Linear Statistical Modeling Methods with SAS

Diagnostic and Remedial Measures for SLR Models

Xuemao Zhang
East Stroudsburg University

February 19, 2024

Outline

- SLR Assumptions
- Analysis of Residuals
- F Test for Lack of Fit

SLR Assumptions

Recall that the SLR model between X and Y has the form

$$Y_i|X = x_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

where the random errors ϵ_is are assumed to be

- independent
- normal with having mean **0** and
- common variance σ^2

The **residuals**, $\mathbf{e_i}, i = 1, \ldots, n$ are the differences between the observed and fitted values for each data value:

$$\mathbf{e_i} = y_i - \hat{Y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

- If the model fits well, the residuals should behave as the random errors ϵ_i would: **independent** random selections from the same distribution (usually assumed normal) with mean 0, i.e $N(0, \sigma^2)$.
- We analyze **the residuals** to check the quality of the model fit.

Some Properties of the Residuals

Among the interesting properties of the residuals are these:

- They sum to zero: $\sum_{i=1}^{n} e_i = 0$. This also implies that their mean $\overline{e} = 0$.
- The Pearson correlation between the residuals and the predictors (independent variables) is zero (we say the residuals and the predictors are **uncorrelated**).
- The residuals and fitted values are also **uncorrelated**.

Residual Plot

- A residual plot is a scatter-plot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $y - \hat{y}$ (where $\hat{y}$ is the predicted value of y).
- That is, a residual plot is a graph of the points

$$(x_i, y_i - \hat{y}_i), i = 1, \ldots, n.$$

- After a SLR model is fit, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ depends on x_i only, the plot

$$(\hat{y}_i, y_i - \hat{y}_i), i = 1, \ldots, n.$$

is equivalent to the above residual plot except that the x-axis values are different.

Checking Model Assumptions

When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

(1) **Independence:** The residual plot should not have any obvious patterns (not even a straight line pattern). This confirms that the scatterplot of the sample data is a straight-line pattern.

(2) **Constance Variance:** The residual plot should not become thicker (or thinner) when viewed from left to right. This confirms the requirement that for different fixed values of x, the distributions of the corresponding y values all have the same standard deviation.

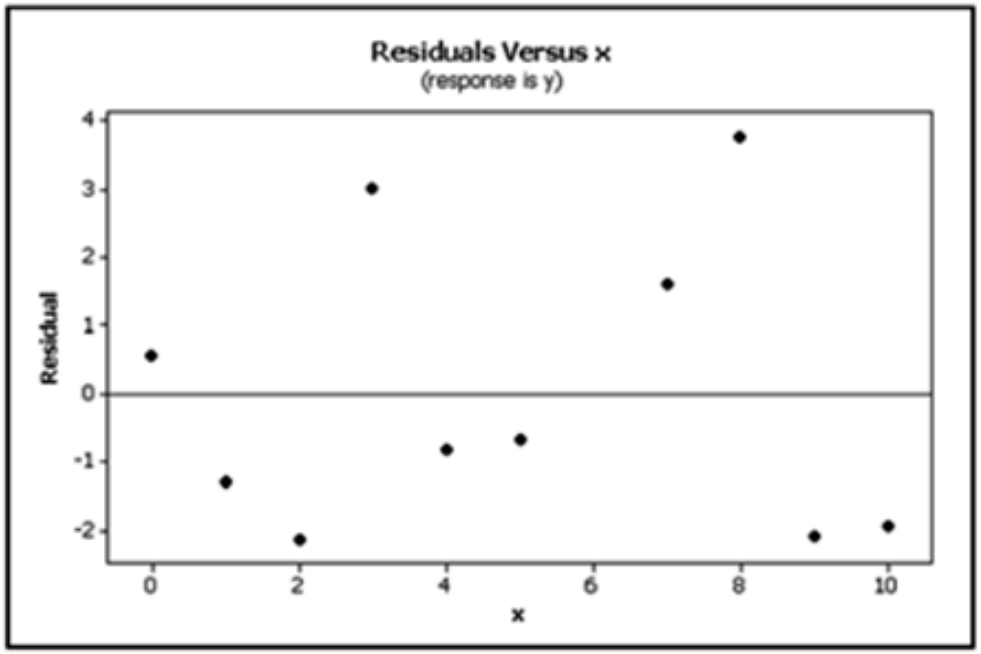
(3) **Normality.** Residuals should be plotted on a normal quantile plot (the SAS procedure univariate will do this). The population distribution is Normal if the pattern of the points is reasonably close to a straight line and the points do not show some **systematic** pattern that is not a straight-line pattern.

