Predictive Analysis for Decision Making

Week 1 and 2
Extending Linear Regression

Dr Issam Malki
School of Finance and Accounting
Westminster Business School

January 2024

Classical Linear Regression
RECAP

Linear Regression

- The multiple linear regression model can be expressed as:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + u_i$$

i: subscript refers to the units and takes values from 1 to n.

x_{ij}: explanatory variables where j=1, 2, …, k with x_{i1}=1.

β_j: unknown parameters to be estimated.

u_i: random error component. Unobserved with variance σ^2.

We aim to estimate β_j and σ^2.

Linear Regression

- The multiple linear in a matrix notation:

$$Y = X\beta + \mathbf{u}$$

where Y is $(n \times 1)$, X is $(n \times k)$, $\mathbf{u}$ is $(n \times 1)$ and β is $(k \times 1)$.

- The OLS estimator is define as follows:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Why? See notes.

LINEAR REGRESSION

- Assumptions

  - X is fixed (non stochastic) with rank k (full column rank, meaning?)

  - $\mathbf{u}$ is random vector with $\boldsymbol{E(u)=0}$ and $\mathbf{var}(\boldsymbol{u}) = \boldsymbol{E(u'u)} = \boldsymbol{\sigma}^{\mathbf{2}}\mathbf{I}$. Or:

$$var[u_i] = var[u] = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & \vdots \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

LINEAR REGRESSION

- Assumptions – Implications

  - Assumption 1:
    - Explanatory variables are strictly exogenous (i.e. $E(X\mathbf{u})=0$).
    - The X's are linearly independent (i.e. no multicollinearity).

  - Assumption 2:
    - The fitted line is indeed in the middle.
    - The errors are homoscedastic (no heteroskedasticity).
    - The errors are serially uncorrelated (no autocorrelation)

  - The validity of these assumptions is a must for the OLS to be a reliable estimator.

Linear Regression

- Key results I: unbiasedness

$$E(\hat{\beta}) = \beta$$

See full proof in the notes

- Key results II:

$$\boldsymbol{var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

where σ^2 is estimated using the sum of squared residuals, $e'e = \sum_{i=1}^{n} e_i^2$, as follows:

$$\widehat{\sigma^2} = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$$

- Key results III: Gauss-Markov Theorem:

  OLS estimator is the Best Linear Unbiased Estimator (BLUE)

LINEAR REGRESSION IN PYTHON

- Data: nls80.xls. See problem set for details on data and variables definitions.
- You need the following libraries for the basic linear regression
  - Pandas
  - Numpy
  - Statsmodels.api
- For today, we need one more: matplotlib.pyplot
- Use the file: computer_seminar 1
- The model we are about to estimate is:

$$\ln(wage_i) = \beta_1 + \beta_2 educ_{i2} + \beta_3 hours_{i3} + u_i$$

Linear Regression in Python

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                logwage   R-squared:                       0.103
Model:                            OLS   Adj. R-squared:                  0.101
Method:                 Least Squares   F-statistic:                     53.62
Date:                Wed, 03 Feb 2021   Prob (F-statistic):           9.12e-23
Time:                        10:54:46   Log-Likelihood:                -466.72
No. Observations:                 935   AIC:                             939.4
Df Residuals:                     932   BIC:                             954.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      6.1504      0.109     56.534      0.000       5.937       6.364
educ           0.0612      0.006     10.243      0.000       0.049       0.073
hours         -0.0044      0.002     -2.448      0.015      -0.008      -0.001
==============================================================================
Omnibus:                       28.032   Durbin-Watson:                   1.765
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               34.519
Skew:                          -0.340   Prob(JB):                     3.19e-08
Kurtosis:                       3.651   Cond. No.                         388.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Classical Linear Regression

Extension I: Relaxing Assumptions About the Error Term

Assumption Violations: General View

- We will now study these assumptions further, and in particular look at:
  - How we test for violations
  - Causes
  - Consequences
- In general we could encounter any combination of 3 problems:
  - the coefficient estimates are wrong
  - the associated standard errors are wrong
  - the distribution that we assumed for the test statistics will be inappropriate
- Solutions
  - The assumptions are no longer violated
  - we work around the problem so that we
  - use alternative techniques which are still valid

Assumption Violations: $E(\varepsilon_t) = 0$

- Assumption that the mean of the disturbances is zero.

