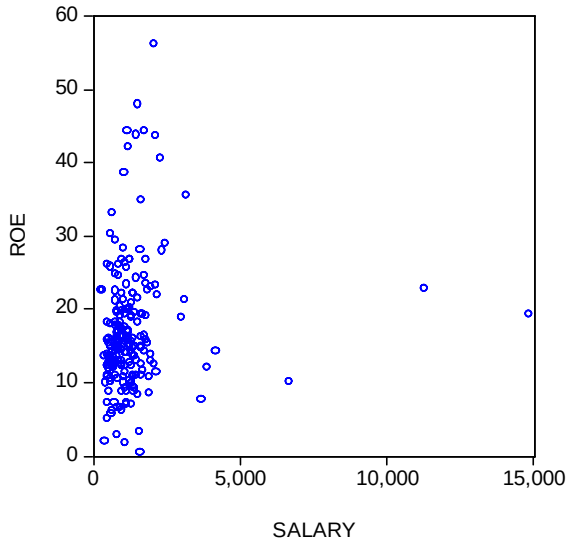


Figure 2: Scatter diagram *roe* and *salary*



window, click on the View button and choose Graph .... Then select Scatter on the left-hand side of the dialog that pops up.

Before we do regression analysis, let us review a summary of combinations of functional forms available from using either the original variable or its natural log.

Model	Independent variable	Dependent variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \Delta x$
Log-level (semi-elasticity)	$log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log (constant elasticity)	$log(y)$	$log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

First, we will look at a level-level model, then at two constant elasticity models.

Exercise 4 ***(Level-level model)***

A) Obtain the OLS regression line relating salary and roe, and the R^2.

B) Interpret the equation.

C) Write the predicted change in salary as a function of roe.

D) Graph the estimated line. On a printout, sketch the population regression function $E(salary|roe)$.

E) How much of the variation in salary is explained by the return to equity?

F) Is this lack of explanatory power surprising?

G) Does the seemingly low R^2 necessarily mean that an OLS regression equation is useless?

Partial solution: The firm's return to equity explains only 0.013189 of the variation in salaries for this sample of CEOs. That means that 0.98681 of the salary variations for these CEOs is left unexplained.

Exercise 5 ***(Constant elasticity model)***

Estimate a constant elasticity model relating CEO salary to firm sales. Let sales be annual firm sales, measured in millions of dollar. The model falls under the ***simple regression model*** *by defining the dependent variable to be $y = log(salary)$ and the independent variable to be $x = log(sales)$.*

A) Estimate the equation $log(salary) = \beta_0 + \beta_1 \log(sales) + u$ where β_1 is the elasticity of salary with respect to sales. Report your results in the usual form. Write down the R^2.

B) What is the coefficient of $log(sales)$? What does it imply for this data?

Partial solution: $log(salary) = 4.821996 + 0.256672log(sales)$, $n = 209$, $R^2 = 0.210817$. The coefficient of $log(sales)$ is the estimated elasticity of *salary* with respect to *sales*. It implies that a 1% increase in firm sales increase CEO salary by about 0.257%.

Exercise 6 ***(Another constant elasticity model)*** *Consider an equation to explain salaries of CEOs in terms of annual sales, return on equity (roe, in percentage form) and return on the firm's stock (ros, in percentage form):* $log(salary) = \beta_0 + \beta_1 log(sales) + \beta_2 roe + \beta_3 ros + u$.

A) In terms of the model parameters, state the null hypothesis that, after controlling for sales and roe, ros has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.

B) Estimate the model.

Partial solution:

Dependent Variable:	LOG(SALARY)			
Method:	Least Squares			
Sample	1 209			
Included observations:	209			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.311712	0.315433	13.66919	0.0000
LOG(SALES)	0.280315	0.035320	7.936426	0.0000
ROE	0.017417	0.004092	4.255977	0.0000
ROS	0.000242	0.000542	0.446022	0.6561

C) Test the null hypothesis that *ros* has no effect on *salary* against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

Solution: The t statistic on ROS is 0.446022 which is well below the critical value. Therefore, we fail to reject the H_0 at the 10% significance level.

