R Lab for Regression Analysis

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

29 Mar, 2019

Contents

Chapter 1

Linear Regression Analysis

```
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,

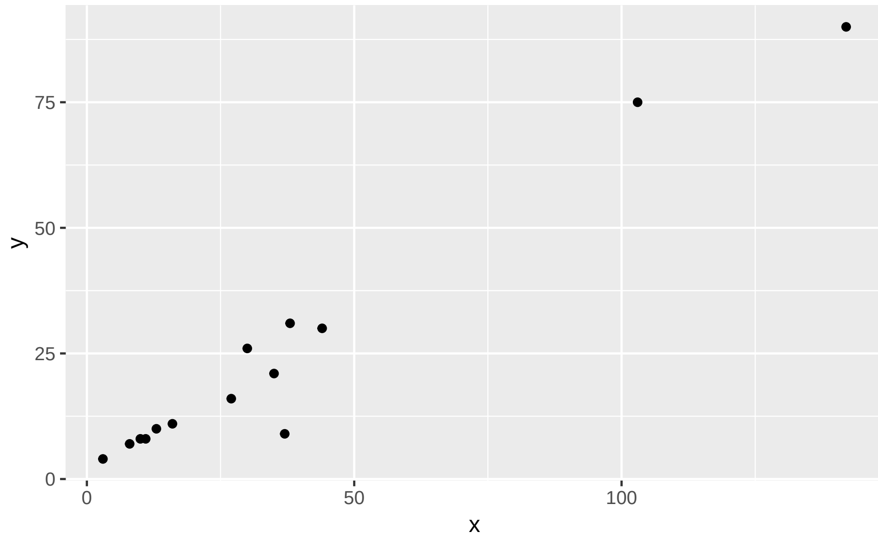
$$y = \beta_0 + \beta_1 x$$

- Statistical relation: embeded with noise

So we try to estimate

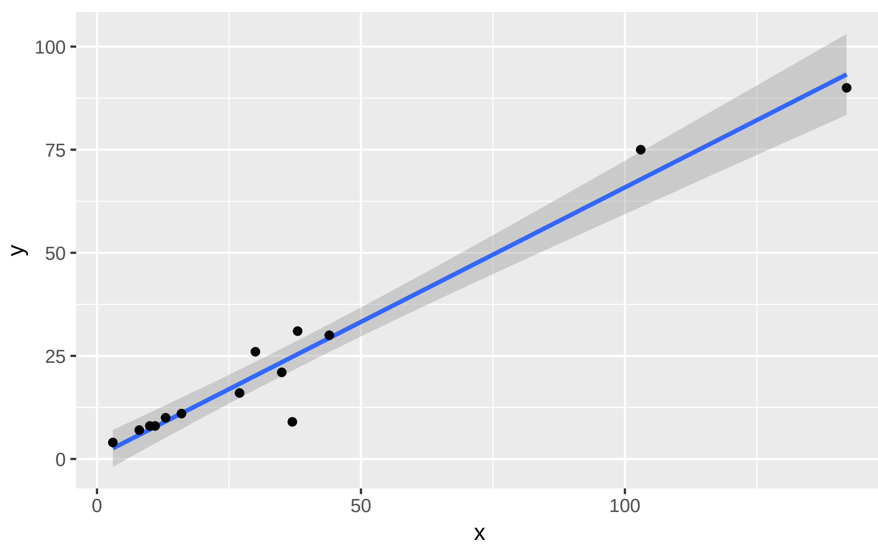
$$y = \beta_0 + \beta_1 x + \epsilon$$

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



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

