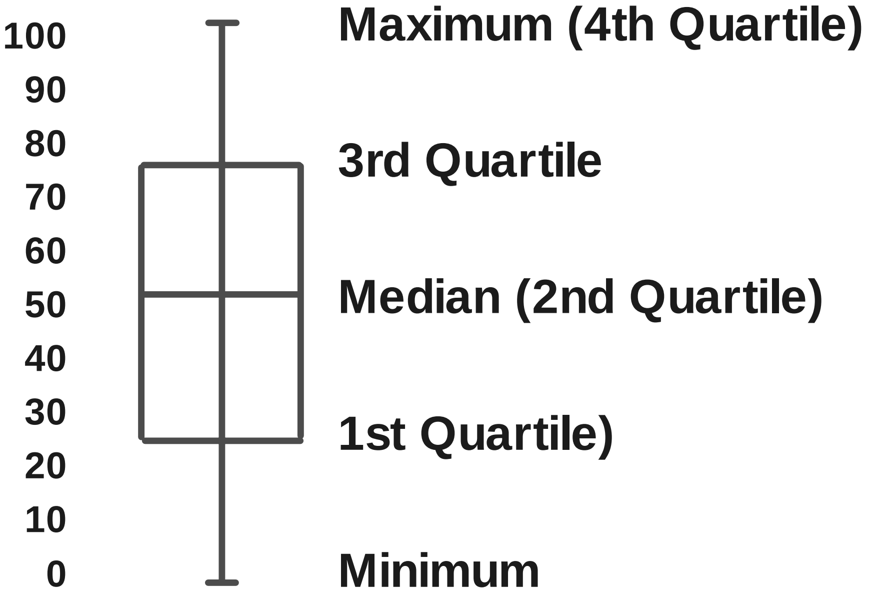
Regression diagnostics – dependent variable Y_i

There are a number of graphic representations which will help with problem

detection and which can be used to obtain a better understanding of the dataset available.

a) Box plot – available in SAS PROC UNIVARIATE

this plot shows the Quartiles : Maximum (100^{th} percentile) and Minimum (0 percentile) values, and the 1^{st} quartile (25^{th} percentile), 2^{nd} (50^{th} percentile or the median) and 3^{rd} quartile (or 75^{th} percentile).



These statistics are “non-parametric" in that they are not influenced by the distribution, but they will help get a feel for the distribution.

1) if the median is centered, and the box centered between the maximum and minimum, then the data is symmetric

2) if the median is NOT centered, this indicates a skew in the data

3) if the median is centered in the box, and the box IS NOT centered between the maximum and minimum, this may indicate an outlier

4) if neither the median nor the box is centered, this is a pretty good indicator of skewness.

Example from Freund and Wilson - Tree Weight on Length done as linear

```
Boxplot
   0    beyond 1.5 interquartile distances

   |    1.5 interquartiles above third quartile
   |
   |
   |
+-----+ third quartile
|  +  | mean
*-----* median or second quartile
+-----+ first quartile
   |
   |
   |
   |    minimum
```

The interquartile distance is the distance from the 25th to the 75th percentile.

The wiskers (vertical bars) extend out 1.5 interquartile distances from the quartiles.

Outside the wiskers, values are repsented with 0 out to 3 interquartile distances

Beyond 3 interquartile distances values are repsented with asterisks

b) Time plot – this plot could be obtained in SAS with PROC PLOT, but some order variable (eg OBS) is required

where applicable, this can be a useful indicator of variation which may not be accounted for by the regression line. The order in which the data was acquired may also be useful as an indicator.

1) very often there is some effect of “time" in the model. The time plot will aid in the determination of such and effect.

2) The data may not have been gathered at random, and some aspects of this can also be detected.

3) If the data must be gathered over time, try to randomize the dependent variable X_i when this can be controlled over time to avoid confounding.

for example, we wish to measure the amount of lactose in a fish's blood, and regress this on the amount of time he spent swimming in a 1 foot/sec current created in an artificial stream. The dependent variable is “time", but not in the sense of a time plot. The order here may be important, as lactose may change if there are slight shifting of the current speed over time, or a buildup of metabolic wastes in the water which could effect lactose levels. Don't do all the short times first, and all the long times later.

c) Stem and leaf plot – available in SAS PROC UNIVARIATE

This plot is useful to give an extra dimension to the information obtained from the Box plot.

