



# Regression Analysis

*R Lab*

**O RLY?**

*Young-geun Kim*



# R Lab for Regression Analysis

*Young-geun Kim*

*Department of Statistics, SKKU*

*dudrms33@g.skku.edu*

*09 Apr, 2019*



# Contents

<b>Welcome</b>	<b>5</b>
Linear Regression Analysis . . . . .	5
<b>1 Simple Linear Regression</b>	<b>7</b>
1.1 Model . . . . .	7
1.2 Least Squares Estimation . . . . .	8
1.3 Maximum Likelihood Estimation . . . . .	17
1.4 Residuals . . . . .	20
1.5 Decomposition of Total Variability . . . . .	23
1.6 Geometric Interpretations . . . . .	26
1.7 Distributions . . . . .	31
1.8 Statistical Inference . . . . .	36
1.9 Analysis of Variance . . . . .	42
<b>2 Multiple Linear Regression</b>	<b>53</b>
2.1 Model . . . . .	53
2.2 Least Square Estimation . . . . .	54



# Welcome

This book aims at covering materials of regression analysis. Also, there will be R programming for regression.

```
library(tidyverse)
```

tidyverse package will be used in every chapter, so loading step will be hidden.

## 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

## 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 1

## Simple Linear Regression

### 1.1 Model

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

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

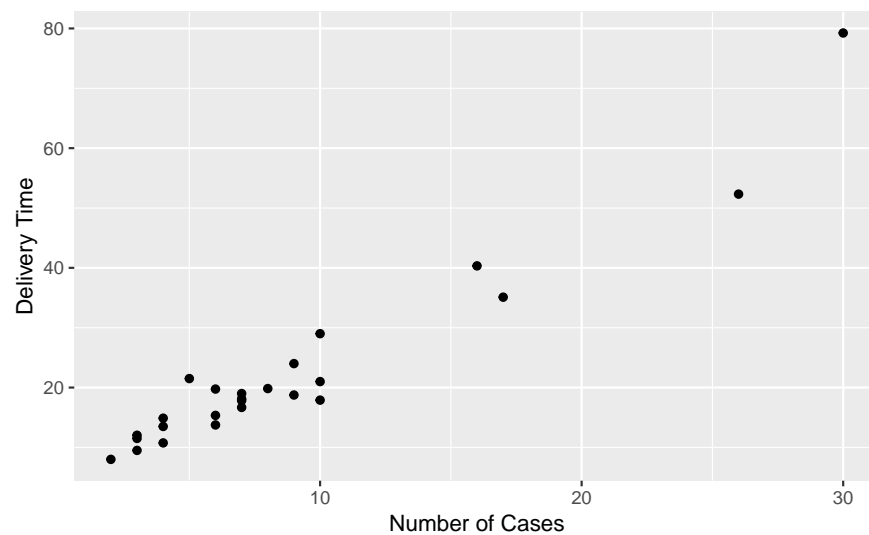


Figure 1.1: The Delivery Time Data

Given data  $(x_1, Y_1), \dots, (x_n, Y_n)$ , 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.

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

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 distributed, i.e. *constant variance*
4. In some setting,  $\epsilon_i \sim N$

## 1.2 Least Squares Estimation



Figure 1.2: 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.

$$(\beta_0, \beta_1) = \arg \min \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 \quad (1.1)$$

### 1.2.1 Normal equations

Denote that Equation (1.1) 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

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) = 0 \quad (1.2)$$

and

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (1.3)$$

From Equation (1.2),

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

Thus,

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

Equation (1.3) gives

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

Thus,

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

*Remark.*

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

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

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

$$\begin{aligned} S_{XX} &= \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n \bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{2}{n} \left( \sum_{i=1}^n x_i \right)^2 + \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \\ &= \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \end{aligned} \quad (1.4)$$

It follows that

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

□

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

Call:

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

Coefficients:

(Intercept)	x
3.32	2.18

### 1.2.2 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 1.1** (Mean response).

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

We can estimate this mean response by

$$\widehat{E(Y | x)} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.5}$$

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

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1.6}$$

Observe that the values of Equations (1.5) and (1.6) are same. However, due to the **error term in the prediction**, it has larger standard error.

### 1.2.3 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 1.1.**

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

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

*Proof.* We already proven the first part of  $S_{XX}$ . See the Equation (1.4). The second part is tivial. Since  $\sum (x_i - \bar{x}) = 0$ ,

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

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

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

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

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

□

**Lemma 1.2** (Linearity). *Each coefficient is a linear estimator.*

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

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

*Proof.* From lemma 1.1,

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

It gives that

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

□

**Proposition 1.1** (Unbiasedness). *Both coefficients are unbiased.*

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

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

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

*Proof.* From lemma 1.1,

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

It follows that

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

□

**Proposition 1.2** (Variances). *Variances and covariance of coefficients*

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

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

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

*Proof.* Proving is just arithmetic.

(a)

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

(b)

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

(c)

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

□

### 1.2.4 Gauss-Markov Theorem

Chapter 1.2.3 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 1.2 are lower than any parameters in this parameter family,  $\beta_0^{LSE}$  and  $\beta_1^{LSE}$  are the **best linear unbiased estimators**.

**Theorem 1.1** (Gauss Markov Theorem).  *$\hat{\beta}_0$  and  $\hat{\beta}_1$  are BLUE, i.e. the best linear unbiased estimator.*

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

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

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

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

$$\begin{cases} \sum b_i = 0 \\ \sum b_i x_i = 1 \end{cases} \quad (1.7)$$

Then

$$\begin{aligned} 0 &\leq Var\left(\sum b_i Y_i - \hat{\beta}_1\right) = Var\left(\sum b_i Y_i - \sum \frac{(x_i - \bar{x})}{S_{XX}} Y_i\right) \\ &\stackrel{indep}{=} \sum \left(b_i - \frac{(x_i - \bar{x})}{S_{XX}}\right)^2 \sigma^2 \\ &= \sum \left(b_i^2 - \frac{2b_i(x_i - \bar{x})}{S_{XX}} + \frac{(x_i - \bar{x})^2}{S_{XX}^2}\right) \sigma^2 \\ &= \sum b_i^2 \sigma^2 - \frac{2\sigma^2}{S_{XX}} \sum b_i x_i + \frac{2\bar{x}\sigma^2}{S_{XX}} \sum b_i + \sigma^2 \frac{\sum (x_i - \bar{x})^2}{S_{XX}^2} \\ &= \sum b_i^2 \sigma^2 - \frac{\sigma^2}{S_{XX}} \quad \because (1.7) \text{ and } S_{XX} = \sum (x_i - \bar{x})^2 \\ &= Var(\sum b_i Y_i) - Var(\hat{\beta}_1) \end{aligned}$$

Hence,

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

□

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

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

$$\begin{cases} \sum a_i = 1 \\ \sum a_i x_i = 0 \end{cases} \quad (1.8)$$



Then

$$\begin{aligned}
0 \leq \text{Var}\left(\sum a_i Y_i - \hat{\beta}_0\right) &= \text{Var}\left[\sum a_i Y_i - \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right) Y_i\right] \\
&= \sum \left(a_i - \frac{1}{n} + \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right)^2 \sigma^2 \\
&= \sum \left[a_i^2 - 2a_i\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right) + \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}\right)^2\right] \sigma^2 \\
&= \sum a_i^2 \sigma^2 - \frac{2\sigma^2}{n} \sum a_i + \frac{2\bar{x}\sigma^2 \sum a_i x_i}{S_{XX}} - \frac{2\bar{x}^2 \sigma^2 \sum a_i}{S_{XX}} \\
&\quad + \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{nS_{XX}} \sum (x_i - \bar{x}) + \frac{\bar{x}^2 \sum (x_i - \bar{x})^2}{S_{XX}^2}\right) \\
&= \sum a_i^2 \sigma^2 - \frac{2\sigma^2}{n} - \frac{2\bar{x}^2 \sigma^2}{S_{XX}} \quad \because (1.8) \\
&\quad + \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \sigma^2 \quad \because \sum (x_i - \bar{x}) = 0 \text{ and } S_{XX} := \sum (x_i - \bar{x})^2 \\
&= \sum a_i^2 \sigma^2 - \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \sigma^2 \\
&= \text{Var}\left(\sum a_i Y_i\right) - \text{Var}\hat{\beta}_0
\end{aligned}$$

Hence,

$$\text{Var}\left(\sum a_i Y_i\right) \geq \text{Var}(\hat{\beta}_0)$$

□

**Example 1.1.** 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 (1.2) and (1.3). Note that  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ .

From the Equation (1.2), we have  $\sum (Y_i - \hat{Y}_i) = 0$ .

From the Equation (1.3), we have  $\sum x_i (Y_i - \hat{Y}_i) = 0$ .

It follows that

$$\begin{aligned}
\sum \hat{Y}_i (Y_i - \hat{Y}_i) &= \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) (Y_i - \hat{Y}_i) \\
&= \hat{\beta}_0 \sum (Y_i - \hat{Y}_i) + \hat{\beta}_1 \sum x_i (Y_i - \hat{Y}_i) \\
&= 0
\end{aligned}$$

### 1.2.5 Estimation of $\sigma^2$

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

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (1.9)$$

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

**Proposition 1.3** (Unbiasedness).

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

*Proof.* Note that

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

Then

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

(a)

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

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

$$\begin{aligned} E \left( \sum (\epsilon_i - \bar{\epsilon})^2 \right) &= E \left( \sum (\epsilon_i^2 + \bar{\epsilon}^2 - 2\epsilon_i \bar{\epsilon}) \right) \\ &= \sum E(\epsilon_i^2) - nE(\bar{\epsilon}^2) \quad \because \sum \epsilon = n\bar{\epsilon} \\ &= \sum (Var(\epsilon_i) + E(\epsilon_i)^2) - n(Var(\bar{\epsilon}) + E(\bar{\epsilon})^2) \\ &= n\sigma^2 - \sigma^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

Thus,

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

(b)

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

It follows that

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

□

## 1.3 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 1.2** (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

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

### 1.3.1 Likelihood equations

**Definition 1.2** (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.

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{1.10}$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i (Y_i - \beta_0 - \beta_1 x_i) = 0 \quad (1.11)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \quad (1.12)$$

Denote that Equations (1.10) and (1.11) 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 Equation (1.12),

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

While  $\hat{\sigma}^{2LSE}$  is an unbiased, above *MLE is not an unbiased estimator*. Since  $\hat{\sigma}^{2MLE} \approx \hat{\sigma}^{2LSE}$  for large  $n$ , however, it is *asymptotically unbiased*.

**Theorem 1.2** (Rao-Cramer Lower Bound, univariate case). *Let  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ . If  $\hat{\theta}$  is an unbiased estimator of  $\theta$ ,*

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

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

To apply this theorem 1.2 in the simple linear regression setting, i.e.  $(\beta_0, \beta_1)$ , we need to look at the *bivariate case*.

**Theorem 1.3** (Rao-Cramer Lower Bound, bivariate case). *Let  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta_1, \theta_2)$  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*

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

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  $\sigma^2$  is **known**. From the Equations (1.10) and (1.11),

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

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,

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

Hence,

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

Since  $\text{Var}(\hat{\beta}) - I^{-1} = 0$  is non-negative definite, each  $\text{Var}(\hat{\beta}_0) = \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \sigma^2$  and  $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{XX}}$  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* 1.1, where the LSE has the smallest 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  $\sigma^2$ .

