F- Test for Lack of Fit

In this part, we take up a formal test for determining whether a specific type of regression function adequately fits the data.

We illustrate this test for ascertaining whether a linear regression function is a good fit for the data under some assumptions.

Assumptions

The lack of fit test assumes that the observations *Y* for given X are:

(1) independent of errors

(2) the error is normally distributed

(3) The distributions of Y~have the same variance σ^2

The lack of fit test requires repeat, observations at one or more X levels. These data can be obtained as:

1- In non-experimental data, these may occur by chance, as when in a productivity study relating workers' output and

age, several workers of the same age happen to be included in the study.

2- In an experiment, one can assure by design that there are repeat observations. Repeat trials for the same level of the predictor variable, of the type described, are called replications. The resulting observations are called replicates.

Example [see text book page.120]

In an experiment involving 12 similar but scattered suburban branch offices of a commercial bank, holders of checking accounts at the offices were offered gifts for setting up money market accounts. Minimum initial deposits in the new money market account were specified to qualify for the gift. The value of the gift was directly proportional to the specified minimum deposit. Various levels of minimum deposit and related gift values were used in the experiment in order to ascertain the relation between the specified minimum deposit and gift value, on the one hand, and number of accounts opened at the office, on the other. Altogether, six levels of minimum deposit and proportional gift value were used, with two of the branch offices assigned at random to each level. One branch office had a fire during the period and was dropped from the study. The data are given below:

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

The simple regression model is estimated as

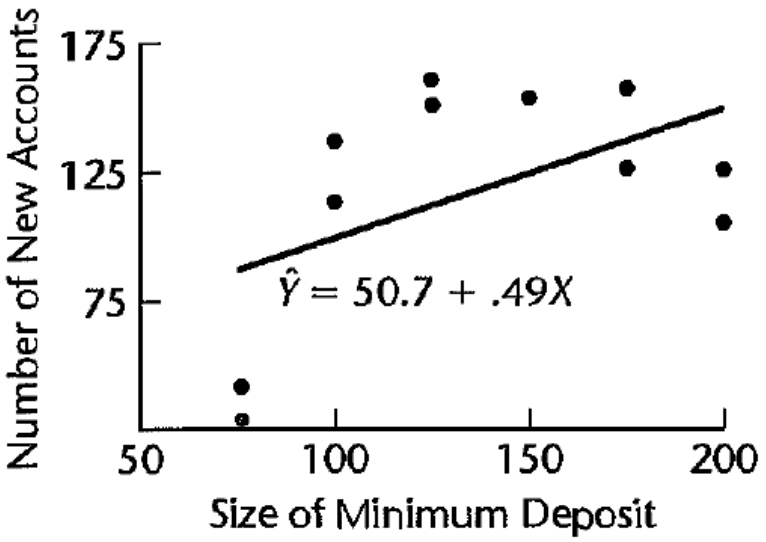
$$\widehat{Y} = 50.72251 + 0.48670X$$

R-results

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.7225   39.3979   1.287    0.23
x              0.4867    0.2747   1.772    0.11

Residual standard error: 40.47 on 9 degrees of freedom
Multiple R-squared:  0.2586,    Adjusted R-squared:  0.1762
F-statistic: 3.139 on 1 and 9 DF,  p-value: 0.1102
```

