Introduction to Regression

Administrivia

- **Homework 6** posted later tonight. Due Friday after Break.

Statistical Modeling

Thus far we've talked about

- **Descriptive Statistics**: This is the way my sample **is**
- **Inferential Statistics**: This is what I can likely **conclude** from my sample

Today we move towards what we might call **Predictive Statistics**

Linear Regression for Prediction

Examples:

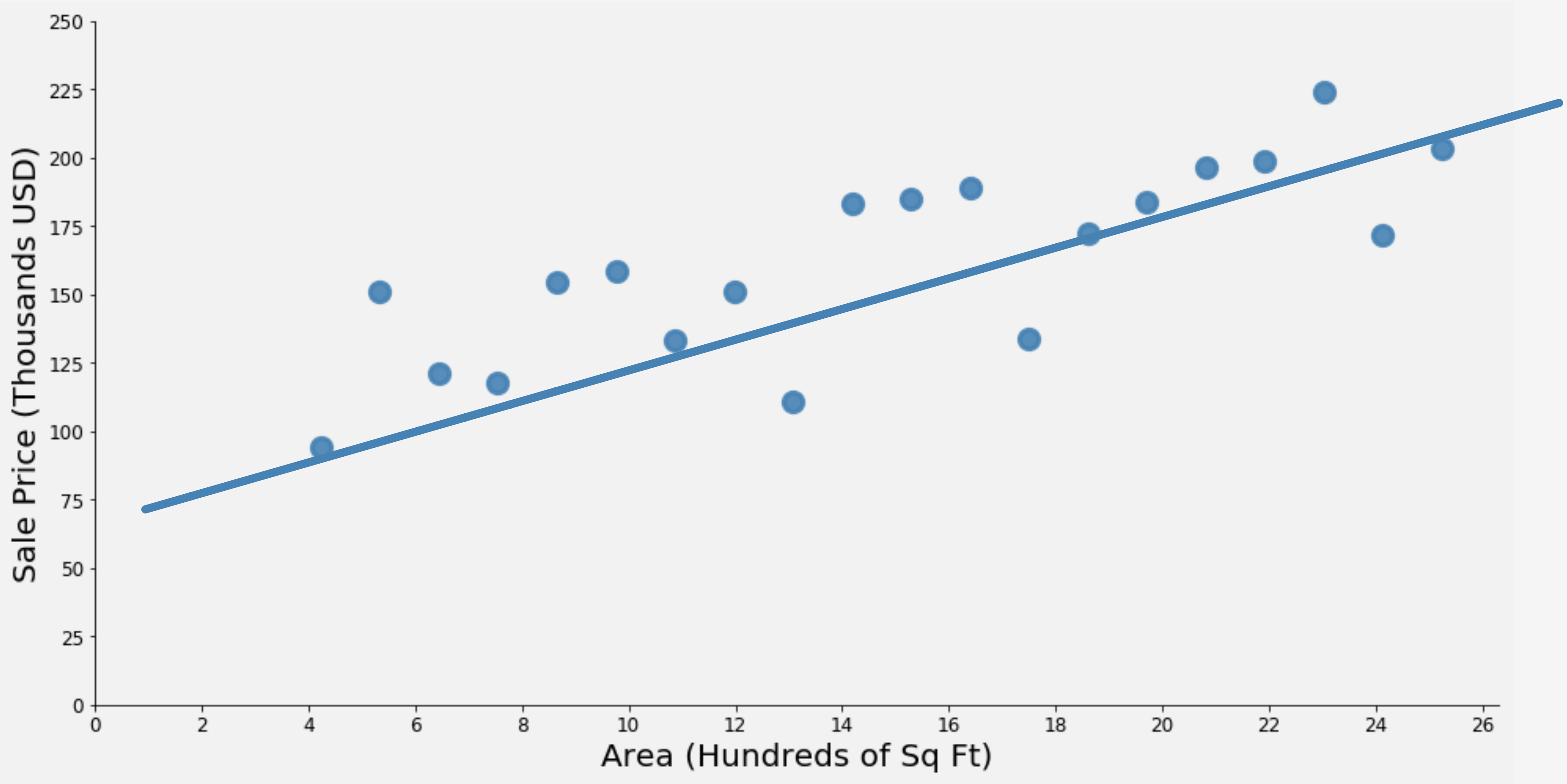
- Given a person's age and gender, predict their height
- **Given the area of a house, predict its sale price**
- Given unemployment, inflation, number of wars, and economic growth, predict the president's approval rating.
- Given a person's browser history, predict how long they'll stay on a product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

