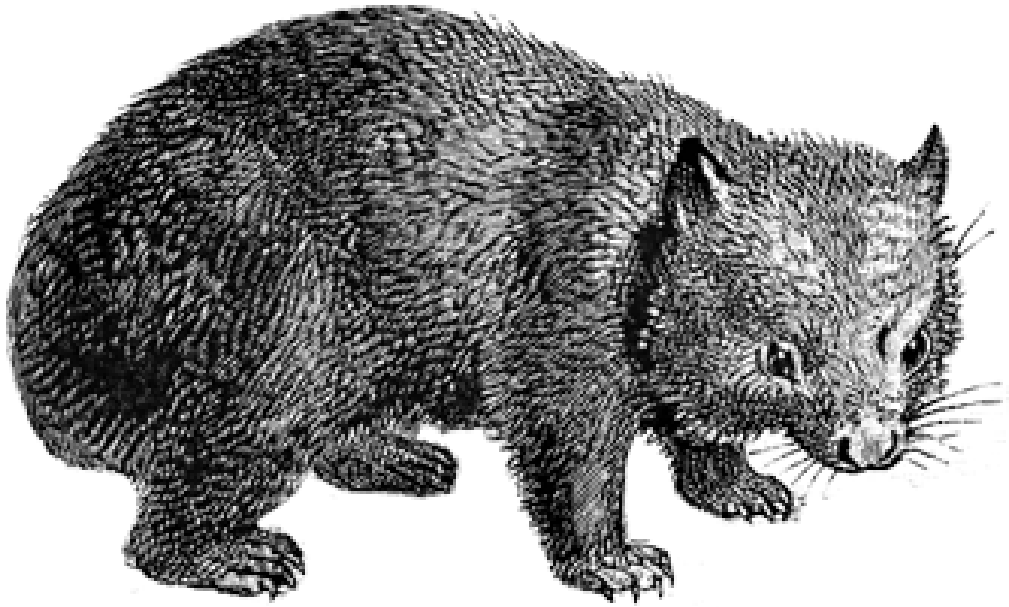
YOUNG-GEUN statistics



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

02 May, 2019

Contents

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,

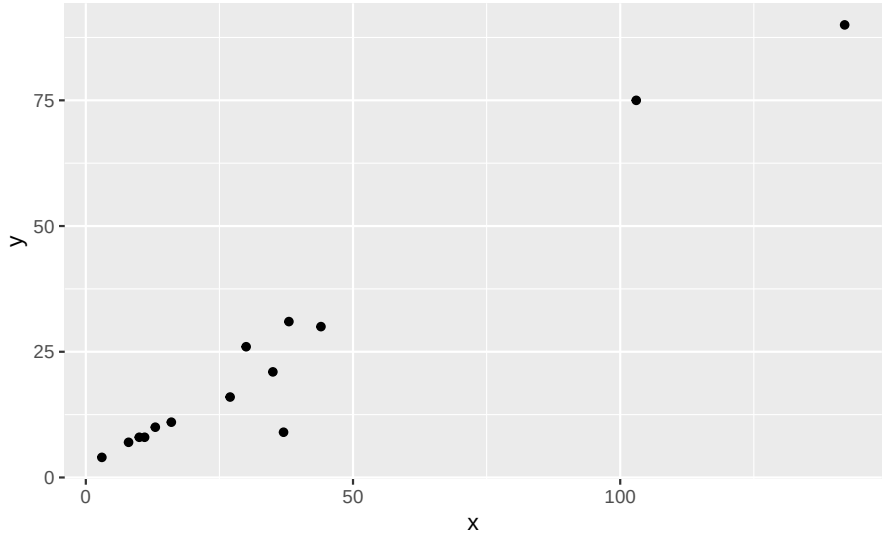
$$y = \beta_0 + \beta_1 x$$

- Statistical relation: embeded with noise

So we try to estimate

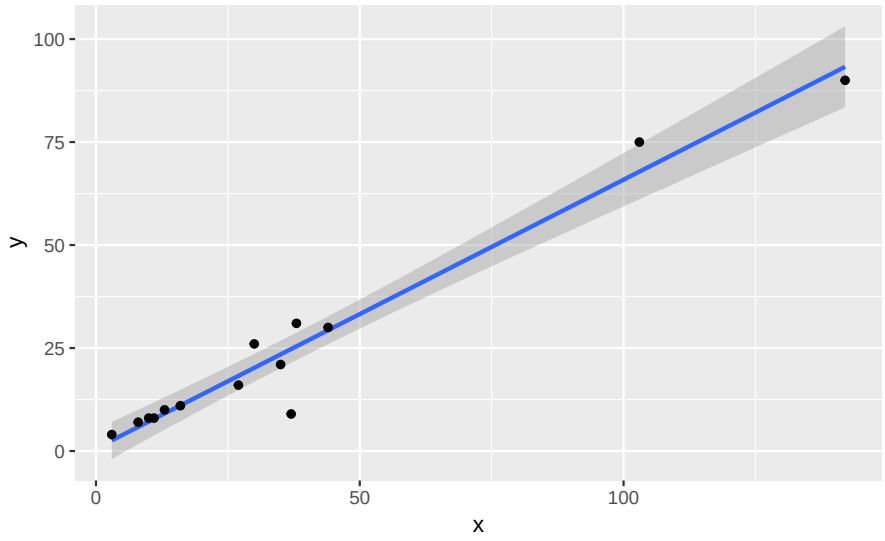
$$y = \beta_0 + \beta_1 x + \epsilon$$

```
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"
  )
```

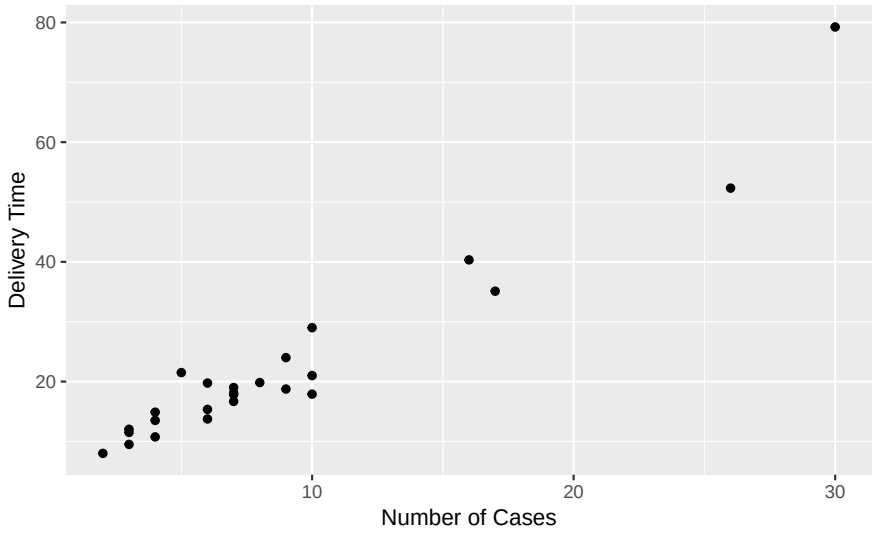


Figure 1.1: The Delivery Time Data

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

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

Here ϵ_i is a error term, which is a random variable.

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

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

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

1.2 Least Squares Estimation

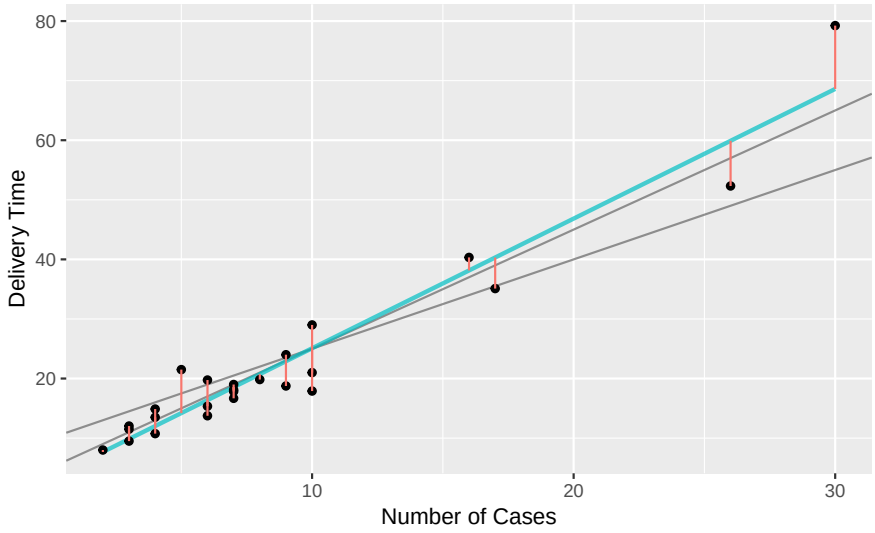


Figure 1.2: Idea of the least square estimation

We try to find β_0 and β_1 that minimize the sum of squares of the vertical distances, i.e.

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

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

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

and

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

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,

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

Equation (1.3) gives

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

Thus,

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

Remark.

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

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

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

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

It follows that

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

□

```
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 β_0 and β_1 is same as estimating the *assumed true model*.

Definition 1.1 (Mean response)**.**

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

We can estimate this mean resonse by

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

However, in practice, the model might not be true, which is included in ϵ term.

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

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

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

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 β_0 and β_1 have some properties related to the expectation and variance. We can notice that these lse's are **unbiased linear estimator**. In fact, these are the *best unbiased linear estimator*. This will be covered in the Gauss-Markov theorem.

Lemma 1.1.

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

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

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

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

For the first part of S_{XY},

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

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

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

Same for the last part. □

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

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

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

Proof. From lemma 1.1,

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

It gives that

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

□

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

1. $E\hat{\beta}_1 = \beta_1$
2. $E\hat{\beta}_0 = \beta_0$

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

Proof. From lemma 1.1,

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

It follows that

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

□

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

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

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

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

Proof. Proving is just arithmetic.

(a)

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

(b)

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

(c)

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

□

1.2.4 Gauss-Markov Theorem

Chapter 1.2.3 shows that the β_0^{LSE} and β_1^{LSE} are the **linear unbiased estimators**. Are these good? Good compared to *what estimators*? Here we consider *linear unbiased estimator*. If variances in the proposition 1.2 are lower than any parameters in this parameter family, β_0^{LSE} and β_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.*

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

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

Bestness of β_1. Consider $\Theta := \left\{ \sum_{i=1}^{n} b_i Y_i \in \mathbb{R} : E\Big(\sum_{i=1}^{n} b_i Y_i \Big) = \beta_1 \right\}$.

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

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

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

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

It gives

$$\begin{cases} \sum b_i = 0 \\ \sum b_i x_i = 1 \end{cases} \tag{1.7}$$

Then

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

Hence,

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

□

Bestness of β_0. Consider $\Theta := \left\{ \sum_{i=1}^{n} a_i Y_i \in \mathbb{R} : E\Big(\sum_{i=1}^{n} a_i Y_i \Big) = \beta_0 \right\}$.

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

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

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

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

It gives

$$\begin{cases} \sum a_i = 1 \\ \sum a_i x_i = 0 \end{cases} \tag{1.8}$$

Then

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

Hence,

$$Var(\sum a_iY_i) \ge Var(\hat{\beta}_0)$$

□

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 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1x_i$.

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

$$
\begin{aligned}
\sum \hat{Y}_i(Y_i - \hat{Y}_i) &= \sum(\hat{\beta}_0 + \hat{\beta}_1x_i)(Y_i - \hat{Y}_i) \\
&= \hat{\beta}_0\sum(Y_i - \hat{Y}_i) + \hat{\beta}_1\sum x_i(Y_i - \hat{Y}_i) \\
&= 0
\end{aligned}
$$

1.2.5 Estimation of σ^2

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

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1x_i)^2 \tag{1.9}$$

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

Proposition 1.3 (Unbiasedness)**.**

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

Proof. Note that

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

Then

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

(a)

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

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

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

Thus,

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

(b)

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

It follows that

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

□

1.3 Maximum Likelihood Estimation

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

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

Example 1.2 (Gaussian Likelihood)**.** Note that $Y_i \overset{indep}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Then the likelihood function is

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

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.

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

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

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

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, howerver, it is *asymptotically unbiased*.

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

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

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

To apply this theorem 1.2 in the simple linear regression setting, i.e. (β_0, β_1), we need to look at the *bivariate case*.

Theorem 1.3 (Rao-Cramer Lower Bound, bivariate case)**.** *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta 1, \theta_2)$ *and let* $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$*. If each* $\hat{\theta}_1$*,* $\hat{\theta}_2$ *is an unbiased estimator of* θ_1 *and* θ_2*, then*

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

where

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

Assume that σ^2 is **known**. From the Equations (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,

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

Hence,

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

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

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

This result is *stronger than Gauss-Markov theorem* 1.1, where the LSE has the smalleset variance among all *linear unbiased* estimators. It can be simply obtained from the *Lehmann-Scheffe Theorem*: If some unbiased estimator is a function of complete sufficient statistic, then this estimator is the unique MVUE (Hogg et al., 2018).

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

Proof. From the example 1.2,

$$\begin{aligned}
L(\beta_0,\beta_1)&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\Big[-\frac{1}{2\sigma^2}\sum(Y_i-\beta_0-\beta_1x_i)^2\Big]\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\Big[-\frac{1}{2\sigma^2}\sum\Big(Y_i^2-(\beta_0+\beta_1x_i)Y_i+(\beta_0+\beta_1x_i)^2\Big)\Big]\\
&=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\Big[-\frac{1}{2\sigma^2}\Big(-\beta_0\sum Y_i-\beta_1\sum x_iY_i\Big)\Big]\exp\Big[-\frac{1}{2\sigma^2}\Big(\sum Y_i^2+(\beta_0+\beta_1x_i)^2\Big)\Big]
\end{aligned}$$

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"
  )
```

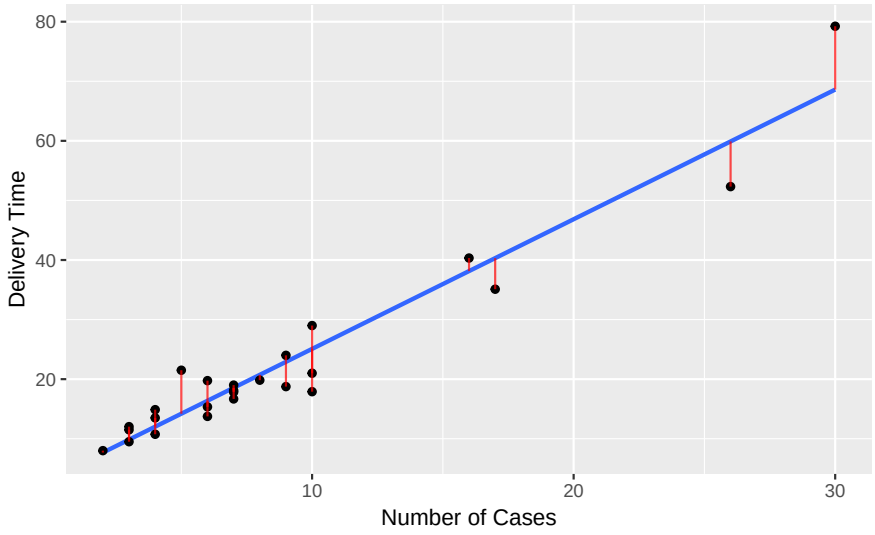


Figure 1.3: Fit and residuals

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

Lemma 1.3. *Covriance between Y and each coefficient*

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

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

Proof. (a)

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

(b)

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

□

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

(a) $E(e_i) = 0$

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

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

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

(b) Lemma 1.3 implies that

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

Then

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

From the same lemma 1.3,

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

These Equations (1.13) and (1.14) give that

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

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

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

□

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)**.**

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

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

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

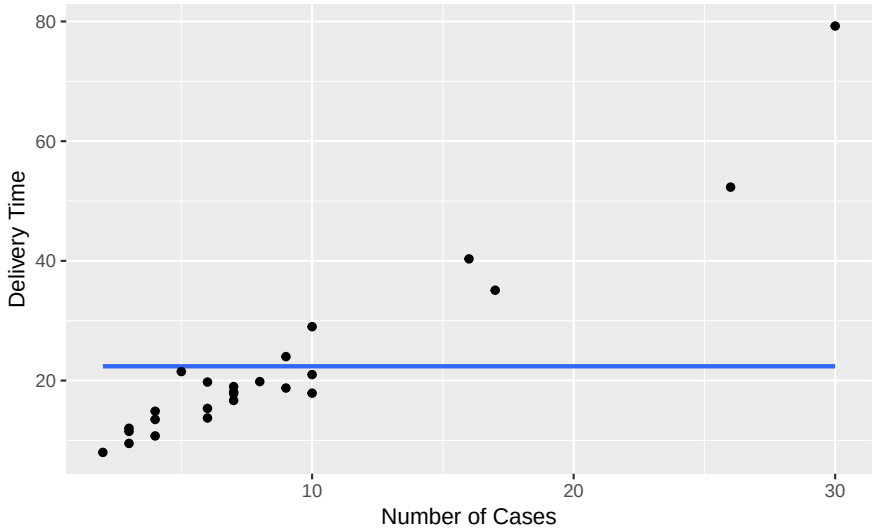


Figure 1.4: Regression without predictor

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,

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

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

1.5.2 Regression sum of squares

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

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

This SSR compares $\hat{Y}_i$ versus $\overline{Y}$, 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 $SST = \sum(Y_i - \overline{Y})^2$*,* $SSR = \sum(\hat{Y}_i - \overline{Y})^2$*, and* $SSE = \sum(Y_i - \hat{Y}_i)$

Proof. From the Example 1.1,

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

□

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

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

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)**.**

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

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.*

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

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

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

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

It follows that

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

□

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

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

The Column space is defined by

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

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 $R(X) = Col(X)$ and $R(X^T) = Col(X^T)$.

