1 Some Theory for Simple Linear Regression

1.1 Distribution Theory for b_1

1.1.1 Mean and Variance of b_1

- To determine the mean and variance of the sampling distribution of b_1, we'll need the following fact regarding mean and variance of a **linear combination** of random variables.

> **Fact 1.1** Suppose $Y_1, Y_2, \ldots, Y_n$ are any random variables with $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \sigma_i^2$. Then if $a_1, a_2, \ldots, a_n$ are any constants,
>
> $$\mathrm{E}\left(\sum_{i=1}^{n} a_i Y_i\right) \;=\; \sum_{i=1}^{n} a_i \mu_i \,, \tag{1}$$
>
> and if $Y_1, Y_2, \ldots, Y_n$ are independent,
>
> $$\mathrm{Var}\left(\sum_{i=1}^{n} a_i Y_i\right) \;=\; \sum_{i=1}^{n} a_i^2 \sigma_i^2 \,. \tag{2}$$

- Recall that under the simple linear regression model, $Y_1, Y_2, \ldots, Y_n$ are independent, with

$$Y_i \;\sim\; N(\beta_0 + \beta_1 X_i,\;\; \sigma^2).$$

Recall also that b_1 can be written as a **linear combination** of $Y_1, Y_2, \ldots, Y_n$,

$$b_1 \;=\; \sum_{i=1}^{n} k_i Y_i \tag{3}$$

where $k_i \;=\; (X_i - \bar{X}) / \sum_i (X_i - \bar{X})^2$.

- By (1) and (3),

$$\begin{aligned} \mathrm{E}(b_1) &= \mathrm{E}\left(\sum_{i=1}^{n} k_i Y_i\right) \\ &= \sum_{i=1}^{n} k_i \mathrm{E}(Y_i) \\ &= \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i) \\ &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\beta_0 + \beta_1 X_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \beta_1 \end{aligned} \tag{4}$$

  where the last line follows from the next to last one because

$$\sum_{i=1}^{n}(X_i - \bar{X}) \;=\; 0$$

  implies in (4) that

$$\beta_0 \sum_{i=1}^{n}(X_i - \bar{X}) \;=\; 0$$

  and that

$$\begin{aligned} \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})X_i &= \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})X_i \;-\; 0 \\ &= \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})X_i \;-\; \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})\bar{X} \\ &= \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 . \end{aligned}$$

  Expressions (4) confirm that b_1 is an **unbiased** estimator of β_1.

- By (2) and (3), letting $\boldsymbol{\sigma^2\{b_1\}}$ denote $\mathbf{Var(b_1)}$, we have

$$\begin{aligned} \sigma^2\{b_1\} &= \mathrm{Var}\left(\sum_{i=1}^{n} k_i Y_i\right) \\ &= \sum_{i=1}^{n} k_i^2 \mathrm{Var}(Y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \end{aligned}$$

where we used the fact that $\text{Var}(Y_i) = \sigma^2$ for $i = 1, 2, \ldots, n$.

- To summarize:

> **Mean and Variance of b_1**: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma)$, the mean and variance of b_1 are
>
> $$\begin{aligned} \text{E}(b_1) &= \beta_1 \\ \sigma^2\{b_1\} &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned} \tag{5}$$

1.1.2 Normality of b_1

- The next fact will be used to establish normality of the sampling distribution of b_1:

> **Fact 1.2** Suppose $Y_1, Y_2, \ldots, Y_n$ are independent, with $Y_i \sim N(\mu_i, \sigma_i^2)$. Then for any constants $a_1, a_2, \ldots, a_n$,
>
> $$\sum_{i=1}^n a_i Y_i \;\sim\; N\left(\sum_{i=1}^n a_i \mu_i,\; \sum_{i=1}^n a_i^2 \sigma_i^2\right) \tag{6}$$

Thus **linear combinations** of normal random variables are again **normal**.

- From (3) and (6), we get the ***sampling distribution*** of b_1:

> **Sampling Distribution of b_1**: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma)$,
>
> $$b_1 \;\sim\; N\left(\beta_1,\, \sigma^2\{b_1\}\right) \tag{7}$$
>
> where $\sigma^2\{b_1\}$ is given by (5).

1.2 Distribution Theory for b_0

1.2.1 Mean and Variance of b_0

- It can be shown that b_0 can be written as a **linear combination** of $Y_1, Y_2, \ldots, Y_n$,

$$b_0 \;=\; \sum_{i=1}^n d_i Y_i \tag{8}$$

where $d_i = 1/n - \bar{X}(X_i - \bar{X})/\sum_i (X_i - \bar{X})^2$.

