



Introduction to MicroEconometrics

Ass. Prof. Andriy Stavytskyy

Agenda

- Econometric review
- Econometric tests

ECONOMETRIC REVIEW



Econometric analysis

- Theoretical approach
- Empirical approach

Types of Data and Notation

- ✓ Time series data
- ✓ Cross-sectional data
- ✓ Panel data, a combination of mentioned above types

Time series data

- **The data may be**
 - quantitative (e.g. exchange rates, stock prices, number of shares outstanding),
 - qualitative (e.g. day of the week).
- **Examples of time series data**

Series

GNP or unemployment
government budget deficit
money supply
value of a stock
market index

Frequency

monthly or quarterly
annually
weekly
as transactions occur




Examples of Problems Using Time Series Regression

1. How the value of a country's stock index has varied with that country's macroeconomic fundamentals.
2. How the value of a company's stock price has varied when it announced the value of its dividend payment.
3. The effect on country's currency of an increase in its interest rate.

Cross-sectional data

- Cross-sectional data is data on one or more variables collected at a single point in time, e.g.
 - A poll of usage of internet stock broking services
 - Cross-section of stock returns on the New York Stock Exchange
 - A sample of bond credit ratings for UK banks



Examples of Problems Using a Cross-Sectional Regression

- The relationship between company **size** and the **return** to investing in its shares
- The relationship between a country's **GDP level** and the **probability** that the government will **default** on its sovereign debt.

Panel Data

- Panel Data has the dimensions of both time series and cross-sections, e.g. the *daily prices of number of blue chip stocks over two years*.
- It is common to denote that each observation by the letter **t** and the total number of observations by **T** for time series data, and to denote each observation by the letter **i** and the total number of observations by **N** for cross-sectional data.

Goal

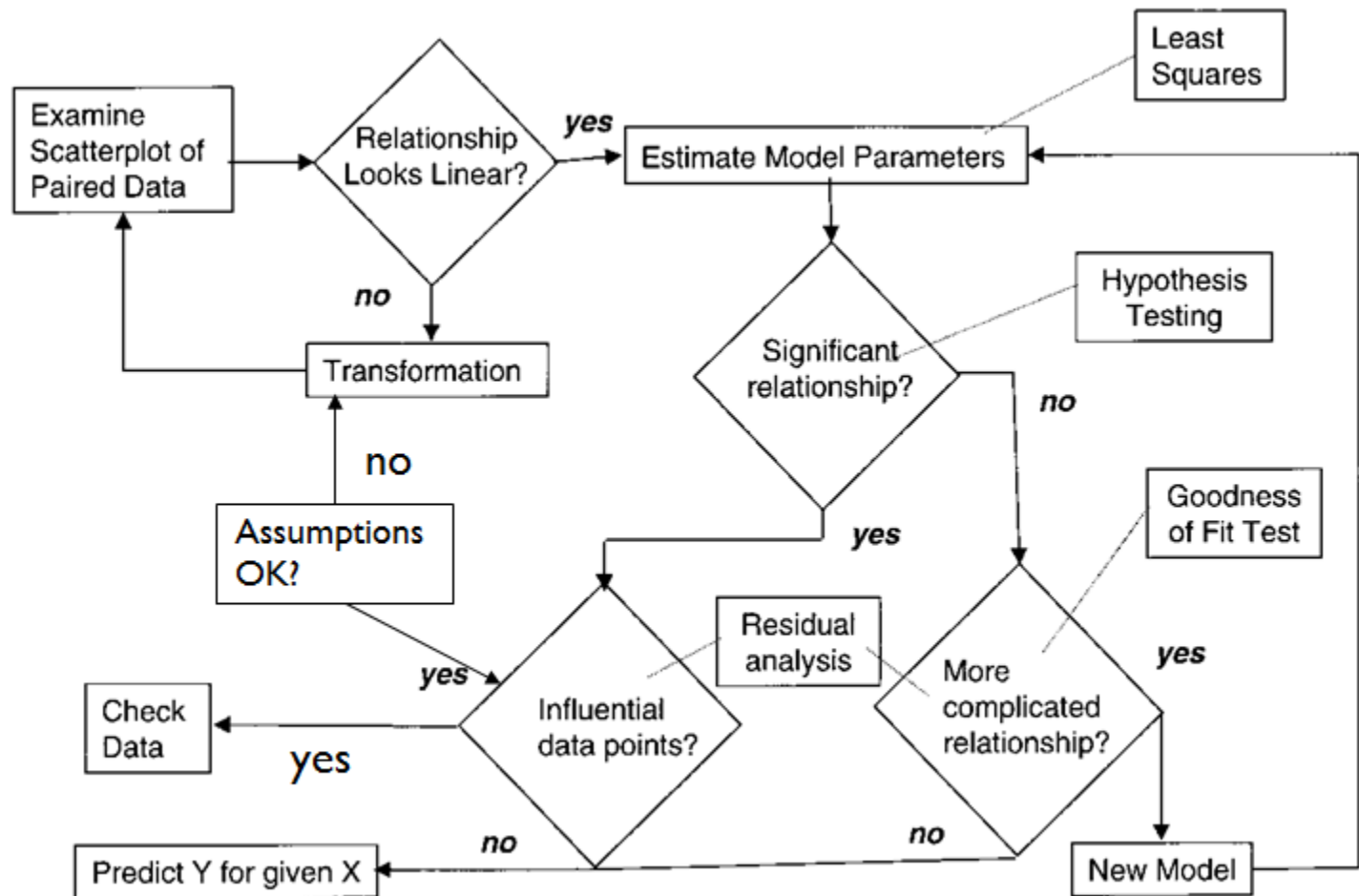
- Develop a **statistical model** that can predict the values of a dependent (response) variable based upon the values of the independent (explanatory) variables.