Linear Regression for Inference

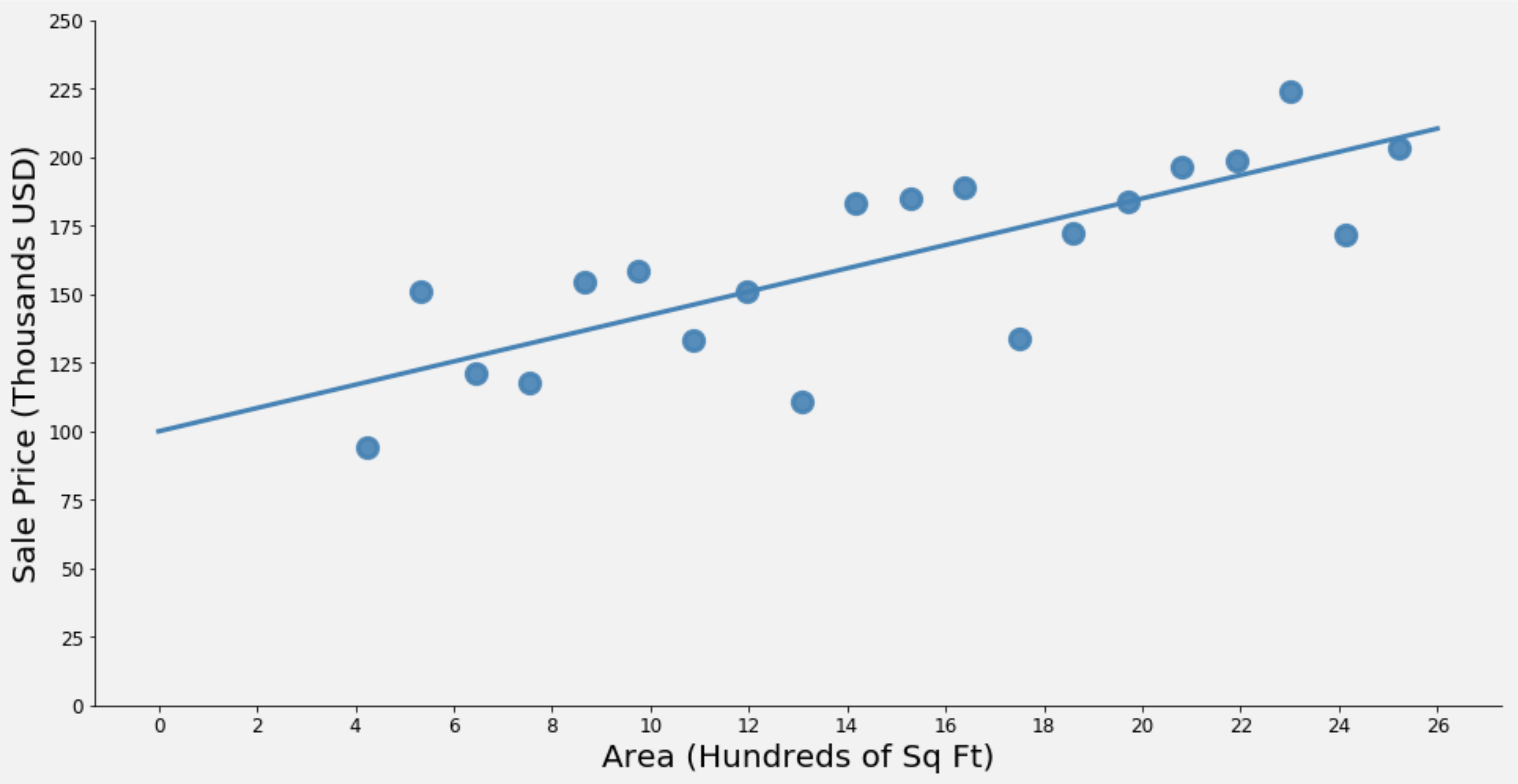
Examples:

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

Area as Predictor for House Price



Area as Predictor for House Price



Exploration

Open up your computer and load the Lecture 20 in-class notebook

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

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

Simple Linear Regression Model

Defs and **Assumptions** of SLR model: (n DATA points)

