Chapter 3

SLR Estimation & Prediction

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

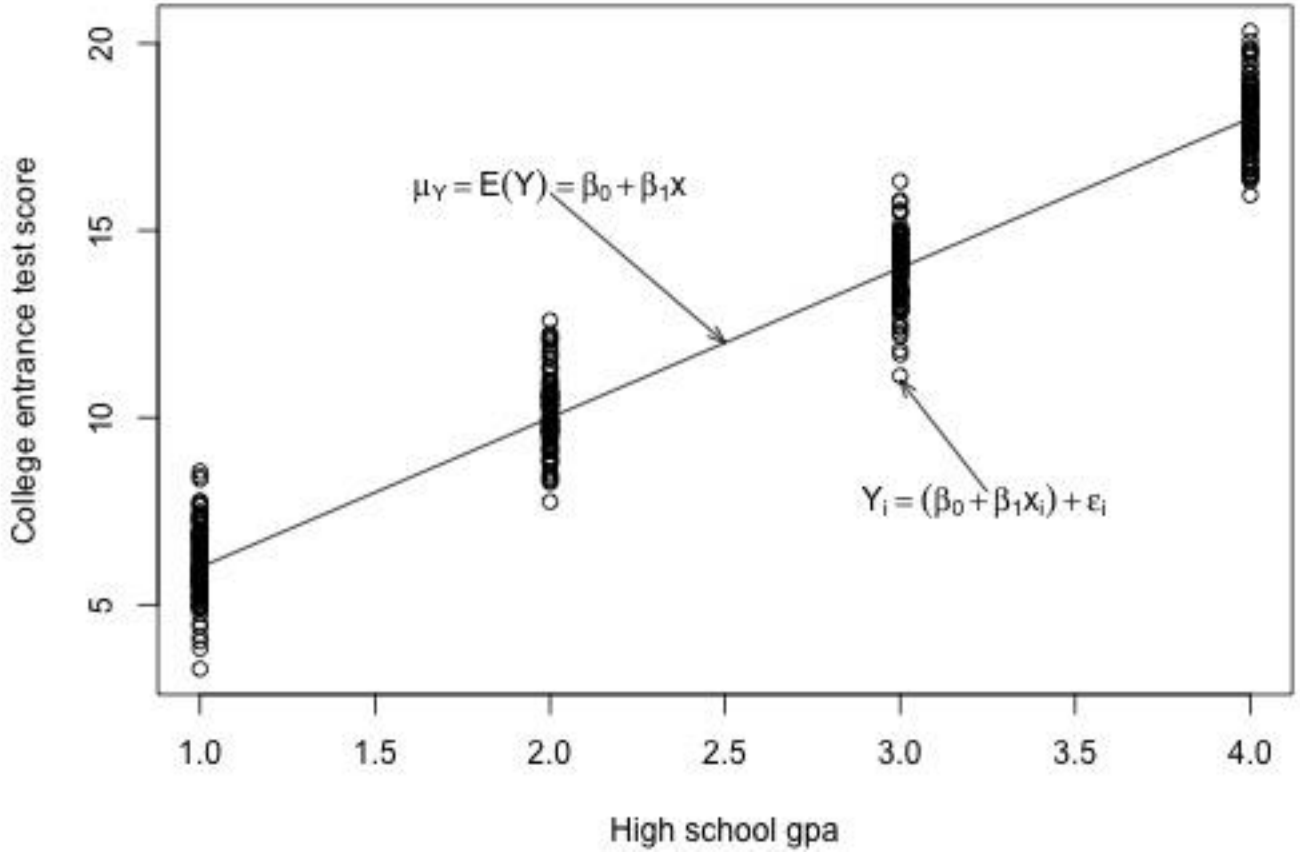
Contents

1. The Research Questions

1. The Research Questions

- What is the mean college entrance test score for the subpopulation of students whose high school gpa is 3? (Answering this question entails estimating the mean response μ_Y when $x = 3$.)
- What college entrance test score can we predict for a student whose high school gpa is 3? (Answering this question entails predicting the response y_{new} when $x = 3$.)

