UNIVERSITÀ DEGLI STUDI DELLA TUSCIA

SAPIENZA UNIVERSITÀ DI ROMA

Statistics for business and decision making course

Dr. Ilaria Benedetti

11. Simple regression model: Goodness of fit

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR1:
The value of *y*, for each value of *x*, is: $y = \beta_1 + \beta_2 x + e$

Assumption SR2:
The expected value of the random error *e* is:

$$E(e) = 0$$

This is equivalent to assuming that

$$E(y) = \beta_1 + \beta_2 x$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR3:

The variance of the random error e is: $\text{var}(e) = \sigma^2 = \text{var}(y)$

- The random variables y and e have the same variance because they differ only by a constant (Homoskedasticity assumption)

Assumption SR4:

The covariance between any pair of random errors, e_i and e_j is:

$$\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$$

The stronger version of this assumption is that the random errors e are statistically independent, in which case the values of the dependent variable y are also statistically independent

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR5:
The variable *x* is not random, and must take at least two different values

Assumption SR6:
(*optional*) The values of *e* are *normally distributed* about their mean if the values of *y* are normally distributed, and *vice versa*

$$e \sim N(0, \sigma^2)$$

Least Square Prediction

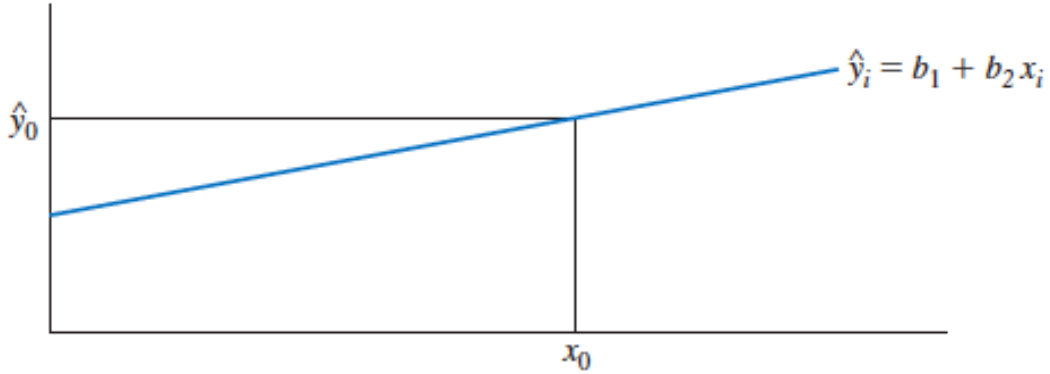
In order to use regression analysis as a basis for prediction, we must assume that y0 and x0 are related to one another by the same regression model that describes our sample of data, so that, in particular, SR1 holds for these observations:

$$\boldsymbol{y_0 = \beta_1 + \beta_2 x_0 + e_0}$$

e_0 is a random error

We assume that $E(y_0) = \beta_1 + \beta_2 x_0$ and $e_0 = 0$

We also assume that e_0 has the same varince as the regression error $\text{var}(e_0) = \sigma^2$ and e_0 is uncorrelated with the random errors that are part of the sample data, so that $cov(e_0; e_1) = 0$ for $i = 1,2, \dots, N$



Least Square Prediction

The task of predicting y_0 is related to the problem of estimating

$$E(y_0) = \beta_1 + \beta_2 x_0$$

The outcome $y_0 = E(y_0) + e_0$ is composed of two parts:

- Systematic (non random part) $E(y_0) = \beta_1 + \beta_2 x_0$
- Random component e_0

We estimate the systematic portion using $\widehat{E(y_0)} = \beta_1 + \beta_2 x_0$ and add an "estimate" of e_0 equal to its expected value (which is 0).

Therefore:

$$\widehat{y_0} = \widehat{E(y_0)} += b_1 + b_2 x_0$$

The least squares predictor of y_0 comes from the fitted regression line

Least Square Prediction

There are two major reasons for analyzing the model:

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

1. to explain how the dependent variable (yi) changes as the independent variable (xi) changes
2. to predict y_0 given an x_0

Closely allied with the prediction problem is the desire to use x_i to explain as much of the variation in the dependent variable y_i as possible.