1. $y_i = \alpha + \beta x_i + \epsilon_i$

2. EACH OF ϵ_i's ARE INDEPENDENT

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

Simple Linear Regression Model

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

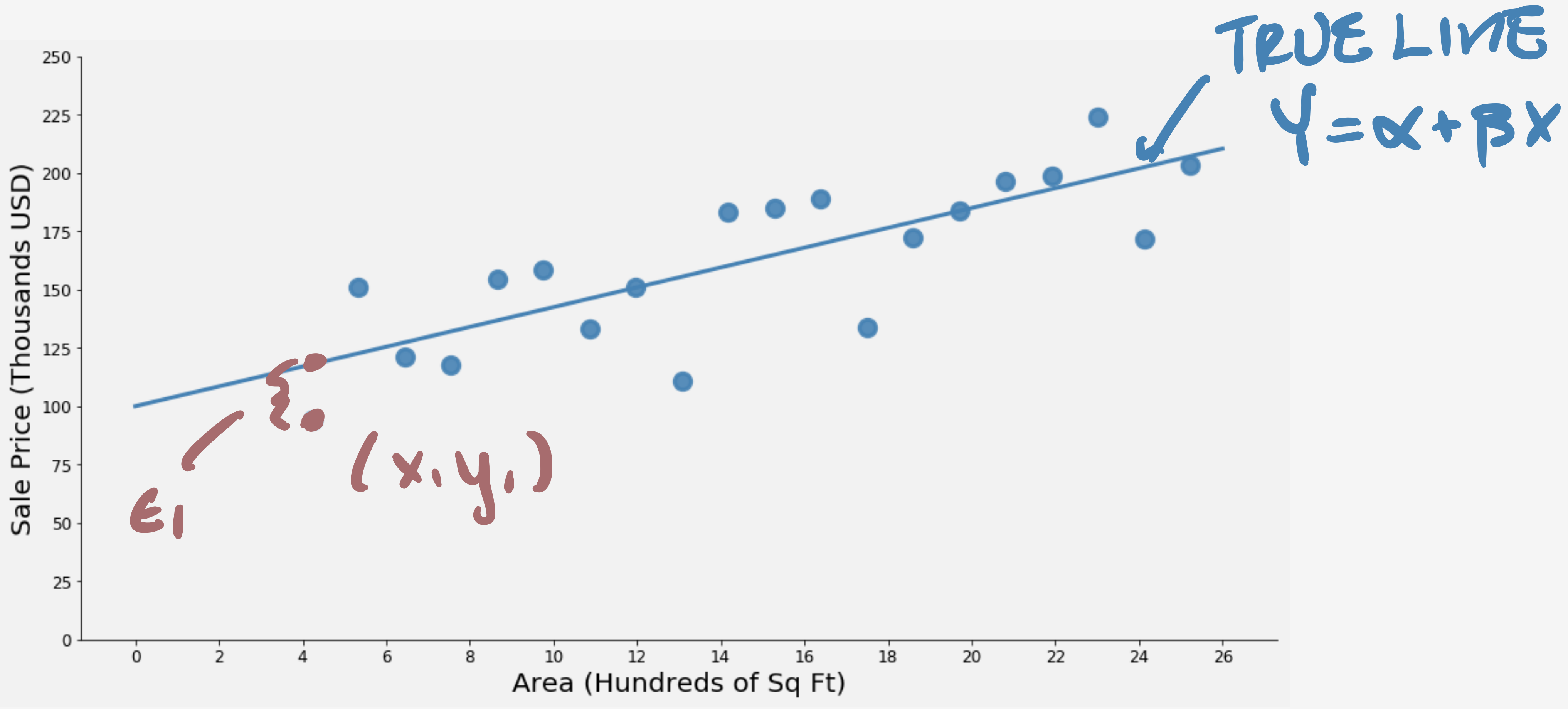
SLR model **vocabulary**:

- X: the independent variable, the predictor, the explanatory variable, the **feature**
  - * FIXED, NOT RANDOM.
- Y: the dependent variable or the **response** variable
  - * RANDOM VARIABLE
- ϵ : the random deviation or **random error**
  - * RANDOM

Question: What exactly is ϵ doing?

Simple Linear Regression Model

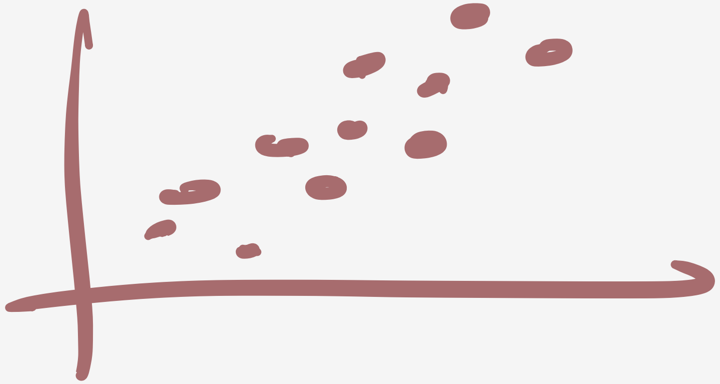
The points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations will be scattered about the true regression line



Simple Linear Regression Theory

Question: How do we know that the simple linear regression is appropriate?

