Lack of Fit and Pure Error

Assumed Model:

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

$$\epsilon_{ji}\text{'s} \sim \text{NID}(0,\sigma^2)$$

Question: Are we sure that $E(Y_{ij}) = \beta_0 + \beta_1 X_j$

Answer: We will NEVER be absolutely sure, we should try to check it.

Procedure:

In order to address this question we will need replication or at least some “approximate" replication).

a) What constitutes a “replicate". Multiple or repeated observations of the dependent variable Y_{ij} at the same (or very nearly the same) value of X_j. **IMPORTANT - we are not talking about repeated sampling, we need independent observations.**

b) From the repeated observations within a single value of X_j, which we will call a RUN, we compute what is called PURE ERROR (ie. We will obtain a SSPure Error or SSPE).

NOTE: It is not necessary that we have runs at EVERY value of X_i, as long as we have runs at some.

c) The difference between the SSResiduals (SSE) and the SSPure Error (SSPE) is called the SSLack of Fit (SSLOF).

d) Under the hypothesis that the model is correct, then both SSPE and SSLOF should be estimating SSE
(ie. if the model is correct SSE = SSPE = SSLOF).

e) We then compare the estimators of Pure Error and Lack of Fit, and if they are not different, then we are safe in using our original model. Otherwise, the model is not useful for our problem.

SKIP

Notation:

$$\begin{bmatrix} Y_{11} & Y_{21} & \cdots & Y_{m1} \\ Y_{12} & Y_{22} & \cdots & Y_{m2} \\ \vdots & & & \\ Y_{1n_1} & Y_{2n_2} & \cdots & Y_{mn_m} \end{bmatrix}$$ the raw data matrix of Y_{ji} values

$\uparrow \quad \uparrow \quad \quad \uparrow$

$X_1 \quad X_2 \quad \dots \quad X_m$ the corresponding X_j values

$\uparrow \quad \uparrow \quad \quad \uparrow$

$n_1 \quad n_2 \quad \dots \quad n_m$ the corresponding sample size for each X_j

$i = 1, 2, ..., n_1$ repeated observations at value of X_1
$i = 1, 2, ..., n_2$ repeated observations at value of X_2
$\vdots$
$i = 1, 2, ..., n_m$ repeated observations at value of X_m

Consider the following sum of squares

at X_1: $\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2$

This sum of squares does not depend on the model, it only depends on the “true" residuals.
To see this let

$Y_{1i} = f(X_1) + r_{1i}$ where r_{1i} are the true residuals
then
$\bar{Y}_1 = f(X_1) + \bar{r}_1$ and

$$\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 = \sum_{i=1}^{n_1}(r_{1i} - \bar{r}_1)^2$$

LOF assumes ϵ_i is NIDrv(0,σ^2)

From basic statistics we know that

$$\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 \quad \text{estimates} \quad (n_1 - 1)\,\sigma^2$$

$$\sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)^2 \quad \text{estimates} \quad (n_2 - 1)\,\sigma^2$$

$$\vdots$$

$$\sum_{i=1}^{n_m}(Y_{mi} - \bar{Y}_m)^2 \quad \text{estimates} \quad (n_m - 1)\,\sigma^2$$

therefore, since all σ^2 are equal,
This is basically ANOVA, where each $X_i = \tau_i$

$$\sum_{j=1}^{m}\ \sum_{i=1}^{n_m}(Y_{ji} - \bar{Y}_j)^2 \quad \text{estimates} \quad \sigma^2 \sum_{j=1}^{m}(n_j - 1)$$

and

$$\frac{\sum_{j=1}^{m}\ \sum_{i=1}^{n_m}(Y_{ji} - \bar{Y}_j)^2}{\sum_{j=1}^{m}(n_j - 1)} \quad \text{estimates} \quad \sigma^2$$

Note that these calculations are wholly independent of whatever model is chosen. These calculations depend only on the deviations of members of a “RUN" from the mean of that RUN. In fact, these deviations would be the same for ANY regression or even for an Analysis of Variance.

Continuing our development of the residuals for a regression
Recall that for regression the **SSError $\equiv$ SSResiduals**

$$\text{SSError} = \sum_{j=1}^{m} \sum_{i=1}^{n_m} (Y_{ji} - \hat{Y}_{ji})^2 =$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n_m} (Y_{ji} - \bar{Y}_j)^2 - \sum_{j=1}^{m} \sum_{i=1}^{n_m} (\bar{Y}_{ji} - \hat{Y}_j)^2$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n_m} (Y_{ji} - \bar{Y}_j)^2 - \sum_{j=1}^{m} n_j (\bar{Y}_{ji} - \hat{Y}_j)^2$$

SSPure Error SSLack of Fit

The above calculations are true regardless of the linear model used.

