Chapter1.

Simple Linear Regression

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

한국방송통신대학교
프라임칼리지

Contents

1. Introduction
2. Best Fitting Line
3. Simple Linear Regression
4. Common Error Variance
5. Coefficient of Determination

1. Introduction

1. Simple Linear Regression

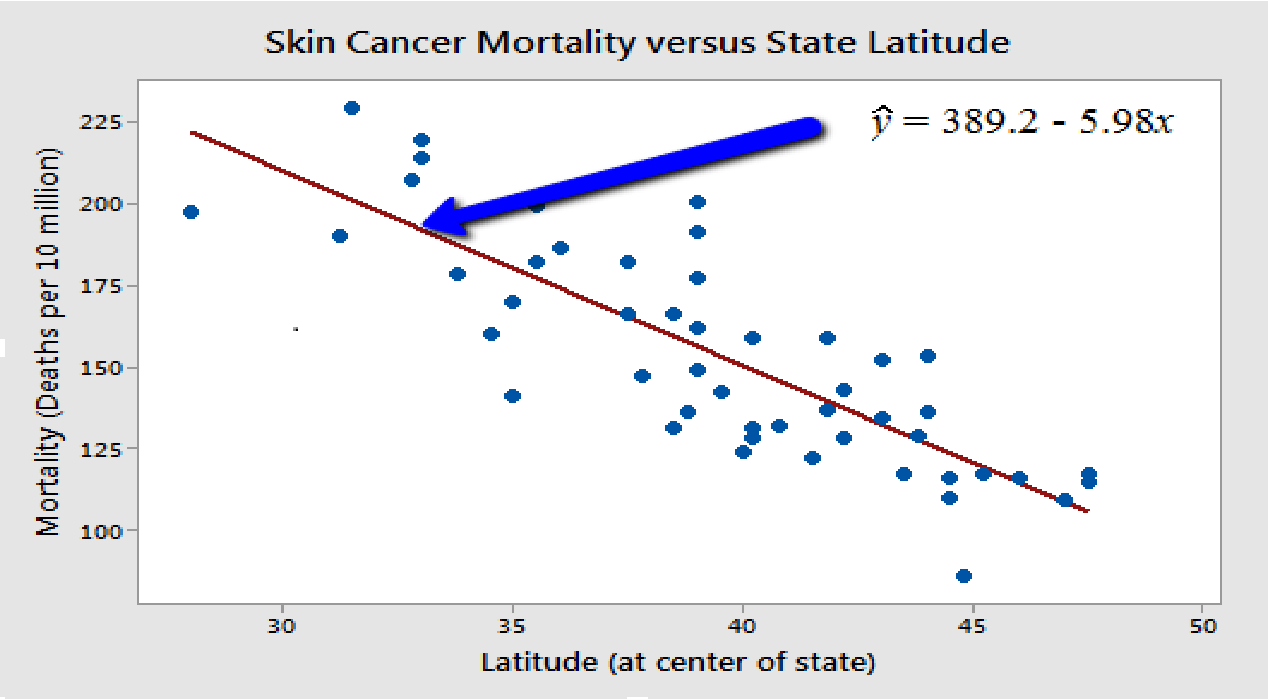
- A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables
  - One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
  - The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

2. Deterministic Relationship

- Equations exactly describe the relationship between the two variables
  - Circumference = π × diameter
  - Ohm's Law: $I = \frac{V}{r}$