*Proof.* From the example 1.2,

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

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. □

## 1.4 Residuals

**Definition 1.3** (Residuals).

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

### 1.4.1 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"
  )
)
```

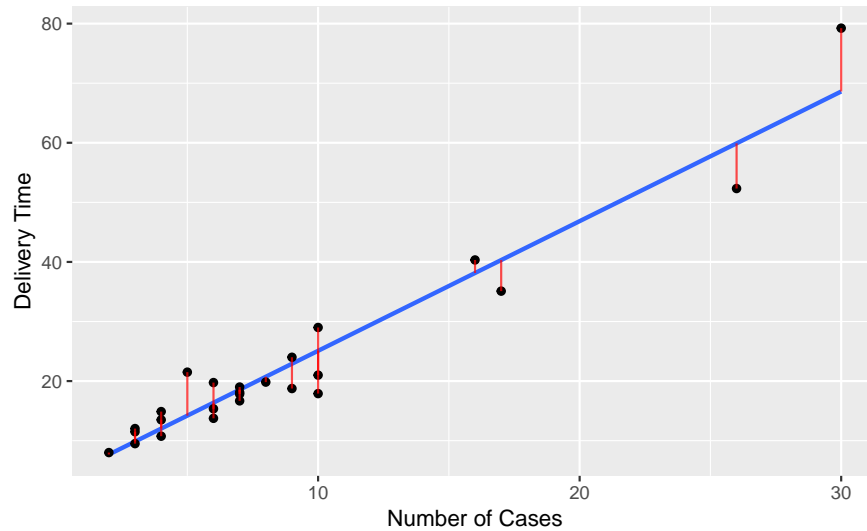


Figure 1.3: Fit and residuals

See Figure 1.3. 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 1.4** (Residual Sum of Squares).

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

### 1.4.2 Residuals and the variance

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

**Lemma 1.3.** *Covariance between  $Y$  and each coefficient*

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

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

*Proof.* (a)

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

(b)

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

□

**Proposition 1.4** (Properties of residuals). *Mean and variance of the residual*

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

$$(b) \text{Var}(e_i) \neq \sigma^2$$

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

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

(b) Lemma 1.3 implies that

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

Then

$$\begin{aligned}
\text{Var}(\hat{Y}_i) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \text{Var}\left[\bar{Y} + (x_i - \bar{x})\hat{\beta}_1\right] \quad \because \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \\
&= \text{Var}(\bar{Y}) + (x_i - \bar{x})^2 \text{Var}(\hat{\beta}_1) + 2(x_i - \bar{x})\text{Cov}(\bar{Y}, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + (x_i - \bar{x})^2 \frac{\sigma^2}{S_{XX}} + 0 \\
&= \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right) \sigma^2
\end{aligned} \tag{1.13}$$

From the same lemma 1.3,

$$\begin{aligned}
\text{Cov}(Y_i, \hat{Y}_i) &= \text{Cov}(Y_i, \bar{Y} + (x_i - \bar{x})\hat{\beta}_1) \\
&= \text{Cov}(Y_i, \bar{Y}) + (x_i - \bar{x})\text{Cov}(Y_i, \hat{\beta}_1) \\
&= \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}} \sigma^2 \quad \because \text{Cov}(Y_i, \hat{\beta}_1) = \frac{(x_i - \bar{x})}{S_{XX}} \sigma^2 \\
&= \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right) \sigma^2
\end{aligned} \tag{1.14}$$

These Equations (1.13) and (1.14) give that

$$\begin{aligned}
\text{Var}(e_i) &= \text{Var}(Y_i) + \text{Var}(\hat{Y}_i) - 2\text{Cov}(Y_i, \hat{Y}_i) \\
&= \sigma^2 + \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right) \sigma^2 - 2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{XX}}\right) \sigma^2 \\
&= \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{XX}}\right) \sigma^2 \\
&\neq \sigma^2
\end{aligned} \tag{1.15}$$

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

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



□

## 1.5 Decomposition of Total Variability

### 1.5.1 Total sum of squares

**Definition 1.5** (Uncorrected Total Sum of Squares).

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

**Definition 1.6** (Corrected Total Sum of Squares).

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

What does this total sum of squares mean? To know this, we should know  $\bar{Y}$  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 1.4: Regression without predictor

See Figure 1.4. 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,

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

From this fact, *SST* implies **total variance**.

### 1.5.2 Regression sum of squares

**Definition 1.7** (Regression Sum of Squares).

$$SSR := \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

This *SSR* compares  $\hat{Y}_i$  versus  $\bar{Y}$ , computing the sum of squares for difference between predicted values from *regression model* and *model not using predictors*.

### 1.5.3 Residual sum of squares

Now consider the *residual sum of squares* *SSE* in the definition 1.4. As mentioned, this is related to the *prediction errors*, which the regression model could not explain the data.

### 1.5.4 Decomposition of total sum of squares

*SST* can be decomposed by construction of sum of squares.

**Proposition 1.5** (Decomposition of SST).

$$SST = SSR + SSE$$

where  $SST = \sum (Y_i - \bar{Y})^2$ ,  $SSR = \sum (\hat{Y}_i - \bar{Y})^2$ , and  $SSE = \sum (Y_i - \hat{Y}_i)^2$

*Proof.* From the Example 1.1,

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad \because \sum (Y_i - \hat{Y}_i) = 0 \text{ and } \sum (Y_i - \hat{Y}_i)\hat{Y}_i = 0 \end{aligned}$$

□

This represents each  $SSR$  and  $SSE$  divides total variability as following.

$$\overset{SST}{\text{total variability}} = \overset{SSR}{\text{explained by regression}} + \overset{SSE}{\text{left unexplained by regression}}$$

Denote that the total variability  $SST$  is *constant given data set*. If our model is good,  $SSR$  grows and  $SSE$  flattens. Thus the larger  $SSR$  is, the better. The lower  $SSE$  is, the better.

### 1.5.5 Coefficient of determination

We have discussed in the previous section 1.5.4 that  $SSR$  and  $SSE$  splits the total variability into *explained part and not-explained part by our regression model*. Our first interest is whether the model works well for the data well, so we can think about the *proportion of explained part to the total variance*. The following measure  $R^2$  computes this kind of value.

**Definition 1.8** (Coefficient of Determination).

$$R^2 := \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

By construction,

$$0 \leq R^2 \leq 1$$

As  $R^2$  goes to 0, the model goes wrong. As  $R^2$  is close to 1, large proportion of variability has been explained. So we prefer large values rather than small.

**Proposition 1.6.**  $R^2$  shows the strength of linear relation between two variables  $x$  and  $Y$  in the simple linear regression.

$$R^2 = \hat{\rho}_{XY}^2$$

where  $\hat{\rho}_{XY} := \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$  is the sample correlation coefficients

*Proof.* Note that  $\hat{Y}_i - \bar{Y} = \hat{\beta}_1(x_i - \bar{x}) = \frac{S_{XY}}{S_{XX}}(x_i - \bar{x})$ . Then

$$\begin{aligned} \sum (\hat{Y}_i - \bar{Y})^2 &= \frac{S_{XY}^2}{S_{XX}^2} \sum (x_i - \bar{x})^2 \\ &= \frac{S_{XY}^2}{S_{XX}} \end{aligned}$$

It follows that

$$\begin{aligned} R^2 &= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{S_{XY}^2}{S_{XX} S_{YY}} \\ &=: \hat{\rho}_{XY}^2 \end{aligned}$$

□

In this relation, we can know that  $R^2$  statistic performs as a measure of the linear relationship in the simple linear regression setting.

## 1.6 Geometric Interpretations

### 1.6.1 Fundamental subspaces

These linear algebra concepts might be more useful for *multiple linear regression*, but let's briefly recap (Leon, 2014).

**Definition 1.9** (Fundamental Subspaces). Let  $X \in \mathbb{R}^{n \times (p+1)}$ .

Then the Null space is defined by

$$N(X) := \{\mathbf{b} \in \mathbb{R}^n \mid X\mathbf{b} = \mathbf{0}\}$$

The Row space is defined by

$$Row(X) := sp(\{\mathbf{r}_1, \dots, \mathbf{r}_{p+1}\}) \quad \text{where } X^T = [\mathbf{r}_1^T, \dots, \mathbf{r}_n^T]$$

The Column space is defined by

$$Col(X) := sp(\{\mathbf{c}_1, \dots, \mathbf{c}_n\}) \quad \text{where } X = [\mathbf{c}_1, \dots, \mathbf{c}_{p+1}]$$

The Range of  $X$  is defined by

$$R(X) := \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} = X\mathbf{b} \quad \text{for some } \mathbf{b} \in \mathbb{R}^{p+1}\}$$

These spaces have some constructional relationship.

**Theorem 1.4** (Fundamental Subspaces Theorem). Let  $X \in \mathbb{R}^{n \times (p+1)}$ . Then

$$N(X) = R(X^T)^\perp = Col(X^T)^\perp = Row(X)^\perp$$

Transposed matrix also satisfy this.

$$N(X^T) = R(X)^\perp = Col(X)^\perp$$

*Proof.* Let  $\mathbf{a} \in N(X)$ . Then  $X\mathbf{a} = \mathbf{0}$ .

Let  $\mathbf{y} \in R(X^T)$ . Then  $X^T\mathbf{b} = \mathbf{y}$  for some  $\mathbf{b} \in \mathbb{R}^{p+1}$ .

Choose  $\mathbf{b} \in \mathbb{R}^{p+1}$  such that  $X^T\mathbf{b} = \mathbf{y}$ . Then

$$\begin{aligned} \mathbf{0} &= X\mathbf{a} \\ &= \mathbf{b}^T X\mathbf{a} \\ &= \mathbf{y}^T \mathbf{a} \end{aligned}$$

Hence,

$$N(X) \perp R(X^T)$$

Since

$$X^T \mathbf{b} = \mathbf{c}_1 \mathbf{b} + \cdots + \mathbf{c}_{p+1} \mathbf{b}$$

it is trivial that  $R(X) = \text{Col}(X)$  and  $R(X^T) = \text{Col}(X^T)$ .

If  $\mathbf{a} \in N(X)$ , then

$$X\mathbf{a} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \vdots \\ \mathbf{r}_n \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_{p+1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Thus,

$$\forall i : \mathbf{a}^T \mathbf{r}_i = 0$$

and so

$$N(X) \subseteq \text{Row}(X)^\perp$$

Conversely, if  $\mathbf{a} \in \text{Row}(X)^\perp$ , then  $\forall i : \mathbf{a}^T \mathbf{r}_i = 0$ . This implies that  $X\mathbf{a} = \mathbf{0}$ . Thus,

$$\text{Row}(X)^\perp \subseteq N(X)$$

and so

$$N(X) = \text{Row}(X)^\perp$$

□

$N(X^T) = R(X)^\perp$  part in Theorem 1.4 will give the geometric insight to *least squares solution*.

**Theorem 1.5.** *Let  $S$  be a subspace of  $\mathbb{R}^n$ . Then*

$$\dim S + \dim S^\perp = n$$

*If  $\{\mathbf{x}_1, \dots, \mathbf{x}_r\}$  is a basis for  $S$  and  $\{\mathbf{x}_{r+1}, \dots, \mathbf{x}_n\}$  is a basis for  $S^\perp$ , then  $\{\mathbf{x}_1, \dots, \mathbf{x}_r, \mathbf{x}_{r+1}, \dots, \mathbf{x}_n\}$  is a basis for  $\mathbb{R}^n$ .*

**Theorem 1.6.** *Let  $S$  be a subspace of  $\mathbb{R}^n$ . Then*

$$\mathbb{R}^n = S \oplus S^\perp$$

### 1.6.2 Simple linear regression

**Theorem 1.7.** *Let  $S$  be a subspace of  $\mathbb{R}^n$ . For each  $\mathbf{y} \in \mathbb{R}^n$ , there exists a unique  $\mathbf{p} \in S$  that is closest to  $\mathbf{y}$ , i.e.*

$$\|\mathbf{y} - \mathbf{p}\| > \|\mathbf{y} - \hat{\mathbf{y}}\|$$

*for any  $\mathbf{p} \neq \hat{\mathbf{y}}$ . Furthermore, a given vector  $\mathbf{p} \in S$  will be the closest to a given vector  $\mathbf{y} \in \mathbb{R}^n$  if and only if*

$$\mathbf{y} - \hat{\mathbf{y}} \in S^\perp$$

Least square estimator  $(\hat{\beta}_0, \hat{\beta}_1)^T$  minimizes

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 = \|\mathbf{Y} - (\beta_0 \mathbf{1} + \beta_1 \mathbf{x})\|^2 \quad (1.16)$$

with respect to  $(\hat{\beta}_0, \hat{\beta}_1)^T \in \mathbb{R}^2$  (where  $\mathbf{1} := (1, 1)^T$ ). Recall that the normal equation gives

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \left( \mathbf{Y} - (\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}) \right)^T \mathbf{1} = 0$$

and

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = \left( \mathbf{Y} - (\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}) \right)^T \mathbf{x} = 0$$

These two relation give

$$\mathbf{Y} - (\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}) \perp \text{sp}(\{\mathbf{1}, \mathbf{x}\})^\perp$$

i.e.  $\hat{\mathbf{Y}} = \hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}$  is the projection of  $\mathbf{Y}$ .

Theorem 1.7 can give the same result.

$$\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x} \in R([\mathbf{1}, \mathbf{x}])^\perp = \text{sp}(\{\mathbf{1}, \mathbf{x}\})^\perp$$

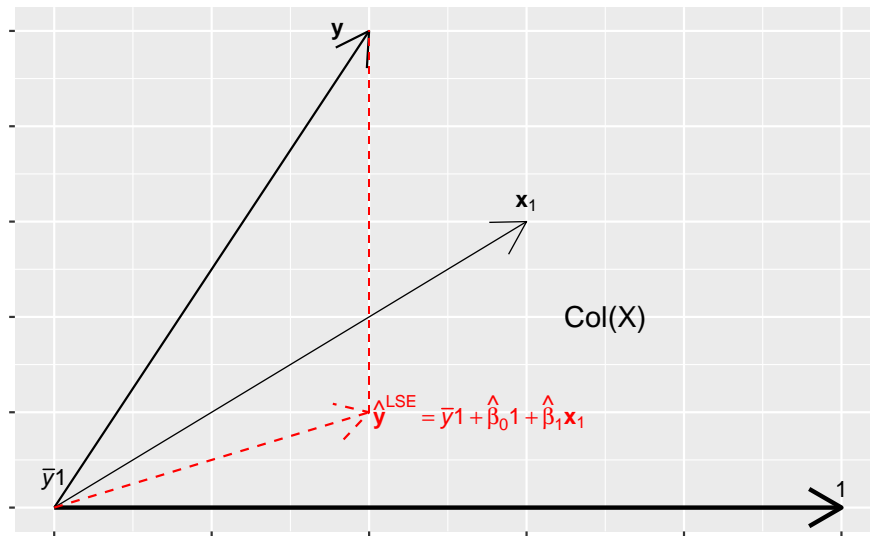


Figure 1.5: Geometric Illustration of Simple Linear Regression

We can see the details from Figure 1.5. In fact, decomposition of  $SST$  and  $R^2$  are also in here.



Figure 1.6: Geometric Illustration of Decomposing SST

See Figure 1.6.

$$\begin{cases} SST = \|\mathbf{Y} - \bar{Y}\mathbf{1}\|^2 \\ SSR = \|\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}\|^2 \\ SSE = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \end{cases}$$

Pythagorean law implies that

$$SST = SSR + SSE$$

Also,

$$R^2 = \frac{SSR}{SST} = \cos^2 \theta = \hat{\rho}_{XY}^2$$

### 1.6.3 Projection mapping

Look again Figure 1.5. Let  $X \equiv [\mathbf{1}, \mathbf{x}] \in \mathbb{R}^{n \times 2}$  and let  $\beta \equiv (\beta_0, \beta_1)^T$ . By the fundamental subspaces theorem 1.4,

$$\mathbf{Y} - X\hat{\beta} \in \text{Col}(X)^\perp = N(X^T)$$

Thus,

$$X^T(\mathbf{Y} - X\hat{\beta}) = \mathbf{0} \tag{1.17}$$

This is the another representation of normal equation. Then we now have

$$\begin{aligned} X^T \mathbf{Y} - X^T X \hat{\beta} &= \mathbf{0} \\ \Leftrightarrow X^T \mathbf{Y} &= X^T X \hat{\beta} \end{aligned}$$

If  $X^T X$  is nonsingular,

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$$

It follows that

$$\hat{\mathbf{Y}} = X \hat{\beta} = X(X^T X)^{-1} X^T \mathbf{Y}$$

Combining this equation and our figure, we can know that  $X(X^T X)^{-1} X^T$  projects  $\mathbf{Y}$  from  $\mathbb{R}^n$  onto  $Col(X) = R(X)$ . This is called projection operator/mapping.

**Definition 1.10** (Projection matrix). Projection operator or mapping from  $\mathbb{R}^n$  to  $W$  is written by

$$\Pi(\cdot | W) := X(X^T X)^{-1} X^T$$

As mentioned,  $X^T X$  should be invertible to get the LSE solution.

**Theorem 1.8.** Let  $\mathbf{Y} = X\beta$  inconsistent and let  $X \in \mathbb{R}^{n \times (p+1)}$  with  $n > p + 1$ .

If  $rank(X) = p + 1$ , i.e. full rank, then  $X^T X$  is invertible.

*Proof.* Suppose that  $(X^T X)\mathbf{b} = \mathbf{0}$ . Then

$$X^T(X\mathbf{b}) = \mathbf{0}$$

By the fundamental subspaces theorem 1.4,

$$X\mathbf{b} \in N(X^T) = Col(X)^\perp$$

By construction,

$$X\mathbf{b} \in Col(X) = N(X^T)^\perp$$

Then

$$X\mathbf{b} \in N(X^T) \cap N(X^T)^\perp = \{\mathbf{0}\}$$

It follows that

$$X\mathbf{b} = \mathbf{0}$$

If  $rank(X) = n$ , then the linear equation system has trivial solution  $\mathbf{b} = \mathbf{0}$  and so does  $X^T(X\mathbf{b}) = \mathbf{0}$ . Hence,  $X^T X$  is invertible.  $\square$

Using projection matrix  $\Pi_W$ , we can re-express each sum of squares. Recall that when we only use  $y_i$  for regression fitting, the result becomes its average. It is because  $\mathbf{Y}$  vector has been projected onto  $sp(\{\mathbf{1}\})$  line.

*Remark.*

$$\bar{Y}\mathbf{1} = \mathbf{1}(\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T \mathbf{Y} = \Pi_1 \mathbf{Y}$$

$$\hat{\mathbf{Y}} = X(X^T X)^{-1} X^T \mathbf{Y} = \Pi_X \mathbf{Y}$$



Intuitively, every projection matrix is idempotent and symmetric. Once projected, the result is same when projecting it again.

**Corollary 1.1** (Sum of squares).  $\Pi_1$  and  $\Pi_X$  can express each  $SS$  as following.

(i)

$$SST = \mathbf{Y}^T(I - \Pi_1)\mathbf{Y}$$

(ii)

$$SSR = \mathbf{Y}^T(\Pi_X - \Pi_1)\mathbf{Y}$$

(iii)

$$SSE = \mathbf{Y}^T(I - \Pi_X)\mathbf{Y}$$

## 1.7 Distributions

### 1.7.1 Mean response and response

We have already look at predicting each mean response and response from equation (1.5) and (1.6).

**Theorem 1.9** (Estimation of the mean response).

$$\hat{\mu}_x \equiv E(\widehat{Y} \mid x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

**Theorem 1.10** ((out of sample) Prediction of a response).

$$\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

Recall that predicting 1.9 targets at

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

which have been assumed to be true model. On the other hand, predicting 1.10 targets at

$$Y = \beta_0 + \beta_1 + \epsilon_x$$

The linearity is not true in reality. So the errors caused by modeling linear model are included in  $\epsilon_x$ . This error term makes difference between properties of 1.9 and 1.10.

To derive their distribution and see the difference, we additionally assume Normality, i.e.

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

### 1.7.2 Regression coefficients

Under Normality, we have

$$Y_i \stackrel{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Then

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \sim MVN_n \left( \boldsymbol{\mu} \equiv \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}, \Sigma \equiv \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \right)$$

Write  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ . From Lemma 1.2,

$$\hat{\beta}_0 = \mathbf{a}^T \mathbf{Y}$$

where  $\mathbf{a} = (a_1, \dots, a_n)^T \in \mathbb{R}^n$  with  $a_i = \left( \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}} \right)$

and

$$\hat{\beta}_1 = \mathbf{b}^T \mathbf{Y}$$

where  $\mathbf{b} = (b_1, \dots, b_n)^T \in \mathbb{R}^n$  with  $b_i = \frac{(x_i - \bar{x})}{S_{XX}}$ .

Let

$$A^T = [\mathbf{a}^T, \mathbf{b}^T]$$

Then

$$\hat{\boldsymbol{\beta}} = A\mathbf{Y}$$

Linearity of the multivariate normal distribution, Proposition 1.1 and 1.2 imply that

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \sim MVN \left( A\boldsymbol{\mu} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, A\Sigma A^T = \sigma^2 A A^T = \begin{bmatrix} \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \sigma^2 & -\frac{\bar{x}}{S_{XX}} \sigma^2 \\ -\frac{\bar{x}}{S_{XX}} \sigma^2 & \frac{\sigma^2}{S_{XX}} \end{bmatrix} \right) \quad (1.18)$$

Since the joint random vector follows multivariate normal distribution, each *partitioned subset follow normal*. For this theorem, see Johnson and Wichern (2013). Hence, we finally get the following result.

**Theorem 1.11** (Distributions of regression coefficients). *Each regression coefficient follows Normal distribution.*

$$\hat{\beta}_0 \sim N \left( \beta_0, \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \sigma^2 \right)$$

$$\hat{\beta}_1 \sim N \left( \beta_1, \frac{\sigma^2}{S_{XX}} \right)$$

### 1.7.3 Mean response

In simple linear regression setting, we assume  $\mu_x = E(Y | x) = \beta_0 + \beta_1 x$  is true.