- Using (8) along with (1), (2), and some algebra, and letting $\boldsymbol{\sigma^2\{b_0\}}$ denote $\mathbf{Var(b_0)}$, it can be shown that:

> **Mean and Variance of b_0**: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma)$, the mean and variance of b_0 are
>
> $$\begin{aligned} \mathrm{E}(b_0) &= \beta_0 \\ \sigma^2\{b_0\} &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) . \end{aligned} \tag{9}$$

Thus b_0 is an **unbiased** estimator of β_0.

1.2.2 Normality of b_0

- From (6) and (8), we get the ***sampling distribution*** of b_0:

> **Sampling Distribution of b_0**: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma)$,
>
> $$b_0 \sim N(\beta_0, \sigma^2\{b_0\}) \tag{10}$$
>
> where $\sigma^2\{b_0\}$ is given by (9).

2 Inference for Regression Parameters

2.1 Background Theory

- The (**estimated**) ***standard error*** of a statistic is (an **estimate** of) the standard deviation of its sampling distribution.

Since MSE estimates σ^2, from (5) and (9) the (estimated) **standard errors** of b_1 and b_0, denoted $\boldsymbol{s\{b_1\}}$ and $\boldsymbol{s\{b_0\}}$, are:

(Estimated) Standard Errors of b_1 and b_0:

$$s\{b_1\} = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (X_i - \bar{X})^2}} \tag{11}$$

$$s\{b_0\} = \sqrt{\text{MSE} \cdot \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)} \tag{12}$$

The values of $s\{b_0\}$ and $s\{b_1\}$ are reported by statistical software when a regression analysis is carried out.

- From (7) and (10),

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \sim N(0,1) \qquad \text{and} \qquad \frac{b_0 - \beta_0}{\sigma\{b_0\}} \sim N(0,1) \tag{13}$$

- When we replace $\sigma\{b_1\}$ and $\sigma\{b_0\}$ in (13) by their estimates $s\{b_1\}$ and $s\{b_0\}$, the resulting random variables both follow a ***t* distibution with $n-2$ degrees of freedom**, i.e.

Fact 2.1 Under the simple linear regression model, with the ϵ_i's independent $N(0,\sigma)$,

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2) \qquad \text{and} \qquad \frac{b_0 - \beta_0}{s\{b_0\}} \sim t(n-2)\,. \tag{14}$$

2.2 Confidence Interval and Hypothesis Test for β_1

- ***Confidence interval for β_1*** with level of confidence $100(1-\alpha)\%$:

Confidence Interval for β_1: Under the simple linear regression model, with the ϵ_i's independent $N(0,\sigma)$, a **$100(1-\alpha)\%$ confidence interval for β_1** is

$$b_1 \pm t(\alpha/2, n-2) s\{b_1\}$$

where $t(\alpha/2, n-2)$ is the $100(1-\alpha/2)$th percentile of the $t(n-2)$ distribution.

We can be $100(1-\alpha)\%$ confident that the true (unknown) slope β_1 will be contained in this interval.

- ***Hypothesis test for β_1***: To test

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

the ***t* test statistic** is

> **Test Statistic**: The test statistic for the *t test for a slope* is
>
> $$t = \frac{b_1 - 0}{s\{b_1\}}.$$

When H_0 is true, by Fact 2.1, the test statistic t follows a $t(n-2)$ distribution. **P-values** are tail areas beyond the observed t value under the $t(n-2)$ distribution.

The results of the t test for β_1 are reported by statistical software when a regression analysis is carried out.

Other null-hypothesized values for β_1 could be used in place of zero, and one-sided alternative hypotheses can be tested.

2.3 Confidence Interval and Hypothesis Test for β_0

- ***Confidence interval for β_0*** with level of confidence $100(1-\alpha)\%$:

> **Confidence Interval for β_0**: Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma)$, a **$100(1-\alpha)\%$ confidence interval for β_0** is
>
> $$b_0 \pm t(\alpha/2, n-2)s\{b_0\}$$

We can be $100(1-\alpha)\%$ confident that the true (unknown) intercept β_0 will be contained in this interval.

- ***Hypothesis test for β_0***: To test

$$\begin{aligned} H_0 : \beta_0 &= 0 \\ H_a : \beta_0 &\neq 0 \end{aligned}$$

the ***t* test statistic** is

Test Statistic: The test statistic for the t *test for an intercept* is

$$t = \frac{b_0 - 0}{s\{b_0\}}.$$

When H_0 is true, by Fact 2.1, the test statistic t follows a $t(n-2)$ distribution. **P-values** are tail areas beyond the observed t value under the $t(n-2)$ distribution.

The results of the t test for β_0 are reported by statistical software when a regression analysis is carried out.

Other null-hypothesized values for β_0 could be used in place of zero, and one-sided alternative hypotheses can be tested.