Simple Linear Regression

- Regression model with one predictor variable (Ch.1)
- Statistical Inference in simple linear regression (Ch.2)
- Checking Assumptions: diagnosis and remedial measure (Ch. 3)
- Simultaneous Intervals and other topics (Ch.4)
- Using Statistical software (R, SPSS, etc.)

Regression model with one predictor

- General strategy for data analysis.
- What is regression?
- The statistical/mathematical model for simple linear regression (SLR).
- Estimating the regression coefficients.
- Pitfalls of regression

General Strategy in Data Analysis

1. Define the problem of interest
2. Collect data
   - Representative samples: simple random samples
   - Experiment vs. observational study
3. Explore the data
   - Graphical and numerical summaries
4. Propose and run candidate models
   - Linear vs. Non-linear, choice of predictors, etc.
5. Compare, diagnose and modify models
6. Choose the “best” model
7. Make inference based on the “best model”

Simple Linear Regression

- Example: Blood Alcohol

  - How much alcohol can one consume before one's blood alcohol content (BAC) is above the legal limit? A statistics project at the University of Western Sydney in Sydney, Australia explored the relationship between BAC and other variables such as amount of alcohol consumed, weight, gender and age.

ID	Gender	Weight	Height	Age	BAC	Wine
1	female	70	167	20	0.025	4
2	female	66	161	21	0.040	4
3	male	67	169	27	0.070	6
4	male	91	187	20	0.065	6
5	female	58	158	25	0.015	3

  - A new variable, alcohol-to-weight ratio, is created:

$$AW = (\text{number of wine})/(\text{weight} - 20\text{kg})$$

  - In this example, we'll focus on BAC and AW.

Simple Linear Regression

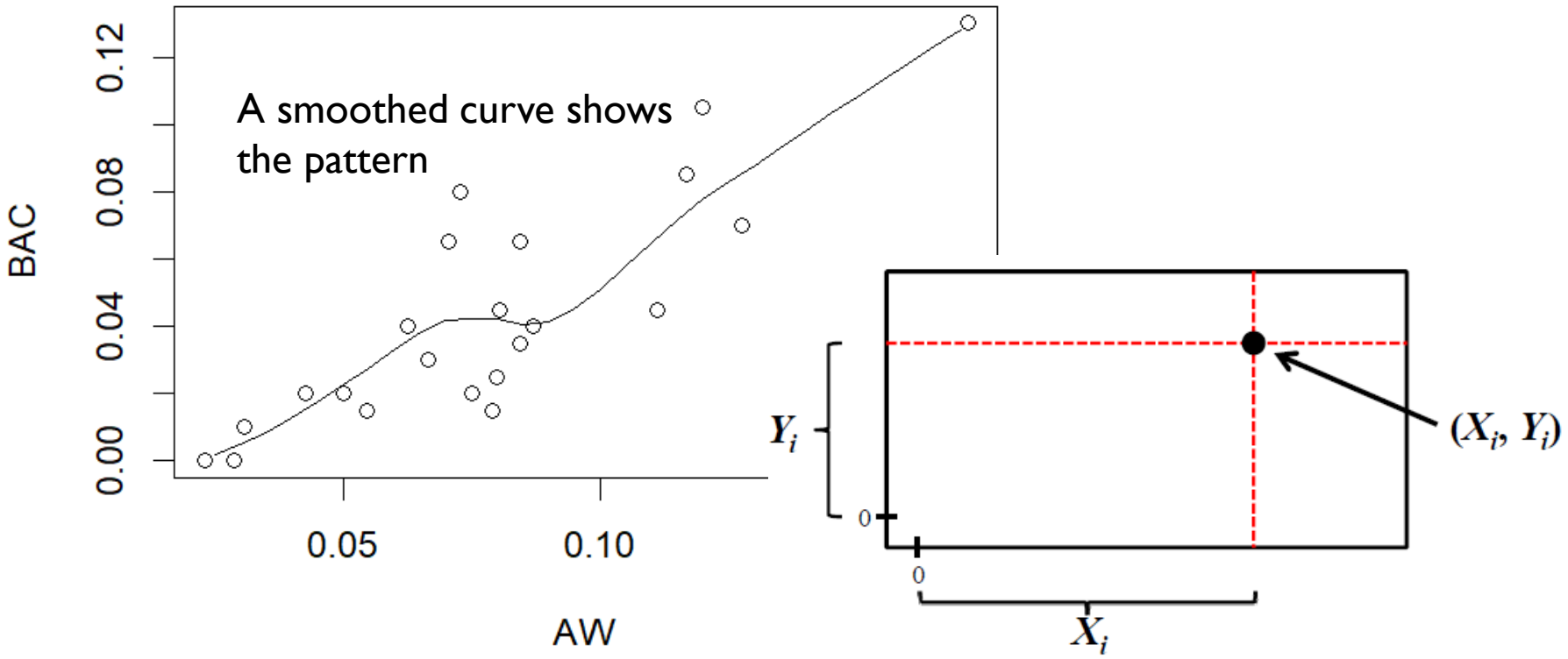
- Questions of interest:
  - How does BAC change over AW?
  - Can we predict the BAC given AW?

