Chapter 4

Lack of Fit Test

Overview

- Pure error sum of squares, SSPE
- Lack of fit sum of squares, SSLF
- Error sum of squares, SSE = SSLF + SSPE
- Repeated measurements
- Lack of fit test
- Use of software to do a lack of fit test

4.1 Introduction

- F-test for the significance of the model only tests if a model with at least one predictors is better than a model without any predictor

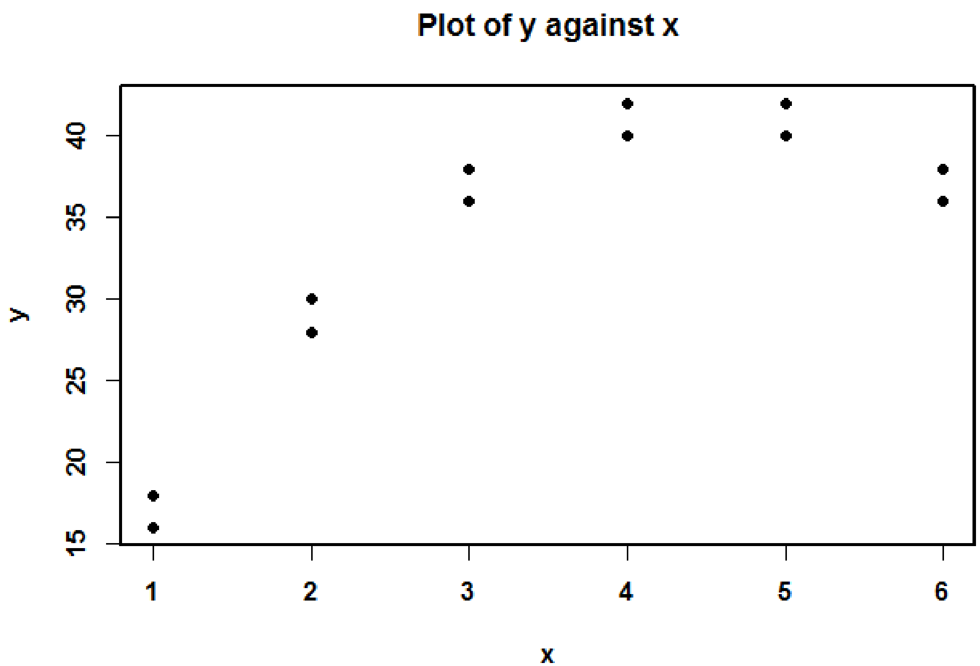
- While the partial F-test only test if some of the predictors contributing to the model that has already included other predictors