Regression Modeling Steps

- Define a problem or question
- Specify model
- Collect data
- Do descriptive data analysis
- Estimate unknown parameters
- Evaluate model
- Use model for prediction

How is a Linear Regression Analysis done?



Linear regression

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_{k-1} x_{k-1t} + \varepsilon_t, t = \overline{1, n}$$

y_t - dependent variable;

$x_{1t}, x_{2t}, \dots, x_{k-1t}$ independent variables;

ε_t - residuals.

Assumptions

- **Linearity** - the Y variable is linearly related to the value of the X variable.
- **Independence of Error** - the error (residual) is independent for each value of X.
- **Homoscedasticity** - the variation around the line of regression be constant for all values of X.
- **Normality** - the values of Y be normally distributed at each value of X.

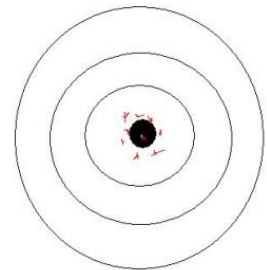
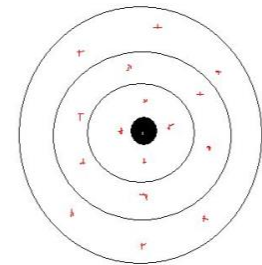
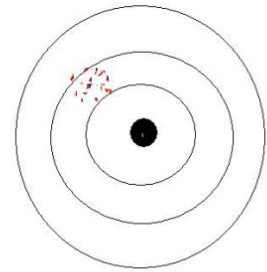
Method of Least Squares

- The **straight line** that best fits the data.
- Determine the straight line for which the differences between the actual values (Y) and the values that would be predicted from the fitted line of regression (Y-hat) are as small as possible.

$$L = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 \rightarrow \min$$

The Three Desirable Characteristics

- **Lack of bias** $E(\hat{\beta}) = \beta$
- **Efficiency**
 - Standard error will be minimum
 - Remember:
$$\text{var}(\hat{\beta}) = \frac{1}{\sum x_i^2} \sigma^2 = \frac{\sigma^2}{\sum x_i^2}$$
 - OLS will minimize σ^2 (the error variance)
- **Consistency**
 - As N increases the standard error decreases
 - Notice: as N increases so does $\sum x_i^2$



Inherently Linear Models

- Non-linear models that can be expressed in linear form
 - *Can be estimated by least square in linear form*
- Require data transformation

Dummy-Variable Regression Model

- Involves categorical X variable with two levels
 - *e.g., female-male, employed-not employed, etc.*
- Variable levels coded 0 & 1
- Assumes only intercept is different
 - *Slopes are constant across categories*



ECONOMETRIC TESTS

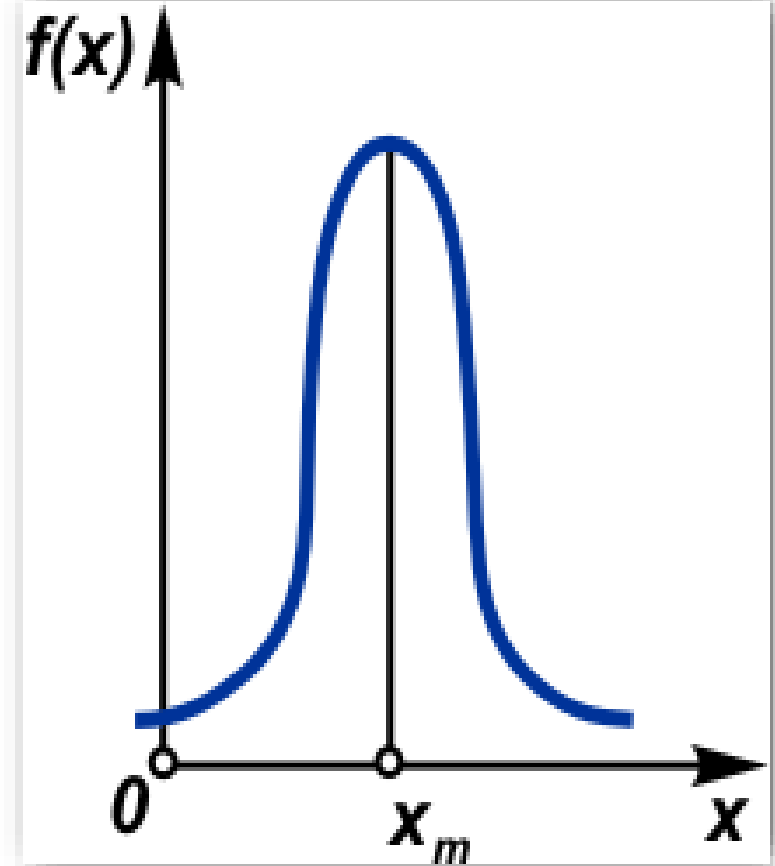
Multiple Regression Tests

- Test residual for normality
- Test parameter significance
 - Overall model
 - Individual coefficients
- Test for multicollinearity
- Test for model stability
- Test for residuals autocorrelation
- Test for residuals homoscedasticity
- Test for specification
- Test for stationary process