In the regression model we call x_i the "explanatory" variable because we hope that its variation will "explain" the variation in y_i

Goodness of fit

- To develop a measure of the variation in yi that is explained by the model, we begin by separating yi into its explainable and unexplainable components.

$$y_i = E(y_i) + e_i$$

  - **$E(y_i)$** is the explainable or systematic part
  - ***e_i*** is the random, unsystematic and unexplainable component

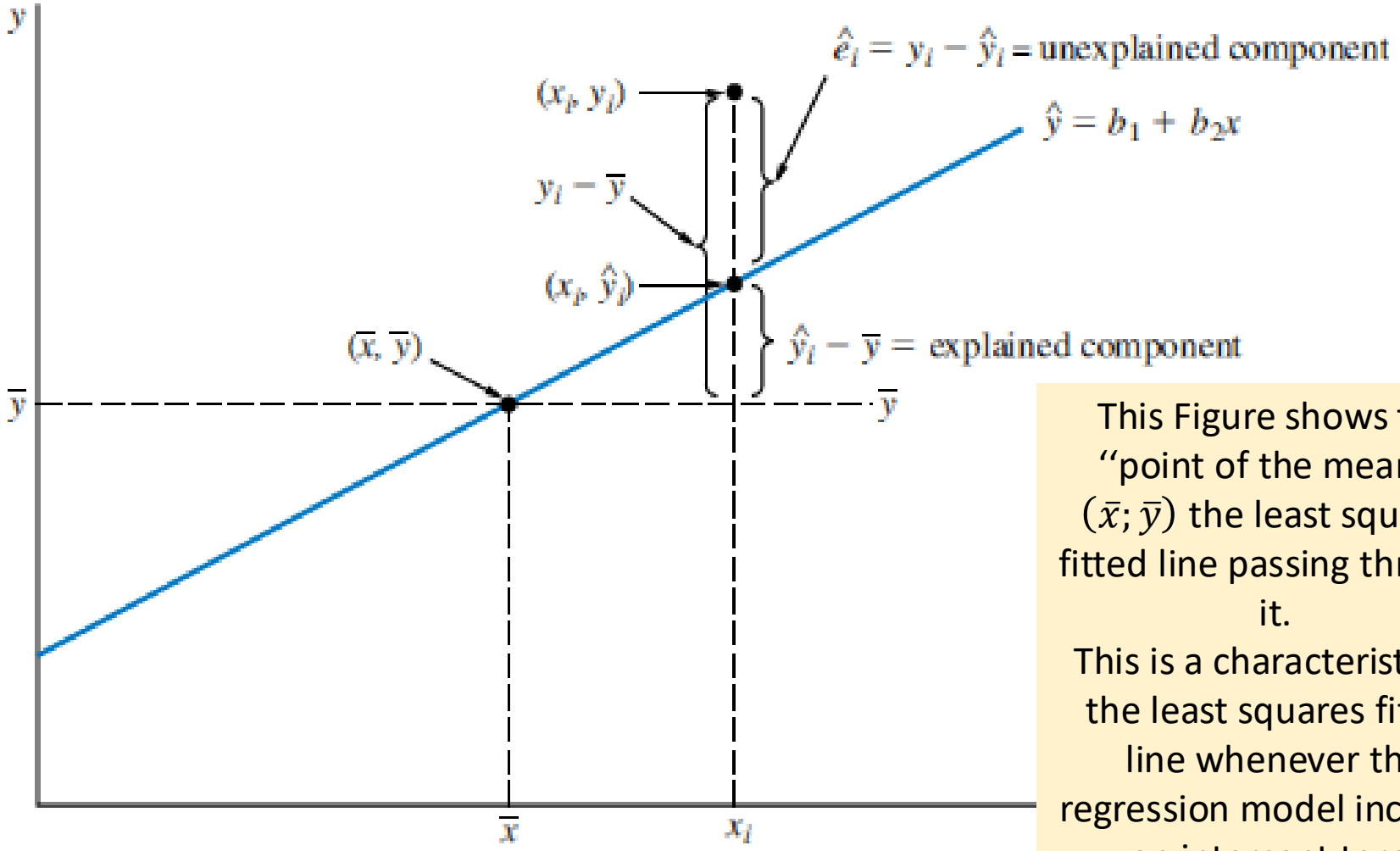
- we can write: $y_i = \hat{y}_i + \hat{e}_i$
- If we consider the "point of the mean" $(\bar{x}; \bar{y})$
- Subtract the sample mean y from both sides of the equation to obtain the total variability of the regression model:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$$

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$$

How good is the estimate?

Goodness of fit



This Figure shows the "point of the means" $(\bar{x}; \bar{y})$ the least squares fitted line passing through it.

This is a characteristic of the least squares fitted line whenever the regression model includes an intercept term

Goodness of fit

- Recall that the sample variance of y_i is $s_y^2 = \frac{\sum(\hat{y}_i - \bar{y})}{N-1}$

- Squaring and summing both sides, and using the fact that: $\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$

We get: $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum\hat{e}_i^2$

total sample Variation SST

explained sample Variation SSR

unexplained sample Variation SSE

decomposition of the "total sample variation" in y into explained and unexplained components .

