STAT5102
Regression in Practice

I. Simple Linear Regression

Department of Statistics
The Chinese University of Hong Kong

I. Simple Linear Regression

In this chapter, we shall cover

- Relations between two variables
- Scatter plots
- Univariate regression model
- Assumptions
- Estimation and method of least squares
- Inferences concerning β_1
- Inferences concerning β_0
- Estimation of the mean of the response variable for a given level of X
- Prediction of new observation
- Analysis of variance approach to regression analysis
- Measures of linear association between X and Y

Relations between variables

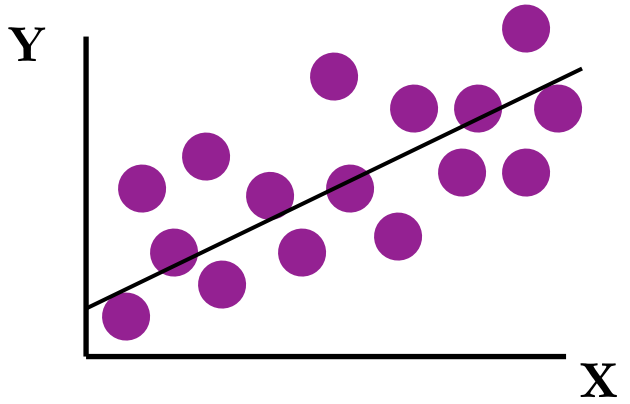
- Functional relation between two variables
  - Independent (predictor, explanatory) variable
  - Dependent (response) variable

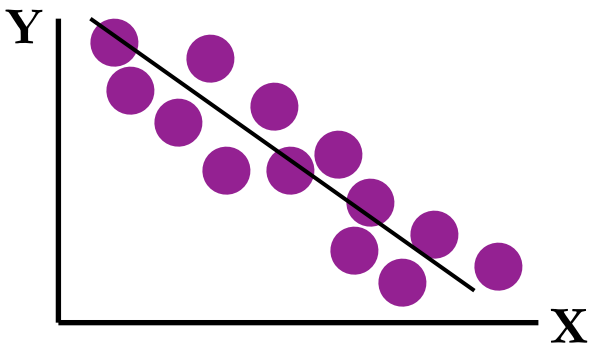
- Statistical relation between two variables

Scatter plots

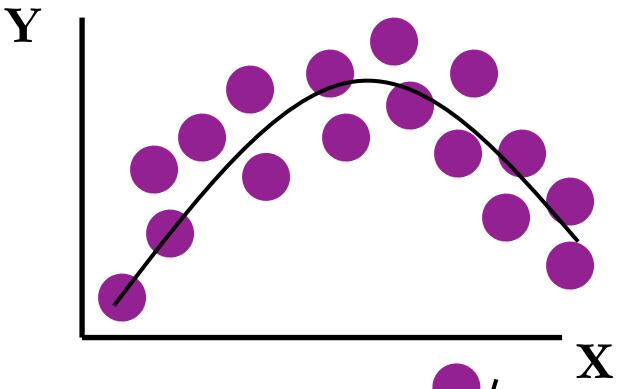
- A *scatter plot* (or scatter diagram) can be used to show the relationship between two variables.