1. The Research Questions



1. The Research Questions

- What is the mean college entrance test score for the subpopulation of students whose high school gpa is 3
  - confidence interval for μ_Y
- What college entrance test score can we predict for a student whose high school gpa is 3?
  - prediction interval for y_{new}

2. Confidence Interval for the Mean Response

1. Confidence Interval for the Mean Response

- $100(1-\alpha)$ percent confidence interval for μ_Y
  - sample estimate $\pm$ (t−multiplier × standard error)
  - $\hat{y}_h \pm t_{\left(\frac{\alpha}{2},n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$

2. Factors affecting the confidence interval for μ_Y

- $2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$
  - As the mean square error (MSE) decreases, the width of the interval decreases.
  - As we decrease the confidence level, the t−multiplier decreases, and hence the width of the interval decreases.

2. Factors affecting the confidence interval for μ_Y

- $2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$
  - As we increase the sample size n, the width of the interval decreases.
  - The more spread out the predictor values, the larger the quantity $\sum(x_i - \bar{x})^2$ and hence the narrower the interval.

2. Factors affecting the confidence interval for μ_Y

- $2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$
  - As we increase the sample size n, the width of the interval decreases.
  - The more spread out the predictor values, the larger the quantity $\sum (x_i - \bar{x})^2$ and hence the narrower the interval.

2. Factors affecting the confidence interval for μ_Y

- $2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$
  - The closer x_h is to the average of the sample's predictor values $\bar{x}$, the smaller the quantity $(x_h - \bar{x})^2$, and hence the narrower the interval.

3. When is it okay to use the formula for the confidence interval for μ_Y?

- When x_h is a value within the range of the x values in the data set — that is, when is a value within the "scope of the model."
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. The formula works okay even if the error terms are only approximately normal.

3. Prediction Interval for a New Response

1. Prediction Interval for a New Response

- $100(1-\alpha)$ percent confidence interval for y_{new}

  - $\hat{y}_h \pm t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$

2. Understanding the difference in the two formulas

- $100(1-\alpha)$ percent confidence interval for μ_Y
  - $\hat{y}_h \pm t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$
- $100(1-\alpha)$ percent confidence interval for y_{new}
  - $\hat{y}_h \pm t_{\left(\frac{\alpha}{2}, n-2\right)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$

2. Understanding the difference in the two formulas

- If we know population parameters μ_Y and σ^2 at specific x_h, and y is normally distributed, it says that 95% of the measurements are in the interval sandwiched by $\mu_Y - 2\sigma$ and $\mu_Y + 2\sigma$

2. Understanding the difference in the two formulas

- The mean μ_Y is typically not known. The logical thing to do is estimate it with the predicted response $\hat{y}$. The cost of using $\hat{y}$ to estimate μ_Y is the variance of $\hat{y}$. That is, different samples would yield different predictions $\hat{y}$, and so we have to take into account this variance of $\hat{y}$.
- The variance σ^2 is typically not known. The logical thing to do is to estimate it with MSE.