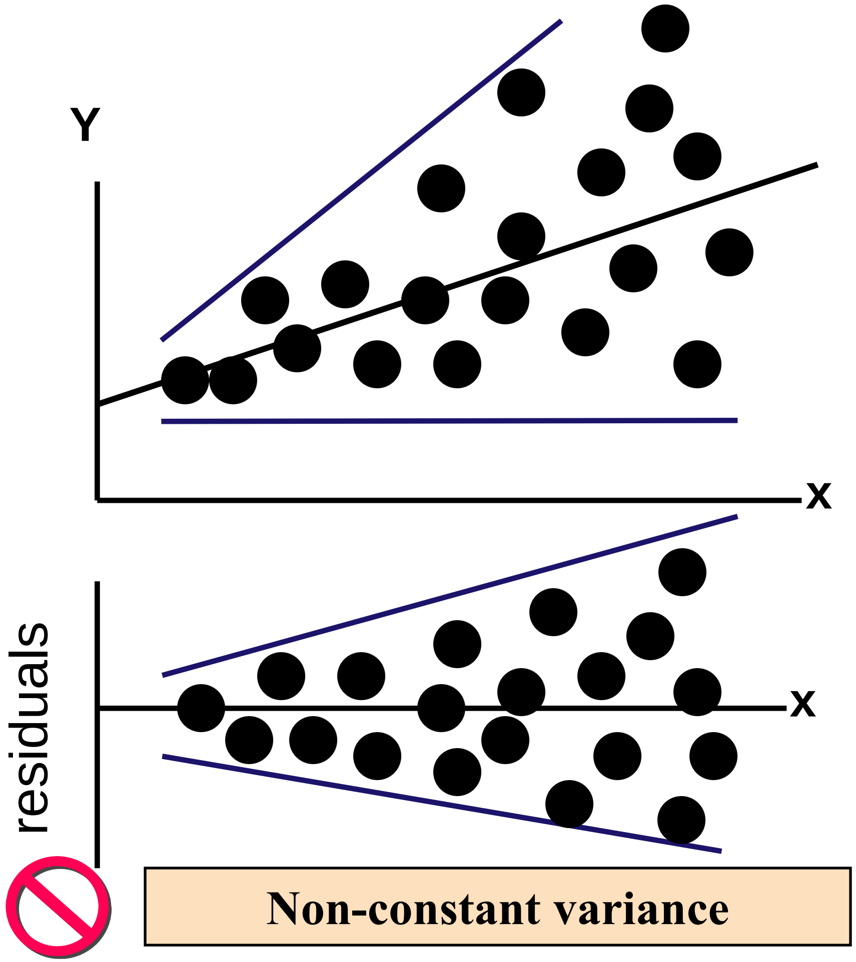
- For all diagnostic tests, we cannot observe the disturbances and so perform the tests of the residuals.

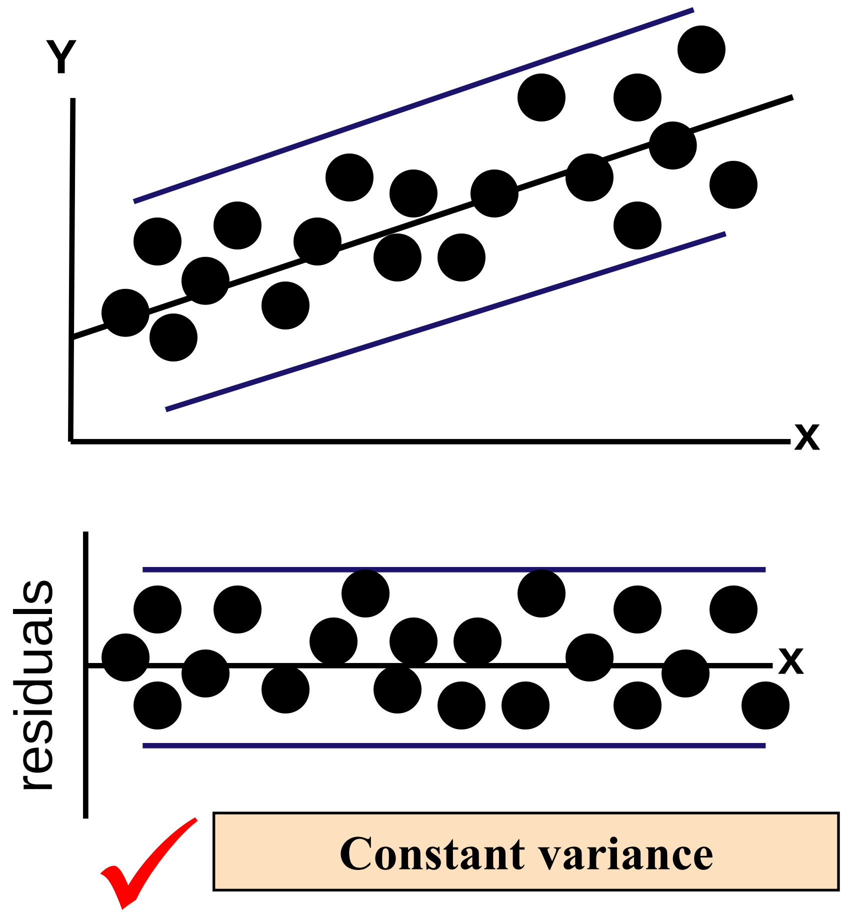
- The mean of the residuals will always be zero provided that there is a constant term in the regression.