These are called "sums of squares"

Goodness of fit

$$\sum (y_i - \bar{y})^2 = \text{total sum of squares} = \text{SST}$$

$$\sum (\hat{y}_i - \bar{y})^2 = \text{sum of squares due to regression} = \text{SSR}$$

$$\sum \hat{e}_i^2 = \text{sum of squares due to error} = \text{SSE}$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2$$

Goodness of fit

This decomposition of the total variation in y into a part that is explained by the regression model and a part that is unexplained allows us to define a measure, called the **coefficient of determination**, or **R^2**

Let's define the ***coefficient of determination, or R^2*** , as the proportion of variation in *y* explained by *x* within the regression model:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Goodness of fit

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

We can see that:

- The closer R^2 is to 1, the closer the sample values y_i are to the fitted regression equation
- If R^2 = 1, then all the sample data fall exactly on the fitted least squares line, so *SSE* = 0, and the model fits the data ''perfectly''
- If the sample data for *y* and *x* are uncorrelated and show no linear association, then the least squares fitted line is ''horizontal,'' and identical to *y*, so that *SSR* = 0 and R^2 = 0
- When **0 < R^2 < 1** then R^2 is interpreted as ''*the proportion of the variation in y about its mean that is explained by the regression model*''

Goodness of fit

- The correlation coefficient ρ_{xy} between *x* and *y* is defined as:

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Substituting sample values, as get the sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})/(N-1)$$

$$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (N-1)}$$

$$s_y = \sqrt{\sum (y_i - \bar{y})^2 / (N-1)}$$

- The sample correlation coefficient r_{xy} has a value between -1 and 1, and it measures the strength of the linear association between observed values of *x* and *y*

Goodness of fit

- Two relationships between R^2 and r_{xy}:
  1. $r^2_{xy} = R^2$ the square of the sample correlation coefficient between the sample data values xi and yi is algebraically equal to R2 in a simple regression model.

  2. R^2 can also be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = b_1 + b_2 x_i$

Example on food expenditure

For the food expenditure example, the sums of squares are:

$$SST = \sum (y_i - \bar{y})^2 = 495132.160$$

$$SSE = \sum (y_i - \hat{y})^2 = \sum \hat{e}_i^2 = 304505.176$$

Goodness of fit

$$R^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{304505.176}{495132.160}$$

$$= 0.385$$

We conclude that 38.5% of the variation in food expenditure (about its sample mean) is explained by our regression model, which uses only income as an explanatory variable

The sample correlation between the y and x sample values is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{478.75}{(6.848)(112.675)}$$

$$= 0.62$$

$$r_{xy}^2 = 0.62^2 = 0.385 = R^2$$

Goodness of fit

- The key ingredients in a report are:
  1. the *coefficient estimates*
  2. the *standard errors* (or *t*-values)
  3. an indication of *statistical significance (ex. P-value)*
  4. R^2
- Avoid using symbols like *x* and *y*
  - Use abbreviations for the variables that are readily interpreted, defining the variables precisely in a separate section of the report.

Goodness of fit: example

- For our food expenditure example, we might have:

  FOOD_EXP = weekly food expenditure by a household of size 3, in dollars

  INCOME = weekly household income, in $100 units

- And:

$$\widehat{Food\_exp} = 83.42 + 10.21 INCOME \qquad R^2 = 0.385$$

$$(\text{se}) \qquad (43.41)(2.09)^{***}$$

where

* indicates significant at the 10% level

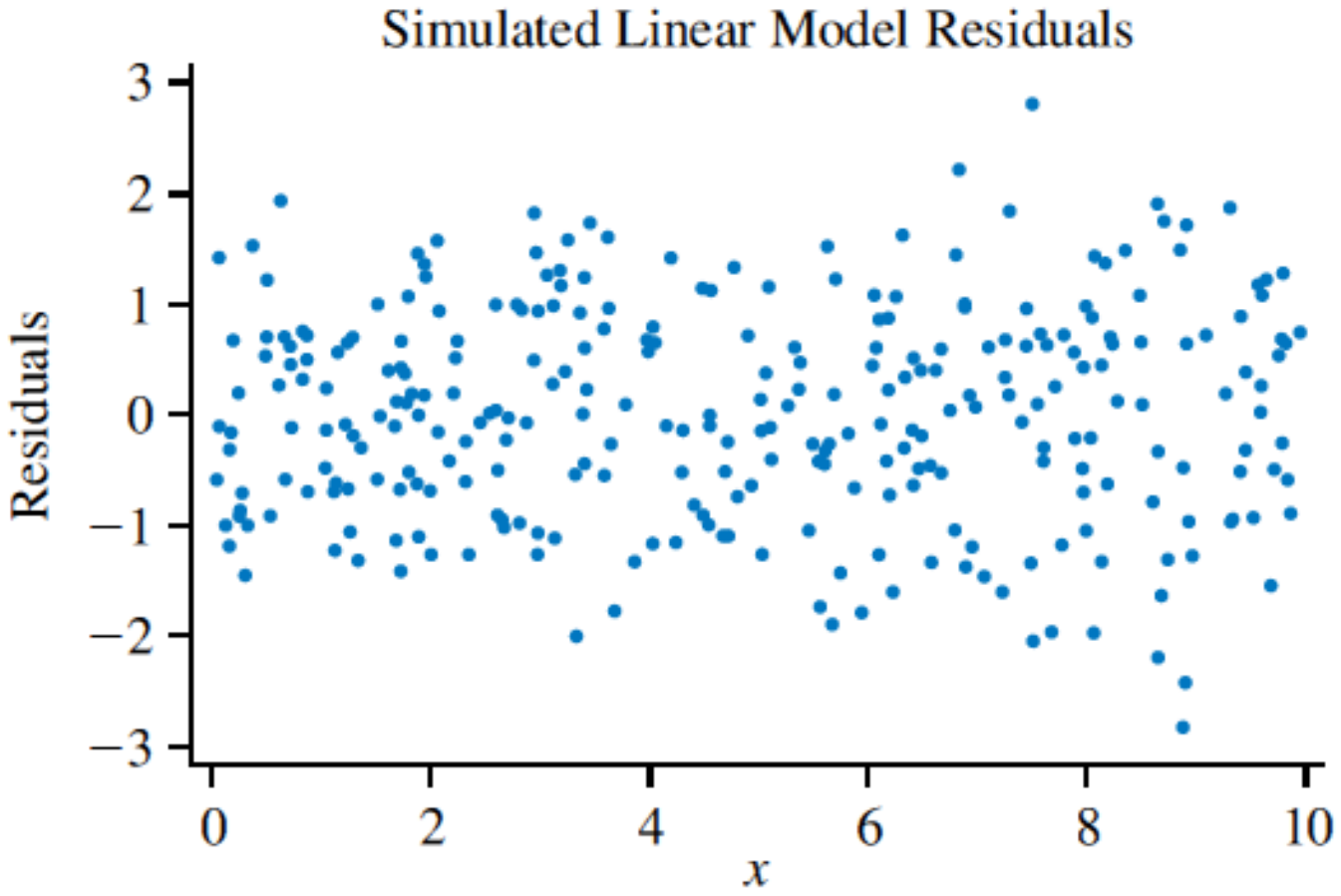
** indicates significant at the 5% level

*** indicates significant at the 1% level

Modelling issues
Using Diagnostic Residual Plots

- When specifying a regression model, we may inadvertently choose an inadequate or incorrect functional form
  1. Examine the regression results
     - There are formal statistical tests to check for:
       - Homoskedasticity
       - Serial correlation
  2. Use residual plots

