Basics with Simple Linear Regression (SLR)

1 Introduction

We will start this module by introducing the simple linear regression model. Simple linear regression uses the term "simple," because it concerns the study of only one predictor variable with one quantitative response variable. In contrast, multiple linear regression, which we will study in future modules, uses the term "multiple," because it concerns the study of two or more predictor variables with one quantitative response variable. We start with simple linear regression as it is much easier to visualize concepts in regression models when there is only one predictor variable.

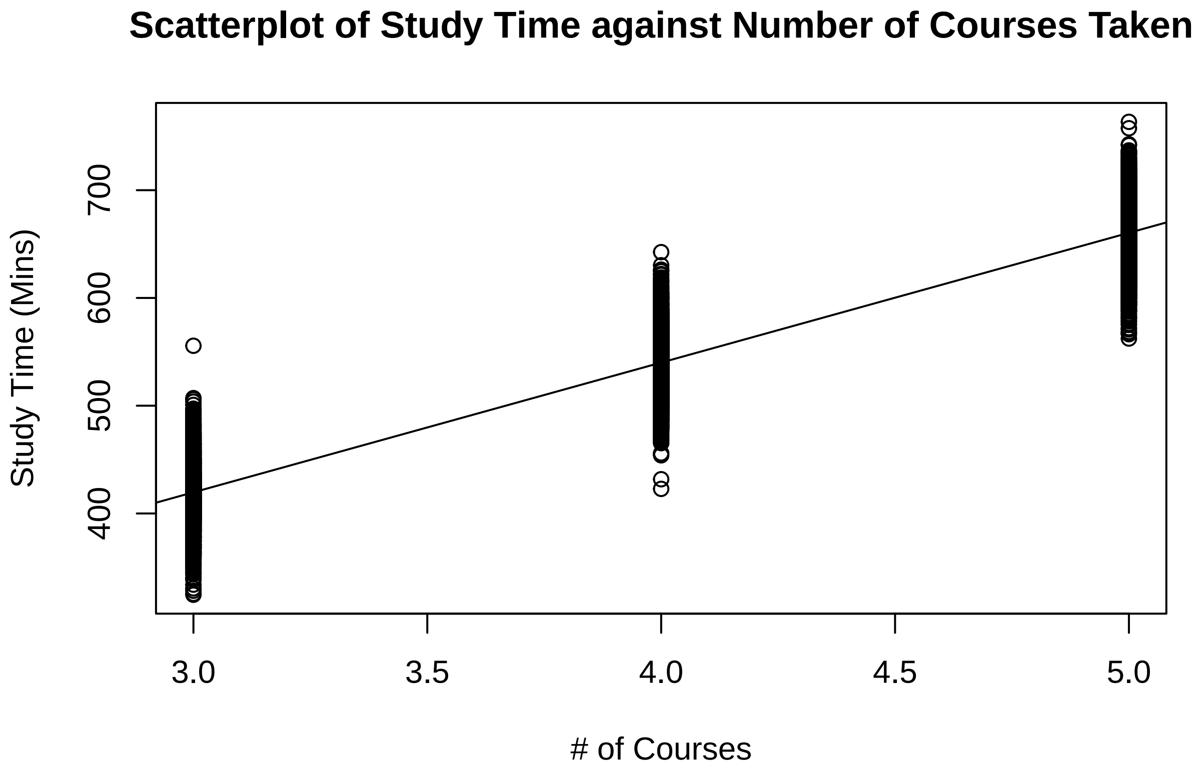
For the time being, we will only consider predictor variables that are quantitative. We will consider predictor variables that are categorical in future modules.

