# R Lab for Regression Analysis

*Young-geun Kim*
*Department of Statistics, SKKU*
*dudrms33@g.skku.edu*

*25 Mar, 2019*

# Contents

# Chapter 1

# Linear Regression Analysis

```r
data(BioOxyDemand, package = "MPV")
(BioOxyDemand <-
  BioOxyDemand %>%
  tbl_df())
```

```
# A tibble: 14 x 2
       x     y
   <int> <int>
 1     3     4
 2     8     7
 3    10     8
 4    11     8
 5    13    10
 6    16    11
 7    27    16
 8    30    26
 9    35    21
10    37     9
11    38    31
12    44    30
13   103    75
14   142    90
```

## 1.1 Relation

We wonder how `x` affects `y`, especially linearly.

- Functional relation: mathematical equation,

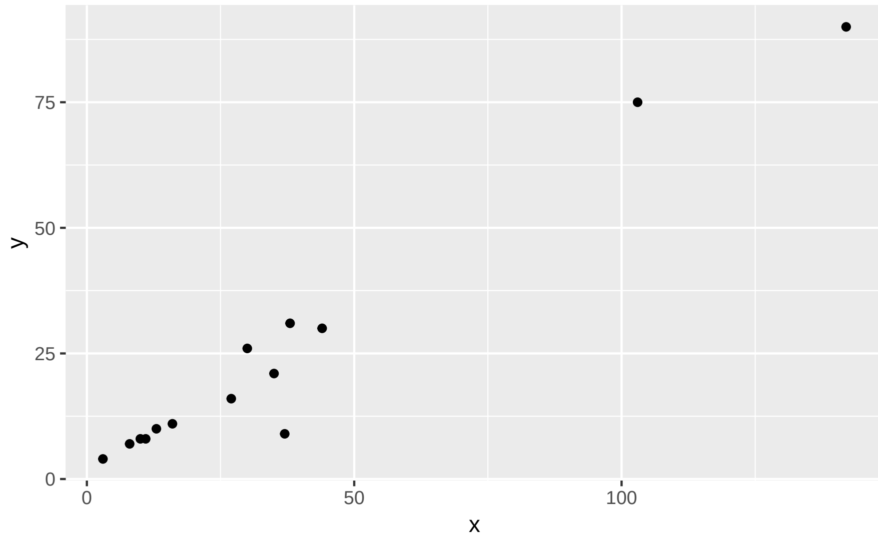
$$y = \beta_0 + \beta_1 x$$

- Statistical relation: embeded with noise

So we try to estimate

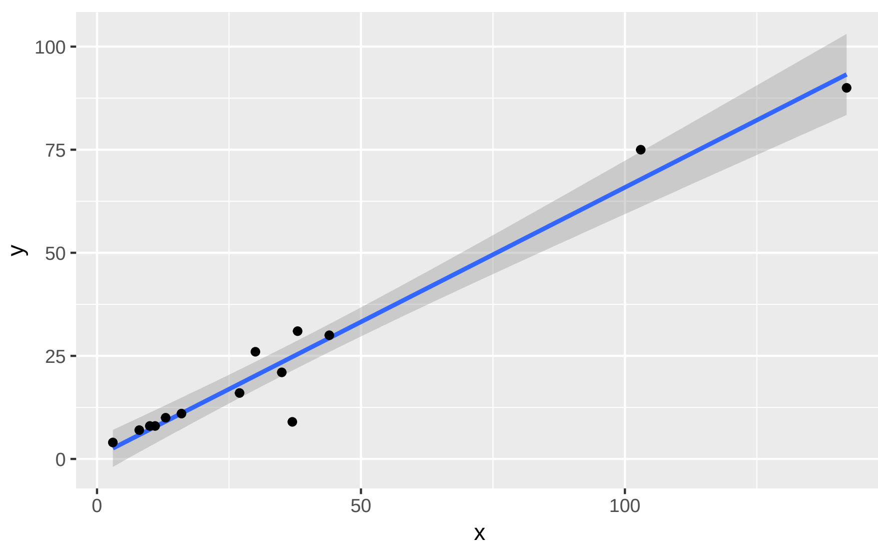
$$y = \beta_0 + \beta_1 x + \epsilon$$

```
BioOxyDemand %>%
  ggplot(aes(x, y)) +
  geom_point()
```