```r
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"
  )
```

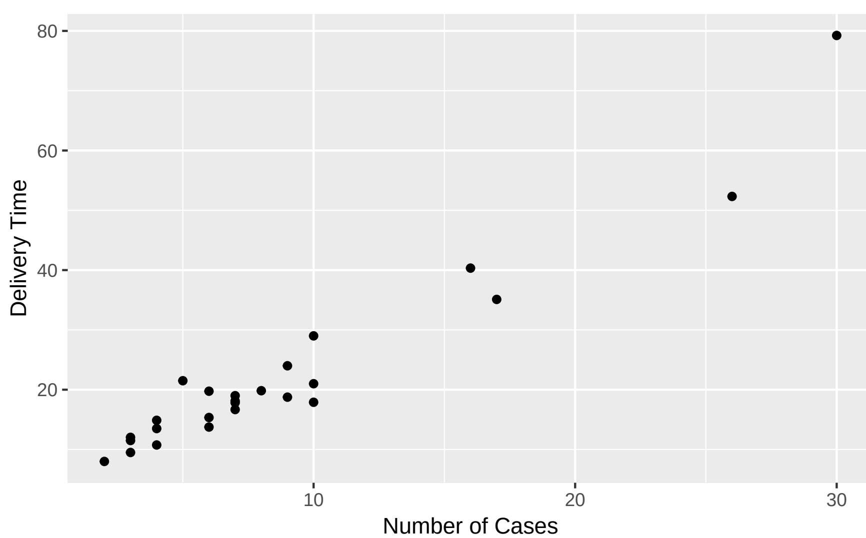


Figure 2.1: The Delivery Time Data

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 ϵ_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. ϵ_is are independent
3. ϵ_is are identically destributed, i.e. *constant variance*
4. In some setting, $\epsilon_i \sim N$

2.2 Least Squares Estimation

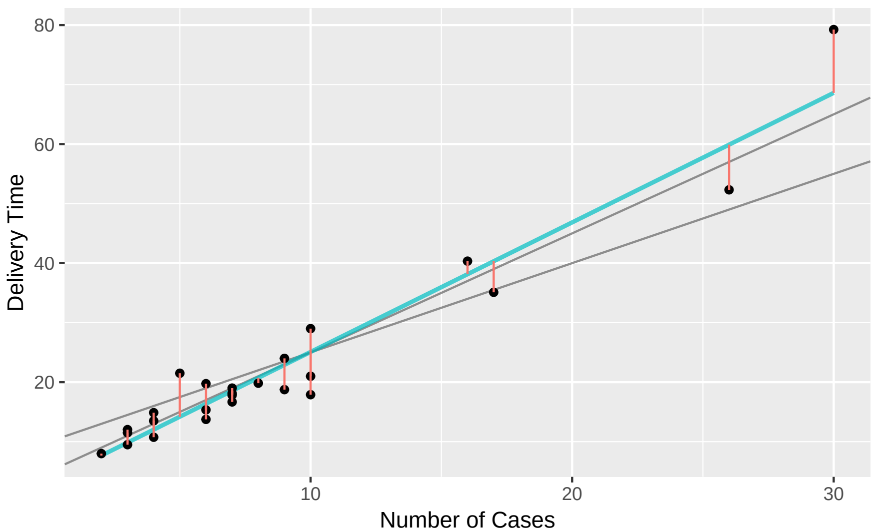


Figure 2.2: Idea of the least square estimation

We try to find β_0 and β_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,

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

Remark.

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

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

Proof. Note that $\overline{x}^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2$. 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 β_0 and β_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 ϵ 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 β_0 and β_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)

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

(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 β_0^{LSE} and β_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, β_0^{LSE} and β_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.*

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

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

Bestness of beta1. Consider $\Theta := \left\{ \sum_{i=1}^{n} b_i Y_i \in \mathbb{R} : E\Big(\sum_{i=1}^{n} b_i Y_i \Big) = \beta_1 \right\}$.

Claim: $Var(\sum b_i Y_i) - Var(\hat{\beta}_1) \geq 0$

Let $\sum b_i Y_i \in \Theta$. Then $E(\sum b_i Y_i) = \beta_1$.

Since $E(Y_i) = \beta_0 + \beta_1 x_i$,

$$\beta_0 \sum b_i + \beta_1 \sum b_i x_i = \beta_1$$

It gives

$$\begin{cases} \sum b_i = 0 \\ \sum b_i x_i = 1 \end{cases} \tag{2.7}$$

Then

$$\begin{aligned}
0 \leq Var\Big(\sum b_i Y_i - \hat{\beta}_1 \Big) &= Var\Big(\sum b_i Y_i - \sum \frac{(x_i - \bar{x})}{S_{XX}} Y_i \Big) \\
&\overset{indep}{=} \sum \left(b_i - \frac{(x_i - \bar{x})}{S_{XX}} \right)^2 \sigma^2 \\
&= \sum \left(b_i^2 - \frac{2b_i(x_i - \bar{x})}{S_{XX}} + \frac{(x_i - \bar{x})^2}{S_{XX}^2} \right) \sigma^2 \\
&= \sum b_i^2 \sigma^2 - \frac{2\sigma^2}{S_{XX}} \sum b_i x_i + \frac{2\bar{x}\sigma^2}{S_{XX}} \sum b_i + \sigma^2 \frac{\sum (x_i - \bar{x})^2}{S_{XX}^2} \\
&= \sum b_i^2 \sigma^2 - \frac{\sigma^2}{S_{XX}} \qquad \because (2.7) \text{ and } S_{XX} = \sum (x_i - \bar{x})^2 \\