The most common way of visualizing the relationship between one quantitative predictor variable and one quantitative response variable is with a scatter plot. In the simulated example below, we have data from 6000 UVa undergraduate students on the amount of time they spend studying in a week (in minutes), and how many courses they are taking in the semester (3 or 4 credit courses).

```
##create dataframe
df<-data.frame(study,courses)

##fit regression
result<-lm(study~courses, data=df)
```

```
##create scatterplot with regression line overlaid
plot(df$courses, df$study, xlab="# of Courses", ylab="Study Time (Mins)",
     main="Scatterplot of Study Time against Number of Courses Taken")
abline(result)
```



Questions that we may have include:

- Are study time and the number of courses taken related to one another?
- How strong is this relationship?
- Could we use the data to make a prediction for the study time of a student who is not in this scatterplot?
- How confident are we of the prediction?

These questions can be answered using simple linear regression.

Note that we will only be learning about models with just one response variable. We will not cover multivariate regression, which is used when there is more than one response variable. There may be some confusion between "multiple" linear regression and "multivariate" regression due to the closeness in terminology.

1.1 Basic Ideas with Statistics

1.1.1 Population vs Sample

Statistical methods are usually used to make inferences about the **population** based on information from a **sample**.

- A sample is the collection of units that is actually measured or surveyed in a study.
- The population includes all units of interest.

In the study time example above, the population is all UVa undergraduate students, while the sample is the 6000 students that we have data on and are displayed on the scatterplot.

1.1.2 Parameters vs Statistics

- **Parameters** are numerical quantities that describe a population.
- **Statistics** are numerical quantities that describe a sample.

In the study time example, an example of a parameter will be the average study time among all UVa undergraduate students (called the population mean), and an example of a statistic will be the average study time among the 6000 UVa students we have data on (called the sample mean).

Notice that in real life, we will rarely know the actual numerical value of a parameter. So we use the numerical value of the statistic to **estimate** the unknown numerical value of the corresponding parameter.

We also have different notation for parameters and statistics. For example,

- the population mean is denoted as μ.
- the sample mean is denoted as $\bar{x}$.

We say that $\bar{x}$ is an **estimator** of μ.

It is important to pay attention to whether we are describing a statistic (a known value that can be calculated) or a parameter (an unknown value).

1.2 Motivation

Linear regression models generally have two primary uses:

1. **Prediction**: Predict a future value of a response variable, using information from predictor variables.
2. **Association**: Quantify the relationship between variables. How does a change in the predictor variable change the value of the response variable?

We always distinguish between a **response variable, denoted by** y, and a **predictor variable, denoted by** x. In most statistical models, we say that the response variable can be approximated by some **mathematical function, denoted by** f, of the predictor variable, i.e.

$$y \approx f(x).$$

Oftentimes, we write this relationship as

$$y = f(x) + \epsilon,$$

where ϵ **denotes a random error term**, with a mean of 0. The error term cannot be predicted based on the data we have.

There are various statistical methods to estimate f. Once we estimate f, we can use our method for prediction and / or association.

Using the study time example above:

- a prediction example: a student intends to take 4 courses in the semester. What is this student's predicted study time, on average?
- an association example: we want to see how taking more courses increases study time.

1.2.1 Practice questions

In the examples below, are we using a regression model for prediction or for association?

1. It is early in the morning and I am heading out for the rest of the day. I want to know the weather forecast for the rest of the day so I know what to wear.

2. An executive for a sports league wants to assess how increasing the length of commercial breaks may impact the enjoyment of sports fans who watch games on TV.

3. The Education Secretary would like to evaluate how certain factors such as use of technology in classrooms and investment in teacher training and teacher pay are associated with reading skills of students.

4. When buying a home, the prospective buyer would like to know if the home is under- or over- priced, given its characteristics.

2 Simple Linear Regression (SLR)

In simple linear regression (SLR), the function f that relates the predictor variable with the response variable is typically $\beta_0 + \beta_1 x$. Mathematically, we express this as

$$y \approx \beta_0 + \beta_1 x,$$

or in other words, that the response variable has an approximately linear relationship with the predictor variable.

