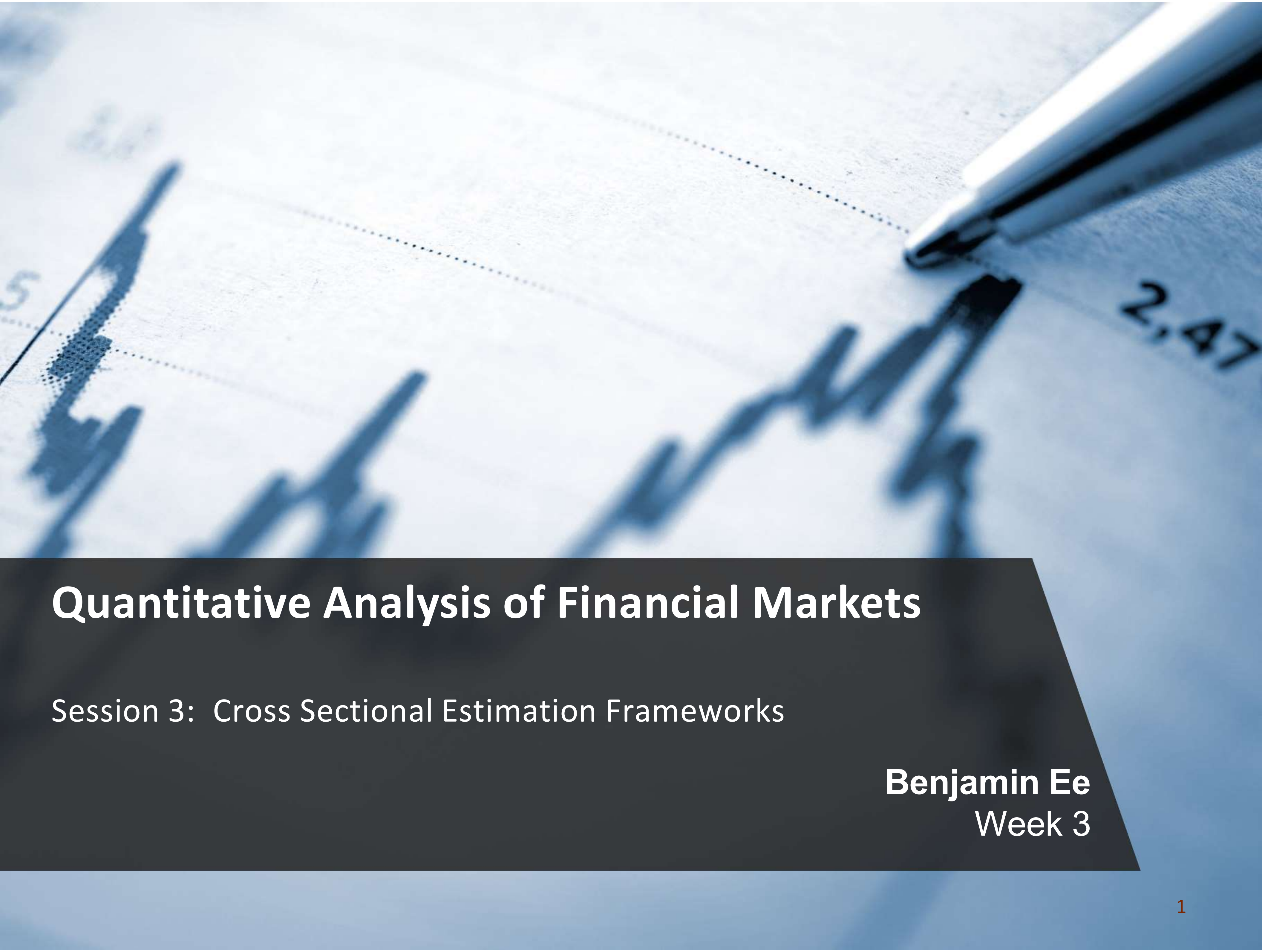


Quantitative Analysis of Financial Markets

Session 3: Cross Sectional Estimation Frameworks

Benjamin Ee
Week 3

Recap of pre-class prep

We **overviewed various distributions**, which will be helpful in testing hypotheses related to non-normal test statistics

Examples (some of these may be new)

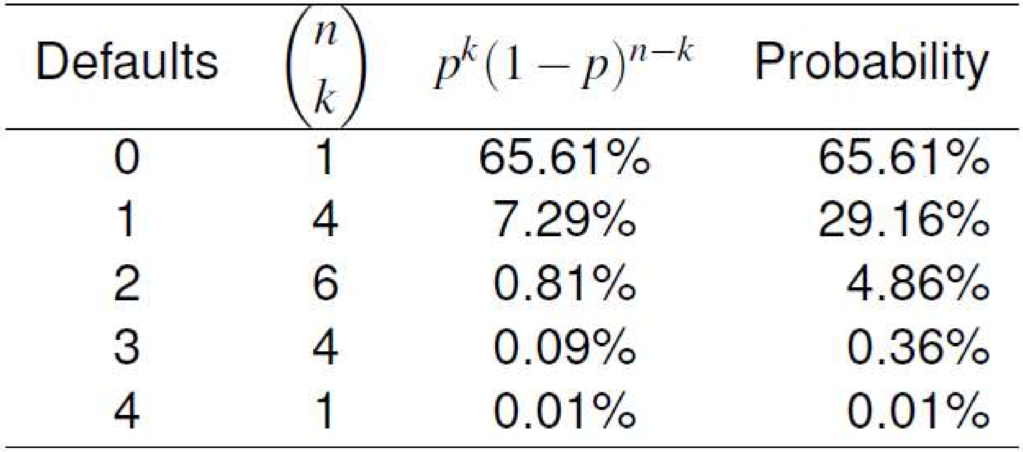
- Variance estimates are **chi-square distributed**
- Ratio of 2 chi-squared variables **is F-distributed**
- Test statistics on ‘regression coefficients’ used in modelling are **t-distributed**

As we go into regression modelling with OLS, we will **make further use of some of the above distributions**

Example 1 [Binomial]: Bond defaults

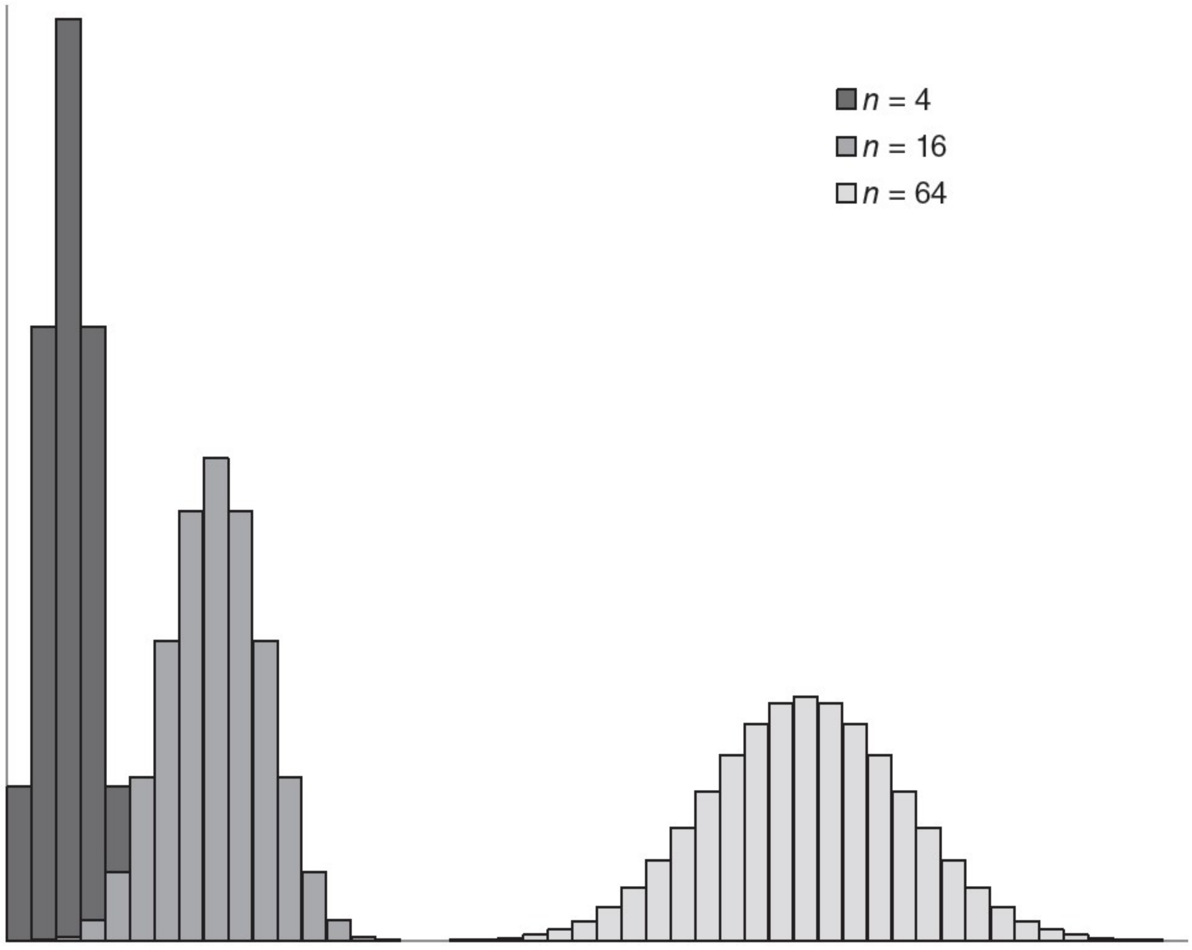
- Suppose you have four bonds, each with a 10% probability of defaulting over the next year. The event of default for any given bond is independent of the other bonds defaulting. What is the probability that zero, one, two, three, or all of the bonds default? What is the mean number of defaults? The standard deviation?

