Linear Regression

C1 Birds

Wing Span cm/in

Beak size mm/cm

Color categorical

Body mass g

wing span → body mass

$X \sim \text{Wingspan} \sim N(\mu, \sigma^2)$

$Y \sim \text{Body Mass} \sim N(\mu, \sigma^2)$

$$Y = f(X) + \varepsilon$$

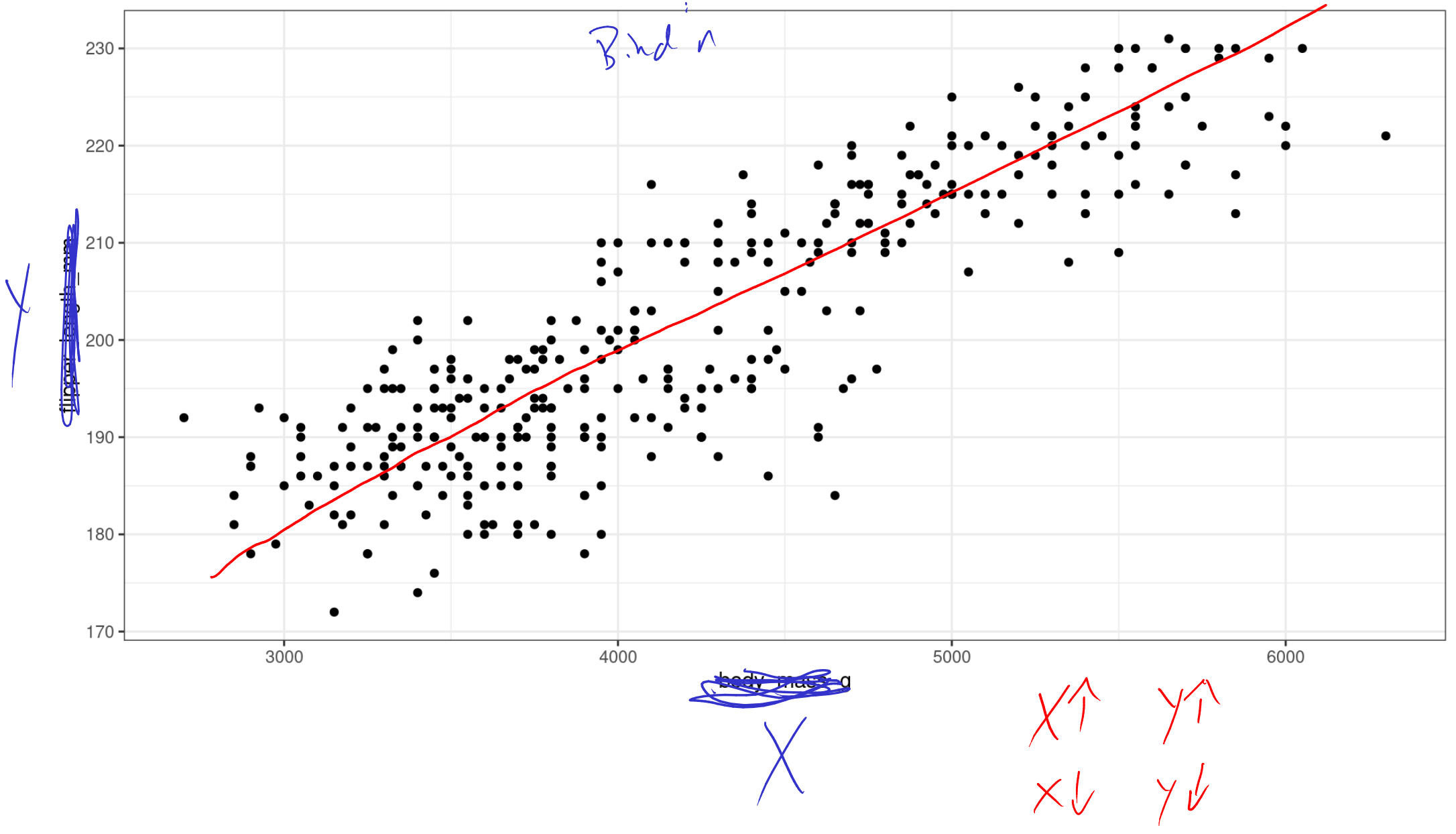
$$\varepsilon \sim N(0, \sigma^2)$$

Learning Outcomes

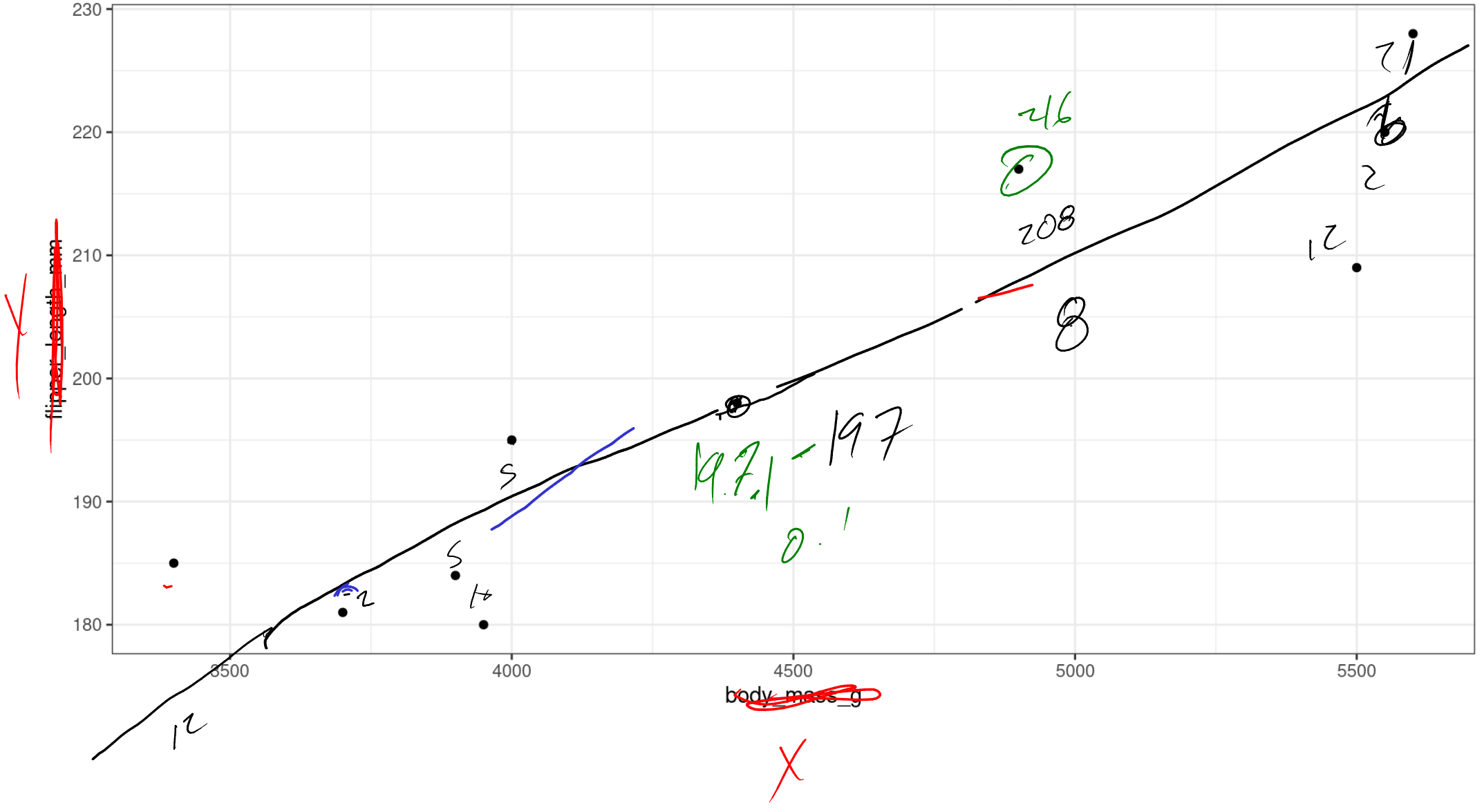
- Scatter Plot
- Linear Regression
- Ordinary Least Squares
- Unbiasedness

Scatter Plot

Scatter Plot



Scatter Plot



Linear Regression

Linear Regression

Linear regression is used to model the association between a set of predictor variables (x's) and an outcome variable (y). Linear regression will fit a line that best describes the data points.

