Chapter 2

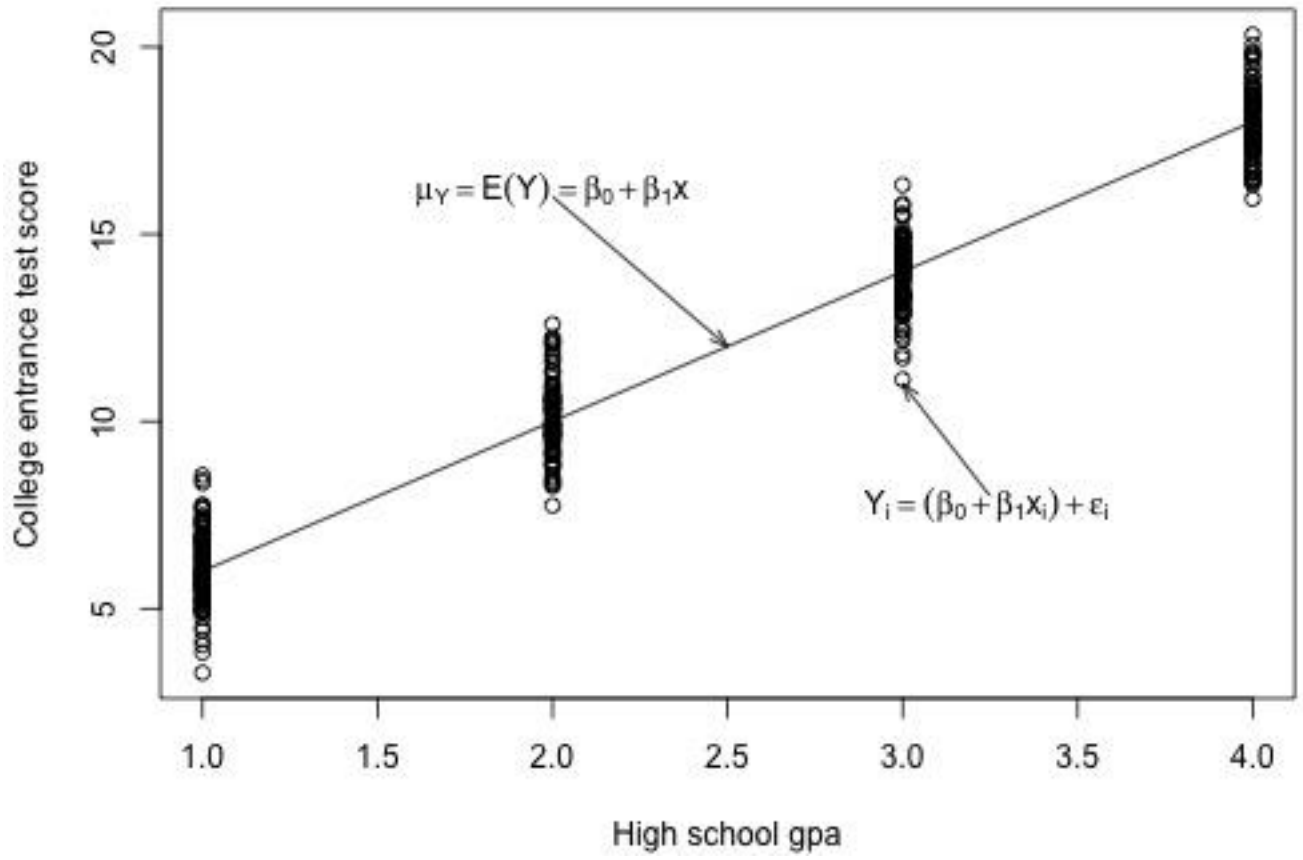
SLR Model Evaluation

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

Contents

1. Common Error Variance

1. Error Variance



- σ^2 quantifies how much the responses (y) vary around the (unknown) mean population regression line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$.

2. Estimation of Error Variance

- We don't know σ^2 because it is a population parameter.
- sample variance
  - $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$
  - Estimation of μ with $\bar{y}$ costs one degree of freedom(df).