In SLR, this relationship is more explicitly formulated as the **simple linear regression equation**:

$$E(y|x) = \beta_0 + \beta_1 x. \tag{1}$$

- β_0 and β_1 are parameters in the SLR equation, and we want to estimate them.
- These parameters are sometimes called **regression coefficients**.
- β_1 is also called the **slope. It denotes the change in y, on average, when x increases by one unit.**
- β_0 is also called the **intercept. It denotes the average of y when $x = 0$.**
- The notation on the left hand side of (1) denotes the **expected value** of the response variable, for a fixed value of the predictor variable. What (1) implies is that, for each value of the predictor variable x, the expected value of the response variable y is $\beta_0 + \beta_1 x$. The expected value is also the population mean. Applying (1) to our study time example, it implies that:
  - for students who take 3 courses, their expected study time is equal to $\beta_0 + 3\beta_1$,
  - for students who take 4 courses, their expected study time is equal to $\beta_0 + 4\beta_1$,
  - for students who take 5 courses, their expected study time is equal to $\beta_0 + 5\beta_1$.

So $f(x) = \beta_0 + \beta_1 x$ gives us the value of the expected value of the response variable for a specific value of the predictor variable. But, for each value of the predictor variable, the value of the response variable is not a constant. We say that for each value of x, the response variable y has some variance. The variance of the response variable for each value of x is the same as the variance of the error term, ϵ. Thus we have the **simple linear regression model**

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{2}$$

We need to make some assumptions for the error term ϵ. Generally, the assumptions are:

1. The errors have mean 0.
2. The **errors have variance denoted by σ^2**. Notice this variance is constant.
3. The errors are independent.
4. The errors are normally distributed.

From (2), notice we have another parameter, σ^2.

We will go into more detail about what these assumptions mean, and how to assess whether they are met, in module 5.

What these assumptions mean is that for each value of the predictor variable x, the response variable:

1. follows a normal distribution,

2. with mean equal to $\beta_0 + \beta_1 x$,
3. and variance equal to σ^2.

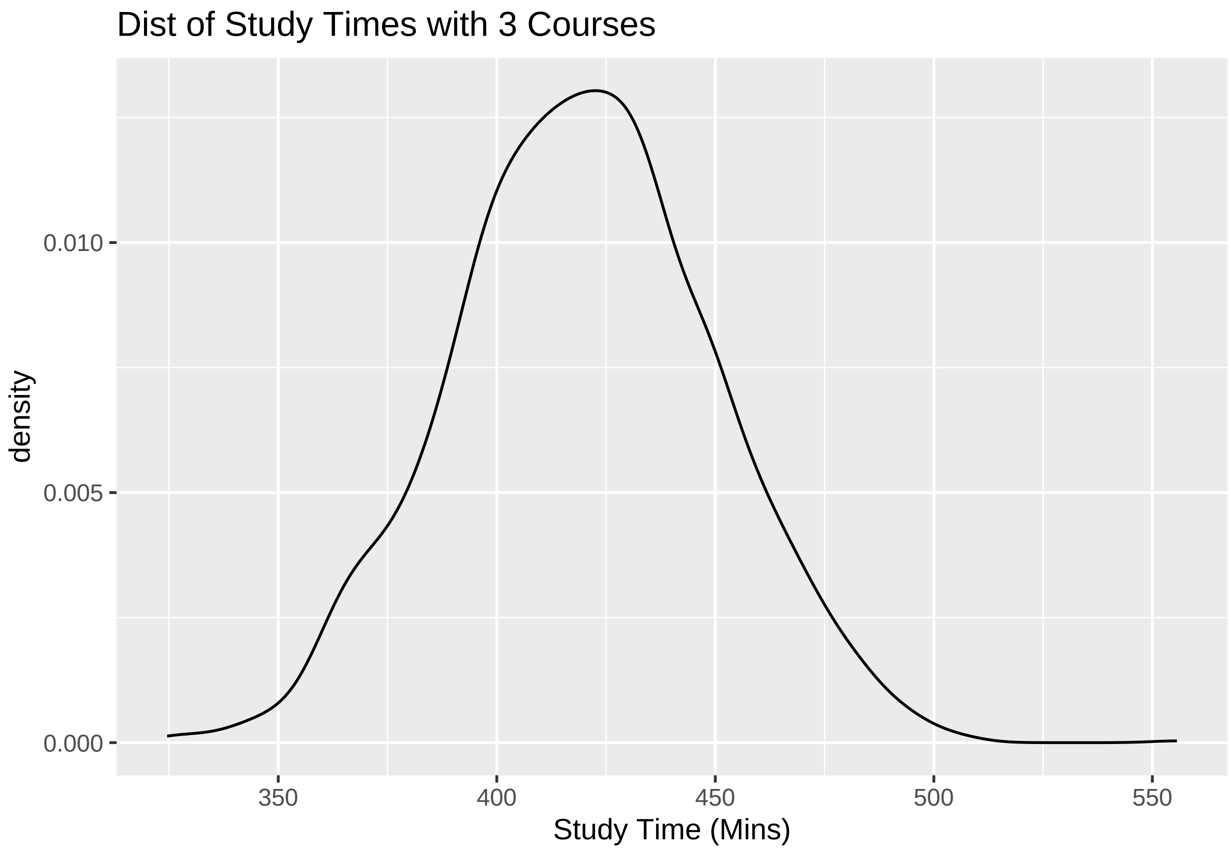
Using our study time example, it means that:

- for students who take 3 courses, the distribution of their study times is $N(\beta_0 + 3\beta_1, \sigma^2)$.
- for students who take 4 courses, the distribution of their study times is $N(\beta_0 + 4\beta_1, \sigma^2)$.
- for students who take 5 courses, the distribution of their study times is $N(\beta_0 + 5\beta_1, \sigma^2)$.

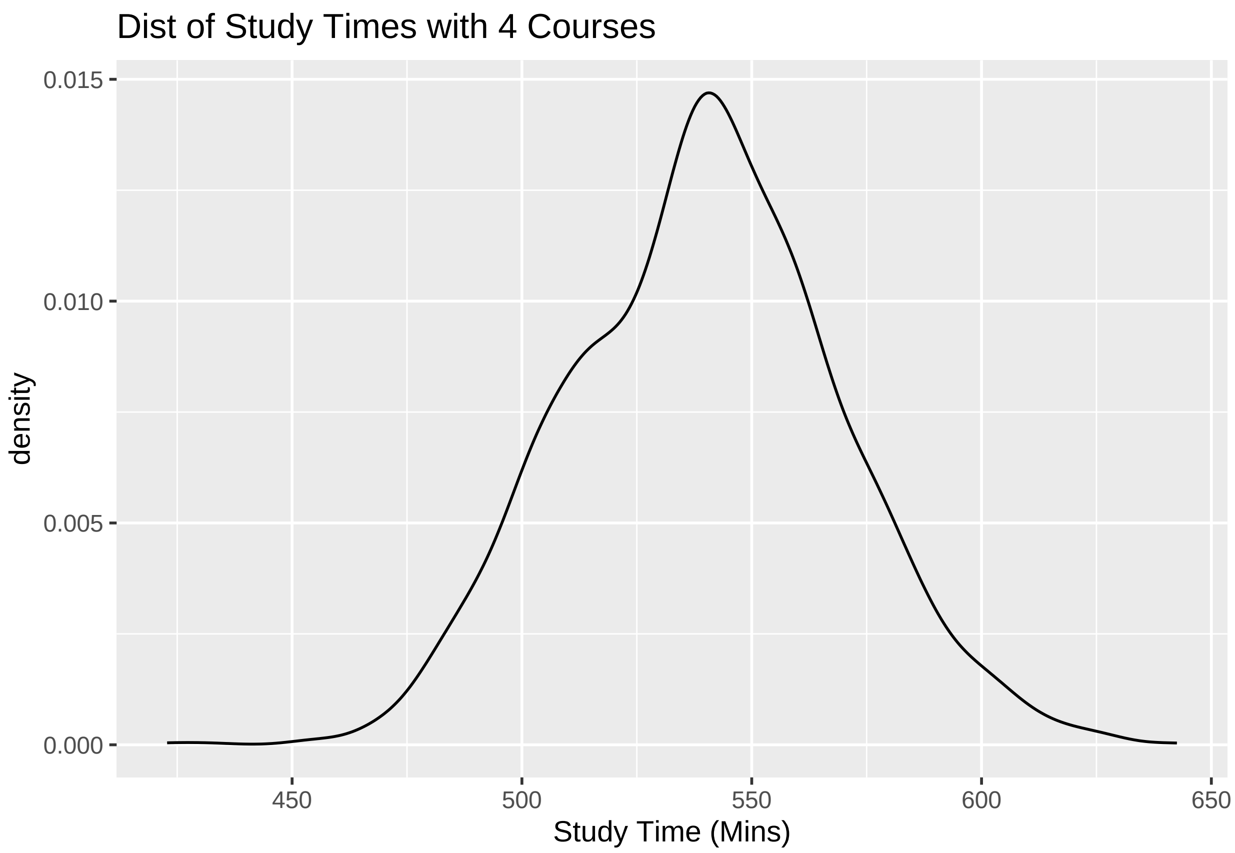
So if we were to subset our dataframe into three subsets, one with students who take 3 courses, another subset for students who take 4 courses, and another subset for students who take 5 courses, and then create a density plot of study times for each subset, each density plot should follow a normal distribution, with different means, and the same spread.

