Statistical Data Mining
MATH 4720
Lecture 7
Weighted Least Squares

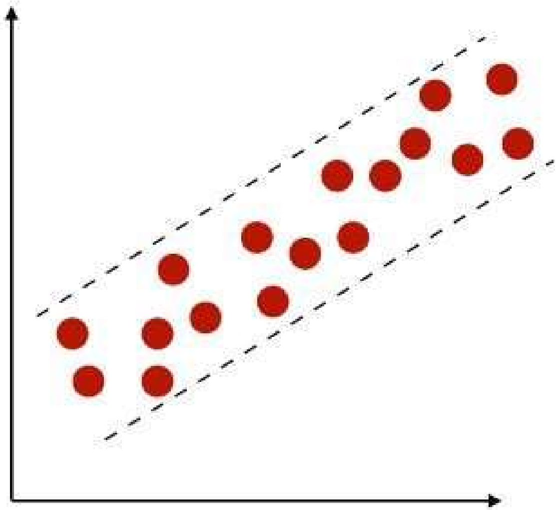
Gaurav Gupta

GEH A401

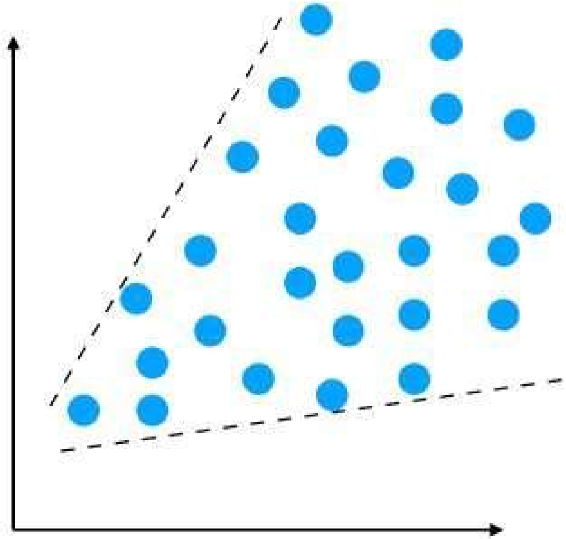
Recall

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^{n} \epsilon_i^2 = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

?



