Econometrics

Francis X. Diebold
University of Pennsylvania

November 2, 2017
Copyright © 2013-2017, by Francis X. Diebold.

All rights reserved.

All materials are freely available for your use, but be warned: they are highly preliminary, significantly incomplete, and rapidly evolving. All are licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. (Briefly: I retain copyright, but you can use, copy and distribute non-commercially, so long as you give me attribution and do not modify. To view a copy of the license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.) In return I ask that you please cite the books whenever appropriate, as: "Diebold, F.X. (year here), Book Title Here, Department of Economics, University of Pennsylvania, http://www.ssc.upenn.edu/ fdiebold/Textbooks.html.”

The painting is *Enigma*, by Glen Josselsohn, from Wikimedia Commons.
Introduction to Econometrics
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 successful analysis.

- Trend and seasonality
- Cycles (“serial correlation”)
- Volatility fluctuations (“heteroskedasticity”)
- Structural change
- Multivariate interactions
- Emphasis on prediction
Let’s Elaborate on the “Emphasis on Prediction”…

Q: What is econometrics is 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.
There are Many Issues Regarding Types of Recorded Economic Data

- Time series
- Continuous recording
- Discrete recording
- Equally-spaced
- Unequally-spaced
- Common-frequency
- Mixed-frequency
- Cross section
- Time series of cross sections
- Balanced panel
- Unbalanced panel
Notational Aside

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

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

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

Without loss of generality, we will use $t = 1, \ldots, T$. 
A Few Leading Econometrics Web Data Resources (Clickable)

Indispensable:

- 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 (CRAN; RStudio; R-bloggers), Python, Julia
- Low-Level: C, C++, Fortran

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

- More than you ever wanted to know about econometric software, broadly defined: The Econometrics Journal software links
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 Government Bond Yield, Levels
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, within 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:

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

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

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

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

- $\sigma$ easier to interpret than $\sigma^2$. 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:
\[ \text{cov}(y, x) = E[(y_t - \mu_y)(x_t - \mu_x)]. \]

So does correlation:
\[ \text{corr}(y, x) = \frac{\text{cov}(y, x)}{\sigma_y \sigma_x}. \]

Correlation is often more convenient. Why?
Sampling and Estimation

Sample: \( \{y_t\}_{t=1}^T \sim iid f(y) \)

Sample mean:

\[
\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t
\]

Sample variance:

\[
\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T}
\]

Unbiased sample variance:

\[
s^2 = \frac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T - 1}
\]
More Sample Moments

Sample skewness:

$$\hat{S} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})^3$$

Sample kurtosis:

$$\hat{K} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})^4$$
Still More Sample Moments

Sample covariance:

$$\hat{\text{cov}}(y, x) = \frac{1}{T} \sum_{t=1}^{T} [(y_t - \bar{y})(x_t - \bar{x})]$$

Sample correlation:

$$\hat{\text{corr}}(y, x) = \frac{\hat{\text{cov}}(y, x)}{\hat{\sigma}_y \hat{\sigma}_x}$$
Exact Finite-Sample Distribution of the Sample Mean
(Requires iid Normality)

Simple random sampling: \( y_t \sim iid \ N(\mu, \sigma^2), t = 1, \ldots, T \)

\( \bar{y} \) is unbiased, consistent, normally distributed with variance \( \sigma^2 / T \),
and minimum variance unbiased (MVUE).

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

(and we estimate \( \sigma^2 \) consistently using \( s^2 \))

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

\[
\mu = \mu_0 \implies \frac{\bar{y} - \mu_0}{s / \sqrt{T}} \sim t(T - 1)
\]
Large-Sample Distribution of the Sample Mean (Requires iid, but not Normality)

Simple random sampling: \( y_t \sim iid (\mu, \sigma^2) \), \( t = 1, \ldots, T \)

\( \bar{y} \) is asymptotically unbiased, consistent, asymptotically normally distributed with variance \( \sigma^2 / T \), and best linear unbiased (BLUE).

\[
\bar{y} \xrightarrow{a} N \left( \mu, \frac{\sigma^2}{T} \right),
\]

and we estimate \( \sigma^2 \) consistently using \( s^2 \). This is an approximate (large-sample) result, due to the central limit theorem. The “a” is for “asymptotically”, which means “as \( T \to \infty \).”

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

\[
As \ T \to \infty, \ \mu = \mu_0 \implies \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{T}}} \sim N(0, 1)
\]
Wages: Distributions
## Wages: Sample Statistics

<table>
<thead>
<tr>
<th></th>
<th>WAGE</th>
<th>log WAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>12.19</td>
<td>2.34</td>
</tr>
<tr>
<td>Sample Median</td>
<td>10.00</td>
<td>2.30</td>
</tr>
<tr>
<td>Sample Maximum</td>
<td>65.00</td>
<td>4.17</td>
</tr>
<tr>
<td>Sample Minimum</td>
<td>1.43</td>
<td>0.36</td>
</tr>
<tr>
<td>Sample Std. Dev.</td>
<td>7.38</td>
<td>0.56</td>
</tr>
<tr>
<td>Sample Skewness</td>
<td>1.76</td>
<td>0.06</td>
</tr>
<tr>
<td>Sample Kurtosis</td>
<td>7.93</td>
<td>2.90</td>
</tr>
<tr>
<td>Jarque-Bera</td>
<td>2027.86</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>$(p = 0.00)$</td>
<td>$(p = 0.53)$</td>
</tr>
<tr>
<td>$t(\mu = 12)$</td>
<td>0.93</td>
<td>-625.70</td>
</tr>
<tr>
<td></td>
<td>$(p = 0.36)$</td>
<td>$(p = 0.00)$</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.40</td>
<td></td>
</tr>
</tbody>
</table>
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:

\[ y_t = \beta_1 + \beta_2 x_t \]

Solve:

\[ \min_{\beta_1, \beta_2} \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_t)^2 \]

“least squares” (LS, or OLS)

“quadratic loss”
Actual Values, Fitted Values and Residuals

Actual values: \( y_t, \ t = 1, \ldots, T \)

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

Fitted values: \( \hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_t, \ t = 1, \ldots, T, \)

(“hats” denote fitted things...)

Residuals: \( e_t = y_t - \hat{y}_t, \ t = 1, \ldots, T. \)
Log Wage vs. Education with Superimposed Regression Line

\[ \hat{LWAGE} = 1.273 + 0.081 \cdot EDUC \]
Multiple Linear Regression ($K$ RHS Variables)

Solve:

$$\min_{\beta_1, \ldots, \beta_k} \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_{2t} - \ldots - \beta_K x_{Kt})^2$$

Fitted hyperplane:

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{2t} + \hat{\beta}_3 x_{3t} + \ldots + \hat{\beta}_K x_{Kt}$$

More compactly:

$$\hat{y}_t = \sum_{i=1}^{K} \hat{\beta}_i x_{it},$$

where $x_{1t} = 1$ for all $t$.

Wage dataset:

$$\widehat{LWAGE} = .867 + .093EDUC + .013EXPER$$
Regression as a Probability Model
An Ideal Situation ("The Ideal Conditions", or IC)

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

\[ y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t \]

\[ \varepsilon_t \sim iidN(0, \sigma^2), \]

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 (\(\beta\)'s) are fixed (whether over space or time, depending on whether we’re working in a time-series or cross-section environment)
   1.4 The relationship is linear
   1.5 The \(\varepsilon_t\)'s have constant variance \(\sigma^2\)
   1.6 The \(\varepsilon_t\)'s are uncorrelated

2. \(\varepsilon_t\) and \(x_{it}\) are independent, for all \(i, t\)
Some Crucial Matrix Notation

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

\[
y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{21} & x_{31} & \ldots & x_{K1} \\ 1 & x_{22} & x_{32} & \ldots & x_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2T} & x_{3T} & \ldots & x_{KT} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}
\]
Elementary Matrices and Matrix Operations

\[ 0 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \]

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.
Ideal Conditions Redux

We Used to Write This:

The DGP is

\[ y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t \]

\[ \varepsilon_t \sim iidN(0, \sigma^2), \]

and the fitted model matches it exactly, and

\[ \varepsilon_t \text{ independent of } x_{it}, \forall i, t, \ t = 1,2,\ldots, T \]

Now, equivalently, we write this:

The DGP is

\[ y = X\beta + \varepsilon \]

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

and the fitted model matches it exactly, and

\[ \varepsilon \text{ independent of } X \]
\[
\begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_T
\end{pmatrix} =
\begin{pmatrix}
  1 & x_{21} & x_{31} & \ldots & x_{K1} \\
  1 & x_{22} & x_{32} & \ldots & x_{K2} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  1 & x_{2T} & x_{3T} & \ldots & x_{KT}
\end{pmatrix}
\begin{pmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_K
\end{pmatrix} +
\begin{pmatrix}
  \varepsilon_1 \\
  \varepsilon_2 \\
  \vdots \\
  \varepsilon_T
\end{pmatrix}
\sim N\left(
\begin{pmatrix}
  0_1 \\
  0_2 \\
  \vdots \\
  0_T
\end{pmatrix},
\begin{pmatrix}
  \sigma^2 & 0 & \ldots & 0 \\
  0 & \sigma^2 & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \ldots & \sigma^2
\end{pmatrix}
\right)
\]

\(\varepsilon\) independent of \(X\)
The OLS Estimator in Matrix Notation:

As always, the LS estimator solves:

$$\min_\beta \left( \sum_{t=1}^{T} \epsilon_t^2 \right) = \min_\beta \left( \sum_{t=1}^{T} (y_t - \beta_1 - \beta_2 x_{2t} - \ldots - \beta_K x_{Kt})^2 \right)$$

In matrix notation:

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

It can be shown that the solution is:

$$\hat{\beta}_{\text{LS}} = (X'X)^{-1}X'y$$
Exact Finite-Sample Distribution of $\hat{\beta}_{LS}$
Under the Ideal Conditions

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

where $V$ is consistently estimated by $s^2(X'X)^{-1}$.