ASSUMPTION VIOLATIONS: Var(u_i) is not constant

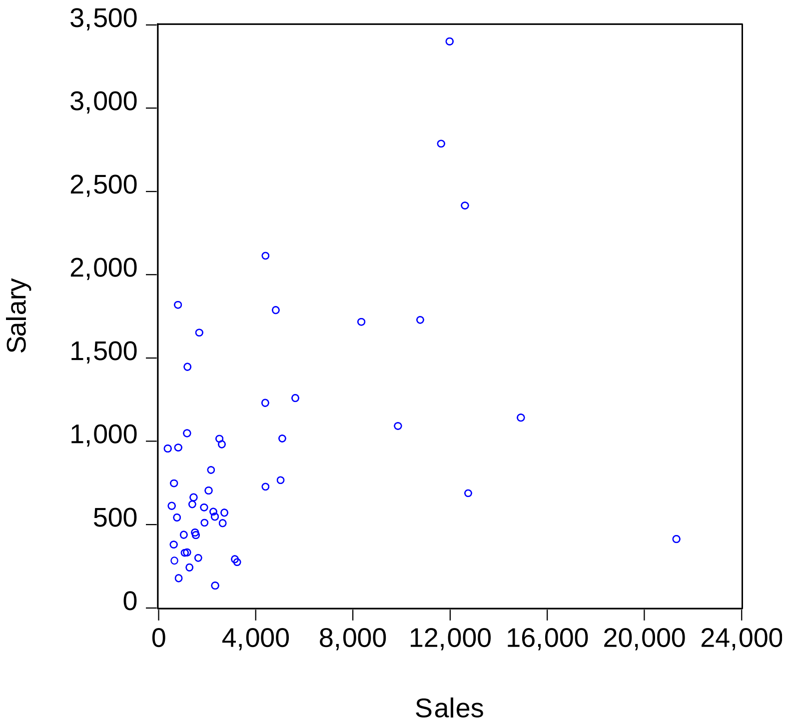
- Assumption that the variance of the error terms is finite and constant.
- Implications of this property
  - Homoscedasticity
  - Deviation from the fitted line is – on average- constant
  - All observed individuals have the same errors on average
- Violation of this assumptions leads to
  - Heteroscedasticity (spread of the variance)
  - Affect the standard errors not the estimated coefficients.
  - Ignoring this produces incorrectly smaller errors and higher t-statistics
  - OLS is no longer efficient and thus no longer BLUE.

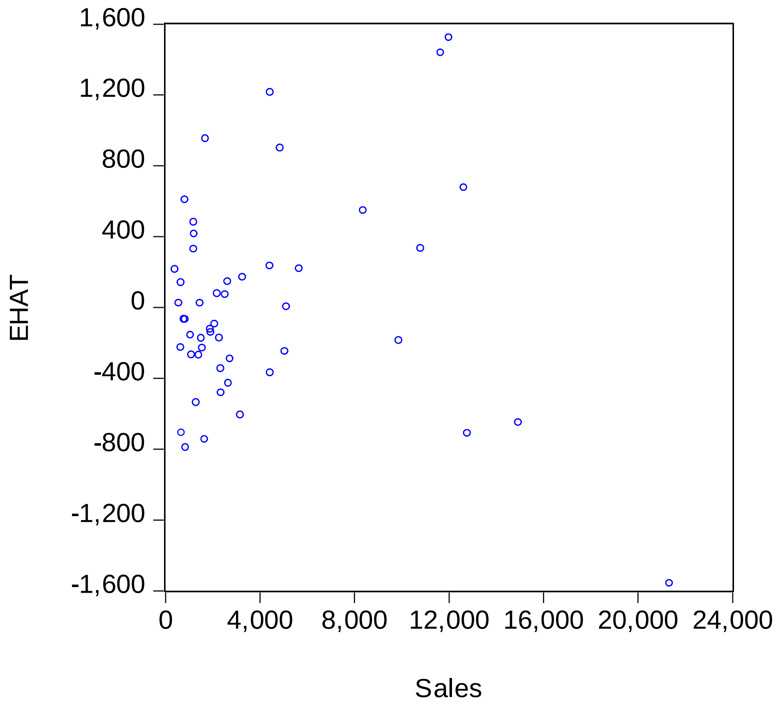
Assumption Violations: Var(u_i) is not constant