* EYEBALL MEASURE
* EXPERIENCE



SPOILERS

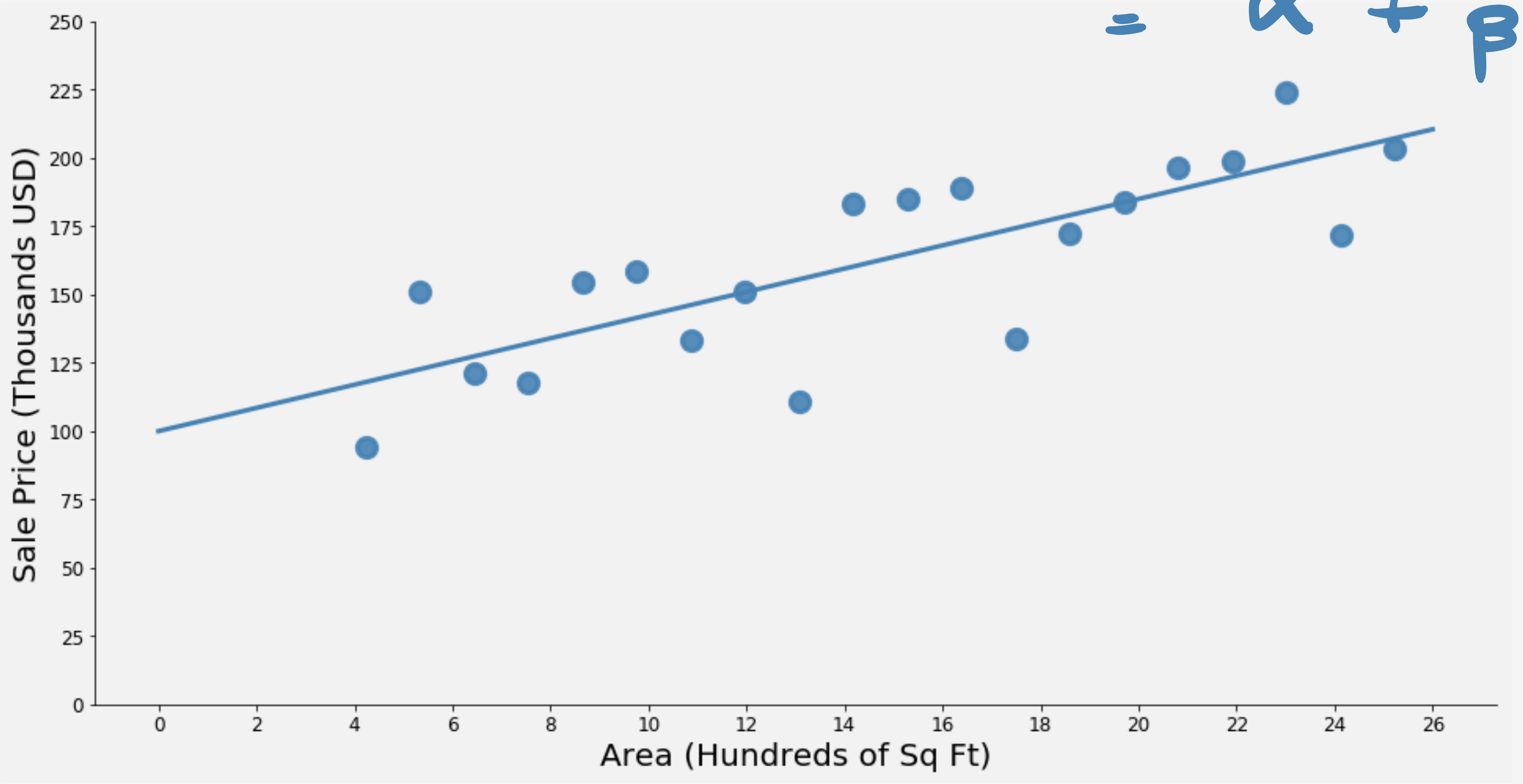
* R^2-VALUE

Interpreting SLR Parameters

- Y is a random variable. What is it's expectation?