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K}$$

"$\hat{\beta}_{LS}$ is unbiased, consistent, normally distributed,
and minimum variance unbiased (MVUE)."

Note the precise parallel with the distribution of the sample mean
in Gaussian iid environments.
Sample mean is just LS regression on nothing but a constant.

(Prove it.)
Notational Review (And Just a Little Bit More)

We wrote:

\[ y_t = \beta_1 + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t, \quad t = 1, 2, \ldots, T \]

and

\[ y = X\beta + \varepsilon \]

We can also write

\[ y_t = x'_t \beta + \varepsilon_t, \quad t = 1, 2, \ldots, T \]
Conditional Implications of the DGP

Conditional mean:

\[ E(y_t \mid x_t = x^*) = x^* \beta \]

Conditional variance:

\[ \text{var}(y_t \mid x_t = x^*) = \sigma^2 \]

Full conditional density:

\[ y_t \mid x_t = x^* \sim N(x^* \beta, \sigma^2) \]
Why All the Talk About Conditional Implications?: The “Point Prediction” Problem

A major goal in econometrics is predicting $y$. The question is “If a new person $t$ arrives with characteristics $x_t=x^*$, what is my minimum-MSE prediction of her $y_t$? The answer under quadratic loss is $E(y_t \mid x_t=x^*) = x^*\beta$.

“The conditional mean is the minimum MSE (point) predictor”

Non-operational version (we don’t know $\beta$):
$E(y_t \mid x_t=x^*)=x^*\beta$

Operational version (use $\hat{\beta}_{LS}$):
$E(y_t \mid x_t=x^*) = x^*\hat{\beta}_{LS} \quad \text{(regression fitted value at $x_t=x^*$)}$

– LS delivers operational optimal predictor with great generality (i.e., even when the IC fail)

– Follows immediately from the LS optimization problem
Interval Prediction

Non-operational:

\[ y_t \in [x^*\beta \pm 1.96\sigma] \quad w.p. \ 0.95 \]

Operational:

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

(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:

\[ y_t \mid x_t = x^* \sim N(x^* \beta, \sigma^2) \]

Operational version:

\[ y_t \mid x_t = x^* \sim N(x^* \hat{\beta}_{LS}, s^2) \]

(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.)
Consider a density forecast for a person $t$ with characteristics $x_t = x^*$. 

1. Take $R$ draws from $N(0, s^2)$. 
2. Add $x^* \hat{\beta}$ to each disturbance draw. 
3. Form a density forecast by making a histogram for the output from step 2. 
4. 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 2.5% and 97.5% values, respectively.

As $R \to \infty$, the algorithmic and analytic results coincide.
“Typical” Regression Analysis of Wages, Education and Experience

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.867382</td>
<td>0.075331</td>
<td>11.51431</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.093229</td>
<td>0.005045</td>
<td>18.48002</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.013104</td>
<td>0.001164</td>
<td>11.26208</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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
“Top Matter”: Background Information

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

54 / 244
“Middle Matter”: Estimated Regression Function

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


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 $y$. 
There are many...
Regression Statistics: Mean dependent var 2.342

\[ \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t \]
The sample, or historical, mean of the dependent variable, $\bar{y}$, an estimate of the *unconditional* mean of $y$, is a benchmark forecast. It is obtained by regressing $y$ on an intercept alone – no conditioning on other regressors.
Regression Statistics: S.D. dependent var .561

\[
SD = \sqrt{\frac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T - 1}}
\]
– The sample standard deviation of $y$ is a measure of the in-sample accuracy of the unconditional mean forecast $\tilde{y}$. 
Regression Statistics: Sum squared resid 319.938

\[ SSR = \sum_{t=1}^{T} e_t^2 \]

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

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

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

$SSR$ measures “total” in-sample accuracy of the regression predictions

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

$$MSE = \frac{1}{T} SSR = \frac{1}{T} \sum_{t=1}^{T} e_t^2$$

(“average” in-sample accuracy of the regression predictions)
Regression Statistics: $F$-statistic 199.626

\[ F = \frac{(SSR_{res} - SSR)/(K - 1)}{SSR/(T - K)} \]
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 \(0.492\)

\[
s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K}
\]

\[
SER = \sqrt{s^2} = \sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T - K}}
\]
\( s^2 \) is just \( SSR \) scaled by \( T - K \), 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

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$
Regression Statistics: Adjusted $R$-squared .231

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-K} \sum_{t=1}^{T} e_t^2}{\frac{1}{T-1} \sum_{t=1}^{T} (y_t - \bar{y})^2}$$
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:

$$R^2 = 1 - \frac{1}{T} \sum_{t=1}^{T} e_t^2 \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})^2.$$ 

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

$$\bar{R}^2 = 1 - \frac{1}{T-K} \sum_{t=1}^{T} e_t^2 \frac{1}{T-1} \sum_{t=1}^{T} (y_t - \bar{y})^2,$$
Regression Statistics: Log likelihood -938.236

- Intimately related to SSR under normality

- Therefore closely related to prediction as well
Background/Detail: Regression Statistics: Log likelihood
-938.236

- Likelihood – joint density of the data (the $y_t$’s)

- Maximum-likelihood estimation – natural estimation strategy: find the parameter configuration that maximizes the likelihood of getting the $y_t$’s that you actually did get.

- Log likelihood – will have same max as the likelihood (why?) but it’s more important statistically

- Hypothesis tests and model selection – based on log likelihood
Linear regression model (under the IC) implies that:

\[ y_t | x_t \sim iidN(x_t' \beta, \sigma^2), \]

so that

\[ f(y_t | x_t) = (2\pi \sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2} (y_t - x_t' \beta)^2}. \]

Now by independence of the \( \varepsilon_t \)'s and hence \( y_t \)'s,

\[ L = f(y_1, \ldots, y_T | x_t) = f(y_1 | x_1) \cdots f(y_T | x_T) = \prod_{t=1}^{T} (2\pi \sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2} (y_t - x_t' \beta)^2} \]
Background/Detail: Log Likelihood

\[
\ln L = \ln \left( (2\pi \sigma^2)^{-\frac{T}{2}} \right) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - x'_t \beta)^2
\]

\[
= -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (y_t - x'_t \beta)^2
\]

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

- Additive constant \( -\frac{T}{2} \ln(2\pi) \) can be dropped

Note that the \( \beta \) vector that maximizes the likelihood is the \( \beta \) vector that minimizes the sum of squared residuals.

“MLE and OLS coincide for linear regression under the IC”

– Normality, in particular, is crucial.
Under conditions, asymptotically as $T \to \infty$:

$$-2(\ln L_0 - \ln L_1) \sim \chi^2_d,$$

where $\ln L_0$ is the maximized log likelihood under the restrictions implied 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. That’s why they can be written in terms of minimized $SSR$’s rather than maximized $\ln L$’s.
Regression Statistics: Schwarz criterion 1.435

We’ll get there shortly...
Regression Statistics: Durbin-Watson stat. 1.926

We’ll get there in six weeks...
Residual Scatter

![Residual Scatter Plot](image)
Residual Plot

Figure: Wage Regression Residual Plot
Predictive Perspectives

– The OLS fitted values, $\hat{y}_t = x_t'\hat{\beta}$, are effectively best in-sample predictions.

– The OLS residuals, $e_t = y_t - \hat{y}_t$, are effectively in-sample prediction errors corresponding to use of the best predictor.

– Residual plots are useful for visually flagging neglected things that impact forecasting. Residual serial correlation indicates that point forecasts could be improved. Residual volatility clustering indicates that interval and density forecasts could be improved.
Misspecification

Do we really believe that the fitted model matches the DGP?
Regression Statistics: Schwarz criterion 1.435

\[ SIC = T \left( \frac{K}{T} \right) \sum_{t=1}^{T} \frac{e_t^2}{T} \]

More general \( \ln L \) version:

\[ SIC = -2 \ln L + K \ln T \]
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 $MSE$.

\[
MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}
\]

\[
s^2 = \left( \frac{T}{T - K} \right) MSE
\]

\[
SIC = \left( T \left( \frac{K}{T} \right) \right) MSE
\]

“Oracle property”
Non-Normality and Outliers

Do we really believe that the disturbances are Gaussian?
What We’ll Do

– Distributional results under non-normality

– Detecting non-normality and outliers

– Dealing with non-normality and outliers (robust estimation)
Recall:

Exact Finite-Sample Distribution of $\hat{\beta}_{LS}$
Under the Ideal Conditions (Including Normality)

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

where $V$ is consistently estimated by $s^2(X'X)^{-1}$.

$$
\begin{pmatrix}
  s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T - K}
\end{pmatrix}
$$

“$\hat{\beta}_{LS}$ is unbiased, consistent, normally distributed, and minimum variance unbiased (MVUE).”

Note the precise parallel with the distribution of the sample mean in Gaussian iid environments.
Now:

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

\[ a \]
\[ \hat{\beta}_{LS} \sim N(\beta, V). \]

where $V$ is consistently estimated by $s^2(X'X)^{-1}$.

“$\hat{\beta}_{LS}$ is asymptotically unbiased, consistent, asymptotically normally distributed, and best linear unbiased (BLUE).”

