CSCI 3022: Intro to Data Science

Lecture 20: Introduction to Regression

Rachel Cox

Department of Computer Science

Announcements & Reminders

- HW4 Due Monday April 6th at 11:59pm
  - normal usage of late days

What will we learn today?

- ❑ Predictive statistics
- ❑ Simple linear regression (SLR) model
- ❑ Residuals
- ❑ Sum of squared-errors
- ❑ Fitted regression line
- ❑ Least-squares line

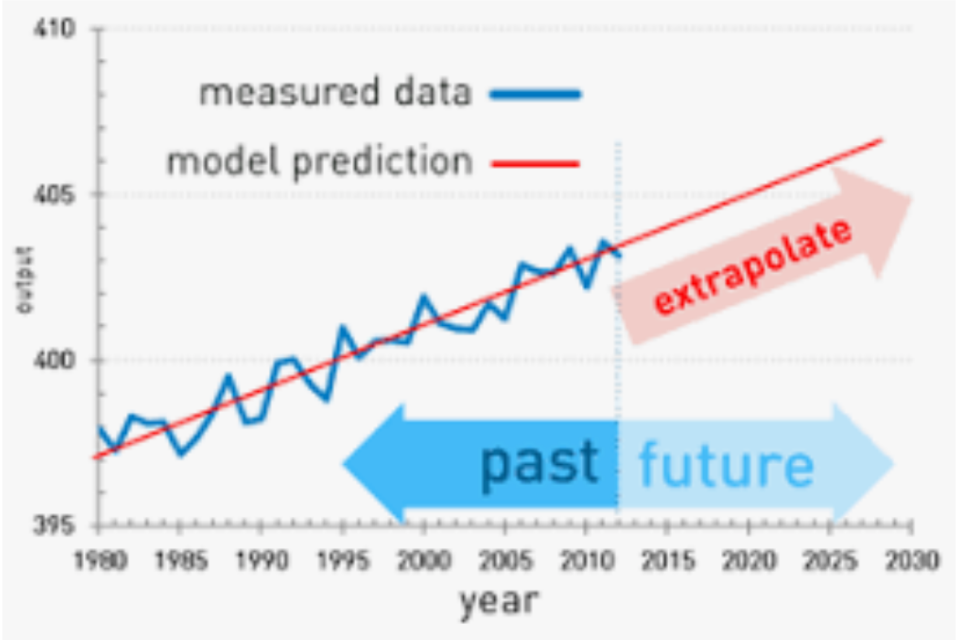
- ❑ *A Modern Introduction to Probability and Statistics, Chapter 22*

Statistical Modeling

So far, we've talked about:

- ❑ Descriptive statistics: "this is the way my sample is"
- ❑ Inferential statistics: "This is what I can conclude from my sample [with $100(1-\alpha)\%$ confidence]"

- ➢ Today: predictive statistics



Linear regression for prediction

Examples:

- Given a person's age and gender, predict their height
- Given the area of a house, predict its sale price
- Given unemployment, inflation, number of wars and economic growth, predict the president's approval rating
- Given a person's browser history, predict how long they'll stay on a product page
- Given the advertising budget expenditures in various media markets, predict the number of products they'll sell.

Simple Linear Regression

line: $y = mx + b$

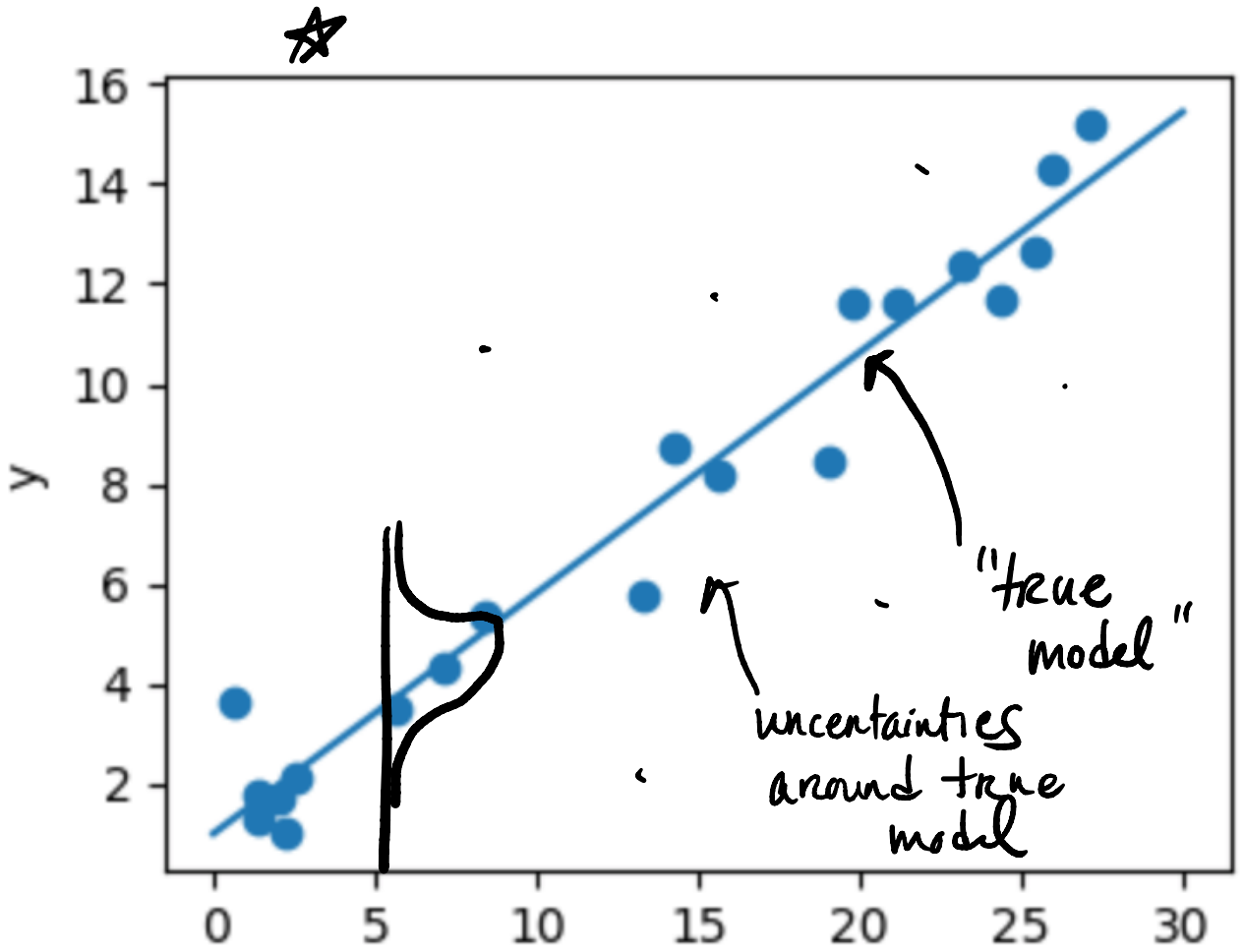
Our goal: figure out the equation of the line through the data.