```
delv %>%
  ggplot(aes(x = x, y = y)) +
  geom_smooth(method = "lm") +
  geom_point(alpha = .7) +
  labs(
    x = "Number of Cases",
    y = "Delivery Time"
  )
```

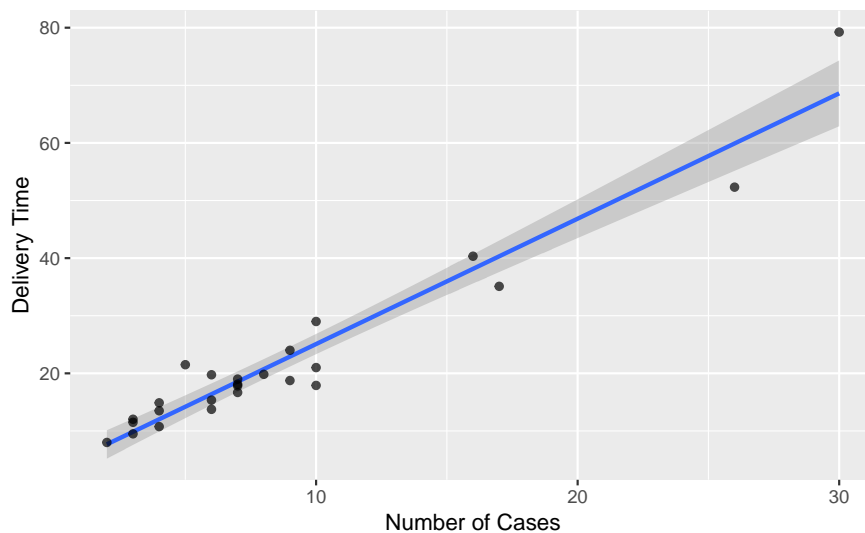


Figure 1.7: Mean response and its standard deviation

For example, in the Figure 1.7, the blue line indicates  $E(Y | X = x)$  for each point  $x$ . Without fitting using `lm()`, `geom_smooth(method = "lm")` let us visualize the fitted line. Since the default method is not the linear regression, the `method` option should be specified.