- **Time series plot** is a plot of residuals versus time if the observations form a time series. The time series plot can be used to detect the dependence of this type. If the points vary randomly around the horizontal axis in the time series plot, everything is fine. However, if adjacent residuals are similar, this indicates that the errors may be **serially correlated**.

Checking Model Assumptions

Example. Regression model is a good model:

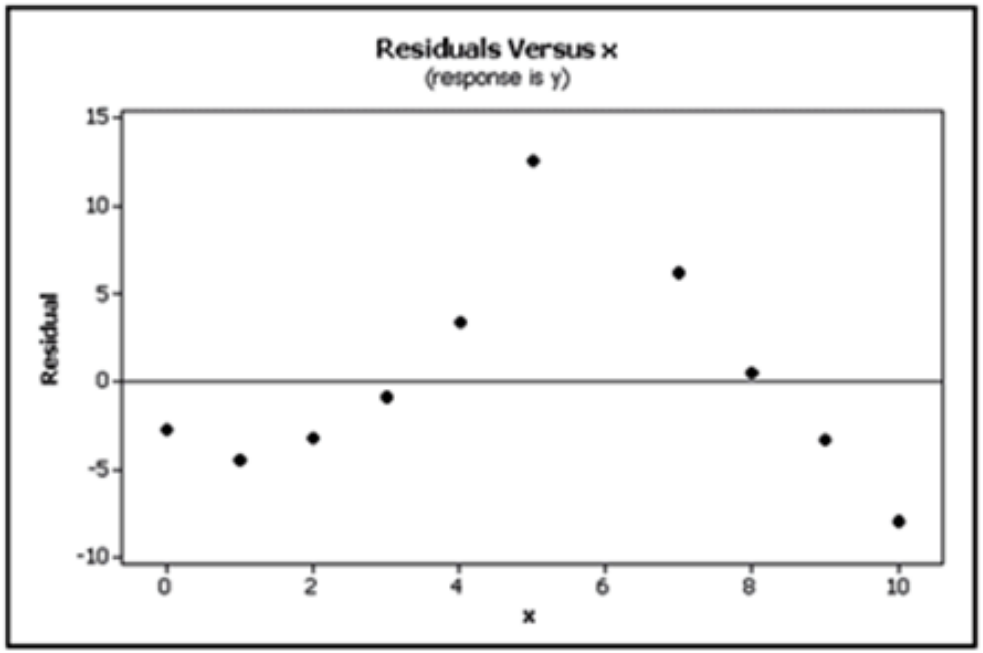
Residual Plot Suggesting That the Regression Equation Is a Good Model



Checking Model Assumptions

Example. Distinct pattern: sample data may not follow a straight-line pattern.

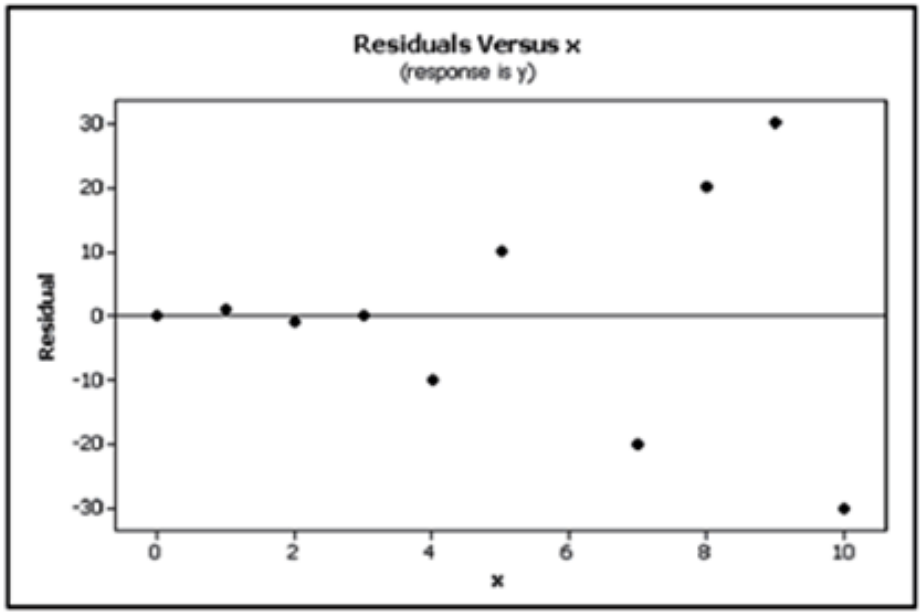
Residual Plot with an Obvious Pattern, Suggesting That the Regression Equation Is Not a Good Model



