Chapter 3: SLR Estimation and Prediction

# Introduction

A typical regression analysis involves the following steps:

1. Model formulation
2. Model estimation
3. Model evaluation
4. Model use

So far, we have learned how to formulate and estimate a simple linear regression model. We have also learned about some methods for evaluating the model (and we will learn further evaluation methods in Chapter 4). In this lesson, we focus our efforts on using the model to answer two specific research questions, namely:

* What is the average response for a given value of the predictor ?
* What is the value of the response likely to be for a given value of the predictor ?

In particular, we will learn how to calculate and interpret:

* A confidence interval for estimating the mean response for a given value of the predictor .
* A prediction interval for predicting a new response for a given value of the predictor .

# Learning Objectives and Outcomes

Upon completion of this lesson, you should be able to do the following:

* Distinguish between estimating a mean response (confidence interval) and predicting a new observation (prediction interval).
* Understanding the various factors that affect the width of a confidence interval for a mean response.
* Understand why a prediction interval for a new response is eider than the corresponding confidence interval for a mean response.
* Know the formula for a prediction interval depends strongly on the condition that the error terms are normally distributed, while the formula for the confidence interval is not so dependent on this condition for large samples.
* Know the types of research questions that can be answered using the materials and methods of this lesson.

# The research questions

In this lesson, we are concerned with answering two different types of research questions. Our goal here – and throughout the practice statistics – is to translate the research questions into reasonable statistical procedures.

Let’s take a look at examples of the two types of research questions we learn how to answer in this lesson:

1. What is the mean weight, , of all American women, aged 18-24?

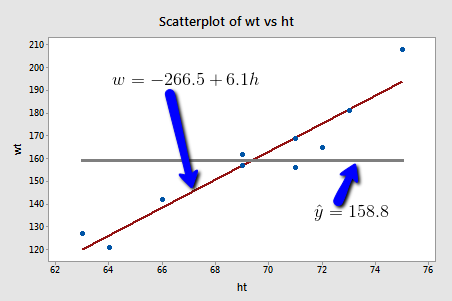
If we wanted to estimate , what would be a good estimate? It seems reasonable to calculate a confidence interval for using , the average weight of a random sample of American women, aged 18-24.

1. What is the weight, , of an individual American woman, aged 18-24?

If we want to predict , what would be a good prediction? It seems reasonable to calculate a “prediction interval” for using, again, , the average weight of a random sample of American women, aged 18-24.

A person’s weight, of course, highly associated with the person’s height. In answering each of the above questions, we likely could do better by taking into account a person’s height. That’s where an estimated regression equation becomes useful.

Here are some weight and height data from ta sample of people. ([student\_height\_weight.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/student_height_weight.txt)):

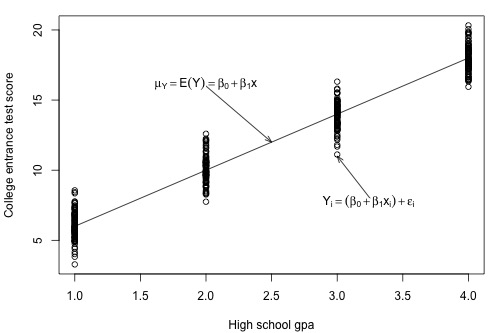


If we used the average weight of the 10 people in the sample to estimate , we could claim that the average weight of all American women aged 18-24 is 158.8 pounds *regardless* of the height of the women. Similarly, if we used the average weight of the 10 people in the sample to predict , we would claim that the weight of an individual American women aged 18-24 is 158.8 pounds *regardless* of the woman’s height.

On the other hand, if we used the estimated regression equation to estimate , we would claim that the average weight of all American women aged 18-24 *who are only 64 inches tall* is -266.5 + 6.1(64) = 123.9 pounds. Similarly, we would predict that the weight of an individual American woman aged 18-24 *who is only 64 inches’ tall* is 123.9 pounds. This example makes it clear that we get significantly different (and better!) answers to our research questions when we take into account a person’s height.

Let’s make it clear that it is one thing to **estimate**  and yet another thing to **predict** . (Note that we subscript with to make it clear that we are talking about the mean of the response , not the mean of the predictor .)

Let’s return to our example in which we consider the potential relationship between the predictor “high school GPA” and the response “college entrance test score.”



For this example, we could ask two different research questions concerning the response:

* What is the mean college entrance test score for the subpopulation of students whose high school GAP is 3? (answering this question entails estimating the mean response when .)
* What college entrance test score can we predict for a student whose high school GPA is 3? (Answering this question entails predicting the response when.)

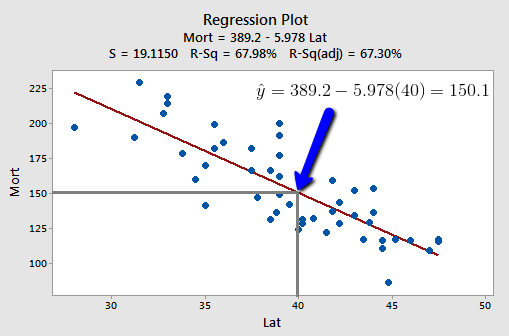
The two research questions can be asked more generally as:

* What is the mean response when the predictor value is ?
* What value will a new response be when the predictor value is ?

Let’s take a look at one more example, namely, the one concerning the relationship between the response “skin cancer mortality” and the predictor “latitude” ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)). Again, we could ask two different research questions concerning the response:

* What is the expected (mean) mortality rate due to skin cancer **for all locations** at 40 degrees north latitude?
* What is the predicted mortality rate **for one individual location** at 40 degrees north, say at Chambersburg, Pennsylvania?

At some level, answering these two research questions is straightforward. Both involve using the estimated regression equation:



That is, is the best answer to each of the research questions. It is the best guess of the mean response and it is the best guess of a new response at :

* Our best estimate of the mean mortality rate due to skin cancer for all location at 40 degrees north latitude is 389.19 – 5.97764(40) = 150 deaths per 10 million people.
* Our best prediction of the mortality rate due to skin cancer in Chambersburg, Pennsylvania is 389.19 – 5.97764(40) = 150 deaths per 10 million people.

The problem with the answers to our two research questions is that we’d have obtained a completely different answer if we had selected a different random sample of data. As always, to be confident in the answer to our research questions, we should put an interval around our best guesses. We learn how to do this in the next two sections. That is, we first learn a “**confidence interval for** ” and then a “**prediction interval for** .”

# Confidence Interval for the Mean Response

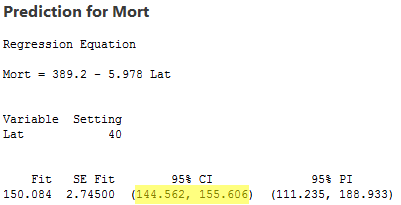
In this section, we are concerned with the confidence interval, called a “-interval,” for the mean response when the predictor value is . Let’s jump right in and learn the formula for the confidence interval. The general formula in words is as always:

Sample estimate (-multiplier standard error)

Where:

* is the “**fitted value**” or “**predicted value**” of the response when the predictor is
* is the “-multiplier has (not ) degrees of freedom, because the confidence interval uses the mean square error () whose denominator is .
* is the “standard error of the fit,” which depends on the mean square error (), the sample size (), how far in squared units the predictor value is from tge average of the predictor values , or , and the sum of the squared distances of the predictor values from the average of the predictor values , or.

Fortunately, we won’t have to use the formula to calculate the confidence interval, since statistical software such as Minitab will do the dirty work for us. Here is some Minitab output for our example with “skin cancer mortality” as the response and “latitude” as the predictor ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)):



Here’s what the output tells us:

* In the section labeled “**Values of Predictors for New Observations**,” Minitab reports the value .
* In the section labeled “**Predicted Values for New Observations**,” Minitab reports the 95% confidence interval. We can be 95% confident that the average skin cancer mortality rate of all locations at 40 degrees north is between 144.562 and 155.606 deaths per 10 million people.
* In the section labeled “**Predicted Values for New Observations**,” Minitab also reports the predicted value , (“**Fit**” = 150.084), the standard error of the fit (“**SE Fit**” = 2.74500), and the 95% prediction interval for a new response (which we discuss in the next section).

## Factors affecting the width of the -interval for the mean response

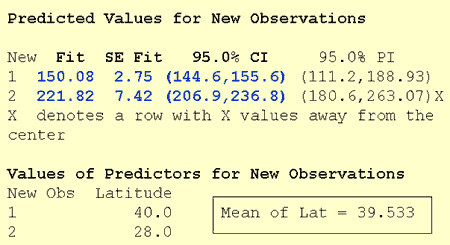
Why do we bother learning the formula for the confidence interval for when we let statistical software calculate it for us anyway? As always, the formula is useful for investigating what factors affect the width of the confidence interval for . Again, the formula is:

And therefore, the width of the confidence interval for is:

So how can we affect the width of our resulting interval for ?

* **As the mean square error () decreases, the width of the interval decreases**. Since is an estimate of how much the data vary naturally around the unknown population regression line, we have little control over other than making sure that we make our measurements as carefully as possible. (We will return to this issue later when we address “model selection.”)
* **As we decrease the confidence level, the -multiplier decreases, and hence the width of the interval decreases**. In practice, we wouldn’t want to set the confidence level below 90%
* **As we increase the sample size , the width of the interval decreases**. We have complete control over the size of our sample – the only limitation being our time and financial constraints.
* The more spread out the predictor values, the larger the quantity and hence the narrower the interval. In general, you should make sure you predictor values are not too clumped together but rather sufficiently spread out.
* The closer is to the average of the sample’s predictor values , the smaller the quantity , and hence the narrower the interval. If you know that you want to use your estimated regression equation to estimate when the predictor’s value is , then you should be aware that the confidence interval will be narrower the closer is to .

Let’s see this last claim for our example with “skin cancer mortality” as the response and “latitude” as the predictor:



The Minitab output reports a 95% confidence interval for for a latitude of 40 degrees north (first row) and 28 degrees north (second row). The average latitude of the 49 states in the data set is 39.533 degrees north. The output tells us:

* We can be 95% confident that the mean skin cancer mortality rate of all locations at 40 degrees north is between 144.6 and 155.6 deaths per 10 million people.
* And, we can be 95% confident that the mean skin cancer mortality rate of all locations at 28 degrees north is between 206.9 and 236.8 deaths per 10 million people.

The width of the 40 degrees north interval (155.6 – 144.6 = 11 deaths) is shorter than the width of the 28 degree north interval (236.8 – 206.9 = 29.9 deaths), because 40 is much closer than 28 is to the sample mean 39.533. note that Minitab is kind enough to warn us (“X denotes a row with X values away from the center”) that 28 degrees north is far from the mean of the sample’s predictor values.

## When is it okay to use the formula for the confidence interval for ?

One thing we haven’t discussed yet is when it is okay to use the formula for the confidence interval for . It is okay:

* When is a value within the range of the values in the data set – that is, when is a value within the “**scope of the model**.” But, note that does not have to be one of the actual values in the data set.
* When the “LINE” conditions – linearity, independent errors, normal errors, equal error variances – are met. The formula works okay even if the error terms are only approximately normal. And, if you have a large sample, the error terms can even deviate substantially from normality.

# Prediction Interval for a New Response

In this section, we are concerned with the prediction interval for a new response when the predictor’s value is . Again, let’s just jump right in and learn the formula for the prediction interval. The general formula in words is as always:

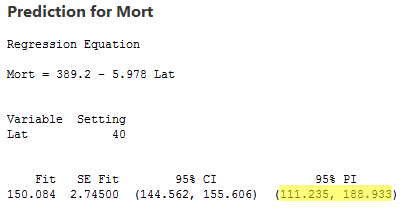
**Sample estimate (-multiplier standard error)**

And the formula in notation is:

Where:

* is the “**fitted value**” or “**predicted value**” of the response when the predictor is
* is the “**-multiplier**.” Note again that the -multiplier has (not ) degrees of freedom, because the prediction interval uses the mean square error () whose denominator is .
* is the “**standard error of the prediction**,” which is very similar to the “standard error of the fit” when estimating . The standard error of the prediction just has an extra term added that the standard error of the fit does not.

Again, we won’t use the formula to calculate our prediction intervals. We’ll let statistical software such as Minitab do the calculation for us. Let’s look at the prediction interval for our example with “skin cancer mortality” as the response and “latitude” as the predictor ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)):



The output reports the 95% prediction interval for an individual location at 40 degrees north. We can be 95% confident that the skin cancer mortality rate at an individual location at 40 degrees north will be between 111.235 and 188.933 deaths per 10 million people.

## When is it okay to use the prediction interval for formula?

The requirements are similar to, but a little more restrictive than, those for the confidence interval. It is okay:

* When is a value within the scope of the model. Again, does not have to be one of the actual values in the data set.
* When the “LINE” conditions – Linearity, independent errors, equal error variances – are met. Unlike the case for the formula for the confidence interval, the formula for the prediction interval depends **strongly** on the condition that the error terms are normally distributed.

## Understanding the difference in the two formulas

In our discussion of the confidence interval for , we used the formula to investigate what factors affect the width of the confidence interval. There’s no need to do it again. Because the formulas are so similar, it turns out that the factors affecting the width of the prediction interval are identical to the factors affecting the width of the confidence interval.

Let’s instead investigate the formula for the prediction interval for

To see how it compares to the formula for the confidence interval for

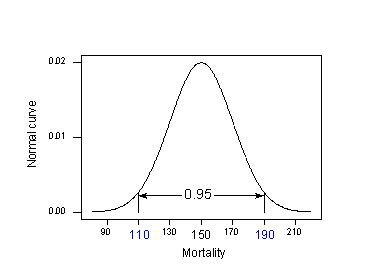
Observe that the only difference in the formulas is that the standard error of the prediction for has an extra term in it that the standard error of the fit for does not.

Let’s try to understand the prediction interval to see what causes the extra term. In doing so, let’s start with an easier problem first. Think about how we could predict a new response at a particular if the mean of the responses at were known. That is, suppose it were known that the mean skin cancer mortality at is 150 deaths per million (with variance 400)? What is the predicted skin cancer mortality in Columbus, Ohio?

Because and are known, we can take advantage of the “**empirical rule**,” which states among other things that 95% of the measurements of normally distributed data are within 2 standard deviations of the mean. That is, it says that 95% of the measurements are in the interval sandwhiched by:

and

Applying the 95% rule to our example with and



95% of the skin cancer mortality rates of locations at 40 degrees north latitude are in the interval sandwhiched by:

and

That is, if someone wanted to know the skin cancer mortality rate for a location at 40 degrees north, our best guess would be somewhere between 110 and 190 deaths per 10 million. The problem is that our calculation used and , population values that we would typically not know. Reality sets in:

* The mean is typically not known. The logical thing to do is estimate it with the predicted response . That is, different samples would yield different predictions , and so we have to take into account this variance of .
* The variance is typically not known. The logical thing to do is to estimate it with the .

Because we have to estimate these unknown quantities, the variation in the prediction of a new response depends on two components:

1. The variation due to estimating the mean with , which we denote “.” (Note that the estimate of this quantity is just the standard error of the fit that appears in the confidence interval formula.)
2. The variation in the response , which we denote “.” (Note that quantity estimated, as usual, with the mean square error .)

Adding the two variance components, we get:

Which is estimated by:

Do you recognize this quantity? It’s just the variance of the prediction that appears in the formula for the prediction interval !

Let’s compare the two intervals again:

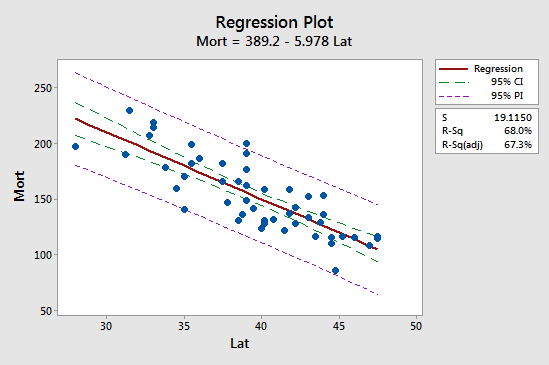
Confidence interval for :

Prediction interval for :

What’s the practical implications of the difference in the two formulas?

* Because the prediction interval has the extra term, a confidence interval for at will always be narrower than the corresponding prediction interval for at .
* By calculating the interval at the sample’s mean of the predictor values and increasing the sample size , the confidence interval’s standard error can approach 0. Because the prediction interval has the extra term, the prediction interval’s standard error cannot get close to 0.

The first implication is seen most easily by studying the following plot for our skin cancer mortality example:

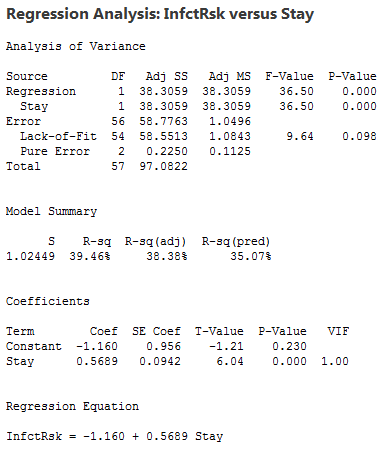


Observe that the prediction interval (in purple) is always wider than the confidence interval (in green). Furthermore, both intervals are narrowest at the mean of the predictor values (about 39.5).

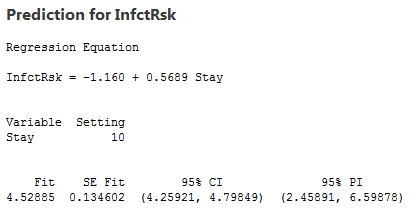
# Further Examples

## Hospital Infection Data

The hospital infection risk data set consists of a sample of hospitals in the east and north central U.S. ([hospitalinfct\_reg1and2.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/hospitalinfct_reg1and2.txt)). The response variable is infection risk (percent of patients who get an infection) and the predictor variables average length of stay (in days). Minitab output for a simple linear regression model fit to these data as follows:



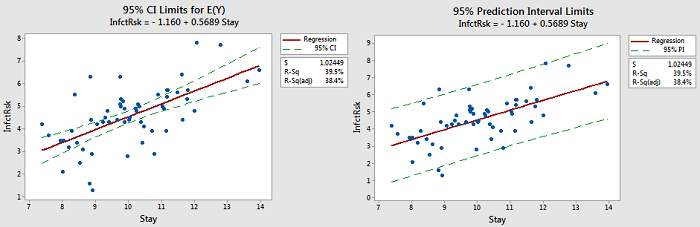
Minitab output with information for .



We can make the following observations:

1. For the interval given under 95% confidence we can estimate that in hospitals in which average length of stay is 10 days, the mean infection risk is between 4.25921 and 4.79849.
2. For the interval given under 95% PI, we say with 95% confidence that for any future hospital where the average length of stay is 10 days, the infection risk is between 2.45891 and 6.59878.
3. The value under Fit is calculated as
4. The value under SE Fit is the standard error of and it measures the accuracy of as an estimate of .
5. Since , the multiplier for 95% confidence is 2.00324. The 95% CI for is calculated as .
6. Since , the 95% PI is calculated as

The following figure provides plots showing the difference between the confidence intervals (CI) and prediction intervals (PI) we have been considering.



There are also some things to note:

1. Notice that the limits for are close to the line. The purpose for those limits is to estimate the “true” location of the line.
2. Notice that the prediction limits (on the right) bracket most of the data. Those limits describe the location of individual -values.
3. Notice that the prediction intervals are wider than the confidence intervals. This is something that can be noted by the formulas.

Quick reference: , , , ,