Financial econometrics

Chapter 5,
Linear regression model diagnostics

Overview I

Introduction I

How well the OLS estimate of the multiple regression line describes, or "fits," the data?

Three commonly used statistics in multiple regression are

- standard error of the regression
- the regression R^2
- the Adjusted R^2

The Standard Error of the Regression (SER) I

The Standard Error of the Regression (SER)

- The standard error of the regression (SER) estimates the standard deviation of the error term u_i.
- Thus the SER is a measure of the spread of the distribution of Y around the regression line

The SER is

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}$$

where

$$s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}^2 g_i = \frac{SSR}{n-k-1}$$

The Standard Error of the Regression (SER) II

Notes:

- the divisor $n - k - 1$ adjusts for the downward bias introduced by estimating$k + 1$ coefficients (the k slope coefficients plus the intercept).
- Using $n - k - 1$ rather than n is called a degrees-of-freedom adjustment.
- If there is a single regressor, then $k = 1$,
- When n is large, the effect of the degrees-of-freedom adjustment is negligible.

The R^2 I

Definition

The regression R^2 is the fraction of the sample variance of Y_i explained by (or predicted by) the regressors.

Equivalently, the R^2 is 1 minus the fraction of the variance of Y_i **not** explained by the regressors.
The R^2 is defined as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where,

- explained sum of squares (EES): $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$
- total sum of squares (TSS): $\sum\left(Y_i - \bar{Y}\right)^2$
- sum of squared residuals (SSR): $\sum\left(Y_i - \hat{Y}_i\right)^2$

The R^2 II

In multiple regression, the R^2 increases whenever a regressor is added unless the estimated coefficient on the added regressor is exactly 0.

To see this, think about starting with one regressor and then adding a second

- When you use OLS to estimate the model with both regressors, OLS finds the values of the coefficients that minimize the sum of squared residuals
- If OLS happens to choose the coefficient on the new regressor to be exactly 0, then the SSR will be the same whether or not the second variable is included in the regression.
- But if OLS chooses any value other than 0, then it must be that this value reduced the SSR relative to the regression that excludes this regressor.

The R^2 III

- In practice, it is extremely unusual for an estimated coefficient to be exactly 0, so in general the SSR will decrease when a new regressor is added
- $\rightarrow$ this means that the R2 generally increases (and never decreases) when a new regressor is added.

The Adjusted R^2 I

Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model.

One way to correct for this is to reduce the R^2 by some factor, and this is what the adjusted R^2, or $\bar{R}^2$, does.

The adjusted is a modified version of the R^2 that does not necessarily increase when a new regressor is added, it is defined as

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y}$$

The Adjusted R^2 II

The difference between this formula and the definition of the R^2 is that the ratio of the sum of squared residuals to the total sum of squares is multiplied by the factor $(n-1)/(n-k-1)$.

Moreover:

1. $(n-1)/(n-k-1)$ is always greater than 1, so $\bar{R}^2$ is always less than R^2 .
2. adding a regressor has two opposite effects on the $\bar{R}^2$:
   - the SSR falls, which increases the R^2
   - the factor $(n-1)/(n-k-1)$ increases

   $\rightarrow$ Whether the $\bar{R}^2$ increases or decreases depends on which of these two effects is stronger.
3. $\bar{R}^2$ can be negative: This happens when the regressors, taken together, reduce the sum of squared residuals by such a small amount that this reduction fails to offset the factor $(n-1)/(n-k-1)$.

Using the R^2 and adjusted R^2

The $\bar{R}^2$ is useful because it quantifies the extent to which the regressors account for, or explain, the variation in the dependent variable. Nevertheless, heavy reliance on the $\bar{R}^2$ (or R^2) can be a trap

- In applications in which the goal is to produce reliable out-of-sample predictions,
- including many regressors can produce a good in-sample fit but can degrade the out-of- sample performance.
- Although the $\bar{R}^2$ improves upon the R^2 for this purpose, simply maximizing the $\bar{R}^2$ still can produce poor out-of-sample forecasts

Violation of the Assumptions of the CLRM

- Recall what we assumed for the Classical linear regression model (CLRM) and diagnostics disturbance terms:
  1. $\mathrm{E}(u_t) = 0$
  2. $\mathrm{var}(u_t) = \sigma^2 < \infty$
  3. $\mathrm{cov}(u_i, u_j) = 0$
  4. The X matrix is non-stochastic or fixed in repeated samples $\mathrm{cov}(u_t, x_t) = 0$
  5. $u_t \sim \mathrm{N}(0, \sigma^2)$