Checking Model Assumptions

Example. Residual plot becoming thicker: equal standard deviations violated.

Residual Plot That Becomes Thicker, Suggesting That the Regression Equation Is Not a Good Model



Checking Independence and Constant Variance

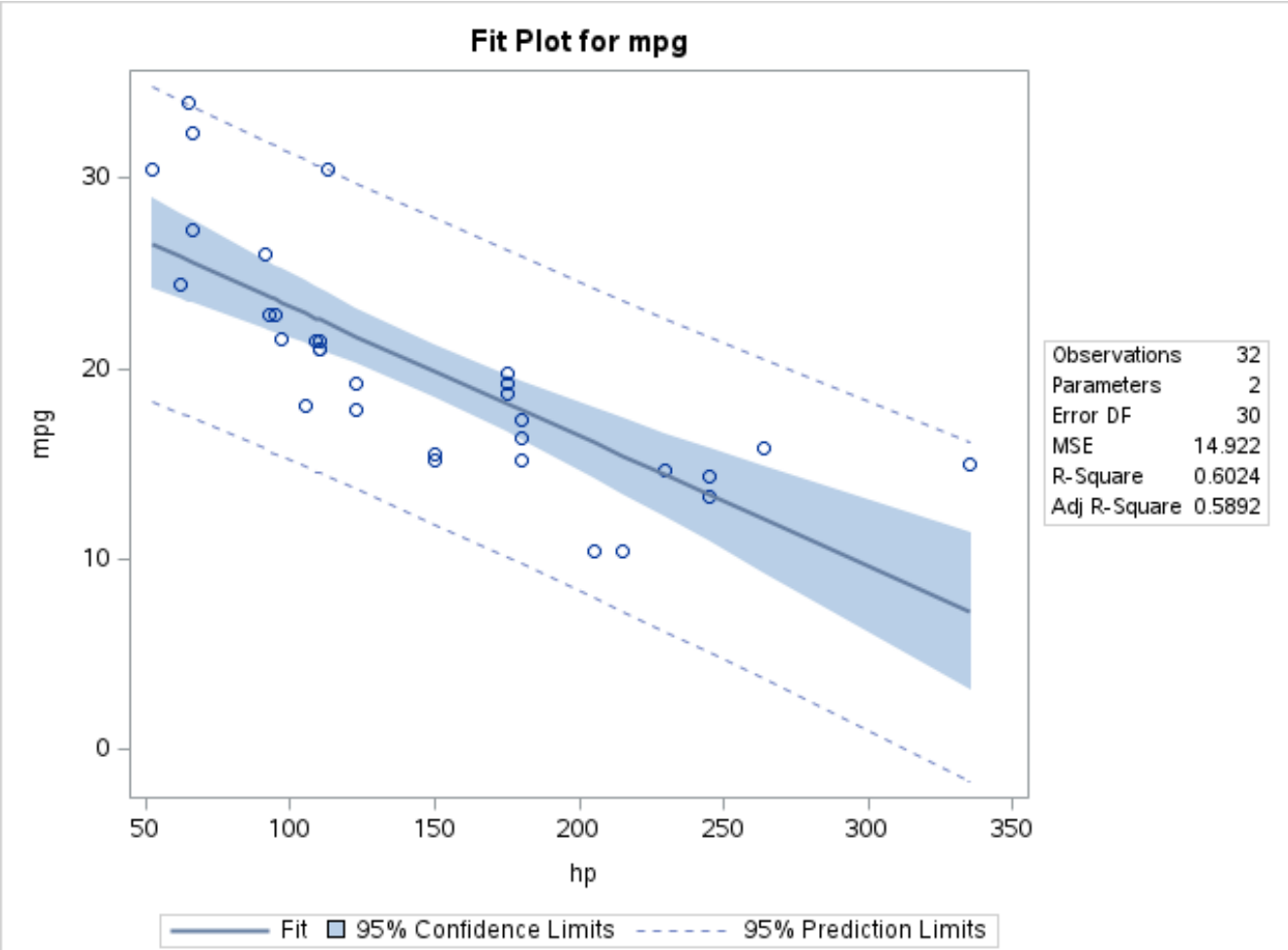
Example. Model fit and residual analysis by SAS: Check out the quality of the fit for the mtcars data:

```
proc reg data=mtcars plots(only) = (residualplot);
   model mpg = hp;
run;
```

