A constant linear model is a type of model that provides us with tools for drawing statistical inferences about means of random variables. Means of random variables are theoretical constructs, so we can never observe them directly but must instead use such tools to gain information about them.

In this chapter, we define constant linear models and explore how to use them to address real-world questions. The sections of the chapter follow the order of the steps in the process that we use to address such questions, a process that we call a constant linear model analysis.

**Constant linear model analysis**

The steps in a constant linear model analysis are:

1. Formulate the questions of interest.
2. Explore the data.
3. Fit the model.
4. Check the Sampling Variability Assumptions.
5. Draw statistical inferences.
6. Interpret the results.

In each section of this chapter, we begin by describing a step in the linear constant model analysis process, discussing aspects that apply to both constant linear models and the linear models defined in later chapters. Later in each section, we focus specifically on constant linear models, the main subject of the chapter.
We use the process given here with only minor modifications for other types of (nonconstant) linear models introduced throughout the book. In this broader context, we refer to the process as a linear model analysis.

Questions of interest

The first step in a linear model analysis is to formulate the questions that we plan to address with the analysis, which leads us to the following definition.

**Definition: Questions of Interest.** A non-statistical question of interest is a real-world question that we plan to address with a linear model analysis. It must be specific enough that we can translate it into a statistical question of interest, which is a question about a linear model that will address the corresponding non-statistical question of interest. Any variable included in a statistical question of interest is called a variable of interest.

These definitions are not technical, statistical definitions but just convenient terms. Note that a non-statistical question of interest may not be answered in full by the linear model analysis, but a statistical question of interest should be. We will learn more about how this works as we go.

With the above terminology at hand, we can now state the first step in a linear model analysis more precisely: Formulate the non-statistical questions of interest. This will generate corresponding statistical questions of interest, which will also determine the variables of interest.

Common questions of interest

A constant linear model analysis involves only a single variable of interest, a random variable often denoted generically by $Y$. For such an analysis, the statistical questions of interest should be questions about the mean of that random variable. Such questions commonly arise as follows:

- We have a random variable $Y$ that is the difference $Y_2 - Y_1$ of two other random variables. In order to determine whether the means of $Y_1$ and $Y_2$ are equal, we ask the question: Is the mean of $Y$ equal to 0?

- Again we have $Y = Y_2 - Y_1$, but now we simply ask: What is the
mean of $Y$? The answer to this question will tell us how different the means of $Y_1$ and $Y_2$ are.

- We have a random variable $Y$ (that is not necessarily a difference of random variables), and we choose a particular value $b_0$ and ask: Is the mean of $Y$ equal to $b_0$?

- We have a random variable $Y$ and we ask: What is the mean of $Y$?

All of these questions are about random variable means. Since we cannot observe random variable means directly, we use tools provided by constant linear models to help us answer these questions.

Non-statistical questions of interest should translate into statistical questions of interest, so common non-statistical questions for constant linear model analyses are similar to the statistical questions in the above list, except that $Y$ is replaced by any quantity that can be modeled by a random variable. Also, for non-statistical questions, we often use the non-technical term average instead of the statistical term mean. These two terms do not have identical meanings, but average is often (although not always) translated into the technical term mean.

Avoiding causation

Formulating proper questions of interest at the outset of a linear model analysis is important. Without this step, we may waste a considerable amount of time exploring irrelevant aspects of a linear model, or (worse yet!) developing an unsuitable model.

In formulating questions of interest, we should avoid involving causation. Linear model analyses do not establish causation, so non-statistical questions involving causation cannot be translated into statistical questions of interest. Consequently, non-statistical questions involving causation cannot be used as non-statistical questions of interest. For example, we cannot use either of the following questions as a non-statistical question of interest:

- Does increasing the amount of sunlight a particular type of plant receives cause the plant to grow taller?

- Does a particular SAT preparation course increase student SAT scores?
The first question asks explicitly about the cause of a plant’s growth. The second question doesn’t actually contain the word “cause” but is effectively asking whether the course causes an increase in SAT score.

On the other hand, we could use the following question as a non-statistical question of interest:

- Does the amount of vitamin C contained in organically grown Granny Smith apples in Washington state differ from the amount contained in non-organically grown ones on average?

This question does not ask whether growing the apples organically causes a different mean amount of vitamin C; it simply asks whether such a difference exists. To demonstrate how this question could be used as a non-statistical question of interest, we must also translate it into a statistical question of interest, but we postpone this until after we have defined constant linear models.

**Examining the data**

Once we have formulated the questions of interest, the next step in a linear model analysis is: *Examine the data*. In order to elaborate on this, we introduce a useful definition.

**Definition: Context of the data.** The context of the data means everything about how the data was collected and what it represents in the real world.