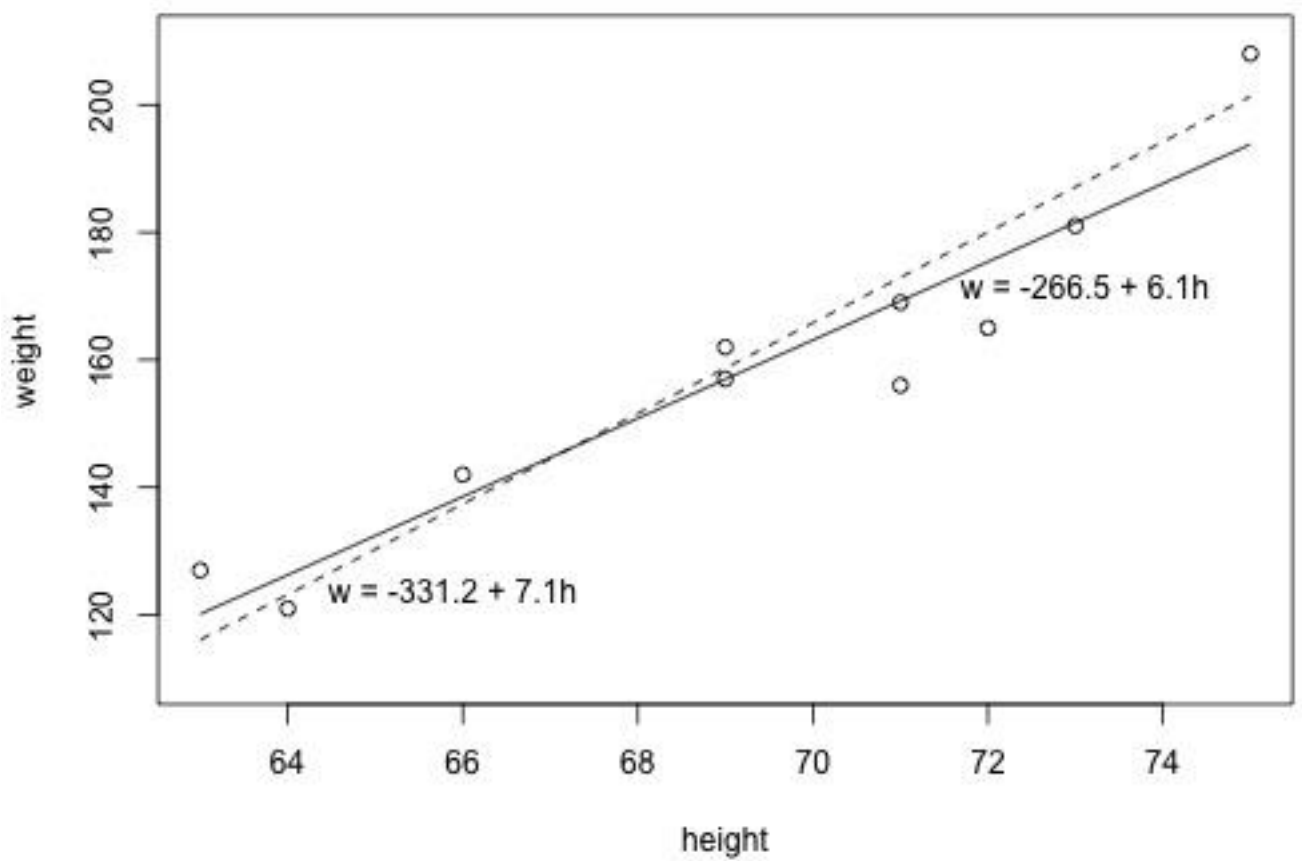
3. Statistical Relationship

- The relationship between the variables is not perfect.



2. Best Fitting Line

1. What is the "Best Fitting Line"?



2. Notation

- y_i: the observed response for experimental unit i
- x_i: the predictor value for experimental unit i
- $\hat{y}_i$: the predicted response (or fitted value) for experimental unit i
- experimental unit: the object or person on which the measurement is made

3. Prediction Error

- $w = -266.5 + 6.1h$
- $x_1 = 63, y_1 = 127$
- $\hat{y}_1 = -266.5 + 6.1 \times 63 = 117.8$
- prediction error (residual error)
  - $e_i = y_i - \hat{y}_i$
  - e.g. $y_1 - \hat{y}_1 = 127 - 117.8 = 9.2$

4. Least Squares

- A line that fits the data "best" will be one for which the n prediction errors — one for each observed data point — are as small as possible in some overall sense. One way to achieve this goal is to invoke the "least squares criterion," which says to "minimize the sum of the squared prediction errors."

4. Least Squares

- $L = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - $\hat{y}_i = b_0 + b_1 x_i$
  - $e_i = y_i - \hat{y}_i$: the prediction error for data point i
  - $e_i{}^2 = (y_i - \hat{y}_i)^2$: the squared prediction error for data point i
- Example
  - (the solid line) $w = -266.5 + 6.1h,\ L = 597.4$
  - (the dashed line) $w = -331.2.5 + 7.1h,\ L = 766.5$

5. Least Squares Regression Line

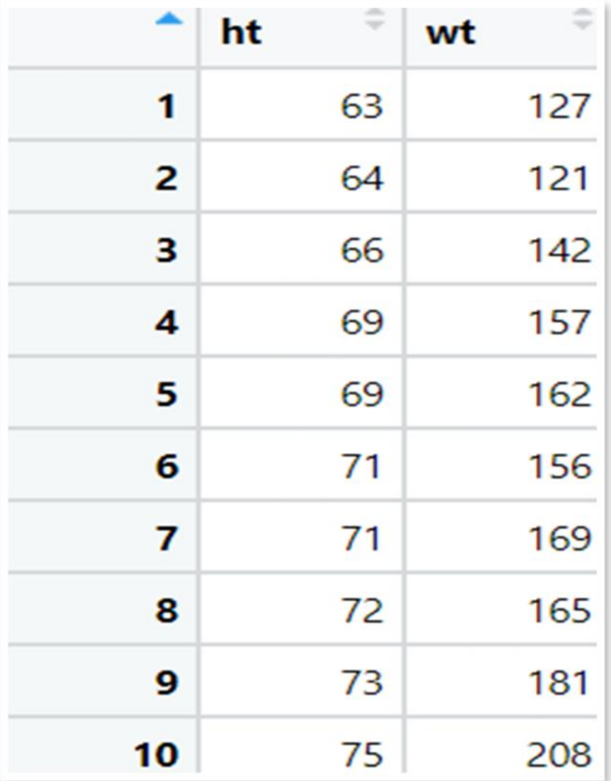
- $\hat{y}_i = b_0 + b_1 x_i$
  - $L = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$
  - (intercept) $b_0 = \bar{y} - b_1 \bar{x}$
  - (slope) $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - The least squares line passes through the point $(\bar{x}, \bar{y})$