Note the precise parallel with the distribution of the sample mean in non-Gaussian iid environments.
Detecting Non-Normality
(In Data or in Residuals)

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

- Jarque-Bera test. Under normality we have:

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

- Many more
Recall Our OLS Wage Regression

![Wage Regression Output](image)

**Dependent Variable:** LWAGE  
**Method:** Least Squares  
**Date:** 06/27/13  **Time:** 16:38  
**Sample (adjusted):** 11323  
**Included observations:** 1323 after adjustments

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.867382</td>
<td>0.075331</td>
<td>11.51431</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.093229</td>
<td>0.005045</td>
<td>18.48002</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.013104</td>
<td>0.001164</td>
<td>11.26208</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**R-squared:** 0.232224  
**Adjusted R-squared:** 0.231061  
**S.E. of regression:** 0.492318  
**Sum squared resid:** 319.9376  
**Log likelihood:** -938.2356  
**F-statistic:** 199.6260  
**Prob(F-statistic):** 0.000000

**Mean dependent var:** 2.341995  
**S.D. dependent var:** 0.561435  
**Akaike info criterion:** 1.422881  
**Schwarz criterion:** 1.434644  
**Hannan-Quinn criter.:** 1.427291  
**Durbin-Watson stat:** 1.926045
OLS Residual Histogram and Statistics

Series: RESID
Sample 1 1323
Observations 1323

- Mean: -5.82e-16
- Median: 0.003600
- Maximum: 1.766517
- Minimum: -1.888482
- Std. Dev.: 0.455104
- Skewness: -0.228689
- Kurtosis: 3.712251
- Jarque-Bera: 39.49685
- Probability: 0.000000
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
Detecting Outliers and Influential Observations: OLS Residual Plot
Detecting Outliers and Influential Observations: Leave-One-Out Plot

Consider:

\[ \left( \hat{\beta}^{(-t)} - \hat{\beta} \right), \ t = 1, ... T \]

“Leave-one-out plot”
Wage Regression

Leave-One-Out Plot

Coefficient (Education)

Leave t out

0.090
0.094

0 200 400 600 800 1000 1200

0 200 400 600 800 1000 1200
Dealing with Outliers:  
Least Absolute Deviations (LAD)

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

$$
\min_{\beta} \sum_{t=1}^{T} |\varepsilon_t|
$$

– Not as simple as OLS, but still simple

– Recall that OLS fits the conditional mean function:

$$
mean(y|X) = x\beta
$$

– LAD fits the conditional median function (50% quantile):

$$
median(y|X) = x\beta
$$

– The two are equal under symmetry as with FIC, but not under asymmetry, in which case the median is a more robust measure of central tendency
LAD Wage Regression Estimation

```
qreg log(wage) c educ exper
```

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.709127</td>
<td>0.087528</td>
<td>8.101740</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.101366</td>
<td>0.006316</td>
<td>16.04818</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.016394</td>
<td>0.001352</td>
<td>12.12388</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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
The environment is:

\[ y_t = x_t' \beta + \varepsilon_t, \quad t = 1, \ldots, T, \]

subject to the IC, except that we allow

\[ \varepsilon_t \sim iid \ D(0, \sigma^2) \]
Recall: Simulation Algorithm for Feasible Density Prediction With Normality:

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

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

As $R \to \infty$, the simulation error vanishes.
Now: Simulation Algorithm for Feasible Density Prediction Without Normality

1. Take $R$ approximate disturbance draws by assigning probability $1/T$ to each regression residual and sampling with replacement.

2. Add $x_t^*/\beta$ to each draw.

3. Form a density forecast by fitting a density to the output from step 2.

4. 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, respectively.

As $R \to \infty$ and $T \to \infty$, the simulation error vanishes.
Indicator Variables in Cross Sections: Group Effects

Effectively a type of Structural change: Do we really believe that intercepts are fixed across groups?
A dummy variable, or indicator variable, is just a 0-1 variable that indicates something, such as whether a person is female:

\[ \text{FEMALE}_t = \begin{cases} 
1 & \text{if person } t \text{ is female} \\
0 & \text{otherwise} 
\end{cases} \]

(It really is that simple.)

“Intercept dummies”
Histograms for Wage Covariates

Note that the sample mean of a dummy 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.867382</td>
<td>0.075331</td>
<td>11.51431</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.093229</td>
<td>0.005045</td>
<td>18.48002</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.013104</td>
<td>0.001164</td>
<td>11.26208</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared: 0.232224
Adjusted R-squared: 0.231061
S.E. of regression: 0.492318
Sum squared resid: 319.9376
Log likelihood: -938.2358
F-statistic: 199.8260
Prob(F-statistic): 0.000000
Controlling for 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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.000385</td>
<td>0.073180</td>
<td>13.67013</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.090809</td>
<td>0.004814</td>
<td>18.83314</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.012707</td>
<td>0.001119</td>
<td>11.35624</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.237535</td>
<td>0.025965</td>
<td>-9.148397</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.085286</td>
<td>0.035766</td>
<td>-2.383199</td>
<td>0.0173</td>
</tr>
<tr>
<td>UNION</td>
<td>0.223392</td>
<td>0.035307</td>
<td>6.327126</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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

**Basic Wage Regression**
- Conditions only on education and experience.
- Intercept is a mongrel combination of those for men, women, union, non-union, white, non-white.
- Comparatively sparse “information set”. Forecasting performance could be improved.

**Wage Regression With Dummies**
- 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

<table>
<thead>
<tr>
<th>obs</th>
<th>Y1</th>
<th>X1</th>
<th>Y2</th>
<th>X2</th>
<th>Y3</th>
<th>X3</th>
<th>Y4</th>
<th>X4</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>5.680000</td>
<td>5.000000</td>
<td>4.740000</td>
<td>5.000000</td>
<td>5.730000</td>
<td>5.000000</td>
<td>6.890000</td>
<td>8.000000</td>
</tr>
</tbody>
</table>
## Anscombe's Quartet: Regressions

### LS // Dependent Variable is Y1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td>X1</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
</tbody>
</table>

R-squared: 0.67  
S.E. of regression: 1.24

### LS // Dependent Variable is Y2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td>X2</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
</tbody>
</table>

R-squared: 0.67  
S.E. of regression: 1.24

### LS // Dependent Variable is Y3

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td>X3</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
</tbody>
</table>

R-squared: 0.67  
S.E. of regression: 1.24

### LS // Dependent Variable is Y4

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td>X4</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
</tbody>
</table>

R-squared: 0.67  
S.E. of regression: 1.24
Anscombe’s Quartet: Graphics

Figure: Anscombe’s Quartet: Graphs
Log-Log Regression

\[ \ln y_t = \beta_1 + \beta_2 \ln x_t + \varepsilon_t \]

For close \( y_t \) and \( x_t \), \( (\ln y_t - \ln x_t) \cdot 100 \) is approximately the percent difference between \( y_t \) and \( x_t \). 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

\[ y_t = A L_t^\alpha K_t^\beta \exp(\varepsilon_t) \]

\[ \ln y_t = \ln A + \alpha \ln L_t + \beta \ln K_t + \varepsilon_t \]

We expect an \( \alpha \)% increase in output in response to a 1% increase in labor input.
Log-Lin Regression

\[ \ln y_t = \beta x_t + \varepsilon \]

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

Example: Exponential growth

\[ y_t = Ae^{rt} \]

\[ \ln y_t = \ln A + rt \]

Coefficient \( r \) gives the expected percent change in \( y \) for a one-unit change in time.

Another 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

– Use non-linear least squares (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

\[ y_t = g(x_t, \varepsilon_t) \]

or slightly less generally as

\[ y_t = f(x_t) + \varepsilon_t \]
Consider Taylor series expansions of \( f(x_t) \).

The linear (first-order) approximation is

\[
f(x_t) \approx \beta_1 + \beta_2 x,
\]

and the quadratic (second-order) approximation is

\[
f(x_t) \approx \beta_1 + \beta_2 x_t + \beta_3 x_t^2.
\]

In the multiple regression case, Taylor approximations also involve interaction terms. Consider, for example, \( f(x_t, z_t) \):

\[
f(x_t, z_t) \approx \beta_1 + \beta_2 x_t + \beta_3 z_t + \beta_4 x_t^2 + \beta_5 z_t^2 + \beta_6 x_t z_t + \ldots.
\]
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 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 upper-right Anscombe panel!
  (Consider both cross-section and time-series cases.)
Assessing Non-Linearity

Use $SIC$ as always.

Use $t$’s and $F$ as always.
Linear Wage Regression (Actually Log-Lin)

![Linear Wage Regression Table]

**Dependent Variable:** LWAGE  
**Method:** Least Squares  
**Date:** 07/03/13  
**Time:** 13:36  
**Sample (adjusted):** 11323  
**Included observations:** 1323 after adjustments

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.000385</td>
<td>0.073180</td>
<td>13.67013</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.090809</td>
<td>0.004814</td>
<td>18.68314</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.012707</td>
<td>0.001119</td>
<td>11.35624</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.237535</td>
<td>0.025965</td>
<td>-9.148397</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.085286</td>
<td>0.035786</td>
<td>-2.383199</td>
<td>0.0173</td>
</tr>
<tr>
<td>UNION</td>
<td>0.223392</td>
<td>0.035307</td>
<td>6.327126</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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

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

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

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 |
```
Dummy Interactions?

Figure: Wage Regression on Education, Experience, Group Dummies, and Interactions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.011503</td>
<td>0.073797</td>
<td>13.70657</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.090805</td>
<td>0.004819</td>
<td>18.84231</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.012674</td>
<td>0.001120</td>
<td>11.31689</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.257621</td>
<td>0.030125</td>
<td>-8.551814</td>
<td>0.0000</td>
</tr>
<tr>
<td>UNION</td>
<td>0.216911</td>
<td>0.047498</td>
<td>4.566771</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.157606</td>
<td>0.056446</td>
<td>-2.792173</td>
<td>0.0053</td>
</tr>
<tr>
<td>FEM_UNI</td>
<td>0.003921</td>
<td>0.072024</td>
<td>0.054439</td>
<td>0.9566</td>
</tr>
<tr>
<td>FEM_NON</td>
<td>0.125567</td>
<td>0.071781</td>
<td>1.749313</td>
<td>0.0805</td>
</tr>
<tr>
<td>UNI_NON</td>
<td>0.017743</td>
<td>0.091336</td>
<td>0.194258</td>
<td>0.8460</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.309487</td>
<td>Mean dep. var</td>
<td>2.341995</td>
<td></td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.305283</td>
<td>S.D. dep. var</td>
<td>0.561435</td>
<td></td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>0.467955</td>
<td>Akaike info criterion</td>
<td>1.325889</td>
<td></td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>287.7418</td>
<td>Schwarz criterion</td>
<td>1.361179</td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-868.0755</td>
<td>Hannan-Quinn critic.</td>
<td>1.339118</td>
<td></td>
</tr>
</tbody>
</table>
Figure: Wage Regression with Continuous Non-Linearities and Interactions, and Discrete Interactions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.482967</td>
<td>0.240926</td>
<td>2.004623</td>
<td>0.0452</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.109522</td>
<td>0.029003</td>
<td>3.776211</td>
<td>0.0002</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.064269</td>
<td>0.007654</td>
<td>8.396570</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC2</td>
<td>0.000517</td>
<td>0.000900</td>
<td>0.573929</td>
<td>0.5661</td>
</tr>
<tr>
<td>EXPER2</td>
<td>-0.000701</td>
<td>8.87E-05</td>
<td>-7.904460</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDU Exp</td>
<td>-0.001796</td>
<td>0.000429</td>
<td>-4.186876</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.252921</td>
<td>0.029659</td>
<td>-8.527539</td>
<td>0.0000</td>
</tr>
<tr>
<td>UNION</td>
<td>0.200937</td>
<td>0.046575</td>
<td>4.314297</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.161501</td>
<td>0.050777</td>
<td>-3.232246</td>
<td>0.0034</td>
</tr>
<tr>
<td>FEM_UNI</td>
<td>-0.012956</td>
<td>0.070470</td>
<td>-0.183153</td>
<td>0.8547</td>
</tr>
<tr>
<td>FEM_NON</td>
<td>0.110319</td>
<td>0.070993</td>
<td>1.573909</td>
<td>0.1157</td>
</tr>
<tr>
<td>UNI_NON</td>
<td>0.033202</td>
<td>0.069258</td>
<td>0.371975</td>
<td>0.7100</td>
</tr>
</tbody>
</table>

R-squared: 0.344357
Adjusted R-squared: 0.338856
S.E. of regression: 0.456507
Akaike info criterion: 1.278805
Sum squared resid: 273.2109
Schwarz criterion: 1.325658
Log likelihood: -833.7970
Hannan-Quinn criter.: 1.296244
Durbin-Watson stat: 1.891544
So Drop Dummy Interactions and Tighten the Rest
Discrete Response Models

What if the dependent variable is binary?

– Ultimately violates the IC in many ways…
Many Names

“discrete response models”

“qualitative response models”

“limited dependent variable models”

“binary (binomial) response models”

“multinomial response models”

“classification models” (machine-learning lingo)

– Just dummy variable modeling, but dummy variable is on the left
Framework

Left-hand-side variable is $y_t = I_t(z)$, where the “indicator variable” $I_t(z)$ indicates whether event $z$ occurs; that is,

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

In that case we have

$$E(y_t|x_t) = E(I_t(z)|x_t) = x'_t \beta.$$ 

A key insight, however, is that

$$E(I_t(z)|x_t) = P(I_t(z) = 1|x_t),$$

so the statement is really

$$P(I_t(z) = 1|x_t) = x'_t \beta. \quad (1)$$

- Leading examples: recessions, bankruptcies, loan or credit card defaults, financial market crises, consumer choices, ...
How to “fit a line” when the LHS variable is binary?

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

Problem: The LPM fails to constrain the fitted probabilities to be in the unit interval.
Solution: Run $x_t'\beta$ through a monotone “squashing function,” $F(\cdot)$, that keeps $P(I_t(z) = 1|x_t)$ in the unit interval.

More precisely, move to models with

$$P(I_t(z) = 1|x_t) = F(x_t'\beta),$$

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,

\[
F(w) = \frac{e^w}{1 + e^w} = \frac{1}{1 + e^{-w}},
\]

so

\[
P(l_t(z) = 1|x_t) = \frac{e^{x_t'\beta}}{1 + e^{x_t'\beta}}.
\]

- The likelihood function can be derived, and the model can be immediately estimated by numerical maximization of the likelihood function.
Logit as a Linear Model for Log Odds

Consider a linear model for log odds

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

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

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

So logit is just linear regression for log odds.

A full statement of the implicit DGP is:

$$y_t \mid x_t \sim Bern(P(I_t(z) = 1 \mid x_t))$$

$$\ln \left( \frac{P(I_t(z) = 1 \mid x_t)}{1 - P(I_t(z) = 1 \mid x_t)} \right) = x_t' \beta.$$
Marginal Effects

Logit marginal effects $\frac{\partial E(y_t|x_t)}{\partial x_{it}}$ are not simply given by the $\beta_i$’s. Instead we have

$$\frac{\partial E(y_t|x_t)}{\partial x_{it}} = f(x_t^\prime \beta) \beta_i,$$

where $f(x) = dF(x)/dx$. So the marginal effect is not simply $\beta_i$; instead it is $\beta_i$ weighted by $f(x_t^\prime \beta)$, which depends on all $\beta_i$’s and $x_i$’s.

– However, signs of $\beta_i$’s are the signs of the effects, because $f$ must be positive. (Recall that $F$ is monotone increasing.)

– In addition, ratios of $\beta_i$’s do give ratios of effects, because the $f$’s cancel.
Recall that traditional $R^2$ for continuous LHS variables is

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

It’s not clear how to define or interpret $R^2$ for binary regression. Several variants have been proposed.

“Effron’s $R^2$”:

$$R^2 = 1 - \frac{\sum(y_t - \hat{P}(l_t(z) = 1|x_t))}{\sum(y_t - \bar{y}_t)^2}.$$
Classification and “0-1 Forecasting”

– Classification maps probabilities into 0-1 forecasts. “Bayes classifier” uses a cutoff of .5.

– Decision boundary. Suppose we use a Bayes classifier. We predict 1 when \( \logit(x'\beta) > 1/2 \). But that’s the same as predicting 1 when \( x'\beta > 0 \). If there are 2 \( x \) variables (potentially plus an intercept), then the condition \( x'\beta > 0 \) 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.
Example: Predicting High-Wage Individuals

Using 1995 CPS data, let’s suppose that we don’t observe exact wages. Instead, we observe whether a person’s wage is high ($WAGE \geq 15$) or low ($WAGE < 15$).

- 357 people with $HIGHWAGE = 1$ (high wage)

966 people with $HIGHWAGE = 0$ (low wage)
### Linear Probability Model

**Table: Linear Probability Model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUC</td>
<td>0.058***</td>
<td>(0.004)</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.008***</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.645***</td>
<td>(0.063)</td>
</tr>
</tbody>
</table>

*Note:* *p* < 0.1; **p** < 0.05; ***p*** < 0.01
Logit Regression of *HIGHWAGE* on *EDUC* and *EXPER*

**Table: Logit Regression**

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>HIGHWAGE</em></td>
<td></td>
</tr>
<tr>
<td>EDUC</td>
<td>0.347*** (0.028)</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.046*** (0.006)</td>
</tr>
<tr>
<td>Constant</td>
<td>−6.608*** (0.455)</td>
</tr>
</tbody>
</table>

**Ratio of Effects *EDUC*/*EXPER*:** 7.54

**Efron’s $R^2$:** 0.146

**Observations:** 1,323

**Note:** *p<0.1; **p<0.05; ***p<0.01
Fitted Probabilities From LPM vs. Logit

Note that LPM has some negative fitted probabilities.
Classification Table for Logit Bayes Classifier

Defining high probability as having a probability greater than 0.5, our logistic model gets the prediction right for about 75% of the time.

Table: Classification Percentages

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>HIGHWAGE</td>
<td>68%</td>
<td>5%</td>
</tr>
<tr>
<td>1</td>
<td>HIGHWAGE</td>
<td>20%</td>
<td>7%</td>
</tr>
</tbody>
</table>

- Type I Error: Model predicts 1 (HIGHWAGE) and truth is 0. “False positive”
- Type II Error: Model predicts 0 and truth is 1. “False negative”
Decision Boundary Scatterplot for Logit Bayes Classifier
Heteroskedasticity in Cross-Section Regression

Do we really believe that disturbance variances are constant over space?
Heteroskedasticity is Another Type of Violation of the IC (This time it’s “non-constant disturbance variances”.)

Consider: \( \varepsilon \sim \mathcal{N}(0, \Omega) \)

Heteroskedasticity corresponds to \( \Omega \) diagonal but \( \Omega \neq \sigma^2 I \)

\[
\Omega = \begin{pmatrix}
\sigma_1^2 & 0 & \ldots & 0 \\
0 & \sigma_2^2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_T^2
\end{pmatrix}
\]
Causes and Predictive Consequences

Causes:

– Can arise for many reasons

– e.g., Engel curve (food expenditure vs. income) is classic example

Predictive consequences:

– Point forecasts remain largely OK.

  OLS parameter estimates remain consistent,

  so we still have $E(y_t \mid x_t=x^*_t) \to E(y_t \mid x_t=x^*_t)$.

– Our earlier (“constant $\sigma$ across people”) interval and density forecasts are now suboptimal.

  Heteroskedasticity enables “different $\sigma$’s for different people”,

  so it represents an opportunity, not a pitfall.
Heteroskedasticity-Robust Standard Errors

– Perhaps you’re only interested in point prediction but still want to do credible inference regarding the contributions of the various $x$ variables to the point prediction.

– Perhaps you’re ultimately interested in interval or density prediction, but still want to do credible inference regarding the contributions of the various $x$ variables to the point prediction.

Then use “heteroskedasticity-robust standard errors”

“White standard errors”

– Mechanically, it’s just a simple regression option

  e.g., in EViews,
  instead of “ls y,c,x”, use “ls(cov=white) y,c,x”
Wage regression with White Standard Errors

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.482967</td>
<td>0.239035</td>
<td>2.020483</td>
<td>0.0435</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.109522</td>
<td>0.030058</td>
<td>3.643689</td>
<td>0.0003</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.064269</td>
<td>0.007604</td>
<td>8.451643</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDUC2</td>
<td>0.000517</td>
<td>0.000965</td>
<td>0.535056</td>
<td>0.5927</td>
</tr>
<tr>
<td>EXPER2</td>
<td>-0.000701</td>
<td>8.89E-05</td>
<td>-7.885181</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDU_EXP</td>
<td>-0.001796</td>
<td>0.000464</td>
<td>-3.866971</td>
<td>0.0001</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.252921</td>
<td>0.030716</td>
<td>-8.234167</td>
<td>0.0000</td>
</tr>
<tr>
<td>UNION</td>
<td>0.200937</td>
<td>0.043029</td>
<td>4.669795</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.161501</td>
<td>0.055951</td>
<td>-2.886490</td>
<td>0.0040</td>
</tr>
<tr>
<td>FEM_UNI</td>
<td>-0.012956</td>
<td>0.062048</td>
<td>-0.208810</td>
<td>0.8346</td>
</tr>
<tr>
<td>FEM_NON</td>
<td>0.110319</td>
<td>0.068631</td>
<td>1.607439</td>
<td>0.1082</td>
</tr>
<tr>
<td>UNI_NON</td>
<td>0.033202</td>
<td>0.084284</td>
<td>0.393923</td>
<td>0.6937</td>
</tr>
</tbody>
</table>

R-squared   | 0.344357    | Mean dependent var | 2.341995
Detecting Heteroskedasticity

- Graphical heteroskedasticity diagnostics
- Formal heteroskedasticity tests
Graph $e_t^2$ against $x_t$, for various regressors
Recall Our “Final” Wage Regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.360535</td>
<td>0.131792</td>
<td>2.735636</td>
<td>0.0063</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.125028</td>
<td>0.009188</td>
<td>13.60791</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.086130</td>
<td>0.007016</td>
<td>9.424974</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER2</td>
<td>-0.000710</td>
<td>8.82E-05</td>
<td>-8.042035</td>
<td>0.0000</td>
</tr>
<tr>
<td>EDU_EXP</td>
<td>-0.001905</td>
<td>0.000375</td>
<td>-5.078046</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.239352</td>
<td>0.025327</td>
<td>-9.450447</td>
<td>0.0000</td>
</tr>
<tr>
<td>UNION</td>
<td>0.202574</td>
<td>0.034553</td>
<td>5.862629</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.094903</td>
<td>0.034921</td>
<td>-2.717655</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

R-squared      | 0.342915    |           |             |       |
Adjusted R-squared | 0.339418    |           |             |       |
S.E. of regression | 0.456313    |           |             |       |
Sum squared resid  | 273.8119    |           |             |       |
Log likelihood    | -835.2503   |           |             |       |
F-statistic       | 98.03775    |           |             |       |
Prob(F-statistic) | 0.000000    |           |             |       |
The Breusch-Pagan-Godfrey Test (BPG)

Limitation of graphing $e_t^2$ against $x_t$: 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 $T \cdot R^2$ from this regression. In large samples $T \cdot R^2 \sim \chi^2_{K-1}$ under the null, where $K$ is the number of regressors in the test regression.
**Heteroskedasticity Test: Breusch-Pagan-Godfrey**

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C</strong></td>
<td>-0.170309</td>
<td>0.097349</td>
<td>-1.749473</td>
<td>0.0804</td>
</tr>
<tr>
<td><strong>EDUC</strong></td>
<td>0.024074</td>
<td>0.006787</td>
<td>3.547204</td>
<td>0.0004</td>
</tr>
<tr>
<td><strong>EXPER</strong></td>
<td>0.011701</td>
<td>0.005183</td>
<td>2.257616</td>
<td>0.0241</td>
</tr>
<tr>
<td><strong>EXPER2</strong></td>
<td>-5.53E-05</td>
<td>6.52E-05</td>
<td>-0.849150</td>
<td>0.3960</td>
</tr>
<tr>
<td><strong>EDU_EXP</strong></td>
<td>-0.000478</td>
<td>0.000277</td>
<td>-1.725513</td>
<td>0.0847</td>
</tr>
<tr>
<td><strong>FEMALE</strong></td>
<td>-0.009757</td>
<td>0.018708</td>
<td>-0.521530</td>
<td>0.6021</td>
</tr>
<tr>
<td><strong>UNION</strong></td>
<td>-0.079848</td>
<td>0.025523</td>
<td>-3.120823</td>
<td>0.0018</td>
</tr>
<tr>
<td><strong>NONWHITE</strong></td>
<td>0.000486</td>
<td>0.025794</td>
<td>0.018829</td>
<td>0.9850</td>
</tr>
</tbody>
</table>
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 $T \cdot R^2$ from this regression. In large samples $T \cdot R^2 \sim \chi^2_{K-1}$ under the null.
White’s Test

Heteroskedasticity Test: White

<table>
<thead>
<tr>
<th>F-statistic</th>
<th>2.431488</th>
<th>Prob. F(29,1293)</th>
<th>0.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs*R-squared</td>
<td>68.41804</td>
<td>Prob. Chi-Square(29)</td>
<td>0.0000</td>
</tr>
<tr>
<td>Scaled explained SS</td>
<td>91.66473</td>
<td>Prob. Chi-Square(29)</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Interval and Density Prediction with Heteroskedastic Disturbances

Consider a person \( t \) with characteristics \( x_t = x^* \).

- An operational interval forecast is:
  
  \[
  y_t \in [x^* \hat{\beta} \pm 1.96 \hat{\sigma}_*] \quad \text{w.p.} \ 0.95,
  \]

  where \( \hat{\sigma}^2_* \) is the fitted value from the BPG or White regression evaluated at \( x_t = x^* \).

- An operational density forecast is:
  
  \[
  y_t \mid x_t = x^* \sim N(x^* \hat{\beta}, \hat{\sigma}^2_*)
  \]
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 intercepts are fixed over time?

- Trend is about \textit{gradual} intercept evolution
Liquor Sales
Log Liquor Sales
Linear Trend

\[ \text{Trend}_t = \beta_1 + \beta_2 \text{TIME}_t \]

where \( \text{TIME}_t = t \)

Simply run the least squares regression \( y \rightarrow c, \text{TIME} \), where

\[
\text{TIME} = \begin{pmatrix}
1 \\
2 \\
3 \\
\vdots \\
T - 1 \\
T
\end{pmatrix}
\]
Various Linear Trends

\[ \text{TREND} = 10 - 0.25 \times \text{TIME} \]

\[ \text{TREND} = -50 + 0.8 \times \text{TIME} \]
### Linear Trend Estimation

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>6.454290</td>
<td>0.017468</td>
<td>369.4834</td>
<td>0.0000</td>
</tr>
<tr>
<td>TIME</td>
<td>0.003809</td>
<td>8.98E-05</td>
<td>42.39935</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

R-squared: 0.843318  
Mean dependent var: 7.096188  
Adjusted R-squared: 0.842849  
S.D. dependent var: 0.402962  
Akaike info criterion: -0.824561  
Schwarz criterion: -0.801840  
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 intercepts are fixed over seasons? (quite apart from, and even after accounting for, time-varying intercepts due to trend)
Seasonal Dummies

\[ \text{Seasonal}_t = \sum_{i=1}^{s} \beta_i \text{SEAS}_{it} \quad (s \text{ seasons per year}) \]

where \( \text{SEAS}_{it} = \begin{cases} 
1 & \text{if observation } t \text{ falls in season } i \\
0 & \text{otherwise} 
\end{cases} \)

Simply run the least squares regression \( y \rightarrow \text{SEAS}_1, ..., \text{SEAS}_s \)
(or blend: \( y \rightarrow \text{TIME}, \text{SEAS}_1, ..., \text{SEAS}_s \))

where (e.g., in quarterly data case, assuming Q1 start and Q4 end):
\[ \text{SEAS}_1 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, ..., 0)' \]
\[ \text{SEAS}_2 = (0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, ..., 0)' \]
\[ \text{SEAS}_3 = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, ..., 0)' \]
\[ \text{SEAS}_4 = (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, ..., 1)' \].
**Linear Trend with Seasonal Dummies**

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>0.003779</td>
<td>6.24E-05</td>
<td>60.57536</td>
<td>0.0000</td>
</tr>
<tr>
<td>D1</td>
<td>6.361233</td>
<td>0.023283</td>
<td>273.2148</td>
<td>0.0000</td>
</tr>
<tr>
<td>D2</td>
<td>6.304412</td>
<td>0.023310</td>
<td>270.4571</td>
<td>0.0000</td>
</tr>
<tr>
<td>D3</td>
<td>6.391653</td>
<td>0.023338</td>
<td>273.8773</td>
<td>0.0000</td>
</tr>
<tr>
<td>D4</td>
<td>6.392737</td>
<td>0.023365</td>
<td>273.6004</td>
<td>0.0000</td>
</tr>
<tr>
<td>D5</td>
<td>6.461768</td>
<td>0.023393</td>
<td>276.2273</td>
<td>0.0000</td>
</tr>
<tr>
<td>D6</td>
<td>6.468189</td>
<td>0.023421</td>
<td>276.1145</td>
<td>0.0000</td>
</tr>
<tr>
<td>D7</td>
<td>6.510789</td>
<td>0.023449</td>
<td>277.6602</td>
<td>0.0000</td>
</tr>
<tr>
<td>D8</td>
<td>6.482457</td>
<td>0.023477</td>
<td>276.1210</td>
<td>0.0000</td>
</tr>
<tr>
<td>D9</td>
<td>6.422551</td>
<td>0.023505</td>
<td>273.2406</td>
<td>0.0000</td>
</tr>
<tr>
<td>D10</td>
<td>6.444589</td>
<td>0.023533</td>
<td>273.8476</td>
<td>0.0000</td>
</tr>
<tr>
<td>D11</td>
<td>6.476504</td>
<td>0.023562</td>
<td>274.8709</td>
<td>0.0000</td>
</tr>
<tr>
<td>D12</td>
<td>6.798519</td>
<td>0.023591</td>
<td>288.1874</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
```

R-squared: 0.927059  
Adjusted R-squared: 0.924350  
S.E. of regression: 0.110833  
Sum squared resid: 3.967734  
Log likelihood: 268.9746  
Durbin-Watson stat: 0.100500

Mean dependent var: 7.096188  
S.D. dependent var: 0.402962  
Akaike info criterion: 1.523658  
Schwarz criterion: 1.375972  
Hannan-Quinn criterion: 1.464786

Path = c:\users\francis.diebo\documents\diebold\files\courses\econ04\old\econ04_2011\sw3e\views
```
Seasonal Pattern

Estimated Seasonal Factors

Factor vs. Months (M1 to M12)
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 $TIME$ and $TIME^2$:

$$Trend_t = \beta_1 + \beta_2 TIME_t + \beta_3 TIME_t^2$$
Figure: Various Quadratic Trends
### Log-Quadratic Liquor Sales Trend Estimation

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>6.231269</td>
<td>0.020653</td>
<td>301.7187</td>
<td>0.0000</td>
</tr>
<tr>
<td>TIME</td>
<td>0.007768</td>
<td>0.000283</td>
<td>27.44987</td>
<td>0.0000</td>
</tr>
<tr>
<td>TIME2</td>
<td>-1.17E-05</td>
<td>8.13E-07</td>
<td>-14.44511</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- **R-squared:** 0.903676  
- **Mean dependent var:** 7.096188  
- **S.D. dependent var:** 0.402962  
- **Akaike info criterion:** -1.305106  
- **Schwarz criterion:** -1.271025  
- **Durbin-Watson stat:** 1.754412  
- **Prob(F-statistic):** 0.000000

---

**Figure:**

---
Residual Plot
Log-Quadratic Liquor Sales Trend Estimation with Seasonal Dummies

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>0.007739</td>
<td>0.000104</td>
<td>74.49828</td>
<td>0.0000</td>
</tr>
<tr>
<td>TIME2</td>
<td>-1.18E-05</td>
<td>2.98E-07</td>
<td>-39.36756</td>
<td>0.0000</td>
</tr>
<tr>
<td>D1</td>
<td>6.138362</td>
<td>0.011207</td>
<td>547.7315</td>
<td>0.0000</td>
</tr>
<tr>
<td>D2</td>
<td>6.081424</td>
<td>0.011218</td>
<td>542.1044</td>
<td>0.0000</td>
</tr>
<tr>
<td>D3</td>
<td>6.168571</td>
<td>0.011229</td>
<td>549.3318</td>
<td>0.0000</td>
</tr>
<tr>
<td>D4</td>
<td>6.169584</td>
<td>0.011240</td>
<td>548.8944</td>
<td>0.0000</td>
</tr>
<tr>
<td>D5</td>
<td>6.238568</td>
<td>0.011251</td>
<td>554.5117</td>
<td>0.0000</td>
</tr>
<tr>
<td>D6</td>
<td>6.243596</td>
<td>0.011261</td>
<td>554.4517</td>
<td>0.0000</td>
</tr>
<tr>
<td>D7</td>
<td>6.287566</td>
<td>0.011271</td>
<td>557.8584</td>
<td>0.0000</td>
</tr>
<tr>
<td>D8</td>
<td>6.259257</td>
<td>0.011281</td>
<td>554.8647</td>
<td>0.0000</td>
</tr>
<tr>
<td>D9</td>
<td>6.199399</td>
<td>0.011290</td>
<td>549.0938</td>
<td>0.0000</td>
</tr>
<tr>
<td>D10</td>
<td>6.221507</td>
<td>0.011300</td>
<td>550.5987</td>
<td>0.0000</td>
</tr>
<tr>
<td>D11</td>
<td>6.253515</td>
<td>0.011309</td>
<td>552.9885</td>
<td>0.0000</td>
</tr>
<tr>
<td>D12</td>
<td>6.575648</td>
<td>0.011317</td>
<td>581.0220</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

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 in Time-Series Regression

Do we really believe that disturbances are uncorrelated over time?
(Not possible in cross sections, so we didn’t study it before...)
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)
Some Important Language and Tools
For Characterizing Serial Correlation

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

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

“Partial autocorrelations”: \( p_{\varepsilon}(\tau) \), \( \tau = 1, 2, ... \)
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: \( \hat{\rho}_{\varepsilon}(\tau) = \hat{\text{corr}}(e_t, e_{t-\tau}) \), \( \tau = 1, 2, ... \)

Sample partial autocorrelations: \( \hat{p}_{\varepsilon}(\tau) \), \( \tau = 1, 2, ... \)
is the coefficient on \( e_{t-\tau} \) in the finite-sample regression:
\[ e_t \rightarrow c, e_{t-1}, ..., e_{t-(\tau-1)}, e_{t-\tau} \]
Why is Neglected Serial Correlation a Problem? 
Minor Answer: Serial Correlation Implies $\Omega \neq \sigma^2 I$

Heteroskedasticity: $\Omega$ diagonal but with different diagonal elements

Now serial correlation: $\Omega$ not even diagonal

$$\Omega = \begin{pmatrix}
\gamma_\varepsilon(0) & \gamma_\varepsilon(1) & \ldots & \gamma_\varepsilon(T - 1) \\
\gamma_\varepsilon(1) & \gamma_\varepsilon(0) & \ldots & \gamma_\varepsilon(T - 2) \\
\vdots & \vdots & \ddots & \vdots \\
\gamma_\varepsilon(T - 1) & \gamma_\varepsilon(T - 2) & \ldots & \gamma_\varepsilon(0)
\end{pmatrix}$$

Note the "band symmetric", or "toeplitz", structure
(illustrated here for $T = 4$):

$$\Omega = \begin{pmatrix}
a & b & c & d \\
b & a & b & c \\
c & b & a & b \\
d & c & b & a
\end{pmatrix}$$

We get inconsistent s.e.'s, just as with heteroskedasticity.
Heteroskedasticity and Autocorrelation Robust Standard Errors

– Perhaps you’re ultimately interested in making forecasts, but as a preliminary step you want to do credible inference regarding the contributions of the various x variables to a point prediction based only on the x’s.

Then use “heteroskedasticity and autocorrelation robust standard errors”

“HAC standard errors”, “Newey-West standard errors”

– Mechanically, just a simple regression option

  e.g., in EViews, instead of “ls y,c,x”, use “ls(cov=hac) y,c,x”
Trend + Seasonal Liquor Sales Regression with HAC Standard Errors

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>0.007739</td>
<td>0.000197</td>
<td>39.26392</td>
<td>0.0000</td>
</tr>
<tr>
<td>TIME2</td>
<td>-1.18E-05</td>
<td>6.34E-07</td>
<td>-18.52563</td>
<td>0.0000</td>
</tr>
<tr>
<td>D1</td>
<td>6.138362</td>
<td>0.013502</td>
<td>454.6369</td>
<td>0.0000</td>
</tr>
<tr>
<td>D2</td>
<td>6.081424</td>
<td>0.012767</td>
<td>476.3219</td>
<td>0.0000</td>
</tr>
<tr>
<td>D3</td>
<td>6.168571</td>
<td>0.012301</td>
<td>501.4707</td>
<td>0.0000</td>
</tr>
<tr>
<td>D4</td>
<td>6.169584</td>
<td>0.010921</td>
<td>564.9371</td>
<td>0.0000</td>
</tr>
<tr>
<td>D5</td>
<td>6.238568</td>
<td>0.011627</td>
<td>536.5598</td>
<td>0.0000</td>
</tr>
<tr>
<td>D6</td>
<td>6.243596</td>
<td>0.011226</td>
<td>556.1570</td>
<td>0.0000</td>
</tr>
<tr>
<td>D7</td>
<td>6.287566</td>
<td>0.010984</td>
<td>572.4238</td>
<td>0.0000</td>
</tr>
<tr>
<td>D8</td>
<td>6.259257</td>
<td>0.012412</td>
<td>504.2842</td>
<td>0.0000</td>
</tr>
<tr>
<td>D9</td>
<td>6.199399</td>
<td>0.011517</td>
<td>538.2667</td>
<td>0.0000</td>
</tr>
<tr>
<td>D10</td>
<td>6.221507</td>
<td>0.011771</td>
<td>528.5468</td>
<td>0.0000</td>
</tr>
<tr>
<td>D11</td>
<td>6.253515</td>
<td>0.013011</td>
<td>480.6421</td>
<td>0.0000</td>
</tr>
<tr>
<td>D12</td>
<td>6.575648</td>
<td>0.013389</td>
<td>491.1133</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Why is Neglected Serial Correlation a Problem?
Major Answer: All Types of Predictions Will be Suboptimal

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 (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...
White Noise Disturbances

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

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

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

We write:

$\varepsilon_t \sim WN(0, \sigma^2)$
Realization of White Noise Process

Time

Process
Population Autocorrelation Function
White Noise Process
AR(1) Disturbances

\[ \varepsilon_t = \phi \varepsilon_{t-1} + \nu_t, \quad |\phi| < 1 \]

\[ \nu_t \sim WN(0, \sigma^2) \]
Realizations of Two AR(1) Processes \((N(0, 1)\) shocks)
Population Autocorrelation Function
AR(1) Process, $\phi=.4$

$$\rho(\tau) = \phi^\tau$$
Population Autocorrelation Function
AR(1) Process, $\phi=.95$

$$\rho(\tau) = \phi^{\tau}$$
$\rho_\varepsilon(\tau)$ and $p_\varepsilon(\tau)$ for Generic AR($p$)

**AR($p$) Process:**

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_p \varepsilon_{t-p} + \nu_t$$

$$\nu_t \sim WN(0, \sigma^2), \quad |\phi| < 1$$

$$\rho_\varepsilon(\tau) \to 0 \text{ as } \tau \to \infty, \text{ gradually}$$

$$p_\varepsilon(\tau) \downarrow 0 \text{ at } \tau = p, \text{ abruptly}$$
Detecting Serial Correlation

- Graphical diagnostics
  - Residual plot
  - Residual scatterplot of \((e_t \text{ vs. } e_{t-\tau})\)
  - Residual correlogram

- Formal tests
  - Durbin-Watson
  - Breusch-Godfrey
Recall Our Log-Quadratic Liquor Sales Model

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

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

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

Sample: 1987M01 2014M12
Included observations: 336

<table>
<thead>
<tr>
<th>Autocorrelation</th>
<th>Partial Correlation</th>
<th>AC</th>
<th>PAC</th>
<th>Q-Stat</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.707</td>