- **Response (Dependent)** variable:
  - Y: BAC (Blood Alcohol Content)

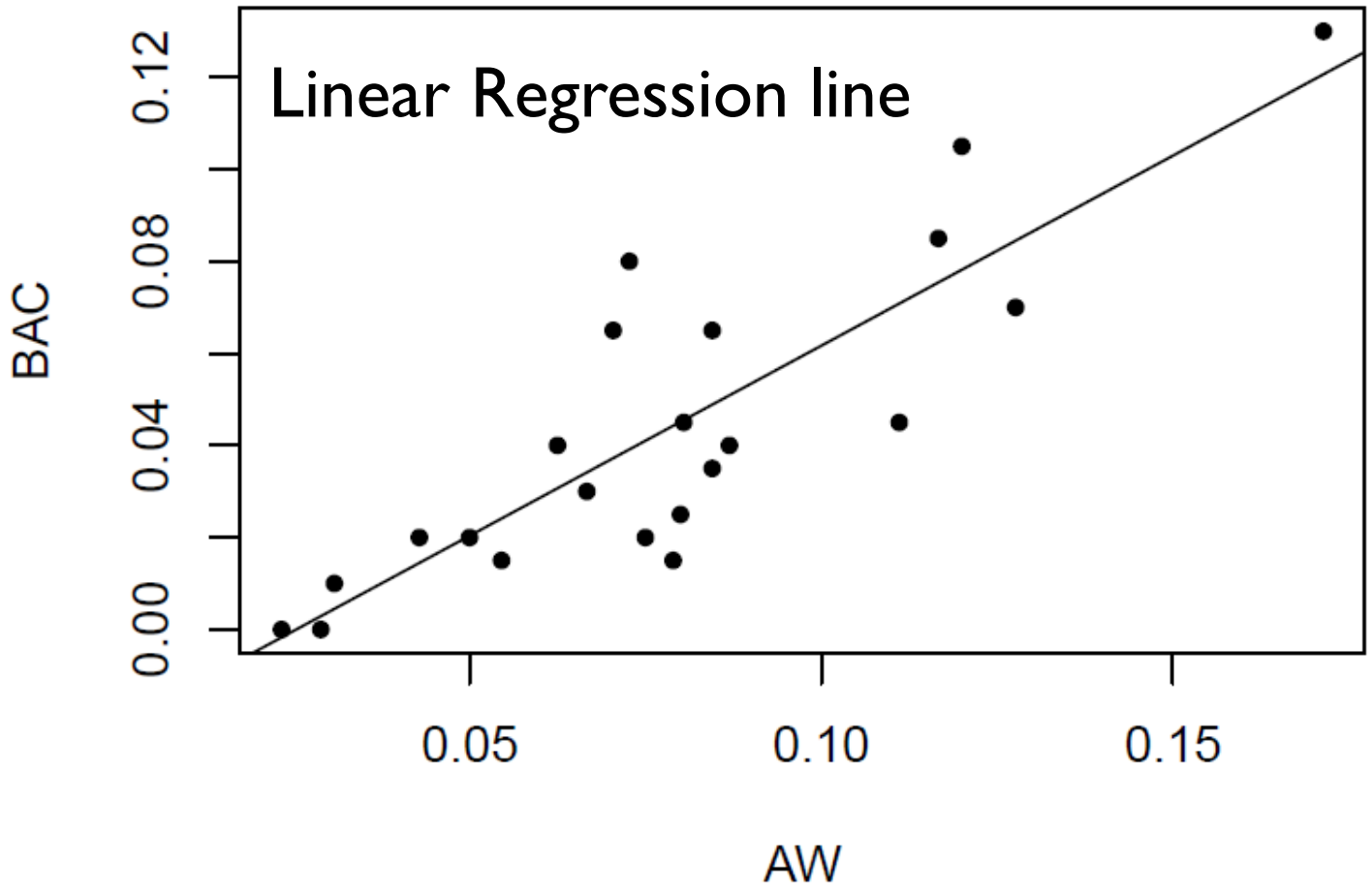
- **Explanatory (Independent, Predictor)** variable:
  - X: AW (Alcohol-to-Weight ratio)