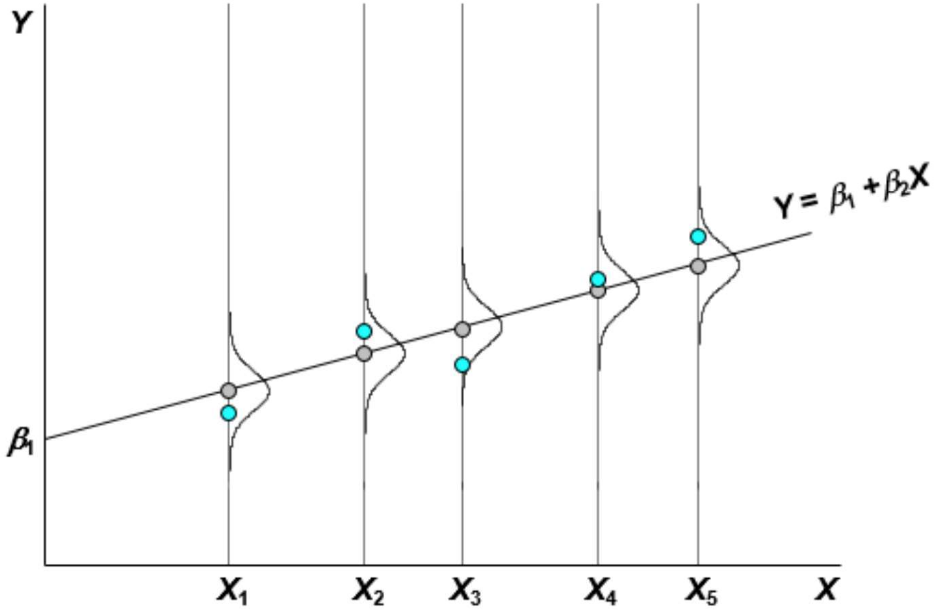
Homoscedasticity

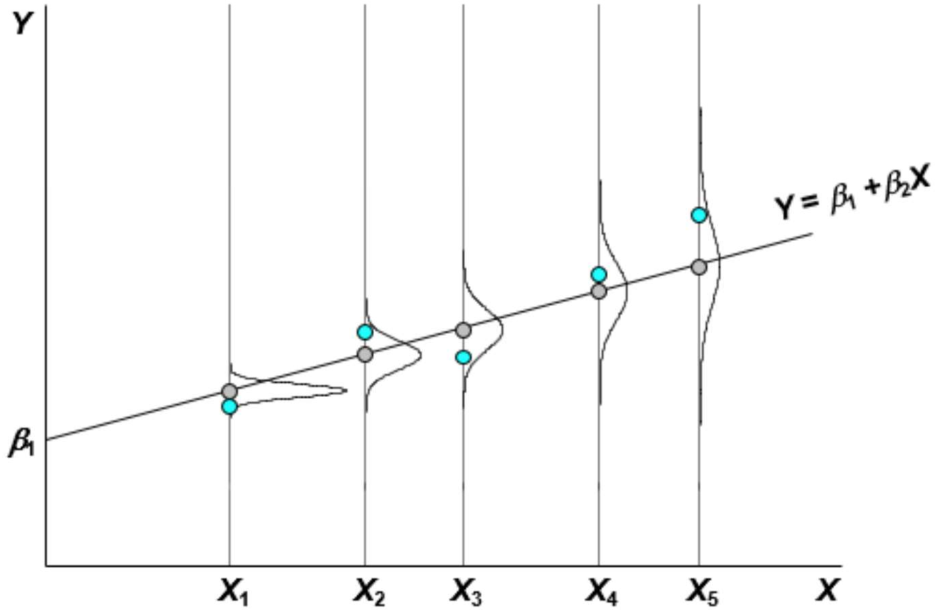


Heteroscedasticity

- **Homoscedasticity:** $\sigma^2_{\epsilon_i} = \sigma^2_{\epsilon}$ the same for all observations
- **Heteroscedasticity:** $\sigma^2_{\epsilon_i}$ is not the same for all observations.

Visual demonstration of homoscedasticity and heteroscedasticity





Concerns:

The variances of the regression coefficients:

- No heteroscedasticity: Lowest variances of all the unbiased estimators
- With Heteroscedasticity: Find other estimators that have smaller variances and are still unbiased.

The estimators of the standard errors of the regression coefficients will be wrong and, as a consequence, the t-tests/F tests will be invalid. It is quite likely that the standard errors will be underestimated, so the t statistics will be overestimated and you will have a misleading impression of the precision of your regression coefficients.

You may be led to believe that a coefficient is significantly different from 0, at a given significance level, when, in fact, it is not.

Statistical tests to detect heteroscedasticity

- The Spearman Rank Correlation Test
- The Goldfeld–Quandt Test
- The Glejser Test
- The Breusch-Pagan test
- The White test

Our test: plotting the residual against the predicted response variable

OLS VS WLS

Ordinary least squares (OLS) does not discriminate between the quality of the observations, giving equal weight to each, irrespective of whether they are good or poor guides to the location of the line.

Thus, it may be concluded that if we can find a way of assigning more weight to high-quality observations and less to the unreliable ones, we are likely to obtain a better fit.

WLS works by incorporating extra nonnegative constants (weights) associated with each data point into the fitting criterion.

Weighted Least Squares

- OLS assumes that there is constant variance in the errors (which is called **homoscedasticity**).
- **Weighted least squares** can be used as one method when the ordinary least squares assumption of constant variance in the errors is violated (which is called **heteroscedasticity).**

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon^*,$$

where $\epsilon*$ is assumed to be (multivariate) normally distributed with mean vector **0** and nonconstant variance-covariance matrix

$$\begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}$$

Weights

Suppose that the variance is not constant:

$$\mathrm{var}(Y_i) = \sigma_i^2.$$

If we use weights

$$w_i \propto \frac{1}{\sigma_i^2},$$

the WLS estimates have smaller standard errors than the ordinary least squares (OLS) estimates.

In fact, using weights proportional to $1/\sigma_i^2$ is optimal: no other weights give smaller standard errors.

When you specify weights, regression software calculates standard errors on the assumption that they are proportional to $1/\sigma_i^2$.

As OLS is unbiased so weights from the residuals of an OLS regression is reasonable, even in the presence of heteroscedasticity.

Nonetheless, those weights are contingent on the original model, and may change the fit of the subsequent WLS model. Thus, we should check our results by comparing the estimated betas from the two regressions. If they are very similar, then it's. If the WLS coefficients diverge from the OLS ones, we should use the WLS estimates to compute residuals manually (the reported residuals from the WLS fit will take the weights into account). Having calculated a new set of residuals, determine the weights again and use the new weights in a second WLS regression. This process should be repeated until two sets of estimated betas are sufficiently similar (even doing this once is uncommon, though).