Test residual for normality

Check normality of residuals:

- Jarque-Bera statistics
- Shapiro–Wilk test



Jarque-Bera statistics

$$JB = \frac{n}{6} \left(S^2 + \frac{1}{4} \left((K - 3)^2 \right) \right)$$

$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$	$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$
---	---

- *S is the sample skewness,*
- *K is the sample kurtosis.*

Example

Dependent Variable: TAX_ENT

Method: Least Squares

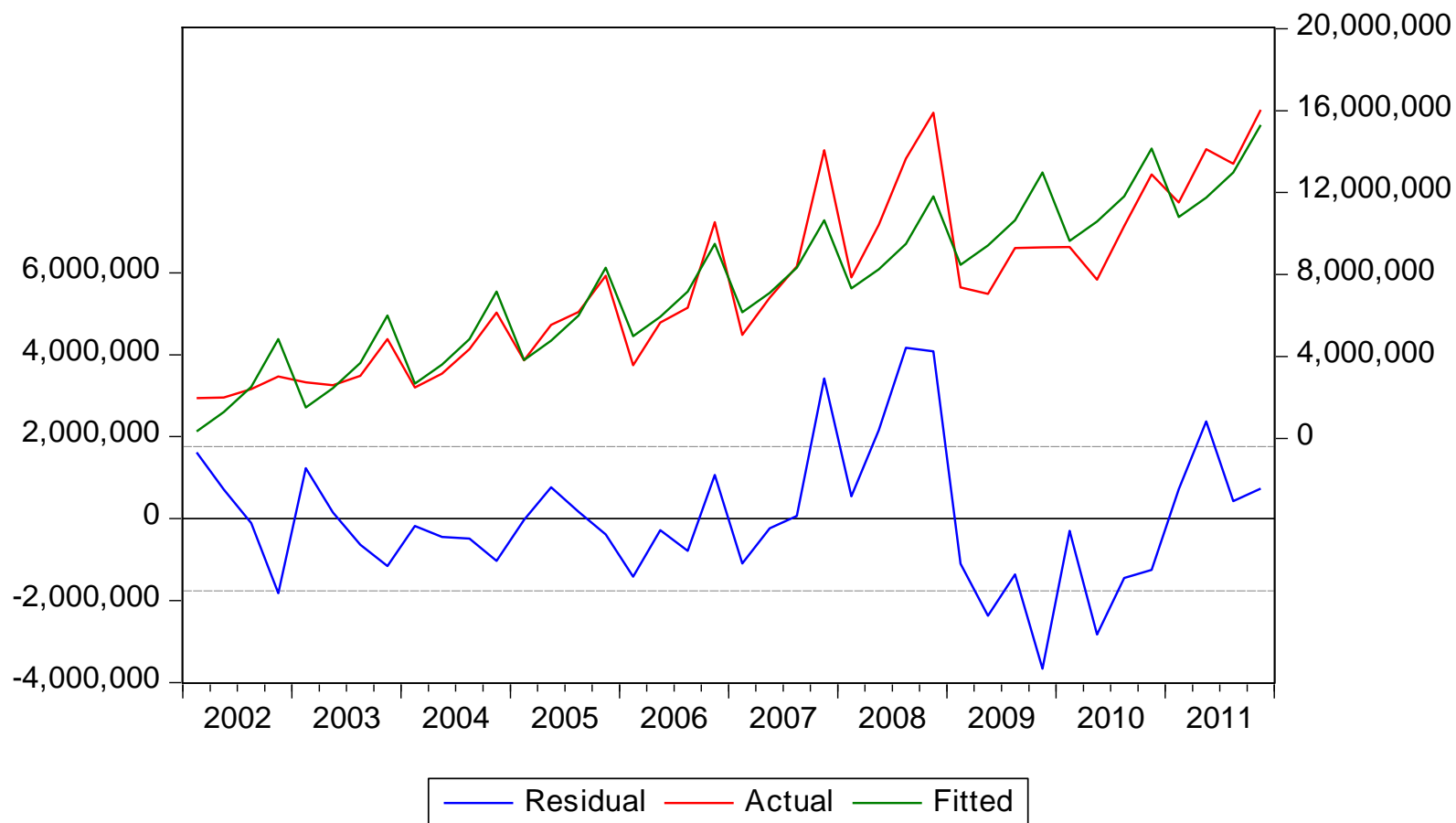
Date: 12/09/12 Time: 20:49

Sample: 2002Q1 2011Q4

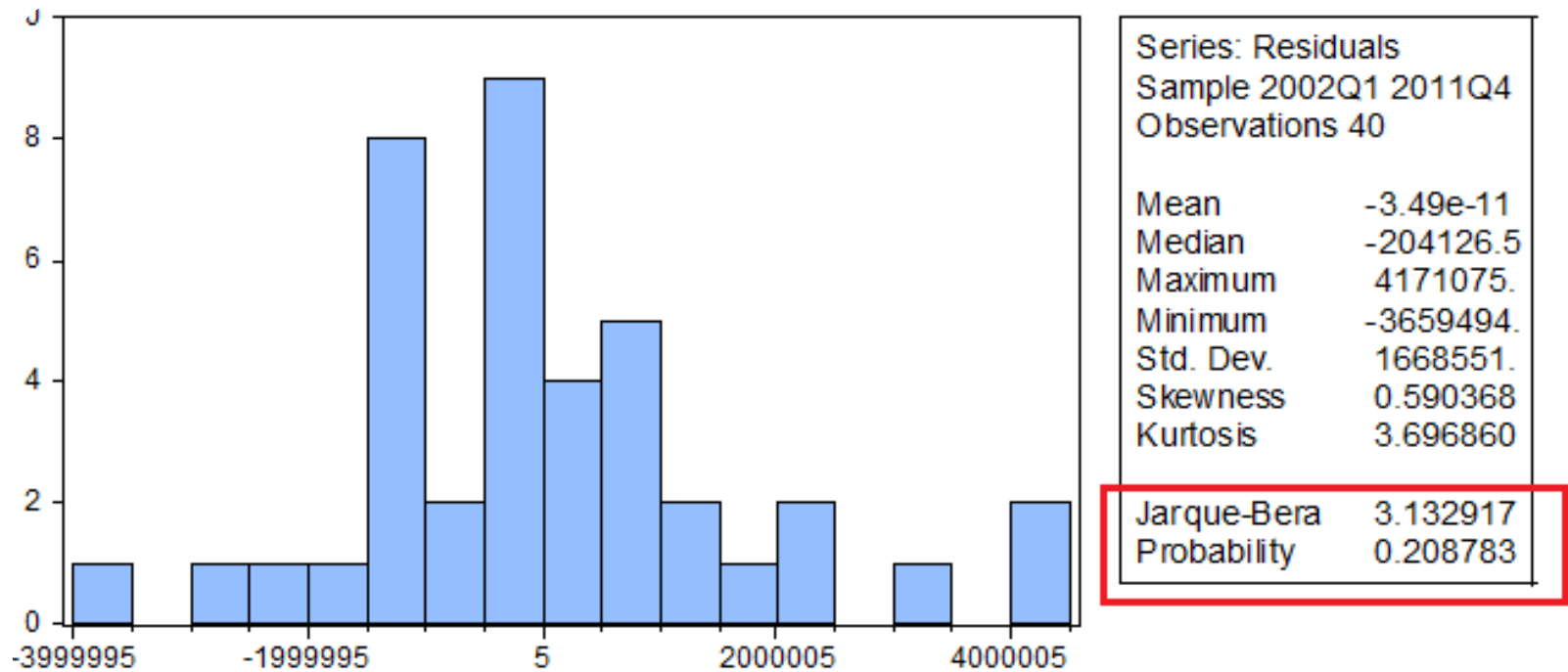
Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143
R-squared	0.837415	Mean dependent var		7480035.
Adjusted R-squared	0.818834	S.D. dependent var		4138083.