Let us take a look at these density plots next.

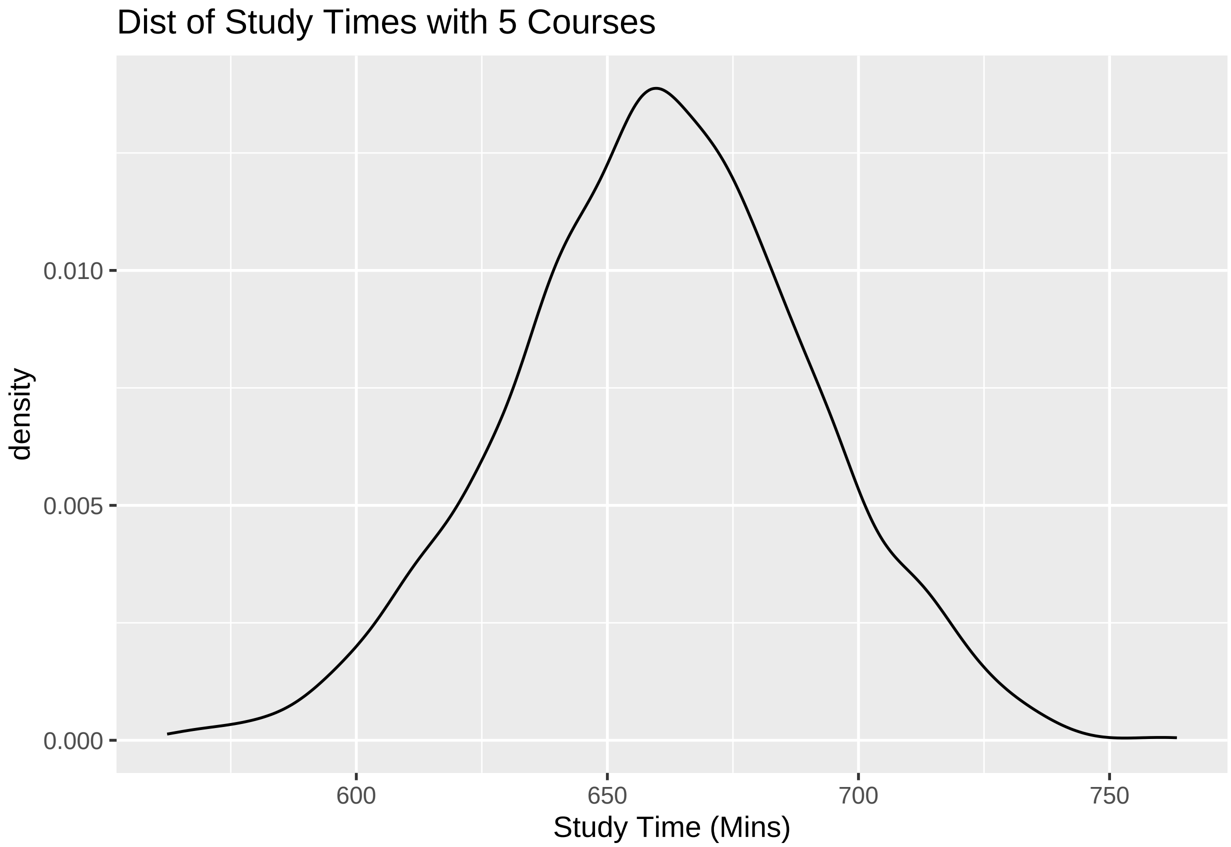
```r
##subset dataframe
x.3<-df[which(df$courses==3),]
##density plot of study time for students taking 3 courses
ggplot(x.3,aes(x=study))+
  geom_density()+
  labs(x="Study Time (Mins)", title="Dist of Study Times with 3 Courses")
```



```r
##subset dataframe
x.4<-df[which(df$courses==4),]
##density plot of study time for students taking 4 courses
ggplot(x.4,aes(x=study))+
  geom_density()+
  labs(x="Study Time (Mins)", title="Dist of Study Times with 4 Courses")
```



```
##subset dataframe
x.5<-df[which(df$courses==5),]
##density plot of study time for students taking 5 courses
ggplot(x.5,aes(x=study))+
  geom_density()+
  labs(x="Study Time (Mins)", title="Dist of Study Times with 5 Courses")
```