General Weighted Least Squares Solution

- In Case of OLS, we minimize the mean squared error

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i\beta)^2$$

- The solution is of course

$$\hat{\beta}_{OLS} = \left(X^T X\right)^{-1} X^T Y$$

- We could instead minimize the weighted mean squared error,

$$WMSE(\beta, w_1, w_2 \ldots\ldots w_n) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \boldsymbol{X}_i\beta)^2$$

$$\text{WMSE} \quad = \frac{1}{n}(\mathrm{Y} - \mathrm{X}\beta)^T w(Y - X\beta)$$
$$= \frac{1}{n}(Y^T wY - Y^T wX\beta - \beta^T X^T wY + \beta^T X^T wX\beta)$$

Differentiating with respect to β, we get as the gradient

$$\nabla WMSE = \frac{2}{n}(-X^T wY + X^T wX\beta)$$

Setting this to zero and solving,

$$\hat{\beta}_{WLS} = \left(X^T wX\right)^{-1} X^T wY$$

Weighted Least Squares as a Transformation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\text{Var}(\varepsilon_i) = x_i^2 \sigma^2$.

* We can transform this into a regular least squares problem by taking

$$y_i' = \frac{y_i}{x_i} \quad x_i' = \frac{1}{x_i} \quad \varepsilon_i' = \frac{\varepsilon_i}{x_i}.$$

* Then the model is

$$y_i' = \beta_1 + \beta_0 x_i' + \varepsilon_i'$$

where $\text{Var}(\varepsilon_i') = \sigma^2$.

Weighted Least Squares as a Transformation

* The residual sum of squares for the transformed model is

$$
\begin{aligned}
S_1(\beta_0, \beta_1) &= \sum_{i=1}^{n} (y_i' - \beta_1 - \beta_0 x_i')^2 \\
&= \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \beta_1 - \beta_0 \frac{1}{x_i} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{1}{x_i^2} \right) (y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}
$$

* This is the weighted residual sum of squares with $w_i = 1/x_i^2$.

* Hence the weighted least squares solution is the same as the regular least squares solution of the transformed model.

* In general suppose we have the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$$

where $\mathsf{Var}(\varepsilon) = \boldsymbol{W}^{-1}\sigma^2$.

* Let $\boldsymbol{W}^{1/2}$ be a diagonal matrix with diagonal entries equal to $\sqrt{w_i}$.

* Then we have $\mathsf{Var}(\boldsymbol{W}^{1/2}\varepsilon) = \sigma^2\boldsymbol{I}_n$.

* Hence we consider the transformation

$$\boldsymbol{Y}' \;=\; \boldsymbol{W}^{1/2}\boldsymbol{Y} \quad \boldsymbol{X}' \;=\; \boldsymbol{W}^{1/2}\boldsymbol{X} \quad \varepsilon' \;=\; \boldsymbol{W}^{1/2}\varepsilon.$$

* This gives rise to the usual least squares model

$$\boldsymbol{Y}' \;=\; \boldsymbol{X}'\beta + \varepsilon'$$

* Using the results from regular least squares we then get the solution

$$\widehat{\boldsymbol{\beta}} \;=\; \left(\left(\boldsymbol{X}'\right)^t \boldsymbol{X}'\right)^{-1} \left(\boldsymbol{X}'\right)^t \boldsymbol{Y}' \;=\; \left(\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^t\boldsymbol{W}\boldsymbol{Y}.$$

* Hence this is the weighted least squares solution.

Key Points

1. The difficulty, in practice, is determining estimates of the error variances (or standard deviations).
2. Weighted least squares estimates of the coefficients will usually be nearly the same as the "ordinary" unweighted estimates. In cases where they differ substantially, the procedure can be iterated until estimated coefficients stabilize (often in no more than one or two iterations); this is called iteratively reweighted least squares.
3. In some cases, the values of the weights may be based on theory or prior research.
4. In designed experiments with large numbers of replicates, weights can be estimated directly from sample variances of the response variable at each combination of predictor variables.
5. Use of weights will (legitimately) impact the widths of statistical intervals.

Weighted Least Squares in Simple Regression

* Suppose that we have the following model

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \varepsilon_i \qquad\qquad i = 1, \ldots, n$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2/w_i)$ for **known** constants $w_1, \ldots, w_n$.

* The weighted least squares estimates of β_0 and β_1 minimize the quantity

$$S_w(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

* The weighted least squares estimates are then given as

$$\begin{aligned}\widehat{\beta}_0 &= \overline{y}_w - \widehat{\beta}_1 \overline{x}_w \\ \widehat{\beta}_1 &= \frac{\sum w_i (x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum w_i (x_i - \overline{x}_w)^2}\end{aligned}$$

where $\overline{x}_w$ and $\overline{y}_w$ are the weighted means

$$\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad \overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}.$$

* Furthermore we can find their variances

$$
\begin{aligned}
\text{Var}(\widehat{\beta}_1) &= \frac{\sigma^2}{\sum w_i(x_i - \overline{x}_w)^2} \\
\text{Var}(\widehat{\beta}_0) &= \left[\frac{1}{\sum w_i} + \frac{\overline{x}_w^2}{\sum w_i(x_i - \overline{x}_w)^2}\right]\sigma^2
\end{aligned}
$$

* Since the estimates can be written in terms of normal random variables, the sampling distributions are still normal.