S.E. of regression	1761318.	Akaike info criterion		31.71749
Sum squared resid	1.09E+14	Schwarz criterion		31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.		31.79382
F-statistic	45.06800	Durbin-Watson stat		1.123746
Prob(F-statistic)	0.000000			

Residuals



Check for normality



Test parameter significance: Overall model

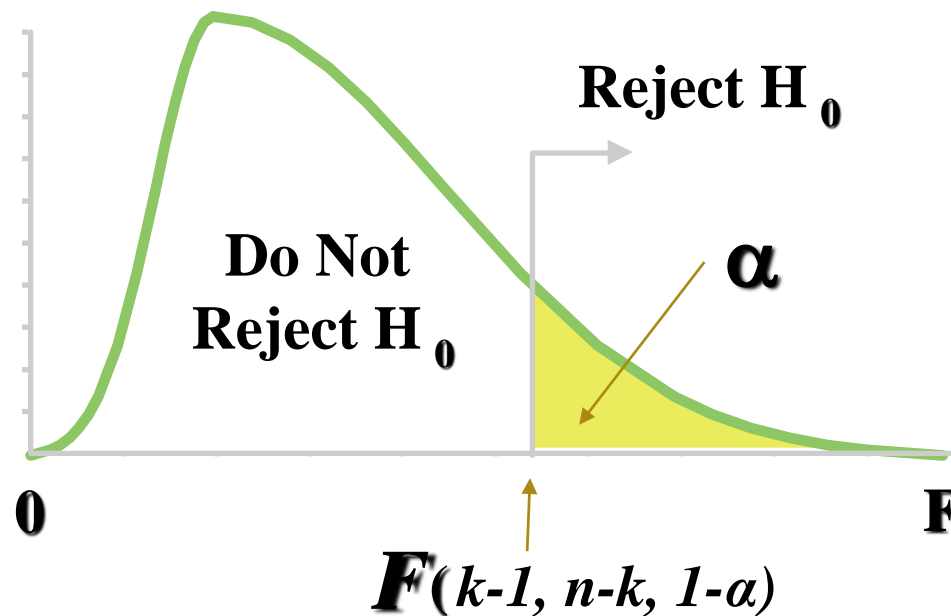
- Hypotheses

- $H_0: \beta_1 = \beta_2 = \dots = \beta_{k-1} = 0$
 - *No Linear Relationship*
- $H_a: \text{At Least One Coefficient Is Not 0}$
 - *At Least One X Variable linearly Affects Y*

$$F = \frac{RSS / (k - 1)}{ESS / (n - k)} = \frac{R^2 / (k - 1)}{(1 - R^2) / (n - k)} \stackrel{H_0}{\sim} F_{k-1, n-k}$$