&= Var(\sum b_i Y_i) - Var(\hat{\beta}_1)
\end{aligned}$$

Hence,

$$Var(\sum b_i Y_i) \geq Var(\hat{\beta}_1)$$

□

Bestness of beta0. Consider $\Theta := \left\{ \sum_{i=1}^{n} a_i Y_i \in \mathbb{R} : E\Big(\sum_{i=1}^{n} a_i Y_i \Big) = \beta_0 \right\}$.

Claim: $Var(\sum a_i Y_i) - Var(\hat{\beta}_0) \geq 0$

Let $\sum a_i Y_i \in \Theta$. Then $E(\sum a_i Y_i) = \beta_0$.

Since $E(Y_i) = \beta_0 + \beta_1 x_i$,

$$\beta_0 \sum a_i + \beta_1 \sum a_i x_i = \beta_0$$

It gives

$$\begin{cases} \sum a_i = 1 \\ \sum a_i x_i = 0 \end{cases} \tag{2.8}$$

Then

$$
\begin{aligned}
0 \le Var\Big(\sum a_i Y_i - \hat\beta_0\Big) &= Var\Big[\sum a_i Y_i - \sum\Big(\frac{1}{n} - \frac{(x_k-\bar x)\bar x}{S_{XX}}\Big)Y_k\Big] \\
&= \sum\left(a_i - \frac{1}{n} + \frac{(x_i-\bar x)\bar x}{S_{XX}}\right)^2\sigma^2 \\
&= \sum\left[a_i^2 - 2a_i\Big(\frac{1}{n} - \frac{(x_i-\bar x)\bar x}{S_{XX}}\Big) + \Big(\frac{1}{n} - \frac{(x_i-\bar x)\bar x}{S_{XX}}\Big)^2\right]\sigma^2 \\
&= \sum a_i^2\sigma^2 - \frac{2\sigma^2}{n}\sum a_i + \frac{2\bar x\sigma^2\sum a_i x_i}{S_{XX}} - \frac{2\bar x^2\sigma^2\sum a_i}{S_{XX}} \\
&\quad + \sigma^2\left(\frac{1}{n} - \frac{2\bar x}{nS_{XX}}\sum(x_i-\bar x) + \frac{\bar x^2\sum(x_i-\bar x)^2}{S_{XX}^2}\right) \\
&= \sum a_i^2\sigma^2 - \frac{2\sigma^2}{n} - \frac{2\bar x^2\sigma^2}{S_{XX}} \qquad \because (2.8) \\
&\quad + \left(\frac{1}{n} + \frac{\bar x^2}{S_{XX}}\right)\sigma^2 \qquad \because \sum(x_i-\bar x) = 0 \text{ and } S_{XX} := \sum(x_i-\bar x)^2 \\
&= \sum a_i^2\sigma^2 - \left(\frac{1}{n} + \frac{\bar x^2}{S_{XX}}\right)\sigma^2 \\
&= Var\Big(\sum a_i Y_i\Big) - Var\hat\beta_0
\end{aligned}
$$

Hence,

$$Var(\sum a_i Y_i) \ge Var(\hat\beta_0)$$

□

Example 2.1. Show that $\sum(Y_i - \hat Y_i) = 0$, $\sum x_i(Y_i - \hat Y_i) = 0$, and $\sum \hat Y_i(Y_i - \hat Y_i) = 0$.

Solution. Consider the two normal equations (2.2) and (2.3). Note that $\hat Y_i = \hat\beta_0 + \hat\beta_1 x_i$.

From the Equation (2.2), we have $\sum(Y_i - \hat Y_i) = 0$.

From the Equation (2.3), we have $\sum x_i(Y_i - \hat Y_i) = 0$.

It follows that

$$
\begin{aligned}
\sum \hat Y_i(Y_i - \hat Y_i) &= \sum(\hat\beta_0 + \hat\beta_1 x_i)(Y_i - \hat Y_i) \\
&= \hat\beta_0\sum(Y_i - \hat Y_i) + \hat\beta_1\sum x_i(Y_i - \hat Y_i) \\
&= 0
\end{aligned}
$$

2.2.5 Estimation of σ^2

There is the last parameter, $\sigma^2 = Var(Y_i)$. In the *least squares estimation literary*, we estimate σ^2 by

$$\hat\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \tag{2.9}$$

Why $n-2$? This makes the estimator unbiased.