Looking just with the eyes, we can see the linear relationship. Regression analysis estimates the relationship statistically.

```
BioOxyDemand %>%
  ggplot(aes(x, y)) +
  geom_smooth(method = "lm") +
  geom_point()
```



# Chapter 2

# Simple Linear Regression

## 2.1 Model

```r
delv <- MPV::p2.9 %>% tbl_df()
```

```r
delv %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  labs(
    x = "Number of Cases",
    y = "Delivery Time"
  )
```

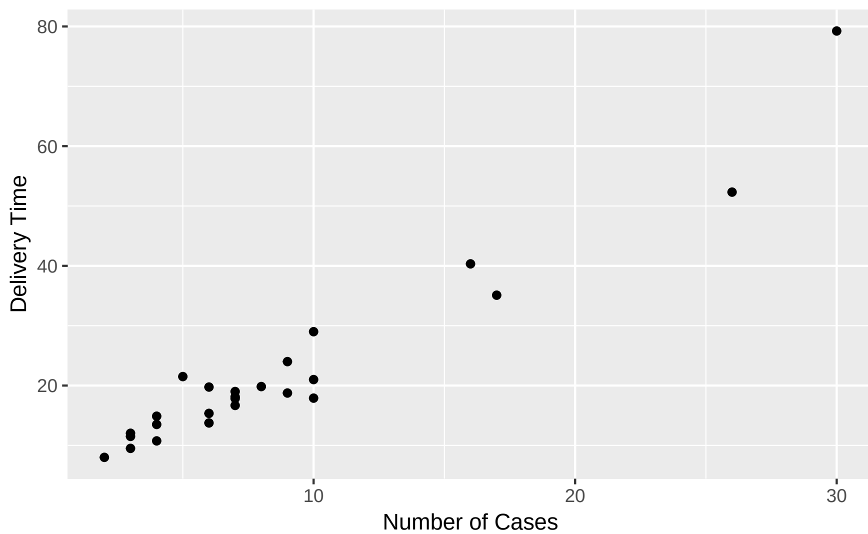



Figure 2.1: The Delivery Time Data

Given data $(x_1, Y_1), \ldots, (x_n, Y_n)$, we try to fit linear model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here $\epsilon_i$ is a error term, which is a random variable.

$$\epsilon \overset{iid}{\sim} (0, \sigma^2)$$

It gives the problem of estimating three parameters $(\beta_0, \beta_1, \sigma^2)$. Before estimating these, we set some assumptions.

1. linear relationship
2. $\epsilon_i$s are independent
3. $\epsilon_i$s are identically destributed, i.e. *constant variance*
4. In some setting, $\epsilon_i \sim N$

## 2.2 Least Squares Estimation

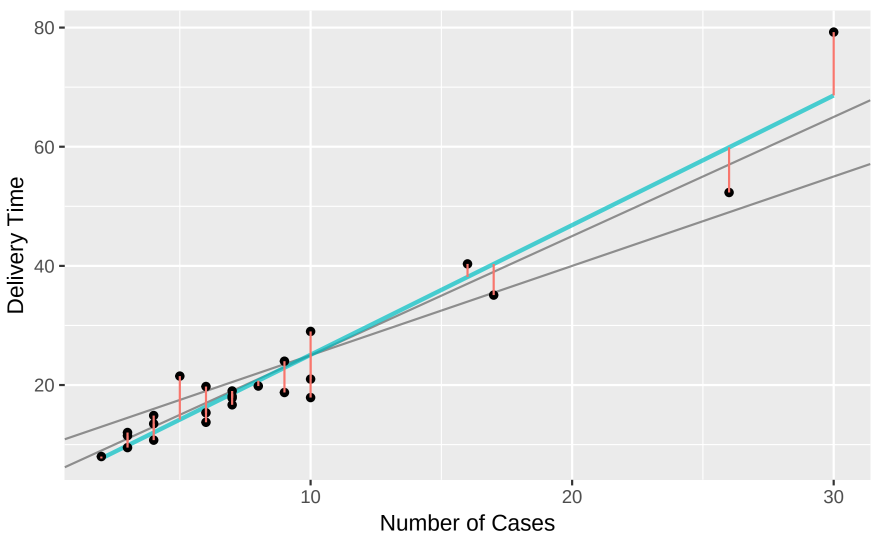



Figure 2.2: Idea of the least square estimation

We try to find $\beta_0$ and $\beta_1$ that minimize the sum of squares of the vertical distances, i.e.

