THE UNIVERSITY OF STRATHCLYDE
BUSINESS SCHOOL

Applied Econometrics

Professor Bernard Fingleton

Causation & Prediction

Causation

- One of the main difficulties in the social sciences is estimating whether a variable has a true causal effect
- Data are given to us by surveys etc where there is uncontrolled variation of a range of possible causes, so estimating individual causal effects when there is <u>other causal variation</u> going on in the background can lead to wrong interpretation of causation
- Usually we cannot carry out controlled experiments in which background variables are eliminated

Causation

- Typically we are interested in the effect of X on Y but we have other variables (Z) also affecting the outcome Y
- If we cannot eliminate Z experimentally (as in laboratory experiments) then we need to eliminate its effect statistically

Causation

» In experimental sciences, the effect of a treatment (X) is identified by eliminating other variables (eg Z) …the experimental subjects are identical

» We use multiple regression to control for some covariates (Z) to isolate the effect of X on Y,

» but some causal variables may still be omitted from the regression, so while regression is an advance on doing nothing, it does not prove causation

» This can only really be done via a controlled randomized experiment (see Stock & Watson,2007)

Granger causality

- One way to obtain a truer estimate of the effect of X on Y is to use that fact that in time series we have temporal order, before comes before after!
- Variable X <u>Granger causes</u> Y if past values of X have explanatory power
- This does not guarantee true causation (which is why we use the term Granger causation) but at least it suggests that X may cause Y
- If X does not Granger cause Y then we can be pretty sure that it does not cause Y.

Granger causality

- Assume stationary variables X and Y
- Include lags of both Y and X to form an ADL (autoregressive distributed lag)
- The Y lags should soak up residual autocorrelation so that we can validly use the F test to test whether X is significant
- if the X lags have any explanatory power, then that is only suggestive of their causal influence
- It does mean that past values of X contain information that is useful for forecasting Y, beyond that contained in past values of Y
  - See Stock & Watson p. 547

Granger causality

Autoregressive distributed lag (ADL) models

'a workhorse of the modern literature on time-series analysis' (Greene, 2003, p.579)

We refer to this as an ADL(1,1) one lag for each of X and Y

$$Y_t = \gamma_0 X_t + \gamma_1 X_{t-1} + \alpha_1 Y_{t-1} + u_t$$

we could have more variables and lags

more general specification is ADL(p,q).

This is defined by Stock and Watson(2007, p.544)

Granger causality

From ADL(1,1) to ADL(p,q)

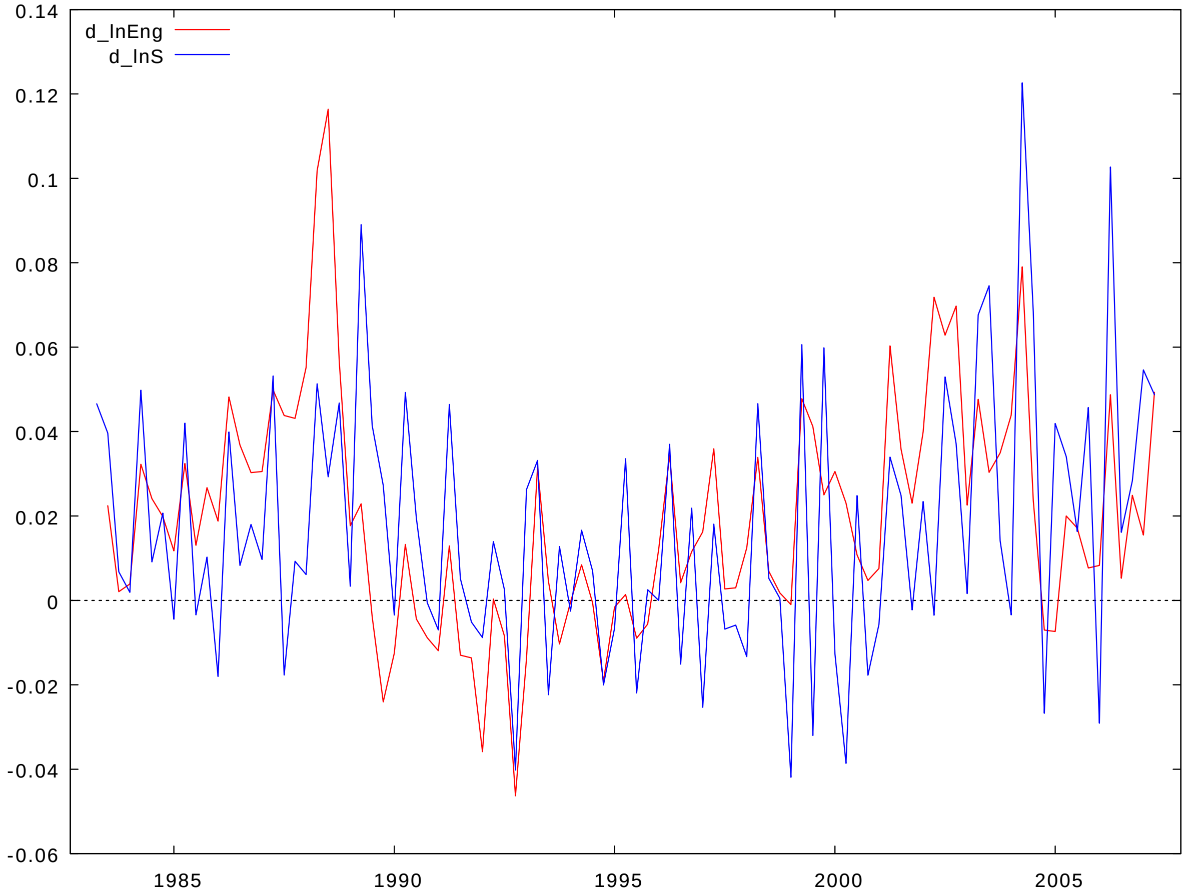
ADL(1,q)

