Econometric Data Science

Francis X. Diebold
University of Pennsylvania

October 22, 2019
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
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.

▶ Emphasis on predictions, guiding decisions
▶ Observational data
▶ Structural change
▶ Volatility fluctuations (”heteroskedasticity”)
▶ Even trickier in time series: Trend, Seasonality, Cycles (”serial correlation”)
Let’s Elaborate on the “Emphasis on Predictions Guiding Decisions” …

Q: What is econometrics about, broadly?

A: Helping people to make better decisions

► Consumers
► Firms
► Investors
► Policy makers
► Courts

Forecasts guide decisions.

Good forecasts promote good decisions.

Hence prediction holds a distinguished place in econometrics, and it will hold a distinguished place in this course.
Types/Arrangements of Economic Data

– Cross section

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

– Time series

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

Much of our discussion will apply to both cross-section and time-series environments, but still we have to pick a notation.
A Few Leading Econometrics Web Data Resources
(Clickable)

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

“High-level” does not mean “best”, and “low-level” does not mean worst. There are many issues.
Graphics Review
Graphics Help us to:

- Summarize and reveal patterns in univariate cross-section data. Histograms and density estimates are helpful for learning about distributional shape. Symmetric, skewed, fat-tailed, ...

- Summarize and reveal patterns in univariate time-series data. Time Series plots are useful for learning about dynamics. Trend, seasonal, cycle, outliers, ...

- Summarize and reveal patterns in multivariate data (cross-section or time-series). Scatterplots are useful for learning about relationships. Does a relationship exist? Is it linear or nonlinear? Are there outliers?
Histogram Revealing Distributional Shape:
1-Year Government Bond Yield
Time Series Plot Revealing Dynamics: 1-Year Government Bond Yield
Scatterplot Revealing Relationship:
1-Year and 10-Year Government Bond Yields
Some Principles of Graphical Style

► Know your audience, and know your goals.
► Appeal to the viewer.
► Show the data, and only the data, 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:

\[ \mu = E(y) = \sum_i p_i y_i \quad \text{(discrete case)} \]

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

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:
\[
\text{cov}(y, x) = E[(y - \mu_y)(x - \mu_x)].
\]

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

Sample mean:

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

Sample variance:

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

Unbiased sample variance:

\[
s^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1}
\]
More Sample Moments

Sample skewness:

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

Sample kurtosis:

\[ \hat{K} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^4 \]
Still More Sample Moments

Sample covariance:

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

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: \( y_i \sim iid \ N(\mu, \sigma^2), \ i = 1, ..., N \)

\( \bar{y} \) is unbiased and normally distributed with variance \( \sigma^2 / N \).

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

and we estimate \( \sigma^2 \) using \( s^2 \), where

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

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

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

where \( t_{1-\frac{\alpha}{2}} (N - 1) \) denotes the appropriate critical value of the Student’s t density with \( N - 1 \) degrees of freedom
Large-Sample Distribution of the Sample Mean
(Requires iid, but not Normality)

Simple random sampling: \( y_i \sim iid (\mu, \sigma^2), i = 1, \ldots, N \)

\( \bar{y} \) is consistent and asymptotically normally distributed with variance \( \nu \).

\[ a \]

\[ \bar{y} \sim N(\mu, \nu), \]

and we estimate \( \nu \) using \( \hat{\nu} = s^2 / N \), where

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

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

As \( N \rightarrow \infty \), \( \mu \in \left[ \bar{y} \pm z_{1-\alpha} \frac{s}{\sqrt{N}} \right] \) w.p. \( 1 - \alpha \)

As \( N \rightarrow \infty \), \( \mu = \mu_0 \quad \Rightarrow \quad \frac{\bar{y} - \mu_0}{s/\sqrt{N}} \sim N(0, 1) \)
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>t($H_0: \mu = 12$)</td>
<td>0.93</td>
<td>-625.70</td>
</tr>
</tbody>
</table>

(p = 0.36) (p = 0.00)

(p = 0.00) (p = 0.53)
Regression
Regression

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

B. As a probabilistic framework for optimal prediction.
Regression as Curve Fitting
Distributions of Log Wage, Education and Experience
Scatterplot: Log Wage vs. Education
Curve Fitting

Fit a line:

\[ y_i = \beta_1 + \beta_2 x_i \]

Solve:

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

“least squares” (LS, or OLS)

“quadratic loss”
Actual Values, Fitted Values and Residuals

Actual values: $y_i, \ i = 1, \ldots, N$

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

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

(“hats” denote fitted things...)

Residuals: $e_i = y_i - \hat{y}_i, \ i = 1, \ldots, N.$
Log Wage vs. Education with Superimposed Regression Line

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

Solve:

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

Fitted hyperplane:

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

More compactly:

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

where \( x_{i1} = 1 \) for all \( i \).

Wage dataset:

\[
\text{LWAGE} = .867 + .093\text{EDUC} + .013\text{EXPER}
\]
Regression as a Probability Model
An Ideal Situation (”The Ideal Conditions”, or IC)

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

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

and the fitted model matches it exactly.

1.1 The fitted model is correctly specified
1.2 The disturbances are Gaussian
1.3 The coefficients (\( \beta_k \)'s) are fixed
1.4 The relationship is linear
1.5 The \( \varepsilon_i \)'s have constant unconditional variance \( \sigma^2 \)
1.6 The \( \varepsilon_i \)'s are uncorrelated

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

2.1 \( E(\varepsilon_i \mid x_{i1}, \ldots, x_{iK}) = 0, \) for all \( i \)
2.2 \( \text{var}(\varepsilon_i \mid x_{i1}, \ldots, x_{iK}) = \sigma^2, \) for all \( i \)

(Written here for cross sections. Slight changes in 2.1, 2.2 for time series.)
Some Concise Matrix Notation
(Useful for Notation, Estimation, Inference)

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

\[
y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1K} \\ 1 & x_{22} & x_{23} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & x_{N3} & \cdots & x_{NK} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}
\]
Elementary Matrices and Matrix Operations

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

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.
The DGP in Matrix Form, Written Out

\[
\begin{pmatrix}
    y_1 \\
    y_2 \\
    \vdots \\
    y_N
\end{pmatrix}
= 
\begin{pmatrix}
    1 & x_{12} & x_{13} & \ldots & x_{1K} \\
    1 & x_{22} & x_{23} & \ldots & x_{2K} \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    1 & x_{N2} & x_{N3} & \ldots & x_{NK}
\end{pmatrix}
\begin{pmatrix}
    \beta_1 \\
    \beta_2 \\
    \vdots \\
    \beta_K
\end{pmatrix}
+ 
\begin{pmatrix}
    \varepsilon_1 \\
    \varepsilon_2 \\
    \vdots \\
    \varepsilon_N
\end{pmatrix}
\sim \mathcal{N}
\begin{pmatrix}
    \begin{pmatrix}
        0 \\
        0 \\
        \vdots \\
        0
    \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}
\end{pmatrix}

y = X\beta + \varepsilon

\varepsilon \sim \mathcal{N}(0, \sigma^2 I)
Three Notations

Original form:

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

\[ i = 1, 2, \ldots, N \]

Intermediate form:

\[ y_i = x_i^\prime \beta + \varepsilon_i, \quad \varepsilon_i \sim iidN(0, \sigma^2) \]

\[ i = 1, 2, \ldots, N \]

