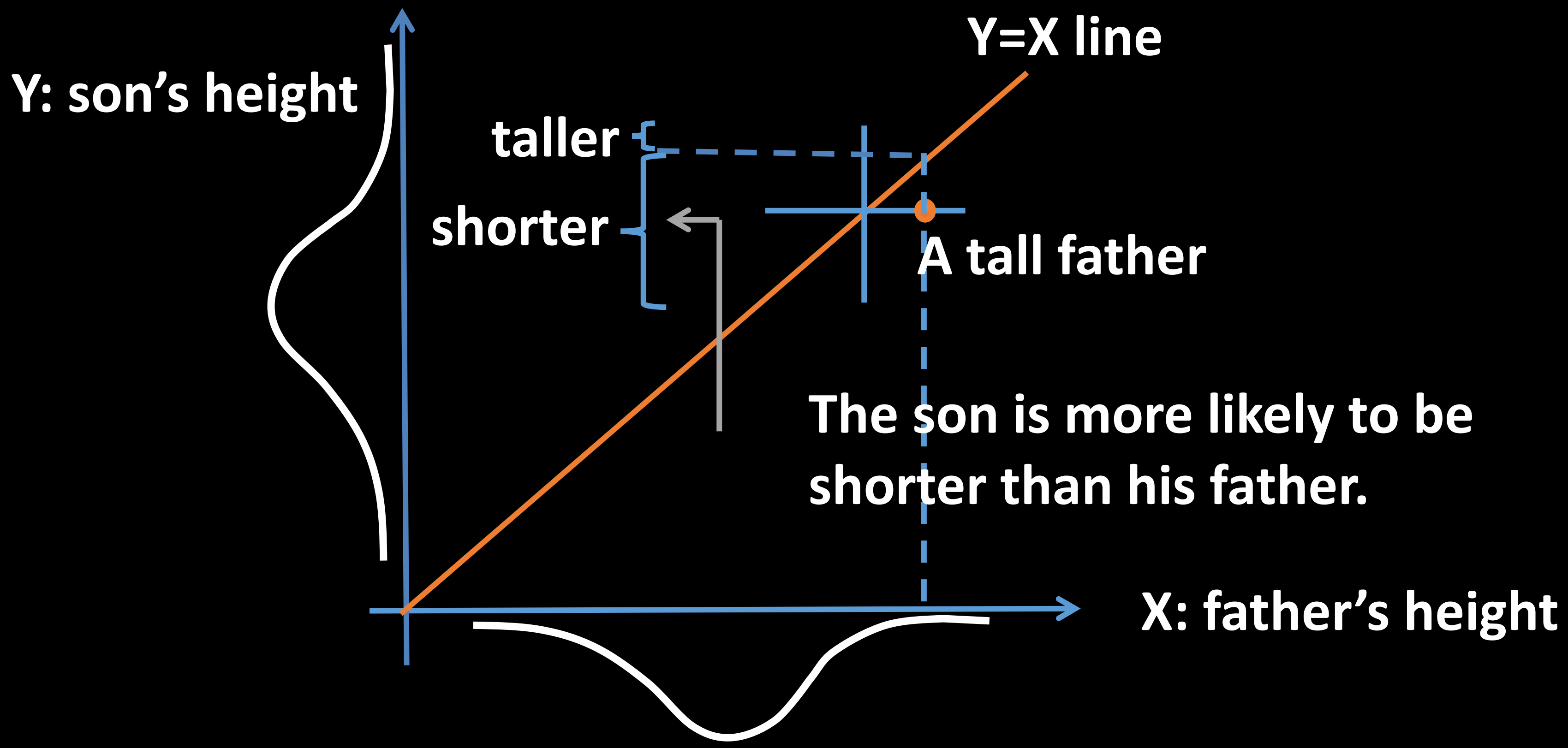
Linear regression

Regression

- The word “regression” comes from the phrase

  “regression towards the mean”.

  - A phenomenon people observed when studying relation of two related quantities.
- Example:
  - father's height (X) versus
  - son's height (Y).



