IDS 702

Linear Regression - 1

September 1, 2022

Dr. Andrea Lane

Agenda

1. Reading poll
2. Big picture review
3. SLR review
4. EDA/SLR activity
5. MLR

Learning Objectives

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

- Identify when SLR and MLR are useful (e.g., what kind of data?)
- Describe ordinary least squares (OLS) estimation
- Generate EDA plots in R
- Generate an SLR model in R

1. Reading poll

Sakai $\longrightarrow$ Polls

2. Big picture review

Data analysis depends on the data

Types of variables	Examples
continous	price, revenue, stock
binary var	test result disease status
>2 categories nominal ordinal	 color rating, grade

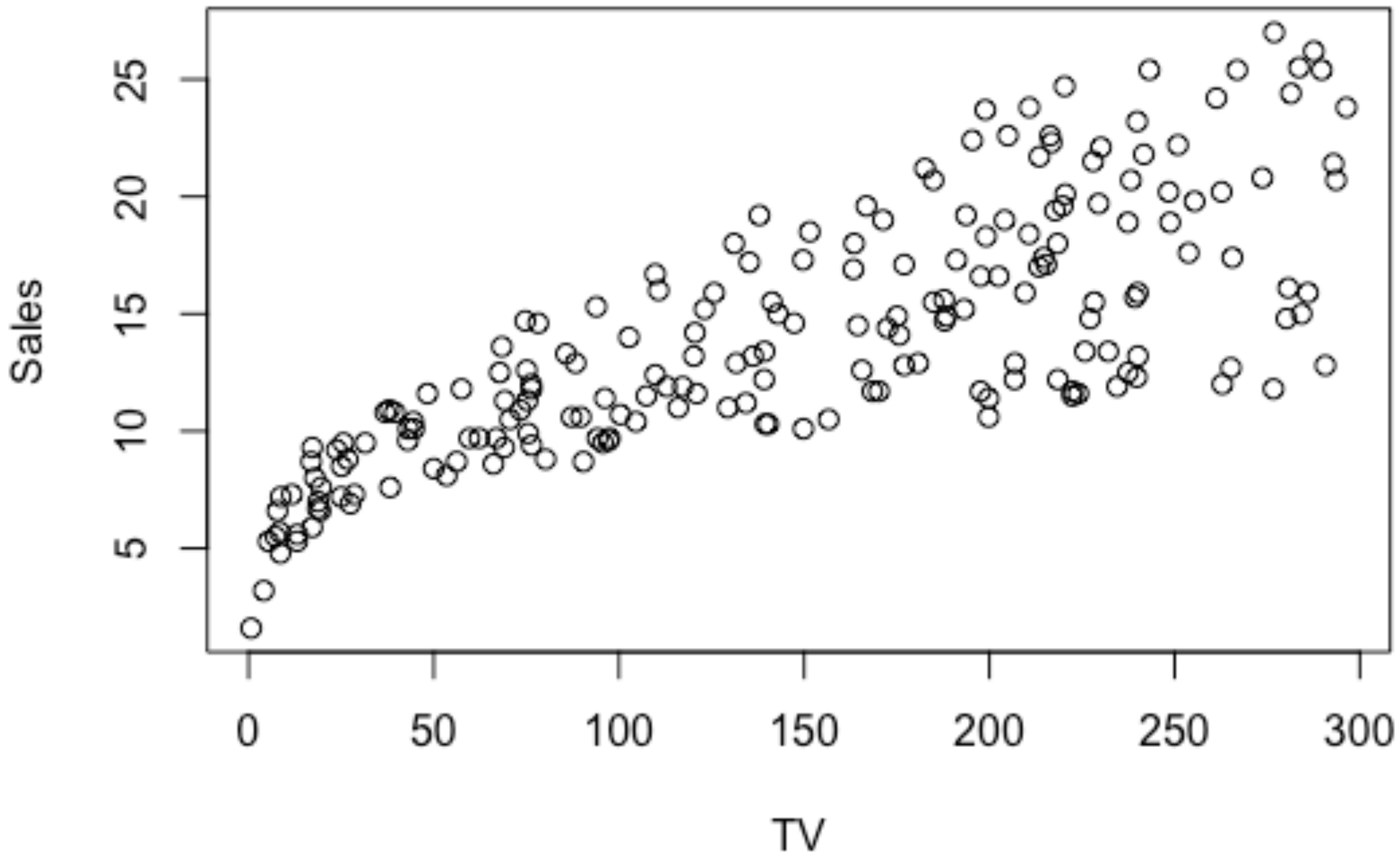
Type of model depends on the response variable