6. Sign of the Slope

- $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

7. Dataset

- Student Height and Weight Dataset



	ht	wt
1	63	127
2	64	121
3	66	142
4	69	157
5	69	162
6	71	156
7	71	169
8	72	165
9	73	181
10	75	208

8. Code: Height and Weight

```
heightweight = read.table("student_height_weight.txt", header=T)

attach(heightweight)

model = lm(formula = wt ~ ht)

summary(model)

detach(heightweight)
```

9. Results: Height and Weight

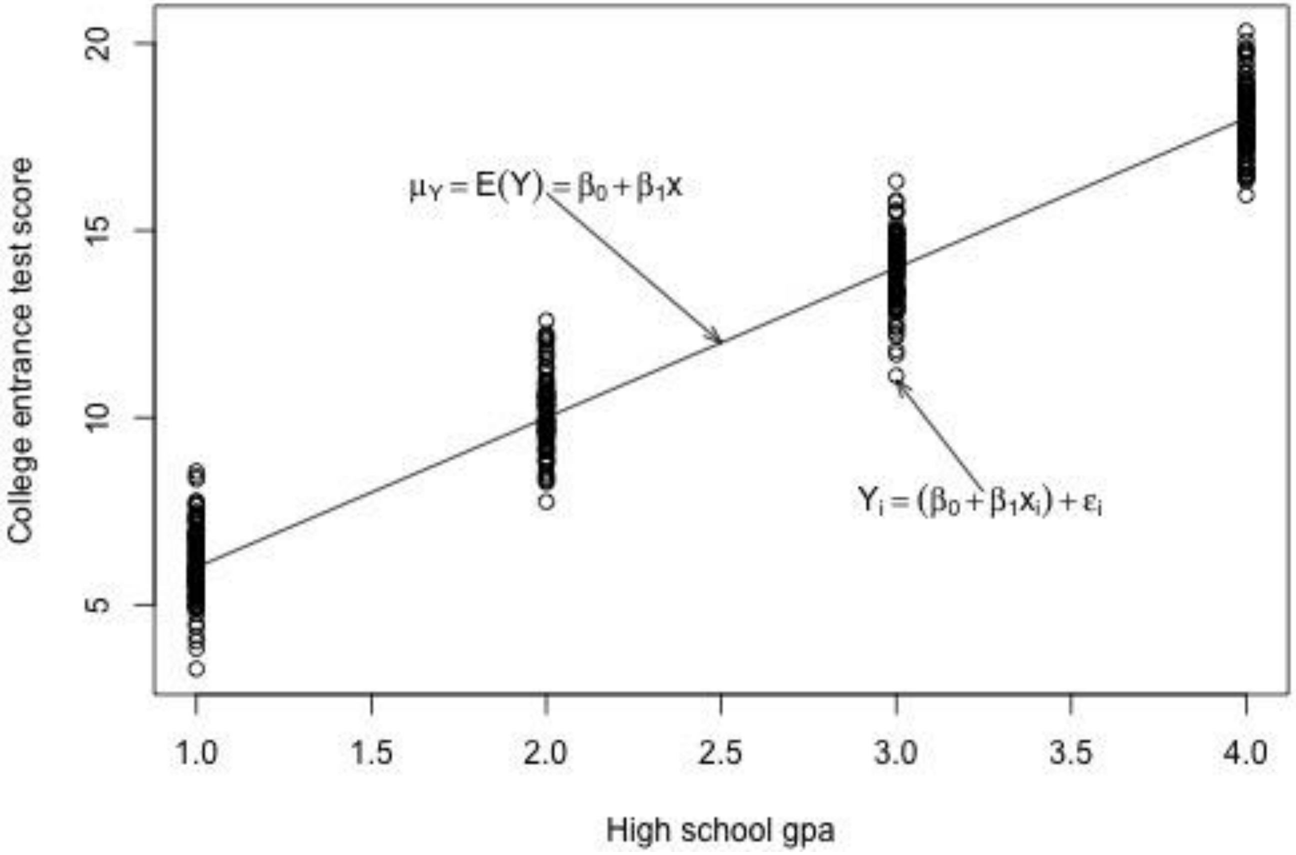
```
Call:
lm(formula = wt ~ ht)

Coefficients:
             Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -266.5344    51.0320  -5.223    8e-04 ***
ht             6.1376     0.7353   8.347 3.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

10. Code: Prediction

