Linear Models for Data Science

Jeffrey Woo

2025-06-09

Contents

Preface

Who is this book for?

There are many books on linear models, with various expectations for different levels of familiarity with statistical, mathematical, and coding concepts. These books generally fall into one of two camps:

1. Little to no familiarity with statistical and mathematical concepts, but fairly familiar to coding. These books tend to be written for programmers who want to get into data science. These books tend to explain linear models while trying to avoid statistical and mathematical concepts as much as possible. These books tend to present linear models in a recipe format giving readers directions on what to do to build their models.

The drawback of such books is that readers do not get much understanding of the underlying concepts of linear models. It is impossible to give directions covering every possible scenario in the real world as real data are messy. Practitioners of data science often have to think outside the box in order to make linear models work for their particular data, and it is difficult to do so without understanding the mathematical framework of linear models.

2. Familiarity with mathematical notation and introductory statistical concepts such as statistical inference, and little to no familiarity with coding. These books tend to be written for mathematicians (or anyone with a strong background in mathematics) who want to get into data science. These books cover the mathematical framework of linear models thoroughly.

The drawback of such books is that readers must be comfortable with mathematical notation. This limits the audience for such books to people with fairly thorough training in mathematics. People without such training will get lost trying to read such books, and do not understand why we need to know the mathematical foundations to use linear models in data science.

This book is meant to be readable by both groups of readers. Some foundational mathematical knowledge will be presented, but will be written so that is readable by anyone. This book will also explain what these knowledge mean

in the context of data science. Practical advice, based on the foundational mathematical knowledge, will also be given.

This book accompanies the course STAT 6021: Linear Models for Data Science, for the Masters of Data Science (MSDS) program at the University of Virginia School of Data Science.

As introductory statistics and introductory programming are pre-requisites for entering the MSDS program, this book assumes basic knowledge of statistical inference and coding. Review materials covering these concepts are provided separately for enrolled students.

Chapters

The chapters for the book is as follows:

- Chapters **??** and **??** focus on core R skills needed: data wrangling and data visualization. These are R skills needed to perform regression analysis.
- Chapters 1, 2, and **??** cover simple linear regression (SLR). This is the simplest scenario in regression when we have one predictor and one response variable that is quantitative. We are using this scenario to be able to more clearly explain concepts in regression before moving into the more practical multiple linear regression (MLR), which involves multiple predictors.
- Chapters **??** to **??** cover multiple linear regression: when we have multiple predictors and one response variable that is quantitative.
- Chapters **??** and **??** cover logistic regression: when we have a response variable that is binary.

How to use this book

- If you are using the provided R code from each chapter, please remember to clear your R environment whenever you move to a new chapter. This can be done by typing `rm(list = ls())`.
- For Chapters 1 to **??**, there is an R tutorial provided in the last section. You should also clear your R environment before running the code in the tutorials.
- Some additional resources are provided for students enrolled in STAT 6021. These include:
  - Learning objectives.
  - Explainer videos.
  - Practice questions.
  - Assignments.

Data sets used

I have tried to use as many open source data sets as much as possible so that readers can work on the various examples I have provided on their own. However, some data sets may not be open source and were shared by other statistics and data science educators over my years of teaching this class (or variations of it) since 2013. It is my goal to eventually use only open source data sets.

Reporting issues with this book

This book is mostly a compilation of course notes that were originally written as separate chapters. While effort has been made to fix typos, issues may still exist. If you find any issues (typos, formatting, etc), please report them at https://github.com/jwooSDS/linear_models/issues. Please be as specific as you can, including providing the specific section and paragraph where the issue is found.

Other resources

Some other resources that readers may want to check out:

- *OpenIntro Statistics*, 4th ed. Diez, Cetinkaya-Rundel, Barr, OpenIntro. Get free PDF version at https://leanpub.com/os, just set the price that you want to pay to $0. This is a good book for introductory statistics.
- *Introduction to Probability for Data Science*, Chan. https://probabilit y4datascience.com/index.html. This book covers the fundamentals of probability and mathematics needed for data science. It does a good job explaining how seemingly abstract mathematical concepts are needed and applied in data science.
- *Linear Models with R*, 2nd ed. Faraway. This is probably one of the few books that balances between the two camps that I wrote about earlier. It does require familiarity with matrices and linear algebra though.
- *Introduction to Linear Regression Analysis*, 5th or 6th ed. Montgomery, Peck, Vining. You may be able to access an e-version of the book through your university library if you are affiliated with a university. This book is mathematically rigorous so is useful to those who are interested in mathematical proofs that is not covered.
- *Applied Linear Statistical Models* (ALSM), Kutner, Nachtsheim, Neter, Li, 5th ed. This book covers a wide range of topics in linear models and is also mathematically rigorous.
- *Applied Linear Regression Models* (ALRM), Kutner, Nachtsheim, Neter, 4th ed. ALRM is the same as the first 14 chapters of ALSM. The second

part of ALSM covers topics in Design of Experiments, which I highly recommend if you are interested in those topics.

Chapter 1

Basics with Simple Linear Regression (SLR)

