Econometric Data Science

Francis X. Diebold
University of Pennsylvania

October 22, 2019

The painting is *Enigma*, by Glen Josselsohn, from Wikimedia Commons.

Introduction

Numerous Communities Use Econometrics

Economists, statisticians, analysts, "data scientists" in:

- Finance (Commercial banking, retail banking, investment banking, insurance, asset management, real estate, ...)
- Traditional Industry (manufacturing, services, advertising, brick-and-mortar retailing, ...)
- e-Industry (Google, Amazon, eBay, Uber, Microsoft, ...)
- Consulting (financial services, litigation support, ...)
- Government (treasury, agriculture, environment, commerce, ...)
- Central Banks and International Organizations (FED, IMF, World Bank, OECD, BIS, ECB, ...)

Econometrics is Special

Econometrics is not just "statistics using economic data". Many properties and nuances of economic data require knowledge of economics for sucessful analysis.

- Emphasis on predictions, guiding decisions
- Observational data
- Structural change
- Volatility fluctuations ("heteroskedasticity")
- Even trickier in time series: Trend, Seasonality, Cycles ("serial correlation")

Let's Elaborate on the "Emphasis on Predictions Guiding Decisions"...

Q: What is econometrics about, broadly?

A: Helping people to make better decisions

- Consumers
- Firms
- Investors
- Policy makers
- Courts

Forecasts guide decisions.

Good forecasts promote good decisions.

Hence prediction holds a distinguished place in econometrics, and it will hold a distinguished place in this course.

Types/Arrangements of Economic Data

– Cross section

Standard cross-section notation: $i = 1, ..., N$

– Time series

Standard time-series notation: $t = 1, ..., T$

Much of our discussion will apply to *both* cross-section and time-series environments, but still we have to pick a notation.

A Few Leading Econometrics Web Data Resources (Clickable)

Indispensible:

- Resources for Economists (AEA)
- FRED (Federal Reserve Economic Data)

More specialized:

- National Bureau of Economic Research
- FRB Phila Real-Time Data Research Center
- Many more

A Few Leading Econometrics Software Environments (Clickable)

- High-Level: EViews, Stata
- Mid-Level: R (also CRAN; RStudio; R-bloggers), Python (also Anaconda), Julia
- Low-Level: C, C++, Fortran

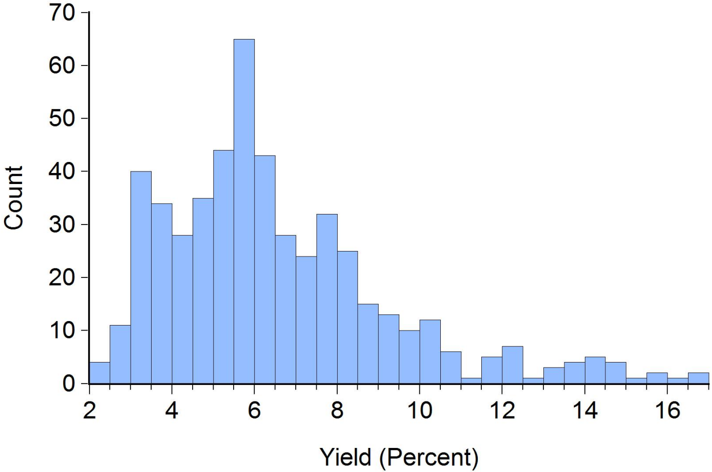
"High-level" does not mean "best", and "low-level" does not mean worst. There are many issues.

Graphics Review

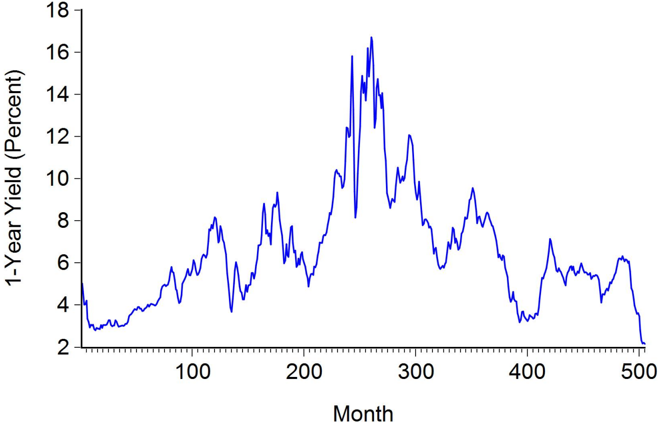
Graphics Help us to:

- Summarize and reveal patterns in univariate cross-section data. Histograms and density estimates are helpful for learning about distributional shape. Symmetric, skewed, fat-tailed, ...
- Summarize and reveal patterns in univariate time-series data. Time Series plots are useful for learning about dynamics. Trend, seasonal, cycle, outliers, ...
- Summarize and reveal patterns in multivariate data (cross-section or time-series). Scatterplots are useful for learning about relationships. Does a relationship exist? Is it linear or nonlinear? Are there outliers?

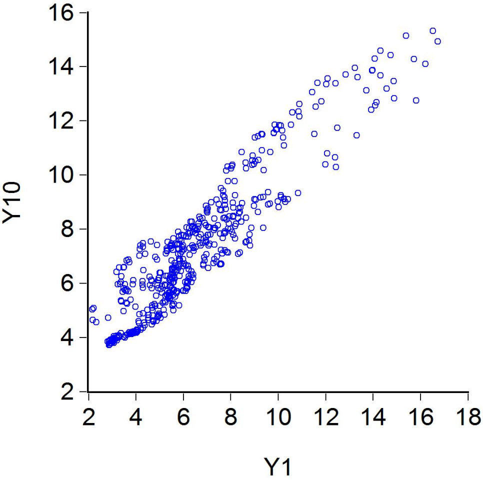
Histogram Revealing Distributional Shape: 1-Year Government Bond Yield



Time Series Plot Revealing Dynamics: 1-Year Goverment Bond Yield



Scatterplot Revealing Relationship: 1-Year and 10-Year Government Bond Yields



Some Principles of Graphical Style

- Know your audience, and know your goals.
- Appeal to the viewer.
- Show the data, and only the data, withing the bounds of reason.
  - Avoid distortion. The sizes of effects in graphics should match their size in the data. Use common scales in multiple comparisons.
  - Minimize, within reason, non-data ink. Avoid chartjunk.
  - Third, choose aspect ratios to maximize pattern revelation. Bank to 45 degrees.
  - Maximize graphical data density.
- Revise and edit, again and again (and again). Graphics produced using software defaults are almost *never* satisfactory.

Probability and Statistics Review

Moments, Sample Moments and Their Sampling Distributions

- Discrete random variable, y
- Discrete probability distribution $p(y)$
- Continuous random variable y
- Probability density function $f(y)$

Population Moments: Expectations of Powers of R.V.'s

Mean measures location:

$$\mu = E(y) = \sum_i p_i y_i \ \ \text{(discrete case)}$$

$$\mu = E(y) = \int y\, f(y)\, dy \ \ \text{(continuous case)}$$

Variance, or standard deviation, measures dispersion, or scale:

$$\sigma^2 = var(y) = E(y-\mu)^2.$$

– σ easier to interpret than σ^2. Why?

More Population Moments

Skewness measures skewness (!)

$$S = \frac{E(y-\mu)^3}{\sigma^3}.$$

Kurtosis measures tail fatness relative to a Gaussian distribution.

$$K = \frac{E(y-\mu)^4}{\sigma^4}.$$

Covariance and Correlation

Multivariate case: Joint, marginal and conditional distributions
$f(x,y)$, $f(x)$, $f(y)$, $f(x|y)$, $f(y|x)$

Covariance measures linear dependence:

$$cov(y,x) = E[(y-\mu_y)(x-\mu_x)].$$

So does correlation:

$$corr(y,x) = \frac{cov(y,x)}{\sigma_y \sigma_x}.$$

Correlation is often more convenient. Why?

Sampling and Estimation

$$\text{Sample}: \ \{y_i\}_{i=1}^{N} \ \sim \ iid \ f(y)$$

Sample mean:

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Sample variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N}$$

Unbiased sample variance:

$$s^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1}$$

More Sample Moments

Sample skewness:

$$\hat{S} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^3}{\hat{\sigma}^3}$$

Sample kurtosis:

$$\hat{K} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^4}{\hat{\sigma}^4}$$

Still More Sample Moments

Sample covariance:

$$\widehat{cov}(y, x) = \frac{1}{N} \sum_{i=1}^{N} [(y_i - \bar{y})(x_i - \bar{x})]$$

Sample correlation:

$$\widehat{corr}(y, x) = \frac{\widehat{cov}(y, x)}{\hat{\sigma}_y \hat{\sigma}_x}$$

Exact Finite-Sample Distribution of the Sample Mean (Requires *iid* Normality)

$$\text{Simple random sampling}: y_i \sim iid\ N(\mu, \sigma^2), i = 1, ..., N$$

$\bar{y}$ is unbiased and normally distributed with variance σ^2/N.

$$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{N}\right),$$

and we estimate σ^2 using s^2, where

$$s^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}.$$

$$\mu \in \left[\bar{y} \pm t_{1-\frac{\alpha}{2}}(N-1)\frac{s}{\sqrt{N}}\right]\ w.p.\ 1-\alpha$$

$$\mu = \mu_0 \implies \frac{\bar{y}-\mu_0}{\frac{s}{\sqrt{N}}} \sim t_{1-\frac{\alpha}{2}}(N-1)$$

where "$t_{1-\frac{\alpha}{2}}(N-1)$" denotes the appropriate critical value of the Student's t density with $N-1$ degrees of freedom

Large-Sample Distribution of the Sample Mean (Requires *iid*, but not Normality)

$$\text{Simple random sampling} : y_i \sim iid\,(\mu, \sigma^2), i = 1, ..., N$$

$\bar{y}$ is consistent and asymptotically normally distributed with variance v.

$$\bar{y} \overset{a}{\sim} N(\mu, v),$$

and we estimate v using $\hat{v} = s^2/N$, where

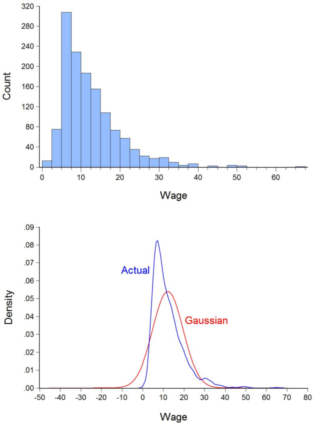
$$s^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}.$$

This is an approximate (large-sample) result, due to the central limit theorem. The "a" is for "asymptotically", which means "as $N \to \infty$".

$$\text{As } N \to \infty,\ \mu \in \left[\bar{y} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{N}}\right] \ w.p.\ 1 - \alpha$$

$$\text{As } N \to \infty,\ \mu = \mu_0 \implies \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{N}}} \sim N(0, 1)$$

Wages: Distributions



Wages: Sample Statistics

	WAGE	log WAGE
Sample Mean	12.19	2.34
Sample Median	10.00	2.30
Sample Maximum	65.00	4.17
Sample Minimum	1.43	0.36
Sample Std. Dev.	7.38	0.56
Sample Skewness	1.76	0.06
Sample Kurtosis	7.93	2.90
Jarque-Bera	2027.86	1.26
	$(p = 0.00)$	$(p = 0.53)$
$\mathrm{t}(H_0 : \mu = 12)$	0.93	-625.70
	$(p = 0.36)$	$(p = 0.00)$

Regression

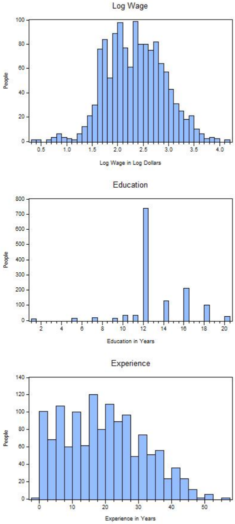
Regression

A. As curve fitting. "Tell a computer how to draw a line through a scatterplot". (Well, sure, but there must be more...)

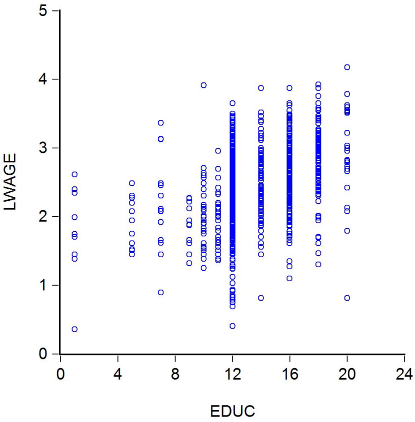
B. As a probabilistic framework for optimal prediction.

Regression as Curve Fitting

Distributions of Log Wage, Education and Experience



Scatterplot: Log Wage vs. Education



Curve Fitting

Fit a line:

$$y_i = \beta_1 + \beta_2 x_i$$

Solve:

$$\min_{\beta_1,\beta_2} \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

“least squares” (LS, or OLS)

“quadratic loss”

Actual Values, Fitted Values and Residuals

Actual values: $y_i, \; i = 1, ..., N$

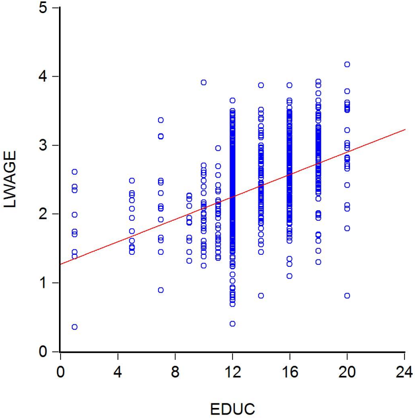
Least-squares fitted parameters: $\hat{\beta}_1$ and $\hat{\beta}_2$

Fitted values: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i, \; i = 1, ..., N,$

("hats" denote fitted things...)

Residuals: $e_i = y_i - \hat{y}_i, \; i = 1, ..., N.$

Log Wage vs. Education with Superimposed Regression Line



$$\widehat{LWAGE} = 1.273 + .081 EDUC$$

Multiple Linear Regression (K RHS Variables)

Solve:

$$\min_{\beta_1,\ldots,\beta_K} \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_{i2} - \ldots - \beta_K x_{iK})^2$$

Fitted hyperplane:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK}$$

More compactly:

$$\hat{y}_i = \sum_{k=1}^{K} \hat{\beta}_k x_{ik},$$

where $x_{i1} = 1$ for all i.

Wage dataset:

$$\widehat{LWAGE} = .867 + .093 EDUC + .013 EXPER$$

Regression as a Probability Model

An Ideal Situation ("The Ideal Conditions", or IC)

1. The data-generating process (DGP) is:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i$$
$$\varepsilon_i \sim iidN(0, \sigma^2)$$
$$i = 1, ..., N,$$

   and the fitted model matches it exactly.
   1.1 The fitted model is correctly specified
   1.2 The disturbances are Gaussian
   1.3 The coefficients (β_k's) are fixed
   1.4 The relationship is linear
   1.5 The ε_i's have constant unconditional variance σ^2
   1.6 The ε_i's are uncorrelated
2. ε_i is independent of $(x_{i1}, ..., x_{iK})$, for all i
   2.1 $E(\varepsilon_i \mid x_{i1}, ..., x_{iK}) = 0$, for all i
   2.2 $var(\varepsilon_i \mid x_{i1}, ..., x_{iK}) = \sigma^2$, for all i

(Written here for cross sections. Slight changes in 2.1, 2.2 for time series.)

Some Concise Matrix Notation (Useful for Notation, Estimation, Inference)

You already understand matrix ("spreadsheet") notation, although you may not know it.

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{12} & x_{13} & \dots & x_{1K} \\ 1 & x_{22} & x_{23} & \dots & x_{2K} \\ \vdots & & & & \\ 1 & x_{N2} & x_{N3} & \dots & x_{NK} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

Elementary Matrices and Matrix Operations

$$0 = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix} \qquad I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Transposition: $A'_{ij} = A_{ji}$

Addition: For A and B $n \times m$, $(A+B)_{ij} = A_{ij} + B_{ij}$

Multiplication: For A $n \times m$ and B $m \times p$, $(AB)_{ij} = \sum_{k=1}^{m} A_{ik} B_{kj}$.

Inversion: For non-singular A $n \times n$, A^{-1} satisfies $A^{-1}A = AA^{-1} = I$. Many algorithms exist for calculation.

The DGP in Matrix Form, Written Out

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & x_{13} & \dots & x_{1K} \\ 1 & x_{22} & x_{23} & \dots & x_{2K} \\ \vdots & & & & \\ 1 & x_{N2} & x_{N3} & \dots & x_{NK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} \right)$$

$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

Three Notations

Original form:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i, \quad \varepsilon_i \sim iidN(0, \sigma^2)$$

$$i = 1, 2, ..., N$$

Intermediate form:

$$y_i = x_i' \beta + \varepsilon_i, \quad \varepsilon_i \sim iidN(0, \sigma^2)$$

$$i = 1, 2, ..., N$$

Full matrix form:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

Ideal Conditions Redux

We used to write this: The DGP is

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i, \quad \varepsilon_i \sim iidN(0, \sigma^2),$$

and the fitted model matches it exactly, and

$$\varepsilon_i \text{ is independent of } (x_{i1}, ..., x_{iK}), \text{ for all } i$$

Now, equivalently, we write this: The DGP is

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

and the fitted model matches it exactly, and

$$\varepsilon_i \text{ is independent of } x_i, \text{ for all } i$$

The OLS Estimator in Matrix Notation

As before, the LS estimator solves:

$$\min_{\beta_1,\ldots,\beta_K} \left(\sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_{i2} - \ldots - \beta_K x_{iK})^2 \right)$$

Now, in matrix notation:

$$\min_{\beta} \left((y - X\beta)'(y - X\beta) \right)$$

It can be shown that the solution is:

$$\hat{\beta}_{LS} = (X'X)^{-1} X'y$$

Large-Sample Distribution of $\hat{\beta}_{LS}$ Under the IC

$\hat{\beta}_{LS}$ is consistent and asymptotically normally distributed with covariance matrix V,

$$\hat{\beta}_{LS} \overset{a}{\sim} N(\beta,\ V),$$

we estimate V using $\hat{V} = s^2(X'X)^{-1}$, where

$$s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N-K}.$$

Note the precise parallel with the large-sample distribution of the sample mean.

Sample Mean, Regression on an Intercept, and Properties of Residuals

– Sample mean is just LS regression on nothing but an intercept. (Why?)

– Intercept picks up a "level effect"

– Regression generalizes the sample mean to include predictors other than just a constant

– If an intercept is included in a regression, the residuals must sum to 0 (Why?)

Conditional Moment Implications of the IC

Conditional mean:

$$E(y_i \mid x_i{=}x^*) \;=\; x^{*\prime}\beta$$

Conditional variance:

$$var(y_i \mid x_i{=}x^*) \;=\; \sigma^2$$

Full conditional density:

$$y_i \mid x_i{=}x^* \;\sim\; N(x^{*\prime}\beta,\; \sigma^2)$$

Why All the Talk About Conditional Moment Implications?

"Point Prediction"

A major goal in econometrics is predicting y. The question is "If a new person i arrives with characteristics $x_i = x^*$, what is my best prediction of her y_i? The answer is $E(y_i \mid x_i = x^*) = x^{*\prime}\beta$.

"The conditional mean is the minimum MSE point predictor"

Non-operational version (remember, in reality we don't know β):
$E(y_i \mid x_i = x^*) = x^{*\prime}\beta$

Operational version (use $\hat{\beta}_{LS}$):
$\widehat{E(y_i \mid x_i = x^*)} = x^{*\prime}\hat{\beta}_{LS}$ (regression fitted value at $x_i = x^*$)

– LS delivers operational optimal predictor with great generality

– Follows immediately from the LS optimization problem

“Interval Prediction”

Non-operational (in reality we don't know β or σ):

$$y_i \in [x^{*\prime}\beta \pm 1.96\,\sigma] \quad w.p.\ 0.95$$

Operational:

$$y_i \in [x^{*\prime}\hat{\beta}_{LS} \pm 1.96\,s] \quad w.p.\ 0.95$$

(Notice that, as is common, this operational interval forecast ignores parameter estimation uncertainty, or equivalently, assumes a large sample, so that that the interval is based on the standard normal distribution rather than Student's t.)

"Density Prediction"

Non-operational version:

$$y_i \mid x_i = x^* \ \sim \ N(x^{*\prime}\beta, \ \sigma^2)$$

Operational version:

$$y_i \mid x_i = x^* \ \sim \ N(x^{*\prime}\hat{\beta}_{LS}, \ s^2)$$

(This operational density forecast also ignores parameter estimation uncertainty, or equivalently, assumes a large sample, as will all of our interval and density forecasts moving forward.)

“Typical” Regression Analysis of Wages, Education and Experience

Equation: UNTITLED Workfile: GRAPHS::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date: 06/27/13 Time: 16:38
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.867382	0.075331	11.51431	0.0000
EDUC	0.093229	0.005045	18.48002	0.0000
EXPER	0.013104	0.001164	11.26208	0.0000

R-squared	0.232224	Mean dependent var	2.341995
Adjusted R-squared	0.231061	S.D. dependent var	0.561435
S.E. of regression	0.492318	Akaike info criterion	1.422881
Sum squared resid	319.9376	Schwarz criterion	1.434644
Log likelihood	-938.2358	Hannan-Quinn criter.	1.427291
F-statistic	199.6260	Durbin-Watson stat	1.926045
Prob(F-statistic)	0.000000		

EXPER

“Top Matter”: Background Information

- Dependent variable
- Method
- Date
- Sample
- Included observations

"Middle Matter": Estimated Regression Function

- Variable
- Coefficient – appropriate element of $(X'X)^{-1}X'y$
- Standard error – appropriate diagonal element of $\sqrt{s^2(X'X)^{-1}}$
- t-statistic – coefficient divided by standard error
- p-value

Predictive Perspectives

– OLS coefficient signs and sizes give the weights put on the various x variables in forming the best in-sample prediction of y.

– The standard errors, t statistics, and p-values let us do statistical inference as to which regressors are most relevant for predicting y.

“Bottom Matter: Statistics”

There are many...

Regression Statistics: Mean dependent var 2.342

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Predictive Perspectives

The sample, or historical, mean of the dependent variable, $\bar{y}$, an estimate of the *unconditional* mean of y, is a naive benchmark forecast. It is obtained by regressing y on an intercept alone – no conditioning on other regressors.

Regression Statistics: S.D. dependent var .561

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}}$$

Predictive Perspectives

– The sample standard deviation of y is a measure of the in-sample accuracy of the unconditional mean forecast $\bar{y}$.

Regression Statistics: Sum squared resid 319.938

$$SSR = \sum_{i=1}^{N} e_i^2$$

– Optimized value of the LS objective; will appear in many places.

Predictive Perspectives

– The OLS fitted values, $\hat{y}_i = x_i'\hat{\beta}$, are effectively in-sample regression predictions.

– The OLS residuals, $e_i = y_i - \hat{y}_i$, are effectively in-sample prediction errors corresponding to use of the regression predictions.

SSR measures "total" in-sample predictive accuracy

"squared-error loss"

"quadratic loss"

SSR is closely related to in-sample *MSE*:

$$MSE = \frac{1}{N} SSR = \frac{1}{N} \sum_{i=1}^{N} e_i^2$$

("average" in-sample predictive accuracy)

Regression Statistics: F-statistic 199.626

$$F = \frac{(SSR_{res} - SSR)/(K-1)}{SSR/(N-K)}$$

Predictive Perspectives

– The F statistic effectively compares the accuracy of the regression-based forecast to that of the unconditional-mean forecast.

– Helps us assess whether the x variables, taken as a set, have predictive value for y.

– Contrasts with the t statistics, which assess predictive value of the x variables one at a time.

Regression Statistics: S.E. of regression .492

$$s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N - K}$$

$$SER = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N - K}}$$

Predictive Perspectives

s^2 is just SSR scaled by $N - K$, so again, it's a measure of the in-sample accuracy of the regression-based forecast.

Like MSE, but corrected for degrees of freedom.

Regression Statistics:
R-squared .232, Adjusted R-squared .231

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N} e_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

"What percent of variation in y is explained by variation in x?"

Predictive Perspectives

R^2 and $\bar{R}^2$ effectively compare the in-sample accuracy of conditional-mean and unconditional-mean forecasts.

R^2 is not corrected for d.f. and has *MSE* on top:

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} e_i^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$\bar{R}^2$ is corrected for d.f. and has s^2 on top:

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N} e_i^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

R_k^2 and "Multicollinearity" (not shown in the computer output)

Perfect multicollinearity (Big problem for LS!):
One x a perfect linear combination of others. $X'X$ singular.

Imperfect multicollinearity (Not a big problem for LS):
One x correlated with a linear combination of others.

We often measure the strength of multicollinearity by "R_k^2", the R^2 from a regression of x_k on all other regressors.