Data Source: Introductory Econometrics, Wooldridge, International Edition, 4th.

Further reading: Rose, N.L. and A. Sheppard (1997), Firm Diversification and CEO Compensation: Managerial Ability or Executive Entrenchment?, RAND Journal of Economics, 28, 489-514

1.2 Appendix: Motivation "Restriction on CEO's salaries"

Are there good reasons for ceilings on the pay of executives at companies receiving assistance from the government?

Examples of ceilings on the pay of executives: **Salary cap** for executives at companies, **restrictions** on when executives can cash in the **stock** they will receive, **limits** to executives' **severance pay** and **monitoring** of **fringe benefits**, like company jets.

We acknowledge that there are no good guides to setting of either the form or the level of compensation to employees in any occupation, including executives.

However, we believe that the main problem with wage and price controls is that controls do not work, although governments have imposed them throughout history.

Reasons against wage controls: 1) Lawyers and accountants will discover fringe benefits that can help circumvent the pay caps. 2) Executives will not want to work for companies in distress.

We have asked an empirical question: which factors actually do affect chief executives officer (CEO) salaries?

Econometrics - Computer Workshops

Juergen Bracht (Ph.D. Economics, Pittsburgh, U.S.A.)

15 March 2009

Abstract

1 Computer Workshop 3 - Weighted Least Squares, Heteroskedasticity, Limited dependent variable

Topics: Growth of currency in the hand of the U.S. public and union membership in the U. S..

Objective: To illustrate how EViews works ... little depth here, actually.

Data sets: currency.wf1, cpsmar2004wa.wf1.

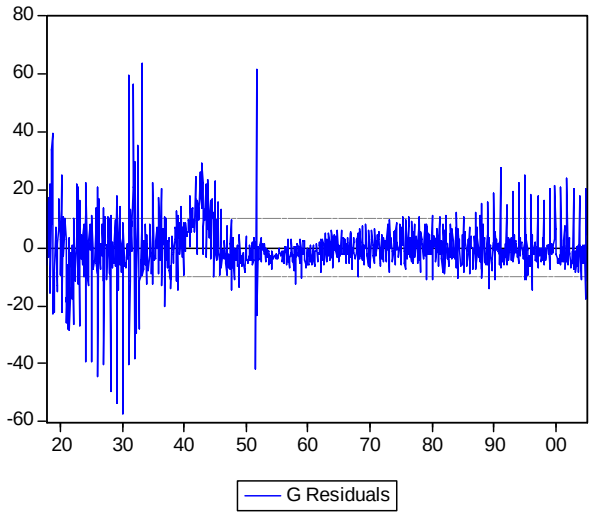
1.1 Weighted Least Squares

Data set: currency.wf1

The version of least squares that attaches weights to each observation is named weighted least squares.

1. Make sure that you include all observations in the sample (type smpl @all in the pane).
2. Type the command "ls g @trend g(-1) @expand(@month)". @expand(@month) is added in estimation to indicate the use of one or more automatically created dummy variables.
3. A new window has opened, titled Equation. The window shows your regression results. Click on the button Name in the menu. Return to the Equation window to plot a figure showing actual and fitted currency and the left over. Maximize the Equation window, go to View, then to Actual, Fitted, Residual and then to Actual, Fitted, Residual Graph. The red line shows the time-series graph of actual currency, the green line shows the fitted currency and the blue line shows the residuals. Note that the left axis shows the scale of the residuals and the right axis shows the scale of the actual and fitted currency. Plot the residuals (go back to Residual Graph). You should obtain figure 1.

Figure 1: Currency: Plot of residuals



Remark: The estimation results look fine but the residuals were much nosier early in the sample than they were later on.

Remark: We might get a better estimate by giving less weight to the early observations. As a rough and ready adjustment after looking at the residual plot, we'll choose to give more weight to observations from 1952 on and less to those earlier.

1. (Advanced) Find the error standard deviations for each subperiod. To do that, use the **Stats By Classification** ... view of residuals. Open the resid series in the workfile. Go to View, Descriptive Statistics & Tests, Stats By Classification. Type "@year<1952" for the series to classify.