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

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

Thus,

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

and so

$$N(X) \subseteq Row(X)^\perp$$

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

$$Row(X)^\perp \subseteq N(X)$$

and so

$$N(X) = Row(X)^\perp$$

□

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

$$dimS + dimS^\perp = n$$

If $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$ is a basis for S and $\{\mathbf{x}_{r+1}, \ldots, \mathbf{x}_n\}$ is a basis for $S^\perp$, then $\{\mathbf{x}_1, \ldots, \mathbf{x}_r, \mathbf{x}_{r+1}, \ldots, \mathbf{x}_n\}$ 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 $\mathbf{y} \in \mathbf{R}^n$, 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

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

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

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

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.

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

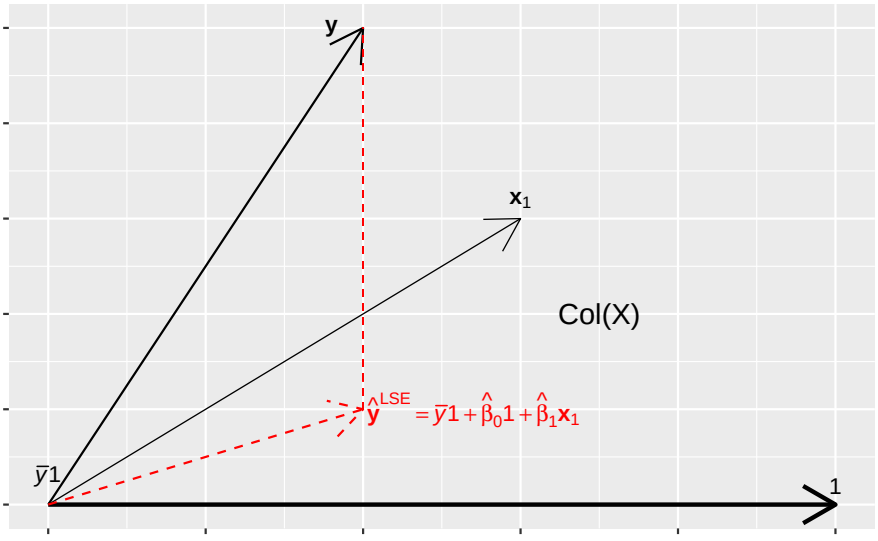


Figure 1.5: Geometric Illustration of Simple Linear Regression

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

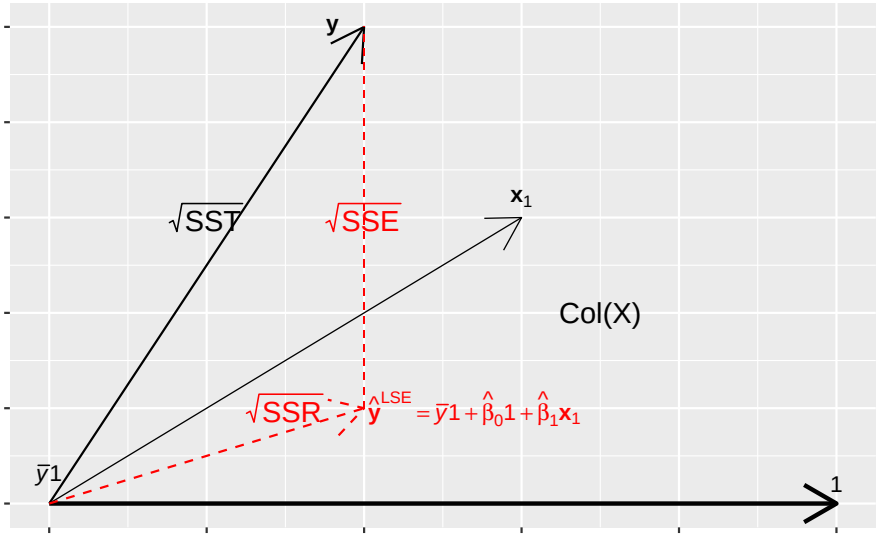


Figure 1.6: Geometric Illustration of Decomposing SST

See Figure 1.6.

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

Pythagorean law implies that

$$SST = SSR + SSE$$

Also,

$$R^2 = \frac{SSR}{SST} = cos^2\theta = \hat{\rho}_{XY}^2 \tag{1.17}$$

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 $\boldsymbol{\beta} \equiv (\beta_0, \beta_1)^T$. By the fundamental subspaces theorem 1.4,

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

Thus,

$$X^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \mathbf{0} \tag{1.18}$$

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

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

If X^TX is nonsingular,

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

It follows that

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

Combining this equation and our figure, we can know that $X(X^TX)^{-1}X^T$ 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

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

It can also be called ***Hat matrix*** and written as H.

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

Theorem 1.8. *Let $\mathbf{Y} = X\boldsymbol{\beta}$ 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^TX is invertible.

Proof. Suppose that $(X^TX)\mathbf{b} = \mathbf{0}$. 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 $rank(X) = \min(n, p+1)$, then the linear equation system has trivial solution $\mathbf{b} = \mathbf{0}$ and so does $X^T(X\mathbf{b}) = \mathbf{0}$. Hence, X^TX is invertible. □

Using projection matrix Π_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.

$$\overline{Y}\mathbf{1} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y} = \Pi_{\mathbf{1}}\mathbf{Y}$$

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

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

Corollary 1.1 (Sum of squares)**.** $\Pi_{\mathbf{1}}$ *and* Π_X *can express each SS as following.*

(i)

$$SST = \mathbf{Y}^T(I - \Pi_{\mathbf{1}})\mathbf{Y}$$

(ii)

$$SSR = \mathbf{Y}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{Y}$$

(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)**.**

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

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 ϵ_x. This error term makes difference between properties of 1.9 and 1.10.

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

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

1.7.2 Regression coefficients

Under Normality, we have

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

Then

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

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

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

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

and

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

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

Let

$$A = \begin{bmatrix} \mathbf{a}^T \\ \hline \mathbf{b}^T \end{bmatrix} \in \mathbb{R}^{2 \times n}$$

Then

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

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

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\beta}_1 \end{bmatrix} \sim MVN\left(A\boldsymbol{\mu} = \begin{bmatrix} \beta_0 \\ \hline \beta_1 \end{bmatrix}, A\Sigma A^T = \sigma^2 AA^T = \left[\begin{array}{cc} \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right)\sigma^2 & -\frac{\overline{x}}{S_{XX}}\sigma^2 \\ \hline -\frac{\overline{x}}{S_{XX}}\sigma^2 & \frac{\sigma^2}{S_{XX}} \end{array}\right]\right) \tag{1.19}$$

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.*

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

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

1.7.3 Mean response

In simple linear regression setting, we assume $\mu_x = E(Y \mid x) = \beta_0 + \beta_1 x$ 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"
  )
```

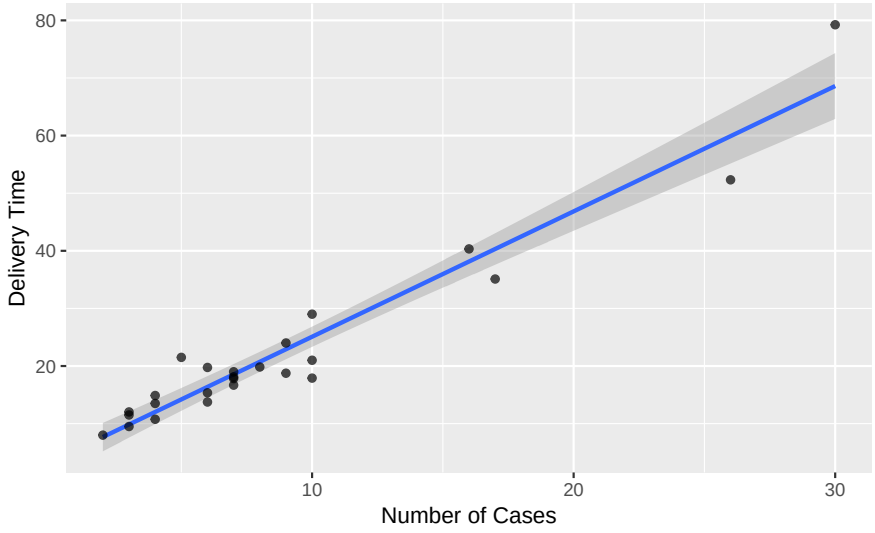


Figure 1.7: Mean response and its standard deviation

For example, in the Figure 1.7, the blue line indicates $E(Y \mid X = x)$ 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.*

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

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

$$\hat{\mu}_x \sim N\Big(E(\hat{\mu}_x), Var(\hat{\mu}_x)\Big)$$

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

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

□

Corollary 1.2.

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

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

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

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

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 $(\epsilon_1, \ldots, \epsilon_n)$. 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 $Y_1, \ldots, Y_n$, not Y_x.

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

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

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

$$\begin{aligned} Cov((\hat{\beta}_0, \hat{\beta}_1)^T, \epsilon_x) &= \Big[Cov(\hat{\beta}_i, \epsilon_x)\Big]_{2\times 1} \\ &= \left[Cov\left(\sum_{i=1}^{n} k_i Y_i, \epsilon_x\right)\right]_{2\times 1} \qquad k_i = \text{each linear coefficient for } \hat{\beta}_0, \hat{\beta}_1 \\ &= \mathbf{0} \end{aligned} \tag{1.20}$$

From Equation (1.19),

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

and from assumption,

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

It follows from Equation (1.20) 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))

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

□

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*

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

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 prosition 1.7,

$$\begin{aligned}\hat{Y_x} - Y_x &\sim MVN\left([1,x,-1]\begin{bmatrix}\beta_0\\ \beta_1\\ 0\end{bmatrix} - \beta_0 - \beta_1 x, [1,x,-1]\left[\begin{array}{c|c} Cov(\hat{\boldsymbol{\beta}}) \in \mathbb{R}^{2\times 2} & \mathbf{0} \in \mathbb{R}^2 \\ \hline \mathbf{0}^T \in \mathbb{R}^{2\times 1} & \sigma^2 \end{array}\right]\begin{bmatrix}1\\ x\\ -1\end{bmatrix}\right) \\ &\overset{d}{=} MVN\left(0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}} - 2\frac{\overline{x}x}{S_{XX}} + \frac{x^2}{S_{XX}}\right) + 1\right) \\ &\overset{d}{=} MVN\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{XX}}\right)\right)\end{aligned} \tag{1.21}$$

□

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

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

We can see this standard error is *always larger than of mean response estimator* due to 1 in the bracket, i.e. σ^2. As mentioned, this is due to ϵ 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 ϵ 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 μ_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)**.**

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

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

1.8 Statistical Inference

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

- β_0
- β_1
- μ_x
- Y_x
- σ^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-or-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.

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

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

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

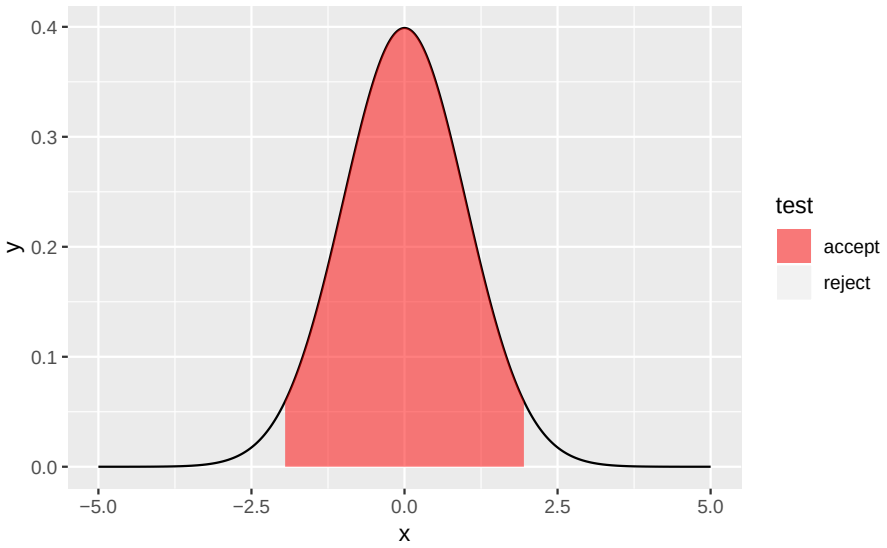


Figure 1.8: Confidence Interval when σ^2 is known

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

Proposition 1.8 (Confidence intervals on β)**.** *With known σ^2, $(1-\alpha)100\%$ confidence intervals on β_0 and β_1 are given as*

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

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

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

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

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

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

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

Proposition 1.10 (Confidence intervals on β when unknown σ^2)**.** *With unknown σ^2, $(1-\alpha)100\%$ confidence intervals on β_0 and β_1 are given as*

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

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

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

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

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* σ^2, $(1-\alpha)100\%$ *confidence interval on* $\hat{\mu}_x$ *is given as*

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

Also, with unknown σ^2,

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

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

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.

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

$$T = \frac{\hat{\beta}_0 - \alpha_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}} \overset{H_0}{\sim} t(n-2) \tag{1.22}$$

For this test statistic (1.22),

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

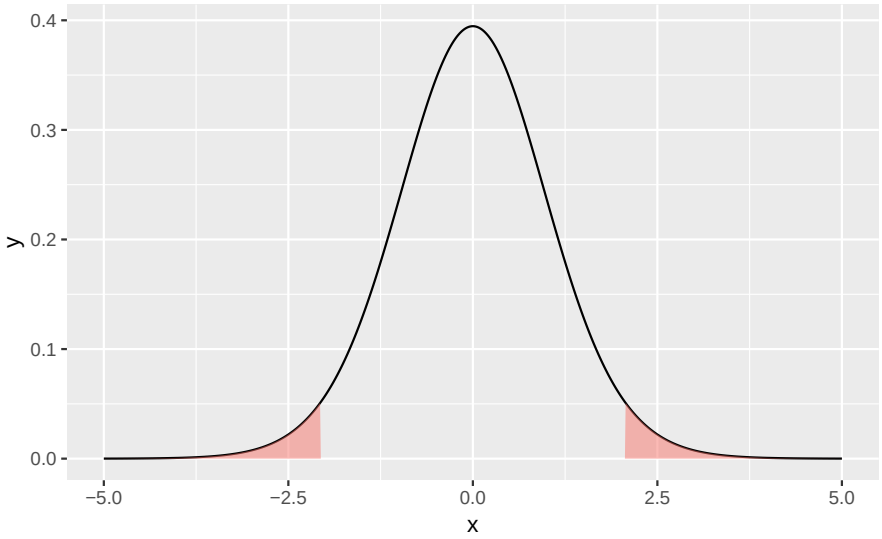


Figure 1.9: Rejection region for β_0

More importantly, we test β_1 which means slope

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

$$T = \frac{\hat{\beta}_1 - \alpha_1}{\hat{\sigma}\sqrt{\frac{1}{S_{xx}}}} \overset{H_0}{\sim} t(n-2) \tag{1.23}$$

For this test statistic (1.23),

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

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 $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. There are some useful distributions around Normal.

Proposition 1.14 (χ^2-distribution). *Square of standard normal follows χ^2-distribution.*

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

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

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

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

Proposition 1.16 (F-distribution). *Let $V \sim \chi^2(m) \perp\!\!\!\perp W \sim \chi^2(n)$. 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 χ^2-distribution). *Square of scaled normal follows non-central χ^2-distribution.*

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

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

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

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

μ is called a non-central parameter.

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

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

δ 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. See Corollary 1.1 for this.

- $SST = \mathbf{Y}^T(I - \Pi_{\mathbf{1}})\mathbf{Y}$
- $SSR = \mathbf{Y}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{Y}$
- $SSE = \mathbf{Y}^T(I - \Pi_X)\mathbf{Y}$

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 $tr(A) = rank(A)$

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 λ 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 λ^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 λ_j

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

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

$$tr(A) = rank(A)$$

□

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

(i) If A and B are symmetric,

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

(ii) If A is symmetric,

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

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

$$Y^T AY \sim \chi^2(K, \delta)$$

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

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

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

$$K = rank(I_n) = n$$

$$\delta = \mathbf{0}$$

□

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

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

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

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

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

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

$$\begin{aligned}\delta &= \left(\frac{X\boldsymbol{\beta}}{\sigma}\right)^T (I - \Pi_X)\left(\frac{X\boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}}{\sigma^2} - \frac{(\boldsymbol{\beta}^T X^T)X(X^T X)^{-1}X^T(X\boldsymbol{\beta})}{\sigma^2} \\ &= \frac{\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}}{\sigma^2} - \frac{\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}}{\sigma^2} \\ &= 0\end{aligned} \tag{1.24}$$

Hence,

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

□

In case of SSE, it always follows $\chi^2(n-2)$ no matter what H_0 is. However, SSR and SST depend on β_1 that we want to test.

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

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

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

Proof. From Corollary 1.1, write

$$\frac{SSR}{\sigma^2} = \left(\frac{\mathbf{Y}}{\sigma}\right)^T (\Pi_X - \Pi_{\mathbf{1}})\left(\frac{\mathbf{Y}}{\sigma}\right)$$

Note that $\Pi_X - \Pi_{\mathbf{1}}$ is symmetric idempotent. One proceeds in a similar way.

$$K = rank(\Pi_X - \Pi_{\mathbf{1}}) = tr(\Pi_X - \Pi_{\mathbf{1}}) = rank(\Pi_X) - rank(\Pi_{\mathbf{1}}) = 2 - 1 = 1$$

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

Since $\mathbf{1} \in Col(X)$,

$$\Pi_X\mathbf{1} = \mathbf{1}$$

It gives that

$$\mathbf{1}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{1} = 0 \tag{1.25}$$

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

$$\mathbf{x}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{x} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = S_{xx} \tag{1.26}$$

Recall that

$$\overline{x}\mathbf{1} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{x} = \Pi_{\mathbf{1}}\mathbf{x}$$

Then we have

$$\mathbf{1}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{x} = \sum x_i - n\overline{x} = 0 \tag{1.27}$$

By symmetry,

$$\mathbf{x}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{1} = n\overline{x} - \sum x_i = 0 \tag{1.28}$$

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

$$\begin{aligned}
\delta &= \frac{\boldsymbol{\beta}^T\Big\{[\mathbf{1} \mid \mathbf{x}]^T(\Pi_X - \Pi_{\mathbf{1}})[\mathbf{1} \mid \mathbf{x}]\Big\}\boldsymbol{\beta}}{\sigma^2} \\
&= \frac{\boldsymbol{\beta}^T\begin{bmatrix}(1.25) & (1.27)\\ (1.28) & (1.26)\end{bmatrix}\boldsymbol{\beta}}{\sigma^2} \\
&= \frac{\boldsymbol{\beta}^T\begin{bmatrix}0 & 0\\ 0 & S_{xx}\end{bmatrix}\boldsymbol{\beta}}{\sigma^2} \\
&= \frac{S_{xx}\beta_1^2}{\sigma^2}
\end{aligned} \tag{1.29}$$

□

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 $\Pi_X - \Pi_{\mathbf{1}}$ is symmetric. Then from Proposition 1.20,

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

It is obvious that

$$\Pi_X\Pi_{\mathbf{1}} = \Pi_{\mathbf{1}}$$

Then

$$\begin{aligned}(I-\Pi_X)(\Pi_X-\Pi_{\mathbf{1}}) &= \Pi_X-\Pi_{\mathbf{1}}-\Pi_X^2+\Pi_X\Pi_{\mathbf{1}} \\ &= \Pi_X-\Pi_{\mathbf{1}}-\Pi_X+\Pi_{\mathbf{1}} \qquad \because \text{idempotent} \\ &= 0\end{aligned}$$

This completes the proof. □

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

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

Proof. Note that

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

Since $I-\Pi_X$ of SSE is symmetric, from Proposition 1.20,
Claim: $((X^TX)^{-1}X^T)(\sigma^2 I)(I-\Pi_X)=0$, i.e. $((X^TX)^{-1}X^T)(I-\Pi_X)=0$
Since $\Pi_X = X(X^TX)^{-1}X^T$,

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

This completes the proof. □

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

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

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

Proof. It proceedes in a similary way from Corollary 1.1

$$\frac{SST}{\sigma^2} = \left(\frac{\mathbf{Y}}{\sigma}\right)^T(I-\Pi_{\mathbf{1}})\left(\frac{\mathbf{Y}}{\sigma}\right)$$

Since $I-\Pi_{\mathbf{1}}$ is symmetric idempotent,

$$K = rank(I-\Pi_{\mathbf{1}}) = tr(I-\Pi_{\mathbf{1}}) = n-rank(\Pi_{\mathbf{1}}) = n-1$$

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

$$\Pi_{\mathbf{1}}\mathbf{1} = \mathbf{1}$$

Hence,

$$\begin{aligned}\delta &= \left(\frac{X\boldsymbol{\beta}}{\sigma}\right)^T (I - \Pi_{\mathbf{1}})\left(\frac{X\boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{S_{xx}\beta_1^2}{\sigma^2} \qquad \because (1.24) \text{ and } (1.29)\end{aligned}$$

□

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

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

(b)

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

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

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* $H_0 : \beta_1 = 0$*,*

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

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 β_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

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

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

$$H_0 : \text{Model is not useful at all} \qquad \text{vs} \qquad H_1 : \text{Model can explain data} \tag{1.30}$$

Remark (F statistic on testing significance). Null hypothesis (1.30) can be tested with F-statistic.

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

Here, it is

$$F_0 = \frac{SSR/1}{SSE/(n-2)}$$

Then we reject H_0 if

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

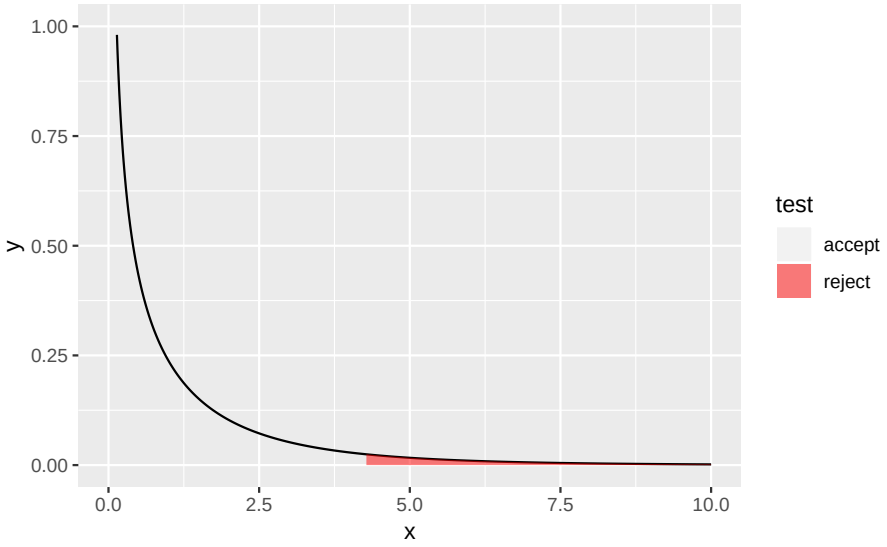


Figure 1.10: Rejection region for significance testing