Definitions and Assumptions:

1) $y_i = \alpha + \beta\, x_i + \epsilon_i$

2) Each of the ϵ_i are independent

3) $\epsilon_i \sim N(0, \sigma^2)$



Simple Linear Regression

noisy process

$$Y = \alpha + \beta X + \epsilon$$

linear equation

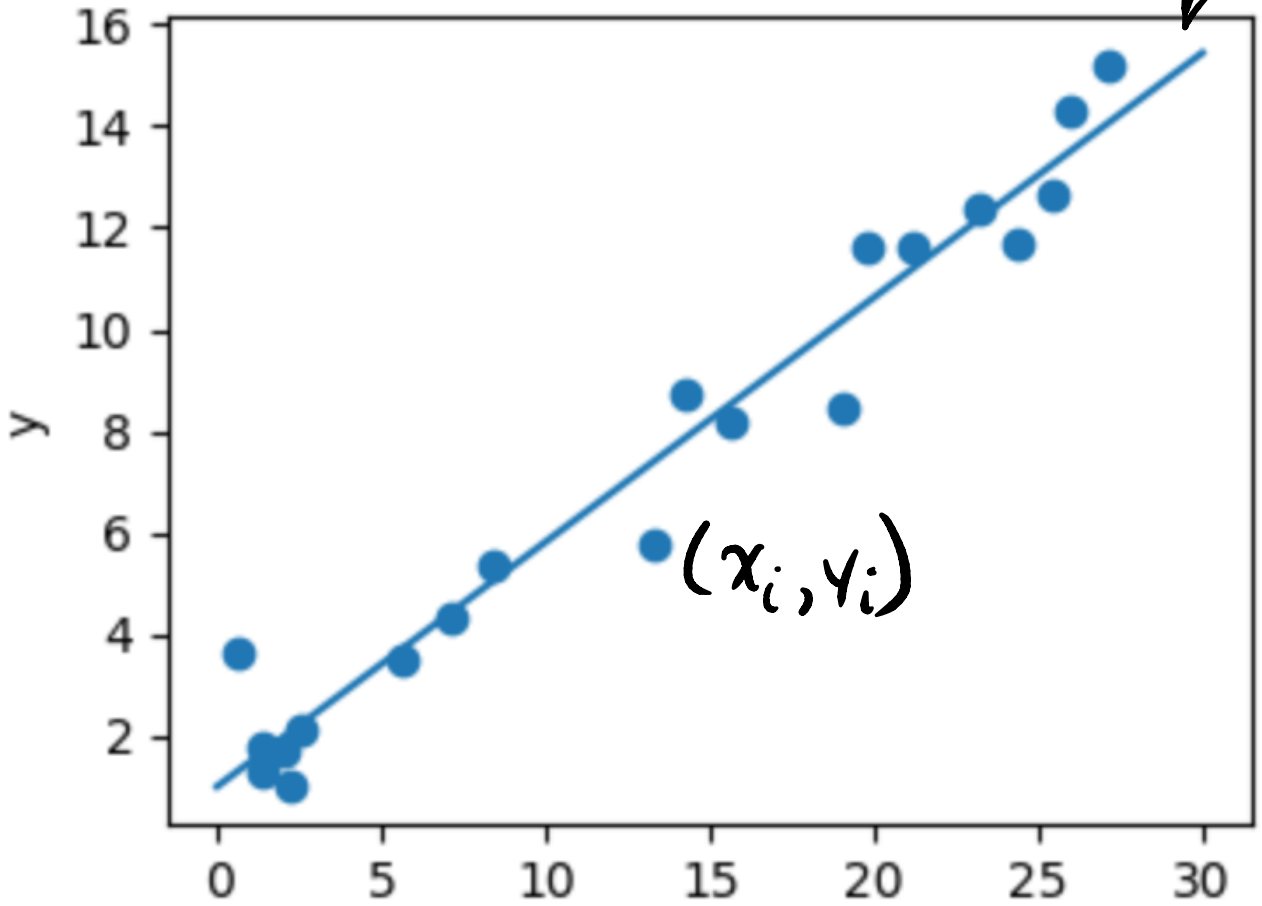
X: the independent variable, the predictor, the explanatory variable, the feature

Y: the dependent variable, the response variable

ϵ: the random deviation, random error – accounts for the fact that the world is uncertain and that there are random deviations around the true process.