```
delv %>%
  mutate(eyx = predict(delv_fit, newdata = data.frame(x = x)))
```

```
# A tibble: 25 x 3
      y      x  eyx
  <dbl> <dbl> <dbl>
1  16.7     7 18.6
2  11.5     3  9.85
3  12.0     3  9.85
4  14.9     4 12.0
5  13.8     6 16.4
6  18.1     7 18.6
7    8      2  7.67
8  17.8     7 18.6
9  79.2    30 68.6
10 21.5     5 14.2
# ... with 15 more rows
```

We have already seen in section 1.7.2 that the estimators  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are random variables. So  $\hat{\mu}_x$  is. In fact,

the ribbon of the line in Figure 1.7 represents upper and lower confidence limits on mean response. In the later section, we get to know that it is  $+t(n-2)\widehat{SE}(\hat{\mu}_x)$  and  $-t(n-2)\widehat{SE}(\hat{\mu}_x)$ . It can be drawn by default with the option of the `geom_smooth(se = TRUE)`.

**Theorem 1.12** (Distribution of mean response estimator).  $\hat{\mu}_x$  is also Normally distributed.

$$\hat{\mu}_x \sim N\left(\mu_x, \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}\right)\right)$$

*Proof.* Since  $\hat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x$  is the linear combination of  $(\hat{\beta}_0, \hat{\beta}_1)^T$ ,

$$\hat{\mu}_x \sim N\left(E(\hat{\mu}_x), \text{Var}(\hat{\mu}_x)\right)$$

From Theorem 1.11,

$$E(\hat{\mu}_x) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x = \beta_0 + \beta_1 x \equiv \mu_x$$

and from Proposition 1.2

$$\begin{aligned} \text{Var}(\hat{\mu}_x) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \text{Var}(\hat{\beta}_0) + x^2 \text{Var}(\hat{\beta}_1) + 2x \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\sigma^2 + \frac{x^2\sigma^2}{S_{XX}} - \frac{2\bar{x}x\sigma^2}{S_{XX}} \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}\right) \end{aligned}$$

□

**Corollary 1.2.**

$$\hat{\mu}_x - \mu_x \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}\right)\right)$$

Denote that in both Theorem 1.12 and Corollary 1.2,  $\sigma^2$  is parameter. So to use  $SE(\hat{\mu}_x) = \sqrt{\text{Var}(\hat{\mu}_x)}$  in practice we plug in its estimator, usually Equation (1.9).

**Corollary 1.3** (Standard error of mean response estimator).

$$\widehat{SE}(\hat{\mu}_x) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}}\right)$$

where  $\hat{\sigma}^2 = MSE$

#### 1.7.4 Response

Our goal is to predict each response at each point, i.e.  $Y_x = \beta_0 + \beta_1 x + \epsilon_x$ .  $\epsilon_x \sim N(0, \sigma^2)$  is independent of the given data  $(\epsilon_1, \dots, \epsilon_n)$ . In this sense, this prediction is called *out of sample prediction*. This setting makes difference between the *residuals, which are correlated to the data*. See Proposition 1.4 for this. This is occurred because each  $\hat{\beta}_0$  and  $\hat{\beta}_1$  is linear combination of  $Y_1, \dots, Y_n$ , not  $Y_x$ .

While  $\text{Cov}(Y_i, \hat{Y}_i) > 0, i = 1, \dots, n$  (See Equation (1.14)), in case of out-of-sample  $Y_x$ ,

$$\text{Cov}(Y_x, \hat{Y}_x) = \text{Cov}(Y_x, \hat{\beta}_0 + \hat{\beta}_1 x) = 0$$

Hence, arithmetically, this *out of sample prediction becomes to have larger standard error*.

**Proposition 1.7** (Joint distribution of coefficients and error term).  $(\hat{\beta}_0, \hat{\beta}_1, \epsilon_x)^T$  is Normally distributed.

*Proof.* Want 1:  $(\hat{\beta}_0, \hat{\beta}_1)^T \perp\!\!\!\perp \epsilon_x$

We have

$$\begin{aligned} \text{Cov}((\hat{\beta}_0, \hat{\beta}_1)^T, \epsilon_x) &= \left[ \text{Cov}(\hat{\beta}_i, \epsilon_x) \right]_{2 \times 1} \\ &= \left[ \text{Cov} \left( \sum_{i=1}^n k_i Y_i, \epsilon_x \right) \right]_{2 \times 1} \quad k_i = \text{each linear coefficient for } \hat{\beta}_0, \hat{\beta}_1 \\ &= \mathbf{0} \end{aligned} \tag{1.19}$$

From Equation (1.18),

$$(\hat{\beta}_0, \hat{\beta}_1)^T \sim MVN$$

and from assumption,

$$\epsilon_x \sim N(0, \sigma^2)$$

It follows from Equation (1.19) that (Johnson and Wichern (2013))

$$(\hat{\beta}_0, \hat{\beta}_1)^T \perp\!\!\!\perp \epsilon_x$$

Want 2:  $(\hat{\beta}_0, \hat{\beta}_1, \epsilon_x)^T \sim MVN$

From independency, we have (Johnson and Wichern (2013))

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \epsilon_x \end{bmatrix} \sim MVN_{2+1} \left( \begin{bmatrix} \beta_0 \\ \beta_1 \\ 0 \end{bmatrix}, \left[ \begin{array}{c|c} \text{Cov}(\hat{\beta}) \in \mathbb{R}^{2 \times 2} & \mathbf{0} \in \mathbb{R}^2 \\ \hline \mathbf{0}^T \in \mathbb{R}^{2 \times 1} & \sigma^2 \end{array} \right] \right)$$

□

This proposition gives clue to distribution of prediction error.

**Theorem 1.13** (Distribution of out-of-sample prediction error). *Out of sample prediction error  $\hat{Y}_x - Y_x$  is Normally distributed*

$$\hat{Y}_x - Y_x \sim N \left( 0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} \right) \right)$$

*Proof.* Note that

$$\begin{aligned} \hat{Y}_x - Y_x &= (\hat{\beta}_0 + \hat{\beta}_1 x) - (\beta_0 + \beta_1 x + \epsilon_x) \\ &= [1, x, -1](\hat{\beta}_0, \hat{\beta}_1, \epsilon_x)^T - \beta_0 - \beta_1 x \end{aligned}$$

i.e.  $\hat{Y}_x - Y_x$  is a linear combination of  $(\hat{\beta}_0, \hat{\beta}_1, \epsilon_x)^T$ . From proposition 1.7,

$$\begin{aligned} \hat{Y}_x - Y_x &\sim MVN\left([1, x, -1] \begin{bmatrix} \beta_0 \\ \beta_1 \\ 0 \end{bmatrix} - \beta_0 - \beta_1 x, [1, x, -1] \left[ \frac{Cov(\hat{\beta}) \in \mathbb{R}^{2 \times 2}}{\mathbf{0}^T \in \mathbb{R}^{2 \times 1}} \middle| \frac{\mathbf{0} \in \mathbb{R}^2}{\sigma^2} \right] \begin{bmatrix} 1 \\ x \\ -1 \end{bmatrix} \right) \\ &\stackrel{d}{=} MVN\left(0, \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} - 2 \frac{\bar{x}x}{S_{XX}} + \frac{x^2}{S_{XX}} \right) + 1 \right) \\ &\stackrel{d}{=} MVN\left(0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} \right) \right) \end{aligned} \quad (1.20)$$

□

Now we know the standard error of this out-of-sample prediction error.

$$SE(\hat{Y}_x - Y_x) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} \right)$$

We can see this standard error is *always larger than of mean response estimator* due to 1 in the bracket, i.e.  $\sigma^2$ . As mentioned, this is due to  $\epsilon$  term. When we estimate or predict the mean response the model have been assumed to be true. In this out-of-sample prediction setting, however, the model can be wrong. This assumption error is also included in  $\epsilon$  term and it is called *irreducible error*, which cannot be reduced anymore.

*Remark.*

$$SE(\hat{\mu}_x - \mu_x) < SE(\hat{Y}_x - Y_x)$$

It might be more clear if we see the inequality in the above remark. We know the fact that  $\hat{Y}_x$  and  $Y_x$  are uncorrelated in this out-of-sample setting.  $Y_x$  is random variable, while  $\mu_x$  is constant. Then we can re-express the inequality as

$$SE(\hat{\mu}_x) < SE(\hat{Y}_x) + SE(Y_x)$$

Actually, both  $\hat{\mu}_x$  and  $\hat{Y}_x$  are estimated as  $\hat{\beta}_0 + \hat{\beta}_1 x$ . Thus,  $SE(Y_x) = \sigma^2$  makes out-of-sample more noisy.

To use standard error practically, we use  $\hat{\sigma}^2$  as in corollary 1.3.

**Corollary 1.4** (Standard error of out-of-sample prediction error).

$$\widehat{SE}(\hat{Y}_x - Y_x) = \hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} \right)$$

where  $\hat{\sigma}^2 = MSE$

## 1.8 Statistical Inference

Based on each distribution of estimator in section 1.7, we can construct various inference for each

- $\beta_0$
- $\beta_1$
- $\mu_x$
- $Y_x$
- $\sigma^2$

We can get the standard error for each coefficient through `summary()` function.

```
summary(delv_fit)
```

Call:

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

Residuals:

```
      Min       1Q   Median       3Q      Max
-7.581 -1.874 -0.349  2.181 10.634
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371    2.42  0.024 *
x              2.176      0.124   17.55 8.2e-15 ***
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.18 on 23 degrees of freedom

Multiple R-squared: 0.93, Adjusted R-squared: 0.927

F-statistic: 308 on 1 and 23 DF, p-value: 8.22e-15

Or more state-of-the-art way, `broom::tidy()` function has a method for each model object to make tidy data: tibble.

```
broom::tidy(delv_fit)
```

# A tibble: 2 x 5

```
  term      estimate std.error statistic  p.value
<chr>      <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  3.32      1.37      2.42 2.37e- 2
2 x           2.18      0.124     17.5 8.22e-15
```

### 1.8.1 Confidence interval

Consider standardization.

$$\frac{\hat{\theta} - \theta}{\sqrt{SE(\hat{\theta})}}$$

Each  $SE$  includes  $\sigma^2$  as we have already seen. First think about **known**  $\sigma^2$  setting. All three estimators follow Normal distribution, and  $SE$  is constant by our the setting. Then we can construct each confidence interval as

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{SE(\hat{\theta})}$$

Figure 1.8: Confidence Interval when  $\sigma^2$  is known

Now just plug in the results of section 1.7. For each regression coefficient,

**Proposition 1.8** (Confidence intervals on  $\beta$ ). *With known  $\sigma^2$ ,  $(1 - \alpha)100\%$  confidence intervals on  $\beta_0$  and  $\beta_1$  are given as*

$$\beta_0 : \hat{\beta}_0 \pm z_{\frac{\alpha}{2}} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \sigma^2}$$

$$\beta_1 : \hat{\beta}_1 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{S_{XX}}}$$

**Proposition 1.9** (Confidence interval on  $\hat{\mu}_x$ ). *With known  $\sigma^2$ ,  $(1 - \alpha)100\%$  confidence interval on  $\hat{\mu}_x$  is given as*

$$\mu_x : \hat{\mu}_x \pm z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}$$

In practice, however, we do not know  $\sigma^2$ . In this case, we replace  $\sigma^2$  with  $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 = MSE$ . Then

$$\frac{\hat{\theta} - \theta}{\sqrt{SE}} = \frac{\frac{\hat{\theta} - \theta}{\sqrt{SE = \sigma^2(\cdot)}}}{\sqrt{\frac{\frac{SSE}{\sigma^2}}{n-2} \left( \cdot \right)}} = \frac{\frac{\hat{\theta} - \theta}{\sqrt{SE = \sigma^2}} \sim N(0, 1)}{\sqrt{\frac{\frac{SSE}{\sigma^2} \sim \chi^2(n-2)}{n-2}}} \sim t(n-2)$$

Thus, we need to replace  $z_{\frac{\alpha}{2}}$  with  $t_{\frac{\alpha}{2}}(n-2)$ .

**Proposition 1.10** (Confidence intervals on  $\beta$  when unknown  $\sigma^2$ ). *With unknown  $\sigma^2$ ,  $(1 - \alpha)100\%$  confidence intervals on  $\beta_0$  and  $\beta_1$  are given as*



$$\beta_0 : \hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2) \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right) \hat{\sigma}^2}$$

$$\beta_1 : \hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

where  $\hat{\sigma}^2 = MSE$

Here we can estimate the intervals. Basically, `confint()` function gives this interval.

```
confint(delv_fit, level = .95)
```

```

                2.5 % 97.5 %
(Intercept) 0.484   6.16
x           1.920   2.43
```

**Proposition 1.11** (Confidence interval on  $\hat{\mu}_x$  when unknown  $\sigma^2$ ). *With unknown  $\sigma^2$ ,  $(1-\alpha)100\%$  confidence interval on  $\hat{\mu}_x$  is given as*

$$\mu_x : \hat{\mu}_x \pm t_{\frac{\alpha}{2}}(n-2) \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}$$

where  $\hat{\sigma}^2 = MSE$

`predict()` provides options for this confidence interval. Specify `interval = "confidence"`. This argument has three option.

1. "none": just compute fitted value, by default.
2. "confidence": confidence interval of mean response
3. "prediction": prediction interval of out-of-sample prediction

Default level is 0.95.

```
predict(delv_fit, interval = "confidence", level = .95) %>% tbl_df()
```

```

# A tibble: 25 x 3
  fit   lwr   upr
<dbl> <dbl> <dbl>
1 18.6  16.8  20.3
2  9.85  7.57  12.1
3  9.85  7.57  12.1
4 12.0   9.91  14.1
5 16.4  14.5  18.2
6 18.6  16.8  20.3
7  7.67  5.22  10.1
8 18.6  16.8  20.3
9 68.6  62.9  74.3
10 14.2  12.2  16.2
# ... with 15 more rows
```

## 1.8.2 Prediction interval

One proceeds in a similar way for out-of-sample  $Y_x$ .