Overall Significance Rejection Rule

- Reject H_0 in favor of H_a if F_{calc} falls in colored area



- Reject H_0 for H_a if P-value = $P(F > F_{\text{calc}}) < \alpha$

Example

Dependent Variable: TAX_ENT

Method: Least Squares

Date: 12/09/12 Time: 20:49

Sample: 2002Q1 2011Q4

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143

R-squared	0.837415	Mean dependent var	7480035.
Adjusted R-squared	0.818834	S.D. dependent var	4138083.
S.E. of regression	1761318.	Akaike info criterion	31.71749
Sum squared resid	1.09E+14	Schwarz criterion	31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.	31.79382
F-statistic	45.06800	Durbin-Watson stat	1.123746
Prob(F-statistic)	0.000000		

Test of slope coefficients

- Hypotheses
 - $H_0: \beta_i = m$
 - $H_a: \beta_i \neq m$

Slope Coefficient Test Statistic

$$t = \frac{\hat{\beta}_i - m}{S_{\hat{\beta}_i}}$$

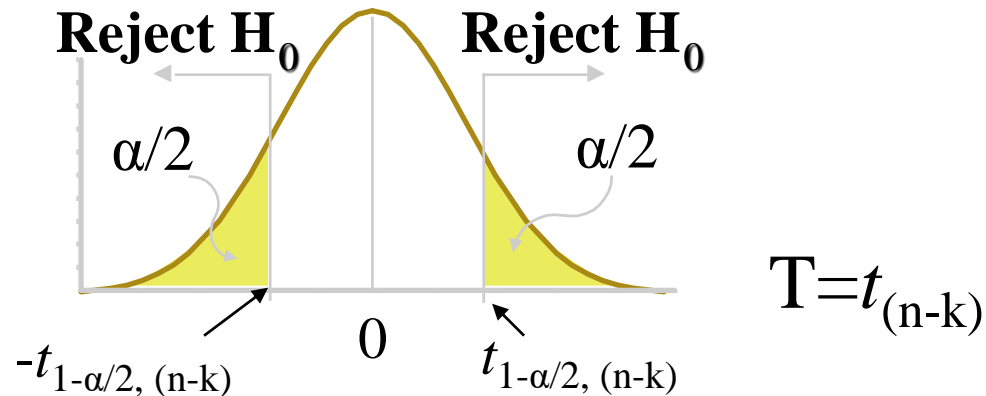
$$\text{where } S_{\hat{\beta}_i} = \frac{S}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}}$$

$$\text{with } S = \hat{\sigma} = \sqrt{\frac{RSS}{n-k}}$$

$$\text{and } RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \left[Y_i - \left(\hat{\beta}_0 + \sum_{i=1}^{k-1} \hat{\beta}_i X_i \right) \right]^2$$

Test of Slope Coefficient Rejection Rule

- Reject H_0 in favor of H_a if t falls in colored area



- Reject H_0 for H_a if P-value = $P(T > |t|) < \alpha$

Special case: significance of coefficient

- Hypotheses
 - $H_0: \beta_i = 0$
 - $H_a: \beta_i \neq 0$

$$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}}$$

Example

Dependent Variable: TAX_ENT

Method: Least Squares

Date: 12/09/12 Time: 20:49

Sample: 2002Q1 2011Q4

Included observations: 40

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3973770.	754540.7	5.266475	0.0000
@TREND	290525.1	24239.34	11.98568	0.0000
@SEAS(1)	-3627516.	791034.8	-4.585786	0.0001
@SEAS(2)	-2975920.	789175.7	-3.770922	0.0006
@SEAS(3)	-2032456.	788058.1	-2.579068	0.0143
R-squared	0.837415	Mean dependent var		7480035.
Adjusted R-squared	0.818834	S.D. dependent var		4138083.
S.E. of regression	1761318.	Akaike info criterion		31.71749
Sum squared resid	1.09E+14	Schwarz criterion		31.92860
Log likelihood	-629.3498	Hannan-Quinn criter.		31.79382
F-statistic	45.06800	Durbin-Watson stat		1.123746
Prob(F-statistic)	0.000000			

Wald test

Null Hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Alternative hypothesis $H_1 : \beta_1$ or β_2 or β_3

or any two of them or all are nonzero.

At least one of them is significant.

In matrix notation

$$\text{Hypothesis: } Rb = r \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Test statistics with J numbers of restriction

$$F = \frac{\frac{(Rb - r)' [R \text{cov}(b) R']^{-1} (Rb - r)}{J}}{\frac{RSS}{n - k}}$$

Calculate F and compare it with the critical values $F(J, n-k)$ from the Table.