Simple Linear Regression (SLR)

$$Y = \alpha + \beta X + \epsilon$$



The points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations will be scattered about the true regression line.

Simple Linear Regression

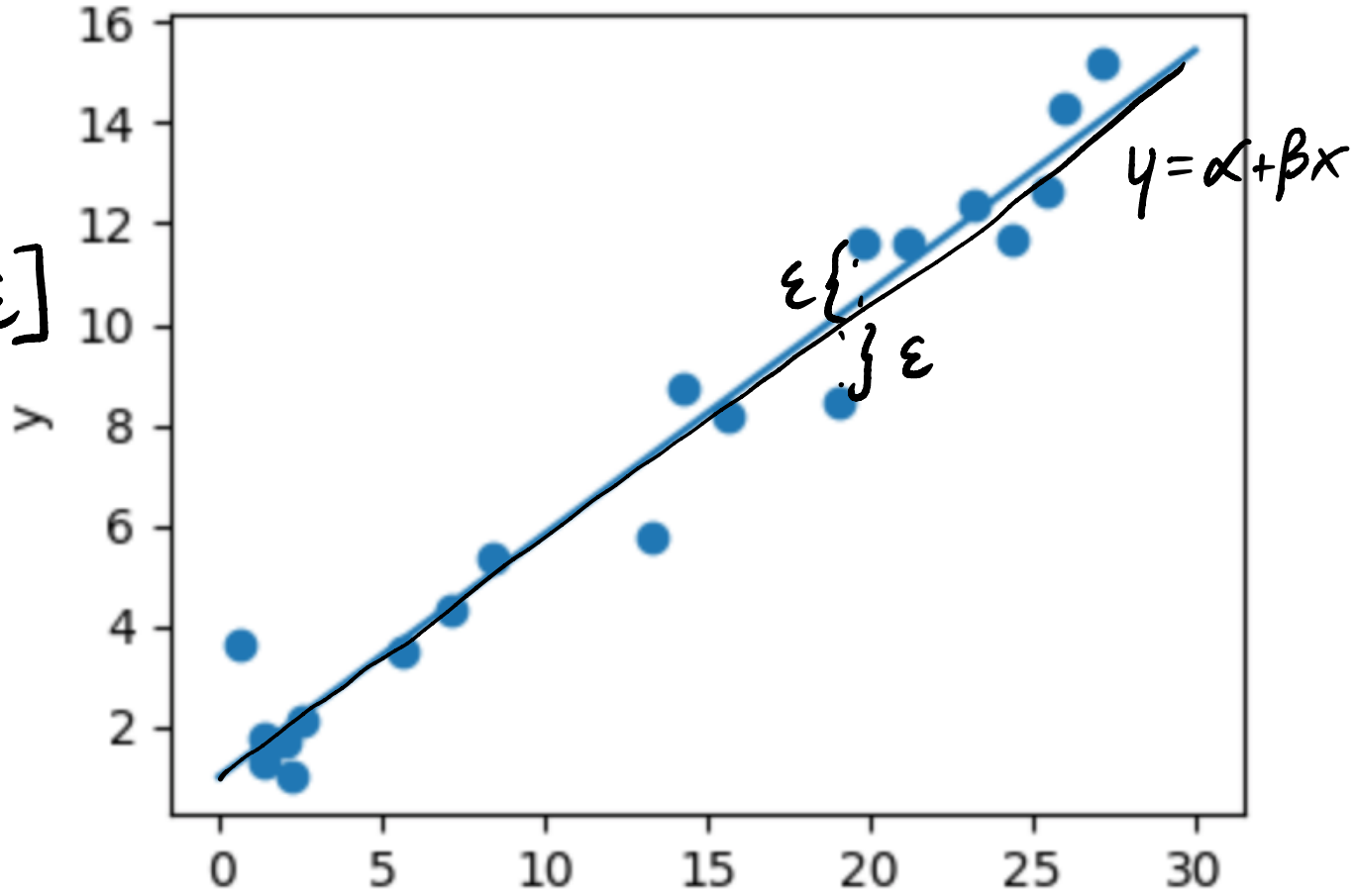
$\varepsilon \sim N(0, \sigma^2)$

✓ model for data set

Y is a random variable.
What is $E[Y]$?

$$E[Y] = E[\alpha + \beta X + \varepsilon]$$
$$= E[\alpha] + E[\beta X] + E[\varepsilon]$$
$$= \alpha + \beta x + 0$$
$$= \alpha + \beta x$$