**Proposition 1.12** (Prediction interval on  $\hat{Y}_x$ ). *With known  $\sigma^2$ ,  $(1 - \alpha)100\%$  confidence interval on  $\hat{\mu}_x$  is given as*

$$Y_x : \hat{Y}_x \pm z_{\frac{\alpha}{2}} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}$$

Also, with unknown  $\sigma^2$ ,

**Proposition 1.13** (Prediction interval on  $\hat{Y}_x$  when unknown  $\sigma^2$ ). *With unknown  $\sigma^2$ ,  $(1 - \alpha)100\%$  confidence interval on  $\hat{\mu}_x$  is given as*

$$Y_x : \hat{Y}_x \pm t_{\frac{\alpha}{2}}(n - 2) \sqrt{\hat{\sigma}^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}$$

where  $\hat{\sigma}^2 = MSE$

Since this is out-of-sample setting, we should also give `newdata` option. Otherwise, we will get warning message. Denote that this argument only receive `data.frame` object with same element names.

```
predict(delv_fit, newdata = data.frame(x = 31:35), interval = "prediction", level = .95)
```

```
      fit   lwr   upr
1 70.8 60.3 81.3
2 73.0 62.3 83.6
3 75.1 64.3 85.9
4 77.3 66.4 88.3
5 79.5 68.4 90.6
```

### 1.8.3 Hypothesis testing

Look again the output of `summary.lm()` and `broom::tidy.lm()`.

```
summary(delv_fit)
```

Call:

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

Residuals:

```
      Min       1Q   Median       3Q      Max
-7.581 -1.874 -0.349  2.181 10.634
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371    2.42   0.024 *
x              2.176      0.124   17.55 8.2e-15 ***
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.18 on 23 degrees of freedom

Multiple R-squared: 0.93, Adjusted R-squared: 0.927

F-statistic: 308 on 1 and 23 DF, p-value: 8.22e-15

We can see `t` value and `Pr(>|t|)`. At the same time, `statistic` and `p.value`. What are these values? These are the results of the following tests.

$$H_0 : \beta_0 = \alpha_0 \quad \text{vs} \quad H_1 : \beta_0 \neq \alpha_0$$

$$T = \frac{\hat{\beta}_0 - \alpha_0}{\hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \stackrel{H_0}{\sim} t(n-2) \quad (1.21)$$

For this test statistic (1.21),

$$\text{reject } H_0 \quad \text{if } |T| > t_{\frac{\alpha}{2}}(n-2)$$

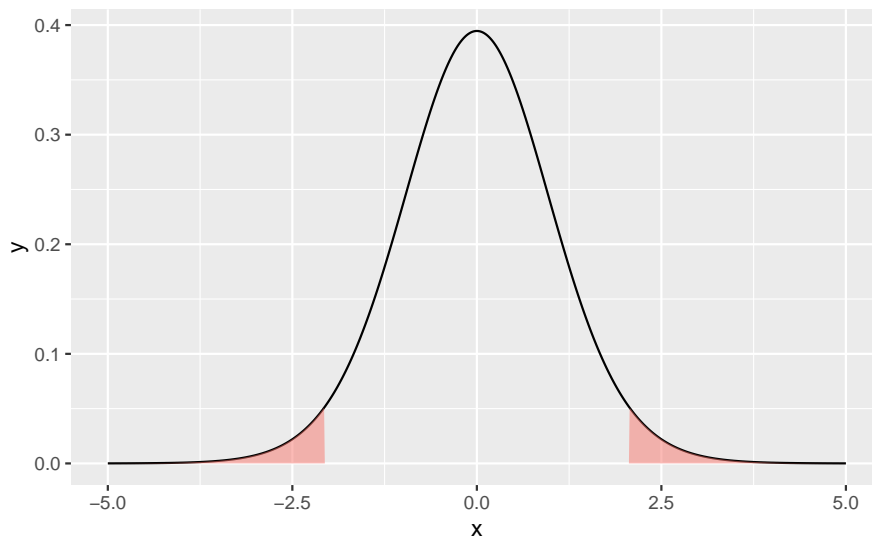


Figure 1.9: Rejection region for  $\beta_0$

More importantly, we test  $\beta_1$  which means slope

$$H_0 : \beta_1 = \alpha_1 \quad \text{vs} \quad H_1 : \beta_1 \neq \alpha_1$$

$$T = \frac{\hat{\beta}_1 - \alpha_1}{\hat{\sigma} \sqrt{\frac{1}{S_{xx}}}} \stackrel{H_0}{\sim} t(n-2) \quad (1.22)$$

For this test statistic (1.22),

$$\text{reject } H_0 \quad \text{if } |T| > t_{\frac{\alpha}{2}}(n-2)$$

Looking at these two statistics, we can intuitively know the meaning. As  $|\hat{\beta}_1 - \alpha_1|$  becomes larger, the data support  $H_1$ .

## 1.9 Analysis of Variance

### 1.9.1 Useful distributions

In linear regression setting, we usually assume  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ . There are some useful distributions around Normal.

**Proposition 1.14** ( $\chi^2$ -distribution). *Square of standard normal follows  $\chi^2$ -distribution.*

If  $Z \sim N(0, 1)$ , then  $Z^2 \sim \chi^2(1)$

If  $Z_i \stackrel{indep}{\sim} N(0, 1)$ , then  $Z_1^2 + \dots + Z_n^2 \sim \chi^2(n)$

**Proposition 1.15** (t-distribution). *Let  $Z \sim N(0, 1) \perp V \sim \chi^2(m)$ . Then*

$$T = \frac{Z}{\sqrt{V/m}} \sim t(m)$$

**Proposition 1.16** (F-distribution). *Let  $V \sim \chi^2(m) \perp W \sim \chi^2(n)$ . Then*

$$F = \frac{V/m}{W/n} \sim F(m, n)$$

Also, there is *non-central analogue* of these three distributions, i.e. starting from  $Z \sim N(\mu, 1)$ .

**Proposition 1.17** (Noncentral  $\chi^2$ -distribution). *Square of scaled normal follows non-central  $\chi^2$ -distribution.*

If  $Z_i \stackrel{indep}{\sim} N(\mu_i, 1)$ , then  $Z_1^2 + \dots + Z_n^2 \sim \chi^2(n, \sum_{i=1}^n \mu_i^2)$

$\sum_{i=1}^n \mu_i^2$  is called a non-central parameter.

**Proposition 1.18** (Noncentral t-distribution). *Let  $X \sim N(\mu, 1) \perp V \sim \chi^2(m)$ . Then*

$$T = \frac{Z}{\sqrt{V/m}} \sim t(m, \mu)$$

$\mu$  is called a non-central parameter.

**Proposition 1.19** (Noncentral F-distribution). *Let  $V \sim \chi^2(m, \delta) \perp W \sim \chi^2(n)$ . Then*

$$F = \frac{V/m}{W/n} \sim F(m, n, \delta)$$

$\delta$  is called a non-central parameter.

### 1.9.2 Quadratic form

Now we can determine the distributions of various quadratic forms. The reason we are taking care of this is ANOVA deals with sum of squares, i.e. quadratic form.

**Theorem 1.14** (Idempotent and symmetric). *Let  $A \in \mathbb{R}^{k \times k}$  be idempotent and symmetric. Then*

(a)  $A^n$  is also idempotent

(b)  $I - A$  is also idempotent

(c) Every eigenvalue of  $A$  is either 0 or 1 so that  $\text{tr}(A) = \text{rank}(A)$

*Proof.* (a) and (b) are trivial.

$$(A^n)^2 = (A^2)^n = A^n$$

$$(I - A)^2 = I - 2A + A^2 = I - A$$

(c)

Fix  $\lambda$  an eigenvalue of  $A$ . Let  $\mathbf{v} \neq \mathbf{0}$  be the corresponding eigenvector.

By definition,

$$A\mathbf{v} = \lambda\mathbf{v}$$

Then

$$A^2\mathbf{v} = \lambda(A\mathbf{v}) = \lambda^2\mathbf{v}$$

and so  $\lambda^2$  is eigenvalue of  $A^2$ .

Since  $A^2 = A$ ,

$$\lambda = \lambda^2$$

Hence,

$$\lambda = 0 \text{ or } 1$$

Note that for every matrix and its eigenvalues  $\lambda_j$

$$\text{tr}(X) = \sum_{j=1}^p \lambda_j, \quad \text{rank}(X) = \text{the number of non-zero } \lambda_j$$

Since  $\lambda = 0, 1$  of  $A$ ,

$$\text{tr}(A) = \text{rank}(A)$$

□

**Proposition 1.20** (Independence). Assume  $\mathbf{Y} \sim MVN(\mu, \Sigma)$ . Then

(i) If  $A$  and  $B$  are symmetric,

$$Y^T A Y \perp\!\!\!\perp Y^T B Y \Leftrightarrow A \Sigma B = 0$$

(ii) If  $A$  is symmetric,

$$Y^T A Y \perp\!\!\!\perp B Y \Leftrightarrow B \Sigma A = 0$$

**Theorem 1.15** (Distribution of quadratic form). *Assume that  $\mathbf{Y} \sim MVN(\mu, \Sigma)$  and that  $A$  is symmetric and idempotent. Then*

$$\mathbf{Y}^T A \mathbf{Y} \sim \chi^2(K, \delta)$$

where  $K = \text{rank}(A)$  and  $\delta = \mu^T A \mu$ . Furthermore,

$$\begin{cases} E(\mathbf{Y}^T A \mathbf{Y}) = K + \delta \\ \text{Var}(\mathbf{Y}^T A \mathbf{Y}) = 2(K + 2\delta) \end{cases}$$

**Corollary 1.5** (Inner product of standard normal vector). *Let  $\mathbf{Z} = (Z_1, \dots, Z_n)^T \sim MVN(\mathbf{0}, I_n)$ . Then*

$$\mathbf{Z}^T \mathbf{Z} = \sum_{i=1}^n Z_i^2 \sim \chi^2(n)$$

*Proof.* From Theorem 1.15 point of view,

$$\mathbf{Z}^T \mathbf{Z} = \mathbf{Z}^T I_n \mathbf{Z}$$

Thus,

$$K = \text{rank}(I_n) = n$$

$$\delta = 0$$

□

Using the above facts, we can now show distributions of sums of squares. First recall that

$$\mathbf{Y} \sim MVN(X\beta, \sigma^2 I)$$

**Proposition 1.21** (Distribution of SSE).

$$\frac{SSE}{\sigma^2} \sim \chi^2(n - 2, 0)$$

*Proof.* From Corollary 1.1, write

$$\frac{SSE}{\sigma^2} = \left( \frac{\mathbf{Y}}{\sigma} \right)^T (I - \Pi_X) \left( \frac{\mathbf{Y}}{\sigma} \right)$$

Note that

$$\frac{\mathbf{Y}}{\sigma} \sim MVN\left(\frac{1}{\sigma} X\beta, I\right)$$

Since  $I - \Pi_X$  is idempotent and symmetric,

$$K = \text{rank}(I - \Pi_X) = \text{tr}(I - \Pi_X) = n - \text{rank}(\Pi_X) = n - 2$$

$$\begin{aligned}
\delta &= \left( \frac{X\beta}{\sigma} \right)^T (I - \Pi_X) \left( \frac{X\beta}{\sigma} \right) \\
&= \frac{\beta^T X^T X \beta}{\sigma^2} - \frac{(\beta^T X^T) X (X^T X)^{-1} X^T (X\beta)}{\sigma^2} \\
&= \frac{\beta^T X^T X \beta}{\sigma^2} - \frac{\beta^T X^T X \beta}{\sigma^2} \\
&= 0
\end{aligned} \tag{1.23}$$

Hence,

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

□

**Proposition 1.22** (Distribution of SSR).

$$\frac{SSR}{\sigma^2} \sim \chi^2(1, \delta)$$

$$\text{where } \delta = \sum_{i=1}^n (x_i - \bar{x})^2 \beta_1^2 = \frac{S_{xx} \beta_1^2}{\sigma^2}$$

*Proof.* From Corollary 1.1, write

$$\frac{SSR}{\sigma^2} = \left( \frac{\mathbf{Y}}{\sigma} \right)^T (\Pi_X - \Pi_1) \left( \frac{\mathbf{Y}}{\sigma} \right)$$

Note that  $\Pi_X - \Pi_1$  is symmetric idempotent. One proceeds in a similar way.

$$K = \text{rank}(\Pi_X - \Pi_1) = \text{tr}(\Pi_X - \Pi_1) = \text{rank}(\Pi_X) - \text{rank}(\Pi_1) = 2 - 1 = 1$$

$$\begin{aligned}
\delta &= \left( \frac{X\beta}{\sigma} \right)^T (\Pi_X - \Pi_1) \left( \frac{X\beta}{\sigma} \right) \quad \because \frac{\mathbf{Y}}{\sigma} \sim MVN\left(\frac{1}{\sigma} X\beta, I\right) \\
&= \frac{\beta^T X^T X \beta}{\sigma^2} - \frac{\beta^T X^T \Pi_1 X \beta}{\sigma^2} \\
&= \frac{\beta^T (X^T X - X^T \Pi_1 X) \beta}{\sigma^2} \\
&= \frac{\beta^T \{X^T (I - \Pi_1) X\} \beta}{\sigma^2}
\end{aligned}$$

Since  $\mathbf{1} \in \text{sp}(\{\mathbf{1}\})$ ,

$$\Pi_1 \mathbf{1} = \mathbf{1}$$

It gives that

$$\mathbf{1}^T(I - \Pi_1)\mathbf{1} = 0$$

If  $\mathbf{x} \neq \mathbf{1}$ , then we have

$$\mathbf{x}^T(I - \Pi_1)\mathbf{x} = \sum_{i=1}^n (x_i - \bar{x})^2 = S_{xx}$$

Recall that

$$\bar{x}\mathbf{1} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{x} = \Pi_1\mathbf{x}$$

Then we have

$$\mathbf{1}^T(I - \Pi_1)\mathbf{x} = \sum x_i - n\bar{x} = 0$$

Similarly,

$$\mathbf{x}^T(I - \Pi_1)\mathbf{1} = n\bar{x} - \sum x_i = 0$$

Hence by partitioning  $X = [\mathbf{1} \mid \mathbf{x}]$ ,

$$\begin{aligned} \delta &= \frac{\beta^T \left\{ [\mathbf{1} \mid \mathbf{x}]^T (I - \Pi_1) [\mathbf{1} \mid \mathbf{x}] \right\} \beta}{\sigma^2} \\ &= \frac{\beta^T \begin{bmatrix} 0 & 0 \\ 0 & S_{xx} \end{bmatrix} \beta}{\sigma^2} \\ &= \frac{S_{xx}\beta_1^2}{\sigma^2} \end{aligned} \tag{1.24}$$

□

**Proposition 1.23** (Independence). *SSE and SSR are independent, i.e.*

$$SSE \perp\!\!\!\perp SSR$$

*Proof.* Note that both  $SSE$  and  $SSR$  are quadratic forms of  $\mathbf{Y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I)$  and that each  $I - \Pi_X$  and  $\Pi_X - \Pi_1$  is symmetric. Then from Proposition 1.20,

Claim:  $(I - \Pi_X)(\sigma^2 I)(\Pi_X - \Pi_1) = 0$ , i.e.  $(I - \Pi_X)(\Pi_X - \Pi_1) = 0$

It is obvious that

$$\Pi_X \Pi_1 = \Pi_1$$

Then

$$\begin{aligned} (I - \Pi_X)(\Pi_X - \Pi_1) &= \Pi_X - \Pi_1 - \Pi_X^2 + \Pi_X \Pi_1 \\ &= \Pi_X - \Pi_1 - \Pi_X + \Pi_1 \quad \because \text{idempotent} \\ &= 0 \end{aligned}$$



This completes the proof.  $\square$

**Proposition 1.24** (Independence). *SSE and  $(\hat{\beta}_0, \hat{\beta}_1)$  are independent, i.e.*

$$SSE \perp (\hat{\beta}_0, \hat{\beta}_1)^T$$

*Proof.* Note that

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^T = (X^T X)^{-1} X^T \mathbf{Y}$$

Since  $I - \Pi_X$  of  $SSE$  is symmetric, from Proposition 1.20,

Claim:  $((X^T X)^{-1} X^T)(\sigma^2 I)(I - \Pi_X) = 0$ , i.e.  $((X^T X)^{-1} X^T)(I - \Pi_X) = 0$

Since  $\Pi_X = X(X^T X)^{-1} X^T$ ,

$$\begin{aligned} ((X^T X)^{-1} X^T)(I - \Pi_X) &= (X^T X)^{-1} X^T - (X^T X)^{-1} X^T X (X^T X)^{-1} X^T \\ &= (X^T X)^{-1} X^T - (X^T X)^{-1} X^T \\ &= 0 \end{aligned}$$

This completes the proof.  $\square$

**Proposition 1.25** (Distribution of SST).

$$\frac{SST}{\sigma^2} \sim \chi^2(n-1, \delta)$$

$$\text{where } \delta = \sum_{i=1}^n (x_i - \bar{x})^2 \beta_1^2 = \frac{S_{xx} \beta_1^2}{\sigma^2}$$

*Proof.* It proceeds in a similiary way from Corollary 1.1

$$\frac{SST}{\sigma^2} = \left( \frac{\mathbf{Y}}{\sigma} \right)^T (I - \Pi_1) \left( \frac{\mathbf{Y}}{\sigma} \right)$$

Since  $I - \Pi_1$  is symmetric idempotent,

$$K = \text{rank}(I - \Pi_1) = \text{tr}(I - \Pi_1) = n - \text{rank}(\Pi_1) = n - 1$$

$$\begin{aligned} \delta &= \left( \frac{X\beta}{\sigma} \right)^T (I - \Pi_1) \left( \frac{X\beta}{\sigma} \right) \\ &= \frac{S_{xx} \beta_1^2}{\sigma^2} \quad \because (1.23) \text{ and } (1.24) \end{aligned}$$

$\square$

### 1.9.3 ANOVA for testing significance of regression

Recall that

$$SST = SSR + SSE$$

- $SST$ : the variation of a response itself
- $SSR$ : the variation of a response *explained by the model*
- $SSE$ : the variation of a response that *cannot be explained by the model*

As mentioned in section 1.5.4, whether the model is useful or not can depend on the proportion of  $SSR$  versus  $SSE$  in constant  $SST$ . When  $SSR$  is large compared to  $SSE$ , we can say that the model is good. On the other hand, when  $SSR$  is not large, the model might be poor. This is what  $R^2$  measures intuitively.

However, this direct comparison sometimes does not work in many times. Both  $SSR$  and  $SSE$  comes from different distribution, which have different degrees of freedom. So we *compare standardized versions*, i.e. divided by the degrees of freedom.

**Definition 1.11** (Degrees of freedom). Degrees of freedom of each sum of squares is

$$df = \text{the number of deviation} - \text{the number of linear constraints}$$

**Corollary 1.6** (df of SS). *df of each sum of square is computed as*

$$(a) df(SST) = n - 1$$

$$(b) df(SSR) = 1$$

$$(c) df(SSE) = n - 2$$

*Proof.* (a)

Since  $\sum(Y_i - \bar{Y}) = 0$ , we have 1 linear constraints. Thus,

$$df(SST) = n - 1$$

(b)

Note that  $\hat{Y}_i - \bar{Y} = \hat{\beta}_1(x_i - \bar{x})$

where  $\sum(x_i - \bar{x}) = 0$ .

Thus,

$$df(SSR) = n - (n - 1) = 1$$

(c)

From Example 1.1,  $\sum(Y_i - \hat{Y}_i) = 0$  and  $\sum x_i(Y_i - \hat{Y}_i) = 0$ .

Thus,

$$df(SSE) = n - 2$$

□

Dividing sum of squares in  $df$ , we can standardize it.

**Definition 1.12** (Mean square). Mean square is a sum of square  $SS$  divided by its degree of freedom  $df$

$$MS := \frac{SS}{df}$$

Using the values of corollary 1.6 we can define each mean square for  $SSR$  and  $SSE$ .

**Definition 1.13** (Regression mean square).

$$MSR := \frac{SSR}{1} = SSR$$

From Proposition 1.22, the following corollary can be drawn.

**Corollary 1.7** (Distribution of MSR). Under  $H_0 : \beta_1 = 0$ ,

$$\frac{SSR}{\sigma^2} \stackrel{H_0}{\sim} \chi^2(1)$$

Now standardize residual sum of square.

**Definition 1.14** (Residual mean square).

$$MSE := \frac{SSE}{n-2}$$

From Proposition 1.22, we can construct same statistic. In fact,  $\frac{SSE}{\sigma^2}$  follows  $\chi^2(n-2)$  whether or not  $\beta_1$  is zero. Its  $\delta = 0$ .

**Corollary 1.8** (Distribution of MSE).

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

Finally, we can now use Proposition 1.16 so that

$$F \equiv \frac{MSR}{MSE} = \frac{\frac{SSR/\sigma^2 \sim \chi^2(1)}{1}}{\frac{SSE/\sigma^2 \stackrel{H_0}{\sim} \chi^2(n-2)}{n-2}} \stackrel{H_0}{\sim} F(1, n-2)$$

By construction, this test statistic is used for

$$H_0 : \beta_1 = 0$$

which means that the predictor does not explain the response anything. In other words, we are testing that

$$H_0 : \text{Model is not useful at all} \quad \text{vs} \quad H_1 : \text{Model can explain data} \quad (1.25)$$

*Remark* (F statistic on testing significance). Null hypothesis (1.25) can be tested with  $F$ -statistic.

$$F_0 = \frac{MSR}{MSE} = \frac{SSR/df(SSR)}{SSE/df(SSE)} \stackrel{H_0}{\sim} F(df(SSR), df(SSE))$$

Then we reject  $H_0$  if

$$F_0 > F_\alpha(df(SSR), df(SSE))$$

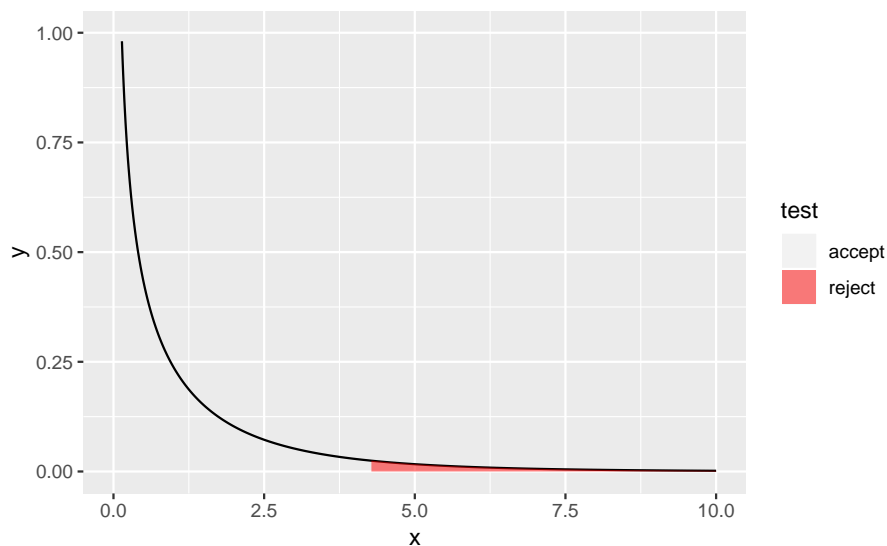


Figure 1.10: Rejection region for significance testing