Proposition 2.3 (Unbiasedness)**.**

$$E(\hat{\sigma}^2) = \sigma^2$$

Proof. Note that

$$(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = (Y_i - \overline{Y}) - \hat{\beta}_1(x_i - \overline{x})$$

Then

$$\begin{aligned}
E(\hat{\sigma}^2) &= \frac{1}{n-2} E\Big[\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\Big] \\
&= \frac{1}{n-2} E\Big[\sum (Y_i - \overline{Y})^2 + \hat{\beta}_1^2 \sum (x_i - \overline{x})^2 - 2\hat{\beta}_1 \sum (Y_i - \overline{Y})(x_i - \overline{x})\Big] \\
&= \frac{1}{n-2} E(S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1 S_{XY}) \\
&= \frac{1}{n-2} E(S_{YY} - \hat{\beta}_1^2 S_{XX}) \qquad \because S_{XY} = \hat{\beta}_1 S_{XX} \\
&= \frac{1}{n-2} \Big(\underbrace{ES_{YY}}_{(a)} - S_{XX} \underbrace{E\hat{\beta}_1^2}_{(b)}\Big)
\end{aligned}$$

(a)

$$\begin{aligned}
ES_{YY} &= E\Big[\sum (Y_i - \overline{Y})^2\Big] \\
&= E\Big[\sum \Big((\beta_0 + \beta_1 x_i + \epsilon_i) - (\beta_0 + \beta_1 \overline{x} + \overline{\epsilon})\Big)^2\Big] \\
&= E\Big[\sum \Big(\beta_1(x_i - \overline{x}) + (\epsilon_i - \overline{\epsilon})\Big)^2\Big] \\
&= \beta_1^2 S_{XX} + E\Big(\sum (\epsilon_i - \overline{\epsilon})^2\Big) + 2\beta_1 \sum (x_i - \overline{x}) E(\epsilon_i - \overline{\epsilon}) \\
&= \beta_1^2 S_{XX} + E\Big(\sum (\epsilon_i - \overline{\epsilon})^2\Big)
\end{aligned}$$

Since $E(\overline{\epsilon}) = 0$ and $Var(\overline{\epsilon}) = \frac{\sigma^2}{n}$,

$$\begin{aligned}
E\Big(\sum (\epsilon_i - \overline{\epsilon})^2\Big) &= E\Big(\sum (\epsilon_i^2 + \overline{\epsilon}^2 - 2\epsilon_i \overline{\epsilon})\Big) \\
&= \sum E(\epsilon_i^2) - nE(\overline{\epsilon}^2) \qquad \because \sum \epsilon = n\overline{\epsilon} \\
&= \sum (Var(\epsilon_i) + E(\epsilon_i)^2) - n(Var(\overline{\epsilon}) + E(\overline{\epsilon})^2) \\
&= n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Thus,

$$ES_{YY} = \beta_1^2 S_{XX} + (n-1)\sigma^2$$

(b)

$$
\begin{aligned}
E\hat{\beta}_1^2 &= Var\hat{\beta}_1 + E(\hat{\beta}_1)^2 \\
&= \frac{\sigma^2}{S_{XX}} + \beta_1^2
\end{aligned}
$$

It follows that

$$
\begin{aligned}
E(\hat{\sigma}^2) &= \frac{1}{n-2}\Big(\underbrace{ES_{YY}}_{(a)} - S_{YY}\underbrace{E\hat{\beta}_1^2}_{(b)}\Big) \\
&= \frac{1}{n-2}\bigg(\Big(\beta_1^2 S_{XX} + (n-1)\sigma^2\Big) - S_{XX}\Big(\frac{\sigma^2}{S_{XX}} + \beta_1^2\Big)\bigg) \\
&= \frac{1}{n-2}((n-2)\sigma^2) \\
&= \sigma^2
\end{aligned}
$$

□

2.3 Maximum Likelihood Estimation

In this section, we add an assumption to an random errors ϵ_i.