1.1 Introduction HI AGAIN!!!!

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)")
```

```
abline(result)
```

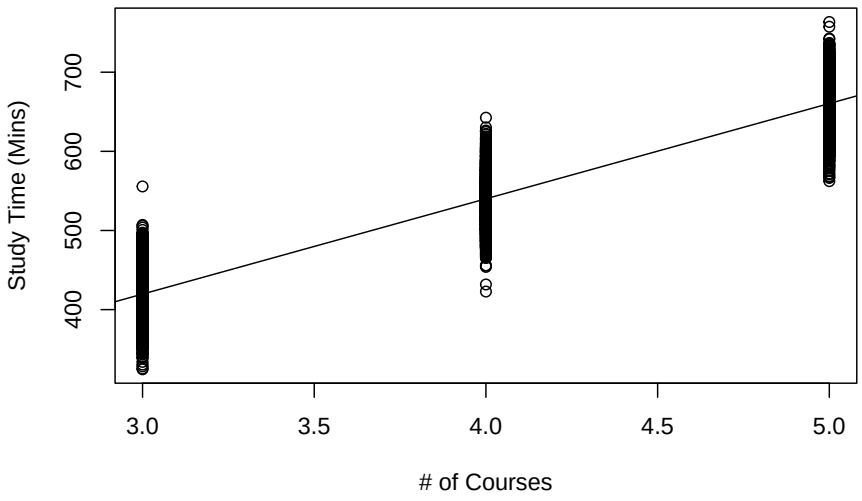


Figure 1.1: Scatterplot of Study Time against Number of Courses Taken

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.

1.1.1 Basic Ideas with Statistics

1.1.1.1 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.

1.1.1.2 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).

1.1.2 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.

1.1.2.1 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.

1.2 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**:

$$E(y|x) = \beta_0 + \beta_1 x. \tag{1.1}$$

- 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 (1.1) denotes the **expected value** of the response variable, for a fixed value of the predictor variable. What (1.1) 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 (1.1) 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**

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1.2}$$

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 (1.2), 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 **??**.

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:

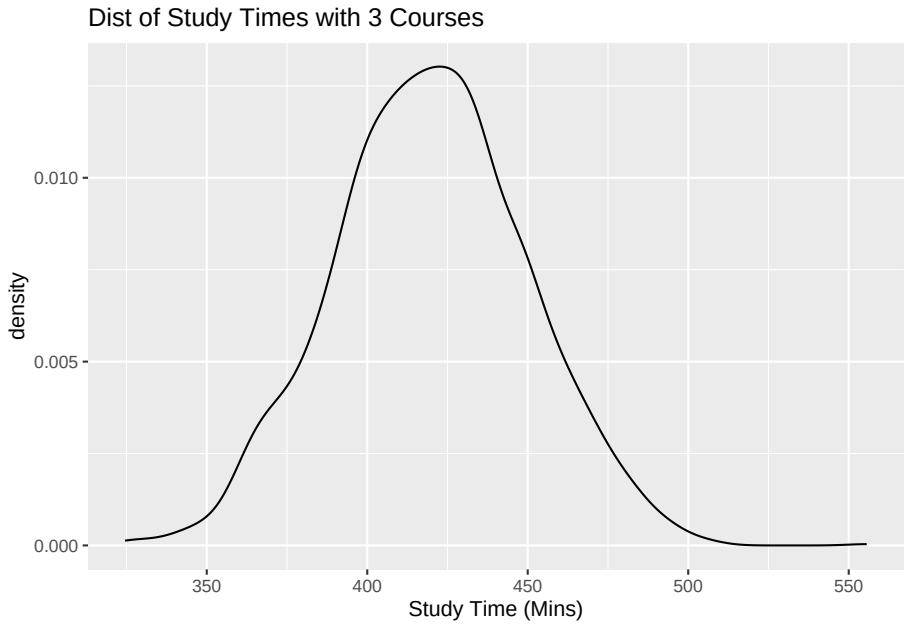
- for students who take 3 courses, the distribution of their study times is $N(\beta_0 + 3\beta_1, \sigma^2)$.
- for students who take 4 courses, the distribution of their study times is $N(\beta_0 + 4\beta_1, \sigma^2)$.
- for students who take 5 courses, the distribution of their study times is $N(\beta_0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

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.

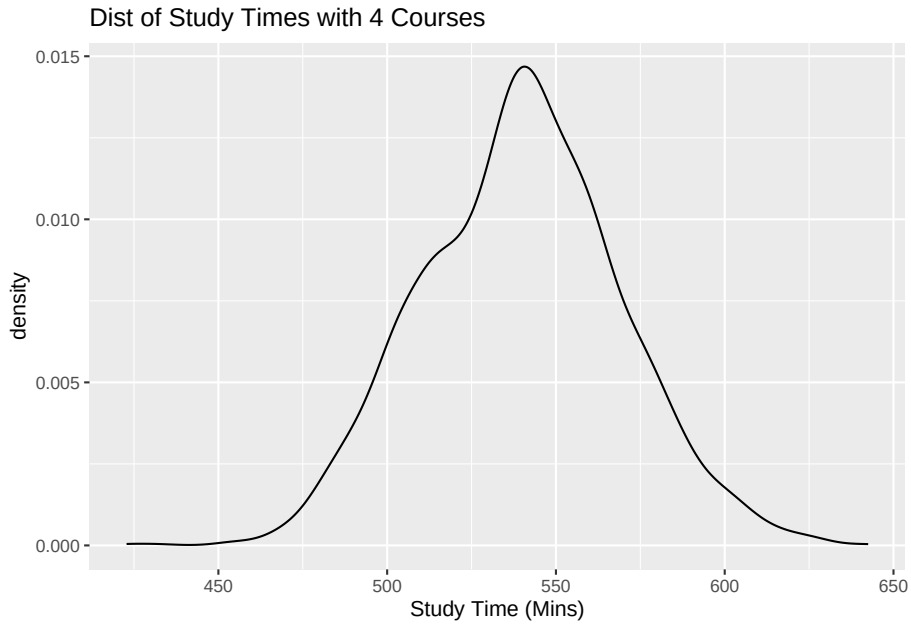
Let us take a look at these density plots next.

```
library(tidyverse)
```

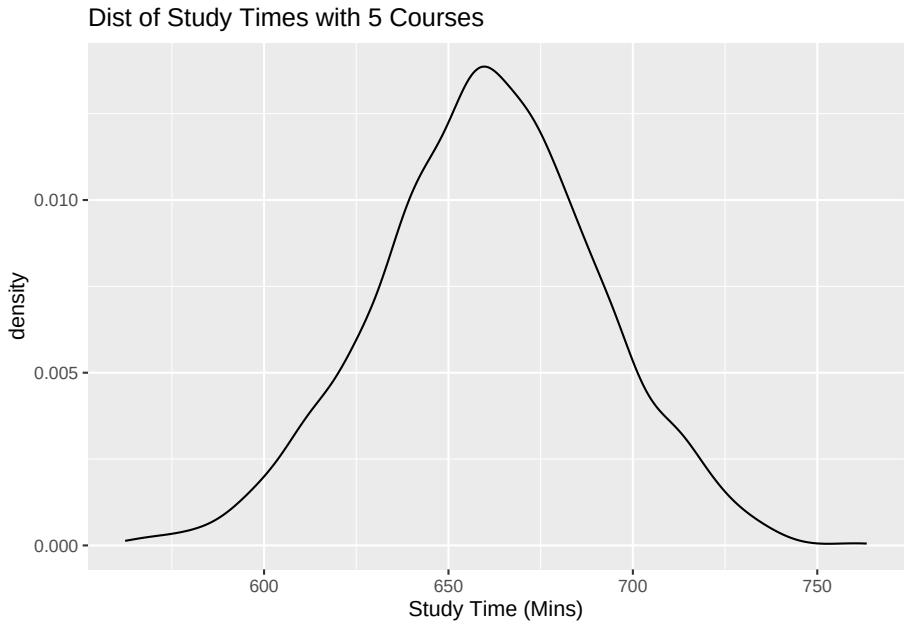
```
##subset dataframe
x.3<-df[which(df$courses==3),]
##density plot of study time for students taking 3 courses
ggplot(x.3,aes(x=study))+
  geom_density()+
  labs(x="Study Time (Mins)", title="Dist of Study Times with 3 Courses")
```