```
summary(delv_fit)
```

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-7.581	-1.874	-0.349	2.181	10.634

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.321	1.371	2.42	0.024 *
x	2.176	0.124	17.55	8.2e-15 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.18 on 23 degrees of freedom

Multiple R-squared: 0.93, Adjusted R-squared: 0.927

F-statistic: 308 on 1 and 23 DF, p-value: 8.22e-15

This statistic is F-statistic included in `summary.lm()` output. This is saved as `$fstatistic`.

```
summary(delv_fit)$fstatistic
```

value	numdf	dendf
308	1	23

We usually summarize these statistic in table form, so called *ANOVA table*.

Source	SS	df	MS	F	p-value
Model	$SSR$	1	$MSR$	$F_0$	p-value
Error	$SSE$	$n - 2$	$MSE$		
Total	$SST$	$n - 1$			

To get this table, just use `anova()` for `lm` object.

```
anova(delv_fit)
```

Analysis of Variance Table

Response: y

```
      Df Sum Sq Mean Sq F value    Pr(>F)
x         1   5382     5382    308 8.2e-15 ***
Residuals 23    402       17
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the last Total row is just sum of the model and error, the function does not give it. To use this table as `data.frame` more easily, just implement `broom::tidy` as before.

```
anova(delv_fit) %>%
  broom::tidy()
```

```
# A tibble: 2 x 6
  term      df sumsq meansq statistic  p.value
<chr>   <int> <dbl>  <dbl>    <dbl>   <dbl>
1 x         1 5382.  5382.    308. 8.22e-15
2 Residuals 23  402.   17.5     NA    NA
```

Denote that here *simple linear regression setting*  $F$ -statistic and  $t$ -statistic of Equation (1.22) perform exactly same thing,  $H_0 : \beta_1 = 0$ . In fact, we know that

$$F(1, k) \stackrel{d}{=} T_k^2$$

*Remark.* In the simple linear regression setting,  $F$ -test for significance and  $t$ -test for no slope are equivalent, i.e. under  $H_0 : \beta_1 = 0$

$$F_0 = \frac{\hat{\beta}_1 S_{xx}}{\sigma^2} = \left( \frac{\hat{\beta}_1}{\sigma / \sqrt{S_{xx}}} \right)^2 = T_0^2$$



## Chapter 2

# Multiple Linear Regression

### 2.1 Model

```
(cem <- MPV::cement %>% tbl_df())
```

```
# A tibble: 13 x 5
   y    x1    x2    x3    x4
<dbl> <dbl> <dbl> <dbl> <dbl>
1  78.5     7    26     6    60
2  74.3     1    29    15    52
3 104.    11    56     8    20
4  87.6    11    31     8    47
5  95.9     7    52     6    33
6 109.    11    55     9    22
7 103.     3    71    17     6
8  72.5     1    31    22    44
9  93.1     2    54    18    22
10 116.    21    47     4    26
11  83.8     1    40    23    34
12 113.    11    66     9    12
13 109.    10    68     8    12
```

Above is a data set about cement and concerning four ingredients from the Montgomery et al. (2015) textbook.