```
> predict(model, newdata=data.frame(ht=c(66, 67)))
       1        2
138.5460 144.6836
```

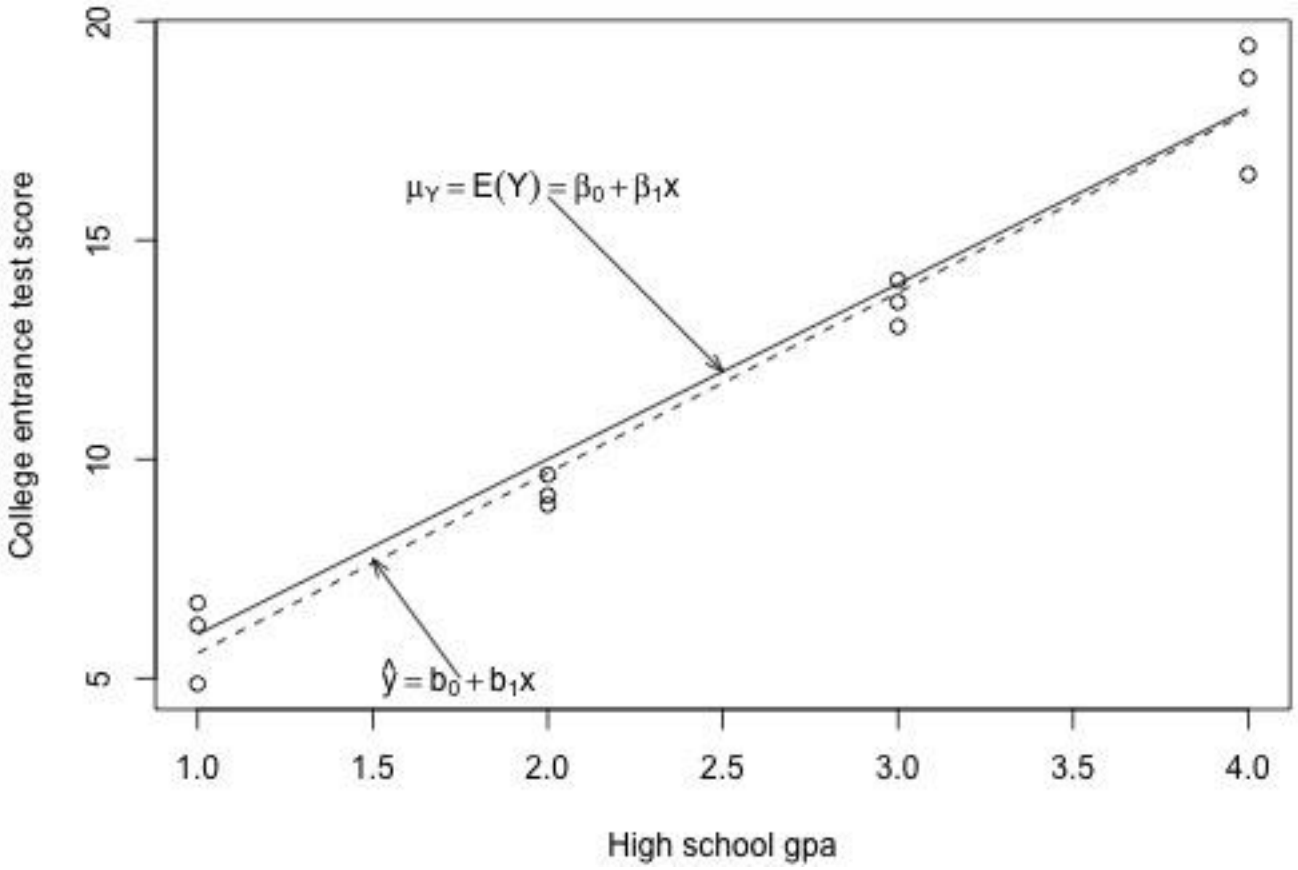
3. Simple Linear Regression

1. Population Regression Line



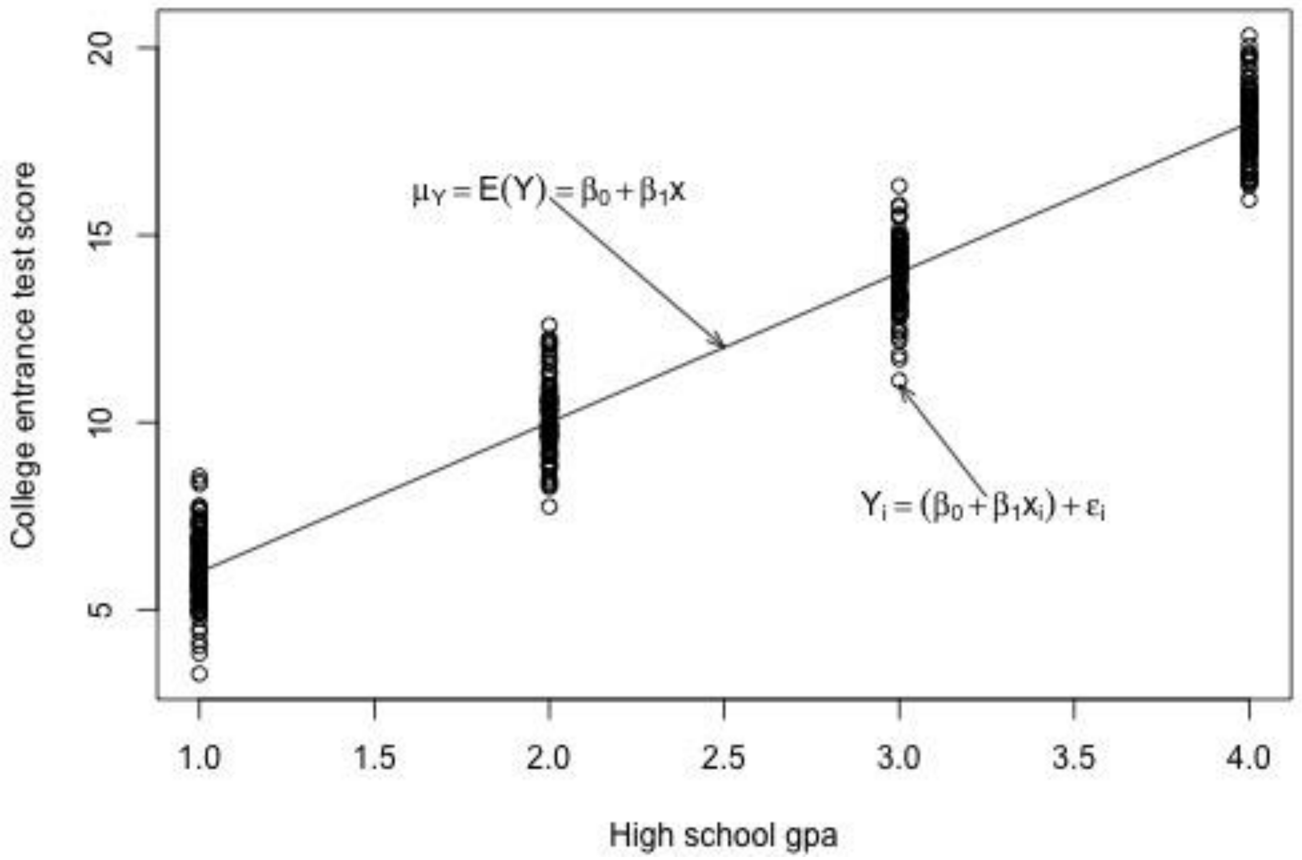
- $\mu_Y = E(Y) = \beta_0 + \beta_1 x$