```
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 (1.23) perform exactly same thing, $H_0 : \beta_1 = 0$. In fact, we know that

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

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

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

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

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

with

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

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

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

Each β_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,

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

where ϵ_i are i.i.d., and

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

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

2.2 Least Square Estimation

Write $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$. Extend Equation (1.16).

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

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

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

Our first question is when X^TX 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^TX is always non-negative definite.*

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

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

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

□

This lemma can also prove our Theorem 1.8.

Theorem 2.1. *Let $\mathbf{Y} = X\boldsymbol{\beta}$ 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^TX is invertible.

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

Suppose that X^TX is positive definite.

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

□

2.2.2 Orthogonal decomposition

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

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

where

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

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

Theorem 2.3. *Under the hypothesis of Theorem 2.2, $\hat{\mathbf{Y}} \in Col(X)$ 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 $Col(X)$

In other words, projection of $\mathbf{Y}$ onto $Col(X)$, $\hat{\mathbf{Y}}$ 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 coved 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

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

From (ii) of Theorem 2.4,

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

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

$$\begin{cases}\hat{\beta}_0 = \overline{Y} - \frac{1}{n}\mathbf{1}^T \mathbb{X}_A \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\beta}}_A = (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1}\mathbb{X}_{A,\perp}^T \mathbf{Y}\end{cases} \tag{2.5}$$

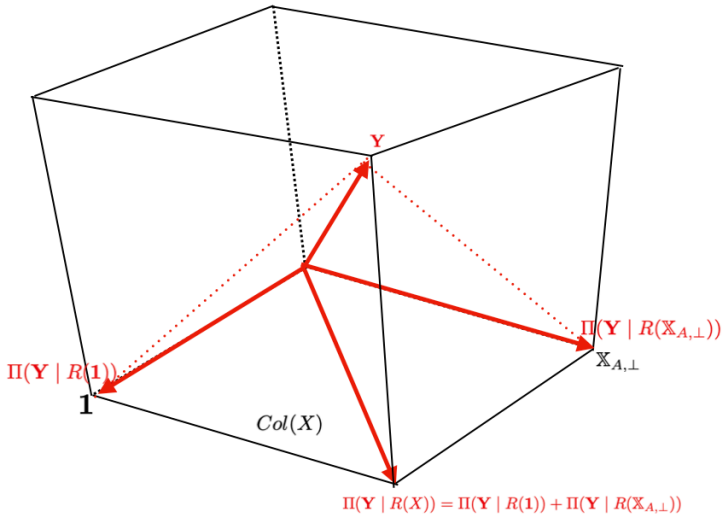


Figure 2.1: Orthogonal decomposition of the column space and LSE

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.

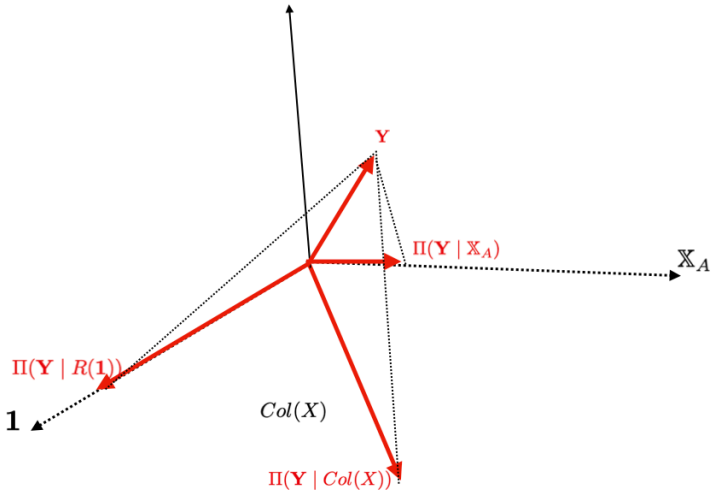


Figure 2.2: Non-orthongality

So what we have done is orthogonalization.

$$\tilde{\mathbb{X}}_A = \Pi_{\mathbf{1}}\mathbb{X}_A + (\mathbb{X}_A - \Pi_{\mathbf{1}}\mathbb{X}_A)$$

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* $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}\}$ *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 $\mathbf{u}_2, \ldots, \mathbf{u}_{p+1}$ *recursively by*

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

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

$$\mathbf{r}_k^* = < \mathbf{x}_{k+1}, \mathbf{u}_1 > \mathbf{u}_1 + < \mathbf{x}_{k+1}, \mathbf{u}_2 > \mathbf{u}_2 + \cdots + < \mathbf{x}_{k+1}, \mathbf{u}_k > \mathbf{u}_k$$

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

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

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

input: basis $\{\mathbf{x}_0, \ldots, \mathbf{x}_p\}$
Initialize $\mathbf{v}_0 = \mathbf{x}_0$;
for $k \leftarrow 1$ **to** p **do**
  $\mathbf{u}_{k-1} = \frac{\mathbf{v}_{k-1}}{\|\mathbf{v}_{k-1}\|}$;
  $\mathbf{r}_k^* = < \mathbf{x}_{k+1}, \mathbf{u}_0 > \mathbf{u}_0 + < \mathbf{x}_{k+1}, \mathbf{u}_1 > \mathbf{u}_1 + \cdots + < \mathbf{x}_{k+1}, \mathbf{u}_k > \mathbf{u}_k$;
  $\mathbf{v}_{k+1} = \mathbf{x}_{k+1} - \mathbf{r}_k^*$
end
$\mathbf{u}_p = \frac{\mathbf{v}_p}{\|\mathbf{v}_p\|}$

Algorithm 1: Gram-schmidt process

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

$$\mathbf{u}_1 = \frac{\mathbf{x}_1}{\|\mathbf{x}_1\|} \Rightarrow \mathbf{x}_1 = \|\mathbf{x}_1\|\mathbf{u}_1$$

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

It procees in a simlar way to the others. Hence,

$$\begin{aligned} X &= \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_{p+1} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{p+1} \end{bmatrix} \begin{bmatrix} \|\mathbf{v}_1\| & < \mathbf{x}_2, \mathbf{u}_1 > & < \mathbf{x}_3, \mathbf{u}_1 > & \cdots & < \mathbf{x}_{p+1}, \mathbf{u}_1 > \\ 0 & \|\mathbf{v}_2\| & < \mathbf{x}_3, \mathbf{u}_2 > & \cdots & < \mathbf{x}_{p+1}, \mathbf{u}_2 > \\ 0 & 0 & \|\mathbf{v}_3\| & \cdots & < \mathbf{x}_{p+1}, \mathbf{u}_3 > \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \|\mathbf{v}_{p+1}\| \end{bmatrix} \\ &\equiv QR \end{aligned} \tag{2.6}$$

□

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 $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p+1}\}$ be a basis for the inner product space V and let $\{\mathbf{q}_1, \ldots, \mathbf{q}_{p+1}\}$ 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 $\{\mathbf{x}_2, \ldots, \mathbf{x}_{p+1}\}$ 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* $\mathbf{q}_2 = \frac{\mathbf{x}^{(1)}}{\|\mathbf{x}_2^{(2)}\|}$. *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 similary 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

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

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.

```
for k ← 1 to (p + 1) do
    r_kk = ‖x_k‖;
    q_k = x_k / r_kk;
    for j ← 1 to (p + 1) do
        r_kj = q_k^T x_j;
        x_j = x_j − r_kj q_k;
    end
end
```

Algorithm 2: 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 1 is equivalent to succesive orthogonalization, i.e. regress(project) $\mathbf{x}_j$ onto the others (Hastie et al., 2013).

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

Initialize $\mathbf{v}_0 = \mathbf{1}$;
for $k \leftarrow 1$ **to** p **do**
  Regress $\mathbf{x}_k$ on $\mathbf{q}_0, \ldots, \mathbf{q}_{k-1}$;
  $\hat{\beta}_{lk} = \frac{<\mathbf{v}_l, \mathbf{x}_k>}{<\mathbf{v}_l, \mathbf{v}_l>}, l = 0, \ldots, k-1$;
  Residual $\mathbf{v}_k = \mathbf{x}_k - \sum_{l=0}^{k-1} \hat{\beta}_{lk} \mathbf{v}_k$;
end
Regress $\mathbf{Y}$ on $\mathbf{v}_p$

Algorithm 3: Successive orthogonalization

$$
\begin{aligned}
&(X^T X)\hat{\boldsymbol{\beta}} = X^T \mathbf{Y} \\
&\Leftrightarrow R^T Q^T Q R \hat{\boldsymbol{\beta}} = R^T Q^T \mathbf{Y} \\
&\Leftrightarrow R^T R \hat{\boldsymbol{\beta}} = R^T Q^T \mathbf{Y} \qquad \because Q^T Q = I \\
&\Leftrightarrow R \hat{\boldsymbol{\beta}} = Q^T \mathbf{Y} \qquad \text{if } R \text{ is invertible}
\end{aligned}
\tag{2.7}
$$

Hence,

$$
\hat{\boldsymbol{\beta}} = R^{-1} Q^T \mathbf{Y} \tag{2.8}
$$

It follows that

$$
\hat{\mathbf{Y}} = (QR)\hat{\boldsymbol{\beta}} = QQ^T \mathbf{Y} \tag{2.9}
$$

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 (2.8), 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.

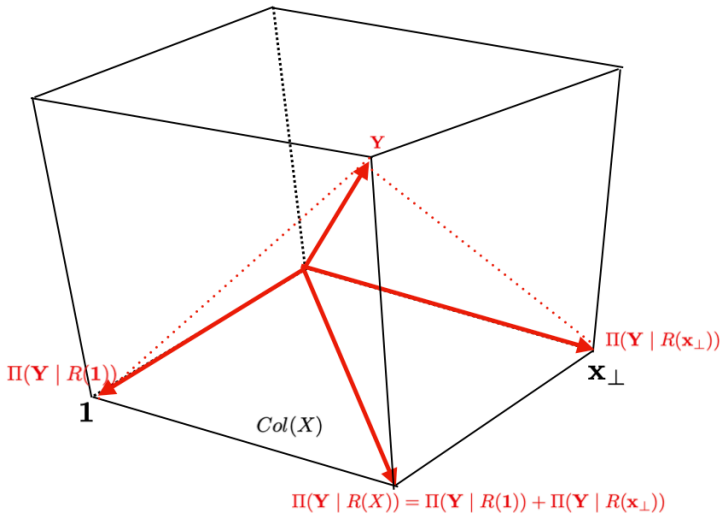


Figure 2.3: Orthogonalized basis

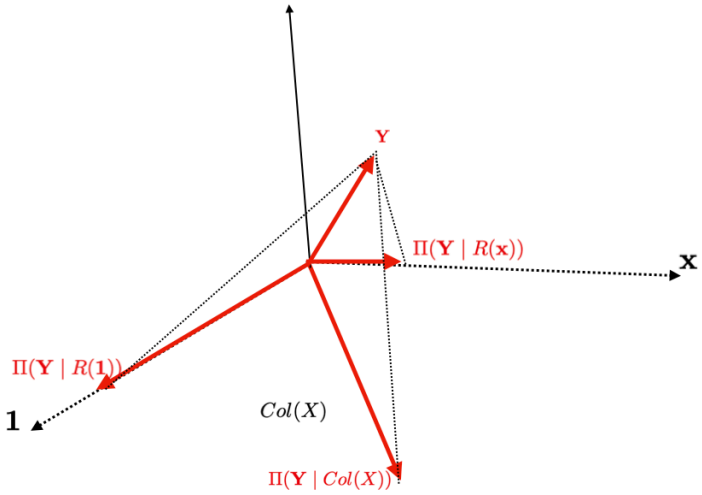


Figure 2.4: Non-orthogonal basis

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.