This is not a technical, statistical definition, but merely convenient terminology. It is mean to help us distinguish between the raw numbers that make up the data and what the raw numbers represent.

Mathematically the context of the data is irrelevant, but statistically it is crucial. We can apply the same mathematical tools to an analysis whether the data set represents sizes of chickens or temperatures of raindrops, but the conclusions that we draw about the real world would be quite different in these two cases! Consequently, as we describe how to conduct statistical analyses, we will frequently refer to the context of the data.

With this definition, we can now state the two parts of examining the data in a linear model analysis:

1. Examine the context of the data.
2. Examine numerical and graphical depictions of the raw data itself. We discuss each of these in turn.

**Examining the context of the data**

For all types of linear model analyses, examining the context of the data includes exploring questions such as:

- **Who collected the data set?** This is important in evaluating the reliability of the data source. We might view a study conducted as part of a major university research project as more reliable than an elementary school class project, for example.

- **What does the data set represent?** We should understand clearly how each number represents something in the real world. This will necessarily include a knowledge of the units used for each variable.

- **When was the data set collected?** Many things change with time, and statistical inferences drawn from an outdated data set may no longer be valid.

- **Where was the data set collected?** Many phenomena vary by location, and we might draw statistical inferences that are not valid for the location we’re interested in if we don’t know where our data set was collected.

- **Why was the data set collected?** Sometimes people’s motivations can lead them to represent a data set as something that it is not, or to mislead the unwary data analyst. There may be a big difference between a data set collected by a disinterested party and one collected by someone with a vested interest in what may be implied by the data.

- **How was the data set collected?** If a data set was not collected appropriately, we may not be able to draw any valid statistical inferences from it.

We may not be able to answer all of these questions. However, we should state clearly in any data analysis what information is and is not available.
That way our readers can evaluate the reliability of our results based on all the information we have.

Understanding what the data set represents, including the units used for each variable, is crucial. Without this, we can conduct only the mathematical parts of a statistical analysis; if we don’t know what the raw numbers represent, we cannot complete the key statistical step where the mathematical results are connected with something in the real world.

The question of how the data set was collected is also particularly important in a linear model analysis. For a constant linear model analysis, we need to know whether it is reasonable to consider our data as a sample of the variable of interest. A sample is a collection of independent observations of a random variable, so we should consider whether the variable of interest can be modeled as a random variable and whether the observations in our data are independent. They will never be truly independent (or if they were, we wouldn’t be able to prove it anyway), but we should think about whether they are at least approximately independent. We should also explain these considerations in our constant linear model analysis to help our readers evaluate our analysis.

Examing depictions of the raw data

As the second part of examining the data, we look at numerical and graphical depictions of the raw data itself. For constant linear model analyses, this usually means examining a stripchart of our sample of the variable of interest, although we might also decide to look at other graphical or numerical summaries of the sample.

Understanding why we explore the data will help us select appropriate graphical and numerical displays. Perhaps the most important reason we explore data is to avoid typographical errors in it, as these can lead to highly embarrassing errors in an analysis. Typographical errors are notoriously difficult to eradicate just by looking at raw data in a spreadsheet, but they often jump right out of the page when we view data graphically. If we don’t explore our data graphically, we might conduct our entire analysis without being aware that one of our data points is off by a factor of a thousand! As you can imagine, such an error can affect an analysis drastically.

In order to detect typographical errors and for many other reasons that we will learn later, we look for two things in graphical displays of our data:
patterns and deviations from those patterns. While computers are good at producing displays and summaries of data, they are not good at detecting patterns. Therefore it is important that we evaluate these displays and summaries ourselves.

For a constant linear model analysis, we have only one variable of interest, and we usually examine a stripchart of our sample of this variable. We might examine a histogram or a boxplot as well, especially if the stripchart is too crowded to reveal any patterns. Histogrames and boxplots don’t display each observation individually though, so we should use them only in conjunction with a stripchart, not alone. Whatever the display, the main pattern to look for is the general shape of the distribution, and the main reason we look for this pattern is to check for outliers. For example, we might notice that our sample appears to be of an approximately normally distributed random variable. In this case, if one point appears to be about twenty standard deviations from the mean, then it is certainly an outlier.

If we don’t find any outliers, then we can proceed with our analysis. If we do notice outliers though, we should try to use the context of the data to figure out why these observations are outliers. Three of many possible reasons for an observation to be an outlier are:

1. The value of the observation is a typographical error. Typographical errors are perhaps the most common reason for outliers, so the first thing to do if we find an outlier is to check the original data source to see if the value is a typographical error. If it is and if we can determine the correct value, then the correct value should be used instead of the flawed entry. If we can’t tell whether the error is typographical or if we can’t determine the correct value, then we need to consider omitting the observation in the analysis. However, this should be done only with great care; we discuss this issue further below.