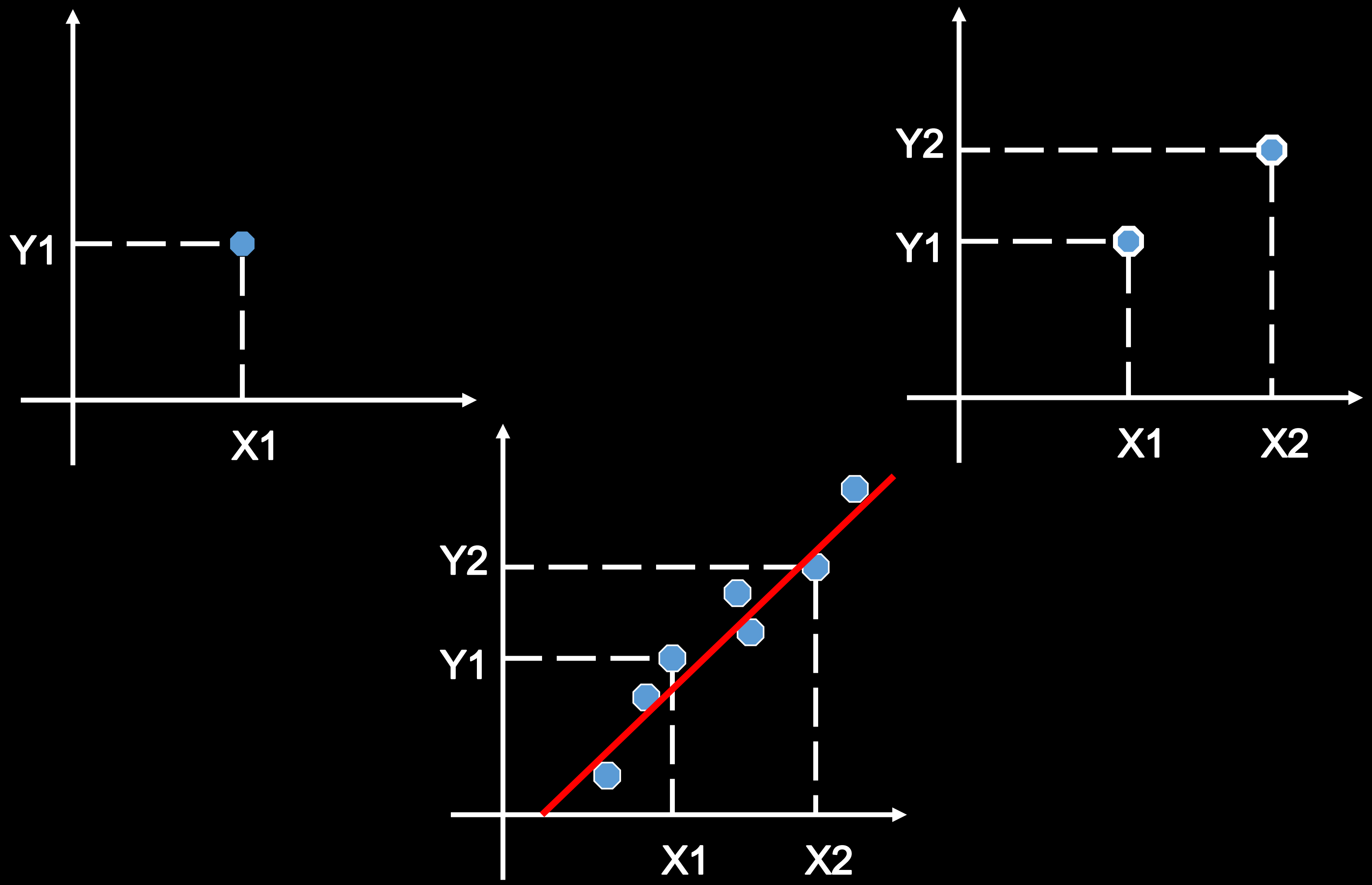
- Suppose genetically, the father and son has the same expected height (expected to be on the Y=X line).
- Both father and son depart from this theoretical height due to random factors such as diet, exercise, etc.

Simple linear regression for quantitative variables

- X variable
  - Explanatory variable
  - Independent variable
- Y variable
  - Response
  - Dependent variable

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

Fit a line to a scatterplot



Evaluate the "fit"

- Data: $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$
- A candidate regression line
$$Y = a + bX$$
- How do we evaluate the fit?
- For any given value of X, the regression line suggests a "predicted value" for Y.
$$\widehat{Y_i} = a + bX_i$$
- Prediction error
$$e_i = Y_i - \widehat{Y_i}$$

Least-square regression

- For X and Y, among all possible linear regression models between X and Y, the “best” regression line is the *minimizer* of

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2$$

- It is the “closest” fit to the observed points.