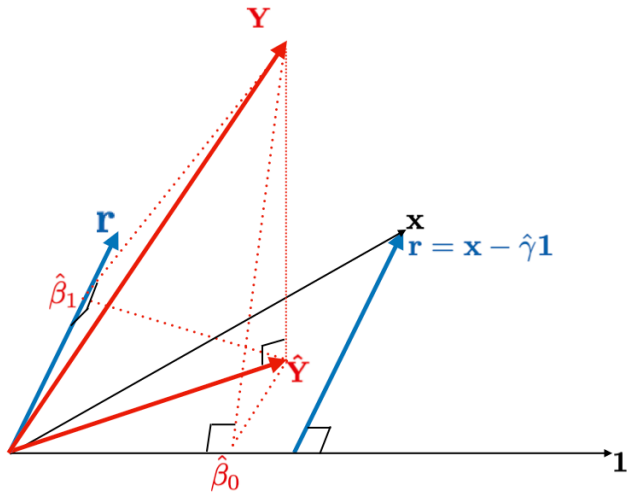


Figure 2.5: QR decomposition for model matrix

Regress **x** onto **1**. Its residual can be a new orthogonalized predictor.

2.2.5 Properties of LSE

We have seen how we extend point estimator $\hat{\boldsymbol{\beta}}$. In turn, we can check this is unbiased, and BLUE.

Proposition 2.1 (Expectation and Variance)**.** *$\hat{\boldsymbol{\beta}}$ is unbiased.*

1. $E\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$
2. $Var\hat{\boldsymbol{\beta}} = \sigma^2(X^TX)^{-1}$

Proof.

$$\begin{aligned} E\hat{\boldsymbol{\beta}} &= E\Big((X^TX)^{-1}X^T\mathbf{Y}\Big) \\ &= (X^TX)^{-1}X^TE\mathbf{Y} \\ &= (X^TX)^{-1}X^TX\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Hence, $\hat{\boldsymbol{\beta}}$ is unbiased.

$$\begin{aligned} Var\hat{\boldsymbol{\beta}} &= Var\Big((X^TX)^{-1}X^T\mathbf{Y}\Big) \\ &= (X^TX)^{-1}X^TVar(\mathbf{Y})X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T(\sigma^2I)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

□

Since the variance of LSE have been revealed, now we want to know if this is the lowest one among estimators. Gauss-Markov theorem states that LSE has the lowest variance among linear unbiased estimators for $\boldsymbol{\beta}$, so called the **best linear unbiased estimator (BLUE)**.

Theorem 2.8 (Gauss-Markov Theorem)**.** *$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$ is BLUE, i.e.*

For any $\tilde{\boldsymbol{\beta}} \in \Omega \equiv \{\tilde{\boldsymbol{\beta}} : \tilde{\boldsymbol{\beta}} = C\mathbf{Y}, E\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\}$,

$$Var(\hat{\boldsymbol{\beta}}) \leq Var(\tilde{\boldsymbol{\beta}})$$

Proof. Let $\tilde{\boldsymbol{\beta}} \in \Omega \equiv \{\tilde{\boldsymbol{\beta}} : \tilde{\boldsymbol{\beta}} = C\mathbf{Y}, E\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\}$

Claim: $Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}})$ is non-negative definite.

Note that $\hat{\boldsymbol{\beta}}$ is the one with $C = (X^TX)^{-1}X^T$.

Set $D := C - (X^TX)^{-1}X^T$. From unbiasedness,

$$\begin{aligned} E\tilde{\boldsymbol{\beta}} &= CE\mathbf{Y} \\ &= CX\boldsymbol{\beta} \\ &= \Big((X^TX)^{-1}X^T + D\Big)X\boldsymbol{\beta} \\ &= \boldsymbol{\beta} + DX\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

Since $\forall \boldsymbol{\beta} \in \mathbb{R}^{p+1} : DX\boldsymbol{\beta} = \mathbf{0}$,

$$DX = 0 \tag{2.10}$$

$$\begin{aligned}
Var\tilde{\boldsymbol{\beta}} &= Var(C\mathbf{Y}) \\
&= \sigma^2 CC^T \\
&= \sigma^2\Big((X^TX)^{-1}X^T + D\Big)\Big((X^TX)^{-1}X^T + D\Big)^T \\
&= \sigma^2\Big((X^TX)^{-1} + DX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T + DD^T\Big) \\
&= \sigma^2\Big((X^TX)^{-1} + DD^T\Big) \qquad \because (2.10) \\
&= Var\hat{\boldsymbol{\beta}} + \sigma^2 DD^T
\end{aligned}$$

Note that DD^T is non-negative definite. Hence,

$$Var\tilde{\boldsymbol{\beta}} - Var\hat{\boldsymbol{\beta}} = \sigma^2 DD^T$$

is non-negative definite. This completes the proof. □

As in simple linear regression setting, we define *residuals* and explain σ^2.

Definition 2.1 (Residuals)**.** Let $\hat{\mathbf{Y}} = (\hat{Y_1}, \ldots, \hat{Y_n})^T$ with $\hat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$. Then the residual is defined by

$$\mathbf{e} := (\ldots, Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}, \ldots)^T = \mathbf{Y} - \hat{\mathbf{Y}} \in \mathbb{R}^n$$

Extending the simple setting, we estimate σ^2 with inner product of residuals divided by its degrees of freedom, i.e. *MSE*. The degrees of freedom becomes $n -$ the number of coefficients. Thus, $n - (p+1) = n - p - 1$.

Proposition 2.2 (Estimation of σ^2)**.** *Let* $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ *be residuals as in Definition 2.1. Then*

$$\hat{\sigma}^2 = \frac{\|\mathbf{e}\|}{n - p - 1}$$

The reason we divide with degrees of freedom is to make $\hat{\sigma}^2$ unbiased.

Proposition 2.3 (Mean of $\hat{\sigma}^2$)**.** $\hat{\sigma}^2$ *is unbiased, i.e.*

$$E\Big(\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2\Big) = (n - p - 1)\sigma^2$$

Proof. Since $\mathbf{Y} = \Pi_X \mathbf{Y}$,

$$\begin{aligned}
\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 &= \|(I - \Pi_X)\mathbf{Y}\|^2 \\
&= \mathbf{Y}^T(I - \Pi_X)^T(I - \Pi_X)\mathbf{Y} \\
&= \mathbf{Y}^T(I - \Pi_X)\mathbf{Y} \qquad \because (I - \Pi_X) : \text{symmetric idempotent}
\end{aligned}$$

Since $\mathbf{Y}^T(I - \Pi_X)\mathbf{Y} \in \mathbb{R}$,

$$\begin{aligned}\mathbf{Y}^T(I-\Pi_X)\mathbf{Y} &= tr\Big(\mathbf{Y}^T(I-\Pi_X)\mathbf{Y}\Big)\\ &= tr\Big((I-\Pi_X)\mathbf{Y}\mathbf{Y}^T\Big)\end{aligned}$$

Then

$$\begin{aligned}E\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2 &= E\Big[\mathbf{Y}^T(I-\Pi_X)\mathbf{Y}\Big]\\ &= E\Big[tr\Big((I-\Pi_X)\mathbf{Y}\mathbf{Y}^T\Big)\Big]\\ &= tr\Big((I-\Pi_X)\underbrace{E\Big[\mathbf{Y}\mathbf{Y}^T\Big]}_{(*)}\Big)\end{aligned}$$

Consider $(*)$.

$$\begin{aligned}E(\mathbf{Y}\mathbf{Y}^T) &= Var\mathbf{Y} + (E\mathbf{Y})(E\mathbf{Y})^T\\ &= \sigma^2 I + X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T\end{aligned}$$

Hence,

$$\begin{aligned}E\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2 &= tr\Big((I-\Pi_X)E\Big[\mathbf{Y}\mathbf{Y}^T\Big]\Big)\\ &= tr\Big((I-\Pi_X)\sigma^2 + (I-\Pi_X)X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T\Big)\\ &= tr\Big((I-\Pi_X)\sigma^2\Big) + tr\Big(\boldsymbol{\beta}^T X^T(I-\Pi_X)X\boldsymbol{\beta}\Big) \qquad \because (I-\Pi_X)X\boldsymbol{\beta} = X\boldsymbol{\beta} - X\boldsymbol{\beta} = 0\\ &= tr\Big((I-\Pi_X)\sigma^2\Big)\\ &= (n-p-1)\sigma^2 \qquad \because \begin{cases} tr(I) = n \\ tr(\Pi_X) = p+1 \end{cases}\end{aligned}$$

□

In this proposition, we have used model matrix directly. Instead, we can use Equation (2.8) for simplicity.

Proposition 2.4 (Variance using QR decomposition)**.** *Let $X = QR$. Then $\hat{\boldsymbol{\beta}} = R^{-1}Q^T\mathbf{Y}$. It follows that*

$$Var\hat{\boldsymbol{\beta}} = \sigma^2(R^TR)^{-1}$$

Proof. It proceeds in a similar way for $\hat{\boldsymbol{\beta}} = R^{-1}Q^T\mathbf{Y}$.

$$\begin{aligned} Var\hat{\boldsymbol{\beta}} &= Var\left(R^{-1}Q^T\mathbf{Y}\right) \\ &= R^{-1}Q^T Var(\mathbf{Y})Q(R^T)^{-1} \\ &= R^{-1}Q^T(\sigma^2 I)Q(R^T)^{-1} \\ &= \sigma^2(R^T R)^{-1} \qquad \because Q^T Q = I \end{aligned}$$

□

Proposition 2.5 (QR representation for residual)**.** *Let $X = QR$. Then*

$$\mathbf{e} = (I - QQ^T)\mathbf{Y}$$

Proof. From Equation (2.9),

$$\Pi_X = QQ^T$$

Hence,

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

□

On Figure 2.5, we can see these relations. Operation $Q^T\mathbf{Y}$ is just projection to each orthogonal basis. Q sums these projection so that we get $\hat{\mathbf{Y}}$ which projection of $\mathbf{Y}$ onto the column space of model matrix.

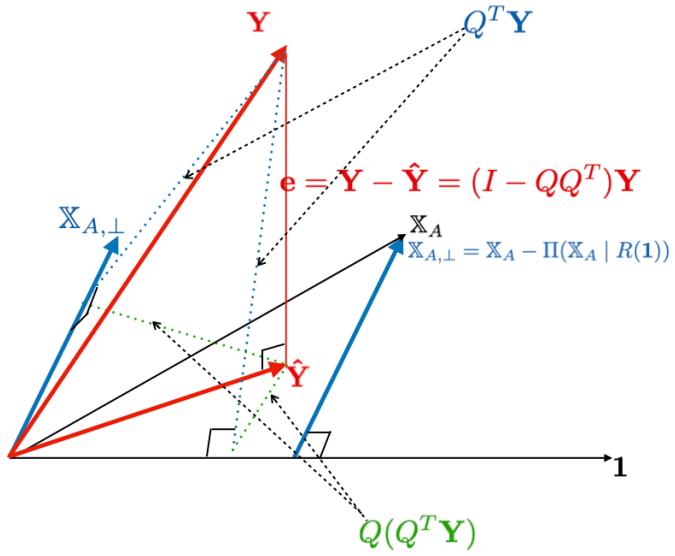


Figure 2.6: Residual vector

2.2.6 Mean response and response

Let $\mathbf{z} = (z_1, \ldots, z_p)^T$.

Theorem 2.9 (Estimation of the mean response)**.**

$$\hat{\mu}_z = \hat{\beta}_0 + \mathbf{z}^T \hat{\boldsymbol{\beta}}_A$$

Theorem 2.10 ((out of sample) Prediction of a response)**.**

$$\hat{Y}_z = \hat{\beta}_0 + \mathbf{z}^T \hat{\boldsymbol{\beta}}_Z$$

Proposition 2.6 (Residual vector)**.** *Let* $\mathbf{e} = (I - \Pi_X)\mathbf{Y}$*. Then*

1. $Var(\mathbf{e}) = \sigma^2(I - \Pi_X)$
2. $\mathbf{e} \perp \hat{\mathbf{Y}}$

$Var(\mathbf{e})$.

$$\begin{aligned} Var(\mathbf{e}) &= Var\Big((I - \Pi_X)\mathbf{Y}\Big) \\ &= (I - \Pi_X) Var(\mathbf{Y}) (I - \Pi_X)^T \\ &= \sigma^2 (I - \Pi_X) \qquad \because (I - \Pi_X) \text{ symmetric itempotent} \end{aligned}$$

□

$\mathbf{e} \perp \hat{\mathbf{Y}}$. Note that

$$\mathbf{e} \in Col(X)^{\perp}$$

From the properties of projection, we have

$$\begin{cases} \mathbf{e} \perp \mathbf{1} \\ \mathbf{e} \perp \mathbf{x}_j \quad \forall j = 1, 2, \ldots, p \end{cases} \tag{2.11}$$

Since $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$,

$$\mathbf{e} \perp \hat{\mathbf{Y}}$$

□

Equation (2.11) is another form of the *normal equation*.

Remark. The least squares regression line $\{(\mathbf{z}, y) : y = \hat{\beta}_0 + \mathbf{z}^T \hat{\beta}_A\}$ always passes through

$$\left(\frac{1}{n}\mathbb{X}_A^T \mathbf{1}, \overline{Y}\right)$$

In simple linear regression setting,

$$(\overline{x}, \overline{Y})$$

Proof. First consider $p = 1$. Normal equation gives directly that

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

Thus,

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

We now give more general proof, i.e. for $p \geq 1$.

Claim: $\overline{Y} = \hat{\beta}_0 + \left(\frac{1}{n}\mathbb{X}_A^T \mathbf{1}\right)^T \hat{\boldsymbol{\beta}}_A$

From Equation (2.5),

$$\hat{\beta}_0 = \overline{Y} - \frac{1}{n}\mathbf{1}^T \mathbb{X}_A \hat{\boldsymbol{\beta}}_A$$

It follows that

$$\overline{Y} = \hat{\beta}_0 + \frac{1}{n}\mathbf{1}^T \mathbb{X}_A \hat{\boldsymbol{\beta}}_A$$

This completes the proof. □

2.3 Analysis of Variance

2.3.1 Decomposition of SST

We have

- $SST = \|\mathbf{Y} - \overline{Y}\mathbf{1}\|^2 = \mathbf{Y}^T(I - \Pi_{\mathbf{1}})\mathbf{Y}$
- $SSR = \|\hat{\mathbf{Y}} - \overline{Y}\mathbf{1}\|^2 = \mathbf{Y}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{Y}$
- $SSE = \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \mathbf{Y}^T(I - \Pi_X)\mathbf{Y}$

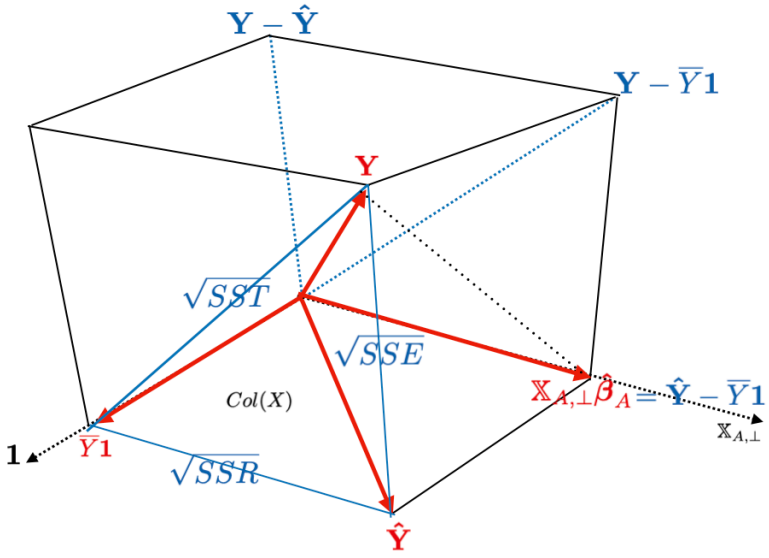


Figure 2.7: Decomposition of SST

In Figure 2.7, Pythagorean law gives

$$SST = SSR + SSE$$

Lemma 2.2. *Let* $X = [\mathbf{1} \mid \mathbb{X}_A]$ *and let* $\mathbb{X}_{A,\perp} = \mathbb{X}_A - \Pi_{\mathbf{1}}\mathbb{X}_A$. *Then*

$$\hat{\mathbf{Y}} - \overline{Y}\mathbf{1} = \mathbb{X}_{A,\perp}\hat{\boldsymbol{\beta}}_A$$

Proof. Note that $\mathbf{1} \perp \mathbb{X}_{A,\perp}$.