$$(\beta_0, \beta_1) = \arg\min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.1}$$

### 2.2.1 Normal equations

Denote that Equation (2.1) is quadratic. Then we can find its minimum by find the zero point of the first derivative. Set

$$Q(\beta_0, \beta_1) := \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Then we have

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.2}$$

and

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)x_i = 0 \tag{2.3}$$

From (2.2),

$$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

Thus,

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

(2.3) gives

$$\sum_{i=1}^{n} x_i(Y_i - \overline{Y} + \hat{\beta}_1\overline{x} - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} x_i(Y_i - \overline{Y}) - \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \overline{x}) = 0$$

Thus,

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n} x_i(x_i - \overline{x})}$$

*Remark.*

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

where $S_{XX} := \sum\limits_{i=1}^{n}(x_i - \overline{x})^2$ and $S_{XY} := \sum\limits_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})$

*Proof.* Note that $\overline{x}^2 = \frac{1}{n^2}\left(\sum\limits_{i=1}^{n} x_i\right)^2$. Then we have

$$\begin{aligned} S_{XX} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\overline{x} + \sum_{i=1}^{n}\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \end{aligned} \tag{2.4}$$

It follows that

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum x_i(Y_i - \overline{Y})}{\sum x_i(x_i - \overline{x})} \\
&= \frac{\sum x_i(Y_i - \overline{Y}) - \overline{x}\sum(Y_i - \overline{Y})}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \qquad \because \sum(Y_i - \overline{Y}) = 0 \\
&= \frac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\
&= \frac{S_{XY}}{S_{XX}}
\end{aligned}
$$

□

```r
lm(y ~ x, data = delv)
```

```

Call:
lm(formula = y ~ x, data = delv)

Coefficients:
(Intercept)            x
       3.32         2.18
```

### 2.2.2 Prediction and Mean response

> "Essentially, all models are wrong, but some are useful."
>
> —George Box

Recall that we have assumed the **linear assumption** between the predictor and the response variables, i.e. the true model. Estimating $\beta_0$ and $\beta_1$ is same as estimating the *assumed true model*.

**Definition 2.1** (Mean response)**.**

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

We can estimate this mean resonse by

$$\widehat{E(Y \mid x)} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.5}$$

However, in practice, the model might not be true, which is included in $\epsilon$ term.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Our real problem is predicting individual $Y$, not the mean. The *prediction* of response can be done by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.6}$$

Observe that the values of Equation (2.5) and (2.6) are same. However, due to the **error term in the prediction**, it has larger standard error.

### 2.2.3 Properties of LSE

Parameters $\beta_0$ and $\beta_1$ have some properties related to the expectation and variance. We can notice that these lse's are **unbiased linear estimator**. In fact, these are the *best unbiased linear estimator*. This will be covered in the Gauss-Markov theorem.

**Lemma 2.1.**

$$S_{XX} = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n x_i(x_i - \overline{x})$$

$$S_{XY} = \sum_{i=1}^n x_i Y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n Y_i(x_i - \overline{x})$$

*Proof.* We already proven the first part of $S_{XX}$. See the Equation (2.4). The second part is tivial. Since $\sum(x_i - \overline{x}) = 0$,

$$S_{XX} = \sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i - \overline{x})x_i$$

For the first part of $S_{XY}$,

$$\begin{aligned} S_{XY} &= \sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) \\ &= \sum_{i=1}^n x_i Y_i - \overline{x}\sum_{i=1}^n Y_i - \overline{Y}\sum_{i=1}^n x_i + n\overline{x}\overline{Y} \\ &= \sum_{i=1}^n x_i Y_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n Y_i\right) \end{aligned}$$

Second part of $S_{XY}$ also can be proven from the definition.

$$\begin{aligned} S_{XY} &= \sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y}) \\ &= \sum_{i=1}^n Y_i(x_i - \overline{x}) - \overline{Y}\sum_{i=1}^n (x_i - \overline{x}) \\ &= \sum_{i=1}^n Y_i(x_i - \overline{x}) \qquad \because \sum_{i=1}^n (x_i - \overline{x}) = 0 \end{aligned}$$