Investigating Violations of the Assumptions of the CLRM

- 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

If the assumptions are still violated:
$\rightarrow$ we work around the problem so that we use alternative techniques which are still valid

Statistical Distributions for Diagnostic Tests

- Often, an *F*- and a χ^2- version of the test are available.
- The *F*-test version involves estimating a restricted and an unrestricted version of a test regression and comparing the *RSS*.
- The χ^2- version is sometimes called an "LM" test, and only has one degree of freedom parameter: the number of restrictions being tested, *m*.
- Asymptotically, the 2 tests are equivalent since the χ^2 is a special case of the *F*-distribution:

$$F(m, T-k) \to \frac{\chi^2(m)}{m} \quad \text{as} \quad (T-k) \to \infty$$

- For small samples, the *F*-version is preferable.

Assumption 1: $E(u_t) = 0$

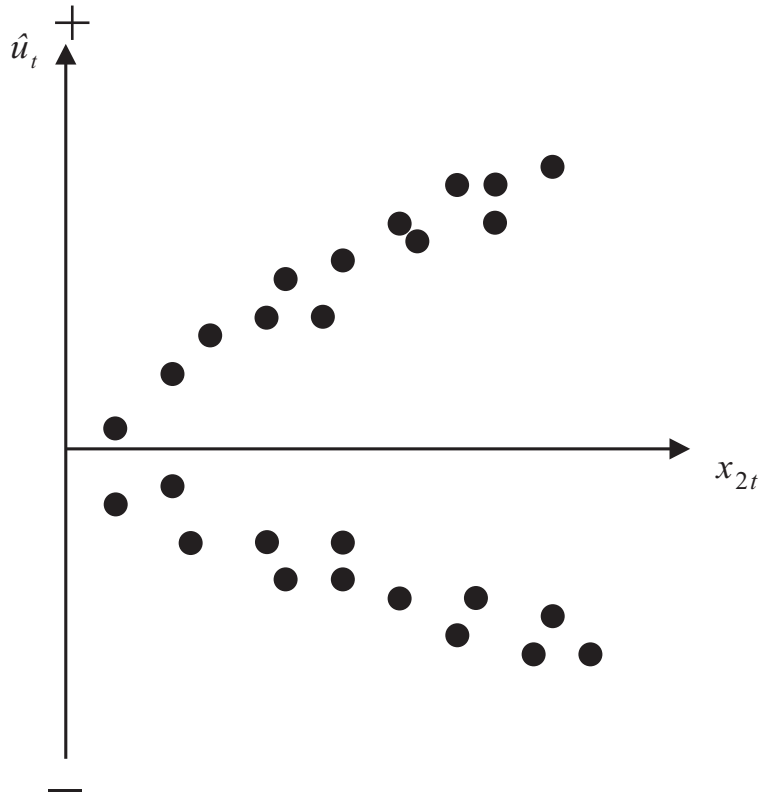
- Assumption that the mean of the disturbances is zero.
- For all diagnostic tests, we cannot observe the disturbances and so perform the tests of the residuals.
- The mean of the residuals will always be zero provided that there is a constant term in the regression.

Consequences of Using OLS in the Presence of Heteroscedasticity

- OLS estimation still gives unbiased coefficient estimates, but they are no longer BLUE.
- This implies that if we still use OLS in the presence of heteroscedasticity, our standard errors could be inappropriate and hence any inferences we make could be misleading.
- Whether the standard errors calculated using the usual formulae are too big or too small will depend upon the form of the heteroscedasticity.

Assumption 2: $\text{var}(u_t) = \sigma^2 < \infty$

- We have so far assumed that the variance of the errors is constant, σ^2 - this is known as homoscedasticity. If the errors do not have a constant variance, we say that they are heteroscedastic e.g. say we estimate a regression and calculate the residuals, $\hat{u}_t$.



Detection of Heteroscedasticity

- Graphical methods
- Formal tests: There are many of them: we will discuss Goldfeld-Quandt (Goldfeld and Quandt, 1965) test and White's test (White, 1980)

[1] Stephen M Goldfeld and Richard E Quandt. "Some tests for homoscedasticity". In: *Journal of the American statistical Association* 60.310 (1965), pp. 539–547

[2] Halbert White. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *Econometrica: journal of the Econometric Society* (1980), pp. 817–838

The GQ test I

The Goldfeld-Quandt (GQ) test is carried out as follows.