2. Estimation of Error Variance

- Mean Square Error(MSE)
  - Estimates of σ^2
  - $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$
  - Estimation of two population parameters(β_0 and β_1) costs two degrees of freedom.

3. Code: MSE

```
> print(model$df.residual)
[1] 8
> print(sum(model$residuals^2))
[1] 597.386
> print(sprintf("MSE=%0.2f",
sum(model$residuals^2)/model$df.residual))
[1] "MSE=74.67"
```

3. Code: MSE

```
> anova(model)
Analysis of Variance Table
Response: wt
          Df Sum Sq Mean Sq F value    Pr(>F)
ht         1 5202.2  5202.2  69.666 3.214e-05 ***
Residuals  8  597.4    74.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. Coefficient of Determination

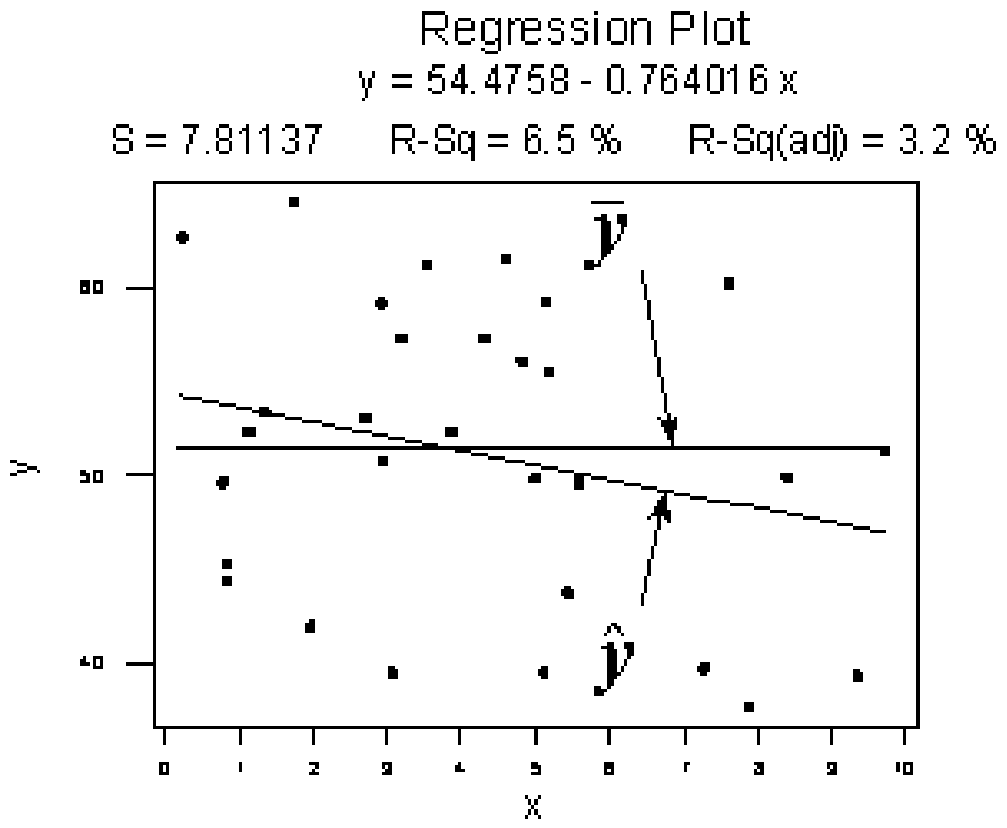
1. Sum of Squares

- SSR(regression sum of squares) quantifies how far the estimated sloped regression line, $\hat{y}_i$, is from the horizontal "no relationship line," the sample mean or $\bar{y}$.
  - $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- SSE(error sum of squares) quantifies how much the data points, y_i, vary around the estimated regression line, $\hat{y}_i$.
  - $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