Test for multicollinearity

- High correlation between X variables
- Coefficients measure combined effect
- Leads to unstable coefficients depending on X variables in model
- Always exists; matter of degree
- *Example*: Using both total number of rooms and number of bedrooms as explanatory variables in same model

Detecting Multicollinearity

- Farrar-Glauber Multicollinearity
- VIF-test
- Few remedies
 - Obtain new sample data
 - Eliminate one correlated X variable
 - Standardize your independent variables.

Example

$$\hat{s}_t = 0.4 + 0.8y_t + 0.2li_t - 0.1si_t$$

(0.9) (1.2) (0.4) (0.1)

$\bar{R}^2 = 0.98$, (standard errors in parentheses)

(n = 60). where :

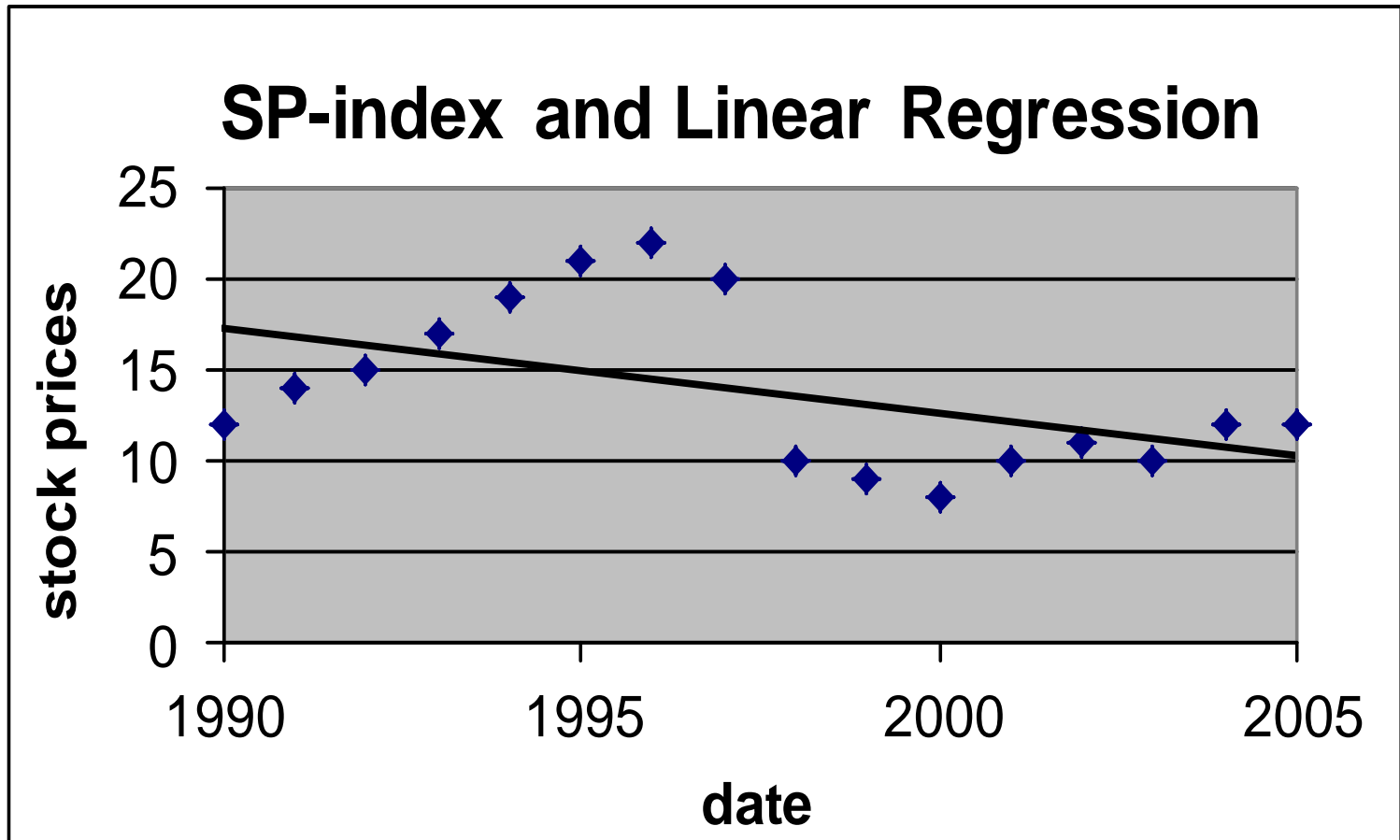
s_t – stock prices

y_t – output

li_t – long - run interest rates

si_t - short - run interest rates

Test for structural breaks



Chow Test

- Tests whether the coefficients in two linear regressions on different data sets are equal.

$$F = \frac{RSS_c - (RSS_1 + RSS_2) / k}{(RSS_1 + RSS_2) / n - 2k} \sim F_{k, n-2k}$$