1. Split the total sample of length T into two sub-samples of length T_1 and T_2. The regression model is estimated on each sub-sample and the two residual variances are calculated.

2. The null hypothesis is that the variances of the disturbances are equal, $H_0 : \sigma_1^2 = \sigma_2^2$

3. 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.

$$GQ = \frac{s_1^2}{s_2^2}$$

4. The test statistic is distributed as an $\mathrm{F}(T_1 - k,\ T_2 - k)$ under the null of homoscedasticity.

The GQ test II

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

The White's Test I

- White's general test for heteroscedasticity is one of the best approaches because it makes few assumptions about the form of the heteroscedasticity.
- The test is carried out as follows:
  1. Assume that the regression we carried out is as follows

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

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

  2. Then run the auxiliary regression

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

The White's Test II

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

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

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.

How Do we Deal with Heteroscedasticity? I

- If the form (i.e. the cause) of the heteroscedasticity is known, then we can use an estimation method that takes this into account (called generalized least squares, GLS).
- A simple illustration of GLS is as follows: Suppose that the error variance is related to another variable z_t by

$$\text{var}(u_t) = \sigma^2 z_t^2$$

- To remove the heteroscedasticity, divide the regression equation by z_t

$$\frac{y_t}{z_t} = \beta_1 \frac{1}{z_t} + \beta_2 \frac{x_{2t}}{z_t} + \beta_3 \frac{x_{3t}}{z_t} + v_t$$

where $v_t = \dfrac{u_t}{z_t}$ is an error term.

How Do we Deal with Heteroscedasticity? II

- Now $\text{var}(u_t) = \sigma^2 z_t^2$, $\text{var}(v_t) = \text{var}\left(\frac{u_t}{z_t}\right) = \frac{\text{var}(u_t)}{z_t^2} = \frac{\sigma^2 z_t^2}{z_t^2} = \sigma^2$ for known z_t.
- So the disturbances from the new regression equation will be homoscedastic.
- **Other solutions include:**
  1. Transforming the variables into logs or reducing by some other measure of "size".
  2. 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.

  This makes us more "conservative" in hypothesis testing, so that we would need more evidence against the null hypothesis before we would reject it.

Autocorrelation I

- We assumed of the CLRM's errors that Cov $(u_i, u_j) = 0$ for $i \neq j$,

  This is essentially the same as saying there is no pattern in the errors.

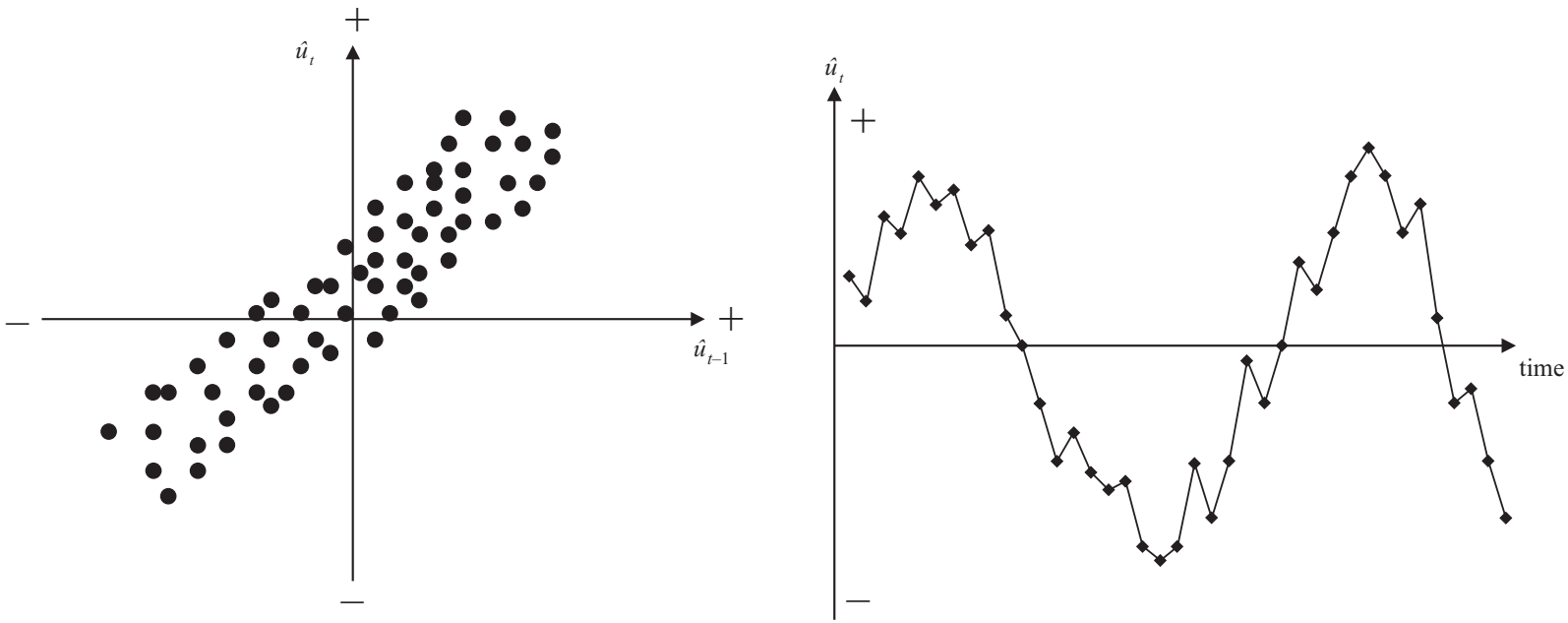
- Obviously we never have the actual u's, so we use their sample counterpart, the residuals (the $\hat{u}_t$'s).

