WEIGHTED LEAST SQUARES - used to give more emphasis to selected points in the analysis

What are weighted least squares?

Recall, in OLS we minimize $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

$$\text{or} \quad Q = (\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$$

In weighted least squares, we minimize

$$Q = \sum_i \sum_j e_i^2 = \sum_i \sum_j w_i [Y_i - \bar{Y} - b_1 X_i)]^2$$

The normal equations become

$$b_0 \Sigma w_i + b_1 \Sigma w_i X_i = \Sigma w_i Y_i$$

$$b_0 \Sigma w_i X_i + b_1 \Sigma w_i X_i^2 = \Sigma w_i X_i Y_i$$

For the intermediate calculations we get (where w_i is the weight)

$$\sum_i^t \sum_j^{n_i} w_i X_i, \ \sum_i \sum_j w_i Y_{ij}, \ \sum_i \sum_j w_i X_i Y_{ij}, \ \sum_i \sum_j w_i X_i^2, \ \sum_i \sum_j w_i Y_{ij}^2, \ \sum_i \sum_j w_i$$

Calculation of the corrected sum of squares is

$$\sum_i \sum_j y_{ij}^2 = \sum_i \sum_j w_i (Y_{ij} - \bar{Y}_{..})^2 = \sum_i \sum_j w_i Y_{ij}^2 - \frac{(\sum_i \sum_j w_i Y_{ij})^2}{\sum_i \sum_j w_i}$$

the slope is

$$\hat{\beta}_1 = b_1 = \frac{\sum_i \sum_j x_i y_{ij}}{\sum_i \sum_j x_i^2} = \frac{\sum_i \sum_j w_i Y_{ij} X_i - \frac{(\sum_i \sum_j w_i Y_{ij} X_i)}{\Sigma w_i}}{\sum_i \sum_j w_i X_i^2 - \frac{(\sum_i \sum_j w_i X_i)^2}{\sum_i \sum_j w_i}}$$

the intercept is calculated with $\bar{X}$ and $\bar{Y}$ as usual, but these are calculated as

$$\bar{Y}_{..} = \frac{\sum_i \sum_j w_i Y_{ij}}{\sum_i \sum_j w_i} \quad \text{and} \quad \bar{X}_{..} = \frac{\sum_i \sum_j w_i X_i}{\sum_i \sum_j w_i}$$

It the weights are 1, then all results are the same as OLS

The variance is $\sigma_i^2 = \sigma_{\epsilon_i}^2 = \sigma_{Y_i}^2 = \frac{\sigma^2}{w_i}$

WEIGHTED LEAST SQUARES handout

In OLS we minimize $Q = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$ or $Q = (\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta})$

In weighted least squares, we minimize

$$Q = \sum_i e_i^2 = \sum_i w_i [Y_i - b_0 - b_1 X_i]^2$$

If the weights are 1, then all results are the same as OLS

The normal equations become

$$b_0 \Sigma w_i + b_1 \Sigma w_i X_i = \Sigma w_i Y_i$$
$$b_0 \Sigma w_i X_i + b_1 \Sigma w_i X_i^2 = \Sigma w_i X_i Y_i$$

For the intermediate calculations we get (where w_i is the weight)

$$\sum_i^t \sum_j^{n_i} w_i X_i,\ \sum_i \sum_j w_i Y_{ij},\ \sum_i \sum_j w_i X_i Y_{ij},\ \sum_i \sum_j w_i X_i^2,\ \sum_i \sum_j w_i Y_{ij}^2,\ \sum_i \sum_j w_i$$