```
##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.

Please see the associated video for more explanation about the distribution of the response variable, for each value of the predictor variable, in an SLR setting.

1.3 Estimating Regression Coefficients in SLR

From (1.1) and (1.2), 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 (1.1) and (1.2), the sample versions are

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.3}$$

and

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + e \tag{1.4}$$

respectively. (1.3) is called the **estimated SLR equation**, or **fitted SLR equation**. (1.4) 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 (1.4), notice we use e **to denote the residual**, or in other words, the "error" in the sample.

From (1.3) and (1.4), we have the following quantities that we can compute:

$$\text{Predicted/Fitted values: } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{1.5}$$

$$\text{Residuals: } e_i = y_i - \hat{y}_i. \tag{1.6}$$

$$\text{Sum of Squared Residuals: } SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{1.7}$$

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 (1.7).

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1.8}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1.9}$$

$\hat{\beta}_1, \hat{\beta}_0$ 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 (1.7) 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:

```
##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.

1.4 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

$$s^2 = MS_{res} = \frac{SS_{res}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}, \tag{1.10}$$

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.

1.4.1 Practice questions

Take a look at the scatterplot of study time against number of courses taken, Figure 1.1. 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.

Try these on your own first, then view the associated video to see if you labeled the plot correctly!

1.5 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.

1.5.1 Sum of squares

$$\text{Total Sum of Squares: } SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{1.11}$$

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.

$$\text{Regression sum of squares: } SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{1.12}$$

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 (1.7). It is defined as the **variance in the response variable that cannot be explained by our regression**.

It can be shown that

$$SS_T = SS_R + SS_{res}. \tag{1.13}$$

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: $df_T = n - 1$

Please see the associated video for more explanation about the concept behind degrees of freedom.

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.

1.5.3 ANOVA F Test

In SLR, 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

$$F = \frac{MS_R}{MS_{res}} \tag{1.14}$$

and is compared with an $F_{1,n-2}$ distribution. Note that $F_{1,n-2}$ is read as an **F distribution with 1 and $n-2$ degrees of freedom**.

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.

1.5.4 Coefficient of determination

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

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}. \tag{1.15}$$

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.

1.5.5 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

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}. \tag{1.16}$$

A few notes about 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).

1.5.5.1 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.

1.6 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.

1.7 R Tutorial

For this tutorial, we will work with the dataset `elmhurst` from the `openintro` package in R.

```
library(tidyverse)
library(openintro)
Data<-openintro::elmhurst
```

Type `?openintro::elmhurst` to read the documentation for datasets in R. Always seek to understand the background of your data! The key pieces of information are:

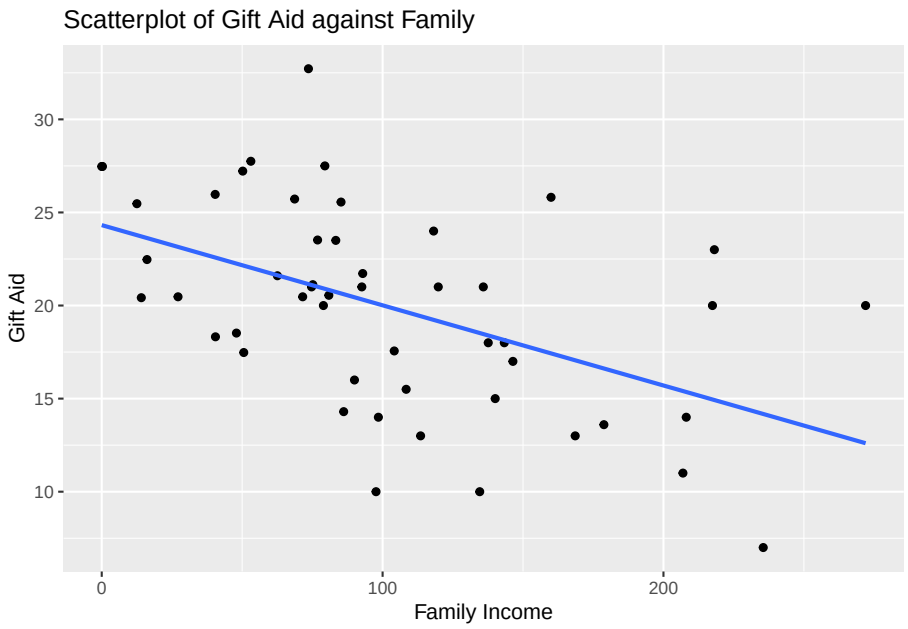
- A random sample of 50 students (all freshman from the 2011 class at Elmhurst College).
- Family income of the student (units are missing).
- Gift aid, in $1000s.

We want to explore how family income may be related to gift aid, in a simple linear regression framework.

Visualization

We should always verify with scatterplot that the relationship is (approximately) linear before proceeding with correlation and simple linear regression!