2. The context of the observation is different. If the observation has a different context than the other observations in the sample, then it may have a value that does not fit the general pattern of the sample. For example, if one observation is of a male and all of the other observations are of females, the value of the variable for the male may not fit the pattern of the females. If an observation has a different context (such as being male), we should ask whether we want to include observations with its context (such as males) in our analysis. If
not, then we can remove the outlier and know that our analysis does not apply to that type of observation (males, in our example). If so, then we should try to obtain more observations with a similar context (more observations of males here) in order to clarify how they fit into the model. If we can’t obtain such data, then we should be aware that the lack of such observations may limit the applicability of our findings, and we should include a statement to this effect in our analysis.

3. The observation’s context is not different, so the observation casts doubt on the pattern. If we find that an observation has a similar context but genuinely does not fit the pattern of the other observations, then we must consider the possibility that such a pattern does not hold in general. For example, suppose that we thought that all of the observations would be positive, but we find an outlier that is negative. If this outlier is not a typographical error and does not result from a different data context, then we are forced to reconsider the positive pattern. More generally, if the pattern is crucial to our analysis but an outlier forces us to re-evaluate it, then we may need to consider using a different model or even a different type of model altogether.

Notice that in this list of common possibilities, outliers that are not typographical errors are often highly informative for our analysis. The above list illustrates that if we remove outliers without considering their context, we may be throwing away the most informative observations in the entire data set! Under no circumstances should we omit an observation because we don’t like its statistical properties; we should omit an observation only if its data context dictates doing so.

We face a difficult decision when we can’t determine the nature of an outlier. For example, we might be almost certain that a point is a typographical error, but if we don’t have access to the original data source, we can’t be sure and we don’t know what the correct value should be. As another example, we might suspect that an outlier is an observation of a male while all the other observations are of females, but if we don’t have access to the genders of the individuals being observed, we can’t be sure that the gender of the outlier is different from the gender of the other observations.

Statisticians have developed many ways to deal with such cases of imperfect information, but these are beyond the scope of this book. As a practical matter, we propose a conservative approach here: when in
doubt, conduct two analyses, one including outliers and one excluding them. If the two analyses yield similar results, then you don’t need to agonize about which one to use. If they yield different results, include both analyses in your overall analysis and be as forthcoming as possible about everything that you were able to determine about the nature of the outliers.

**Fitting the model**

Once we have determined the questions of interest and examined the data, the next step in a linear model analysis is to fit the model to data. In order to define this term, we now introduce several concepts, beginning with that of a constant linear model.

**Constant linear models**

A constant linear model is a model to describe a random variable whose mean does not depend systematically on the values of any other variables. More specifically, we have the following definition.

**Definition:** Constant Linear Model. A constant linear model consists of a random variable $Y$ with mean $\mu(Y)$ and a real number $\beta_0$, related by the equation:

$$\mu(Y) = \beta_0.$$

The random variable $Y$ is called the response variable, and the real number $\beta_0$ is called the coefficient of the model. The above equation is called the model equation.

When we conduct constant linear model analyses, we are often interested in the deviation of the response variable from its mean, which leads us to the following definition.

**Definition:** Error Terms and Residuals in Constant Models. For a constant linear model whose response variable $Y$ has mean $\mu(Y)$, the error term $e(Y)$ is defined to be:

$$e(Y) := Y - \mu(Y).$$
Any sample $y_1, y_2, \ldots, y_n$ of $Y$ gives rise to a sample $e_1, e_2, \ldots, e_n$ of $e(Y)$ defined by:

$$e_i := y_i - \mu(Y)$$

for $i = 1, 2, \ldots, n$. The real number $e_i$ is called the $i$-th residual of the model relative to the sample.

Because $\mu(Y) = \beta_0$, the above equations can be rewritten as:

$$e(Y) := Y - \beta_0$$
$$e_i := y_i - \beta_0.$$ 

However, these equations are specific to regression on a constant. The equations given in the definition apply to many other types of regression as well, as we will see later.

When we conduct a constant linear model analysis, we usually want to interpret the model coefficient in terms of what we are modeling. This is straightforward because the coefficient $\beta_0$ of the model is (by definition) simply the mean $\mu(Y)$ of the response variable. In the broader context of linear model analyses though, we often go further than this in our interpretation. We think of $\mu(Y)$ as the deterministic part of the model and $e(Y)$, which is what is left after we subtract off the deterministic part, as the random part of the model. The deterministic part of the model is systematic and should be something that can be predicted exactly once the model is known. The random part represents the variability in the response variable that cannot be predicted exactly with the model.

**Fitting the model**

In a constant linear model analysis, rather than using only a single constant linear model, we usually consider a set of closely related constant linear models. This leads us to the following definition.