Full matrix form:

\[ y = X \beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I) \]
We used to write this: The DGP is

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

and the fitted model matches it exactly, and

\[ \varepsilon_i \text{ is independent of } (x_{i1}, \ldots, x_{iK}), \text{ for all } i \]

Now, equivalently, we write this: The DGP is

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

and the fitted model matches it exactly, and

\[ \varepsilon_i \text{ is independent of } x_i, \text{ for all } i \]
The OLS Estimator in Matrix Notation

As before, the LS estimator solves:

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

Now, in matrix notation:

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

It can be shown that the solution is:

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

Under the IC

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

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

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

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

Note the precise parallel with the large-sample distribution of the sample mean.
Sample Mean, Regression on an Intercept, and Properties of Residuals

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

– Intercept picks up a “level effect”

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

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

Conditional mean:

\[ E(y_i \mid x_i=x^*) = x^*\beta \]

Conditional variance:

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

Full conditional density:

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

Why All the Talk About Conditional Moment Implications?
“Point Prediction”

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

“The conditional mean is the minimum MSE point predictor”

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

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

– LS delivers operational optimal predictor with great generality
– Follows immediately from the LS optimization problem
“Interval Prediction”

Non-operational (in reality we don’t know $\beta$ or $\sigma$):

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

Operational:

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

(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:

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

Operational version:

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

(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.)
“Typical” Regression Analysis of Wages, Education and Experience

![Wage Regression Output Image]

<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**: 199.6260
- **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 criterion**: 1.427291
- **Durbin-Watson stat**: 1.926045
“Top Matter”: Background Information

- Dependent variable
- Method
- Date
- Sample
- Included observations
“Middle Matter”: Estimated Regression Function

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

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

– The standard errors, $t$ statistics, and $p$-values let us do statistical inference as to which regressors are most relevant for predicting $y$. 
There are many...
Regression Statistics: Mean dependent var 2.342

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

\[ SD = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N - 1}} \]
Predictive Perspectives

– The sample standard deviation of $y$ is a measure of the in-sample accuracy of the unconditional mean forecast $\bar{y}$. 
Regression Statistics: Sum squared resid 319.938

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

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

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

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

$SSR$ measures “total” in-sample predictive accuracy

“squared-error loss”

“quadratic loss”

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

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

(“average” in-sample predictive accuracy)
Regression Statistics: \( F \)-statistic 199.626

\[
F = \frac{(SSR_{\text{res}} - SSR)/(K - 1)}{SSR/(N - K)}
\]
– 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 .492

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

\[ SER = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N - K}} \]
$s^2$ is just $SSR$ scaled by $N - K$, so again, it’s a measure of the in-sample accuracy of the regression-based forecast.

Like MSE, but corrected for degrees of freedom.
Regression Statistics:
*R*-squared .232, Adjusted *R*-squared .231

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

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

“What percent of variation in \( y \) is explained by variation in \( x \)”
Predictive Perspectives

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

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

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

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

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^{N} e_i^2}{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2}$$
$R_k^2$ and “Multicollinearity” (not shown in the computer output)

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

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

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

It can be shown that:

$$\text{var}(\hat{\beta}_k) = f \left( \frac{\sigma^2}{+}, \frac{\sigma_{x_k}^2}{-}, \frac{R_k^2}{+} \right)$$
Predictive Perspectives

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

– But we still might see evidence of a strong overall predictive relationship.
  (High $F$-stat)
Regression Statistics: Log likelihood -938.236

Understanding this requires some background / detail:

- **Likelihood** – joint density of the data (the \( y_i \)’s)

- **Maximum-likelihood estimation** – natural estimation strategy: find the parameter configuration that maximizes the likelihood of getting the \( y_i \)’s that you actually *did* get.

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

- **Hypothesis tests** are based on log likelihood
Detail: Maximum-Likelihood Estimation

Linear regression DGP (under the IC) implies that:

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

so that

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

Now by independence of the \( \varepsilon_i \)'s and hence \( y_i \)'s,

\[ L = f(y_1, ..., y_N | x_i) = f(y_1 | x_1) \cdots f(y_N | x_N) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - x_i' \beta)^2} \]
Detail: Log Likelihood

\[
\ln L = \ln \left( (2\pi\sigma^2)^{-\frac{N}{2}} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x'_i\beta)^2
\]

\[
= -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - x'_i\beta)^2
\]

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

– The \( \beta \) vector that maximizes the likelihood is the \( \beta \) vector that minimizes the sum of squared residuals

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

– “MLE and OLS coincide for linear regression under the IC” (Normality, in particular, is crucial)
It can be shown that, under the null hypothesis (that is, if the restrictions imposed under the null are true):

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

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

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

- Gaussian $L$ is intimately related to $SSR$

- Therefore $L$ is closely related to prediction (and measuring predictive accuracy) as well

- Small $SSR \iff$ large $L$
Regression Statistics: Schwarz criterion 1.435
Akaike info criterion 1.422

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

Figure: Wage Regression Residual Scatter
Residual Plot

Figure: Wage Regression Residual Plot
Predictive Perspectives

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

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

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

  For example:

  1. The true DGP may be nonlinear
  2. $\varepsilon$ may be non-Gaussian
  3. $\varepsilon$ may have non-constant variance
Misspecification and Model Selection

Do we really believe that the fitted model matches the DGP?
Regression Statistics:
Akaike info criterion 1.422, Schwarz criterion 1.435

SSR versions:

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

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

More general \textit{lnL} versions:

$$AIC = -2\text{lnL} + 2K$$

$$SIC = -2\text{lnL} + K\text{ln}N$$
Penalties
Predictive Perspectives

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

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

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

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

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

“Oracle property”
Non-Normality and Outliers

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

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

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

– Dealing with non-normality, outliers, and influential observations
  (LAD regression, simulation-based density forecasts, ...)
Large-Sample Distribution of $\hat{\beta}_{LS}$
Under the Ideal Conditions (Except Normality)

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

$$a$$

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

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

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

No change from asymptotic result under IC!
So why worry about normality?

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

- Interval and density prediction change fundamentally
Jarque-Bera Normality Test

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

– Jarque-Bera test. Under normality we have:

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

– Many more
Recall Our OLS Wage Regression

**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:** -938.2358  
- **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 citer.: **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: 1.765517
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

Quantiles of Normal vs Quantiles of RESID
Residual Scatter

Figure: Wage Regression Residual Scatter
OLS Residual Plot
Leave-One-Out Plot

Consider:

\[
\left( \hat{\beta}^{(-i)} - \hat{\beta} \right), \quad i = 1, \ldots, N
\]

“Leave-one-out plot”
Wage Regression

Leave-One-Out Plot

Coefficient (Education)

0.090

0.094

Leave t out

0

200

400

600

800

1000

1200
Robust Estimation: Least Absolute Deviations (LAD)

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

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

- Not as simple as OLS, but still simple

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

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

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

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

Equation: UNTITLED  Workfile: CPS 1995 E View:::Output95_update

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:

\[ y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \ldots, N, \]

subject to the IC, except that we allow

\[ \varepsilon_i \sim iid D(0, \sigma^2) \]
Consider a density forecast for a person $t$ with characteristics $x_i = x_i^*$.