Remark: You obtain the descriptive statistics for the residual series, categorized by values of @YEAR<1952, Sample (adjusted): 1917M10 2005M04, Included observations: 1045 after adjustments.

@YEAR<1952	Std. Dev.
0	5.592399
1	14.00910
All	9.785594

The residual standard deviation is around 6 in the years prior to 1952 and around 14 in 1952 and thereafter. Look at the figure and you can can see that the standard deviation falls in 1952. We'll use this information to create a new series. Here is how we go about this.

1. Create a series rough_w for weighting observations.

   Type "series rough_w =14*(@year<1952)+6*(@year>=1952)" in the command pane.

Remark: That's the heart of the trick in instructing EViews to do weighted least squares - you need to create a series which holds the weight for every observation. EViews then multiplies each observation by the weight you supply. Essentially, this is equivalent to replicating each observation in proportion to its weight. Let us now use EViews' **Weighted Option**.

1. Open the least squares equation EQ01 in the workfile. Click the Estimate button and switch to the **Options** tab. Check the **Weighted LS/TSLS** box and enter the weight series "1/rough_w" in the **Weight** field.

Remark: That's because we want the weights to be inversely proportional to the error standard deviation (well, that is the idea!!!). The weighted least squares estimates include two summary statistics panels. The first panel is calculated from the residuals from the weighted regression, while the second is based on unweighted residuals (ok then). Well, this is really advanced stuff, and should be covered under the headline "homoskedasticity assumption for time series models".

1.2 Heteroskedasticity

Data sets: currency.wf1

EViews offers both tests for heteroskedasticity and methods for producing correct standard errors in the presence of heteroskedasticity.

1.2.1 Test for Heteroskedastic Residuals

1. Type the command "ls c @trend @trend^2 g(-1) g(-1)^2 @expand(@month)" in the command pane.

The Residual tests/Heteroskedasticity tests ... view of an equation offers two variants of the White heteroskedasticity test. The White test is essentially a test of whether values of the right-hand side variables - and/or their cross terms - help explain the squared residuals. To perform a White test with only the squared terms (no cross terms), uncheck the **INCLUDE WHITE CROSS TERMS**.

Remark: The results of the White test on our currency growth equation show that the null hypothesis of homoskedasticity is rejected (see F and χ^2).

1.2.2 Heteroskedasticity Robust Standard Errors

One approach to dealing with heteroskedasticity is to weight observations such that the weighted data are homoskedastic. That's essentially what we did in the previous section. A different approach is to stick with least squares estimation, but to correct standard errors to account for heteroskedasticity.

1. Type the command "ls g @trend g(-1) @expand(@month)" in the command pane.

   Click the Estimate button in the equation window.

   Switch to the **OPTIONS** tab.

   Check **HETEROSKEDASTICITY CONSISTENT COEFFICIENT COVARIANCE** and choose **WHITE**.

Remark (Advanced): Note that some standard errors are smaller and some a larger (Ok then).

Remark: Hypothesis tests computed using **COEFFICIENT TESTS/WALD-COEFFICIENT RESTRICTIONS** ... correctly account for the adjusted standard errors. The **OMITTED VARIABLES** and **REDUNDANT VARIABLES** tests do not use the adjusted standard errors.

1.3 Limited dependent variable problem

Logit: Think of the model having two parts. First, an index s is created, which is a weighted combination of the explanatory variables. Then, the probability of observing the outcome depends on the cumulative distribution function of the index. Logit uses the cdf of the logistic distribution; probit uses the normal distribution instead.

Data set: cpsmar2004wa.wf1

1. To do a Logit, type "logit union c age" in the command pane.

Dependent Variable:	UNION			
Method:	ML - Binary Logit	(Quadratic hill climbing)		
Sample:	1 1759			
Included observations:	1759			
Convergence achieved after 5 iterations				
Covariance matrix computed using second derivatives				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-4.234841	0.447392	-9.465621	0.0000
AGE	0.028066	0.010102	2.778217	0.0055
McFadden R-squared	0.012502	Mean dependent var		0.043775
S.D. dependent var	0.204652	S.E. of regression		0.204371
Akaike info criterion	0.357301	Sum squared resid		73.38559
Schwarz criterion	0.363523	Log likelihood	-312.2459	
Hannan-Quinn criter.	0.359600	Restr. log likelihood	-316.1991	
LR statistic	7.906377	Avg. log likelihood	-0.177513	
Prob(LR statistic)	0.004926			
Obs with Dep=0	1682	Total obs	1759	
Obs with Dep=1	77			

The coefficients shown in the output are the coefficients for constructing the index. In this case, our estimated model says:

$s = -4.23 + 0.0028 * age$

$prob(union = 1) = 1 - F(-s)$, $F(s) = \frac{e^s}{1+e^s}$.