This subdivision of SS is accompanied by a subdivision in df.

$$\text{df for SSError} = \sum_{j=1}^{m} n_j - p = n - p$$

$$\text{df for SSPure Error} = \sum_{j=1}^{m} (n_j - 1) = n - m$$

$$\text{df for SSLack of Fit} = m - p$$

ANOVA TABLE

SOURCE	d.f.	SS	MS
Regression	p-1	SSReg	
Residual or Error	n-p	SSRes	
Lack of Fit	m-p	SSRes-SSPE	MSLOF
Pure Error	n-m	SSPE	S^2_{Error}
Total corrected	n-1		

To test Lack of Fit, compute an F test statistic as

$$F_{LOF} = \frac{MS_{LOF}}{S^2_{\text{Error}}} ,$$

and compare this with $F_{(m\text{-}p,\ n\text{-}m,\ 1\text{-}\alpha)}$. If $F_{LOF} > F_{(m\text{-}p,\ n\text{-}m,\ 1\text{-}\alpha)}$, there is evidence that your model is inappropriate.

As a result, you cannot test H_o: $\beta_0 = 0$ or H_o: $\beta_1 = 0$ because your estimators (b_0 and b_1) are biased and are inappropriate. Confidence intervals also will not be meaningful.

However, S^2_{Error} is still an unbiased estimator of σ^2 (since it does not depend on the model).

If $F_{LOF} > F_{(m\text{-}p,\ n\text{-}m,\ 1\text{-}\alpha)}$, then MS_{LOF} and S^2_{Error} are both estimators of σ^2, and may be pooled. The model is acceptable under these conditions.

The text looks at LOF testing as a Full and Reduced Model, where

REDUCED = SSE_{Reg}

FULL = $SSE_{\text{ANOVA}} = SSE_{\text{PureError}}$

Difference = SSLOF

and this is tested with the full model, SSPE

This test gives exactly the same results as any test of $F = \frac{\text{SSLOF}}{\text{SSPE}}$

Lack of Fit in Simple Linear Regression - Numerical Example

Recall our Assumptions of model and residuals

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$\epsilon_i\text{'s} \sim \text{NID}(0,\sigma^2)$$

Question: Are we sure that $E(Y_i) = \beta_0 + \beta_1 X_i$

SSPure error and SSLack of fit in SAS

The pure error is obtained by adjusting the variation about each individual value of X_j.

This is basically ANOVA, where each X_j is a “treatment"

In the event that there is only one observation at an X_j, then $\bar{X}_j = X_j$, and the SS contribution is 0 handout from document HO1994

Lack of Fit and Pure Error (Data: Draper and Smith, Page 60, Problem “F".)

X (coded)	Y (coded)	$\bar{Y}$
4.7	3, 2	2.5
5.0	3, 4	3.5
5.2	4, 5, 3	4.0
5.3	7	7.0
5.6	6	6.0
5.9	10, 9, 6	8.33

Intermediate Calculations

$\Sigma X_i = 63.6$; $\Sigma X_i^2 = 339.18$; $\Sigma(X_i-\bar{X})^2 = 2.10$; $\bar{X} = 5.30$

$\Sigma Y_i = 62.0$; $\Sigma Y_i^2 = 390.00$; $\Sigma(Y_i-\bar{Y})^2 = 69.67$; $\bar{Y} = 5.17$

$\Sigma X_i Y_i = 339.1$; $\Sigma(X_i-\bar{X})(Y_i-\bar{Y}) = 10.5$

Computations

$b_1 = 5.0$

$b_0 = \bar{Y} - b_1\bar{X} = -21.33$

SSRegression = 52.5

Total (Corrected) = 69.67

Residual = 17.17

SSPE (SS Pure Error) $= (3-2.5)^2+(2-2.5)^2+(3-3.5)^2+(4-3.5)^2+(4-4)^2$

$+(5-4)^2+(3-4)^2+(7-7)^2+(6-6)^2+(10-8.33)^2+(9-8.33)^2+(6-8.33)^2$

$= .5^2+.5^2+.5^2+.5^2+0^2+1^2+1^2+0^2+0^2+1.67^2+.67^2+2.33^2 = 11.66$

SSLOF (SS Lack of Fit) = 17.17 - 11.66 = 5.51

ANOVA TABLE

SOURCE	d.f.	SS	MS	F
Regression	1	52.50	52.50	30.52
Residual or Error	10	17.17	1.72	
Lack of Fit	(4)	5.50	1.375	0.706
Pure Error	(6)	11.67	1.945	
Total	11	69.97		

Tabular values: $F_{0.05,\,4,6\,df} = 4.53$, $F_{0.05,\,1,10\,df} = 4.96$

WE CONCLUDE THAT THERE IS NO EVIDENCE OF LACK OF FIT.

Note: We do not have to calculate the SS components by hand. There are several ways to do this on the computer.

1) Recall that the Lack of Fit is the SSDeviations of the $\bar{Y}._j$ from the regression line. We could therefore regress the $\bar{Y}._j$ on the X_j.
The residuals for this line would then be the $SS_{\text{Lack of Fit}}$.

2) Recall that the Pure Error is the SSDeviation of the individual points from the $\bar{Y}._j$. We could then do a simple ANOVA (one way), treating each X_j as a separate treatment. Since ANOVA calculates the deviations from the treatment means as its error, we would obtain the $SS_{\text{Pure Error}}$

3) If the independent variable is included in the model as both a quantitative variable, and as a categorical variable, then SAS will give both the Pure Error and the Lack of Fit.

See SAS Handout of Examples