Assumptions in OLS Regression

1. $\varepsilon$ 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. $E[\varepsilon] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All $\varepsilon_i$ are independent of each other (uncorrelated for the population, but not for a sample)
4. All $\varepsilon_i$ have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim N(0, \sigma)$ (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 ($\sigma_x$)
- 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 $\sigma_{y_i} \gg \sigma_{x_i}$ 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 = $H_0 \times$ 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!

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

  $$S = \sum_{i=1}^{n} \left( \frac{(\hat{y}_i - y_i)^2}{\sigma_{y_i}^2} + \frac{(\hat{x}_i - x_i)^2}{\sigma_{x_i}^2} \right)$$

  $\hat{y}_i$ = predicted $y$ value
  $\hat{x}_i$ = predicted $x$ value

  - Example: $\hat{y}_i = \beta_0 + \beta_1 x_i$
  - 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 \( \hat{y}_i = f(x_i) \):
  \[
  \sigma_{y_i}^2 = \sigma_{y_i}^2 + \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_x^2
  \]
  - Use a weighted least-squares regression with weights \( w_i = 1/\sigma_{y_i}^2 \)
  - 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 \( \hat{y}_i = f(x_i) \),
  \[
  \hat{x}_i = x_i + \frac{(y_i - \hat{y}_i) \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_y^2}{\frac{\partial f}{\partial x_i} \sigma_{y_i}^2}
  \]
  - 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
  \[
  \sigma_{y_i}^2 = \sigma_y^2 + \left( \frac{\partial f}{\partial x_i} \right)^2 \sigma_x^2
  \]
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?