1) as with the Box plot, we get an additional idea of symmetry

2) This plot will also indicate bimodality or polymodality, which the box plot will not.

d) Dot plot – similar to a histogram in SAS, PROC CHART

This plot is similar to a stem and leaf plot, plotted horizontally instead of vertically.

1) as with the Box plot, we get an additional idea of symmetry

2) This plot will also indicate bimodality or polymodality, which the box plot

Direct observation of Y_i is frequently not a useful undertaking. Y_i is assumed to be normally distributed at EACH value of X_i. Even perfectly normally distributed data could appear polymodal or asymmetric if the data is taken as widely separate X_i values or if most of the data is taken a high or low values of X_i.

Observations of e_i are generally more useful. Y_i alone can be misleading.

Residual Analysis and RESIDUAL PLOTS –

help in determining if the ASSUMPTIONS are met, and if the model is correct

Given the term population residuals,

$$\epsilon_i = Y_i - E(Y_i)$$

which are assumed to be NIDrv(0,σ^2).

These are then estimated by the observed deviations,

$$e_i = Y_i - \hat{Y}_i$$

where we know that

$$\bar{e}_i = \frac{\Sigma e_i}{n} = 0 \text{ since } \Sigma e_i = 0$$

and we define

$$\frac{\Sigma(e_i - \bar{e})}{n-2} = \frac{\Sigma(e_i)}{n-2} = \frac{SSE}{n-2} = MSE$$

where,

$$E(MSE) = \sigma^2$$

Note that the residuals measure the actual deviation of the point from the regression line, and as a result will have the same units as the variable Y_i. For example, if we regress vial breakage on transfers, then a residual of 2 would be 2 vials.

For some examinations, residuals are standardized to a “Z" distribution. Since the residuals are assumed to be normal, we know that (for large samples) about 65% of the standardized residuals would fall between -1 and 1, about 95% would fall between -1.96 and 1.96, and about 99% would fall between -2.576 and 2.576.

If we standardize the residuals, we can examine a residual and see if it appears to be “unusually large" or within the usual bounds.

Standardized residuals are calculated as

$$\text{standard } e_i = \frac{e_i - \bar{e}}{\hat{\sigma}} = \frac{e_i}{\sqrt{MSE}}$$

Residual diagnostics : the plots previously mentioned can also be used for residuals with similar interpretations.

a) Box plot – shows the Quartiles

b) Time plot – or the order the observations were taken in

Weld example : later welds stronger due to some learning process. Could not determine if not done in random order, “learning" would be fitted in with the functional relationship

c) Stem and leaf plot – can show modes

d) Dot plot – similar information stem and leaf plot

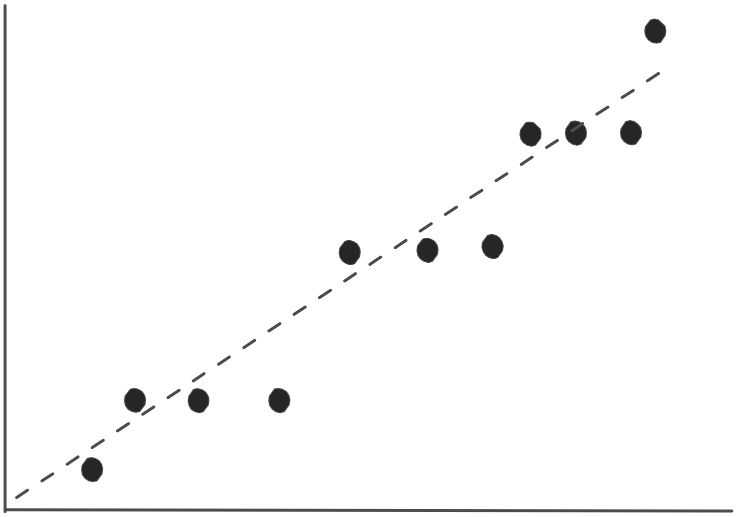
e) Normal Probability Plot – available in SAS in PROC UNIVARIATE

this is a plot of the values of the residuals of the ordered observations from smallest to largest. Untransformed, this should be sigmoid for a normal distribution.