Regression estimates

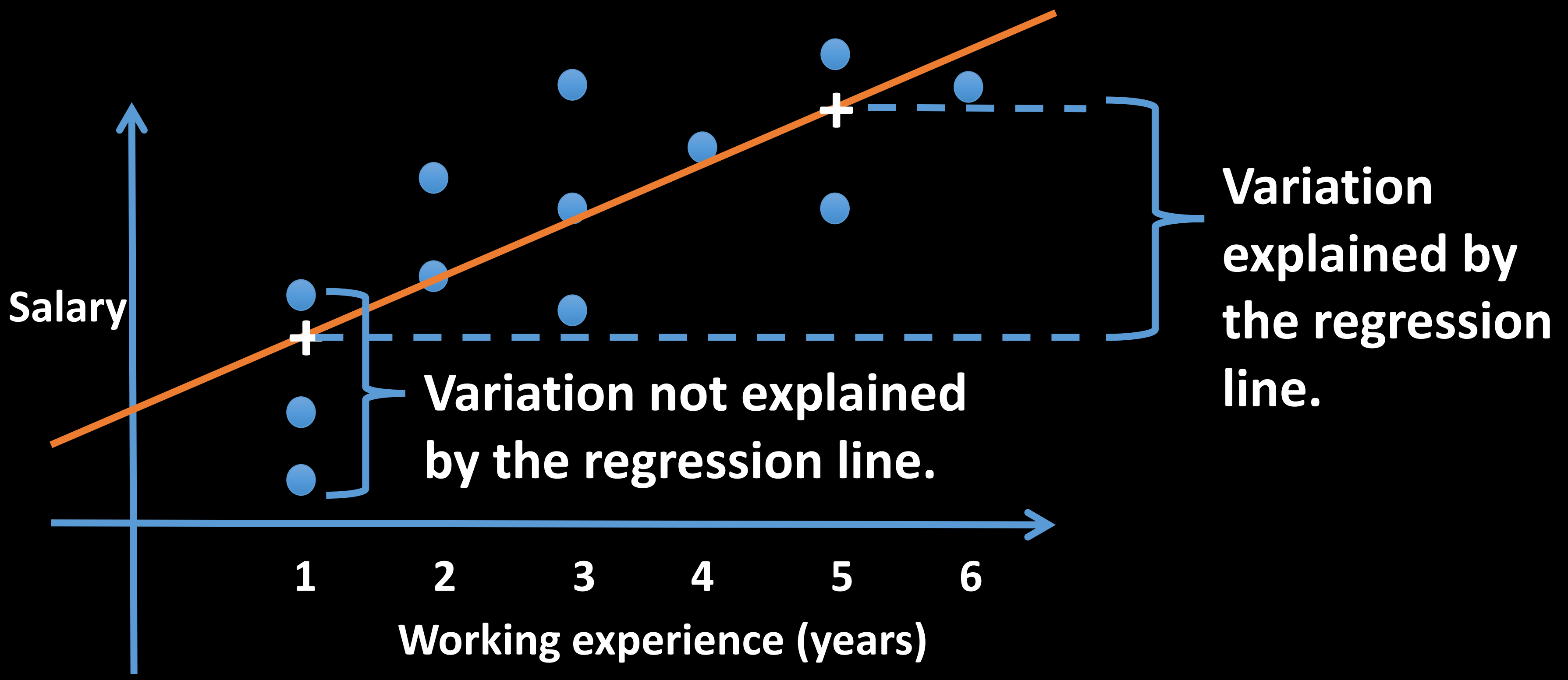
- **The estimated least square regression line**

$$\hat{Y} = b_0 + b_1 X$$

- b_0 **is the intercept, predicted Y value at X=0.**
- b_1 **is the slope, which estimates the increment of Y when X increases one unit.**
- **For example, price of Apartment (Y)**

$$\hat{Y} = 100{,}000 + 200\ (sqrt\ ft)$$

Analysis of variance



Analysis of variance

- Sum of squares
  - $SSE = \sum_{i=1}^{n}\left(Y_i - \widehat{Y_i}\right)^2$
  - $SSR = \sum_{i=1}^{n}\left(\widehat{Y_i} - \overline{Y}\right)^2$
  - $SSTO = SSE + SSR = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$
- $R^2 = SSR/SSTO$ measures the fraction of variation in Y can be explained by X.
- *Still not necessarily causation.*

