P8130: Biostatistical Methods I

Lecture 12: Simple Linear Regression

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Outline

- Purpose of linear regression
- Simple linear regression
  - Model formulation
  - Methods of estimation

Statistical Learning

- Refers to a vast set of understanding data
- **Supervised learning** – building a model for prediction or estimating an output based on one or more inputs
  e.g., Linear Regression
- **Unsupervised learning** – there are inputs, but no supervising output
  e.g., clustering
- **Semi-supervised learning** – a data subset contains predictors and response, the rest does not

Linear Regression

- Very simple approach of supervised learning
- Somewhat ‘dull’ compared to other fancy methods, but still useful for:
  - Describing associations and NOT CAUSATION between predictors (X) and response/outcome (Y)
  - Predicting a quantitative response (how precise is our Y estimate for a given X?)
  - Adjusting – quantify the association between Y and a main predictor X, adjusting for other factors (covariates)

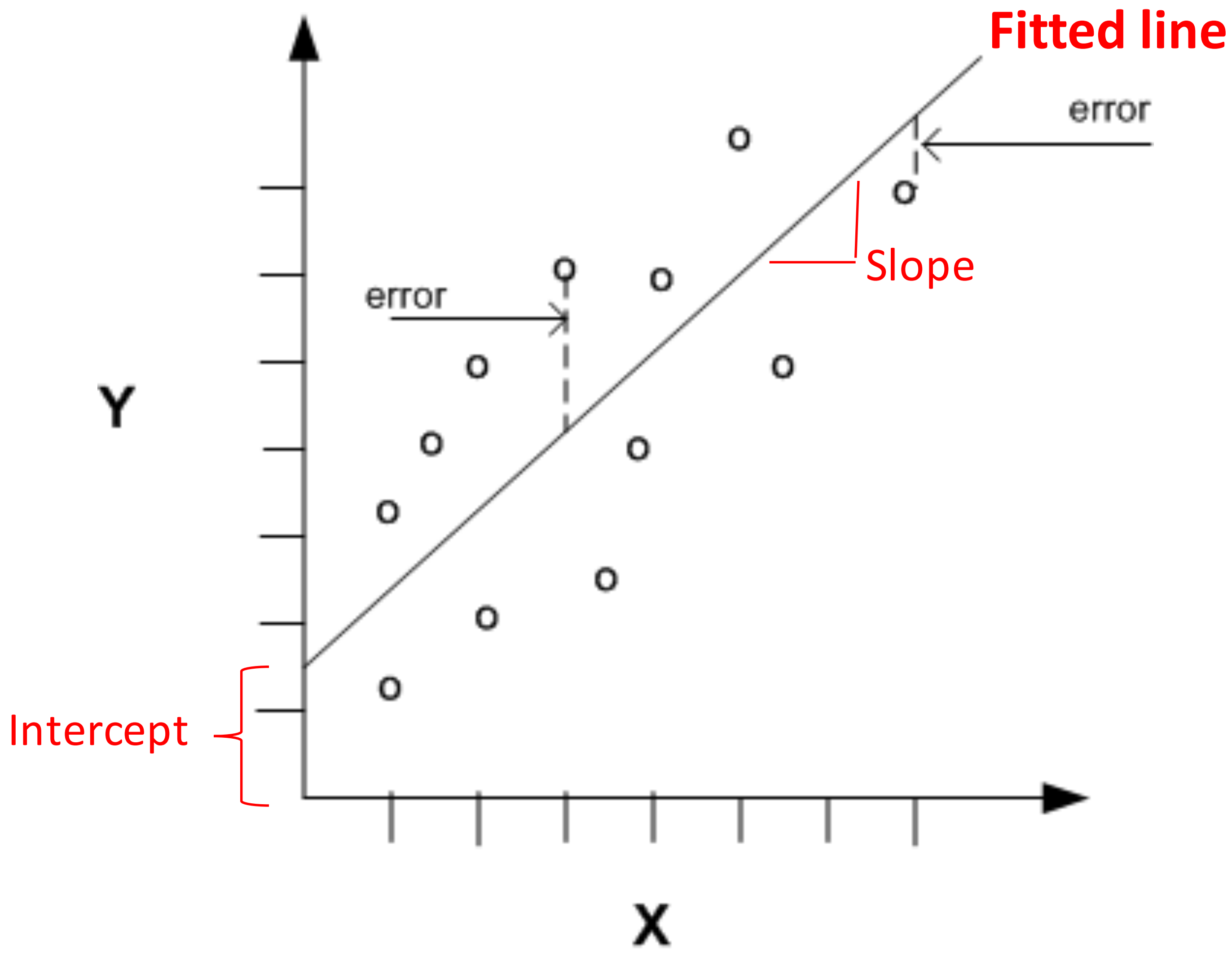
Linear Regression

- Why linear?
- Outcome (Y) is a continuous (dependent) variable
- Predictor(s) (X) can be continuous or categorical (independent) variable(s)
- Assumes a LINEAR relationship between Y and X: (approximately) straight line
- Approximately?! Well, there is always ‘error’, ‘noise’, or ‘unexplained variation’