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

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



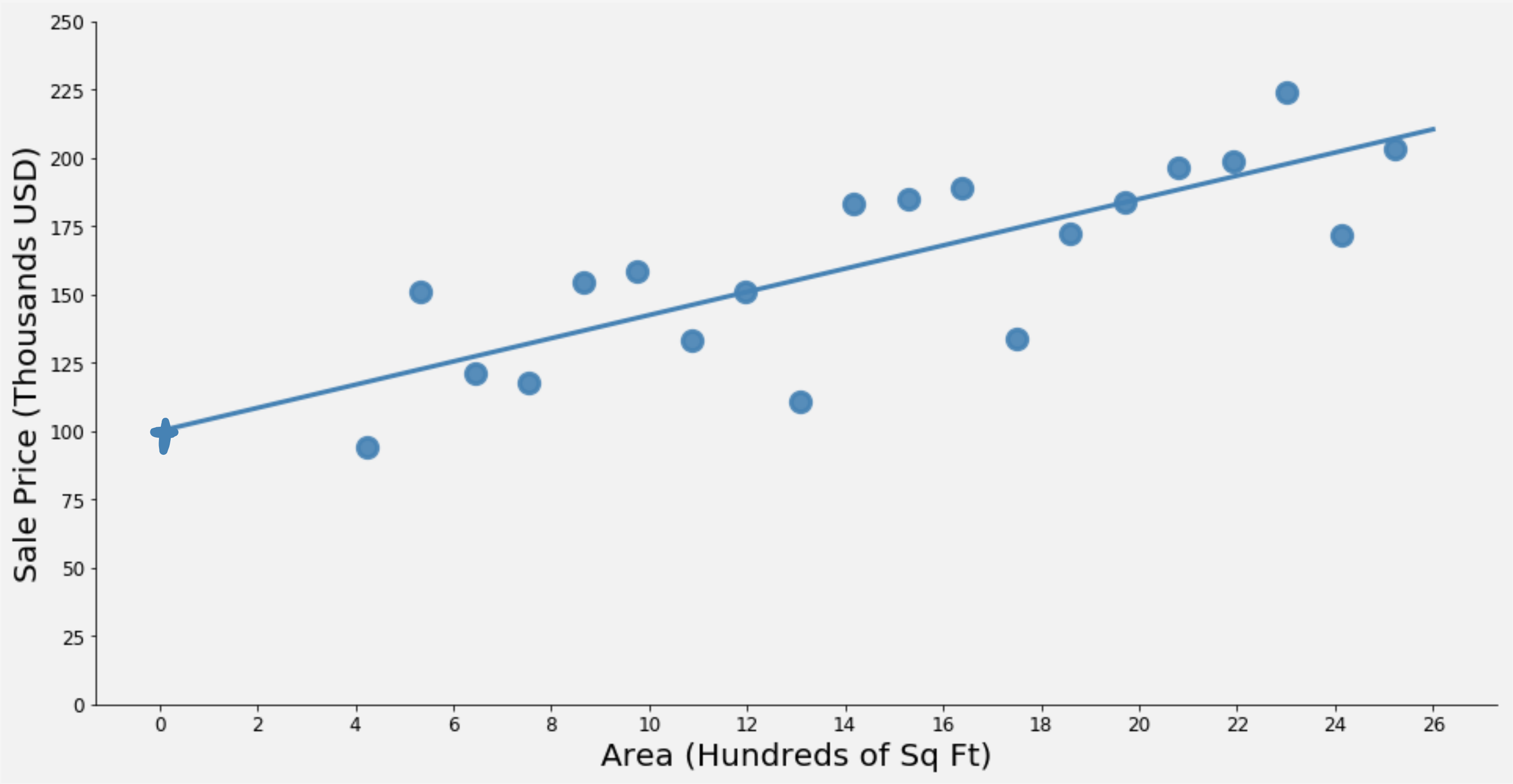
$\epsilon \sim N(0, \sigma^2)$

Interpreting SLR Parameters

$Y = \alpha + \beta X$

- α is the intercept of the true regression line (the so-called baseline average)

