

Chapter 10

Examination of Residuals

Overview

- Assumptions on the error terms:
 - Constant variance
 - Normality distributed
 - Independence
- Residual plots
 - Plot residual against sequence number
 - Plot residual against fitted value
 - Plot residual against predictor, x_k
 - Plot residual against other predictor variables not in the model
- Patterns of residual plots

Overview (Continued)

- Test for independence
 - Runs test
 - Durbin-Watson Test for serial correlation, ρ_s
- Test for normality
 - Normal probability plot
 - Kolmogorov-Smirnov Test

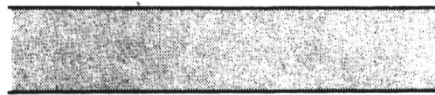
10.1 Introduction

- We assume that the data are represented by the model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$
- The errors ϵ_j 's are assumed to be normally and independently distributed with mean zero and variance σ^2 .
- In order to check the above assumptions on the error terms, ϵ_j 's, we examine the residuals e_j 's where $e_j = y_j - \hat{y}_j$ with $\hat{y}_j = \underline{\hat{\beta}}' \underline{x}_j$

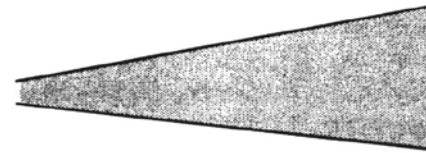
10.2 Residual Plots

- (I) Residual plot in **time order**,
i.e. e_j against j ;
- (II) Plot residual against **predicted value**
i.e. e_j against \hat{y}_j ;
- (III) Plot residual against **predictor x_k**
i.e. e_j against x_{kj} , $k = 1, \dots, p$;
- (IV) Plot residual against **other predictors** that are
not in the regression model.

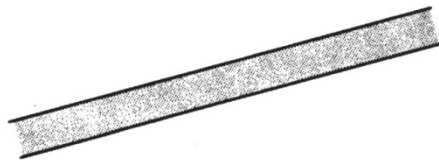
10.2.1 Patterns of Plots



Type 0



Type 1



Type 2



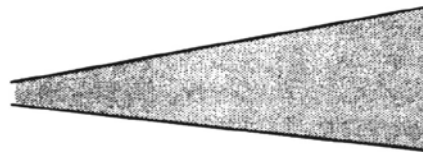
Type 3

10.2.2 Residual plot in Time Order

- **Pattern 0:** If the residual plot does not indicate any pattern, then the assumption of independence is not violated.

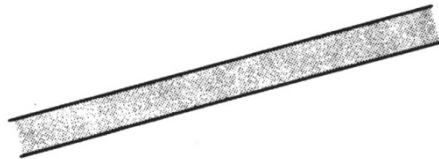


- **Pattern 1:** It implies that the variance is not constant, but increases with time. We may consider to transform the data to stabilize the variance



10.2.2 Residual plot in Time Order

- **Pattern 2:** It implies the model is inappropriate, a linear term in time should be included in the model.

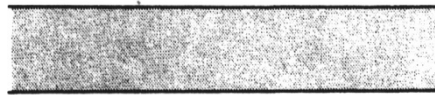


- **Pattern 3:** It implies that linear and quadratic terms in time should be included in the model.

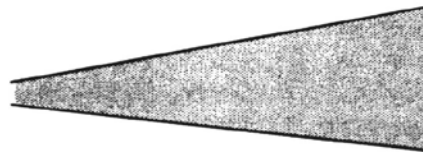


10.2.3 Residuals against Predicted Values

- **Pattern 0:** If the residual plot does not indicate any pattern, then the model seems fine.

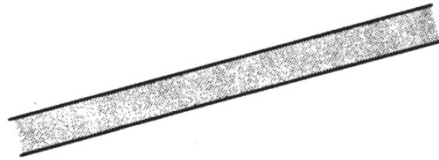


- **Pattern 1:** It implies the variance is not constant. A transformation on the responses, y , is needed.



10.2.3 Residuals against Predicted Values

- **Pattern 2:** It implies there is a mistake in the analysis or a wrongfully omission of the intercept term, β_0 .

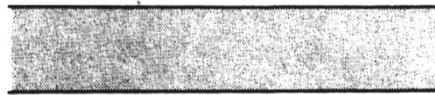


- **Pattern 3:** It implies the model is inadequate. A transformation on y_j may be needed or some other terms should be included in the model.