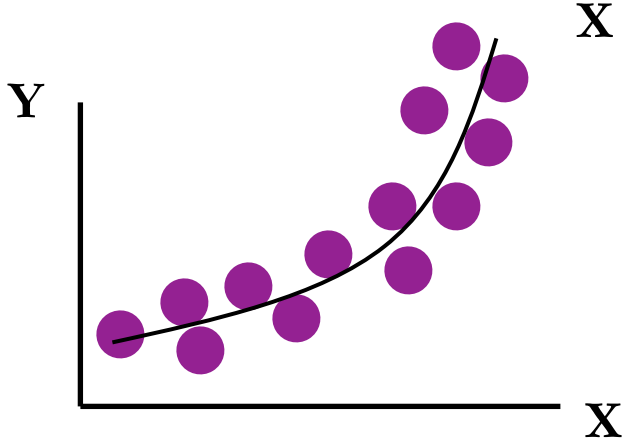
Linear relations



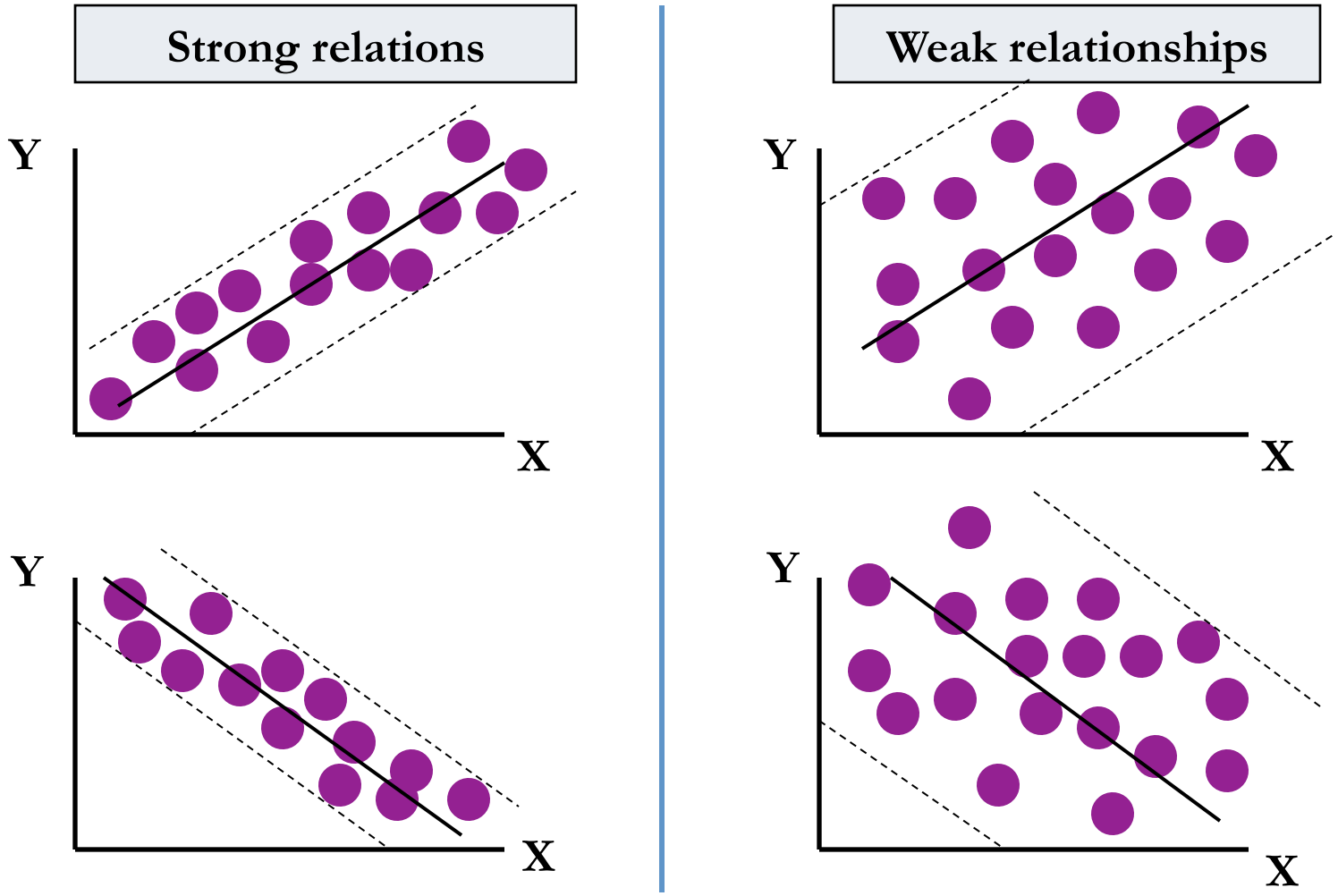


Curvilinear relations

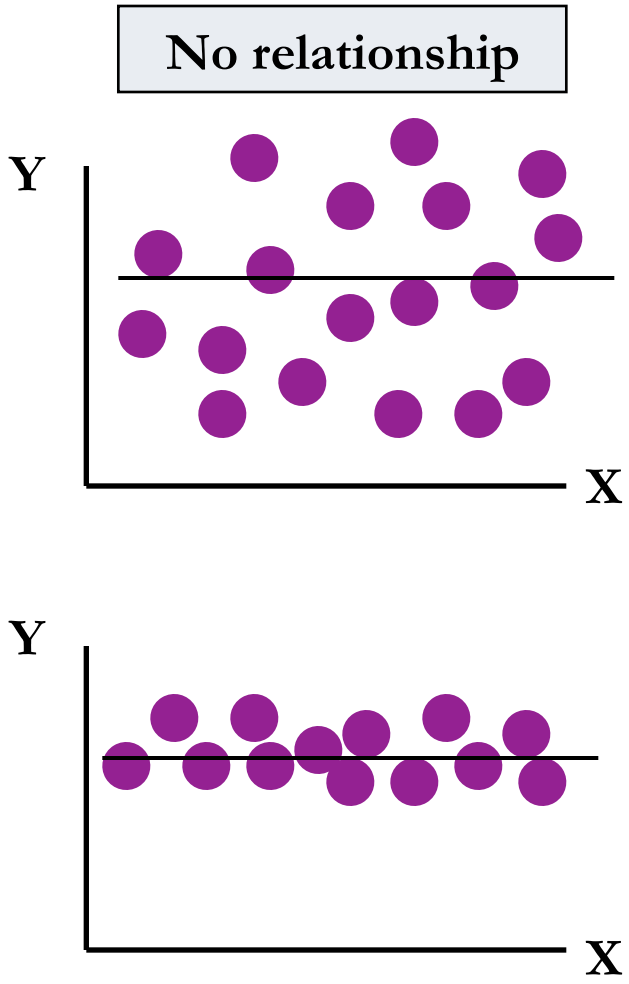




Types of Relations

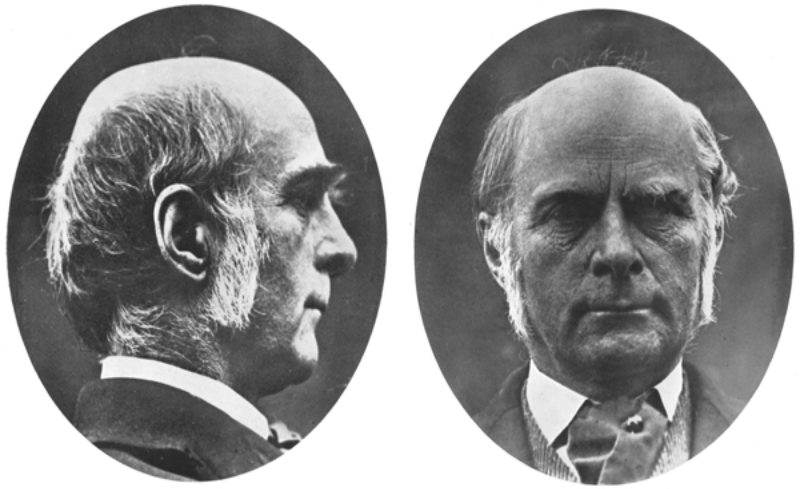


Types of Relations (Cont'd)



Historical Origins of Regression Models

- It was first introduced by Sir Francis Galton in the late 19th century.
- Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to "revert" or "regress" to the mean of the group.

