STAT 462

Applied Regression Analysis

Lesson 3: SLR Evaluation

Overview of this Lesson

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

$$H_0: \beta_1 = 0 \text{ versus } H_A: \beta_1 \neq 0.$$

One is the ***t*-test for the slope** while the other is an **analysis of variance (ANOVA) *F*-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 t-test to determine if the slope, β_1, of the population regression line is 0.
   - Alternatively, we could perform an (analysis of variance) F-test.
2. Is there a (linear) relationship between height and grade point average?
   - What statistical procedure answers this research question? We could estimate the regression line and then use the t-test to see if the slope, β_1, of the population regression line is 0.
   - Again, we could alternatively perform an (analysis of variance) F-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 y for at least one (preferably more) values of x, and concerns the following hypotheses:

- H_0: There is no lack of linear fit.
- H_A: There is lack of linear fit.

Key Learning Goals for this Lesson:
▪ Be able to calculate confidence intervals and conduct hypothesis tests for the population intercept β_0 and population slope β_1 using statistical software. ▪ Be able to draw research conclusions about the population intercept β_0 and population slope β_1 using the above confidence intervals and hypothesis tests.

- Know the six possible outcomes about the slope β_1 whenever we test whether there is a linear relationship between a predictor x and a response y.
- Understand the "derivation" of the analysis of variance F-test for testing H_0: $\beta_1 = 0$. That is, understand how the total variation in a response y is broken down into two parts — a component that is due to the predictor x 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 variance table is calculated.
- Know what scientific questions can be answered with the analysis of variance F-test.
- Conduct the analysis of variance F-test to test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$.
- Know the similarities and distinctions of the t-test and F-test for testing H_0: $\beta_1 = 0$.
- Know the t-test for testing that $\beta_1 = 0$ and the F-test for testing that $\beta_1 = 0$ yield similar results, but understand when it makes sense to report the results of each one.
- Calculate all of the values in the lack of fit analysis of variance table.
- Conduct the F-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.

3.1 - Inference for the Population Intercept and Slope › (/stat462/node/102)

STAT 462

Applied Regression Analysis

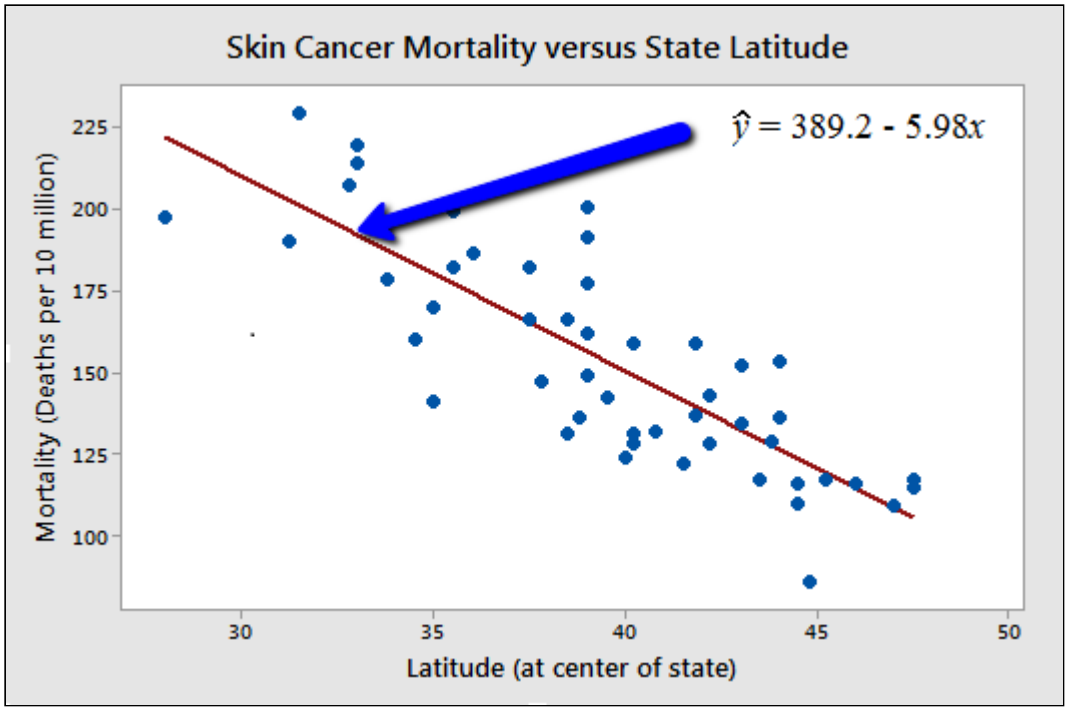
3.1 - Inference for the Population Intercept and Slope

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 β_0 and the population slope β_1. 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 β_0 and β_1.