RSS_c – combined _RSS

RSS_1 – pre – break _RSS

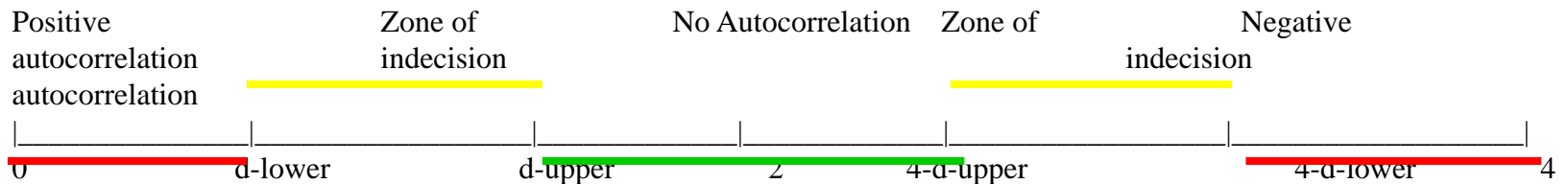
RSS_2 – post – break _RSS

Test for residuals autocorrelation

- **Durbin-Watson test** (only checks for first order serial correlation in residuals)
- **Breusch-Godfrey Test** (checks for higher order autocorrelation $AR(q)$ in residuals)

Durbin-Watson statistic

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}, \text{ for } n \text{ and } K - 1 \text{ d.f.}$$



- Autocorrelation is clearly evident
- Ambiguous – cannot rule out autocorrelation
- Autocorrelation is not evident

Breusch-Godfrey Test

Higher Order Autocorrelation model : AR(p)

$$\mu_t = \rho_1 \mu_{t-1} + \rho_2 \mu_{t-2} + \dots + \rho_p \mu_{t-p} + \varepsilon_t$$

Null Hypothesis

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_p = 0$$

Test Model :

$$\hat{\mu}_t = \delta_1 + \delta_2 X_{2t} + \dots + \delta_k X_{kt} + \lambda_1 \hat{\mu}_{t-1} + \dots + \lambda_p \hat{\mu}_{t-p} + \omega_t$$

Test Statistic

$$LM = (n - p) * R_{aux}^2 \sim \chi_p^2$$

Tests for Heteroskedasticity

- There are two types of tests:
 - Tests for continuous changes in variance:
White test, Breusch–Pagan tests, etc.
 - Tests for discrete (lumpy) changes in variance:
the Goldfeld–Quandt test

The White Test

- The White test for heteroskedasticity has a basic premise: *if disturbances are homoskedastic, then squared errors are on average roughly constant.*
- Explanators should **NOT** be able to predict squared errors, or their proxy, squared residuals.
- The White test is the most general test for heteroskedasticity.

Steps of the White Test

- Regress Y against your various explanators using OLS, compute the OLS residuals, $\varepsilon_1, \dots, \varepsilon_n$
- Regress ε_i^2 against a constant, all of the explanators, the squares of the explanators, and all possible interactions between the explanators (p slopes total)
- Compute R^2 from the “auxiliary equation” in step 2
- Compare nR^2 to the critical value from the Chi-squared distribution with p degrees of freedom.

The Breusch–Pagan Test – 1

- The Breusch–Pagan test is very similar to the White test.
- The White test specifies exactly which explanators to include in the auxiliary equation. Because the test includes cross-terms, the number of slopes (p) increases very quickly.
- In the Breusch–Pagan test the econometrician selects which explanators to include. Otherwise, the tests are the same.

The Breusch–Pagan Test – 2

- In the Breusch–Pagan test, the econometrician selects **m** explanators to include in the auxiliary equation.
- Which explanators to include is a **judgment call**.
- A **good** judgment call leads to a more powerful test than the White test.
- A **poor** judgment call leads to a poor test.

The Goldfeld–Quandt Test – 1

- Both the *White* test and *the Breusch–Pagan* test focus on smoothly changing variances for the disturbances.
- *The Goldfeld–Quandt* test compares the variance of error terms across discrete subgroups.
- Under homoskedasticity, all subgroups should have the same estimated variances.

The Goldfeld-Quandt Test – 2

Divide the n observations into h groups, of sizes $n_1..n_h$

Choose two groups, say 1 and 2.

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ against } H_a : \sigma_1^2 \neq \sigma_2^2$$

Regress Y against the explanators for group 1.

Regress Y against the explanators for group 2.

Goldfeld-Quandt Test – 3

Relabel the groups as L and S , such that $\frac{RSS_L}{n_L - k} > \frac{RSS_S}{n_S - k}$

$$\text{Compute } F_{calc} = \frac{\frac{RSS_L}{n_L - k}}{\frac{RSS_S}{n_S - k}} > 1$$

Compare F_{calc} to the critical value for an F -statistic with $(n_L - k)$ and $(n_S - k)$ degrees of freedom.

Test for specification

$$F_{n-m-k+1}^k \sim \frac{\frac{R_1^2 - R_0^2}{k}}{\frac{1 - R_1^2}{n - m - k}}$$

Ramsey's RESET

- RESET relies on a trick similar to the special form of the White test
- Instead of adding functions of the x 's directly, we add and test functions of \hat{y}
- So, estimate $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \delta_1 \hat{y}_2 + \delta_2 \hat{y}_3 + \varepsilon$ and test

$H_0: \delta_1 = 0, \delta_2 = 0$ using $F \sim F_{2, n-k-3}$ or
 $LM \sim \chi^2(2)$.