Recall that

$$\Pi(\mathbf{Y} \mid R(\mathbf{1})) = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y} = \overline{Y}\mathbf{1}$$

From Theorem 2.4,

$$\begin{aligned}\hat{\mathbf{Y}} &= \Pi(\mathbf{Y} \mid R(\mathbf{1})) + \Pi(\mathbf{Y} \mid R(\mathbb{X}_{A,\perp})) \\ &= \overline{Y}\mathbf{1} + \mathbb{X}_{A,\perp}\hat{\boldsymbol{\beta}}_A\end{aligned}$$

Hence,

$$\hat{\mathbf{Y}} - \overline{Y}\mathbf{1} = \mathbb{X}_{A,\perp}\hat{\boldsymbol{\beta}}_A$$

□

2.3.2 Distributions

Proposition 2.7 (Distribution of SS)**.** *Extending for $p > 1$, we can get each result.*

1. $\frac{SSE}{\sigma^2} \sim \chi^2(n-p-1)$
2. $\frac{SSR}{\sigma^2} \sim \chi^2(p, \delta), \quad \delta = \frac{\boldsymbol{\beta}^T X^T (\Pi_X - \Pi_{\mathbf{1}}) X\boldsymbol{\beta}}{\sigma^2}$
3. $SSR \perp\!\!\!\perp SSE$
4. $SSE \perp\!\!\!\perp \hat{\boldsymbol{\beta}}$
5. $\frac{SST}{\sigma^2} \sim \chi^2(n-1, \delta), \quad \delta = \frac{\boldsymbol{\beta}^T X^T (I - \Pi_{\mathbf{1}}) X\boldsymbol{\beta}}{\sigma^2}$

Distribution of SSE. Note that

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

From Theorem 1.15,

$$K = rank(I - \Pi_X) = tr(I - \Pi_X) = n - rank(\Pi_X) = n - p - 1$$

δ proof is exactly same as Proposition 1.21.

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

Hence, $\delta = 0$. □

Distribution of SSR. Note that

$$SSR = \mathbf{Y}^T (\Pi_X - \Pi_{\mathbf{1}})\mathbf{Y}$$

From Theorem 1.15,

$$K = rank(\Pi_X - \Pi_{\mathbf{1}}) = tr(\Pi_X) - tr(\Pi_{\mathbf{1}}) = rank(\Pi_X) - rank(\Pi_{\mathbf{1}}) = p + 1 - 1 = p$$

δ proof is exactly same as Proposition 1.22.

$$\begin{aligned}
\delta &= \left(\frac{X\boldsymbol{\beta}}{\sigma}\right)^T (\Pi_X - \Pi_{\mathbf{1}})\left(\frac{X\boldsymbol{\beta}}{\sigma}\right) \qquad \because \frac{\mathbf{Y}}{\sigma} \sim MVN(\frac{1}{\sigma} X\boldsymbol{\beta}, I) \\
&= \frac{\boldsymbol{\beta}^T \left\{ X^T (\Pi_X - \Pi_{\mathbf{1}}) X \right\} \boldsymbol{\beta}}{\sigma^2}
\end{aligned} \tag{2.12}$$

This completes the proof. □

In univariate setting, we have seen that δ is expressed in terms of $\hat{\beta}_1$ excluding $\hat{\beta}_0$. How about in multivariate? In Equation (2.12), we can block design matrix X into

$$X = \left[\mathbf{1} \mid \mathbb{X}_A\right]$$

One proceeds in a similar way. Since both $\mathbf{1}, \mathbb{X}_A \in Col(X)$,

$$\Pi_X \mathbf{1} = \mathbf{1}, \qquad \Pi_X \mathbb{X}_A = \mathbb{X}_A$$

Since $\mathbf{1} \in R(\{\mathbf{1}\})$,

$$\Pi_{\mathbf{1}} \mathbf{1} = \mathbf{1}$$

Since $\Pi_{\mathbf{1}} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$,

$$\Pi_{\mathbf{1}} \mathbb{X}_A = \mathbf{1}\overline{\mathbf{x}}^T = \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \\ \vdots & \vdots & \vdots & \vdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{bmatrix}$$

Using these facts, we have following

$$\mathbf{1}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{1} = \mathbf{1}^T(\mathbf{1} - \mathbf{1}) = 0 \tag{2.13}$$

$$\begin{aligned} \mathbf{1}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbb{X}_A &= \mathbf{1}^T(\mathbb{X}_A - \mathbf{1}\overline{\mathbf{x}}^T) \\ &= \mathbf{1}^T \left[x_{ij} - \overline{x}_j\right]_{1\times j} \\ &= \left[\textstyle\sum_i x_{ij} - n\overline{x}_j\right]_{1\times j} \\ &0 \end{aligned} \tag{2.14}$$

$$\mathbb{X}_A^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{1} = \mathbb{X}_A^T(\mathbf{1} - \mathbf{1}) = 0 \tag{2.15}$$

From Lemma 1.1,

$$\begin{aligned} \mathbb{X}_A^T(\Pi_X - \Pi_{\mathbf{1}})\mathbb{X}_A &= \mathbb{X}_A^T \left[x_{ij} - \overline{x}_j\right]_{1\times j} \\ &= \left[\textstyle\sum_i x_{ij}(x_{jk} - \overline{x}_k)\right]_{j\times k} \\ &= \left[\textstyle\sum_i (x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)\right]_{j\times k} \\ &= (n-1)Var(\mathbb{X}_A) \\ &\equiv S \end{aligned} \tag{2.16}$$

Hence,

$$\begin{aligned}\delta &= \frac{\boldsymbol{\beta}^T\Big\{\left[\mathbf{1} \mid \mathbb{X}_A\right]^T(\Pi_X - \Pi_{\mathbf{1}})\left[\mathbf{1} \mid \mathbb{X}_A\right]\Big\}\boldsymbol{\beta}}{\sigma^2} \\ &= \frac{\boldsymbol{\beta}^T\left[\begin{array}{c|c} (2.13) & (2.14) \\ \hline (2.15) & (2.16)\end{array}\right]\boldsymbol{\beta}}{\sigma^2} \\ &= \frac{\boldsymbol{\beta}^T\left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & S\end{array}\right]\boldsymbol{\beta}}{\sigma^2} \\ &= \frac{\boldsymbol{\beta}_A^T S\boldsymbol{\beta}_A}{\sigma^2}\end{aligned} \tag{2.17}$$

Independence between SSE and SSR. Since SSE and SSR are quadratic form of $\mathbf{Y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I)$ and each $I - \Pi_X$ and $\Pi_X - \Pi_{\mathbf{1}}$ is symmetric,

Claim: $(I - \Pi_X)(\Pi_X - \Pi_{\mathbf{1}}) = 0$

We have already shown this in Proposition 1.23. Using the fact that $\Pi_X\Pi_{\mathbf{1}} = \Pi_{\mathbf{1}}$,

$$\begin{aligned}(I - \Pi_X)(\Pi_X - \Pi_{\mathbf{1}}) &= \Pi_X - \Pi_{\mathbf{1}} - \Pi_X^2 + \Pi_X\Pi_{\mathbf{1}} \\ &= \Pi_X - \Pi_{\mathbf{1}} - \Pi_X + \Pi_{\mathbf{1}} \qquad \because \text{idempotent} \\ &= 0\end{aligned}$$

□

Independence between SSE and regression vector. The proof is same as 1.24, by showing that $((X^TX)^{-1}X^T)(I - \Pi_X) = 0$.

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

$$((X^TX)^{-1}X^T)(I - \Pi_X) = (X^TX)^{-1}X^T - (X^TX)^{-1}X^T = 0$$

This completes the proof. □

Distribution of SST. Note that

$$SST = \mathbf{Y}^T(I - \Pi_{\mathbf{1}})\mathbf{Y}$$

From Theorem 1.15,

$$K = rank(I - \Pi_{\mathbf{1}}) = tr(I - \Pi_{\mathbf{1}}) = n - 1$$

and

$$\begin{aligned}\delta &= \left(\frac{X\boldsymbol{\beta}}{\sigma}\right)^T(I - \Pi_{\mathbf{1}})\left(\frac{X\boldsymbol{\beta}}{\sigma}\right) \\ &= \frac{\boldsymbol{\beta}^T X^T(I - \Pi_{\mathbf{1}})X\boldsymbol{\beta}}{\sigma^2}\end{aligned}$$

□

2.3.3 ANOVA for testing significance of regression

Under the normality of error term

$$\boldsymbol{\epsilon}_i \sim MVN(\mathbf{0}, \sigma^2 I)$$

a test statistic can follow F-distribution as in univariate setting.

Corollary 2.1 (F-test)**.** *Under normality,*

$$F = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1, \delta)$$

where
$\delta = \frac{\boldsymbol{\beta}^T X^T (\Pi_X - \Pi_1) X \boldsymbol{\beta}}{\sigma^2}$

In the proof of 1.22, we can understand the structure that $\delta = 0$ when all coefficients corresponding to predictors are zero.

$$F \overset{H_0}{\sim} F(p, n-p-1)$$

where

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

As other ordinary test, we just reject this H_0 if observed F_0 is large, i.e.

$$F_0 > F_\alpha(p, n-p-1)$$

See Figure 1.10. **ANOVA table** summarizes these statistics in table form.

Source	SS	df	MS	F	p-value
Model	SSR	p	$MSR = \frac{SSR}{p}$	$F_0 = \frac{MSR}{MSE}$	p-value
Error	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$		
Total	SST	$n-1$			

Everything is same in R.

```
cem_fit <- lm(y ~ ., data = cem)
summary(cem_fit)
#>
#> Call:
#> lm(formula = y ~ ., data = cem)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -3.175 -1.671  0.251  1.378  3.925
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   62.405     70.071    0.89    0.399
```

```
#> x1             1.551      0.745    2.08    0.071 .
#> x2             0.510      0.724    0.70    0.501
#> x3             0.102      0.755    0.14    0.896
#> x4            -0.144      0.709   -0.20    0.844
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 2.45 on 8 degrees of freedom
#> Multiple R-squared:  0.982,  Adjusted R-squared:  0.974
#> F-statistic:  111 on 4 and 8 DF,  p-value: 4.76e-07
```

You can see `F-statistic` with degrees of freedom 4 8.

```
summary(cem_fit)$fstatistic
#> value numdf dendf
#>   111     4     8
```