□

**Lemma 2.2** (Linearity)**.** *Each coefficient is a linear estimator.*

$$\hat{\beta}_1 = \sum_{i=1}^n \frac{(x_i - \overline{x})}{S_{XX}} Y_i$$

$$\hat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \frac{(x_i - \overline{x})}{S_{XX}}\right) Y_i$$

*Proof.* From lemma 2.1,

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} \\
&= \frac{1}{S_{XX}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i
\end{aligned}$$

It gives that

$$\begin{aligned}
\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{x} \\
&= \frac{1}{n} \sum_{i=1}^{n} Y_i - \overline{x} \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{S_{XX}} Y_i \\
&= \sum_{i=1}^{n} \left( \frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{XX}} \right) Y_i
\end{aligned}$$

□

**Proposition 2.1** (Unbiasedness)**.** *Both coefficients are unbiased.*

*(a)* $E\hat{\beta}_1 = \beta_1$

*(b)* $E\hat{\beta}_0 = \beta_0$

From the model, $Y_1, \ldots, Y_n \overset{indep}{\sim} (\beta_0 + \beta_1 x_i, \sigma^2)$.

*Proof.* From lemma 2.1,

$$\begin{aligned}
E\hat{\beta}_1 &= \sum_{i=1}^{n} \left[ \frac{(x_i - \overline{x})}{S_{XX}} E(Y_i) \right] \\
&= \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{S_{XX}} (\beta_0 + \beta_1 x_i) \\
&= \frac{\beta_1 \sum (x_i - \overline{x}) x_i}{\sum (x_i - \overline{x}) x_i} \qquad \because \sum (x_i - \overline{x}) = 0 \\
&= \beta_1
\end{aligned}$$

It follows that

$$\begin{aligned}
E\hat{\beta}_0 &= E(\overline{Y} - \hat{\beta}_1 \overline{x}) \\
&= E(\overline{Y}) - \overline{x} E(\hat{\beta}_1) \\
&= E(\beta_0 + \beta_1 \overline{x} + \overline{\epsilon}) - \beta_1 \overline{x} \\
&= \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} \\
&= \beta_0
\end{aligned}$$

□

**Proposition 2.2** (Variances)**.** *Variances and covariance of coefficients*

*(a)* $Var\hat{\beta}_1 = \frac{\sigma^2}{S_{XX}}$

*(b)* $Var\hat{\beta}_0 = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2$

*(c)* $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}}{S_{XX}}\sigma^2$

*Proof.* Proving is just arithmetic.

(a)

$$
\begin{aligned}
Var\hat{\beta}_1 &= \frac{1}{S_{XX}^2}\sum_{i=1}^n\Big[(x_i-\overline{x})^2Var(Y_i)\Big] + \frac{1}{S_{XX}^2}\sum_{j\neq k}^n\Big[(x_j-\overline{x})(x_k-\overline{x})Cov(Y_j,Y_k)\Big] \\
&= \frac{\sigma^2}{S_{XX}} \qquad \because Cov(Y_j,Y_k) = 0 \text{ if } j \neq k
\end{aligned}
$$

(b)

$$
\begin{aligned}
Var\hat{\beta}_0 &= \sum_{i=1}^n\left(\frac{1}{n} - \frac{(x_i-\overline{x})\overline{x}}{S_{XX}}\right)^2 Var(Y_i) + \sum_{j\neq k}\left(\frac{1}{n} - \frac{(x_j-\overline{x})\overline{x}}{S_{XX}}\right)\left(\frac{1}{n} - \frac{(x_k-\overline{x})\overline{x}}{S_{XX}}\right)Cov(Y_j,Y_k) \\
&= \frac{\sigma^2}{n} - 2\sigma^2\frac{\overline{x}}{S_{XX}}\sum_{i=1}^n(x_i-\overline{x}) + \frac{\sigma^2\overline{x}^2\sum(x_i-\overline{x})^2}{S_{XX}^2} \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2 \qquad \because \sum(x_i-\overline{x}) = 0
\end{aligned}
$$