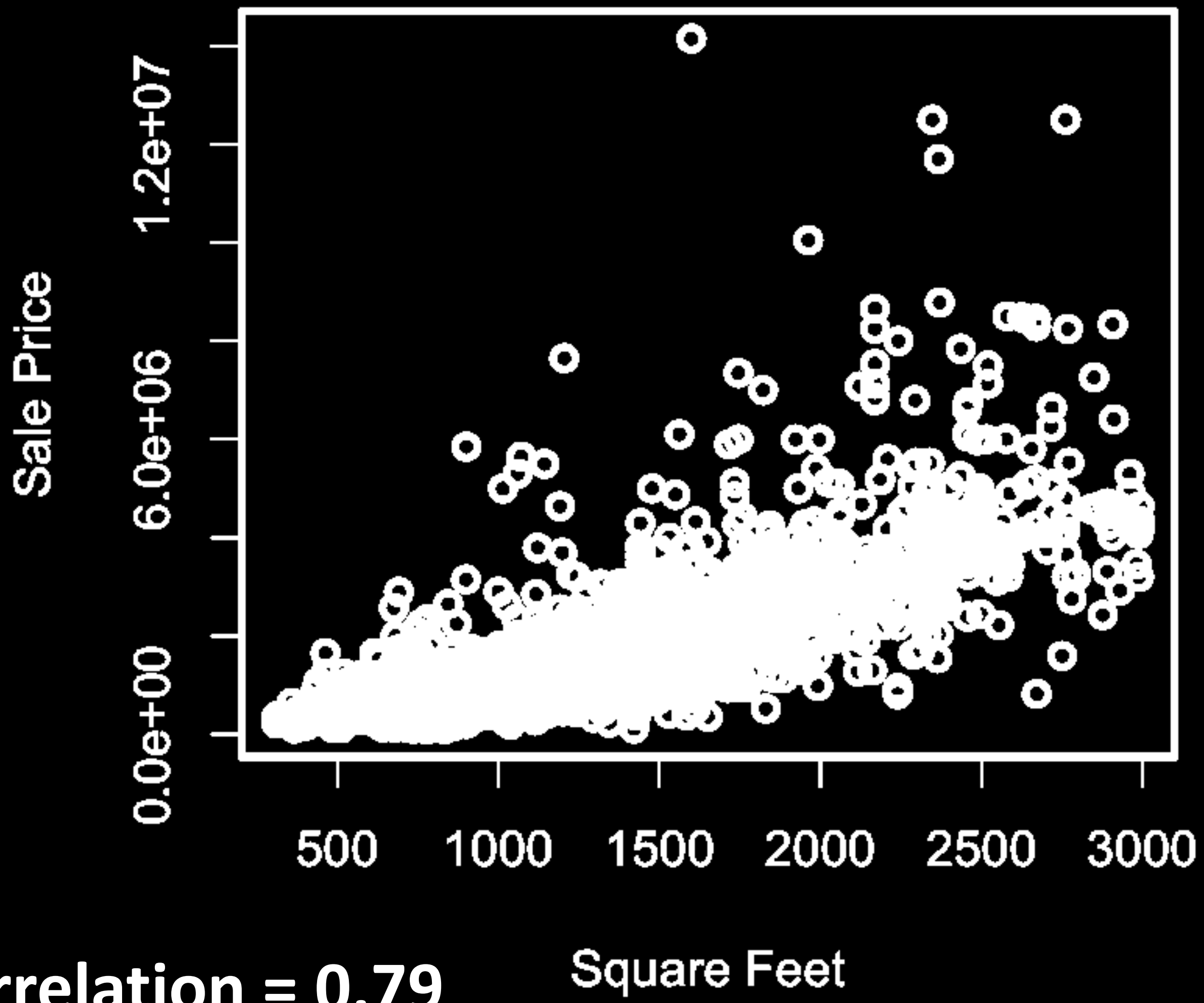
Normal Error Regression Model

- A probability model for linear regression
- $Y = \beta_0 + \beta_1 X + \varepsilon$
- Here ε is a random error that follows normal distribution with a constant variance.
- Under this model, least square regression estimates also maximize the “likelihood” function.
