IDS 702

Linear Regression - 2 (MLR Estimation)

September 6, 2022

Andrea Lane, PhD

Agenda

1. Pre-class reading questions
2. Multiple Linear Regression
   - Motivation
   - Estimation
   - Matrix formulation
   - Hypothesis testing
3. MLR Activity

Learning Objectives

By the end of today's class, you should be able to:

- Identify the need for multiple linear regression
- Describe the difference between ordinary least squares and maximum likelihood estimation
- Write the MLR model in matrix notation
- Generate a MLR model in R

1. Pre-class Reading Questions

Discuss the following with a partner:

- Fill in the blank: The model $y_i = X_i\beta + \epsilon_i; \epsilon_i \sim N(0,\sigma^2), i = 1,...,n$ can also be written as $y_i \sim N($ _____, _____$), i = 1,...,n$
- Describe the difference between coefficient estimate standard error and residual standard error
- True or False: The R^2 will always increase after adding a predictor to a linear model

Always true, no matter what predictor you add

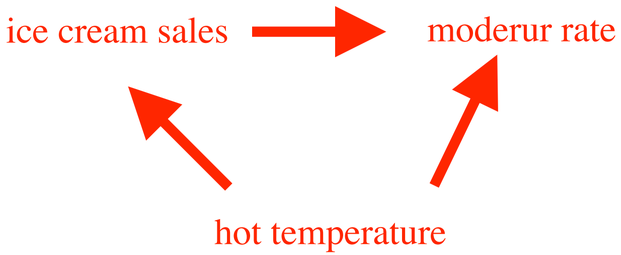
Discuss the following with a partner:

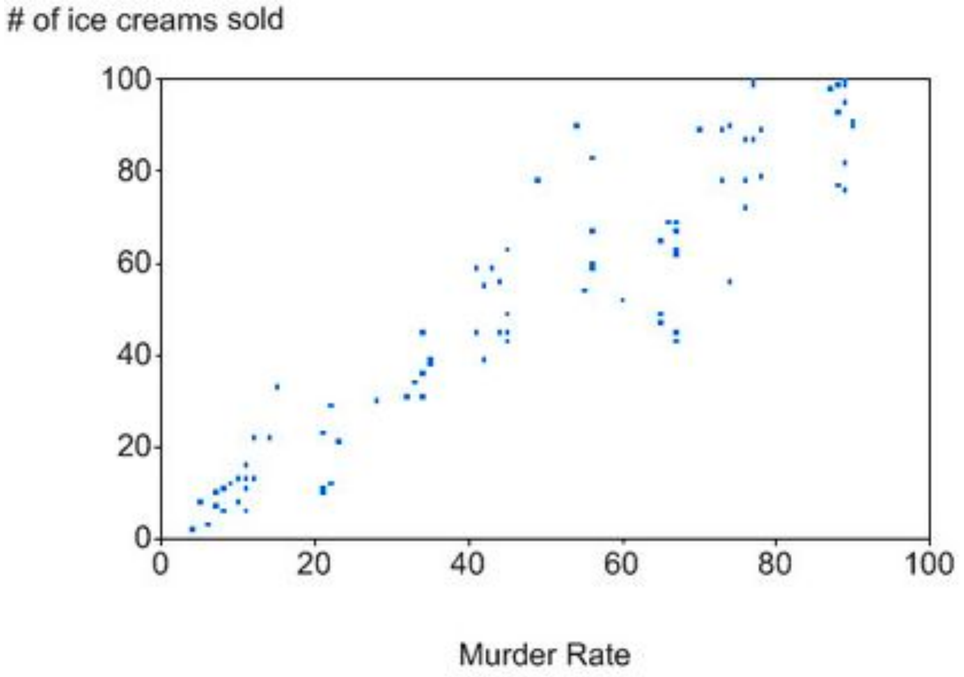
- Fill in the blank: The model $y_i = X_i\beta + \epsilon_i; \epsilon_i \sim N(0,\sigma^2), i = 1,...,n$ can also be written as $y_i \sim N(\underline{\quad X_i\beta \quad}, \underline{\quad \sigma^2 \quad}), i = 1,...,n$
- Describe the difference between coefficient estimate standard error and residual standard error $\hat{\sigma}$ $SE(\hat{\beta})$
- True or False: The R^2 will always increase after adding a predictor to a linear model