$$Y = \alpha + \beta X + \epsilon$$



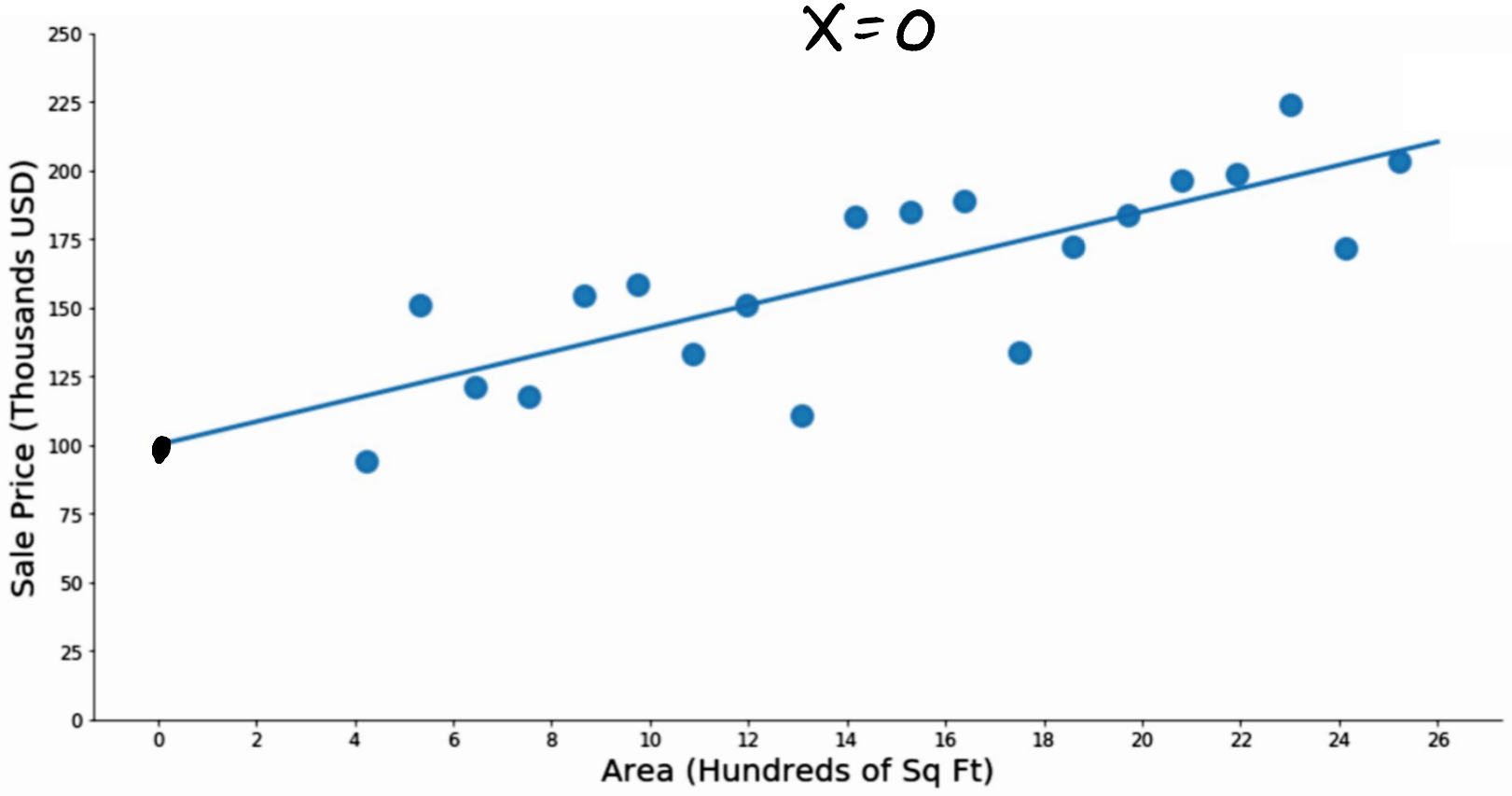
Simple Linear Regression – Interpreting parameters

true model

α is the intercept of the true regression line (aka the baseline average)