**Definition: Constant Linear Model Class.** The model class $Y \sim 1$ (read “$Y$ modeled on $1$”) is defined to be the set of all linear models on a constant with response variable $Y$.

When we use the model class $Y \sim 1$, we are assuming that the random variable $Y$ can be correctly described by an equation of the form given in the definition on page 77. We don’t know the value of $\beta_0$ though, so at the outset of a constant linear model analysis,
we don’t know which model in this model class is actually a correct description of the random variable $Y$. Instead, we use a particular criterion to help us estimate $\beta_0$ from the available data, a process called fitting a linear model. To fit a model is to choose a particular model from among a specified class of models according to a specified criterion. The criterion used in linear model analyses is called the least squares criterion, which we now define.

**Definition: Least Squares Criterion.** A model satisfies the least squares criterion among a specified class of models if it has the smallest sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

among all models in the specified class.

Thanks to the following mathematical theorem, we can be more specific about how to use this criterion in the context of a constant linear model analysis.

**Theorem: Constant Linear Model Least Squares Fitting.** If $Y$ be a random variable and $y_1, y_2, \ldots, y_n$ is a sample of $Y$, then there is a unique linear model of the class $Y \sim 1$ satisfying the least squares criterion. Furthermore, the value of the coefficient in this model is the sample mean $\bar{y}$.

The proof of this theorem is a straightforward exercise in calculus that we leave to the interested reader.

The existence and uniqueness established by this theorem allow us to make the following definition.

**Definition: Constant Linear Model Fitting.** Let $Y$ be a random variable. To fit a linear model of class $Y \sim 1$ to a given sample of $Y$ is to select the unique model satisfying the least squares criterion among the class. The model selected this way is called the fitted model. For the fitted model, we denote the mean of $Y$ by $\hat{\mu}(Y)$ and the coefficient by $\hat{\beta}_0$. This means that the equation of the fitted model is:

$$\hat{\mu}(Y) = \hat{\beta}_0.$$

By the uniqueness guaranteed by the above theorem, there is exactly one fitted model of class $Y \sim 1$, so we can correctly refer to that model as the fitted model of that class.
Consistent with the notation introduced in this definition, we will denote quantities associated with fitted models with hats throughout this text. We will interpret these quantities as estimates of the corresponding true quantities, which we ordinarily denote without hats. For example, in the context of constant linear models, we interpret the fitted model coefficient $\hat{\beta}_0$ as an estimate of the true (unknown) model coefficient $\beta_0$. Similarly, we think of the random variable mean $\hat{\mu}(Y)$ if the fitted model is correct as an estimate of the true (unknown) random variable mean $\mu(Y)$.

From this point of view, it makes sense that $\hat{\beta}_0 = \bar{y}$, since $\hat{\beta}_0$ is an estimate of $\beta_0$, which equals the true random variable mean $\mu(Y)$. We are already used to thinking of the sample mean as an estimate of the random variable mean in other contexts.

**Sampling Variability Assumptions**

To fit a constant linear model is completely straightforward; we simply compute a sample mean. If that were as far as linear model analysis went though, this would be a very short book indeed!

However, a prevailing philosophy in statistics dictates that making an estimate from a sample is not particularly interesting unless we can say something about the way that our estimate would vary if we took other samples. Of course, the nicest way to be able to say such a thing would be if we could take lots of samples, but that is ordinarily not a luxury that we have. Instead, we make some assumptions about the way that the response variable $Y$ is distributed (or perhaps more often about the way that the fitted model error term $\hat{e}(Y)$ is the model is distributed), and from these assumptions we can conclude things about how our estimates would vary if we could take more samples.

We refer to these assumptions as *sampling variability assumptions*. They are an important part of regression analysis, even though they are not part of the definitions of linear models themselves. For the particular case of regression analysis involving linear models on a constant, there is only a single sampling variability assumption: normality of the error term.
We now elaborate on this assumption and how to verify it.

**Normality of the error term**

When we work with linear models on a constant, the response variable $Y$ and the fitted model error term $\hat{e}(Y)$ differ only by a constant (namely $\hat{\mu}(Y)$, which equals $\hat{\beta}_0$). This means that the error term is normally distributed if and only if the response variable is normally distributed. However, for various reasons that will become clear when we explore more general types of linear models, we focus on the error term’s distribution rather than that of the response variable. Suffice it to say at this point that considering the distribution of the response variable itself would be a bad habit to get into, one that won’t be appropriate for linear models in general.

To verify the normality of the random variable $\hat{e}(Y)$, we examine the only data that we have for it: the residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$, which are the sample of $\hat{e}(Y)$ derived from a sample $y_1, y_2, \ldots, y_n$ of $Y$. Using the residuals, we check the normality of $\hat{e}(Y)$ graphically and numerically as we would for any random variable, namely with:

- A normal quantile plot of the residuals.
- A histogram of the residuals.
- The Shapiro-Wilk test applied to the residuals.