- Neither of these 2 tests tells us whether the regression is appropriate or not

4.1 Introduction (Continued)

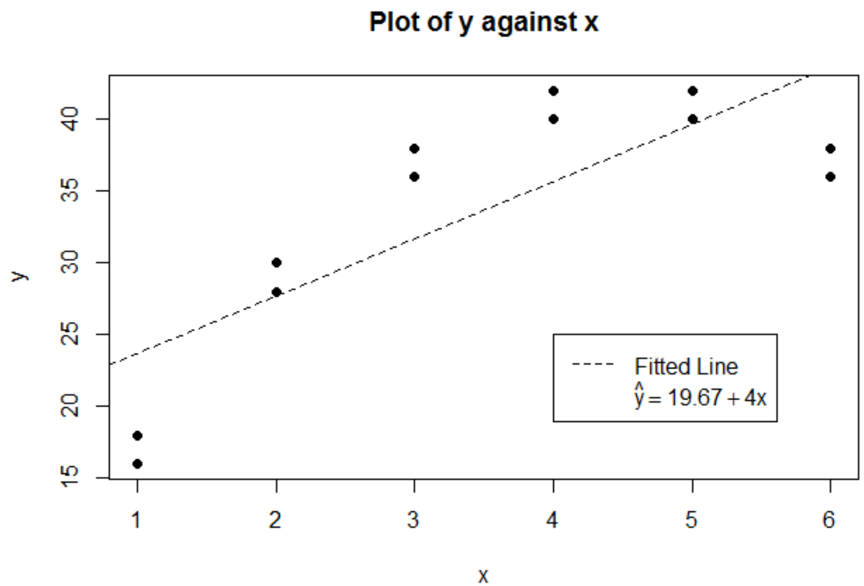
- Consider the following data set

x	1	2	3	4	5	6	1	2	3	4	5	6
y	18	30	38	42	42	38	16	28	36	40	40	36



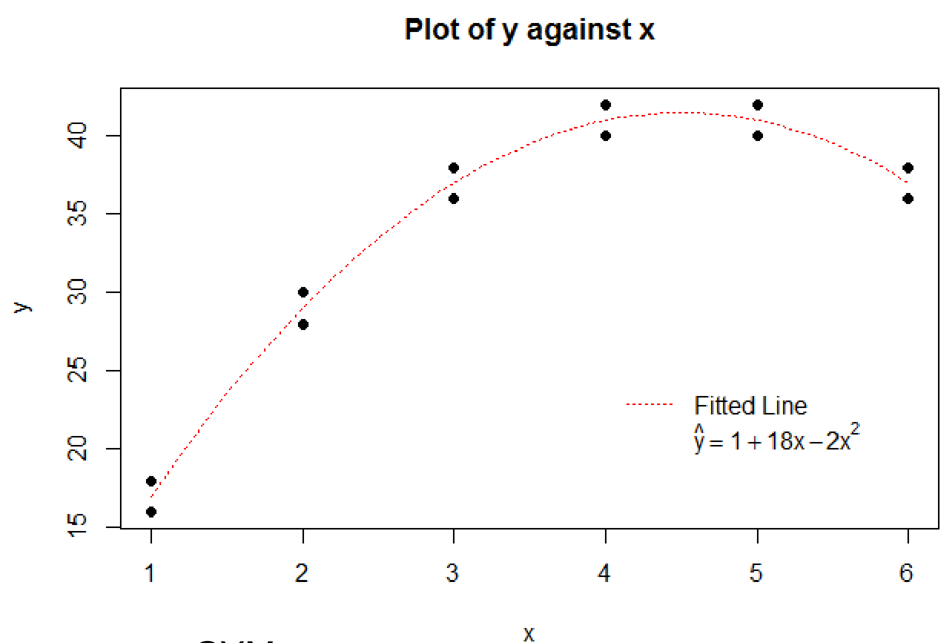
4.1 Introduction (Continued)

- Fitting a simple regression model $y = \beta_0 + \beta_1 x + \epsilon$ to the data gives the following results
- The fitted model is $\hat{y} = 19.67 + 4\,x$
- F_{obs} = 18.03, p-value = 0.002
- Hence the simple regression model is significant.



4.1 Introduction (Continued)

- Question:
  - Is the simple regression model appropriate?
  - Is it possible to get a better model?
- We may try to fit a quadratic polynomial model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ to the data



4.1 Introduction (Continued)

- The fitted model is $\hat{y} = 1 + 18x - 2x^2$
- F_{obs} = 322, p-value = $4.23(10)^{-9}$
- Hence the quadratic polynomial model is significant.
- Partial F-tests are significant
  - $F_{obs} = SSR(x^2 \mid x)/MSE$ = 224 with a p-value = $1.15(10)^{-7}$
  - $F_{obs} = SSR(x \mid x^2)/MSE$ = 354.84 with a p-value = $1.54(10)^{-8}$
- Hence both x and x^2 terms contributing significantly to the model

4.1 Introduction (Continued)

- Question:
  - Is the quadratic polynomial model appropriate?
  - Is it possible to get a better model?
- Answer:
  - Perform a lack of fit test if there are repeated measurements