Types of variables	Model
continous	SLR/MLR linear Regression
binary	logistic regression
> 2 categories	multinominal proportional odds

3. SLR Review

Simple Linear Regression

Goal: Examine the relationship between two continuous variables



SLR Model

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

SLR Model

-predictor
-independent
-covariance

independently identically distributed

intercept

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

slope

error

-response
-dependent variable

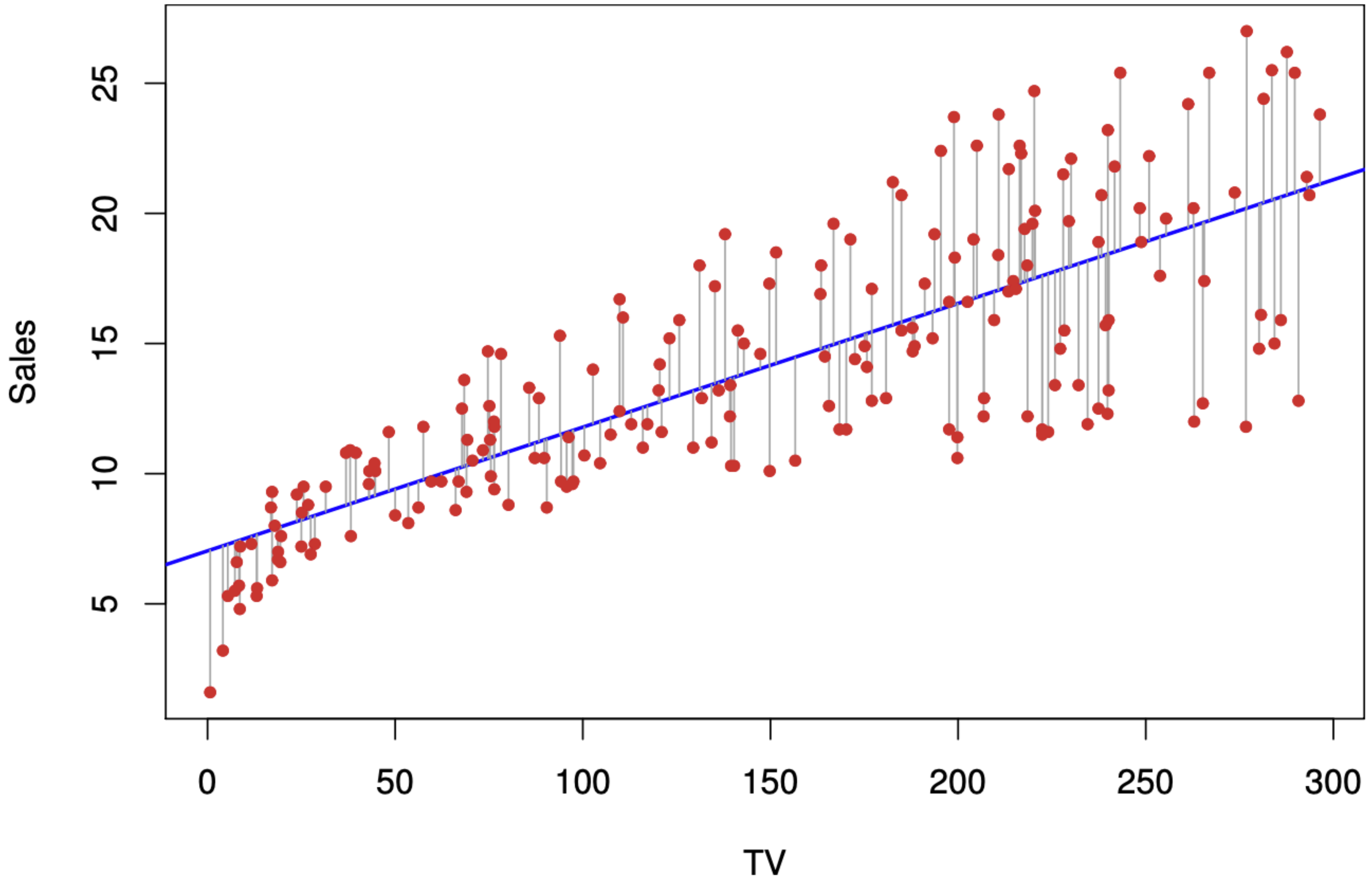
Goals: Estimation and Inference

Assumptions for Linear Regression

- Linear relationship between X and Y
- Independence of errors
- Equal variance of errors
- Normality of errors