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

Example 2.2 (Gaussian Likelihood)**.** Note that $Y_i \overset{indep}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. 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.

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.10}$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{2.11}$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{2.12}$$

Denote that Equations (2.10) and (2.11) 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 (2.12),

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

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 $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta)$. If $\hat{\theta}$ is an unbiased estimator of θ,*

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

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

To apply this theorem 2.2 in the simple linear regression setting, i.e. (β_0, β_1), we need to look at the *bivariate case.*

Theorem 2.3 (Rao-Cramer Lower Bound, bivariate case)**.** *Let $X_1, \dots, X_n \overset{iid}{\sim} f(x;\theta 1, \theta_2)$ and let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. If each $\hat{\theta}_1$, $\hat{\theta}_2$ is an unbiased estimator of θ_1 and θ_2, then*

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

where

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

Assume that σ^2 is **known**. From the Equations (2.10) and (2.11),

$$\begin{cases} \frac{\partial^2 l}{\partial \beta_0^2} = -\frac{n}{\sigma^2} \\ \frac{\partial^2 l}{\partial \beta_1^2} = -\frac{\sum x_i^2}{\sigma^2} \\ \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = -\frac{\sum x_i}{\sigma^2} \end{cases}$$

Thus,

$$I_n(\beta_0, \beta_1) = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{\sum x_i}{\sigma^2} \\ \frac{\sum x_i}{\sigma^2} & \frac{\sum x_i^2}{\sigma^2} \end{bmatrix}$$

Applying gaussian elimination,

$$\begin{aligned}
\left[\begin{array}{cc|cc} \frac{n}{\sigma^2} & \frac{\sum x_i}{\sigma^2} & 1 & 0 \\ \frac{\sum x_i}{\sigma^2} & \frac{\sum x_i^2}{\sigma^2} & 0 & 1 \end{array}\right] &\leftrightarrow \left[\begin{array}{cc|cc} \frac{n}{\sigma^2} & \frac{\sum x_i}{\sigma^2} & 1 & 0 \\ \frac{\sum x_i}{\sigma^2}\left(\frac{n}{\sum x_i}\right) & \frac{\sum x_i^2}{\sigma^2}\left(\frac{n}{\sum x_i}\right) & 0 & \frac{1}{\overline{x}} \end{array}\right] \\
&\leftrightarrow \left[\begin{array}{cc|cc} \frac{n}{\sigma^2} & \frac{\sum x_i}{\sigma^2} & 1 & 0 \\ 0 & \frac{\sum x_i^2 - \overline{x}\sum x_i}{\sigma^2 \overline{x}} = \frac{S_{XX}}{\sigma^2 \overline{x}} & -1 & \frac{1}{\overline{x}} \end{array}\right] \\
