JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Lecture 5

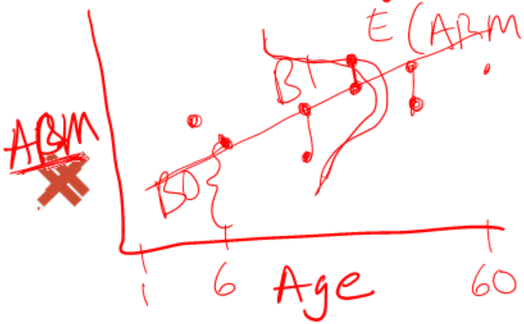
The classical linear regression model

Review of key concepts from Lecture 3 and 4

- Simple linear regression model
  - ARM = B0 + B1 (age – 6) + e, e~N(0,σ^2), independent

systematic component → random component / error

E(ARM | age)



B0 = average ARM at age = 6 months

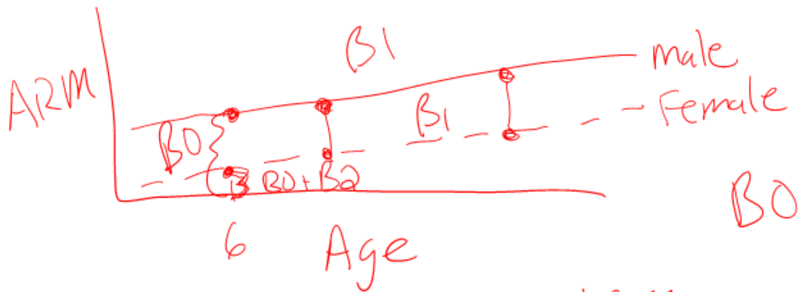
B1 = expected # difference in ARM comparing two children who differ in age by 1 month

Review of key concepts from Lecture 3 and 4

- Sex adjusted relationship between ARM and age
  - ARM = B0 + B1 (age – 6) + B2 Female + e, e~N(0,σ^2), independent

$B0 + B1(age-6) + B2$

$(B0 + B2) + B1(age-6)$



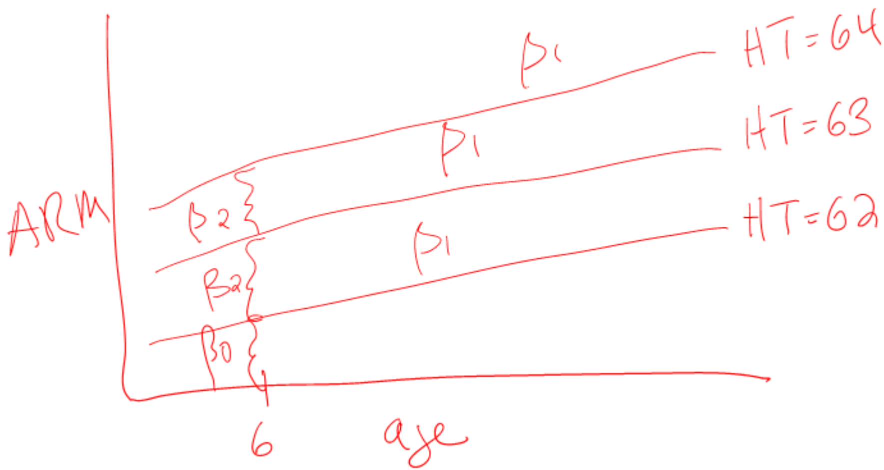
B0 = Averge ARM among male children 6 months of age

B1 = difference in mean ARM comparing children of the same gender but whom differ by 1 month of age

B2 = Difference in average ARM comparing females to males at any Age

Review of key concepts from Lecture 3 and 4

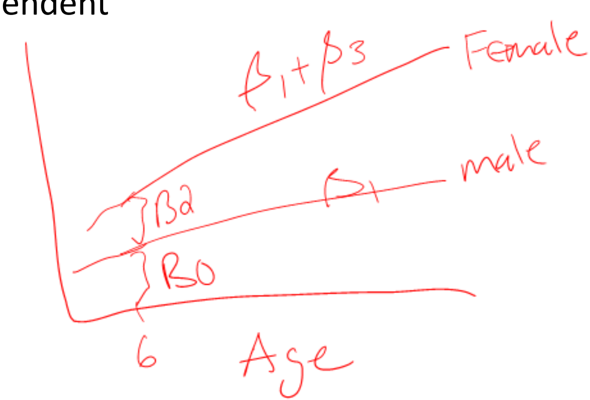
- Height adjusted relationship between ARM and age
  - ARM = B0 + B1 (age – 6) + B2 (HT – 62) + e, e~N(0,σ^2), independent



B1 = difference in the mean ARm comparing childen who differ in age by 1 month but whom have the same HT.