1. Sum of Squares

- SSTO(total sum of squares) quantifies how much the data points, y_i, vary around their mean, $\bar{y}$.
  - $SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SSTO = SSR + SSE$
  - $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

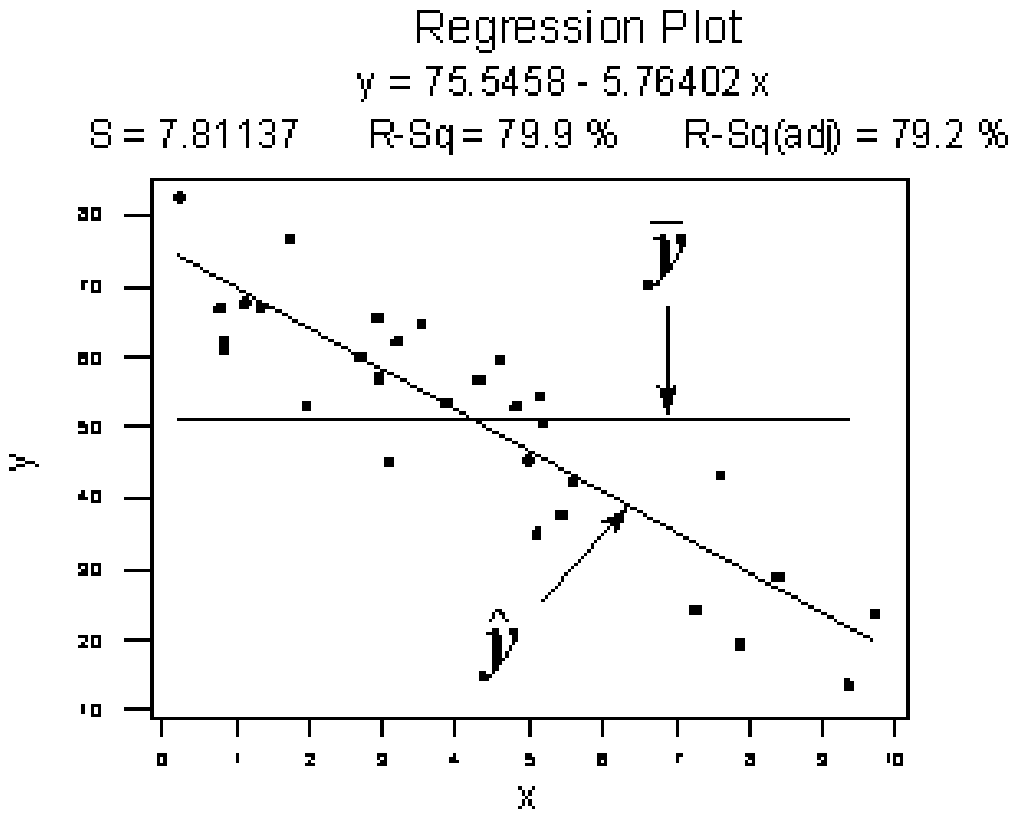
2. Situation 1



- $SSR = 119.1$
- $SSE = 1708.5$
- $SSTO = 1827.6$

$$\frac{SSR}{SSTO} = \frac{119.1}{1827.6} = 0.065$$

3. Situation 2



- $SSR = 6679.3$
- $SSE = 1708.5$
- $SSTO = 8487.8$

$$\frac{SSR}{SSTO} = \frac{6679.3}{8487.8} = 0.799$$

4. r^2

- The "coefficient of determination" or "r-squared value," denoted r^2, is the regression sum of squares divided by the total sum of squares.
  - $r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$

5. Interpretation of r^2

- $r^2 \times 100$ percent of the variation in y is 'explained by' the variation in predictor x.
  - Caution: Large r-squared value does not imply that x causes the changes in y.
  - Causation is different from association.
    - e.g. ice cream and shark

5. Interpretation of r^2

- What's considered a large r-squared value?
  - It depends on the research area.

6. Results: Height and Weight