(c)

$$
\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\overline{Y} - \hat{\beta}_1\overline{x}, \hat{\beta}_1) \\
&= -\overline{x}Var\hat{\beta}_1 \\
&= -\frac{\overline{x}}{S_{XX}}\sigma^2
\end{aligned}
$$

□

### 2.2.4 Gauss-Markov Theorem

Chapter 2.2.3 shows that the $\beta_0^{LSE}$ and $\beta_1^{LSE}$ are the **linear unbiased estimators**. Are these good? Good compared to *what estimators*? Here we consider *linear unbiased estimator*. If variances in the proposition 2.2 are lower than any parameters in this parameter family, $\beta_0^{LSE}$ and $\beta_1^{LSE}$ are the **best linear unbiased estimators**.

**Theorem 2.1** (Gauss Markov Theorem)**.** *$\hat{\beta}_0$ and $\hat{\beta}_1$ are BLUE, i.e. the best linear unbiased estimator.*

$$
Var(\hat{\beta}_0) \leq Var\Big(\sum_{i=1}^n a_iY_i\Big) \; \forall a_i \in \mathbb{R} \; s.t. \; E\Big(\sum_{i=1}^n a_iY_i\Big) = \beta_0
$$

$$
Var(\hat{\beta}_1) \leq Var\Big(\sum_{i=1}^n b_iY_i\Big) \; \forall b_i \in \mathbb{R} \; s.t. \; E\Big(\sum_{i=1}^n b_iY_i\Big) = \beta_1
$$

## 2.3 Maximum Likelihood Estimation

In this section, we add an assumption to an random errors $\epsilon_i$.

$$\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

**Example 2.1** (Gaussian Likelihood). Note that $Y_i \stackrel{indep}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Then the likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( - \frac{(Y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) \right)$$

and so the log-likelihood function can be computed as

$$l(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

### 2.3.1 Likelihood equations

**Definition 2.2** (Maximum Likelihood Estimator).

$$(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}, \hat{\sigma}^{2MLE}) := \arg\sup L(\beta_0, \beta_1, \sigma^2)$$

Since $l(\cdot) = \ln L(\cdot)$ is monotone,

*Remark.*

$$(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}, \hat{\sigma}^{2MLE}) = \arg\sup l(\beta_0, \beta_1, \sigma^2)$$

We can find the maximum of this *quadratic* function by making first derivative.

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.7}$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.8}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{2.9}$$

Denote that Equations (2.7) and (2.8) given $\hat{\sigma}^2$ are equivalent to the normal equations. Thus,

$$\hat{\beta}_0^{MLE} = \hat{\beta}_0^{LSE}, \quad \hat{\beta}_1^{MLE} = \hat{\beta}_1^{LSE}$$

From (2.9),

$$\hat{\sigma}^{2MLE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 = \frac{n-2}{n} \hat{\sigma}^{2LSE}$$

Recall that $\hat{\sigma}^{2LSE}$ is an unbiased, i.e. this *MLE is not an unbiased estimator*. Since $\hat{\sigma}^{2MLE} \approx \hat{\sigma}^{2LSE}$ for large $n$, howerver, it is *asymptotically unbiased*.

**Theorem 2.2** (Rao-Cramer Lower Bound, univariate case)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$. *If* $\hat{\theta}$ *is an unbiased estimator of* $\theta$*,*

$$Var(\hat{\theta}) \ge \frac{1}{I_n(\theta)}$$

*where* $I_n(\theta) = -E\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right)$

To apply this theorem @(thm:rclb) in the simple linear regression setting, i.e. $(\beta_0, \beta_1)$, we need to look at the *bivariate case.*

**Theorem 2.3** (Rao-Cramer Lower Bound, bivariate case)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta 1, \theta_2)$ *and let* $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. *If each* $\hat{\theta}_1$, $\hat{\theta}_2$ *is an unbiased estimator of* $\theta_1$ *and* $\theta_2$*, then*

$$Var(\boldsymbol{\theta}) := \begin{bmatrix} Var(\hat{\theta}_1) & Cov(\hat{\theta}_1, \hat{\theta}_2) \\ Cov(\hat{\theta}_1, \hat{\theta}_2) & Var(\hat{\theta}_2) \end{bmatrix} \ge I_n^{-1}(\theta_1, \theta_2)$$

