Poorly fitting model diagnostics

CHRIS SUTTON, FEBRUARY 2024

Can you list?

3 TYPES OF RESIDUAL PLOTS

Plot 1
- What we plot
- R code

Plot 2
- What we plot
- R code

Plot 3
- What we plot
- R code

4 THINGS WE ARE LOOKING FOR FROM THEM

Check 1	Check 2
Check 3	Check 4

Pure Error and Lack of Fit

We have listed the scenarios where a simple linear regression model might not be appropriate:

- residuals not from Normal distribution
- variance not constant

But so far, we have relied on looking at residual plots to assess this

❑ Is there some more formal way to show this lack of fit?

○ we will now look at one type of case of poorly fitting model

Replications

Replications are where we have more than one observation with the same *x* value but they have different *y* values

we use y_{ij} to be the jth observation at x_i

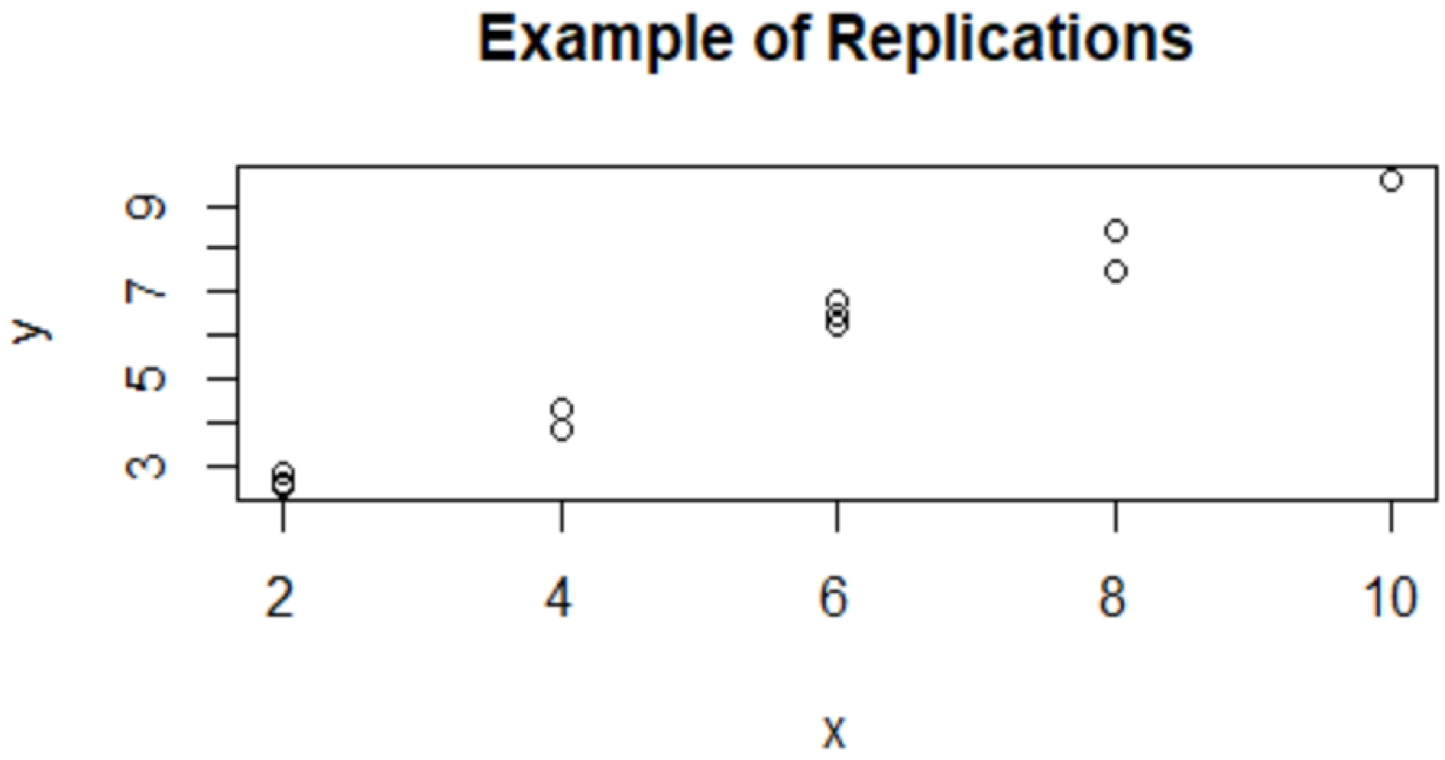
- where $i = 1, 2, \ldots m$ and $j = 1, 2, \ldots n_i$

In our linear regression model, although each of the y_{ij} observations might be different at a certain x_i, the fitted value will be the same $\hat{y}_i$ for all

Residuals

The residuals are now

$$e_{ij} = y_{ij} - \hat{y}_i$$



Two sources of Residual Error

1
- random variation in y_{ij} where observations at the same x_i can produce different *y* values

2
- lack of fit in the model which does not capture all that is found in the observed data