Simple Linear Regression

Simple linear regression will model the association between one predictor variable and an outcome:

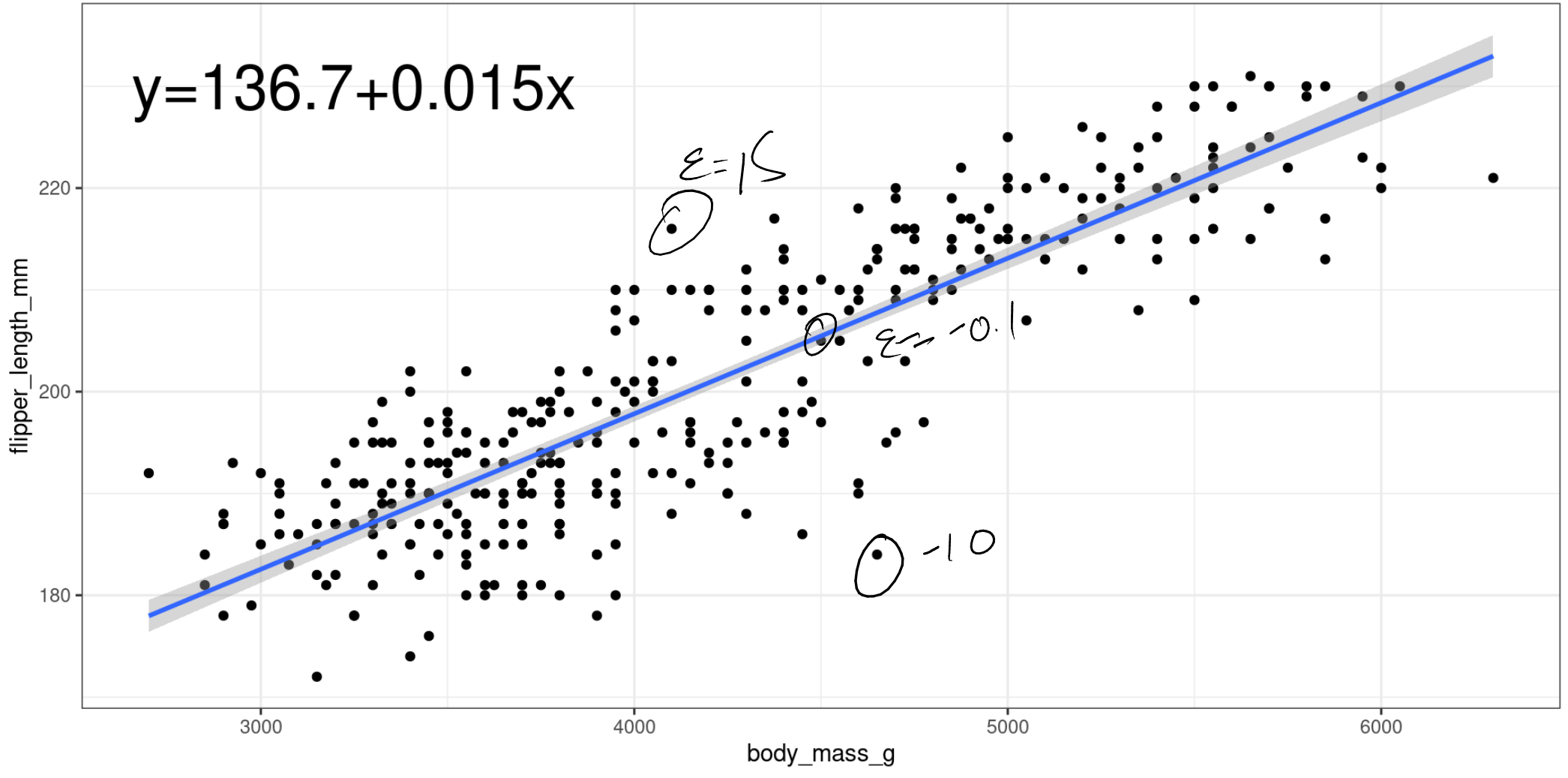
$$Y = \beta_0 + \beta_1 X + \epsilon$$

intercept + slope + error term

- β_0: Intercept term
- β_1: Slope term
- $\epsilon \sim N(0, \sigma^2)$

Fitting a Line

$y = \beta_0 + \beta_1 x$



Interpretation

$$\hat{y} = 136.73 + 0.015x$$

Body Mass is 0, the flipper length is 136.73

X Y

As X increase 1 unit, Y increase by an average of 0.015 units

Ordinary Least Squares

Ordinary Least Squares

For a data pair $(X_i, Y_i)_{i=1}^n$, the ordinary least squares estimator will find the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following function:

$$\ell = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

Estimating β's

$$\frac{d\ell}{d\beta_0}$$

$$\frac{d\ell}{d\beta_1}$$

Estimating β_1

$$\sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

$$\frac{d\ell}{d\beta_1} = \sum_{i=1}^{n} 2\left(Y_i - (\beta_0 + \beta_1 X_i)\right)(-X_i)$$

$$= -2 \sum \left(Y_i - \beta_0 - \beta_1 X_i\right) X_i \qquad \bar{X} = \frac{\sum X_i}{n}$$

$$= -2 \sum \left(X_i Y_i - \beta_0 X_i - \beta_1 X_i^2\right)$$

$$= -2 \left[\sum X_i Y_i - \beta_0 n\bar{X} - \beta_1 \sum X_i^2\right]$$

$$0 = -2 \left[\sum X_i Y_i - \beta_0 n\bar{X} - \beta_1 \sum X_i^2\right]$$

$$0 = \sum X_i Y_i - \beta_0 n\bar{x} - \beta_1 \sum x_i^2$$

$$\frac{\beta_1 \sum x_i^2}{\sum x_i^2} = \frac{\sum X_i Y_i - \beta_0 n\bar{x}}{\sum X_i^2}$$

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - \beta_0 n\bar{x}}{\sum X_i^2}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Estimating β_0

Estimates

intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

slope

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

model error

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

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

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Unbiasedness of β's

Unbiasedness of β's

Both β_0 and β_1 are unbiased estimators.

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

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

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

$$E(\varepsilon) = 0$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 X$$

$$Var(Y) = \sigma^2$$

$$Var(\beta_0 + \beta_1 X + \varepsilon)$$

$$Var(\beta_0) + Var(\beta_1 X) + Var(\varepsilon)$$

$$0 + 0 + \sigma^2$$

$E(\hat{\beta}_0)$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ $\quad E(\hat{\beta}_1) = \beta_1$

$\frac{\sum x_i}{n} = \bar{x}$

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - E(\hat{\beta}_1 \bar{x})$$

$$= E(\bar{y}) - \bar{x}\, E(\hat{\beta}_1) \qquad E(y_i) = \beta_0 + \beta_1 x_i$$

$$E(\bar{y}) - \bar{x}\beta_1$$

$$E\left(\frac{1}{n}\sum y_i\right) - \bar{x}\beta_1$$

$$\frac{1}{n}\sum E(x_i) - \bar{x}\beta_1$$

$$\frac{1}{n}\sum (\beta_0 + \beta_1 x_i) - \bar{x}\beta_1$$

$$\frac{1}{n}\left[n\beta_0 + \beta_1 \Sigma x_i\right] - \bar{x}\beta_1$$

$$\beta_0 + \cancel{\beta_1 \bar{x}} - \cancel{\bar{x}\beta_1}$$

$$\beta_0$$

$E(\hat{\beta}_1)$

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_1) = E\left(\frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)$$

$$\frac{1}{\sum (x_i - \bar{x})^2} E\left(\sum (y_i - \bar{y})(x_i - \bar{x}) \right)$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum_{i=1}^{n} E\left[(y_i - \bar{y})(x_i - \bar{x}) \right]$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) E(y_i - \bar{y})$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x})\left(E(y_i) - E(\bar{y})\right)$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x})\left(\cancel{\beta_0} + \beta_1 x_i - \cancel{\beta_0} - \beta_1 \bar{x}\right)$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x})\left(\beta_1 x_i - \beta_1 \bar{x}\right)$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x})\beta_1 (x_i - \bar{x})$$

$$\frac{1}{\sum (x_i - \bar{x})^2} \beta_1 \sum \cancel{(x_i - \bar{x})^2}$$

$$\frac{1}{\cancel{\sum (x_i - \bar{x})^2}} \beta_1 \cancel{\sum (x_i - \bar{x})^2}$$

$$= \beta_1$$