Calculation of the corrected sum of squares is

$$\sum_i \sum_j y_{ij}^2 = \sum_i \sum_j w_i (Y_{ij} - \bar{Y}_{..})^2 = \sum_i \sum_j w_i Y_{ij}^2 - \frac{(\sum_i \sum_j w_i Y_{ij})^2}{\sum_i \sum_j w_i}$$

the slope is

$$\hat{\beta}_1 = b_1 = \frac{\sum_i \sum_j w_i Y_{ij} X_i - \frac{(\sum_i \sum_j w_i Y_{ij} X_i)}{\Sigma w_i}}{\sum_i \sum_j w_i X_i^2 - \frac{(\sum_i \sum_j w_i X_i)^2}{\sum_i \sum_j w_i}}$$

the intercept is calculated with $\bar{X}$ and $\bar{Y}$ as usual, but the means are calculated as

$$\bar{Y}_{..} = \frac{\sum_i \sum_j w_i Y_{ij}}{\sum_i \sum_j w_i} \quad \text{and} \quad \bar{X}_{..} = \frac{\sum_i \sum_j w_i X_i}{\sum_i \sum_j w_i}$$

The variance is $\sigma_i^2 = \sigma_{\epsilon_i}^2 = \sigma_{Y_i}^2 = \frac{\sigma^2}{w_i}$

For multiple regression

$$Q = \sum_i \sum_j e_i^2 = \sum_i \sum_j w_i [Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - \ldots - b_{p-1} X_{p-1,i})]^2$$

the normal equations $(X'WX)B = (X'WY)$

the regression coefficients $B = (X'WX)^{-1}(X'WY)$

the variance-covariance matrix; $\sigma_b^2 = \sigma^2 (X'WX)^{-1}$

σ^2 is estimated by MSE_w, based on weighted deviations; $MSE_w = \frac{\Sigma w_i (Y_i - \hat{Y}_i)^2}{n-p}$

In a multiple regression, we minimize

$$Q = \sum_i\sum_j e_i^2 = \sum_i\sum_j w_i[Y_i - \bar{Y} - b_1X_{1i} - b_2X_{2i} - \ldots - b_{p-1}X_{p-1,i})]^2$$

The matrix containing weights is $W_{nxn} = \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & w_n \end{bmatrix}$

The matrix equations for the regression solutions are,

the normal equations $(X'WX)B = (X'WY)$

the regression coefficients $B = (X'WX)^{-1}(X'WY)$

the variance-covariance matrix; $\sigma_b^2 = \sigma^2(X'WX)^{-1}$

where σ^2 is estimated by MSE_w, based on weighted deviations

$$MSE_w = \frac{\Sigma w_i(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\Sigma w_i e_i^2}{n-p}$$

Estimating weights

For ANOVA this may be relatively easy if enough observations are available in each group. Variances can be estimated directly.

However, in regression we usually have a "smooth" function of changing residuals with changing X_i values.

To estimate these we note that σ_i can be estimated by the residuals $|e_i|$.

We can then estimate the residuals with OLS, and regress the residuals (squared or unsquared) on a suitable variable (usually one of the X_i variables or $\hat{Y}_i$).

The procedure can be iterated. That is, if the new regressions coefficients differ substancially from the old, the residuals can be estimated again from the weighted regression and the weights can be calcalated again and the regression fit again with the new weights.

Weighting may be used for various cases

1) One common use is to adjust for non-homogeneous variance.

One common approach is assume that the variance is non-homogeneous because it si a function of X_i.

Then we must determine what that function is, for simple linear regression, the function is commonly

$$\sigma_i^2 = \sigma^2 X_i, \qquad \sigma_i^2 = \sigma^2 X_i^2, \qquad \sigma_i^2 = \sigma^2 \sqrt{X_i}$$

to adjust for this, we would weight by the inverse of the function

2) Another common case is where the function is not known, but the data can be subset into smaller groups. These may be separate samples, or they may be cells from an analysis of variance (not regression, but weighting also works for (ANOVA).