- Answer: We calculate the probability of each possible outcome as follows:



Defaults	$\binom{n}{k}$	$p^k(1-p)^{n-k}$	Probability
0	1	65.61%	65.61%
1	4	7.29%	29.16%
2	6	0.81%	4.86%
3	4	0.09%	0.36%
4	1	0.01%	0.01%

Binomial Probability Mass Function



Example 2: Goodness of Fit test [Chi-sq]

- In a goodness of fit test, we want to test if our sample is consistent with an assumed distribution (including its parameter values)

- For example, we want to test:
  - A coin is fair (binomial distribution with p = 0.5)
  - Expected returns are 0 (normal distribution with u = 0)

- Test statistic (O_i is observed while E_i is expected). We will compare the resulting value with the **chi-square distribution** (k-c degrees of freedom where c is number of parameters and k is number of independent trials) to determine if there is a significant difference

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Example 2: Goodness of Fit test (“Chi-square test”)

- Recall:

$$\sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \frac{(X_i - \mu_i)^2}{\mu_i} \overset{d}{\sim} \chi^2_{k-1}.$$

- Example: There are 1500 returns and 55 returns are smaller than the fifth percentile. What is the chi-square statistic?
- Solution:
  1. There are only two bins: either below or above the fifth percentile.
  2. Fifth percentile means that there is a 5% chance of falling into the “below” bin.
  3. So $\mu_1 = 1500 \times 0.05 = 75$, and $\mu_2 = 1500 \times 0.95 = 1425$

$$\frac{(55-75)^2}{75} + \frac{((1500-55)-1425)^2}{1425} = 5.61.$$

Chi-Square Test: General Case

Issues	NYSE	Nasdaq	NYSE MKT	Total
Advances	2,164	1,794	228	4,186
Declines	987	997	150	2,134
Unchanged	102	69	27	198
Total	3,253	2,860	405	6,518

Source: Wall Street Journal

1 Set up hypotheses and determine level of significance.

H_0 : stock movement and exchange listing are independent.

H_1 : H_0 is false. $\alpha = 0.05$

2 Compute the expected frequency

$$\text{Expected Frequency}_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{\text{Grand Total}}.$$

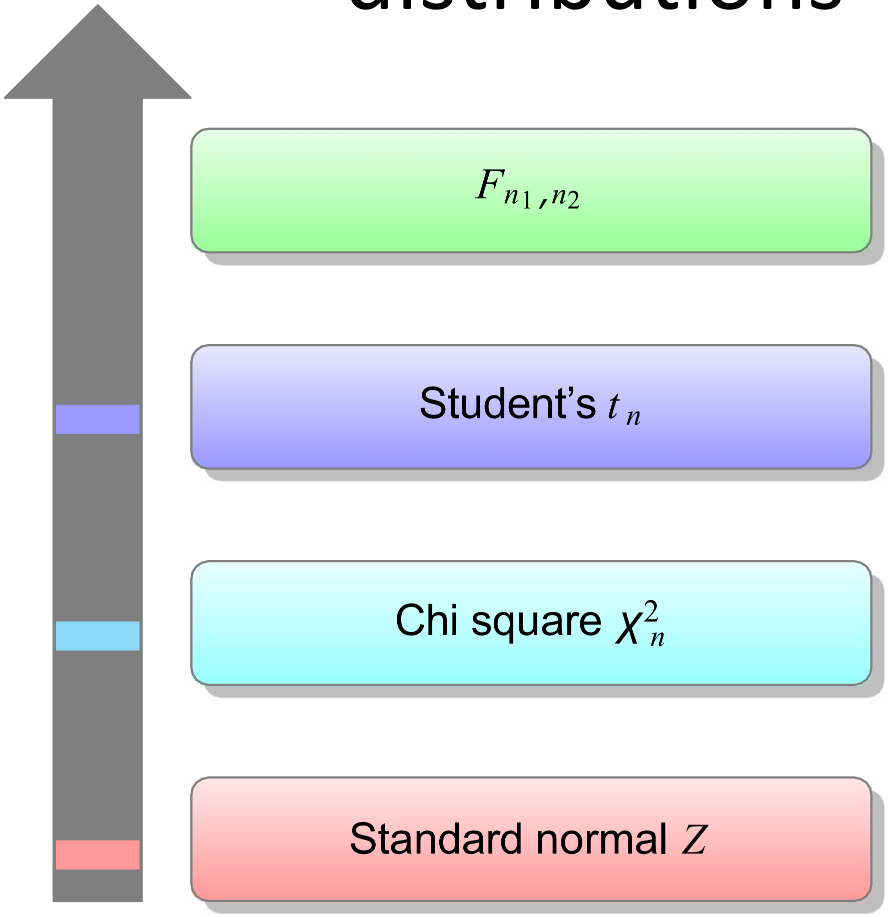
Real-Life Example: Results

Issues	NYSE	Nasdaq	NYSE MKT	Total
Advances	2,089.1	1,836.8	260.1	4,186.0
Declines	1,065.0	936.4	132.6	2,134.0
Unchanged	98.8	86.9	12.3	198.0
Total	3,253.0	2,860.0	405.0	6,518.0

$$\begin{aligned}\chi^2 =& \frac{(2164-2089.1)^2}{2089.1}+\frac{(1794-1836.8)^2}{1836.8}+\frac{(228-260.1)^2}{260.1}\\ &+\frac{(987-1065.0)^2}{1065.0}+\frac{(997-936.4)^2}{936.4}+\frac{(150-132.6)^2}{132.6}\\ &+\frac{(102-98.8)^2}{98.8}+\frac{(69-86.9)^2}{86.9}+\frac{(27-12.3)^2}{12.3}=40.91.\end{aligned}$$

Degrees of Freedom = (rows -1) × (columns-1) = 4.

Recap of distributions



F_{n_1,n_2}

Student's t_n

Chi square χ^2_n

Standard normal Z

$$F_{n_1,n_2} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$$

$$\lim_{n\to\infty} t_n \longrightarrow Z$$

$$V := \sum_{i=1}^{n} Z_i^2 \overset{d}{\sim} \chi^2_n$$

$$Z = \frac{r-\mu}{\sigma}$$

Overview of Ordinary Least Squares (OLS) and checking of prerequisite conditions

Ordinary Least Squares (OLS) Motivation

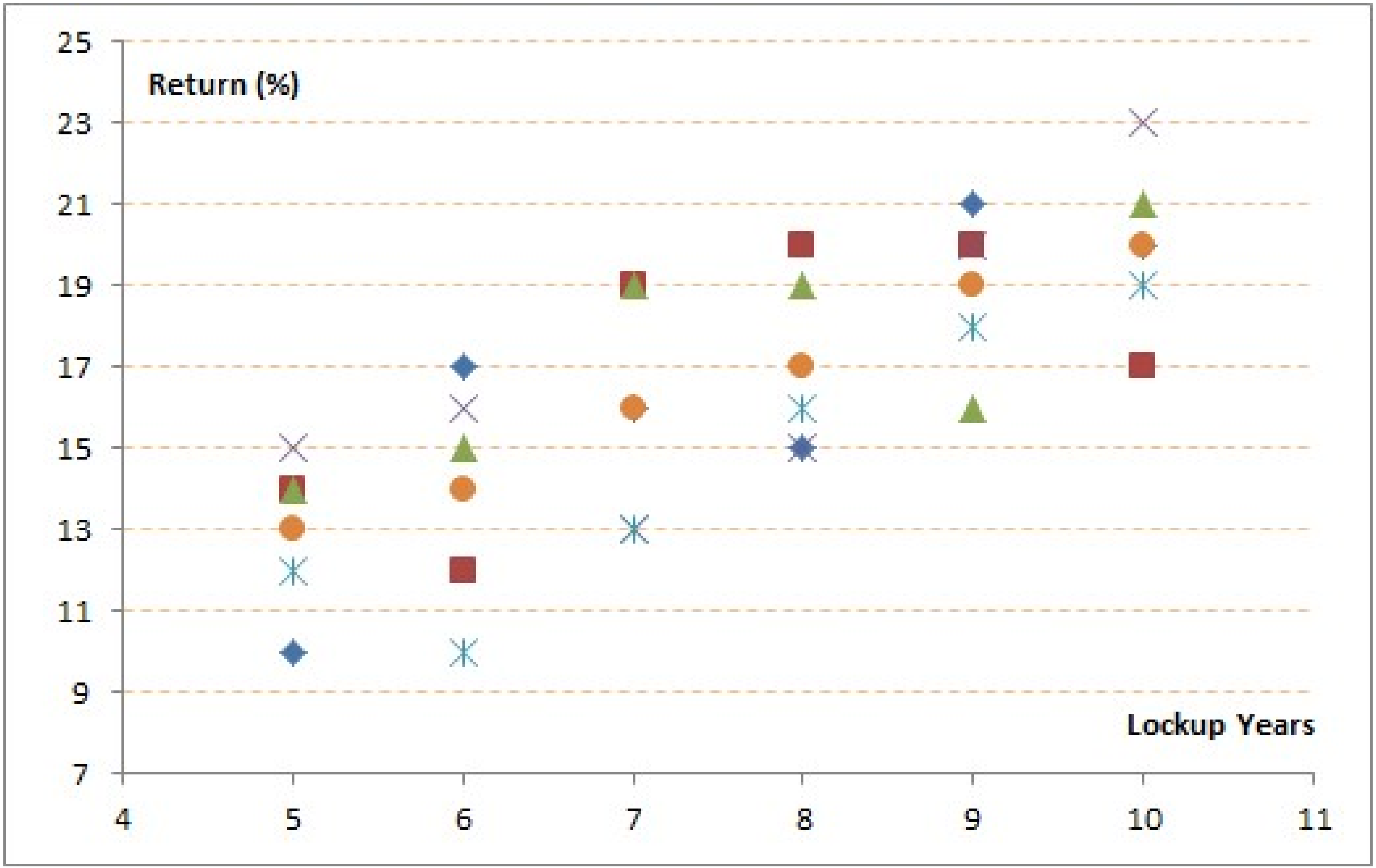
- So far we look at one time series or one set of data X.
- What about two sets of data X and Y?

Example: Annual Returns of 30 Hedge Funds

The population consists of 30 hedge funds that follow the same strategy, but of different length of the lockup period (minimum number of years an investor must keep funds invested).

Lockup (years)	Return (% per year)					Average Return
5	10	14	14	15	12	13
6	17	12	15	16	10	14
7	16	19	19	13	13	16
8	15	20	19	15	16	17
9	21	20	16	20	18	19
10	20	17	21	23	19	20

Scatter Plot



- The scatter plot indicates that there is a positive relationship between the hedge fund returns and the lockup period.