Outliers and Influential Points

- In a residual plot, an **outlier** is a point lying far away from the other data points.
- **Influential points** are points that strongly affect the graph of the SLR line.

The following is the SLR fit of the mtcars data



Outliers and Influential Points

Observe what happens if we include this additional data point:

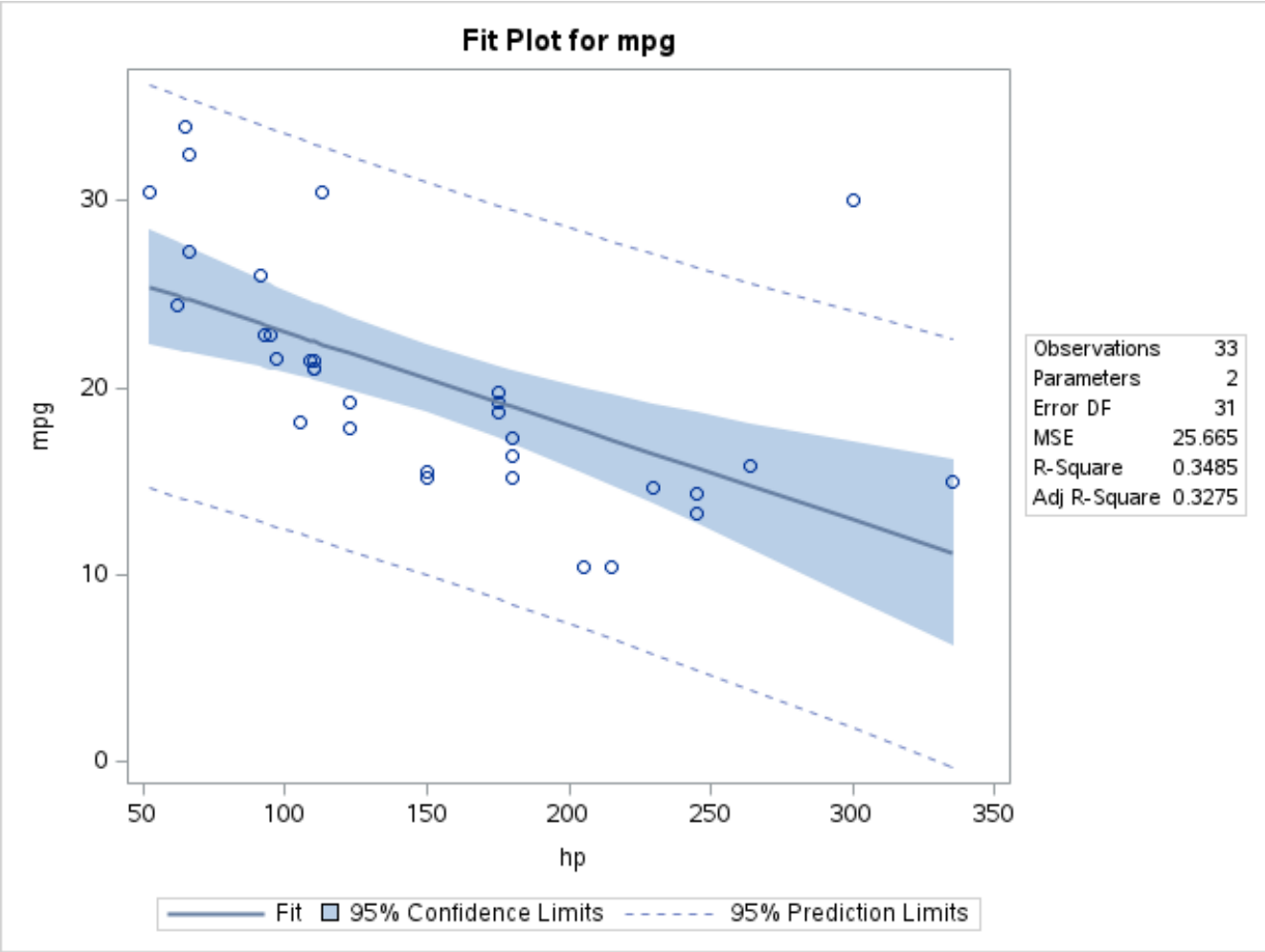
$$\text{mpg} = 30, \text{hp} = 300.$$

```
data newobs;
input hp mpg@@;
datalines;
300 30
;
run;

data mtcars;
set mtcars newobs;
run;

proc reg data=mtcars;
model mpg = hp ;
run;
```