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

Estimation: Ordinary Least Squares



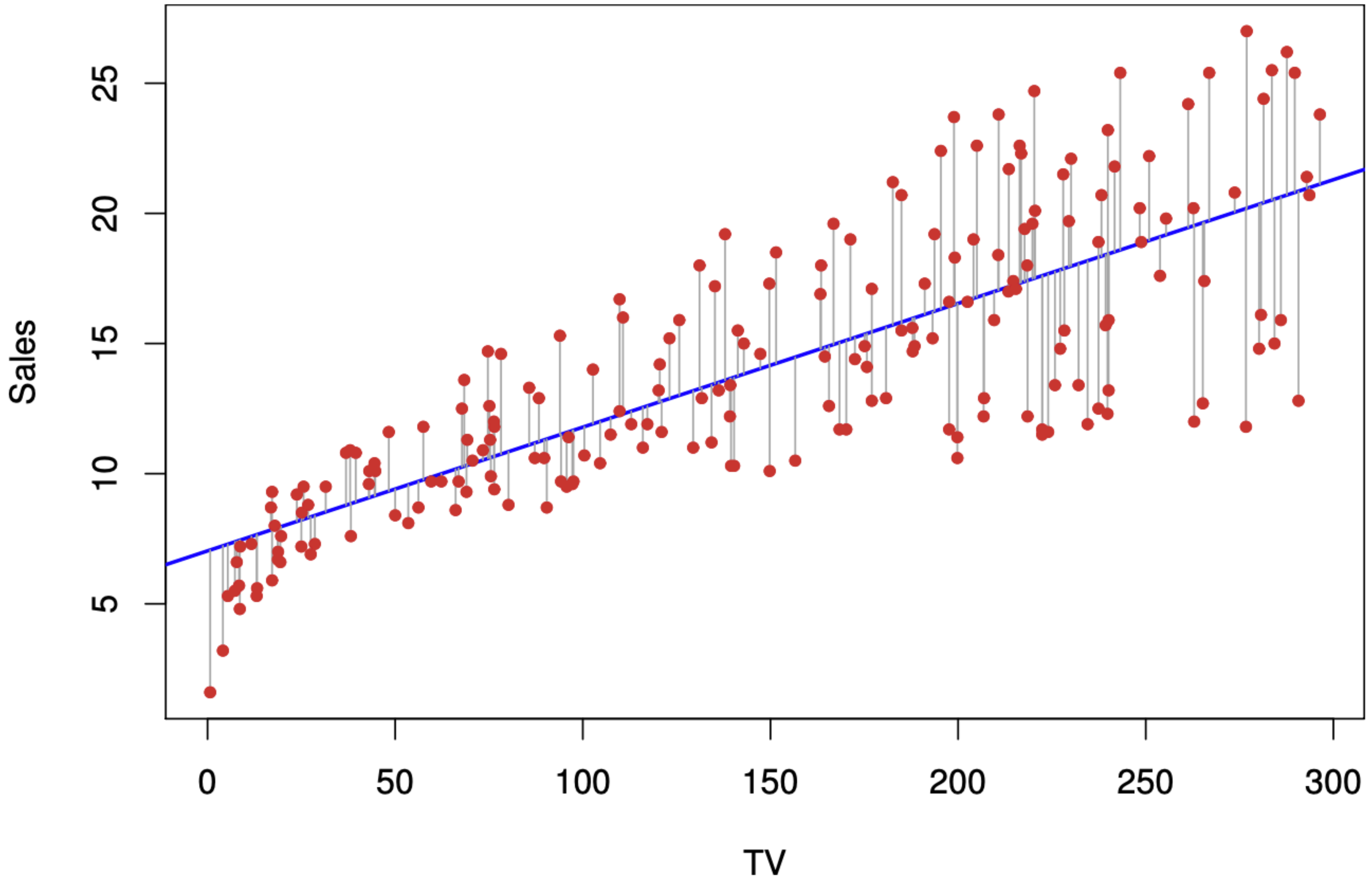
OLS chooses estimates that minimize the **residual sum of squares**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

observed

beta0 + beta1x

Estimated Regression Line



$\hat{\beta}_0$ $\hat{\beta}_1$

$$\hat{Y} = 7.032 + 0.048X$$

Inference

hypothesis is based on the population— so not beta^

$$H_0 : \beta_1 = 0$$

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Assessing the accuracy of the model

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

- Residual standard error (RSE)
  - Estimate of the standard error of ϵ
  - $\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

unbaised

beta1 beta2

- R^2 statistic
  - Proportion of variance explained
  - $1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

residual sum of squares

total sum of squares

[0,1], 越靠近1 越好

0.75 can be explained yby the model

Let’s see it in R!