$$Y_t = \alpha_1 Y_{t-1} + \gamma_0 X_t + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \ldots + \gamma_q X_{t-q} + u_t$$

ADL(p,q)

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_3 Y_{t-3} \ldots + \alpha_p Y_{t-p} + \gamma_0 X_t + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \ldots + \gamma_q X_{t-q} + u_t$$

Scottish and English house prices : growth



Granger causality

Does English house price growth Granger cause Scottish house price growth?

Use difference in logs (growth) which we assume are stationary

```
Model 7: OLS, using observations 1984:3-2007:2 (T = 92)
Dependent variable: d_lnS

              coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        0.00113920   0.00420682    0.2708   0.7872
  d_lnS_1     -0.425580     0.102801     -4.140    8.30e-05 ***
  d_lnS_2      0.0316127    0.112305      0.2815   0.7790
  d_lnS_3      0.369217     0.114033      3.238    0.0017   ***
  d_lnS_4      0.367178     0.112187      3.273    0.0016   ***
  d_lnEng_1    0.426773     0.160431      2.660    0.0094   ***
  d_lnEng_2    0.0444523    0.198087      0.2244   0.8230
  d_lnEng_3   -0.308842     0.194168     -1.591    0.1155
  d_lnEng_4    0.352070     0.162601      2.165    0.0332   **

Restriction set
 1: b[d_lnEng_1] = 0
 2: b[d_lnEng_2] = 0
 3: b[d_lnEng_3] = 0
 4: b[d_lnEng_4] = 0

Test statistic: F(4, 83) = 5.20779, with p-value = 0.000858687
```

Granger causality

Does Scottish house price growth Granger cause English house price growth?

```
Model 8: OLS, using observations 1984:3-2007:2 (T = 92)
Dependent variable: d_lnEng

               coefficient   std. error   t-ratio   p-value
  -----------------------------------------------------------
  const         0.00579451   0.00292438    1.981    0.0509   *
  d_lnEng_1     0.757613     0.111524      6.793    1.54e-09 ***
  d_lnEng_2     0.0769121    0.137700      0.5585   0.5780
  d_lnEng_3    -0.0872252    0.134976     -0.6462   0.5199
  d_lnEng_4     0.228962     0.113032      2.026    0.0460   **
  d_lnS_1      -0.375584     0.0714620    -5.256    1.12e-06 ***
  d_lnS_2      -0.181442     0.0780688    -2.324    0.0226   **
  d_lnS_3       0.0685379    0.0792702     0.8646   0.3897
  d_lnS_4       0.178741     0.0779867     2.292    0.0244   **

Restriction set
 1: b[d_lnS_1] = 0
 2: b[d_lnS_2] = 0
 3: b[d_lnS_3] = 0
 4: b[d_lnS_4] = 0

Test statistic: F(4, 83) = 8.06225, with p-value = 1.52969e-005
```