- If there are patterns in the residuals from a model, we say that they are autocorrelated.

- Some stereotypical patterns we may find in the residuals are given on the next 3 slides.

Autocorrelation II

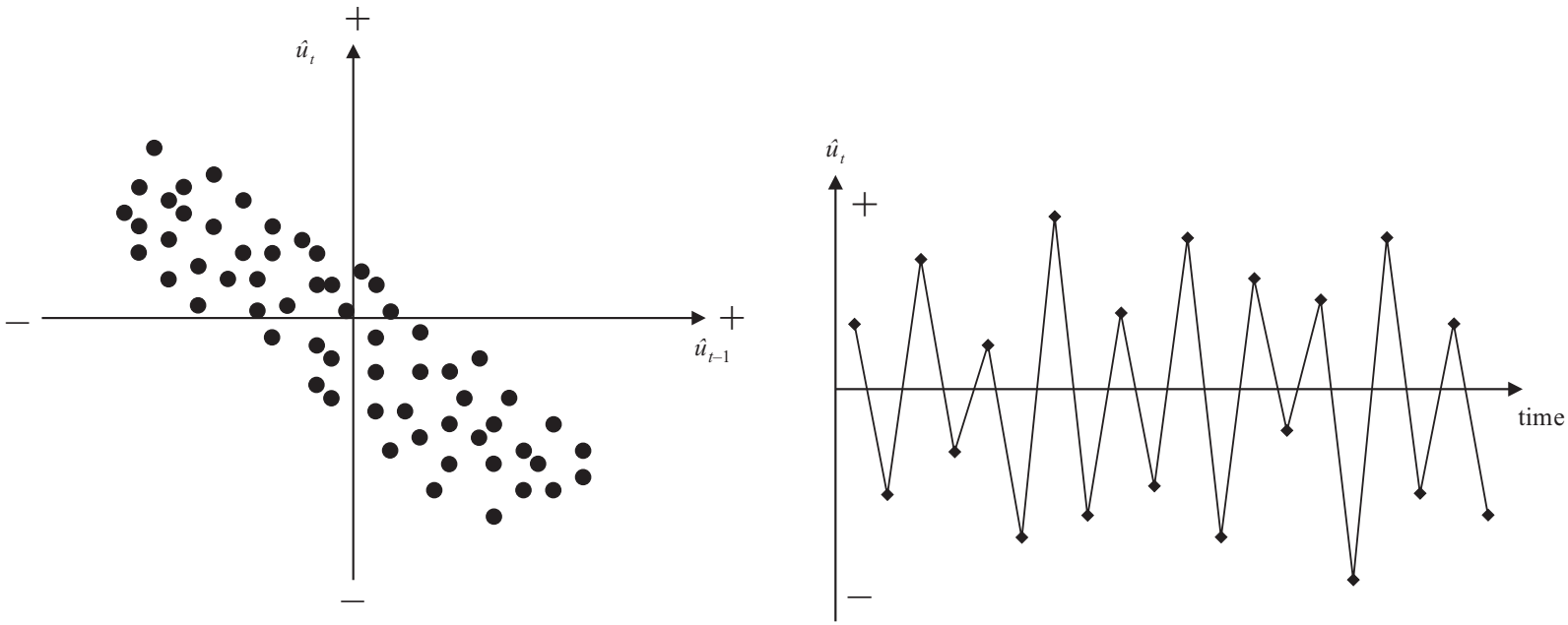
Positive Autocorrelation



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

Autocorrelation III

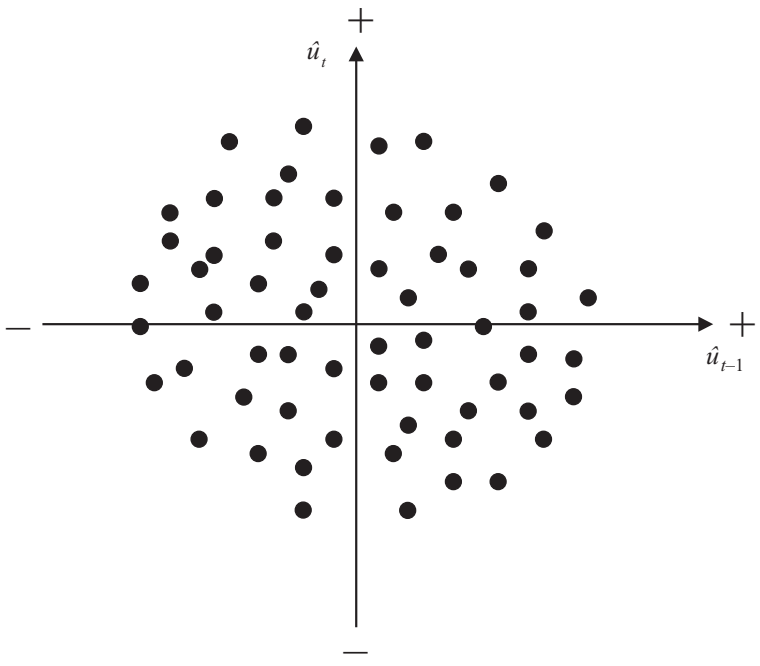
Negative Autocorrelation

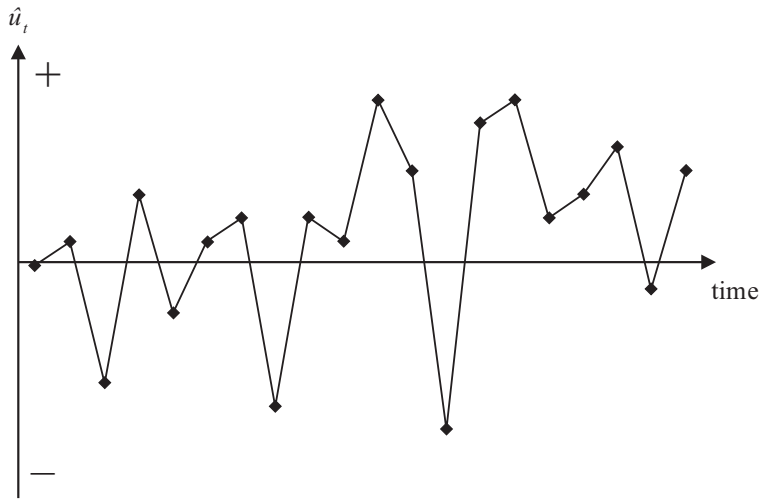


Negative autocorrelation is indicated by an alternating pattern where the residuals cross the time axis more frequently than if they were distributed randomly

Autocorrelation IV

No pattern in residuals – No autocorrelation





No pattern in residuals at all: this is what we would like to see

The Durbin-Watson Test: Critical Values I

[1] James Durbin and Geoffrey S Watson. "Testing for serial correlation in least squares regression. I". In: *Breakthroughs in Statistics*. Springer, 1992, pp. 237–259