```
> summary(aov(model))
             Df Sum Sq Mean Sq F value Pr(>F)
x             1   5141    5141   3.139   0.11
Residuals     9  14742    1638
```



A scatter plot, together with the fitted regression line, and the other results indications that a linear regression function is inappropriate. Here we can us the **for Lack of Fit** to test this.

F-Test for lake of fit

Let the simple regression model (the reduced model) in the from

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n,$$

While the true model (Full model) in the form

$$Y_i = \mu_i + \varepsilon_i, i = 1, \ldots, n,$$

$Y_i = \mu_i + \varepsilon_i, i = 1, \ldots, n \Rightarrow$ the true model and μ_i might not be

$\beta_0 + \beta_1 x_i$.

Now if the used model is not the true model ($\mu_i \neq \beta_0 + \beta_1 x_i$), then $\hat{y}_i = b_0 + b_1 x_i$ based on the simple linear regression model cannot be an accurate predicted. value of u_i .

Thus, $e_i = y_i - \hat{y}_i = \mu_i - b_0 - b_1 x_i + \varepsilon_i = s_i \neq .\varepsilon_i$.

Then, the mean residual sum of squares

$$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} s_i^2}{n-2} \neq \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-2}$$

is no longer a sensible estimate of σ^2 .

To resolve this problem, we could try to obtain repeat observations with respect to the same covariate. Let

$y_{11}, y_{12}, \ldots, y_{1n_1} \Rightarrow$ n_1 repeated observation at x_1;

$y_{21}, y_{22}, \ldots, y_{2n_2} \Rightarrow$ n_2 repeated observation at x_2;

$\vdots$

$y_{c1}, y_{c2}, \ldots, y_{cn_c} \Rightarrow$ n_m repeated observation at x_c;

Note: $\sum_{j=1}^{c} n_j = n$.

In view of the repeated observation the full model can be formatted as:

$Y_{ij} = \mu_j + \varepsilon_{ij}$

where:

μ_j are parameters $j = 1, \ldots, c$
ε_{ij} are independent $N(0, \sigma^2)$

Since the error terms have expectation zero, it follows that:

$$E\{Y_{ij}\} = \mu_j$$

Thus, the parameter μ_j $(j = 1, \ldots, c)$ is the mean response when $X = X_j$.

To fit the full model to the data, we require the least squares or maximum likelihood estimators for the parameters μ_j . It can be shown that these estimators of μ_j are simply the sample means $\bar{Y}_j$ as follows

$$Q = \sum_{i=1}^{n_j} \sum_{j=1}^{c} \varepsilon_{ij}^2 = \sum_{i=1}^{n_j} \sum_{j=1}^{c} (Y_{ij} - \mu_j)^2$$

$$\frac{\partial}{\partial \mu_j} Q = 0 \Rightarrow -2 \sum_{i=1}^{n_j} (Y_{ij} - \mu_j) = 0 \Rightarrow n_j \mu_j = \sum_{i=1}^{n_j} (Y_{ij})$$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_{ij}) = \bar{Y}_j$$

Thus, the estimated expected value for observation Y_{ij} is $\bar{Y}_j$, and the error sum of squares for the full model therefore is:

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = SSPE$$

In the context of the test for lack of fit, the full model error sum of squares is called the *pure error sum of squares* and is denoted by *SSPE.*

Note that *SSPE* is made up of the sums of squared deviations at each X level. At level $X = X_j$, this sum of squared deviations is:

$$\sum_i (Y_{ij} - \bar{Y}_j)^2$$

The degrees of freedom of the SSPE is

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

The reduced model

The general linear test approach next requires consideration of the reduced model under H_0. For testing the appropriateness of a linear regression relation, the alternatives are:

$$H_0: E\{Y\} = \beta_0 + \beta_1 X$$
$$H_a: E\{Y\} \neq \beta_0 + \beta_1 X$$

Thus, H_0 postulates that μ_j in the full model is linearly related to X_j:

$$\mu_j = \beta_0 + \beta_1 X_j$$

The reduced model under H_0 therefore is:

$$Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij} \qquad \text{Reduced model}$$

Hence, the error sum of squares for the reduced model is the usual error sum of squares *SSE*:

$$SSE(R) = \sum\sum [Y_{ij} - (b_0 + b_1 X_j)]^2$$
$$= \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 = SSE$$

We also know that the degrees of freedom associated with $SSE(R)$ are:

$$df_R = n - 2$$

Test Statistic

The general linear test statistic

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

here becomes:

$$F^* = \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c}$$

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by *SSLF*:

$$SSLF = SSE - SSPE$$

We can then express the test statistic as follows:

$$F^* = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

where *MSLF* denotes the *lack of fit mean square* and *MSPE* denotes the *pure error mean square.*

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule

If $F^* \leq F(1-\alpha; c-2, n-c)$, conclude H_0

If $F^* > F(1-\alpha; c-2, n-c)$, conclude H_a

Then, we summarize the test steps as follows:

Let the given data as follows

X_1	X_1	.....	X_c

Y_{11} Y_{12} $\vdots$ Y_{1n_1}	Y_{21} Y_{22} $\vdots$ Y_{2n_2}		Y_{c1} Y_{c2} $\vdots$ Y_{cn_c}
$\bar{Y}_1$	$\bar{Y}_2$	....	$\bar{Y}_c$

From the data we calculate the following quantities

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{Error deviation}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\text{Pure error deviation}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\text{Lack of fit deviation}}$$

hence

$$\sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = \sum\sum(Y_{ij} - \bar{Y}_j)^2 + \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$$

$$SSE = SSPE + SSLF$$

$$\hat{Y}_{ij} = b_0 + b_1 X_j$$

$$\bar{y} = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ji}}{n}, \qquad \bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}, \qquad SSR = \sum_{j=1}^{c} n_j(\hat{y}_j - \bar{y})^2$$

we can define the lack of fit sum of squares directly as follows:

$$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$$

Since all Y_{ij} observations at the level X_j have the same fitted value, which we can denote by $\hat{Y}_j$, we can express

$$SSLF = \sum_j n_j(\bar{Y}_j - \hat{Y}_j)^2$$

Step 1:

$$H_0 : E\left(y_{ji}\right) = \beta_0 + \beta_1 x_j \quad vs \quad H_1 : E\left(y_{ji}\right) \neq \beta_0 + \beta_1 x_j$$

Step 2:

Calculate the test statistic F^* from the ANOVA table

Step 3:

Calculate the critical region F(1-α, c-2, n-c), where c-2 and n-c are the degrees of freedom (from AONVA table)

Step 4:

Decision:

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule

$$\text{If } F^* \leq F(1-\alpha; c-2, n-c), \text{ conclude } H_0$$
$$\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ conclude } H_a$$

All what we need now is to construct the ANOVA Table as follows:

The ANOVA table

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

Example [see text book page.120]

In this example, we see

	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

These sums of squares are then added over all of the X levels (j = **1,** ... , *c).*

For the bank example, we have:

SSPE == (28 - 35)2 + (42 – 35)^2 + (112 - 124)2 + (136 - 124)2 + (160 - 155)2+ (150 - 155)2 + (152 - 152)2 + (156 - 140)2

+ (124 - 140)2+ (124 - 114)2 + (104 - 114)2
= 1148
df_f=n-c=11-6=5

$$SSE(R) = SSE = 14741.6, \qquad df_R = n - 2 = 9$$

SSLF=SSE-SSPE= 14741.6 – 1148=13593.6

Then ANOVA Table is

Source	df	SS	MS	F
SSR	1	5141	5141	F1=5141/1637.956=3.138668
SSE	n-2=9	14741.6	1637.956	
Lack of fit	c-2=6-4=4	13593.6	3398.4	F^*=3398.4/229.6=14.80139
Pure error	n-c=11-6=5	1148	229.6	
SSTOT	10	19882.6		

(b) Bank Example

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

If the level of significance is to be a = .01, we require $F(.99;4, 5) = 11.4$. Since
F^* =MSLF/MSPE=13593.6/1148= 14.80 > 11.4, we conclude *H0,* that the regression function is not linear.
This, of course, accords with our visual impression from scatter plot. The P-value for the test is .006 → reject → simple linear model is not good fit for the given data.

Using R