As with any random variable, these methods may give conflicting signals as to the possible normality of $\hat{e}(Y)$. In this case, decision-making precedence should be given to the methods in the order that they are listed here. All such conflicts should be noted in the analysis, whatever decision regarding the possible normality of $\hat{e}(Y)$ is reached.

If the fitted model error term seems to be at least approximately normally distributed, then we are free to proceed with statistical inference to
address our questions of interest. If not though, another approach to answering the questions of interest should be taken. One possibility would be to try transforming the data, as we will discuss later in this text. Another possibility, which we will not discuss in this text, is to consider using a non-parametric model (meaning a model that does not make assumptions about the distributions of the variables involved).

**Statistical inference**

If the Sampling Variability Assumption is (at least approximately) satisfied for the model class $Y \sim 1$, then we can draw statistical inferences from the fitted model. This amounts to addressing questions about the true model based on what we know about the fitted model and what we assume about model variability under sampling. The two standard types of statistical inference that we apply in regression analysis are conducting of hypothesis tests and computation of confidence intervals.

Before we introduce the key mathematical theorem that allows us to carry out such statistical inferences, we need some preliminary definitions.

**Definition: Statistic.** A statistic is a random variable whose value is based on sampling some other random variable.

Examples of statistics include the sample mean, the sample standard deviation, and (of particular relevance for our purposes) the model coefficient $\hat{\beta}_0$.

When we concern ourselves with the distribution of a statistic, we ordinarily restrict our attention to the values that the statistic takes over all possible samples of a given size of a given random variable. When this set of possible samples needs clarification, we will do so. In some cases though, we will leave the set of possible samples to be determined by the context, as in the following definition.

**Definition: Standard Error.** An estimate of the standard deviation of a statistic is called a standard error of the statistic.

The main reason that we have introduced this definition is conceptual, not computational. The definition asserts that the standard error of a statistic is a measure of its variability over all possible samples of the random variable that it is derived from. For our purposes at the moment, we are particularly interested in the variability of $\hat{\beta}_0$ as different samples
of $Y$ are observed.

**Definition: Standard Error of Constant Model Coefficient.** Suppose $Y$ is a random variable, and $y = \{y_1, y_2, \ldots, y_n\}$ is a sample of $Y$ with sample mean $s(y)$. If the Sampling Variability Assumption for linear models on a constant is satisfied, then the **standard error** $se(\hat{\beta}_0)$ of $\hat{\beta}_0$ is defined to be:

$$se(\hat{\beta}_0) := \frac{s(y)}{\sqrt{n}}$$

This is an estimate of the standard deviation of $\hat{\beta}_0$ over all samples of $Y$ of size $n$.

Having introduced such terminology, we can now state the key mathematical result underlying our ability to carry out statistical inferences on linear models on a constant.

**Theorem: Test Statistic Distribution for Constant Models.** Suppose $Y$ is a random variable, and $y = \{y_1, y_2, \ldots, y_n\}$ is a sample of $Y$. Also suppose that the Sampling Variability Assumption for linear models on a constant is satisfied. If $\mu(Y) = \beta_0$ is the true linear model, then when considered over all samples of $Y$ of size $n$, the statistic

$$T_0 := \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}$$

has a $t$ distribution with $n - 1$ degrees of freedom.

With this result at our disposal, we can discuss how to carry out statistical inferences on linear models on a constant.

**Hypothesis Tests**

As discussed previously, a hypothesis test provides the answer to the question: If the null hypothesis is true, what is the probability of observing a test statistic value at least as extreme as one for our sample? The meaning of the term “extreme” depends on what the alternative hypothesis is, and the probability computed in a hypothesis test is called the $p$-value of the test. If the $p$-value of the test is less than a pre-determined significance level, traditionally 0.05, then we conclude that ....
In the context of linear models on a constant, the most common hypothesis test is of the form

\[ H_{null} : \beta_0 = b_0 \quad \text{and} \quad H_{alt} : \beta_0 \neq b_0, \]

where \( \beta_0 \) is the unknown (and not directly observable) model coefficient and \( b_0 \) is some particular value of interest, often 0. In conducting such a hypothesis test, we are trying to determine whether our data provides evidence that the random variable mean of \( Y \) (which is \( \beta_0 \)) is different from the specified value \( b_0 \). The test statistic for such a hypothesis test is

\[ T_0 := \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}. \]

The theorem \textit{TEST STATISTIC DISTRIBUTION FOR CONSTANT MODELS} on page 83 guarantees that if the Sampling Variability Assumption for linear models on a constant is satisfied, then we know the distribution of \( T_0 \) over size \( n \) samples of \( Y \) under the null hypothesis: it is \( t_{n-1} \).