4.2 SS Pure Error and SS LOF

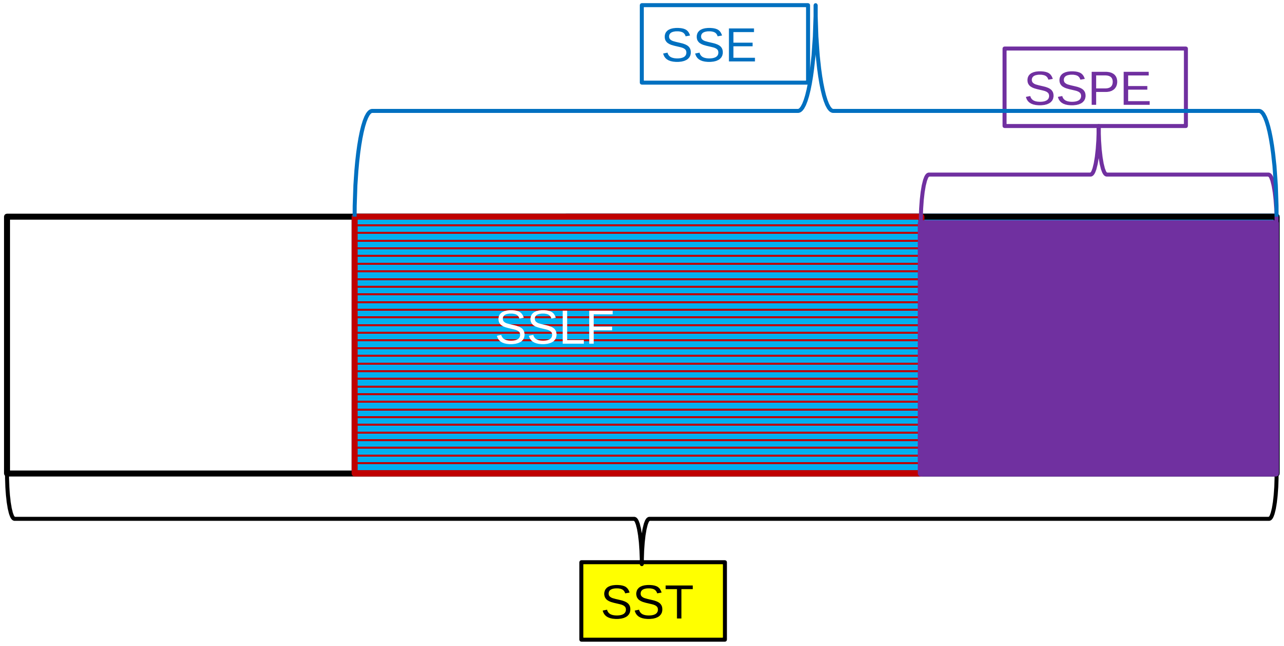
- To test for the appropriateness of a particular multiple regression model, we perform a lack of fit test
- To test for lack of fit, we need to have some independent repeated measurements of y
- Example

y	9.73	11.19	8.75	6.25	9.10	9.71	8.5
x_1	0	0	5	5	10	10	5
x_2	20	20	5	5	10	10	10

  – 9.73 and 11.19 are 2 repeated measurements of y for $x_1 = 0$ and $x_2 = 20$

SS Pure Error and SS LOF (Continued)

- *Error Sum of Squares, SSE*, can be decomposed into 2 components, **sum of squares pure error** (*SSPE*) and **sum of squares due to lack of fit** (*SSLF*).



SS Pure Error and SS LOF (Continued)

- SSE measures the variability of y which cannot be explained by the given model.
- The pure error component, SSPE measures the inherent variability of y which cannot be explained by **ANY** model.
- The lack of fit component, SSLF, represents the variability of y that cannot be explained by the given model and may be reduced if a “better” model is used.
- That is,

$$SSE = SSPE + SSLF.$$

4.3 Repeated Measurements

- Suppose there are m groups of repeated measurements each has n_j, $j = 1, \dots, m$ observations.

- Repeated measurements are the measurements taken at the same combination of levels of $x_1, \dots, x_p$.

