1.5 Properties of LS Estimators

Recall the SLR Model defined as before:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

Assumptions: The **errors** $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ are assumed to

- have *mean zero*: $\mathbb{E}(\varepsilon_i) = 0$
- be *uncorrelated*: $Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j,$
- be *homoscedastic*: $Var(\varepsilon_i) = \sigma^2$ does not depend on i.

We can combine the last two and write it as

$$Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij},\ \text{where } \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

The assumptions on the error terms, imply the following assumptions on the moments of Y **conditional on X**:

Assumptions on $Y|X$

$$\begin{aligned} &\mathbb{E}\left(y_i|x_i\right) = \beta_0 + \beta_1 x_i \\ &Var\left(y_i|x_i\right) = \sigma^2 \\ &Cov\left(y_i, y_j|x_i, x_j\right) = 0,\ i \neq j \end{aligned}$$

Remark: When we evaluate expectation, only y_i's are random and x_i's are treated as known, non-random constants.

1.5.1 Unbiasedness of the LS Estimators

Proposition

Both LS estimators $\hat{\beta}_1$, $\hat{\beta}_0$ are ***unbiased***, i.e.

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1, \quad \mathbb{E}\left(\hat{\beta}_0\right) = \beta_0.$$

Proof for the Slope

Recall that

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) \cdot y_i}{\sum_i (x_i - \bar{x})^2}$$

So, we have

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_1\right) &= \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right] \\
&= \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}\left(y_i\right)}{\sum_i (x_i - \bar{x})^2}, \qquad \text{since the } x_i' s \text{ are known} \\
&= \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}\left(\beta_0 + \beta_1 x_i\right)}{\sum_i (x_i - \bar{x})^2} \\
&= \sum_i c_i \left(\beta_0 + \beta_1 x_i\right), \text{ where } c_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\
&= \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i = \beta_1
\end{aligned}$$

The last result is true since

(i)

$$\begin{aligned}
\sum_i c_i &= \sum_i \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \left(\sum_i x_i - n\bar{x}\right) = 0,
\end{aligned}$$

and (ii)

$$
\begin{aligned}
\sum_i c_i x_i &= \sum_i \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} x_i \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) x_i \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \left(\sum_i (x_i - \bar{x}) x_i + \sum_i (x_i - \bar{x}) \bar{x} \right) \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i \left((x_i - \bar{x}) x_i + (x_i - \bar{x}) \bar{x} \right) \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x})^2 \\
&= 1.
\end{aligned}
$$

where we used the fact that $\sum_i (x_i - \bar{x}) \bar{x} = \bar{x} \sum_i x_i - n\bar{x}^2 = n\bar{x}^2 - n\bar{x}^2 = 0.$

■

Proof for the Intercept

Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. So, we have

$$
\begin{aligned}
\mathbb{E}\left(\hat{\beta}_0\right) &= \mathbb{E}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathbb{E}\left(\bar{y}\right) - \bar{x} \cdot \mathbb{E}\left(\hat{\beta}_1\right) \\
&= \frac{1}{n} \sum_i \mathbb{E}(y_i) - \bar{x} \cdot \beta_1 \\
&= \frac{1}{n} \sum_i (\beta_0 + \beta_1 x_i) - \bar{x} \cdot \beta_1 \\
&= \beta_0 + \bar{x} \cdot \beta_1 - \bar{x} \cdot \beta_1 = \beta_0
\end{aligned}
$$

■

1.5.2 MSE of LS Estimators

Since both estimators are unbiased $\Rightarrow MSE = Variance.$

For the Slope, the variance computes as

$$\begin{aligned}
Var\left(\hat{\beta}_1\right) &= Var\left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}\right] = Var\left(\sum_i c_i y_i\right) \ (c_i \text{ as before}) \\
&= \sum_i c_i^2 \cdot Var(y_i) = \sum_i c_i^2 \sigma^2 \ (\text{from model assumption}) \\
&= \sigma^2 \cdot \sum_i \left(\frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}\right)^2 \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{1}{S_{xx}},
\end{aligned}$$

where we used the fact that

$$\begin{aligned}
\sum_i \left(\frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}\right)^2 &= \frac{\sum_i (x_i - \bar{x})^2}{\left(\sum_i (x_i - \bar{x})^2\right)^2} \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