With this, conducting such a hypothesis test is routine. We first compute the value \( t_0 \) of the test statistic \( T_0 \) for our sample of \( Y \). The \( p \)-value is simply the area of the outer tails of \( t_{n-1} \) at \( t_0 \).

\textit{Confidence intervals}

In the context of linear models on a constant, another common type of statistical inference is the computation of a confidence interval for the true model coefficient \( \beta_0 \). A level \( C \) confidence interval for \( \beta_0 \) has two main interpretations:

1. It is a set of values that are consistent with our data at the specified level, meaning that it consists of precisely those values \( b_0 \) of \( \hat{\beta}_0 \) that would not be rejected in a 2-sided hypothesis test of \( H_{null} : \beta_0 = b_0 \) with significance level \( 1 - C \).

2. If we consider all possible size \( n \) samples of \( Y \), the probability that the true \( \beta_0 \) will lie in the confidence interval constructed from the sample this way equals the confidence level \( C \).

The theorem \textit{TEST STATISTIC DISTRIBUTION FOR CONSTANT MODELS} on page 83 provides us with a test statistic \( T_0 \) with which we can compute
confidence intervals for the true model coefficient $\beta_0$. Following the previously discussed method of computing confidence intervals, we arrive at the following result.

**Theorem: Confidence Interval for Constant Model Coefficient.**

Let $Y$ be a random variable satisfying the Sampling Variability Assumption for the model class $Y \sim 1$. Also, let $t^*$ denote the central $C$-quantile for a $t$ distribution with $n - 1$ degrees of freedom. Then a level $C$ confidence interval for the true model coefficient $b_0$ is from

$$\hat{\beta}_0 - se(\hat{\beta}_0) t^* \text{ to } \hat{\beta}_0 + se(\hat{\beta}_0) t^*.$$ 

With this, we can now readily compute a confidence interval for the true model coefficient in the model class $Y \sim 1$. Since the true model coefficient is the random variable mean of $Y$, we now have a method of computing a confidence interval for this mean.

**Worked examples**

Now that we have discussed the constant linear model analysis process, we work through some examples to illustrate it.

**Paired $t$-inferences**

As listed on page 70, two common types of statistical questions of interest for constant linear model analyses arise as follows:

- We have a random variable $Y$ that is the difference $Y_2 - Y_1$ of two other random variables. In order to determine whether the means of $Y_1$ and $Y_2$ are equal, we ask the question: Is the mean of $Y$ equal to 0?

- Again we have $Y = Y_2 - Y_1$, but now we simply ask: What is the mean of $Y$? The answer to this question will tell us how different the means of $Y_1$ and $Y_2$ are.

The process used to answer the first of these is traditionally called a **paired $t$-test**. The method used to solve the second isn’t usually called by name, but it might be referred to as computing a **paired $t$ confidence interval**. The term paired refers to the fact that we collect observations in pairs (one
of $Y_1$ and one of $Y_2$), which we then compare by subtraction (in investigating $Y$).

The following example illustrates both of these types of statistical questions of interest.

**Example: Comparing Mean Precipitation in Two Cities.** Does it rain more in Tacoma, Washington or Portland, Oregon? Use the Pacific Northwest precipitation data 1966–2006 data set to compare Tacoma’s mean annual precipitation with that of Portland.

We translate the statement of the problem into two non-statistical questions of interest:

1. Is the difference in annual precipitation amounts (Tacoma’s minus Portland’s) equal to 0 on average?
2. What is the average magnitude of the difference in annual precipitation amounts (Tacoma’s minus Portland’s)?

These in turn give us two corresponding statistical questions of interest. We state these questions in terms of the variable of interest $\Delta \text{precipitation}$, which we define to be Tacoma’s annual precipitation amount minus Portland’s annual precipitation amount. We also use the notation that $\beta_0$ is the coefficient in the (unknown) true linear model of class

$$\Delta \text{precipitation} \sim 1.$$ 

With this notation, the statistical questions of interest are:

1. Conduct a hypothesis test of:

   $$H_{null} : \beta_0 = 0 \text{ inches and } H_{alt} : \beta_0 \neq 0 \text{ inches},$$

   evaluating the result at the usual significance level of 0.05.

2. Compute a 95% confidence interval for $\Delta \text{precipitation}$. 

Having formulated the questions of interest, we next examine the data. According to its description, the Pacific Northwest precipitation data 1966–2006 data set was compiled from data from the Western Regional Climate Center, a reliable source. However, we should consider the extent to which contiguous years of precipitation data are independent observations of an annual precipitation random variable. While we might not
think that precipitation one year is highly correlated with annual precipitation the next year, each rainy season (primarily Fall through Spring) overlaps two calendar years. Also, there could be some correlation between precipitation in certain non-contiguous years, such as El Niño years. In our estimation, such possible correlations are small and therefore indicate only a small deviation from independence for observations in this data set. So we can reasonably think of the data as a sample of Tacoma’s annual precipitation random variable and one of Portland’s.