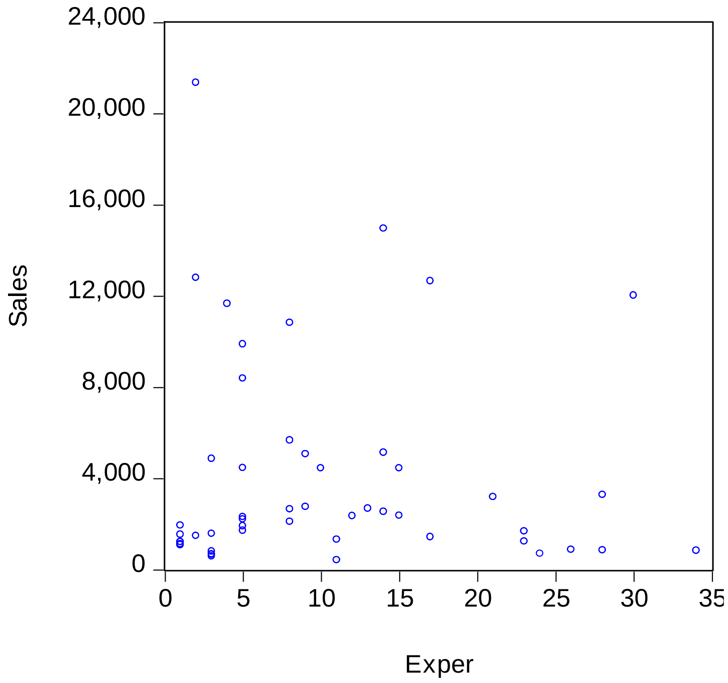


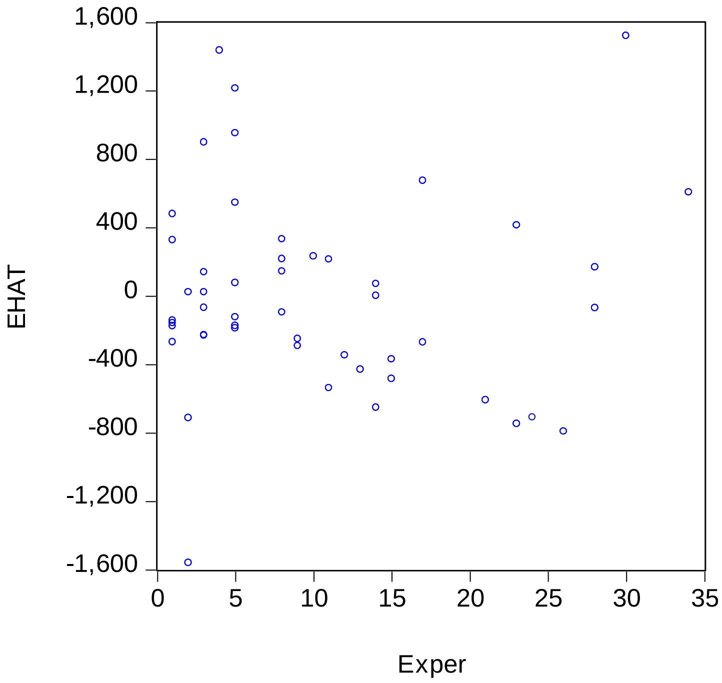
EXAMPLES: GRAPHICAL INSPECTION





Examples: Graphical Inspection





ASSUMPTION VIOLATIONS: $\mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty$

White's Test for Heteroscedasticity

- White's general test for heteroscedasticity is one of the best approaches because it makes few assumptions about the form of the heteroscedasticity.
- The hypotheses

$$H_0\text{: Residuals are homoscedastic}$$

$$H_1\text{: Residuals are heteroskedastic}$$

- The test is carried out as follows:

1. Assume that the regression we carried out is as follows

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

And we want to test $\mathrm{Var}(\varepsilon_t) = \sigma^2$. We estimate the model, obtaining the residuals, $\hat{\varepsilon}_t$

2. Then run the auxiliary regression

$$\hat{\varepsilon}_t^2 = \alpha_0 + \alpha_1 x_{2t} + \alpha_2 x_{3t} + \alpha_3 x_{2t}^2 + \alpha_4 x_{3t}^2 + \alpha_4 x_{2t} x_{3t} + v_t$$

ASSUMPTION VIOLATIONS: $\text{Var}(\varepsilon_t) = \sigma^2 < \infty$

White's Test for Heteroscedasticity

3. Obtain R^2 from the auxiliary regression and multiply it by the number of observations, T. It can be shown that

$$T\,R^2 \sim \chi^2\,(m)$$

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

4. If the χ^2 test statistic from step 3 is greater than the corresponding value from the statistical table then reject the null hypothesis that the disturbances are homoscedastic.