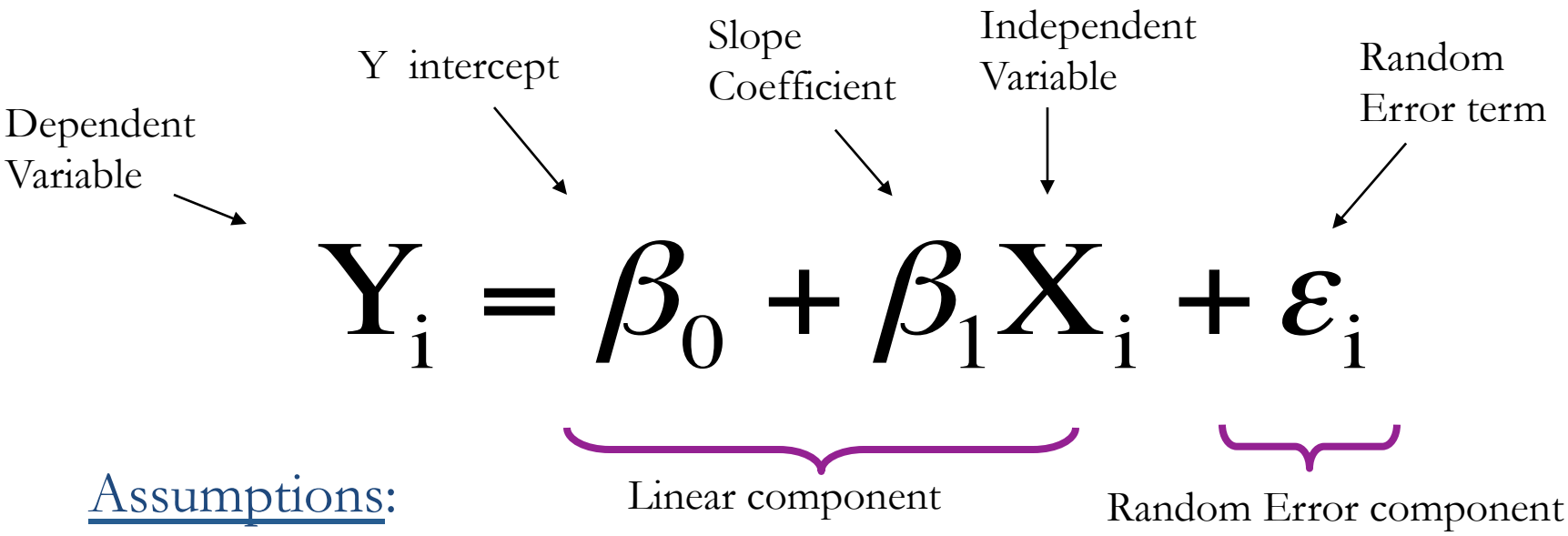


geographer, meteorologist, tropical explorer, founder of differential psychology, inventor of fingerprint identification, pioneer of statistical correlation and regression …

Basic concepts

- A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
- A scattering of points around the curve of statistical relationship.
- There is a probability distribution of Y for each level of X.
- The means of these probability distributions vary in some systematic fashion with X.

Simple Linear Regression Models



Assumptions:

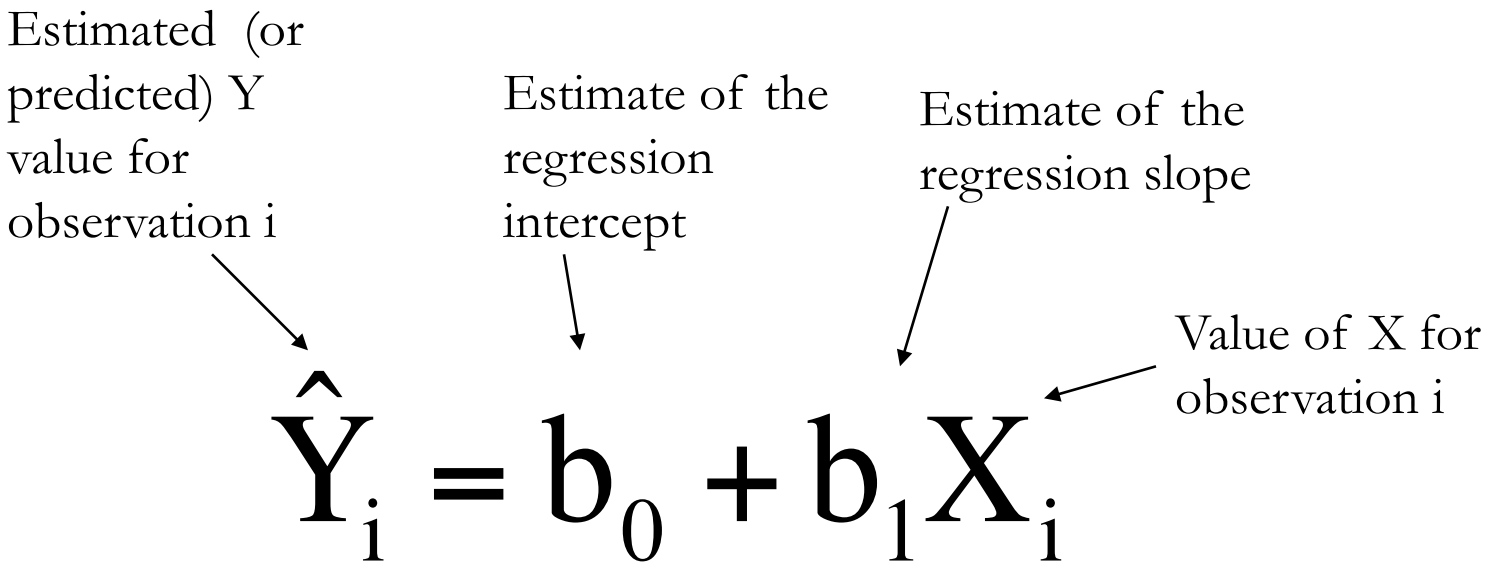
- $E(\varepsilon_i) = 0$
- Variance $(\varepsilon_i) = \sigma^2$
- Covariance $(\varepsilon_i, \varepsilon_j) = 0$

Model Assumptions

- Normality of Error
  - Error values (ε) are (normally) distributed for any given value of X.
- Homoscedasticity
  - The probability distribution of the errors has constant variance.
- Independence of Errors
  - Error values are statistically independent.