- Example (Continued)

y	9.73	11.19	8.75	6.25	9.10	9.71	8.5
x_1	0	0	5	5	10	10	5
x_2	20	20	5	5	10	10	10

4.4 Pure Error Sum of Squares, SSPE

Definition

$$SSPE = \sum_{j=1}^{m} \sum_{k=1}^{n_j} \left(y_{jk} - \bar{y}_j\right)^2$$

where $\bar{y}_j$ is the mean of y's for the j-th combination of levels of $x_1, \ldots, x_p$, which has n_j repeated measurements.

- $SSPE$ has $\sum_{j=1}^{m}(n_j - 1)$ degrees of freedom, where m is the number of levels of $x_1, \ldots, x_p$ that have repeated measurements.

4.5 Lack of Fit Sum of Squares, SSLF

- SS Lack of fit = the difference between SS Error and SS Pure Error. i.e.

$$SSLF = SSE - SSPE$$

- The degrees of freedom of SSLF is

d.f. for SSE

d.f. for SSPE

$$\boxed{n - (p+1)} - \boxed{\sum_{j=1}^{m} (n_j - 1)}$$

4.6 Lack of Fit Test

- Test H_0: There is no lack of fit in the model, against H_1: There is lack of fit.

Let

$$F_{LOF} = \frac{MSLF}{MSPE}$$

where $MSLF = \frac{SSLF}{n-(p+1)-\sum_{j=1}^{m}(n_j-1)}$

and $MSPE = \frac{SSPE}{\sum_{j=1}^{m}(n_j-1)}$

Let a be the d.f. of $SSPE$, i.e. $a = \sum_{j=1}^{m}(n_j - 1)$

Lack of Fit Test (Continued)

- It can be shown that under H_0

$$F_{LOF} \sim F(n - (p + 1) - a\ , a).$$

- Reject H_0 at the α level of significance if

$$F_{LOF,\,obs} > F_\alpha(n - (p + 1) - a\ , a)$$

Lack of Fit Test (Continued)

Remarks:

- If F_{LOF} is significant, then we should look for an alternate model for the relationship between y and the x's.

- If F_{LOF} is not significant, then it is not necessary to find a more complicated model.

- However, this fact does not ensure that the given model is a useful model for the purpose of prediction.

4.7 Example 1

- The marketing department for a large manufacturer of electronic games would like to measure the effectiveness of different types of advertising media in promotion of its products.

- Specifically, two types of media are to be considered: radio and television advertising, and newspaper advertising (including the cost of discount coupons).

Example 1 (Continued)

- A sample of 22 cities with approximately equal populations is selected for the study during a test period of 1 month. Each city is to allocate a specific expenditure level for both types of advertising.

- The sales for electronic games during the test month are recorded in the following table.

Example 1 (Continued)

City	1	2	3	4	5	6	7	8
y	9.73	11.19	8.75	6.25	9.10	9.71	9.31	11.77
x_1	0	0	5	5	10	10	15	15
x_2	20	20	5	5	10	10	15	15

Repeated Measurements

City	9	10	11	12	13	14	15
y	8.82	9.82	16.28	15.77	10.44	9.14	13.29
x_1	20	20	25	25	30	30	35
x_2	5	5	25	25	0	0	5

City	16	17	18	19	20	21	22
y	13.30	14.05	14.36	15.21	17.41	18.66	17.17
x_1	35	40	40	45	45	50	50
x_2	5	10	10	15	15	20	20

y: sales in million dollars

x_1: radio and TV advertising ($000)

x_2: newspaper advertising ($000)

Example 1 (Continued)