It can be shown that:

$$var(\hat{\beta}_k) = f\left(\underbrace{\sigma^2}_{+},\ \underbrace{\sigma^2_{x_k}}_{-},\ \underbrace{R_k^2}_{+}\right)$$

Predictive Perspectives

– Multtollinearity makes it hard to identify the contributions of the individual x's to the overall predictive relationship.
(Low t-stats)

– But we still might see evidence of a strong overall predictive relationship.
(High F-stat)

Regression Statistics: Log likelihood -938.236

Understanding this requires some background / detail:

- Likelihood – joint density of the data (the y_i's)
- Maximum-likelihood estimation – natural estimation strategy: find the parameter configuration that maximizes the likelihood of getting the y_i's that you actually *did* get.
- Log likelihood – will have same max as the likelihood (why?) but it's more important statistically
- Hypothesis tests are based on log likelihood

Detail: Maximum-Likelihood Estimation

Linear regression DGP (under the IC) implies that:

$$y_i|x_i \sim iidN(x_i'\beta, \sigma^2),$$

so that

$$f(y_i|x_i) = (2\pi\sigma^2)^{\frac{-1}{2}} e^{\frac{-1}{2\sigma^2}(y_i - x_i'\beta)^2}$$

Now by independence of the ε_i's and hence y_i's,

$$L = f(y_1, ..., y_N|x_i) = f(y_1|x_1) \cdots f(y_N|x_N) = \prod_{i=1}^{N} (2\pi\sigma^2)^{\frac{-1}{2}} e^{\frac{-1}{2\sigma^2}(y_i - x_i'\beta)^2}$$

Detail: Log Likelihood

$$\ln L = ln\left((2\pi\sigma^2)^{\frac{-N}{2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\beta)^2$$

$$= \frac{-N}{2} ln(2\pi) - \frac{N}{2} ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\beta)^2$$

– Log turns the product into a sum and eliminates the exponential

– The β vector that maximizes the likelihood is the β vector that minimizes the sum of squared residuals

– Additive constant $\frac{-N}{2} ln(2\pi)$ can be dropped

– "MLE and OLS coincide for linear regression under the IC" (Normality, in particular, is crucial)

Detail: Likelihood-Ratio Tests

It can be shown that, under the null hypothesis (that is, if the restrictions imposed under the null are true):

$$-2(\ln L_0 - \ln L_1) \overset{a}{\sim} \chi^2_d,$$

where $\ln L_0$ is the maximized log likelihood under the restrictions imposed by the null hypothesis, $\ln L_1$ is the unrestricted log likelihood, and d is the number of restrictions imposed under the null hypothesis.

– t and F tests are likelihood ratio tests under a normality assumption, which of course is part of the IC. That's why they can be written in terms of minimized SSR's in addition to maximized $\ln L$'s.

Predictive Perspectives

- Gaussian L is intimately related to SSR
- Therefore L is closely related to prediction (and measuring predictive accuracy) as well
- Small $SSR \iff$ large L

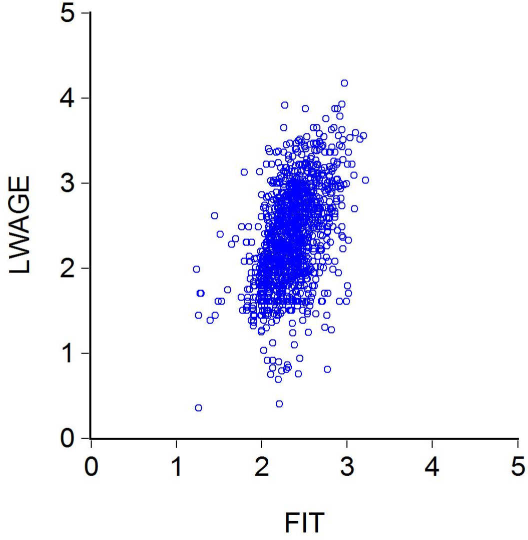
Regression Statistics: Schwarz criterion 1.435 Akaike info criterion 1.422

We'll get there shortly...

Regression Statistics: Durbin-Watson stat. 1.926

We'll get there in six weeks...

Residual Scatter



Residual Plot

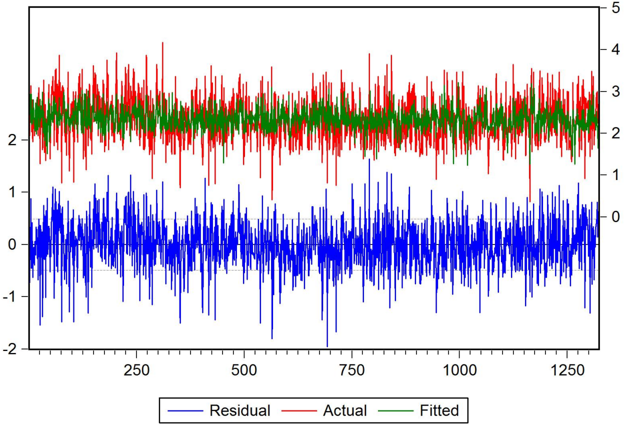


Figure: Wage Regression Residual Plot

Predictive Perspectives

– The LS fitted values, $\hat{y}_i = x_i'\hat{\beta}$, are effectively best in-sample predictions.

– The LS residuals, $e_i = y_i - \hat{y}_i$, are effectively in-sample prediction errors corresponding to use of the best predictor.

– Residual plots are useful for visually flagging violations of the IC that can impact forecasting.

For example:

1. The true DGP may be nonlinear
2. ε may be non-Gaussian
3. ε may have non-constant variance

Misspecification and Model Selection

Do we really believe that the fitted model matches the DGP?

Regression Statistics:
Akaike info criterion 1.422, Schwarz criterion 1.435

SSR versions:

$$AIC = e^{\frac{2K}{N}} \frac{\sum_{i=1}^{N} e_i^2}{N}$$

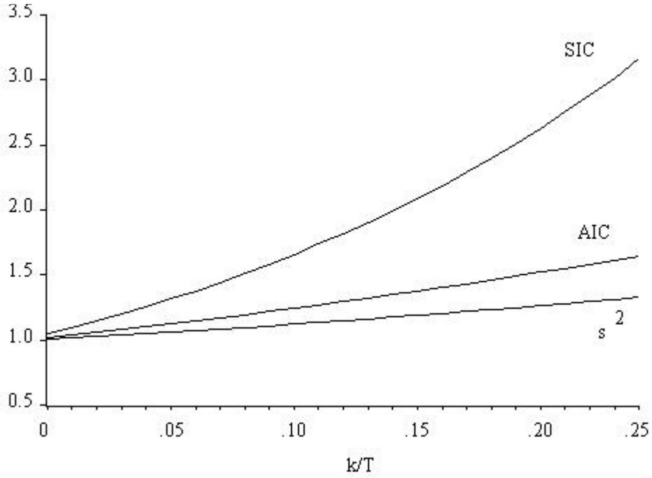
$$SIC = N^{\left(\frac{K}{N}\right)} \frac{\sum_{i=1}^{N} e_i^2}{N}$$

More general *lnL* versions:

$$AIC = -2lnL + 2K$$

$$SIC = -2lnL + KlnN$$

Penalties



Predictive Perspectives

– Estimate *out-of-sample* forecast accuracy (which is what we really care about) on the basis of in-sample forecast accuracy. (We want to select a forecasting model that will perform well for out-of-sample forecasting, quite apart from its in-sample fit.)

– We proceed by inflating the in-sample mean-squared error (MSE), in various attempts to offset the deflation from regression fitting, to obtain a good estimate of out-of-sample MSE.

$$MSE = \frac{\sum_{i=1}^{N} e_i^2}{N}$$

$$s^2 = \left(\frac{N}{N-K}\right) MSE$$

$$SIC = \left(N^{\left(\frac{K}{N}\right)}\right) MSE$$

"Oracle property"

Non-Normality and Outliers

Do we really believe that the disturbances are Gaussian?

What We'll Do

– Problems caused by non-normality and outliers
(Large sample estimation results don't change,
LS results can be distorted or fragile, and
density prediction changes)

– Detecting non-normality, outliers, and influential observations
(JB test, residual histogram, residual QQ plot,
residual plot and scatterplot, leave-one-out plot, ...)

– Dealing with non-normality, outliers, and influential observations
(LAD regression, simulation-based density forecasts, ...)

Large-Sample Distribution of $\hat{\beta}_{LS}$ Under the Ideal Conditions (Except Normality)

$\hat{\beta}_{LS}$ is consistent and asymptotically normally distributed with covariance matrix V,

$$\hat{\beta}_{LS} \overset{a}{\sim} N(\beta,\ V),$$

and we estimate V using $\hat{V} = s^2(X'X)^{-1}$, where

$$s^2 = \frac{\sum_{i=1}^{N} e_i^2}{N - K}$$

No change from asymptotic result under IC!

So why worry about normality?

– Non-normality and resulting outliers can distort finite-sample estimates

– Interval and density prediction change fundamentally

Jarque-Bera Normality Test

- Sample skewness and kurtosis, $\hat{S}$ and $\hat{K}$

- Jarque-Bera test. Under normality we have:

$$JB = \frac{N}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{K}-3)^2\right) \sim \chi^2_2$$

- Many more

Recall Our OLS Wage Regression

Equation: UNTITLED Workfile: GRAPHS::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date: 06/27/13 Time: 16:38
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

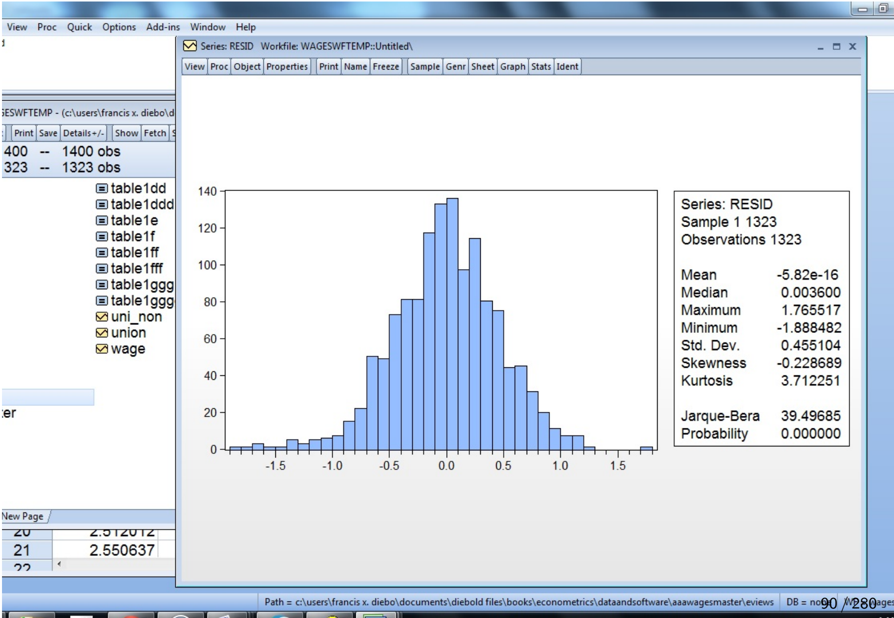
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.867382	0.075331	11.51431	0.0000
EDUC	0.093229	0.005045	18.48002	0.0000
EXPER	0.013104	0.001164	11.26208	0.0000

R-squared	0.232224	Mean dependent var	2.341995
Adjusted R-squared	0.231061	S.D. dependent var	0.561435
S.E. of regression	0.492318	Akaike info criterion	1.422881
Sum squared resid	319.9376	Schwarz criterion	1.434644
Log likelihood	-938.2358	Hannan-Quinn criter.	1.427291
F-statistic	199.6260	Durbin-Watson stat	1.926045
Prob(F-statistic)	0.000000		

EXPER

Path = c:\users\francis x. diebo\documents\diebold files\courses\econ104\old\econ104_2011\sw3e\eviews materials

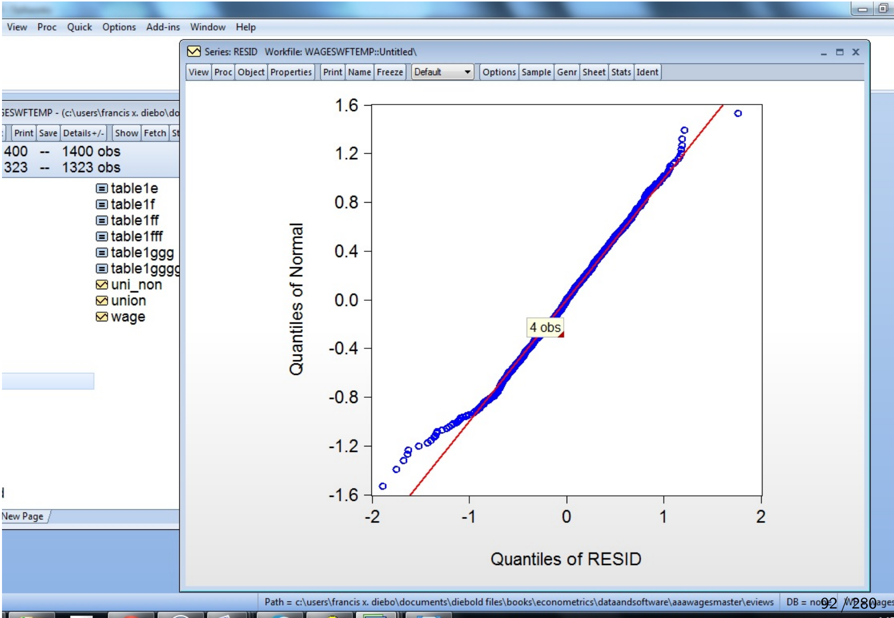
OLS Residual Histogram and Statistics



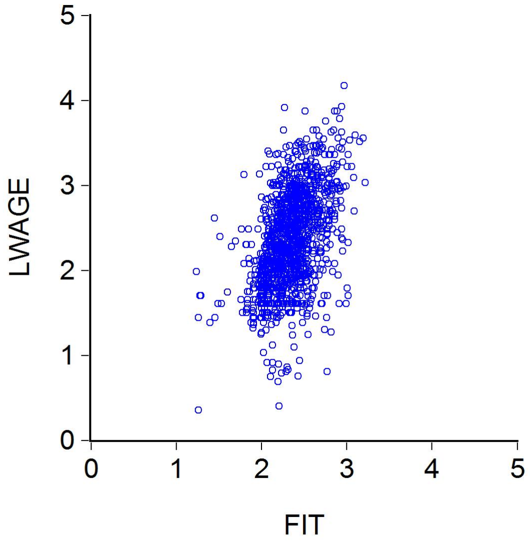
QQ Plots

- We introduced histograms earlier...
- ...but if interest centers on the *tails* of distributions, QQ plots often provide sharper insight as to the agreement or divergence between the actual and reference distributions
- QQ plot is quantiles of the standardized data against quantiles of a standardized reference distribution (e.g., normal)
- If the distributions match, the QQ plot is the 45 degree line
- To the extent that the QQ plot does not match the 45 degree line, the nature of the divergence can be very informative, as for example in indicating fat tails

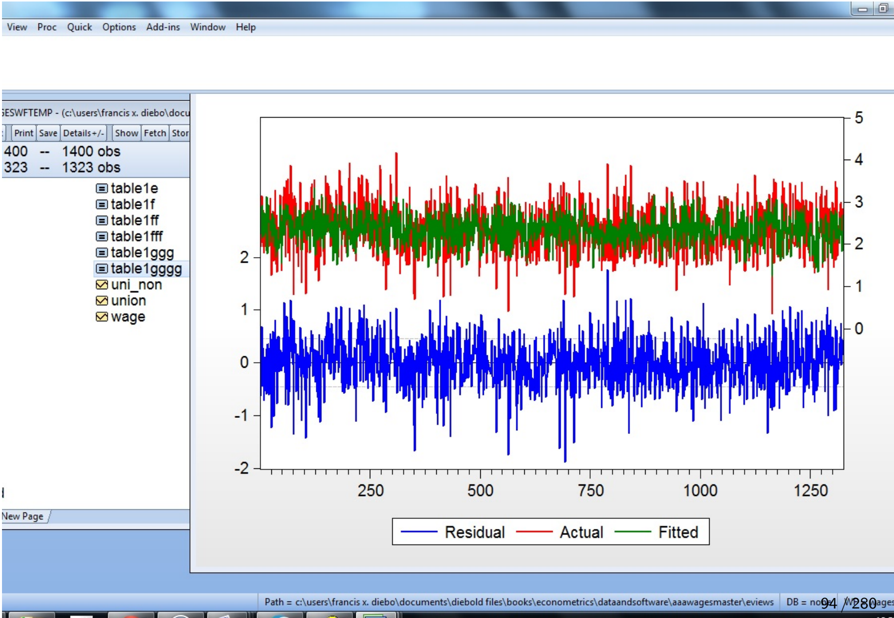
OLS Wage Regression Residual QQ Plot



Residual Scatter



OLS Residual Plot



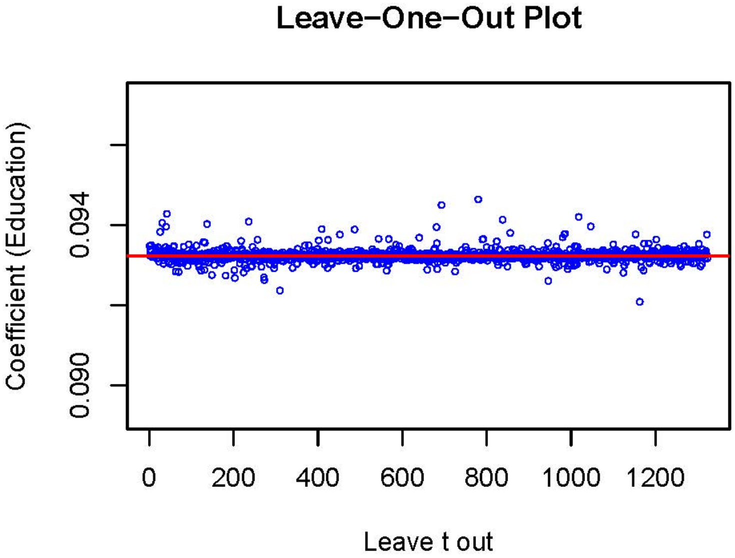
Leave-One-Out Plot

Consider:

$$\left(\hat{\beta}^{(-i)} - \hat{\beta}\right), \;\; i = 1, \ldots N$$

"Leave-one-out plot"

Wage Regression



Robust Estimation: Least Absolute Deviations (LAD)

The LAD estimator, $\hat{\beta}_{LAD}$, solves:

$$\min_{\beta} \sum_{i=1}^{N} |\varepsilon_i|$$

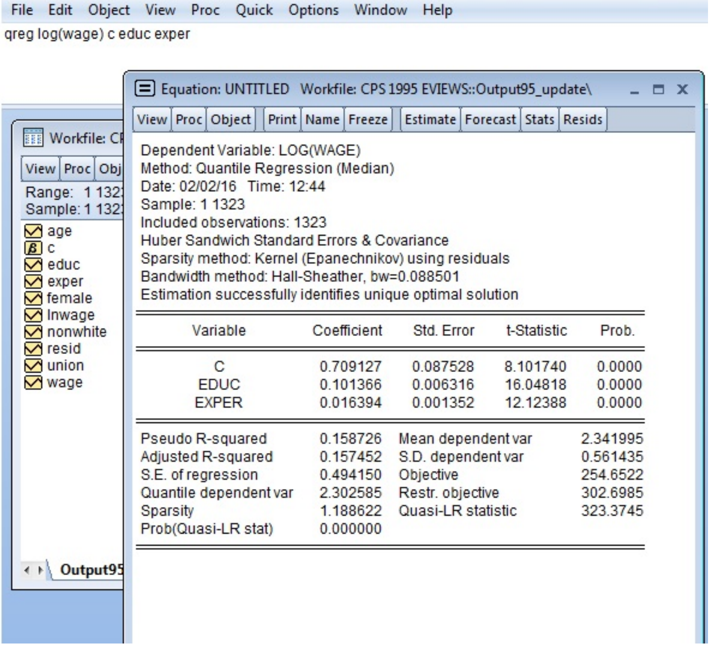
– Not as simple as OLS, but still simple

$x_i'\hat{\beta}_{OLS}$ is an estimate of $E(y_i|x_i)$
"OLS fits the conditional mean function"

$x_i'\hat{\beta}_{LAD}$ is an estimate of $median(y_i|x_i)$
"LAD fits the conditional median function"

– The two are equal with symmetric disturbances, but not with asymmetric disturbances, in which case the median is a more robust measure of central tendency of the conditional density

LAD Wage Regression Estimation



Dependent Variable: LOG(WAGE)
Method: Quantile Regression (Median)
Date: 02/02/16 Time: 12:44
Sample: 1 1323
Included observations: 1323
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kernel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.088501
Estimation successfully identifies unique optimal solution

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.709127	0.087528	8.101740	0.0000
EDUC	0.101366	0.006316	16.04818	0.0000
EXPER	0.016394	0.001352	12.12388	0.0000

Pseudo R-squared	0.158726	Mean dependent var	2.341995
Adjusted R-squared	0.157452	S.D. dependent var	0.561435
S.E. of regression	0.494150	Objective	254.6522
Quantile dependent var	2.302585	Restr. objective	302.6985
Sparsity	1.188622	Quasi-LR statistic	323.3745
Prob(Quasi-LR stat)	0.000000		

Digging into Prediction (Much) More Deeply (Again)

The environment is:

$$y_i = x_i'\beta + \varepsilon_i, \;\; i = 1, ..., N,$$

subject to the IC, except that we allow

$$\varepsilon_i \sim \;\; iid \;\; D(0, \sigma^2)$$

Simulation Algorithm for Feasible Density Prediction With Normality

Consider a density forecast for a person t with characteristics $x_i = x_i^*$.

1. Take R draws from $N(0, s^2)$.
2. Add $x_i^{*\prime}\hat{\beta}$ to each disturbance draw.
3. Form a density forecast by making a histogram for the output from step 2.

[If desired, form an interval forecast (95%, say) by sorting the output from step 2 to get the empirical cdf, and taking the left and right interval endpoints as the the 2.5% and 97.5% values.]

As $R \to \infty$ and $N \to \infty$, all error vanishes.

Now: Simulation Algorithm for Feasible Density Prediction Without Normality

1. Take R disturbance draws by assigning probability $1/N$ to each regression residual and sampling with replacement.
2. Add $x_i^{*\prime}\hat{\beta}$ to each draw.
3. Form a density forecast by fitting a density to the output from step 2.

[If desired, form a 95% interval forecast by sorting the output from step 2, and taking the left and right interval endpoints as the the .025% and .975% values.]

As $R \to \infty$ and $N \to \infty$, all error vanishes.

Indicator Variables in Cross Sections: Group Effects

Effectively a type of structural change in cross sections
(Different intercepts for different groups of people)

Do we really believe that intercepts are identical across groups?

Dummy Variables for Group Effects

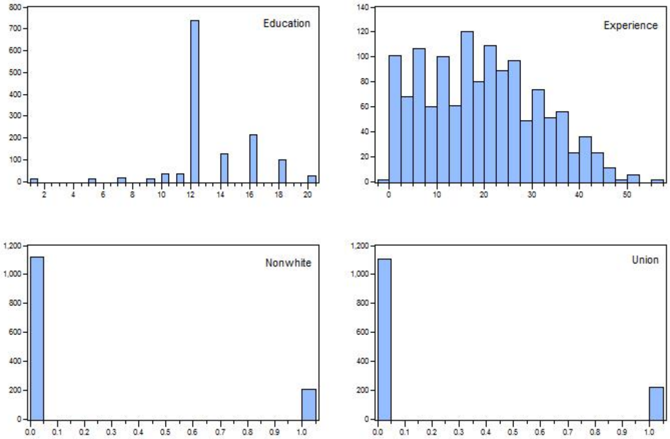
A dummy variable, or indicator variable, is just a 0-1 variable that indicates something, such as whether a person is female:

$$FEMALE_i = \begin{cases} 1 \text{ if person } i \text{ is female} \\ 0 \text{ otherwise} \end{cases}$$

(It really is that simple.)

"Intercept dummies"

Histograms for Wage Covariates

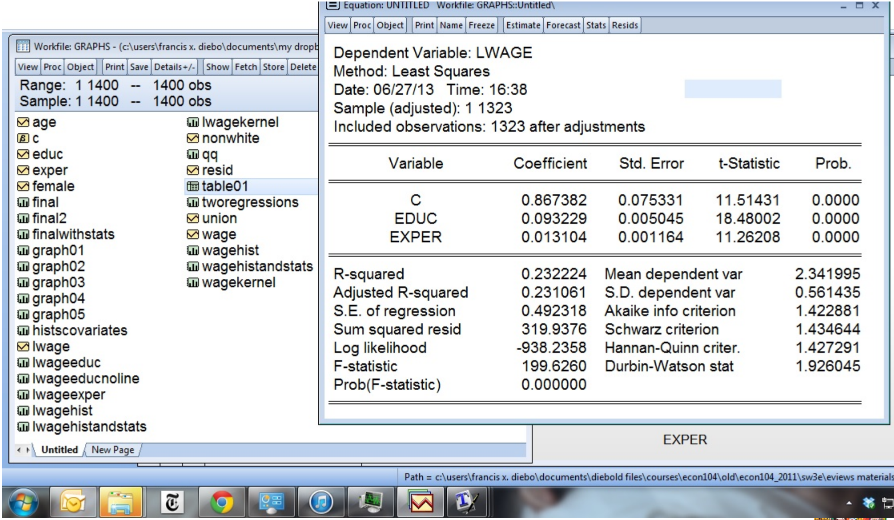


Notice that the sample mean of an indicator variable is the fraction of the sample with the indicated attribute.