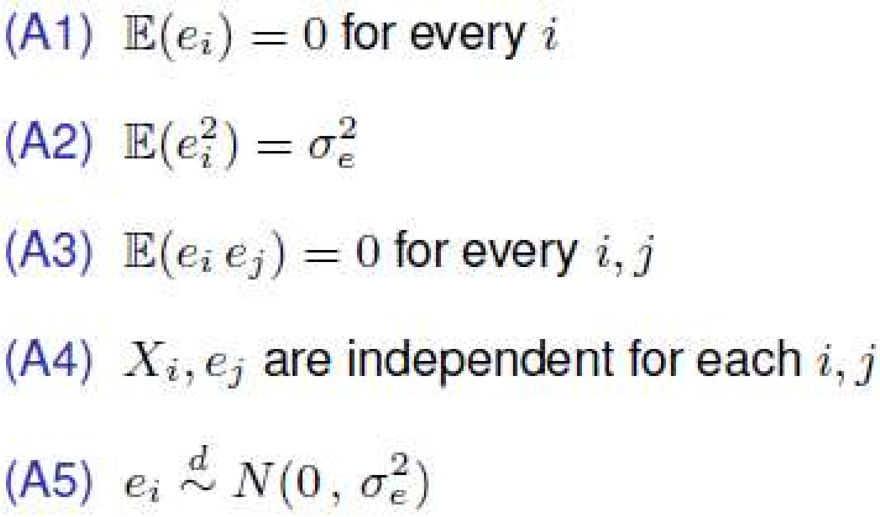
OLS Assumptions

- Given n pairs of observations on explanatory variable X_i and dependent variable Y_i, we can have **a linear model** postulating that

$$Y_i = a + bX_i + e_i\,, \qquad i = 1, 2, \ldots, n,$$

where e_i is the noise.

- Assumptions:



(A1) $\mathbb{E}(e_i) = 0$ for every i

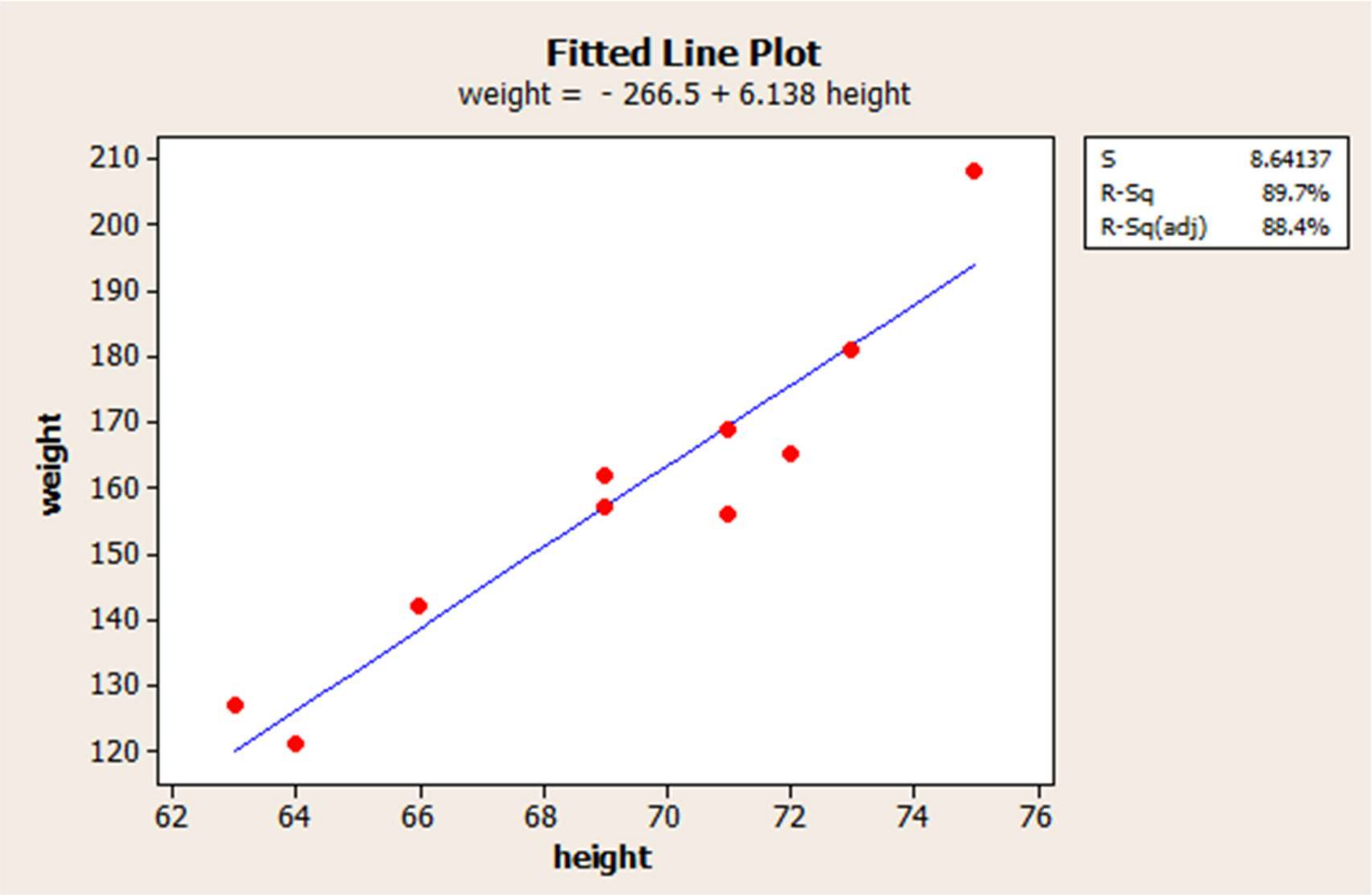
(A2) $\mathbb{E}(e_i^2) = \sigma_e^2$

(A3) $\mathbb{E}(e_i\, e_j) = 0$ for every i, j

(A4) X_i, e_j are independent for each i, j

(A5) $e_i \overset{d}{\sim} N(0\,,\, \sigma_e^2)$

With these assumptions, **we are able to generate a linear equation for the variable of interest in terms of predictors**

