Biostatistics Preparatory Course: Methods and Computing

Lecture 5

Linear Regression

Linear Regression

- Linear regression is a way to model the association between some continuous outcome, Y, with a set of predictors, $X_1 \dots X_p$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \text{ where } E(\epsilon_i) = 0$$

Linear Regression

- Linear regression is a way to model the association between some continuous outcome, Y, with a set of predictors, $X_1 \ldots X_p$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \text{ where } E(\epsilon_i) = 0$$

- This can also be expressed as,

$$E[Y_i | X_{i1}, ..., X_{ip}] = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

Linear Regression

- Linear regression is a way to model the association between some continuous outcome, Y, with a set of predictors, $X_1 \ldots X_p$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \text{ where } E(\epsilon_i) = 0$$

- This can also be expressed as,

$$E[Y_i | X_{i1}, ..., X_{ip}] = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

- or as,

$$E[Y_i | \mathbf{X}_i] = \mathbf{X}_i^T \boldsymbol{\beta}$$

Linear Regression

- Linear regression is a way to model the association between some continuous outcome, Y, with a set of predictors, $X_1 \ldots X_p$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i, \text{ where } E(\epsilon_i) = 0$$

- This can also be expressed as,

$$E[Y_i | X_{i1}, ..., X_{ip}] = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}$$

- or as,

$$E[Y_i | \mathbf{X}_i] = \mathbf{X}_i^T \boldsymbol{\beta}$$

- β_j is the change in mean value of Y corresponding to a one unit change in X_1, holding all other variables constant.

Assumptions for Ordinary Least Squares

We will make the following assumptions,

- **Linearity**: the expectation of Y is linear in $X_1 \ldots X_p$
- **Independence**: the ϵ_i are independent
- **Mean zero errors**: the ϵ_i have mean zero, i.e. $E[\epsilon_i] = 0$
- **Equal variance (homoscedasticity)**: the ϵ_i have the same variance, i.e. $Var[\epsilon_i] = \sigma^2$

Note: We are not making any assumptions about X_j; they can be continuous, binary, or categorical. This will change interpretations!

Estimating $\boldsymbol{\beta}$ with OLS

- The goal is to estimate $\boldsymbol{\beta} = \{\beta_0, \beta_1, ..., \beta_p\}$
- We want to minimize the distance between the observed Y_i's and their fitted values (i.e. the residuals)
- For the ith observation, the residual is:

$$\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1}$$

- Thus, the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are those values of β_0 and β_1 that minimize,

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1})^2$$

Estimating β with OLS

- When there is more than one covariate, we want to minimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Fortunately, this has a closed-form solution,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix} ; \boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$$

- Another way to think of this is as the projection of Y onto the linear subspace spanned by the columns of $\mathbf{X}$.

Maximum Likelihood Estimation vs. OLS

- We did not place any distributional assumptions on the outcome,
  - We only required that $E[\epsilon_i] = 0$ with constant variance
  - In other words, OLS is a *semiparametric* method

Maximum Likelihood Estimation vs. OLS

- We did not place any distributional assumptions on the outcome,
  - We only required that $E[\epsilon_i] = 0$ with constant variance
  - In other words, OLS is a *semiparametric* method
- Sometimes, people assume that $\epsilon_i \sim N(0, \sigma^2)$, which means

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_1 X_{ip}, \sigma^2)$$

Maximum Likelihood Estimation vs. OLS

- We did not place any distributional assumptions on the outcome,
  - We only required that $E[\epsilon_i] = 0$ with constant variance
  - In other words, OLS is a *semiparametric* method
- Sometimes, people assume that $\epsilon_i \sim N(0, \sigma^2)$, which means

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + ... + \beta_1 X_{ip}, \sigma^2)$$

  - If this additional assumption is made, then we can instead use maximum likelihood estimation for $\boldsymbol{\beta}$
  - This connects to a whole other class of models called generalized linear models (GLMs)
  - Interestingly, in this case, you will end up with the same estimates for $\boldsymbol{\beta}$

Inference

Once you have estimated values for $\boldsymbol{\beta}$, you can perform inference:

- Estimate the standard error for $\hat{\boldsymbol{\beta}}$
- With a large enough sample, asymptotic normality kicks in which makes it easy to do:
  - Hypothesis tests of the form: $H_0 : \beta_j = 0$
  - Construct confidence intervals for β_j
- If you have a small sample size, you will need to rely on other methods for hypothesis testing and confidence intervals

Notes on interpreting regression coefficients

After you have estimated β_j, you will typically be tasked with interpretation. Be sure to mention:

- The parameter of interest (mean, odds, risk, etc.)
- The value of the *estimated* parameter
- The groups you are comparing (this depends on how you have coded your covariate!)
- What is happening with other variables in the model (i.e. adjusting for x,y,z)

Exercises: Implementation in R