Let's revisit the example concerning the relationship between skin cancer mortality and state latitude (skincancer.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt)). The response variable y 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 x is the latitude (degrees North) at the center of each of 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, $b1$, is -5.98, not 0, there is a relationship between state latitude and skin cancer mortality *in the sample* of 49 data points. But, we want to 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 β_1 is unlikely to be 0.

An α-level hypothesis test for the slope parameter β_1

We follow standard hypothesis test procedures in conducting a hypothesis test for the slope β_1. **First**, we specify the null and alternative hypotheses:

Null hypothesis $H_0 : \beta_1 =$ some number β
Alternative hypothesis $H_A : \beta_1 \neq$ some number β

The phrase "some number β" means that you can test whether or not the population slope takes on any value. Most often, however, we are interested in testing whether β_1 is 0. By default, statistical software conducts the hypothesis test with null hypothesis, β_1 is equal to 0, and alternative hypothesis, β_1 is not equal to 0. However, we can test values other than 0 and the alternative hypothesis can also state that β_1 is less than (<) some number β or greater than (>) some number β.

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

$$t^* = \frac{b_1 - \beta}{\left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)} = \frac{b_1 - \beta}{se(b_1)}$$

Third, we use the resulting test statistic to calculate the *P*-value. As always, the *P*-value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if the null hypothesis were true?" The *P*-value is determined by referring to a *t*-distribution with n-2 degrees of freedom.

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

Drawing conclusions about the slope parameter β_1 using statistical software

Let's see how we can use statistical software to calculate conduct a hypothesis test for the slope β_1. Minitab's regression analysis output for our skin cancer mortality and latitude example appears below. The output for other software will be similar.

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 b_1, under the column labeled **Coef**, is **-5.9776**. The estimated standard error of b_1, denoted $se(b_1)$, in the column labeled **SE Coef** for "standard error of the coefficient," is **0.5984**.

```
The regression equation is Mort = 389 - 5.98 Lat

Predictor        Coef     SE Coef          T        P
Constant       389.19       23.81      16.34    0.000
Lat           -5.9776      0.5984      -9.99    0.000

S = 19.12       R-Sq = 68.0%     R-Sq(adj) = 67.3%

Analysis of Variance

Source          DF      SS      MS        F        P
Regression       1   36464   36464    99.80    0.000
Residual Error  47   17173     365
Total           48   53637
```

By default, the test statistic is calculated assuming the user wants to test that the slope is 0. Divide the estimated coefficient -5.9776 by the estimated standard error 0.5984 to obtain a test statistic **T** = **-9.99**.

By default, the *P*-value is calculated assuming the alternative hypothesis is a "two-tailed, not-equal-to" hypothesis. Calculate the probability that a *t*-random variable with $n–2 = 47$ degrees of freedom would be larger than 9.99 to get the probability in the upper tail. Then multiply this probability by 2 to get the two-tailed *P*-value, **P** = **0.000** (to three decimal places). In other words, the *P*-value is less than 0.001. (Note we multiply the probability by 2 since this is a two-tailed test. See this video for the reason why.)

Because the *P*-value is so small (less than 0.001), we can reject the null hypothesis and conclude that β_1 does not equal 0. There is sufficient evidence, at the $\alpha = 0.05$ level, to conclude that there is a linear relationship in the population between skin cancer mortality and latitude.

Software Note. The *P*-value in statistical software regression analysis output is always calculated assuming the alternative hypothesis is testing the two-tailed $\beta_1 \neq 0$. If your alternative hypothesis is the one-tailed $\beta_1 < 0$ or $\beta_1 > 0$, you have to divide the *P*-value that the software 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-tail 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're not sure why.)

Six possible outcomes concerning slope β_1

There are six possible outcomes whenever we test whether there is a linear relationship between the predictor *x* and the response *y*, that is, whenever we test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_A : \beta_1 \neq 0$.

When we don't reject the null hypothesis $H_0 : \beta_1 = 0$, any of the following three realities are possible:

1. We committed a Type II error. That is, in reality $\beta_1 \neq 0$ and our sample data just didn't provide enough evidence to conclude that $\beta_1 \neq 0$.
2. There really is not much of a linear relationship between *x* and *y*.
3. There is a relationship between *x* and *y* — it is just not linear.

When we do reject the null hypothesis $H_0 : \beta_1 = 0$ in favor of the alternative hypothesis $H_A : \beta_1 \neq 0$, any of the following three realities are possible:

1. We committed a Type I error. That is, in reality $\beta_1 = 0$, but we have an unusual sample that suggests that $\beta_1 \neq 0$.
2. The relationship between *x* and *y* is indeed linear.
3. A linear function fits the data okay, but a curved ("curvilinear") function would fit the data even better.

(1-α)100% *t*-interval for the slope parameter β_1

e interval for β_1, in words, is:

Sample estimate ± (t-multiplier × standard error)

and, in notation, is:

$$b_1 \pm t_{(\alpha/2,n-2)} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i-\bar{x})^2}} \right)$$

The resulting confidence interval not only gives us a range of values that is likely to contain the true unknown value β_1. It also allows us to answer the research question "is the predictor *x* linearly related to the response *y*?" If the confidence interval for β_1 contains 0, then we conclude that there is no evidence of a linear relationship between the predictor *x* and the response *y* in the population. On the other hand, if the confidence interval for β_1 does not contain 0, then we conclude that there is evidence of a linear relationship between the predictor *x* and the response *y* in the population.

It's easy to calculate a 95% confidence interval for β_1 using the information in the software output. You just need to find the *t*-multiplier, either from a table or by using statistical software. In this case it is *t*(0.025, 47) = 2.0117. Then, the 95% confidence interval for β_1 is -5.9776 ± 2.0117(0.5984) or (-7.2, -4.8). [Alternatively, some statistical software will display the interval directly.]

We can be 95% confident that the population slope 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 10 million people.

Factors affecting the width of a confidence interval for β_1

Recall that, in general, we want our confidence intervals to be as narrow as possible to be the most informative. If we know what factors affect the length of a confidence interval for the slope β_1, 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:

$$b_1 \pm t_{\alpha/2,n-2} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i-\bar{x})^2}} \right)$$

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

$$\text{Width } = 2 \times t_{\alpha/2,n-2} \times \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i-\bar{x})^2}} \right)$$

So, how can we affect the width of our resulting interval for β_1?

- **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 *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. Clearly, you can't 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 ion function fits the data as well as it can.

- **The more spread out the predictor *x* values, the narrower the interval.** The quantity $\sum(x_i - \bar{x})^2$ in the denominator summarizes the spread of the predictor *x* values. The more spread out the predictor values, the larger the denominator, and hence the narrower the interval. Therefore, we can decrease the width 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 *t*-multiplier depends on the sample size through *n*-2. Therefore, as the sample size increases, the *t*-multiplier decreases, the length of the interval decreases. Second, the denominator $\sum(x_i - \bar{x})^2$ also depends on *n*. 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.

An α-level hypothesis test for intercept parameter β_0

Conducting hypothesis tests and calculating confidence intervals for the intercept parameter β_0 is not done as often as it is for the slope parameter β_1. The reason for this becomes clear upon reviewing the meaning of β_0. The intercept parameter β_0 is the mean of the responses at $x = 0$. If $x = 0$ is meaningless, as it would be, for example, if your predictor variable was height, then β_0 is not meaningful. For the sake of completeness, we present the methods here for those rare situations in which β_0 is meaningful.

To conduct a hypothesis test for the intercept parameter β_0 we again follow standard hypothesis test procedures. **First**, we specify the null and alternative hypotheses:

Null hypothesis H_0 : β_0 = some number β
Alternative hypothesis H_A: $\beta_0 \neq$ some number β

The phrase "some number β" means that you can test whether or not the population intercept takes on any value. By default, statistical software conducts the hypothesis test for testing whether or not β_0 is 0. But, the alternative hypothesis can also state that β_0 is less than (<) some number β or greater than (>) some number β.

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

$$t^* = \frac{b_0 - \beta}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}} = \frac{b_0 - \beta}{se(b_0)}$$

Third, we use the resulting test statistic to calculate the *P*-value. Again, the *P*-value is the answer to the question "how likely is it that we'd get a test statistic *t** as extreme as we did if the null hypothesis were true?" The *P*-value is determined by referring to a *t*-distribution with *n*-2 degrees of freedom.

Finally, we make a decision. If the *P*-value is smaller than the significance level α, 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 α level to conclude that the mean of the responses is not 0 when $x = 0$." If the *P*-value is larger than the significance level α, we fail to reject the null hypothesis.

Drawing conclusions about intercept parameter β_0 using statistical software

Let's see how we can use statistical software to conduct a hypothesis test for the intercept β_0. Statistical software regression analysis output for our skin cancer mortality and latitude example appears below. The work involved is e β_1.

The line pertaining to the intercept is labeled **Constant** (but in other software may be labeled **Intercept**). It tells us that the estimated intercept coefficient b_0, under the column labeled **Coef**, is **389.19**. The estimated standard error of b_0, denoted $se(b_0)$, in the column labeled **SE Coef** is **23.81**.

```
The regression equation is Mort = 389 - 5.98 Lat

Predictor        Coef     SE Coef          T        P
Constant       389.19       23.81      16.34    0.000
Lat           -5.9776      0.5984      -9.99    0.000

S = 19.12       R-Sq = 68.0%     R-Sq(adj) = 67.3%

Analysis of Variance

Source          DF      SS      MS        F        P
Regression       1   36464   36464    99.80    0.000
Residual Error  47   17173     365
Total           48   53637
```

By default, the test statistic is calculated assuming the user wants to test that the mean response is 0 when $x = 0$. Note that this 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, the test statistic **T** is **16.34**. By default, the *P*-value is calculated assuming the alternative hypothesis is a "two-tailed, not-equal-to-0" hypothesis. The probability that a t random variable with n-2 = 47 degrees of freedom would be larger than 16.34, and multiplying that probability by 2, the *P*-value is **P** = **0.000** (to three decimal places). That is, the *P*-value is less than 0.001.

Because the *P*-value is so small (less than 0.001), we can reject the null hypothesis and conclude that β_0 does not equal 0 when $x = 0$. There is sufficient evidence, at the $\alpha = 0.05$ 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 beyond the scope of the model in order to arrive at this conclusion, which in general is not advisable.)

(1-α)100% *t*-interval for intercept parameter β_0

The formula for the confidence interval for β_0, in words, is:

Sample estimate ± (t-multiplier × standard error)

and, in notation, is:

$$b_0 \pm t_{\alpha/2,n-2} \times \sqrt{MSE}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}}$$

The resulting confidence interval gives us a range of values that is likely to contain the true unknown value β_0. The factors affecting the length of a confidence interval for β_0 are identical to the factors affecting the length of a confidence interval for β_1.

Proceed as previously described to calculate a 95% confidence interval for β_0. Find the *t*-multiplier using a table or statistical software. Again, it is $t(0.025, 47) = 2.0117$. Then, the 95% confidence interval for β_0 is $389.19 \pm 2.0117(23.81) = (341.3, 437.1)$. [Alternatively, if possible, use statistical software to display the interval directly.] 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 hypothesis testing procedures and confidence interval formulas for β_0 and β_1. 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 far from normality.

Regression through the origin

In rare circumstances it may make sense to consider a simple linear regression model in which the intercept, β_0, 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 were produced, then there are obviously 0 rejects. Such a situation may indicate deleting β_0 from the model since β_0 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

$$y_i = \beta_1 x_i + \epsilon_i,$$

which is called a **regression through the origin** (or **RTO**) model. The estimate for β_1 when using the regression through the origin model is:

$$\hat{\beta}_{\text{RTO}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Thus, the estimated regression equation is

$$\hat{y}_i = \hat{\beta}_{\text{RTO}} x_i.$$

Note that this formula no longer centers (or "adjusts") the x_i's and y_i's by their sample means (compare this estimate for $\hat{\beta}_1$ 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 (0,0)).

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

1. Removal of β_0 is a strong assumption which forces the line to go through the point (0,0). 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, $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$. However, in regression through the origin, generally $\sum_{i=1}^{n} e_i \neq 0$. Because of this, the SSE could actually be larger than the SSTO, thus resulting in $r^2 < 0$.
3. Since r^2 can be negative, the usual interpretation of this value as a measure of the strength of the linear relationship through the origin model cannot be used here.

If you strongly believe that a regression through the origin model is appropriate for your situation, then statistical testing can help justify your decision. However, if no data has been collected near $x = 0$, 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.

Statistical software generally includes an option to fit a "regression through the origin" model.

‹ Lesson 3: SLR Evaluation (/stat462/node/80) up (/stat462/node/80) 3.2 - Another Example of Slope Inference › (/stat462/node/103)

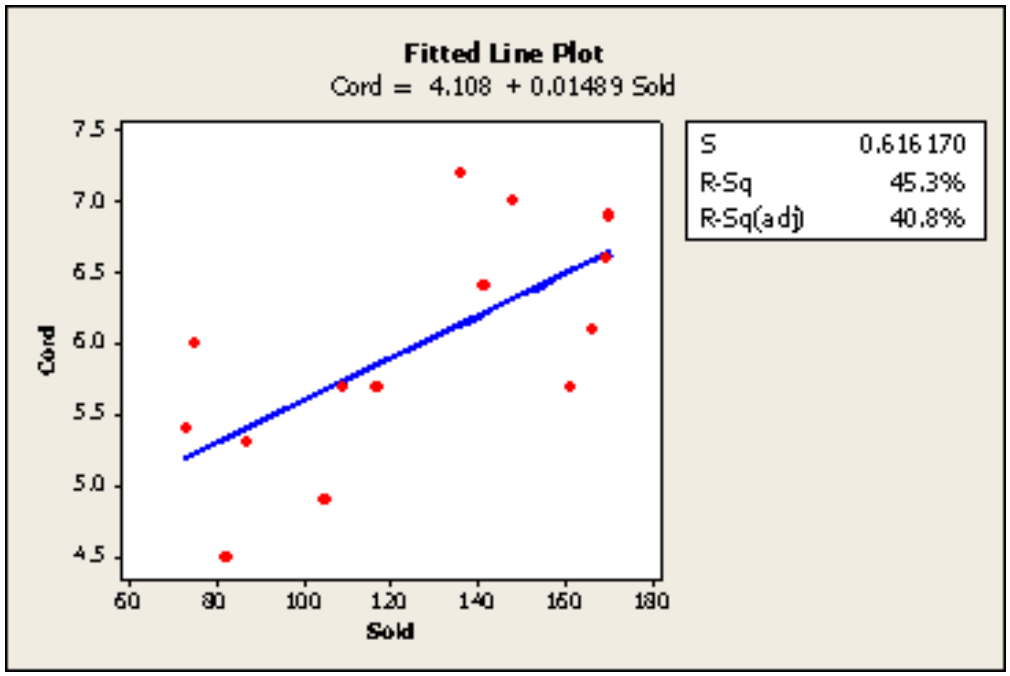
STAT 462

Applied Regression Analysis

3.2 - 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 (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/leadcord.txt)) on the monthly gasoline lead sales (in metric tons) in Massachusetts and mean lead concentrations (µl/dl) 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 fitted line plot:



and standard regression analysis output:

```
The regression equation is Cord = 4.11 + 0.0149 Sold

Predictor      Coef   SE Coef     T      P
Constant     4.1082    0.6088  6.75  0.000
Sold       0.014885  0.004719  3.15  0.008

S = 0.616170   R-Sq = 45.3%   R-Sq(adj) = 40.8%

Analysis of Variance

Source          DF      SS      MS     F      P
Regression       1  3.7783  3.7783  9.95  0.008
Residual Error  12  4.5560  0.3797
Total           13  8.3343
```

The *P*-value for testing $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_A : \beta_1 \neq 0$ is 0.008. Therefore, since the test statistic is positive, the *P*-value for testing $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_A : \beta_1 > 0$ is $0.008 \div 2 = 0.004$. The *P*-value is less than 0.05. There is sufficient statistical evidence, at the 0.05 level, to conclude that $\beta_1 > 0$.

Furthermore, since the 95% *t*-multiplier is $t(0.025, 12) = 2.1788$, a 95% confidence interval for β_1 is:

$$0.014885 \pm 2.1788(0.004719) \text{ or } (0.0046, 0.0252).$$

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 µl/dl 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.

‹ 3.1 - Inference for the Population Intercept and Slope (/stat462/node/102) | up (/stat462/node/80) | 3.3 - Sums of Squares › (/stat462/node/104)

STAT 462

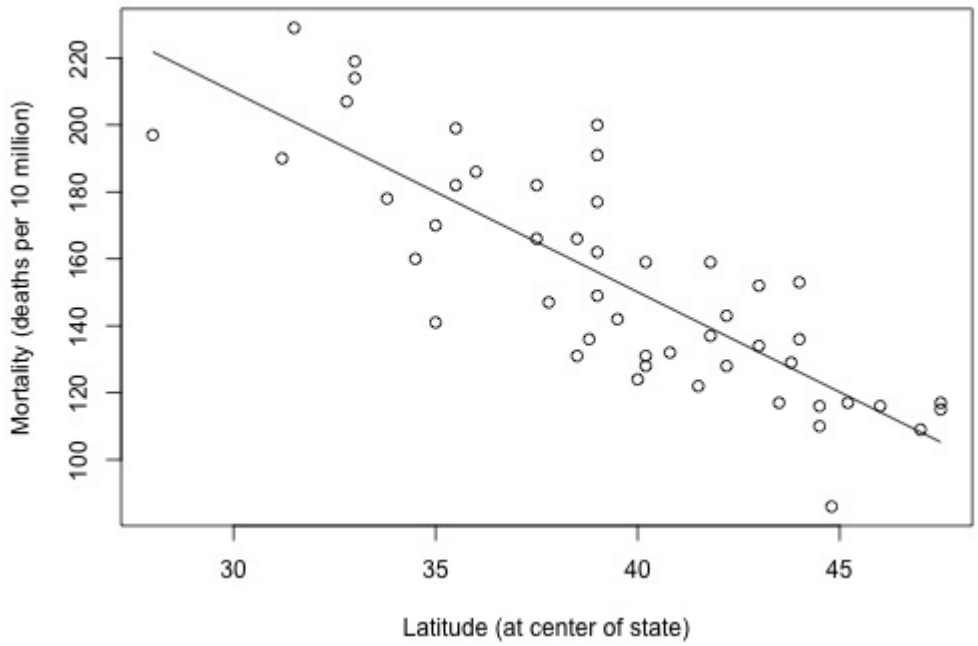
Applied Regression Analysis

3.3 - Sums of Squares

Example 1

Let's return to the skin cancer mortality example (skincancer.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/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 the research question? The linear relationship looks fairly strong. The estimated slope is negative, not equal to 0.



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

- the null hypothesis H_0: $\beta_1 = 0$
- against the alternative hypothesis H_A: $\beta_1 \neq 0$.

As the statistical software output below suggests, the *P*-value of the *t*-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 *F*-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 software output below. There is a column labeled **F**, which contains the *F*-test statistic, and there is a column labeled **P**, which contains the *P*-value associated with the *F*-test. Notice that the *P*-value, 0.000, appears to be the same as the *P*-value, 0.000, for the *t*-test for the slope. The *F*-test similarly tells us that there is enough statistical evidence to conclude that there is a linear relationship between skin cancer mortality and latitude.

```
The regression equation is Mort = 389 - 5.98 Lat

Predictor        Coef     SE Coef          T        P
Constant       389.19       23.81      16.34    0.000
Lat           -5.9776      0.5984      -9.99    0.000

S = 19.12       R-Sq = 68.0%     R-Sq(adj) = 67.3%

Analysis of Variance

Source           DF      SS     MS         F        P
Regression        1   36464  36464     99.80    0.000
Residual Error   47   17173    365
Total            48   53637
```

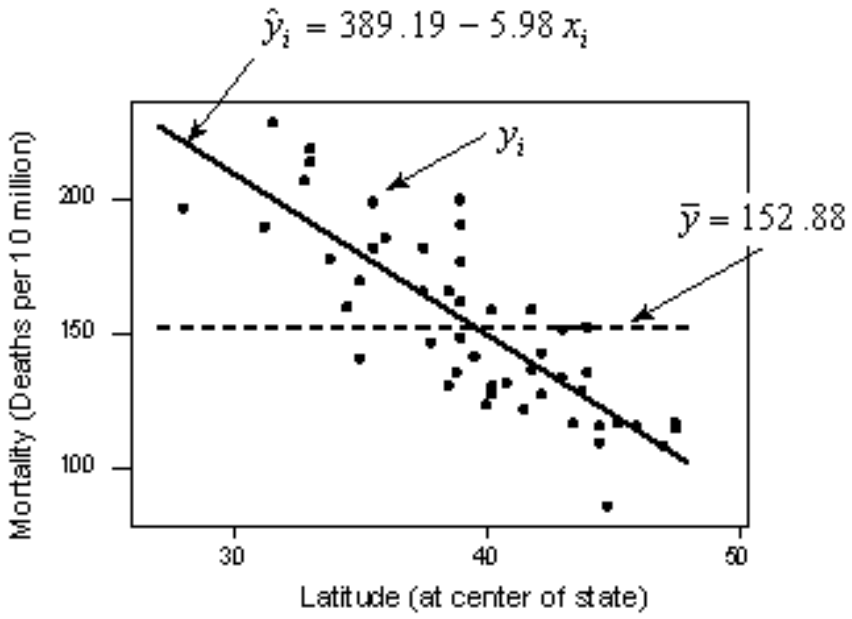
Now, let's investigate what all the numbers in the table represent. Let's start with the column labeled **SS** for "**sums of squares**." We considered sums of squares in Lesson 2 when we defined the coefficient of determination, r^2, but now we consider them again in the context of the analysis of variance table.

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

- y_i denotes the observed mortality for state *i*
- $\hat{y}_i$ is the estimated regression line (solid line) and therefore denotes the estimated (or "fitted") mortality for the latitude of state *i*
- $\bar{y}$ 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.

	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = 36464$ $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = 17173$ $\sum_{i=1}^{n}(y_i-\bar{y})^2 = 53637$



The distance of each observed value y_i from the no regression line $\bar{y}$ is $y_i - \bar{y}$. If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 53637$$

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 $\hat{y}_i$ from the no regression line $\bar{y}$ is $\hat{y}_i - \bar{y}$. If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 36464$$

Called the "**regression sum of squares**," it quantifies how far the estimated regression line is from the no relationship line.

The distance of each observed value y_i from the estimated regression line $\hat{y}_i$ is $y_i - \hat{y}_i$. If you determine this distance for each data point, square each distance, and add up all of the squared distances, you get:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 17173$$