- Advertising.csv located in Sakai (Resources — datasets or Lessons)

4. EDA/SLR Activity

Your turn!

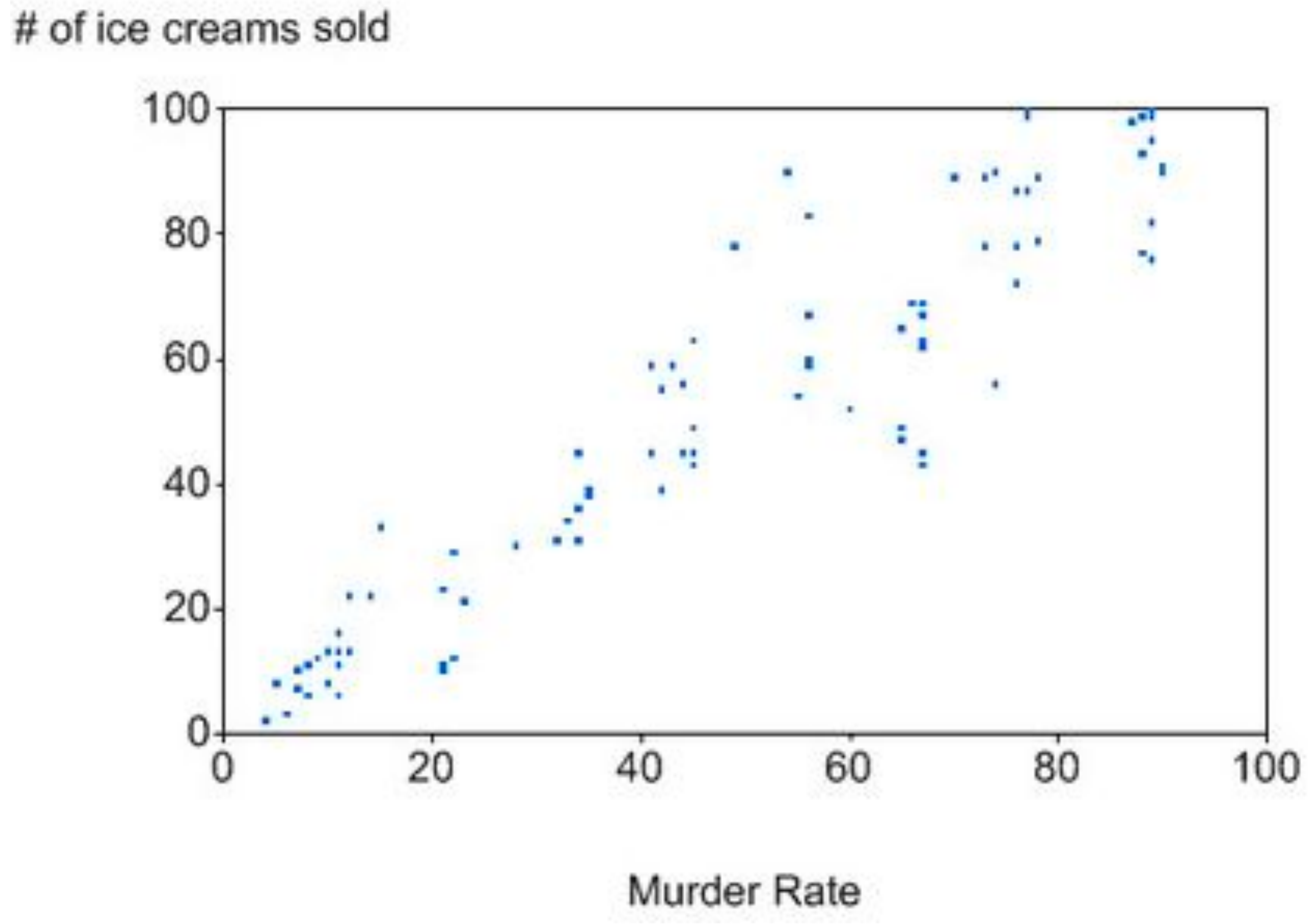
In a group of 2-3, generate a SLR model with the “Boston” dataset

- Install the “ISLR2” package
- data(“Boston”)
- Select two (continuous) variables you are interested in
    - Generate a histogram for each variable
    - Generate a scatter plot to visually assess the relationship
    - Generate the SLR model and note the estimates, relevant p-value, RSE, and R^2