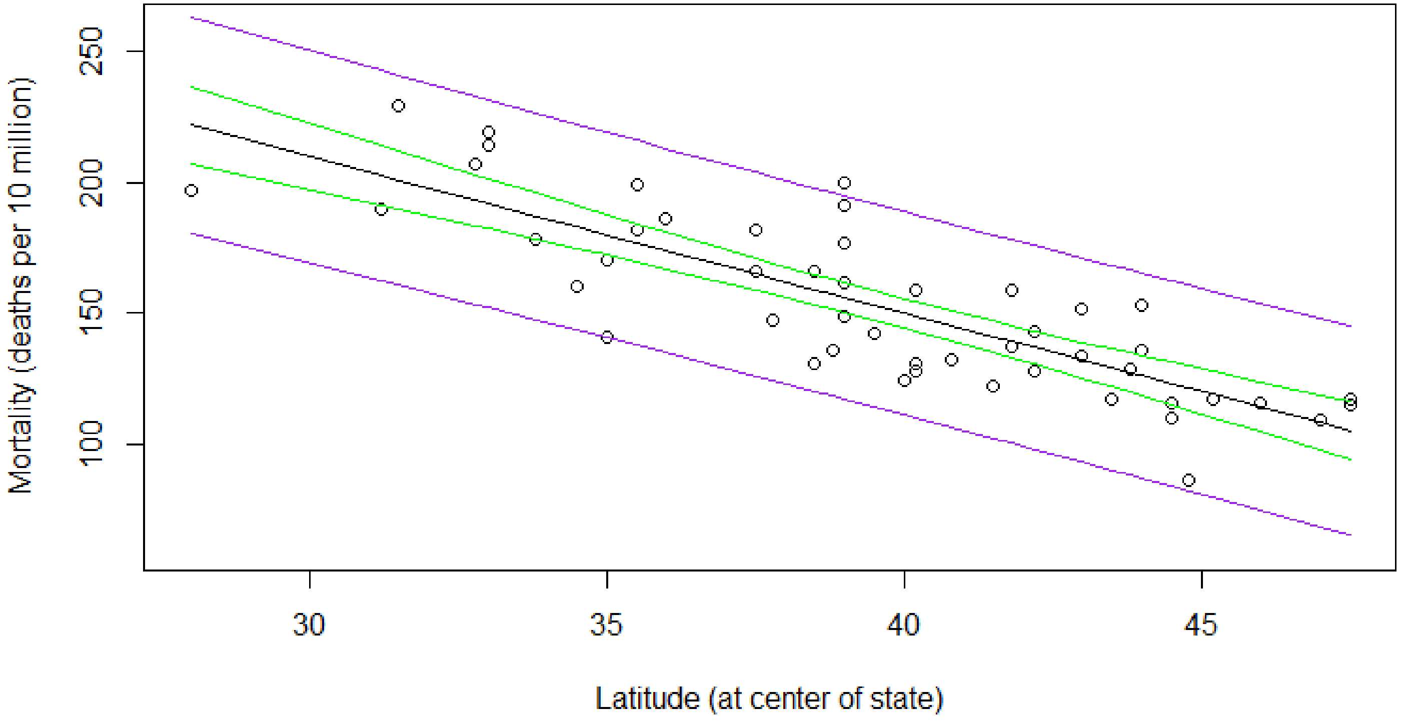
2. Understanding the difference in the two formulas

- The variation in the prediction of a new response depends on two components:
  - The variation due to estimating the mean μ_Y with $\hat{y}_h$, which we denote "$\sigma^2(\hat{Y}_h)$".
  - The variation in the responses y, which we denote as " σ^2"

2. Understanding the difference in the two formulas

- $\sigma^2 + \sigma^2(\hat{Y}_h)$
  - $MSE + MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right) = MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$

3. Confidence Interval and Prediction Interval



4. Code: Confidence Interval

```
skincancer <- read.table("skincancer.txt", header=T)
attach(skincancer)
model <- lm(Mort ~ Lat)
predict(model, interval="confidence", se.fit=T, level = 0.95,
        newdata=data.frame(Lat=40))
predict(model, interval="prediction", level = 0.95,
        newdata=data.frame(Lat=40))
detach(skincancer)
```

5. Results: Confidence Interval for μ_Y

```
> predict(model, interval="confidence", se.fit=T, level = 0.95,
+         newdata=data.frame(Lat=40))
$fit
       fit      lwr      upr
1 150.0839 144.5617 155.6061

$se.fit
[1] 2.745

$df
[1] 47
```

6. Results: Prediction Interval for y_{new}

```
> predict(model, interval="prediction", level = 0.95,
+         newdata=data.frame(Lat=40))
       fit     lwr      upr
1 150.0839 111.235 188.9329
```

7. When is it okay to use the formula for the prediction interval for y_{new}?

- When x_h is a value within the scope of the model. Again, x_h does not have to be one of the actual x values in the data set.
- When the "LINE" conditions — linearity, independent errors, normal errors, equal error variances — are met. Unlike the case for the formula for the confidence interval, the formula for the prediction interval depends **strongly** on the condition that the error terms are normally distributed.

Next

Chapter 4
SLR Model Assumptions