Outliers and Influential Points



Outliers and Influential Points

- The additional point is an influential point because the graph of the regression line did change considerably.
- The additional point is also an outlier because it is far from the other points.

Remark. An influential point must be an outlier. An outlier may not be an influential point.

Time series plot

If the errors are dependent, the levels of the tests and confidence intervals are no longer correct.

Time series plot: If the observations form a time series, their covariance often is a (usually monotonely decreasing) function of the time between observations. Dependence of this type can be detected by plotting the residuals e_i against the observation times t_i, $i = 1, \ldots, n$.

- If the points vary randomly around the horizontal axis in the time series plot, everything is fine.
- However, if adjacent e_i are similar, this indicates that the errors may be serially correlated.
- Sometimes we even observe a jump in the level of the residuals. In such a case, the model has evidently changed suddenly at a particular point in time.

Durbin-Watson test

It is possible to test independence against an alternative of serial correlation. Two such tests are:

- The **run test**, which counts the number of continuous sub-sequences (runs) in which the residuals have identical signs. When independence is assumed, there should not be too many or too few runs.
- The Durbin-Watson test, which uses the test statistic

$$T = \frac{\sum_{i=1}^{n-1}(e_{i+1} - e_i)^2}{\sum_{i=1}^{n} e_i^2} \approx 2\left(1 - \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sum_{i=1}^{n} e_i^2}\right)$$

The quotient in this formula is an estimate of the correlation of ε_i and ε_{i+1} (assuming that all the ε_i have the same variance). If the ε_i are independent, T is approximately 2; small values of T indicate positive dependence and large values of T indicate negative dependence.

Durbin-Watson test

```
proc reg data=mtcars;
model mpg = hp / DW; /* DW  computes a Durbin-Watson statistic */
run;
```

F Test for Lack of Fit

We illustrate this test for ascertaining whether or not a linear regression function is a good fit for the data. This lack of fit test assumes that the observations Y for given $X = x$ are

- independent
- normally distributed
- the distributions of Y have the same variance

- In order to do a lack of fit test we need to have repeated observations at one or more levels of X.
- We perform the following hypothesis test:

$$H_0 : E(Y_i) = \beta_0 + \beta_1 x_i$$
$$H_a : E(Y_i) \neq \beta_0 + \beta_1 x_i$$

Note: The model under the alternative uses the mean response at each level of X to predict Y

F Test for Lack of Fit

Repeated trials for the same level of the independent variable, of the type described, are called **replications**. The resulting observations are called **replicates**.