```
install.packages("alr3")
library(alr3)
x=c(125,100,200,75,150,175,75,175,125,200,100)
y=c(160,112,124,28,152,156,42,124,150,104,136)
fit=lm(y~x)
pureErrorAnova(fit)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	5141.3	5141.3	22.393	0.005186 **
Residuals	9	14741.6	1638.0		
Lack of fit	4	13593.6	3398.4	14.801	0.005594 **
Pure Error	5	1148.0	229.6		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$F = \frac{\left(RSS(\text{model 2}) - RSS(\text{model m})\right)/m-2}{RSS(\text{model m})/n-m} = \frac{\dfrac{\sum_{j=1}^{m} n_j(\bar{y}_j - \hat{y}_j)^2}{m-2}}{\dfrac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ji} - \bar{y}_j)^2}{n-m}}$$

In general, we use the following procedure to fit simple regression model when the data contain repeated observations.

1. Fit the model, write down the usual analysis of variance table. Do

not **perform an F-test for regression ($H_0: \beta_1 = 0$).**

2. Perform the F-test for lack of fit. There are two possibilities.

(a) If significant lack of fit, stop the analysis of the model fitting and seek ways to improve the model by examining residuals.

(b) If lack of fit test is not significant, carry out an F-test for regression, obtain confidence interval and so on. The residuals should still be plotted and examined for peculiarities.

Example:

X	Y
90	81,83
79	75
66	68,60,62
51	60,64
35	51,53

$$\sum_{i=1}^{10} x_i = 629, \sum_{i=1}^{10} y_i = 657, \sum_{i=1}^{10} x_i^2 = 43161, \sum_{i=1}^{10} y_i^2 = 44249, \sum_{i=1}^{10} x_i y_i = 43189.$$

Thus, total sum of squares:

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 44249 - 10(65.7)^2 = 1084.1.$$

$$\Rightarrow b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2} = \frac{43189 - 10 * 62.9 * 65.7}{43161 - 10 * (62.9)^2} = 0.51814$$

$$\Rightarrow \text{regression sum of squares} = b_1^2 S_{XX} = 965.65636$$

$$\Rightarrow \text{residual sum of squares (reduced model)} = 1084.1 - 956.66 = 118.44$$

Pure error sum of squares:

X

90: $\bar{Y}_1 = \frac{81+83}{2} = 82, \sum_{i=1}^{2} (Y_{1i} - \bar{Y}_1)^2 = (81-82)^2 + (83-82)^2 = 2.$

79: $(75-75)^2 = 0$

66: $\bar{Y}_3 = \frac{68+60+62}{3} = 63.33, \sum_{i=1}^{3} (Y_{3i} - \bar{Y}_3)^2 = (68-63.33)^2 + (60-63.33)^2 + (62-63.33)^2 = 34.67$

51: $\bar{Y}_4 = \frac{60+64}{2} = 62, \sum_{i=1}^{2} (Y_{4i} - \bar{Y}_4)^2 = (60-62)^2 + (64-62)^2 = 8.$

35: $\bar{Y}_5 = \frac{51+53}{2} = 52, \sum_{i=1}^{2}(Y_{5i} - \bar{Y}_5)^2 = (51-52)^2 + (53-52)^2 = 2$

Then, pure error sum of squares=2+0+34.67+8+2=46.67

Lack of fit sum of squares=118.44-46.67=71.77

Source	df	SS	MS
Regression	1	965.66	965.66
Lack of fit	3	71.77	23.92
Pure error	5	46.67	9.33
Total	9	1084.1	

$\Rightarrow F = \frac{23.92}{9.33} = 2.56 < f_{3,5,0.05} = 5.41$

$\Rightarrow$ Not significant!! That is, the simple linear regression is adequate. The standard F-test for regression can be carried out.