&\leftrightarrow \left[\begin{array}{cc|cc} 1 & \overline{x} & \frac{\sigma^2}{n} & 0 \\ 0 & 1 & -\frac{\overline{x}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{array}\right] \\
&\leftrightarrow \left[\begin{array}{cc|cc} 1 & 0 & \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2 & -\frac{\overline{x}}{S_{XX}}\sigma^2 \\ 0 & 1 & -\frac{\overline{x}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{array}\right]
\end{aligned}$$

Hence,

$$I_n^{-1}(\beta_0, \beta_1) = \begin{bmatrix} \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2 & -\frac{\overline{x}}{S_{XX}}\sigma^2 \\ -\frac{\overline{x}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{bmatrix} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}$$

Since $Var(\hat{\boldsymbol{\beta}}) - I^{-1} = 0$ is non-negative definite, each $Var(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2$ and $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$ is a theoretical bound.

Remark. This says that $\hat{\beta}_0^{LSE} = \hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{LSE} = \hat{\beta}_1^{MLE}$ have the smallest variance among all unbiased estimator.

This result is *stronger than Gauss-Markov theorem* 2.1, where the LSE has the smalleset variance among all *linear unbiased* estimators. It can be simply obtained from the *Lehmann-Scheffe Theorem*: If some unbiased estimator is a function of complete sufficient statistic, then this estimator is the unique MVUE (Hogg et al., 2018).

Remark (Lehmann and Scheffe for regression coefficients). $u\left(\sum Y_i, S_{XY}\right)$ is CSS in this regression problem, i.e. known σ^2.

Proof. From the example 2.2,

$$\begin{aligned}
L(\beta_0, \beta_1) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum (Y_i - \beta_0 - \beta_1 x_i)^2\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum \left(Y_i^2 - (\beta_0 + \beta_1 x_i)Y_i + (\beta_0 + \beta_1 x_i)^2\right)\right] \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(-\beta_0 \sum Y_i - \beta_1 \sum x_i Y_i\right)\right] \exp\left[-\frac{1}{2\sigma^2}\left(\sum Y_i^2 + (\beta_0 + \beta_1 x_i)^2\right)\right]
\end{aligned}$$

By the Factorization theorem, both $\sum Y_i$ and $\sum x_i Y_i$ are sufficient statistics. Since S_{XY} is one-to-one function of $\sum x_i Y_i$, it is also a sufficient statistic.

Denote that the normal distribution is in exponential family.

Hence, $(\sum Y_i, S_{XY})$ are CSS. □

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"
  )
```