*where*

$$I_n(\theta_1, \theta_2) = -\begin{bmatrix} E\left(\frac{\partial^2 l(\theta_1,\theta_2)}{\partial \theta_1^2}\right) & E\left(\frac{\partial^2 l(\theta_1,\theta_2)}{\partial \theta_1 \partial \theta_2}\right) \\ E\left(\frac{\partial^2 l(\theta_1,\theta_2)}{\partial \theta_1 \partial \theta_2}\right) & E\left(\frac{\partial^2 l(\theta_1,\theta_2)}{\partial \theta_2^2}\right) \end{bmatrix}$$

## 2.4 Residuals

**Definition 2.3** (Residuals)**.**

$$e_i := Y_i - \hat{Y}_i$$

### 2.4.1 Prediction error

```
delv %>%
  mutate(yhat = predict(lm(y ~ x))) %>%
  ggplot(aes(x = x, y = y)) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point() +
  geom_linerange(aes(ymin = y, ymax = yhat), col = I("red"), alpha = .7) +
  labs(
    x = "Number of Cases",
    y = "Delivery Time"
  )
```

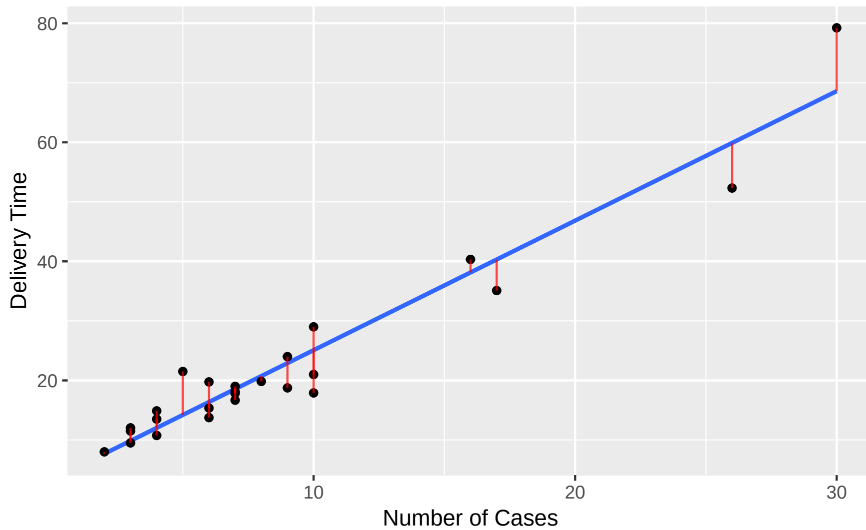



Figure 2.3: Fit and residuals

See Figure 2.3. Each red line is $e_i$. As we can see, $e_i$ represents the difference between *observed* response and *predicted* response. A large $|e_i|$ indicates a large prediction error. You can call this $e_i$ for each $Y_i$ by `lm()$residuals` or `residuals()`.

```
delv_fit <- lm(y ~ x, data = delv)
delv_fit$residuals
```

```
     1      2      3      4      5      6      7      8      9     10
-1.874  1.651  2.181  2.855 -2.628 -0.444  0.327 -0.724 10.634  7.298
    11     12     13     14     15     16     17     18     19     20
 2.191 -4.082  1.475  3.372  1.094  3.918 -1.028  0.446 -0.349 -5.216
    21     22     23     24     25
-7.182 -7.581 -4.156 -0.900 -1.275
```

$\sum e_i^2$, which has been minimized in the procedure of LSE, can be used to see *overall size of prediction errors.*

**Definition 2.4** (Error Sums of Squares)**.**

$$SSE := \sum_{i=1}^{n} e_i^2$$

### 2.4.2 Residuals and the variance

$e_i$ contains the information for $\epsilon_i$. $\sum e_i^2$ can give information about $\sigma^2 = Var(\epsilon_i)$. For this, it is expected that $e_i$ and $\epsilon_i$ have similar feature.

**Proposition 2.3** (Properties of residuals)**.** *Mean and variance of the residual*

*(a)* $E(e_i) = 0$

*(b)* $Var(e_i) \neq \sigma^2$

*(c)* $\forall i \neq j : Cov(e_i, e_j) \neq 0$

## 2.5 Decomposition of Total Variability

## 2.6 Geometric Interpretations

# References