As part of examining the data, we make a stripchart of our sample of the random variable of interest \( \Delta \text{precipitation} \), looking for anything unusual or noteworthy.

We don’t see any outliers or anything else needing particular attention here, so we continue with our analysis.

The next step in the analysis is to fit the model of class

\[ \Delta \text{precipitation} \sim 1. \]

We have a computer do this.

In order to be able to draw the desired statistical inferences, we next check the Sampling Variability Assumption for models of this class, namely the normality of the error term. For this, we examine a normal quantile plot of the residuals.
We next conduct the desired hypothesis test of

\[ H_{\text{null}} : \beta_0 = 0 \text{ inches} \quad \text{and} \quad H_{\text{alt}} : \beta_0 \neq 0 \text{ inches}. \]

For this, we compute the value \( t_0 \) of the test statistic \( T_0 \) for our sample of the random variable \( \Delta \text{precipitation} \). The theorem \textsc{test statistic distribution for constant models} on page 83 gives us the formula for \( t_0 \) as

\[
t_0 = \frac{\hat{\beta}_0 - 0}{\text{se}(\hat{\beta}_0)} = \frac{0.550}{0.948} = 0.580.
\]

Since we have 41 observations in our sample, the theorem also tells us that the distribution of \( T_0 \) under the null hypothesis is \( t_{40} \). Computing the area
of the outer tails at 0.580 of $t_{40}$, we obtain a $p$-value of:

$$p = 0.565.$$  

Since this is above our chosen significance level of 0.05, we do not have statistically significant evidence that the mean of $\Delta$precipitation is nonzero.

To answer the second statistical question of interest, we give the least squares estimate $\hat{\beta}_0$ along with a 95% confidence interval for the true model coefficient $\beta_0$, which is also the true mean of $\Delta$precipitation. In this example, the computer tells us that

$$\hat{\beta}_0 = 0.55 \text{ inches}.$$  

To compute a 95% confidence interval for $\beta_0$, we use the the formula given in the theorem [CONFIDENCE INTERVAL FOR CONSTANT MODEL COEFFICIENT](#) on page 85. The computer gives us the following values for the terms in the formula:

$$\hat{\beta}_0 = 0.55$$
$$se(\hat{\beta}_0) = 0.95$$
$$t^{*} = 2.023,$$

where $t^{*}$ is the central 0.95 quantile for a $t_{39}$ distribution. Putting these values into the formula, we find that a 95% confidence interval for $\hat{\beta}_0$ is from $-1.4$ inches to 2.5 inches.

As we expect from the statistically insignificant result of our hypothesis test, this confidence interval includes negative values, positive values, and zero, so at this confidence level we cannot conclude anything in particular about whether the mean difference in annual precipitation amounts is negative, positive, or zero.

To interpret our results, with our hypothesis test we did not find statistically significant evidence that the mean annual precipitation amount in Tacoma was different from that in Portland. We estimate the mean annual precipitation amount in Tacoma to be 0.55 inches more than the mean in Portland, with a 95% confidence interval for the mean of Tacoma’s annual precipitation minus Portland’s being from $-1.4$ inches to 2.5 inches. ♦

Of course, we draw the same conclusions (or lack thereof) from the hypothesis test and the confidence interval computation. However, these two statistical methods are designed to address the questions of interest
in different ways. A hypothesis test addresses the question of interest by answering the question of how extreme our data would be under the null hypothesis. A confidence interval addresses the questions of interest by producing a range of coefficient values that are consistent with our data at the specified level.

Estimating a mean

As listed on page 70, two more common types of statistical questions of interest for constant linear model analyses arise as follows:

- We have a random variable $Y$ (that is not necessarily a difference of random variables), and we choose a particular value $b_0$ and ask: Is the mean of $Y$ equal to $b_0$?

- We have a random variable $Y$ and we ask: What is the mean of $Y$?

We traditionally address the first of these with what is called a 1-sample $t$ test. The method ordinarily used to address the second doesn’t really have a name but might be called computing a 1-sample $t$ confidence interval.

The next example illustrates both of these questions.

**Example: estimating mean Peanut M and M mass.** A particular bag of Peanut M&Ms states that the total mass of the M&Ms that it contains equals 396.9 grams. Since the bag contains 153 M&Ms, we might guess that the true mean mass of a Peanut M&M in general is $396.9/153 = 2.594$ grams. Use the M&M candies data set to test this and to estimate the true mean mass.