Prediction

- One of the main advantages of regression analysis is its ability to PREDICT or FORECAST
- Given a model, we can use the model to estimate the value of the dependent variable that would occur if the independent variable(s) takes a specific value
- Typically we would estimate a model and then use that model to predict the dependent variable for some points in time in the future
- However this assume that the relationship between dependent and independent variable(s) remains the same over the forecast period

Typically, with one independent variable we estimate a regression

$$\hat{Y}_{it} = \hat{b}_0 + \hat{b}_1 X_{it}$$

and use this estimated model to predict what Y will be for future X values

$$\hat{Y}_{it+1} = \hat{b}_0 + \hat{b}_1 X_{it+1}$$

and so on, forecasting to p periods ahead

$$\hat{Y}_{it+p} = \hat{b}_0 + \hat{b}_1 X_{it+p}$$

Prediction

- however, we need to take account of the inherent uncertainty in this prediction
  - as a result of the fact that b_0 and b_1 are estimates, not the true values

Prediction

- assume that we fit the regression
  - $Y_t = b_0 + b_1 time_t + e_t \quad t = 1,\ldots,T$
- with estimates of b_0 and b_1 it is possible to

predict Y_{T+1} given $time_{T+1}$

$$\hat{Y}_{T+1} = \hat{b}_0 + \hat{b}_1 time_{T+1}$$

$$\hat{Y}_{T+2} = \hat{b}_0 + \hat{b}_1 time_{T+2}$$

The $\hat{Y}$ ‘s are estimates of E(*Y*), the mean of *Y* at *T*+1, *T*+2 etc

Real UK consumers' expenditure

quarter	real UK	ln cex	DLcex
1	49.514	3.9023	*
2	48.386	3.8792	-0.0230
3	48.971	3.8912	0.0120
4	48.189	3.8751	-0.0161
5	48.710	3.8859	0.0108

q1 = 1980q1 q38 = 1989q2

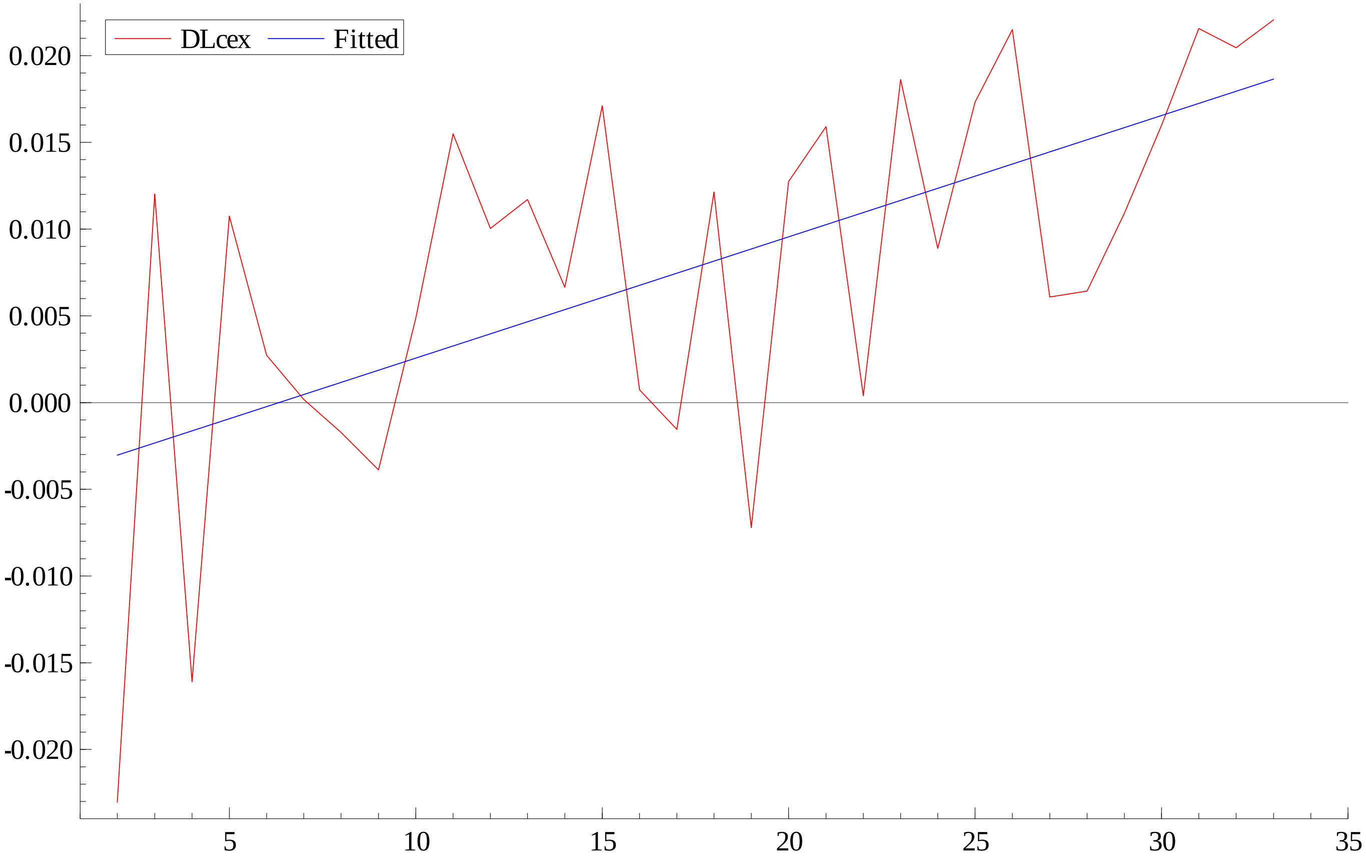
Model fitted to data for q2 to q33

```
EQ( 2) Modelling DLcex by OLS (using CEXPDI.xls)
       The estimation sample is: 2 to 33

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant          -0.00443574   0.003287    -1.35   0.187   0.0572
Trend             0.000699500  0.0001661     4.21   0.000   0.3715

sigma              0.00867655  RSS              0.00225847624
R^2                  0.371473  F(1,30) =     17.73 [0.000]**
log-likelihood        107.535  DW                        2.41
no. of observations        32  no. of parameters            2
```