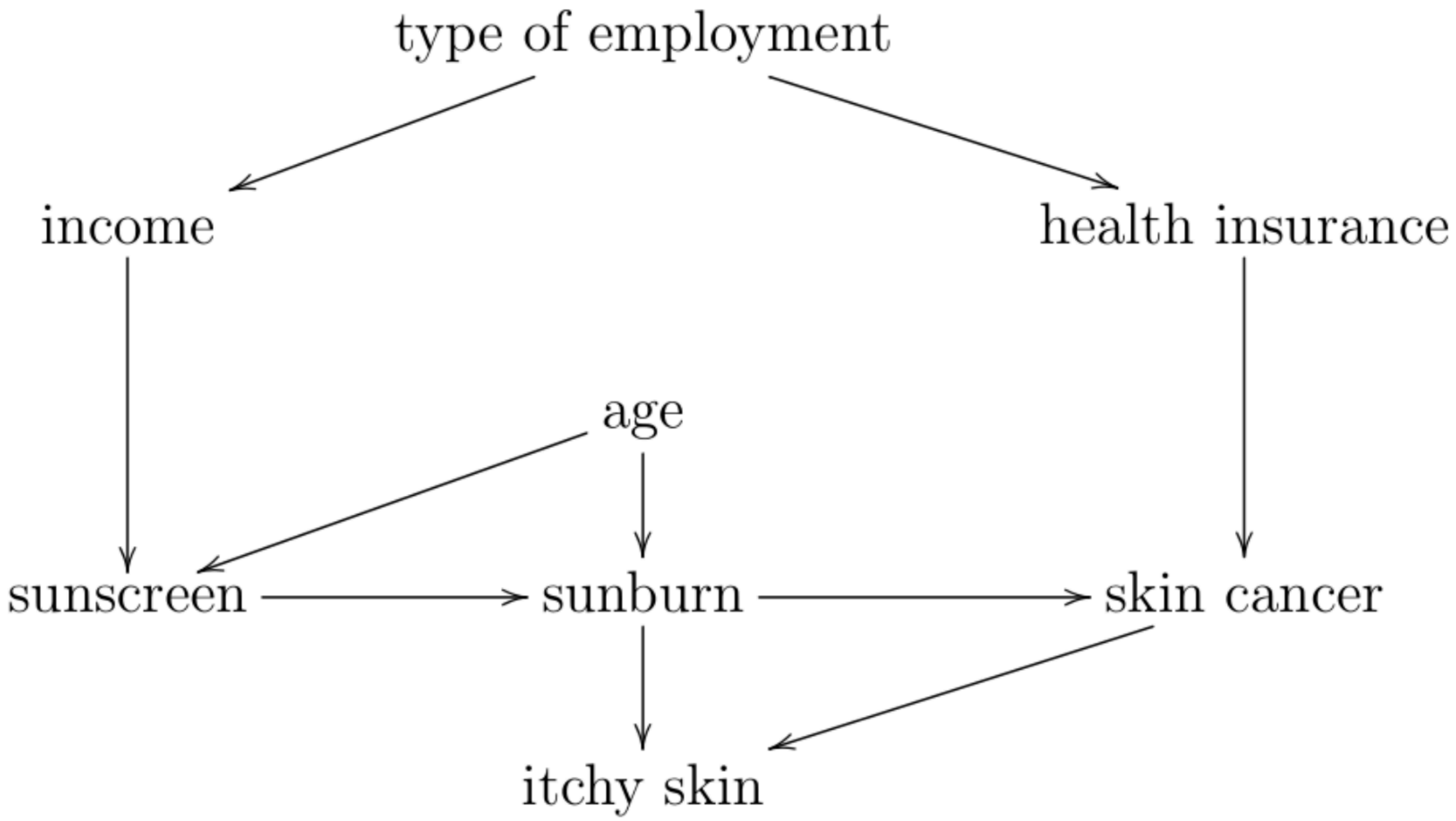
5. MLR

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

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

We can also write the model as:

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

$$p(y_i \,|\, x_i) = N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$$

MLR Assumptions

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

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

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

Matrix Representation

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

Matrix Representation

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

Then the OLS estimates are:

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

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:

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

Hat matrix/Projection matrix:

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

Matrix Representation: SE

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

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

Wrap-up

- Statistical Reflection I due Friday (9/2) 11:55 PM
- Reading for next week will be posted by Friday (9/2) 11:55 PM
- First data analysis assignment will be posted by Tues (9/6) at the latest
  - Due Sept 16