- $E(Y_i) = \beta_0 + \beta_1 x_i$
- $y_i = E(Y_i) + \epsilon_i$

2. Least Squares Regression Line



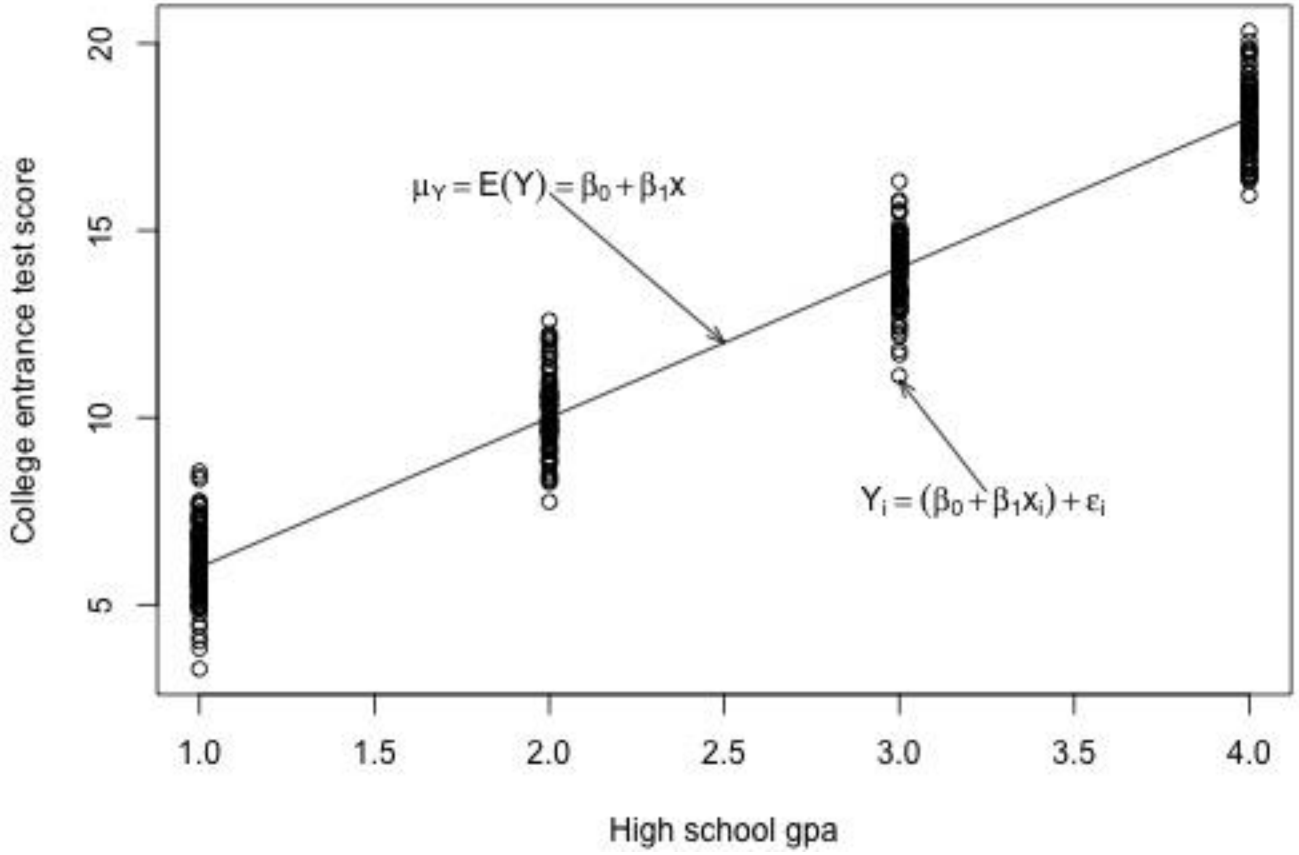
- $\mu_Y = E(Y) = \beta_0 + \beta_1 x$
- $\hat{y}_i = b_0 + b_1 x_i$

3. Assumptions: Linear Function



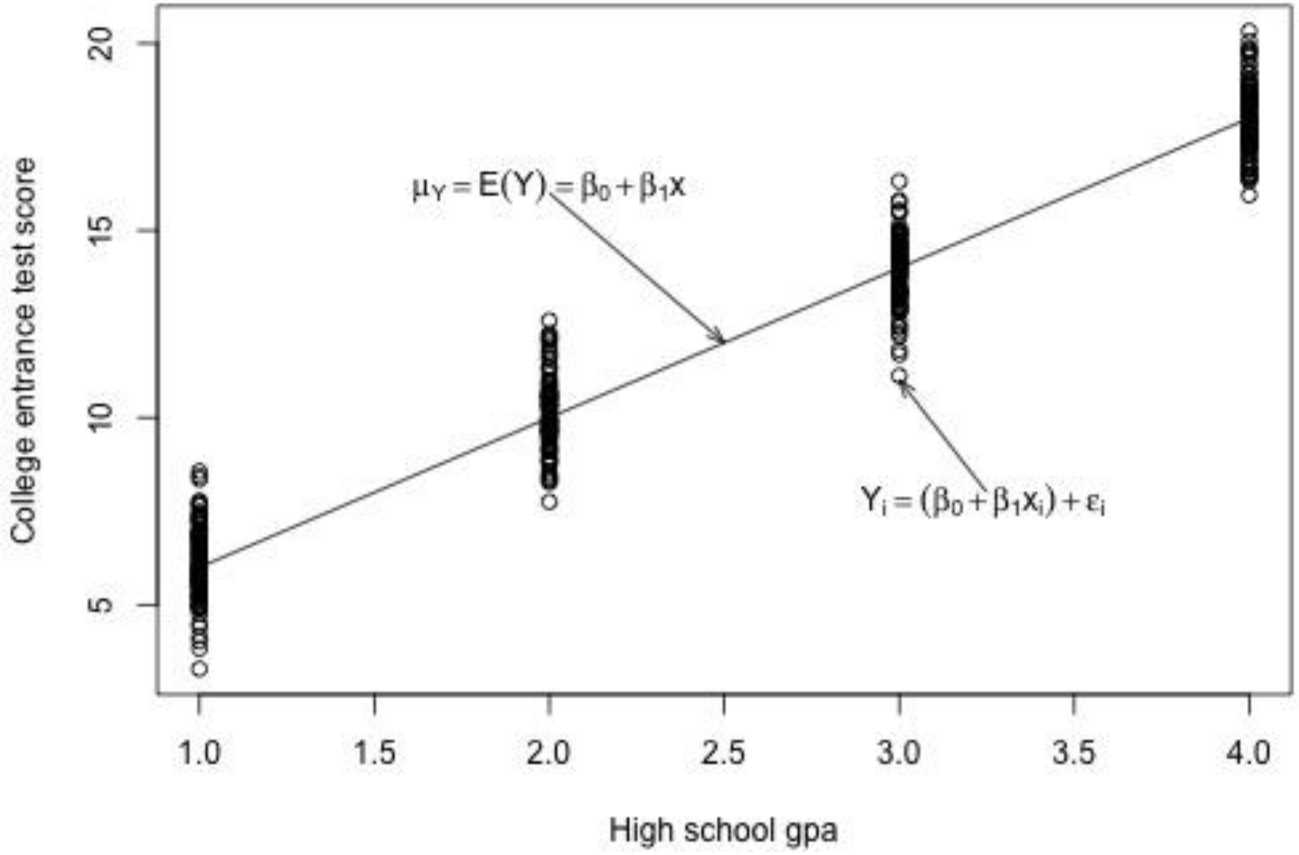