Simple Linear Regression Equation

The simple linear regression equation provides an *estimate* of the population regression line.



Slopes and the Intercept

- b_0 is the estimated average value of Y when the value of X is zero.
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X.

Example

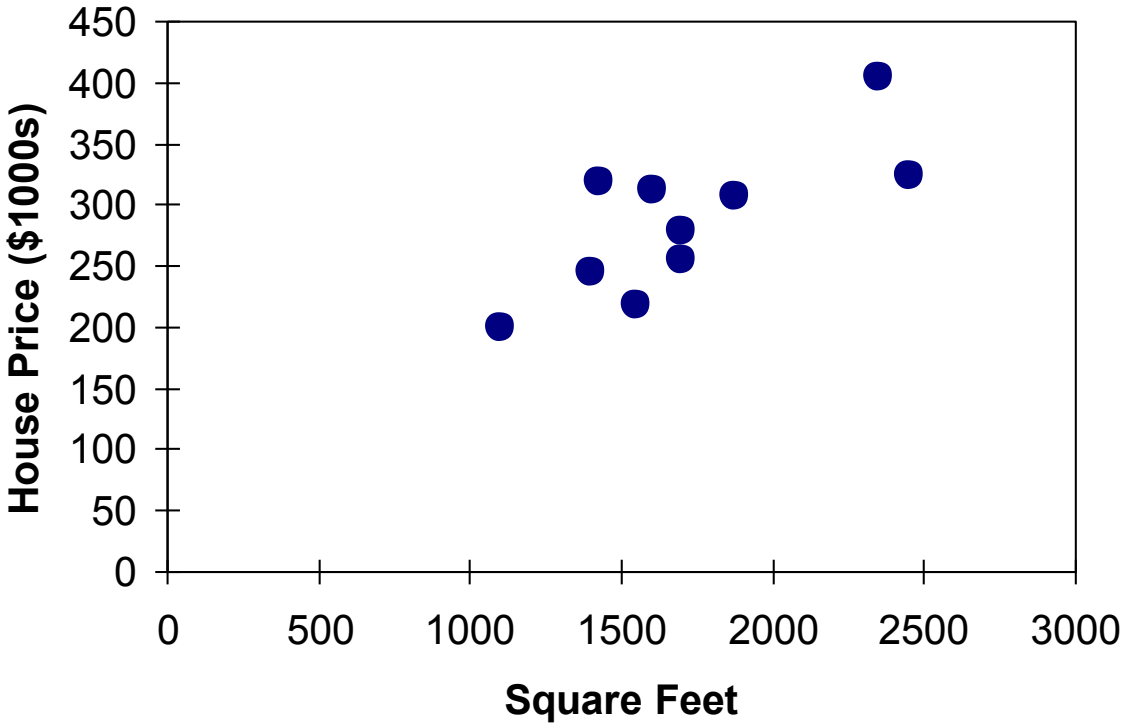
- A real estate agent wishes to examine the relation between the selling price of a home and its size (measured in square feet).

- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in \$'000s
  - Independent variable (X) = square feet

Sample Data for House Price Model

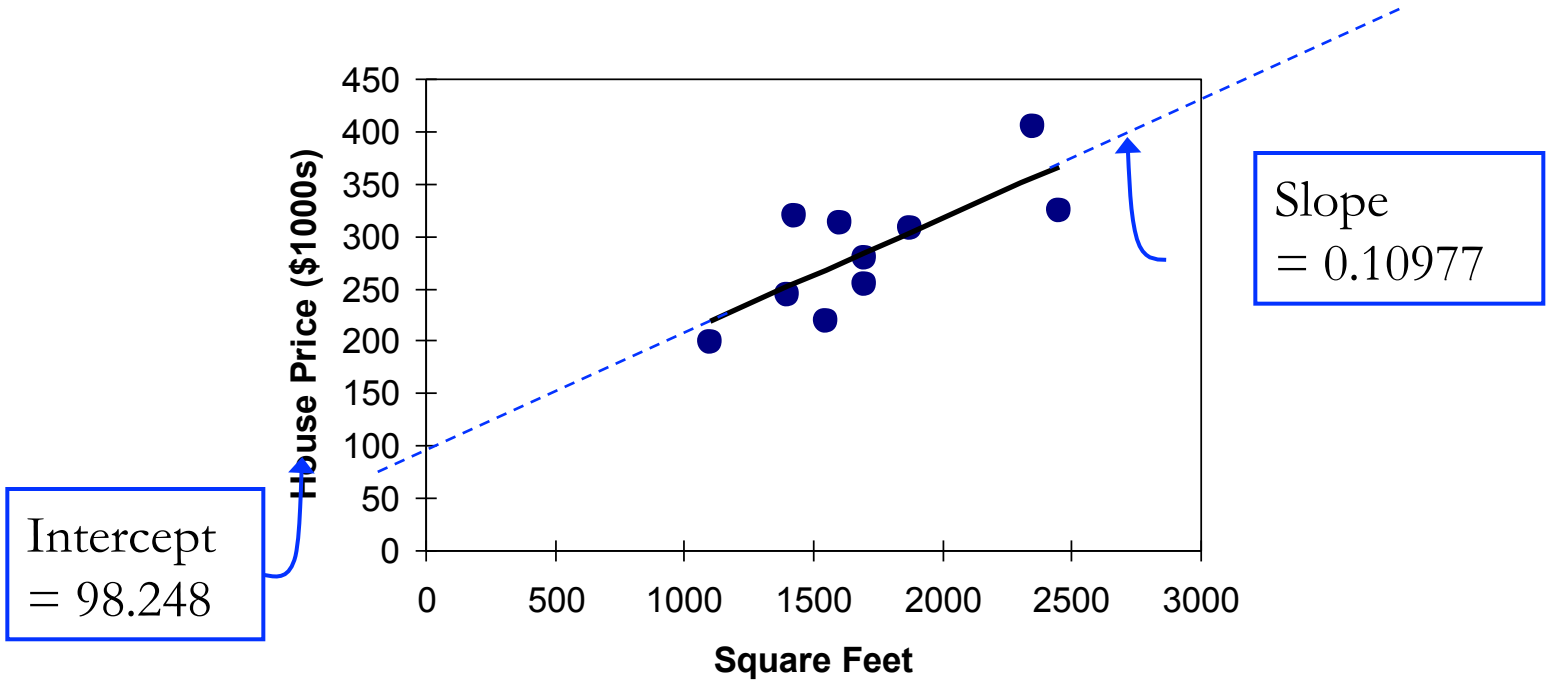
House Price in $'000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

House price model: scatter plot



Graphical Presentation

House price model: scatter plot and regression line



$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)

- Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977\,(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X.