```
##scatterplot of gift aid against family income
ggplot2::ggplot(Data, aes(x=family_income,y=gift_aid))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Family Income", y="Gift Aid", title="Scatterplot of Gift Aid against Family")
```



We note that the observations are fairly evenly scattered on both sides of the regression line, so a linear association exists. We see a negative linear association. As family income increases, the gift aid, on average, decreases.

We also do not see any observation with weird values that may warrant further investigation.

Regression

We use the `lm()` function to fit a regression model:

```
##regress gift aid against family income
result<-lm(gift_aid~family_income, data=Data)
```

Use the `summary()` function to display relevant information from this regression:

```
##look at information regarding regression
summary(result)
```

```
##
## Call:
## lm(formula = gift_aid ~ family_income, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1128  -3.6234  -0.2161   3.1587  11.5707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.31933    1.29145  18.831  < 2e-16 ***
## family_income -0.04307    0.01081  -3.985 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.783 on 48 degrees of freedom
## Multiple R-squared:  0.2486, Adjusted R-squared:  0.2329
## F-statistic: 15.88 on 1 and 48 DF,  p-value: 0.0002289
```

We see the following values:

- $\hat{\beta}_1 =$ -0.0430717. The estimated slope informs us the the predicted gift aid decreases by 0.0430717 thousands of dollars (or $43.07) per unit increase in family income.
- $\hat{\beta}_0 = 24.319329$. For students with no family income, their predicted gift aid is $24 319.33. Note: from the scatterplot, we have an observation with 0 family income. We must be careful in not extrapolating when making predictions with our regression. We should only make predictions for family incomes between the minimum and maximum values of family incomes in our data.
- $s = 4.7825989$, is the estimate of the standard deviation of the error terms. This is reported as residual standard error in R. Squaring this gives the estimated variance.

- $F = 15.8772043$. This is the value of the ANOVA F statistic. The corresponding p-value is reported. Since this p-value is very small, we reject the null hypothesis. The data support the claim that there is a linear association between gift aid and family income.
- $R^2 = 0.2485582$. The coefficient of determination informs us that about 24.86% of the variation in gift aid can be explained by family income.

Extract values from R objects

We can actually extract these values that are being reported from `summary(result)`. To see what can be extracted from an R object, use the `names()` function:

```
##see what can be extracted from summary(result)
names(summary(result))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

To extract the estimated coefficients:

```
##extract coefficients
summary(result)$coefficients
```

```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   24.31932901 1.29145027 18.831022 8.281020e-24
## family_income -0.04307165 0.01080947 -3.984621 2.288734e-04
```

Notice the information is presented in a table. To extract a specific value, we can specify the row and column indices:

```
##extract slope
summary(result)$coefficients[2,1]
```

```
## [1] -0.04307165
```

```
##extract intercept
summary(result)$coefficients[1,1]
```

```
## [1] 24.31933
```

On your own, extract the values of the residual standard error, the ANOVA F statistic, and R^2.

Prediction

A use of regression models is prediction. Suppose I want to predict the gift aid of a student with family income of 50 thousand dollars (assuming the unit is in thousands of dollars). We use the `predict()` function:

```
##create data point for prediction
newdata<-data.frame(family_income=50)
##predicted gift aid when x=50
predict(result,newdata)
```

```
##        1
## 22.16575
```

This student's predicted gift aid is \$22 165.75. Alternatively, you could have calculated this by plugging $x = 50$ into the estimated SLR equation:

```
summary(result)$coefficients[1,1] + summary(result)$coefficients[2,1]*50
```

```
## [1] 22.16575
```

ANOVA table

We use the `anova()` function to display the ANOVA table

```
anova.tab<-anova(result)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: gift_aid
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## family_income  1  363.16  363.16  15.877 0.0002289 ***
## Residuals     48 1097.92   22.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The report F statistic is the same as the value reported earlier from `summary(result)`.

The first line of the output gives SS_R, the second line gives SS_{res}. The function doesn't provide SS_T, but we know that $SS_T = SS_R + SS_{res}$.

Again, to see what can be extracted from `anova.tab`:

```
names(anova.tab)
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

So SS_T can be easily calculated:

```
SST<-sum(anova.tab$"Sum Sq")
SST
```

```
## [1] 1461.079
```

The R^2 was reported to be 0.2485582. To verify using the ANOVA table:

```
anova.tab$"Sum Sq"[1]/SST
```

```
## [1] 0.2485582
```

Correlation

We use the `cor()` function to find the correlation between two quantitative variables:

```
##correlation
cor(Data$family_income,Data$gift_aid)
```

```
## [1] -0.4985561
```

The correlation is -0.4985561. We have a moderate, negative linear association between family income and gift aid.

Chapter 2

Inference with Simple Linear Regression (SLR)

2.1 Introduction

Oftentimes, the data we collect come from a random sample that is representative of the population of interest. A common example is an election poll before a presidential election. Random sampling allows the sample to be representative of the population. However, if we obtain another random sample, the characteristics of the new sample are unlikely to be exactly the same as the first sample. For example, the sample proportion who will vote for a certain party is unlikely to be the same for both random samples. What this tells us is that even with representative samples, sample proportions are unlikely to be equal to the population proportion, and sample proportions vary from sample to sample.

Dr. W. Edwards Deming's Red Bead experiment illustrates this concept. A video of this experiment can be found here.

In this video, the number of red beads, which represent bad products, varies each time the worker obtains a random sample of 50 beads. The fact that the number of red beads increases in his second sample does not indicate that he performed his task any worse, as this increase is due to the random variation associated with samples.

Note: Deming's Red Bead experiment was developed to illustrate concepts associated with management. He is best known for his work in developing the Japanese economy after World War II. You will be able to find many blogs/articles discussing the experiment on the World Wide Web. Although many of the articles discuss how this experiment applies in management, it can be used to illustrate concepts of variation.

The same idea extends to the slope and intercept of a regression line. The estimated slope and intercept will vary from sample to sample and are unlikely to be equal to the population slope and intercept. In inferential statistics, we use hypothesis tests and confidence intervals to aid us in accounting for this random variation. In this module, you will learn how to account for and quantify the random variation associated with the estimated regression model, and how to interpret the estimated regression model while accounting for random variation.

2.1.1 Review from previous module

The **simple linear regression model** is written as

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{2.1}$$

We make some assumptions for the error term ϵ. They 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.

These assumptions allow us to derive the distributional properties associated with our least squares estimators $\hat{\beta}_0, \hat{\beta}_1$, which then enables us to compute reliable confidence intervals and perform hypothesis tests on our SLR reliably.