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

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

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

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

2. Add $x_i^* \hat{\beta}$ to each draw.

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

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

As $R \to \infty$ and $N \to \infty$, all error vanishes.
Indicator Variables in Cross Sections: Group Effects

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

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

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

(It really is that simple.)

“Intercept dummies”
Histograms for Wage Covariates

Notice that the sample mean of an indicator variable is the fraction of the sample with the indicated attribute.
Recall Basic Wage Regression on Education and Experience

$LWAGE \rightarrow C, EDUC, EXPER$
Basic Wage Regression Results

Dependent Variable: LWAGE
Method: Least Squares
Date: 06/27/13 Time: 16:38
Sample (adjusted): 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 | 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 | 319.9378 | Schwarz criterion | 1.434644|
Log likelihood | -938.2358 | Hannan-Quinn criter. | 1.427291|
F-statistic | 199.6260 | Durbin-Watson stat | 1.926045|
Prob(F-statistic) | 0.000000 |
Introducing Sex, Race, and Union Status in the Wage Regression

Now:

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

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

![Regression Output](attachment:image.png)

**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>18.86314</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**  
**Adjusted R-squared**  
**S.E. of regression**  
**Sum squared resid**  
**Log likelihood**  
**F-statistic**  
**Prob(F-statistic)**
Predictive Perspectives

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

**Wage Regression With Dummies**
- Conditions on education, experience, \textit{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

<table>
<thead>
<tr>
<th>LS // Dependent Variable is Y1</th>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>X1</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>R-squared</td>
<td>0.67</td>
<td></td>
<td>1.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LS // Dependent Variable is Y2</th>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>X2</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>R-squared</td>
<td>0.67</td>
<td></td>
<td>1.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LS // Dependent Variable is Y3</th>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>X3</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>R-squared</td>
<td>0.67</td>
<td></td>
<td>1.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LS // Dependent Variable is Y4</th>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>T-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>3.00</td>
<td>1.12</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>X4</td>
<td>0.50</td>
<td>0.12</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>R-squared</td>
<td>0.67</td>
<td></td>
<td>1.24</td>
</tr>
</tbody>
</table>
Anscombe’s Quartet Graphics: Dataset 1
Anscombe’s Quartet Graphics: Dataset 2
Anscombe’s Quartet Graphics: Dataset 3
Anscombe’s Quartet Graphics: Dataset 4

![Graph of Y4 vs. X4 showing a linear relationship with scattered data points.](image)
Log-Log Regression

\[ \ln y_i = \beta_1 + \beta_2 \ln x_i + \varepsilon_i \]

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

Example: Cobb-Douglas production function

\[ y_i = AL_i^\alpha K_i^\beta \exp(\varepsilon_i) \]

\[ \ln y_i = \ln A + \alpha \ln L_i + \beta \ln K_i + \varepsilon_i \]

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

\[ \ln y_i = \beta x_i + \varepsilon \]

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

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

One example is the “S-curve” model,

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

(\(0 < r < 1\))

– No way to transform to linearity

– Minimize the sum of squared errors numerically
  “Nonlinear least squares”
  \(\hat{\beta}_{NLS}\)
Really no such thing as an intrinsically non-linear model...

In the bivariate case we can think of the relationship as

\[ y_i = g(x_i, \varepsilon_i) \]

or slightly less generally as

\[ y_i = f(x_i) + \varepsilon_i \]
Consider Taylor series expansions of $f(x_i)$. The linear (first-order) approximation is

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

and the quadratic (second-order) approximation is

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

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

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

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

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

– One can always fit a linear model

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

– Just see the earlier Dataset 2 Anscombe graph!
Assessing Non-Linearity
(i.e., deciding on higher-order Taylor terms)

Use SIC as always.

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

Figure: Basic Linear Wage Regression

```
Dependent Variable: LWAGE
Method: Least Squares
Date: 07/03/13   Time: 13:36
Sample (adjusted): 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>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>18.86314</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    | 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 criterion | 1.332532 |
F-statistic      | 117.1568   | Durbin-Watson stat  | 1.910120 |
Prob(F-statistic) | 0.000000  |
```

Path = c:\users\francis x. diebo\documents\diebold files\courses\econ104\old\econ104_2011\sw3e\views materials
Quadratic 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>0.473236</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 |
### Quadratic Wage Regression with Dummy Interactions

#### Figure: Wage Regression with Continuous Non-Linearities and Interactions, and Discrete Interactions

- **Date:** 10/02/13
- **Sample:** 1,1323
- **Included observations:** 1,1323

<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>0.482967</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>0.5611</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_EXPR</td>
<td>-0.001796</td>
<td>0.000429</td>
<td>-4.185787</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>0.055077</td>
<td>-2.932246</td>
<td>0.0034</td>
</tr>
<tr>
<td>FEM_UNI</td>
<td>-0.012956</td>
<td>0.070740</td>
<td>-0.183153</td>
<td>0.8547</td>
</tr>
<tr>
<td>FEM_NON</td>
<td>0.110319</td>
<td>0.070093</td>
<td>1.573909</td>
<td>0.1157</td>
</tr>
<tr>
<td>UNI_NON</td>
<td>0.033202</td>
<td>0.089258</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
- **Sum squared resid:** 273.2109
- **Schwarz criterion:** 1.325858
- **Hannan-Quinn criter.:** 1.296244
- **Durbin-Watson stat:** 1.891544
Final Specification

Dependent Variable: LWAGE
Method: Least Squares
Date: 10/03/13  Time: 11:19
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>0.360535</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>0.066130</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>0.0067</td>
</tr>
</tbody>
</table>

R-squared
Adjusted R-squared
S.E. of regression
Sum squared resid
Log likelihood
F-statistic
Prob(F-statistic)

Mean dependent var 2.341995
S.D. dependent var 0.561435
Akaike info criterion 1.274755
Schwarz criterion 1.306124
Hannan-Quinn criter. 1.286514
Durbin-Watson stat 1.894273

Figure: “Final” Wage Regression
Discrete Response Models

What if the dependent variable is binary?

– Ultimately violates the IC in multiple ways...

(Nonlinear, non-Gaussian)
Many Names

“discrete response models”

“qualitative response models”

“limited dependent variable models”

“binary (binomial) response models”

“classification models”

“logistic regression models” (a leading case)

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

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

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

The usual linear regression setup,

\[
E(y_i|x_i) = x_i'\beta
\]

becomes

\[
E(I_i(z) \mid x_i) = x_i'\beta.
\]

A key insight, however, is that

\[
E(I_i(z) \mid x_i) = P(I_i(z)=1 \mid x_i),
\]

so the setup is really

\[
P(I_i(z)=1 \mid x_i) = x_i'\beta. \quad (1)
\]

– 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 \( l_i(z) \rightarrow x_i \).

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

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

More precisely, move to models with

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

where $F(\cdot)$ is monotone increasing, with $\lim_{w \to \infty} F(w) = 1$ and $\lim_{w \to -\infty} F(w) = 0$. 
In the “logit” model, the squashing function $F(\cdot)$ is the logistic function,

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

so the logit model is

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

- Logit is a nonlinear model for the event probability.
Logit as a Linear Model for the Log Odds

Consider a linear model for log odds

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

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

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

So logit is just linear regression for log odds.
Logit Estimation

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

For linear regression we had:

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

from which we derived the likelihood and the MLE.