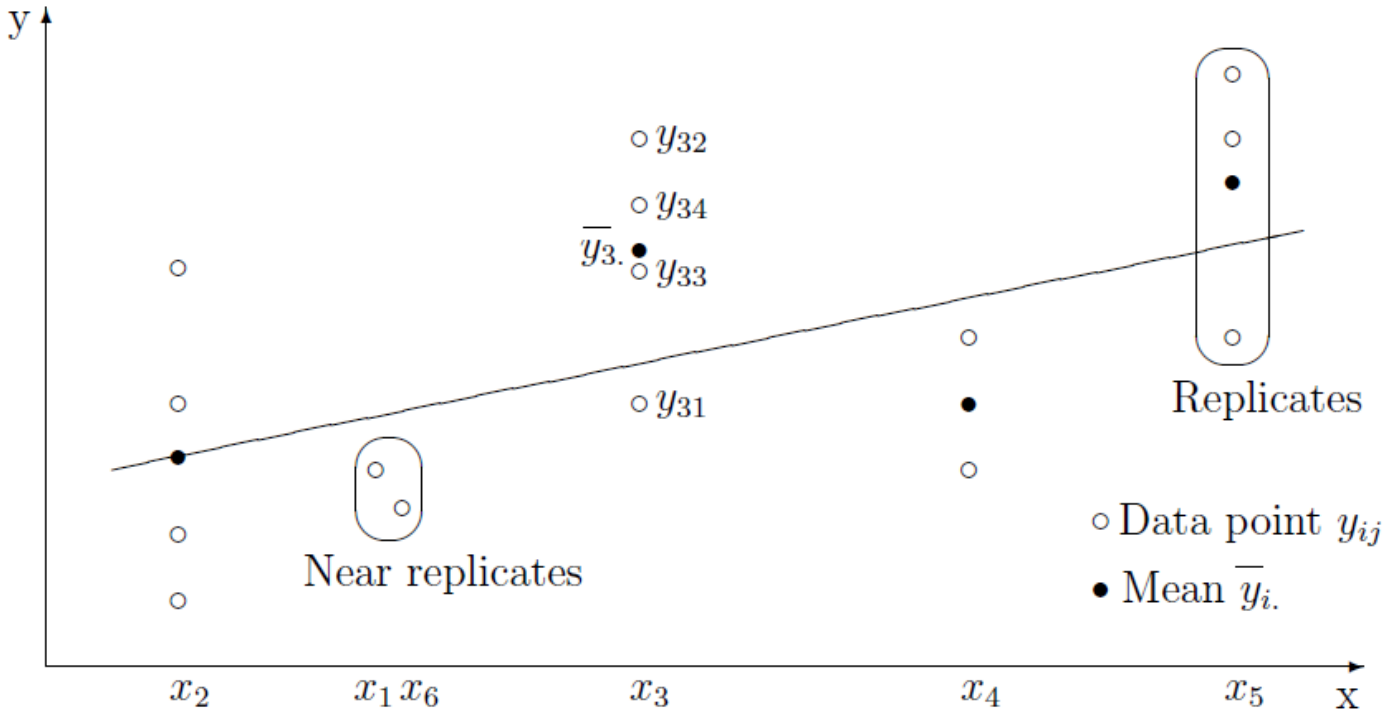


Figure : Replicates

F Test for Lack of Fit

Data: $(x_i, y_{ij}), i = 1, \ldots, k; j = 1, \ldots, n_i$ and let $n = \sum_{i=1}^{k} \sum_{j=1}^{n_i}$.

The model under H_a then is

$$Y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \ldots, k; j = 1, \ldots, n_i,$$

where $\varepsilon_{ij} iid \sim N(0, \sigma^2)$.

The least squares estimate of μ_i is simply the (arithmetic) mean of the observations at x_i:

$$\widehat{\mu_i} = \overline{y_{i.}} = \sum_{j=1}^{n_i} y_{ij}/n_i, i = 1, \ldots, k.$$

And the SSE then can be split as

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \widehat{y_i})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_{i.}})^2 + \sum_{i=1}^{k} n_i (\overline{y_{i.}} - \widehat{y_i})^2$$

- The first part measures the random error, and has $\sum_{i=1}^{k} (n_i - 1) = n - k$ degrees of freedom
- The second part measures the random error and the "lack of fit", and has $k - 2$ degrees of freedom.

F Test for Lack of Fit

Table 1: ANOVA Table for Lack of Fit

Source	df	SS	MS(Mean Squares)
Model(R)	1	$\sum_{i=1}^{k} n_i(\widehat{y_i} - \overline{y})^2$	MSR=SSR/1
Error (E)	$n-2$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y_i})^2$	MSE = SSE/(n-2) $=\widehat{\sigma^2}$
LF	$k-2$	$\sum_{i=1}^{k} n_i(\overline{y_{i.}} - \widehat{y_i})^2$	SSLF/(k-2)
PE	$n-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})^2$	SSPE/(n-k)
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$	

$$SSPE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_{i.}})^2$$

Here SSPE stands for pure error sum of squares. we see that X_i levels for which $n_i = 1$ do not contribute to the degrees of freedom since $n_i - 1 = 0$ then.

$$MSPE = \frac{SSPE}{n-k}$$

The reason for the term "pure error" is that MSPE is an unbiased estimator of the error variance σ^2 no matter what is the nature of the regression function. MSPE measures the variability of the distributions of Y without relying on any assumptions about the nature of the regression relation; hence, it is a "pure" measure of the error variance.