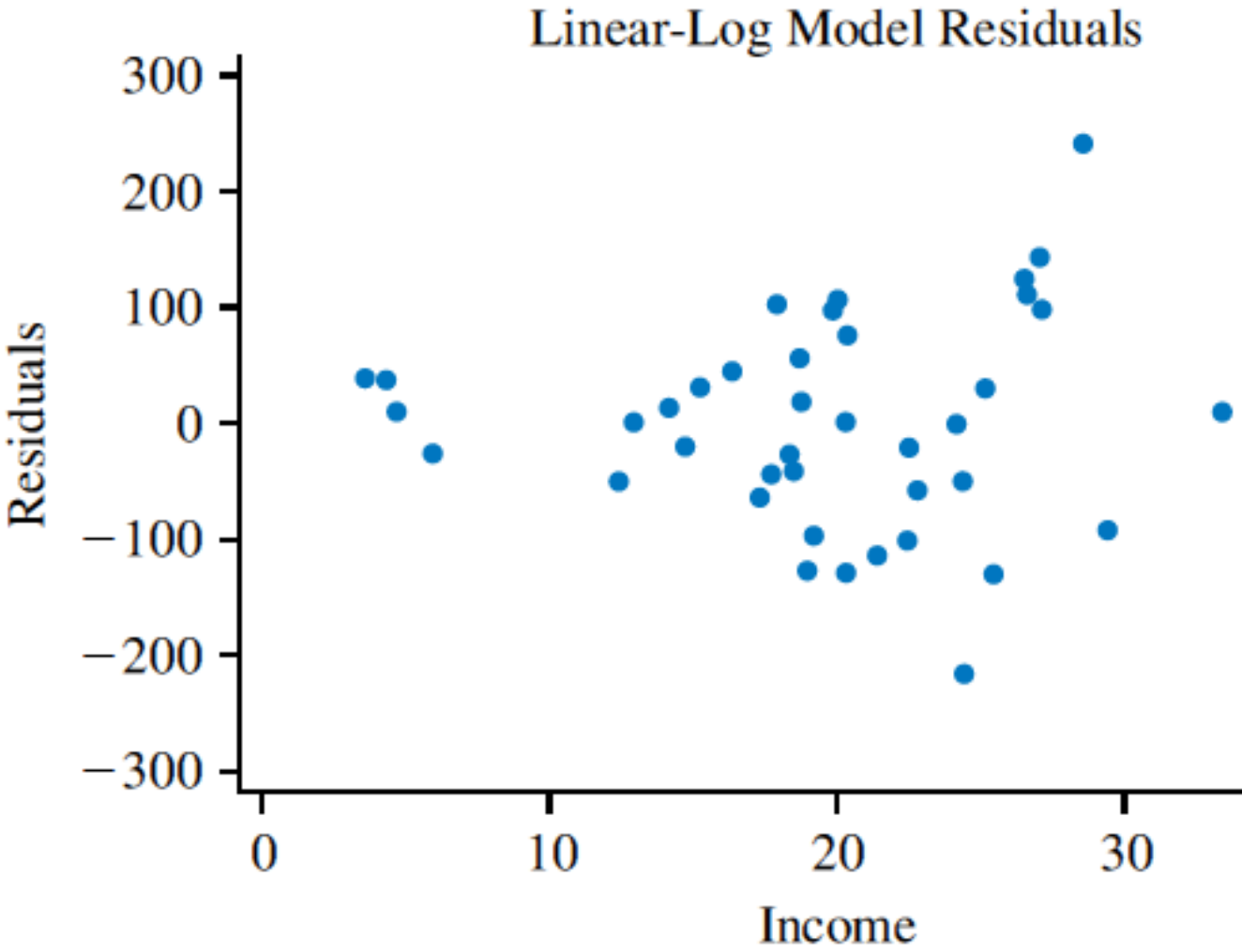
Modelling issues
Using Diagnostic Residual Plots



The **error term** e is uncorrelated, homoskedastic, and from a standard normal distribution, e∼N(0, 1)

Assumption SR3:
The variance of the random error e is: $var(e) = \sigma^2 = var(y)$
The random variables y and e have the same variance because they differ only by a constant.

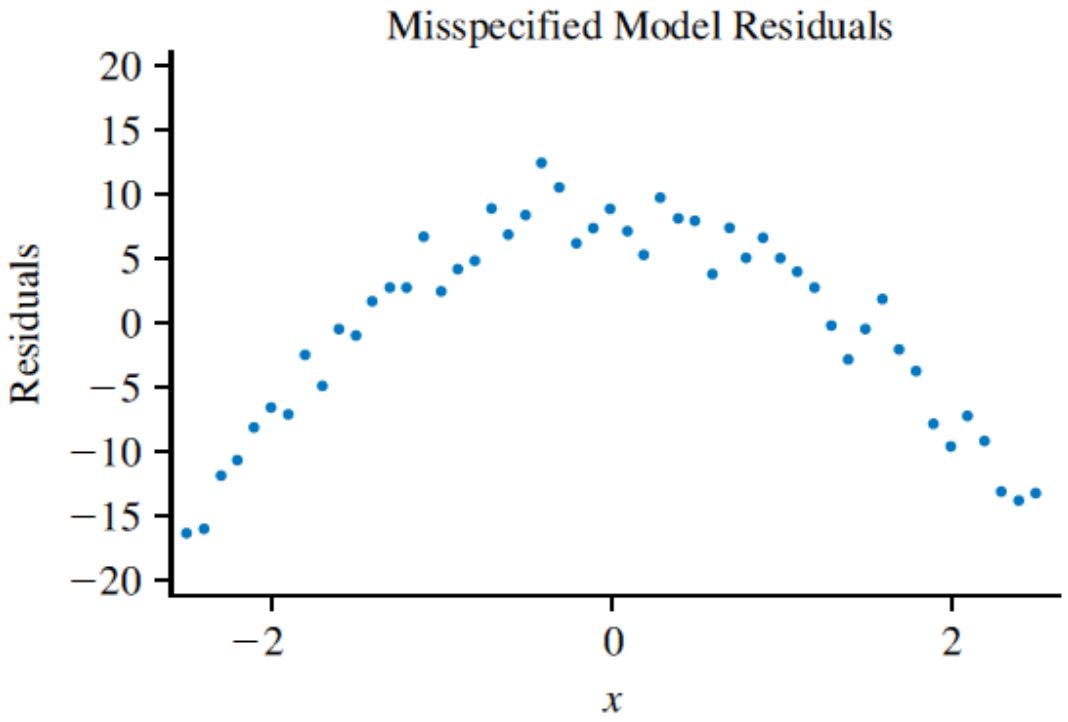
Modelling issues
Using Diagnostic Residual Plots