Recall Basic Wage Regression on Education and Experience

$$LWAGE \rightarrow C, EDUC, EXPER$$

Basic Wage Regression Results



Dependent Variable: LWAGE
Method: Least Squares
Date: 06/27/13 Time: 16:38
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.867382	0.075331	11.51431	0.0000
EDUC	0.093229	0.005045	18.48002	0.0000
EXPER	0.013104	0.001164	11.26208	0.0000

R-squared	0.232224	Mean dependent var	2.341995
Adjusted R-squared	0.231061	S.D. dependent var	0.561435
S.E. of regression	0.492318	Akaike info criterion	1.422881
Sum squared resid	319.9376	Schwarz criterion	1.434644
Log likelihood	-938.2358	Hannan-Quinn criter.	1.427291
F-statistic	199.6260	Durbin-Watson stat	1.926045
Prob(F-statistic)	0.000000		

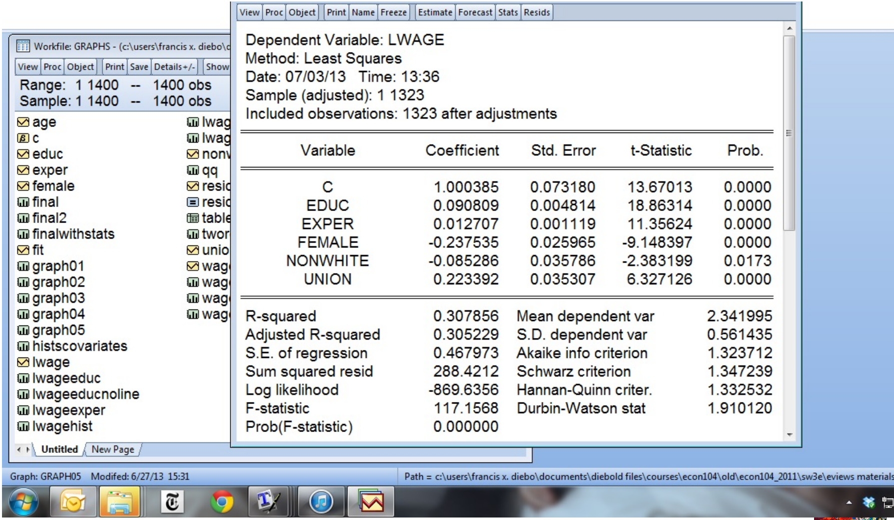
Introducing Sex, Race, and Union Status in the Wage Regression

Now:

$$LWAGE \;\rightarrow\; C, EDUC, EXPER, FEMALE, NONWHITE, UNION$$

The estimated intercept corresponds to the “base case” across all dummies (i.e., when all dummies are simultaneously 0), and the estimated dummy coefficients give the estimated extra effects (i.e., when the respective dummies are 1).

Wage Regression on Education, Experience, and Group Dummies


Dependent Variable: LWAGE
Method: Least Squares
Date: 07/03/13 Time: 13:36
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.000385	0.073180	13.67013	0.0000
EDUC	0.090809	0.004814	18.86314	0.0000
EXPER	0.012707	0.001119	11.35624	0.0000
FEMALE	-0.237535	0.025965	-9.148397	0.0000
NONWHITE	-0.085286	0.035786	-2.383199	0.0173
UNION	0.223392	0.035307	6.327126	0.0000

R-squared	0.307856	Mean dependent var	2.341995
Adjusted R-squared	0.305229	S.D. dependent var	0.561435
S.E. of regression	0.467973	Akaike info criterion	1.323712
Sum squared resid	288.4212	Schwarz criterion	1.347239
Log likelihood	-869.6356	Hannan-Quinn criter.	1.332532
F-statistic	117.1568	Durbin-Watson stat	1.910120
Prob(F-statistic)	0.000000		

Predictive Perspectives

<u>Basic Wage Regression</u>

– Conditions only on education and experience.

– Intercept is a mongrel combination of those for men, women; white, non-white; union, non-union.

– Comparatively sparse "information set". Forecasting performance could be improved.

<u>Wage Regression With Dummies</u>

– Conditions on education, experience, *and* sex, race, and union status.

– Now we have different, "customized", intercepts by sex, race, and union status.

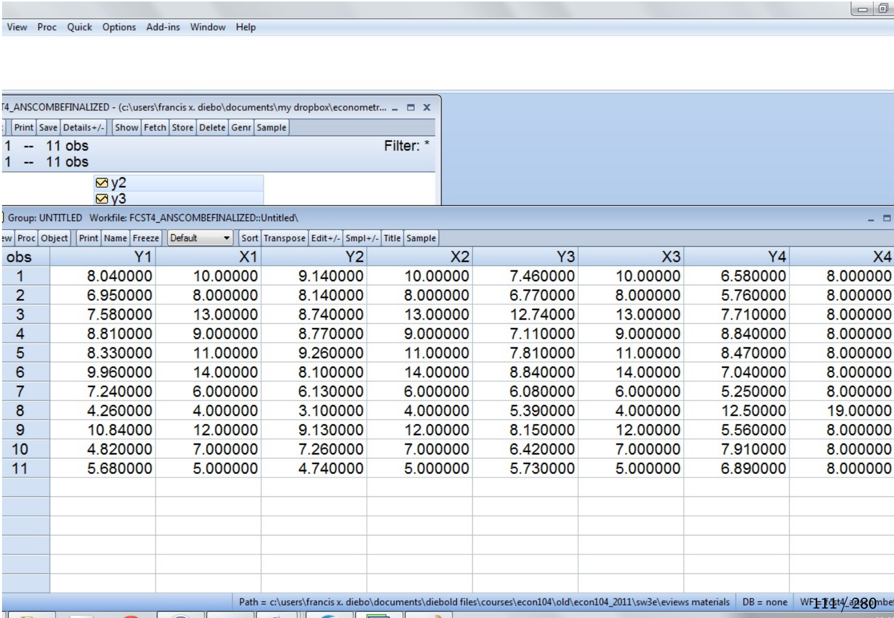
– Comparatively rich information set. Forecasting performance should be better.

e.g., knowing that someone is female, non-white, and non-union would be very valuable (in addition to education and experience) for predicting her wage!

Nonlinearity

Do we really believe that the relationship is linear?

Anscombe's Quartet



obs	Y1	X1	Y2	X2	Y3	X3	Y4	X4
1	8.040000	10.00000	9.140000	10.00000	7.460000	10.00000	6.580000	8.000000
2	6.950000	8.000000	8.140000	8.000000	6.770000	8.000000	5.760000	8.000000
3	7.580000	13.00000	8.740000	13.00000	12.74000	13.00000	7.710000	8.000000
4	8.810000	9.000000	8.770000	9.000000	7.110000	9.000000	8.840000	8.000000
5	8.330000	11.00000	9.260000	11.00000	7.810000	11.00000	8.470000	8.000000
6	9.960000	14.00000	8.100000	14.00000	8.840000	14.00000	7.040000	8.000000
7	7.240000	6.000000	6.130000	6.000000	6.080000	6.000000	5.250000	8.000000
8	4.260000	4.000000	3.100000	4.000000	5.390000	4.000000	12.50000	19.00000
9	10.84000	12.00000	9.130000	12.00000	8.150000	12.00000	5.560000	8.000000
10	4.820000	7.000000	7.260000	7.000000	6.420000	7.000000	7.910000	8.000000
11	5.680000	5.000000	4.740000	5.000000	5.730000	5.000000	6.890000	8.000000

Anscombe's Quartet: Regressions

LS // Dependent Variable is Y1

Variable	Coefficient	Std. Error	T-Statistic	
C	3.00	1.12	2.67	
X1	0.50	0.12	4.24	
R-squared	0.67		S.E. of regression	1.24

LS // Dependent Variable is Y2

Variable	Coefficient	Std. Error	T-Statistic	
C	3.00	1.12	2.67	
X2	0.50	0.12	4.24	
R-squared	0.67		S.E. of regression	1.24

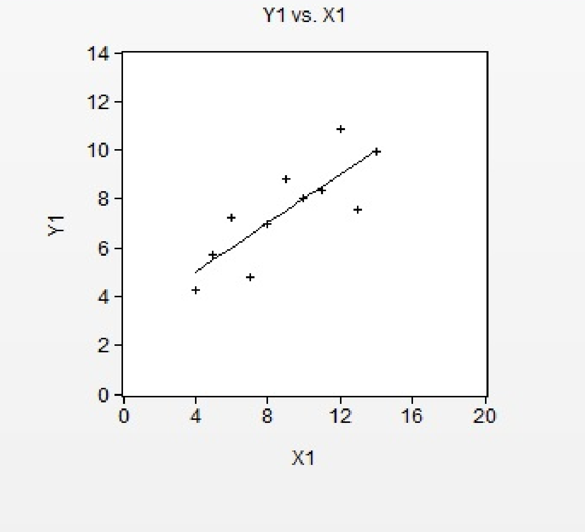
LS // Dependent Variable is Y3

Variable	Coefficient	Std. Error	T-Statistic	
C	3.00	1.12	2.67	
X3	0.50	0.12	4.24	
R-squared	0.67		S.E. of regression	1.24

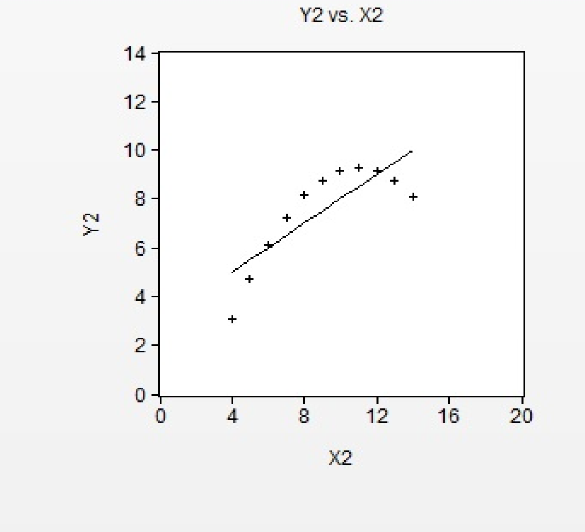
LS // Dependent Variable is Y4

Variable	Coefficient	Std. Error	T-Statistic	
C	3.00	1.12	2.67	
X4	0.50	0.12	4.24	
R-squared	0.67		S.E. of regression	1.24

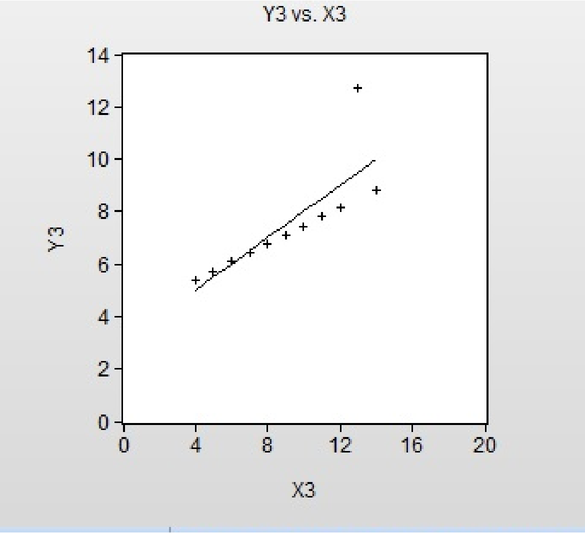
Anscombe's Quartet Graphics: Dataset 1



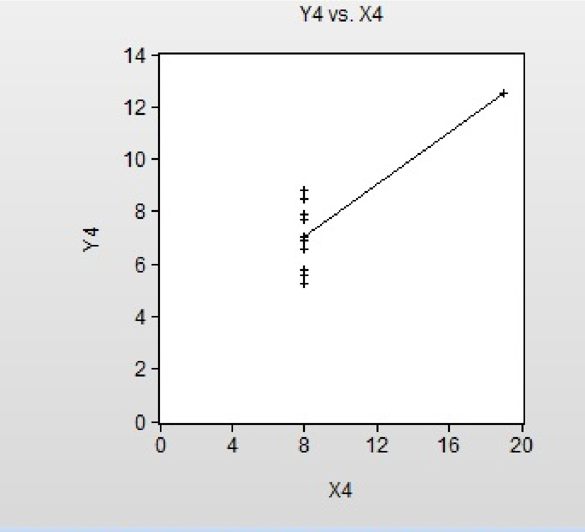
Anscombe's Quartet Graphics: Dataset 2



Anscombe's Quartet Graphics: Dataset 3



Anscombe's Quartet Graphics: Dataset 4



Log-Log Regression

$$lny_i = \beta_1 + \beta_2 lnx_i + \varepsilon_i$$

For close y_i and x_i, $(ln\, y_i - ln\, x_i) \cdot 100$ is approximately the percent difference between y_i and x_i. Hence the coefficients in log-log regressions give the expected percent change in y for a one-percent change in x. That is, they give the *elasticity of y with respect to x*.

Example: Cobb-Douglas production function

$$y_i = AL_i^{\alpha} K_i^{\beta} exp(\varepsilon_i)$$

$$lny_i = lnA + \alpha lnL_i + \beta lnK_i + \varepsilon_i$$

We expect an α% increase in output
in response to a 1% increase in labor input

Log-Lin Regression

$$lny_i = \beta x_i + \varepsilon$$

The coefficients in log-lin regressions give the expected percentage change in y for a one-unit (not 1%!) change in x.

Example: LWAGE regression
Coefficient on education gives the expected percent change in WAGE arising from one more year of education.

Intrinsically Non-Linear Models

One example is the "S-curve" model,

$$y = \frac{1}{a + br^x}$$

$$(0 < r < 1)$$

– No way to transform to linearity

– Minimize the sum of squared errors numerically
"Nonlinear least squares"
$\hat{\beta}_{NLS}$

Taylor Series Expansions

Really no such thing as an intrinsically non-linear model...

In the bivariate case we can think of the relationship as

$$y_i = g(x_i, \varepsilon_i)$$

or slightly less generally as

$$y_i = f(x_i) + \varepsilon_i$$

Taylor Series, Continued

Consider Taylor series expansions of $f(x_i)$.
The linear (first-order) approximation is

$$f(x_i) \approx \beta_1 + \beta_2 x_i,$$

and the quadratic (second-order) approximation is

$$f(x_i) \approx \beta_1 + \beta_2 x_i + \beta_3 x_i^2.$$

In the multiple regression case, Taylor approximations also involve interaction terms. Consider, for example, $y_i = f(x_{i2}, x_{i3})$. Then:

$$y_i = f(x_{i2}, x_{i3}) \approx \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i3}^2 + \beta_6 x_{i2} x_{i3} + \ldots.$$

A Key Insight

The ultimate point is that so-called "intrinsically non-linear" models are themselves linear when viewed from the series-expansion perspective. In principle, of course, an infinite number of series terms are required, but in practice nonlinearity is often quite gentle (e.g., quadratic) so that only a few series terms are required.

– So omitted non-linearity is ultimately
an omitted-variables problem

Predictive Perspectives

– One can always fit a linear model

– But if DGP is nonlinear, then potentially-important Taylor terms are omitted, potentially severely degrading forecasting performance

– Just see the earlier Dataset 2 Anscombe graph!

Assessing Non-Linearity
(i.e., deciding on higher-order Taylor terms)

Use *SIC* as always.

Use t's and F as always.

Linear Wage Regression (Actually Log-Lin)

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LWAGE
Method: Least Squares
Date: 07/03/13 Time: 13:36
Sample (adjusted): 1 1323
Included observations: 1323 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.000385	0.073180	13.67013	0.0000
EDUC	0.090809	0.004814	18.86314	0.0000
EXPER	0.012707	0.001119	11.35624	0.0000
FEMALE	-0.237535	0.025965	-9.148397	0.0000
NONWHITE	-0.085286	0.035786	-2.383199	0.0173
UNION	0.223392	0.035307	6.327126	0.0000

R-squared	0.307856	Mean dependent var	2.341995
Adjusted R-squared	0.305229	S.D. dependent var	0.561435
S.E. of regression	0.467973	Akaike info criterion	1.323712
Sum squared resid	288.4212	Schwarz criterion	1.347239
Log likelihood	-869.6356	Hannan-Quinn criter.	1.332532
F-statistic	117.1568	Durbin-Watson stat	1.910120
Prob(F-statistic)	0.000000		

Path = c:\users\francis x. diebo\documents\diebold files\courses\econ104\old\econ104_2011\sw3e\eviews materials DB =

Quadratic Wage Regression

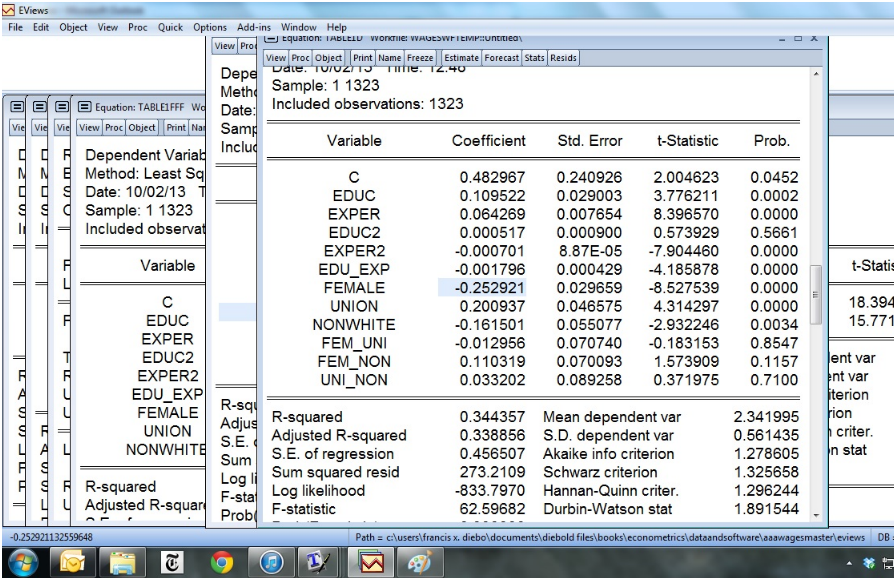
Dependent Variable: LWAGE
Method: Least Squares
Date: 10/02/13 Time: 12:37
Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.473236	0.240586	1.967017	0.0494
EDUC	0.109673	0.028918	3.792608	0.0002
EXPER	0.064422	0.007652	8.419060	0.0000
EDUC2	0.000501	0.000895	0.559994	0.5756
EXPER2	-0.000705	8.86E-05	-7.962263	0.0000
EDU_EXP	-0.001789	0.000429	-4.173423	0.0000
FEMALE	-0.237696	0.025506	-9.319335	0.0000
UNION	0.202955	0.034569	5.870998	0.0000
NONWHITE	-0.095028	0.034931	-2.720476	0.0066

R-squared	0.343072	Mean dependent var	2.341995
Adjusted R-squared	0.339073	S.D. dependent var	0.561435
S.E. of regression	0.456433	Akaike info criterion	1.276028
Sum squared resid	273.7465	Schwarz criterion	1.311318
Log likelihood	-835.0925	Hannan-Quinn criter.	1.289257
F-statistic	85.77745	Durbin-Watson stat	1.894409

Path = c:\users\francis x. diebo\documents\diebold files\books\econometrics\dataandsoftware\aaawagesmaster\

Quadratic Wage Regression with Dummy Interactions

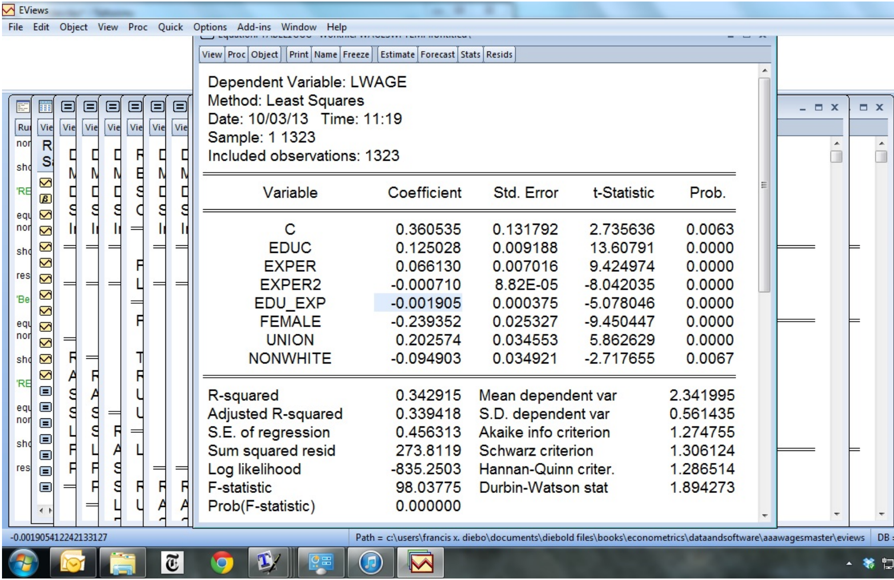

Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.482967	0.240926	2.004623	0.0452
EDUC	0.109522	0.029003	3.776211	0.0002
EXPER	0.064269	0.007654	8.396570	0.0000
EDUC2	0.000517	0.000900	0.573929	0.5661
EXPER2	-0.000701	8.87E-05	-7.904460	0.0000
EDU_EXP	-0.001796	0.000429	-4.185878	0.0000
FEMALE	-0.252921	0.029659	-8.527539	0.0000
UNION	0.200937	0.046575	4.314297	0.0000
NONWHITE	-0.161501	0.055077	-2.932246	0.0034
FEM_UNI	-0.012956	0.070740	-0.183153	0.8547
FEM_NON	0.110319	0.070093	1.573909	0.1157
UNI_NON	0.033202	0.089258	0.371975	0.7100

R-squared	0.344357	Mean dependent var	2.341995
Adjusted R-squared	0.338856	S.D. dependent var	0.561435
S.E. of regression	0.456507	Akaike info criterion	1.278605
Sum squared resid	273.2109	Schwarz criterion	1.325658
Log likelihood	-833.7970	Hannan-Quinn criter.	1.296244
F-statistic	62.59682	Durbin-Watson stat	1.891544

Figure: Wage Regression with Continuous Non-Linearities and

Final Specification



Dependent Variable: LWAGE
Method: Least Squares
Date: 10/03/13 Time: 11:19
Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.360535	0.131792	2.735636	0.0063
EDUC	0.125028	0.009188	13.60791	0.0000
EXPER	0.066130	0.007016	9.424974	0.0000
EXPER2	-0.000710	8.82E-05	-8.042035	0.0000
EDU_EXP	-0.001905	0.000375	-5.078046	0.0000
FEMALE	-0.239352	0.025327	-9.450447	0.0000
UNION	0.202574	0.034553	5.862629	0.0000
NONWHITE	-0.094903	0.034921	-2.717655	0.0067

R-squared	0.342915	Mean dependent var	2.341995
Adjusted R-squared	0.339418	S.D. dependent var	0.561435
S.E. of regression	0.456313	Akaike info criterion	1.274755
Sum squared resid	273.8119	Schwarz criterion	1.306124
Log likelihood	-835.2503	Hannan-Quinn criter.	1.286514
F-statistic	98.03775	Durbin-Watson stat	1.894273
Prob(F-statistic)	0.000000		

Discrete Response Models

What if the dependent variable is binary?

– Ultimately violates the IC in multiple ways...

(Nonlinear, non-Gaussian)

Many Names

"discrete response models"

"qualitative response models"

"limited dependent variable models"

"binary (binomial) response models"

"classification models"

"logistic regression models" (a leading case)

– Another appearance of a dummy variable,
but the dummy is on the left

Framework

Left-hand-side variable is $y_i = I_i(z)$, where the "indicator variable" $I_i(z)$ indicates whether event z occurs; that is,

$$I_i(z) = \begin{cases} 1 \text{ if event } z \text{ occurs} \\ 0 \text{ otherwise.} \end{cases}$$

The usual linear regression setup,

$$E(y_i|x_i) = x_i'\beta$$

becomes

$$E(I_i(z)\,|\,x_i) = x_i'\beta.$$

A key insight, however, is that

$$E(I_i(z)\,|\,x_i) = P(I_i(z){=}1\,|\,x_i),$$

so the setup is really

$$P(I_i(z){=}1\,|\,x_i) = x_i'\beta. \tag{1}$$

– Leading examples: recessions, bankruptcies, loan or credit card defaults, financial market crises, consumer choices, ...