```
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

Residual standard error: 8.641 on 8 degrees of freedom
Multiple R-squared:  0.897,   Adjusted R-squared:  0.8841
```

3. Inference for the Parameter

1. Confidence Interval for β_1

- $100(1-\alpha)$ percent confidence interval for β_1
  - sample estimate $\pm$ (t-multiplier $\times$ standard error)
  - $b_1 \pm t_{\left(\frac{\alpha}{2}, n-2\right)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i-\bar{x})^2}}\right)$
    - $n-2$: degree of freedom(df)

1. Confidence Interval for β_1

- If the confidence interval for β_1 contains 0, then we conclude that there is no evidence of a linear relationship between the predictor x and the response y in the population.
- If the confidence interval for β_1 does not contain 0, then we conclude that there is evidence of a linear relationship between the predictor x and the response y in the population.

2. Hypothesis Test for β_1

- Null hypothesis $H_0: \beta_1 = \beta$
- Alternative hypothesis $H_A: \beta_1 \neq \beta$
- $\beta = 0$

2. Hypothesis Test for β_1

- $t^* = \frac{b_1 - \beta}{se(b_1)} = \frac{b_1 - \beta}{\left(\frac{\sqrt{MSE}}{\sqrt{\Sigma(x_i - \bar{x})^2}}\right)}$

2. Hypothesis Test for β_1

- We use the resulting test statistic to calculate the P-value. As always, the P-value is the answer to the question "how likely is it that we'd get a test statistic t* as extreme as we did if the null hypothesis were true?" The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.

2. Hypothesis Test for β_1

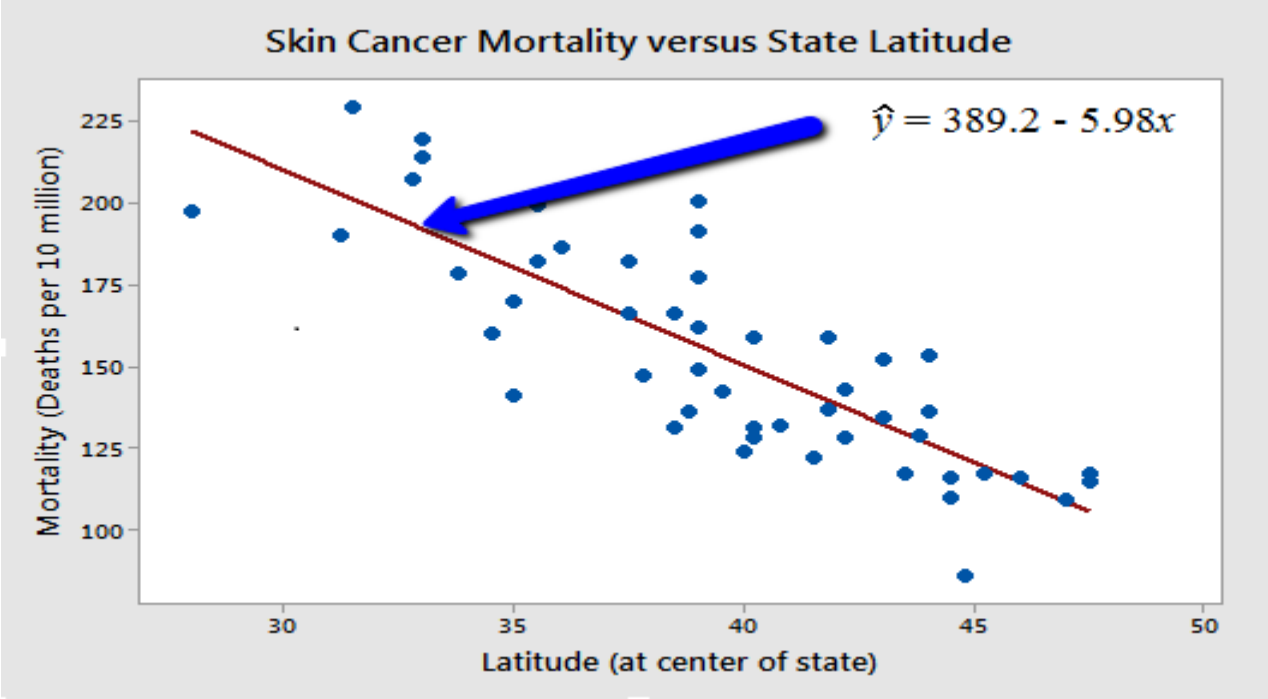
- If the P-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y."

2. Hypothesis Test for β_1

- If the P-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor x and response y."

3. Dataset

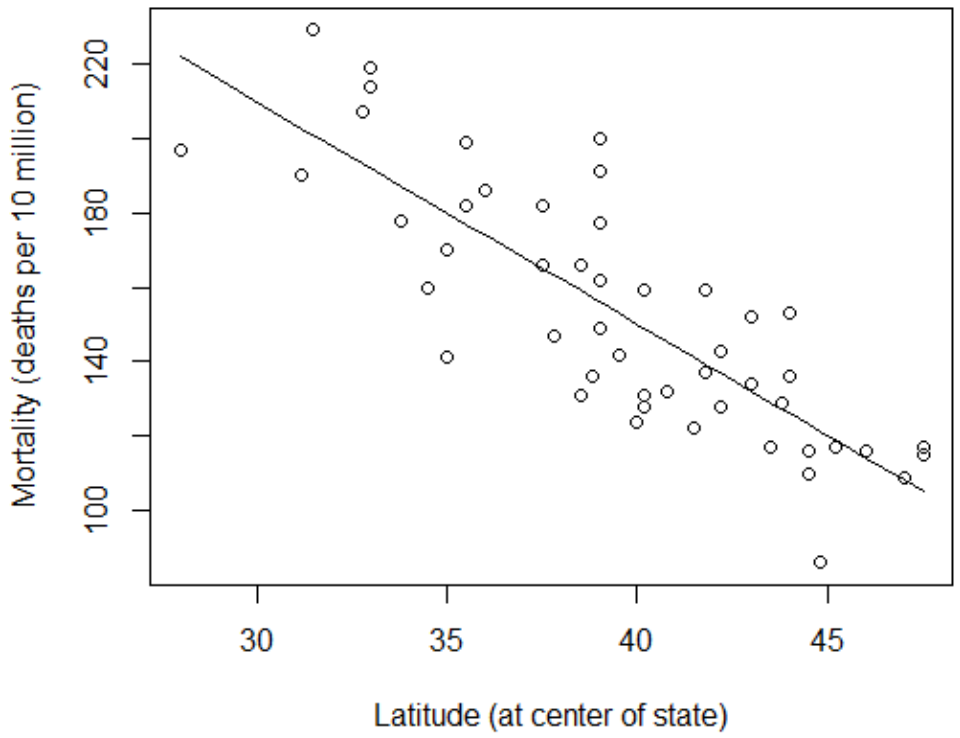
- Skin Cancer Dataset



4. Code: Skin Cancer