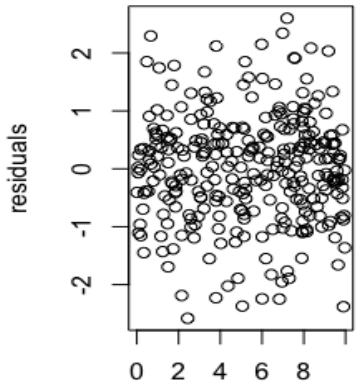
When the residuals are plotted against the independent variable, they show a funnel shape **Homoskedasticity** assumption does **not hold**!

Modelling issues
Using Diagnostic Residual Plots

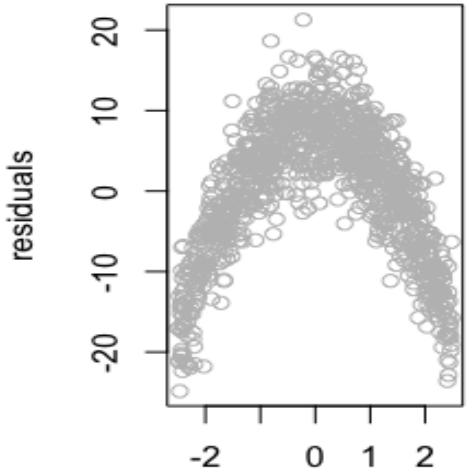
- The well-defined quadratic pattern in the least squares residuals indicates that something is wrong with the linear model specification
  - The linear model has ‘‘missed’’ a curvilinear aspect of the relationship



```
#############################################################################
# Let us draw a residual plot generated with a simulated model that satisfies the regression assumptions:
# EXAMPLE 1 :
set.seed(12345)   #sets the seed for the random number generator
x <- runif(300, 0, 10)
e <- rnorm(300, 0, 1)
y <- 1+x+e
mod3 <- lm(y~x)
ehat <- resid(mod3)
plot(x,ehat, xlab="x", ylab="residuals")
```




```
# EXAMPLE 2:
# The next example illustrates how the residuals look like when a linear functional form is used when
#the true relationship is quadratic. The data generating equation
set.seed(12345)
x <- runif(1000, -2.5, 2.5)
e <- rnorm(1000, 0, 4)
y <- 15-4*x^2+e
mod3 <- lm(y~x)
ehat <- resid(mod3)
ymi <- min(ehat)
yma <- max(ehat)
plot(x, ehat, ylim=c(ymi, yma),
     xlab="x", ylab="residuals",col="grey")
```



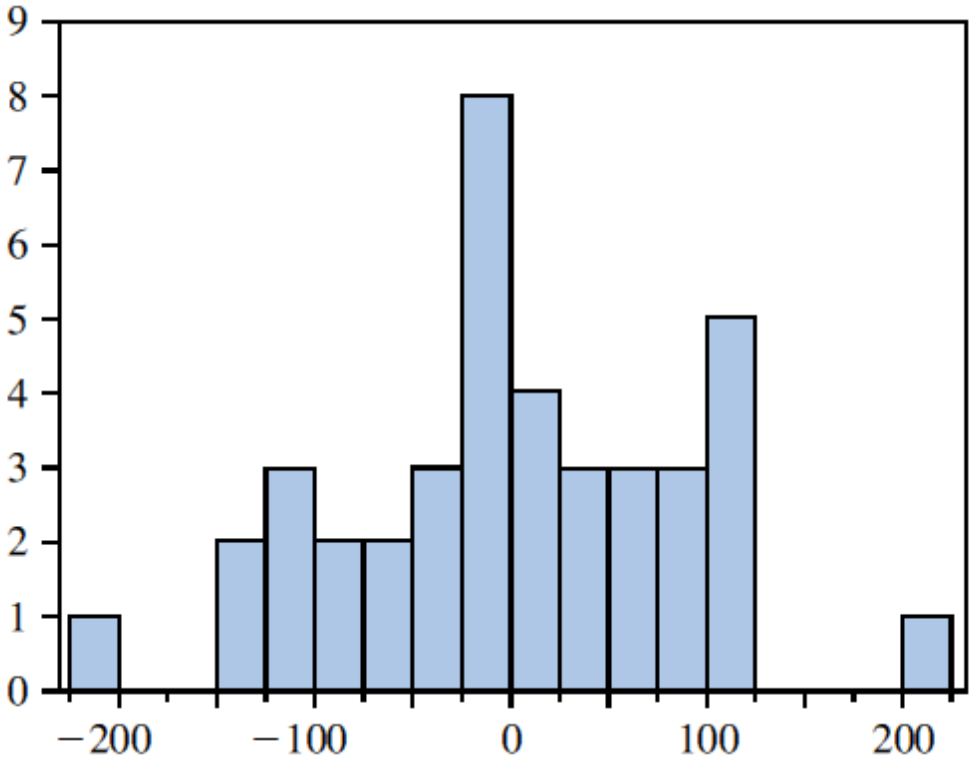

```
# Figure shows the residuals from estimating an incorrectly specified,
#linear econometric model when the correct specification should be quadratic.
```