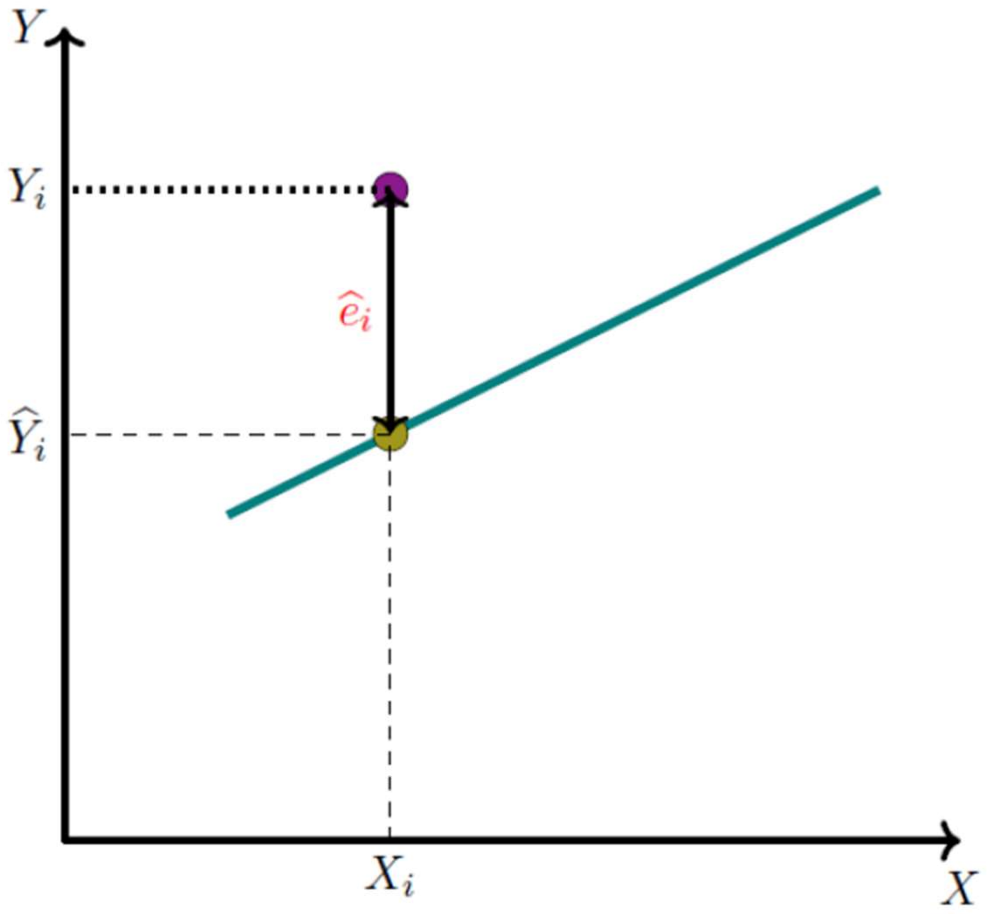


Source: https://online.stat.psu.edu/stat415/book/export/html/880

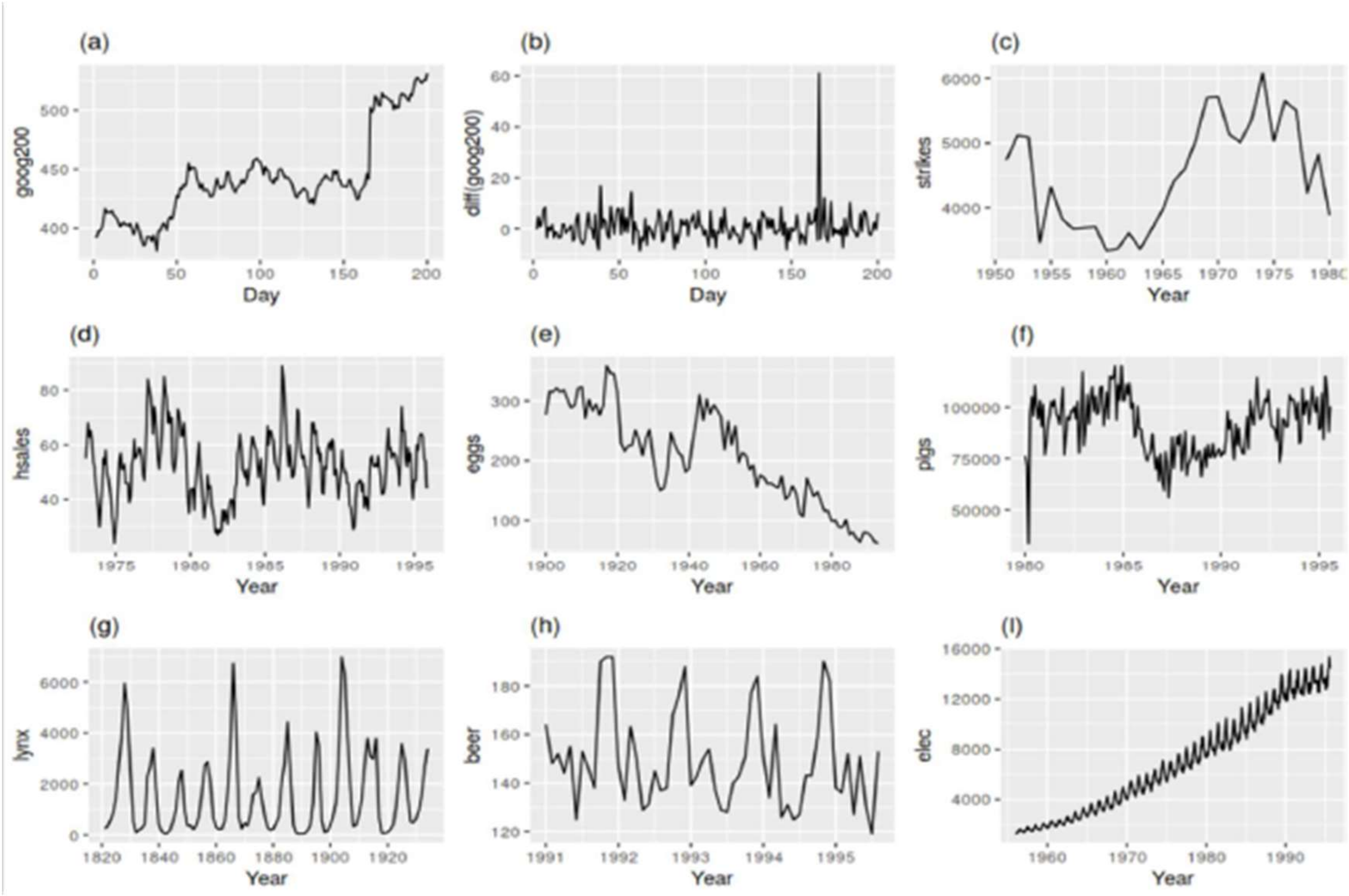
OLS Condition A1 means average error is 0

- A1 means that **the y-intercept is correctly estimated**
- If y-intercept were incorrectly estimated such that E(e_i) != 0, **we adjust y-intercept by the difference to render the error to be 0**

Residual is the Vertical Length



OLS Condition A2 means the volatility of data has no predictable patterns w.r.t. the x variable (homoskedasticity)



Which one of these graphs do you think show a predictable pattern of volatility w.r.t the x variable?

OLS Condition A2: Implications of constant variance assumption

- Constant variance means **OLS assigns an equal importance to all data points in the optimization**

- A2 is not necessarily always satisfied for financial data because there is **volatility clustering**

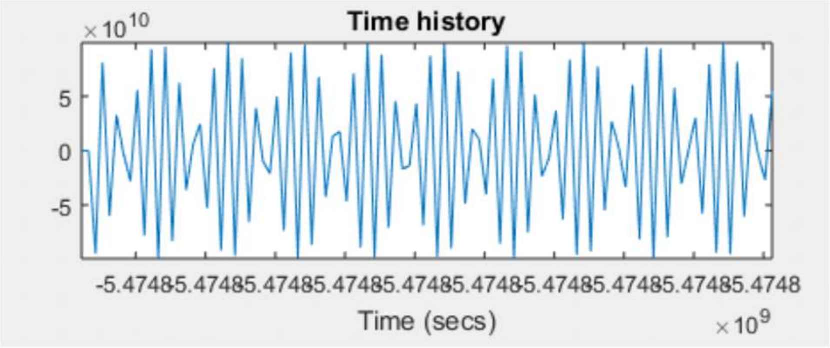
- We can account for predictable volatility using **“weighted least squares”** later on, which is a more general version of ordinary least squares

A3 and A4 both mean that there is no time series content in the data

- A3 means there is no predictable relationship between the error from 1 time period and the error from another time period. If indeed we can use the error from today to “forecast” what the error will be tomorrow, then **we should put that into the model**

- A4 means there is no predictable **relationship between the error from 1 time period and value of the predictor variable**. Otherwise same as above

- If we built an OLS model for the time series history shown on the right, **do you think A3 and/or A4 will be satisfied?**



Source: https://www.researchgate.net/figure/a-First-order-perturbed-wave-autocorrelation-b-Output-SHG-wave-autocorrelation-using_fig9_323434355

When to use OLS:

- Overall, **when do we use OLS**:
  - When A3 and A4 are not met**, we should use time series model**s (which we will cover later in the term)
  - When A2 is not met, we should consider whether to use **Weighted Least Squares (WLS)**
    - Note that estimation of a volatility model may itself **not be easy and introduce noise**, which is why we do not just use WLS "all the time"
- When **A2, A3 and A4 are met we can use OLS**

Agenda

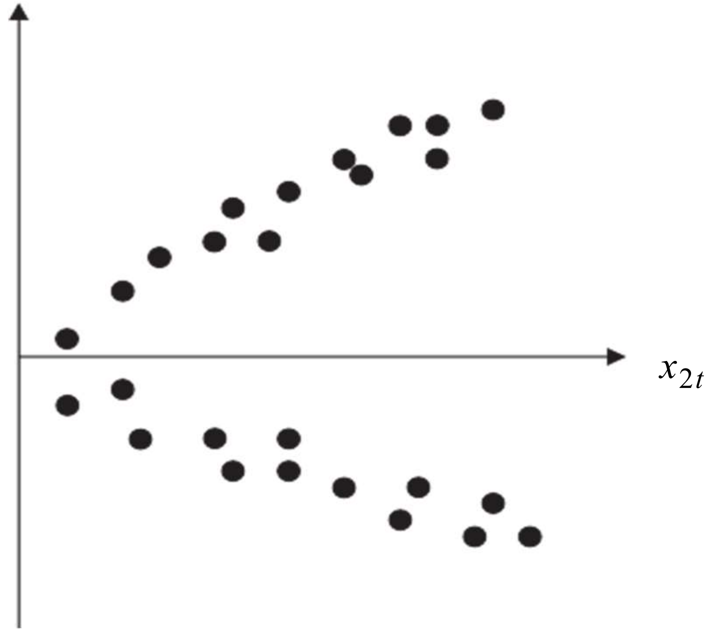
- **Today**: **Statistical tests to decide if OLS conditions are met**
  - These ‘statistical tests’ are **formally hypotheses tests**
  - Many of them make use of distributions we learnt before, **such as F and Chi-squared distributions**

- **Next week**:
  - **Derivation of OLS estimators** [pre-class prep]
  - **Application of OLS estimators**
  - OLS case study: **CAPM**

A2: Homoscedasticity versus Heteroscedasticity

- We have so far assumed that the variance σ^2 of the errors is constant

- How do we test this?



The Goldfeld-Quandt (GQ) Test

- Split the total sample of length T into two sub-samples oflength T_1 and T_2. The regression model is estimated on each sub-sample and the two residual variances are calculated.
- The null hypothesis is that the variances of the disturbances are equal, $H_0 : \sigma^2{}_1 = \sigma^2{}_2$
- The test statistic, denoted GQ, is simply the ratio of the two residual variances where the larger of the two variances must be placed in the numerator.

$$\text{GQ} = \frac{s_1^2}{s_2^2} \sim F(T_1 - K, T_2 - K)$$

- A problem with the test is that the choice of where to split the sample is usually arbitrary and may crucially affect the outcome of the test.

White’s Test

- White’s general test for heteroscedasticity is one of the best approaches because it makes few assumptions about the form of the heteroscedasticity.
- Suppose the regression we have carried out is as follows

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + u_t$$

- We want to test $\mathrm{V}(u_t) = \sigma^2$. We estimate the model, obtaining the residuals, $\hat{u}_t$.
- Then run the auxiliary regression

$$\hat{u}_t^2 = \alpha_1 + \alpha_2 x_{2,t} + \alpha_3 x_{3,t} + \alpha_4 x_{2,t}^2 + \alpha_5 x_{3,t}^2 + \alpha_6 x_{2,t} x_{3,t} + v_t.$$

White's Test and χ^2 Statistic

- Obtain R^2 from the auxiliary regression and multiply it by the number of observations, T.