- Here, $b_1 = 0.10977$ tells us that the average value of a house increases by 0.10977($1000) = $109.77, on average, for each additional one square foot of size.

Predictions using the Fitted Model

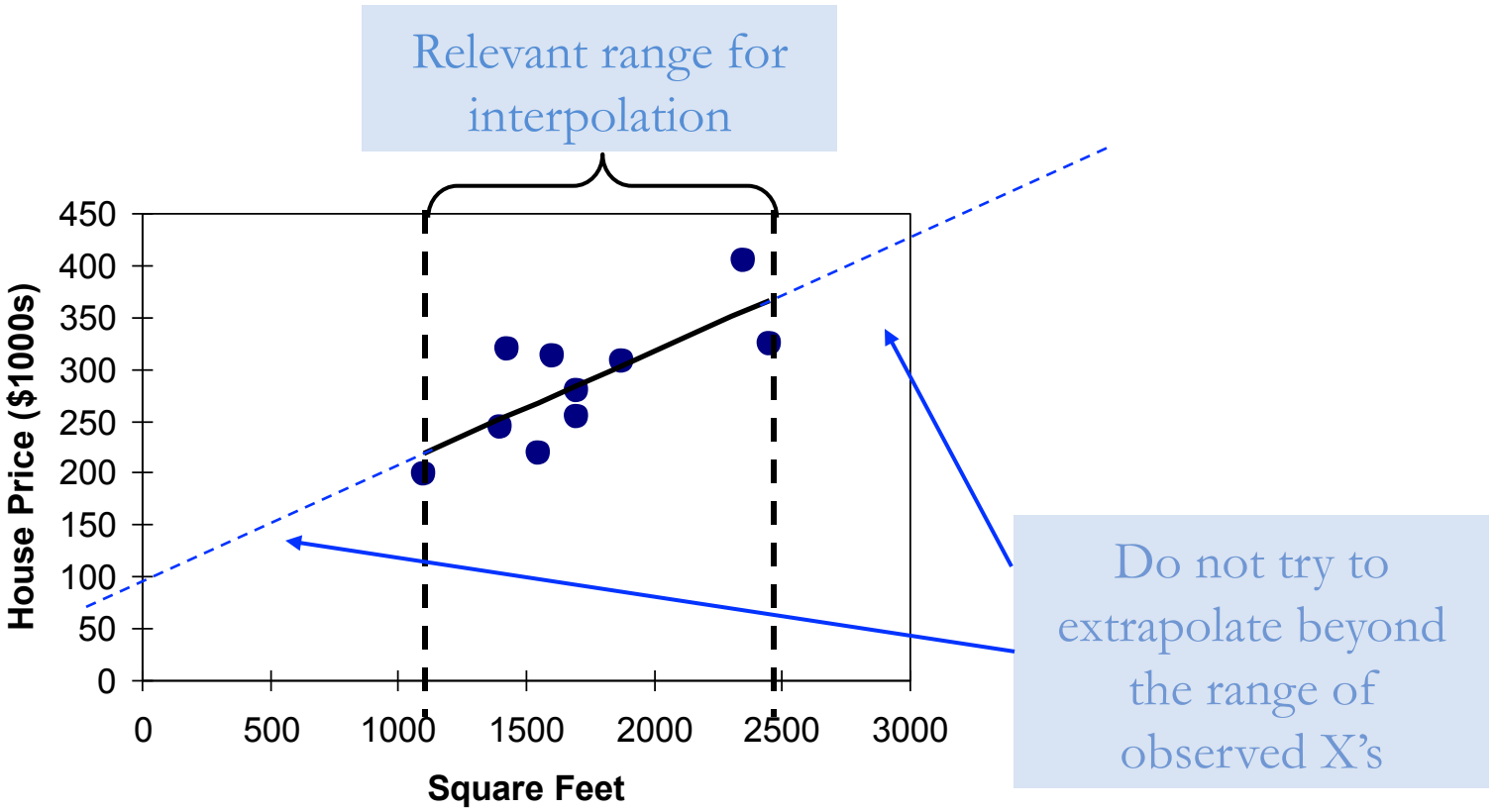
To predict the price for a house with 2,000 square feet, we write:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$
$$= 98.25 + 0.1098\,(2000)$$
$$= 317.85$$