For the linear probability model we have:

\[ y_i | x_i \sim Bernoulli (x_i' \beta). \]

For logit we have:

\[ y_i | x_i \sim Bernoulli \left( \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}} \right). \]
Note that the individual RHS variable effects, $\partial E(y_i|x_i)/\partial x_{ik}$, are not simply given by the $\beta_k$’s as in standard linear regression. Instead we have

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

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

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

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

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

For binary regression we proceed similarly:

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

“Efron’s $R^2$”
The Logit Classifier

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

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

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

– Note the “linear decision boundary”. We can generalize to nonlinear decision boundaries in various ways.
Example: High-Wage Individuals

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

– 357 people with $HIGHWAGE_i = 1$

– 966 people with $HIGHWAGE_i = 0$

We will fit a logit model using education and experience and see how it performs as a Bayes classifier.
Logit Regression of *HIGHWAGE* on *EDUC* and *EXPER*

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th><em>HIGHWAGE</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>EDUC</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>(se=0.03)</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(se=0.01)</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.61</td>
</tr>
<tr>
<td></td>
<td>(se=0.46)</td>
</tr>
</tbody>
</table>

Ratio of Effects *EDUC/EXPER*: 7.54

Efron’s $R^{2}$: 0.15
Covariates and Decision Boundary for Logit Bayes Classifier

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

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

Do we really believe that disturbance variances are constant over space?
“Unconditional Heteroskedasticity” is Occasionally Relevant...

Consider IC1:

\[ \varepsilon_i \sim iidN(0, \sigma^2), \quad i = 1, \ldots, N \]

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

Violation of IC1, in particular IC1.5:

“The \( \varepsilon_i \)’s have constant variance \( \sigma^2 \)”
... But *Conditional* Heteroskedasticity is Often Highly Relevant

Consider IC2.2:

\[
var(\varepsilon_i \mid x_{i1}, \ldots, x_{iK}) = \sigma_i^2, \text{ for all } i
\]

Conditional heteroskedasticity occurs when \(\sigma_i^2\) varies systematically with \(x_{i1}, \ldots, x_{iK}\), so that IC2.2 is violated

e.g., Consider the regression

\[
\text{fine wine consumption} \rightarrow \text{income}
\]
Consequences for Estimation and Inference

- Estimation: OLS estimation remains largely OK. Parameter estimates remain consistent and asymptotically normal.

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

– Earlier point forecasts remain largely OK.

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

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

It is no longer appropriate to base interval and density forecasts on “identical \( \sigma \)’s for different people”. Now we need to base them on “different \( \sigma \)’s for different people”.
Detecting Conditional Heteroskedasticity

- Graphical heteroskedasticity diagnostics
- Formal heteroskedasticity tests
Graphical Diagnostics

Graph $e_i^2$ against $x_{ik}$, for various regressors $(k)$
Recall Our “Final” Wage Regression
Squared Residual vs. EDUC
The Breusch-Pagan-Godfrey Test (BPG)

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

In contrast, BPG blends information from all regressors

BPG test:

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

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

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

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

▶ Estimate the OLS regression, and obtain the squared residuals

▶ Regress the squared residuals on all regressors, squared regressors, and pairwise regressor cross products

▶ To test the null hypothesis of no relationship, examine \( N \cdot R^2 \) from this regression. In large samples \( N \cdot R^2 \sim \chi^2_{K-1} \) under the null.
White’s Test

```
ls l wage c educ exper exper2 edu_exp female union nonwhite
```
Dealing with Heteroskedasticity

- Adjusting standard errors
- Adjusting density forecasts
Adjusting Standard Errors

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

“Heteroskedasticity-robust standard errors”
“White standard errors”

Before, under the IC:

\[
V = \text{cov}(\hat{\beta}_{LS}) \text{ estimated by } \\
\hat{V} = s^2 (X'X)^{-1},
\]

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

Now, under heteroskedasticity, \( V \) estimated by

\[
\hat{V}_{\text{White}} = (X'X)^{-1} (X' \text{diag}(e_1^2, ..., e_N^2)X) (X'X)^{-1}
\]

– Mechanically, it’s just a simple OLS regression option.
Final Wage Regression with Robust 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>0.360535</td>
<td>0.130391</td>
<td>2.765026</td>
<td>0.0058</td>
</tr>
<tr>
<td>EDUC</td>
<td>0.125028</td>
<td>0.009890</td>
<td>12.64118</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.066130</td>
<td>0.006967</td>
<td>9.491284</td>
<td>0.0000</td>
</tr>
<tr>
<td>EXPER2</td>
<td>-0.000710</td>
<td>0.007657</td>
<td>-0.040820</td>
<td>0.9668</td>
</tr>
<tr>
<td>EDU_EXP</td>
<td>-0.001905</td>
<td>0.000412</td>
<td>-4.623006</td>
<td>0.0000</td>
</tr>
<tr>
<td>FEMALE</td>
<td>-0.239352</td>
<td>0.025499</td>
<td>-9.386559</td>
<td>0.0000</td>
</tr>
<tr>
<td>UNION</td>
<td>0.202574</td>
<td>0.031386</td>
<td>6.454196</td>
<td>0.0000</td>
</tr>
<tr>
<td>NONWHITE</td>
<td>-0.094903</td>
<td>0.034074</td>
<td>-2.785164</td>
<td>0.0054</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
- Durbin-Watson stat: 1.894273
Adjusting Density Forecasts

Recall non-operational version for Gaussian homoskedastic disturbances:

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

Recall operational version for Gaussian homoskedastic disturbances:

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

Now: Operational version for Gaussian heteroskedastic disturbances:

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

Q: Where do we get \( \hat{\sigma}^2 \)?
Time Series
Misspecification and Model Selection

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

Do we really believe that the disturbances are Gaussian?
No major changes in time series
Same tools and techniques...
Indicator Variables in Time Series I: Trend

Trend is effectively a type of structural change

Do we really believe that intercepts are fixed over time?

– Trend is about \textit{gradual} intercept evolution
From now on we will take logs of liquor sales. When we say “liquor sales”, logs are understood.
Linear Trend

\[ Trend_t = \beta_1 + \beta_2 \, TIME_t \]

where \( TIME_t = t \)

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

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

TREND=10-.25*TIME

TREND=-50+.8*TIME

Time

Trend
**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  
S.E. of regression 0.159743 Akaike info criterion -0.824561  
Sum squared resid 8.523001 Schwarz criterion -0.801840  
Log likelihood 140.5262 Hannan-Quinn criter. -0.815504  
F-statistic 1797.705 Durbin-Watson stat 1.078573  
Prob(F-statistic) 0.000000
Residual Plot
Indicator Variables in Time Series II: Seasonality

Seasonality is effectively a type of structural change

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

\[ \text{Seasonal}_s = \sum_{s=1}^{S} \beta_s \text{SEAS}_{st} \quad (S \text{ seasons per year}) \]

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

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

where (e.g., in quarterly data case, assuming Q1 start and Q4 end):
\[
\begin{align*}
\text{SEAS}_1 &= (1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, \ldots, 0)'
\text{SEAS}_2 &= (0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, \ldots, 0)'
\text{SEAS}_3 &= (0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, \ldots, 0)'
\text{SEAS}_4 &= (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, \ldots, 1)'.
\end{align*}
\]