Two sources of Residual Error

1 Pure Error

- the amount of random variation at x_i
- the difference between an observation y_{ij} and the mean of observations taken at the same x_i

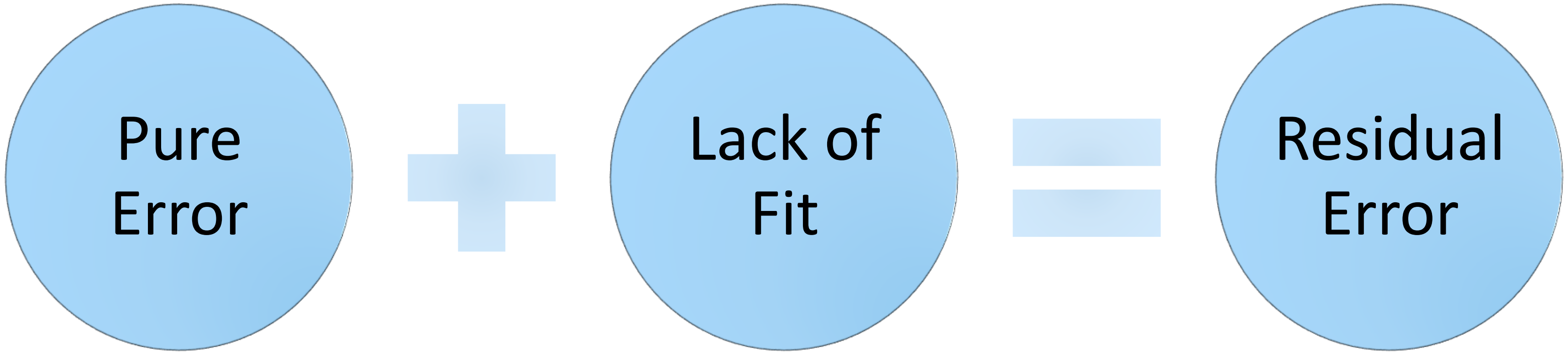
2 Lack of Fit

- the difference between the mean observed value and the model fitted value at x_i

Two sources of Residual Error

Pure Error = $y_{ij} - \bar{y}_i$

Lack of Fit = $\bar{y}_i - \hat{y}_i$



Residual Sum of Squares

We can split the residual sum of squares SS_E (from our ANOVA table) into:

- *Pure Error sum of squares* SS_{PE}
  - measures overall random variation
- *Lack of Fit sum of squares* SS_{LoF}
  - measures overall model lack of fit

Residual Sum of Squares

With the i, j notation SS_E becomes

$$SS_E = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$$

And this can be split between:

$$SS_{PE} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

$$SS_{LoF} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$$

Residual Sum of Squares

In the simple linear regression model with replications

$$SS_{PE} + SS_{LoF} = SS_E$$

Expanded ANOVA table with replications

Using *Pure Error* and *Lack of Fit* we can expand the ANOVA table

Where there are replications

Splitting the Residual Sum of Squares SS_E

Note the Regression Sum of Squares entry is unchanged

Degrees of freedom

To calculate pure error we need m means for the $\bar{y}_i$ (i = 1, 2, …, m)

Each of these mean calculations takes a degree of freedom

Therefore degrees of freedom for Pure Error = $n - m$

Previously Residuals had $n - 2$ d.f.

Therefore degrees of freedom for Lack of Fit = $(n - 2) - (n - m) = m - 2$

Mean Squares

We will see later that

$E[SS_{PE}] = (n - m)\sigma^2$ whether the model assumptions are true or not

$E[SS_{LoF}] = (m - 2)\sigma^2$ if the model assumptions are true

Which means that:

- MS_{PE} gives an unbiased estimator of σ^2
- MS_{LoF} gives an unbiased estimator of σ^2 if the model assumptions are true

Variance Ratio

Therefore in all cases

$$\frac{(n-m)MS_{PE}}{\sigma^2} \sim \chi^2_{n-m}$$

and if the model assumptions are true

$$\frac{(m-2)MS_{LoF}}{\sigma^2} \sim \chi^2_{m-2}$$

We can now use the ratio of these two divided by their respective d.f. to calculate another Variance Ratio

Variance Ratio for residuals

If the regression model assumptions are true

$$\frac{MS_{LoF}}{MS_{PE}} \sim F_{n-m}^{m-2}$$

We are now able to construct an expanded ANOVA table for the case where there are replications

Source of variation	d.f.	SS	MS	VR
Regression	1	SS_R	MS_R	$\frac{MS_R}{MS_E}$
Residual	n – 2	SS_E	$MS_E = \frac{SS_E}{n-2}$	
Lack of Fit	$m-2$	SS_{LoF}	$MS_{LoF} = \frac{SS_{LoF}}{m-2}$	$\frac{MS_{LoF}}{MS_{PE}}$
Pure Error	$n-m$	SS_{PE}	$MS_E = \frac{SS_{PE}}{n-m}$	
Total	n – 1	SS_T		

Expanded ANOVA table