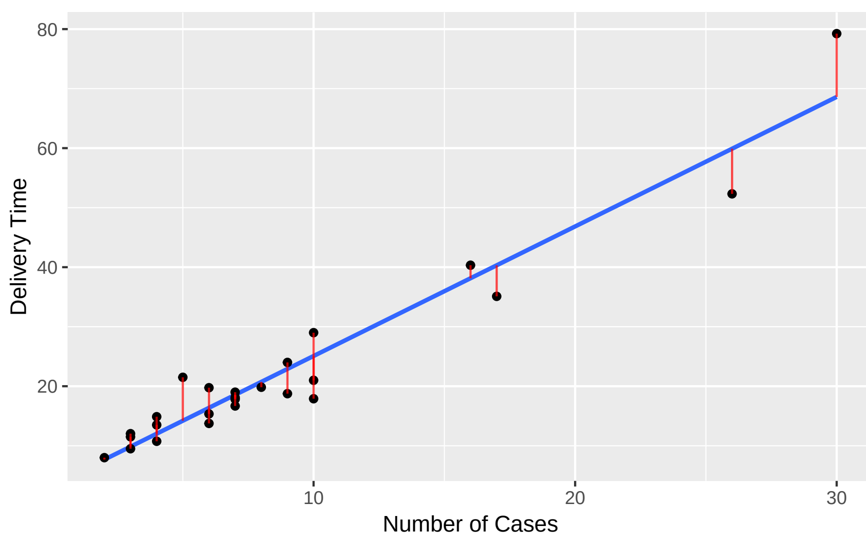


Figure 2.3: Fit and residuals

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

Lemma 2.3. *Covriance between* Y *and each coefficient*

(a) $Cov(\hat{\beta}_0, Y_i) = \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{XX}}\right)\sigma^2$

(b) $Cov(\hat{\beta}_1, Y_i) = \frac{(x_i - \overline{x})}{S_{XX}}\sigma^2$

Proof. (a)

$$\begin{aligned} Cov(\hat{\beta}_0, Y_i) &= Cov(\sum a_i Y_i, Y_i) \\ &= \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{XX}}\right)\sigma^2 \end{aligned}$$

(b)

$$\begin{aligned} Cov(\hat{\beta}_1, Y_i) &= Cov(\sum b_i Y_i, Y_i) \\ &= \frac{(x_i - \overline{x})}{S_{XX}}\sigma^2 \end{aligned}$$

□

Proposition 2.4 (Properties of residuals)**.** *Mean and variance of the residual*

(a) $E(e_i) = 0$

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

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

Proof. (a) Recall that this is the assumption of the regression model.

(b) Lemma 2.3 implies that

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

Then

$$\begin{aligned}
Var(\hat{Y}_i) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= Var\Big[\overline{Y} + (x_i - \overline{x})\hat{\beta}_1\Big] \qquad \because \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \\
&= Var(\overline{Y}) + (x_i - \overline{x})^2 Var(\hat{\beta}_1) + 2(x_i - \overline{x}) Cov(\overline{Y}, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + (x_i - \overline{x})^2 \frac{\sigma^2}{S_{XX}} + 0 \\
&= \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\right)\sigma^2
\end{aligned} \tag{2.13}$$

From the same lemma 2.3,

$$\begin{aligned}
Cov(Y_i, \hat{Y}_i) &= Cov(Y_i, \overline{Y} + (x_i - \overline{x})\hat{\beta}_1) \\
&= Cov(Y_i, \overline{Y}) + (x_i - \overline{x}) Cov(Y_i, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\sigma^2 \qquad \because Cov(Y_i, \hat{\beta}_1) = \frac{(x_i - \overline{x})}{S_{XX}}\sigma^2 \\
&= \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\right)\sigma^2
\end{aligned} \tag{2.14}$$

These (2.13) and (2.14) give that

$$\begin{aligned}
Var(e_i) &= Var(Y_i) + Var(\hat{Y}_i) - 2Cov(Y_i, \hat{Y}_i) \\
&= \sigma^2 + \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\right)\sigma^2 - 2\left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{XX}}\right)\sigma^2 \\
&= \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{XX}}\right)\sigma^2 \\
&\neq \sigma^2
\end{aligned}$$