Hence, we could calculate and plot the probabilities, conditional on age, with ease.

Well, we will have a closer look at this stuff in the lecture.

Textbook hint: Textbooks describe the relation between probability and index in a logit with $prob(union = 1) = F(-s)$. The two expressions are equivalent.

Source: EViews Illustrated
Data source: http://www.eviews.com/download/download.html

1.4 Appendix: Union membership in the U.S.

- Union membership had been declining since 1983.
- Union share has been declining since the late 1940s. Almost 36% of the nonagricultural workers were represented by unions in 1945. Today that figure is about 12%.
- The rapid growth of public employee unions since the 1960s has served to mask an even more dramatic decline in private-sector union membership. At the apex of union share in the 1940s, only about 9.8% of public employees were represented by unions, while 33.9% of private, non-agricultural workers had such representation. In this decade, those proportions have essentially reversed, with 36% of public workers being represented by unions while private sector union density has plummeted to around 7%.
- We see the overall decline in union membership from its sharp decline with the age of the individuals: the union share of workers is highest among those aged 45-64 at almost 15% which exceeds the 11% for workers aged 25-44 and only 5% union share for workers under age 25.

Candidate factors: Social forces. Politics of Executive. Deregulation of communication, transportation and utilities industries. Shift in importance of manufacturing to service. Growth in imports, globalization. Codification of personnel relations.

Econometrics - Computer Workshops

Juergen Bracht (Ph.D. Economics, Pittsburgh, U.S.A.)

2 April 2009

Abstract

1 Advanced Computer Workshop

Topics: Creating an EViews workfile, entering data. Panel and Pool.

1.1 Creating an EViews workfile

I have collected this data in a previous introductory statistics course. Attendance is the number of tutorials attended (possible values are 0, 1,..., 8). Mark is the final grade (possible values are 0, 1,.., 20 with twenty being the top score).

Student ID	Attendance	Mark	Student ID	Attendance	Mark
1	7	11	11	5	11
2	8	15	12	8	19
3	8	17	13	6	10
4	8	17	14	8	20
5	7	16	15	7	15
6	8	19	16	7	20
7	7	16	17	7	19
8	6	6	18	8	12
9	5	17	19	8	19
10	8	19	20	6	18

As you see, data comes arranged in rows and columns. Every column holds a series of data; every row holds one observation. When data comes arranged in a neat rectangle, statisticians call the arrangement a **data rectangle**. When thinking of an econometric model, a data series is often called a **variable**.

The observations come in order and are often numbered ("**observation numbers**"). Sometimes the entire set of observation numbers is called **identifier** or **id series**. Hint: Series (columns) don't have any inherent order, but observation numbers (rows) do. EViews needs to know how observations are numbered. When you set up a workfile, the first thing you need to do is tell EViews how the identifier of your data is structured: monthly, quarterly, annual, just numbered 1, 2, 3, ..., etc.. Your second task

is to specify the range your observations take: For instance, January 1990 through January 2000, 1966 to 1994, etc ... .

1. To create a **New Workfile**, open EViews and use the menu to choose File/New/Workfile ... . The **Workfile Create** dialog pops up. The default for the workfile structure type is Dated - regular frequency and Annual. However, the data shown in the table are just numbered sequentially. They aren't dated. The **Workfile Structure Type** drop-down menu offers three choices: **Unstructured / Undated**, **Dated-regular frequency** and **Balanced Panel**. Our data are "Unstructured/ Undated". Select this option. "Unstructured/Undated" instructs EViews to number the observations from 1 through however-many-observations-you-have. In our example we have 20 observations. So enter "20" in the field marked "Observations". If you would like to give your workfile a name you can enter the name in the "**WF:**" field at the lower right. You can also name the workfile when you save it. Click on Ok. Hint: Windows stores your file in the My documents folder.