For the Intercept we have:

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

because $Cov\left(\bar{y}, \hat{\beta}_1\right) = 0$. Let us check the last one:

$$
\begin{aligned}
Cov\left(\bar{y}, \hat{\beta}_1\right) &= Cov\left(\frac{1}{n}\sum_j y_j, \sum_i c_i y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i Cov\left(y_j, y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i Cov\left(y_j, y_i\right) \\
&= \sigma^2 \sum_{i=1}^{n} c_i = 0,
\end{aligned}
$$

The last equation is true since the covariance is non-zero and equal to σ^2 when $i = j$, and zero when $i \neq j$. We also used the fact that $\sum_i c_i = 0$.

MSE of LS Estimators

$$Var\left(\hat{\beta}_1\right) = \sigma^2 \frac{1}{S_{xx}}$$

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

1.5.3 Normal Error Regression Model

So far, we have derived the LS estimators and proved that they are unbiased, without making **any** assumptions on the distribution of $\mathbf{y}$ (or that of the error terms). However, to do inference, we need to impose additional assumptions on the *distribution of the ε's*.

So, for our SLR Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

we assume that

Normality Assumption

$$\varepsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$$

This implies that

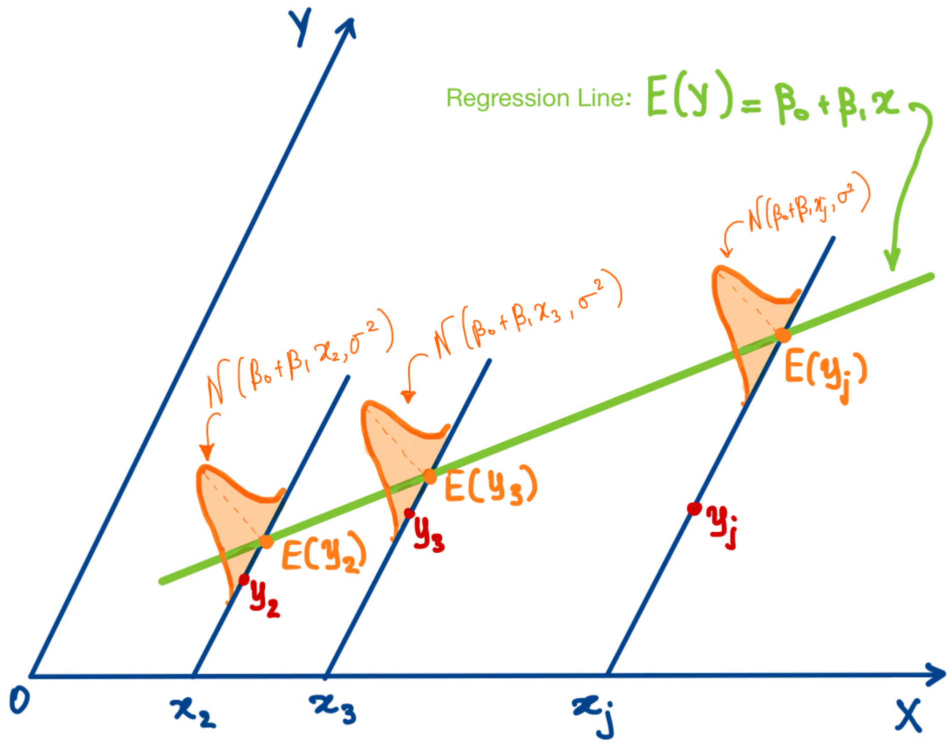
$$y_i \sim^{iid} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Recall that the error terms ε_i are *independent*, *normally distributed* with mean 0 and variance σ^2. Based on that, we can prove the following properties for the y_i's.

Properties of y_i

1. $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$, since the ε_i's have mean zero.
2. y_i's are independent, since ε_i's are independent.
3. $Var(y_i) = Var(\varepsilon_i) = \sigma^2$.
4. y_i's are a linear shift of the ε_i's, so they are also *normally distributed*.
5. The y_i's are **jointly normal**, and so are linear combinations of the y_i's, since the errors are *normally distributed* and *uncorrelated/independent*.