  - Likelihood is the probability for the observed data under a specified model—a function for models given observed data.

Manhattan Condo Prices

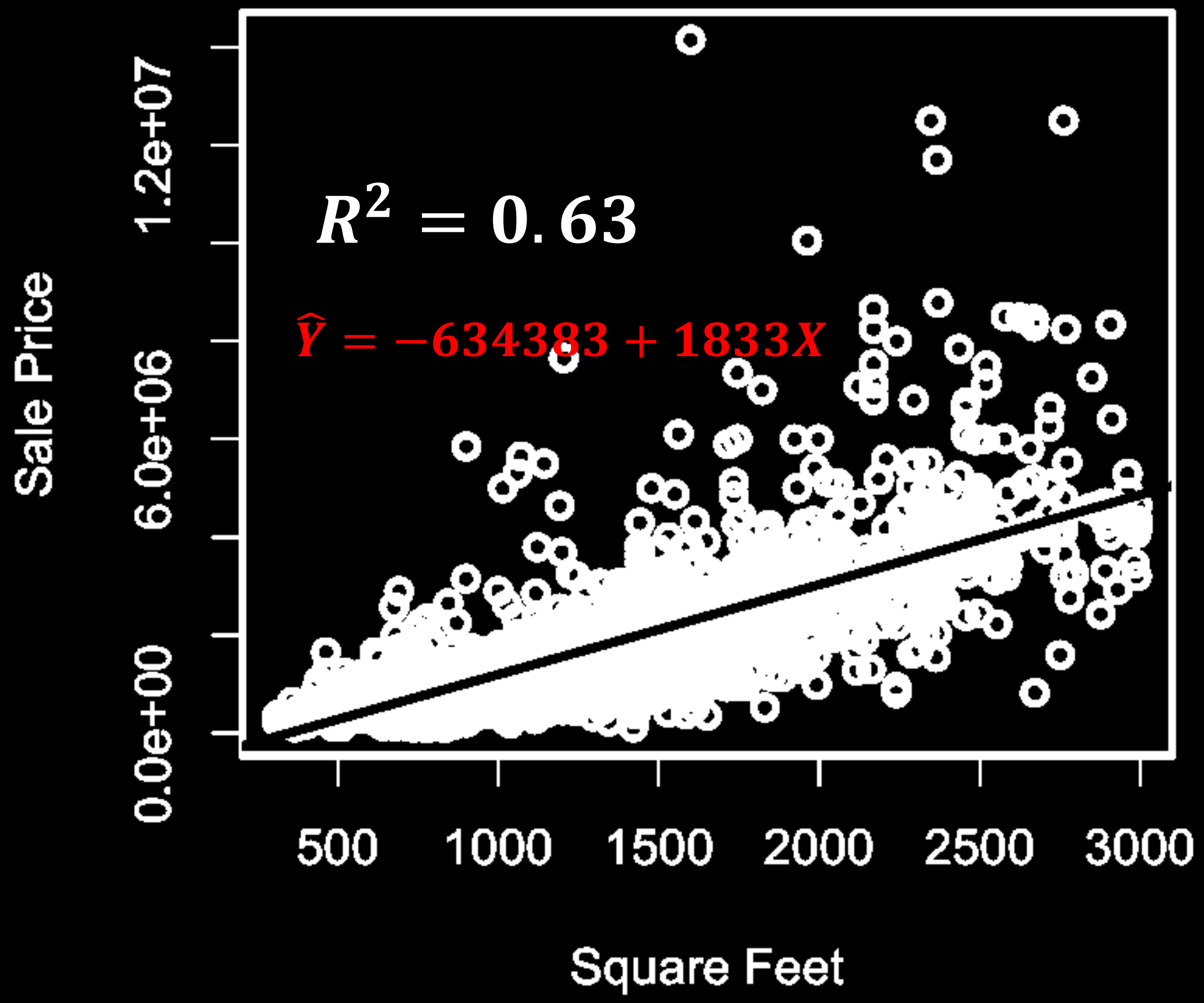
- From NYC Open Data
- Year 2009
- Condo building with elevator
- 4656 apartments