It's called the "baseline" because it's the response when the feature = 0

$X = 0$



Simple Linear Regression – Interpreting parameters

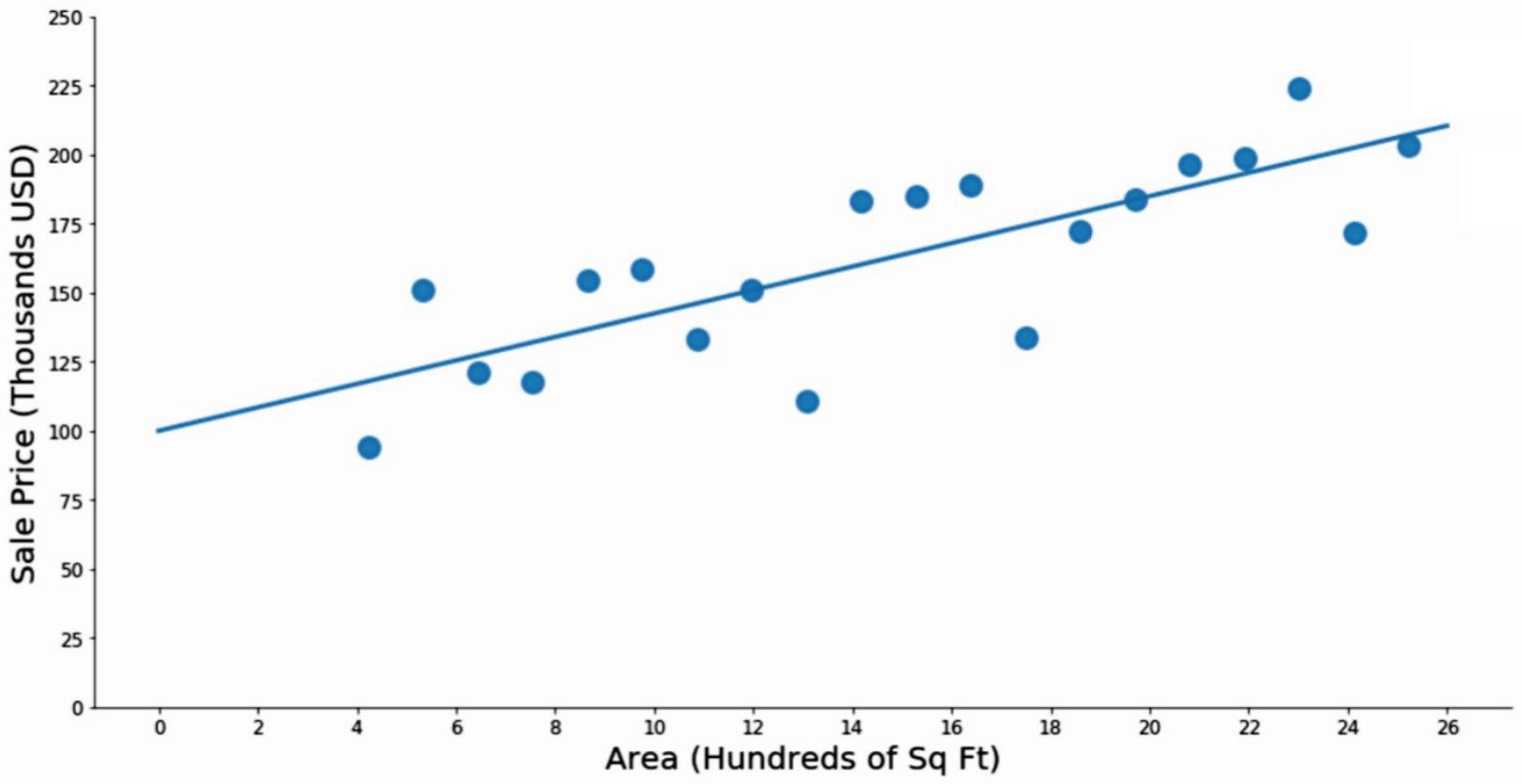
β is the slope of the true regression line

It's the increase in our response from a unit increase in our feature

$$Y(x+1) = \alpha + \beta(x+1) + \varepsilon$$

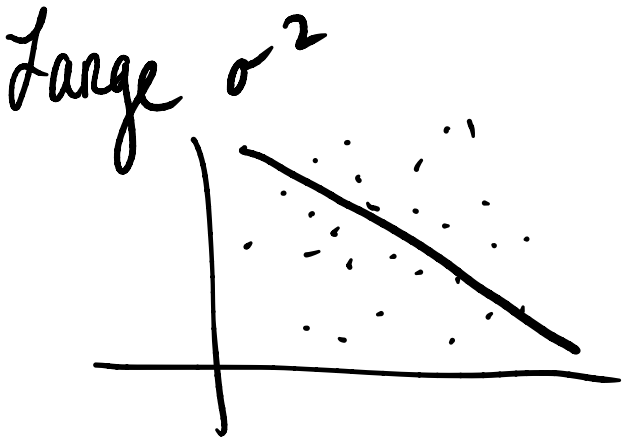
$$\bar{Y}(x) = \alpha + \beta x + \varepsilon$$