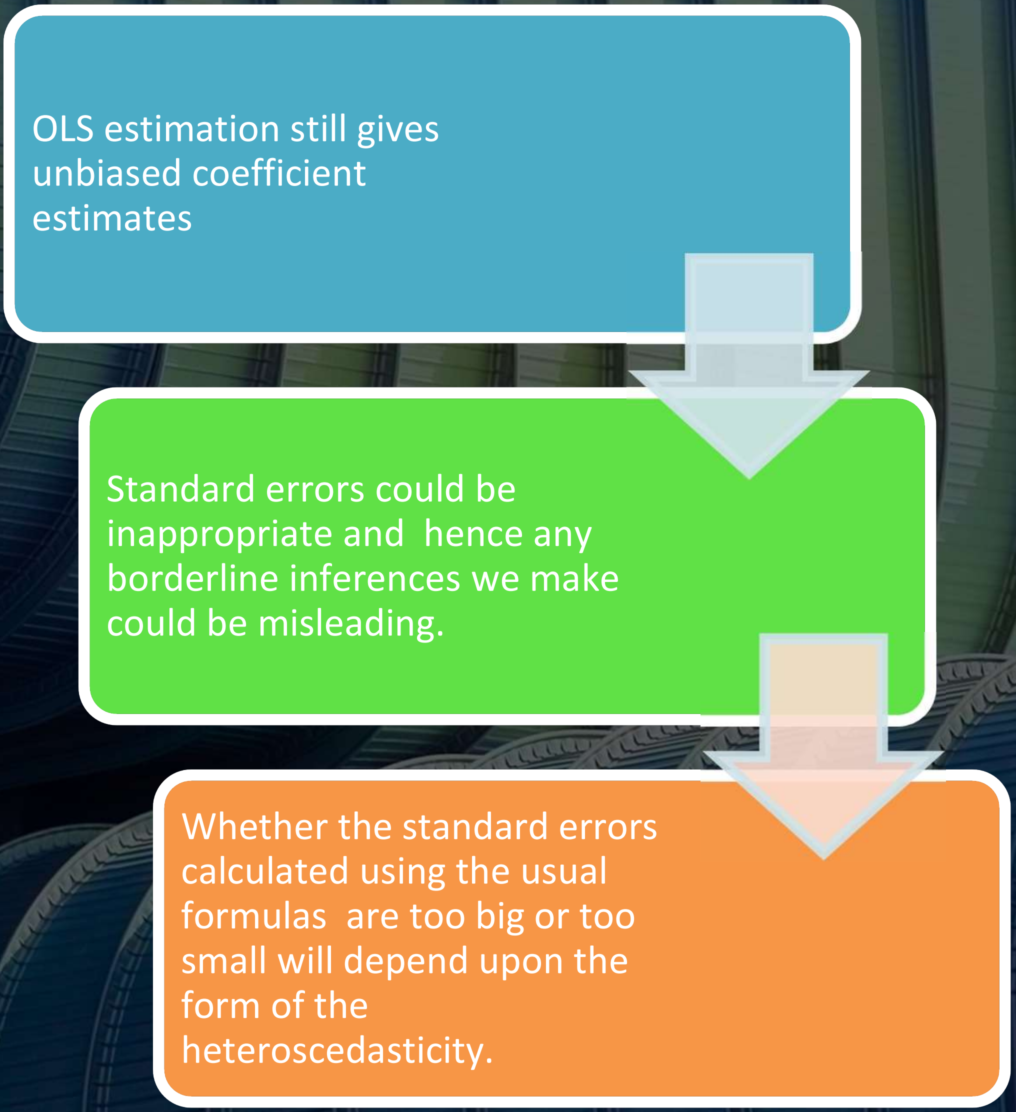
- White shows that

$$T \times R^2 \sim x^2(m).$$

where m is the number of regressors in the auxiliary regression excluding the constant term.

- If the x^2 test statistic is greater than the corresponding value from the statistical table then reject the null hypothesis that the disturbances are homoscedastic.

Consequences of Heteroscedasticity



OLS estimation still gives unbiased coefficient estimates

Standard errors could be inappropriate and hence any borderline inferences we make could be misleading.

Whether the standard errors calculated using the usual formulas are too big or too small will depend upon the form of the heteroscedasticity.

Mitigation measures

- Transform the variables into logs or reducing by some other measure of “size”.

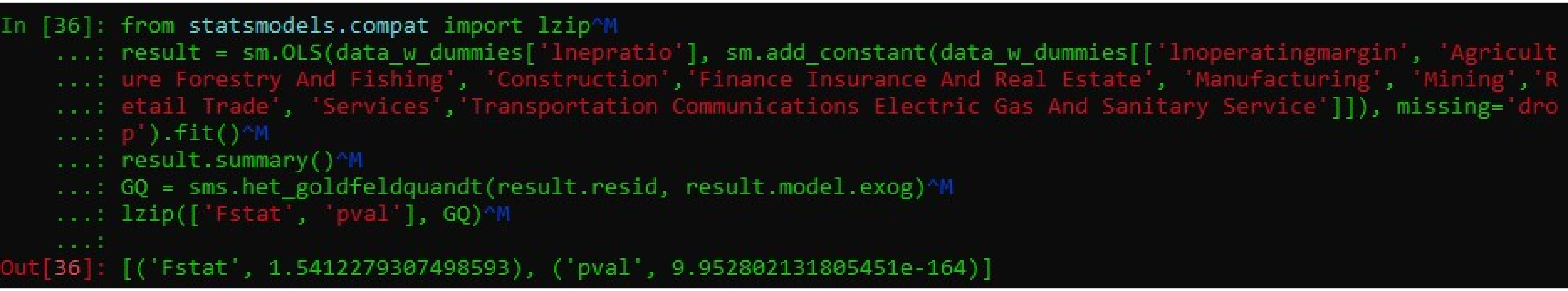
- Use White’s heteroscedasticity consistent standard error estimates.

- The effect of using White’s correction is that in general the standard errors for the slope coefficients are increased relative to the usual OLS standard errors.

- The goal is to be “conservative” in hypothesis testing, so that we would need more evidence against the null hypothesis before we could reject it.

Goldfeld-Quandt Test in python

(code and data uploaded to eLearn as ‘ols_tests.py’– see line 174)



```
In [36]: from statsmodels.compat import lzip^M
    ...: result = sm.OLS(data_w_dummies['lnepratio'], sm.add_constant(data_w_dummies[['lnoperatingmargin', 'Agricult
    ...: ure Forestry And Fishing', 'Construction','Finance Insurance And Real Estate', 'Manufacturing', 'Mining','R
    ...: etail Trade', 'Services','Transportation Communications Electric Gas And Sanitary Service']]), missing='dro
    ...: p').fit()^M
    ...: result.summary()^M
    ...: GQ = sms.het_goldfeldquandt(result.resid, result.model.exog)^M
    ...: lzip(['Fstat', 'pval'], GQ)^M
    ...:
Out[36]: [('Fstat', 1.5412279307498593), ('pval', 9.952802131805451e-164)]
```

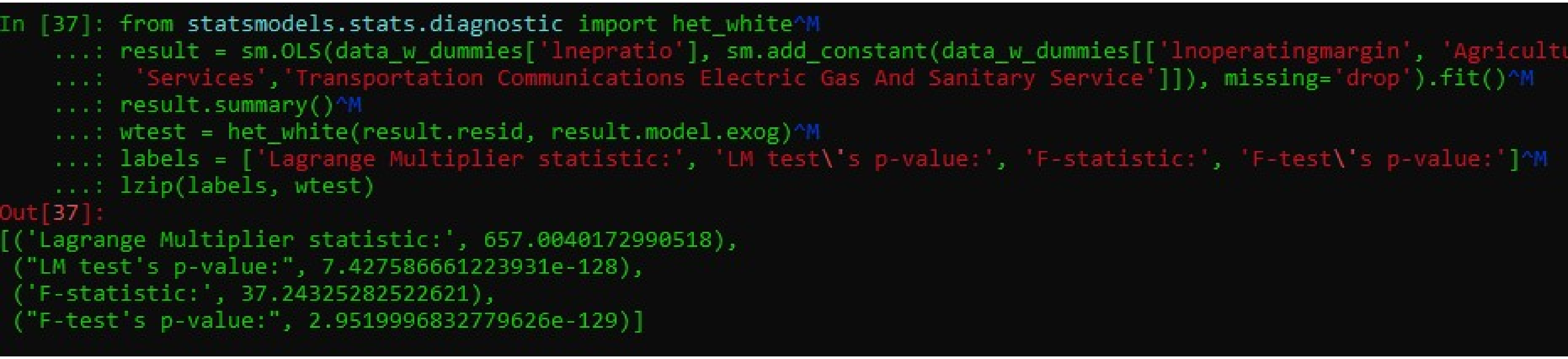
Based on p-value, we reject hypothesis of constant variance in dataset

By default, python just splits the sample evenly at the midpoint index

We can also specify a split index by using “split=X” parameter

Note: unzip data.zip and run ols_tests.py in same directory. The code also executes other tests, please ignore those for now

White’s test in python [line 181]



```
In [37]: from statsmodels.stats.diagnostic import het_white^M
    ...: result = sm.OLS(data_w_dummies['lnepratio'], sm.add_constant(data_w_dummies[['lnoperatingmargin', 'Agricultu
    ...:  'Services','Transportation Communications Electric Gas And Sanitary Service']]), missing='drop').fit()^M
    ...: result.summary()^M
    ...: wtest = het_white(result.resid, result.model.exog)^M
    ...: labels = ['Lagrange Multiplier statistic:', 'LM test\'s p-value:', 'F-statistic:', 'F-test\'s p-value:']^M
    ...: lzip(labels, wtest)
Out[37]:
[('Lagrange Multiplier statistic:', 657.0040172990518),
 ("LM test's p-value:", 7.427586661223931e-128),
 ('F-statistic:', 37.24325282522621),
 ("F-test's p-value:", 2.9519996832779626e-129)]
```

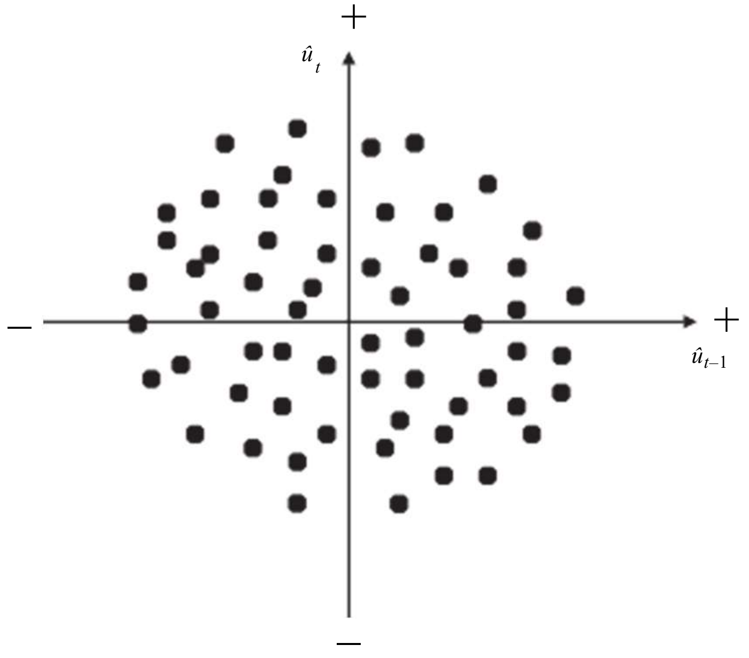
Based on p-value, we reject hypothesis of constant variance in dataset

