**Residual analysis**

The discussion below highlights simple linear regression (one predictor) but the approach generalizes directly to more than one indicator variable.

The *basic regression model* has the following conditions:

L: There is a linear relationship between the *means* at each *x* and *y*: E(Y|X=*x*) = 0 + 1*x*

I: The observations (errors) are independent (uncorrelated)

N: The responses follow a normal distribution for each value of *x* Y|X ~Normal

E: The variance of the responses is the same for each value of *x* V(Y|X=*x*) = 

 *yi* = 0 + *xi* + *i* where *i* ~N(0, )

**How do we verify the distribution of the *i*?**

* Can tell a lot from the scatterplot of *y* vs. *x*
* Can tell even more from a scatterplot of the residuals (*ei = yi −* $\hat{y}$*i* ) vs. *x*.

 Though not uncorrelated and have to worry about estimating variability…

Explanatory variable (*x*)

Explanatory variable (*x*)

Response variable (*y*)

















E(Y at *x*)=0+1*x*

**Residual Plots:**

|  |  |
| --- | --- |
| * L: If the relationship between E(Y) and X is linear, there should not be any pattern in the residuals (what’s leftover after you take out the linear component) vs. X graph.
* E: If the variability is constant, the graph of residuals vs. X should show a band of equal width. That is, the distances of the residuals away from the 0 line should be the same across the length of the graph (as you change *x* values). Don’t worry about a few values outside the box.
 | *An almost ideal residual plot:*  |

|  |  |
| --- | --- |
| * N: If the distribution of the response variable at each *x* value is normal, then the residuals themselves should follow a roughly normal distribution.

“It is a matter of determining what constitutes an extreme configuration of residuals relative to the boundaries, and that determination has to be somewhat of an art.” | *Ideal behavior:*  |

***Residuals vs. observation number*** – We want this to look random or else Independence is violated.

Only meaningful if data are in a particular order, e.g., when collected.

Also see Durbin-Watson test.

When you do have multiple predictor variables, it can also be useful to look at the residuals vs. each of these variables individually, especially after checking residuals vs. fits first and finding a problem.

From *Statistical Sleuth* (Ramsey & Schafer, 2002)





**Definition:** There are several methods for *scaling residuals* including (the *hi* are the “leverage” values – a measure of how far the observation is from the mean of the *x* variable ‘cloud’):

* Dividing by *s: ei*/*s*
* The *standardized residual* stdres*i =*  because V(*ei*) = 2(1-*hi*).
	+ Recognizes that the variability in the residuals depends on the *x* values.
	+ Can’t simply divide by *s* because *s* overestimates estimates the variability in the residuals.
	+ The *stdresi* have mean 0 with Var(*stdresi*) = 1.
	+ Values larger than 2 (Minitab) or 3 indicate a large residual for that case.
	+ Minitab (see options under Storage) refers to this as “SRES” or “St Resid”; others *internally studentized residual* *.* In R, save the model and then use rstandard(model). In JMP, under the Response hot-spot Save Columns > Studentized residuals.
* The *studentized residuals* studre*si* = 
	+ *s*(*i*) is the value of *s* for the regression model with the *i*th observation removed
	+ Minitab refers to this as “Deleted *t* residual”; others *externally studentized residual, standardized deletion residual*. For JMP, you can calculate these from the standardized residual *ri*: . In R, save the model and then use rstudent(model).
	+ The stud res*i* have a *t* distribution with *n* − *p* − 2 degrees of freedom
* *PRESS or deleted residuals* identify observations that are not well predicted *e*(*i*) = = *ei*/(1 − *hi*)
	+ Standardizing the PRESS residual equals the standardized residual.

With simple linear regression

*hi* =  *=* 