Actual and fitted growth rates, q2 to q33



Predicted growth q34 to q38
With 95% confidence interval (without acknowledging parameter uncertainty)

1-step forecasts for DLcex (SE based on error variance only)

Horizon	Forecast	SE	Actual	Error	t-value
34	0.0193472	0.008677	0.00169766	-0.0176496	-2.034
35	0.0200467	0.008677	0.0162947	-0.00375203	-0.432
36	0.0207462	0.008677	0.0191307	-0.00161555	-0.186
37	0.0214457	0.008677	0.00168125	-0.0197645	-2.278
38	0.0221452	0.008677	0.0156462	-0.00649904	-0.749

Prediction

- however, we need to take account of the inherent uncertainty in this prediction
  - a result of the fact that b_0 and b_1 are estimated

$$\hat{Y}_{T+1} = \hat{b}_0 + \hat{b}_1 Time_{T+1}$$

$$Y_{T+1} = b_0 + b_1 Time_{T+1} + e_{T+1} \qquad e \sim N(0, \sigma^2)$$

$$Y_{T+1} - \hat{Y}_{T+1} = (b_0 - \hat{b}_0) + (b_1 - \hat{b}_1) Time_{T+1} + e_{T+1}$$

$$\text{var}(Z) = E\left[(Z - E(Z))^2 \right]$$

$$\text{var}(Z) = E\left[(Z - 0)^2 \right] = E\left[Z^2 \right]$$

$$\text{if } Z = Y_{T+1} - \hat{Y}_{T+1} \text{ and } E(Y_{T+1} - \hat{Y}_{T+1}) = 0$$

$$\text{MeanSquareForecastError} = E\left[(Y_{T+1} - \hat{Y}_{T+1})^2 \right] = \text{var}(Y_{T+1} - \hat{Y}_{T+1})$$

$$\text{MSFE} = \sigma^2 + \text{var}\left[(b_0 - \hat{b}_0) + (b_1 - \hat{b}_1) Time_{T+1} \right]$$

$$\text{RMSFE} = \sqrt{\text{MSFE}}$$

The 95% confidence interval for E(Y)