- **y**: heat evolved in calories per gram of cement
- **x1**: tricalcium aluminate
- **x2**: tricalcium silicate
- **x3**: tetracalcium alumino ferrite
- **x4**: dicalcium silicate

Given data  $(x_{11}, x_{12}, \dots, x_{1p}, Y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{np}, Y_n)$  ( $p = 4$ ), we try to fit linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i$$

with

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

Compared to simple linear regression problem 1, we have more parameters for coefficients

$$(\beta_0, \beta_1, \dots, \beta_p, \sigma^2)$$

Each  $\beta_j$  is a change of  $Y$  when each predictor variable  $x_j$  increases in 1 unit while the others fixed. In this part, we use *matrix notation*. Extending our former matrix work 1.6,

$$\begin{array}{c} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\ \mathbf{Y} \end{array} = \begin{array}{c} \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \\ X \end{array} \begin{array}{c} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \\ \boldsymbol{\beta} \end{array} + \begin{array}{c} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \\ \boldsymbol{\epsilon} \end{array}$$

where  $\epsilon_i$  are i.i.d., and

$$E\boldsymbol{\epsilon} = \mathbf{0}$$

$$Var\boldsymbol{\epsilon} = \sigma^2 I$$

## 2.2 Least Square Estimation

Write  $\boldsymbol{\beta} \equiv (\beta_0, \dots, \beta_p)^T \in \mathbb{R}^{p+1}$ . Extend Equation (1.16).

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \\ &= \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{Y} - \beta_0 \mathbf{1} - \beta_1 \mathbf{x}_1 - \cdots - \beta_p \mathbf{x}_p\|^2 \\ &= \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \end{aligned} \tag{2.1}$$

As discussed, the solution  $\hat{\boldsymbol{\beta}}$  is related to the projection.  $X\hat{\boldsymbol{\beta}}$  is a projection of  $\mathbf{Y}$  onto  $Col(X)$ .

### 2.2.1 Normal equation

Now recap the section 1.6.3. Fundamental subspaces theorem 1.4 implies that

$$\mathbf{Y} - X\hat{\boldsymbol{\beta}} \in Col(X)^\perp = N(X^T)$$

From the second part of subset, i.e.  $N(X^T)$ , we now have *Normal equation*

$$X^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{2.2}$$

This is equivalent to

$$X^T \mathbf{Y} = X^T X \hat{\boldsymbol{\beta}}$$



Hence, if  $X^T X$  is invertible, the equation gives unique solution

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$$

Our first question is when  $X^T X$  is invertible, and Theorem 1.8 have said that it is when the model matrix  $X$  is full rank.

**Lemma 2.1.** *Let  $X \in \mathbb{R}^{n \times (p+1)}$  be any model matrix. Then  $X^T X$  is always non-negative definite.*

$$\forall \mathbf{v} \in \mathbb{R}^{p+1} : \mathbf{v}^T (X^T X) \mathbf{v} \geq 0$$

*Proof.* Let  $\mathbf{v} \in \mathbb{R}^{p+1}$ . Then

$$\mathbf{v}^T (X^T X) \mathbf{v} = (X \mathbf{v})^T (X \mathbf{v}) = \|X \mathbf{v}\|^2 \geq 0$$

□

This lemma can also prove our Theorem 1.8.

**Theorem 2.1.** *Let  $\mathbf{Y} = X\beta$  inconsistent and let  $X \in \mathbb{R}^{n \times (p+1)}$  with  $n > p + 1$ .*

*If  $\text{rank}(X) = p + 1$ , i.e. full rank, then  $X^T X$  is invertible.*

*Proof.* Let  $\mathbf{c} \in \mathbb{R}^{(p+1)}$

Suppose that  $X^T X$  is positive definite.

$$\begin{aligned} \Leftrightarrow \mathbf{c}^T X^T X \mathbf{c} = 0 & \text{ implies } \mathbf{c} = \mathbf{0} \\ \Leftrightarrow X \mathbf{c} = \mathbf{0} & \text{ implies } \mathbf{c} = \mathbf{0} \\ \Leftrightarrow \text{columns of } X & \text{ linearly independent} \\ \Leftrightarrow \text{rank}(X) = p + 1 \end{aligned}$$

□

### 2.2.2 Orthogonal decomposition

**Theorem 2.2.** *Let  $\text{Col}(X)$  be a subspace of  $\mathbb{R}^n$ , let  $\mathbf{Y} \in \mathbb{R}^n$ , and let  $\{\mathbf{u}_0, \dots, \mathbf{u}_p\}$  be an orthonormal basis for  $\text{Col}(X)$ . If*

$$\hat{\mathbf{Y}} = \sum_{j=0}^p \hat{\beta}_j \mathbf{u}_j$$

where

$$\hat{\beta}_j = \Pi(\mathbf{Y} \mid R(\mathbf{u}_j)) \quad \text{for each } i$$

then  $\hat{\mathbf{Y}} - \mathbf{Y} \in \text{Col}(X)^\perp$ .

**Theorem 2.3.** *Under the hypothesis of Theorem 2.2,  $\hat{\mathbf{Y}} \in \text{Col}(X)$  is the closest to  $\mathbf{Y}$  amongst its any element  $\mathbf{p}$ , i.e.*

$$\|\mathbf{p} - \mathbf{Y}\| > \|\hat{\mathbf{Y}} - \mathbf{Y}\|$$

for any  $\mathbf{p} \neq \hat{\mathbf{Y}}$  in  $\text{Col}(X)$

In other words, projection of  $\mathbf{Y}$  onto  $\text{Col}(X)$ ,  $\hat{\mathbf{Y}}$  can be represented as sum of projections of  $\mathbf{Y}$  onto each (orthogonal) individual variable. Before looking at individual basis, consider two-block space.

Write

$$X = \left[ \begin{array}{c|ccc} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{array} \right] = [\mathbf{1}, \mathbb{X}_A]$$

Consider  $R(X)$ ,  $R(\mathbf{1})$ , and  $R(\mathbb{X}_A)$ .

To decompose subspace  $R(X)$ , we try to orthogonalize  $\mathbf{1}$  and  $\mathbb{X}_A$ . By Theorem 2.2, we have

$$\mathbf{1} \perp \mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A$$

In fact, the right one  $\mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A$  is the *residual after simple linear regression*  $\mathbb{X}_A$  onto  $\mathbf{1}$ . We have seen in Figure 1.6 of section 1.6 that the *residual is orthogonal to predictor vector*. In this procedure, we choose residual as new predictor instead of response in simple linear regression, i.e.  $\mathbb{X}_A$ . If this is done to individual predictor variables, it is called *successive orthogonalization* and it will be covered next section with QR decomposition.

Theorem 1.6 implies that

$$R(X) = R(\mathbf{1}) \oplus R(\mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A)$$

**Theorem 2.4** (Orthogonal decomposition). *Let  $X = [\mathbf{1}, \mathbb{X}_A]$ . Then*

(i)

$$R(X) = R(\mathbf{1}) \oplus R(\mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A)$$

(ii)

$$\Pi(\cdot \mid R(X)) = \Pi(\cdot \mid R(\mathbf{1})) + \Pi(\cdot \mid R(\mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A))$$

Write

$$\mathbb{X}_{A,\perp} := \mathbb{X}_A - \Pi_{\mathbf{1}} \mathbb{X}_A$$

Note that

$$\Pi_{\mathbf{1}} = \mathbf{1}(\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T = \frac{1}{n} \mathbf{1} \mathbf{1}^T$$

Then

$$\begin{aligned}
X\hat{\beta} &= \hat{\beta}_0 \mathbf{1} + \mathbb{X}_A \hat{\beta}_A \\
&= \hat{\beta}_0 \mathbf{1} + (\mathbb{X}_{A,\perp} + \Pi_1 \mathbb{X}_A) \hat{\beta}_A \\
&= \left( \hat{\beta}_0 + \frac{1}{n} \mathbf{1}^T \mathbb{X}_A \hat{\beta}_A \right) \mathbf{1} + \mathbb{X}_{A,\perp} \hat{\beta}_A \quad \because \hat{\beta}_0 + \frac{1}{n} \mathbf{1}^T \mathbb{X}_A \hat{\beta}_A \in \mathbb{R}
\end{aligned} \tag{2.3}$$

From (ii) of Theorem 2.4,

$$\begin{aligned}
\Pi(\mathbf{Y} \mid R(X)) &= \Pi(\mathbf{Y} \mid R(\mathbf{1})) + \Pi(\mathbf{Y} \mid R(\mathbb{X}_{A,\perp})) \\
&= \bar{Y} \mathbf{1} + \mathbb{X}_{A,\perp} (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1} \mathbb{X}_{A,\perp}^T \mathbf{Y}
\end{aligned} \tag{2.4}$$

Since  $\mathbf{1} \perp \mathbb{X}_{A,\perp}$ , Equations (2.3) and (2.4) imply that

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \frac{1}{n} \mathbf{1}^T \mathbb{X}_A \hat{\beta}_A \\ \hat{\beta}_A = (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1} \mathbb{X}_{A,\perp}^T \mathbf{Y} \end{cases}$$

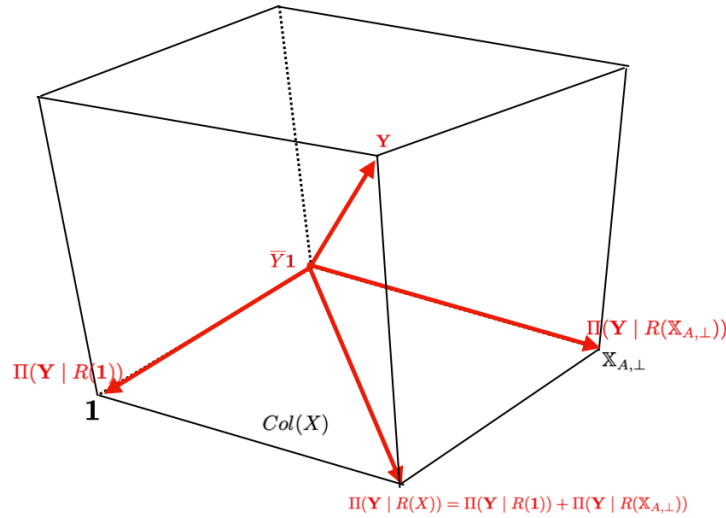


Figure 2.1: Orthogonal decomposition of the column space and LSE

See Figure 2.1. Two are orthogonal, so sum of projections onto them become LSE. In fact, *each projection indicate each regression coefficient*. When we do not have orthogonal basis, however, each projection is nothing.

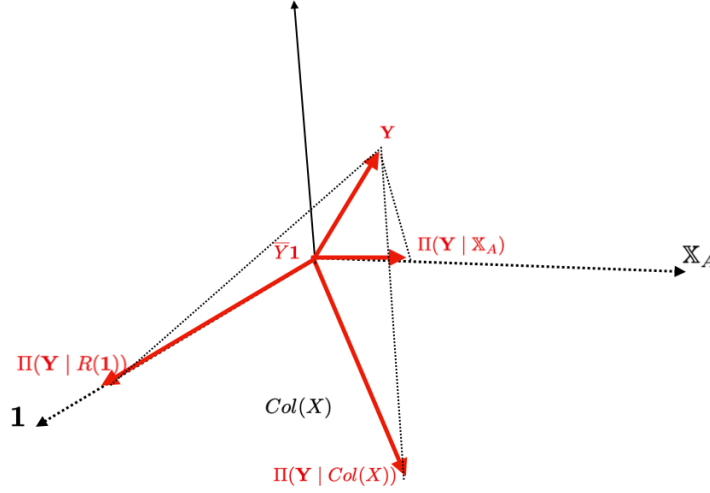


Figure 2.2: Non-orthogonality

In this situation, we have to do orthogonalization.

$$\tilde{X}_A = \Pi_1 X_A + (X_A - \Pi_1 X_A)$$

### 2.2.3 Gram-Schmidt QR factorization

Let's briefly look at orthogonalization process. From Theorem 2.2, we can derive following *orthonormalization process*.

**Theorem 2.5** (Gram-Schmidt Process). *Let  $\{\mathbf{x}_1, \dots, \mathbf{x}_{p+1}\}$  be a basis for the inner product space  $V$ . Let*

$$\mathbf{u}_1 = \left( \frac{1}{\|\mathbf{x}_1\|} \right) \mathbf{x}_1$$

*and define next  $\mathbf{u}_2, \dots, \mathbf{u}_{p+1}$  recursively by*

$$\mathbf{u}_{k+1} = \frac{1}{\|\mathbf{x}_{k+1} - \mathbf{r}_k^*\|} (\mathbf{x}_{k+1} - \mathbf{r}_k^*)$$

*for  $k = 1, \dots, p$ , where*

$$\mathbf{r}_k^* = \langle \mathbf{x}_{k+1}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{x}_{k+1}, \mathbf{u}_2 \rangle \mathbf{u}_2 + \dots + \langle \mathbf{x}_{k+1}, \mathbf{u}_k \rangle \mathbf{u}_k$$

*is the projection of  $\mathbf{x}_{k+1}$  onto  $sp(\{\mathbf{u}_1, \dots, \mathbf{u}_k\})$ .*

*Hence, we get  $\{\mathbf{u}_1, \dots, \mathbf{u}_{p+1}\}$  is an orthonormal basis for  $V$ .*

Our interest is  $Col(X)$ , and we can factorize this model matrix so that it represents orthonormalization process 2.5.