<td>169.30</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.689</td>
<td>330.75</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.729</td>
<td>512.06</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.569</td>
<td>-0.160</td>
<td>622.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0.577</td>
<td>0.043</td>
<td>737.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0.583</td>
<td>0.092</td>
<td>854.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>0.469</td>
<td>-0.067</td>
<td>930.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.502</td>
<td>0.068</td>
<td>1017.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.489</td>
<td>0.047</td>
<td>1100.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>0.361</td>
<td>-0.165</td>
<td>1145.8</td>
</tr>
</tbody>
</table>
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 \( y_t \rightarrow x_t \) 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 \( T \to \infty \):

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

\[ \implies DW \in [0, 4], \quad DW \to 2 \text{ as } \phi \to 0, \text{ and } DW \to 0 \text{ as } \phi \to 1 \]
Formal Tests: Breusch-Godfrey

General AR($p$) environment:

\[ y_t = x_t' \beta + \varepsilon_t \]

\[ \varepsilon_t = \phi_1 \varepsilon_{t-1} + \ldots + \phi_p \varepsilon_{t-p} + \nu_t \]

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

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

Does this sound familiar?
BG for AR(1) Disturbances

\( TR^2 = 168.5, \ p = 0.0000 \)
BG for AR(4) Disturbances

\( TR^2 = 216.7, \ p = 0.0000 \)
BG for AR(8) Disturbances

\( TR^2 = 219.0, \ p = 0.0000 \)
Fixing the Serial Correlation Problem: 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.
“Lagged dependent variables”

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.

\[ y_t = \beta_1 + \beta_2 x_t + \varepsilon_t \]
\[ \varepsilon_t = \phi \varepsilon_{t-1} + \nu_t \]
\[ \nu_t \sim iid \ N(0, \sigma^2) \]
Trend + Seasonal Model with Four Lags of y
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

<table>
<thead>
<tr>
<th>Autocorrelation</th>
<th>Partial Correlation</th>
<th>AC</th>
<th>PAC</th>
<th>Q-Stat</th>
<th>Prob*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>0.005</td>
<td>0.005</td>
<td>0.0077</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.003</td>
<td>0.003</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-0.057</td>
<td>-0.057</td>
<td>1.0926</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>-0.009</td>
<td>-0.008</td>
<td>1.1176</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>-0.002</td>
<td>-0.002</td>
<td>1.1190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0.017</td>
<td>0.014</td>
<td>1.2161</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>-0.025</td>
<td>-0.026</td>
<td>1.4316</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.068</td>
<td>0.068</td>
<td>2.9954</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.059</td>
<td>0.061</td>
<td>4.1934</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>-0.123</td>
<td>-0.128</td>
<td>9.4373</td>
</tr>
</tbody>
</table>
For now assume known parameters. Of course to make it feasible we replace parameters with estimates.

\[ y_t = x_t' \beta + \varepsilon_t \implies y_{t+h} = x_{t+h}' \beta + \varepsilon_{t+h} \]

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 an $AR(1)$

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

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

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

There seems to be a FRHS variables problem for $h > 1$.
But there’s not...

We can build the multi-step forecast recursively. How?
“Wold’s chain rule of forecasting”
What is the optimal 1-step-ahead error? “innovations”
Interval and Density Forecasting in Time Series: Background

Consider a generic $AR(p)$ Process:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2), \ |\phi| < 1$$

It can always be expressed as a linear function of its past innovations (why?):

$$y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$$

where $b_0 = 1$.

It can be shown that the variance of the optimal $h$-step-ahead forecast error is:

$$\sigma_{t+h,t}^2 \equiv \text{var}(y_{t+h} - y_{t+h,t}) = \sigma^2 \sum_{i=0}^{h-1} b_i^2$$
Assuming Gaussian innovations, we immediately have

\[ y_{t+h} \sim N(y_{t+h}, \sigma^2_{t+h}) \]

Done.

Of course to make it operational we replace parameters with estimates and use \( N(\hat{y}_{t+h}, \hat{\sigma}^2_{t+h}) \).
(Is the Normality Assumption OK for Liquor Sales?)

Trend + Seasonal Model with Four Lags of $y$

Residual Histogram and Normality Test

![Graph showing residual histogram with overlayed table of statistics]

- **Observations**: 312
- **Mean**: $3.77 \times 10^{-16}$
- **Median**: $-0.000160$
- **Maximum**: $0.078468$
- **Minimum**: $-0.109856$
- **Std. Dev.**: $0.026635$
- **Skewness**: $0.077911$
- **Kurtosis**: $3.740378$
- **Jarque-Bera**: $7.441714$
- **Probability**: $0.024213$
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...
More General Structural Change in Time Series: Drifts or Breaks in Any or all Parameters

Again, do we really believe that parameters are fixed over time?
Structural Change
Sharp Breakpoint Exogenously Known

For simplicity of exposition, consider a bivariate regression:

\[ y_t = \begin{cases} 
\beta_1^1 + \beta_2^1 x_t + \epsilon_t, & t = 1, \ldots, T^* \\
\beta_1^2 + \beta_2^2 x_t + \epsilon_t, & t = T^* + 1, \ldots, T 
\end{cases} \]

Let

\[ D_t = \begin{cases} 
0, & t = 1, \ldots, T^* \\
D_t = 1, & t = T^* + 1, \ldots, T 
\end{cases} \]

Then we can write the model as:

\[ y_t = (\beta_1^1 + (\beta_2^2 - \beta_1^1)D_t) + (\beta_2^1 + (\beta_2^2 - \beta_2^1)D_t)x_t + \epsilon_t \]

We run:

\[ y_t \rightarrow c, \ D_t, \ x_t, \ D_t \cdot x_t \]

Use regression to test for structural change (\textit{F} test)
Use regression to accommodate structural change if present.
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

\[ \text{MaxChow} = \max_{\tau_{\min} \leq \tau \leq \tau_{\max}} \text{Chow}(\tau), \]

where \( \tau \) denotes potential break location as fraction of sample

(Typically we take \( \tau_{\min} = .15 \) and \( \tau_{\max} = .85 \))

The null distribution of \( \text{MaxChow} \) has been tabulated.
Rolling-Window Regression for Generic Structural Change Assessment

Calculate and examine

$$\hat{\beta}_{t-w:t}$$

for $t = w + 1, \ldots, T$

$w$ is “window width”

What does window width govern?
Expanding-Window ("Recursive") Regression for Generic Structural Change Assessment

Model:

\[ y_t = \sum_{k=1}^{K} \beta_k x_{kt} + \varepsilon_t \]

\[ \varepsilon_t \sim iidN(0, \sigma^2), \]

\[ t = 1, \ldots, T. \]

OLS estimation uses the full sample, \( t = 1, \ldots, T. \)

Recursive least squares uses an expanding sample. Begin with the first \( K \) observations and estimate the model. Then estimate using the first \( K + 1 \) observations, and so on. At the end we have a set of recursive parameter estimates: \( \hat{\beta}_{k,t} \), for \( k = 1, \ldots, K \) and \( t = K, \ldots, T. \)
Recursive Residuals

At each $t$, $t = K, \ldots, T - 1$, compute a 1-step forecast,

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

The corresponding forecast errors, or recursive residuals, are

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

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

where $r_t > 1$ for all $t$
Standardized Recursive Residuals and CUSUM

\[ w_{t+1,t} \equiv \frac{\hat{e}_{t+1,t}}{\sigma \sqrt{r_t}} , \]

\[ t = K, ..., T - 1. \]

Under the maintained assumptions,

\[ w_{t+1,t} \sim iidN(0, 1). \]

Then

\[ CUSUM_{t^*} \equiv \sum_{t=K}^{t^*} w_{t+1,t}, \quad t^* = K, ..., T - 1 \]

is just a sum of iid \( N(0, 1) \)'s.
Recursive Analysis, Constant-Parameter DGP
Recursive Analysis, Breaking-Parameter DGP
Heteroskedasticity in Time Series

Do we really believe that disturbance variances are constant over time?
Varieties of Random (White) Noise

White noise: \( \varepsilon_t \sim WN(\mu, \sigma^2) \) (serially uncorrelated)

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

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

Gaussian white noise: \( \varepsilon_t \sim iid N(0, \sigma^2) \)
Linear Models (e.g., $AR(1)$)

\[ r_t = \phi r_{t-1} + \varepsilon_t \]

\[ \varepsilon_t \sim iid(0, \sigma^2), \quad |\phi| < 1 \]

- Uncond. mean: $E(r_t) = 0$ (constant)
- Uncond. variance: $E(r_t^2) = \sigma^2 / (1 - \phi^2)$ (constant)
- Cond. mean: $E(r_t \mid \Omega_{t-1}) = \phi r_{t-1}$ (varies)