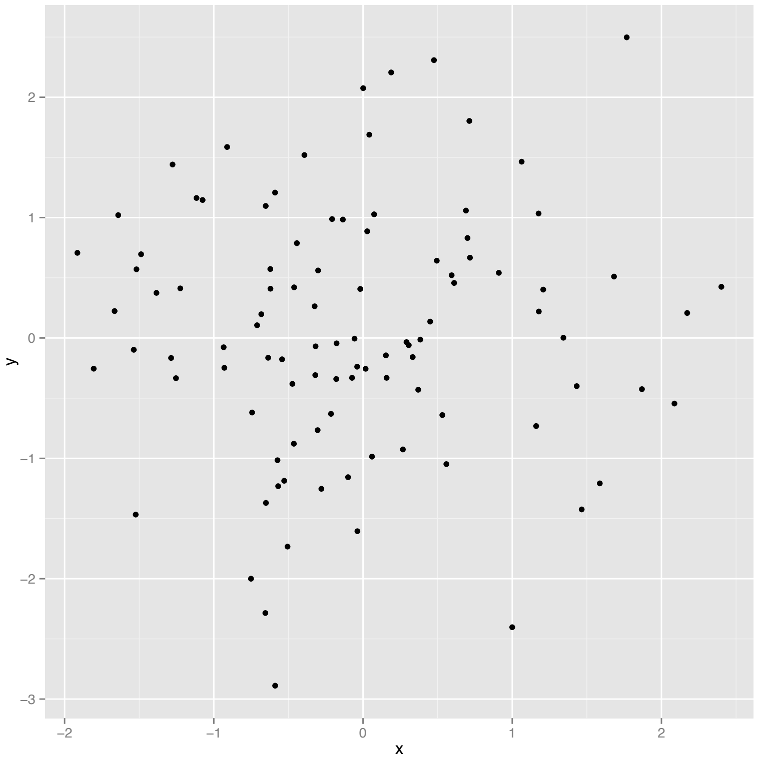
Linear Regression

- What do (should) we want to know?
  - Is there a relationship between Y and X? Is it linear?
  - How strong is the relationship? Can we predict Y with a high level of accuracy or the prediction is only slightly better than a random guess?
  - Which predictor(s) are associated with Y?
  - How accurately can we predict future outcome(s)?
  - Is there any synergy (interaction effect) between the predictors?