- The Durbin-Watson (DW) is a test (Durbin and G. S. Watson, 1992) for first order autocorrelation - i.e. it assumes that the relationship is between an error and the previous one

$$u_t = \rho u_{t-1} + v_t \tag{1}$$

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

The Durbin-Watson Test: Critical Values II

- The DW test statistic actually tests

$$\mathrm{H}_0 : \rho = 0 \quad \text{and} \quad \mathrm{H}_1 : \rho \neq 0$$

- The test statistic is calculated by

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

The Durbin-Watson Test: Critical Values III

- We can also write

$$DW \approx 2(1 - \hat{\rho}) \tag{2}$$

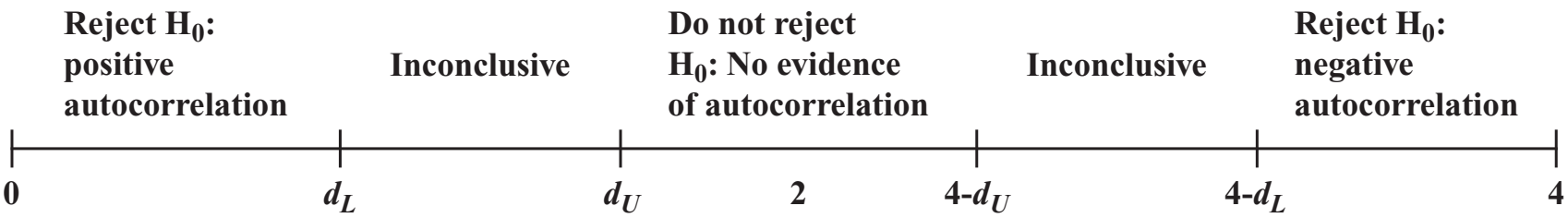
  where $\hat{\rho}$ is the estimated correlation coefficient. Since $\hat{\rho}$ is a correlation, it implies that $-1 \leq \hat{\rho} \leq 1$.

- Rearranging for DW from (2) would give $0 \leq DW \leq 4$.

- If $\hat{\rho} = 0$, DW=2. So roughly speaking, do not reject the null hypothesis if DW is near 2 $\rightarrow$ i.e. there is little evidence of autocorrelation

- Unfortunately, DW has 2 critical values, an upper critical value (d_U) and a lower critical value (d_L), and there is also an intermediate region where we can neither reject nor not reject H_0.

The Durbin-Watson Test: Critical Values IV



Conditions which Must be Fulfilled for DW to be a Valid Test

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

The Breusch-Godfrey Test I

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

$$\begin{aligned} u_t &= \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \cdots + \rho_r u_{t-r} + v_t, \\ & v_t \sim N\left(0, \sigma_v^2\right) \end{aligned}$$

- The null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \rho_1 = 0 \quad \text{and} \quad \rho_2 = 0 \quad \text{and} \ldots \text{and} \ \rho_r = 0 \\ H_1 &: \rho_1 \neq 0 \quad \text{or} \quad \rho_2 \neq 0 \quad \text{or} \ldots \text{or} \ \rho_r \neq 0 \end{aligned}$$

- The test is carried out as follows:
  1. Estimate the linear regression using OLS and obtain the residuals, $\hat{u}_t$.
  2. Regress $\hat{u}_t$ on all of the regressors from stage 1 (the xs) plus $\hat{u}_{t-1}$, $\hat{u}_{t-2}, \ldots, \hat{u}_{t-r}$;

The Breusch-Godfrey Test II

Obtain R^2 from this regression.

3. It can be shown that

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

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

Consequences of Ignoring Autocorrelation if it is Present

- 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.

"Remedies" for Autocorrelation

- If the form of the autocorrelation is known, we could use a GLS procedure – i.e. an approach that allows for autocorrelated residuals e.g., Cochrane-Orcutt estimation (Cochrane and Orcutt, 1949).
- But such procedures that "correct" for autocorrelation require assumptions about the form of the autocorrelation.
- If these assumptions are invalid, the cure would be more dangerous than the disease! - see Hendry and Mizon, 1978.
- However, it is unlikely to be the case that the form of the autocorrelation is known, and a more "modern" view is that residual autocorrelation presents an opportunity to modify the regression.

Multicollinearity

- This problem occurs when the explanatory variables are very highly correlated with each other.
- Perfect multicollinearity

  Cannot estimate all the coefficients
  - e.g. suppose $x_3 = 2x_2$
    and the model is $y_t = \beta_1 + \beta_2 x_{2t} + \beta x_{3t} + \beta_4 x_{4t} + u_t$
- Problems if Near Multicollinearity is Present but Ignored
  - R^2 will be high but the individual coefficients will have high standard errors.
  - The regression becomes very sensitive to small changes in the specification.
  - Thus confidence intervals for the parameters will be very wide, and significance tests might therefore give inappropriate conclusions.