- Explore the relation in a graph:
  - Can we use a line to describe the relationship?

Explore the relationship: Scatter plot



Explore the relationship: Scatter plot



Simple Linear Regression Model

- Recall the mathematical function for a straight-line:

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

- A statistical model will consider the randomness:

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

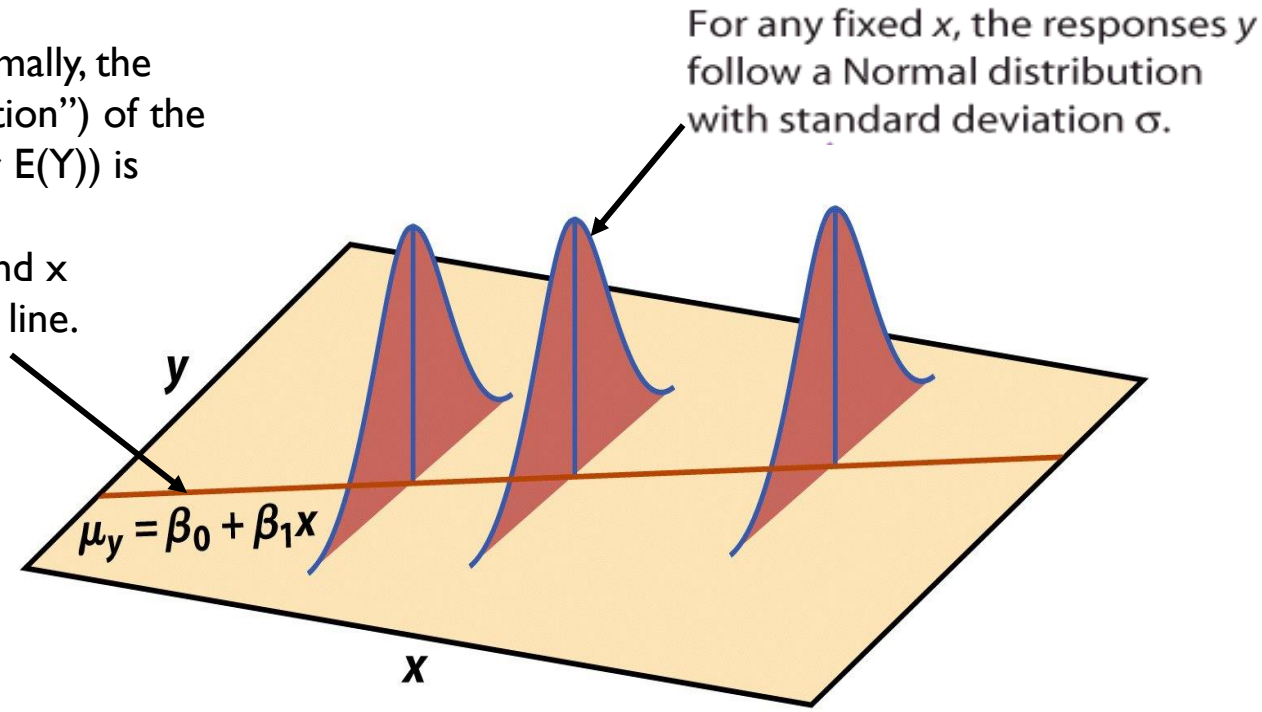
  - β_0 is the intercept (meaning ……)
  - β_1 is the slope (meaning ……)
  - ε_i is a random error term
    - E(ε_i)=0, Var(ε_i)=s^2 , and ε_i and ε_j are uncorrelated
    - More strictly, ε_i is a "normally distributed" random variable