Background information (Workfile): The main window area is divided into an upper pane holding information about the workfile and a lower pane displaying information about the objects-series, equations, etc. that are held in the workfile. The **range** is identifying numbers or dates of the first and last observation in the workfile. The **sample** is the subset of the observations range being used for current operations. Since all we've done so far is to set up a workfile with 20 observations, both range and sample are telling us that we have 20 observations. Let's move to the lower panel. The new workfile comes with two objects that were preloaded. The series resid is designated to hold the residuals from the last regression. Since we have not yet run an estimation procedure, resid series' values are set to NA.

Recall that an EViews workfile holds a collection of objects and that each kind of object is identified by its own icon. The most important icons are the series icons, because that's where our data are stored. Notice that the object c has a different icon, a Greek letter β. Instead of a data series, c holds values of coefficients. Right now c is filled with zeros, but if you ran a regression and then double-clicked on c you would find it had been filled with estimated coefficients from the last regression.

There are two methods to pop open an Untitled series:

(i) Type the command "series" in the command pane.
(ii) Use the menu Object/New Object/Series.

Let us use these type of methods to create a series named "attendance". However, we want to place the series in the workfile.

1. Type the command "series attendance" in the command window. Alternatively, use the menu commands Object/New Object/Series and then enter "attendance" in the "Name for object" field. The methods place the series safely in the workfile. Then double-click to open a series window.

2. In contrast, the former two methods open a window automatically, but don't name it. These methods open an untitled series window. To name the untitled series, click on the Name button and enter "attendance".

Remember: EViews doesn't care about capitalization of name.

Note: Named objects are saved and Untitled ones aren't. This design lets you try out things without cluttering the workfile.

We are now ready to type away numbers. But there's is a trick to entering your data. To protect against accidents, EViews locks the series window so that it can't be edited.

1. To unlock the series window, click on the Edit+/- button.

Unlocked windows have an **edit field** just below the button bar. One way to know that window is locked against editing is to observe the absence of the edit field. Alternatively, if you start typing and nothing happens, you'll remember that you meant to click the Edit+/- button. Now let's turn to entering data in the form of a table. As an example, we'll enter the name of "attendance" and "mark" together. Hint: You can name a group and store it in the workfile just as you can with a series. A group is a list of series names. It is not a separate copy of the data. A series can be a member of as many different groups as you like.

1. Use the menu Quick/Empty Group (Edit series). When the window opens, scroll up one line. Then type "attendance" in the cell next to the cell marked obs. A dialog pops up so that you can tell what sort of series it is going to be. Choose **Numeric series**.

If "attendance" were a text rather than numbers, then choose **Alpha series**. EViews initializes alpha series with blank cells.

1. Move to the cell to the right of "ATTENDANCE" and enter "mark", using the radio button to indicate a Numeric series. EViews fills out the series with NAs.

So, this was really easy.

1. Click the Edit+/- button and type away. When editing a group window, the Enter key moves across the row rather than down a column. This lets you enter a table of data an observation at a time rather than one variable after the other.

Note: Both series, attendance and mark, have been created automatically in the workfile, along the group object. Well, this was really easy.

Here are some regression results:

Dependent Variable: MARK
Method: Least Squares
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.106061	5.682619	0.546590	0.5914
ATTENDANCE	1.787879	0.792624	2.255647	0.0368
R-squared	0.220372	Mean dependent var	15.80000	
Adjusted R-squared	0.177060	S.D. dependent var	3.887903	
S.E. of regression	3.526953	Akaike info criterion	5.453385	
Sum squared resid	223.9091	Schwarz criterion	5.552958	
Log likelihood	-52.53385	Hannan-Quinn criter.	5.472823	
F-statistic	5.087942	Prob(F-statistic)	0.036776	

1.2 Panel and Pool

A **panel** is a set of cross-section (countries, firms, people, etc.) where each place, institution or person is followed over time. A **pool** is a set of time-series on a single variable, observed for a number of places, institutions or people. Pools are very simple to use in EViews because all you need to do is be sure that series names follow a consistent pattern that tells EViews how to connect them with one another.