The "Linear Probability Model"

How to "fit a line" when the LHS variable is binary?

The linear probability model (LPM) does it by brute-force OLS regression $I_i(z) \to x_i$.

Problem: The LPM fails to constrain the fitted probabilities to be in the unit interval.

Squashing Functions

Solution: Run $x_i'\beta$ through a monotone "squashing function," $F(\cdot)$, that keeps $P(I_i(z){=}1\,|\,x_i)$ in the unit interval.

More precisely, move to models with

$$E(y_i|x_i) = P(I_i(z){=}1\,|\,x_i) = F(x_i'\beta),$$

where $F(\cdot)$ is monotone increasing, with $lim_{w\to\infty}F(w) = 1$ and $lim_{w\to-\infty}F(w) = 0$.

The Logit Model

In the "logit" model, the squashing function $F(\cdot)$ is the logistic function,

$$F(w) = logit(w) = \frac{e^{w}}{1+e^{w}} = \frac{1}{1+e^{-w}},$$

so the logit model is

$$P(I_i(z){=}1 \,|\, x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}.$$

– Logit is a nonlinear model for the event probability.

Logit as a Linear Model for the Log Odds

Consider a linear model for log odds

$$ln\left(\frac{P(I_i(z)=1 \mid x_i)}{1-P(I_i(z)=1 \mid x_i)}\right) = x_i'\beta.$$

Solving the log odds for $P(I_i(z)=1 \mid x_i)$ yields the logit model,

$$P(I_i(z)=1 \mid x_i) = \frac{1}{1+e^{-x_i'\beta}} = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}.$$

So logit is just linear regression for log odds.

Logit Estimation

The likelihood function can be derived, and the model can be estimated by numerical maximization of the likelihood function.

For linear regression we had:

$$y_i | x_i \sim N(x_i'\beta, \sigma^2),$$

from which we derived the likelihood and the MLE.

For the linear probability model we have:

$$y_i | x_i \sim Bernoulli\left(x_i'\beta\right).$$

For logit we have:

$$y_i | x_i \sim Bernoulli\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right).$$

Logit RHS Variable Effects

Note that the individual RHS variable effects, $\partial E(y_i|x_i)/\partial x_{ik}$, are not simply given by the β_k's as in standard linear regression. Instead we have

$$\frac{\partial E(y_i|x_i)}{\partial x_{ik}} = \frac{\partial F(x_i'\beta)}{\partial x_{ik}} = f(x_i'\beta)\beta_k,$$

where $f(x) = dF(x)/dx$. So the marginal effect is not simply β_k; instead it is β_k weighted by $f(x_i'\beta)$, which depends on all β_k's and x_{ik}'s, $k = 1, ..., K$.

– However, signs of the β_k's *are* the signs of the effects, because f must be positive. (Recall that F is monotone increasing.)

– In addition, ratios of β_k's do give ratios of effects, because the f's cancel.

Logit R^2

Recall that traditional R^2 for continuous LHS variables is

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}.$$

For binary regression we proceed similarly:

$$R^2 = 1 - \frac{\sum(y_i - \hat{P}(I_i(z) = 1|x_i))^2}{\sum(y_i - \bar{y}_i)^2}.$$

"Efron's R^2"

The Logit Classifier

– Classification maps probabilities into 0-1 classifications. "Bayes classifier" uses a cutoff of .5.

– Decision boundary:
Suppose we use a Bayes classifier.
We predict 1 when $logit(x_i'\beta){>}1/2$. But that's the same as predicting 1 when $x_i'\beta{>}0$ since $logit(0){=}1/2$. If there are 2 x_i variables (potentially plus an intercept), then the condition $x_i'\beta{=}0$ defines a line in $\mathbb{R}^2$. Points on one side will be classified as 0, and points on the other side will be classified as 1. That line is the "decision boundary".

– In higher dimensions the decision boundary will be a plane or hyperplane.

– Note the "linear decision boundary". We can generalize to nonlinear decision boundaries in various ways.

Example: High-Wage Individuals

We now use a new wage data set that contains education and experience data for each person, but not wage data. Instead it contains only an indicator for whether the person is "high-wage" or "low-wage". (The binary indicator $HIGHWAGE_i$=1 if the hourly wage of person i is ≥ 15.)

– 357 people with $HIGHWAGE_i$=1

– 966 people with $HIGHWAGE_i$=0

We will fit a logit model using education and experience and see how it performs as a Bayes classifier.

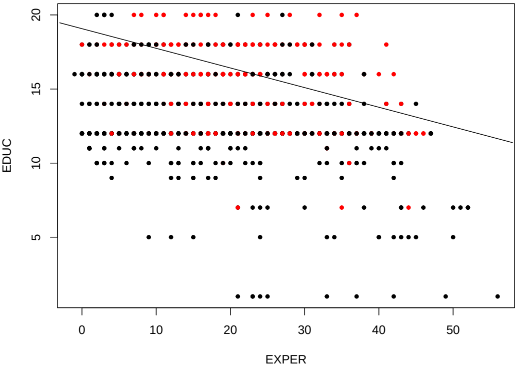
Logit Regression of *HIGHWAGE* on *EDUC* and *EXPER*

Table: Logit Regression

	Dependent variable:
	HIGHWAGE
EDUC	0.35
	(se=0.03)
EXPER	0.04
	(se=0.01)
Constant	-6.61
	(se=0.46)
Ratio of Effects *EDUC*/*EXPER*:	7.54
Efron's R^2:	0.15

Covariates and Decision Boundary for Logit Bayes Classifier

In-Sample:
(Red denotes high-wage people)



Out-of-sample: For a new person with covariates x^*,
predict $HIGHWAGE$=1 if $logit(x^{*\prime}\hat{\beta})$>1/2.
That is, if $x^{*\prime}\hat{\beta}>0$

Heteroskedasticity in Cross-Sections

Do we really believe that disturbance variances are constant over space?

"Unconditional Heteroskedasticity" is Occasionally Relevant...

Consider IC1:

$$\varepsilon_i \sim iidN(0, \sigma^2), \quad i = 1, ..., N$$

Unconditional heteroskedasticity occurs when the unconditional disturbance variance varies across people for some unknown reason.

Violation of IC1, in particular IC1.5:

"The ε_i's have constant variance σ^2"

... But *Conditional* Heteroskedasticity is Often Highly Relevant

Consider IC2.2:

$$var(\varepsilon_i \mid x_{i1}, ..., x_{iK}) = \sigma^2, \text{ for all } i$$

Conditional heteroskedasticity occurs when σ_i^2 varies systematically with $x_{i1}, ..., x_{iK}$, so that IC2.2 is violated

e.g., Consider the regression
fine wine consumption $\rightarrow$ *income*

Consequences for Estimation and Inference

– Esimation: OLS estimation remains largely OK.
Parameter estimates remain consistent and asymptotically normal.

– Inference: OLS inference is badly damaged.
Standard errors are inconsistent. t statistics do not have the t distribution in finite samples and do not even have the $N(0, 1)$ distribution asymptotically.

Consequences for Prediction

– Earlier point forecasts remain largely OK.

OLS parameter estimates remain consistent,
so $E(\widehat{y_i | x_i = x_i^*})$ is still consistent for $E(y_i | x_i = x_i^*)$.

– Earlier density (and hence interval) forecasts not OK.

It is no longer appropriate to base interval and density forecasts on "identical σ's for different people". Now we need to base them on "different σ's for different people".

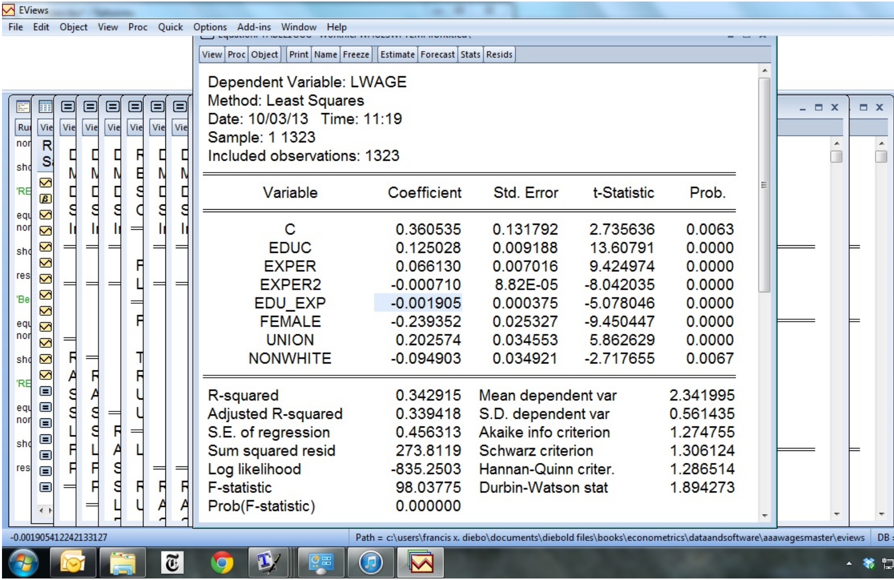
Detecting Conditional Heteroskedasticity

- Graphical heteroskedasticity diagnostics
- Formal heteroskedasticity tests

Graphical Diagnostics

Graph e_i^2 against x_{ik}, for various regressors (k)

Recall Our “Final” Wage Regression

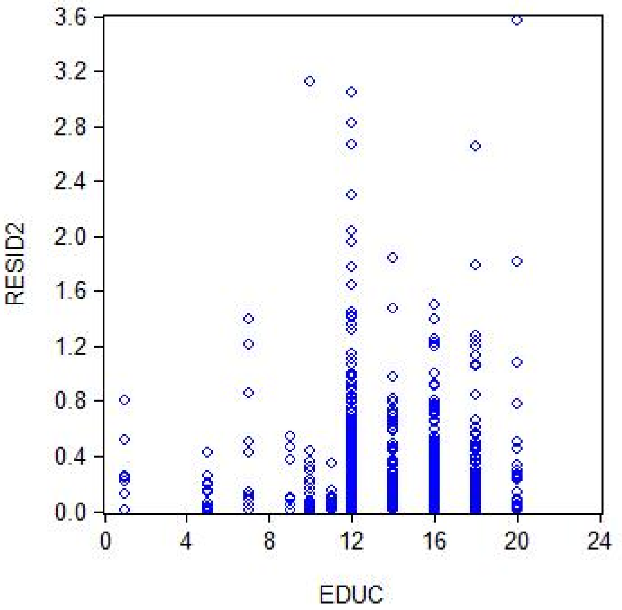


Dependent Variable: LWAGE
Method: Least Squares
Date: 10/03/13 Time: 11:19
Sample: 1 1323
Included observations: 1323

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.360535	0.131792	2.735636	0.0063
EDUC	0.125028	0.009188	13.60791	0.0000
EXPER	0.066130	0.007016	9.424974	0.0000
EXPER2	-0.000710	8.82E-05	-8.042035	0.0000
EDU_EXP	-0.001905	0.000375	-5.078046	0.0000
FEMALE	-0.239352	0.025327	-9.450447	0.0000
UNION	0.202574	0.034553	5.862629	0.0000
NONWHITE	-0.094903	0.034921	-2.717655	0.0067

R-squared	0.342915	Mean dependent var	2.341995
Adjusted R-squared	0.339418	S.D. dependent var	0.561435
S.E. of regression	0.456313	Akaike info criterion	1.274755
Sum squared resid	273.8119	Schwarz criterion	1.306124
Log likelihood	-835.2503	Hannan-Quinn criter.	1.286514
F-statistic	98.03775	Durbin-Watson stat	1.894273
Prob(F-statistic)	0.000000		

Squared Residual vs. EDUC



The Breusch-Pagan-Godfrey Test (BPG)

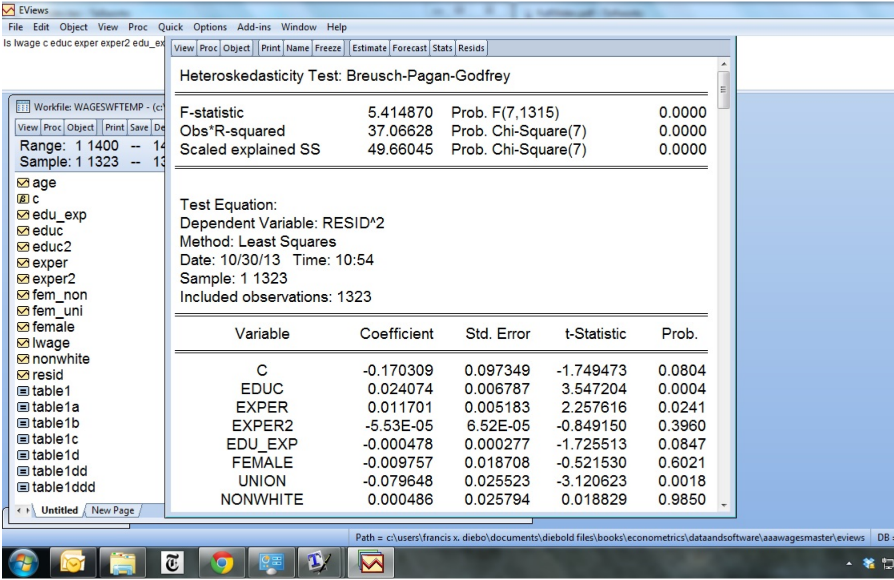
Limitation of graphing e_i^2 against x_{ik}: Purely pairwise

In contrast, BPG blends information from all regressors

BPG test:

- Estimate the OLS regression, and obtain the squared residuals
- Regress the squared residuals on all regressors
- To test the null hypothesis of no relationship, examine $N \cdot R^2$ from this regression. In large samples $N \cdot R^2 \sim \chi^2_{K-1}$ under the null of no conditional heteroskedasticity, where K is the number of regressors in the test regression.

BPG Test



Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	5.414870	Prob. F(7,1315)	0.0000
Obs*R-squared	37.06628	Prob. Chi-Square(7)	0.0000
Scaled explained SS	49.66045	Prob. Chi-Square(7)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 10/30/13 Time: 10:54
Sample: 1 1323
Included observations: 1323

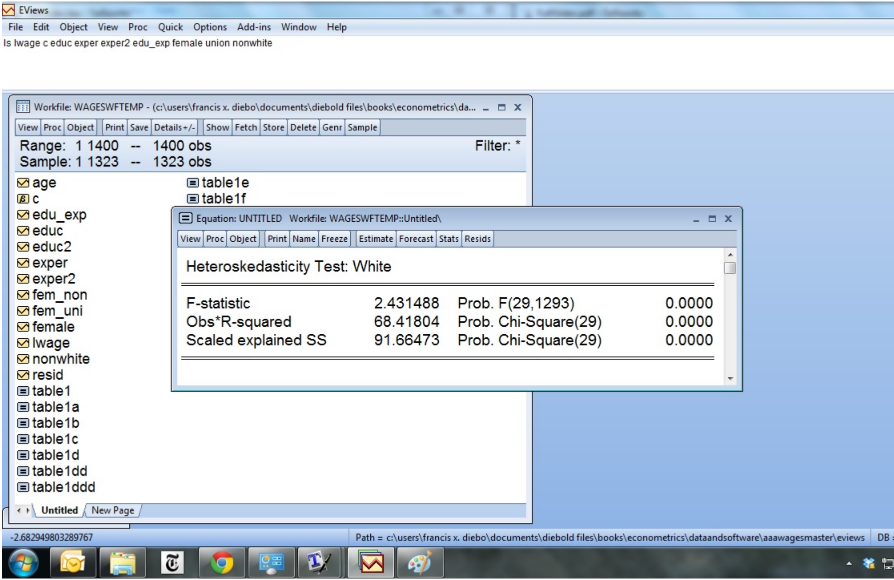
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.170309	0.097349	-1.749473	0.0804
EDUC	0.024074	0.006787	3.547204	0.0004
EXPER	0.011701	0.005183	2.257616	0.0241
EXPER2	-5.53E-05	6.52E-05	-0.849150	0.3960
EDU_EXP	-0.000478	0.000277	-1.725513	0.0847
FEMALE	-0.009757	0.018708	-0.521530	0.6021
UNION	-0.079648	0.025523	-3.120623	0.0018
NONWHITE	0.000486	0.025794	0.018829	0.9850

White's Test

Like BGP, but replace BGP's linear regression
with a more flexible (quadratic) regression

- Estimate the OLS regression, and obtain the squared residuals
- Regress the squared residuals on all regressors, squared regressors, and pairwise regressor cross products
- To test the null hypothesis of no relationship, examine $N \cdot R^2$ from this regression. In large samples $N \cdot R^2 \sim \chi^2_{K-1}$ under the null.

White's Test


EViews
File Edit Object View Proc Quick Options Add-ins Window Help
ls lwage c educ exper exper2 edu_exp female union nonwhite
Workfile: WAGESWFTEMP - (c:\users\francis x. diebo\documents\diebold files\books\econometrics\da...
View Proc Object Print Save Details+/- Show Fetch Store Delete Genr Sample
Range: 1 1400 -- 1400 obs
Sample: 1 1323 -- 1323 obs
Filter: *
age
c
edu_exp
educ
educ2
exper
exper2
fem_non
fem_uni
female
lwage
nonwhite
resid
table1
table1a
table1b
table1c
table1d
table1dd
table1ddd
table1e
table1f
Untitled
New Page
Equation: UNTITLED Workfile: WAGESWFTEMP::Untitled\
View Proc Object Print Name Freeze Estimate Forecast Stats Resids
Heteroskedasticity Test: White
F-statistic 2.431488 Prob. F(29,1293) 0.0000
Obs*R-squared 68.41804 Prob. Chi-Square(29) 0.0000
Scaled explained SS 91.66473 Prob. Chi-Square(29) 0.0000
-2.682949803289767
Path = c:\users\francis x. diebo\documents\diebold files\books\econometrics\dataandsoftware\aaawagesmaster\eviews
DB

Dealing with Heteroskedasticity

- Adjusting standard errors
- Adjusting density forecasts

Adjusting Standard Errors

Using advanced methods, one can obtain estimators for standard errors that are consistent even when heteroskedasticity is present.

"Heteroskedasticity-robust standard errors"
"White standard errors"

Before, under the IC:
$V = cov(\hat{\beta}_{LS})$ estimated by

$$\hat{V} = s^2(X'X)^{-1},$$

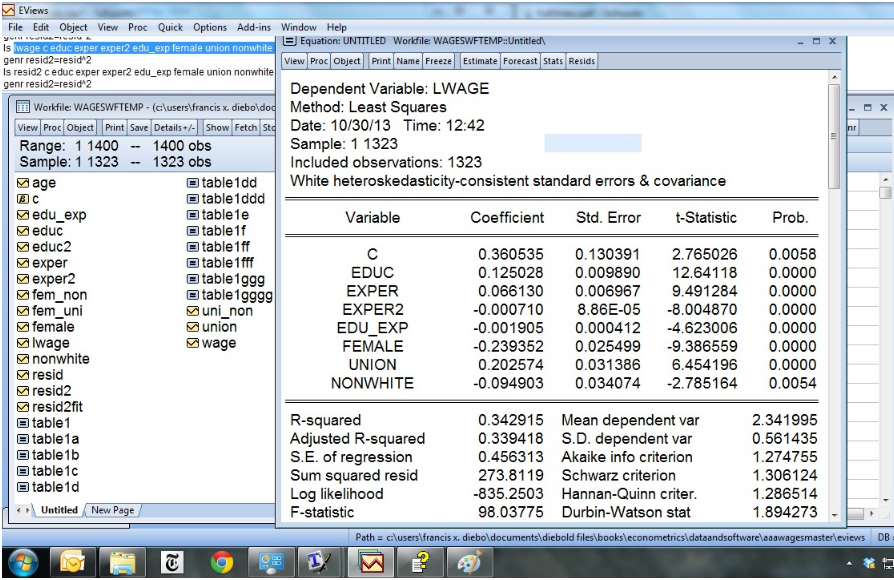
where $s^2 = \sum_{i=1}^{N} e_i^2/(N-K)$.

Now, under heteroskedasticity, V estimated by

$$\hat{V}_{White} = (X'X)^{-1}\left(X' diag(e_1^2, ..., e_N^2)X\right)(X'X)^{-1}$$

– Mechanically, it's just a simple OLS regression option.

Final Wage Regression with Robust Standard Errors



Dependent Variable: LWAGE
Method: Least Squares
Date: 10/30/13 Time: 12:42
Sample: 1 1323
Included observations: 1323
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.360535	0.130391	2.765026	0.0058
EDUC	0.125028	0.009890	12.64118	0.0000
EXPER	0.066130	0.006967	9.491284	0.0000
EXPER2	-0.000710	8.86E-05	-8.004870	0.0000
EDU_EXP	-0.001905	0.000412	-4.623006	0.0000
FEMALE	-0.239352	0.025499	-9.386559	0.0000
UNION	0.202574	0.031386	6.454196	0.0000
NONWHITE	-0.094903	0.034074	-2.785164	0.0054

R-squared	0.342915	Mean dependent var	2.341995
Adjusted R-squared	0.339418	S.D. dependent var	0.561435
S.E. of regression	0.456313	Akaike info criterion	1.274755
Sum squared resid	273.8119	Schwarz criterion	1.306124
Log likelihood	-835.2503	Hannan-Quinn criter.	1.286514
F-statistic	98.03775	Durbin-Watson stat	1.894273

Adjusting Density Forecasts

Recall non-operational version for Gaussian homoskedastic disturbances:

$$y_i \mid x_i = x^* \;\sim\; N(x^{*\prime}\beta,\; \sigma^2)$$

Recall operational version for Gaussian homoskedastic disturbances:

$$y_i \mid x_i = x^* \;\sim\; N(x^{*\prime}\hat{\beta}_{LS},\; s^2)$$

Now: Operational version for Gaussian *hetero*skedastic disturbances:

$$y_i \mid x_i = x^* \;\sim\; N(x^{*\prime}\hat{\beta}_{LS},\; \hat{\sigma}_*^2)$$

Q: Where do we get $\hat{\sigma}_*^2$?

Time Series

Misspecification and Model Selection

Do we really believe that the fitted model matches the DGP?
No major changes in time series
Same tools and techniques...

Non-Normality and Outliers

Do we really believe that the disturbances are Gaussian?
No major changes in time series
Same tools and techniques...

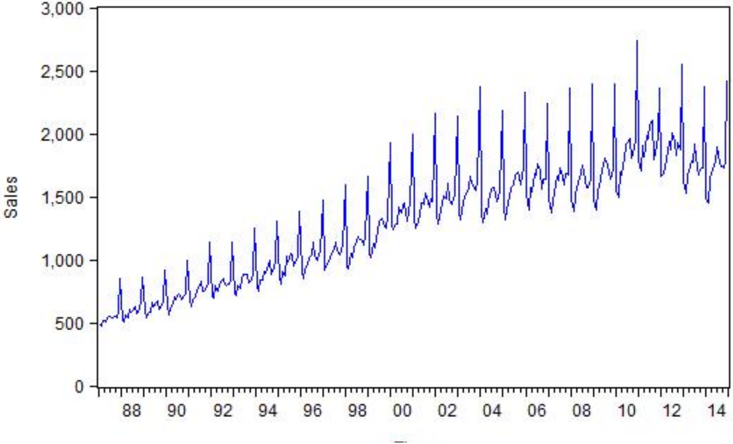
Indicator Variables in Time Series I: Trend

Trend is effectively a type of structural change

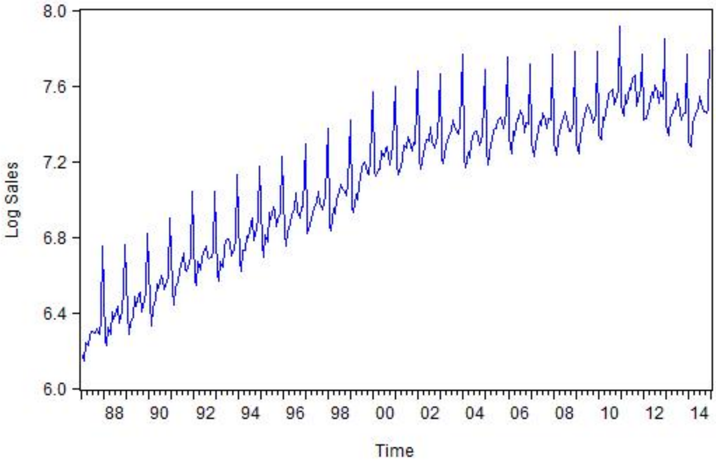
Do we really believe that intereps are fixed over *time*?

– Trend is about *gradual* intercept evolution

Liquor Sales



Log Liquor Sales



From now on we will take logs of liquor sales.
When we say "liquor sales", logs are understood.

