Regression through the Origin

The model is

$$Y_i = \beta_1 X_i + \epsilon_i$$

where;

β_1 is a parameter

X_i are known constants (no error)

ϵ_i are assumed NID$(0,\sigma^2)$

Following a development similar to that of the SLR,

a) $E(Y) = \beta_1 X$

b) the least squares estimator minimizes $Q = \Sigma(Y_i - \beta_1 X_i)^2$

c) the normal equation is $\Sigma X_i(Y_i - b_1 X_i) = 0$

and the estimator for the single parameter is $b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}$
which is an unbiased estimator.

NOTE that this is an UNCORRECTED SS of X and an UNCORRECTED crossproduct of X and Y. Since there is "no intercept", there is no correction factor and no adjustment for the mean.

Also not that the reference to "no intercept" refers to the MODEL, not to a graph. Graphically, the intercept is the origin.

The resulting prediction equation is

$$\hat{Y} = b_1 X$$

and the residuals are calculated as usual,

$$e_i = Y_i - b_1 X_i$$

And an unbiased estimator of σ^2 is given by (note that df $= n - 1$)

$$MSE = \frac{\Sigma e_i^2}{n-1} = \frac{\Sigma(Y_i - \hat{Y})^2}{n-1} = \frac{\Sigma(Y_i - b_1 X_i)^2}{n-1}$$

All development of confidence intervals parallels that of the SLR. The variances for each of the likely candidates for confidence intervals and tests of hypothesis are,

β_1 $\quad s^2_{b_1} = \frac{MSE}{\Sigma X_i^2}$

$\hat{Y}_h$ $\quad s^2_{Y_h} = \frac{X_h^2 * MSE}{\Sigma X_i^2}$

$\hat{Y}_{hnew}$ $\quad s^2_{Y_{hnew}} = MSE\left(1 - \frac{X_h^2}{\Sigma X_i^2}\right)$

There are a few curious things about regression through the origin,

a) The residuals will not sum to zero

b) The confidence band for the regression line and for individual observations do not widen from the center of the line (ie. $\bar{X}$), but instead expand out from the origin where the variance is zero.

Advice :

a) Before I would force a regression line through the origin, I would recommend that two conditions be met, (in addition to checking residuals and other diagnostics)

1) that the regression line be one that theoretically SHOULD go through the origin, and

2) that a SLR be fitted first, and the hypothesis that $\beta_0 = 0$ be tested

because, even if the line DOES go through the origin, it may be curved. Most of the cases which call for the previously discussed model

$$Y_i = b_0 X_i^{b_1} e_i \qquad \log - \log \text{ model}$$

call for the line to go through the origin. However, small segments of data may appear linear, but if fitted with SLR will not go through the origin.

c) For regression through the origin, b_1 estimates the ratio $\frac{Y}{X}$. ACTUALLY there are THREE UNBIASED ESTIMATORS of this RATIO

$$\frac{Y}{X}$$

(1) all can be done as regressions through the origin

– usual assumptions apply, except homogeneity of variance assumption differs with the model

(a) if we can assume HOMOGENEOUS VARIANCE for the model then

$$Y_i = b_1 X_i + e_i \quad \text{and}$$

$$b_1 = \frac{\Sigma XY}{\Sigma X^2} \text{ as before, UNBIASED}$$

$$S^2_{b_1} = \left[\frac{\Sigma Y^2 - \frac{(\Sigma XY)^2}{\Sigma X}}{(n-1)\,\Sigma X^2}\right]$$

(b) if the variance is not homogeneous, one possibility is that the Variance is proportional to X. Then estimate

$$b_1 = \frac{\Sigma Y}{\Sigma X} \text{ or } \frac{\bar{Y}}{\bar{X}} \text{ is } \textbf{also an UNBIASED estimate}$$

$$S^2_{b_1} = \left[\frac{\Sigma \frac{Y^2}{X} - \frac{(\Sigma Y)^2}{\Sigma X}}{(n-1)\,\Sigma X^2}\right]$$

(c) variance not homogeneous, another possibility is that the Variance is proportional to X^2, or the Standard deviation is proportional to X.

$$b_1 = \frac{\sum\left[\frac{Y_i}{X_i}\right]}{n} \quad \text{is } \textbf{also an UNBIASED estimate}$$

$$S^2_{b_1} = \left[\frac{\Sigma\left[\frac{Y_i}{X_i}\right]^2 - \left[\Sigma\frac{Y_i}{X_i}\right]^2}{n\,(n-1)}\right]$$

ALL THREE OF THE RATIO ESTIMATES ARE UNBIASED – which is correct depends on the nature of the variance

d) NOTES on regression through origin

1) can be used for catch per unit of standard effort (or per unit area) when effort (or unit area) not a fixed quantity

2) DO NOT USE $R^2 = \frac{SSR}{SSTOTAL}$

When UNCORRECTED SUMMATIONS are used – this is only useful for comparisons within a data set when all comparisons are based on a common SSTOTAL.

R^2 are normally corrected for the mean, so the value of the mean and CF are irrelevant. However, this is not true in the calculations above.

3) also used as **Direct proportions model** – this is one model which we will discuss as a candidate for biological models. It's use implies that

Y = “a constant fraction or multiple of" X

4) in SAS a direct proportion model can be run as

PROC GLM; MODEL Y = X / NOINT;

SAS will provide a warning that the SSTotal is UNCORRECTED when the NOINT option is used.