Data: "Pop_Pool_Panel.wf1".

1. Open POPCAN and POPUSA as a Group in the page POOL. Open POP as a single series in the page PANEL. Look at labeling of rows.

Note that there are two papes. One page shows the pool approach, with a pooled object ISOCODE. ISOCODE holds the words "CAN" and "USA" to tell EViews that POPCAN and POPUSA are series measuring "POP" for the USA and Canada.

The other page shows the panel approach which introduces a kind of structure in the workfile that we haven't seen before. The range field now reads "1950 2000 x 2". The data is still annual from 1950 through 2000, but the workfile is structured to contain two cross sections (Canada and USA) in a singole series POP.

1.2.1 Panel Data

Panel Data allows you to control for unobservables that would otherwise mess up your regression estimation. It's helpful to think of the observations in a time series as being numbered from 1 to T, even though EViews typically uses dates like "1966q3" rather than $1, 2, 3$ ... as identifier. Cross section data are numbered 1 to N, it being a convention to use T for time series and N for cross section. Using i to subscript the cross section and t to subscript the time period, we can write the equation for a regression line as: $y_{it} = \alpha + \beta x_{it} + u_{it}$. With a panel, we are able to estimate the regression line using $N \times T$ observations. Our data is from the Penn World Table.

1.2.2 Using Panels to Control for Unobservables (Background)

A key assumption in most applications of least squares regression is that there aren't any omitted variables which are correlated with the included explanatory variables. Recall that omitted variables cause least squares estimates to be biased. The usual problem is that if you don't observe a variable, you don't have much choice but to omit it from the regression. When the unobserved variable varies across one dimension of the panel but not across the other, we can use a trick called fixed effects to make up for the omitted variable. As an example, suppose y depends on both x and z and z is unobserved but constant for a given country. The regression equation can be written as $y_{it} = \alpha + \beta x_{it} + [\gamma z_i + u_{it}]$ where the variable z is stuffed inside the square brackets as a reminder that, just like the error term, z is unobservable. The **trick** of fixed effects is to think of there being a unique constant for each country. If we call this constant α_i and use the definition $\alpha_i = \alpha + \gamma z_i$, we can rewrite the equation with the unobservable z replaced by a separate intercept for each country: $y_{it} = \alpha_i + \beta x_{it} + u_{it}$. EViews calls α_i a **cross section fixed effect**. The advantage of including the fixed effect is that by eliminating the unobservables from the equation we can now safely use least squares. The presence of multiple observations for each country makes estimation of the fixed effect possible.

We could have told the story for a variable that was constant over time while varying across countries. This would lead to a **period fixed effect**.

1.2.3 Setting Up Panel Data

Data: PWT61Extract.wf1, PWT61PanelExtract.wf1.

The easiest way to set up a **panel workfile** is to start with an ordinary workfile. Now we want that one series identifies the period and another series identifies the cross section. Our workfile has information for half a century.on both real GDP relative to the USA (!), Y, and on population for a large number of countries, POP. It also contains a series, ISOCODE, that holds an abbreviation for each country and a series, YR, for the year.

1. Open the group Y, ISOCODE, YR. The **range** is 1 10208 – 10208 observations. Now we need to tell EViews how the this data looks like. This exercise is called **structuring a panel workfile**.

2. To change from a regular to a panel structure, use the **Workfile structure** dialog. Double click on **Range** in the upper pane in the ordinary workfile. Choose **Dated panel** for the **Workfile structure type** and then specify the series containing the cross section, i, and date, t, identifiers. In our set, these variables are ISOCODE and YR.

EViews now re-organizes the workfile to have a panel structure. Verify the statement but looking at the **Range**, 1950 2000 x 208. Compare your workfile with the master set "PWT61PanelExtract.wf1".

1. More information about the structure of the workfile is available by pushing the View button in the workfile, then choose Statistics from the menu.

1.2.4 Panel Estimation

Data: PWT61PanelExtract.wf1.