Then we determine a value of σ_i^2 for each cell i, and the weighting function is the reciprocal (inverse) of the variance for each subgroup $\left[\frac{1}{\sigma_i^2}\right]$

3) The values analyzed are means, then mean will differ between “observations" if the sample size are not equal. Since we may still assume that the variance of the original observations is homogeneous, The the variance of each point is a simple function of n_i (where the variance of a mean is $\frac{\sigma^2}{n}$, or $\sigma^2 \frac{1}{n}$)

we would weight by the inverse of the coefficient, of simply n

If we could not assume that the variances were equal then for each mean we would have the variance of $\frac{\sigma_i^2}{n_i}$, and the weight would be $\frac{n_i}{\sigma_i^2}$

NOTE: the actual value of a weight is not important, only the proportional value between weights, so all weights could be multiplied or divided by some constant value. Some people recommend weighting by $\frac{n_i}{\sigma_i^2}$ without concern for the homogeneity of variance. Since all σ_i^2 are equal to σ^2 when the variance is homogeneous, this is the same as taking all n_i weights, and multiplying by a constant $\frac{1}{\sigma^2}$

8. ROBUST REGRESSION – There are many regressions developed that call themselves "robust". Some are based on the median deviation, others on trimmed analyses. The one we will discuss uses weighted least squares, and is called ITERATIVELY REWEIGHTED LEAST SQUARES.

a) ordinary least squares is considered "robust",

but is sensitive to points which deviate greatly from the line

this occurs –
(1) with data which is not normal or symmetrical
– data with a large tail
(2) when there are outliers (contamination)

b) ROBUST REGRESSION is basically weighted least squares where the where the weights are an inverse function of the residuals
(1) points within a range "close" to the regression line get a weight of 1 if all weights are 1 then the result is ordinary least squares regression

(2) points outside the range get a weight less than 1

(3) ordinary least squares minimize

$$\Sigma(Y_{ij} - X_i\beta)^2 = \Sigma(Y_{ij} - \hat{Y}_{..})^2 = (\text{residual})^2$$

robust regression minimizes

$$\frac{\Sigma w_i[(Y_{ij} - X_i\beta)^2}{c\sigma}$$

where w_i is some function of the standardized residual and w_i should be chosen to minimize the influence of **large residuals**. The w_i chosen is a weight, and

if the proper function is chosen, the calculations may be done as a weighted least squares, minimizing

$$\Sigma w_{ij}(Y_{ij} - X_i\beta)^2$$

(4) the function chosen by Huber was chosen such that if the residual is small ($\leq 1.345\hat{\sigma}$) there is no weight

First we define a robust estimate of variance as called the median absolute deviation (MAD) such that MAD = $\frac{median\{|e_i - median(e_i)|\}}{0.6745}$, where the constant 0.6745 will give an unbiased estimate of σ for independent observations from a normal distribution.

MAD is an alternative estimate of $\sqrt{\text{MSE}}$

Then we define a scaled residual as μ, where $\mu = \frac{e_i}{MAD}$

then weight $w_{ij} = 1$ if $|e_{ij}| \leq 1.345\hat{\sigma}$

$w_{ij} = \frac{1.345}{|\mu|}$ for $|e_{ij}| > 1.345\hat{\sigma}$

where 1.345 is called the tuning constant, and is chosen to make the technique robust and 95% efficient. Efficiency refers to the ratio of variances from this model relative to normal regression models.

(5) The solution is iterative

(a) do regression, obtain e_{ij} values

(b) compute the weights

(c) run the weighted least squares analysis using the weights

(d) iteratively recalculate the weights and rerun the weighted least squares analysis – UNTIL

(1) there is no change in the regression coefficients
– to some predetermined level of precision

(2) note that the solution may not be the one with the minimum sum of squares deviations

(6) PROBLEMS

(a) this is not an ordinary least squares, but the hypotheses are likely to be tested as if they were

the distributional properties of the estimator are not well documented

(b) Schreuder, Boos and Hafley (Forest Resources Inventories Workshop Proc., Colorado State U., 1979)

suggest running both regressions

if "similar", use ordinary least squares

if "different" find out why; if because of an outlier then robust regression is "probably" better

they further suggest using robust regression as "a tool for analysis, not a cure all for bad data"

The technique is useful for detecting outlierS.