**Theorem 2.6** (Gram-Schmidt QR factorization). *Let  $X \in \mathbb{R}^{n \times (p+1)}$ . Then  $X$  can be factored into*

$$X = QR$$

where  $Q \in \mathbb{R}^{n \times (p+1)}$  is an orthogonal matrix, i.e. its column vectors are orthonormal and  $R \in \mathbb{R}^{(p+1) \times (p+1)}$  is an upper triangular matrix whose diagonal entries are all positive.

*Proof.* Denote that this is just the representation of Gram-schmidt orthogonalization. Then it gives

$$\begin{aligned} \mathbf{u}_1 &= \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \Rightarrow \mathbf{x}_1 = \|\mathbf{x}_1\| \mathbf{u}_1 \\ \mathbf{v}_2 &= \mathbf{x}_2 - \langle \mathbf{x}_2, \mathbf{u}_1 \rangle \mathbf{u}_1, \quad \mathbf{u}_2 = \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|} \\ \Rightarrow \mathbf{x}_2 &= \langle \mathbf{x}_2, \mathbf{u}_1 \rangle \mathbf{u}_1 + \|\mathbf{v}_2\| \mathbf{u}_2 \\ \Rightarrow \mathbf{x}_2 &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \langle \mathbf{x}_2, \mathbf{u}_1 \rangle \\ \|\mathbf{v}_2\| \end{bmatrix} \end{aligned}$$

It proceeds in a similar way to the others. Hence,

$$\begin{aligned} X &= [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_{p+1}] \\ &= [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_{p+1}] \begin{bmatrix} \|\mathbf{v}_1\| & \langle \mathbf{x}_2, \mathbf{u}_1 \rangle & \langle \mathbf{x}_3, \mathbf{u}_1 \rangle & \cdots & \langle \mathbf{x}_{p+1}, \mathbf{u}_1 \rangle \\ 0 & \|\mathbf{v}_2\| & \langle \mathbf{x}_3, \mathbf{u}_2 \rangle & \cdots & \langle \mathbf{x}_{p+1}, \mathbf{u}_2 \rangle \\ 0 & 0 & \|\mathbf{v}_3\| & \cdots & \langle \mathbf{x}_{p+1}, \mathbf{u}_3 \rangle \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \|\mathbf{v}_{p+1}\| \end{bmatrix} \\ &\equiv QR \end{aligned} \tag{2.5}$$

□

Look again the equation in Theorem 2.5. In each process  $k$ , the projection is done to the  $(k-1)$ -dimensional space. In other words, as process goes through, dimension increases. So we try to project each vector only in 1-dimension each step.

**Theorem 2.7** (Modified Gram-Schmidt Process). *Let  $\{\mathbf{x}_1, \dots, \mathbf{x}_{p+1}\}$  be a basis for the inner product space  $V$  and let  $\{\mathbf{q}_1, \dots, \mathbf{q}_{p+1}\}$  be an orthonormal basis.*

*Set  $\mathbf{q}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|}$ . Then consider  $sp(\{\mathbf{q}_1\})$ .*

*In the first step, make every  $\{\mathbf{x}_2, \dots, \mathbf{x}_{p+1}\}$  orthogonal to  $\mathbf{q}_1$ .*

$$\mathbf{x}_k^{(1)} = \mathbf{x}_k - (\mathbf{q}_1^T \mathbf{x}_k) \mathbf{q}_1, \quad k = 2, \dots, p+1$$

*So we get orthogonal set  $\{\mathbf{q}_1, \mathbf{x}_2^{(1)}, \dots, \mathbf{x}_{p+1}^{(1)}\}$ . Next, set  $\mathbf{q}_2 = \frac{\mathbf{x}_2^{(1)}}{\|\mathbf{x}_2^{(1)}\|}$ . Consider  $sp(\{\mathbf{q}_2\})$ . Since we have  $\mathbf{q}_1 \perp \mathbf{q}_2$ ,*

$$\mathbf{x}_k^{(2)} = \mathbf{x}_k^{(1)} - (\mathbf{q}_2^T \mathbf{x}_k^{(1)}) \mathbf{q}_2 \perp \mathbf{q}_2, \quad k = 3, \dots, p+1$$

*Thus, get  $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{x}_3^{(2)}, \dots, \mathbf{x}_{p+1}^{(2)}\}$ .  $\mathbf{q}_3, \dots, \mathbf{q}_{p+1}$  are successively determined in a similiary way.*

At the last step, set

$$\mathbf{q}_{p+1} = \frac{\mathbf{x}_{p+1}^{(p)}}{\|\mathbf{x}_{p+1}^{(p)}\|}$$

Since each projection is done in 1-dimension, the algorithm becomes more understandable. Consider

$$Q = [\mathbf{q}_1 \quad \mathbf{q}_2 \quad \cdots \quad \mathbf{q}_{p+1}] \in \mathbb{R}^{n \times (p+1)} \quad \text{orthogonal}$$

and

$$R = [r_{kj}] = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1,p+1} \\ 0 & r_{22} & \cdots & r_{2,p+1} \\ 0 & 0 & \cdots & r_{3,p+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & r_{p+1,p+1} \end{bmatrix} \in \mathbb{R}^{(p+1) \times (p+1)}$$

We can perform  $QR$  factorization by following step.

```

1 for  $k \leftarrow 1$  to  $(p+1)$  do
2    $r_{kk} = \|\mathbf{x}_k\|$ ;
3    $\mathbf{q}_k = \frac{\mathbf{x}_k}{r_{kk}}$ ;
4   for  $j \leftarrow 1$  to  $(p+1)$  do
5      $r_{kj} = \mathbf{q}_k^T \mathbf{x}_j$ ;
6      $\mathbf{x}_j = \mathbf{x}_j - r_{kj} \mathbf{q}_k$ ;
7   end
8 end
```

**Algorithm 1:** QR decomposition for modified G-S process

This *orthonormal basis* gives some useful facts with least squares problem (Leon, 2014).

### 2.2.4 Successive orthogonalization

In fact, G-S process 2.5 is equivalent to successive orthogonalization, i.e. regress(project)  $\mathbf{x}_j$  onto the others (Hastie et al., 2013).

```

1 Initialize  $\mathbf{v}_0 = \mathbf{1}$ ;
2 for  $k \leftarrow 1$  to  $p$  do
3   Regress  $\mathbf{x}_k$  on  $\mathbf{q}_0, \dots, \mathbf{q}_{k-1}$ ;
4    $\hat{\beta}_{lk} = \frac{\langle \mathbf{v}_l, \mathbf{x}_k \rangle}{\langle \mathbf{v}_l, \mathbf{v}_l \rangle}, l = 0, \dots, k-1$ ;
5   Residual  $\mathbf{v}_k = \mathbf{x}_k - \sum_{l=0}^{k-1} \hat{\beta}_{lk} \mathbf{v}_l$ ;
6 end
7 Regress  $\mathbf{Y}$  on  $\mathbf{v}_p$ 
```

**Algorithm 2:** Successive orthogonalization

Now we can solve least squares problem using QR decomposition. Recall that

$$X = QR$$

as specified in Theorem 2.6. Then normal equation implies that

$$\begin{aligned}
 (X^T X) \hat{\beta} &= X^T \mathbf{Y} \\
 \Leftrightarrow R^T Q^T Q R \hat{\beta} &= R^T Q^T \mathbf{Y} \\
 \Leftrightarrow R^T R \hat{\beta} &= R^T Q^T \mathbf{Y} \quad \because Q^T Q = I \\
 \Leftrightarrow R \hat{\beta} &= Q^T \mathbf{Y} \quad \text{if } R \text{ is invertible}
 \end{aligned}
 \tag{2.6}$$

Hence,

$$\hat{\beta} = R^{-1} Q^T \mathbf{Y} \tag{2.7}$$

It follows that

$$\hat{\mathbf{Y}} = (QR) \hat{\beta} = Q Q^T \mathbf{Y} \tag{2.8}$$

Let's compare the result. Base function `qr()` give the QR factorization. Given this object, we can get each  $Q$  and  $R$  by `qr.Q()` and `qr.R()`.

```
cem_qr <-
  cem %>%
  model.matrix(y ~ ., data = .) %>%
  qr()
cem_q <- qr.Q(cem_qr)
cem_r <- qr.R(cem_qr)
```

Using Equation (2.7), we get each coefficient as follow.

```
solve(cem_r) %*% t(cem_q) %*% cem$y
```

```
      [,1]
(Intercept) 62.405
x1          1.551
x2          0.510
x3          0.102
x4         -0.144
```

On the other hand, `lm()` gives the following result.

```
lm(y ~ ., data = cem)
```

Call:

```
lm(formula = y ~ ., data = cem)
```

Coefficients:

```
(Intercept)      x1      x2      x3      x4
    62.405    1.551    0.510    0.102   -0.144
```

We can check the result is same. In fact, `lm()` fits the model by default `method = "qr"`.

the method to be used; for fitting, currently only `method = "qr"` is supported; `method = "model.frame"` returns the model frame (the same as with `model = TRUE`, see below).

By default and only way, `lm()` fits the model using  $QR$  factorization. What does this orthogonal basis mean? For simplicity, consider simple linear regression problem.

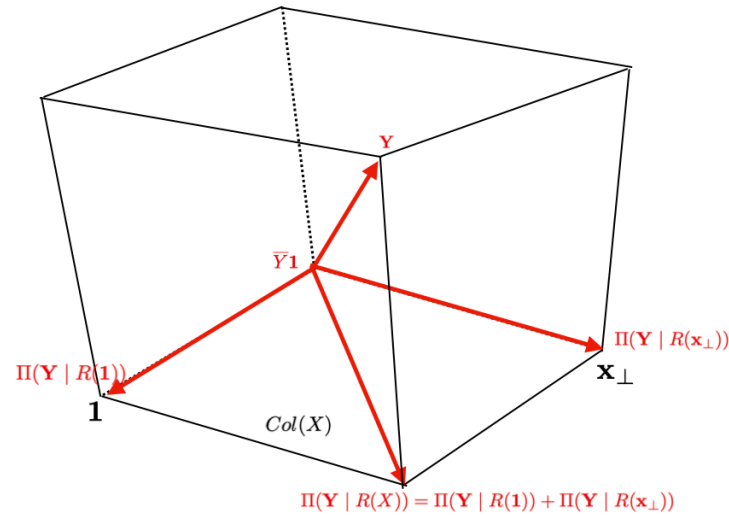


Figure 2.3: Orthogonalized basis

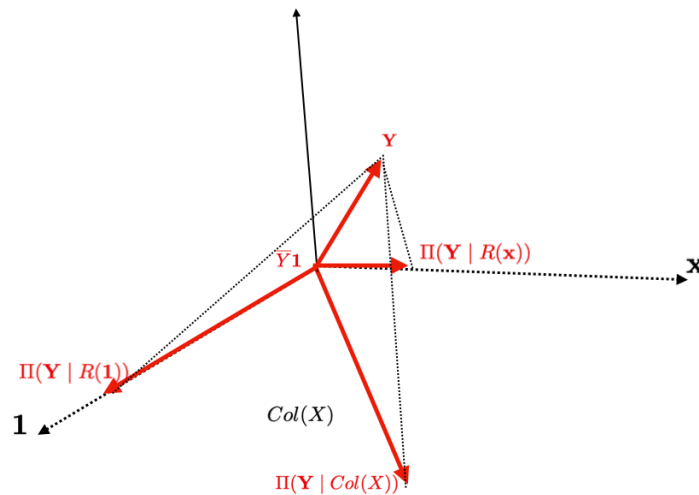


Figure 2.4: Non-orthogonal basis

See Figure 2.3. By construction, projection onto each basis is same as  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . In Figure 2.4, however, each projection is not regression coefficient.



# Bibliography

- Hastie, T., Tibshirani, R., and Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
- Hogg, R. V., McKean, J. W., and Craig, A. T. (2018). *Introduction to Mathematical Statistics*. Pearson College Division, 8 edition.
- Johnson, R. A. and Wichern, D. W. (2013). *Applied Multivariate Statistical Analysis*.
- Leon, S. (2014). *Linear Algebra with Applications*. Pearson Higher Ed.
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2015). *Introduction to Linear Regression Analysis*. John Wiley & Sons.