There is the notion from the Solow growth model that high population growth leads to lower per capita output, conditional on available technologies. We test this theory by regressing real gross domestic product per capita relative to the United States, Y, on the rate of population growth, measured as the change in the log of population: Dependent variable Y, independent variable D(log(POP)).

1. Type the command smpl if isocode<>"USA", then ls y c d(log(pop)).

Dependent Variable:	Y			
Method:	Panel Least Squares			
Sample:	1950 2000	IF ISOCODE<>"USA"		
Periods included:	50			
Cross-sections included:	153			
Total panel (unbalanced) obs:	5622			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	44.93176	0.546861	82.16302	0.00
D(LOG(POP))	-901.0468	23.83234	-37.80773	0.00
R-squared	0.202772	Mean dependent var	27.76428	
Adjusted R-squared	0.202630	S.D. dependent var	25.58963	
S.E. of regression	22.85041	Akaike info criterion	9.096170	
Sum squared resid	2934433.	Schwarz criterion	9.098531	
Log likelihood	-25567.34	Hannan-Quinn criter	9.096993	
F-statistic	1429.424	Durbin-Watson stat	0.102652	
Prob(F-statistic)	0.000000			

Our result: Coefficient on d(log(pop)) is significant.

Now, let us focus on the Central African Republic and Canada.

1. Type the command smpl if isocode="CAF" or isocode="CAN".

Reminder: Variable names aren't case sensitive in EViews but string comparisons using "=" are.

1. Type the command show d(log(pop)). Example: CHE - 51 0.01295914532437499.

Hint: The function d() takes the first difference of a series, and the first difference of a log is approximately the percentage change. Hence, "d(log(pop))" gives the percentage growth of population.

1. In the series window, use the View button to choose **Descriptive Statistics & Tests/Stats by Classification** ... Use "isocode" as the classifying variable. Click on OK.

Result: Population growth was 2.2% (exactly, 0.022474) in CAF and 1.6% (exactly, 0.016092) in Canada. If we multiply the difference in population growth rates, $0.022474 - 0.016092 = 0.006382$, by the estimated regression coefficient, -901.0468, we predict that the relative GDP in the CAF should be $0.006382 * (-901.0468) = -5.7505$ % points lower than in Canada. Hence, population growth has a very large, negative effect on per capita output.

Remark: However, it is easy to imagine that population growth is picking up the effect of omitted variables that we can't measure. To the extent that the omitted variables are constant for each country, fixed effects estimation will control for the omissions.

1. Set the sample back to everything except the United States. Type smpl if isocode<>"USA". Type ls y c D(log(POP)). Click the Estimate button in the equation window. Then choose the Panel Options tab. Set Effects specification to Cross-section Fixed. This instructs EViews to include a separate intercept for each country. Click on OK.

Dependent Variable:	Y			
Method:	Panel Least Squares			
Sample:	1950 2000 IF ISOCODE<>"USA"			
Periods included: 50				
Cross-sections included:	153			
Total panel (unbalanced) obs:	5622			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	27.92379	0.219137	127.4263	0.00
D(LOG(POP))	-8.371774	10.57512	-0.791648	0.43
Effects Specification	Cross-section fixed (dummy variables)			

Hint (Advanced): The menu item Fixed/Random Effects Testing in the equation window offers a formal test for the presence of fixed effects. Look for it in the View menu.

Remark: When using fixed effects, the constant term reported in regression output is the average value of the country intercept. In the new regression results, the effect of population growth is reduced to about 1/100th of the previous estimate (now -8.371774, was -901.0468).

1. Take a look at estimated fixed effects. In the Equation window, click on View, then Fixed/Random Effects/Cross-section Effects. The reported values of the cross-section fixed effects are the intercept for country i less the average intercept. Canada positive (56.17477), CAF negative (-19.15951).

1.2.5 Quick Review

The panel feature lets you analyze two dimensional data. Convenient features include prettier identification of your data in spreadsheet views and some extra graphic capabilities. The use of fixed effects in regression is straightforward, and often critical to getting meaningful estimates from regression by washing out unobservables. The examples used cross-section fixed effect but you can use period fixed effects just as easily.

Source: EViews Illustrated