Stationary process

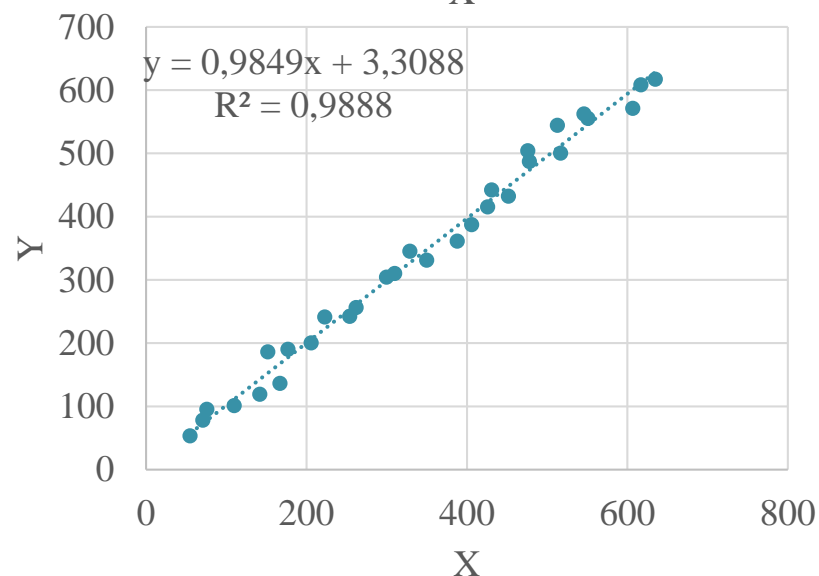
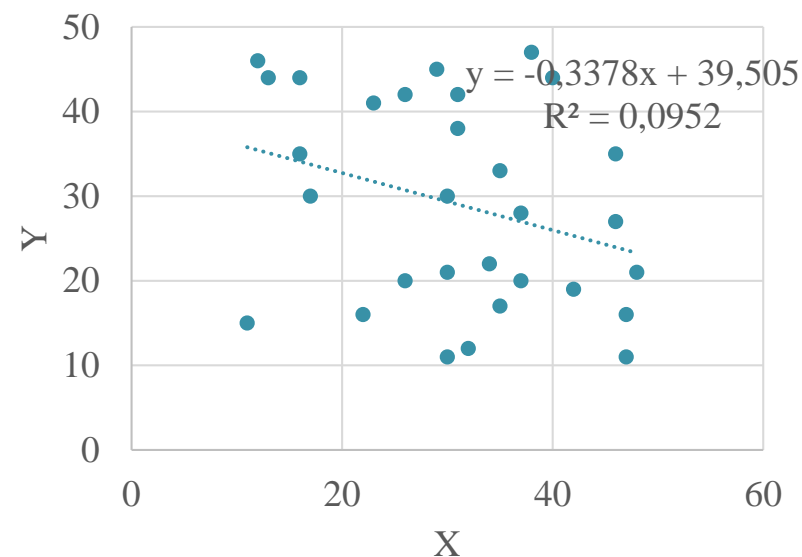
- A **stationary process** is a stochastic process whose joint probability distribution does not change when shifted in time.
- Parameters such as the mean and variance, if they are present, also do not change over time and do not follow any trends.

Solutions:

- Taking differences (Dickey-Fuller test)
- Trend-stationary processes

Why stationarity is important?

Nº	X	Y	T	X*	Y*
1	35	33	20	55	53
2	31	38	40	71	78
3	16	35	60	76	95
4	30	21	80	110	101
5	42	19	100	142	119
6	47	16	120	167	136
7	12	46	140	152	186
8	17	30	160	177	190
9	26	20	180	206	200
10	23	41	200	223	241
11	34	22	220	254	242
12	22	16	240	262	256
13	40	44	260	300	304
14	30	30	280	310	310
15	29	45	300	329	345
16	30	11	320	350	331
...					
30	35	17	600	635	617



Question

- What should we do, if we fail to find an appropriate model that satisfy all tests?



REVIEW

Linear regression

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t} + \dots + \beta_{k-1} x_{k-1t} + \varepsilon_t, t = \overline{1, n}$$

y_t - dependent variable;

$x_{1t}, x_{2t}, \dots, x_{k-1t}$ independent variables;

ε_t - residuals.

Assumptions

- **Linearity** - the Y variable is linearly related to the value of the X variable.
- **Independence of Error** - the error (residual) is independent for each value of X.
- **Homoscedasticity** - the variation around the line of regression be constant for all values of X.
- **Normality** - the values of Y be normally distributed at each value of X.

Regression Modeling Steps

- Define problem or question
- Specify model
- Collect data
- Do descriptive data analysis
- Estimate unknown parameters
- Evaluate model
- Use model for prediction

Multiple Regression Tests

- Test residual for normality
- Test parameter significance
 - Overall model
 - Individual coefficients
- Test for multicollinearity
- Test for model stability
- Test for residuals autocorrelation
- Test for residuals homoscedasticity
- Test for specification
- Test for stationary process



QUESTIONS?



**THANK YOU FOR
YOUR ATTENTION!**