$\hat{\beta}_1, \hat{\beta}_0$ 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** $x = 0$**.**

How do the values of these estimators vary from sample to sample?

2.2 Hypothesis Testing in SLR

2.2.1 Distribution of least squares estimators

Gauss Markov Theorem: Under assumptions for a regression model, the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased and have minimum variance among all unbiased linear estimators.

Thus, the least squares estimators have the following properties:

1. $\mathrm{E}(\hat{\beta}_1) = \beta_1$, $\mathrm{E}(\hat{\beta}_0) = \beta_0$

Note: An estimator is **unbiased** if its expected value is exactly equal to the parameter it is estimating.

2. The variance of $\hat{\beta}_1$ is

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2.2}$$

3. The variance of $\hat{\beta}_0$ is

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \tag{2.3}$$

4. $\hat{\beta}_1$ and $\hat{\beta}_0$ both follow a normal distribution.

Note that in (2.2) and (2.3), we use $s^2 = MS_{res}$ to estimate σ^2 since σ^2 is a unknown value.

What these imply is that if we standardize $\hat{\beta}_1$ and $\hat{\beta}_0$, these standardized quantities will follow a t_{n-2} distribution, i.e.

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2} \tag{2.4}$$

and

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t_{n-2}, \tag{2.5}$$

where

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{\sum (x_i - \bar{x})^2}} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{2.6}$$

and

$$se(\hat{\beta}_0) = \sqrt{MS_{res} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \tag{2.7}$$

Note:

- $se(\hat{\beta}_1)$ is read as the **standard error of** $\hat{\beta}_1$. The standard error of any estimator is essentially the sample standard deviation of that estimator, and measures the spread of that estimator.
- A t_{n-2} distribution is read as a t **distribution with** $n-2$ **degrees of freedom**.

2.2.2 Testing regression coefficients

Hypothesis testing is used to investigate if a population parameter is **different from a specific value**. In the context of SLR, we usually want to test if β_1 is 0 or not. If $\beta_1 = 0$, there is no linear relationship between the variables.

The general steps in hypothesis testing are:

- Step 1: State the null and alternative hypotheses.
- Step 2: A test statistic is calculated using the sample, assuming the null is true. The value of the test statistic measures how the **sample deviates from the null**.
- Step 3: Make conclusion, using either critical values or p-values.

In the previous module, we introduced the ANOVA F test. In SLR, this tests if the slope of the SLR equation is 0 or not. It turns out that we can also perform a t test for the slope. In the t test for the slope, the null and alternative hypotheses are:

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

The test statistic is

$$t = \frac{\hat{\beta}_1 - \text{ value in null}}{se(\hat{\beta}_1)} \tag{2.8}$$

which is compared with a t_{n-2} distribution. Notice that (2.8) comes from (2.4).

Let us go back to our simulated example that we saw in the last module. 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)
##look at regression coefficients
summary(result)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  58.44829  1.9218752  30.41211 4.652442e-189
## courses     120.39310  0.4707614 255.74125  0.000000e+00
```

The t statistic for testing $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$ is reported to be 255.7412482, which can be calculated using (2.8): $t = \frac{120.39310-0}{0.4707614}$. The reported p-value is virtually 0, so we reject the null hypothesis. The data support the claim that there is a linear association between study time and the number of courses taken.

2.3 Confidence Intervals for Regression Coefficients

Confidence intervals (CIs) are similar to hypothesis testing in the sense that they are also based on the distributional properties of an estimator. CIs may differ in their use in the following ways:

1. We are not assessing if the parameter is different from a specific value.
2. We are more interested in exploring a plausible **range of values for an unknown parameter**.

Because CIs and hypothesis tests are based on the distributional properties of an estimator, their conclusions will be consistent (as long as the significance level is the same).

Recall the general form for CIs:

$$\text{estimator} \pm (\text{multiplier} \times \text{s.e of estimator}). \tag{2.9}$$

We have the following components of a CI

- **estimator (or statistic)**: numerical quantity that describes a sample
- **multiplier**: determined by confidence level and relevant probability distribution
- **standard error of estimator**: measure of variance of estimator (basically the square root of the variance of estimator)

Following (2.9) and (2.4), the $100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 \pm t_{1-\alpha/2;n-2} se(\hat{\beta}_1) = \hat{\beta}_1 \pm t_{1-\alpha/2;n-2} \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}. \tag{2.10}$$

Going back to our study time example, the 95% CI for β_1 is (119.470237, 121.3159601).

```
##CI for coefficients
confint(result,level = 0.95)[2,]
```

```
##    2.5 %   97.5 %
## 119.4702 121.3160
```

An interpretation of this CI is that we have 95% confidence that the true slope β_1 lies between (119.470237, 121.3159601). In other words, for each additional course taken, the predicted study time increases between 119.470237 and 121.3159601 minutes.

2.3.1 Thought questions

- Is the conclusion from this 95% CI consistent with the hypothesis test for $H_0 : \beta_1 = 0$ in the previous section at 0.05 significance level?
- I have presented hypothesis tests and CIs for the slope, β_1.
  - How would you calculate the t statistic if you wanted to test $H_0 : \beta_0 = 0, H_0 : \beta_0 \neq 0$?
  - How would you calculate the 95% CI for the intercept β_0?

Generally, we are usually more interested in the slope than the intercept.

2.4 CI of the Mean Response

We have established that the least squares estimators $\hat{\beta}_1, \hat{\beta}_0$ have their associated variances. Since the estimated SLR equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2.11}$$

it stands to reason that $\hat{y}$ has an associated variance as well, since it is a function of $\hat{\beta}_1, \hat{\beta}_0$.

There are two interpretations of $\hat{y}$:

1. it **estimates the mean of y when $x = x_0$**;
2. it **predicts the value of y for a new observation when $x = x_0$**.

Note: x_0 denotes a specific numerical value for the predictor variable.

Depending on which interpretation we want, there are two different intervals based on $\hat{y}$. The first interpretation is associated with the **confidence interval for the mean response, $\hat{\mu}_{y|x_0}$, given the predictor**. This is used when we are interested in the average value of the response variable, when the predictor is equal to a specific value. This CI is

$$\hat{\mu}_{y|x_0} \pm t_{1-\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{2.12}$$

Going back to our study time example, suppose we want the average study time for students who take 5 courses, the 95% CI is