Usually the data is transformed so that a normal distribution, when plotted, would be a straight line. The transformation for each residual is

$$\sqrt{\text{MSE}}*Z*\left[\frac{i-0.375}{n+0.25}\right]$$

where, n is the number of observations, and
i is the observation number from smallest to largest

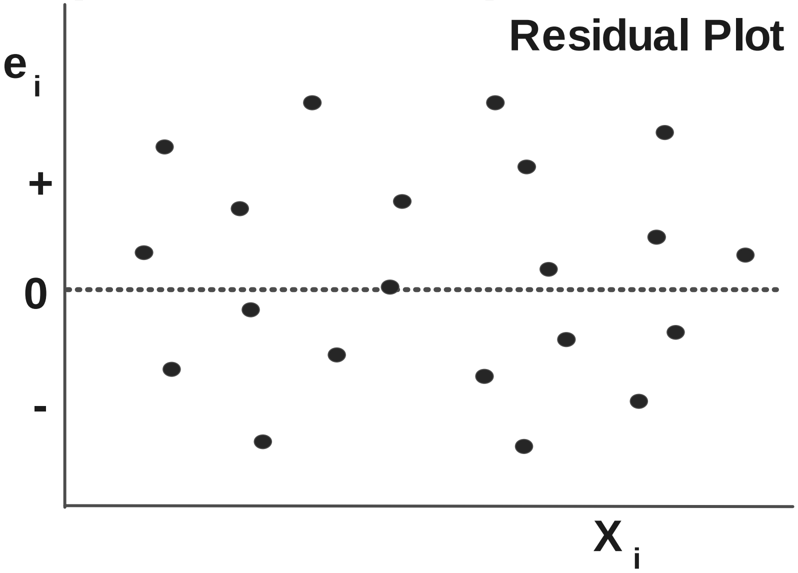


f) Residual plot – plot the observed residuals against the value of X. For the general case, where various independent values are included, plot the value of e_i against $\hat{Y}$

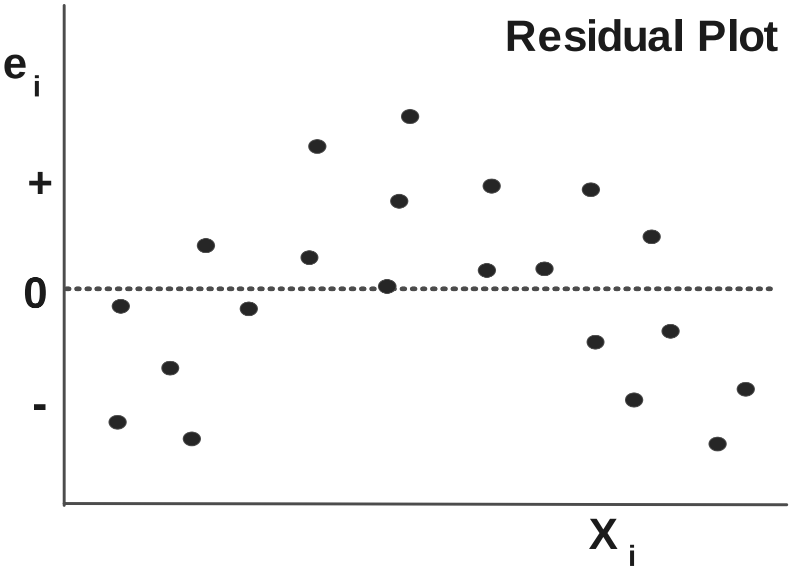
Residuals can be plotted on some other X_i which is a candidate for multiple regression

Residual plots : USUALLY PLOT RESIDUALS ON X FOR A SINGLE INDEPENDENT VARIABLE (**OTHERWISE USE $\hat{Y}$**)