However, `anova.lm()` gives a bit different format This is related to *extra sum of squares*, which will be covered later.

```
anova(cem_fit)
#> Analysis of Variance Table
#>
#> Response: y
#>           Df Sum Sq Mean Sq F value  Pr(>F)
#> x1         1   1450    1450  242.37 2.9e-07 ***
#> x2         1   1208    1208  201.87 5.9e-07 ***
#> x3         1     10      10    1.64    0.24
#> x4         1      0       0    0.04    0.84
#> Residuals  8     48       6
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3.4 Coefficient of determination

In univariate setting, coefficient of determination measures linear relationship based on the SST decomposition.

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

Moreover, it becomes to be same as sample correlation $\hat{\rho}$ and angle between two vectors. In multivariate setting, we also define this kind of measure.

Definition 2.2 (Coefficient of Determination)**.** For $\mathbf{Y}_i = X\boldsymbol{\beta} + \epsilon_i$,

$$R^2 := \max_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \hat{\rho}(\mathbf{Y}, X\boldsymbol{\beta})$$

where $\hat{\rho}$ means sample correlation.

We have mentioned about $R^2 = (\cos\theta)^2$ in simple linear regression. See Equation (1.17). Now we try to see detail of this relation. First, in Leon (2014), you might see the relation of $\cos\theta$ and inner product.

Lemma 2.3. *If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are two nonzero vectors and θ is the angle between them, then*

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta$$

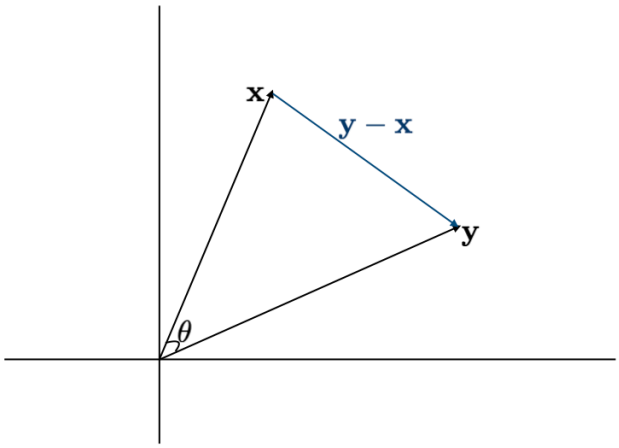


Figure 2.8: Two vectors in $\mathbb{R}^2$

Proof. See Figure 2.8. We have a triangle. Law of cosines gives that

$$\|\mathbf{y}-\mathbf{x}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta$$

It follows that

$$\begin{aligned}\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta &= \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{y}-\mathbf{x}\|^2)\\ &= \frac{1}{2}(\mathbf{x}^T\mathbf{x} + \mathbf{y}^T\mathbf{y} - (\mathbf{y}-\mathbf{x})^T(\mathbf{y}-\mathbf{x}))\\ &= \mathbf{x}^T\mathbf{y}\end{aligned}$$

□

This implies the *relationship between sample correlation and the angle.*

Theorem 2.11 (Sample correlation and the Angle)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_n)^T$ *and* $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ *be random variables. Then*

$$\hat{\rho}(\mathbf{X}, \mathbf{Y}) = cos\theta$$

where θ *is the angle between* $\mathbf{X} - \overline{X}\mathbf{1}$ *and* $\mathbf{Y} - \overline{Y}\mathbf{1}$.

Proof. Note that

$$\widehat{Cov}(\mathbf{X},\mathbf{Y}) = \frac{1}{n-1}(\mathbf{X}-\overline{X}\mathbf{1})^T(\mathbf{Y}-\overline{Y}\mathbf{1})$$

$$\hat{\sigma}_{\mathbf{X}} = \sqrt{\frac{1}{n-1}(\mathbf{X}-\overline{X}\mathbf{1})^T(\mathbf{X}-\overline{X}\mathbf{1})}$$

and

$$\hat{\sigma}_{\mathbf{Y}} = \sqrt{\frac{1}{n-1}(\mathbf{Y}-\overline{Y}\mathbf{1})^T(\mathbf{Y}-\overline{Y}\mathbf{1})}$$

and hence it follows that

$$\begin{aligned}
\hat{\rho}(\mathbf{X},\mathbf{Y}) &= \frac{\widehat{Cov}(\mathbf{X},\mathbf{Y})}{\hat{\sigma}_{\mathbf{X}}\hat{\sigma}_{\mathbf{Y}}} \\
&= \frac{(\mathbf{X}-\overline{X}\mathbf{1})^T(\mathbf{Y}-\overline{Y}\mathbf{1})}{\sqrt{(\mathbf{X}-\overline{X}\mathbf{1})^T(\mathbf{X}-\overline{X}\mathbf{1})}\sqrt{(\mathbf{Y}-\overline{Y}\mathbf{1})^T(\mathbf{Y}-\overline{Y}\mathbf{1})}} \\
&= \frac{<\mathbf{X}-\overline{X}\mathbf{1},\mathbf{Y}-\overline{Y}\mathbf{1}>}{\|\mathbf{X}-\overline{X}\mathbf{1}\|\|\mathbf{Y}-\overline{Y}\mathbf{1}\|} \\
&= cos\theta
\end{aligned}$$

where θ is the angle between $\mathbf{X}-\overline{X}\mathbf{1}$ and $\mathbf{Y}-\overline{Y}\mathbf{1}$. □

Using this fact, we can finally derive that

$$\hat{\rho}(\mathbf{Y}, X\boldsymbol{\beta}) = \cos\theta \tag{2.18}$$

where θ is the angle between $\mathbf{Y}-\overline{Y}\mathbf{1}$ and

$$\beta_1(\mathbf{x}_1-\overline{x}_1\mathbf{1})+\cdots\beta_p(\mathbf{x}_p-\overline{x}_p\mathbf{1}) = \mathbb{X}_{A,\perp}\boldsymbol{\beta}_A \in Col(\mathbb{X}_{A,\perp})$$

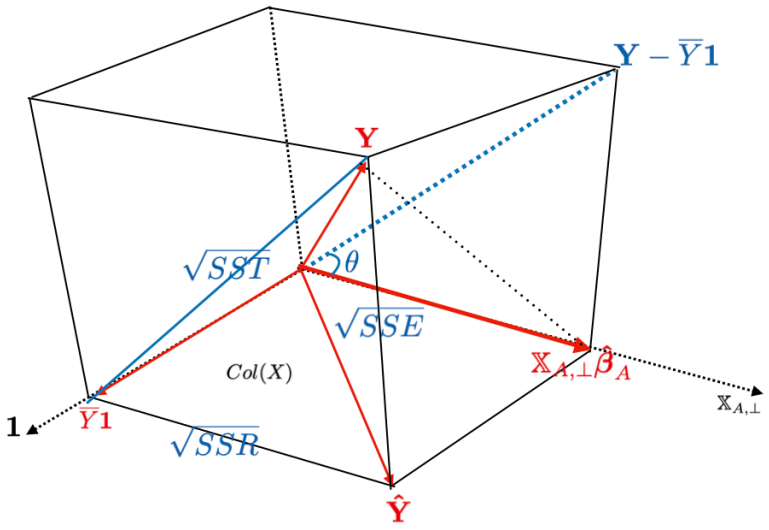


Figure 2.9: R^2 and Projection

See Figure 2.9. θ is marked on Figure 2.7 setting. In this setting, we know that $\theta < \frac{\pi}{2}$ is minimized by projection onto $\mathbb{X}_{A,\perp}$. This means that $\cos\theta$ is maximized. In other words,

$$R^2 = \frac{SSR}{SST} = (\cos\theta)^2$$

is maximized by the projection of $\mathbf{Y} - \overline{Y}\mathbf{1}$ onto $Col(\mathbb{X}_{A,\perp})$. Thus, R^2 can be interpreted as *proportion of variability of Y that is explained by the set of x_js*. It is obvious that $0 \leq R^2 \leq 1$.

R^2 becomes larger if a set of $\mathbb{X}_{A,\perp}$ explains the response well. Is it proper to use this measure as judging *goodness-of-fit*? However, this is not a good measure for model comparison. Model comparison includes different number of predictors. *SSE, however, always decreases when new X_j is added,* while $SST = \sum(Y_i - \overline{Y})$ never changes given Y data. This leads

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

always becomes larger by more predictors. For example,

$$\begin{cases} Y = \beta_0 + \beta_1 X_1 + \epsilon \\ Y = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \epsilon \end{cases}$$

Whether X_2 additionaly contributes to Y significantly, R^2 increases and we could judge that second model is better than first one. Hence to use this properly, we need some adjustment. As $p + 1$ increases, this adjustment should become smaller:

$$\frac{n-1}{n-p-1}$$

Definition 2.3 (Adjusted Rsquared)**.**

$$R_a^2 := 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Remark (Adjustment)*.*

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Proof. Note that

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

and hence,

$$\begin{aligned} R_a^2 &:= 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} \\ &= 1 - \frac{n-1}{n-p-1}\left(\frac{SSE}{SST}\right) \\ &= 1 - \frac{n-1}{n-p-1}(1-R^2) \end{aligned}$$

□

So R_a^2 becomes a useful measure for the goodness-of-fit. On the contrary, *it cannot be interpreted as the proportion of total variation in Y that is explained by $X_1, \ldots, X_p$.*

2.4 Distributions

2.4.1 Individual Regression coefficients

Under Normality,

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

Since $\hat{\boldsymbol{\beta}}$ is an unbiased estimator, Proposition 2.1 implies that

$$\hat{\boldsymbol{\beta}} \sim MVN\Big(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1}\Big) \tag{2.19}$$

Let $C \equiv (X^T X)^{-1}$. Then

$$\begin{bmatrix} \hat{\beta}_0 \\ \hline \hat{\beta}_1 \\ \hline \vdots \\ \hline \hat{\beta}_p \end{bmatrix} \sim MVN\left(\begin{bmatrix} \beta_0 \\ \hline \beta_1 \\ \hline \vdots \\ \hline \beta_p \end{bmatrix}, \sigma^2 \begin{bmatrix} c_{00} & \cdots & \cdots \\ \hline \cdots & c_{11} & \cdots \\ \hline \vdots & \vdots & \vdots \\ \hline \cdots & \cdots & c_{pp} \end{bmatrix} \right)$$

i.e. Individual coefficient follows

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), \quad j = 0, 1, \ldots, p$$

where c_{jj} is $j+1$th diagonal element of $C = (X^TX)^{-1}$. Then

$$\frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{c_{kk}}} \sim N(0,1) \tag{2.20}$$

From Proposition 2.7,

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-p-1)$$

Since $\hat{\sigma}^2 = \frac{SSE}{n-p-1}$ and $\hat{\beta}_k \perp\!\!\!\perp \hat{\sigma}^2$,

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{c_{kk}}} = \frac{(\hat{\beta}_k - \beta_k)/(\sigma\sqrt{c_{kk}}) \sim N(0,1)}{\sqrt{\frac{SSE}{\sigma^2}/(n-p-1) \sim \chi^2(n-p-1)}} \sim t(n-p-1), \quad k = 0, 1, \ldots, p \tag{2.21}$$

2.4.2 Mean response

Consider prediction at $\mathbf{z} = (1, z_1, \ldots, z_p)^T$

Mean response targets

$$\hat{\mu}_{\mathbf{z}} = \mathbf{z}^T\hat{\boldsymbol{\beta}}$$

From Equation (2.19),

$$\hat{\mu}_{\mathbf{z}} \sim N(\mathbf{z}^T\boldsymbol{\beta}, \sigma^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z})$$

Set

$$C_{\mathbf{z}} := \mathbf{z}^T(X^TX)^{-1}\mathbf{z} \tag{2.22}$$

Then by standardization,

$$\frac{\hat{\mu}_{\mathbf{z}} - \mu_{\mathbf{z}}}{\sqrt{C_{\mathbf{z}}\sigma^2}} \sim N(0,1) \tag{2.23}$$

Hence,

$$\frac{\hat{\mu}_{\mathbf{z}} - \mu_{\mathbf{z}}}{\sqrt{C_{\mathbf{z}}\hat{\sigma}^2}} \sim t(n-p-1) \tag{2.24}$$

2.4.3 Response

Now we target $\mathbf{Y}$ at $\mathbf{z} = (1, z_1, \ldots, z_p)^T$ *out-of-sample.* $\boldsymbol{\epsilon}_\mathbf{z}$ at this point is independent of the data set.

Consider

$$\hat{Y}_\mathbf{z} - Y_\mathbf{z} = \mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \epsilon_\mathbf{z} \tag{2.25}$$

As in Proposition 1.7, it can be proven that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \epsilon_\mathbf{z} \end{bmatrix} \sim MVN$$

by showing marginal follows Normal and two are independent. First part - marginal follows normal distribution - has been already shown and assumed. Since these are Normal, it is enough to show covariance is zero.

$$\begin{aligned} Cov((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \epsilon_\mathbf{z}) &= Cov\Big((X^TX)^{-1}X^TY, \epsilon_\mathbf{z}\Big) \\ &= (X^TX)^{-1}X^TCov(Y, \epsilon_\mathbf{z}) \\ &= 0 \end{aligned}$$

Hence, the joint distribution

$$\left[\begin{array}{c} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hline \epsilon_\mathbf{z} \end{array}\right] \sim MVN\left(\mathbf{0}, \left[\begin{array}{c|c} \sigma^2(X^TX)^{-1} & 0 \\ \hline 0 & \sigma^2 \end{array}\right]\right)$$

From Equation (2.25),

$$\begin{aligned} \hat{Y}_\mathbf{z} - Y_\mathbf{z} &= \left[\begin{array}{c|c} \mathbf{z}^T & 1 \end{array}\right] \left[\begin{array}{c} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hline \epsilon_\mathbf{z} \end{array}\right] \\ &\sim N\left(\mathbf{0}, \left[\begin{array}{c|c} \mathbf{z}^T & 1 \end{array}\right] \left[\begin{array}{c|c} \sigma^2(X^TX)^{-1} & 0 \\ \hline 0 & \sigma^2 \end{array}\right] \left[\begin{array}{c} \mathbf{z} \\ 1 \end{array}\right]\right) \\ &\overset{d}{=} N\Big(\mathbf{0}, \sigma^2\mathbf{z}^T(X^TX)^{-1}\mathbf{z} + \sigma^2\Big) \end{aligned}$$

Using notation from Equation (2.22), we get

$$\hat{\mathbf{Y}}_\mathbf{z} - \mathbf{Y}_\mathbf{z} \sim N\Big(\mathbf{0}, (C_\mathbf{z} + 1)\sigma^2\Big) \tag{2.26}$$

Then standardization gives that

$$\frac{\hat{Y}_\mathbf{z} - Y_\mathbf{z}}{\sqrt{(C_\mathbf{z} + 1)\sigma^2}} \sim N(0, 1) \tag{2.27}$$

and hence,

$$\frac{\hat{Y}_\mathbf{z} - Y_\mathbf{z}}{\sqrt{(C_\mathbf{z} + 1)\hat{\sigma}^2}} \sim t(n - p - 1) \tag{2.28}$$

Compare Statistic (2.28) with Statistic (2.24). We can see that out-of-sample one has larger standard error by σ^2 with same degrees of freedom. This is same as simple regression setting. Error of mean response only comes from $\hat{\boldsymbol{\beta}}$. When we predict out-of-sample individual, however, $Var(\epsilon_{\mathbf{z}})$ is added. Denote that this $\epsilon_{\mathbf{z}}$ should be independent with given data. Otherwise, we cannot get the distribution as ordinary (Johnson and Wichern, 2013).

2.5 Statistical Inference

We have derived basic distributions, so we try to test or build a confidence interval.

2.5.1 Individual Regression coefficients

From Statistic (2.21), we can easily make $(1-\alpha)100\%$ confidence interval $\hat{\theta} \pm t_{\frac{\alpha}{2}} SE$ and test statistic.

Theorem 2.12 ($(1-\alpha)100\%$ Confidence interval)**.** *$(1-\alpha)100\%$ confidence interval for each individual β_j is*

$$\left[\hat{\beta}_j \pm t_{\frac{\alpha}{2}}(n-p-1)\hat{\sigma}\sqrt{c_{kk}}\right]$$

In fact, we have already seen the test statistic form.

Theorem 2.13 (Partial t-test)**.** *Test $H_0 : \beta_k = \alpha_k$ vs $H_1 : \beta_k \neq \alpha_k$. For given data, partial t-test computes*

$$T_0 = \frac{\hat{\beta}_k - \alpha_k}{\hat{\sigma}\sqrt{c_{kk}}} \sim t(n-p-1)$$

where $(X^TX)^{-1} = (c_{ij})_{0 \leq i,j \leq p}$

As usual test, we reject H_0 when

$$|T_0| > t_{\frac{\alpha}{2}}(n-p-1)$$

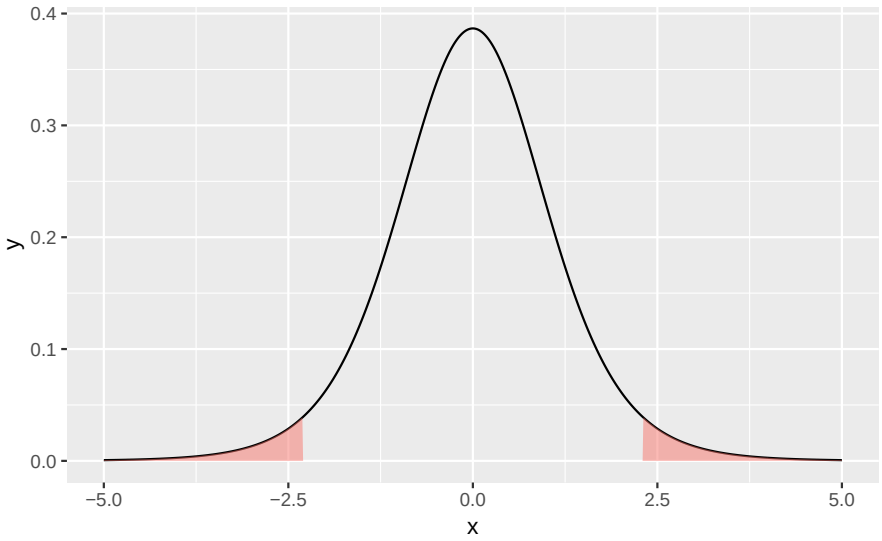


Figure 2.10: Rejection region for β_k

If we use `summary()` to `lm` object, we can get each T statistic(`t value`), standard error(`Std. Error`), and p-value(`Pr(>|t|)`). These are the results of partial t-test.

```
summary(cem_fit)
#>
#> Call:
#> lm(formula = y ~ ., data = cem)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -3.175 -1.671  0.251  1.378  3.925
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   62.405     70.071    0.89    0.399
#> x1             1.551      0.745    2.08    0.071 .
#> x2             0.510      0.724    0.70    0.501
#> x3             0.102      0.755    0.14    0.896
#> x4            -0.144      0.709   -0.20    0.844
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 2.45 on 8 degrees of freedom
#> Multiple R-squared:  0.982,  Adjusted R-squared:  0.974
#> F-statistic:  111 on 4 and 8 DF,  p-value: 4.76e-07
```

If the test is significant, it means that additional contribution of that variable is significant after all other variables are already in the model. This might be understood well with extra sum of squares concept, later.

2.5.2 Prediction interval

Consider prediction at $\mathbf{z} = (1, z_1, \ldots, z_p)^T$. Write

$$C_{\mathbf{z}} := \mathbf{z}^T (X^T X)^{-1} \mathbf{z}$$

Theorem 2.14 (Prediction or Confidence interval for mean response)**.** $(1-\alpha)100\%$ *prediction interval for* $\mu_{\mathbf{z}}$ *is*

$$\left[\hat{\mu}_{\mathbf{z}} \pm t_{\frac{\alpha}{2}}(n-p-1)\sqrt{C_{\mathbf{z}}\hat{\sigma}^2}\right]$$

Theorem 2.15 (Out-of-sample prediction interval)**.** $(1-\alpha)100\%$ *prediction interval for* $Y_{\mathbf{z}}$ *is*

$$\left[\hat{Y}_{\mathbf{z}} \pm t_{\frac{\alpha}{2}}(n-p-1)\sqrt{(C_{\mathbf{z}}+1)\hat{\sigma}^2}\right]$$

See standard error part of Theorem 2.14 and Theorem 2.15. As mentioned, Out of sample prediction interval *always has larger standard error.* This leads to wider interval.

2.5.3 Regression coefficient vector

Now we consider the coefficients simultaneously. For example, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

Note that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim MVN(\mathbf{0}, \sigma^2 (X^T X)^{-1})$$

Then standardization gives

$$\mathbf{Z} \equiv \frac{(X^T X)^{\frac{1}{2}}}{\sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim MVN(\mathbf{0}, I)$$

It follows that

$$\mathbf{Z}^T \mathbf{Z} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2(p+1)$$

Since $\frac{SSE}{\sigma^2} \sim \chi^2(n-p-1)$ and $\hat{\sigma}^2 = MSE$,

$$\begin{aligned} \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{\sigma}^2 (p+1)} &= \frac{(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^T (X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})}{MSE} \\ &= \frac{SSR/(p+1)}{SSE/(n-p-1)} \\ &\sim F(p+1, n-p-1, \delta) \end{aligned} \tag{2.29}$$

where $\delta = \frac{\boldsymbol{\beta}^T X^T (I - \Pi_1) X \boldsymbol{\beta}}{\sigma^2}$.

Corollary 2.2 (F-test). *Test* $H_0 : \boldsymbol{\beta} = \mathbf{0}$ *vs* $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. *For given data, F-test computes*

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\hat{\sigma}^2 (p+1)} = \frac{SSR/(p+1)}{MSE} \overset{H_0}{\sim} F(p+1, n-p-1)$$

From the first part, we can get the confidence region for $\boldsymbol{\beta}$.

Theorem 2.16 (Confidence region). $(1-\alpha)100\%$ *confidence interval for* $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ *is*

$$\left\{ \boldsymbol{\beta} : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1) \hat{\sigma}^2 F_{1-\alpha}(p+1, n-p-1) \right\}$$

Remark. The confidence region for the vector $\boldsymbol{\beta}$ is the ellipsoid that is centered at $\hat{\boldsymbol{\beta}}$. Eigenvectors and eigenvalues of $X^T X$ determines its orientation and size, respectively. See Johnson and Wichern (2013) for details.

`ellipse::ellipse()` has method for `lm` object. So if you provides the regression object, it will give ellipsoid coordinate as `matrix`. However, this function only supports two variables. By specifying `which` argument, you can select which variable to get coordinates. By default, first two variables `c(1, 2)`.