ASSUMPTION VIOLATIONS: $\text{Var}(\varepsilon_t) = \sigma^2 < \infty$

White's Test for Heteroscedasticity

- If we confirm the presence of heteroscedasticity:
  - Estimates are still unbiased, but not efficient
  - This implies that he standard errors and therefore the t statistics are incorrect
  - OLS estimator is not BLUE.
- Use White's Heteroscedasticity Consistent Standard Errors

ASSUMPTION VIOLATIONS: Var(ε_t) = $\sigma^2 < \infty$

Table A.4 χ^2 Distribution: critical values of χ^2 at 5%, 1%, and 0.1% significance levels

Degrees of freedom	5%	1%	0.1%
1	3.8415	6.6349	10.828
2	5.9915	9.2103	13.816
3	7.8147	11.3449	16.266
4	9.4877	13.2767	18.467
5	11.0705	15.0863	20.515
6	12.5916	16.8119	22.458
7	14.0671	18.4753	24.322
8	15.5073	20.0902	26.125
9	16.9190	21.6660	27.877
10	18.3070	23.2093	29.588
11	19.6751	24.7250	31.264
12	21.0261	26.2170	32.909
13	22.3620	27.6882	34.528
14	23.6848	29.1412	36.123
15	24.9958	30.5779	37.697
16	26.2962	31.9999	39.252
17	27.5871	33.4087	40.790
18	28.8693	34.8053	42.312
19	30.1435	36.1909	43.820
20	31.4104	37.5662	45.315
21	32.6706	38.9322	46.797
22	33.9244	40.2894	48.268
23	35.1725	41.6384	49.728
24	36.4150	42.9798	51.179
25	37.6525	44.3141	52.618
26	38.8851	45.6417	54.052
27	40.1133	46.9629	55.476
28	41.3371	48.2782	56.892
29	42.5570	49.5879	58.301
30	43.7730	50.8922	59.703
40	55.7585	63.6907	73.402
50	67.5048	76.1539	86.661
60	79.0819	88.3794	99.607
70	90.5312	100.425	112.317
80	101.879	112.329	124.839
90	113.145	124.116	137.208
100	124.342	135.807	149.449

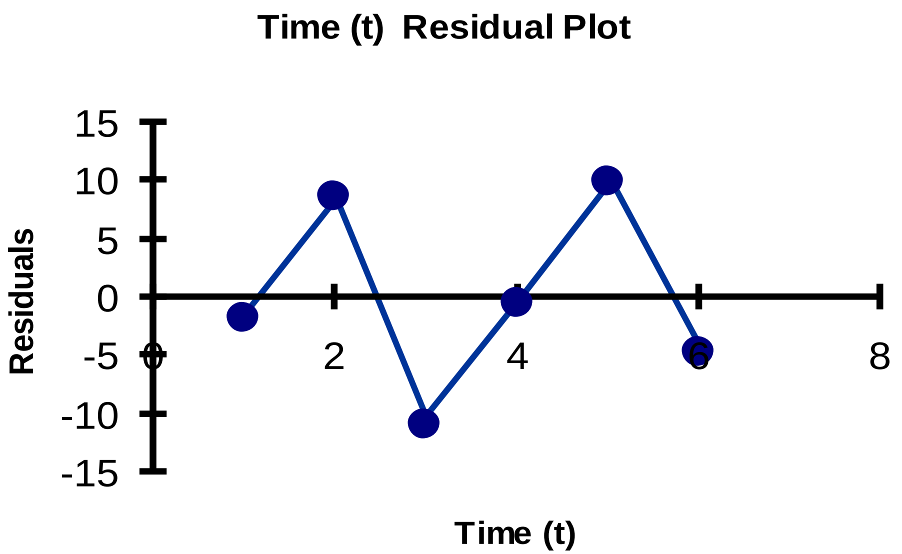
Reprinted from E. S. Pearson and H. O. Hartley (editors), *Biometrika Tables for Statisticians*, Cambridge, Cambridge University Press, 1970 with kind permission of the Biometrika Trustees.

ASSUMPTION VIOLATIONS: $E[\varepsilon_i \varepsilon_j]=0$

- This assumption implies that the errors are independent.
- If violated, then the errors are autocorrelated.
  - Errors are not independent
  - Errors in one period (for one individual observation) is correlated with another.
- Autocorrelation is a prominent feature of time series data. Could be found in cross section too.

ASSUMPTION VIOLATIONS: $E[\varepsilon_i \varepsilon_j]=0$

- Autocorrelation is correlation of the errors (residuals) over time

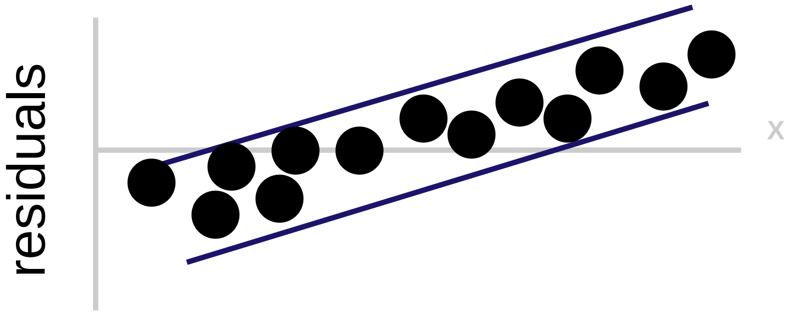


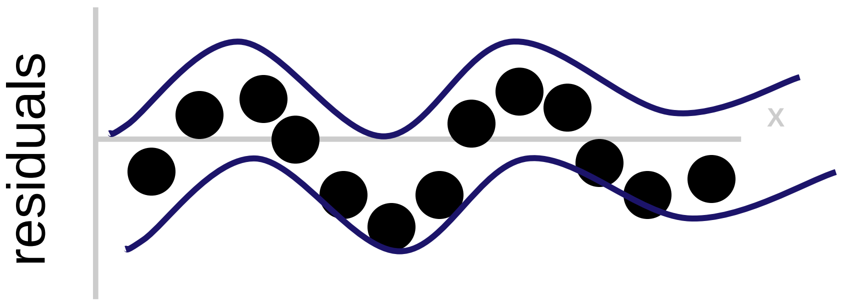
- Here, residuals show a cyclic pattern, not random