Linear Trend

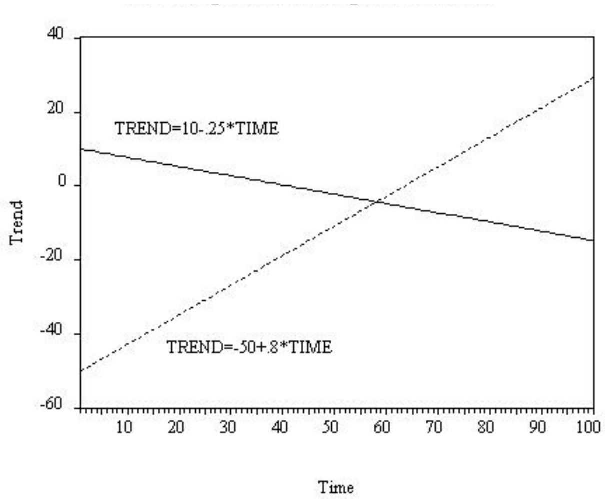
$$Trend_t = \beta_1 + \beta_2 TIME_t$$

where $TIME_t = t$

Simply run the least squares regression $y \rightarrow c, TIME$, where

$$TIME = \begin{pmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ T-1 \\ T \end{pmatrix}$$

Various Linear Trends



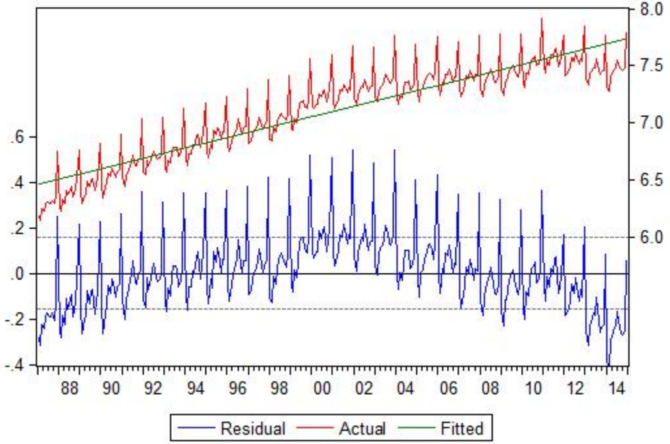
Linear Trend Estimation

Method: Least Squares
Date: 08/08/13   Time: 08:53
Sample: 1987M01 2014M12
Included observations: 336

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.454290	0.017468	369.4834	0.0000
TIME	0.003809	8.98E-05	42.39935	0.0000

R-squared	0.843318	Mean dependent var	7.096188
Adjusted R-squared	0.842849	S.D. dependent var	0.402962
S.E. of regression	0.159743	Akaike info criterion	-0.824561
Sum squared resid	8.523001	Schwarz criterion	-0.801840
Log likelihood	140.5262	Hannan-Quinn criter.	-0.815504
F-statistic	1797.705	Durbin-Watson stat	1.078573
Prob(F-statistic)	0.000000		

Residual Plot



Indicator Variables in Time Series II: Seasonality

Seasonality is effectively a type of structural change

Do we really believe that interepts are fixed over *seasons*?
(quite apart from, and even after accounting for,
time-varying intercepts due to trend)

Seasonal Dummies

$$Seasonal_s = \sum_{s=1}^{S} \beta_s SEAS_{st} \quad (S \text{ seasons per year})$$

$$\text{where } SEAS_{st} = \begin{cases} 1 \text{ if observation } t \text{ falls in season s} \\ 0 \text{ otherwise} \end{cases}$$

Simply run the least squares regression $y \rightarrow SEAS_1, ..., SEAS_S$
(or blend: $y \rightarrow TIME, SEAS_1, ..., SEAS_S$)

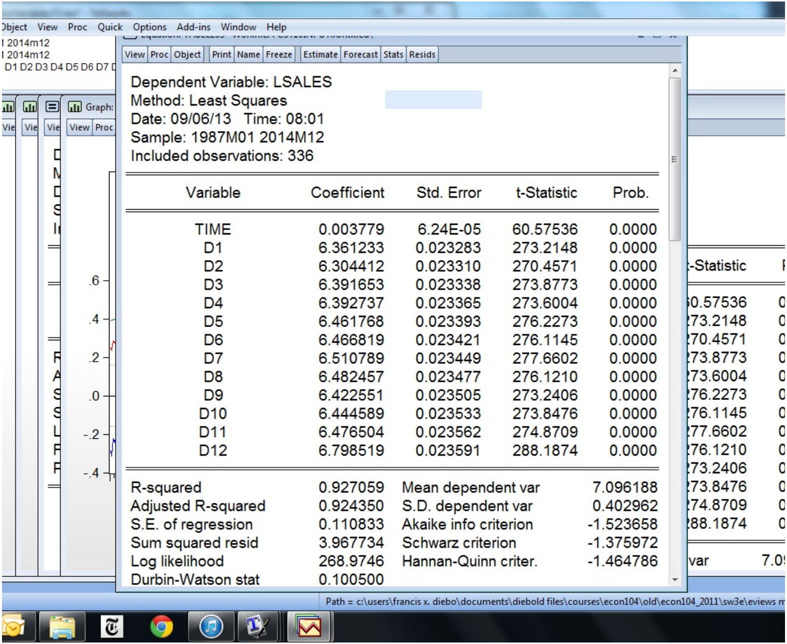
where (e.g., in quarterly data case, assuming Q1 start and Q4 end):

$$SEAS_1 = (1,0,0,0,1,0,0,0,1,0,0,0,...,0)'$$
$$SEAS_2 = (0,1,0,0,0,1,0,0,0,1,0,0,...,0)'$$
$$SEAS_3 = (0,0,1,0,0,0,1,0,0,0,1,0,...,0)'$$
$$SEAS_4 = (0,0,0,1,0,0,0,1,0,0,0,1,...,1)'.$$

– Full set of dummies ("all categories") and hence no intercept.

– In CS case we dropped a category for each dummy (e.g., included "UNION" but not "NONUNION") and included an intercept.

Linear Trend with Seasonal Dummies

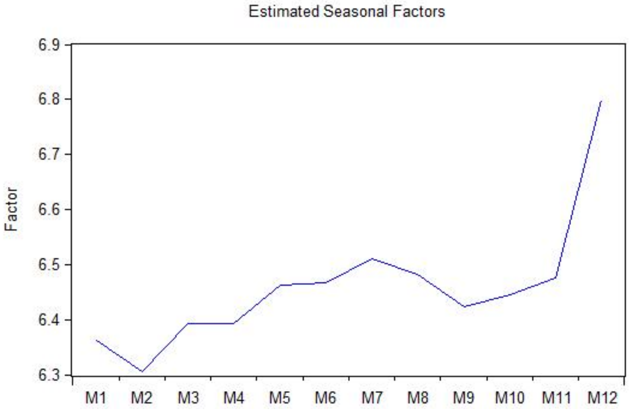


Dependent Variable: LSALES
Method: Least Squares
Date: 09/06/13 Time: 08:01
Sample: 1987M01 2014M12
Included observations: 336

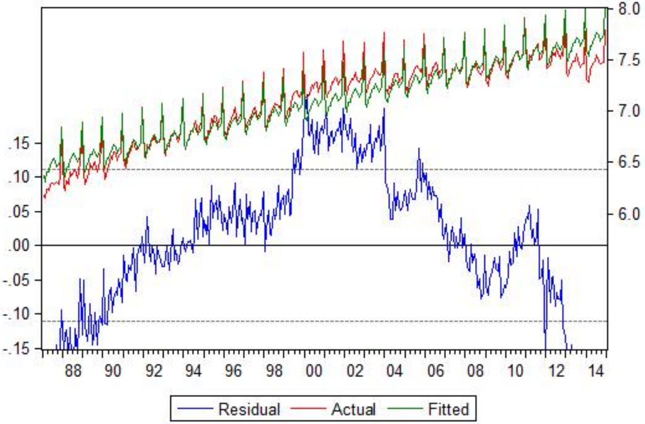
Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.003779	6.24E-05	60.57536	0.0000
D1	6.361233	0.023283	273.2148	0.0000
D2	6.304412	0.023310	270.4571	0.0000
D3	6.391653	0.023338	273.8773	0.0000
D4	6.392737	0.023365	273.6004	0.0000
D5	6.461768	0.023393	276.2273	0.0000
D6	6.466819	0.023421	276.1145	0.0000
D7	6.510789	0.023449	277.6602	0.0000
D8	6.482457	0.023477	276.1210	0.0000
D9	6.422551	0.023505	273.2406	0.0000
D10	6.444589	0.023533	273.8476	0.0000
D11	6.476504	0.023562	274.8709	0.0000
D12	6.798519	0.023591	288.1874	0.0000

R-squared	0.927059	Mean dependent var	7.096188
Adjusted R-squared	0.924350	S.D. dependent var	0.402962
S.E. of regression	0.110833	Akaike info criterion	-1.523658
Sum squared resid	3.967734	Schwarz criterion	-1.375972
Log likelihood	268.9746	Hannan-Quinn criter.	-1.464786
Durbin-Watson stat	0.100500		

Seasonal Pattern



Residual Plot



Nonlinearity in Time Series

Do we really believe that trends are linear?

Non-Linear Trend: Exponential (Log-Linear)

$$Trend_t = \beta_1 e^{\beta_2 TIME_t}$$

$$\ln(Trend_t) = \ln(\beta_1) + \beta_2 TIME_t$$

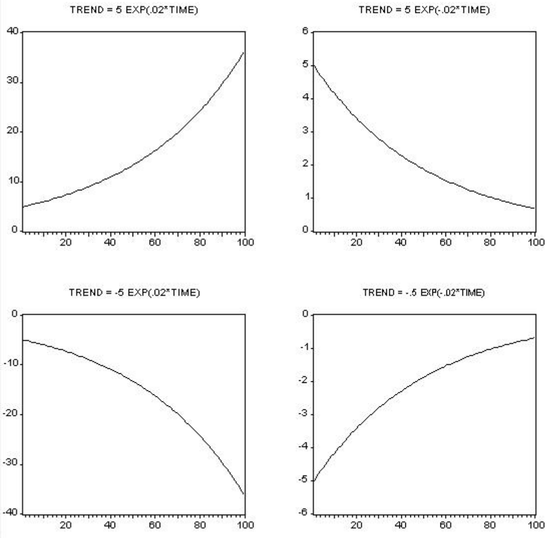


Figure: Various Exponential Trends

Non-Linear Trend: Quadratic

Allow for gentle curvature by including $\mathit{TIME\ and\ TIME}^2$:

$$Trend_t = \beta_1 + \beta_2 \, TIME_t + \beta_3 \, TIME_t^2$$

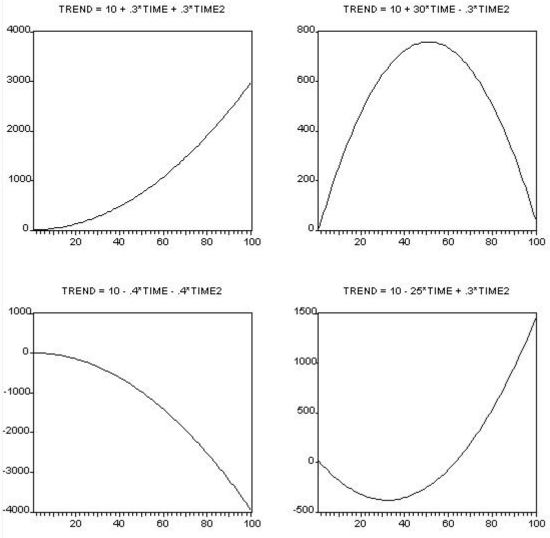


Figure: Various Quadratic Trends

Liquor Sales Quadratic Trend Estimation

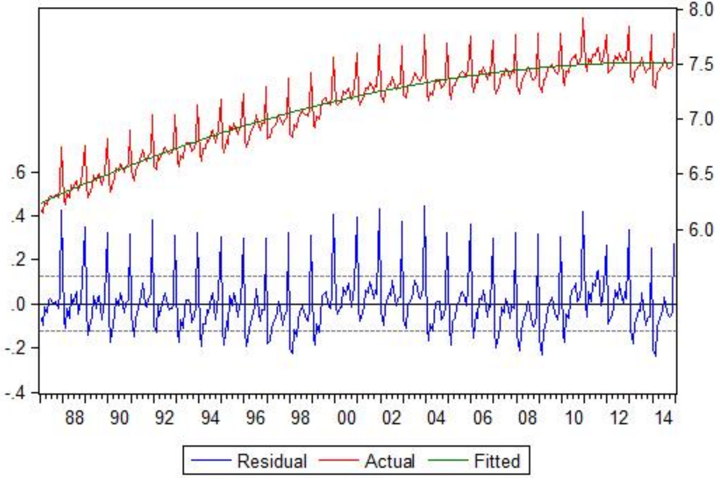
Dependent Variable: LSALES
Method: Least Squares
Date: 08/08/13   Time: 08:53
Sample: 1987M01 2014M12
Included observations: 336

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.231269	0.020653	301.7187	0.0000
TIME	0.007768	0.000283	27.44987	0.0000
TIME2	-1.17E-05	8.13E-07	-14.44511	0.0000

R-squared	0.903676	Mean dependent var	7.096188
Adjusted R-squared	0.903097	S.D. dependent var	0.402962
S.E. of regression	0.125439	Akaike info criterion	-1.305106
Sum squared resid	5.239733	Schwarz criterion	-1.271025
Log likelihood	222.2579	Hannan-Quinn criter.	-1.291521
F-statistic	1562.036	Durbin-Watson stat	1.754412
Prob(F-statistic)	0.000000		

Figure:

Residual Plot



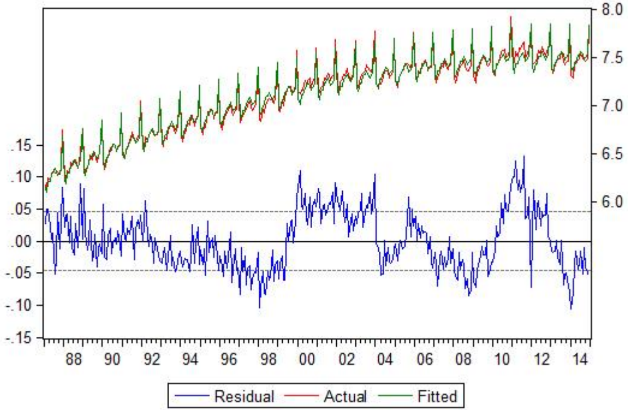
Liquor Sales Quadratic Trend Estimation with Seasonal Dummies

Dependent Variable: LSALES
Method: Least Squares
Date: 08/08/13 Time: 08:53
Sample: 1987M01 2014M12
Included observations: 336

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.007739	0.000104	74.49828	0.0000
TIME2	-1.18E-05	2.98E-07	-39.36756	0.0000
D1	6.138362	0.011207	547.7315	0.0000
D2	6.081424	0.011218	542.1044	0.0000
D3	6.168571	0.011229	549.3318	0.0000
D4	6.169584	0.011240	548.8944	0.0000
D5	6.238568	0.011251	554.5117	0.0000
D6	6.243596	0.011261	554.4513	0.0000
D7	6.287566	0.011271	557.8584	0.0000
D8	6.259257	0.011281	554.8647	0.0000
D9	6.199399	0.011290	549.0938	0.0000
D10	6.221507	0.011300	550.5987	0.0000
D11	6.253515	0.011309	552.9885	0.0000
D12	6.575648	0.011317	581.0220	0.0000

R-squared	0.987452	Mean dependent var	7.096188
Adjusted R-squared	0.986946	S.D. dependent var	0.402962
S.E. of regression	0.046041	Akaike info criterion	-3.277812
Sum squared resid	0.682555	Schwarz criterion	-3.118766
Log likelihood	564.6725	Hannan-Quinn criter.	-3.214412
Durbin-Watson stat	0.581383		

Residual Plot



Serial Correlation

Do we really believe that disturbances are
uncorrelated over time?
(Not possible in cross sections, so we didn't study it before...)

Serial Correlation is Another Type of Violation of the IC (This time it's "correlated disturbances".)

Consider: $\varepsilon \sim N(0, \Omega)$

Serial correlation is relevant in time-series environments.
It corresponds to non-diagonal Ω.
(Violates IC 1.6.)

Key cause: Omission of serially-correlated x's,
which produces serially-correlated ε

Serially Correlated Regression Disturbances

Disturbance serial correlation, or autocorrelation,
means *correlation over time*
– Current disturbance correlated with past disturbance(s)

Leading example
("$AR(1)$" disturbance serial correlation):

$$y_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t, \quad |\phi| < 1$$

$$v_t \sim iid\ N(0,\ \sigma^2)$$

(Extension to "$AR(p)$" disturbance serial correlation is immediate)

Consequences for β Estimation and Inference: As with Heteroskedasticity, Point Estimation is OK, but Inference is Damaged

– Esimation: OLS estimation of β remains largely OK.
Parameter estimates remain consistent and asymptotically normal

– Inference: OLS inference is damaged.
Standard errors are biased and inconsistent.

Consequences for y Prediction: Unlike With Heteroskedasticity, Even *Point* Predictions are Damaged/

Serial correlation is a bigger problem for prediction than heteroskedasticity.

Here's the intuition:

Serial correlation in disturbances/residuals implies that the included "x variables" have missed something that could be exploited for improved **point** *forecasting of y (and hence also improved interval and density forecasting). That is,* **all** *types of forecasts are sub-optimal when serial correlation is neglected.*

Put differently:
Serial correlation in forecast errors means that you can forecast your forecast errors! So something is wrong and can be improved...

Some Important Language and Tools For Characterizing Serial Correlation

"Autocovariances": $\gamma_{\varepsilon}(\tau) = cov(\varepsilon_t, \varepsilon_{t-\tau}), \ \tau = 1, 2, ...$

"Autocorrelations": $\rho_{\varepsilon}(\tau) = \gamma_{\varepsilon}(\tau)/\gamma_{\varepsilon}(0), \ \tau = 1, 2, ...$

"Partial autocorrelations": $p_{\varepsilon}(\tau), \ \tau = 1, 2, ...$

$p_{\varepsilon}(\tau)$ is the coefficient on $\varepsilon_{t-\tau}$ in the population regression:

$$\varepsilon_t \rightarrow c, \varepsilon_{t-1}, ..., \varepsilon_{t-(\tau-1)}, \varepsilon_{t-\tau}$$

Sample autocorrelations: $\hat{\rho}_{\varepsilon}(\tau) = \widehat{corr}(e_t, e_{t-\tau}), \ \tau = 1, 2, ...$

Sample partial autocorrelations: $\hat{p}_{\varepsilon}(\tau), \ \tau = 1, 2, ...$

$\hat{p}_{\varepsilon}(\tau)$ is the coefficient on $e_{t-\tau}$ in the finite-sample regression:

$$e_t \rightarrow c, e_{t-1}, ..., e_{t-(\tau-1)}, e_{t-\tau}$$

White Noise Disturbances

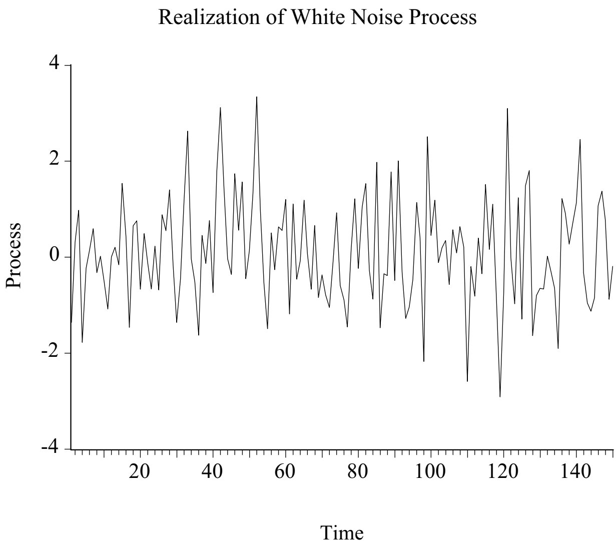
Zero-mean white noise: $\varepsilon_t \sim WN(0,\ \sigma^2)$ (serially uncorrelated)

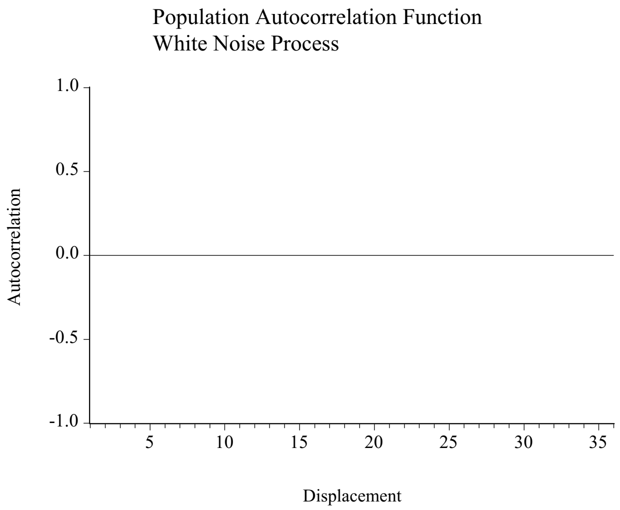
Independent (strong) white noise: $\varepsilon_t \overset{iid}{\sim} (0, \sigma^2)$

Gaussian white noise: $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$

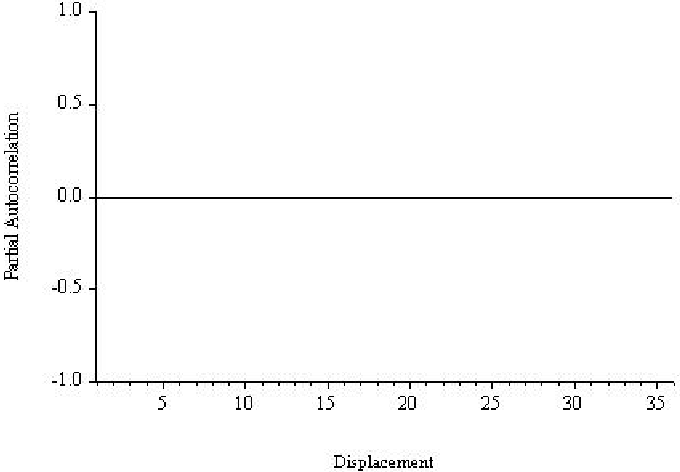
We write:

$$\varepsilon_t \sim WN(0,\ \sigma^2)$$


Realization of White Noise Process
4
2
0
-2
-4
Process
20
40
60
80
100
120
140
Time


Population Autocorrelation Function
White Noise Process
1.0
0.5
0.0
-0.5
-1.0
Autocorrelation
5
10
15
20
25
30
35
Displacement

Population Partial Autocorrelation Function
White Noise Process

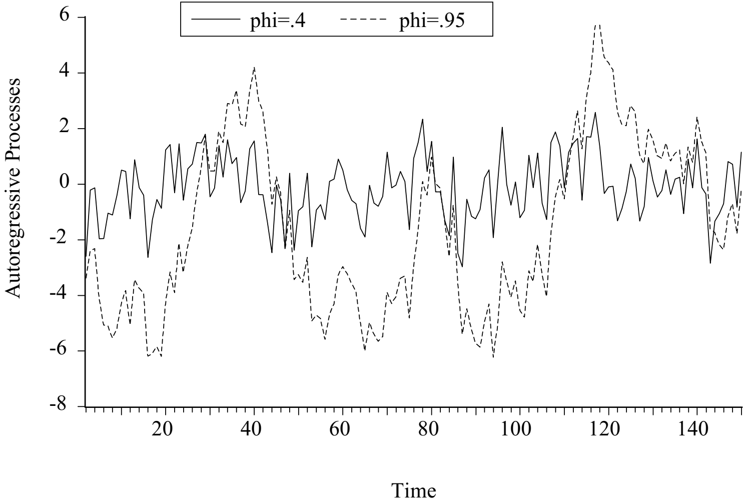


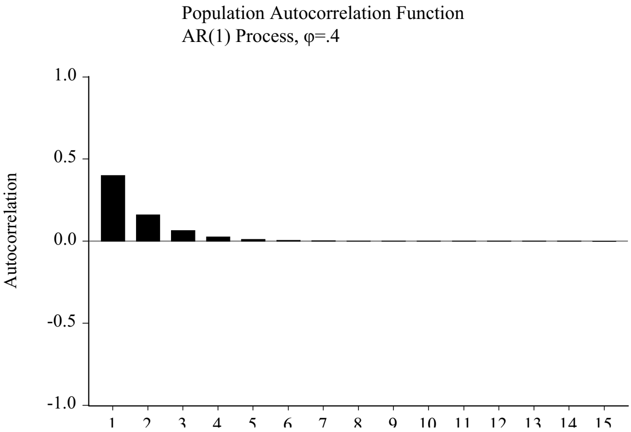
$AR(1)$ Disturbances

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t, \quad |\phi| < 1$$

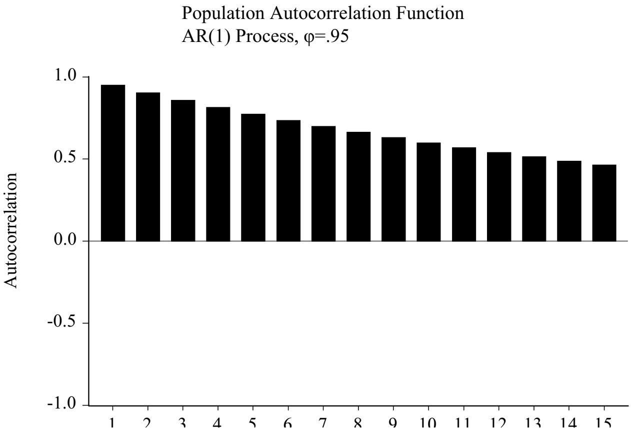
$$v_t \sim WN(0,\ \sigma^2)$$

Realizations of Two $AR(1)$ Processes ($N(0, 1)$ shocks)





$$\rho(\tau) = \phi^{\tau}$$



$$\rho(\tau) = \phi^\tau$$

Detecting Serial Correlation

- Graphical diagnostics
  - Residual plot
  - Residual scatterplot of (e_t vs. $e_{t-\tau}$)
  - Residual correlogram
- Formal tests
  - Durbin-Watson
  - Breusch-Godfrey

Recall Our Log-Quadratic Liquor Sales Model

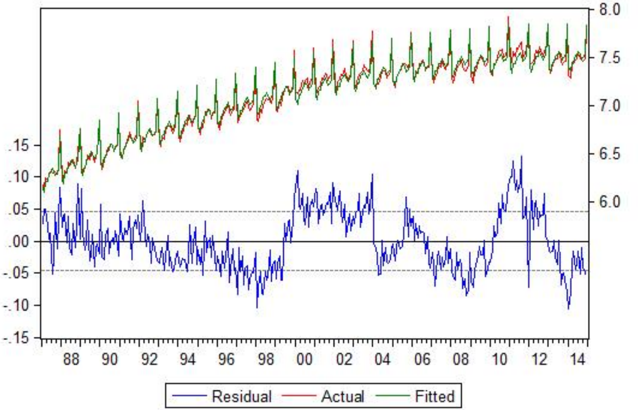
Included observations: 336

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.007739	0.000104	74.49828	0.0000
TIME2	-1.18E-05	2.98E-07	-39.36756	0.0000
D1	6.138362	0.011207	547.7315	0.0000
D2	6.081424	0.011218	542.1044	0.0000
D3	6.168571	0.011229	549.3318	0.0000
D4	6.169584	0.011240	548.8944	0.0000
D5	6.238568	0.011251	554.5117	0.0000
D6	6.243596	0.011261	554.4513	0.0000
D7	6.287566	0.011271	557.8584	0.0000
D8	6.259257	0.011281	554.8647	0.0000
D9	6.199399	0.011290	549.0938	0.0000
D10	6.221507	0.011300	550.5987	0.0000
D11	6.253515	0.011309	552.9885	0.0000
D12	6.575648	0.011317	581.0220	0.0000

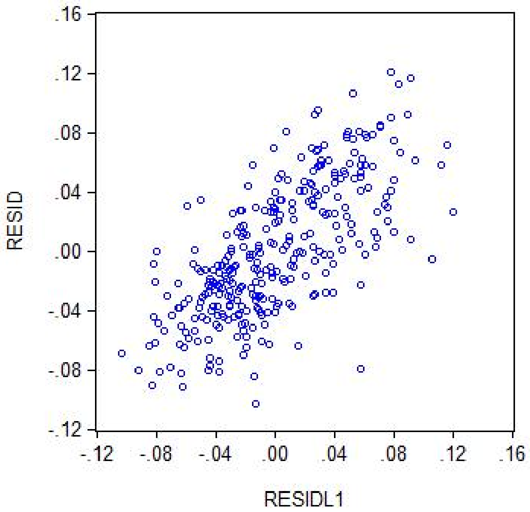
R-squared	0.987452	Mean dependent var	7.096188
Adjusted R-squared	0.986946	S.D. dependent var	0.402962
S.E. of regression	0.046041	Akaike info criterion	-3.277812
Sum squared resid	0.682555	Schwarz criterion	-3.118766
Log likelihood	564.6725	Hannan-Quinn criter.	-3.214412
Durbin-Watson stat	0.581383		

Figure: Liquor Sales Log-Quadratic Trend + Seasonal Estimation

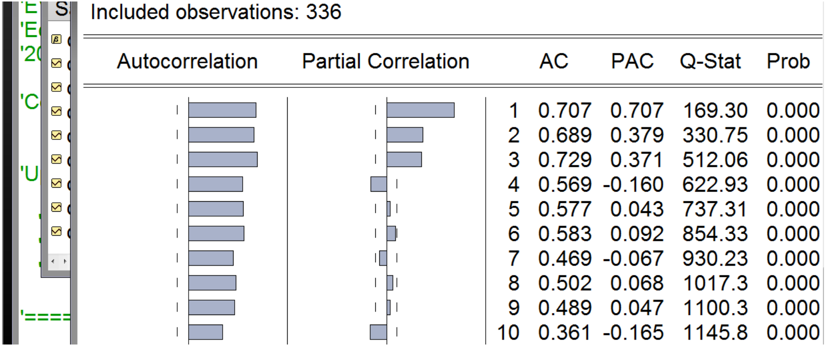
Residual Plot



Residual Scatterplot (e_t vs. e_{t-1})



Residual Correlogram



Included observations: 336

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.707	0.707	169.30	0.000
		2	0.689	0.379	330.75	0.000
		3	0.729	0.371	512.06	0.000
		4	0.569	-0.160	622.93	0.000
		5	0.577	0.043	737.31	0.000
		6	0.583	0.092	854.33	0.000
		7	0.469	-0.067	930.23	0.000
		8	0.502	0.068	1017.3	0.000
		9	0.489	0.047	1100.3	0.000
		10	0.361	-0.165	1145.8	0.000

$$\text{Bartlett standard error } (= 1/\sqrt{T}) = 1/\sqrt{336}) = .055$$

$$95\ \%\ \text{Bartlett band } (= \pm 2/\sqrt{T}) = \pm .11$$

Formal Tests: Durbin-Watson (0.59)

Simple $AR(1)$ environment:

$$y_t = x_t'\beta + \varepsilon_t$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + v_t$$

$$v_t \sim iid\ N(0,\ \sigma^2)$$

We want to test $H_0: \ \phi = 0$ against $H_1: \ \phi \neq 0$

Regress $y_t \to x_t$ and obtain the residuals e_t

Then form:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Understanding the Durbin-Watson Statistic

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} = \frac{\frac{1}{T}\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\frac{1}{T}\sum_{t=1}^{T} e_t^2}$$

$$= \frac{\frac{1}{T}\sum_{t=2}^{T} e_t^2 + \frac{1}{T}\sum_{t=2}^{T} e_{t-1}^2 - 2\ \frac{1}{T}\sum_{t=2}^{T} e_t e_{t-1}}{\frac{1}{T}\sum_{t=1}^{T} e_t^2}$$

Hence as $T \to \infty$:

$$DW \approx \frac{\sigma^2 + \sigma^2 - 2cov(\varepsilon_t, \varepsilon_{t-1})}{\sigma^2} = 1+1-2corr(\varepsilon_t, \varepsilon_{t-1}) = 2(1-corr(\varepsilon_t, \varepsilon_{t-1}))$$

$$\implies DW \in [0,4],\ DW \to 2 \text{ as } \phi \to 0, \text{ and } DW \to 0 \text{ as } \phi \to 1$$

Formal Tests: Breusch-Godfrey

General $AR(p)$ environment:

$$y_t = x_t'\beta + \varepsilon_t$$

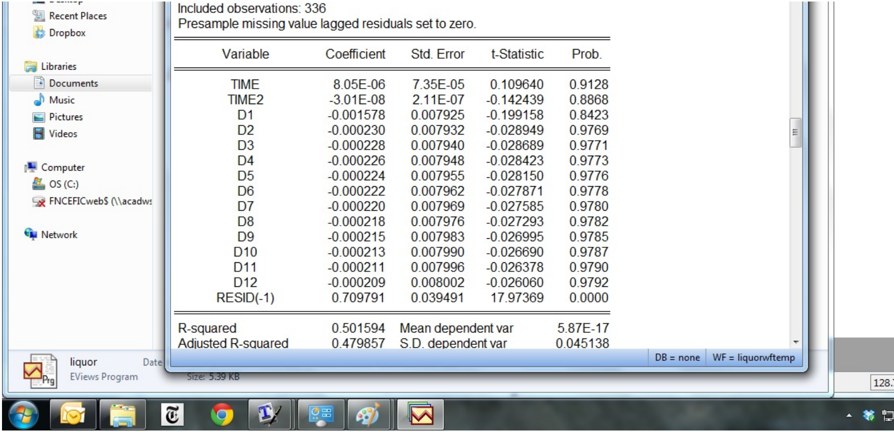
$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + ... + \phi_p \varepsilon_{t-p} + v_t$$

$$v_t \sim iid\ N(0,\ \sigma^2)$$

We want to test $H_0 : \ (\phi_1, ..., \phi_p) = 0$ against $H_1 : \ (\phi_1, ..., \phi_p) \neq 0$

- Regress $y_t \rightarrow x_t$ and obtain the residuals e_t
- Regress $e_t \rightarrow x_t, e_{t-1}, ..., e_{t-p}$
- Examine TR^2. In large samples $TR^2 \sim \chi_p^2$ under the null.

BG for $AR(1)$ Disturbances ($TR^2 = 168.5$, $p = 0.0000$)



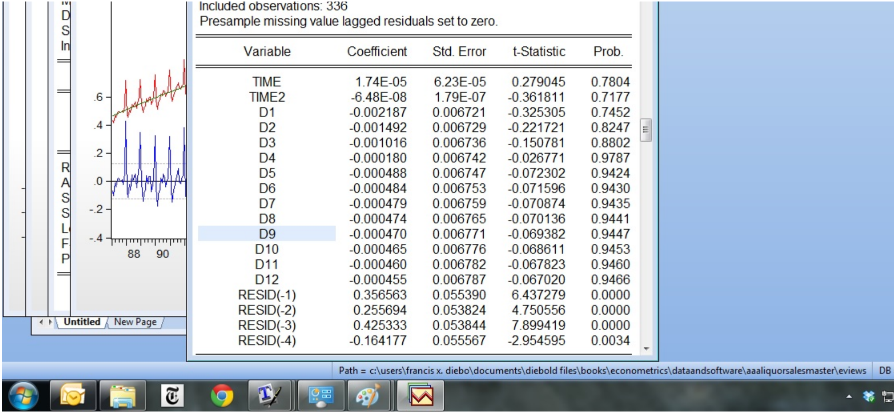
Included observations: 336
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	8.05E-06	7.35E-05	0.109640	0.9128
TIME2	-3.01E-08	2.11E-07	-0.142439	0.8868
D1	-0.001578	0.007925	-0.199158	0.8423
D2	-0.000230	0.007932	-0.028949	0.9769
D3	-0.000228	0.007940	-0.028689	0.9771
D4	-0.000226	0.007948	-0.028423	0.9773
D5	-0.000224	0.007955	-0.028150	0.9776
D6	-0.000222	0.007962	-0.027871	0.9778
D7	-0.000220	0.007969	-0.027585	0.9780
D8	-0.000218	0.007976	-0.027293	0.9782
D9	-0.000215	0.007983	-0.026995	0.9785
D10	-0.000213	0.007990	-0.026690	0.9787
D11	-0.000211	0.007996	-0.026378	0.9790
D12	-0.000209	0.008002	-0.026060	0.9792
RESID(-1)	0.709791	0.039491	17.97369	0.0000

R-squared	0.501594	Mean dependent var	5.87E-17
Adjusted R-squared	0.479857	S.D. dependent var	0.045138

Figure: BG Test Regression, $AR(1)$

BG for $AR(4)$ Disturbances ($TR^2 = 216.7$, $p = 0.0000$)



Included observations: 336
Presampe missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	1.74E-05	6.23E-05	0.279045	0.7804
TIME2	-6.48E-08	1.79E-07	-0.361811	0.7177
D1	-0.002187	0.006721	-0.325305	0.7452
D2	-0.001492	0.006729	-0.221721	0.8247
D3	-0.001016	0.006736	-0.150781	0.8802
D4	-0.000180	0.006742	-0.026771	0.9787
D5	-0.000488	0.006747	-0.072302	0.9424
D6	-0.000484	0.006753	-0.071596	0.9430
D7	-0.000479	0.006759	-0.070874	0.9435
D8	-0.000474	0.006765	-0.070136	0.9441
D9	-0.000470	0.006771	-0.069382	0.9447
D10	-0.000465	0.006776	-0.068611	0.9453
D11	-0.000460	0.006782	-0.067823	0.9460
D12	-0.000455	0.006787	-0.067020	0.9466
RESID(-1)	0.356563	0.055390	6.437279	0.0000
RESID(-2)	0.255694	0.053824	4.750556	0.0000
RESID(-3)	0.425333	0.053844	7.899419	0.0000
RESID(-4)	-0.164177	0.055567	-2.954595	0.0034

Figure: BG Test Regression, $AR(4)$

BG for $AR(8)$ Disturbances ($TR^2 = 219.0$, $p = 0.0000$)

D7	-0.000514	0.006738	-0.076323	0.9392
D8	-0.000689	0.006744	-0.102103	0.9187
D9	-0.000547	0.006748	-0.081066	0.9354
D10	-0.000542	0.006753	-0.080187	0.9361
D11	-0.000536	0.006759	-0.079289	0.9369
D12	-0.000530	0.006764	-0.078372	0.9376
RESID(-1)	0.369859	0.056379	6.560221	0.0000
RESID(-2)	0.246228	0.059906	4.110231	0.0001
RESID(-3)	0.360805	0.061281	5.887687	0.0000
RESID(-4)	-0.165236	0.064757	-2.551620	0.0112
RESID(-5)	0.000506	0.064799	0.007815	0.9938
RESID(-6)	0.097952	0.061505	1.592575	0.1123
RESID(-7)	-0.089030	0.060271	-1.477162	0.1406
RESID(-8)	0.071770	0.056707	1.265623	0.2066

R-squared	0.651708	Mean dependent var	5.87E-17
Adjusted R-squared	0.628414	S.D. dependent var	0.045138
S.E. of regression	0.027515	Akaike info criterion	-4.284907
Sum squared resid	0.237729	Schwarz criterion	-4.034977
Log likelihood	741.8643	Hannan-Quinn criter.	-4.185278
Durbin-Watson stat	2.000282		

Untitled New Page

Durbin-Watson stat 1.978691

Path = c:\users\francis x. diebo\documents\diebold files\books\econometrics\dataandsoftware\aaaliquorsalesmaster\eviews DB

Figure: BG Test Regression, $AR(8)$

Robust Estimation with Serial Correlation

Recall our earlier "heteroskedasticity robust s.e.'s"

We can also consider "serial correlation robust s.e.'s"

The simplest way is to include lags of y as regressors...

Modeling Serial Correlation: Including Lags of y as Regressors

Serial correlation in disturbances means that the included x's don't fully account for the y dynamics.

Simple to fix by modeling the y dynamics directly: Just include lags of y as additional regressors.

More precisely, $AR(p)$ disturbances "fixed" by including p lags of y and x. (Select p using the usual SIC, etc.)

Illustration:
Convert the DGP below to one with white noise disturbances.

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$$

$$\varepsilon_t = \phi \varepsilon_{t-1} + v_t$$

$$v_t \sim iid\ N(0,\ \sigma^2)$$

Liquor Sales: Everything Consistent With $AR(4)$ Dynamics

Trend+seasonal residual plot
Trend+seasonal residual scatterplot
Trend+seasonal DW
Trend+seasonal BG
Trend+seasonal residual correlogram

Also trend+seasonal+$AR(p)$ SIC:

$AR(1) = -3.797$
$AR(2) = -3.941$
$AR(3) = -4.080$
$AR(4) = -4.086$
$AR(5) = -4.071$
$AR(6) = -4.058$
$AR(7) = -4.057$
$AR(8) = -4.040$

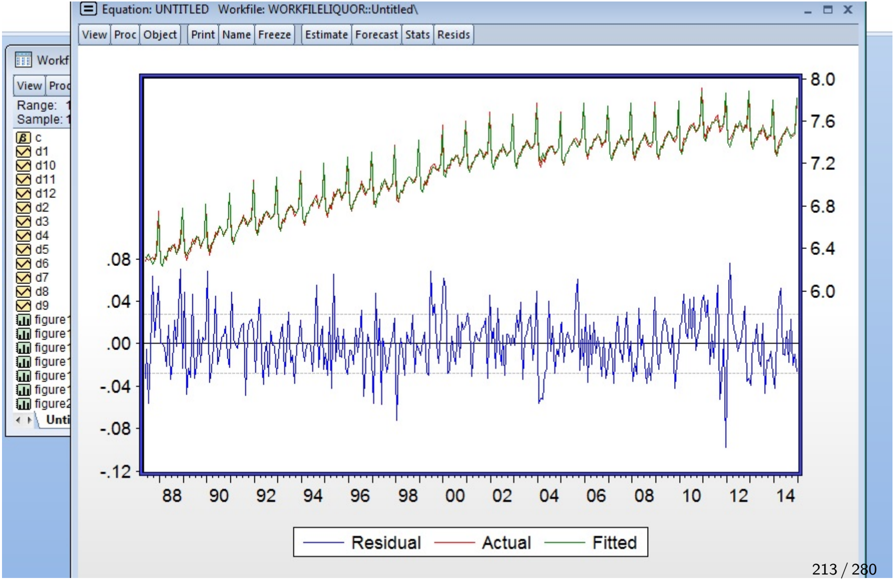
Trend + Seasonal Model with Four Lags of y

Dependent Variable: LSALES
Method: Least Squares
Date: 03/28/16 Time: 06:35
Sample (adjusted): 1987M05 2014M12
Included observations: 332 after adjustments

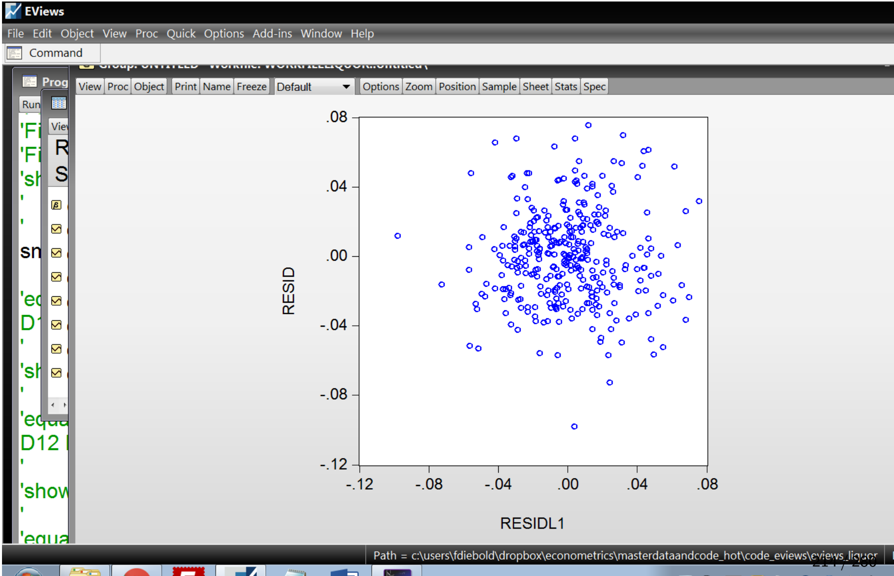
Variable	Coefficient	Std. Error	t-Statistic	Prob.
TIME	0.000993	0.000303	3.274924	0.0012
TIME2	-1.63E-06	4.82E-07	-3.379250	0.0008
D1	0.577182	0.240084	2.404080	0.0168
D2	0.579618	0.239820	2.416883	0.0162
D3	0.667059	0.238627	2.795401	0.0055
D4	0.894665	0.237447	3.767847	0.0002
D5	0.893728	0.232717	3.840401	0.0001
D6	0.827871	0.233806	3.540838	0.0005
D7	0.865982	0.235247	3.681158	0.0003
D8	0.791626	0.236419	3.348398	0.0009
D9	0.739295	0.237199	3.116777	0.0020
D10	0.771468	0.236858	3.257093	0.0012
D11	0.830449	0.236573	3.510331	0.0005
D12	1.156867	0.236231	4.897183	0.0000
LSALES(-1)	0.348107	0.055751	6.243965	0.0000
LSALES(-2)	0.257435	0.053823	4.783041	0.0000
LSALES(-3)	0.429234	0.053804	7.977784	0.0000
LSALES(-4)	-0.161633	0.055771	-2.898162	0.0040

R-squared	0.995335	Mean dependent var	7.107025
Adjusted R-squared	0.995082	S.D. dependent var	0.392974
S.E. of regression	0.027559	Akaike info criterion	-4.292292
Sum squared resid	0.238480	Schwarz criterion	-4.085990
Log likelihood	730.5205	Hannan-Quinn criter.	-4.210019
Durbin-Watson stat	1.982921		

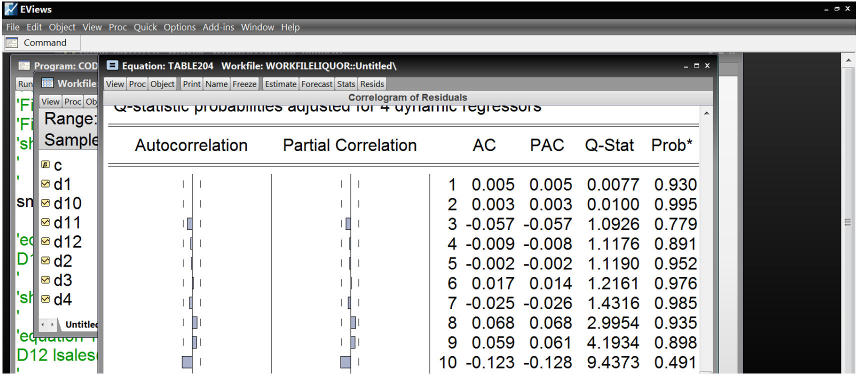
Trend + Seasonal Model with Four Lags of y Residual Plot



Trend + Seasonal Model with Four Lags of y
Residual Scatterplot

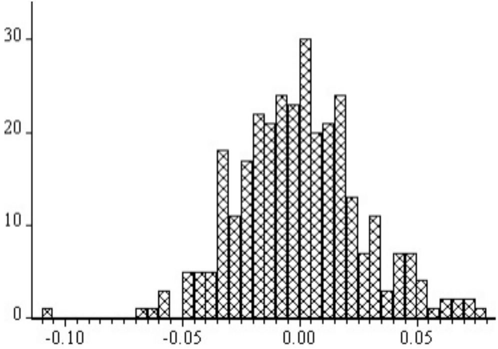


Trend + Seasonal Model with Four Lags of y
Residual Autocorrelations



	AC	PAC	Q-Stat	Prob*
1	0.005	0.005	0.0077	0.930
2	0.003	0.003	0.0100	0.995
3	-0.057	-0.057	1.0926	0.779
4	-0.009	-0.008	1.1176	0.891
5	-0.002	-0.002	1.1190	0.952
6	0.017	0.014	1.2161	0.976
7	-0.025	-0.026	1.4316	0.985
8	0.068	0.068	2.9954	0.935
9	0.059	0.061	4.1934	0.898
10	-0.123	-0.128	9.4373	0.491

Trend + Seasonal Model with Four Lags of y Residual Histogram and Normality Test



Observations 312	
Mean	3.77E-16
Median	-0.000160
Maximum	0.078468
Minimum	-0.109856
Std. Dev.	0.026635
Skewness	0.077911
Kurtosis	3.740378
Jarque-Bera	7.441714
Probability	0.024213

Forecasting Time Series

It's more interesting than in cross sections...

The "Forecasting the Right-Hand-Side Variables Problem"

For now assume known parameters.

$$y_t = x_t' \beta + \varepsilon_t$$

$$\Longrightarrow \; y_{t+h} = x_{t+h}' \beta + \varepsilon_{t+h}$$

Projecting on current information,

$$y_{t+h,t} = x_{t+h,t}' \beta$$

"Forecasting the right-hand-side variables problem" (FRVP):

We don't have $x_{t+h,t}$

But FRVP is not a Problem for Us!

FRVP no problem for trends. Why?

FRVP no problem for seasonals. Why?