Therefore, if we want to illustrate the **Normal Error SLR model**, we have the following:



1.5.4 Distribution of LS Estimators

Proposition: Joint Distribution of $(\hat{\beta}_1, \hat{\beta}_0)$

Under the Normal error regression model, $\hat{\beta}_1$ and $\hat{\beta}_0$ are *jointly normally distributed* with

$$\mathbb{E}(\hat{\beta}_1) = \beta_1 \quad Var(\hat{\beta}_1) = \sigma^2 \frac{1}{S_{XX}}$$

$$\mathbb{E}(\hat{\beta}_0) = \beta_0 \quad Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)$$

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

Proposition: Estimator for σ^2

Under the Normal error regression model, $RSS = \sum_i (y_i - \hat{y}_i)^2 \sim \sigma^2 \chi^2_{n-2}$ which implies that

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{RSS}{n-2}\right) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$$

i.e. $RSS/(n-2)$ is an *unbiased* estimator of σ^2.

Proposition

Under the Normal error regression model, $(\hat{\beta}_0, \hat{\beta}_1)$ and RSS are *independent*.

Remark: The proof to these statements is beyond the scope of the course.

1.5.5 Hypothesis Testing

Testing for the Slope

Assume we want to test whether the slope takes a specific value c or not.

$$\begin{cases} H_0 : \beta_1 = c \ (null) \\ H_\alpha : \beta_1 \neq c \ (alternative) \end{cases}$$

where c is a known constant – in most cases $c = 0$.

Then, the **test statistic** is formulated as

$$t = \frac{\hat{\beta}_1 - c}{\sqrt{Var(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{xx}}}$$

Under the null, the ***distribution** of t is T_{n-2}* (Student's distribution with $n-2$ degrees of freedom). The p**-value** is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

Remark: By default, `R` outputs the p-value for testing β_1 against 0, i.e. $c = 0$.

Testing for the Intercept

Similarly, we can construct a hypothesis test for the intercept:

$$\begin{cases} H_0 : \beta_0 = c \ (null) \\ H_\alpha : \beta_0 \neq c \ (alternative) \end{cases}$$

The **test statistic** is

$$t = \frac{\hat{\beta}_0 - c}{\sqrt{Var(\hat{\beta}_0)}}$$

Under the null, the *distribution of t is T_{n-2}*. The **p-value** is twice the area under the T_{n-2} distribution more extreme than the observed statistic t.

Remark: By default, `R` ouputs the p-value for testing β_0 against 0, i.e. $c = 0$.

1.5.6 Fitted Regression Line in `R`

In the following example, we are going to fit a regression line to the `admissions` data set (that we also discussed before) and discuss the results using the theory we just discussed.

University Admissions Example (Revisited)

We run the SLR with `gpa` as the response and `entrance-score` as the predictor and we obtain:

```r
admissions.lm = lm(gpa~entrance_score, admissions)
summary(admissions.lm)
```

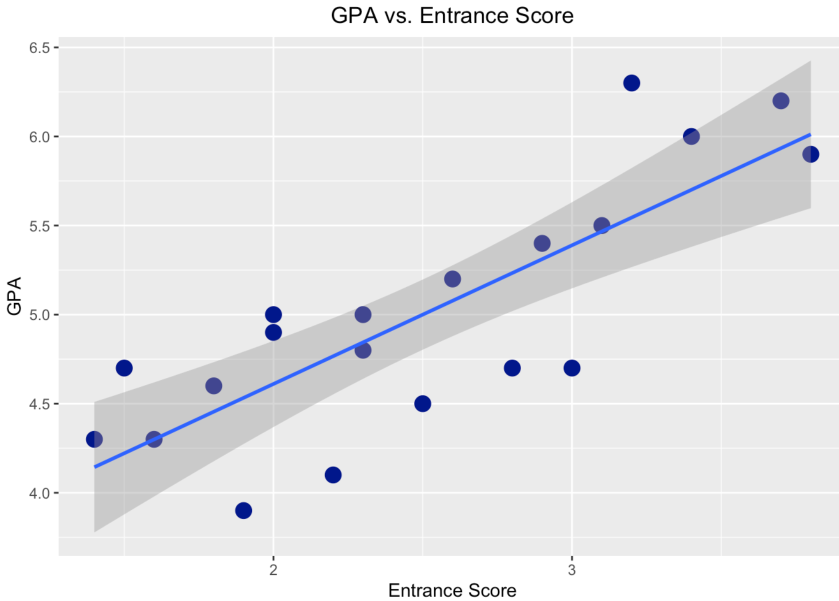
```
##
## Call:
## lm(formula = gpa ~ entrance_score, data = admissions)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6892 -0.2090  0.1054  0.2717  0.7551
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0539     0.3467   8.809 6.05e-08 ***
## entrance_score   0.7785     0.1335   5.831 1.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4188 on 18 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6346
## F-statistic:    34 on 1 and 18 DF,  p-value: 1.597e-05
```