Measuring Multicollinearity

- The easiest way to measure the extent of multicollinearity is simply to look at the matrix of correlations between the individual variables. e.g.

corr	x_2	x_3	x_4
x_2	–	0.2	<u>0.8</u>
x_3	0.2	–	0.3
x_4	<u>0.8</u>	0.3	–

- But another problem: if 3 or more variables are linear
  - e.g. $x_{2t} + x_{3t} = x_{4t}$

- Note that high correlation between y and one of the x's is not muticollinearity.

Solutions to the Problem of Multicollinearity

- "Traditional" approaches, such as ridge regression or principal components. But these usually bring more problems than they solve.
- Some econometricians argue that if the model is otherwise OK, just ignore it
- The easiest ways to "cure" the problems are
  - drop one of the collinear variables
  - transform the highly correlated variables into a ratio
  - go out and collect more data e.g.
    - a longer run of data
    - switch to a higher frequency

Adopting the Wrong Functional Form I

- We have previously assumed that the appropriate functional form is linear.
- This may not always be true.

The RESET test

- We can formally test this using Ramsey's RESET test (Ramsey, 1969), which is a general test for mis-specification of functional form.
- Essentially the method works by adding higher order terms of the fitted values (e.g. $\hat{y}_t^2, \hat{y}_t^3$, etc.) into an auxiliary regression:

  Regress $\hat{u}_t$ on powers of the fitted values:

$$\hat{u}_t = \beta_0 + \beta_1 \hat{y}_t^2 + \beta_2 \hat{y}_t^3 + \cdots + \beta_{p-1} \hat{y}_t^p + v_t$$

  Obtain R^2 from this regression. The test statistic is given by TR^2 and is distributed as a $\chi^2(p-1)$.
- So if the value of the test statistic is greater than a $\chi^2(p-1)$ then reject the null hypothesis that the functional form was correct.
- The RESET test gives us no guide as to what a better specification might be.

- *In class we saw three variants of the RESET test, for the exam know the above.*
- One possible cause of rejection of the test is if the true model is

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{2t}^2 + \beta_4 x_{2t}^3 + u_t$$

  In this case the remedy is obvious.

- Another possibility is to transform the data into logarithms. This will linearise many previously multiplicative models into additive ones:

$$y_t = A x_t^{\beta} e^{u_t} \Leftrightarrow \ln(y_t) = \alpha + \beta \ln(x_t) + u_t$$

Normality

- Why did we need to assume normality for hypothesis testing?
- What happens if normality does not hold?

Non-normality: consequences for skewness and kurtosis:

- A normal distribution is not skewed and is defined to have a coefficient of kurtosis of 3.
- The kurtosis of the normal distribution is 3 so its excess kurtosis is zero.
- Skewness and kurtosis are the (standardized) third and fourth moments of a distribution.

The Jarque-Brera test

- Bera and Jarque formalize this by testing the residuals for normality by testing whether the coefficient of skewness and the coefficient of excess kurtosis are jointly zero.
- It can be proved that the coefficients of skewness and kurtosis can be expressed respectively as:

$$b_1 = \frac{E[u^3]}{(\sigma^2)^{3/2}} \quad \text{and} \quad b_2 = \frac{E[u^4]}{(\sigma^2)^2}$$