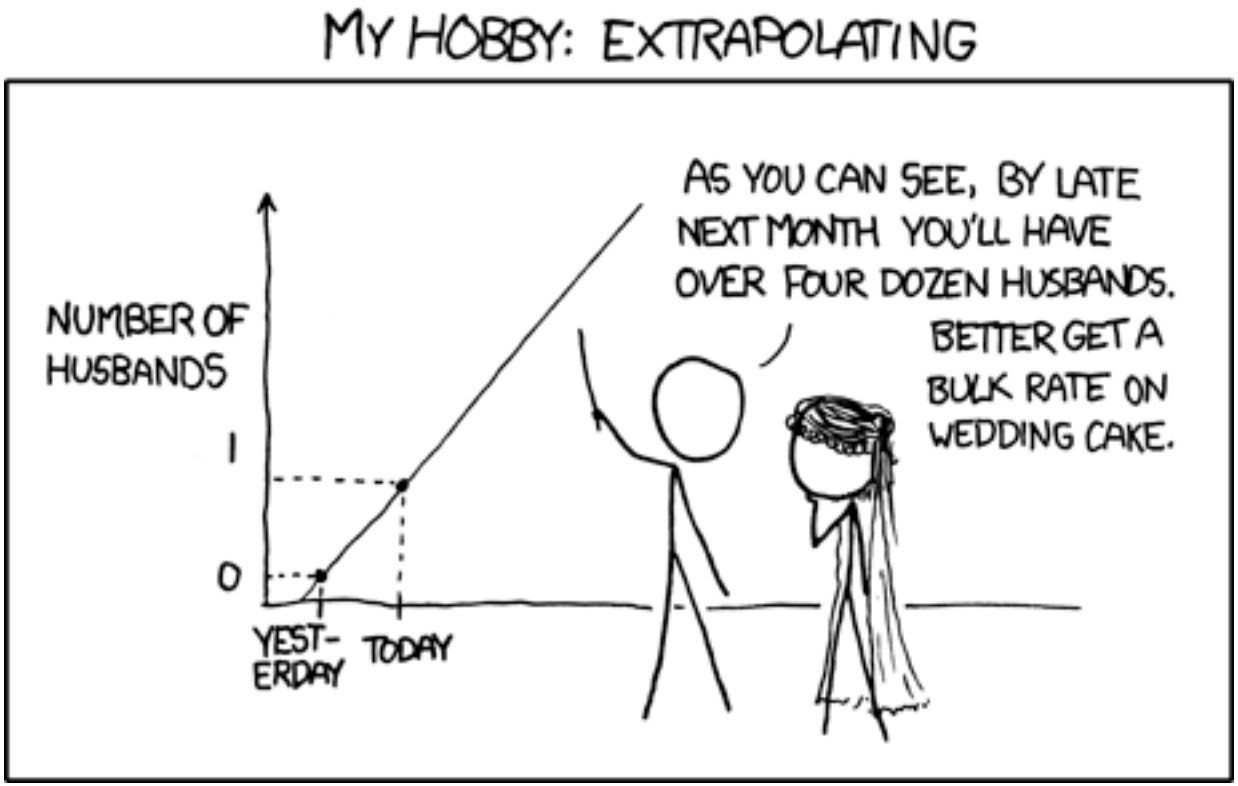
The predicted price for a house with 2,000 square feet is 317.85($1,000s) = $317,850.

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Interpolation vs. Extrapolation



Estimation (Method of Least Squares)

- b_0 and b_1 can be obtained by choosing the values of b_0 and b_1 so that they minimise the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum \{Y_i - \hat{Y}_i(b)\}^2 = \min \sum \{Y_i - (b_0 + b_1 X_i)\}^2$$

Estimation (Method of Least Squares)

- In fact, b_0 and b_1 have nice closed-form solutions.

- $b_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

- $b_0 = \overline{Y} - b_1 \overline{X}$

Estimation of error terms variance σ^2

- The estimator of σ^2 is

$$S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- Note that S^2 is an unbiased estimator of σ^2.

Method of Maximum Likelihood

- $b_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

- $b_0 = \overline{Y} - b_1 \overline{X}$

- The estimator of σ^2 is $\dfrac{SSE}{n} = \dfrac{n-2}{n} S^2$

Normal Error Model

- MLE of b_0 = LSE of b_0 (unbiased)
- MLE of b_1 = LSE of b_1 (unbiased)
- MLE of σ^2 < unbiased estimator of σ^2 [asymptotically unbiased]

Example

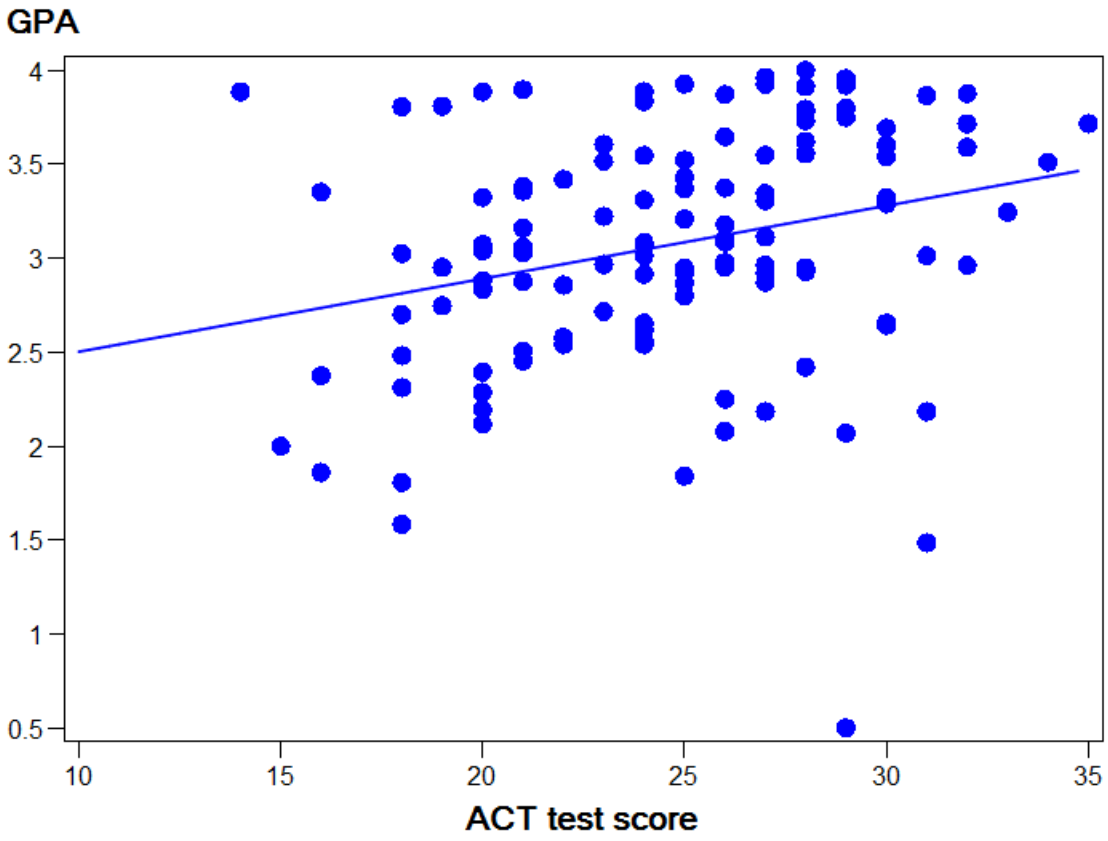
The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The data are given in the disk.