Review of key concepts from Lecture 3 and 4

- Effect modification: Is the ARM vs. age relationship the same or different by sex
  - ARM = B0 + B1 (age – 6) + B2 Female + B3 (age – 6) Female + e, e~N(0,σ^2), independent



B1 = expected change in mean ARM per month among males

B1+B3 = " among females

B3 = difference in monthly growth of ARM comparing females to males

Multiple Linear Regression Model

- Y is a random variable representing the outcome of interest in the population
- The explanatory variables, $X_1, X_2, \ldots, X_p$ are fixed/known (not random or measured with error)
- Sample of size n is observed, data are: $(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi})$

$$\Rightarrow \underbrace{Y_i}_{\text{outcome random variable}} = \underbrace{\mu_i(\beta, X_i)}_{\text{systematic component}} + \varepsilon_i \rightarrow \text{random component}$$

- X is the design matrix or table combining all the explanatory variables $[1_s, X_1, X_2, \ldots, X_p]$
- X_i is the row of the design matrix corresponding to subject i $\hookrightarrow$ column vector of length n

$(1, X_{1i}, X_{2i}, \ldots, X_{pi})$

Multiple Linear Regression Model

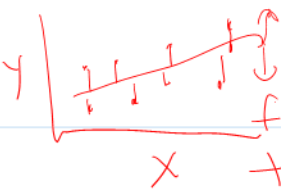
$$Y_i = \mu_i(\underset{\sim}{\beta}, X_i) + \varepsilon_i$$

$$\underset{\sim}{\beta} = \begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix} \quad p+1 \text{ rows} \times 1$$

- Systematic component:
  - $\mu_i(\underset{\sim}{\beta}, X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots \beta_p X_{pi}$
- ε_i is the random components: $\varepsilon_i \sim N(0, \sigma^2)$, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$
- The least squares solution finds the values of $\underset{\sim}{\beta}$ that minimize:

$$\sum_{i=1}^{n} \Big(Y_i - \underbrace{\mu_i(\underset{\sim}{\beta}, X_i)}_{\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}} \Big)^2$$

Find $\underset{\sim}{\beta}$ s.t. the residual sums of squares are minimized



find β_0, β_1 that minimize SSR

Least squares solution: simple linear regression

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

Least squares solution is:

$$\hat{\beta}_1 = \frac{\sum (Y_i - \bar{y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

$$\frac{\text{Covariance between } Y \text{ and } X}{\text{Variance of } X}$$

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

If we were to standardize Y and X,
mean(Y) = mean$(X) = 0$
$SD(Y) = SD(X) = 1$
$\hat{\beta}_1 = r$ Pearson correlation coefficient

Maximum likelihood inference in MLR

- Start with the MLR:

$$\Rightarrow \underset{\rightarrow}{Y_i} = \underbrace{\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \underline{\sigma^2})$$

independent

Data: (y_i, X_i)

- Other notation:

Y_i = Random variable, y_i = observation

$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$ $\underset{\sim}{Y}$, $\underset{\sim}{y}$ = vectors (lists) of independent random variables or observations $\begin{bmatrix} y_1 \\ y_n \end{bmatrix}$