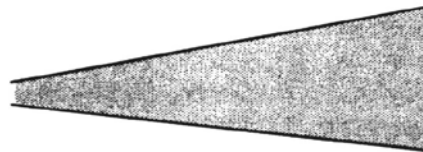


10.2.4 Residuals against predictor x_i

- **Pattern 0:** If the residual plot does not indicate any pattern, then the model seems fine.

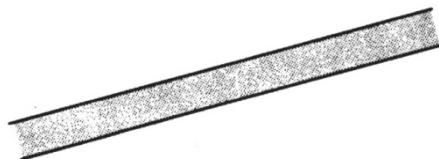


- **Pattern 1:** It implies the variance is not constant. A transformation on the responses, y , or on the predictor x_i , is needed.

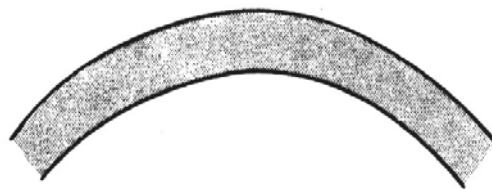


10.2.4 Residuals against predictor x_i

- **Pattern 2:** It implies there is error in calculations. The linear effect of the k -th predictor, X_k not removed.



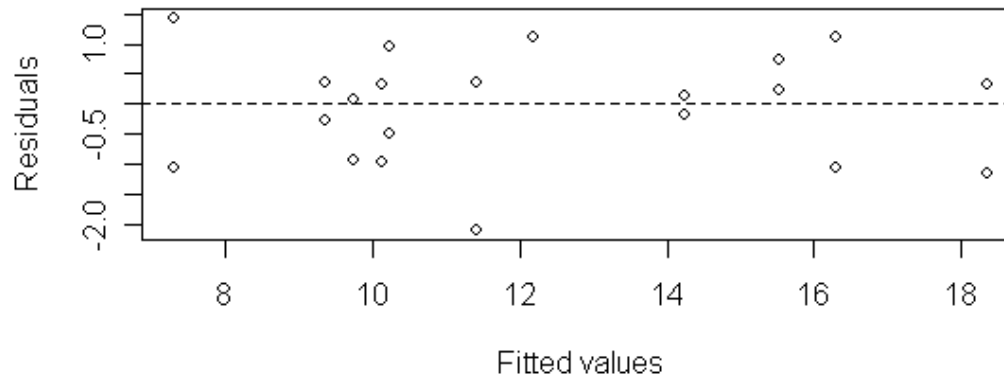
- **Pattern 3:** It implies the model is inadequate. A transformation on y may be needed or some other terms should be included in the model.



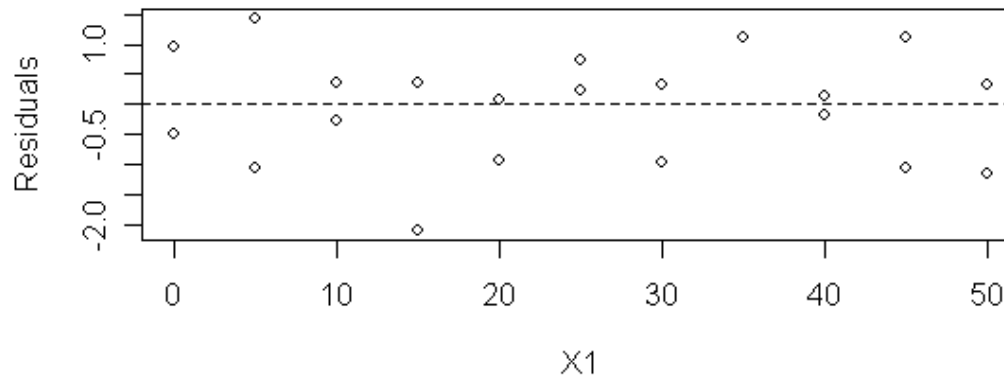
Residual Plots

- Refer to Example 1 in Chapter 4 (p4.20)

Plot of Residuals against the Fitted values



Plot of Residuals against X1



X1
CYM

Residual Plots (Continued)

R program to get the above plots

```
> model1=lm(y~x1+x2)
> res=model1$res
> fv=model1$fit
> par(mfrow=c(2,1))
> #Plot residuals against fitted values
> plot(fv, res, xlab="Fitted values",
      ylab="Residuals", main="Plot of Residuals against
      the Fitted values")
> abline(h=0,lty=2)
> #Plot residuals against the order x1
> plot(x1, res, xlab="X1", ylab="Residuals",
      main="Plot of Residuals against X1")
> abline(h=0,lty=2)
> par(mfrow=c(1,1))
```

Residual Plots (Continued)

SAS program to get the above plots

```

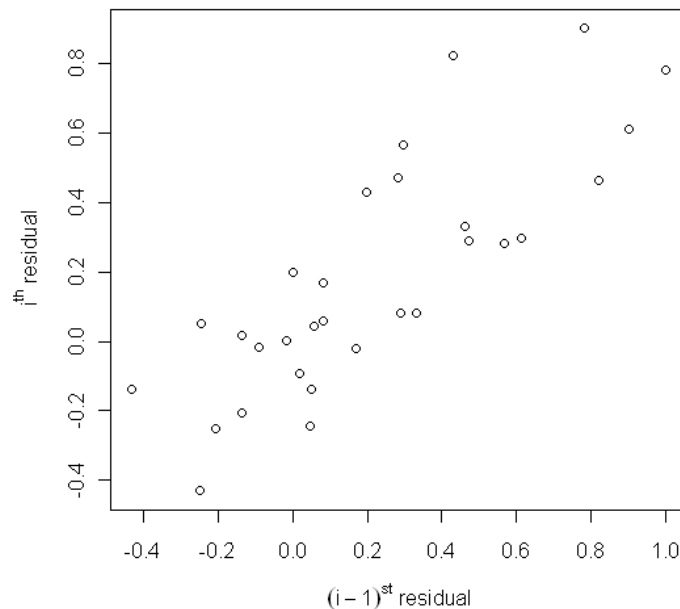
data ch10ex1;
    infile "d:\ST3131\ch10ex1.txt" firstobs=2;
    input y x1 x2;
run;

proc reg data=ch10ex1;
    model y = x1 x2/noprint;
    plot residual.*predicted.
           residual.*x1;
run;
quit;
  
```

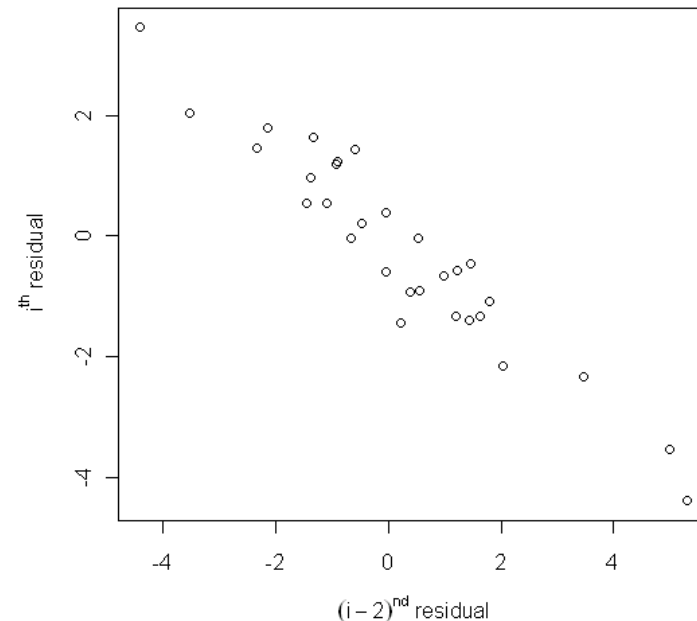
10.3 Serial Correlation

- A lag- s serial correlation is the correlation between observations s steps apart and is denoted by ρ_s .

Positive lag-1 serial correlation presents in the data



Negative lag-2 serial correlation presents in the data



- Two methods to check the presence of serial correlation

10.3.1 Runs Test

- Let U = number of runs in a sequence of residuals that has n_1 positive signs and n_2 negative signs.

- For example

$=$ $+$ $+$ $=$ $+$ $=$ $+$ $=$ $-$ $-$ $-$ $-$ $+$ $+$ $+$ $=$ $+$ $+$ $=$ $+$ $=$ $-$ $-$ $+$ $+$
 $-$ $-$ $+$ $=$ $+$ $+$ $+$

Number of runs = $r = 18$.

Number of positive sign = $n_1 = 16$ and

number of negative sign = $n_2 = 14$.

Runs Test (Continued)

- When $n_1 > 10$ and $n_2 > 10$, then

U approximately $\sim N(\mu_U, \sigma_U^2)$

where

$$\mu_U = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\sigma_U^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

Runs Test (Continued)

- Consider the test statistic

$$Z = \frac{U - \mu_U}{\sigma_U}$$

- Test H_0 : no run pattern against
 - (a) H_1 : too many or too few runs (two sided test)
 - (b) H_1' : too many runs (one sided test)
 - (c) H_1'' : too few runs (one sided test)
- Reject H_0 if
 - (a) $|Z| > z_{\alpha/2}$
 - (b) $Z > z_{\alpha}$
 - (c) $Z < -z_{\alpha}$

10.3.2 The Durbin-Watson Test

- Consider the model

$$y_j = \beta_0 + \sum_{i=1}^k \beta_i x_{ij} + \epsilon_j, \quad j = 1, \dots, n$$

- If $\epsilon_j \sim N(0, \sigma^2)$ independently, then $\rho_s = 0$ for all s .

The Durbin-Watson Test (Continued)

- Test $H_0: \rho_s = 0$ for all s against $H_1: \rho_s = \rho^s$.
- The alternative hypothesis arises from the assumption that

$$\epsilon_j = \rho \epsilon_{j-1} + z_j,$$

where $z_j \sim N(0, \sigma^2)$ and is independent of $\epsilon_{j-1}, \epsilon_{j-2}, \dots$, and of z_{j-1}, z_{j-2}, \dots

The Durbin-Watson Test (Continued)

- We also assume that $E(\epsilon_j)$ and $V(\epsilon_j)$ are constant.
- Therefore $\epsilon_i \sim N(0, \sigma^2 / (1 - \rho^2))$.
- Under H_0 , $\epsilon_i \sim N(0, \sigma^2)$
- The Durbin-Watson statistic is defined as

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

The Durbin-Watson Test (Continued)

- The value of the Durbin-Watson statistic **ranges from 0 to 4**.
- As a general rule of thumb, the residuals are not correlated if the DW statistic is approximately 2, and an acceptable range for the DW statistic is 1.50 to 2.50.

The Durbin-Watson Test (Continued)

- For a fixed significance level α , a pair of critical values are used for the test procedure
- The detail tables of the critical values are given in the paper "Testing for serial correlation in least squares regression II" in Biometrika (1951) p159-178 written by Durbin and Watson.
- The critical values (d_L, d_U) for $\alpha = 0.05$ and 0.025 are given on p10-25 and p10-26.

The Durbin-Watson Test (Continued)

Critical Values (d_L , d_U) at 5%

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
20	1.2	1.41	1.1	1.54	1	1.68	0.9	1.83	0.79	1.99
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
35	1.4	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.8
40	1.44	1.54	1.39	1.6	1.34	1.66	1.29	1.72	1.23	1.79
50	1.5	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

n : the number of observations

k : the number of predictors

The Durbin-Watson Test (Continued)

Critical Values (d_L , d_U) at 2.5%

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	0.95	1.23	0.83	1.4	0.71	1.61	0.59	1.84	0.48	2.09
20	1.08	1.28	0.99	1.41	0.89	1.55	0.79	1.7	0.7	1.87
25	1.18	1.34	1.1	1.43	1.02	1.54	0.94	1.65	0.86	1.77
30	1.25	1.38	1.18	1.46	1.12	1.54	1.05	1.63	0.98	1.73
35	1.3	1.42	1.25	1.48	1.19	1.55	1.33	1.63	1.07	1.7
40	1.34	1.45	1.3	1.51	1.25	1.57	1.2	1.63	1.21	1.69
50	1.42	1.5	1.38	1.54	1.34	1.59	1.3	1.64	1.3	1.69
60	1.47	1.54	1.44	1.57	1.4	1.61	1.37	1.65	1.33	1.69
80	1.54	1.59	1.52	1.62	1.49	1.65	1.47	1.67	1.44	1.7
100	1.59	1.63	1.57	1.65	1.55	1.67	1.53	1.7	1.51	1.72

n : the number of observations

k : the number of predictors

The Durbin-Watson Test (Continued)

The test procedures are given as follows.

(a) 1-sided test ($\rho > 0$)

- If $d < d_L$, we conclude d is significant and **reject** H_0 at the α significance level.
- If $d > d_U$, we conclude d is not significant and **do not reject** H_0 .
- If $d_L \leq d \leq d_U$, the test is said to be **inconclusive**.

(b) 1-sided test ($\rho < 0$)

- Repeat (a) using $(4 - d)$ in place of d .

The Durbin-Watson Test (Continued)

(c) 2-sided test ($\rho \neq 0$)

- If $d < d_L$ or $4 - d < d_L$, we conclude that d is significant and reject H_0 at the 2α significance level.
- If $d > d_U$ and $4 - d > d_U$, we conclude d is not significant and do not reject H_0 .
- Otherwise, the test is said to be inconclusive.

10.3.3 Example 1

- Refer to Example 1 in Chapter 4 (p 4.20)
- From the data, we obtain the regression equation

$$\hat{y} = 5.2573824 + 0.1621127x_1 + 0.2488677x_2$$

- Hence we have the following residuals:

i	1	2	3	4	5	6
e_i	-0.504734	0.95526	1.43772	-1.06228	-0.26719	0.34281
i	7	8	9	10	11	12
e_i	-2.11209	0.34791	-0.92397	0.07603	0.74811	0.23811
i	13	14	15	16	17	18
e_i	0.31924	-0.98076	1.11433	1.12433	-0.18057	0.12943
i	19	20	21	22		
e_i	-1.07547	1.12453	0.3196	-1.17037		

Example 1 (Continued)

Runs Test

- Runs pattern: $- \underline{++} - \underline{+} - \underline{+} - \underline{++++} - \underline{++} - \underline{+} - \underline{++} -$
- No. of runs: $U = 15$
- No. of positive sign: $n_1 = 13$
- No. of negative sign: $n_2 = 9$
- Hence $\mu_U = 2n_1n_2/(n_1 + n_2) = 11.64$ and
 $\sigma^2 = 2n_1n_2(2n_1n_2 - n_1 - n_2)/[(n_1 + n_2)^2(n_1 + n_2 - 1)] = 4.88$
- Therefore U approximately $\sim N(11.64, 4.88)$

Example 1 (Continued)

- Test H_0 : no run pattern against
 H_1 : too many or too few runs (two sided test)

- From the data, we have

$$Z = (13 - 11.64)/4.88^{0.5} = 1.521$$

- Since $Z_{\text{obs}} = 1.521 < z_{0.025} = 1.96$ (or p-value = 0.1282), therefore we do not reject H_0 and conclude that no serial correlation exists in the data.

Example 1 (Continued)

Durbin-Watson Test

- From the data, we have

$$d = \frac{\sum_{i=2}^{22} (e_i - e_{i-1})^2}{\sum_{i=1}^{22} e_i^2} = 2.3216$$

- Suppose we want to test $H_0: \rho = 0$ against $H_1: \rho \neq 0$.
- We have $n = 22, k = 2$.
- Hence $d_U = 1.42$ at the 0.025 sig. level.
- Since $d > d_U$ and $4 - d > d_U$, therefore we do not reject H_0 and conclude that no serial correlation exists.

10.3.4 Programs

SAS Program for Durbin-Watson Statistic

```
proc reg data=ch10ex1;
  model y = x1 x2/dwProb;
run;
```

Partial Output

Durbin-Watson D

2.322

Pr < DW

0.6613

Pr > DW

0.3387

Number of Observations

22

1st Order Autocorrelation

-0.206

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value testing negative autocorrelation

Programs (Continued)

R Program for Runs Test

```
> ch10ex1=read.table("d:/ST3131/ch4ex1.txt",header=T)
> attach(ch10ex1)

> model1=lm(y~x1+x2)

> library(tseries) # A package for analysis of time
  series data
> res <- model1$res
> runs.test(factor(sign(res)))
```

Runs Test

```
data:  factor(sign(res))
Standard Normal = 1.5225, p-value = 0.1279
alternative hypothesis: two.sided
```

Programs (Continued)

R Program for Durbin-Watson Test

```
> model1=lm(y~x1+x2)
> library(car)
> durbinWatsonTest(model1)
```

lag	Autocorrelation	D-W Statistic	p-value
1	-0.2056157	2.321586	0.69

Alternative hypothesis: $\rho \neq 0$

The p-value is obtained using resampling methods such as bootstrapping. Hence it may vary.

10.4 Tests for Normality Assumption

10.4.1 Normal Probability Plot

- Let $Z_{(1)} < \dots < Z_{(n)}$ be the ordered values of n independent and identically distributed $N(0, 1)$ random variables Z_1, \dots, Z_n . Then

$$E(Z_{(i)}) \approx \gamma_i = \Phi^{-1} \left[\frac{i - 3/8}{n + 1/4} \right] \quad (\text{Eq. 10.4.1})$$

where Φ^{-1} is the inverse standard normal cumulative distribution function.

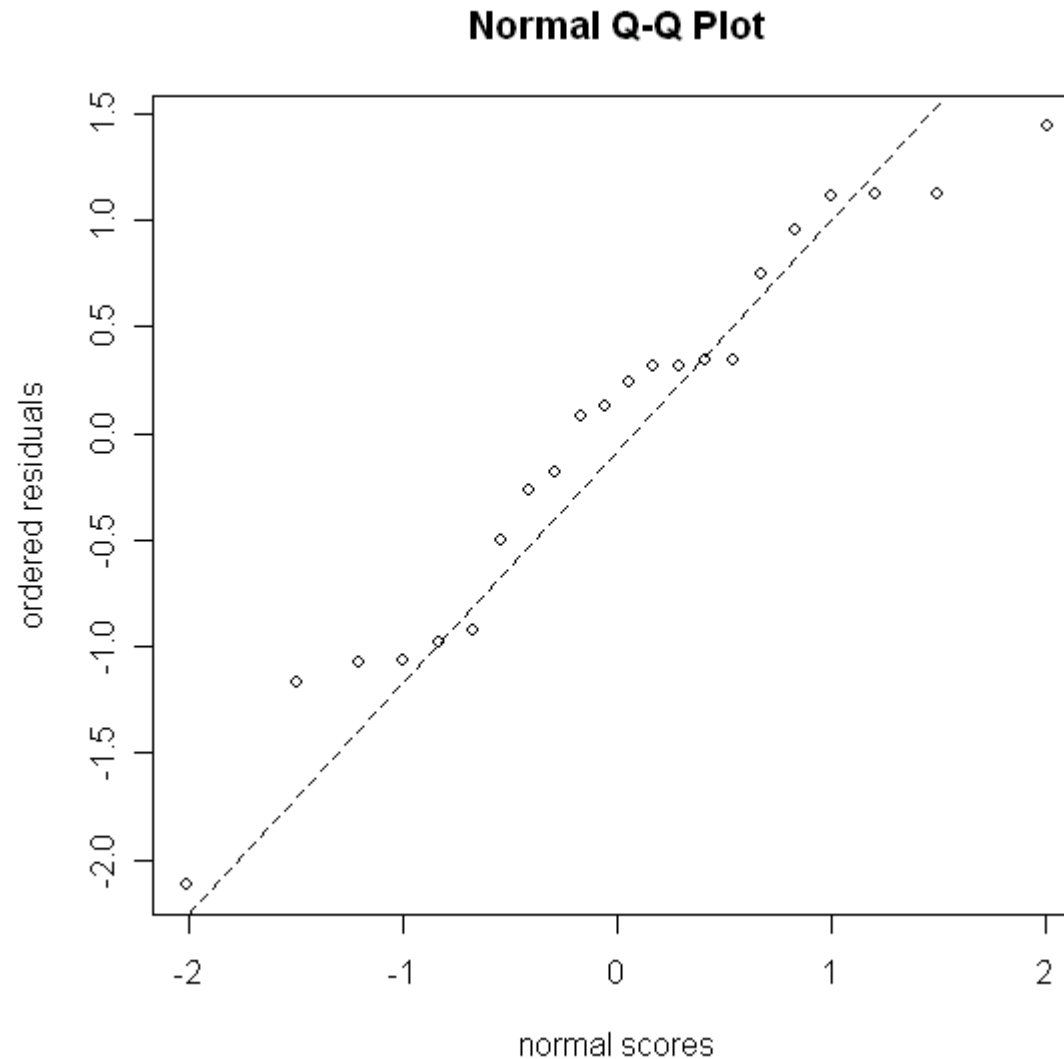
Normality Probability Plot (Continued)

- Let $e_{(1)} < \dots < e_{(n)}$ denote the ordered values of e_i 's.
- A plot of $e_{(i)}$'s against γ_i 's in Eq. 10.4.1 should follow a straight line if the error terms ϵ 's follow a normal distribution.

Example 1 (Continued)

- R program to get the normal probability plot
- ```
> qqnorm(res, xlab="normal scores",
 ylab="ordered residuals")
> qqline(res, lty=2)
```

# Normal Probability Plot (Continued)

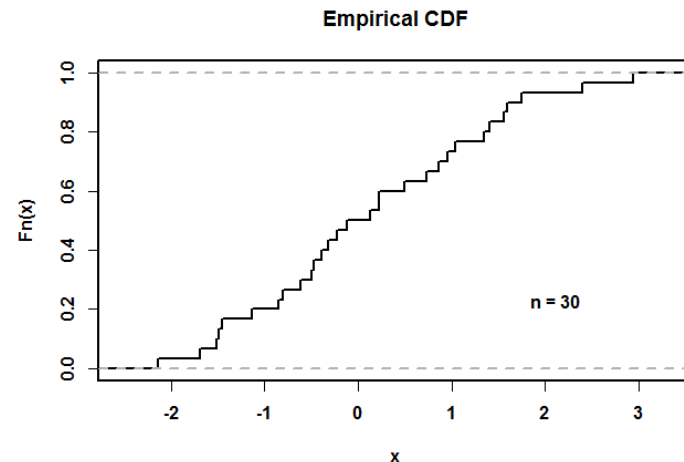


## 10.4.2 Kolmogorov-Smirnov Test

- Kolmogorov-Smirnov Test can be used to test for normality.
- Let  $F_n(x)$  be the empirical distribution of the  $u_i$ 's
- That is,

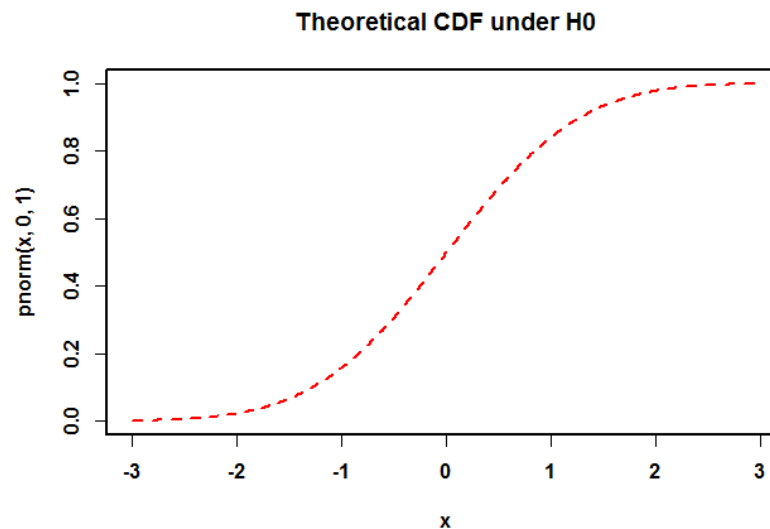
$$F_n(x) = \frac{u_x}{n}$$

where  $u_x$  is the number of  $u_i$ 's that are not greater than  $x$ .



# Kolmogorov-Smirnov Test (Continued)

- Let  $F_H(x)$  be the cumulative distribution function of  $F$  under  $H_0$



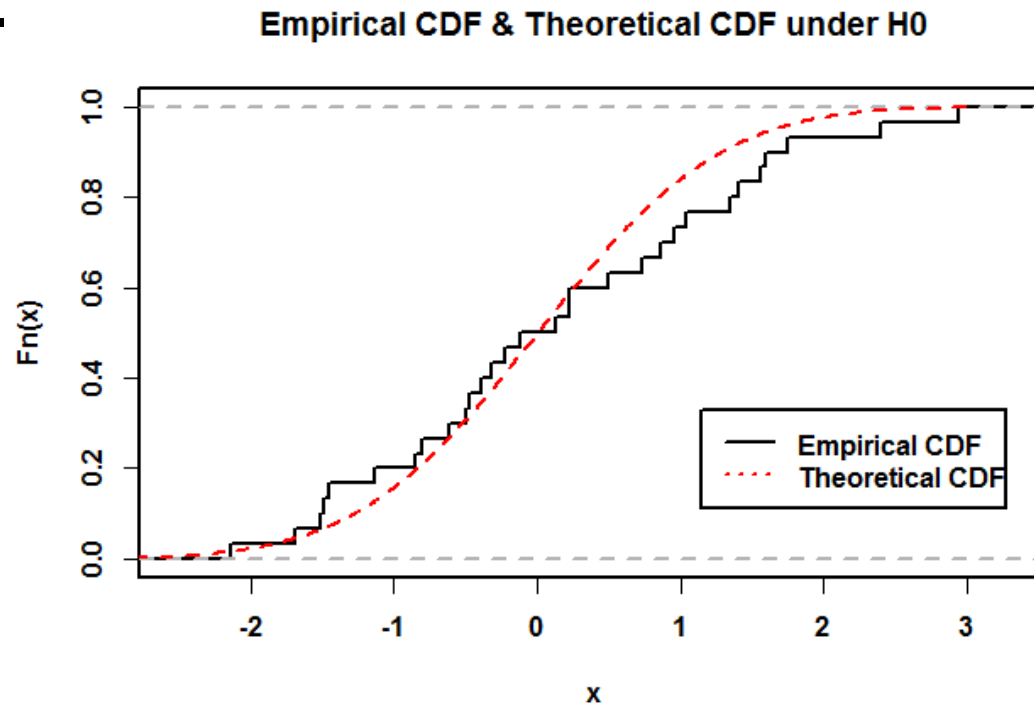


# Kolmogorov-Smirnov Test (Continued)

- Kolmogorov-Smirnov statistic is given by

$$D = \sup_x |F_n(x) - F_H(x)|$$

where  $F_H$  is the cumulative distribution function of  $F$  under  $H_0$ .



# Kolmogorov-Smirnov Test (Continued)

- The hypothesis that the  $u_i$ 's follow the distribution given by  $F_H$  is rejected for large values of  $D$ .
- In our case,  $F_H$  is the c.d.f. of  $N(\mu, \sigma^2)$ .
- The value of  $D$  and the corresponding p-value are usually available for many statistical packages.

# Kolmogorov-Smirnov Test (Continued)

- R program for Example 1

```
> res=model1$res
> ks.test(res,"pnorm",mean(res),sd(res))
 One-sample Kolmogorov-Smirnov test
data: res
D = 0.1235, p-value = 0.8502
alternative hypothesis: two-sided
```

# SAS Program for Testing Normality

```
* Obtain the residuals;
proc glm data=ch10ex1;
 model y = x1 x2;
 output out=ch10ex1out p=yhat r=res;

*Test for normality;
proc univariate data=ch10ex1out;
 var res;
 histogram res/ normal noplot;
 qqplot res;
run;

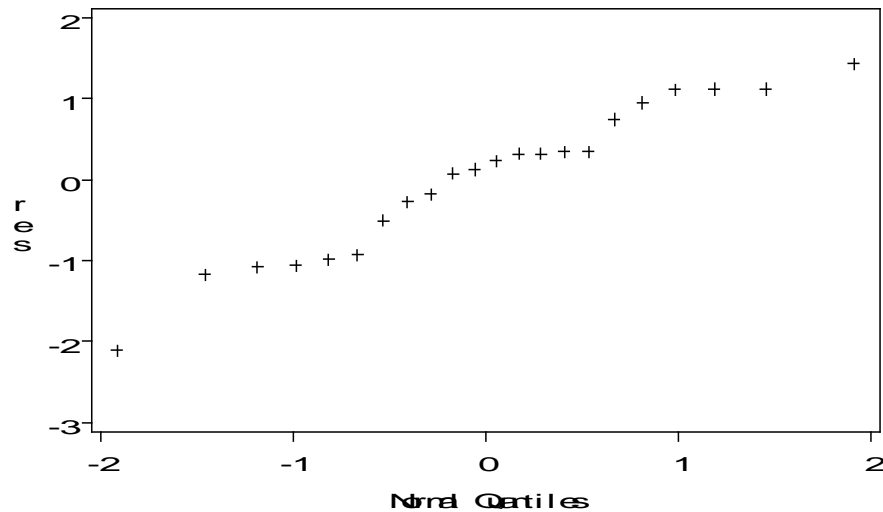
*Residual plots;
proc gplot data=ch10ex1out;
 title "Residual Plots";
 plot res*yhat;
run;
quit;
```

# Partial SAS Printout

## Goodness-of-Fit Tests for Normal Distribution

| Test               | ---Statistic--- |            | ----p Value---- |        |
|--------------------|-----------------|------------|-----------------|--------|
| Kolmogorov-Smirnov | D               | 0.12352227 | Pr > D          | >0.150 |
| Cramer-von Mis     | W-Sq            | 0.06203    | Pr > W-Sq       | >0.250 |
| Anderson-Darling   | A-Sq            | 0.39731064 | Pr > A-Sq       | >0.250 |

Residual Plots



Residual Plots