Regression output from SAS

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	2.11405	0.32089	6.59	<.0001
ACT	ACT test score	**1**	0.03883	0.01277	3.04	0.0029

$$b_0 = 2.11405,\ b_1 = 0.03883,\ \hat{Y} = 2.11405 + .03883X$$

A Scatter Plot with Regression Line



The Slope

- Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

  Answer = 3.27895

- What is the point estimate of the change in the mean response when the entrance test score increase by one point?

  Answer = 0.03883

More on Inference

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

Assumptions:

- X_i are known constants, $i = 1,\ldots n$ (Fixed Design)
- $\varepsilon_i \sim N(0, \sigma^2)$, *i.i.d.*

Therefore, $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, *ind.*

Distribution of b_1

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \sum_{i=1}^{n} k_i Y_i$$

where

$$k_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Distribution of b_1

Hence, we can deduce that

$b_1 \sim N(\beta_1, \sigma^2/S_{xx})$, where $S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2$

It follows that

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{(n-2)}$$

with

$$s^2(b_1) = \frac{MSE}{S_{xx}}$$

Proof will be discussed in class

Testing (Two-sided test of β_1)

H_0: $\beta_1 = 0$ (no linear relationship)
H_1: $\beta_1 \neq 0$ (linear relationship exists between X and Y)

Test statistics: $T = \dfrac{b_1}{s\{b_1\}}$

Decision rule: Reject H_0 if $|T| \geq t_{\alpha/2,n-2}$

Testing (Two-sided confidence interval of β_1)

Two-sided 100(1-α)% C.I. for β_1 :

$$\left(b_1 - t_{\alpha/2,n-2}s\{b_1\}, \quad b_1 + t_{\alpha/2,n-2}s\{b_1\}\right)$$

One-sided Test for $\beta \downarrow 1$

$H_0: \beta_1 \leq k$
$H_1: \beta_1 > k$

Test statistics: T = $\dfrac{b_1 - k}{s\{b_1\}}$

Decision rule: Reject H_0 if $T \geq t_{\alpha, n-2}$

One-sided Test for $\beta \downarrow 1$

$H_0: \beta_1 \geqq k$
$H_1: \beta_1 < k$

Test statistics: T = $\dfrac{b_1 - k}{s\{b_1\}}$

Decision rule: Reject H_0 if $T \leq -t_{\alpha, n-2}$

Distribution of b_0

Recall that $b_0 = \overline{Y} - b_1 \overline{X}$

Hence, we can write

$$b_0 \sim N(\beta_0, \ \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right])$$

$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{(n-2)}$$

Proof will be discussed in class

Two-sided Tests for $\beta \downarrow 0$

$$H_0: \beta_0 = k$$
$$H_1: \beta_0 \neq k$$

Test statistics: $T = \dfrac{b_0 - k}{s\{b_0\}}$

Decision rule: Reject H_0 if $|T| \geq t_{\alpha/2, n-2}$

Two-sided Confidence Interval for β_1 0

Two-sided 100(1-α)% C.I. for β_0 :

$$\left(b_0 - t_{\alpha/2,n-2}s\{b_0\}, \quad b_0 + t_{\alpha/2,n-2}s\{b_0\}\right)$$

One-sided Tests for $\beta \downarrow 0$

$$H_0: \beta_0 \leq k$$
$$H_1: \beta_0 > k$$

Test statistics: $T = \dfrac{b_0 - k}{s\{b_0\}}$

Decision rule: Reject H_0 if $T \geq t_{\alpha, n-2}$

One-sided Test for $\beta \downarrow 0$

$H_0: \beta_0 \geqq k$
$H_1: \beta_0 < k$

Test statistics: $T = \dfrac{b_0 - k}{s\{b_0\}}$

Decision rule: Reject H_0 if $T \leq -t_{\alpha, n-2}$

Estimate the mean of the response variable for given X

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in \$000) and assessed value (X, in \$000). The houses in the city had been reassessed at full value 1 year prior to the study.

Estimate of *the average* selling price for houses with an assessed value of \$70,000.

Estimate the mean of the response variable for given X

Let X_h denote the level of X for which we wish to estimate the mean response [to be estimated by $\hat{Y}_h$]. (Note: Given X_h, the mean response is $E(Y_h) = \beta_0 + \beta_1 X_h$ according to the model)

For estimation, Given X_h: $\hat{Y}_h = b_0 + b_1 X_h$

Distribution of $E(Y_h) - \hat{Y}_h$:

$$E(Y_h) - \hat{Y}_h \sim N(0, \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right])$$

Confidence Interval for $E(Y|h)$

Two-sided 100(1-α)% C.I. for $E(Y_h)$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}} \right)$$

Prediction of a new observation $Y|h(new)$

Example

Suppose you want to develop a model to predict selling price of homes based on assessed value. A sample of 30 recently sold single-family houses in a small city is selected to study the relationship between selling price (Y, in \$000) and assessed value (X, in \$000). The houses in the city had been reassessed at full value 1 year prior to the study.

Estimate the selling price of ***an individual*** house with an assessed value of \$70,000.