```
##CI for mean y when x=5
newdata<-data.frame(courses=5)
predict(result, newdata, level=0.95, interval="confidence")
```

```
##        fit      lwr      upr
## 1 660.4138 659.2224 661.6052
```

We have 95% confidence that the average study time for students who take 5 courses is between 659.2223688 and 661.605187 minutes.

2.5 PI of a New Response

Previously, we found a CI for the mean of y given a specific value of x, (2.12). This CI gives us an idea about the location of the regression line at a specific of x.

Instead, we may have interest in finding an interval for a new value of $\hat{y}_0$, when we have a new observation $x = x_0$. This is called a **prediction interval (PI) for a future observation y_0 when the predictor is a specific value**. This interval follows from the second interpretation of $\hat{y}$.

The PI for $\hat{y}_0$ takes into account:

1. Variation in location for the distribution of y (i.e. where is the center of the distribution of y?).
2. Variation **within the probability distribution of** y.

By comparison, the confidence interval for the mean response (2.12) only takes into account the first element. The PI is

$$\hat{y}_0 \pm t_{1-\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{2.13}$$

Going back to our study time example, suppose we have a newly enrolled student who wishes to take 5 courses, and the student wants to predict his study time

```
##PI for y when x=5
predict(result, newdata, level=0.95, interval="prediction")
```

```
##        fit      lwr      upr
## 1 660.4138 602.0347 718.7928
```

We have 95% confidence that the study time for this student is between 602.0347305 and 718.7928253 minutes.

2.5.1 Thought questions

- In the following two scenarios, decide if we are more interested in the CI for the mean response given the predictor (2.12), or the PI for a future response given the predictor (2.13).
  - We wish to estimate the waiting time, on average, of DMV customers if there are 10 people in line at the DMV.
  - I enter the DMV and notice 10 people in line. I want to estimate my waiting time.

- Look at the standard errors associated with the intervals given in (2.12) and (2.13). How are they related to each other?

2.6 Supplemental Notes on Statistical Inference

2.6.1 Hypothesis statements

Let's consider a t test for the regression parameter, β_1. Depending on context, the following could be null and alternative hypotheses

- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- $H_0 : \beta_1 = 0, H_a : \beta_1 > 0$.
- $H_0 : \beta_1 = 0, H_a : \beta_1 < 0$.

The null hypothesis should be stated as a statement of **equality**. This concept holds true for hypothesis tests in general. Some other books / resources might state them as

- $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$.
- $H_0 : \beta_1 \leq 0, H_a : \beta_1 > 0$.
- $H_0 : \beta_1 \geq 0, H_a : \beta_1 < 0$.

I prefer using the equality statement for the null hypothesis for the following reasons (theoretical, pedagogical, practical):

1. The null hypothesis being an equality aligns with the definition of the p-value.

   - The p-value is the probability of observing our sample estimate (or a value more extreme), if the null hypothesis is true (i.e. β_1 is truly 0). This is what we are assuming in the calculation for the test statistic.

2. People tend to get confused between the null and alternative hypotheses if both involve inequalities (the alternative is the hypothesis you are trying to support).
3. Conclusions are made in terms of supporting (or not supporting) the alternative hypothesis.

2.6.2 Sample size and statistical inference

Generally speaking, there is a relationship between sample size and statistical inference (assuming other characteristics remain the same and our sample was randomly obtained or representative of the population of interest):

- Larger sample sizes (typically) lead to narrower confidence intervals (more precise intervals).
- Sample estimates based on larger samples are more likely to be closer to the true parameters.

- Larger sample (typically) lead to more evidence against the null hypothesis.
  - This means a larger sample size leads to a more powerful test. The power of a test is the probability a hypothesis test is able to correctly reject the null hypothesis.

2.6.2.1 Small sample sizes

Small sample sizes tend to result in:

- Confidence intervals that are wide.
- Sample estimates that are more likely to be further away from the true parameters.
- Hypothesis tests that are more likely to incorrectly fail to reject the null hypothesis when the alternative hypothesis is true.

While larger sample sizes have their advantages, there are also some disadvantages with sample sizes that are extremely large.

2.6.2.2 Large sample sizes

A "statistically significant" result does not necessarily mean that the result has practical consequences. Suppose a 95% confidence interval for β_1 is $(0.001, 0.002)$. The interval excludes 0, so it is "statistically significantly" different from 0 (because it is!), but does this result have practical consequences? A narrow CI that barely excludes the null value can happen when we have a large sample size.

If one was to conduct the corresponding hypothesis test, we would reject the null hypothesis that $\beta_1 = 0$. With large sample sizes, hypothesis tests are sensitive to small departures from the null hypothesis.

In such instances, it may be worth considering hypothesis tests involving a different value in the null hypothesis, one that makes sense for your question. For example, a practically significant slope may need to be greater than a specific numerical value for a certain context.

- Statistical inference to assess statistical significance.
- Subject area knowledge to assess practical significance.

2.6.2.3 Questions

Are the following results statistically significant? If so, are the results also practically significant? Assume a two-sided test with a null value of 0 (These are made up examples):

1. In assessing if studying more is associated with better test scores, a SLR is carried out with test scores (out of 100 points) against study time (in hours). The 95% confidence interval for the slope β_1 is (5.632, 7.829).

2. A SLR is carried out to explore the linear relationship between number of years in school with income (in thousands of dollars). The 95% confidence interval for the slope β_1 is (0.051, 0.243).

2.6.3 Cautions using SLR and Correlation

Simple linear regression and correlation are meant for assessing **linear** relationships. If the relationship is not linear, we will need to transform the variable(s) (so the transformed variables have a linear relationship. Will explore this in Module **??**).

- Always verify via a scatterplot that the relationship is at least approximately linear.
- A high correlation or a significant estimated slope by themselves do not prove that we have a strong linear relationship between the variables. Conversely, a correlation close to 0 or an insignificant estimated slope is also not proof that we do not have a relationship between the variables.

2.6.3.1 Outliers

SLR and correlation are sensitive to outliers / influential observations. Generally speaking, these are data that are "far away" or very different from the rest of the observations. These data points can be visually inspected from a scatterplot. Some potential considerations when dealing with such data points:

- Investigate these observations. There is usually something that is making them "stand out" from the rest of the data.
- Data entry errors that can be corrected. Be sure to mention in the report.
- Revisit how the data were sampled. Perhaps the data point is is not part of the population of interest. If so, data point can be removed (this is legitimate), but be sure to mention in the report.

With regards to regression analysis:

- Exclusion of data points must be clearly documented.
- Fit the regression with and without the data points in question, and see how similar or different the conclusions become.
- If the data points have large value(s) on the predictor and/or response, a log transformation on the variable can pull in the large values.
- Consider subsetting your data and create separate models for each subset; or focus on a subset and make it clear your analysis is for a subset.
- Knowing your data and context can help a lot in these decisions.

2.6.3.2 Association and causation

Two correlated variables do not mean that one variable causes the other variable to change. For example, consider a plot of ice cream consumption and deaths by drowning during various months. There may be some positive correlation, and

clearly, eating more ice cream does not cause more drownings. The correlation can be explained by a third (lurking) variable: the weather.

A **lurking variable** is a variable that has an impact on the relationship between the variables being studied, but is itself not studied.

A carefully designed **randomized experiment** can control for lurking variables, and causal relationships can be established. Typically, such experiments include:

- A control group and a treatment group.
- Random assignment of large number of observations into the treatment and control groups. Due to the random assignment, the general characteristics of of subjects in each group are similar.

Lurking variables are always an issue with **observational studies**. Researchers in observational studies do not intervene with the observations and simply observe the data that the observations generate. Causal relationships are much more difficult to establish with observational studies.

2.6.3.3 Questions

1. Consider the `palmerpenguins` dataset that we have been working on. The data contain size measurements for three different species of penguins on three islands in the Palmer Archipelago, Antarctica over three years. Is this an observational study or randomized experiment?
2. A fertilizer company wishes to evaluate how effective a new fertilizer is in terms of improving the yield of crops. A large field is divided into many smaller plots, and each smaller plot is randomly assigned to receive either the new fertilizer or the standard fertilizer. Is this an observational study or randomized experiment?
3. A professor wishes to evaluate the effectiveness of various teaching methods (traditional vs flipped classroom). The professor uses the traditional approach for a section that meets on Mondays, Wednesdays, and Fridays from 9 to 10am and uses the flipped classroom approach for a section that meets on Mondays, Wednesdays, and Fridays from 2 to 3pm. Students were free to choose whichever section that wanted to register for, with no knowledge of the teaching method being used. What are some potential lurking variables in this study?

2.7 R Tutorial

For this tutorial, we will continue to work with the dataset `elmhurst` from the `openintro` package in R.