- Violates the regression assumption that residuals are random and independent

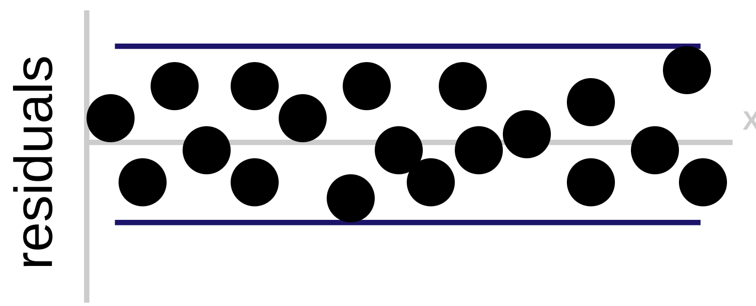
Assumption Violations: E[$\varepsilon_i \varepsilon_j$]=0

Not Independent





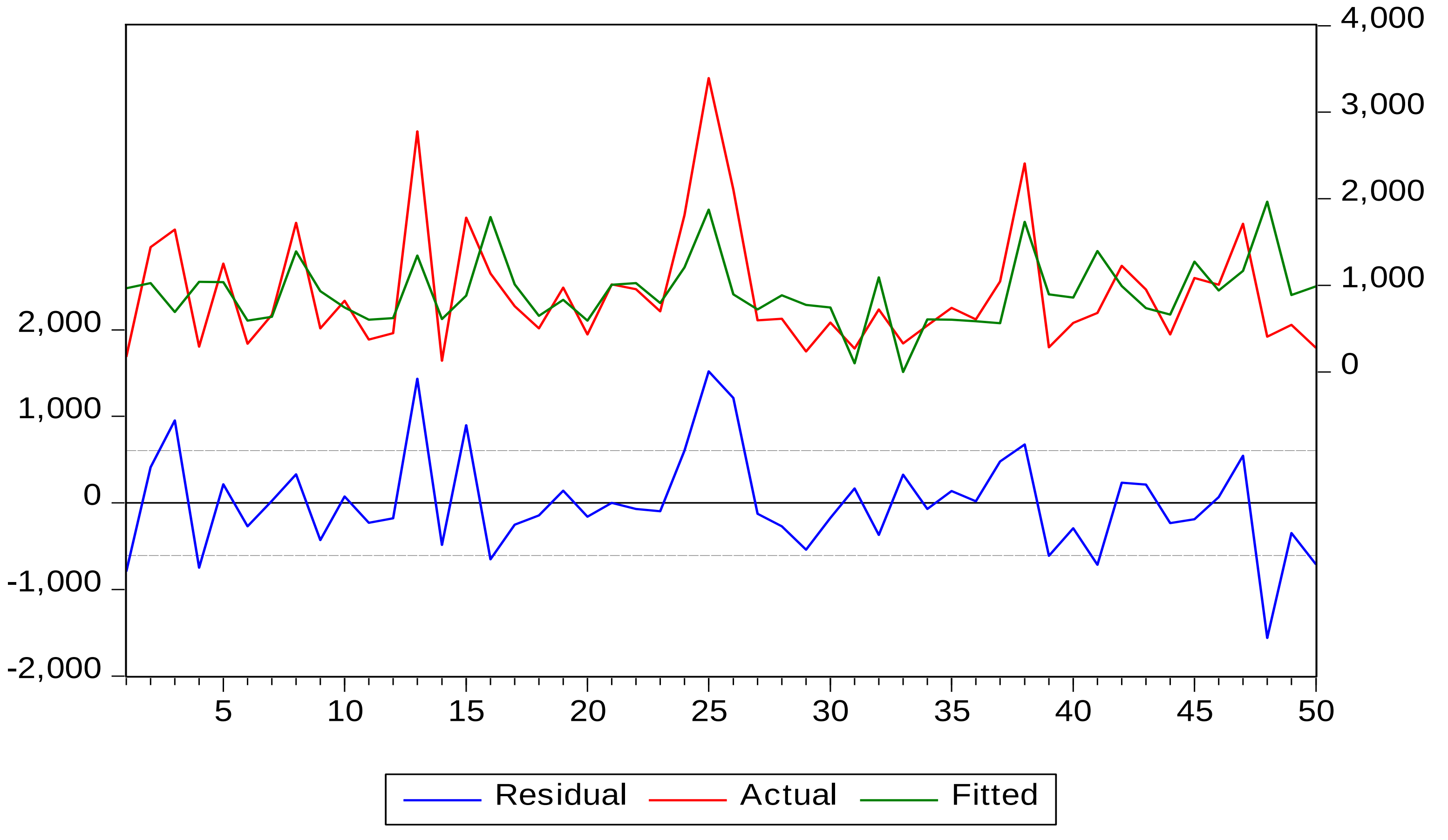
Independent



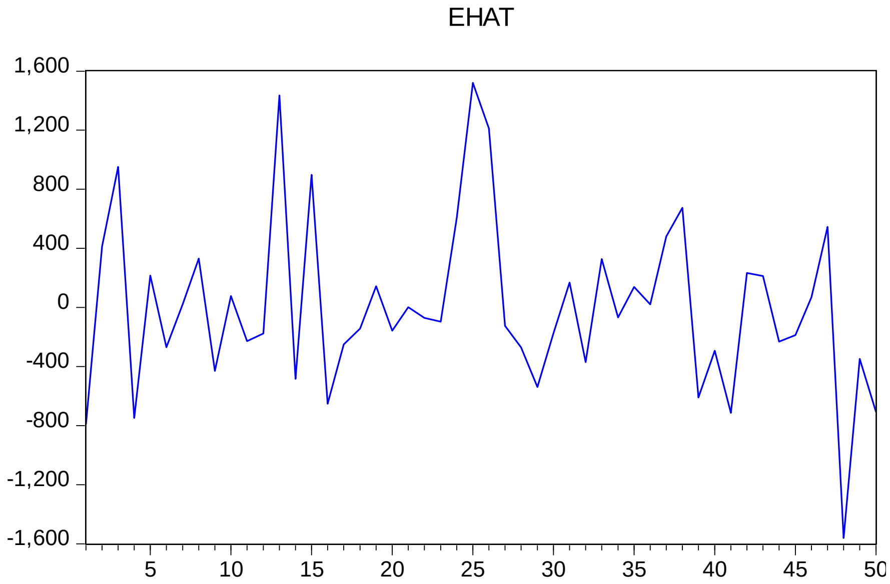
Assumption Violations: E[$\varepsilon_i \varepsilon_j$]=0

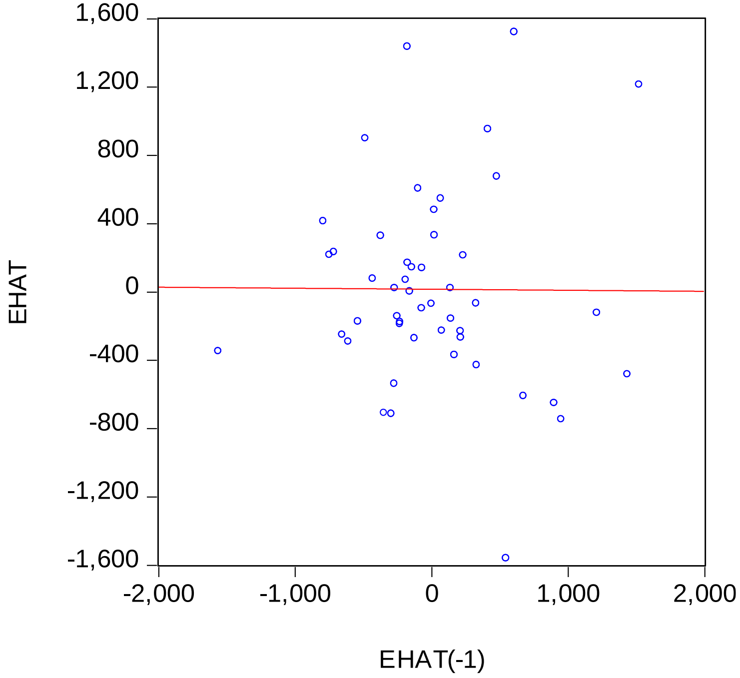
- The consequences of autocorrelation
  - The coefficient estimates derived using OLS are still unbiased, but they are inefficient, i.e. they are not BLUE, even in large sample sizes.
  - Thus, if the standard error estimates are inappropriate, there exists the possibility that we could make the wrong inferences.
  - R^2 is likely to be inflated relative to its "correct" value for positively correlated residuals.

APPLICATION 6
DETECTION OF AUTOCORRELATION



APPLICATION 6
DETECTION OF AUTOCORRELATION





ASSUMPTION VIOLATIONS: E[$\varepsilon_i \, \varepsilon_j$]=0

The Durbin-Watson Test

- The Durbin-Watson (DW) is a test for first order autocorrelation - i.e. it assumes that the relationship is between an error and the previous one

$$\varepsilon_t = \rho u_{t-1} + v_t \qquad (1)$$

where $v_t \sim N(0, \sigma_v^2)$.