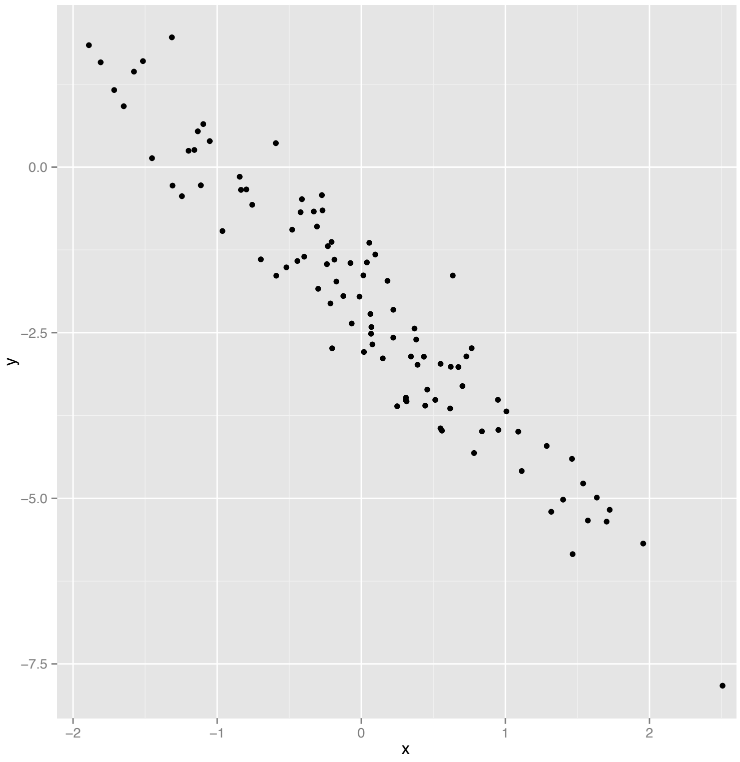
Scatterplot of Y vs. X



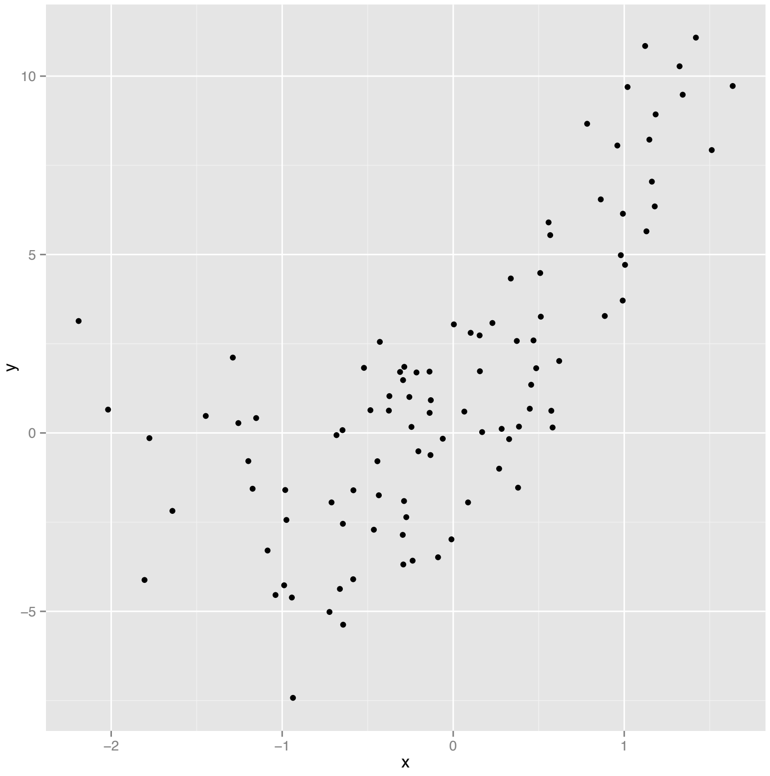
More data patterns



Random pattern



Linear decreasing



Linear? increasing

Linear Regression - Graphics

- Usually scatterplots of Y vs. (each) X
- Before estimation:
  - Identify the pattern: linear?, multimodal?
  - Notice potential usual observations (outliers)
- Other graphical displays: histograms, density plots, box-plots

Simple Linear Regression (SLR)

- Only one predictor X
- Basic regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y_i represents the response variable from the i^{th} individual
- β_0, β_1 represent the model parameters to be estimated
- X_i is a known constant, the value of the predictor variable from the i^{th} individual
- ε_i represents the random error term, described by a probability distribution

Simple Linear Regression (SLR)

- Basic regression model is said to be:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- *Simple:* there is only one predictor

- *Linear in parameters*: no parameter appears as an exponent or is multiplied/divided by another parameter

- *Linear in the predictor variable*: the predictor (X) appears only in the first power

SLR: Assumptions

- Model linear in parameters
- Attributes of error terms $\varepsilon_i, i = 1, 2, \dots, n$:
  - Mean of error terms is 0: $E(\varepsilon_i) = 0$
  - Constant variance: $\sigma^2(\varepsilon_i) = \sigma^2$
  - Error terms are uncorrelated: $cov(\varepsilon_i, \varepsilon_j) = \sigma(\varepsilon_i, \varepsilon_j) = 0$ for all $i, j; i \neq j$.
  - (Normal) distribution of the errors is assumed for inferences

Regression Parameters

- β_0, β_1 represent the model coefficients to be estimated
- β_0 is the Y intercept of the regression line
  - When the scope of the model includes $X = 0$, β_0 gives the mean value of the regression function at $X = 0$. Does not always make sense!
- β_1 is the slope of the regression line
  - Expected increase/decrease in Y for one unit increase in X
  - Expected difference in Y when comparing individuals that differ by one unit or category (for categorical predictors)

Estimation

- Given a dataset composed of $(X_i, Y_i), i = 1, 2, \dots, n$ pairs

Find the 'best' values for the intercept and slope of the linear relationship.

- One option is to use the mathematical criterion Least Squares (LS)
  - Minimize the errors
  - Minimize the vertical distance between the fitted (estimated) Y values and the observed data
  - It reduces to minimizing the criterion denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

LS Estimation

- Minimize: $Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$
- Set derivatives to 0: $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$
- After some math, get the system of normal equations:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

$\rightarrow$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

LS Estimation

- The estimated regression model is given by:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i, i = 1, 2, \dots, n.$$

- For a given i^{th} individual we have the following:

  Observed value (obtained from the model): Y_i

  Fitted value (obtained from the model): $\widehat{Y_i}$

  Residual: $e_i = Y_i - \widehat{Y_i}$

- Distinguish between the model error term value ε_i and residual e_i:

$$\varepsilon_i = Y_i - E(Y_i) \quad \text{vs} \quad e_i = Y_i - \widehat{Y_i}$$

Deviation of Y_i from the TRUE (unknown) regression line vs Deviation of Y_i from the ESTIMATED (known) regression line

Residual Variance Estimation

- Error sum of squares or residual sum of squares:

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- In a single population estimation, we divide by *n-1 df* to estimate the sample variance. Why loose 2 *df* in this case?

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- *MSE* is also called *mean square error* and it can be shown that:

$$E(MSE) = \sigma^2.$$

Readings

Kutner *et al.*, Applied Linear Statistical Models

- Chapter 1, Sections: 1.1 – 1.7

Next class:

- Properties of model coefficients: expected values and variance estimates
- Matrix notation
- Beyond LS estimation - maximum likelihood (ML) estimation