- Full set of dummies ("all categories") and hence no intercept.
- In CS case we dropped a category for each dummy (e.g., included "UNION" but not "NONUNION") and included an intercept.
Linear Trend with Seasonal Dummies

<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>6.46819</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  Mean dependent var: 7.096188
Adjusted R-squared: 0.924350  S.D. dependent var: 0.402962
S.E. of regression: 0.110833  Akaike info criterion: -1.523658
Sum squared resid: 3.967734  Schwarz criterion: -1.375972
Log likelihood: 268.9746  Hannan-Quinn criterion: -1.464876
Durbin-Watson stat: 0.100500
Seasonal Pattern
Residual Plot
Nonlinearity in Time Series

Do we really believe that trends are linear?
Non-Linear Trend: Exponential (Log-Linear)

\[ Trend_t = \beta_1 e^{\beta_2 \text{TIME}_t} \]

\[ \ln(Trend_t) = \ln(\beta_1) + \beta_2 \text{TIME}_t \]
Figure: Various Exponential Trends
Non-Linear Trend: Quadratic

Allow for gentle curvature by including \( \text{TIME} \) and \( \text{TIME}^2 \):

\[
\text{Trend}_t = \beta_1 + \beta_2 \text{TIME}_t + \beta_3 \text{TIME}_t^2
\]
Figure: Various Quadratic Trends
#### Liquor Sales Quadratic Trend Estimation

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

<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  
- **Adjusted R-squared:** 0.903097  
- **S.D. dependent var:** 0.402962  
- **Akaike info criterion:** -1.305106  
- **Schwarz criterion:** -1.271025  
- **Hannan-Quinn criter.:** -1.291521  
- **F-statistic:** 1562.036  
- **Durbin-Watson stat:** 1.754412  
- **Prob(F-statistic):** 0.000000

---

**Figure:**

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

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

<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
Serial Correlation

Do we really believe that disturbances are uncorrelated over time?
(Not possible in cross sections, so we didn’t study it before...)
Serial Correlation is Another Type of Violation of the IC (This time it’s “correlated disturbances”.)

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

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

Key cause: Omission of serially-correlated $x$’s, which produces serially-correlated $\varepsilon$
Disturbance serial correlation, or autocorrelation, means *correlation over time*
– Current disturbance correlated with past disturbance(s)

Leading example
(“AR(1)” disturbance serial correlation):

\[ y_t = x_t' \beta + \varepsilon_t \]
\[ \varepsilon_t = \phi \varepsilon_{t-1} + v_t, \quad |\phi| < 1 \]
\[ v_t \sim iid \ N(0, \sigma^2) \]

(Extension to “AR(p)” disturbance serial correlation is immediate)
Consequences for $\beta$ Estimation and Inference:
As with Heteroskedasticity, Point Estimation is OK, but Inference is Damaged

– Estimation: OLS estimation of $\beta$ remains largely OK. Parameter estimates remain consistent and asymptotically normal

– Inference: OLS inference is damaged. Standard errors are biased and inconsistent.
Consequences for $y$ Prediction:
Unlike With Heteroskedasticity,
Even \textit{Point} Predictions are Damaged/

Serial correlation is a bigger problem for prediction than heteroskedasticity.

Here’s the intuition:

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

Put differently:
Serial correlation in forecast errors means that you can forecast your forecast errors! So something is wrong and can be improved...
Some Important Language and Tools
For Characterizing Serial Correlation

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

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

“Partial autocorrelations”: $p_{\varepsilon}(\tau)$, $\tau = 1, 2, \ldots$

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

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

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

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

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

$$e_t \rightarrow c, e_{t-1}, \ldots, e_{t-(\tau-1)}, e_{t-\tau}$$
White Noise Disturbances

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

\[\text{iid}\]

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

\[\text{iid}\]

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

We write:

\(\varepsilon_t \sim \mathcal{WN}(0, \sigma^2)\)
Realization of White Noise Process
Population Autocorrelation Function
White Noise Process

Autocorrelation

Displacement
Population Partial 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)

\[ y_t = \phi y_{t-1} + \epsilon_t \]

where $\phi = 0.4$ and $\phi = 0.95$.
Population Autocorrelation Function
AR(1) Process, $\phi = .4$

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

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

```
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
```

**Figure:** Liquor Sales Log-Quadratic Trend + Seasonal Estimation
Residual Plot
Residual Scatterplot \((e_t \text{ vs. } e_{t-1})\)
Residual Correlogram

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>0.707</td>
<td>169.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.689</td>
<td>0.379</td>
<td>330.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>0.729</td>
<td>0.371</td>
<td>512.06</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>

Bartlett standard error \(\left(=\frac{1}{\sqrt{T}}\right)=\frac{1}{\sqrt{336}}\) = .055

95 % Bartlett band \(\left(=\pm\frac{2}{\sqrt{T}}\right)=\pm.11\)
Formal Tests: Durbin-Watson (0.59)

Simple AR(1) environment:

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

\[ \varepsilon_t = \phi \varepsilon_{t-1} + \nu_t \]

\[ \nu_t \sim iid \ N(0, \sigma^2) \]

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

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

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

Hence as \( T \to \infty \):

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

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

General $AR(p)$ environment:

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

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

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

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 $TR^2 \sim \chi^2_p$ under the null.
BG for AR(1) Disturbances

\( TR^2 = 168.5, \ p = 0.0000 \)

Figure: BG Test Regression, AR(1)
BG for $AR(4)$ Disturbances ($TR^2 = 216.7, \ p = 0.0000$)

Figure: BG Test Regression, $AR(4)$
BG for $AR(8)$ Disturbances
($TR^2 = 219.0, \ p = 0.0000$)

Figure: BG Test Regression, $AR(8)$
Robust Estimation with Serial Correlation

Recall our earlier “heteroskedasticity robust s.e.’s”

We can also consider “serial correlation robust s.e.’s”

The simplest way is to include lags of $y$ as regressors...
Modeling Serial Correlation: Including Lags of $y$ as Regressors

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

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

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

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

\[
y_t = \beta_1 + \beta_2 x_t + \epsilon_t \\
\epsilon_t = \phi \epsilon_{t-1} + \nu_t \\
\nu_t \sim iid \ N(0, \sigma^2)\]
Liquor Sales: Everything Consistent With $AR(4)$ Dynamics

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

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

$AR(1) = -3.797$
$AR(2) = -3.941$
$AR(3) = -4.080$
$AR(4) = -4.086$
$AR(5) = -4.071$
$AR(6) = -4.058$
$AR(7) = -4.057$
$AR(8) = -4.040$
Trend + Seasonal Model with Four Lags of $y$

![Image of Screenshot of Statistical Analysis Software]