From the statement of the problem, we determine two non-statistical questions of interest:

1. Is the average mass of a Peanut M&M equal to 2.594 grams?
2. What is the average mass of a Peanut M&M?

We translate these into two corresponding statistical questions of interest. We state these questions in terms of the variable of interest mass, which we define to be the mass of a Peanut M&M. We also use the notation that $\beta_0$ is the coefficient of the (unknown) true linear model of class $mass \sim 1$.

With this notation, the statistical questions of interest are:
1. What are the results of a hypothesis test of:

   \[ H_{null} : \beta_0 = 2.594 \text{ grams} \quad \text{and} \quad H_{alt} : \beta_0 \neq 2.594 \text{ grams}, \]

   evaluated at the usual significance level of 0.05?

2. What are an estimate and a 95\% confidence interval for mass?

Looking at the description of the data set, the source seems objective and reliable enough. The sample consists of a single bag of Peanut M&Ms, meaning they are certainly not a simple random sample of all Peanut M&Ms in the world. We should note this, but for the purposes of our analysis, we will need to make the simplifying assumption that they do form a simple random sample, so that the masses constitute a sample of the random variable that gives the mass of a randomly chosen Peanut M&M.

With this simplifying assumption, we can proceed. We next make a density plot of our sample, looking for anything unusual or noteworthy.

This plot reveals no unusual patterns or outliers, so we proceed with our analysis.

We use a computer to fit the model of class mass \( \sim 1 \).
There are two things we would like to do in this problem: test whether the true mean Peanut M&M mass is 2.594 grams, and estimate the true mean Peanut M&M mass. For the first, we conduct a hypothesis test, and for the second, we compute a confidence interval. For both of these, we must first verify the Sampling Variability Assumption for the model class.

For this, we use a computer to fit the model of this class to our sample of mass and

To verify the Sampling Variability Assumption, which is the normality of the error term, we examine a normal quantile plot of the residuals.

![Quantile plot](insert-quantile-plot.png)

Although there is a slight swing at both ends of the plot, well over 95% of the points are within the 95% confidence bands and there are no other disturbing patterns. This tells us that mass does appear to be normally distributed, or at least very nearly so.

In addition, a Shapiro-Wilk test gives a $p$-value of 0.082. This means that, using the traditional significance level of 0.05, we do not have statistically significant evidence that mass is non-normally distributed, which is in keeping with what we see in the normal quantile plot.

In short, both graphically and numerically, the random variable mass appears to be at least approximately normally distributed, meaning that the Sampling Variability Assumption is satisfied. This allows us to proceed to the desired statistical inferences.
To answer the first statistical question of interest, we conduct the specified hypothesis test. For this, we first compute the value $t_0$ of the test statistic $T_0$ for our sample of mass. The theorem TEST STATISTIC DISTRIBUTION FOR CONSTANT MODELS on page 83 gives us the formula for $t_0$ as

$$t_0 = \frac{\hat{\beta}_0 - 2.594}{se(\hat{\beta}_0)} = \frac{2.598 - 2.594}{0.027} = 0.136.$$ 

Since we have 153 observations in our sample, the theorem also tells us that the distribution of $T_0$ under the null hypothesis is $t_{152}$. Computing the area of the outer tails at 0.136 of $t_{152}$, we obtain a $p$-value of:

$$p = 0.892.$$ 

Since this is above our chosen significance level of 0.05, we do not have statistically significant evidence that the true coefficient $\beta_0$ is different from 2.594 grams.

We next address the second statistical question of interest. Using least squares as our optimality criterion, our best estimate of this true model coefficient is the fitted model coefficient $\hat{\beta}_0$, which equals 2.60 grams.

To compute a 95% confidence interval for the true model coefficient $\beta_0$, we use the formula given in the theorem CONFIDENCE INTERVAL FOR CONSTANT MODEL COEFFICIENT on page 85. The computer gives us the values that we need for this formula:

$$\hat{\beta}_0 = 2.60$$
$$se(\hat{\beta}_0) = 0.03$$
$$t^* = 1.976,$$

where $t^*$ is the central .95 quantile for a $t_{151}$ distribution. Putting these numbers into the formula, we compute that a 95% confidence interval for the true mean Peanut M&M mass is from 2.544 grams to 2.652 grams.

In conclusion, with our hypothesis test we did not find statistically significant evidence that the average mass of a Peanut M&M is different from 2.594 grams. We estimate the mean mass of a Peanut M&M to be 2.60 grams, with a 95% confidence interval from 2.544 grams to 2.652 grams. We should remember, however, our sample was only a single bag of Peanut M&Ms, far from a simple random sample of all the Peanut M&Ms in the world. To make our statistical inferences, we made the false
but simplifying assumption that this bag was a simple random sample, which could affect whether our results are correct.

With these two worked examples, we have illustrated the most common applications of constant linear model analyses. Once you understand these examples, you may also find less common uses for such analyses as well.