```
library(tidyverse)
library(openintro)
Data<-openintro::elmhurst
```

The key pieces of information are:

- A random sample of 50 students (all freshman from the 2011 class at Elmhurst College).
- Family income of the student (units are missing).
- Gift aid, in $1000s.

We want to explore how family income may be related to gift aid, in a simple linear regression framework.

Hypothesis test for β_1 (and β_0)

Applying the `summary()` function to `lm()` gives the results of hypothesis tests for β_1 and β_0:

```
##Fit a regression model
result<-lm(gift_aid~family_income, data=Data)

##look at t stats and F stat
summary(result)
```

```
##
## Call:
## lm(formula = gift_aid ~ family_income, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1128  -3.6234  -0.2161   3.1587  11.5707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.31933    1.29145  18.831  < 2e-16 ***
## family_income -0.04307    0.01081  -3.985 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.783 on 48 degrees of freedom
## Multiple R-squared:  0.2486, Adjusted R-squared:  0.2329
## F-statistic: 15.88 on 1 and 48 DF,  p-value: 0.0002289
```

Under coefficients, we can see the results of the hypothesis tests for β_1 and β_0. Specifically, for β_1:

- $\hat{\beta}_1$ = -0.0430717
- $se(\hat{\beta}_1)$ = 0.0108095
- the test statistic is t = -3.984621
- the corresponding p-value is 2.2887345×10^{-4}

You can work out the p-value using R (slight difference due to rounding):

```
##pvalue
2*pt(-abs(-3.985), df = 50-2)
```

```
## [1] 0.0002285996
```

Or find the critical value using R:

```
##critical value
qt(1-0.05/2, df = 50-2)
```

```
## [1] 2.010635
```

Either way, we end up rejecting the null hypothesis. The data support the claim that there is a linear association between gift aid and family income.

Note:

- the t tests for regression coefficients are based on $H_0 : \beta_j = 0, H_a : \beta_j \neq 0$. The reported p-value is based on this set of null and alternative hypotheses. If your null and alternative hypotheses are different, you will need to compute your own test statistic and p-value.
- For SLR, the two-sided t test for β_1 gives the exact same result as the ANOVA F test. Notice the p-values are the same. The F statistic of 15.88 is the squared of the t statistic, $(-3.985)^2$.

Confidence interval for β_1 (and β_0)

To find the 95% confidence intervals for the coefficients, we use the `confint()` function:

```
##to produce 95% CIs for all regression coefficients
confint(result,level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept)   21.72269421 26.91596380
## family_income -0.06480555 -0.02133775
```

The 95% CI for β_1 is (-0.0648056, -0.0213378). We have 95% confidence that for each additional thousand dollars in family income, the predicted gift aid decreases between \$21.3378 and \$64.8056.

Confidence interval for mean response for given x

Suppose we want a confidence interval for the average gift aid for Elmhurst College students with family income of 80 thousand dollars. We can use the `predict()` function:

```
##to produce 95% CI for the mean response when x=80,
newdata<-data.frame(family_income=80)
predict(result,newdata,level=0.95, interval="confidence")
```

```
##       fit      lwr      upr
## 1 20.8736 19.43366 22.31353
```

The 95% CI for the mean gift aid for students with family income of 80 thousand dollars is (19.4336609, 22.3135327). We have 95% confidence the mean gift aid for students with family income of 80 thousand dollars is between $19 433.66 and $22 313.53.

Prediction interval for a response for a given x

For a prediction interval for the gift aid of an Elmhurst College student with family income of 80 thousand dollars:

```
##and the 95% PI for the response of an observation when x=80
predict(result,newdata,level=0.95, interval="prediction")
```

```
##       fit      lwr      upr
## 1 20.8736 11.15032 30.59687
```

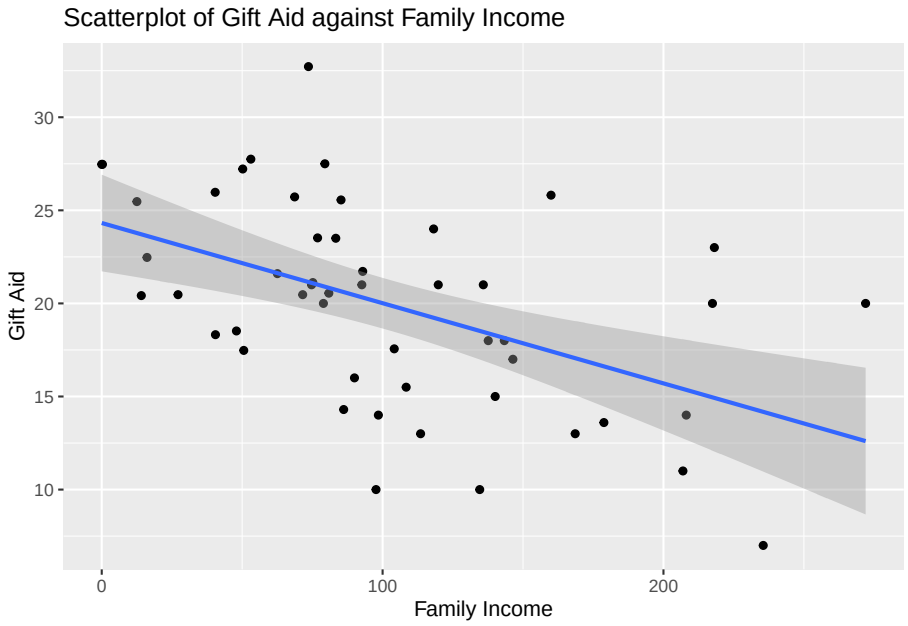
We have 95% confidence that for an Elmhurst College student with family income of 80, this student's gift aid is between $11 150.32 and $30 596.87.

Visualization of CI for mean response given x and PI of response given x

When using the `ggplot()` function to create a scatterplot, we can overlay the SLR equation by adding a layer via `geom_smooth(method = lm)`. By default, the CI for the mean response for each value of the predictor gets overlaid as well. In the previous tutorial, we removed this by adding `se=FALSE` inside `geom_smooth()`:

```
##regular scatterplot
##with regression line overlaid, and bounds of CI for mean y
ggplot2::ggplot(Data, aes(x=family_income, y=gift_aid))+
  geom_point() +
  geom_smooth(method=lm)+
  labs(x="Family Income",
       y="Gift Aid",
       title="Scatterplot of Gift Aid against Family Income")
```

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



Overlaying prediction intervals require a bit more work. We need to compute the lower and upper bounds of the PI for each value of the predictor:

```
##find PIs for each observation
preds <- predict(result, interval="prediction")
```

```
## Warning in predict.lm(result, interval = "prediction"): predictions on current data refer to _
```

Previously, when we used the `predict()` function, we provided the numerical value of x to make a prediction on. If this is not supplied, the function will use all the current values of x to make predictions, and will actually print out a warning message. For our purpose, this is not an issue since this is exactly what we want.

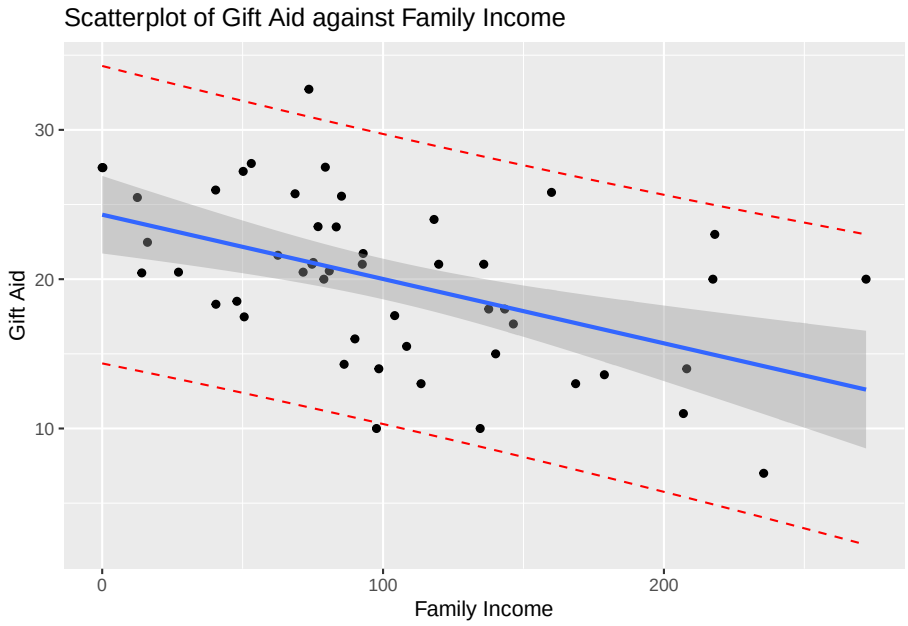
We then add `preds` to the data frame in order to overlay the lower and upper bounds on the scatterplot, by adding extra layers via `geom_line()` in the `ggplot()` function:

```
##add preds to data frame
Data<-data.frame(Data,preds)

##overlay PIs via geom_line()
ggplot2::ggplot(Data, aes(x=family_income, y=gift_aid))+
  geom_point() +
  geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y=upr), color = "red", linetype = "dashed")+
  geom_smooth(method=lm)+
```

```
  labs(x="Family Income",
       y="Gift Aid",
       title="Scatterplot of Gift Aid against Family Income")
```

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



As mentioned in the notes, the CI captures the location of the regression line, whereas the PI captures the data points.