CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 28
Weighted Linear Regression

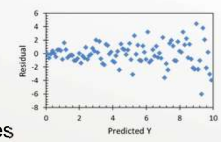
Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Assumptions in OLS Regression

1. ε is a random variable that does not depend on *x* (i.e., the model is perfect, it properly accounts for the role of x in predicting y)
2. $\mathrm{E}[\varepsilon_i] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All ε_i are independent of each other (uncorrelated for the population, but not for a sample)
4. All ε_i have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim \mathrm{N}(0,\sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

When Variance Varies



- Changing variance makes OLS regression inefficient
  - There is no bias in the model parameters, but there is bias in the standard error estimates for those parameters
  - Statistical tests on the parameters will not be as accurate as for constant variance
  - Standard deviations must vary by > 2 before the effect is significant
- If you know how the variance changes with each y_i, use a weighted regression

Weighted MLE

Potentially a different variance for each data point

- Let $y_i = \beta_0 + \beta_1 x + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma_i)$
- Since each ε_i is independent, the likelihood function for the entire data set is

$$L = \prod_{i=1}^{n} \mathrm{P}(\varepsilon_i) \propto exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{\varepsilon_i^2}{\sigma_i^2}\right]$$

- We want to minimize chi-square

$$Let\ w_i = \frac{1}{\sigma_i^2} \qquad \chi^2 = \sum_{i=1}^{n} w_i \varepsilon_i^2 \qquad \frac{\partial \chi^2}{\partial \beta_k} = 0\ for\ all\ k$$

Weighted MLE (2)

- For our line model, $\varepsilon_i = y_i - (\beta_0 + \beta_1 x)$

$$\chi^2 = \sum_{i=1}^{n} w_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Intercept: $$\frac{\partial \chi^2}{\partial \beta_0} = 0 = \sum_{i=1}^{n} 2w_i\left(y_i - (\beta_0 + \beta_1 x_i)\right)(-1)$$

$$b_0 = \bar{y}_w - b_1 \bar{x}_w \qquad \bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

(weighted means)

Weighted MLE (3)

- Substitute our estimate for β_0 into χ^2

$$\chi^2 = \sum_{i=1}^{n} w_i \left((y_i - \bar{y}_w) - \beta_1 (x_i - \bar{x}_w)\right)^2$$

Slope: $$\frac{\partial \chi^2}{\partial \beta_1} = 0 = \sum_{i=1}^{n} 2w_i\left((y_i - \bar{y}_w) - \beta_1(x_i - \bar{x}_w)\right)\left(-(x_i - \bar{x}_w)\right)$$

$$b_1 = \frac{\sum_{i=1}^{n} w_i (y_i - \bar{y}_w)(x_i - \bar{x}_w)}{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2}$$

If w_i = 1 for all i we have OLS

Standard Errors (one predictor)

- Since w_i is the weighting used for χ^2, we can define a weighted residual as $\varepsilon_{wi} = \sqrt{w_i}\varepsilon_i$

$$SE(\varepsilon_w) = s_{\varepsilon_w} = \sqrt{\frac{\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2}{(n-p)}} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_{wi}{}^2}{(n-p)}}$$

$$SE(b_1) = \frac{SE(\varepsilon_w)}{\sqrt{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2}}$$

$$SE^2(b_0) = \frac{SE^2(\varepsilon_w)}{\sum_{i=1}^{n} w_i} + SE^2(b_1)\bar{x}_w{}^2$$

Weighted Regression

- Most statistics packages allow seamless weighted least-squares regression
  - In Excel, LINEST does not support weighted least-squares regression
  - We can use an Excel Add-In for this:
    - Real-Statistics: www.real-statistics.com/
- Note that outliers can be dealt with by assuming they have a greater variance
  - Assign a low weight to the outlier rather than deleting the data point

Estimating Weights

- If we don't know the variance of each data point *a priori*, what can we do?
  - Option 1: Assume a functional form, such as standard deviation or variance of residuals proportional to the response, or to a predictor variable
  - Option 2: Perform OLS, and plot ε_i^2 versus $\hat{y}_i$ or x_i, fit to a straight line
    - Weight will be the inverse of this fit line

Weighted Residuals

- For weighted regression, it is the weighted residuals that we try to make homoscedastic

$$\varepsilon_{wi} = \sqrt{w_i}\varepsilon_i$$

  - Perform residual analysis on the weighted residuals
  - We calculate the isr, esr, etc., using the ε_{wi}
  - Plot studentized residuals versus $\sqrt{w_i}\hat{y}_i$ or $\sqrt{w_i}x_i$ (or just versus $\hat{y}_i$ or x_i)

Real-Statistics Excel Add-In

- Download Resource Pack from www.real-statistics.com
  - Download RealStats.xlam file
  - Move file to C:\Users\user-name\AppData\Roaming\Microsoft\AddIns
  - Select File > Help|Options > Add-Ins and click on the Go button
  - Check the Realstats option and click OK

Lecture 28: What have we learned?

- Why would we ever want to do weighted regression?
- What is a weighted mean?
- How do the weights relate to the variance of each y value?
- How do we estimate weights?
- How does weighted regression affect our analysis of the residuals?