- The DW test statistic actually tests

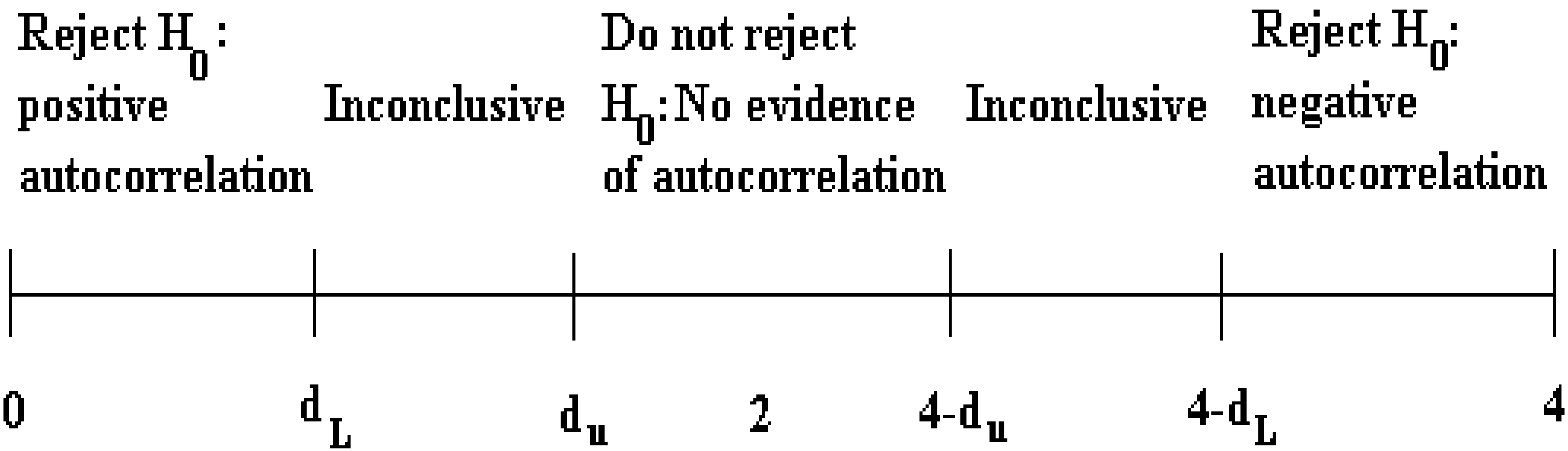
$$H_0 : \rho=0 \text{ (no autocorrelation) and } H_1 : \rho\neq 0$$

- The test statistic is calculated by

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T} \hat{u}_t^2}$$

Assumption Violations: $E[\varepsilon_i \varepsilon_j]=0$

The Durbin-Watson Test



<u>Conditions which Must be Fulfilled for DW to be a Valid Test</u>

1. Constant term in regression
2. Regressors are non-stochastic
3. No lags of dependent variable

ASSUMPTION VIOLATIONS: $E[\varepsilon_i \varepsilon_j]=0$

Table A.5 Durbin–Watson d statistic: d_L and d_U, 5% significance level

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Assumption Violations: $E[\varepsilon_i \varepsilon_j]=0$

The Breusch-Godfrey Test (LM Test)

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

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + \rho_r \varepsilon_{t-r} + v_t,\ v_t \sim iid(0, \sigma^2)$$

- The null and alternative hypotheses are:

$$H_0 : \rho_1 = 0 \text{ and } \rho_2 = 0 \text{ and } \dots \text{ and } \rho_r = 0 \text{ (No autocorrelation)}$$
$$H_1 : \rho_1 \neq 0 \text{ or } \rho_2 \neq 0 \text{ or } \dots \text{ or } \rho_r \neq 0$$

- The test is carried out as follows:
  1. Estimate the linear regression using OLS and obtain the residuals, $\hat{\varepsilon}_t$.
  2. Regress $\hat{\varepsilon}_t$ on all of the regressors from stage 1 (the x's) plus
     Obtain R^2 from this regression.
  3. It can be shown that $(T\text{-}r)R^2 \sim \chi^2(r)$
- If the test statistic exceeds the critical value from the statistical tables, reject the null hypothesis of no autocorrelation.

Assumption Violations: E[$\varepsilon_i \varepsilon_j$]=0

- There are many remedies to correct for serial correlation
  - Using lagged dependent variable (will be revisited when we deal with time series)
  - Use Methods such as Cochrane-Orcutt if the form is known.
  - Use Autocorrelation Consistent Standard Errors such as Newey-West

- Time Series Data
  - Autocorrelation is a key feature
  - Easily fixed if the dynamic process is stationary – by adding lags.
  - Has crucial implications on the stability of the process and long-run analysis
  - We will revisit this issue when dealing with time series

THANK YOU