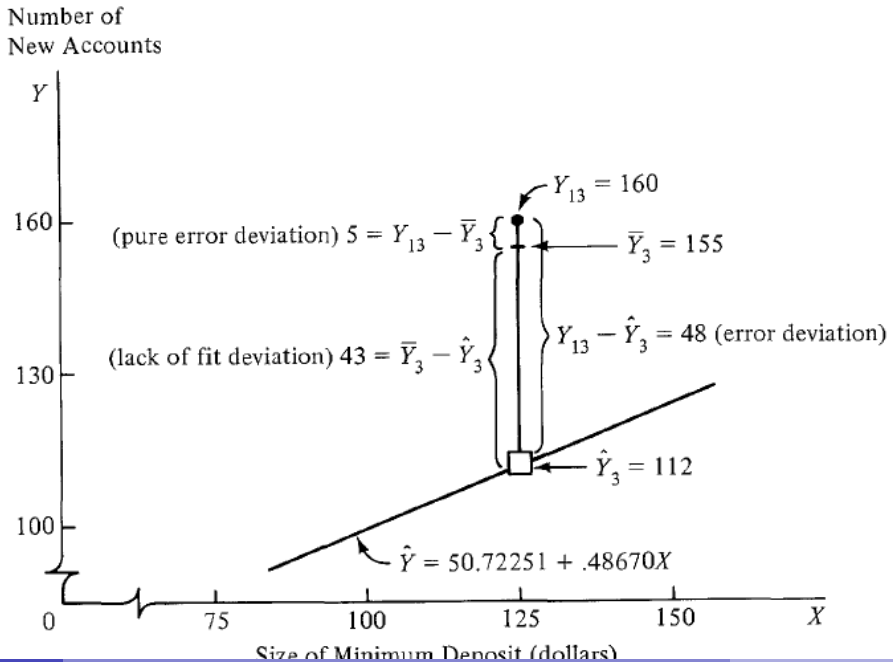
F Test for Lack of Fit

$$SSLF = \sum_{i=1}^{k} n_i(\overline{y_{i.}} - \widehat{y_i})^2$$

Here SSLF denotes lack of fit sum of squares. It is a weighted sum of squares (the weights are the sample sizes n_i) of the deviations: $\overline{y_{i.}} - \widehat{y_i}$.

Note that these deviations represent the difference between the mean $\overline{y_{i.}}$ and the fitted value $\widehat{y_i}$ based on the regression model. The closer the $\overline{y_{i.}}$ are to the $\widehat{y_i}$, the greater is the evidence that the fitted regression function is a good fit and therefore appropriate. The further the $\overline{y_{i.}}$ deviate from the $\widehat{y_i}$, the more the indication that the fitted regression function is inappropriate.

FIGURE 4.11 Illustration of decomposition of $Y_{ij} - \hat{Y}_j$



F Test for Lack of Fit

Test Statistic:

$$F^* = \frac{MSLF}{MSPE}.$$

It can be shown that

$$E(MSLF) = \sigma^2 + \frac{\sum n_i[E(Y_i) - (\beta_0 + \beta_1 x_i)]^2}{k-2}$$

- Under H_0, F^* follows the $F(k-2, n-k)$ distribution if the regression function is linear and all other three model assumptions hold. This test is right-tailed. Large values of the test statistic let us reject H_0 and indicate that regression function is not linear.
- In SAS, you can test for lack of fit by specifying the **LACKFIT** option in the MODEL statement in proc reg.

F Test for Lack of Fit

```
proc reg data= mtcars;
model mpg = hp/ lackfit;
run;
```

- It is not a good example because F Test for Lack of Fit needs responses from same x values which is like repeated measures.

Remedial Measures

Remedial Measures

If a simple linear regression model is not appropriate for the data at hand, there are two basic choices:

- Search for a more appropriate model
- Use some transformation on the data so that a SLR model is appropriate for the transformed data

Some transformations

- $Y^{'} = \sqrt{Y}, X^{'} = \sqrt{X}$
- $Y^{'} = log(Y), X^{'} = log(X)$
- $Y^{'} = \frac{1}{Y}, X^{'} = \frac{1}{X}$
- or transform either variable.

For more information, please read section 3.9 of the textbook.

License