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

Lecture 30
Total Regression, part 1

Chris A. Mack

Adjunct Associate Professor

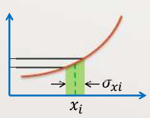
http://www.lithoguru.com/scientist/statistics/

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. $E[\varepsilon_i] = 0$ (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. $\varepsilon \sim N(0,\sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

Uncertainty in X

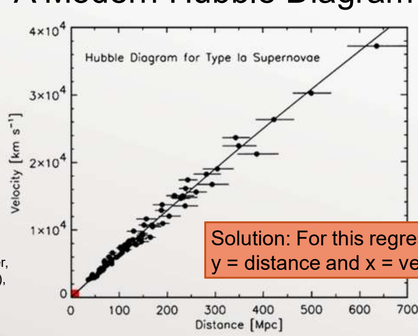
- For most experiments, the predictor variable values (x_i) are themselves the results of measurements
  - All measurements have uncertainty (σ_{xi})
- If the uncertainty in each x_i has only a very small impact on the uncertainty in y_i, it may be OK to ignore it
  - For $\hat{y}_i = f(x_i)$, is $\sigma_{yi} \gg \sigma_{xi} \frac{\partial f}{\partial x_i}$ for each *i* ?



Example: Hubble Constant

- Edwin Hubble noted that the rate galaxies were moving away from us was proportional to their distance from us
  - Model: *Velocity* = H_0 * *Distance*
- He performed a linear regression to obtain the Hubble constant H_0
- But, most of the uncertainty in his data was in the x-variable!

A Modern Hubble Diagram



Solution: For this regression, set y = distance and x = velocity

R. P. Kirshner, *PNAS*,**101**(1), 8-13 (2004).

Total Regression

- If X and Y have non-negligible uncertainty, we must find not only the predicted y values but the predicted x values as well (x and y are interchangeable)
  - Also called Errors-in-Variables regression or Measurement Error Modeling (W.A. Fuller, *Measurement Error Models*, Wiley, 2006)
  - We want values that minimize

$$S = \sum_{i=1}^{n} \left[\left(\frac{\hat{y}_i - y_i}{\sigma_{yi}} \right)^2 + \left(\frac{\hat{x}_i - x_i}{\sigma_{xi}} \right)^2 \right] \qquad \hat{y}_i = predicted\ y\ value \quad \hat{x}_i = predicte\ \ x\ value$$

- Example: $\hat{y}_i = \beta_0 + \beta_1 \hat{x}_i$
  - There are n + 2 best fit parameters
  - Requires a nonlinear least-squares regression

Different Total Regression Approximations

- Effective Variance Approximation
- Orthogonal Regression
- Geometric Mean
- Method of Moments
- Deming Regression
- Full Total Regression

Interpreting Total Regression

- Structural Model
  - The X's are fixed, but unknown, and so must be estimated
- Functional Model
  - The X's are random variables, to be represented by their mean and standard deviation (pdf)
- The difference between these two is subtle

Effective Variance Approximation

- We can simplify the regression for the case of small errors in x
  - Let $\hat{x}_i = x_i$
  - Define an effective variance in y using the model $\hat{y}_i = f(x_i)$:

$$\sigma_{yi-eff}^2 = \sigma_{yi}^2 + \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{xi}^2$$

  - Use a weighted least-squares regression with weights $w_i = 1/\sigma_{yi-eff}^2$
  - What value of $\partial f/\partial x_i$ should we use?

Effective Variance Approximation

How to estimate the model slope ($\partial f/\partial x_i$)?

1. Run a linear regression ignoring the x-variance
2. Use this model fit to calculate $\partial f/\partial x_i$ for each i
3. Calculate the effective variance for each y_i
4. Run a weighted least-squares regression using 1/effective variance to weight the y_i
5. Repeated steps 2-4 until the parameters converge (usually only 1-2 iterations)

Improving the Effective Variance

- We can also improve our estimate of $\hat{x}_i$
  - For $\hat{y}_i = f(x_i)$,

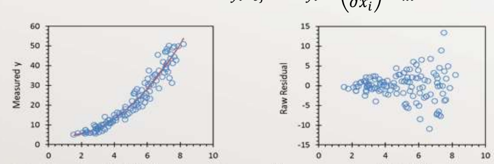
$$\hat{x}_i = x_i + \frac{(y_i - \hat{y}_i)}{\partial f/\partial x_i} \frac{(\partial f/\partial x_i)^2 \sigma_{xi}^2}{\sigma_{yi-}^2}$$

  - Again, iterate and repeat the weighted linear regression, using the better estimates for $\hat{x}_i$ (iteratively reweighted least squares)

Impact of Errors in Predictor Variables

- For a straight line model, errors in x will bias the OLS estimate of the slope towards zero
- For a higher order model, errors in x will look like heteroscedasticity

$$\sigma_{yi-ef}^2 = \sigma_{yi}^2 + \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{xi}^2$$



Lecture 30: What have we learned?

- When do I have to worry about error in the x-variable?
- What is total regression (also called errors-in-variables regression)?
- Explain the effective variance approximation
- How does x uncertainty affects our OLS slope estimate for a straight-line model?
- When does error in the x-variable result in heteroscedasticity?