(c) Let $i \neq j$. Then

$$\begin{aligned}
Cov(e_i, e_j) &= Cov\Big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), Y_j - (\hat{\beta}_0 + \hat{\beta}_1 x_j)\Big) \\
&= Cov(Y_i, Y_j) - Cov\Big(Y_i, (\hat{\beta}_0 + \hat{\beta}_1 x_j)\Big) - Cov((\hat{\beta}_0 + \hat{\beta}_1 x_i), Y_j) + Cov((\hat{\beta}_0 + \hat{\beta}_1 x_i), (\hat{\beta}_0 + \hat{\beta}_1 x_j)) \\
&= 0 - \left(\frac{1}{n} - \frac{(x_i - \overline{x})\overline{x}}{S_{XX}}\right)\sigma^2 - \frac{(x_i - \overline{x})x_j}{S_{XX}}\sigma^2 \\
&\quad - \left(\frac{1}{n} - \frac{(x_j - \overline{x})\overline{x}}{S_{XX}}\right)\sigma^2 - \frac{(x_i - \overline{x})x_i}{S_{XX}}\sigma^2 \\
&\quad + \left(\frac{1}{n} + \frac{\overline{x}^2 + x_i x_j - \overline{x}(x_i + x_j)}{S_{XX}}\right)\sigma^2 \\
&= -\left(\frac{1}{n} + \frac{\overline{x}^2 + x_i x_j - \overline{x}(x_i + x_j)}{S_{XX}}\right)\sigma^2 \\
&\neq 0
\end{aligned}$$

□

2.5 Decomposition of Total Variability

Definition 2.5 (Uncorrected Total Sum of Squares)**.**

$$SST_{uncor} := \sum_{i=1}^{n} Y_i^2$$

Definition 2.6 (Corrected Total Sum of Squares)**.**

$$SST := \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

What does this total sum of squares mean? To know this, we should know $\overline{Y}$ first.

```
delv %>%
  ggplot(aes(x = x, y = y)) +
  geom_smooth(method = "lm", formula = y ~ 1, se = FALSE) +
  geom_point() +
  labs(
    x = "Number of Cases",
    y = "Delivery Time"
  )
```

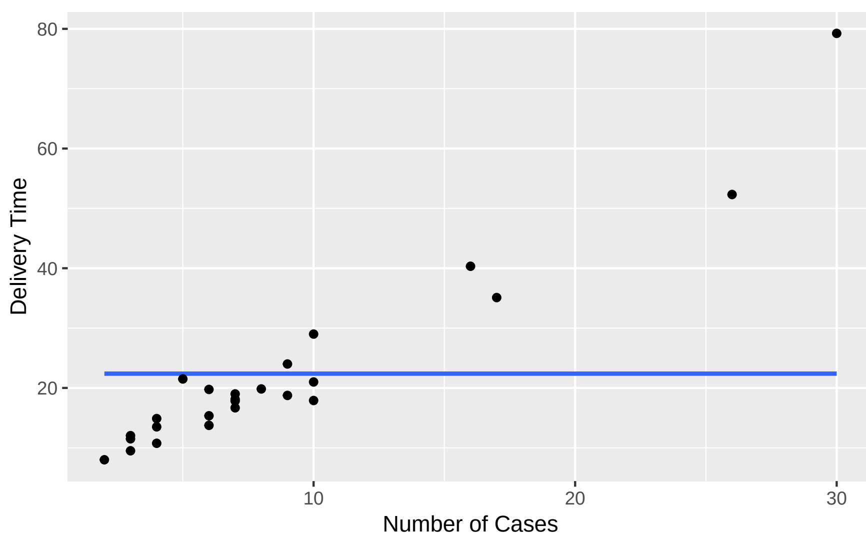


Figure 2.4: Regression without predictor

See Figure 2.4. The line represents the closest line when we use only intercept term for the regression model. In other words, *if we use no information for the response*, i.e. no predictor variables, we will get just average of the response variable. Consider

$$Y_i = \beta_0 + \epsilon_i$$

Then we can get only one normal equation

$$\sum (Y_i - \hat{\beta}_0) = 0$$

Hence,

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i \equiv \overline{Y}$$

From this fact, SST implies

Proposition 2.5 (Decomposition of SST)**.**

$$SST = SSR + SSE$$

where $SSR = \sum(\hat{Y}_i - \overline{Y})^2$ *and* $SSE = \sum(Y_i - \hat{Y}_i)$

2.6 Geometric Interpretations

Bibliography

Hogg, R. V., McKean, J. W., and Craig, A. T. (2018). *Introduction to Mathematical Statistics*. Pearson College Division, 8 edition.