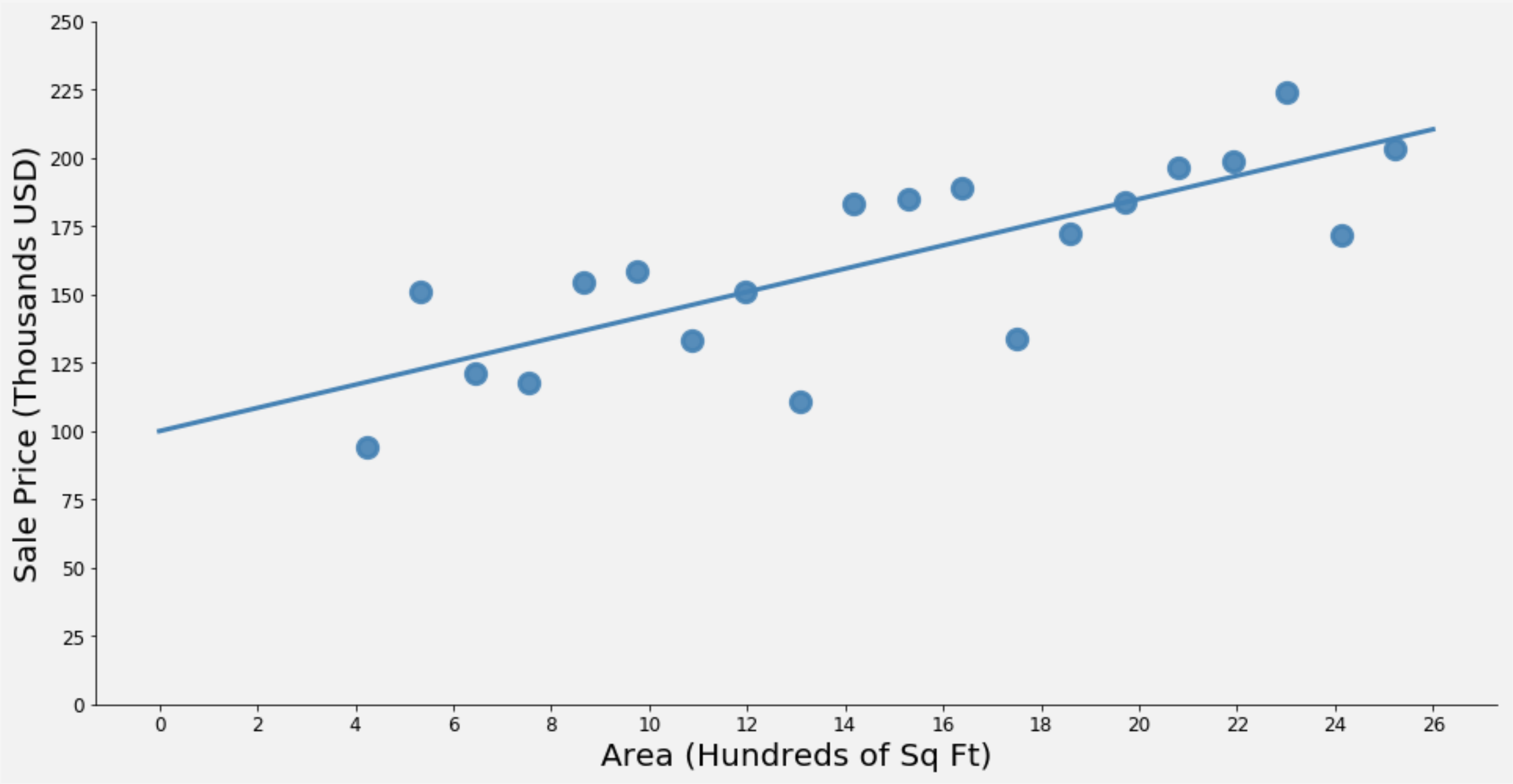
$\alpha = 100$



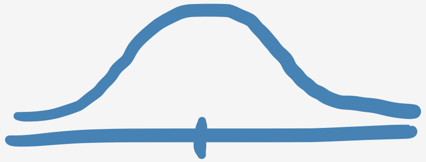
Interpreting SLR Parameters

- β is the slope of the true regression line

$$Y_1 = \alpha + \beta(X+1)$$
$$-\;Y_2 = \alpha + \beta X$$
$$\overline{}$$
$$\beta$$

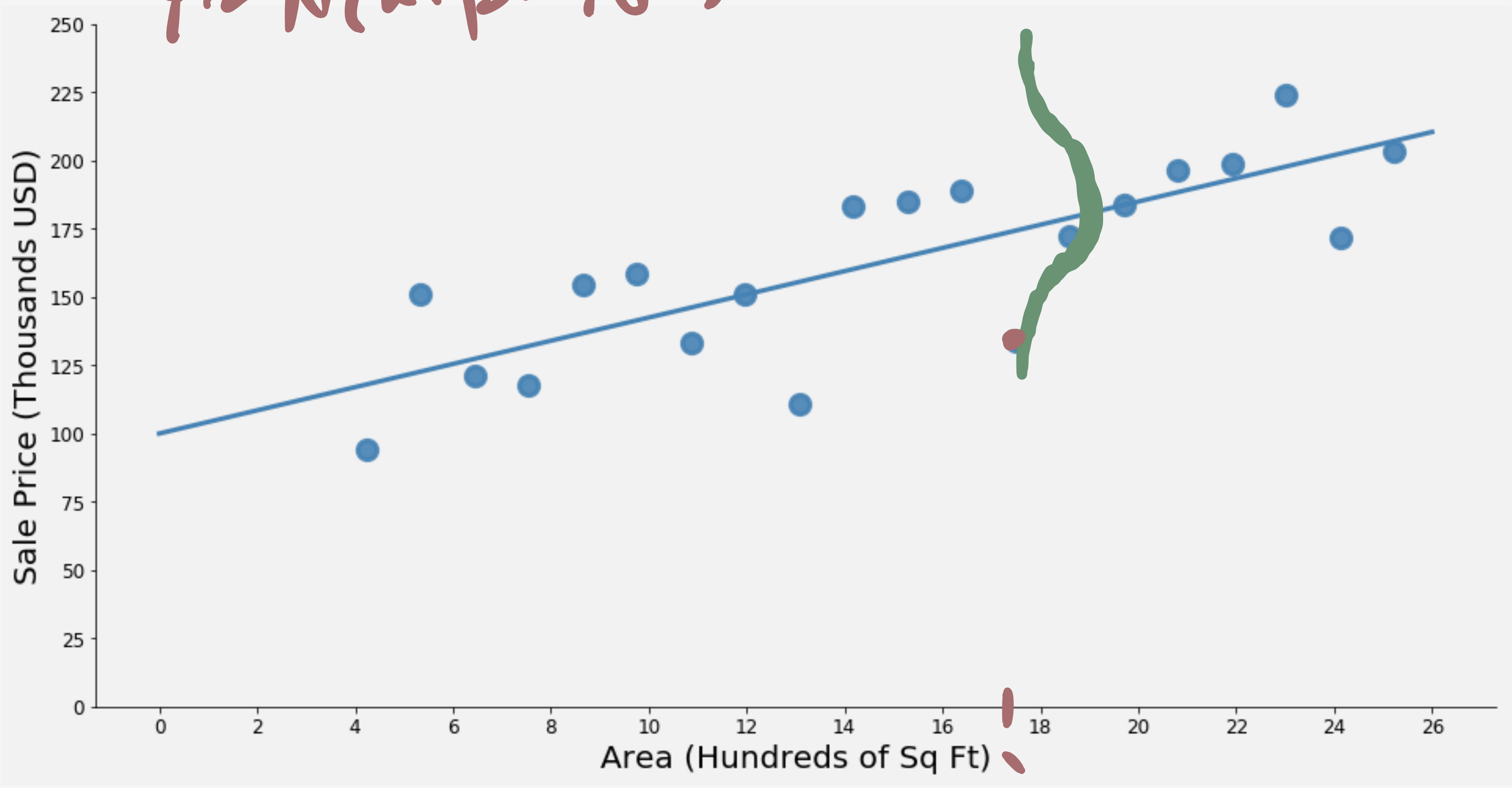


Interpreting the Error Term



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

$Y \sim N(\alpha+\beta X, \sigma^2)$ $\quad Y = \alpha + \beta X + \epsilon$



Directional Considerations

- So far we've come up with a framework where we can choose the model parameters and then generate random data. This is called a **generative model**.