```
ellipse::ellipse(cem_fit) %>%
  tbl_df() %>% # change to data frame
  ggplot(aes(x = `(Intercept)`, y = x1)) +
  geom_path()
```

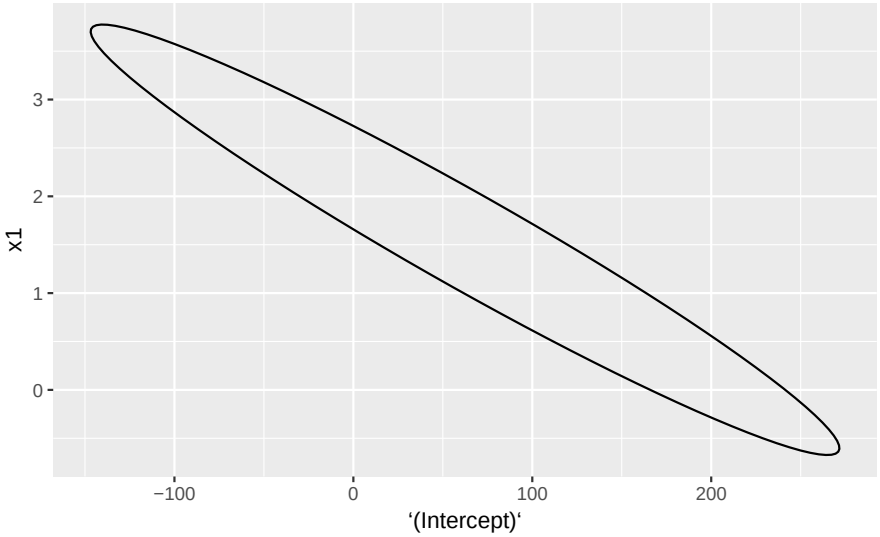


Figure 2.11: Confidence region for (β_0, β_1)

Look again corollary 2.2. Compare this to ANOVA. Something is different. When testing significance in ANOVA, degrees of freedom of SSR was p. Because when testing regression relation, we only need β_1 to β_p, i.e. p parameters. How can we do this here?

2.5.4 Regression coefficient vector $\boldsymbol{\beta}_A$

Consider $\boldsymbol{\beta}_A = (\beta_1, \ldots, \beta_p)^T$. In fact, these tell us significance of variables. We can use $\mathbb{X}_{A,\perp}$ defined before. From Equation (2.5),

$$\hat{\boldsymbol{\beta}}_A = (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1} \mathbb{X}_{A,\perp}^T \mathbf{Y}$$

The reason using $\mathbb{X}_{A,\perp}$ is to decomposing the space orthogonally.

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \begin{bmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \sigma^2 (X^T X)^{-1} \right)$$

with

$$Var(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A) = \sigma^2 (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1}$$

Theorem 2.17 (Confidence region)**.** $(1-\alpha)100\%$ *confidence interval for* $\boldsymbol{\beta}_A = (\beta_1, \ldots, \beta_p)^T$ *is*

$$\left\{ \boldsymbol{\beta}_A : (\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A)^T \mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp} (\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A) \leq p\hat{\sigma}^2 F_\alpha(p, n-p-1) \right\}$$

where $\hat{\boldsymbol{\beta}}_A = (\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp})^{-1} \mathbb{X}_{A,\perp}^T \mathbf{Y}$

This tests the same hypothesis as ANOVA for significance.

Remark. As for X, the confidence region for the vector $\boldsymbol{\beta}_A$ is the ellipsoid that is centered at $\hat{\boldsymbol{\beta}}_A$. Eigenvectors and eigenvalues of $\mathbb{X}_{A,\perp}^T \mathbb{X}_{A,\perp}$ determines its orientation and size, respectively.

2.6 Nested Models

We have seen t-test and F-test testing $\beta_j = 0$ and $\beta_1 = \cdots = \beta_p = 0$, respectively. Here, we generalize the hypothesis in terms of *nested model*. Consider several types of hypothesis.

Example 2.1 (Examples of three types of hypothesis)**.** We can test if every coefficient is zero or some coefficient is zero, or just test if coefficients are same not specific value.

1. $H_0 : \beta_1 = \cdots = \beta_p = 0$
2. $H_0 : \beta_{p-1} = \beta_p = 0$
3. $H_0 : \beta_{p-1} = \beta_p$

What is the unified approach to these kind of test?

2.6.1 Full model and reduced model

To test some forms of H_0, consider two nested model.

Definition 2.4 (Nested models)**.** We name nested models for the test as follow.

1. ***Full model*** (FM) represents the basic starting model
2. ***Reduced model*** (RM) represents the null model under H_0

In Example 2.1, the basic starting model, i.e. FM is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

See the second case.

$$H_0 : \beta_{p-1} = \beta_p = 0 \qquad \text{vs} \qquad H_1 : \beta_{p-1} \neq 0 \text{ or } \beta_p \neq 0$$

This is equivalent to choosing a model between

$$\begin{cases} FM : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\ RM : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-2} x_{i,p-2} + \epsilon_i \end{cases}$$

Let us define SSR and SSE for each model.

Definition 2.5 (SS of nested models)**.** For each full model and reduced model, compute SSR and SSE.

1. $SSR(FM)$ regression sum of squares after fitting the FM
2. $SSE(FM)$ error sum of squares after fitting the FM
3. $SSR(RM)$ regression sum of squares after fitting the RM
4. $SSR(RM)$ error sum of squares after fitting the RM

By construction, $SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ does not depend on model. This measures variability when no predictor is used. It jus the variation of a response itself. So it becomes consistent with given data.

Remark.

$$SST = SSR(FM) + SSE(FM) = SSR(RM) + SSE(RM)$$

and hence

$$SSE(RM) - SSE(FM) = SSR(FM) - SSR(RM)$$

SSE indicates a variation that cannot be explained by the model. SSR represents a variation explained by the model. We can compare each between the two model. If $SSE(RM) - SSE(FM) = SSR(FM) - SSR(RM)$ is large, then we can think that the full model is explaining given data better than reduced model. Or, SSE says that full model can explain some parts that reduced model have failed to explained. In terms of $H_0 : RM$ vs $H_1 : FM$, it can be said that it is a strong evidene supporting FM against RM. In turn, we conclude that $H_1 : \beta_{p-1} \neq 0$ or $\beta_p \neq 0$ from our first hypothesis set.

Conjecture 2.1.

$$SSR(FM) - SSR(RM) > c \Leftrightarrow \textit{reject}\, H_0$$

We can test any hypothesis using RM-FM pair as in Conjecture 2.1. The question is

1. What is the distribution of $SSR(FM) - SSR(RM)$?
2. Then what is c?

Define two design matrices to build full model and reduced model.

$$Z_1 := \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-2} \\ 1 & x_{21} & \cdots & x_{2,p-2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-2} \end{bmatrix}, \quad \text{and} \quad Z_2 := \begin{bmatrix} x_{1,p-1} & x_{1p} \\ x_{2,p-1} & x_{2p} \\ \vdots & \vdots \\ x_{n,p-1} & x_{np} \end{bmatrix}$$

Z_1 is a design matrix of reduced model, while Z_2 consists of column vector that was not in Z_2, i.e. which were specified in H_0. It is intended to make

$$X^* = \begin{bmatrix} Z_1 \mid Z_2 \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

where X^* can have different column order with original X. It follows that

$$\begin{cases} FM : \mathbf{Y} = X\boldsymbol{\beta}_F + \boldsymbol{\epsilon} \\ RM : \mathbf{Y} = Z_1\boldsymbol{\beta}_R + \boldsymbol{\epsilon} \end{cases}$$

where $\boldsymbol{\beta}_F = (\beta_0, \ldots, \beta_p)^T$ and $\boldsymbol{\beta}_R = (\beta_0, \beta_1, \ldots, \beta_{p-2})^T$. Thus,

$$\begin{cases} SSR(FM) = \mathbf{Y}^T(\Pi_X - \Pi_{\mathbf{1}})\mathbf{Y} \\ SSR(RM) = \mathbf{Y}^T(\Pi_{Z_1} - \Pi_{\mathbf{1}})\mathbf{Y} \end{cases} \tag{2.30}$$

and so

$$\begin{aligned} SSR(FM) - SSR(RM) &= \mathbf{Y}^T(\Pi_X - \Pi_{\mathbf{1}} - \Pi_{Z_1} + \Pi_{\mathbf{1}})\mathbf{Y} \\ &= \mathbf{Y}^T(\Pi_X - \Pi_{Z_1})\mathbf{Y} \end{aligned} \tag{2.31}$$

Lemma 2.4. *$SSR(FM) - SSR(RM)$ satisfies following properties.*

1. *$\Pi_X - \Pi_{Z_1}$ is symmetric and idempotent*
2. *$tr(\Pi_X - \Pi_{Z_1}) = (p+1) -$ number of parameters in reduced model*
3. *$SSR(FM) - SSR(RM) \perp\!\!\!\perp SSE(FM)$*

Proof. Propery of prjection implies that

$$\Pi_X \Pi_{Z_1} = \Pi_{Z_1}$$

Note that

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

Then

$$(\Pi_X - \Pi_{Z_1})(I - \Pi_X) = 0$$

Hence, $SSR(FM) - SSR(RM) \perp\!\!\!\perp SSE(FM)$ □

Go back to our Example 2.1. Denote that $\Pi_X - \Pi_{Z_1}$ implies that $R^2_{p-1} > R^2_{p-2}$. It is easy to get

$$tr(\Pi_X - \Pi_{Z_1}) = (p+1) - (p-1) = 2$$

Then

$$\frac{SSR(FM) - SSR(RM)}{\sigma^2} \overset{H_0}{\sim} \chi^2(2) \tag{2.32}$$

From Proposition 2.7

$$\frac{SSE(FM)}{\sigma^2} \overset{H_0}{\sim} \chi^2(n-p-1)$$

Since $SSR(FM) - SSR(RM) \perp\!\!\!\perp SSE(FM)$,

$$\begin{aligned} \frac{(SSR(FM) - SSR(RM))/2}{SSE(FM)/(n-p-1)} &= \frac{\frac{SSR(FM)-SSR(RM)}{\sigma^2}/2}{\frac{SSE(FM)}{\sigma^2}/(n-p-1)} \\ &\overset{H_0}{\sim} F(2, n-p-1) \end{aligned} \tag{2.33}$$

Hence, for our

$$H_0 : \beta_{p-1} = \beta_p = 0 \qquad \text{vs} \qquad H_1 : \beta_{p-1} \neq 0 \text{ or } \beta_p \neq 0$$

reject H_0 if $\frac{(SSR(FM)-SSR(RM))/2}{SSE(FM)/(n-p-1)} > F_\alpha(2, n-p-1)$.

See the numerator part $SSR(FM) - SSR(RM)$ we have been explored. This can be also written as

$$SSR(X_1, \ldots, X_p) - SSR(X_1, \ldots, X_{p-2})$$

and this is called *extra sum of squares*.

2.6.2 Extra sum of squares

Definition 2.6 (Extra sum of squares)**.** Extra sum of squares measures the amount of the additional contribution of variables added after the variables that were already considered. For example,

$$SSR(X_2 \mid X_1) := SSR(X_1, X_2) - SSR(X_1)$$

This measures marginal contribution of X_2 after X_1 is considred. Here, the contribution means marginal increase of SSR. Equivalently, marginal decrease of SSE.

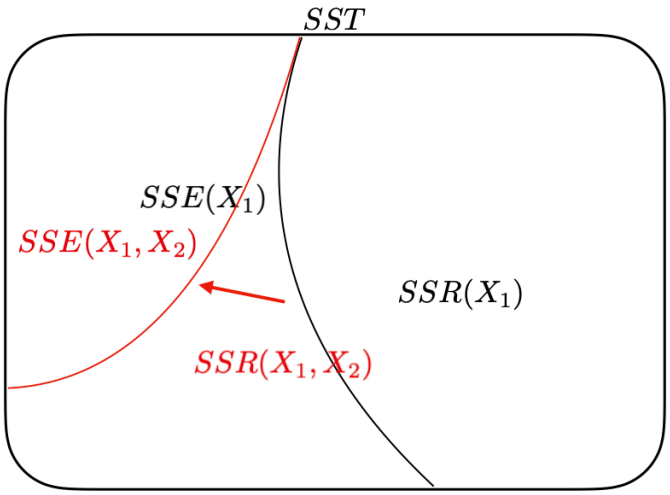


Figure 2.12: Extra sum of squares - SSR increases and SSE decreases

Similarly,

$$SSR(X_3, X_4 \mid X_1, X_2) := SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2)$$

measures of X_3 and X_4 after X_1 and X_2 are already in the model.

SST does not change given data. By adding some variables, SSR increases or SSE decreases. *We interpret this change as additional contribution of those variables.* If it is large enough, they are significant. Using this idea, we might test any kind of sets of coefficients. Partial F-test and sequential F-test are typical forms that we can think of and useful.

2.6.3 Partial F test

Focus on marginal contribution given that every other predictor is already in the model.

Definition 2.7 (Partial sum of squares)**.** Partial SS or Type III Sum of squares are following.

- $SSR(X_1 \mid X_2, \ldots, X_p)$
- $SSR(X_2 \mid X_1, X_3, \ldots, X_p)$
- $SSR(X_3 \mid X_1, X_2, X_4, \ldots, X_p)$

- $\vdots$
- $SSR(X_p \mid X_1, \ldots, X_{p-1})$

Remark. Partial SS satisties following properties.

1. $SSR(X_1 \mid X_2, \ldots X_p) + \cdots + SSR(X_p \mid X_1, \ldots, X_{p-1}) \neq SSR(X_1, \ldots, X_p)$
2. Partial SS (Type III SS) represents the additional contribution of each predictor after the other predictor variables are considred.

Partial F-test aims at testing single coefficient.

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0$$

How to make test statistic is same as previous session: using full model and reduced model. Here, one with and without β_j.

$$F_0 = \frac{SSR(X_j \mid X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)/1}{SSE/(n-p-1)} \overset{H_0}{\sim} F(1, n-p-1) \tag{2.34}$$

Hence, reject H_0 if $F_0 > F_\alpha(1, n-p-1)$. Since the first degrees of freedom is 1, we have

$$F_0 = \left(\frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \right)^2 = T_0^2$$

Remark. Partial F-test is equivalent to partial t-test. Both tests the additional contribution of X_j after the other variables are already considered, not individual significance.

2.6.4 Sequential F test

This sequential F-test adds variable sequentially.

Definition 2.8 (Sequential sum of squares)**.** Sequential SS or Type I Sum of squares are following.

1. $SSR(X_1)$
2. $SSR(X_2 \mid X_1)$
3. $SSR(X_3 \mid X_1, X_2)$
4. $\vdots$
5. $SSR(X_p \mid X_1, \ldots, X_{p-1})$

Remark. Sequential SS satisfies following properties.

1. $SSR(X_1) + SSR(X_2 \mid X_1) + \cdots + SSR(X_p \mid X_1, \ldots, X_{p-1}) = SSR(X_1, \ldots, X_p)$
2. Sequential SS is useful when we consider nested model. That is, when we know that X_1 is the most important, X_2 is the second important, and so on.

Following Definition 2.8, start from intercept-only model.

$j \leftarrow 1$;
repeat
  $\begin{cases} FM : Y_i = \beta_0 + \cdots + \beta_j x_{1j} + \epsilon_i \\ RM : Y_i = \beta_0 + \ldots + \beta_{j-1} x_{1,j-1} + \epsilon_i \end{cases}$;
  $H_0 : \beta_j = 0 (RM) \quad \text{vs} \quad H_1 : \beta_j \neq 0 (FM)$;
  $F_0 = \frac{SSR(X_j | X_1, \ldots, X_{j-1})/1}{SSE(X_1, \ldots, X_j)/(n-j-1)}$;
  reject H_0 if $F_0 > F_{1-\alpha}(1, n-j-3)$;
  $j \leftarrow j + 1$;
until *accept* H_0;
output: accepted model, i.e. final RM

Algorithm 4: Sequential F Test from beginning

Algorithm 4 shows the logic how to add a variable sequentially. It continues until rejecting H_0. In general form, we can test

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0 \qquad \text{vs} \qquad H_1 : \text{not } H_0$$

Here, reject H_0 if

$$\frac{\Big(SSR(X_1, \ldots, X_p) - SSR(X_1, \ldots, X_q)\Big)/(p-q)}{SSE(X_1, \ldots, X_p)/(n-p-1)} > F_\alpha(p-q, n-p-1) \tag{2.35}$$

This form of test might be useful when see the significance of categorical predictor. In regression model, quantative predictor is modeled as multiple dummy variables. In this case, we have to see these coefficients all at once and sequential F-test form might be helpful (Hastie et al., 2013).

2.6.5 General linear hypothesis

See Example 2.1 again. The third one, $H_0 : \beta_{p-1} = \beta_p$ can be also tested using F-test by changing model representation.

$$H_0 : \beta_{p-1} = \beta_p = \alpha \quad \forall \alpha$$

Then null model can be set as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-2} x_{i,p-2} + \alpha(x_{i,p-1} + x_{ip}) + \epsilon_i$$

However, there is a more general test procedure: linear combination. For instance, $\beta_{p-1} = \beta_p$ can be represented as

$$\beta_{p-1} - \beta_p = 0 \Leftrightarrow (0, \ldots, 0, 1, -1)^T \boldsymbol{\beta} = 0$$

Now we analyze this form of hypothesis.

$$H_0 : C\boldsymbol{\beta} = \mathbf{d} \qquad \text{vs} \qquad H_1 : C\boldsymbol{\beta} \neq \mathbf{d}$$

Since $\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$,

$$C\hat{\boldsymbol{\beta}} \sim MVN(C\boldsymbol{\beta}, \sigma^2 C(X^TX)^{-1}C^T)$$

and under $H_0 : C\boldsymbol{\beta} = \mathbf{d}$,

$$C\hat{\boldsymbol{\beta}} \overset{H_0}{\sim} MVN(\mathbf{d}, \sigma^2 C(X^TX)^{-1}C^T)$$

Set

$$\mathbf{Z} \equiv \frac{1}{\sigma}(C(X^TX)^{-1}C^T)^{-\frac{1}{2}}(C\hat{\boldsymbol{\beta}} - \mathbf{d}) \overset{H_0}{\sim} MVN(\mathbf{0}, I)$$

It follows that

$$\mathbf{Z}^T\mathbf{Z} = \frac{1}{\sigma^2}(C\hat{\boldsymbol{\beta}} - \mathbf{d})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\boldsymbol{\beta}} - \mathbf{d}) \overset{H_0}{\sim} \chi^2(q = rank(C))$$

Since $SSE \perp\!\!\!\perp \hat{\boldsymbol{\beta}}$ (See Proposition 2.7),

$$\begin{aligned} \frac{\frac{(C\hat{\boldsymbol{\beta}}-\mathbf{d})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\boldsymbol{\beta}}-\mathbf{d})}{\sigma^2}/q}{\frac{SSE}{\sigma^2}/(n-p-1)} &= \frac{(C\hat{\boldsymbol{\beta}} - \mathbf{d})^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\boldsymbol{\beta}} - \mathbf{d})/q}{\hat{\sigma}^2} \\ &\overset{H_0}{\sim} F(q, n-p-1) \end{aligned} \tag{2.36}$$

2.6.6 Extra SS in R

`anova.lm()` gives extra sum of squres by default, which is *sequential sum of squares*, i.e. type I SS.