Simple Linear Regression Model

- Under the assumptions that the errors, ε_i, $i = 1, 2, \ldots, n$, are independently and identically distributed as N(0, σ^2):

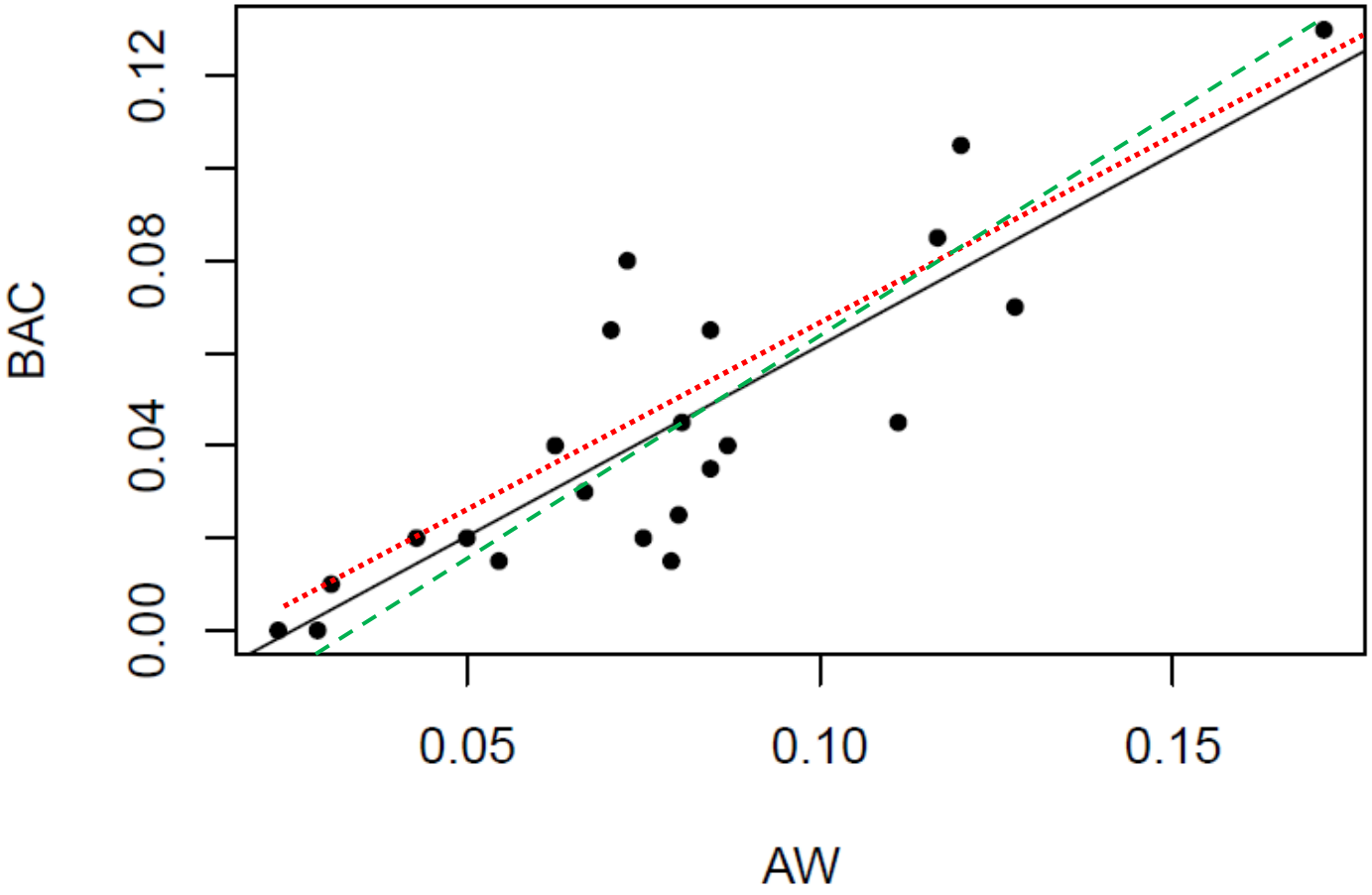
$$Y_i \text{ is } N(\beta_0 + \beta_1 X_i, \sigma^2)$$