- The mean of the response, $E(Y_i)$, at each value of the predictor, x_i, is a linear function of the x_i.

4. Assumptions: Independent



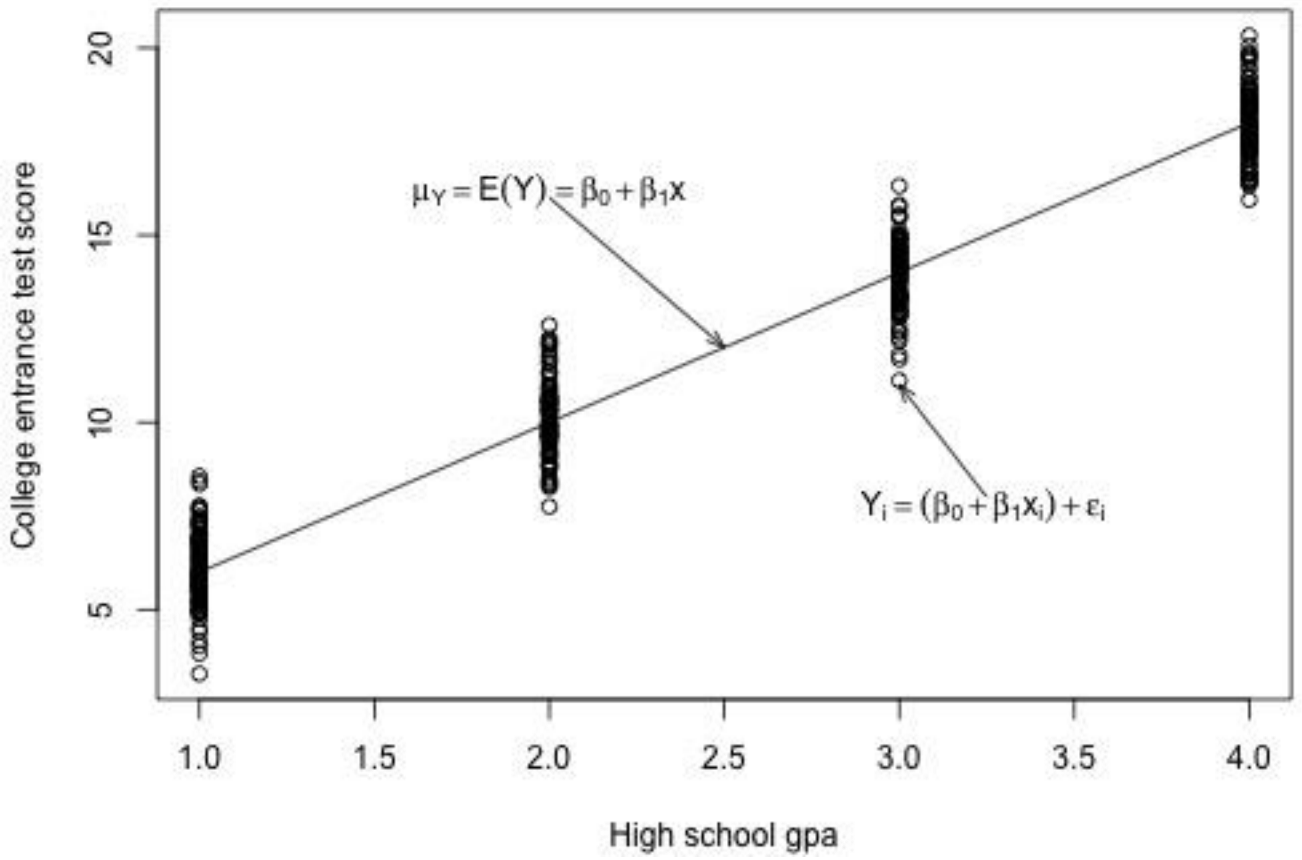
- The errors, ϵ_i, are independent if a random sample from the population is taken.
- $\epsilon_i = y_i - E(Y_i)$