- Cond. variance: $E([r_t - E(r_t \mid \Omega_{t-1})]^2 \mid \Omega_{t-1}) = \sigma^2$ (constant)

- Conditional mean adapts, but conditional variance does not
ARCH(1) Process

\[ r_t | \Omega_{t-1} \sim \mathcal{N}(0, h_t) \]

\[ h_t = \omega + \alpha r_{t-1}^2 \]

\[ E(r_t) = 0 \]

\[ E(r_t^2) = \frac{\omega}{(1 - \alpha)} \]

\[ E(r_t | \Omega_{t-1}) = 0 \]

\[ E([r_t - E(r_t | \Omega_{t-1})]^2 | \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 \]
GARCH(1,1) Process ("Generalized ARCH")

\[ r_t \mid \Omega_{t-1} \sim N(0, h_t) \]

\[ h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} \]

\[ E(r_t) = 0 \]

\[ E(r_t^2) = \frac{\omega}{(1 - \alpha - \beta)} \]

\[ E(r_t \mid \Omega_{t-1}) = 0 \]

\[ E([r_t - E(r_t \mid \Omega_{t-1})]^2 \mid \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} \]

Well-defined and covariance stationary if
\[ 0 < \alpha < 1, \ 0 < \beta < 1, \ \alpha + \beta < 1 \]
GARCH(1,1) and Exponential Smoothing

Exponential smoothing recursion:

\[ \hat{\sigma}^2_t = \lambda \hat{\sigma}^2_{t-1} + (1 - \lambda) r_t^2 \]

\[ \Rightarrow \hat{\sigma}^2_t = (1 - \lambda) \sum_j \lambda^j r_{t-j}^2 \]

But in GARCH(1,1) we have:

\[ h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} \]

\[ h_t = \frac{\omega}{1 - \beta} + \alpha \sum \beta^{j-1} r_{t-j}^2 \]
Tractable Maximum-Likelihood Estimation

\[ L(\theta; r_1, \ldots, r_T) = f(r_T|\Omega_{T-1}; \theta)f((r_{T-1}|\Omega_{T-2}; \theta) \ldots, \]

where \( \theta = (\omega, \alpha, \beta)' \)

If the conditional densities are Gaussian,

\[ f(r_t|\Omega_{t-1}; \theta) = \frac{1}{\sqrt{2\pi} h_t(\theta)} \exp \left( -\frac{1}{2} \frac{r_t^2}{h_t(\theta)} \right), \]

so

\[ \ln L = \text{const} - \frac{1}{2} \sum_t \ln h_t(\theta) - \frac{1}{2} \sum_t \frac{r_t^2}{h_t(\theta)}. \]
Variations on the GARCH Theme

- Regression with GARCH disturbances
- Fat-tailed conditional densities: t-GARCH
- Asymmetric response and the leverage effect: T-GARCH
Regression with GARCH Disturbances

\[ y_t = x_t' \beta + \varepsilon_t \]

\[ \varepsilon_t | \Omega_{t-1} \sim N(0, h_t) \]
Fat-Tailed Conditional Densities: t-GARCH

If \( r \) is conditionally Gaussian, then

\[
r_t = \sqrt{h_t} \ N(0, 1)
\]

But often with high-frequency data,

\[
\frac{r_t}{\sqrt{h_t}} \sim \text{leptokurtic}
\]

So take:

\[
r_t = \sqrt{h_t} \ \frac{t_d}{\text{std}(t_d)}
\]

and treat \( d \) as another parameter to be estimated
Asymmetric Response and the Leverage Effect: T-GARCH

Standard GARCH: \( h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} \)

T-GARCH: \( h_t = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 D_{t-1} + \beta h_{t-1} \)

\[ D_t = \begin{cases} 
1 & \text{if } r_t < 0 \\
0 & \text{otherwise} 
\end{cases} \]

- positive return (good news): \( \alpha \) effect on volatility
- negative return (bad news): \( \alpha + \gamma \) effect on volatility

\( \gamma \neq 0 \): Asymmetric news response
\( \gamma > 0 \): “Leverage effect”
A Useful Specification Diagnostic

\[ \varepsilon_t | \Omega_{t-1} \sim N(0, h_t) \]

\[ \varepsilon_t = \sqrt{h_t} \nu_t, \quad \nu_t \sim iidN(0, 1) \]

\[ \frac{\varepsilon_t}{\sqrt{h_t}} = \nu_t, \quad \nu_t \sim iidN(0, 1) \]

Infeasible: examine \( \nu_t = \varepsilon_t / \sqrt{h_t} \). iid? Gaussian?

Feasible: examine \( \hat{\nu}_t = \varepsilon_t / \sqrt{\hat{h}_t} \). iid? Gaussian?

Key potential deviation from iid is volatility dynamics:
- Examine correlogram of squared standardized returns, \( \hat{\nu}_t^2 \)
- Examine normality of standardized returns, \( \hat{\nu}_t \)
Conditional Mean Estimation

- Dependent Variable: LSALES
- Method: ML ARCH - Normal distribution (BFGS / Marquardt steps)
- Date: 07/05/17  Time: 19:29
- Sample (adjusted): 1987M05 2014M12
- Included observations: 332 after adjustments
- Convergence achieved after 90 iterations
- Coefficient covariance computed using outer product of gradients
- Presample variance: backcast (parameter = 0.7)
- GARCH = C(19) + C(20) RESID(-1)^2 + C(21) GARCH(-1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME</td>
<td>0.000806</td>
<td>0.000290</td>
<td>2.70885</td>
<td>0.0054</td>
</tr>
<tr>
<td>TIME2</td>
<td>-1.34E-06</td>
<td>4.57E-07</td>
<td>-2.933581</td>
<td>0.0034</td>
</tr>
<tr>
<td>D1</td>
<td>0.444664</td>
<td>0.231312</td>
<td>1.922356</td>
<td>0.0546</td>
</tr>
<tr>
<td>D2</td>
<td>0.432966</td>
<td>0.233265</td>
<td>1.856109</td>
<td>0.0634</td>
</tr>
<tr>
<td>D3</td>
<td>0.516899</td>
<td>0.234081</td>
<td>2.208207</td>
<td>0.0272</td>
</tr>
<tr>
<td>D4</td>
<td>0.753770</td>
<td>0.227416</td>
<td>3.314500</td>
<td>0.0009</td>
</tr>
<tr>
<td>D5</td>
<td>0.759088</td>
<td>0.223997</td>
<td>3.388824</td>
<td>0.0007</td>
</tr>
<tr>
<td>D6</td>
<td>0.692614</td>
<td>0.226058</td>
<td>3.063883</td>
<td>0.0022</td>
</tr>
<tr>
<td>D7</td>
<td>0.727432</td>
<td>0.227332</td>
<td>3.198683</td>
<td>0.0014</td>
</tr>
<tr>
<td>D8</td>
<td>0.651924</td>
<td>0.228963</td>
<td>2.847291</td>
<td>0.0044</td>
</tr>
<tr>
<td>D9</td>
<td>0.599579</td>
<td>0.229510</td>
<td>2.612426</td>
<td>0.0090</td>
</tr>
<tr>
<td>D10</td>
<td>0.630021</td>
<td>0.229460</td>
<td>2.745668</td>
<td>0.0060</td>
</tr>
<tr>
<td>D11</td>
<td>0.687792</td>
<td>0.229319</td>
<td>2.999274</td>
<td>0.0027</td>
</tr>
<tr>
<td>D12</td>
<td>1.017883</td>
<td>0.228308</td>
<td>4.458423</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-1)</td>
<td>0.326042</td>
<td>0.064864</td>
<td>5.026551</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-2)</td>
<td>0.266719</td>
<td>0.055166</td>
<td>4.834889</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-3)</td>
<td>0.450762</td>
<td>0.062146</td>
<td>7.253268</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-4)</td>
<td>-0.147672</td>
<td>0.056173</td>
<td>-2.628880</td>
<td>0.0086</td>
</tr>
</tbody>
</table>
Conditional Variance Estimation

Variance Equation

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>RESID(-1)^2</th>
<th>GARCH(-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.000306</td>
<td>0.000129</td>
<td>2.368004</td>
</tr>
<tr>
<td></td>
<td>0.155340</td>
<td>0.087963</td>
<td>1.765959</td>
</tr>
<tr>
<td></td>
<td>0.421656</td>
<td>0.204401</td>
<td>2.062885</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.995317</td>
<td>Mean dependent var</td>
<td>7.107025</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.995063</td>
<td>S.D. dependent var</td>
<td>0.392974</td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>0.027611</td>
<td>Akaike info criterion</td>
<td>-4.294400</td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>0.239377</td>
<td>Schwarz criterion</td>
<td>-4.053713</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>733.8703</td>
<td>Hannan-Quinn criter.</td>
<td>-4.198414</td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>1.935988</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Autocorrelations of Squared Standardized Residuals

Date: 07/05/17   Time: 19:36
Sample: 1987M01 2014M12
Included observations: 332

<table>
<thead>
<tr>
<th>Autocorrelation</th>
<th>Partial Correlation</th>
<th>AC</th>
<th>PAC</th>
<th>Q-Stat</th>
<th>Prob*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>-0.015</td>
<td>0.0762</td>
<td>0.783</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.039</td>
<td>0.5927</td>
<td>0.744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-0.050</td>
<td>1.4339</td>
<td>0.698</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.019</td>
<td>1.5523</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>-0.031</td>
<td>1.8686</td>
<td>0.867</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>-0.059</td>
<td>3.0604</td>
<td>0.801</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>-0.021</td>
<td>3.2176</td>
<td>0.864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.047</td>
<td>3.9859</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.014</td>
<td>4.0534</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>-0.070</td>
<td>5.7601</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>0.017</td>
<td>5.8646</td>
<td>0.882</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>0.036</td>
<td>6.3182</td>
<td>0.899</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>-0.065</td>
<td>7.7802</td>
<td>0.858</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>0.042</td>
<td>8.4003</td>
<td>0.867</td>
</tr>
</tbody>
</table>
Distribution of Standardized Residuals
Time Series of Estimated Conditional Standard Deviations