Notice all of these plots are normal, with different means (centers), and similar spreads.

3 Estimating Regression Coefficients in SLR

From (1) and (2), we noted that we have to estimate the regression coefficients β_0, β_1 as well as the parameter σ^2 associated with the error term. As mentioned earlier, we are unable to obtain numerical values of these parameters as we do not have data from the entire population. So what we do is use the data from our sample to estimate these parameters.

We estimate β_0, β_1 using $\hat{\beta}_0, \hat{\beta}_1$ based on a sample of observations (x_i, y_i) of size n.

The subscripts associated with the response and predictor variables denote which data point that value belongs to. Let us take a look at the first few rows of the data frame for the study time example:

```
head(df)
```

```
##      study courses
## 1 429.8311       3
## 2 458.4588       3
## 3 391.9406       3
## 4 378.0196       3
## 5 397.9856       3
## 6 405.7145       3
```

For example, x_1 denotes the number of courses taken by student number 1 in the dataframe, which is 3. y_4 denotes the study time for student number 4 in the dataframe, which is 378.0196456.

Following (1) and (2), the sample versions are

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

and

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + e \tag{4}$$

respectively. (3) is called the **estimated SLR equation**, or **fitted SLR equation**. (4) is called the **estimated SLR model**.

$\hat{\beta}_1, \hat{\beta}_0$ are the estimators for β_1, β_0 respectively. These estimators can be interpreted in the following manner:

- **$\hat{\beta}_1$ denotes the change in the predicted y when x increases by 1 unit. Alternatively, it denotes the change in y, on average, when x increases by 1 unit.**
- **$\hat{\beta}_0$ denotes the predicted y when $x = 0$. Alternatively, it denotes the average of y when $x = 0$.**

From (4), notice we use e **to denote the residual**, or in other words, the "error" in the sample.

From (3) and (4), we have the following quantities that we can compute:

$$\text{Predicted/Fitted values: } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{5}$$

$$\text{Residuals: } e_i = y_i - \hat{y}_i. \tag{6}$$

$$\text{Sum of Squared Residuals: } SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{7}$$

We compute the estimated coefficients $\hat{\beta}_1, \hat{\beta}_0$ using the **method of least squares**, i.e. choose the numerical values of $\hat{\beta}_1, \hat{\beta}_0$ that minimize SS_{res} as given in (7).