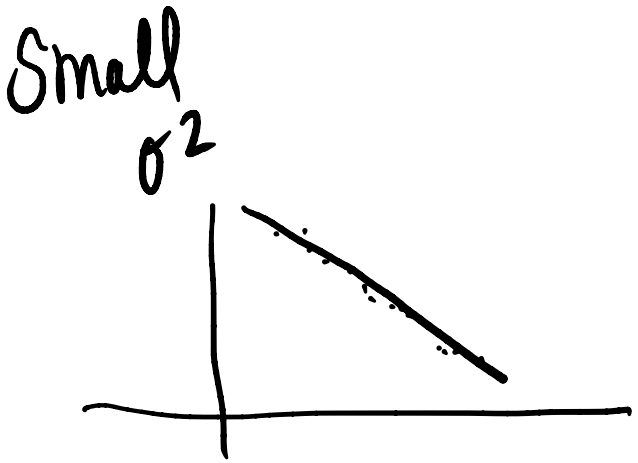
$$\Delta y = \beta$$

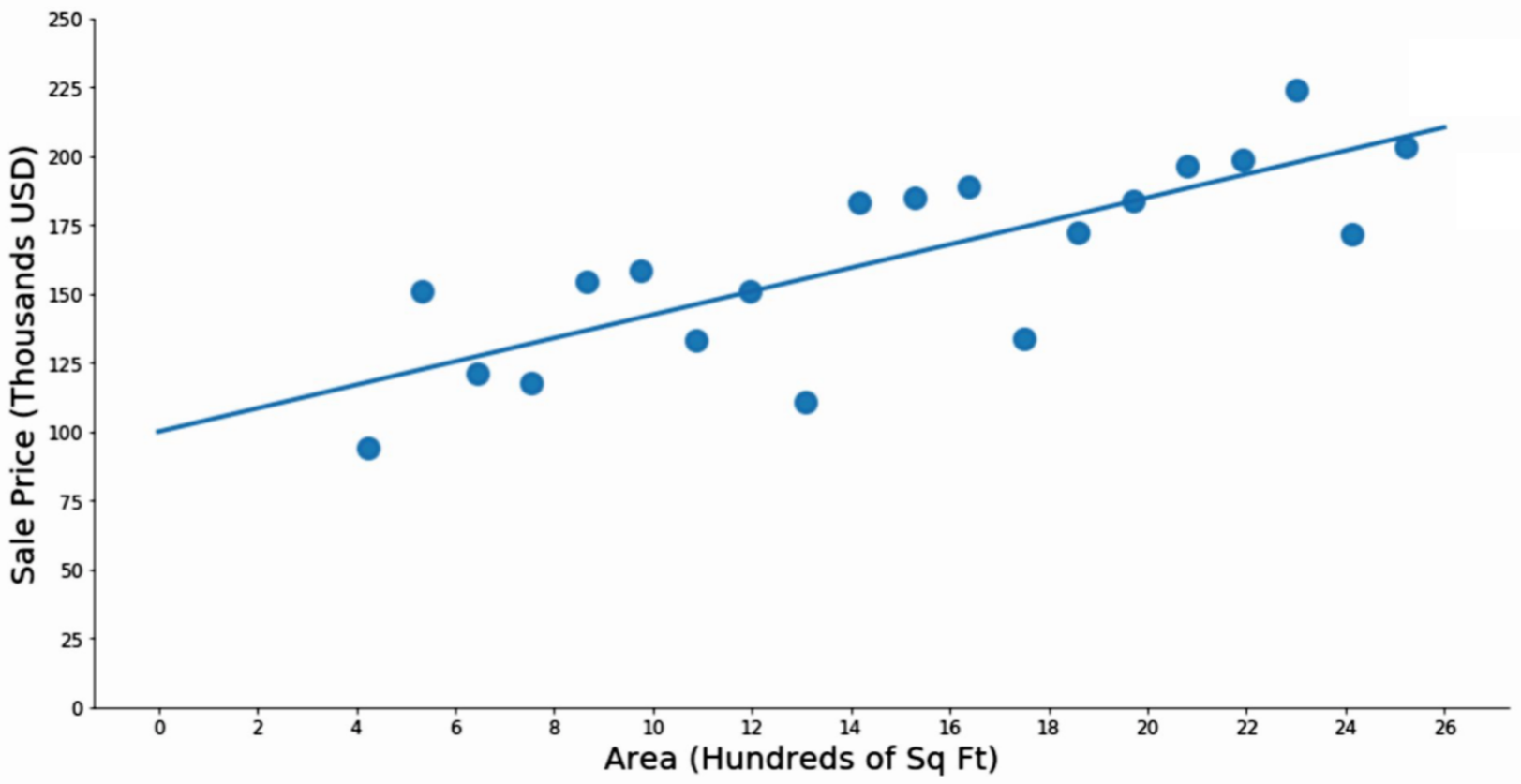


Simple Linear Regression – Interpreting parameters

The variance parameter σ^2 determines the extent to which each normal curve spreads about the true regression line.







Simple Linear Regression – Estimating model parameters

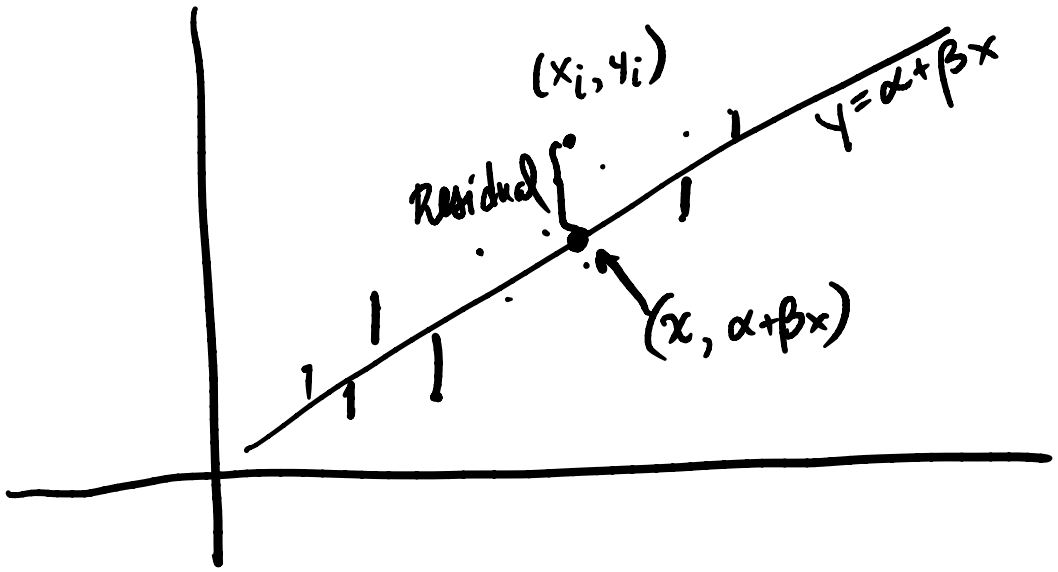
We've got data: $(x_1, y_1), (x_2, y_2), \ldots . (x_n, y_n)$

We want to figure out our regression line: $y = \alpha + \beta x$ vs. modeled line $\hat{y} = \hat{\alpha} + \hat{\beta} x$

α, β - true model parameters

$\hat{\alpha}, \hat{\beta}$ - estimated parameters

How can we minimize the residuals, $\epsilon_i = y_i - (\alpha + \beta x_i)$?