<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>0.000993</td>
<td>0.000303</td>
<td>3.274924</td>
<td>0.0012</td>
</tr>
<tr>
<td>TIME2</td>
<td>-1.63E-06</td>
<td>4.82E-07</td>
<td>-3.379250</td>
<td>0.0008</td>
</tr>
<tr>
<td>D1</td>
<td>0.577182</td>
<td>0.240084</td>
<td>2.404080</td>
<td>0.0168</td>
</tr>
<tr>
<td>D2</td>
<td>0.579618</td>
<td>0.239820</td>
<td>2.416883</td>
<td>0.0162</td>
</tr>
<tr>
<td>D3</td>
<td>0.667059</td>
<td>0.236827</td>
<td>2.795401</td>
<td>0.0055</td>
</tr>
<tr>
<td>D4</td>
<td>0.894665</td>
<td>0.237447</td>
<td>3.767847</td>
<td>0.0002</td>
</tr>
<tr>
<td>D5</td>
<td>0.893728</td>
<td>0.232717</td>
<td>3.840401</td>
<td>0.0001</td>
</tr>
<tr>
<td>D6</td>
<td>0.827871</td>
<td>0.233806</td>
<td>3.540838</td>
<td>0.0005</td>
</tr>
<tr>
<td>D7</td>
<td>0.865982</td>
<td>0.235247</td>
<td>3.681158</td>
<td>0.0003</td>
</tr>
<tr>
<td>D8</td>
<td>0.791626</td>
<td>0.236419</td>
<td>3.348398</td>
<td>0.0009</td>
</tr>
<tr>
<td>D9</td>
<td>0.739295</td>
<td>0.237199</td>
<td>3.116777</td>
<td>0.0020</td>
</tr>
<tr>
<td>D10</td>
<td>0.771468</td>
<td>0.236858</td>
<td>3.257093</td>
<td>0.0012</td>
</tr>
<tr>
<td>D11</td>
<td>0.830449</td>
<td>0.236573</td>
<td>3.510331</td>
<td>0.0005</td>
</tr>
<tr>
<td>D12</td>
<td>1.156887</td>
<td>0.236231</td>
<td>4.897183</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-1)</td>
<td>0.348107</td>
<td>0.055751</td>
<td>6.243965</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-2)</td>
<td>0.257435</td>
<td>0.053823</td>
<td>4.783041</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-3)</td>
<td>0.429234</td>
<td>0.053804</td>
<td>7.977784</td>
<td>0.0000</td>
</tr>
<tr>
<td>LSALES(-4)</td>
<td>-0.161633</td>
<td>0.055771</td>
<td>-2.898162</td>
<td>0.0040</td>
</tr>
</tbody>
</table>

- R-squared: 0.995335
- Mean dependent var: 7.107025
- Adjusted R-squared: 0.995082
- S.D. dependent var: 0.392974
- S.E. of regression: 0.027559
- Akaike info criterion: -4.292292
- Sum squared resid: 0.238480
- Schwarz criterion: -4.085990
- Log likelihood: 730.5205
- Hannan-Quinn criter.: -4.210019
- Durbin-Watson stat: 1.982921
Trend + Seasonal Model with Four Lags of $y$

Residual Plot
Trend + Seasonal Model with Four Lags of $y$
Residual Scatterplot
Trend + Seasonal Model with Four Lags of $y$
Residual Autocorrelations

<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>0.0077</td>
<td>0.930</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.003</td>
<td>0.0100</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>-0.057</td>
<td>1.0926</td>
<td>0.779</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>-0.009</td>
<td>1.1176</td>
<td>0.891</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>-0.002</td>
<td>1.1190</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>0.017</td>
<td>1.2161</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>-0.025</td>
<td>1.4316</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.068</td>
<td>2.9954</td>
<td>0.935</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9</td>
<td>0.059</td>
<td>4.1934</td>
<td>0.898</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>-0.123</td>
<td>9.4373</td>
<td>0.491</td>
</tr>
</tbody>
</table>
Trend + Seasonal Model with Four Lags of $y$
Residual Histogram and Normality Test

![Residual Histogram](image)

**Observations:** 312

- **Mean:** $3.77E-16$
- **Median:** $0.000160$
- **Maximum:** $0.078468$
- **Minimum:** $-0.109856$
- **Std. Dev.:** $0.026635$
- **Skewness:** $0.077911$
- **Kurtosis:** $3.740378$
- **Jarque-Bera:** $7.441714$
- **Probability:** $0.024213$
Forecasting Time Series

It's more interesting than in cross sections...
The “Forecasting the Right-Hand-Side Variables Problem”

For now assume known parameters.

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

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

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 a pure AR(1)

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

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

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

No FRVP for \( h = 1 \). There seems to be an FRVP for \( h > 1 \).
But there’s not...

We build the multi-step forecast recursively.

First 1-step, then 2-step, etc.
“Wold’s chain rule of forecasting”
Assuming Gaussian shocks, we immediately have

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

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

The question is how to get

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

[Of course to make things operational we eventually replace parameters with estimates and use \( \mathcal{N}(\hat{y}_{t+h}, t, \hat{\sigma}^2_{t+h}, t) \).]
Interval and Density Forecasting
(1-Step-Ahead, AR(1))

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

Back substitution yields

\[ y_t = \varepsilon_t + \phi \varepsilon_{t-1} + \phi^2 \varepsilon_{t-2} + \phi^3 \varepsilon_{t-3} + \ldots \]

\[ \implies y_{t+1} = \varepsilon_{t+1} + \phi \varepsilon_t + \phi^2 \varepsilon_{t-1} + \phi^3 \varepsilon_{t-2} + \ldots \]

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

\[ y_{t+1,t} = \phi \varepsilon_t + \phi^2 \varepsilon_{t-1} + \phi^3 \varepsilon_{t-2} + \ldots \]

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

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

with variance

\[ \sigma^2_{t+1,t} = var(e_{t+1,t}) = \sigma^2 \]
Interval and Density Forecasting
(*h*-Step-Ahead, \(AR(p)\))

\[ y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t \quad \varepsilon_t \sim WN(0, \sigma^2) \]

Back substitution yields

\[ y_t = \varepsilon_t + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + b_3 \varepsilon_{t-3} + \ldots \]

\[ \implies y_{t+h} = \varepsilon_{t+h} + b_1 \varepsilon_{t+h-1} + \ldots + b_{h-1} \varepsilon_{t+1} + b_h \varepsilon_t + b_{h+1} \varepsilon_{t-1} + \ldots \]

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

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

\[ y_{t+h,t} = b_h \varepsilon_t + b_{h+1} \varepsilon_{t-1} + b_{h+2} \varepsilon_{t-2} + \ldots \]

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

\[ e_{t+h,t} = y_{t+h} - y_{t+h,t} = \varepsilon_{t+h} + b_1 \varepsilon_{t+h-1} + \ldots + b_{h-1} \varepsilon_{t+1} \]

with variance (non-decreasing in \(h\)):

\[ \sigma_{t+h,t}^2 = var(e_{t+h,t}) = \sigma^2 (1 + b_1^2 + \ldots + b_{h-1}^2) \]
Liquor Sales History and 1- Through 12-Month-Ahead Point and Interval Forecasts From Trend + Seasonal Model with Four Lags of y
Now With Realization Superimposed...
SIC Estimates of Out-of-Sample Forecast Error Variance

\[ \text{LSALES}_t \rightarrow c, \text{TIME}_t \]
\[ \text{SIC} = 0.45 \]

\[ \text{LSALES}_t \rightarrow c, \text{TIME}_t, \text{TIME}^2_t \]
\[ \text{SIC} = 0.28 \]

\[ \text{LSALES}_t \rightarrow \text{TIME}_t, \text{TIME}^2_t, D_{t,1}, \ldots, D_{t,12} \]
\[ \text{SIC} = 0.04 \]

\[ \text{LSALES}_t \rightarrow \text{TIME}_t, \text{TIME}^2_t, D_{t,1}, \ldots, D_{t,12}, \text{LSALES}_{t-1}, \ldots, \text{LSALES}_{t-4} \]
\[ \text{SIC} = 0.02 \]