- But really, we want to run this process in reverse. We have data, and we want to **find/learn/estimate** the parameters that explain the data.

General Model + Parameters $\xrightarrow{\text{SAMPLE}}$ Data

General Model + Parameters $\xleftarrow{\text{INFERENCE}}$ Data

How Can We Estimate Params from Data?

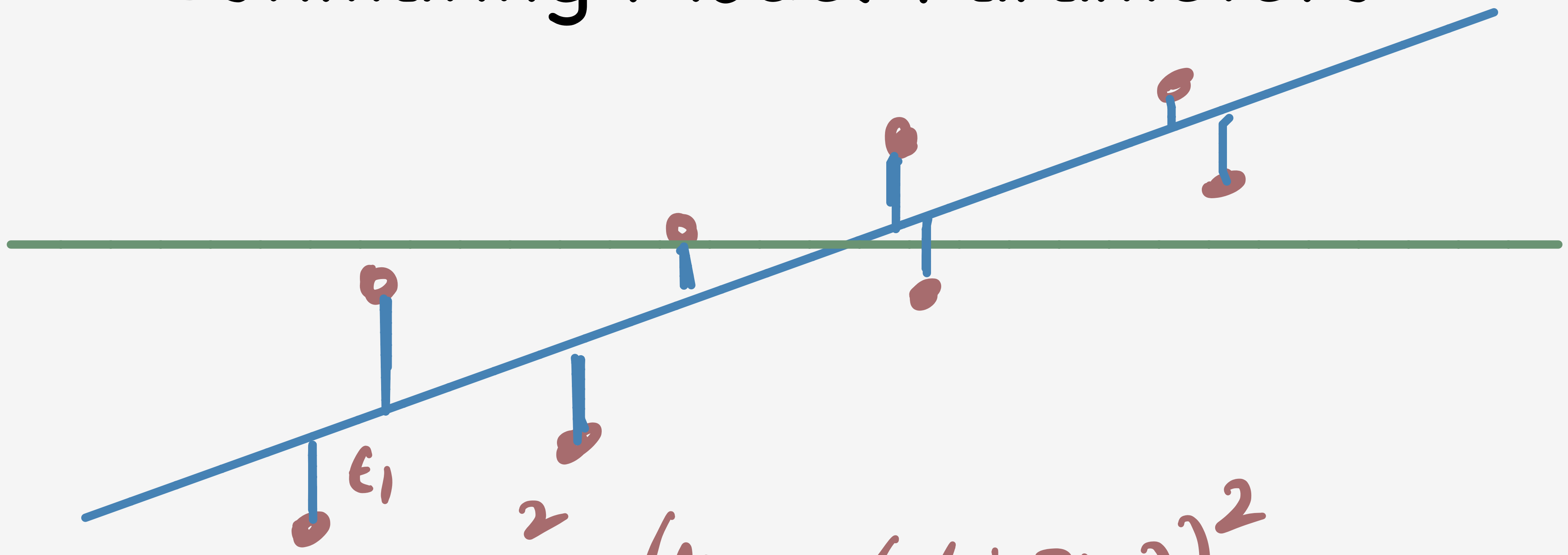
- **Plan of Attack**: The variance of our model σ^2 will be smallest if the differences between between the estimate of the true regression line and each point is the smallest. This is our **goal**: **minimize** σ^2

- We'll use our sample data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, to estimate the parameters of the regression line

- What are we assuming about each of the observations?

(x_1, y_1) (x_2, y_2) ← OBTAINED INDEPENDENTLY

Estimating Model Parameters



$$\epsilon_1^2 = \left(y_1 - (\alpha + \beta x_1)\right)^2$$

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

Estimating Model Parameters

- The sum of the 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

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

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

$$\frac{\partial SSE}{\partial \alpha} = 0 \quad , \quad \frac{\partial SSE}{\partial \beta} = 0 \quad \Big\} \text{ SOLVE simultan...}$$

Estimating Model Parameters

- The **fitted regression line** or the least-squares line is then the line given by

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

Question: How do we actually find the parameter estimates?