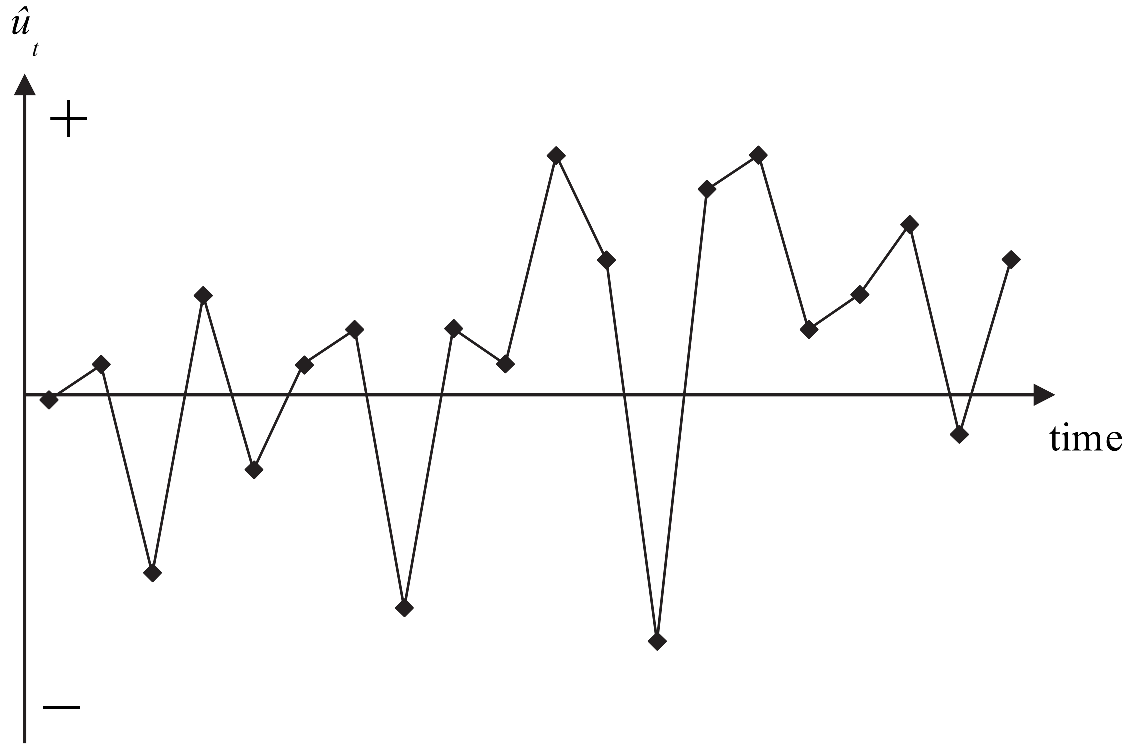
Results from White’s test are not dependent on arbitrary definitions of breakpoints

TESTING FOR SERIAL CORRELATON (A3 AND A4):

- DURBIN-WATSON TEST
- BREUSCH-GODFREY TEST

No Pattern in OLS Residuals





- Ideal Case: No pattern in residuals at all.

Motivation

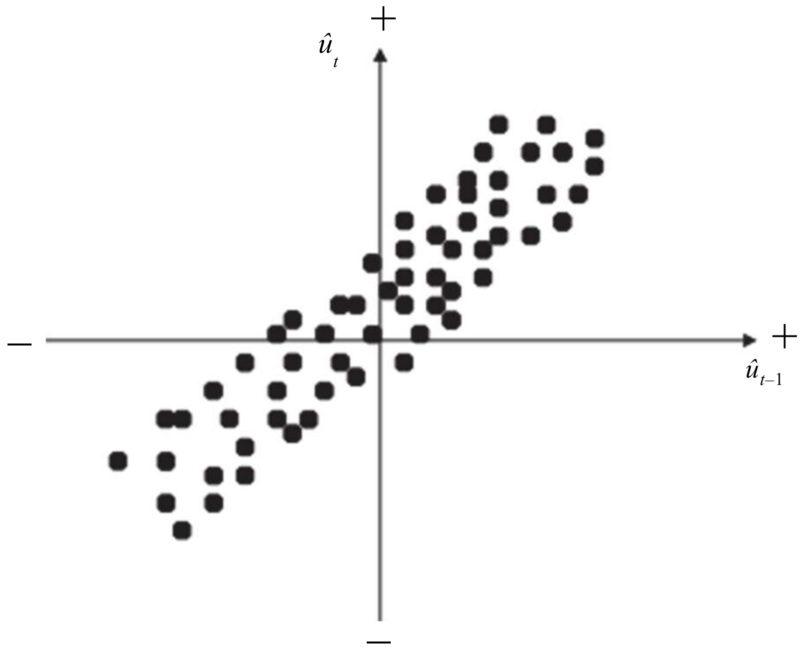
- The residuals $u_i := \widehat{e}_i$from an OLS regression may have a correlation structure:

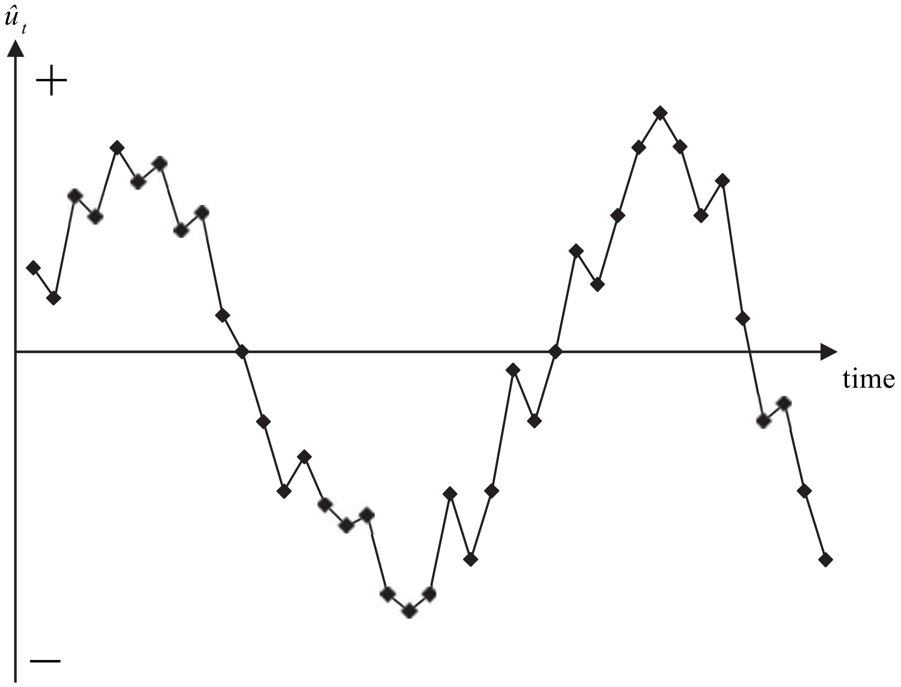
$$u_i = \phi\, u_{i-1} + \varepsilon_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

where ε_i is the noise.

- Ifϕ is 0, then we say that the residual u_i has no correlation structure. Otherwise, if $\varphi$is statistically non-zero, then there is evidence of serial correlation, and **it will make OLS estimators inefficient**.

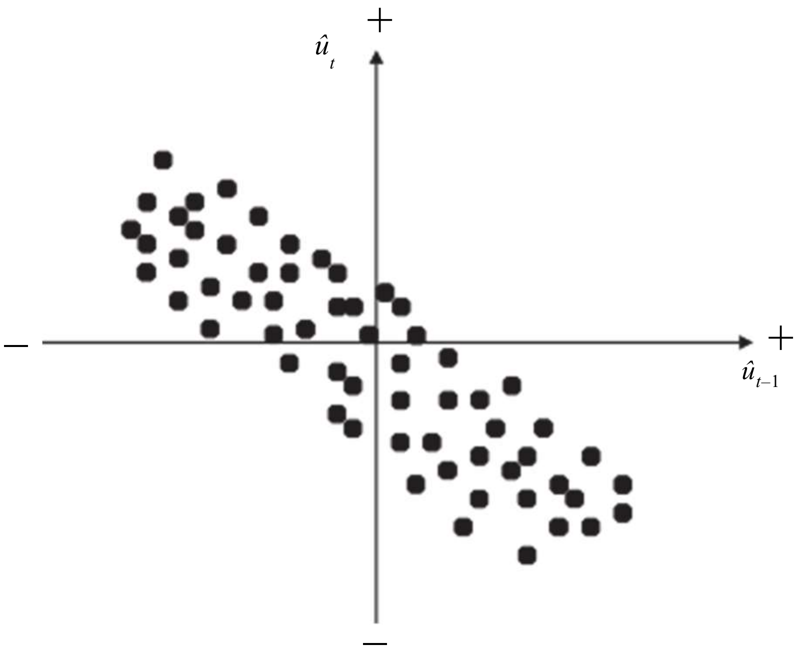
Positive Autocorrelation in OLS Residuals

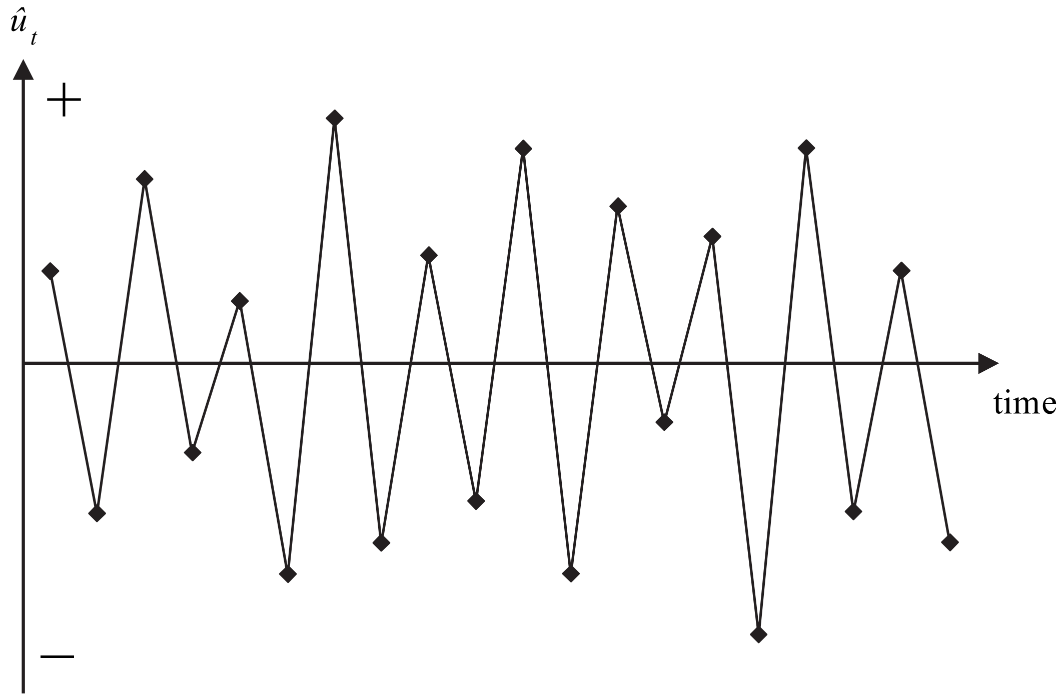




Positive autocorrelation is indicated by a cyclical residual plot over time.