a. Can we say that the regression line has a good fit to the data?

We can plot the regression line along with the connected "point-wise" confidence intervals using `ggplot`:

```
library(ggplot2)
scatterplot = ggplot(admissions, aes(entrance_score, gpa)) +
  geom_point(size=4, color='darkblue') +
  labs(title="GPA vs. Entrance Score",y="GPA",x="Entrance Score") +
  theme(legend.position = "none") +
  theme(plot.title = element_text(hjust = 0.5))+
   geom_smooth(method=lm)
plot(scatterplot)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The regression line has a relatively good fit to the data. There is some variation of the data around the line, but we have to keep in mind that the sample size in this example is relatively small ($n = 20$).

b. By how much relatively is the total variation in the `gpa` *reduced* when the `entrance_score` is introduced into the analysis? Is this a relatively small or large reduction?

The total variation in student's `gpa` when the `entrance_score` is introduced into the analysis is measured by R^2. In this example, the R^2 is approximately 65% with is a medium-sized reduction.

```
summary(admissions.lm)$r.square
```

```
## [1] 0.6538342
```

c. The regression line that we fitted is:

$$[\text{GPA}] = 3.0539 + 0.7785\,[\text{Entrance Score}]$$

As we can see, from the `R` Output, the fitted values for the coefficients are:

$$\hat{\beta}_0 = 3.0539,\ \hat{\beta}_1 = 0.7785$$

```
admissions.coef = summary(admissions.lm)$coef
admissions.coef
```

```
##                Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)    3.0538618  0.3466571 8.809460 6.049507e-08
## entrance_score 0.7784553  0.1335075 5.830797 1.596677e-05
```

d. How do we interpret the slope coefficient *in the context of the problem*?

The difference in the GPA for two students whose Entrance Scores differ by 1 point will be 0.8.

e. How do we interpret the intercept *in the context of the problem*?

In this data set, the intercept does not have a meaningful interpretation. Why? One can argue that in this case a zero score in an entrance exam is a valid score (although I am not sure if anyone with 0 zero score is admitted!)

However, zero is a value that we did not observe in our data, and "plugging-in $x = 0$" to the regression line is an extrapolation that is not necessarily correct. To be more specific, we do not know if the line we fitted will also be valid **outside the range of the data we observed**. Therefore, for *out-of-sample predictions*, we need to be more careful and be aware that the prediction error is significantly larger than within-sample predictions^[More on that later on.

f. How can we test whether or not there is a linear association between `gpa` and `entrance_score_` ? In other words, is the `entrance_score` *statistically significant variable*?

The hypothesis we want to test is formulated as follows:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{cases}$$

If we use the t-test, then t=5.83 corresponding p-value $= 1.59e - 05$ which implies that we reject the null and conclude that the coefficient is *statistically significant* .

We can also compute the p-values *"by hand"*: For the slope:

```
2*pt(-admissions.coef[2,1]/admissions.coef[2,2], 18)
```

```
## [1] 1.596677e-05
```

and for the intercept:

```
2*pt(-admissions.coef[1,1]/admissions.coef[1,2], 18)
```

```
## [1] 6.049507e-08
```

where `18` are the degrees of freedom associated with the residuals. Indeed,

```
admissions.lm$df
```

```
## [1] 18
```