FRVP also no problem for autoregressive effects
(lagged dependent variables)

e.g., consider a pure $AR(1)$

$$y_t = \phi y_{t-1} + \varepsilon_t$$

$$y_{t+h} = \phi y_{t+h-1} + \varepsilon_{t+h}$$

$$y_{t+h,t} = \phi y_{t+h-1,t}$$

No FRVP for $h = 1$. There seems to be an FRVP for $h > 1$.
But there's not...

We build the multi-step forecast recursively.
First 1-step, then 2-step, etc.
"Wold's chain rule of forecasting"

Interval and Density Forecasting

Assuming Gaussian shocks, we immediately have

$$y_{t+h} \mid y_t, y_{t-1}, \ldots \ \sim \ N(y_{t+h,t}, \ \sigma^2_{t+h,t}).$$

We know how to get $y_{t+h,t}$ (Wold's chain rule).

The question is how to get

$$\sigma^2_{t+h,t} = var(e_{t+h,t}) = var(y_{t+h} - y_{t+h,t}).$$

[Of course to make things operational we eventually replace parameters with estimates and use $N(\hat{y}_{t+h,t}, \ \hat{\sigma}^2_{t+h,t})$.]

Interval and Density Forecasting (1-Step-Ahead, AR(1))

$$y_t = \phi y_{t-1} + \varepsilon_t \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Back substitution yields

$$y_t = \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + ...$$

$$\Longrightarrow y_{t+1} = \varepsilon_{t+1} + \phi\varepsilon_t + \phi^2\varepsilon_{t-1} + \phi^3\varepsilon_{t-2} + ...$$

Projecting y_{t+1} on time-t information $(\varepsilon_t, \ \varepsilon_{t-1}, ...)$ gives:

$$y_{t+1,t} = \phi\varepsilon_t + \phi^2\varepsilon_{t-1} + \phi^3\varepsilon_{t-2} + ...$$

Corresponding 1-step-ahead error (zero-mean, unforecastable):

$$e_{t+1,t} = y_{t+1} - y_{t+1,t} = \varepsilon_{t+1}$$

with variance

$$\sigma^2_{t+1,t} = var(e_{t+1,t}) = \sigma^2$$

Interval and Density Forecasting (h-Step-Ahead, $AR(p)$)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \qquad \varepsilon_t \sim WN(0, \sigma^2)$$

Back substitution yields

$$y_t = \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + b_3 \varepsilon_{t-3} + ...$$

$$\implies y_{t+h} = \varepsilon_{t+h} + b_1 \varepsilon_{t+h-1} + ... + b_{h-1} \varepsilon_{t+1} + b_h \varepsilon_t + b_{h+1} \varepsilon_{t-1} + ...$$

(Note that the $b's$ are functions of the $\phi's$.)

Projecting y_{t+h} on time-t information $(\varepsilon_t, \ \varepsilon_{t-1}, ...)$ gives:

$$y_{t+h,t} = b_h \varepsilon_t + b_{h+1} \varepsilon_{t-1} + b_{h+2} \varepsilon_{t-2} + ...$$

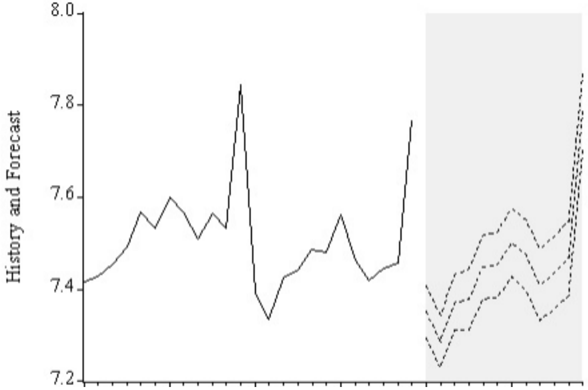
Corresponding h-step-ahead error (zero-mean, unforecastable):

$$e_{t+h,t} = y_{t+h} - y_{t+h,t} = \varepsilon_{t+h} + b_1 \varepsilon_{t+h-1} + ... + b_{h-1} \varepsilon_{t+1}$$

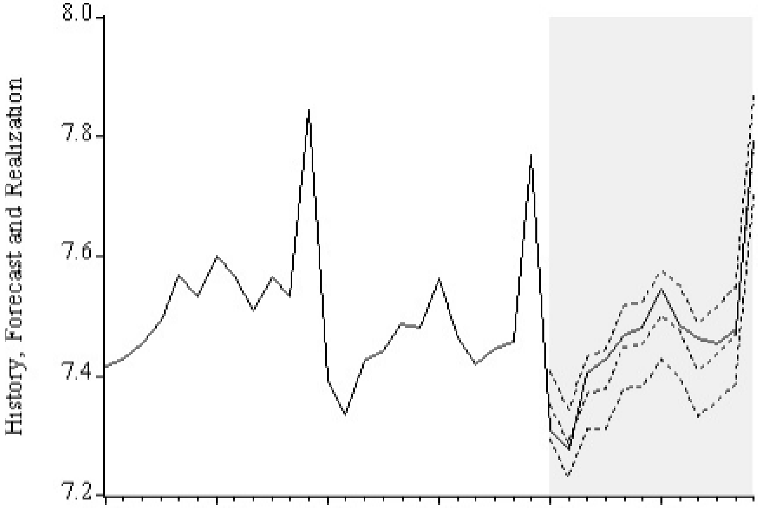
with variance (non-decreasing in h):

$$\sigma^2_{t+h,t} = var(e_{t+h,t}) = \sigma^2 (1 + b_1^2 + ... + b_{h-1}^2)$$

Liquor Sales History and 1- Through 12-Month-Ahead Point and Interval Forecasts From Trend + Seasonal Model with Four Lags of y



Now With Realization Superimposed...



SIC Estimates of Out-of-Sample Forecast Error Variance)

$$LSALES_t \rightarrow c, TIME_t$$
$$SIC = 0.45$$

$$LSALES_t \rightarrow c, TIME_t, TIME_t^2$$
$$SIC = 0,28$$

$$LSALES_t \rightarrow TIME_t, TIME_t^2, D_{t,1}, ..., D_{t,12}$$
$$SIC = 0.04$$

$$LSALES_t \rightarrow TIME_t, TIME_t^2, D_{t,1}, ..., D_{t,12}, LSALES_{t-1}, ..., LSALES_{t-4}$$
$$SIC = 0.02$$

(We report exponentiated SIC's because the software actually reports $ln(SIC)$)

Structural Change in Time Series: Evolution or Breaks in Any or all Parameters

Do we really believe that parameters are fixed over time?

Structural Change
Single Sharp Break, Exogenously Known

For simplicity of exposition, consider a bivariate regression:

$$y_t = \begin{cases} \beta_1^1 + \beta_2^1 x_t + \varepsilon_t, \;\; i = 1, ..., T^* \\ \beta_1^2 + \beta_2^2 x_t + \varepsilon_t, \;\; t = T^* + 1, ..., T \end{cases}$$

Let

$$D_t = \begin{cases} 0, \;\; t = 1, ..., T^* \\ 1, \;\; t = T^* + 1, ... T \end{cases}$$

Then we can write the model as:

$$y_t = (\beta_1^1 + (\beta_1^2 - \beta_1^1) D_t) + (\beta_2^1 + (\beta_2^2 - \beta_2^1) D_t) x_t + \varepsilon_t$$

We run:

$$y_t \rightarrow c, \;\; D_t, \;\; x_t, \;\; D_t \cdot x_t$$

Use regression to test for structural change (F test)
Use regression to accommodate structural change if present.

Structural Change
Single Sharp Break, Exogenously Known, Continued

The "Chow test" is what we're really calculating:

$$Chow = \frac{(e'e - (e_1'e_1 + e_2'e_2))/K}{(e_1'e_1 + e_2'e_2)/(T - 2K)}$$

Distributed F under the no-break null (and the rest of the IC)

Structural Change
Sharp Breakpoint, Endogenously Identified

$$MaxChow = \max_{\tau_{min} \leq \delta \leq \tau_{max}} Chow(\delta),$$

where δ denotes potential break location as fraction of sample

(e.g., we might take $\delta_{min} = .15$ and $\delta_{max} = .85$)

The null distribution of *MaxChow* has been tabulated.

Recursive Parameter Estimates

For generic parameter β, calculate and examine

$$\hat{\beta}_{1:t}$$

for $t = 1, ..., T$

– Note that you have to leave room for startup.
That is, you can't really start at $t = 1$.
Why?

Recursive Residuals

At each t, $t = 1, ..., T-1$ (leaving room for startup), compute a 1-step forecast,

$$\hat{y}_{t+1,t} = \sum_{k=1}^{K} \hat{\beta}_{k,1:t} x_{k,t+1}$$

The corresponding forecast errors, or recursive residuals, are

$$\hat{e}_{t+1,t} = y_{t+1} - \hat{y}_{t+1,t}$$

Under the IC (including structural stability),

$$\hat{e}_{t+1,t} \sim N(0, \sigma^2 r_{t+1,t})$$

where $r_{t+1,t} = 1 + x'_{t+1}(X'_t X'_t)^{-1} x_{t+1}$

Standardized Recursive Residuals and CUSUM

$$\hat{w}_{t+1,t} \equiv \frac{\hat{e}_{t+1,t}}{\sigma\sqrt{r_{t+1,t}}},$$

$t = 1, ..., T-1$ (leaving room for startup)

Under the IC,

$$\hat{w}_{t+1,t} \sim iidN(0,1).$$

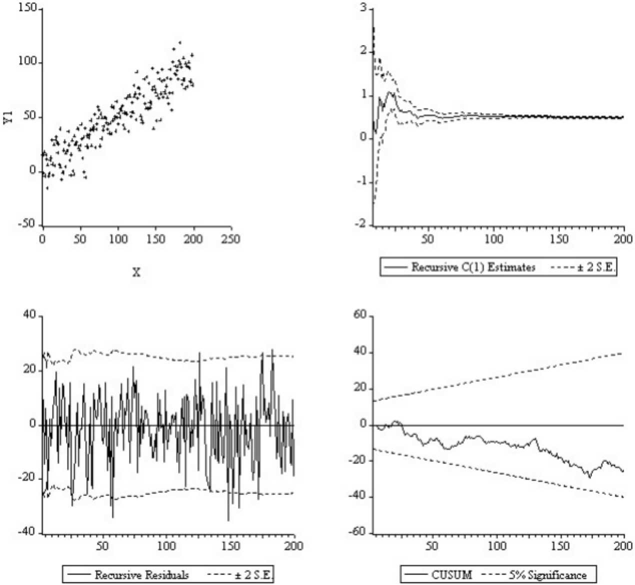
Then

$$CUSUM_{t^*} \equiv \sum_{t=1}^{t^*} w_{t+1,t}, \;\; t^* = 1, ..., T-1$$

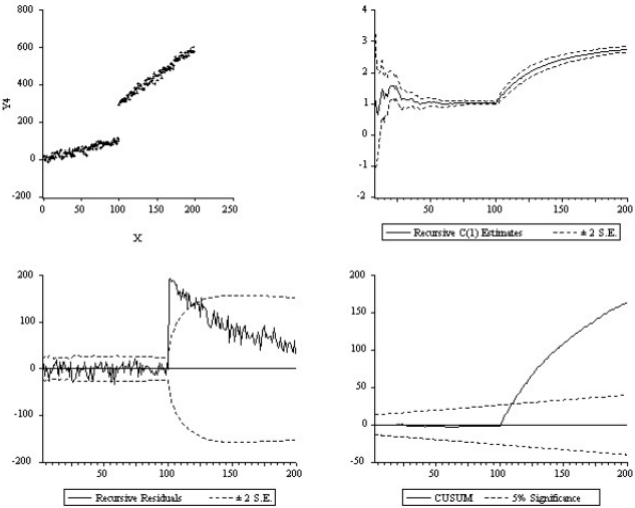
(leaving room for startup)

is just a sum of *iid* $N(0,1)$'s,
and its 95% bounds have been tabulated

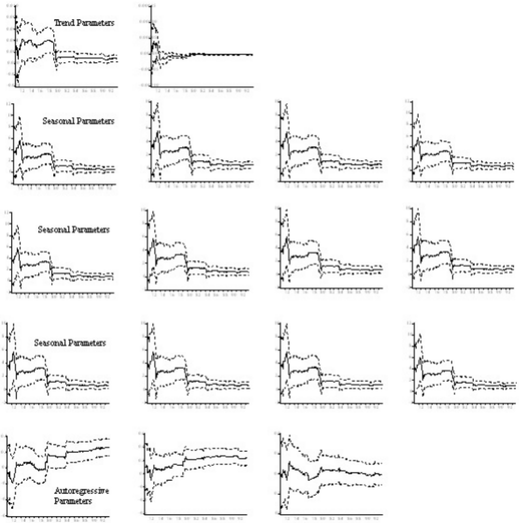
Recursive Analysis, Constant-Parameter DGP



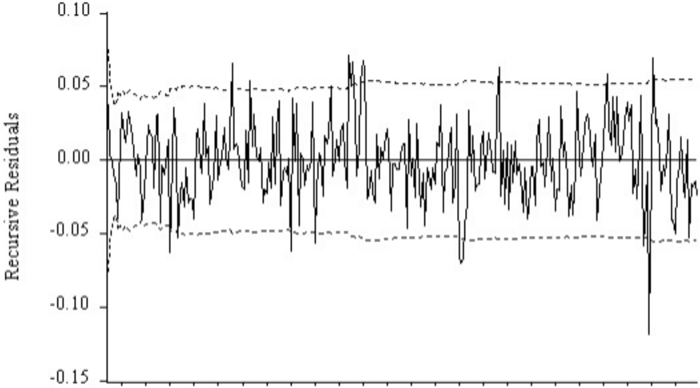
Recursive Analysis, Breaking-Parameter DGP



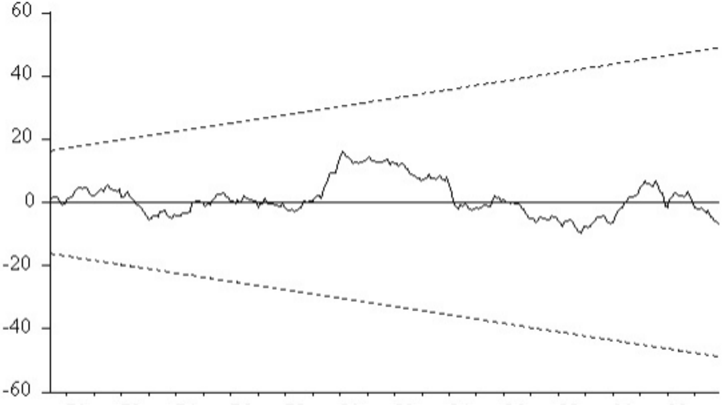
Liquor Sales Model: Recursive Parameter Estimates



Liquor Sales Model: Recursive Residuals With Two Standard Error Bands



Liquor Sales Model: CUSUM



Vector Autoregressions

What if we have more than one time series?

Basic Framework

e.g., bivariate (2-variable) *VAR*(1)

$$
\begin{aligned}
y_{1,t} &= c_1 + \phi_{11} y_{1,t-1} \; + \; \phi_{12} y_{2,t-1} \; + \; \varepsilon_{1,t} \\
y_{2,t} &= c_2 + \phi_{21} y_{1,t-1} \; + \; \phi_{22} y_{2,t-1} \; + \; \varepsilon_{2,t}
\end{aligned}
$$

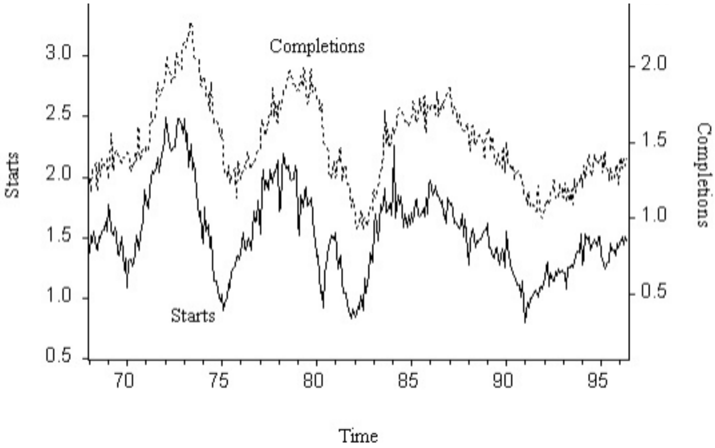
$$\varepsilon_{1,t} \sim WN(0, \; \sigma_1^2)$$

$$\varepsilon_{2,t} \sim WN(0, \; \sigma_2^2)$$

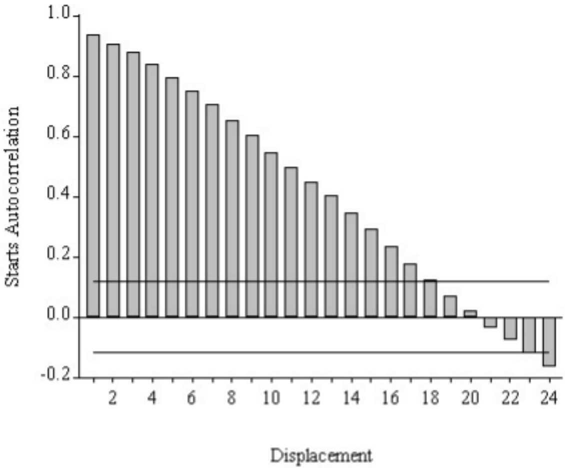
$$cov(\varepsilon_{1,t}, \; \varepsilon_{2,t}) = \sigma_{12}$$

- Can extend to N-variable $VAR(p)$
- Estimation by OLS (as before)
- Can include trends, seasonals, etc. (as before)
- Forecasts via Wold's chain rule (as before)
- Order selection by information criteria (as before)
- Can do predictive causality analysis (coming)

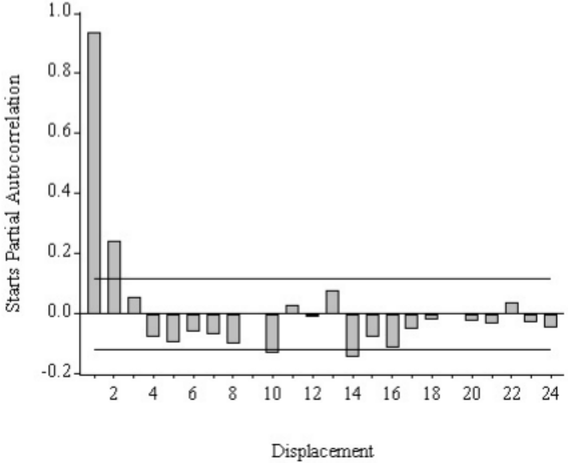
U.S. Housing Starts and Completions, 1968.01-1996.06



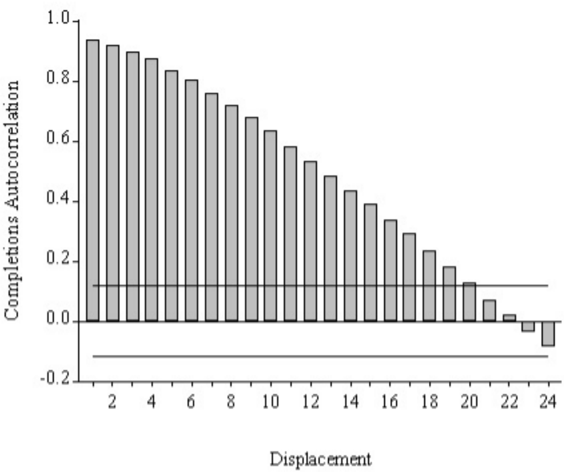
Starts Sample Autocorrelations



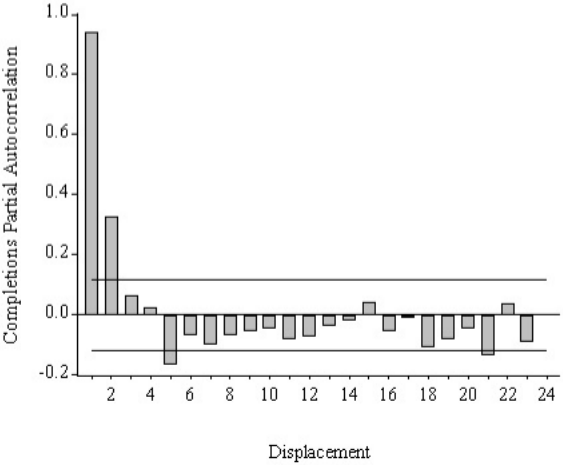
Starts Sample Partial Autocorrelations



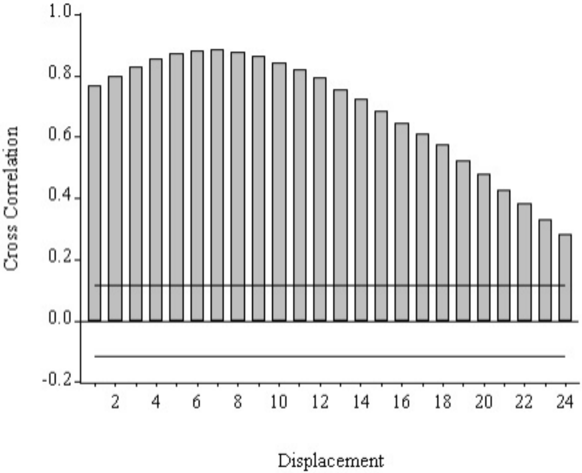
Completions Sample Autocorrelations



Completions Sample Partial Autocorrelations



Starts and Completions: Sample Cross Correlations



VAR Starts Equation

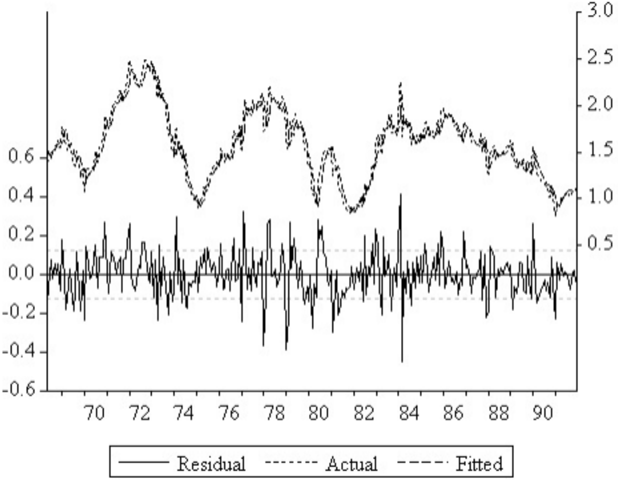
VAR Starts Equation

LS // Dependent Variable is STARTS
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

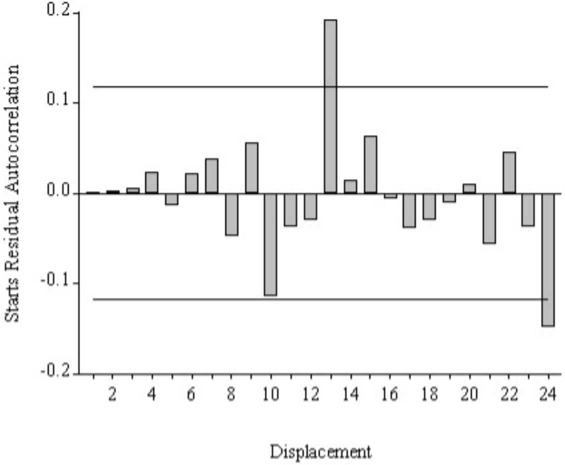
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.146871	0.044235	3.320264	0.0010
STARTS(-1)	0.659939	0.061242	10.77587	0.0000
STARTS(-2)	0.229632	0.072724	3.157587	0.0018
STARTS(-3)	0.142859	0.072655	1.966281	0.0503
STARTS(-4)	0.007806	0.066032	0.118217	0.9060
COMPS(-1)	0.031611	0.102712	0.307759	0.7585
COMPS(-2)	-0.120781	0.103847	-1.163069	0.2458
COMPS(-3)	-0.020601	0.100946	-0.204078	0.8384
COMPS(-4)	-0.027404	0.094569	-0.289779	0.7722

R-squared	0.895566	Mean dependent var	1.574771
Adjusted R-squared	0.892528	S.D. dependent var	0.382362
S.E. of regression	0.125350	Akaike info criterion	-4.122118
Sum squared resid	4.320952	Schwarz criterion	-4.006482
Log likelihood	191.3622	F-statistic	294.7796
Durbin-Watson stat	1.991908	Prob(F-statistic)	0.000000