For any fixed x, the mean (formally, the "expected value" or "expectation") of the responses y (denoted as μ_y or E(Y)) is determined by x.
The association between μ_y and x is described by the regression line.



- Does NOT imply the collection of Y_i are normally distributed

Estimating the regression line



Which line is better?

Estimating the regression line: OLS

- Least Square Estimator (OLS):
  - Let b_0 (or $\hat{\beta}_0$) be the estimated intercept, b_1 (or $\hat{\beta}_1$) be the estimated slope. The predicted value is:

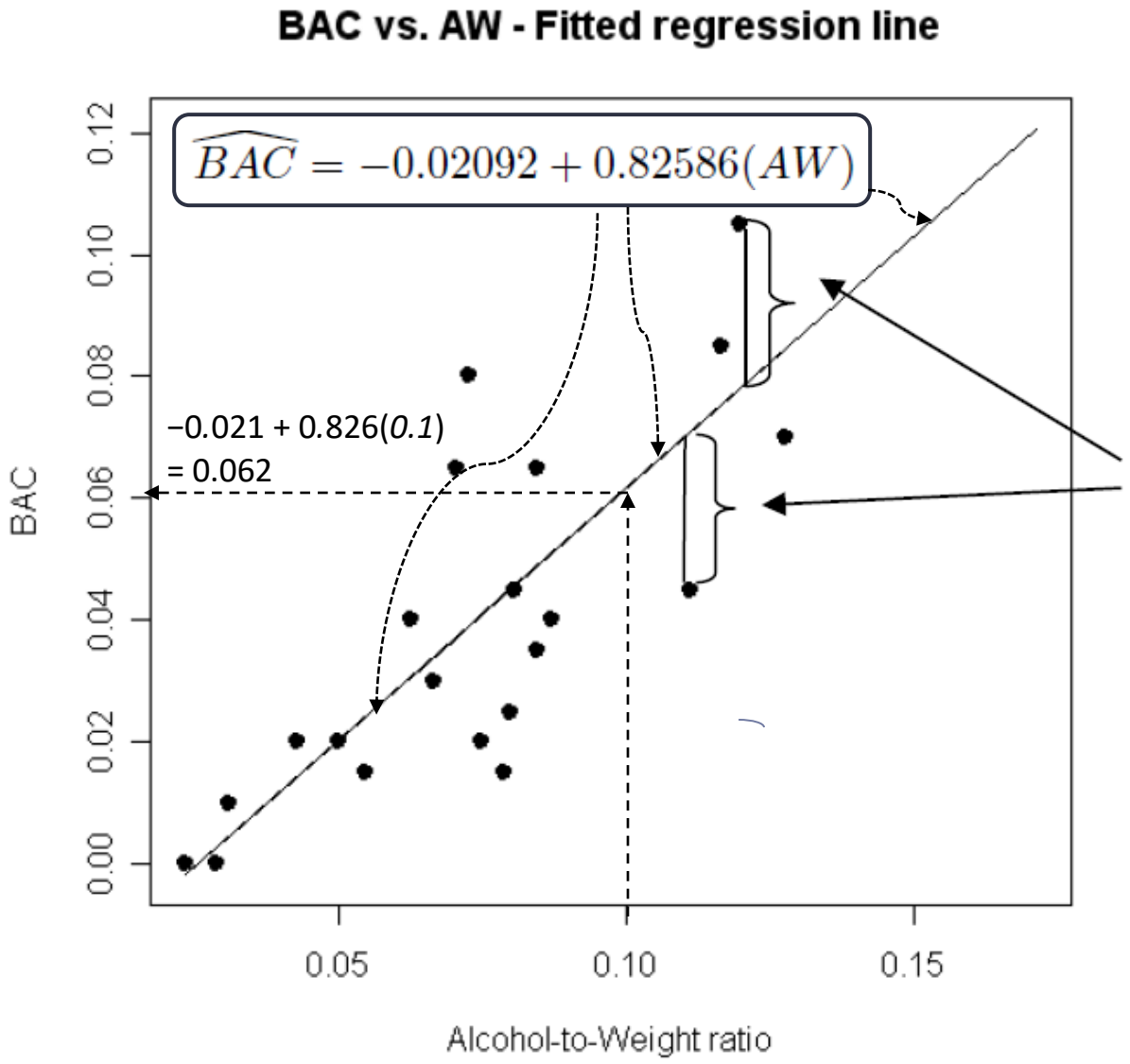
$$\hat{y}_i = b_0 + b_1 x_i$$

  - Choose b_0 *and* b_1 such that the sum of the squared vertical (y) distances between the data points and the line is as small as possible. I.e., minimize Q, where

$$Q = \sum \left(y_i - \hat{y}_i \right)^2 = \sum \left(y_i - (b_0 + b_1 x_i) \right)^2$$

  - The vertical distance is also called the **residual**

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$



Residuals:
Deviations from the observed response to the estimated response (regression line.)
for two data points.

$e_i = y_i - \hat{y}_i$

Estimating the regression line: OLS

- The math of Least Square Estimator (OLS):
  - A mathematical problem: find b_0 *and* b_1 to minimize:

$$Q = \sum \left(y_i - \hat{y}_i \right)^2 = \sum \left(y_i - (b_0 + b_1 x_i) \right)^2$$