Manhattan Condo Prices

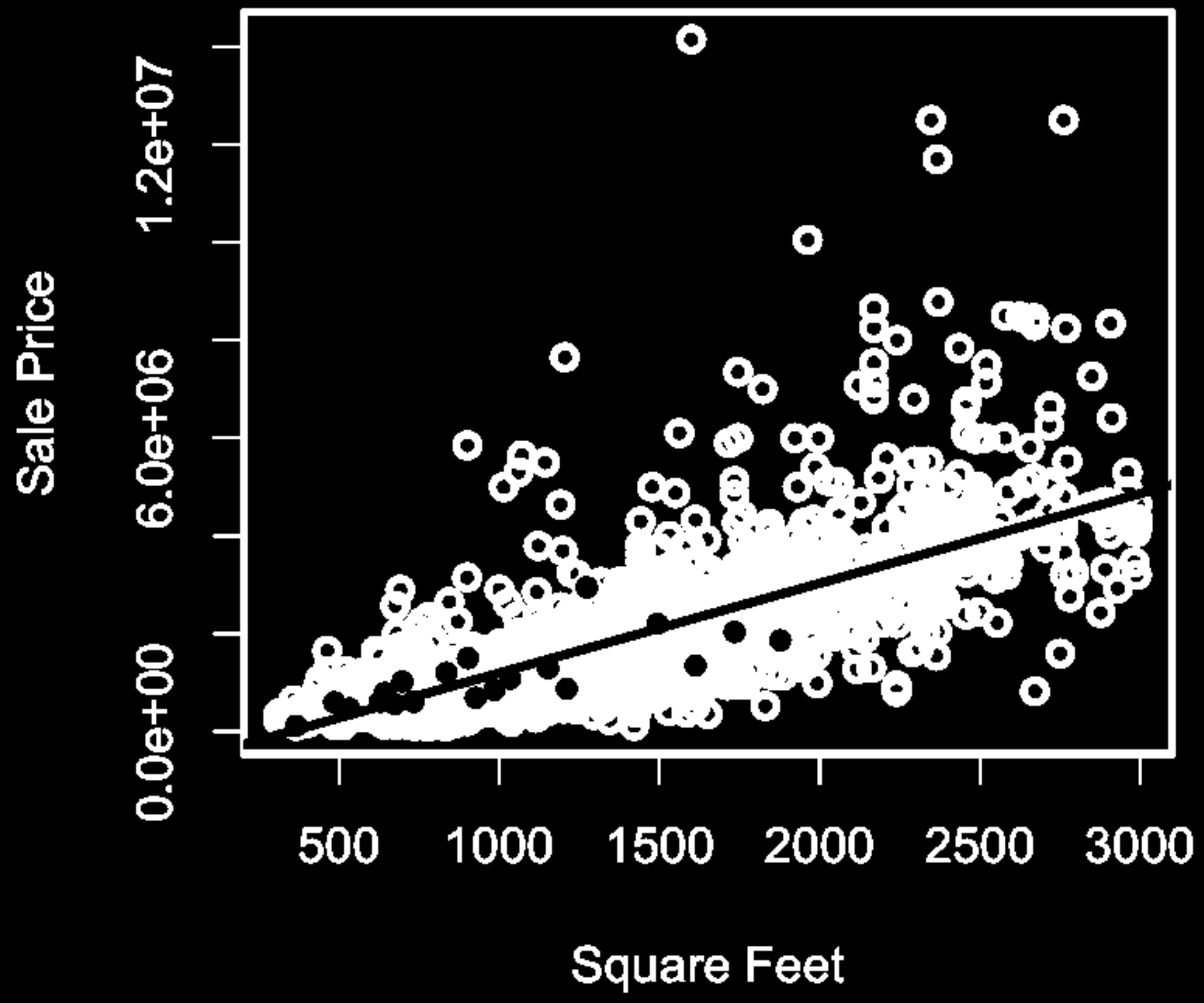


Correlation = 0.79

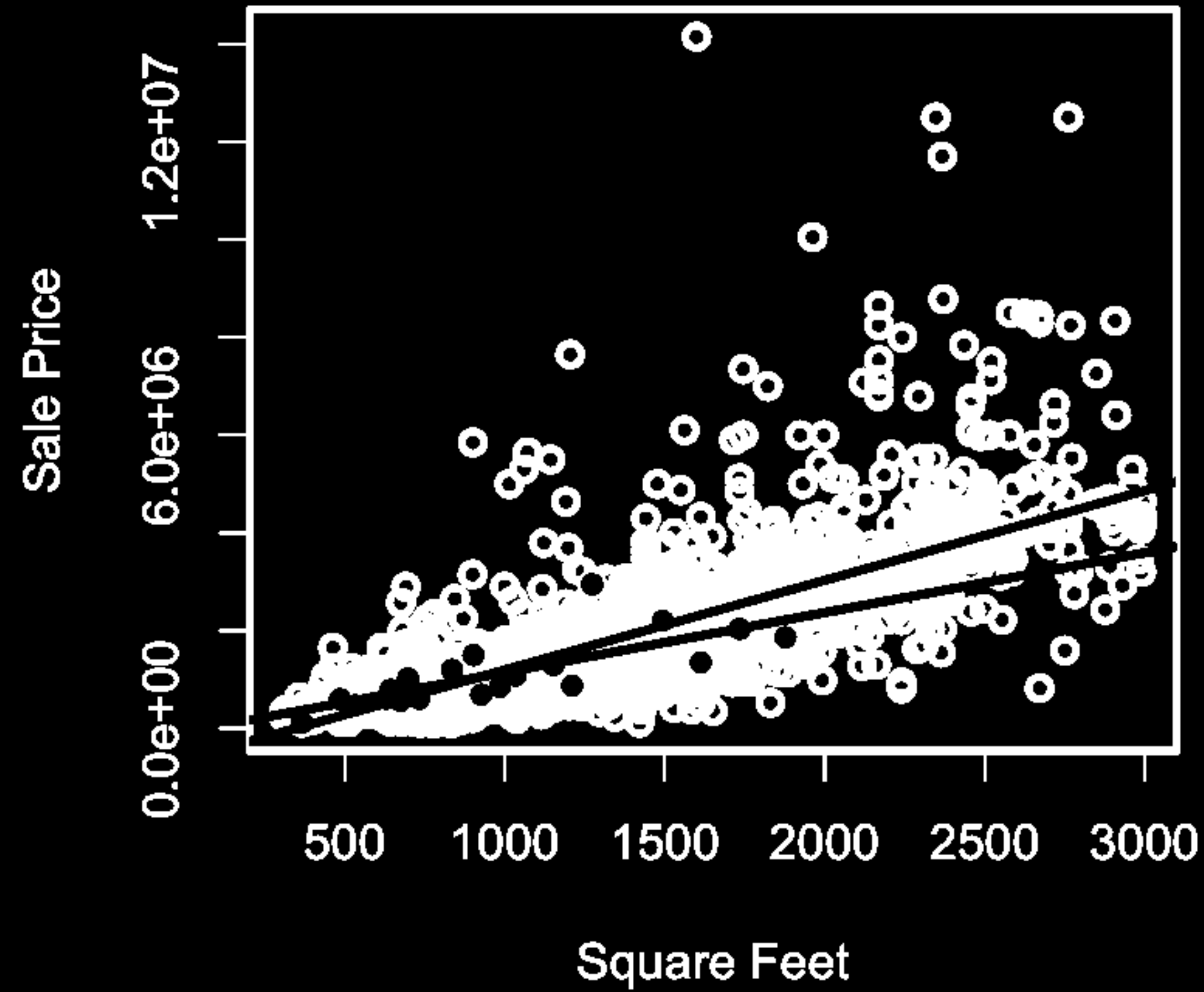
Manhattan Condo Prices



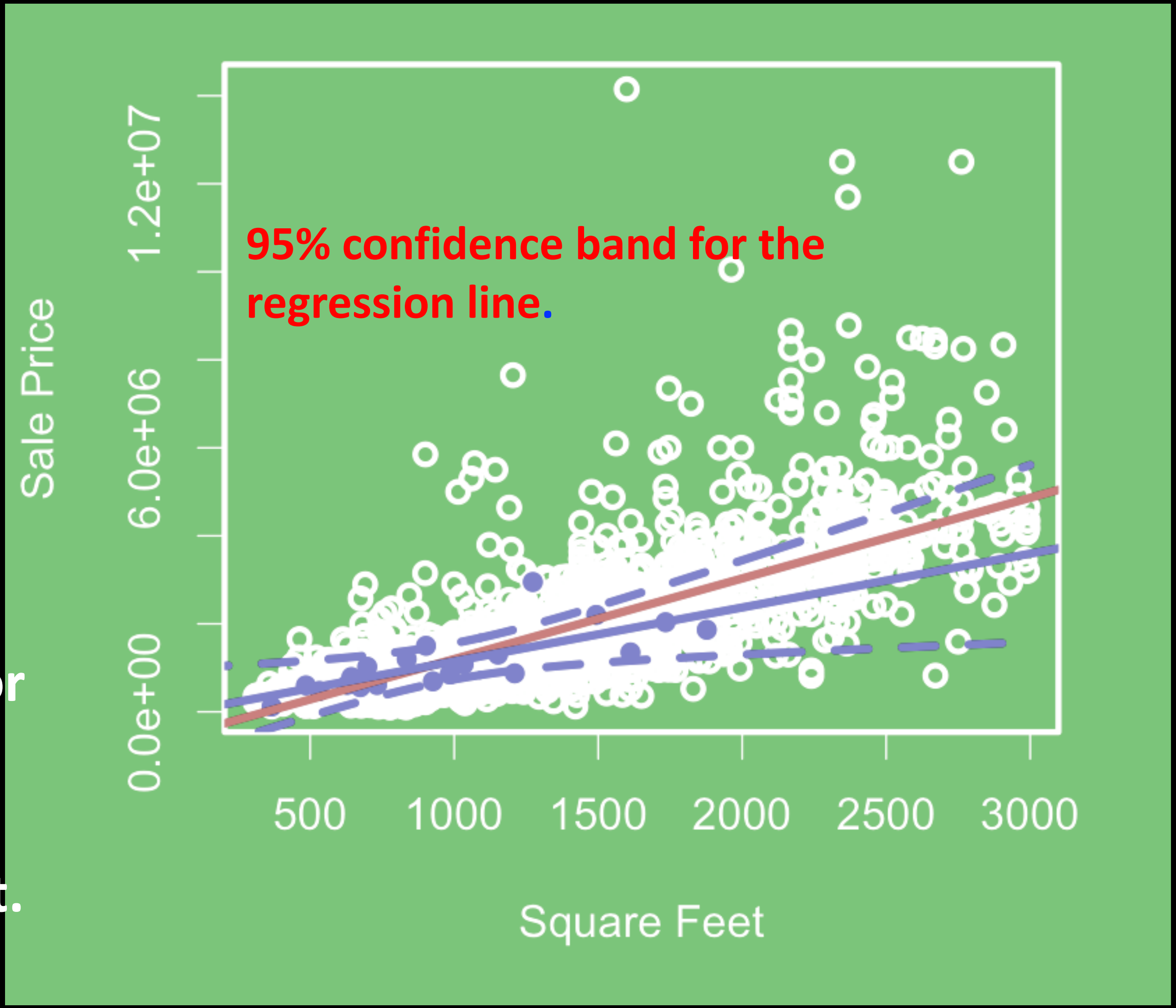
Sampling variability in regression estimates



Sampling variability in regression estimates



Sampling variability in regression estimates



- The confidence band centers at the sample estimate.
- It represents interval estimate for the regression line.
- Other inference on regression estimates can also be carried out.

Prediction

- Given a value of X
- The predicted value is $\widehat{Y} = b_0 + b_1 X$
- It is an estimate for the mean (average) value for Y given the X value.
- Most of the time, prediction is different from what is actually observed.
  - $Y - mean\ of\ Y$ (random variation)
  - $mean\ of\ Y\ - \widehat{Y}$ (estimation error)

Prediction

- **Extrapolation happens when one tries to give prediction on values of X outside the data range.**