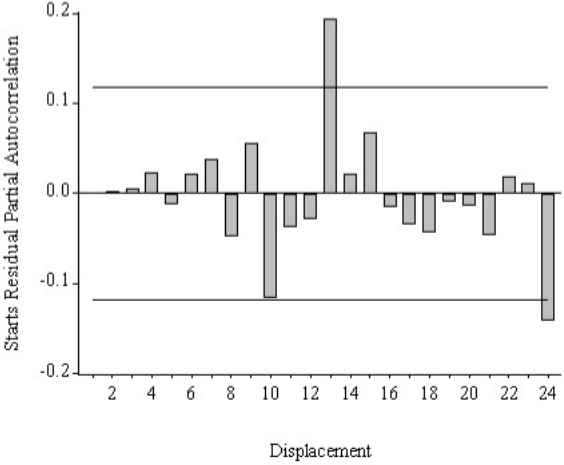
VAR Starts Equations Residual Plot



VAR Starts Equation Residual Sample Autocorrelations



VAR Starts Equation Residual Sample Partial Autocorrelations



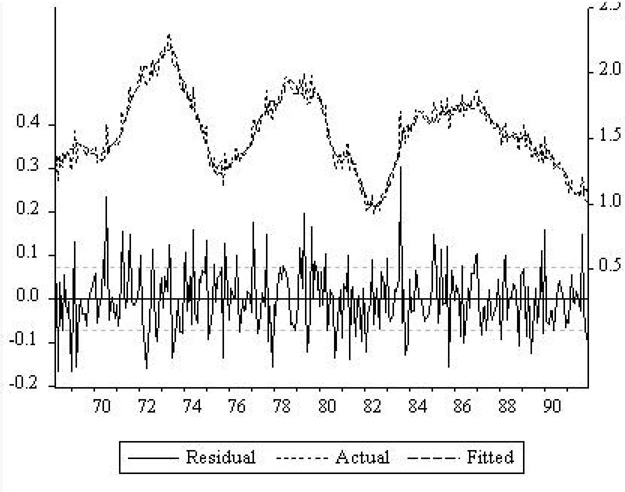
VAR Completions Equation

LS // Dependent Variable is COMPS
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

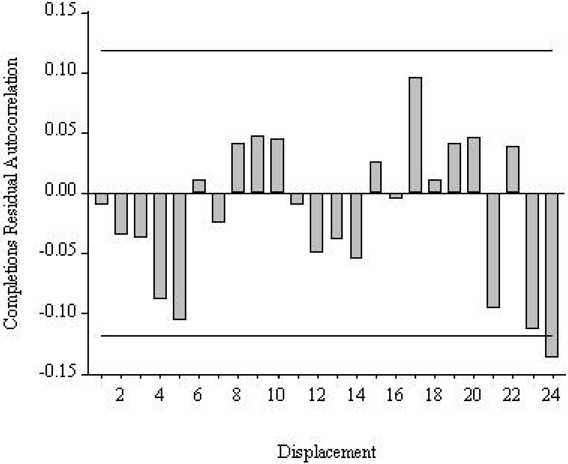
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.045347	0.025794	1.758045	0.0799
STARTS(-1)	0.074724	0.035711	2.092461	0.0373
STARTS(-2)	0.040047	0.042406	0.944377	0.3458
STARTS(-3)	0.047145	0.042366	1.112805	0.2668
STARTS(-4)	0.082331	0.038504	2.138238	0.0334
COMPS(-1)	0.236774	0.059893	3.953313	0.0001
COMPS(-2)	0.206172	0.060554	3.404742	0.0008
COMPS(-3)	0.120998	0.058863	2.055593	0.0408
COMPS(-4)	0.156729	0.055144	2.842160	0.0048

R-squared	0.936835	Mean dependent var	1.547958
Adjusted R-squared	0.934998	S.D. dependent var	0.286689
S.E. of regression	0.073093	Akaike info criterion	-5.200872
Sum squared resid	1.469205	Schwarz criterion	-5.085236
Log likelihood	344.5453	F-statistic	509.8375
Durbin-Watson stat	2.013370	Prob(F-statistic)	0.000000

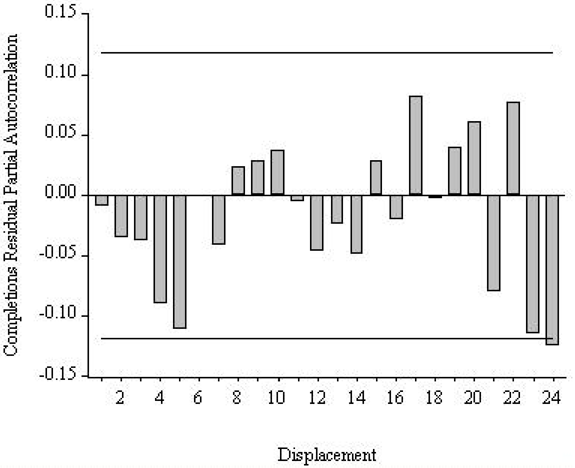
VAR Completions Equation Residual Plot



VAR Completions Equation Residual Sample Autocorrelations



VAR Completions Equation Residual Sample Partial Autocorrelations



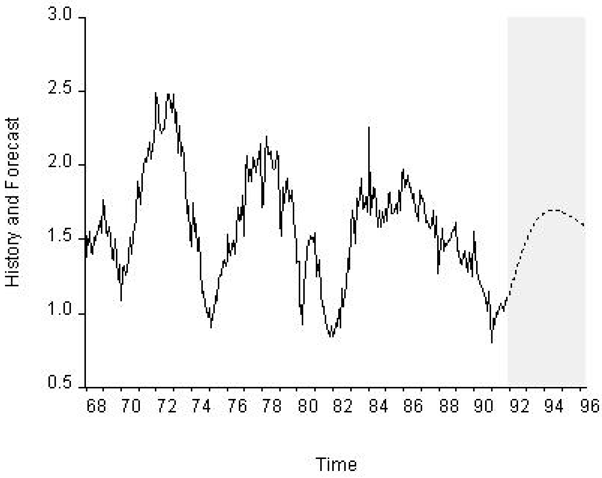
Predictive Causality Analysis

Table 8
Housing Starts and Completions
Causality Tests

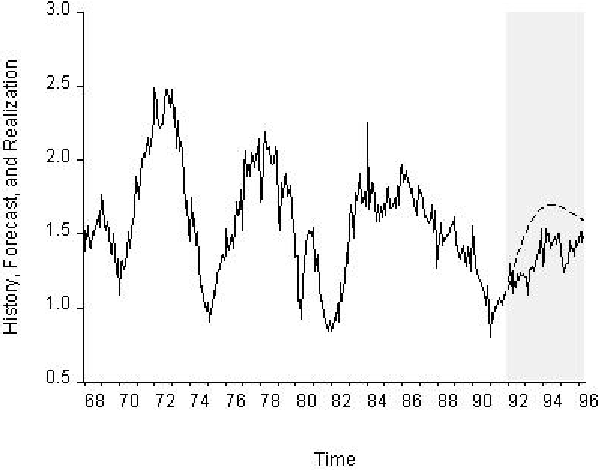
Sample: 1968:01 1991:12
Lags: 4
Obs: 284

Null Hypothesis:	F-Statistic	Probability
STARTS does not Cause COMPS	26.2658	0.00000
COMPS does not Cause STARTS	2.23876	0.06511

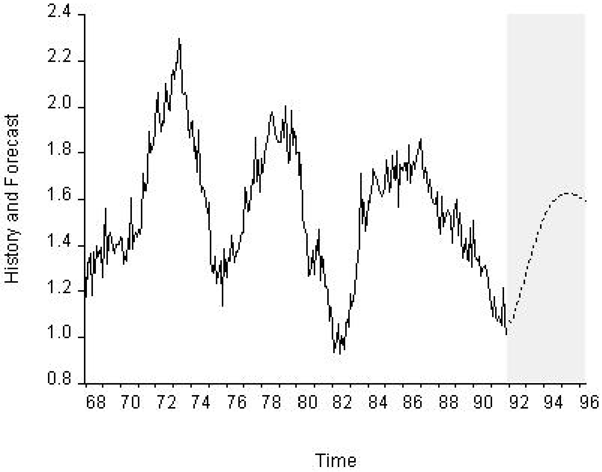
Starts History and Forecast



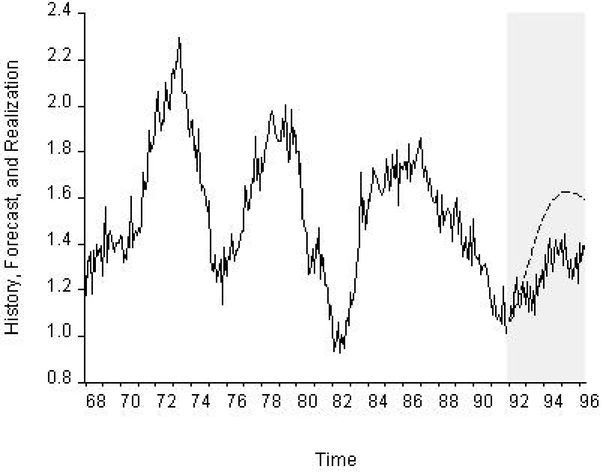
…Now With Starts Realization



Completions History and Forecast



...Now With Completions Realization



Heteroskedasticity in Time Series

Do we really believe that
disturbance variances are constant over time?

Dynamic Volatility is the Key to Finance and Financial Economics

- Risk management
- Portfolio allocation
- Asset pricing
- Hedging
- Trading

Financial Asset Returns

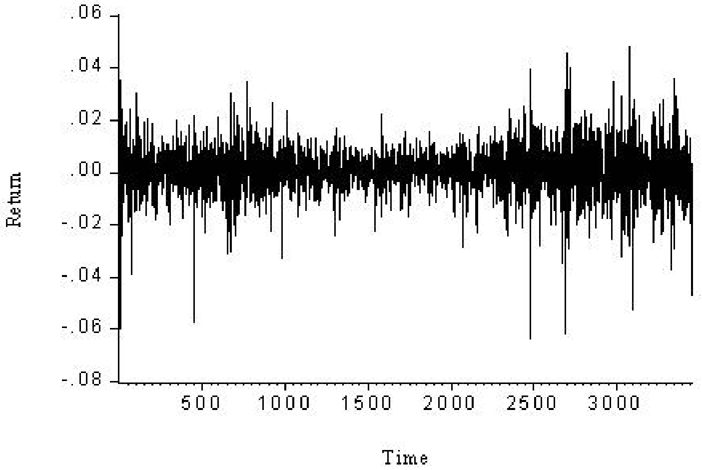


Figure: Time Series of Daily NYSE Returns.

Returns are Approximately Serially Uncorrelated

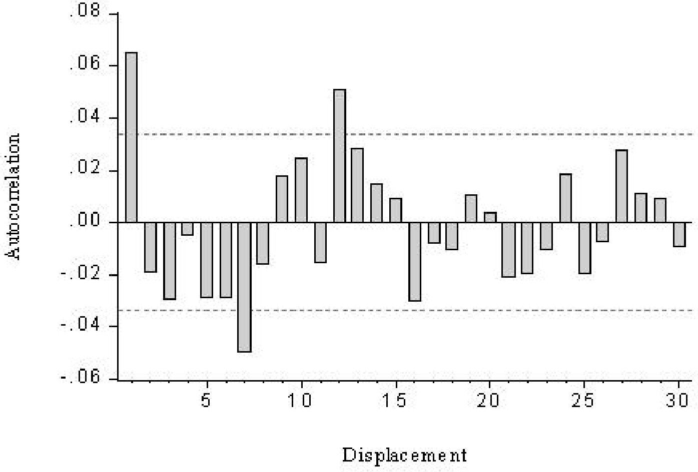


Figure: Correlogram of Daily NYSE Returns.

So returns are approximately white noise. But...

Returns are not Unconditionally Gaussian...

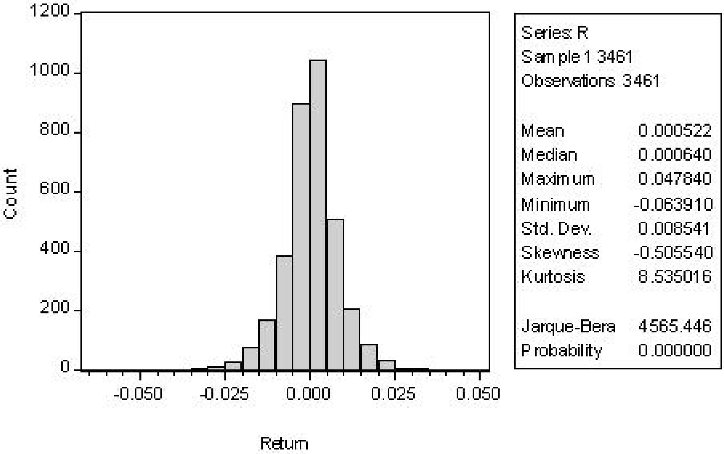


Figure: Histogram and Statistics for Daily NYSE Returns.

Unconditional Volatility Measures

Variance: $\sigma^2 \ = \ E(r_t - \mu)^2$ (or standard deviation: σ)

Kurtosis: $K \ = \ E(r - \mu)^4/\sigma^4$

Mean Absolute Deviation: $MAD \ = \ E|r_t - \mu|$

Interquartile Range: $IQR \ = \ 75\% \ - \ 25\%$

Outlier probability: $P|r_t - \mu| \ > \ 5\sigma$ (for example)

... Returns are Not Homoskedastic

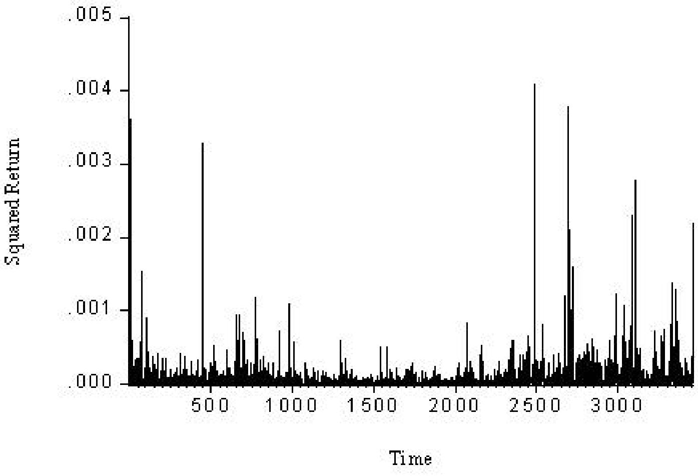


Figure: Time Series of Daily Squared NYSE Returns.

Indeed Returns are Highly Conditionally Heteroskedastic...

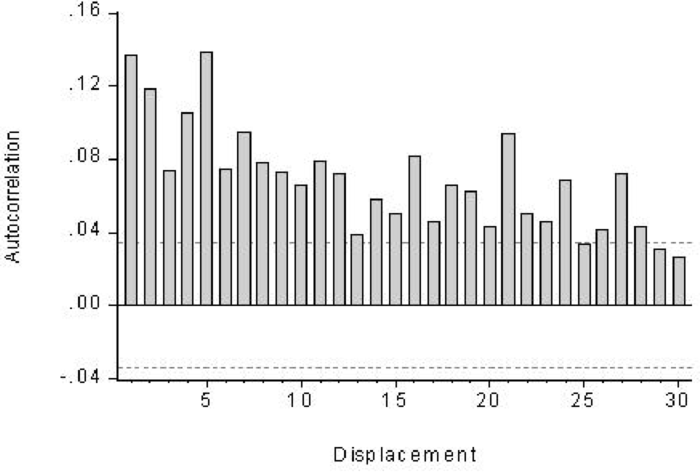


Figure: Correlogram of Daily Squared NYSE Returns.

Standard Models (e.g., $AR(1)$) Fail to Capture the Conditional Heteroskedasticity...

$$r_t = \phi r_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iidN(0,\ \sigma^2)$$

Equivalently, $r_t | \Omega_{t-1} \sim N(\phi r_{t-1}, \sigma^2)$

Conditional mean:
$E(r_t \mid \Omega_{t-1}) = \phi r_{t-1}$ (varies)

Conditional variance:
$var(r_t \mid \Omega_{t-1}) = \sigma^2$ (constant)

...So Introduce Special Heteroskedastic Disturbances

$$r_t = \phi r_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iidN(0,\ \sigma_t^2)$$

$$\text{Equivalently, } r_t \mid \Omega_{t-1} \sim N(\phi r_{t-1},\ \sigma_t^2)$$

Now consider:

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$$

$$\omega > 0, \quad \alpha \geq 0, \quad \beta \geq 0, \quad \alpha + \beta < 1$$

"GARCH(1,1) Process"

$$E(r_t | \Omega_{t-1}) = \phi r_{t-1} \quad (varies)$$

$$var(r_t \mid \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (varies)$$

For modeling daily asset returns we can simply use:

$$r_t \mid \Omega_{t-1} \sim N(0,\ \sigma_t^2)$$

GARCH(1,1) and "Exponential Smoothing"

GARCH(1,1):

$$\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$$

Solving backward:

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha \sum_{j=1}^{\infty} \beta^{j-1} r_{t-j}^2$$

Unified Framework

- Conditional variance dynamics (of course, by construction)
- Conditional variance dynamics produce unconditional leptokurtosis, even in our conditionally Gaussian setup
  (So conditional variance dynamics and unconditional fat tails are intimately related)
- Returns are non-Gaussian weak white noise
  (Serially uncorrelated but nevertheless dependent, due to conditional variance dynamics – today's conditional variance depends on the past.)

Extension: Regression with GARCH Disturbances (GARCH-M)

Standard GARCH regression:

$$r_t = x_t' \beta + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

GARCH-in mean (GARCH-M) regression:

$$r_t = x_t' \beta + \gamma \sigma_t + \varepsilon_t$$

$$\varepsilon_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

Extension: Fat-Tailed Conditional Densities (t-GARCH)

If r is conditionally Gaussian, then

$$r_t|\Omega_{t-1} = \ N(0, \sigma_t^2)$$

or

$$\frac{r_t}{\sigma_t} \sim iid\ N(0, 1)$$

But often with high-frequency data,

$$\frac{r_t}{\sigma_t} \sim iid\ fat{-}tailed$$

So take:

$$\frac{r_t}{\sigma_t} \sim iid\ \frac{t_d}{std(t_d)}$$

and treat d as another parameter to be estimated

Extension: Asymmetric Response and the Leverage Effect (Threshold GARCH)

$$\text{Standard GARCH: } \sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$$

$$\text{Threshold GARCH: } \sigma_t^2 = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 D_{t-1} + \beta \sigma_{t-1}^2$$

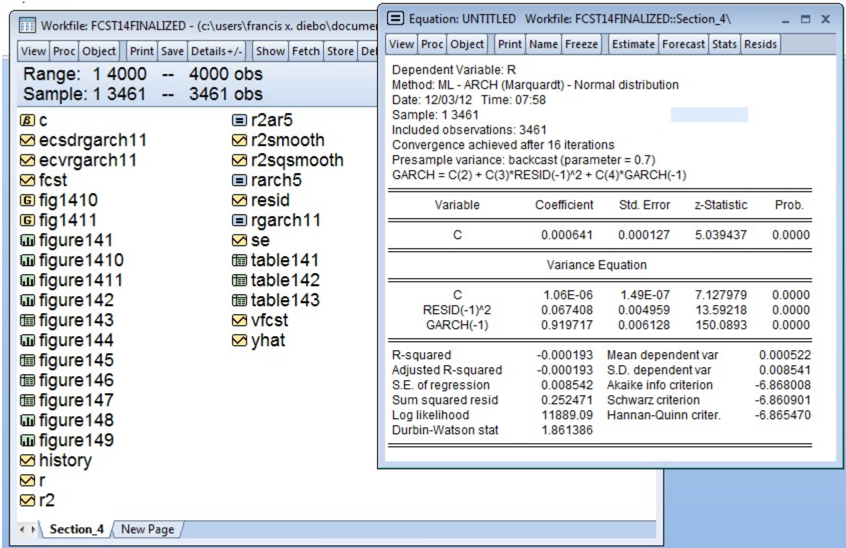
$$D_t = \begin{cases} 1 \textit{ if } \ r_t < 0 \\ 0 \textit{ otherwise} \end{cases}$$

positive return (good news): α effect on volatility

negative return (bad news): $\alpha + \gamma$ effect on volatility

$\gamma \neq 0$: Asymetric news response
$\gamma > 0$: "Leverage effect"

GARCH(1,1) MLE for Daily NYSE Returns



"Fancy" GARCH(1,1) MLE

Dependent Variable: R
Method: ML - ARCH (Marquardt) - Student's t distribution
Date: 04/10/12 Time: 13:48
Sample (adjusted): 2 3461
Included observations: 3460 after adjustments
Convergence achieved after 19 iterations
Presample variance: backcast (parameter = 0.7)
GARCH = C(4) + C(5)*RESID(-1)^2 + C(6)*RESID(-1)^2*(RESID(-1)<0)
+ C(7)*GARCH(-1)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
@SQRT(GARCH)	0.083360	0.053138	1.568753	0.1167
C	1.28E-05	0.000372	0.034443	0.9725
R(-1)	0.073763	0.017611	4.188535	0.0000
Variance Equation				
C	1.03E-06	2.23E-07	4.628790	0.0000
RESID(-1)^2	0.014945	0.009765	1.530473	0.1259
RESID(-1)^2*(RESID(-1)<0)	0.094014	0.014945	6.290700	0.0000
GARCH(-1)	0.922745	0.009129	101.0741	0.0000
T-DIST. DOF	5.531579	0.478432	11.56188	0.0000

Fitted GARCH Volatility

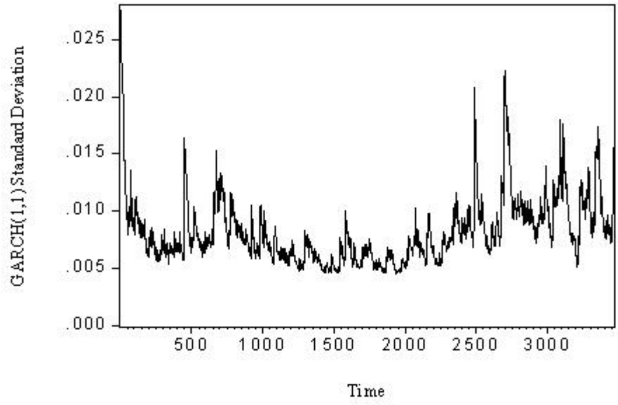


Figure: Estimated Conditional Standard Deviation, Daily NYSE Returns.

A Useful Specification Diagnostic

$$r_t | \Omega_{t-1} \sim N(0, \sigma_t^2)$$

$$\frac{r_t}{\sigma_t} \sim iid\ N(0, 1)$$

Infeasible: examine $\frac{r_t}{\sigma_t}$. iid? Gaussian?

Feasible: examine $\frac{r_t}{\hat{\sigma}_t}$. iid? Gaussian?

Key deviation from iid is volatility dynamics. So examine correlogram of squared standardized returns, $\left(\frac{r_t}{\hat{\sigma}_t}\right)^2$

GARCH Specification Diagnostic

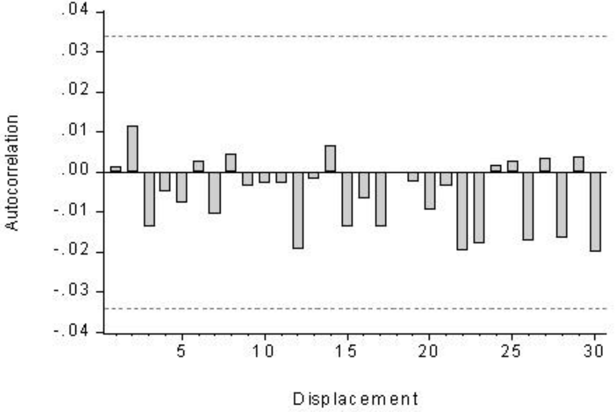


Figure: Correlogram of Squared Standardized Returns

GARCH Volatility Forecast

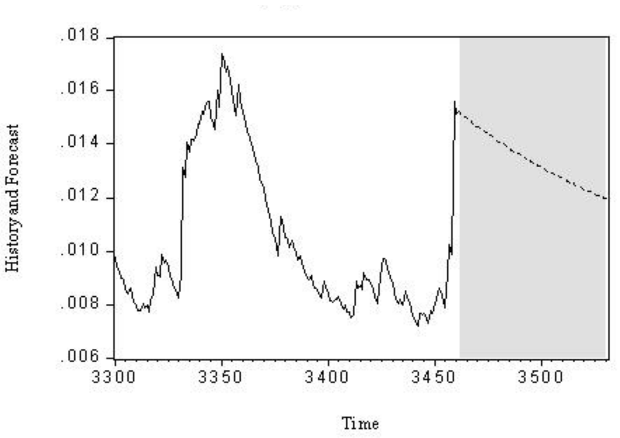


Figure: Conditional Standard Deviation, History and Forecast

Volatility Forecasts Feed Into Return Density Forecasts

In earlier linear (AR) environment we wrote:

$$y_{t+h}|\Omega_t \ \sim \ N(y_{t+h,t},\ \sigma_h^2)$$

(h-step forecast error variance depended only on h, not t)

Now we have:

$$y_{t+h}|\Omega_t \ \sim \ N(y_{t+h,t},\ \sigma_{t+h,t}^2)$$

(h-step forecast error variance now depends on both h and t)