- The Bera Jarque test statistic is given by

$$W = T\left[\frac{b_1^2}{6} + \frac{(b_2 - 3)^2}{24}\right] \sim \chi_2^2$$

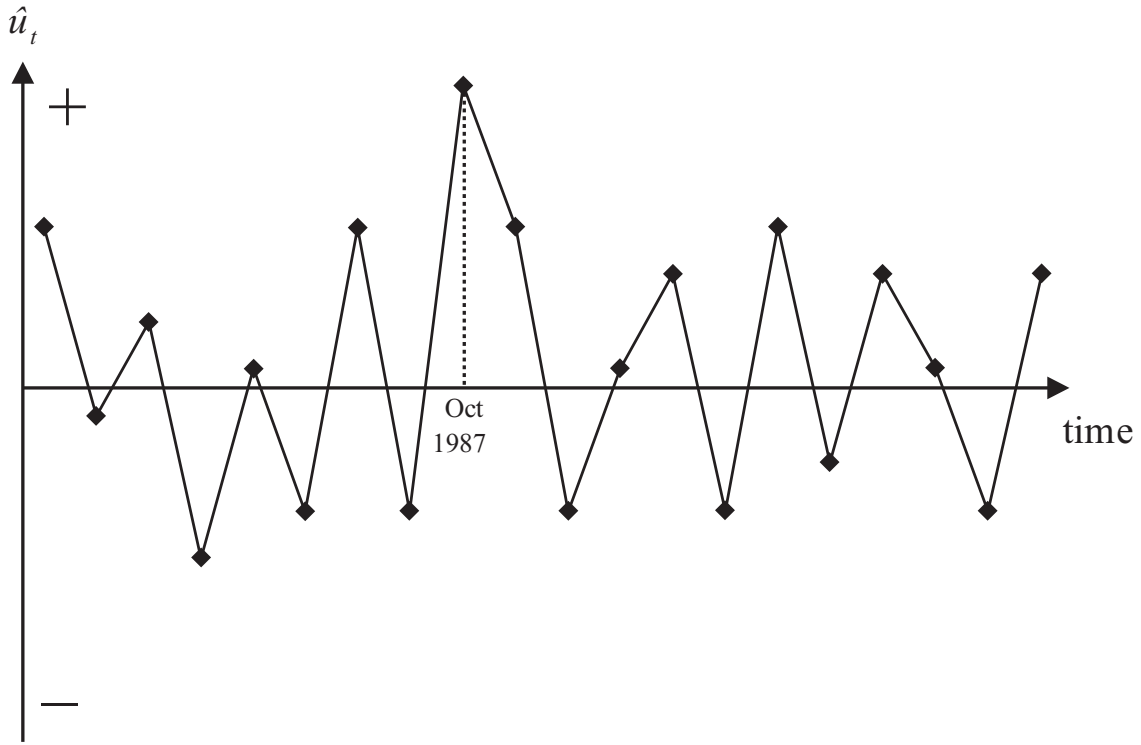
- We estimate b_1 and b_2 using the residuals from the OLS regression: (i) run the regression, (ii) estract the residuals $\hat{u}$, (iii) do the JB test.

What do we do if we find evidence of Non-Normality? I

- It is not obvious what we should do!
- Could use a method that does not assume normality, but is difficult, and what are its properties?
- Often the case that one or two very extreme residuals causes us to reject the normality assumption.
- An alternative is to use dummy variables.

  e.g. say we estimate a monthly model of asset returns from 1980-1990, and we plot the residuals, and find a particularly large outlier for October 1987:

What do we do if we find evidence of Non-Normality? II



- Create a new variable:

  $D87M10_t = 1$ during October 1987 and zero otherwise.

  This effectively knocks out that observation. But we need a theoretical reason for adding dummy variables.

Slides

Disclaimer:

- Some slides are original, based on (Stock and M. W. Watson, 2020).
- Some slides rearranged from Chris Brooks' slides from (Brooks, 2014) (copyrighted)

Biblio I

[CO49] Donald Cochrane and Guy H Orcutt. "Application of least squares regression to relationships containing auto-correlated error terms". In: *Journal of the American statistical association* 44.245 (1949), pp. 32–61.

[GQ65] Stephen M Goldfeld and Richard E Quandt. "Some tests for homoscedasticity". In: *Journal of the American statistical Association* 60.310 (1965), pp. 539–547.

[Ram69] James Bernard Ramsey. "Tests for specification errors in classical linear least-squares regression analysis". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 31.2 (1969), pp. 350–371.

Biblio II

[HM78] David F Hendry and Grayham E Mizon. "Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the Bank of England". In: *The Economic Journal* 88.351 (1978), pp. 549–563.

[Whi80] Halbert White. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *Econometrica: journal of the Econometric Society* (1980), pp. 817–838.

[DW92] James Durbin and Geoffrey S Watson. "Testing for serial correlation in least squares regression. I". In: *Breakthroughs in Statistics*. Springer, 1992, pp. 237–259.

[Bro14] Chris Brooks. *Introductory Econometrics for Finance*. 3rd ed. Cambridge University Press, 2014. DOI: 10.1017/CBO9781139540872.

Biblio III

[SW20] James H Stock and Mark W Watson. *Introduction to econometrics*. Pearson, 2020.