Estimating Model Parameters

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

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

$$\Rightarrow \frac{1}{n}\sum y_i - \alpha - \beta x_i = 0$$

$$\Rightarrow \bar{y} - \alpha - \beta \bar{x} = 0$$

$$\frac{\partial SSE}{\partial \beta} \Rightarrow \sum_{i=1}^{n} -2 x_i \left(y - (\alpha + \beta x_i)\right) = 0$$

How Can We Do This in Practice?

Get your laptops back out and let's figure it out!

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

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

Residuals

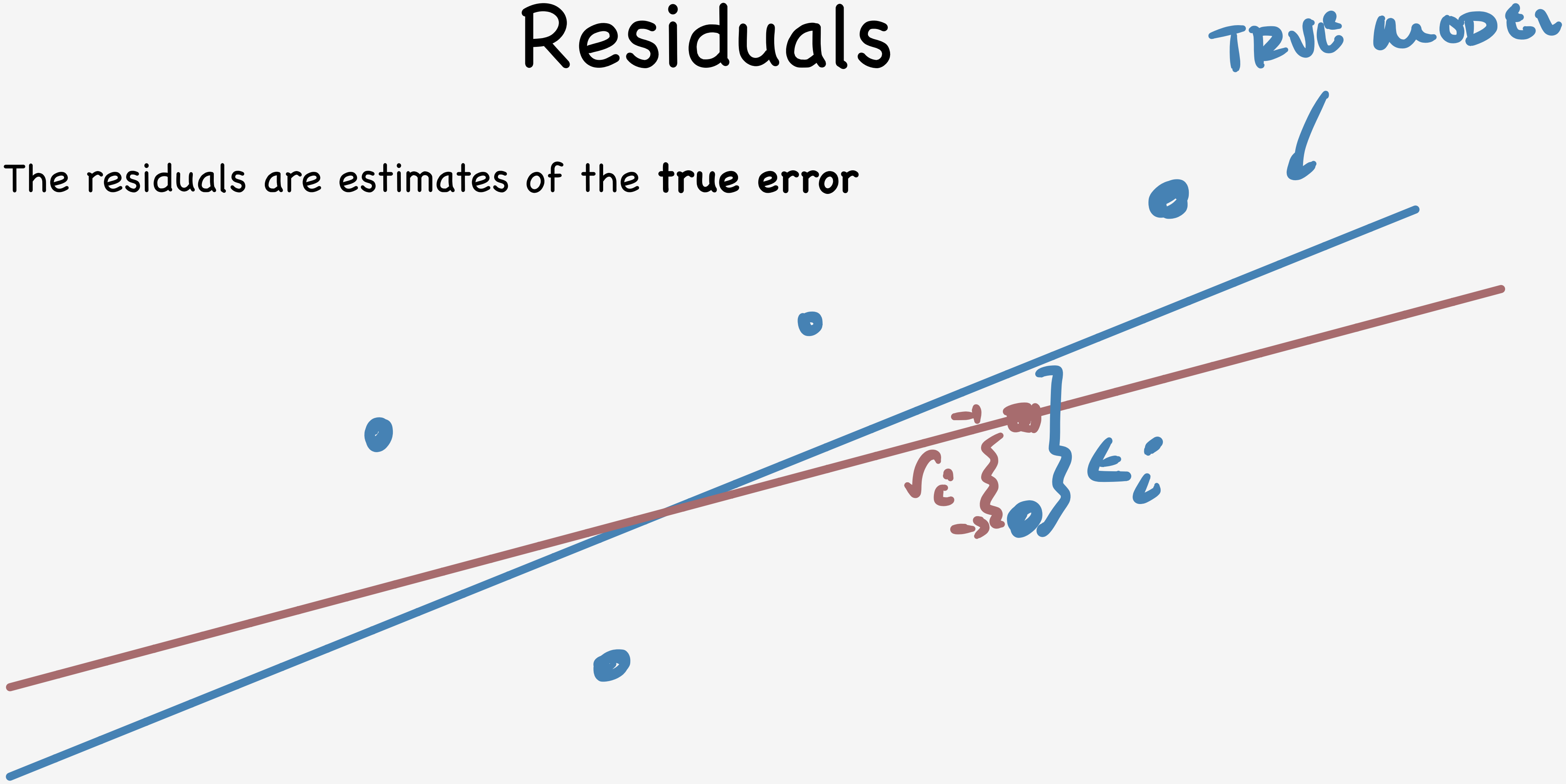
- The fitted or predicted values $\hat{y} = \hat{\alpha} + \hat{\beta} x$ are obtained by plugging in the independent data variables into the fitted model

- The **residuals** are the differences between the observed and predicted responses:

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

Residuals

Claim: The residuals are estimates of the **true error**



Maximum Likelihood Estimation

- An alternate method for estimating model parameters is to create a likelihood function involving the model parameters and the data, and choose the value of the parameter that maximizes it

- We've done this before, just haven't called it Maximum Likelihood Estimation

Example: Suppose you have a biased coin, you flip it 6 times and get 5 Heads and 1 Tails. Estimate the parameter p for the coin.

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Maximum Likelihood Estimation

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 20 In-Class Notebook

Let's:

- Do some stuff!