(We report exponentiated SIC’s because the software actually reports \( \ln(\text{SIC}) \))
Structural Change in Time Series: Evolution or Breaks in Any or all Parameters

Do we really believe that parameters are fixed over time?
Structural Change
Single Sharp Break, Exogenously Known

For simplicity of exposition, consider a bivariate regression:

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

Let

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

Then we can write the model as:

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

We run:

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

Use regression to test for structural change (F test)
Use regression to accommodate structural change if present.
The “Chow test” is what we’re really calculating:

\[
Chow = \frac{(e'e - (e'_1e_1 + e'_2e_2))/K}{(e'_1e_1 + e'_2e_2)/(T - 2K)}
\]

Distributed $F$ under the no-break null (and the rest of the IC)
Structural Change
Sharp Breakpoint, Endogenously Identified

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

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

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

The null distribution of \( MaxChow \) has been tabulated.
Recursive Parameter Estimates

For generic parameter $\beta$, calculate and examine

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

for $t = 1, \ldots, T$

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

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

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

The corresponding forecast errors, or recursive residuals, are

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

Under the IC (including structural stability),

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

where $r_{t+1,t} = 1 + x_{t+1}'(X_t'X_t)^{-1}x_{t+1}$
Standardized Recursive Residuals and CUSUM

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

\[ t = 1, \ldots, T - 1 \text{ (leaving room for startup)} \]

Under the IC,

\[ \hat{w}_{t+1,t} \sim iid N(0,1). \]

Then

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

(leaving room for startup)

is just a sum of iid \( N(0,1) \)'s, and its 95% bounds have been tabulated.
Recursive Analysis, Constant-Parameter DGP
Recursive Analysis, Breaking-Parameter DGP
Liquor Sales Model: Recursive Parameter Estimates
Liquor Sales Model: Recursive Residuals With Two Standard Error Bands
Liquor Sales Model: CUSUM
Vector Autoregressions

What if we have more than one time series?
Basic Framework

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

\[
\begin{align*}
y_{1,t} & = c_1 + \phi_{11}y_{1,t-1} + \phi_{12}y_{2,t-1} + \varepsilon_{1,t} \\
y_{2,t} & = c_2 + \phi_{21}y_{1,t-1} + \phi_{22}y_{2,t-1} + \varepsilon_{2,t}
\end{align*}
\]

\[
\begin{align*}
\varepsilon_{1,t} & \sim \text{WN}(0, \sigma_1^2) \\
\varepsilon_{2,t} & \sim \text{WN}(0, \sigma_2^2) \\
\text{cov}(\varepsilon_{1,t}, \varepsilon_{2,t}) & = \sigma_{12}
\end{align*}
\]

- Can extend to \(N\)-variable VAR(\(p\))
- Estimation by OLS (as before)
- Can include trends, seasonals, etc. (as before)
- Forecasts via Wold’s chain rule (as before)
- Order selection by information criteria (as before)
- Can do predictive causality analysis (coming)
U.S. Housing Starts and Completions, 1968.01-1996.06
Starts Sample Autocorrelations
Starts Sample Partial Autocorrelations
Completions Sample Autocorrelations
Completions Sample Partial Autocorrelations
Starts and Completions: Sample Cross Correlations

![Cross Correlation Chart](image)

- **Cross Correlation**
- **Displacement**
VAR Starts Equation

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

<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>0.146871</td>
<td>0.044235</td>
<td>3.320264</td>
<td>0.0010</td>
</tr>
<tr>
<td>STARTS(-1)</td>
<td>0.659939</td>
<td>0.061242</td>
<td>10.77587</td>
<td>0.0000</td>
</tr>
<tr>
<td>STARTS(-2)</td>
<td>0.229632</td>
<td>0.072724</td>
<td>3.157587</td>
<td>0.0018</td>
</tr>
<tr>
<td>STARTS(-3)</td>
<td>0.142859</td>
<td>0.072655</td>
<td>1.966281</td>
<td>0.0503</td>
</tr>
<tr>
<td>STARTS(-4)</td>
<td>0.007806</td>
<td>0.066032</td>
<td>0.118217</td>
<td>0.9060</td>
</tr>
<tr>
<td>COMPS(-1)</td>
<td>0.031611</td>
<td>0.102712</td>
<td>0.307759</td>
<td>0.7585</td>
</tr>
<tr>
<td>COMPS(-2)</td>
<td>-0.120781</td>
<td>0.103847</td>
<td>-1.163069</td>
<td>0.2458</td>
</tr>
<tr>
<td>COMPS(-3)</td>
<td>-0.020601</td>
<td>0.100946</td>
<td>-0.204078</td>
<td>0.8384</td>
</tr>
<tr>
<td>COMPS(-4)</td>
<td>-0.027404</td>
<td>0.094569</td>
<td>-0.289779</td>
<td>0.7722</td>
</tr>
</tbody>
</table>

R-squared    | 0.895566    | Mean dependent var | 1.574771 |
Adjusted R-squared | 0.892528 | S.D. dependent var | 0.382362 |
S.E. of regression | 0.125350 | Akaike info criterion | -4.122118 |
Sum squared resid  | 4.320952 | Schwarz criterion | -4.006482 |
Log likelihood    | 191.3622   | F-statistic      | 294.7796  |
Durbin-Watson stat | 1.991908 | Prob(F-statistic) | 0.000000 |
VAR Starts Equations Residual Plot
VAR Starts Equation Residual Sample Autocorrelations
VAR Starts Equation Residual Sample Partial Autocorrelations
VAR Completions Equation

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

<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>0.045347</td>
<td>0.025794</td>
<td>1.758045</td>
<td>0.0799</td>
</tr>
<tr>
<td>STARTS(-1)</td>
<td>0.074724</td>
<td>0.035711</td>
<td>2.092461</td>
<td>0.0373</td>
</tr>
<tr>
<td>STARTS(-2)</td>
<td>0.040047</td>
<td>0.042406</td>
<td>0.944377</td>
<td>0.3458</td>
</tr>
<tr>
<td>STARTS(-3)</td>
<td>0.047145</td>
<td>0.042366</td>
<td>1.112805</td>
<td>0.2668</td>
</tr>
<tr>
<td>STARTS(-4)</td>
<td>0.082331</td>
<td>0.038504</td>
<td>2.138238</td>
<td>0.0334</td>
</tr>
<tr>
<td>COMPS(-1)</td>
<td>0.236774</td>
<td>0.059893</td>
<td>3.953313</td>
<td>0.0001</td>
</tr>
<tr>
<td>COMPS(-2)</td>
<td>0.206172</td>
<td>0.060554</td>
<td>3.404742</td>
<td>0.0008</td>
</tr>
<tr>
<td>COMPS(-3)</td>
<td>0.120998</td>
<td>0.058863</td>
<td>2.055593</td>
<td>0.0408</td>
</tr>
<tr>
<td>COMPS(-4)</td>
<td>0.156729</td>
<td>0.055144</td>
<td>2.842160</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