  - There is a unique solution:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where r is the correlation, s_y, s_x are the standard deviation of y and x, respectively.

(We will let software take care of the mathematics.)

Estimating the regression line: MLE

- Maximum Likelihood Estimator (MLE):

$$y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}\right)^2}$$

$$L = f_1 \cdot f_2 \cdot \ldots \cdot f_n \text{ (likelihood function)}$$

Find β_0 and β_1 which maximizes L

Estimating the regression line: variance

- To estimate σ^2, the variance of the error, use

$$s^2 = \frac{\sum e_i^2}{\mathrm{n}-2} = \frac{\sum (y_i - \hat{y}_i)^2}{\mathrm{n}-2} = \frac{SSE}{df_E} = MSE$$

$$s = \sqrt{s^2} = \sqrt{MSE} = RMSE(RootMSE)$$

Properties of OLS regression line

- The line always goes through $(\bar{x}, \bar{y})$
- The sum of the residuals is 0, i.e, $\sum e_i = 0$
- The sum of the squared residuals is minimized by the definition of least squares.
- More can be found in text, p.23, 24

Common pitfalls of regression

- Correlation vs causation
- Knowledge of subject matter
- Model assumptions and how to check them
- Alternatives to OLS
- Extrapolation

Statistical Inference for SLR

- Assumptions
- Inferences concerning regression coefficients
- Confidence Interval and Prediction Interval
- The ANOVA table
- Coefficient of determination
- Other considerations