By minimizing SS_{res} with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated coefficients in the simple linear regression equation are

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \tag{8}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9}$$

$\hat{\beta}_1, \hat{\beta}_0$ are called **least squares estimators**.

The minimization of SS_{res} with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ is done by taking the partial derivatives of (7) with respect to $\hat{\beta}_1$ and $\hat{\beta}_0$, setting these two partial derivatives equal to 0, and solving these two equations for $\hat{\beta}_1$ and $\hat{\beta}_0$.

Let's take a look at the estimated coefficients for our study time example:

```r
##fit regression
result<-lm(study~courses, data=df)
##print out the estimated coefficients
result
```

```
##
## Call:
## lm(formula = study ~ courses, data = df)
##
## Coefficients:
## (Intercept)      courses
##       58.45       120.39
```

From our sample of 6000 students, we have

- $\hat{\beta}_1 = 120.3930985$. The predicted study time increases by 120.3930985 minutes for each additional course taken.
- $\hat{\beta}_0 = 58.4482853$. The predicted study time is 58.4482853 when no courses are taken. Notice this value does not make sense, as a student cannot be taking 0 courses. If you look at our data, the number of courses taken is 3, 4, or 5. So we should only use our regression when $3 \leq x \leq 5$. We cannot use it for values of x outside the range of our data. Making predictions of the response variable for predictors outside the range of the data is called **extrapolation** and should not be done.

4 Estimating Variance of Errors in SLR

The estimator of σ^2, the variance of the error terms (also the variance of the probability distribution of y given x) is

$$s^2 = MS_{res} = \frac{SS_{res}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}, \tag{10}$$

where MS_{res} is the called the **mean squared residuals**.

σ^2, the variance of the error terms, measures the spread of the response variable, for each value of x. The smaller this is, the closer the data points are to the regression equation.

4.1 Practice questions

Take a look at the scatterplot of study time against number of courses taken (top of page 2). On this plot, label the following:

- estimated SLR equation
- the fitted value when $x = 3$, $x = 4$, and $x = 5$.
- the residual for any data point on the plot of your choosing.

5 Assessing Linear Association

As noted earlier, the variance of the error terms inform us how close the data points are to the estimated SLR equation. The smaller the variance of the error terms, the closer the data points are to the estimated SLR equation. This in turn implies the linear relationship between the variables is stronger.

We will learn about some common measures that are used to quantify the strength of the linear relationship between the response and predictor variables. Before we do that, we need to define some other terms.

5.1 Sum of squares

$$\text{Total Sum of Squares: } SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{11}$$

Total sum of squares is defined as the **total variance in the response variable**. The larger this value is, the larger the spread is of the response variable.

$$\text{Regression sum of squares: } SS_R = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2. \tag{12}$$

Regression sum of squares is defined as the **variance in the response variable that can be explained by our regression**.

We also have residual sum of squares, SS_{res}. Its mathematical formulation is given in (7). It is defined as the **variance in the response variable that cannot be explained by our regression**.

It can be shown that

$$SS_T = SS_R + SS_{res}. \tag{13}$$

Each of the sums of squares has its associated **degrees of freedom (df)**:

- df for SS_R: $df_R = 1$
- df for SS_{res}: $df_{res} = n - 2$
- df for SS_T: $df_T = n - 1$

5.2 ANOVA Table

Information regarding the sums of squares is usually presented in the form of an **ANOVA (analysis of variance) table**:

Source of Variation	SS	df	MS	F
Regression	$SS_R = \sum (\hat{y_i} - \bar{y})^2$	$df_R = 1$	$MS_R = \frac{SS_R}{df_R}$	$\frac{MS_R}{MS_{res}}$
Error	$SS_{res} = \sum (y_i - \hat{y_i})^2$	$df_{res} = n - 2$	$MS_{res} = \frac{SS_{res}}{df_{res}}$	***
Total	$SS_T = \sum (y_i - \bar{y})^2$	$df_T = n - 1$	***	***

Note:

- dividing each sum of square with its corresponding degrees of freedom gives the corresponding mean square.
- In the last column, we report an F statistic, which equal to $\frac{MS_R}{MS_{res}}$. This F statistic is associated with an **ANOVA F test**, which we will look at in more detail in the next subsection.