5. Assumptions: Normally Distributed



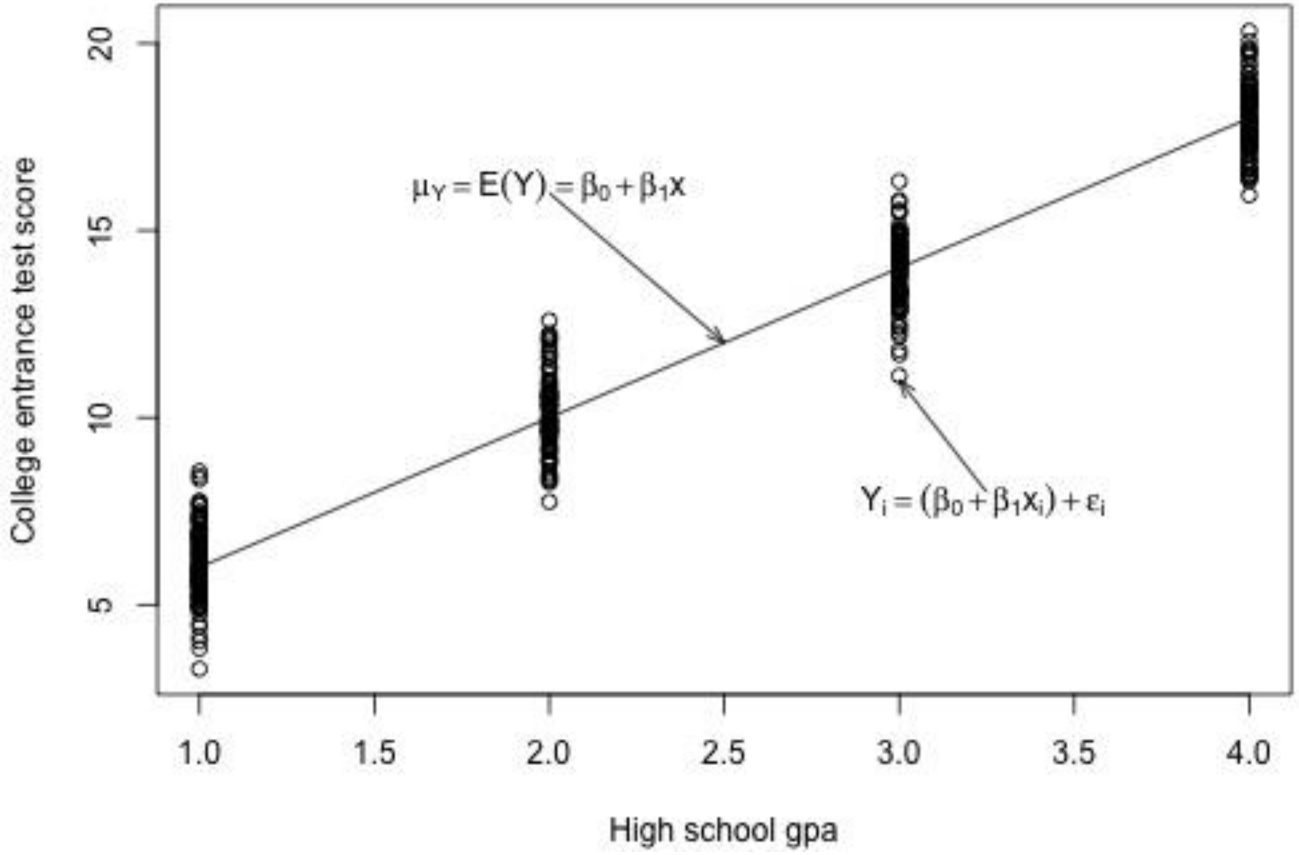
- The errors, ϵ_i, at each value of the predictor, x_i, are normally distributed.

6. Assumptions: Equal Variances



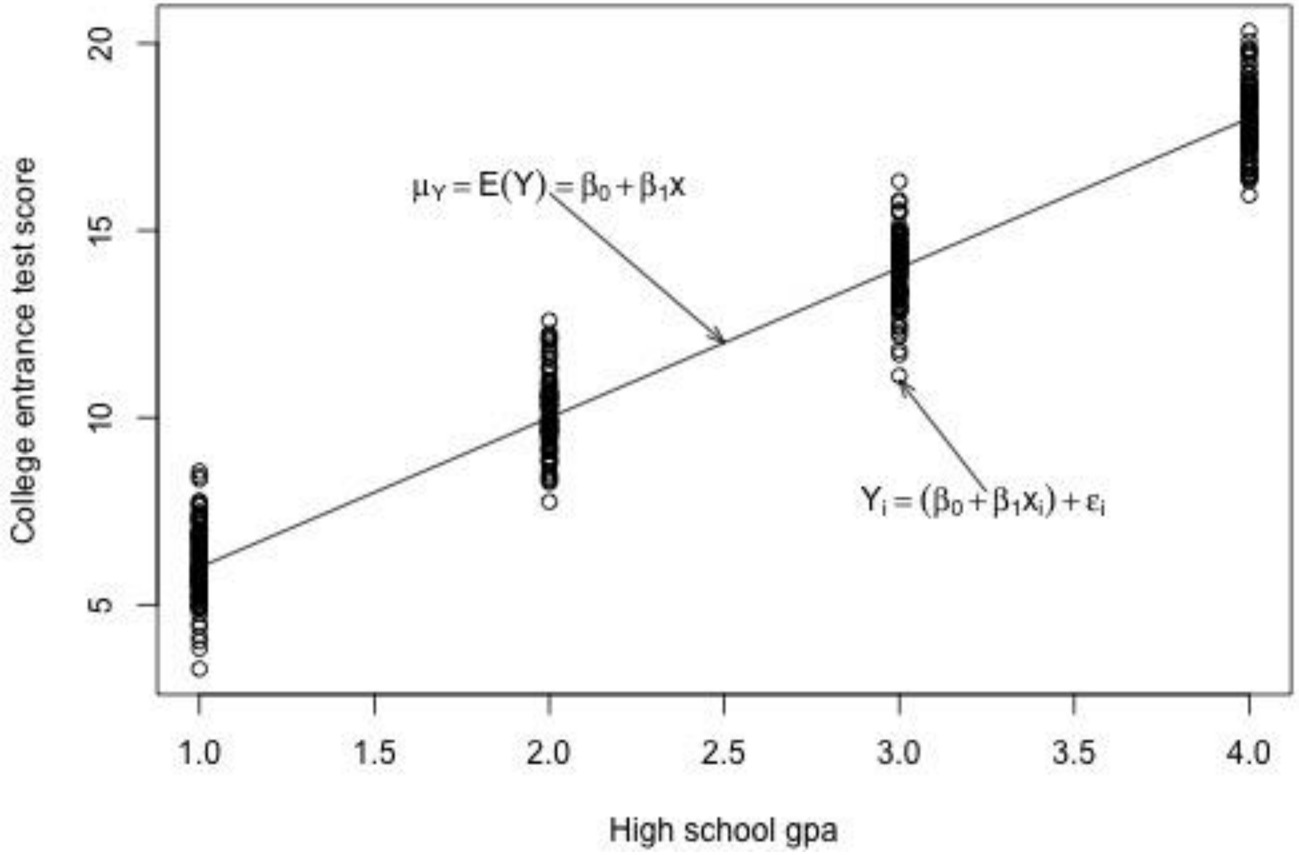
- The errors, ϵ_i, at each value of the predictor, x_i, have equal variances (denoted σ^2).

7. Assumptions: LINE



- Linear Function
- Independent
- Normally Distributed
- Equal Variances

8. Assumptions: Summary



- An alternative way to describe all four assumptions is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2.

Next

Chapter 2
SLR Model Evaluation