(x_1, x_2)	$\sum_{k=1}^{n_j}\left(y_{jk} - \bar{y}_j\right)^2 = \sum_{k=1}^{n_j} y_{jk}^2 - n\bar{y}_j^2$	df
(0, 20)	$9.73^2 + 11.19^2 – 2(10.46)^2 = 1.0658$	1
(5, 5)	$8.75^2 + 6.25^2 – 2(7.5)^2 = 3.125$	1
(10, 10)	$9.10^2 + 9.71^2 – 2(9.045)^2 = 0.18605$	1
(15, 15)	$9.31^2 + 11.77^2 – 2(10.54)^2 = 3.0258$	1
(20, 5)	$8.82^2 + 9.82^2 – 2(9.32)^2 = 0.50$	1
(25, 25)	$16.28^2 +15.77^2 – 2(16.025)^2 = 0.13005$	1
(30, 0)	$10.44^2 + 9.14^2 – 2(9.79)^2 = 0.845$	1
(35, 5)	$13.29^2 + 13.30^2 – 2(13.295)^2 = 0.00005$	1
(40, 10)	$14.05^2 + 14.36^2 – 2(14.205)^2 = 0.13005$	1
(45, 15)	$15.21^2 + 17.41^2 – 2(16.31)^2 = 2.42$	1
(50, 20)	$18.66^2 +17.17^2 – 2(17.915)^2 = 1.11005$	1
	$SSPE = 12.45585$	11

Example 1 (Continued)

- It can be shown that SSE = 18.12167 with 19 d.f.

- Therefore

  $SSLF$ = 18.12167 – 12.45585 = 5.66582 with 8 d.f.

  $MSLF$ = 5.66582/8 = 0.708228,

  $MSPE$ = 12.45585/11 = 1.13235.

Example 1 (Continued)

- Hence

$$F_{LOF} = 0.708228/1.13235 = 0.625.$$

- Since the observed $F_{LOF} = 0.625 < F_{0.05}(8,11) = 2.95$, we do not reject H_0 and conclude that there is no significant evidence of any lack of fit in the multiple regression model.

4.8 Use of SAS to Test for LOF

The following SAS program can be used to test the lack of fit of the model in Example 1

```
data a;
  input y x1 x2;
  datalines;
9.73 0 20
...    ....
17.17 50 20
proc reg data=ch4ex1 lackfit;
   model y = x1 x2
run;
```

Proc reg has the option "lackfit"

Option for Lack of Fit Test

Use of SAS to Test for LOF (Continued)

Partial Output:

Analysis of Variance

Source	DF	Sum of Squares		Mean Square	F Value	Pr > F
Model	2	232.65759		116.32879	121.97	<.0001
Error	19	18.12167		0.95377		
Lack of Fit	8	5.665822		0.708228	0.63	0.7419
Pure Error	11	12.455850	SSPE	1.132350	F_{LOF}	
Corrected Total	19	18.121672		0.953772		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.257382	0.498437	10.55	<.0001
x1	1	0.162113	0.013191	12.29	<.0001
x2	1	0.248868	0.027924	8.91	<.0001

4.9 Use of R to Test for Lack of Fit

```
> ch4ex1 <- read.table("d:/ST3131/ch4ex1.txt", header=T)
> attach(ch4ex1)
>
> #Get SSE
> model1 <- lm(y~x1+x2)
> anova(model1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 156.900 156.900  164.50 8.351e-11 ***
x2         1  75.758  75.758   79.43 3.251e-08 ***
Residuals 19  18.122   0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
  ' ' 1
```

SSE

Use of R to Test for Lack of Fit (Continued)

```
> #Test for Lack of Fit
> fac.x1=factor(x1)
> fac.x2=factor(x2)
> model2=lm(y~fac.x1*fac.x2)
> anova(model1,model2)
Analysis of Variance Table
Model 1: y ~ x1 + x2
Model 2: y ~ fac.x1 * fac.x2
    Res.Df       RSS    Df   Sum of Sq        F     Pr(>F)
1       19   18.1217
2       11   12.4559     8      5.6658   0.6254     0.7419
```

Create new categorical variables for x1 and x2. Each value becomes one categorical variable

SSE

The best model that we can fit to the data

SSPE

SSLF

F_{LOF}