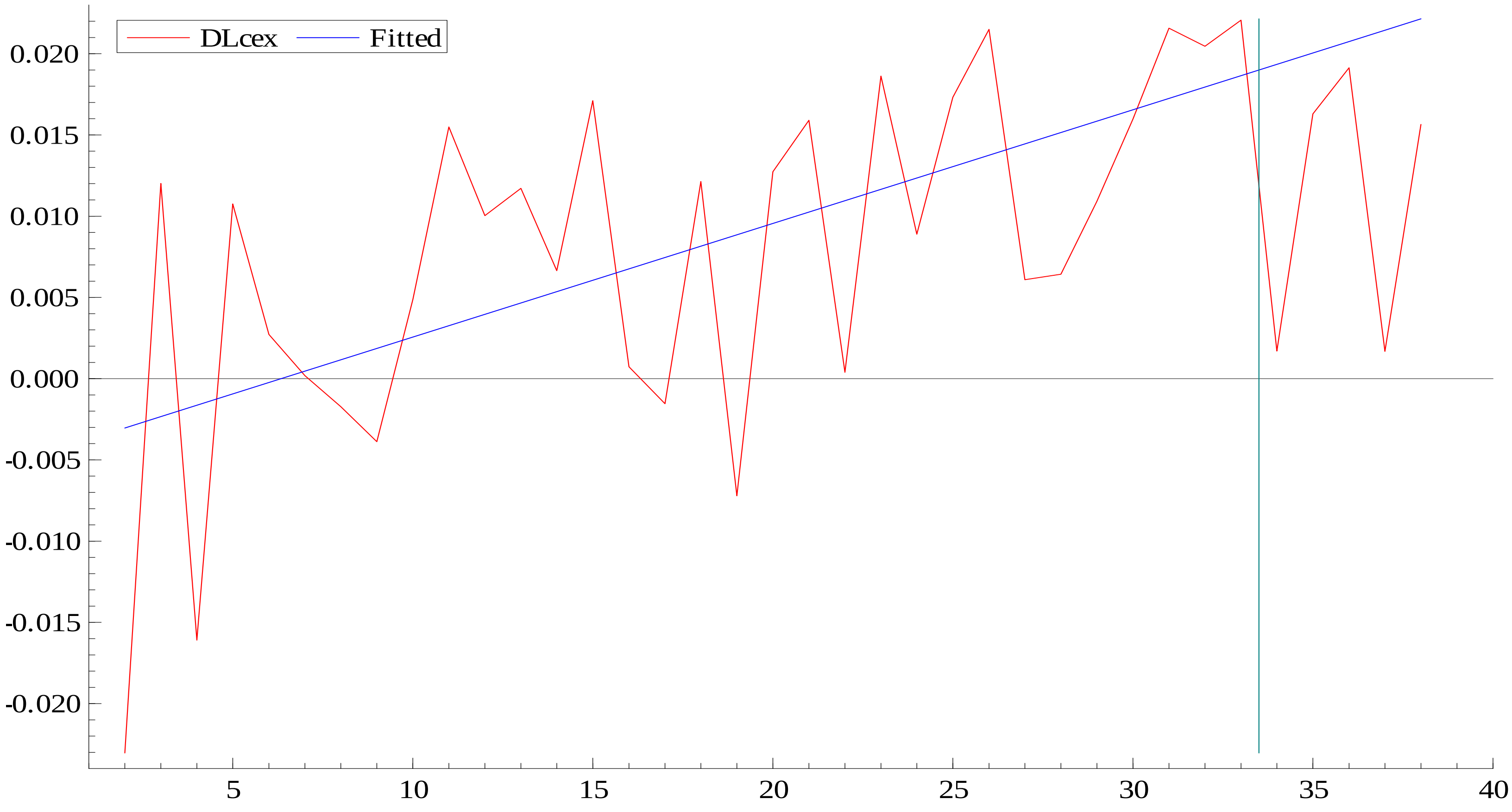
Assume the forecast errors are normally distributed, then