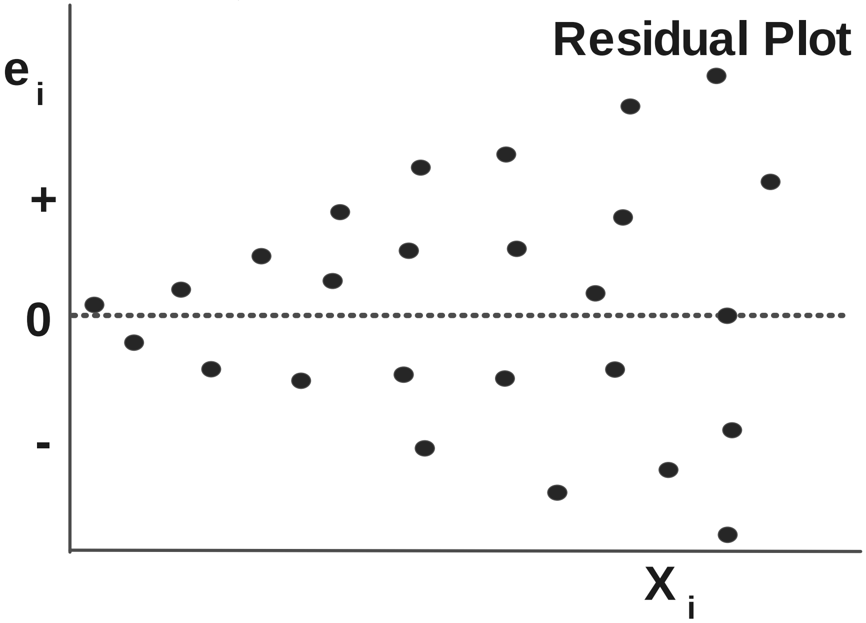
Normal, expected results of a residual plot



POSSIBLE PROBLEMS WE CAN DETECT
a) regression not linear

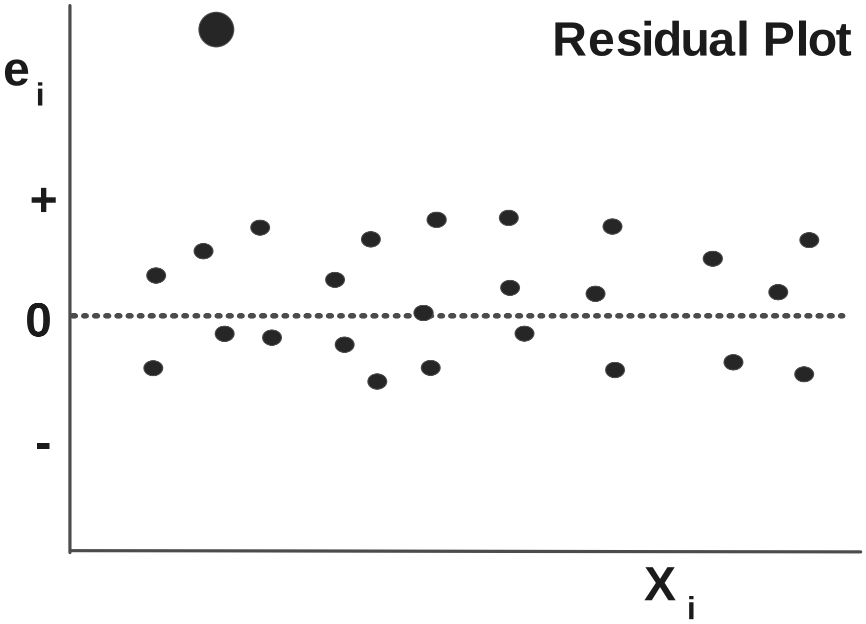


b) non - homogeneous variance - or non-constancy of errors could also be ANACOV (discussed below)

