Examining Residuals in Simple Linear Regression

Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + e_i$ where the error terms $e_1, \ldots, e_n$ are independent normal random variables with mean 0 and standard deviation σ.

For the sake of illustration suppose $\beta_0 = 5.0$, $\beta_1 = 2.0$ and $\sigma = 1.0$. A random set of 18 observations generated according to this model is given below. (Note that all terms have been rounded to two decimal places to make things easier to look at.)

i	X_i	$\mu\{Y_i\|X_i\}$	e_i	Y_i	$\hat{Y}_i = \hat{\mu}\{Y_i\|X_i\}$	$\hat{e}_i = Y_i - \hat{Y}_i$
1	2	5+2*2=9	0.86	9.86	8.85	1.01
2	2	5+2*2=9	0.29	9.29	8.85	0.44
3	2	5+2*2=9	0.00	9.00	8.85	0.15
4	3	5+2*3=11	-1.96	9.04	10.86	-1.82
5	3	5+2*3=11	-0.19	10.81	10.86	-0.05
6	3	5+2*3=11	-0.96	10.04	10.86	-0.82
7	4	5+2*4=13	0.83	13.83	12.87	0.96
8	4	5+2*4=13	0.36	13.36	12.87	0.49
9	4	5+2*4=13	0.66	13.66	12.87	0.79
10	5	5+2*5=15	-1.11	13.89	14.88	-0.99
11	5	5+2*5=15	-1.12	13.88	14.88	-1.00
12	5	5+2*5=15	0.14	15.14	14.88	0.26
13	6	5+2*6=17	-1.37	15.63	16.89	-1.26
14	6	5+2*6=17	-0.35	16.65	16.89	-0.24
15	6	5+2*6=17	1.18	18.18	16.89	1.29
16	7	5+2*7=19	-0.44	18.56	18.90	-0.34
17	7	5+2*7=19	-0.10	18.90	18.90	-0.00
18	7	5+2*7=19	1.18	20.18	18.90	1.28

Using the observed (X_i, Y_i) data points, the least-squares estimates of β_0 and β_1 are

$$\hat{\beta}_0 = 4.83 \text{ and } \hat{\beta}_1 = 2.01,$$

respectively. Thus

$$\hat{Y}_i = \hat{\mu}\{Y_i|X_i\} = 4.83 + 2.01 X_i.$$

The error standard deviation σ was estimated to be $\hat{\sigma} = 0.94$.

If we want to construct prediction intervals for Y values at a particular value of X or conduct tests or compute confidence intervals for β_0, β_1, or $\beta_0 + \beta_1 X$; we need to verify that $e_1, \ldots, e_n$ are indeed independent and normally distributed with mean 0 and constant standard deviation.

In a real problem we don't get to observe the errors $e_1, \ldots, e_n$; but we do observe the residuals $\hat{e}_1, \ldots, \hat{e}_n$. If the residuals appear to be roughly independent and normally distributed with mean 0 and constant standard deviation, we assume the same is roughly true for the errors. Hence we

- make histograms and normal probability plots to check normality of the residuals, and
- make residual plots (residuals vs. fitted values, i.e., $\hat{e}_i$ vs. $\hat{Y}_i$) to see if the standard deviation of the residuals remains relatively constant as the fitted values change.

Residual plots also help us verify whether a linear relationship between the mean of Y and X seems appropriate.

Residual plots can be used with other information to check the assumption that all Y values are independent of one another. The independence assumption is often violated when multiple observations are taken from each of several experimental or observational units. (See example of an experiment with repeated-measures.) When the independence assumption is violated, our estimate of σ may overestimate or underestimate the true value of σ. Thus our tests, confidence intervals, prediction intervals, etc. cannot be trusted. Analysis of dependent data is covered in Statistics 402.

Of all the assumptions of simple linear regression, normality is least important. Close adherence to a normal distribution is only required for accurate prediction intervals.