To obtain the ANOVA table for our study time example:

```
anova(result)
```

```
## Analysis of Variance Table
##
## Response: study
##             Df   Sum Sq  Mean Sq F value    Pr(>F)
## courses      1 57977993 57977993   65404 < 2.2e-16 ***
## Residuals 5998  5317017      886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that R does not print out the information for the line regarding SS_T.

5.3 ANOVA F Test

The ANOVA F statistic from the ANOVA table can be used to test if the slope of the SLR equation is 0 or not. In words, this means that whether there is a linear association between the variables or not. If the slope is 0, it means that changes in the value of the predictor variable do not change the value of the response variable, on average; hence the variables are not linearly associated.

The null and alternative hypotheses are:

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0.$$

The test statistic is

$$F = \frac{MS_R}{MS_{res}} \tag{14}$$

and is compared with an $F_{1,n-2}$ distribution. Going back to the study time example, the F statistic is 6.5403586×10^4. The critical value can be found using

```
qf(1-0.05, 1, 6000-2)
```

```
## [1] 3.84301
```

Since our test statistic is larger than the critical value, we reject the null hypothesis. Our data support the claim that the slope is different from 0, or in other words, that there is a linear association between study time and number of courses taken.

5.4 Coefficient of determination

The **coefficient of determination, R^2,** is

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}. \tag{15}$$

R^2 is an indication of how well the data fits our model. In the context of simple linear regression, it denotes **the proportion of variance in the response variable that is explained by the predictor**.

A few notes about R^2:

- $0 \leq R^2 \leq 1$.
- Values closer to 1 indicate a better fit; values closer to 0 indicate a poorer fit.
- Sometimes reported as a percentage.

To obtain R^2 for our study time example:

```
anova.tab<-anova(result)
##SST not provided, so we add up SSR and SSres
SST<-sum(anova.tab$"Sum Sq")
##R2
anova.tab$"Sum Sq"[1]/SST
```

```
## [1] 0.9159963
```

This implies that the proportion of variance in study time that can be explained by the number of courses taken is 0.9159963.

5.5 Correlation

A measure used to quantify the strength of the linear association between two quantitative variables is the **sample correlation**. The sample correlation, $\text{Corr}(x, y)$ or r, is given by

$$r = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}. \tag{16}$$

A few notes about the sample correlation r:

- $-1 \leq r \leq 1$.
- Sign of correlation indicates direction of association. A positive value indicates a positive linear association: as the predictor variable increases, so does the response variable, on average. A negative value indicates a negative linear association: as the predictor variable increases, the response variable decreases, on average.
- Values closer to 1 or -1 indicate a stronger linear association; values closer to 0 indicate a weaker linear association.
- In SLR, it turns out that $r^2 = R^2$.

Using our study time example, the correlation between study time and number of courses taken is

```
cor(df$study, df$courses)
```

```
## [1] 0.9570769
```

This value indicates a very strong and positive linear association between study time and number of courses taken (remember that this is simulated data and is not real).

5.5.1 How strong is strong?

A question that is often raised is how large should the magnitude of the sample correlation be for it to be considered strong? The answer is: it depends on the context. If you are conducting an experiment that is governed by scientific laws (e.g an experiment verifying Newton's 2nd law that $F = ma$), we should expect an extremely high correlation. A correlation of 0.9 in such an instance may be considered weak. The value of the correlation you have should be compared with correlations from similar studies in that domain to determine if it is strong or not.

6 A Word of Caution

To be able to use the measures we have learned (such as correlation, R^2) and to interpret the estimated regression coefficients, we must verify via a scatterplot that the association between the two variables is approximately linear. If we see a non linear pattern in the scatterplot, we should not use or interpret these values. We will learn how to remedy the situation if we see a non linear pattern in the scatterplot in module 5.

Please see the associated video for a demonstration on how not looking at the scatterplot can lead to misleading interpretations.