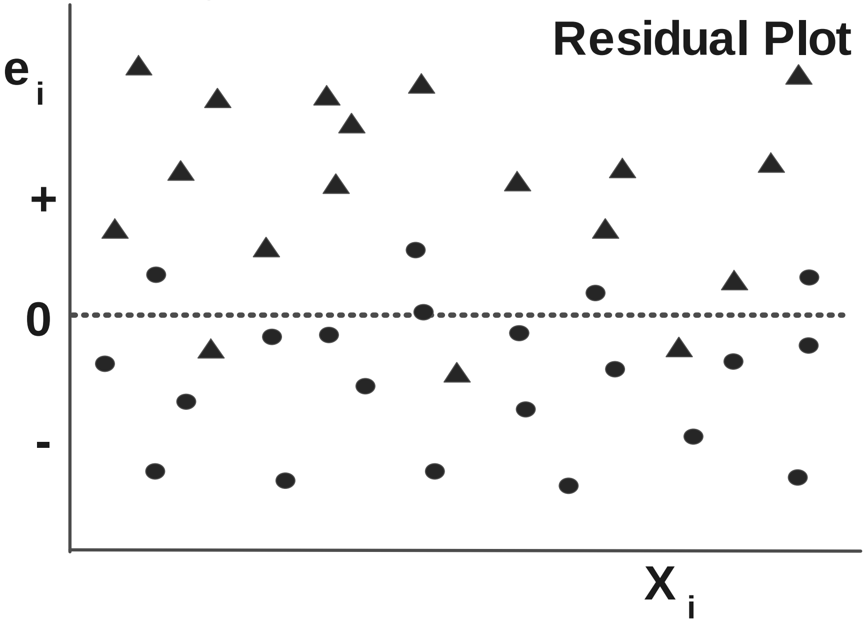


Note that more dense data can also result in a wider distribution, so if there are only a few values at one end, and many at the other, this may also look like non-homogeneous data.

c) outliers



e) missing variables : categorical or class variables



5) RESIDUAL PLOTS IN SAS

```
DATA ONE;   INPUT Y X AGE;
PROC GLM DATA = ONE;    ID AGE;
MODEL  Y = X / P;
```

Use "P" only to obtain a list of predicted values, values can be output for plotting without listing

```
OUTPUT OUT = TWO     RESIDUAL = E     PREDICTED = YHAT;
PROC PLOT DATA = TWO;
PLOT  E * X / VREF = 0;
PLOT  P * X='P' Y * X='O' / OVERLAY;
PLOT  Y * X=AGE / OVERLAY;
```

Overview of Residual Diagnostics

Problem	Detection
1) Not linear	Residual plot, LOF (soon)
2) Nonhomogeneous Variance	Res plot, F test of high/low, Rank correlation
3) Not independent	Residual plot
4) Outliers	Box plot, stem & leaf plot, residual plot
5) Not normally distributed	Box plot, stem & leaf plot, residual plot
6) Variables omitted	Residual plot

Additional tests and cures (most of which will come up later) :

1) Not linear : Transform to curve, polynomial, multiple regression, etc.

2) Non-homogeneous variance :

a) Visual evidence can be compelling

b) Fit lower half of X_i and upper half of X_i, do a two tailed F test of

$F = \frac{\text{MSE upper}}{\text{MSE lower}}$ (or larger over smaller with $\frac{\alpha}{2}$)

unless there is some theoretical justification of a one tailed test

c) Rank test of $|e_i|$ on X_i, to determine if the size of the residual is “correlated" to the value of X_i.

This requires first ranking $|e_i|$. There are various types of test (Kendall, Spearman, and others). The interpretation is similar to the usual Pearson Correlation, but these test are “distribution free".

d) some transformed models and other transformations will address this problem

3) Lack of independence or randomness :

a) runs test on time ordered, or in the order that observations are taken
The idea is to test the hypothesis that a $+$ or $-$ residual is equally likely. If the data is not taken randomly, then we may get long series of $+$'s or $-$'s in a row. Runs are located by finding all strings of like sign (incl runs of 1);

Need a table of runs statistics : Table indicates that with 10pluses and 10 minus, we should have between 6 and 16 runs

eg. To few runs : only 4 in 20 observations

+++++ – – – – – +++++ – – – – –

To many runs 18 in 20 observations

+ – ++ – – + – + – + – + – + – + – + –

b) Durban-Watson discussed later

4) Outliers : a number of additional diagnostics covered later,

Quick & easy test : Omit the point, and refit the regression without the point. Calculate the CLI for the value of X for the suspect observation. View the suspect observation as a new point, and calculate the probability of that value occurring.

5) Tests of Normality : Generally need the predicted value of e_i under the assumption of normality. These are calculated as in the Normality plot above.

a) χ^2 goodness of fit : $\sum_{i=1}^{n} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

b) Kolmogorov - Smirnov test : calculates “Supremum" (basically the largest observed difference from something like normality plot,

need table of KS test statistic
This test used to be available in, and is still used for n>2000 SAS

c) available now in SAS for $n \leq 2000$

Shapiro-Wilk statistic (W):

Calculated as a a ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator.

This value falls between 0 and 1, where small values lead to rejection of the hypothesis of normality