Week 3: Residual Analysis (Chapter 3)

Properties of Residuals

Here's some properties of residuals, some of which you learned in previous lectures:

1. Definition of Residual: $e_i = Y_i - \hat{Y}_i$
2. $\varepsilon_i = Y_i - E(Y_i)$
3. $\varepsilon_i \sim NID(0, \sigma^2)$
4. mean: $\sum e_i = 0$ ➔ $\bar{e} = 0$
5. variance: $\frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE$

Semistudentized Residuals

We will use standardized residuals in some of our analyses of residuals. The standardization formula, $e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$, is often used to standardize residuals. KNNL on page 103 explain this standardization would create a "studentized" residual if $\sqrt{MSE}$ were an estimate of the standard deviation of the residual. However, this formula does not produce truly studentized residuals because $\sqrt{MSE}$ is only an approximation of the standard deviation of e_i. They will discuss how to calculate studentized residuals in chapter 10. Still, this formula is the basis of many residual analysis techniques, and it is the formula used by many statistical software packages, such as SAS, to standardize the residuals.

Six Areas in Which We Will Use Residual Analysis

1. regression not linear
2. non-constant variance
3. independence of residuals
4. outliers
5. normality of errors
6. important predictor (independent) variables.

Visual Diagnostics for Residuals

Residual plots are quite useful for examining residuals in the above six categories (Figure 3.4, page 106). You can also use a *normal probability plot* to examine if the residuals are normally distributed (normal probability plots are often standard output in statistical analysis software).

Statistical Tests for Residuals

<u>Normality</u>: You can use the *Shapiro-Wilks statistic* (this is standard output in the Proc Freq procedure in SAS). KNNL also mention the *Correlation test for Normality*, which they say is easier to use than the Shapiro-Wilks test. You can also use standard goodness of fit tests, like the chi-square or Kolmogorov-Smirnov tests.

<u>Autocorrelation (randomness)</u>: You can use the *Durbin-Watson statistic* to determine if you have significant autocorrelation (this is standard output in SAS). You can also use a *Runs test*.

Non-constant variance (heteroscedasticity): KNNL present two tests that can be used to check for constant variance: 1) the *Modified Levene or Brown-Forsythe test*, and 2) the *Breusch-Pagan test*. The Modified Levene test does not require that the errors be normally distributed, unlike the Breusch-Pagan test. KNNL report that the Modified Levene test is actually quite robust against severe departures from normality. The sample size, though, does need to be large. KNNL present an example of the Modified Levene test on page 117 and an example of the Breusch-Pagan test on page 119.

Lack of Fit Test

A Lack of Fit test is used to determine if a regression function adequately fits the data. This test assumes that the Yi's are independent, normally distributed, and have constant variance. It also requires that replicates at one or more levels of X are available. When replicates are available, the error can be divided into two components: 1) pure error, and 2) lack of fit error.

The pure error component recognizes that replications exist for some levels of X. The sums of squares for the pure error can be expressed as:

$$\mathrm{SSPE} = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 ,$$

where c = number of levels of X and j = number of observations for a given level of X.

The degrees of freedom associated with SSPE = n – c. With this information, we can compute an unbiased estimator of the error variance:

$$\text{MSPE} = \frac{\text{SSPE}}{n - c}.$$

The lack of fit component is simply the difference between the overall error, SSE, and the pure error component, SSPE:

SSLF = SSE – SSPE, or

$$Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \overline{Y}_j \quad + \quad \overline{Y}_j - \hat{Y}_{ij}.$$

SSLF can be expressed directly as:

$$\text{SSLF} = \sum_{j=1}^{c} n_j \left(\overline{Y}_j - \hat{Y}_j\right)^2 \text{, with } c - 2 \text{ degrees of freedom.}$$