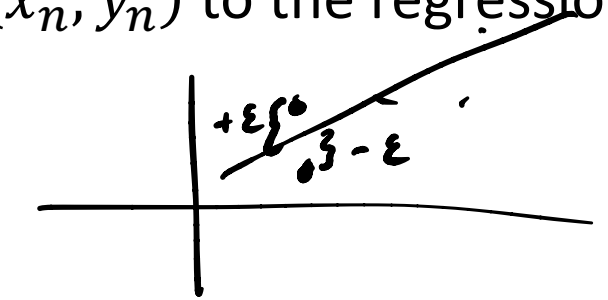


- goal is to fit a line between the data so that we minimize the distance between our data and our line

Simple Linear Regression – Estimating model parameters

The sum of squared-errors for the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to the regression line is given by

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



The point-estimates (single value estimates from the data) of the slope and intercept parameters are called the least-squares estimates, and are defined to be the values that minimize the SSE.

our goal is to minimize the SSE

Simple Linear Regression – Estimating model parameters

How do we actually find the parameter estimates?

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2$$

Finding

$$\frac{\partial SSE}{\partial \hat{\alpha}} = 0$$

$$\frac{\partial SSE}{\partial \hat{\beta}} = 0$$

set equal to 0 to find minimum

$f(x) = 3x^3$

$\frac{df}{dx} = 9x^2$

$$\frac{\partial SSE}{\partial \alpha} = \sum_{i=1}^{n} 2\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)(-1)$$

$$= -2\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\hat{\alpha} - \sum_{i=1}^{n}\hat{\beta}x_i = 0$$

$$\bar{y}\cdot n - \hat{\alpha}\cdot n - \hat{\beta}\bar{x}\cdot n = 0$$

$$\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0 \longrightarrow \boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

Recall: $\frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$

$\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$

Simple Linear Regression – Estimating model parameters

How do we actually find the parameter estimates?

set equal to 0 to find minimum

$$\frac{\partial SSE}{\partial \beta} = \sum_{i=1}^{n} 2(y_i - \hat{\alpha} - \hat{\beta} x_i)(-x_i) = 0$$

$$-2 \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} (x_i y_i - \hat{\alpha} x_i - \hat{\beta} x_i^2) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \hat{\alpha} \sum_{i=1}^{n} x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 = 0$$

$$n\overline{xy} - \hat{\alpha} \cdot n \cdot \bar{x} - \hat{\beta} \cdot n \cdot \overline{x^2} = 0$$

$$\frac{\sum_{i=1}^{n} x_i y_i}{n} = \overline{xy}$$

$$\frac{\sum_{i=1}^{n} x_i^2}{n} = \overline{x^2}$$

Simple Linear Regression – Estimating model parameters

How do we actually find the parameter estimates?

$$\overline{xy} - \hat{\alpha}\bar{x} - \hat{\beta}\overline{x^2} = 0$$

$$-\hat{\beta}\overline{x^2} = +\alpha\bar{x} - \overline{xy}$$

$$\hat{\beta} = \frac{\hat{\alpha}\bar{x}}{-\overline{x^2}} - \frac{\overline{xy}}{-\overline{x^2}}$$

$$\hat{\beta} = \frac{\overline{xy}}{\overline{x^2}} - \alpha\frac{\bar{x}}{\overline{x^2}}$$

Two equations, two unknowns : $\alpha = \bar{y} - \beta\bar{x}$

$$\hat{\beta} = \frac{\overline{xy}}{\overline{x^2}} - \frac{\bar{x}}{\overline{x^2}}(\bar{y} - \beta\bar{x})$$

$$\hat{\beta} = \frac{\overline{xy}}{\overline{x^2}} - \frac{\bar{x}\bar{y}}{\overline{x^2}} + \frac{\beta(\bar{x})^2}{\overline{x^2}}$$

$$\hat{\beta} - \hat{\beta}\frac{(\bar{x})^2}{\overline{x^2}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}}$$

$$\hat{\beta}\left(1 - \frac{(\bar{x})^2}{\overline{x^2}}\right) = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}}$$

$$\hat{\beta}\left(\frac{\overline{x^2} - (\bar{x})^2}{\overline{x^2}}\right) = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}}$$

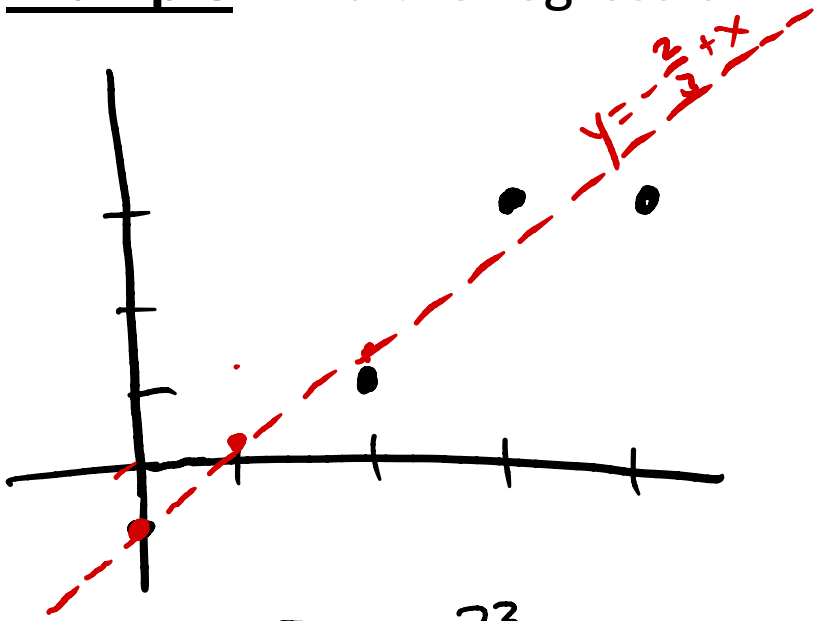
$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}} \cdot \frac{\overline{x^2}}{\overline{x^2} - (\bar{x})^2}$$

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Simple Linear Regression – Estimating model parameters

Example: Find the regression line for the following data: (2,1), (3,3), (4,3)



$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\bar{x}\,\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}$$

$$\bar{x} = \frac{2+3+4}{3} = \frac{9}{3} = 3$$

$$\bar{y} = \frac{1+3+3}{3} = \frac{7}{3}$$

$$\overline{xy} = \frac{2\cdot 1 + 3\cdot 3 + 4\cdot 3}{3} = \frac{23}{3}$$

$$\overline{x^2} = \frac{2^2+3^2+4^2}{3} = \frac{29}{3}$$

$$(\bar{x})^2 = 3^2 = 9$$

$$\hat{\beta} = \frac{3\cdot\frac{7}{3} - \frac{23}{3}}{9 - \frac{29}{3}} = \frac{\frac{21}{3} - \frac{23}{3}}{\frac{27}{3} - \frac{29}{3}} = \frac{-\frac{2}{3}}{-\frac{2}{3}} = 1$$

$$\hat{\beta} = 1$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{7}{3} - 1\cdot 3 = \frac{7}{3} - \frac{9}{3} = -\frac{2}{3}$$

⇒ Simple Regression Line

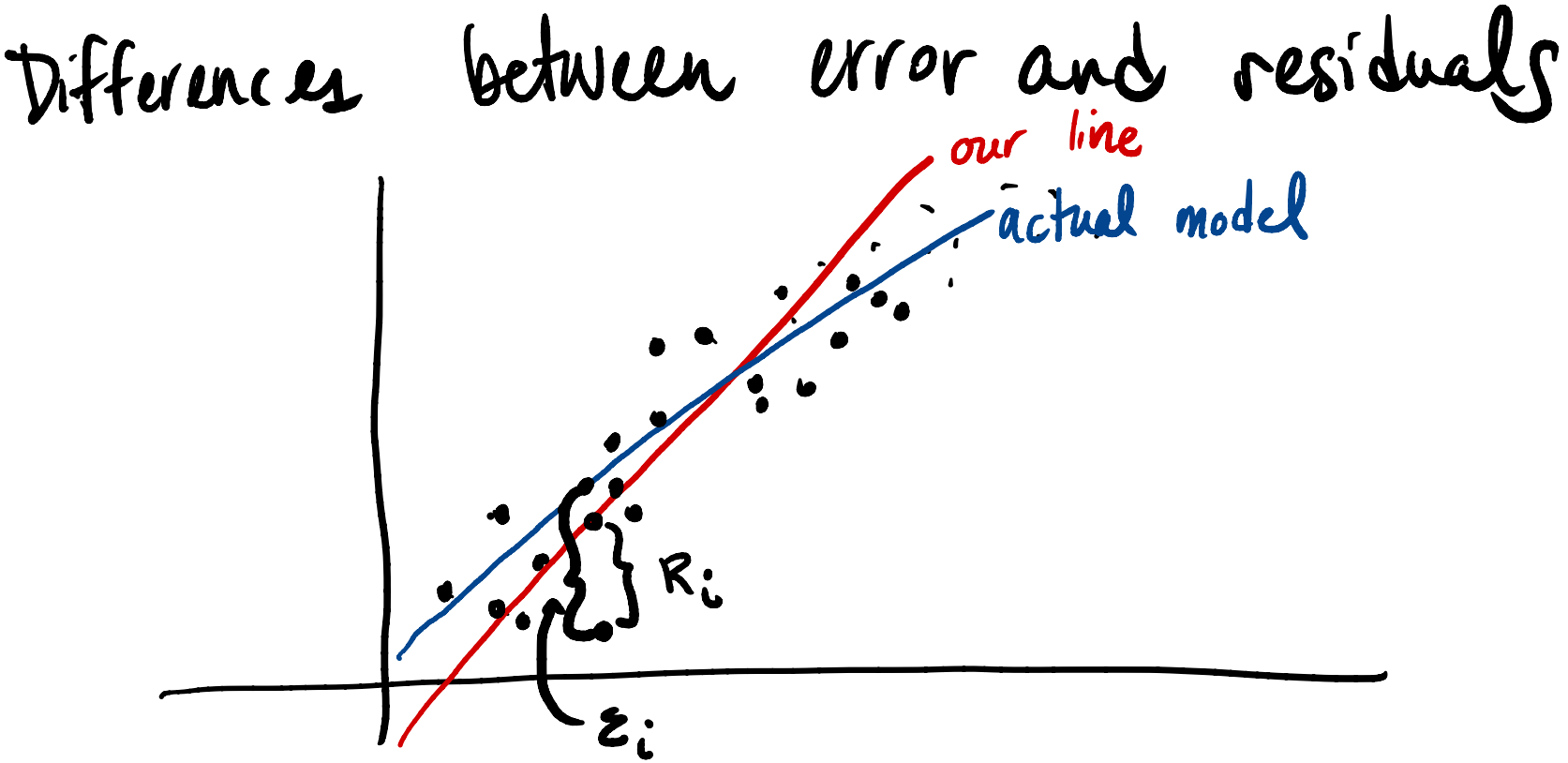
$$\hat{y} = -\frac{2}{3} + x$$

Residuals

The residuals are the difference between the observed and the predicted responses:

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

The residuals r_i are estimates of the unknown true error ϵ_i



Next Time:

- Inference in Simple Linear Regression