* The weighted error mean square $S_w(\widehat{\beta}_0, \widehat{\beta}_1)/(n-2)$ also gives us an unbiased estimator of σ^2 so we can derive t tests for the parameters etc.

For this example the weights were known. There are other circumstances where the weights are known:

- If the *i*-th response is an average of n_i equally variable observations, then $Var(y_i) = \sigma^2/n_i$ and $w_i = n_i$.
- If the *i*-th response is a total of n_i observations, then $Var(y_i) = n_i\sigma^2$ and $w_i = 1/n_i$.
- If variance is proportional to some predictor x_i , then $Var(y_i) = x_i\sigma^2$ and $w_i = 1/n_i$.

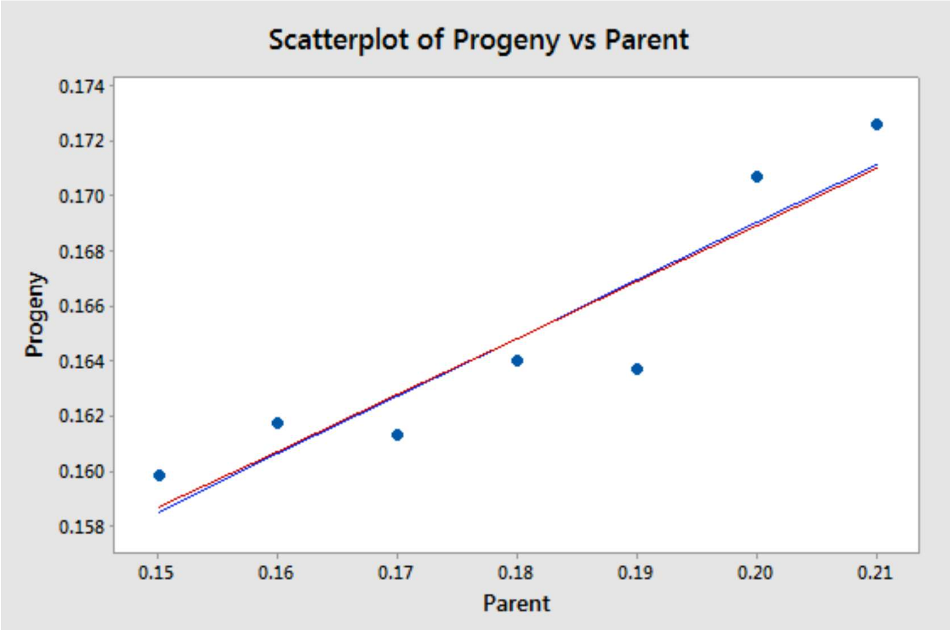
In practice, the structure of **W** is usually unknown, so we have to perform an ordinary least squares (OLS) regression first. Provided the regression function is appropriate, the i-th squared residual from the OLS fit is an estimate of σi2 and the i-th absolute residual is an estimate of σ_i (which tends to be a more useful estimator in the presence of outliers). The residuals are much too variable to be used directly in estimating the weights, w_i, so instead we use either the squared residuals to estimate a variance function or the absolute residuals to estimate a standard deviation function. We then use this variance or standard deviation function to estimate the weights.

weighted least squares

Progeny = 0.12796 + 0.2048 Parent

ordinary least squares model:

Progeny = 0.12703 + 0.2100 Parent



Parent	Progeny	SD
0.21	0.1726	0.01988
0.2	0.1707	0.01938
0.19	0.1637	0.01896
0.18	0.164	0.02037
0.17	0.1613	0.01654
0.16	0.1617	0.01594
0.15	0.1598	0.01763

Blue line: OLS

Red Line: WLS

Some key points regarding **weighted least squares** are:

1. The difficulty, in practice, is determining estimates of the error variances (or standard deviations).
2. Weighted least squares estimates of the coefficients will usually be nearly the same as the "ordinary" unweighted estimates. In cases where they differ substantially, the procedure can be iterated until estimated coefficients stabilize (often in no more than one or two iterations); this is called *iteratively reweighted least squares*.
3. In some cases, the values of the weights may be based on theory or prior research.
4. In designed experiments with large numbers of replicates, weights can be estimated directly from sample variances of the response variable at each combination of predictor variables.
5. Use of weights will (legitimately) impact the widths of statistical intervals.

Exercise

1. Perform Linear Regression
2. Test for Heteroscedasticity

- residual vs. fitted values plot
- Can use perform a Breusch-Pagan test

3. Weighted Least Squares Regression
4. Compare the coefficient estimate and residual error

Hours	Score
1	48
1	78
1	72
2	70
2	66
3	92
4	93
4	75
4	75
5	80
5	95
5	97
6	90
6	96
7	99
8	99