X_i = row vector containing explanation variables for subject i

$\hat{Y}_i = RV = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_2 X_{pi}$

$R_i = RV = Y_i - \hat{Y}_i$ residual

Likelihood function definition

- Model: under the assumption $e_i \sim N(0, \sigma^2)$, $Y_i = RV$, X_i = fixed

$$Y_i \sim N(\mu_i(\beta, X_i), \sigma^2)$$

- Probability density function:

$$f(y \mid \underset{\sim}{\mu}(\beta, X), \sigma^2) = \prod_{i=1}^{n} f(y_i \mid \mu_i(\beta, X_i), \sigma^2)$$

pdf is a function of y with $\mu_i(\beta, X_i)$ and σ^2 fixed

- Likelihood function: $L(\mu(\beta, X), \sigma^2 \mid y) = \prod_{i=1}^{n} L(\mu_i(\beta, X_i), \sigma^2 \mid y_i)$

likelihood function is viewed as a function of $\mu_i(\beta, X_i)$ and σ^2 for fixed y_i

Identify the values of β and σ^2 that maximize L given fixed y_i

Maximum likelihood estimation under gaussian residuals

- Likelihood function, rely on the normality assumptn and independence

$$L(\beta, \sigma^2 \mid y) = \prod_{i=1}^{n} L(\mu_i(\beta, X_i), \sigma^2 \mid y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \underline{\mu_i(\beta, X_i)}\right)^2\right)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \underline{\beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots - \beta_p X_{pi}}\right)^2\right)$$

Maximum likelihood estimation under gaussian residuals

- Log Likelihood Function

$$\ell(\beta, \sigma^2 \mid y) = \log L(\beta, \sigma^2 \mid y)$$

$$\Rightarrow \quad = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}\left(y_i - \mu_i(\beta, x_i)\right)^2 \right)$$

Find $\hat{\beta}$ and σ^2 that maximize $\ell(\beta, \sigma^2 \mid y)$ by differentiating with respect to β and σ^2, setting these derivatives $= 0$ and solving for β and σ^2

Maximum likelihood estimation under gaussian residuals

- Solution for β_j $\ell(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}\left(y_i - \mu_i(\beta, X_i)\right)^2 \right)$

$j = 0, \ldots, p$

wrt β $\ell(\beta, \sigma^2 | y) \propto \sum_{i=1}^{n} \frac{-1}{2\sigma^2}\left(y_i - \mu_i(\beta, X_i)\right)^2$

Score equation for β_j

$$U_{\beta_j}(\beta | \sigma^2) = \frac{\partial}{\partial \beta_j} \ell(\beta, \sigma^2 | y)$$

$$= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \frac{-1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi})\right)^2$$

$$= \frac{-2}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu_i(\beta, X_i)\right)\left(-X_{ij}\right)$$

$\hookrightarrow$ a score equation for each $\beta \Rightarrow p+1$ score equations

$p+1$ unknowns

$\Rightarrow 0$ solve for β_j

Maximum likelihood estimation under gaussian residuals

- Solution for β_j

$$U_{\beta} = \sum_{i=1}^{n} \left(y_i - \mu_i(\beta, X_i)\right) \begin{bmatrix} 1 \\ X_{1i} \\ X_{2i} \\ \vdots \\ X_{pi} \end{bmatrix}_{p+1 \times 1} = 0$$

$p+1 \times 1$

Maximum likelihood estimation under gaussian residuals

- Solution for σ^2

Given the MLEs $\underset{\sim}{\beta}$, derive score equation for σ^2

$$U_{\sigma^2}(\hat{\underset{\sim}{\beta}}) = \sum_{i=1}^{n} \left(-\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \left(y_i - \mu_i(\hat{\underset{\sim}{\beta}}, X_i)\right)^2 \right)$$

$\Rightarrow 0$ and solve for σ^2

$$\Rightarrow \quad \hat{\sigma}^2_{MLE} = \underbrace{\frac{1}{n}}\sum_{i=1}^{n} \left(y_i - \mu_i(\hat{\underset{\sim}{\beta}}, X_i)\right)^2$$

$$E(\hat{\sigma}^2_{MLE}) = \frac{n-(p+1)}{n}\sigma^2$$

in practice, we use

$$\tilde{\sigma}^2 = \frac{1}{\underbrace{n-(p+1)}} \sum \left(y_i - \mu_i(\hat{\underset{\sim}{\beta}}, X_i)\right)^2$$

MLEs for simple linear regression

MLEs for simple linear regression

MLEs for simple linear regression

Take away messages

Take away messages

Next time....

- Vector / Matrix representation of MLR
- Geometry of least squares
- Distribution of MLEs for regression parameters