Prediction of a new observation $Y_{h(new)}$

Prediction of $Y_{h(new)}$ corresponding to a given level X of the predictor variable by $\hat{Y}_h$.

$$\hat{Y}_h = b_0 + b_1 X_h$$

Distribution of $Y_{h(new)} - \hat{Y}_h$:

$$Y_{h(new)} - \hat{Y}_h \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\right)$$

Confidence Interval for $Y_{h(new)}$

Two-sided 100(1-α)% C.I. for $Y_{h(new)}$:

$$\left(\hat{Y}_h - t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]}, \quad \hat{Y}_h + t_{\alpha/2,n-2} s \sqrt{\left[1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right]} \right)$$

ANOVA and Regression Analysis

- Partitioning of Total Sum of Squares
- Mean Squares
- Analysis of Variance (ANOVA) Table

ANOVA and Regression

The difference between the observed value of the response variable and the mean value of the response variable is called the *total deviation* and is equal to

$$y - \bar{y}$$

The difference between the predicted value of the response variable and the mean value of the response variable is called the *explained deviation* and is equal to

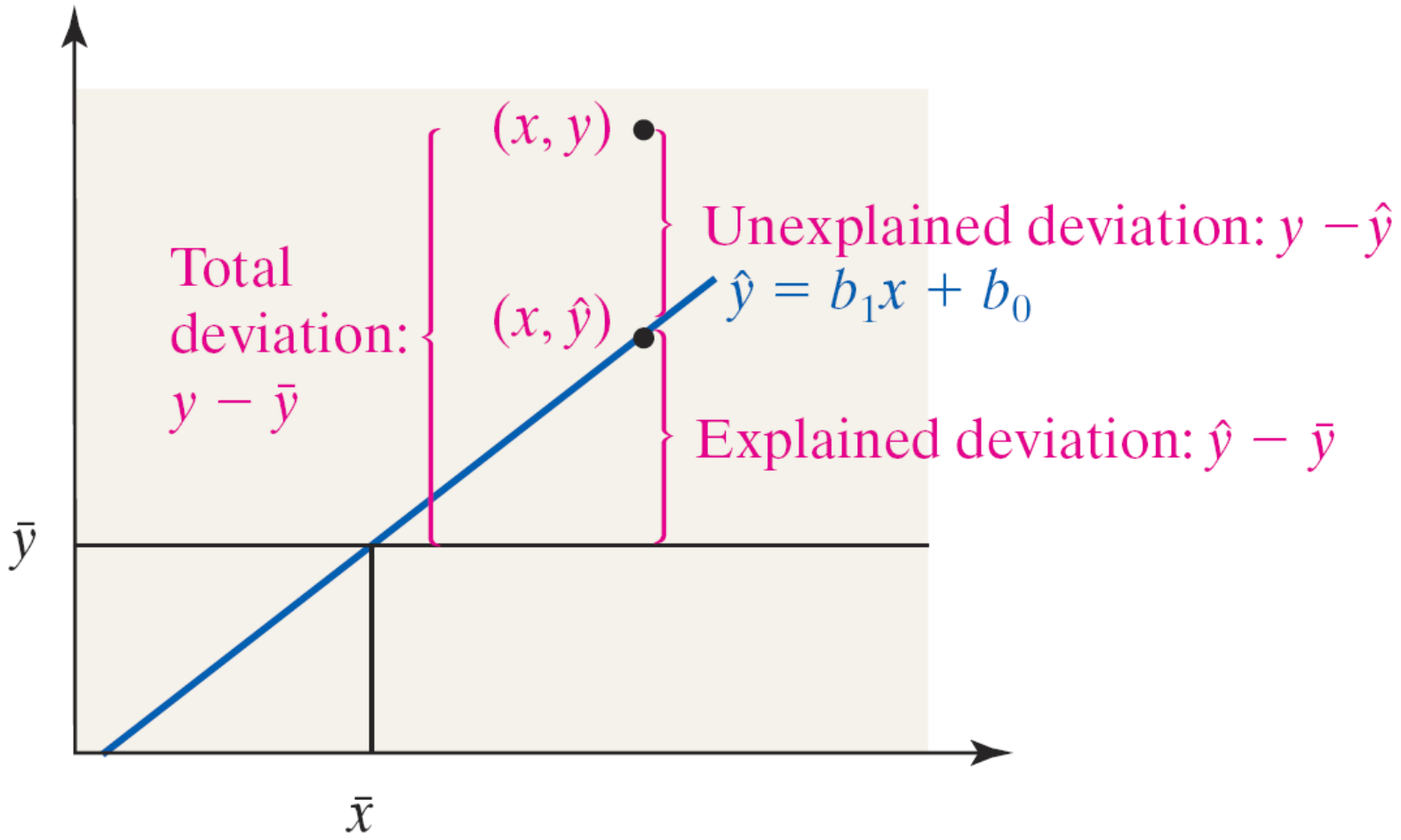
$$\hat{y} - \bar{y}$$

ANOVA and Regression

The difference between the observed value of the response variable and the predicted value of the response variable is called the **unexplained deviation** and is equal to

$$y - \hat{y}$$

ANOVA and Regression



Total Variation

= Unexplained Variation + Explained Variation

$$(y-\bar{y})^2=(y-\hat{y})^2+(\hat{y}-\bar{y})^2$$

ANOVA and Regression

Total Variation = Unexplained Variation + Explained Variation

$$1 = \frac{\text{Unexplained Variation}}{\text{Total Variation}} + \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$\frac{\text{Explained Variation}}{\text{Total Variation}} = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

ANOVA Table

	Sum of Squares (SS)	Degrees of freedom (df)	Mean squares (MS)	F
Regression	SSR	1	MSR = SSR/1	MSR/MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	SST	n-1		

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist between X and Y)

Test Statistics: $F = \text{MSR}/\text{MSE}$

Rejection Rule: reject the null hypothesis if $F > F_{(\alpha, 1, n-2)}$

Measures of Linear Association between X and Y

- Coefficient of Determination R^2

  a) $R^2 = SSR/SSTO = 1 - SSE/SSTO$

  b) $0 \leq R^2 \leq 1$

- Coefficient of Correlation

  a) $r = \pm\sqrt{R^2}$

  b) A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative