Chapter 2: SLR Model Evaluation

# Introduction

This lesson presents two alternative methods for testing whether a linear association exists between the predictor and the response in a simple linear regression:

versus

One is the **-test for the slope** while the other is an **analysis of variance (ANOVA) -test**.

As you know, one of the primary goals of this course is to be able to translate a research question into a statistical procedure. Here are two examples of research questions and the alternative statistical procedures that could be used to answer them:

1. Is there a (linear) relationship between skin cancer mortality and latitude?
   * What statistical procedure answers this research question? We could estimate the regression line and then use the -test to determine slope, , of the population regression line is 0.
   * Alternatively, we could perform an (analysis of variance) -test.
2. Is there a (linear) relationship between height and GPA?
   * What statistical procedure answers this research question? We could estimate the regression line and then use the -test to see if the slope, , of the population regression line is 0.
   * Again, we could alternatively perform an (analysis of variance) -test

We also learn a way to check for linearity – the “L” in the “LINE” conditions – **using the linear lack of fit test**. This test requires replicates, that is multiple observations of for at least one (preferably more) values of , and concerns the following hypotheses:

* : There is no lack of linear fit
* : There is lack of linear fit

Learning Objectives & Outcomes

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

* Be able to calculate confidence intervals and conduct hypothesis tests for the population intercept and population slope using Minitab’s regression analysis output
* Be able to draw research conclusions about the population intercept and population slope using the above confidence intervals and hypothesis tests.
* Know the six possible outcomes about the slope whenever we test whether there is a linear relationship between a predictor and a response .
* Understand the “derivation” of the analysis of variance -test for testing . That is, understand how the total variation in a response is broken down into two parts – a component that is due to the predictor and a component that is just due to random error. And, understand how the expected mean squares tell us to use the ratio MSR/MSE to conduct the test.
* Know how each element of the analysis of the variance table is calculated.
* Know what scientific questions can be answered with the analysis of variance -test
* Conduct the analysis of variance -test to test versus .
* Know the similarities and distinctions of the -test and -test for testing
* Know the -test for testing that . The -test for testing , and the -test for testing that yield similar results, but understand when it makes sense to report results of each one.
* Calculate all of the values in the lack of fit analysis of variance table.
* Conduct the -test for lack of fit
* Know that the (linear) lack of fit test only gives you evidence against linearity. If you reject the null, and conclude lack of linear fit, it doesn’t tell you what (non-linear) regression function would work.
* Understand the “derivation” of the linear lack of fit test. That is, understand the decomposition of the error sum of squares, and how the expected mean squares tell us to use the ratio MSLF/MSPE to test for lack of linear fit.

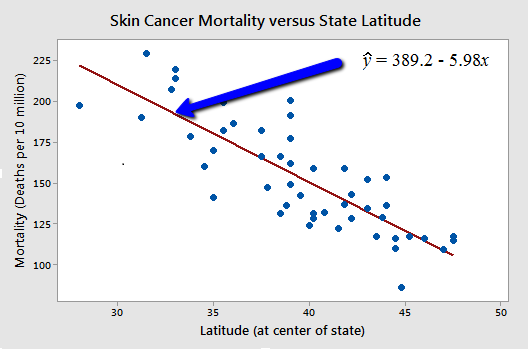
# Inference for the Population Intercept and Slope

Recall that we are ultimately always interested in drawing *conclusions about the population,* not the *particular sample we observed*. In the simple regression setting, we are often interested in learning about the population intercept and the population slope . As you know, confidence intervals and hypothesis tests are two related, but different, ways of learning about the values of population parameters. Here, we will learn how to calculate confidence intervals and conduct hypothesis tests for both and

Let’s revisit the example concerning the relationship between skin cancer mortality and state latitude ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)). The response variable is the mortality rate (number of deaths per 10 million people) of white males due to malignant skin melanoma from 1950 – 1959. The predictor variable is the latitude (degrees North) at the center of each 49 states in the United States. A subset of the data looks like this:

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **State** | **Latitude** | **Mortality** |
| 1 | Alabama | 33.0 | 219 |
| 2 | Arizona | 34.5 | 160 |
| 3 | Arkansas | 35.0 | 170 |
| 4 | California | 37.5 | 182 |
| 5 | Colorado | 39.0 | 149 |
| … | … | … | … |
| 49 | Wyoming | 43.0 | 134 |

And a plot of the data with the estimated regression equation looks like:



IS there a relationship between state latitude and skin cancer mortality? Certainly, since the estimated slope of the line, , not 0, there is a relationship between state latitude and skin cancer mortality in the sample of 49 data points. But, we do not know if there is a relationship between the population of all of the latitudes and skin cancer mortality rates. That is, we want to know if the population slope is also unlikely to be 0.

## -interval for the slope parameter

The formula for the confidence interval for , in words, is:

Sample estimate (-multiplier \* standard error)

And the notation is:

The resulting confidence interval not only gives us a range of values that is likely to contain the true unknown value , it also allows us to answer the research question “Is the predictor linearly related to the response ?” If the confidence interval for contains 0, then we conclude that there is no evidence of a linear relationship between the predictor and the response in the population. On the other hand, if the confidence interval for does not contain 0, then we conclude that there is evidence of a linear relationship between the predictor and the response in the population.

## An -level hypothesis test for the slope parameter

We follow standard hypothesis test procedures in conducting a hypothesis test for the slope : First we specify the null and alternative hypothesis:

Null hypothesis some number .

Alternative hypothesis some number .

The phrase “some number ” means that you can test whether or not the population slope takes on any value, however, we are interested in testing whether is equal to 0, and the alternative hypothesis, is not equal to 0. However, we can test values other than 0 and the alternative hypothesis can also state that is less than (<) some number or greater than some number .

Second, we calculate the value of the test statistic using the following formula:

Third, we use the resulting test statistic to calculate the -value. As always, the -value is the answer to the question “how likely is it that we’d get a test statistic as extreme as we did if the null hypothesis were true?” The -value is determined by referring to a -distribution with degrees of freedom.

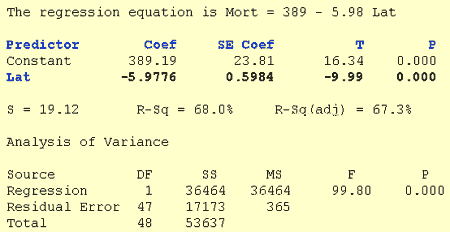
Finally, we make a decision:

* If the -value is smaller than the significance level , we reject the null hypothesis in favor of the alternative. We conclude “there is sufficient evidence at the level to conclude that there is a linear relationship in the population between the predictor and the response .
* If the -value is large than the significance level , we fail to reject the null hypothesis. We conclude “there is not enough evidence at the level to conclude that there is a linear relationship in the population between the predictor and the response .”

## Drawing conclusions about the slope parameter using Minitab

Let’s see how we can use Minitab to calculate confidence intervals and conduct hypothesis tests for the slope . Minitab’s regression analysis output for our skin cancer mortality and latitude example appears below.

The line pertaining to the latitude predictor, **Lat**, in the summary table of predictors has been bolded. It tells us that the estimated slope coefficient , under the column labeled **Coef**, is **-5.9776**. The estimated standard error of , denoted , in the column labeled **SE Coef** for “standard error of the coefficient,” is **0.5984**.



By default, the test statistic is calculated assuming the user wants to test that the slope is 0. Dividing the estimated coefficient -5.9776 by the estimated standard error 0.5984, Minitab reports that the test statistic **T** is **-9.99**.

By default, the -value is calculated assuming the alternative hypothesis is a “two-tailed, not-equal-to” hypothesis. Upon calculating the probability that a -random variablewith degrees of freedom would be larger than 9.99, and multiplying the probability by 2, Minitab reports that is 0.000 (to three decimal places). That is, the -value is less than 0.001. (Note that we multiply the probability by 2 since this is a two-tailed test.)

Because the -value is so small (less than 0.001), we can reject the null hypothesis and conclude that does not equal to 0. There is sufficient evidence, at the level, to conclude that there is a linear relationship in the population between skin cancer mortality and latitude.

**Minitab Note**: The -value in Minitab’s regression analysis output is always calculated assuming the alternative hypothesis is testing the two-sided . If you’re a;ternative hypothesis is the one-tailed or , you have to divide the -value that Minitab reports in the summary table of predictors by 2. (However, be careful if the test statistic is negative for an upper-tailed test or positive for a lower-tailed test, in which case you have to divide by 2 and then subtract from 1. Draw a picture of an appropriately shaded density curve if you are not sure why.)

It’s easy to calculate a 95% confidence interval for using the information in the Minitab output. You just need to use Minitab to find the -multiplier for you. It is .Then , the 95% confidence interval for is or . (Alternatively, Minitab can display the interval directly if you click the “Results” tab in the regression dialogue box, select “Expand Table” and check “Coefficients.”)

We can be 95% confident that the population is between -7.2 and -4.8. That is, we can be 95% confident that for every additional one-degree increase in latitude, the mean skin cancer mortality rate decreases between 4.8 and 7.2 deaths per million people.

## Factors affecting the width of a confidence interval for

Recall that, in general, we want our confidence intervals to be as narrow as possible. If we know what factors affect the length of a confidence interval for the slope , we can control them to ensure that we obtain a narrow interval. The factors can be easily determined by studying the formula for the confidence interval:

First, subtracting the lower endpoint of the interval from the upper endpoint of the interval, we determine that the width of the interval is:

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

* **As the confidence level decreases, the width of the interval decreases**. Therefore, if we decrease our confidence level, we decrease the width of our interval. Clearly, we don’t want to decrease the confidence level too much. Typically, confidence levels are never set below 90%.
* **As MSE decreases, the width of the interval decreases**. The value of the MSE depends on only two factors – how much the responses vary naturally around the estimated regression line, and how well your regression function (line) fits the data. You cannot control the first factor all that much other than to ensure that you are not adding any unnecessary error in your measurement process. Throughout this course, we’ll learn ways to make sure that the regression function fits the data as well as it can.
* **The more spread out the predictor values, the narrower the interval**. The quantity in the denominator summarizes the predictor values. The more spread out the predictor values, the larger the denominator, and hence the narrower the interval. Therefore, we can decrease the width of our interval by ensuring that our predictor values are sufficiently spread out.
* **As the sample size increases, the width of the interval decreases**. The sample size plays a role in two ways. First, recall that the -multiplier depends on the sample size through . Therefore, as the sample size increases, the -multiplier decreases, the length of the interval decreases. Second, the denominator also depends on . The larger the sample size, the more terms you add to this sum, the larger the denominator, the narrower the interval. Therefore, in general, you can ensure that your interval is narrow by having a large enough sample.

## Six possible outcomes concerning the slope

There are six possible outcomes whenever we test whether there is a linear relationship between the predictor and the response , that us, whenever we test the null hypothesis against alternative hypothesis

**When we don’t reject the null hypothesis** , any of the following three realities are possible:

1. We committed a Type II error. That is, in reality and our sample data just didn’t provide enough evidence to conclude that .
2. There really is not much of a linear relationship between and .
3. There is a relationship between and – it is just not linear.

**When we do reject the null hypothesis** in favor of the alternative hypothesis , any of the following three realities are possible:

1. We committed a Type I error. That is, in reality , but we have an unusual sample that suggests that .
2. The relationship between and is indeed linear.
3. A linear function fits the data okay, but a curved (“curvilinear”) function would fit the data better.

## -interval for the intercept parameter

Calculating confidence intervals and conducting hypothesis tests for the intercept parameter is not done as often as it is for the slope parameter . The reason for this becomes clear upon reviewing the meaning of . The intercept parameter is the mean of the responses at . If is meaningless, as it would be, for example, if your predictor variable was height, then is not meaningful. For the sake of completeness, we present the methods here for those situations in which is meaningful.

The formula for the confidence interval , in words, is:

Sample estimate (-multiplier standard error)

And, in notation, is:

The resulting confidence interval gives us a range of values that is likely to contain the true unknown value . The factors affecting the length of a confidence interval for are identical to the factors affecting the length of a confidence interval for .

## An -level hypothesis test for the intercept parameter

Again, we follow standard hypothesis test procedures. First, we specify the null and alternative hypothesis:

Null hypothesis some number

Alternative hypothesis some number

The phrase “some number ” mean that you can test whether or not the population intercept takes on any value. By default, Minitab conducts the hypothesis test for testing whether or not is 0. But, the alternative hypothesis can also state that is less than ( < ) some number or greater than ( > ) some number .

Second, we calculate the value of the test statistic using the following formula:

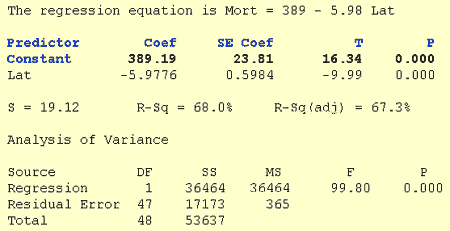
Third, we use the resulting test statistic to calculate the -value. Again, the -value is the answer to the question “how likely is it that we’d get a test statistic as extreme as we did if the null hypothesis were true?” The -value is determined by referring to a -distribution with degrees of freedom.

Finally, we make a decision. If the -value is smaller than the significance level , we reject the null hypothesis in favor of the alternative. If we conduct a “two-tailed, not-equal-to-0” test, we conclude “there is sufficient evidence at the level to conclude that the mean of the responses is not 0 when .” If the -value is larger than the significance level , we fail to reject the null hypothesis.

## Drawing conclusions about the intercept parameter using Minitab

Let’s see how we can use Minitab to calculate confidence intervals and conduct hypothesis tests for the intercept . Minitab’s regression analysis output for our skin cancer mortality and latitude example appears below. The work involved is very similar to that for the slope .

The line pertaining to the intercept, which Minitab always refers to as **Constant**, in the summary table of predictors has been bolded. It tells us that the estimated intercept coefficient , under the column label **Coef**, is **389.19**. The estimated standard error of , denoted , in the column labeled **SE Coef** is **23.81**.



By default, the test statistic is calculated assuming the user wants to test that the mean response is 0 when . Note that there is an ill-advised test here, because the predictor values in the sample do not include a latitude of 0. That is, such a test involves extrapolating outside the scope of the model. Nonetheless, for the sake of illustration, let’s proceed assuming that it is an okay thing to do.

Dividing the estimated coefficient 389.19 by the estimated standard error 23.81, Minitab reports that the test statistic is 16.34. By default, the -value is calculated assuming the alternative hypothesis is a “two-tailed, not-equal-to-0” hypothesis. Upon calculating the probability that a random variable with degrees of freedom would be larger than 16.34, and multiplying the probability by 2., Minitab reports that is 0.000 (to three decimal places). That is, the -value is less than 0.001.

Because the -value is so small (less than 0.001), we can reject the null hypothesis and conclude that does not equal 0 when . There is sufficient evidence, at the level, to conclude that the mean mortality rate at a latitude of 0 degrees North is not 0. (again note that we have to extrapolate in order to arrive at this conclusion, which in general is not advisable.)

Proceed as previously described to calculate 95% confidence interval for . Use Minitab to find the -multiplier for you. Again, it is . Then, the 95% confidence interval for (Alternatively, Minitab can display the interval directly if you click the “Results” tab in the regression dialog box, select “Expanded Table” and check “Coefficients.”) We can be 95% confident that the population intercept is between 341.3 and 437.1. That is, we can be 95% confident that the mean mortality rate at a latitude of 0 degrees North is between 341.3 and 437.1 deaths per 10 million people. (Again, it is probably not a good idea to make this claim because of the severe extrapolation involved.)

## Statistical Inference Conditions

We’ve made no mention yet of the conditions that must be true in order for it to be okay to use the above confidence interval formulas and hypothesis testing procedures for and . In short, the “LINE” assumptions we discussed earlier- linearity, independence, normality, and equal variance – must hold. It is not a big deal if the error terms (and thus responses) are only approximately normal. If you have a large sample, then the error terms can even deviate somewhat from normality.

## Regression Through the Origin

In rare circumstancews it may make sense to consider a simple linear regression model in which the intercept , is assumed to be exactly 0. For example, suppose we have data on the number of items produced per hour along with the number of rejects in each of those time-spans. If we have a period where no items where produced, then there are obviously 0 rejects. Such a situation may indicate deleting from the model since reflects the amount of the response (in this case, the number of rejects) when the predictor is assumed to be 0 (in this case the number of items produced.) Thus, the model to estimate becomes,

Which is called a regression through the origin (or RTO) model. The estimate for when using the regression through the origin model is:

Thus, the estimated regression equation is

Note that we no longer have to center (or “adjust”) the ’s and ’s by their sample means (compare this estimate for to that of the estimate found for the simple linear regression model). Since there is no intercept there is no correction factor and no adjustment for the mean (I.e. the regression line can only pivot about the point ).

Generally, a regression through the origin is not recommended due to the following:

1. Removal of is a strong assumption which forces the line to go through the point . Imposing this restriction does not give ordinary least squares as much flexibility in finding the line of best fit for the data.
2. In a simple linear regression model, . However, in regression through the origin generally . Because of this, the SSE could actually be larger than the SSTO, thus resulting in
3. Since can be negative, the usual interpretation of this value as a measure of the strength of the linear component in the simple linear regression model cannot be used here.

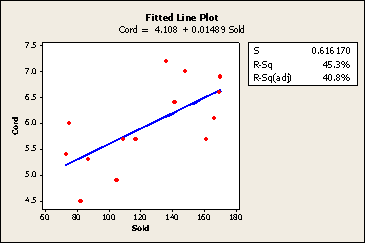
If you strongly believe that a regression through the origin is appropriate for your situation, then the resting ca help justify your decision. Moreover, if data has not been collected near , then forcing the regression line through the origin is likely to make for a worse-fitting model. So again, this model is not usually recommended unless there is a strong belief that it is appropriate.

To fit a regression through the origin model in Minitab click “Options” in the regular regression window and then uncheck the “Fit Intercept” option (v16) or click “Model” and uncheck “Include the constant term in the Model” (v17).

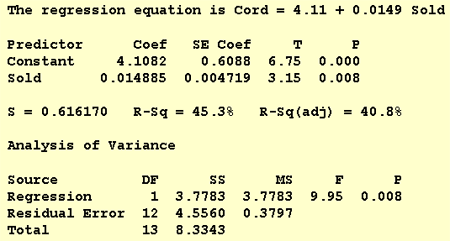
# Another Example of Slope Inference

IS there a positive relationship between sales of leaded gasoline and lead burden in the bodies of newborn infants? Researchers (Rabinowitz, et al, 1984) who were interested in answering this research question compiled data ([leadcord.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/leadcord.txt)) on the monthly gasoline lead sales (in metric tonnes) in Massachusetts and mean lead concentrations (/) in umbilical-cord blood of babies born at a major Boston hospital over 14 months in 1980-1981

Analyzing their data, the researchers obtained the following Minitab fitted line plot:



And the standard regression analysis output:



Minitab reports that the -value for testing against the alternative hypothesis is 0.008. Therefore, since the test statistic is positive, the -value is less than 0.05. There is sufficient evidence at the 0.05 level to conclude that .

Furthermore, since the 95% confident 95% -multiplier is , a 95% confidence interval for is:

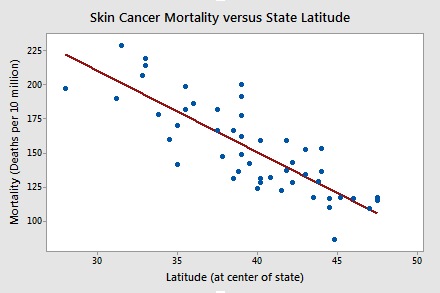
or

The researchers can be 95% confident that the mean lead concentrations in umbilical-cord blood of Massachusetts babies increases between 0.0046 and 0.0252 / for every one-metric ton increase in monthly gasoline lead sales in Massachusetts. It is up to the researchers to debate whether or not this is a meaningful increase.

# Sums of Squares

Let’s return to the skin cancer mortality example ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)) and investigate the research question, “Is there a (linear) relationship between skin cancer mortality and latitude?”

Review the following scatter plot and estimated regression line. What does the plot suggest is the answer to this question? The linear relationship is fairly strong. The estimated slope is negative, not equal to 0.

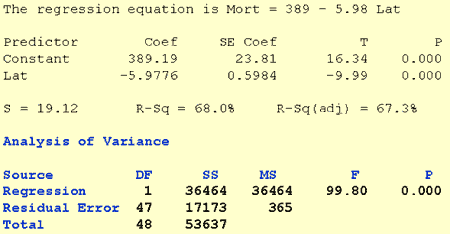


We can answer the research question using the -value of the -test for testing:

* The null hypothesis
* Against the alternative hypothesis

As the Minitab output below suggests, the -value of the -test for “Lat” is less than 0.001. There is enough statistical evidence to conclude that the slope is not 0. That is, that there is a linear relationship between skin cancer mortality and latitude.

There is an alternative method for answering the research question, which uses the analysis of variance -test. Let’s first look at what we are working towards understanding. The (standard) “**analysis of variance**” table for this data set is highlighted in the Minitab output below. There is a column labeled , which contains the -test statistic, and there is a column labeled , which contains the -value associated with the -test. Notice that the -value, 0.000, appears to be the same as the -value, 0.000, for the -test for the slope. The -test similarly tells us that there is enough statistical evidence to conclude that there is a linear relationship between skin cancer mortality and latitude.



Now let’s investigate what all the numbers in the table represent. Let’s start with the column labeled **SS** for “**sums of squares**.” We consider sums of squares in Chapter 1 when we defined the coefficient of determination , but now we consider them again in the context of the analysis of variance table.

The scatter plot of mortality and latitude appears again below, but now it is adorned with three labels:

* denotes the observed mortality for state
* is the estimated regression line (solid line) and therefore denotes the estimated (or “fitted”) mortality for the latitude of state .
* represents what the line would look like if there were no relationship between mortality and latitude. That is, it denotes the “no relationship” line (dashed line). It is simply the average mortality of the sample.

If there is a linear relationship between mortality and latitude, then the estimated regression line should be “far” from the no relationship line. We just need a way of quantifying “far.” The above three elements are useful in quantifying how far the estimated regression line is from the no relationship line. As illustrated by the plot, the two lines are quite far apart.

|  |  |
| --- | --- |
| mortality vs latitude plot |  |

The distance of each observed value from the no regression line is . If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get

Called the “**Total Sum of Squares**.” It quantifies how much the observed responses vary if you don’t take into account their latitude.

The distance of each fitted value from the no regression line is . If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get

Called the “**Regression Sum of Squares**,” it quantifies how far the estimated regression line is from the no relationship line.

This distance of each observed value from the estimated regression line is . If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get:

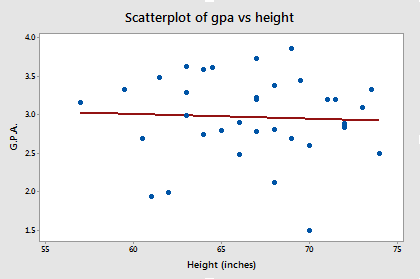
Called the “**Error Sum of Squares**,” as you know, it quantifies how much data points carry around the estimated regression line.

In short, we have illustrated that the total variation in observed mortality (53637) is the sum of two parts – variation “due to” latitude and variation just due to random error (17173). (we are careful to put “due to” in quotes in order to emphasize that a change in latitude does not necessarily cause the change in mortality. All we could conclude is that latitude is “associated with” mortality.)

# Sum of squares (continued…)

Now let’s do a similar analysis to investigate the research question, “Is there a (linear) relationship between height and GPA?” ([heightgpa.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/heightgpa.txt))

Review the following scatterplot and estimated regression line. What does the plot suggest is the answer to the research question? In this case, it appears as if there is almost no relationship whatsoever. The estimated slope is almost 0.

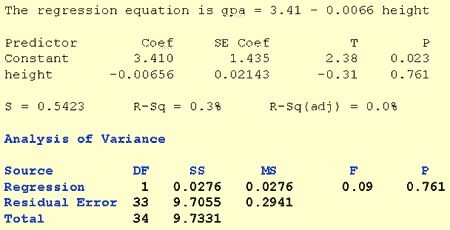


Again, we can answer the research question using the -value of the -test for:

* Testing the null hypothesis
* Against the alternative hypothesis

As the Minitab output below suggests, the -value of the -test for height is 0.761. There is not enough statistical evidence to conclude that the slope is not 0. We conclude that there is no linear relationship between height and GPA.

The Minitab output also shows the analysis of variable table for this data set. Again, the -value associated with the analysis of variance -test, 0.761, appears to be the same as the -value for the -test for the slope, 0.761. The -test similarly tells us that there is insufficient statistical evidence to conclude that there is a linear relationship between height and GPA.



The scatter plot of GPA and height appears below, now adorned with the three labels:

* denotes the observed GPA for student .
* is the estimated regression line (solid line) and therefore denotes the estimated GPA of the sample.
* represents the “no relationship” line (dashed line) between height and GPA. It is simply the average GPA of the sample.

For this data set, note that the estimated regression line and the “no relationship” line are very close together. Let’s see how the sum of squares summarize this point

|  |  |
| --- | --- |
| gpa vs height plot |  |

* The “**Total Sum of Squares**,” which again quantifies how much the observed GPA vary if you don’t take into account height, is .
* The “**Regression Sum of Squares**,” which again quantifies how far the estimated regression line is from the no relationship line, is .
* The “**Error Sum of Squares**,” which again quantifies how much the data points vary around the estimated regression line, is .

In short, we have illustrated that the total variation in the observed GPA’s (9.7331) is the sum of two parts – variation “due to“ height (0.0276) and variation due to random error (9.7055). Unlike the last example, most of the variation in the observed GPA’s is just due to random error. It appears as if very little of the variation can be attributed to the predictor heights.

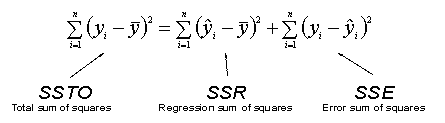
# Analysis of Variance: The Basic Idea

* Break down the total variation in (“**Total Sum of Squares**”) into two components:
  1. A component that is “due to” the change in (“**Regression Sum of Squares**”)
  2. A component that is just due to random error (“**Error Sum of Squares**”)
* If the Regression Sum of Squares is a “large” component of the Total Sum of Squares, it suggests that there *is* a linear association between the predictor and the response .

Here is a simple picture illustrating how the distance is decomposed into the sum of two distances and :

|  |  |  |
| --- | --- | --- |
|  |  |  |

Although the derivation isn’t as simple as it seems, the decomposition holds for the sum of the squared distances too:



The degrees of freedom associated with each of these sums of squares follow a similar decomposition

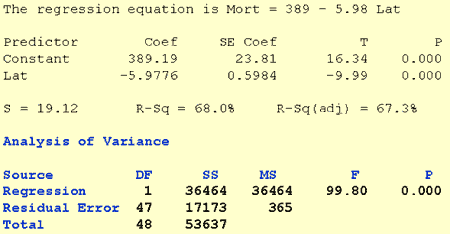
* You might recognize SSTO as being the numerator of the sample variance. Recall that the denominator of the sample variance is . Therefore is the degrees of freedom associated with SSTO.
* Recall that the mean squared error *MSE* is obtained by dividing *SSE* by . Therefore, is the degrees of freedom associated with *SSE*.

Then we obtain the following breakdown of the degrees of freedom:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | = |  | + |  |
| Degrees of freedom associated with SSTO |  | Degrees of freedom associated with SSR |  | Degrees of freedom associated with SSE |

# The Analysis of Variance (ANOVA) table and the -test

We’ve covered quite a bit of ground. Let’s review the analysis of variance table for the example concerning skin cancer mortality and latitude ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)).



Recall that there were 49 states in the data set.

* The degrees of freedom associated with will always be 1 for the simple linear regression model. The degrees of freedom associated with is . The degrees of freedom associated with is . And the degrees of freedom add up: .
* The sums of squares add up: . That is, here: 53637 = 36464 + 17173.

Let’s tackle a few more columns of the analysis of variance table, namely the “**mean square**” column, labeled , and the -statistic column, labeled .

## Definitions of mean squares

We already know the “**mean square error (MSE)**” is defined as:

That is, we obtain the mean squared error by dividing the error sum of squares by its associated degrees of freedom . Similarly, we obtain the “**regression mean square (MSR)**” by dividing the regression sum of squares by its degrees of freedom 1:

Of course, that means the regression sum of squares (SSR) and the regression mean squares (MSR) are always identical for the simple linear regression model

Now, why do we care about mean squares? Because their expected values suggest how to test the null hypothesis against the alternative hypothesis .

## Expected Mean Squares

Imagine taking many, many random samples of size from some population, and estimating the regression line and determining MSR amd MSE for each data set obtained. It has been shown that the average (that is, the expected value) of all of the MSRs you can obtain equals:

Similarly, it has been shown that the average (that is, the expected value) of all of the MSEs you can obtain equals:

These expected values suggest how to test versus

* If then we’d expect the ratio to equal 1.
* If then we’d expect the ratio to be greater than 1.

These two facts suggest that we should use the ratio, to determine whether or not .

Note that because is squared in , we cannot use the ratio .

* To test versus
* Or to test versus

We can only use to test versus

We have now completed our investigation of all of the entries of a standard analysis variance table. The formula for each entry is summarized for you in the following analysis of variance table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source of Variation | DF | SS | MS | F |
| Regression | 1 |  |  |  |
| Residual Error |  |  |  |  |
| Total |  |  |  |  |

However, we will always let Minitab do the dirty work of calculating the values for us. Why is the ratio labeled in the analysis of the variance table? That’s because the ratio is known to follow an distribution with 1 numerator degree of freedom and denominator degrees of freedom. For this reason, it is often referred to as the analysis of variance -test. The following section summarizes the formal -test.

## The Formal -test for the slope parameter

The null hypothesis is

The alternative hypothesis is

The test statistics is

As always, the -value is obtained by answering the question: “What is the probability that we’d get an statistic as large as we did, if the null hypothesis is true?”

The -value is determined by comparing to an distribution with 1 numerator degree of freedom and denominator degrees of freedom.

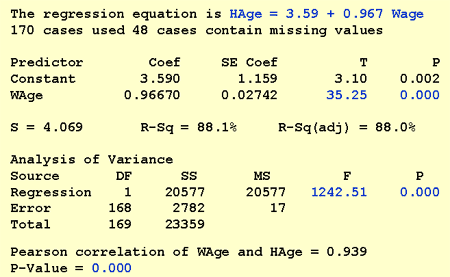
In reality, we are going to let Minitab calculate the statistic and the -value for us. Let’s try it out on a new example!

# Example: Are Men Getting Faster?

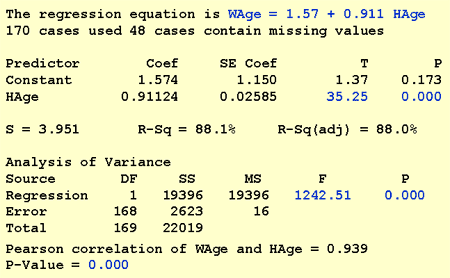
Obtained text from file on work computer

# Equivalent Linear Relationship Tests

It should be noted that the three hypothesis tests we have learned for testing the existence of a linear relationship – the -test for , the ANOVA -test for , and the -test for – will always yield the same results. For example, when evaluating whether or not a linear relationship exists between a husband’s age and his wife’s age, if we treat husband’s age (“HAge”) as the response and wife’s age (“Wage”) as the predictor, each test yields a -value of 0.000… < 0.001 ([husbandwife.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/husbandwife.txt)):



And similarly, if we treat wife’s age (“Wage”) as the response and husband’s age (“HAge”) as the predictor, each test yields of -value of 0.000… < 0.001:



Technically, then, it doesn’t matter what test you use to obtain the -value. You will always get the same -value. But, you should report the results of the test that make sense for your particular situation:

* If one of the variables can be clearly identified as the response, report that you conducted a -test of -test results for testing . (Does it make sense to use to predict ?)
* If it is not obvious which variable is the response, report that you conducted a -test for testing . (Does it only make sense to look for an association between and ?)

# Notation for the Lack of Fit Test

To conclude this lesson, we’ll digress slightly to consider the lack of fit test for linearity – the “L” in the “LINE” conditions. The reason we consider this here is that, like the ANOVA test earlier, this test is an -test based on decomposing sums of squares.

However, before we “derive” the lack of fit -test, it is important to note that the test requires repeat observations – called “replicates” – for at least one of the values of the predictor . That is, if each value in the data set is unique, then the lack of fit test can’t be conducted on the data set. Even when we do have replicates, we typically need quite a few for the test to have any power. As such, this test generally only applies to specific types of data sets with plenty of replicates.

As is often the case before we learn a new hypothesis test, we have to get some new notation under our belt. In doing so, we’ll look at some (contrived) data that purports to describe the relationship between the size of the minimum deposit required when opening a new checking account at a bank () and the number of new accounts at the bank () ([newaccounts.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/newaccounts.txt)). Suppose the trend in the data looks curved, but we fit a line through the data nonetheless:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Look at the above graphs for specific values (75, 100,125, 150, 175, 200) on the -axis, you will see the standard notation used for the lack of fit -test. Let’s take the case where dollars:

* denotes the first measurement (28) made at the first -value () in the data set
* denotes the second measurement (42) made at the first -value () in the data set
* denotes the average (35) of all of the values at the first -value (
* denotes the predicted response (87.5) for the first measurement made at the first -value ()
* denotes the predicted response (87.5) for the second measurement made at the first -value ()

You should now understand the notation that corresponds with the other values (100, 125, and so on). In general:

* denotes the th measurement made at the th -value in the data set
* denotes the average of all the values at the th -value
* denotes the predicted response for the th measurement made at the th -value

# Decomposing the Error

If you think about it, there are two different explanations for why our data points might not fall right on the estimated regression line. One possibility is that our regression model doesn’t describe the trend in the data well enough. That is, the model may exhibit “**lack of fit**.” The second possibility is that, as is often the case, there is just random variation in the data. This realization suggests that we should decompose the error into two components – one part due to lack of fit of the model and the second part just due to random error. If most of the error is due to lack of fit, and not just random error, it suggests that we should scrap our model and try a different one.

## An Example

Let’s try decomposing the error in the checking account example, (newaccounts.txt). Recall that the prediction error for any data point is the distance of the observed response from the predicted response, i.e., . (Can you identify these distances on the plot of the data below?) To quantify the total error of prediction, we determine this distance for each data point, square the distance, and add up all of the distances to get:

Not surprisingly, this quantity is called the “**error sum of squares**” and is denoted . The error sum of squares for our checking account example is .

If a line fits the data well, then the average of the observed responses at each -value should be close to the predicted response for that -value. Therefore, to determine how much of the total error is due to lack of model fit, we determine how far the average observed response at each -value is from the predicted response of each data point. That is, we calculate the distance . To quantify the total lack of fit, we determine this distance for each data point, square the distance, and add up all of the distances to get:

Not surprisingly, this quantity is called the “**lack of fit sum of squares**” and is denoted . The lack of fit sum of squares for our checking account example is .

To determine how much of the total error is due to just random error, we determine how far each observed response is from the average observed response at each -value. That is, we calculate the distance . To quantify the total pure error, we determine this distance for each data point, square the distance, and add up all of the distances to get:

Not surprisingly, this quantity is called the “**pure error sum of squares**” and is denoted . The pure error sum of squares for our checking account example is .

|  |  |
| --- | --- |
| new accounts vs size of minimum desposit plot |  |

In summary, we’ve shown in this checking account example that most of the error () is attributed to the lack of a linear fit () and not just to random error ().

## Another Example

Let’s see how our decomposition of the error works with a different example – one in which a line fits the data well. Suppose the relationship between the size of the minimum deposit required when opening a new checking account at a bank () and the number of new accounts at the bank () instead looks like this:

|  |  |
| --- | --- |
| new accounts vs size of minimum desposit plot |  |

In this case, as we would expect based on the plot, very little of the total error () is due to lack of a linear fit (). Most of the error appears to be due to just random variation in the number of checking accounts ().

## In Summary

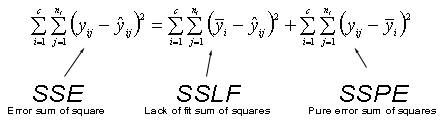
The basic idea behind decomposing the total error is:

* We break down the residual error (“**error sum of squares**” – denoted ) into two components:
  + A component that is due to lack of model fit (“**lack of fit sum of squares**” – denoted )
  + A component that is due to pure random error (“**pure error sum of squares**” – denoted )
* If the lack of fit sum of squares is a large component of the residual error, it suggests that a linear function is inadequate.

Here is a simple picture illustrating how distance is decomposed into the sum of two distances and . You can see the geometric representation of the three components of the equation below:

|  |  |  |
| --- | --- | --- |
|  |  |  |

Although the derivation isn’t as simple as it seems, the decomposition holds for the sum of the squared distances as well ():



The degrees of freedom associated with each of these sums of squares follow a similar decomposition.

* As before, the degrees of freedom associated with is . (The 2 comes from the fact that you estimate 2 parameters – the slope and the intercept – whenever you fit a line to a set of data.)
* The degrees of freedom associated with is , where denotes the number *distinct*  values you have.
* The degrees of freedom associated with is , where denotes the number of *distinct*  values that you have.

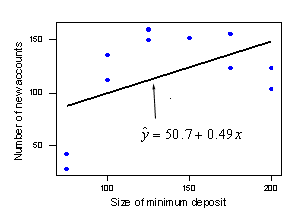
You might notice that the degrees of freedom breakdown as:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | = |  | + |  |
| Degrees of freedom associated with |  | Degrees of freedom associated with |  | Degrees of freedom associated with |

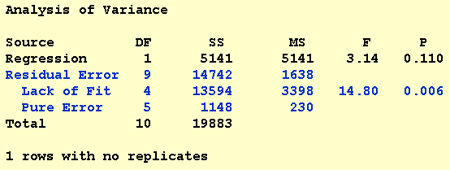
Where again denotes the number of *distinct*  values you have.

# The Lack of Fit F-Test

We’re almost there! We just need to determine an objective way of deciding when too much error in our prediction is due to lack of model fit. That’s where the lack of fit -test comes into play. Let’s return to the first checking account example, ([newaccounts.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/newaccounts.txt)):



Jumping ahead to the punchline, here’s Minitab’s output for the lack of fit -test for this data set:



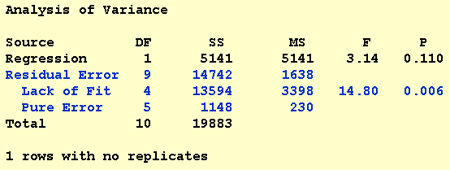
As you can see, the lack of fit output appears as a portion of the analysis of variance table. In the **Sum of Squares** (“**SS**”) column, we see – as we previously calculated – that and sum to . We also see in the Degrees of Freedom (“DF”) column that – since there are data points and distinct values (75, 100, 125, 150, 175, and 200) – the lack of fit degrees of freedom .

Just as is done for the sums of squares in the basic analysis of variance table, the lack of fit sum of squares and the error sum of squares are used to calculate “mean squares.” They are even calculated similarly, namely, by dividing the sum of squares by its associated degrees of freedom. Here are the formal definitions of the mean squares:

The “**lack of fit mean square**” is

The “**pure error mean square**” is

In the Mean Squares (“MS”) column, we see that the lack of fit mean square is 13594 divided by 4, or 3398. The pure error mean square is 1148 divided by 5, or 230:



You might notice that the lack of fit -statistic is calculated by dividing the lack of fit mean square () by the pure error mean square () to get 14.80. How do we know that this -statistic helps us in testing the hypotheses:

* : The relationship assumed in the model is reasonable, i.e. there is no lack of fit.
* : The relationship assumed in the model is not reasonable, i.e., there is lack of fit.

The answer lies in the “**expected mean squares**.” In our sample of newly opened checking accounts, we obtained . If we had taken a different random sample of size , we would have obtained a different value for . Theory tells us that the average of all of the possible values we could obtain is:

That is, we should expect , on average, to equal the above quantity - plus another messy-looking term. Think about that messy term. If the null hypothesis is true, i.e., if the relationship between the predictor and the response is linear, then equals and the messy term becomes 0 and goes away. That is, if there is no lack of fit, we should expect the lack of fit mean square to equal .

What should we expect to equal? Theory tells us it should, on average, always equal to :

Aha – there we go! The logic behind the calculation of the -statistic is now clear:

* If there is a linear relationship between and , then . That is, there is no lack of fit in the simple linear regression model. We would expect to be close to 1.
* If there is not a linear relationship between and , then . That is, there is lack of fit in the simple linear regression model. We would expect the ratio to be large, i.e., a value greater than 1.

So, to conduct the lack of fit test, we calculate the value of the -statistic:

And determine if it is large. To decide if it is large, we compare the -statistic to an -distribution with numerator degrees of freedom and denominator degrees of freedom.

## In Summary

We follow standard hypothesis test procedures in conducting the lack of fit -test. First, we specify the null and alternative hypotheses:

* : The relationship assumed in the model is reasonable, i.e. there is no lack of fit in the model .
* : The relationship assumed in the model is not reasonable, i.e., there is lack of fit in the model .

Second, we calculate the value of the -statistic:

To do so, we complete the analysis of variance table using the following formulas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source of Variation** | **DF** | **SS** | **MS** | **F** |
| Regression |  |  |  |  |
| Residual Error |  |  |  |  |
| Lack of Fit |  |  |  |  |
| Pure Error |  |  |  |  |
| Total |  |  |  |  |

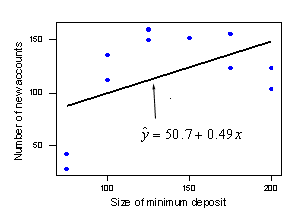
In reality, we let statistical software, such as Minitab, determine the analysis of variance table for us.

Thirdly, we use the resulting -statistic to calculate the -value. As always, the -value is the answer to the question “how likely is it that we’d get an -statistic as extreme as we did if the null hypothesis were true?” The -value is determined by referring to an -distribution with numerator degrees of freedom and denominator degrees of freedom.

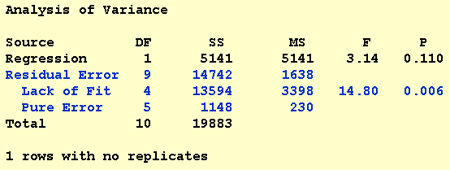
Finally, we make a decision:

* If the -value is smaller than the significance level , we reject the null hypothesis in favor of the alternative. We conclude “there is sufficient evidence at the level to conclude that there is lack of fit *in the simple linear regression model*.”
* If the -value is larger than the significance level , we fail to reject the null hypothesis. We conclude “there is not enough evidence at the to conclude that there is lack of fit *in the simple linear regression model*.”

For our checking account example:



In which we obtain:



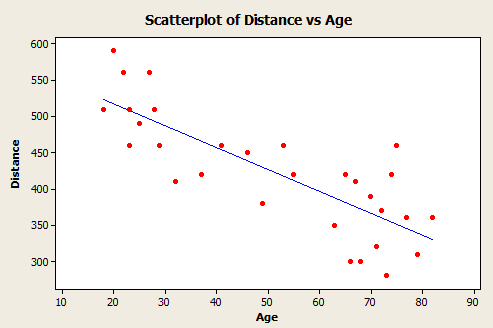
the -statistic is 14.80 and the -value is 0.006. The -value is smaller than the significance level – we reject the null hypothesis in favor of the alternative. There is sufficient evidence at the level to conclude that there is lack of fit in the simple linear regression model. In light of the scatterplot, the lack of fit test provides the answer we expected.

# Further Examples

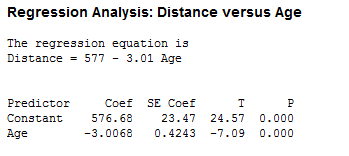
## Example 1: Highway Sign Reading Distance and Driver Age

The data are observations on driver age and the maximum distance (feet) at which individuals can read a highway sign ([signdist.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/examples/signdist.txt)). (Data source: Mind on Statistics, 3rd edition, Utts and Heckard).

The plot below gives a scatterplot of the highway sign along with the least squares regression line.



Here is the accompanying Minitab output, which is found by performing Stat >> Regression >> Regression on the highway sign data.



Hypothesis Test for the Intercept ()

This test is rarely a test of interest, but does show up when one is interested in performing a regression through the origin (which we touched on earlier in this lesson). In the Minitab output above, the row labeled Constant gives the information used to make inferences about the intercept. The null and alternative hypothesis for a hypothesis test about the intercept are written as:

In other words, the null hypothesis is testing if the population intercept is equal to 0 versus the alternative hypothesis that the population intercept is not equal to 0. In most problems, we are not particularly interested in hypotheses about the intercept. For instance, in our example, the intercept is the mean distance when the age is 0, a meaningless age. Also, the intercept does not give information about the value of changes when the value of changes. Nevertheless, to test whether the population intercept is 0, the information from the Minitab output is used as follows:

1. The sample intercept is , the value under **Coef**.
2. The standard error (SE) of the sample intercept, written as , is , the value under SE Coef. The SE of any statistic is a measure of its accuracy. In this case, the SE of gives, very roughly, the average difference between the sample and the true population intercept , for random samples of this size (and with these -values).
3. The test statistics is , the value under .
4. The -value for the test is and is given under . The -value is actually very small and *not* exactly 0.
5. The decision rule at the 0.05 significance level is to reject the null hypothesis since our . Thus, we conclude that there is statistically significant evidence that the population intercept is not equal to 0.

So how exactly is the -value found? For simple regression, the -value is determined using a distribution with degrees (), which is written as , and is calc

Quick reference , , , , ,