The lack of fit aspect can be seen in the difference, $\overline{Y}_j - \hat{Y}_j$. If the difference is small, then you can conclude that the regression model is a better fit than if the difference is large. With SSLF and its degrees of freedom, we can calculate the mean square for the lack of fit:

$$\text{MSLF} = \frac{\text{SSLF}}{c - 2}.$$

Now that we have expressions for two mean squares, we can construct an F-statistic for our lack of fit test:

$$F = \frac{\text{MSLF}}{\text{MSPE}}.$$

We can carry out the test as an ANOVA in the usual fashion. Note: KNNL present the lack of fit test in the context of the general linear test on page 71. Beginning on page 121, they present the Full Model and the Reduced Model, which provide the appropriate sums of squares to construct the F-statistic for the lack of fit test.

Example: Problem 3.15 on page 150.

In this example, a chemist measured the concentration of a solution (Y) over time (X) (n = 15 solutions). The n = 15 solutions were randomly divided into five sets of three, and the five sets were measured after 1, 3, 5, 7, and 9 hours, respectively. Here are the data:

i	X_i	Y_i
1	9	0.07
2	9	0.09
3	9	0.08
4	7	0.16
5	7	0.17
6	7	0.21
7	5	0.49
8	5	0.58
9	5	0.53
10	3	1.22
11	3	1.15
12	3	1.07
13	1	2.84
14	1	2.57
15	1	3.10

Hypotheses

$$\text{Ho} : E(Y_i) = \beta_0 + \beta_1 X_i$$

$$\text{Ha} : E(Y_i) \neq \beta_0 + \beta_1 X_i$$

$\alpha = 0.05$

Decision Rule

If $F \geq F_{1-\alpha,\, c-2,\, n-c} = F_{.95,\, 3,\, 10} = 3.71$, then reject Ho and conclude that the regression function does not adequate fit the data (i.e., a significant lack of fit exists).

Results

Regression Equation: $Y = 2.5753 - 0.324 * X$.

Since $F = 58 >> 3.71$ ($p << 0.0001$), reject Ho and conclude there is a lack of fit.

ANOVA table

Source	df	SS	MS	F	P - value
Regression	1	12.5971	12.5971		
Error	n – 2 = 13	2.9246	0.225		
Lack of Fit	c – 2 = 3	2.7672	0.9224	58.75	< 0.0001
Pure Error	n – c = 10	0.1574	0.0157		
Total	n – 1 = 14	15.5218			

Remedial Measures

If the SLR model does not fit the data, then you have two choices:

1. Find a new model form; i.e., pick a nonlinear model.
2. Transform the data.

Data Transformations

The following transformations on X will often linearize a nonlinear relationship. If they do not adequately work, then you should use a nonlinear regression model, which we will discuss later in the class.

- $X' = \sqrt{X}$
- $X' = \frac{1}{X}$
- $X' = \log X$ or $\ln X$
- $X' = X^2$

The following transformations of Y (similar to the ones above for X) will often stabilize the variance and/or fix non-normality in the errors. Often, these transformations on Y_i will fix non-normality and stabilize the variance simultaneously.

- $Y^{'} = \sqrt{Y}$ ➔ Y ~ Poisson
- $Y^{'} = \arcsin\sqrt{Y}$ ➔ Y ~ Binomial
- $Y^{'} = \frac{1}{Y}$
- $Y^{'} = \log Y \text{ or } \ln Y$ ➔ Y ~ Lognormal.

You can also use weighted regression to stabilize the variance. We will discuss this topic later in Chapter 11. Also, KNNL mention Box-Cox transformations (page 134) as remedial measures for unequal variances, nonnormality, etc. Box-Cox transformations are useful when you don't know exactly which transformation on Y to use. The Box-Cox procedure utilized Maximum Likelihood estimation to identify the appropriate transformation on Y from a family of power functions. They also present an approximation technique that does not require you to minimize the likelihood function since many statistical analysis packages typically do not allow you to use Box-Cox transformations (see page 136).