Modelling issues
Using Diagnostic Residual Plots

- Hypothesis tests and interval estimates for the coefficients rely on the assumption that the errors, and hence the dependent variable *y*, are normally distributed
  - **Are they normally distributed?**

- We can check the **distribution of the residuals** using:
  - A histogram
  - Formal statistical test
    - Merely checking a histogram is not a formal test
    - Many formal tests are available
      - A good one is the **Jarque–Bera test** for normality

Modelling issues
Using Diagnostic Residual Plots



Series: Residuals	
Sample 140	
Observations 40	
Mean	6.93e-15
Median	−6.324473
Maximum	212.0440
Minimum	−223.0255
Std. Dev.	88.36190
Skewness	−0.097319
Kurtosis	2.989034
Jarque-Bera	0.063340
Probability	0.968826

Modelling issues
Using Diagnostic Residual Plots

- The Jarque–Bera test for normality is based on two measures, **skewness and kurtosis**
  - Skewness (S) refers to how symmetric the residuals are around zero
    - Perfectly symmetric residuals will have a skewness of zero
    - The skewness value for the food expenditure residuals is -0.097
  - Kurtosis (K) refers to the ‘‘peakedness’’ of the distribution.
    - For a normal distribution the kurtosis value is 3

The Jarque–Bera statistic is given by: $$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

where

N = sample size

S = skewness

K = kurtosis

Modelling issues
Using Diagnostic Residual Plots

- When the residuals are normally distributed, the Jarque–Bera statistic has a

  chi-squared distribution with two degrees of freedom

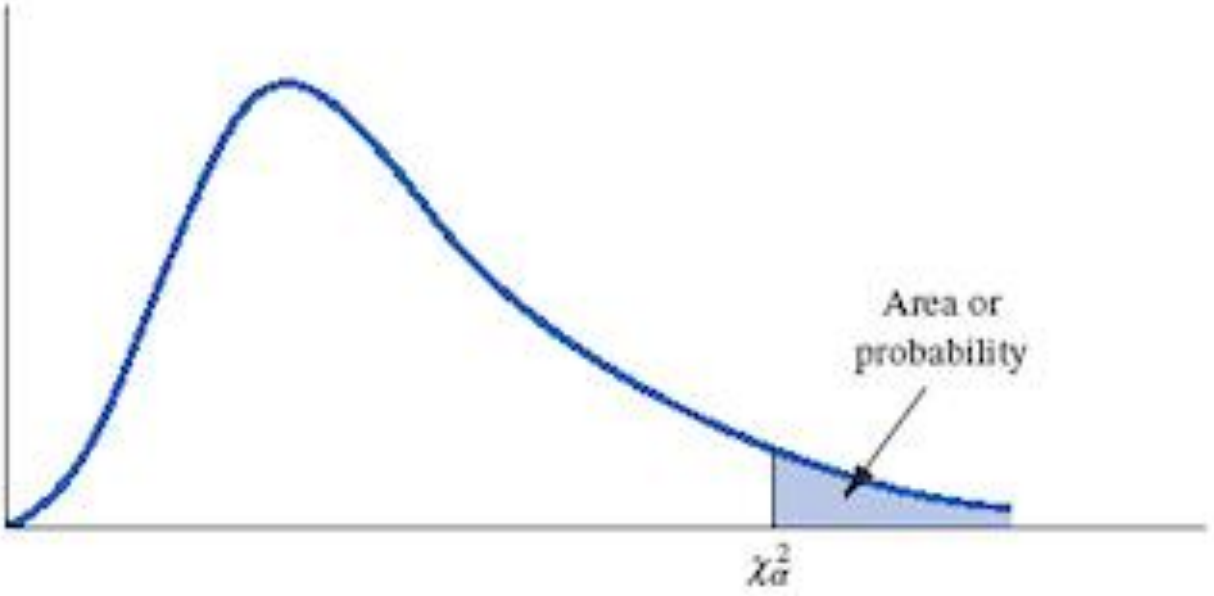
  - We reject the hypothesis of normally distributed errors if a calculated value of the statistic exceeds a critical value selected from the chi-squared distribution with two degrees of freedom
    - The 5% critical value from a χ^2-distribution with two degrees of freedom is 5.99, and the 1% critical value is 9.21

- For the food expenditure example, the Jarque–Bera statistic is:

$$\text{JB} = \frac{40}{6}\left(-0.097^2 + \frac{(2.99-3)^2}{4}\right) = 0.063$$

  - Because 0.063 < 5.99 there is insufficient evidence from the residuals to conclude that the normal distribution assumption is unreasonable at the 5% level of significance
- We could reach the same conclusion by examining the *p*-value
  - The *p*-value described as ‘‘Probability’’
  - Thus, we also fail to reject the null hypothesis on the grounds that 0.9688 > 0.05

TABLE 11.1 SELECTED VALUES FROM THE CHI-SQUARE DISTRIBUTION TABLE*



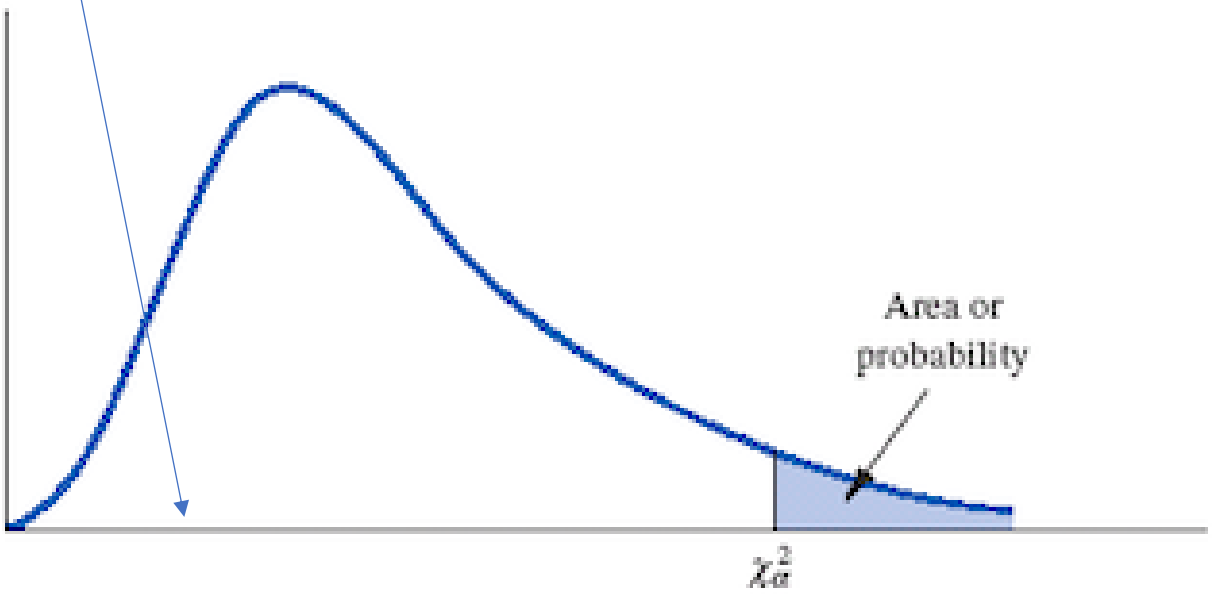
Degrees of Freedom	Area in Upper Tail							
	.99	.975	.95	.90	.10	.05	.025	.01
1	.000	.001	.004	.016	2.706	3.841	5.024	6.635
2	.020	.051	.103	.211	4.605	5.991	7.378	9.210
3	.115	.216	.352	.584	6.251	7.815	9.348	11.345
4	.297	.484	.711	1.064	7.779	9.488	11.143	13.277
5	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086
6	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209

Jarque Bera test

```
> jarque.bera.test(ehat) #(in package 'tseries')

        Jarque Bera Test

data:  ehat
X-squared = 0.06334, df = 2, p-value = 0.9688
```



Accept Hypothesis of normality distribution