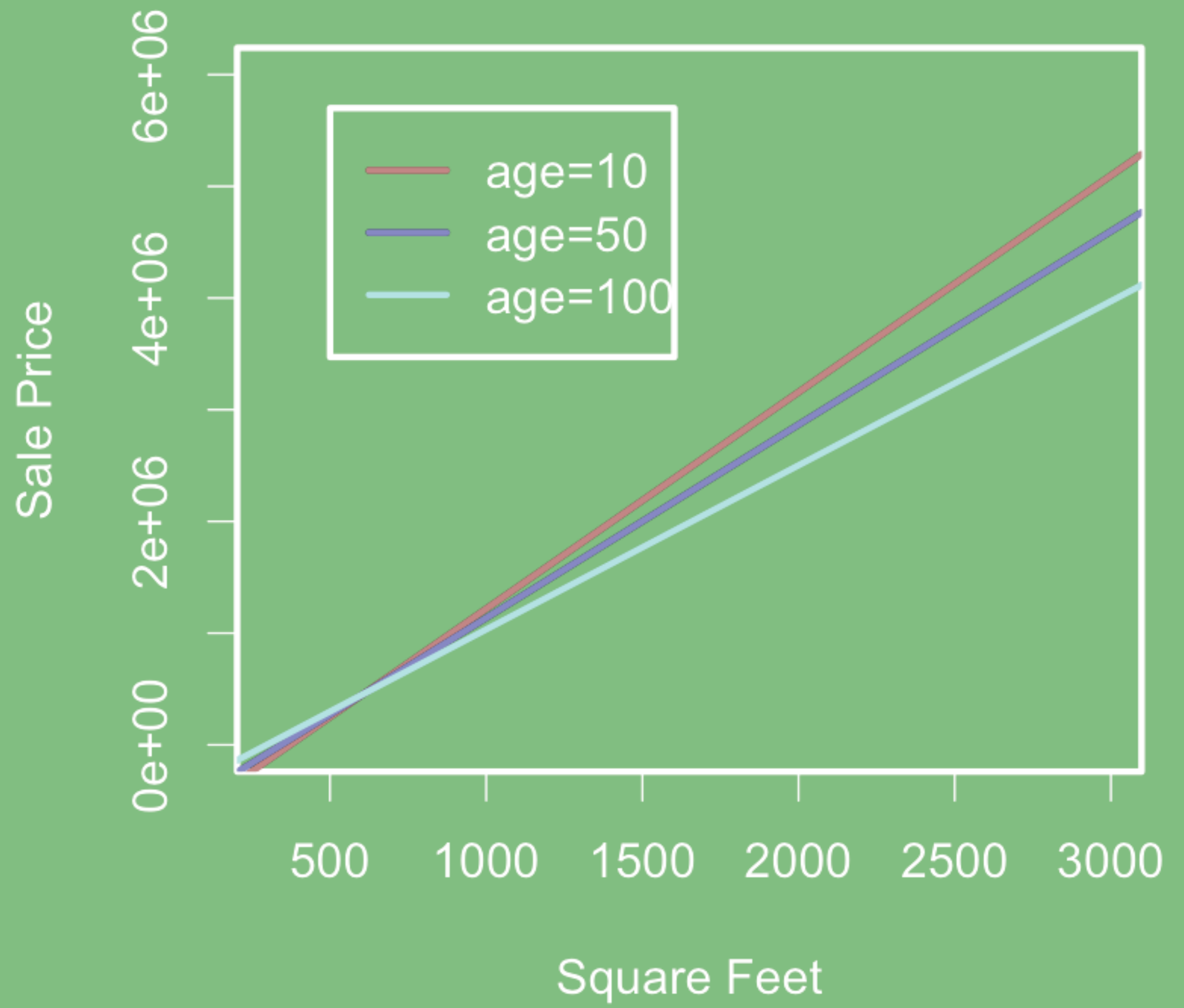
Multiple regression

- Y: response
- Multiple X variables
- $\hat{Y} = -546944 - 3265\ Age + 1770\ SQFT$

Multiple regression

- Y: response
- Multiple X variables
- $\hat{Y} = -546944 - 3265\ Age + 1770\ SQFT$
- Consider interaction
- $\hat{Y} = -753800 + 3173\ Age + 1992\ SQFT - 5.223\ Age \times SQFT$
- $\hat{Y} = (-753800 + 3173\ Age) + (1992 - 5.223\ Age)\ SQFT$

Multiple regression



Other considerations in regression analysis

- Outliers and influential observations.
- Model evaluation and comparison
- Model selection
- Hidden extrapolation
- Multiple testing *or* Multiple comparison

Extending linear regression

- Linear regression can be extended to nonlinear regression using transformed variables such as X^2, log Y , etc.
- Generalized linear models (GLM) are linear regression models for non-Gaussian Y variables such as categorical variables.
- Local regression applies linear regression using observations close to individual X values.
- Regression models have also been extended to more complex types of Y variables.

Concluding remarks

- Association patterns are everywhere.
- They represent information we can utilize
  - To explain
  - To estimate
  - To predict
- Association does not equal causation.
- Using a single set of data and search for various association patterns among a large number of variables is dangerous.
- Models, when used correctly, can be very useful.

Context Module

Lauren Hannah

Descriptive analytics of text

Statistical Thinking for Data Science and Analytics

Week 1: Introduction

		Lecture Modules		Context Modules
Week 2		Statistics and Probability I		Observational health studies
Week 3		Statistics and Probability II		Descriptive analytics of text
Week 4		Exploratory Data Analysis and Visualization		EDAV case studies
Week 5		Introduction to Bayesian Modeling		Bayesian modeling in Marketing