```
skincancer <- read.table("skincancer.txt", header=T)
attach(skincancer)
model <- lm(Mort ~ Lat)
plot(x=Lat, y=Mort,
     xlab="Latitude (at center of state)", ylab="Mortality
(deaths per 10 million)",
     panel.last = lines(sort(Lat), fitted(model)[order(Lat)]))
summary(model)
confint(model, level=0.95)
detach(skincancer)
```

5. Results: Skin Cancer



5. Results: Skin Cancer

```
> summary(model)

Call:
lm(formula = Mort ~ Lat)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5. Results: Skin Cancer

```
> confint(model, level=0.95)
                2.5 %     97.5 %
(Intercept) 341.285151 437.093552
Lat          -7.181404  -4.773867
```

6. Factors affecting the confidence interval for β_1

- $width = 2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$
- As the confidence level decreases, the width of the interval decreases.
- As MSE decreases, the width of the interval decreases.

6. Factors affecting the confidence interval for β_1

- $width = 2 \times t_{\left(\frac{\alpha}{2}, n-2\right)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$
- The more spread out the predictor x values, the narrower the interval.
- As the sample size increases, the width of the interval decreases.

7. Possible outcomes concerning β_1

- When we don't reject the null hypothesis,
  - We committed a Type II error. That is, in reality $\beta_1 \neq 0$ and our sample data just didn't provide enough evidence to conclude that $\beta_1 \neq 0$.
  - There really is not much of a linear relationship between x and y.
  - There is a relationship between x and y — it is just not linear.

7. Possible outcomes concerning β_1

- When we do reject the null hypothesis,
  - We committed a Type I error. That is, in reality $\beta_1 = 0$, but we have an unusual sample that suggests that $\beta_1 \neq 0$.
  - The relationship between x and y is indeed linear.
  - A linear function fits the data okay, but a curved ("curvilinear") function would fit the data even better.

Next

Chapter 3
SLR Estimation & Prediction