Negative Autocorrelation in Residuals





Negative autocorrelation is indicated by an alternating pattern where the residuals cross the time axis more frequently than if they were distributed randomly.

Assumptions and Test for Serial Correlation

Assumptions

- Suppose $\mathbb{E}\left(\varepsilon_t\right) = 0$, $\mathbb{V}\left(\varepsilon_t\right) = \sigma_\varepsilon^2$, and $\mathbb{C}\left(\varepsilon_t, \varepsilon_s\right) = 0$ if $t \neq s$.
- Suppose $|\phi| < 1.$

The Durbin-Watson Test

- $H_0 : \phi = 0$ versus $H_A : \phi \neq 0$
- The Durbin-Watson test statistic for the residual u_t:

$$\mathrm{DW} = \frac{\sum_{t=2}^{T}\left(u_t - u_{t-1}\right)^2}{\sum_{t=1}^{T} u_t^2} \tag{2}$$

Breaking Down Durbin-Watson Test

- Expansion of numerator

$$\text{DW} = \frac{\sum_{t=2}^{T} u_t^2}{\sum_{t=1}^{T} u_t^2} + \frac{\sum_{t=2}^{T} u_{t-1}^2}{\sum_{t=1}^{T} u_t^2} - 2\frac{\sum_{t=2}^{T} u_t u_{t-1}}{\sum_{t=1}^{T} u_t^2}$$

- The first term is approximately 1. So is the second term.
- The third term is an estimator of the correlation with itself!

$$\phi = \frac{\mathbb{C}\left(u_t, u_{t-1}\right)}{\mathbb{V}(u_t)},$$

i.e., since $\overline{u} = \frac{1}{T}\sum_{t=1}^{T} u_t,$

$$\widehat{\phi} = \frac{\sum_{t=2}^{T} u_t u_{t-1}}{\sum_{t=1}^{T} u_t^2}.$$

The Durbin-Watson Test Hypotheses

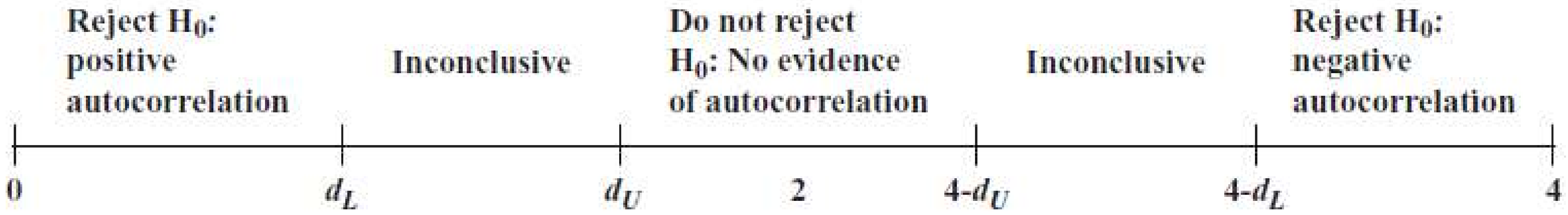
- The approximation for the Durbin-Watson statistic therefore is

$$\text{DW} \simeq 1 + 1 - 2\widehat{\phi} = 2\left(1 - \widehat{\phi}\right)$$

- Test with null hypothesis: $H_0\phi = 0$ versus $H_A : \phi > 0$
- If H_0 is true, DW '.: 2. Else if $\varphi > 0$, DW < 2
- Accept H_0 (no autocorrelation) when DW ≈ 2, and reject H_0 (positive serial correlation) when DW is significantly smaller than 2.
- Likewise, accept H_0 (no autocorrelation) when DW ≈ 2, and reject H_0 (negative serial correlation) when DW is significantly larger than 2.

Caveat of the Durbin-Watson Test and Inference

- Since $-1 \leq \varphi \leq 1$, $0 \leq$ DW ≤ 4.
- DW has 2 critical values, an upper critical value (d_U) and a lower critical value (d_L), and an intermediate region where you can neither reject nor not reject H_0.


Reject H_0: positive autocorrelation
Inconclusive
Do not reject H_0: No evidence of autocorrelation
Inconclusive
Reject H_0: negative autocorrelation
0
d_L
d_U
2
4-d_U
4-d_L
4

- Conditions that must be fulfilled for DW to be a valid test

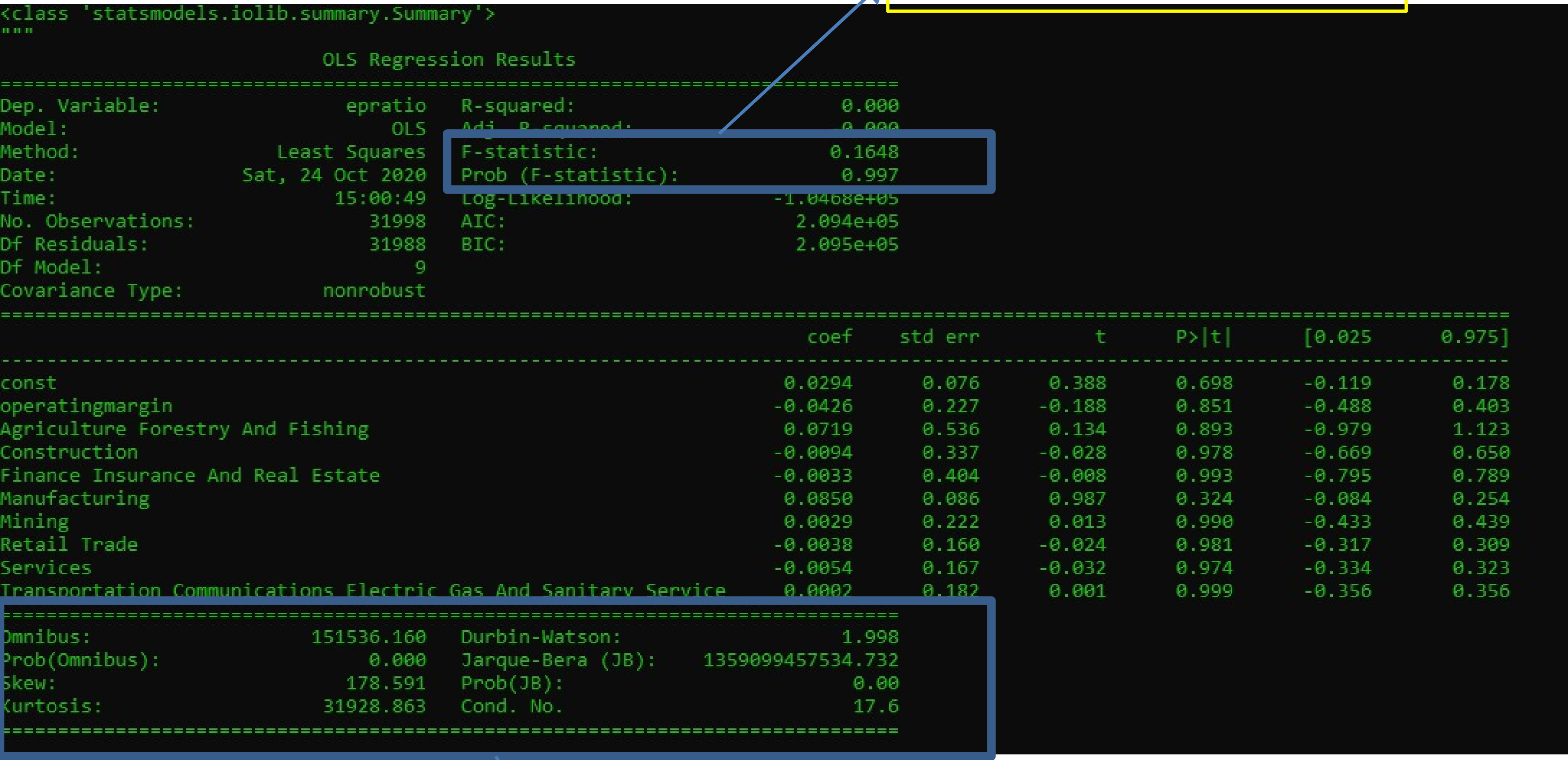
1. y-intercept must be in the regression.