```
anova(cem_fit)
#> Analysis of Variance Table
#>
#> Response: y
#>           Df Sum Sq Mean Sq F value  Pr(>F)
#> x1         1   1450    1450  242.37 2.9e-07 ***
#> x2         1   1208    1208  201.87 5.9e-07 ***
#> x3         1     10      10    1.64    0.24
#> x4         1      0       0    0.04    0.84
#> Residuals  8     48       6
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we change the order,

```
anova(lm(y ~ x2 + x1 + x3 + x4, data = cem))
#> Analysis of Variance Table
#>
#> Response: y
#>           Df Sum Sq Mean Sq F value  Pr(>F)
#> x2         1   1809    1809  302.43 1.2e-07 ***
#> x1         1    848     848  141.81 2.3e-06 ***
#> x3         1     10      10    1.64    0.24
#> x4         1      0       0    0.04    0.84
#> Residuals  8     48       6
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SS differs. To get Type III SS, we can use `car::Anova()`. This function can compute type II or type III sum of squares. Since it gives type II by default, we should specify `type = "III"` or `type = 3`.

```
car::Anova(cem_fit, type = 3)
#> Anova Table (Type III tests)
#>
#> Response: y
#>             Sum Sq Df F value Pr(>F)
#> (Intercept)    4.7  1    0.79  0.399
#> x1            26.0  1    4.34  0.071 .
#> x2             3.0  1    0.50  0.501
#> x3             0.1  1    0.02  0.896
#> x4             0.2  1    0.04  0.844
#> Residuals     47.9  8
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As base `anova.lm()`, we can also tidy this object.

```
car::Anova(cem_fit, type = 3) %>%
  broom::tidy()
#> # A tibble: 6 x 5
#>   term         sumsq    df statistic p.value
#>   <chr>        <dbl> <dbl>     <dbl>   <dbl>
#> 1 (Intercept)  4.75      1    0.793   0.399
#> 2 x1          26.0       1    4.34    0.0708
#> 3 x2           2.97      1    0.497   0.501
#> 4 x3           0.109     1    0.0182  0.896
#> 5 x4           0.247     1    0.0413  0.844
#> 6 Residuals   47.9       8   NA      NA
```

Look p-value of the type III SS, i.e. partial F-test. Comparing to partial t-test, we can see both are equivalent.

```
car::Anova(cem_fit, type = 3) %>%
  broom::tidy() %>%
  na.omit() %>%
  select(p.value) %>%
  bind_cols(t_test = broom::tidy(cem_fit)$p.value)
#> # A tibble: 5 x 2
#>   p.value t_test
#>     <dbl>  <dbl>
#> 1  0.399  0.399
#> 2  0.0708 0.0708
#> 3  0.501  0.501
#> 4  0.896  0.896
#> 5  0.844  0.844
```

As we can see in the argument of `car::Anova()`, there are also type II and type IV SS. These are not popular ones, though.

`car::linearHypothesis()` function peforms test for linear combination. Just specify C matrix to `hypothesis.matrix`. If you want additional $\mathbf{d}$, specify `rhs`. If this is `NULL` (by default), test will be done with zero. Try $H_0 : \beta_3 = \beta_4$.

```
car::linearHypothesis(cem_fit, hypothesis.matrix = c(0, 0, 0, 1, -1))
#> Linear hypothesis test
#>
```

```
#> Hypothesis:
#> x3 - x4 = 0
#>
#> Model 1: restricted model
#> Model 2: y ~ x1 + x2 + x3 + x4
#>
#>   Res.Df  RSS Df Sum of Sq    F Pr(>F)
#> 1      9 57.2
#> 2      8 47.9  1      9.38 1.57   0.25
```

2.7 Qualitative Variables as Predictors

See the example data set from Chatterjee and Hadi (2015).

```
salary <- read_delim("data/P124.txt", delim = "\t")
(salary <-
  salary %>%
  mutate_at(.vars = vars("E", "M"), .funs = list(~factor)))
#> # A tibble: 46 x 4
#>        S     X E     M
#>    <dbl> <dbl> <fct> <fct>
#>  1 13876     1 1     1
#>  2 11608     1 3     0
#>  3 18701     1 3     1
#>  4 11283     1 2     0
#>  5 11767     1 3     0
#>  6 20872     2 2     1
#>  7 11772     2 2     0
#>  8 10535     2 1     0
#>  9 12195     2 3     0
#> 10 12313     3 2     0
#> # ... with 36 more rows
```

- `S`: salary (*response variable*)
- `X`: year of experiece
- `E`: education level
  - `1`: high school (HS)
  - `2`: bachelor degree (BS)
  - `3`: advanced degree (ADV)
- `M`: management status
  - `1`: person with management responsibility (MGT)
  - `0`: otherwise (None)

Unlike with previous setting, we have two qualitative variables `E` and `M`.

```
salary %>%
  ggplot(aes(x = X, y = S)) +
  geom_point(aes(colour = E, shape = M)) +
  labs(
    x = "Experience (year)",
    y = "Salary"
  ) +
  scale_colour_discrete(
    name = "Education",
```

```
    label = c("High School", "Bachelor", "Advanced")
  ) +
  scale_shape_discrete(
    name = "Management",
    label = c("None", "Management")
  )
```

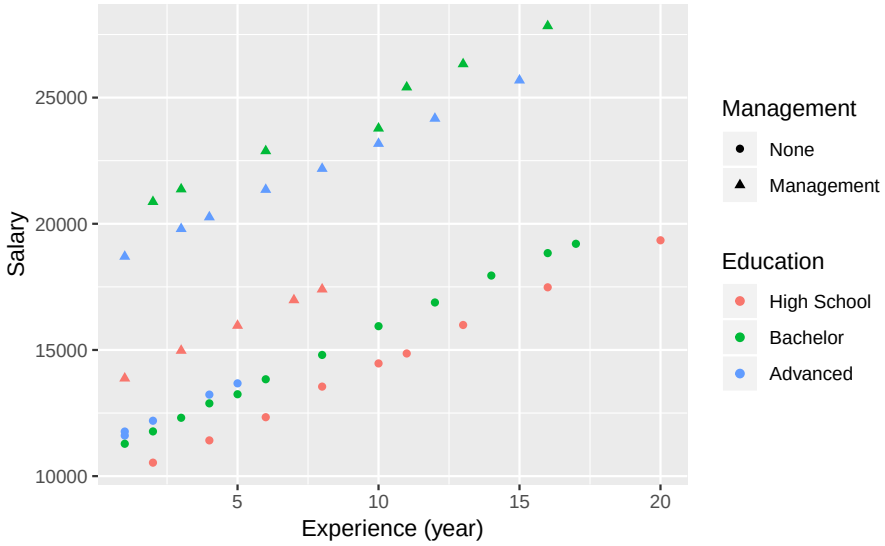


Figure 2.13: Salary Survey Data

See Figure 2.13. There are 6 combinations of `E-M`. For each level, we can see obvious linear relationship between `S` (salary) and `X` (experience). Also, person in management statement has larer salary than not. How can we deal with these variables?

2.7.1 Separate models

The easiest way that we can think of is fit separate regression models for

1. different levels of the qualitative predictor (in case there is only one qualitative predictor)
2. *different combinations of the levels.*

In the above salary survey dataset, we can make 6 models.

levels	HS	BS	ADV
MGT	$Y_i = \beta_{01} + \beta_{11}x_i + \epsilon_{i1}$	$Y_i = \beta_{02} + \beta_{12}x_i + \epsilon_{i2}$	$Y_i = \beta_{03} + \beta_{13}x_i + \epsilon_{i3}$
None	$Y_i = \beta_{04} + \beta_{14}x_i + \epsilon_{i4}$	$Y_i = \beta_{05} + \beta_{15}x_i + \epsilon_{i5}$	$Y_i = \beta_{06} + \beta_{16}x_i + \epsilon_{i6}$

2.7.2 Dummy variables

However, we want to explain linear relationship between response variable and qualitative predictors. Commonly, this is done by defining **dummy variables** or **indicator variables**. For instance, Define E_{i1} and E_{i2} by

$$E_{i1} = \begin{cases} 1 & E = HS \\ 0 & \text{o/w} \end{cases}$$
$$E_{i2} = \begin{cases} 1 & E = BS \\ 0 & \text{o/w} \end{cases}$$

In other words, the last level ADV(3) has value of $E_{i1} = E_{i2} = 0$. This is called *dummy coding* of the last level as baseline. The following function gives how the coding is happened. In R, baseline is set to be the first level, i.e. `base = 1` by default.

```
C(salary$E, contr = contr.treatment, base = 3)
#>  [1] 1 3 3 2 3 2 2 1 3 2 1 2 3 1 3 3 2 2 3 1 1 3 2 2 1 2 1 3 1 1 2 3 2 2 1
#> [36] 2 3 1 2 2 3 2 2 1 2 1
#> attr(,"contrasts")
#>   1 2
#> 1 1 0
#> 2 0 1
#> 3 0 0
#> Levels: 1 2 3
```

For management status, define MGT_i by

$$MGT_i = \begin{cases} 1 & M = MGT \\ 0 & M = None \end{cases}$$

This is the case when `base = 1`, i.e. 0(None) in our data set.

```
C(salary$M, contr = contr.treatment)
#>  [1] 1 0 1 0 0 1 0 0 0 0 1 1 1 0 1 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0 1 1 1 0
#> [36] 0 1 0 1 0 1 1 0 0 0 0
#> attr(,"contrasts")
#>   2
#> 0 0
#> 1 1
#> Levels: 0 1
```

The we now have regression model as

$$Y_i = \beta_0 + \beta_1 x_i + \gamma_1 E_{i1} + \gamma_2 E_{i2} + \delta MGT_i + \epsilon_i \tag{2.37}$$

```
salary %>%
  mutate(
    E = C(E, contr = contr.treatment, base = 3) # different with default
  ) %>%
  lm(S ~ ., data = .)
#>
#> Call:
#> lm(formula = S ~ ., data = .)
#>
#> Coefficients:
#> (Intercept)            X           E1           E2           M1
#>       11032          546        -2996          148         6884
```

Plug in each value. Then we can make 6 models as previous section.

$$\begin{cases} Y_i = (\beta_0 + \gamma_1) + \beta_1 x_i + \epsilon_i & \text{HS-None} & E_1 = 1, E_2 = 0, MGT = 0 \\ Y_i = (\beta_0 + \gamma_1 + \delta) + \beta_1 x_i + \epsilon_i & \text{HS-MGT} & E_1 = 1, E_2 = 0, MGT = 1 \\ Y_i = (\beta_0 + \gamma_2) + \beta_1 x_i + \epsilon_i & \text{BS-None} & E_1 = 0, E_2 = 1, MGT = 0 \\ Y_i = (\beta_0 + \gamma_2 + \delta) + \beta_1 x_i + \epsilon_i & \text{BS-MGT} & E_1 = 0, E_2 = 1, MGT = 1 \\ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i & \text{ADV-None} & E_1 = 0, E_2 = 0, MGT = 0 \\ Y_i = (\beta_0 + \delta) + \beta_1 x_i + \epsilon_i & \text{ADV-MGT} & E_1 = 0, E_2 = 0, MGT = 1 \end{cases} \tag{2.38}$$

Observe that every line has same slope β_1.

```
salary %>%
  mutate(
    E = C(E, contr = contr.treatment, base = 3)
  ) %>%
  ggplot(aes(x = X, y = S, colour = E, linetype = M)) +
  geom_point(aes(shape = M), alpha = .7, show.legend = FALSE) +
  geom_smooth(method = "lm", se = FALSE, fullrange = TRUE) +
  labs(
    x = "Experience (year)",
    y = "Salary"
  ) +
  scale_colour_discrete(
    name = "Education",
    label = c("High School", "Bachelor", "Advanced")
  ) +
  scale_linetype_discrete(
    name = "Management",
    label = c("None", "Management")
  )
```

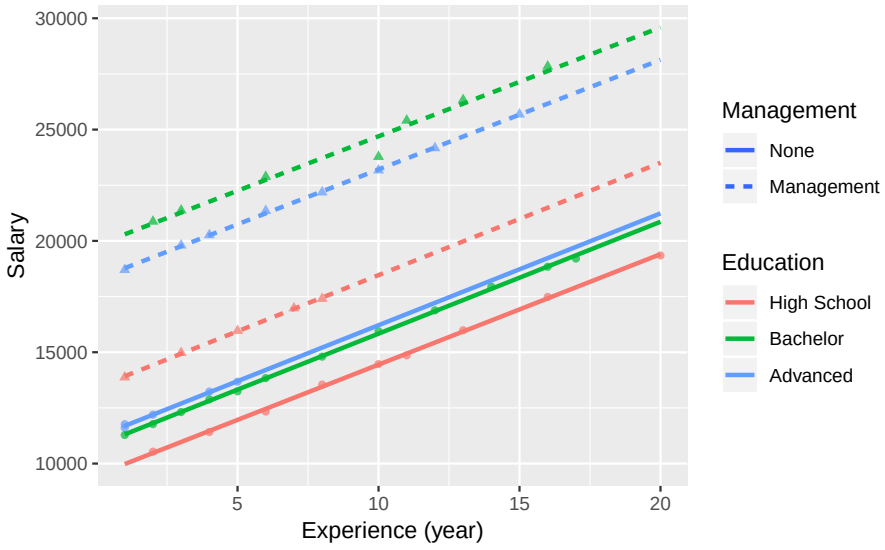


Figure 2.14: Salary Survey Data with Dummy variables

Every line is parallel to each other. In the set of Equations (2.38), we can interpret β_1 as the increment of salary when experience increases in 1 year *when the other predictors are fixed*.

γ_1, the coefficient of E_{i1}, can be interpreted as the increment of salary for `HS` compared to `ADV`(baseline) when the other predictors are fixed. Similarly, we interpret γ_2, the coefficient of E_{i2} as the increment of salary for `BS` compared to the baseline level `ADV` when the other predictors are fixed. The coefficient of MGT δ means the increment of salary for management status for none when the other predictors fixed.

In sum, each coefficient for dummy variables is the increment for corresponding level versus baseline level.

2.7.3 Interaction variables

See Figure 2.13 and focus on education level. Among people in management responsibility, ones with bachelor degrees have the largest salaries and next advanced dgrees. On the other hand, advanced degrees seem to be more important than bachelor among people that are not in management status. The *magnitude of the salary difference between education level also depends management status*. To explain this, we add *interaction term* in the previous model (2.37).

$$Y_i = \beta_0 + \beta_1 x_i + \gamma_1 E_{i1} + \gamma_2 E_{i2} + \delta MGT_i + \alpha_1 (E_{i1} MGT_i) + \alpha_2 (E_{i2} MGT_i) + \epsilon_i \tag{2.39}$$

```
salary %>%
  mutate(
    E = C(E, contr = contr.treatment, base = 3)
  ) %>%
  lm(S ~ X + E * M, data = .)
#>
#> Call:
#> lm(formula = S ~ X + E * M, data = .)
#>
#> Coefficients:
#> (Intercept)            X           E1           E2           M1
#>       11203          497        -1731         -349         7047
#>       E1:M1        E2:M1
#>       -3066         1836
```

$$\begin{cases} Y_i = (\beta_0 + \gamma_1) + \beta_1 x_i + \epsilon_i & \text{HS-None} & E_1 = 1, E_2 = 0, MGT = 0 \\ Y_i = (\beta_0 + \gamma_1 + \delta + \alpha_1) + \beta_1 x_i + \epsilon_i & \text{HS-MGT} & E_1 = 1, E_2 = 0, MGT = 1 \\ Y_i = (\beta_0 + \gamma_2) + \beta_1 x_i + \epsilon_i & \text{BS-None} & E_1 = 0, E_2 = 1, MGT = 0 \\ Y_i = (\beta_0 + \gamma_2 + \delta + \alpha_2) + \beta_1 x_i + \epsilon_i & \text{BS-MGT} & E_1 = 0, E_2 = 1, MGT = 1 \\ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i & \text{ADV-None} & E_1 = 0, E_2 = 0, MGT = 0 \\ Y_i = (\beta_0 + \delta) + \beta_1 x_i + \epsilon_i & \text{ADV-MGT} & E_1 = 0, E_2 = 0, MGT = 1 \end{cases} \tag{2.40}$$

Bibliography

Chatterjee, S. and Hadi, A. S. (2015). *Regression Analysis by Example*. John Wiley & Sons.

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.