2. Multiple Linear Regression

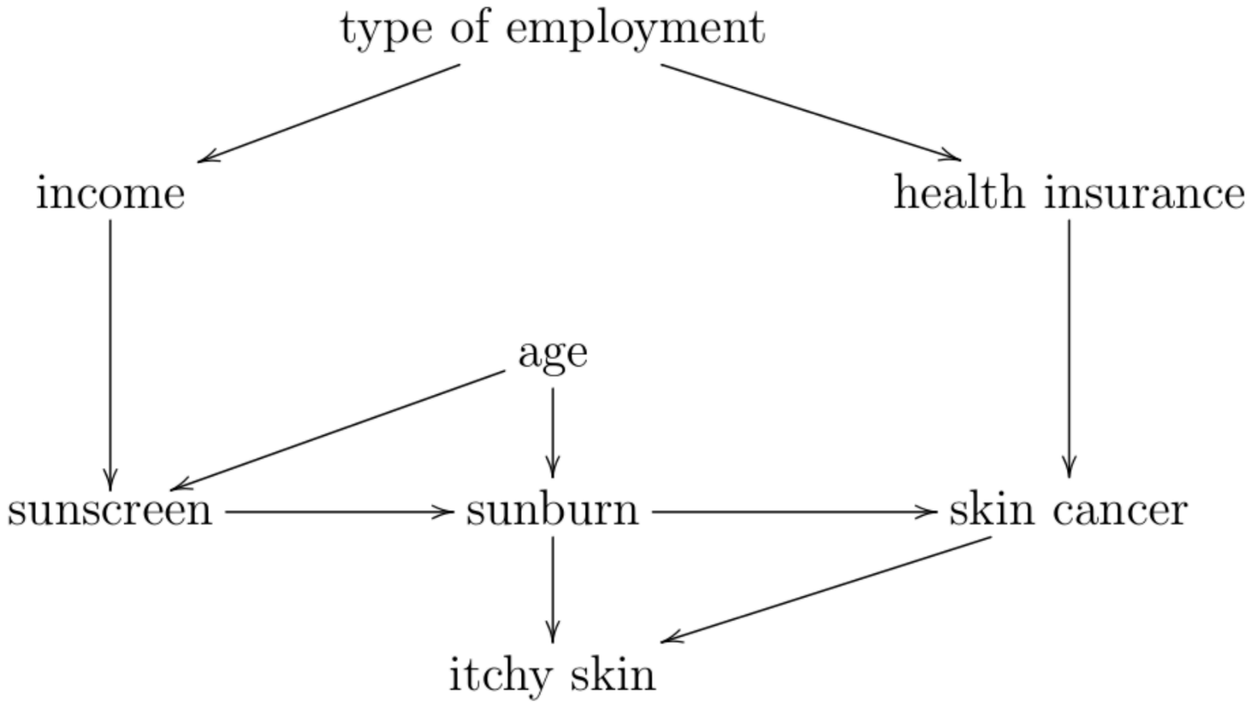
Most relationships cannot be fully explained by two variables

- **Confounding variables** are related to both variables of interest and explain (at least) some of the relationship between them





Directed Acyclic Graph (DAG)



https://health.ucdavis.edu/ctsc/area/Resource_Library/documents/directed-acyclic-graphs20220209.pdf

Multiple Linear Regression Model

slope coefficient

continuous outcome

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i; \epsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2), i = 1,...n$$

intercept

p = number of predictors

We can also write the model as:

$$y_i \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$$

$$E[Y \mid X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

E[epsilon] = 0

conditional expectation/mean

MLR Assumptions

- Linear relationship between EACH X and Y
- Independence of errors
- Equal variance of errors
- Normality of errors
- No multicollinearity

those predictor can't be too correlated with each other

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i; \epsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2), i = 1,\ldots n$$