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

![Graph showing VAR completions equation residual plot with lines for residual, actual, and fitted data.]
VAR Completions Equation Residual Sample Autocorrelations
VAR Completions Equation Residual Sample Partial Autocorrelations

![Graph showing completions residual partial autocorrelation versus displacement](image-url)
Predictive Causality Analysis

Table 8
Housing Starts and Completions
Causality Tests

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

<table>
<thead>
<tr>
<th>Null Hypothesis</th>
<th>F-Statistic</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>STARTS does not Cause COMPS</td>
<td>26.2658</td>
<td>0.00000</td>
</tr>
<tr>
<td>COMPS does not Cause STARTS</td>
<td>2.23876</td>
<td>0.06511</td>
</tr>
</tbody>
</table>
Starts History and Forecast
...Now With Starts Realization
Completions History and Forecast
Now With Completions Realization
Heteroskedasticity in Time Series

Do we really believe that disturbance variances are constant over time?
Dynamic Volatility is the Key to Finance and Financial Economics

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

Figure: Time Series of Daily NYSE Returns.
Returns are Approximately Serially Uncorrelated

Figure: Correlogram of Daily NYSE Returns.

So returns are approximately white noise. But...
Returns are not Unconditionally Gaussian...

**Figure:** Histogram and Statistics for Daily NYSE Returns.

**Series:** R  
**Sample:** 3461  
**Observations:** 3461

- **Mean:** 0.000522  
- **Median:** 0.000640  
- **Maximum:** 0.047840  
- **Minimum:** -0.063910  
- **Std. Dev.:** 0.008541  
- **Skewness:** -0.505540  
- **Kurtosis:** 8.535016

- **Jarque-Bera:** 4565.446  
- **Probability:** 0.000000
Unconditional Volatility Measures

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

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

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

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

Outlier probability: $P|r_t - \mu| > 5\sigma$ (for example)
Returns are Not Homoskedastic

Figure: Time Series of Daily Squared NYSE Returns.
Indeed Returns are Highly Conditionally Heteroskedastic...

Figure: Correlogram of Daily Squared NYSE Returns.
Standard Models (e.g., AR(1)) Fail to Capture the Conditional Heteroskedasticity…

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

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

Conditional mean:
\[ E(r_t | \Omega_{t-1}) = \phi r_{t-1} \text{ (varies)} \]

Conditional variance:
\[ var(r_t | \Omega_{t-1}) = \sigma^2 \text{ (constant)} \]
...So Introduce Special Heteroskedastic Disturbances

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

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

Now consider:

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

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

“GARCH(1,1) Process”

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

\[ \text{var}(r_t | \Omega_{t-1}) = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (\text{varies}) \]

For modeling daily asset returns we can simply use:

\[ r_t | \Omega_{t-1} \sim N(0, \sigma_t^2) \]
GARCH(1,1) and “Exponential Smoothing”

GARCH(1,1):

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

Solving backward:

\[ \sigma_t^2 = \frac{\omega}{1 - \beta} + \alpha \sum_{j=1}^{\infty} \beta^{j-1} r_{t-j}^2 \]
Unified Framework

▶ Conditional variance dynamics (of course, by construction)

▶ Conditional variance dynamics produce unconditional leptokurtosis, even in our conditionally Gaussian setup (So conditional variance dynamics and unconditional fat tails are intimately related)

▶ Returns are non-Gaussian weak white noise (Serially uncorrelated but nevertheless dependent, due to conditional variance dynamics – today’s conditional variance depends on the past.)
Extension: Regression with GARCH Disturbances (GARCH-M)

Standard GARCH regression:

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

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

GARCH-in mean (GARCH-M) regression:

\[ r_t = x_t' \beta + \gamma \sigma_t + \varepsilon_t \]

\[ \varepsilon_t | \Omega_{t-1} \sim \mathcal{N}(0, \sigma_t^2) \]
Extension: Fat-Tailed Conditional Densities (t-GARCH)

If $r$ is conditionally Gaussian, then

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

or

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

But often with high-frequency data,

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

So take:

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

and treat $d$ as another parameter to be estimated.
Extension: Asymmetric Response and the Leverage Effect (Threshold GARCH)

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

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

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

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”
GARCH(1,1) MLE for Daily NYSE Returns

Figure: GARCH(1,1) Estimation

- Dependent Variable: R
- Method: ML - ARCH (Marquardt) - Normal distribution
- Date: 12/03/12  Time: 07:58
- Sample: 1 3461
- Included observations: 3461
- Convergence achieved after 16 iterations
- Presample variance: backcast (parameter = 0.7)
- GARCH = C(2) + C(3)*RESID(-1)*2 + C(4)*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>C</td>
<td>0.000641</td>
<td>0.000127</td>
<td>5.039437</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- Variance Equation

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>1.06E-06</td>
<td>1.49E-07</td>
<td>7.127979</td>
<td>0.0000</td>
</tr>
<tr>
<td>RESID(-1)*2</td>
<td>0.067408</td>
<td>0.004959</td>
<td>13.59218</td>
<td>0.0000</td>
</tr>
<tr>
<td>GARCH(-1)</td>
<td>0.919717</td>
<td>0.006128</td>
<td>150.0893</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- R-squared: -0.000193
- Adjusted R-squared: -0.000193
- S.E. of regression: 0.008542
- Akaike info criterion: -6.868008
- Schwarz criterion: -6.860901
- Hannan-Quinn criterion: -6.865470
- Durbin-Watson stat: 1.861386
“Fancy” GARCH(1,1) MLE

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

<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>@SQRT(GARCH)</td>
<td>0.083360</td>
<td>0.053138</td>
<td>1.568753</td>
<td>0.1167</td>
</tr>
<tr>
<td>C</td>
<td>1.28E-05</td>
<td>0.000372</td>
<td>0.034443</td>
<td>0.9725</td>
</tr>
<tr>
<td>R(-1)</td>
<td>0.073763</td>
<td>0.017611</td>
<td>4.188535</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Variance Equation

<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>C</td>
<td>1.03E-06</td>
<td>2.23E-07</td>
<td>4.628790</td>
<td>0.0000</td>
</tr>
<tr>
<td>RESID(-1)^2</td>
<td>0.014945</td>
<td>0.009765</td>
<td>1.530473</td>
<td>0.1259</td>
</tr>
<tr>
<td>RESID(-1)^2*(RESID(-1)&lt;0)</td>
<td>0.094014</td>
<td>0.014945</td>
<td>6.290700</td>
<td>0.0000</td>
</tr>
<tr>
<td>GARCH(-1)</td>
<td>0.922745</td>
<td>0.009129</td>
<td>101.0741</td>
<td>0.0000</td>
</tr>
<tr>
<td>T-DIST. DOF</td>
<td>5.531579</td>
<td>0.478432</td>
<td>11.56188</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Fitted GARCH Volatility

Figure: Estimated Conditional Standard Deviation, Daily NYSE Returns.
A Useful Specification Diagnostic

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

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

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

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

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

Figure: Correlogram of Squared Standardized Returns
GARCH Volatility Forecast

Figure: Conditional Standard Deviation, History and Forecast
Volatility Forecasts Feed Into Return Density Forecasts

In earlier linear (AR) environment we wrote:

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

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

Now we have:

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

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