2. Regressor X_i is non-stochastic.

3. The error e_i is normally distributed.

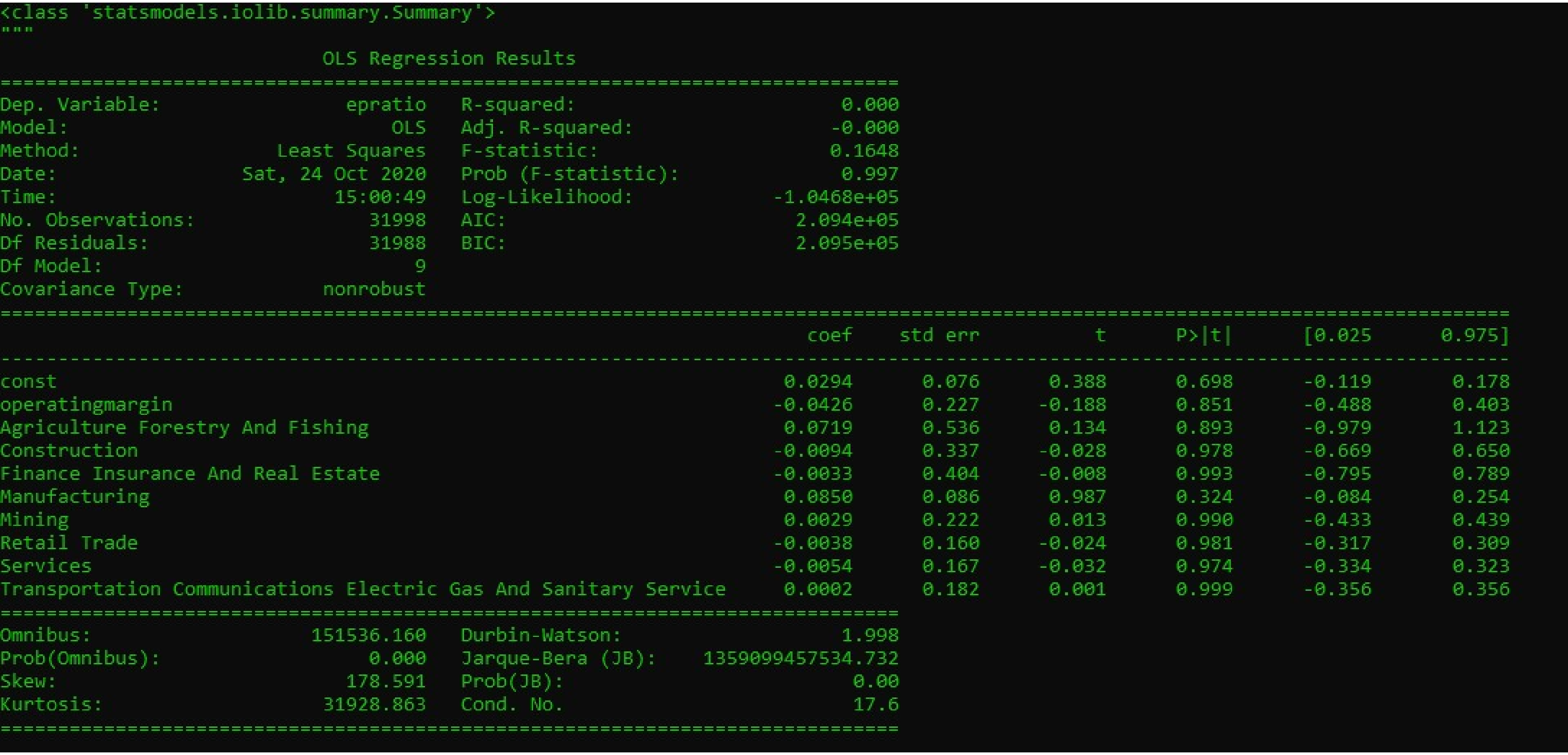
Durbin Watson test in python – line 158

Null is all coefficients are 0



```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                epratio   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                    0.1648
Date:                Sat, 24 Oct 2020   Prob (F-statistic):              0.997
Time:                        15:00:49   Log-Likelihood:            -1.0468e+05
No. Observations:               31998   AIC:                         2.094e+05
Df Residuals:                   31988   BIC:                         2.095e+05
Df Model:                           9
Covariance Type:            nonrobust
====================================================================================================================================
                                                                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------------------
const                                                                0.0294      0.076      0.388      0.698      -0.119       0.178
operatingmargin                                                     -0.0426      0.227     -0.188      0.851      -0.488       0.403
Agriculture Forestry And Fishing                                     0.0719      0.536      0.134      0.893      -0.979       1.123
Construction                                                        -0.0094      0.337     -0.028      0.978      -0.669       0.650
Finance Insurance And Real Estate                                   -0.0033      0.404     -0.008      0.993      -0.795       0.789
Manufacturing                                                        0.0850      0.086      0.987      0.324      -0.084       0.254
Mining                                                               0.0029      0.222      0.013      0.990      -0.433       0.439
Retail Trade                                                        -0.0038      0.160     -0.024      0.981      -0.317       0.309
Services                                                            -0.0054      0.167     -0.032      0.974      -0.334       0.323
Transportation Communications Electric Gas And Sanitary Service      0.0002      0.182      0.001      0.999      -0.356       0.356
==============================================================================
Omnibus:                   151536.160   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):    1359099457534.732
Skew:                         178.591   Prob(JB):                         0.00
Kurtosis:                   31928.863   Cond. No.                         17.6
==============================================================================
```

Python automatically runs a battery of 6 tests with every OLS estimation (lines 151 to 154 of code under header "#Durbin Watson and other diagnostics#)



```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                epratio   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.1648
Date:                Sat, 24 Oct 2020   Prob (F-statistic):              0.997
Time:                        15:00:49   Log-Likelihood:            -1.0468e+05
No. Observations:               31998   AIC:                         2.094e+05
Df Residuals:                   31988   BIC:                         2.095e+05
Df Model:                           9
Covariance Type:            nonrobust
===================================================================================================================================
                                                                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------------------------
const                                                               0.0294      0.076      0.388      0.698      -0.119       0.178
operatingmargin                                                    -0.0426      0.227     -0.188      0.851      -0.488       0.403
Agriculture Forestry And Fishing                                    0.0719      0.536      0.134      0.893      -0.979       1.123
Construction                                                       -0.0094      0.337     -0.028      0.978      -0.669       0.650
Finance Insurance And Real Estate                                  -0.0033      0.404     -0.008      0.993      -0.795       0.789
Manufacturing                                                       0.0850      0.086      0.987      0.324      -0.084       0.254
Mining                                                              0.0029      0.222      0.013      0.990      -0.433       0.439
Retail Trade                                                       -0.0038      0.160     -0.024      0.981      -0.317       0.309
Services                                                           -0.0054      0.167     -0.032      0.974      -0.334       0.323
Transportation Communications Electric Gas And Sanitary Service     0.0002      0.182      0.001      0.999      -0.356       0.356
==============================================================================
Omnibus:                   151536.160   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):    1359099457534.732
Skew:                         178.591   Prob(JB):                         0.00
Kurtosis:                   31928.863   Cond. No.                         17.6
==============================================================================
```

In this estimation:

1. The Durbin-Watson statistic appears almost completely typical (around 2), so serial correlation is not a concern

The Breusch-Godfrey Test Hypotheses

- It is a more general test for r^{th} order autocorrelation:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \cdots + \rho_r u_{t-r} + v_t,$$

$$v_t \sim N\left(0, \sigma_v^2\right)$$

- The null and alternative hypotheses are:

$$H_0 : \rho_1 = 0 \text{ and } \rho_2 = 0 \text{ and} \ldots \text{and } \rho_r = 0$$
$$H_1 : \rho_1 \neq 0 \text{ or } \rho_2 \neq 0 \text{ or} \ldots \text{or } \rho_r \neq 0$$

The Breusch-Godfrey Test Procedures

- The test is carried out as follows:

  1 Estimate the linear regression using OLS and obtain the residuals, $\widehat{e}_t$.

  2 Regress e on X plus $\widehat{e}_{t-1}, \widehat{e}_{t-2}, \ldots, \widehat{e}_{t-r}$;

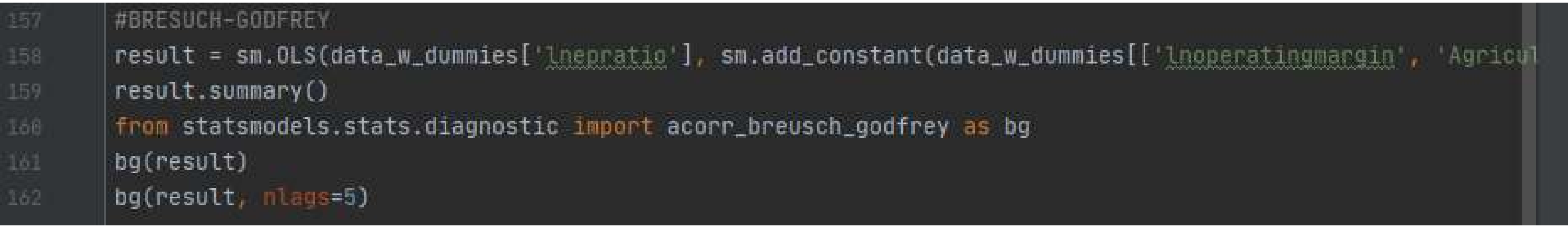
  3 Obtain R^2 from this regression.

  4 It can be shown that

$$(T - r)R^2 \sim \chi^2_r$$

- If the test statistic exceeds the critical value, reject the null hypothesis of no autocorrelation.

Breusch Godfrey test in python – line 166



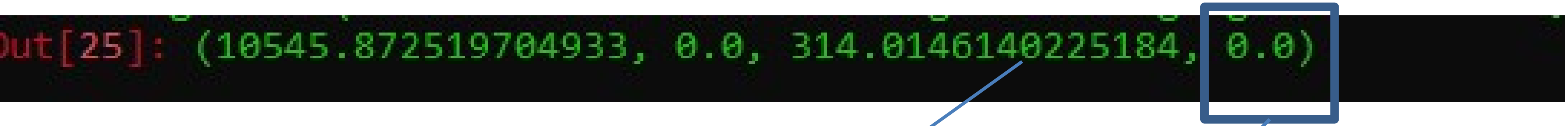
```
157    #BRESUCH-GODFREY
158    result = sm.OLS(data_w_dummies['lnepratio'], sm.add_constant(data_w_dummies[['lnoperatingmargin', 'Agricul
159    result.summary()
160    from statsmodels.stats.diagnostic import acorr_breusch_godfrey as bg
161    bg(result)
162    bg(result, nlags=5)
```

Lines 166 / 167: Run a sample estimation for us to do the Breusch Godfrey (BG) test on

Line 169: Do BG test with default number of lags (which is min(10, nobs/5)

Line 171: BG test with user specified number of lags

Recall that BG tests for serial correlation at an arbitrary number of lags, which is different from Durbin Watson which is confined to just a single lag



```
Out[25]: (10545.872519704933, 0.0, 314.0146140225184, 0.0)
```

F statistic

P value of test statistic is second item returned, and indicates in this case that we reject null hypothesis of no serial correlation

Consequences of Ignoring Autocorrelation

- The coefficient estimates derived using OLS are **still unbiased**
-
- Standard error estimates are inappropriate, there is a possibility that we could make the wrong inferences.
- R^2 is likely to be inflated relative to its “correct” value for positively correlated residuals.

Next week

OLS estimators, applications and model specification tests