$$c.i. = 95\% \text{ confidence interval}$$

$$c.i. = \hat{Y}_{T+1} \pm 1.96\, s.e.(Y_{T+1} - \hat{Y}_{T+1})$$

$$c.i. \approx \hat{Y}_{T+1} \pm 1.96\text{RMSFE}$$

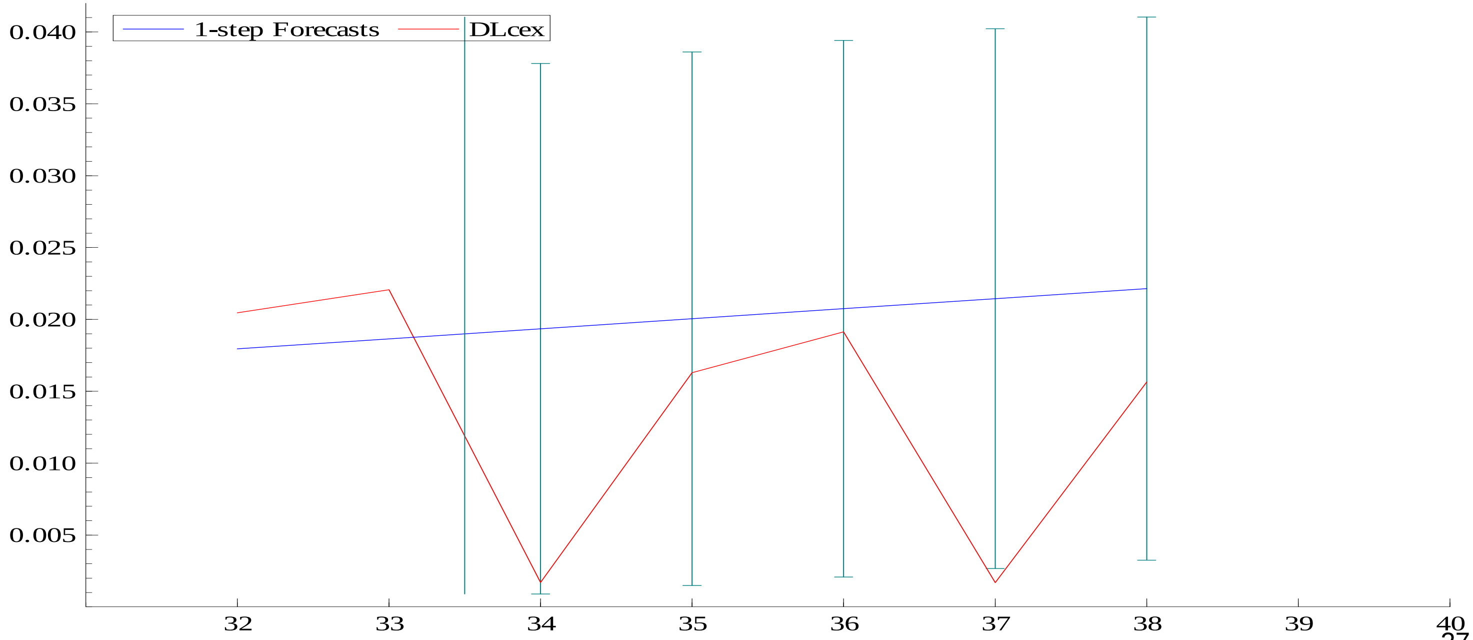
Predicted growth q34 to q38



Predicted growth q34 to q38
With 95% confidence interval (with parameter uncertainty)

Given *time* = 34...38 we predict
the growth of consumers' expenditure

We place a 95% confidence interval around these estimates
just as for the ordinary sample mean



Predicted growth q34 to q38
With 95% confidence interval (with and without parameter uncertainty)

1-step forecasts for DLcex (SE based on error variance only)

Horizon	Forecast	SE	Actual	Error	t-value
34	0.0193472	0.008677	0.00169766	-0.0176496	-2.034
35	0.0200467	0.008677	0.0162947	-0.00375203	-0.432
36	0.0207462	0.008677	0.0191307	-0.00161555	-0.186
37	0.0214457	0.008677	0.00168125	-0.0197645	-2.278
38	0.0221452	0.008677	0.0156462	-0.00649904	-0.749

1-step forecasts for DLcex (SE with parameter uncertainty)

Horizon	Forecast	SE	Actual	Error	t-value
34	0.0193472	0.009228	0.00169766	-0.0176496	-1.913
35	0.0200467	0.009278	0.0162947	-0.00375203	-0.404
36	0.0207462	0.009332	0.0191307	-0.00161555	-0.173
37	0.0214457	0.009388	0.00168125	-0.0197645	-2.105
38	0.0221452	0.009446	0.0156462	-0.00649904	-0.688

Multiple regression

- it is a possible (using statistical software) to find the c.i. in MULTIPLE regression (ie with known X_1, X_2 etc)
  - one has to specify the actual values of X_1, X_2 etc for which one requires estimated E(Y) and the c.i.

Summary

- beware of predicting *Y* beyond the domain of the variables
- always accompany any prediction of E(*Y*) with its (say 95%) confidence interval

Prediction

- Notice that we are making several big assumptions
- The regression parameters, and the error variance, estimated for the period 1….*T* remain constant over the period *T*+1….*T*+*p*
- The functional form (a straight line) remains the same
- We know the values of *X* over the period *T*+1….*T*+*p*
- If we do, then we are making what is known as an ex post prediction
- If we do not, then we are making an ex ante prediction. This requires us to first predict *X* in order to predict *Y*
- Needless to say, ex post prediction is much safer than ex ante prediction