Two ways of looking at estimation: Least Squares and Maximum Likelihood

Least Squares	Maximum Likelihood
Geometric Formulation something like find the line, show the line specific to the regression problem does not require distribution assumption	statistical formulation General require a distribution assumption

Estimation: Ordinary Least Squares

Coefficient estimates are obtained by taking partial derivatives of the sum of squares of the errors with respect to each parameter

$$\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}])^2$$

Estimation: Maximum likelihood

each y is independent so that we can multiply(assumed)

u = b0 (beta 下标0） + b1x1 +…+ bpXp

$$L(\mu, \sigma^2 \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

PDF of a normal distribution

PDF : Probability density function

input: random variable

output: probability

what is the most likely parameter given the data and the statistical model we chose

Likelihood Function describes the joint probability of the observed data as a function of the parameters of the chosen statistical model

In linear regression, OLS and MLE provide the same estimates.

Why?

because in OLS we still assume the normal distribution by making epsilon ~ N(0, sigma^2)

$\varepsilon_i \sim N(0, \sigma^2)$

Matrix Representation

Multivariate Normal

vector of 0s

covariance matrix

vector

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon; \epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

identity matrix n*n

Matrix Representation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon; \epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & & & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$n \times 1$ $\quad$ $n \times p$ $\quad$ $p \times 1$ $\quad$ $n \times 1$

Matrix Representation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon; \epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Then the OLS estimates are:

$$\hat{\beta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

transpose inverse

For more on LR matrix representation: “Plane Answers to Complex Questions” by Ronald Christensen

Matrix Representation

The predictions can be written as:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}[(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}]$$

And the residuals can be written as:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - [\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}]\mathbf{Y} = [\mathbf{1_n} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}]\mathbf{Y}$$

Hat matrix/Projection matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$$

Matrix Representation: SE

recall SLR

$$s_e^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n - (p+1)} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X}\hat{\beta})}{n - (p+1)} = \frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{n - (p+1)}$$

之前的是 n-2

p = number of predictors

1 = intercept

The variance of the OLS estimates of all (p+1) coefficients is

$$\mathbf{V}[\hat{\beta}] = \sigma^2(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$$

Note that this is a **covariance matrix**; the square root of the diagonal elements give us the standard errors for each coefficient, which we can use for hypothesis testing

Interpreting Coefficients

- Each estimated slope is the amount y is expected to increase when the value of the corresponding predictor is increased by one unit, *holding the values of the other predictors constant*.

- What if the predictor is not continuous?

difference between variables
man = 0 women = 1

Inference

- Is there a relationship between the predictors and the response?

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \qquad H_A = \text{at least one predictor not = 0}$$

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \qquad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Inference

$H_0 : \beta_j = 0$ (GIVEN all the other variables in the model)

$H_A : \beta_j \neq 0$

$$T = \frac{estimate - Null}{SE} = \frac{\hat{\beta}_j - 0}{\sqrt{[s_e^2(\mathbf{X^T X})^{-1}]_{\mathbf{jj}}}}$$

index of the variance matrix p18页的V[B]

$$CI = \hat{\beta}_j \pm SE(\hat{\beta}_j) C_\alpha$$

Which variable is the strongest predictor of the outcome?

- The coefficient that has the strongest linear association with the outcome variable is the one with the largest absolute value of T (test statistic), which equals the coefficient estimate over the corresponding SE
- Note: T is NOT the size of the coefficient
- Coefficients are sensitive to the scale of the predictors, but T is not

Notes about tests and CIs

- When sample size is large enough, you will probably reject the null hypothesis
    - As the sample size increases, SE decreases, test statistic increases

SE作为分母

    - Consider **practical significance**
- When sample size is small, there may not be enough evidence to reject the null hypothesis
    - This does not mean the null is true
    - Can consider a **power calculation**

reject the null when the null is not true

Let’s do it in R

- Advertising dataset from last week

Your turn: In a group of 2-3, complete “In-Class Analysis 1” (Sakai — Lessons)

Wrap-up

- Thursday reading: ISL 3.3.3 "Potential Problems" #1-3, pgs 92-97
- Data Analysis Assignment 1 posted - due Sept 16 11:55 PM