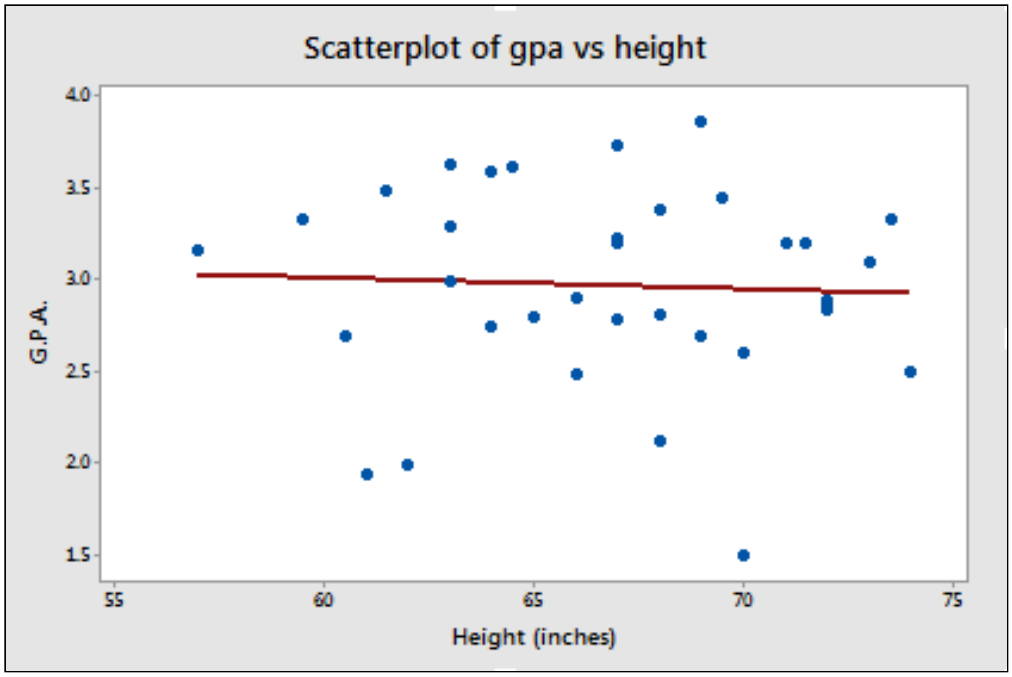
Called the "**error sum of squares**," as you know, it quantifies how much the data points vary around the estimated regression line.

In short, we have illustrated that the total variation in observed mortality y (53637) is the sum of two parts — variation "due to" latitude (36464) 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 is "associated with" mortality.)

Example 2

Now, let's do a similar analysis to investigate the research question, "Is there a (linear) relationship between height and grade point average?" (heightgpa.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/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 *P*-value of the *t*-test for:

- testing the null hypothesis H_0: $\beta_1 = 0$
- against the alternative hypothesis H_A: $\beta_1 \neq 0$.

As the statistical software output below suggests, the *P*-value of the *t*-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 grade point average.

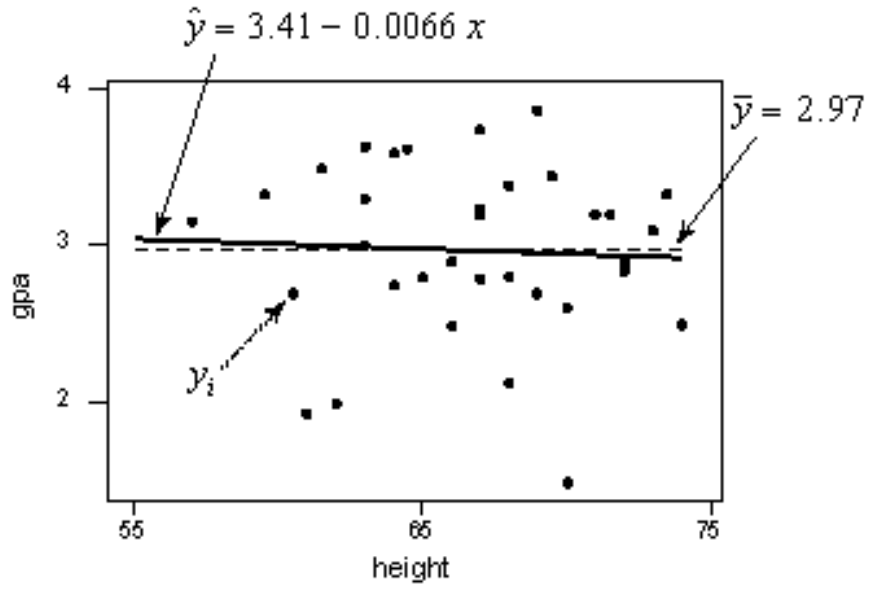
The software output also shows the analysis of variable table for this data set. Again, the *P*-value associated with the analysis of variance *F*-test, 0.761, appears to be the same as the *P*-value, 0.761, for the *t*-test for the slope. The *F*-test similarly tells us that there is insufficient statistical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

```
The regression equation is gpa = 3.41 - 0.0066 height

Predictor        Coef     SE Coef           T        P
Constant        3.410       1.435        2.38    0.023
height       -0.00656     0.02143       -0.31    0.761

S = 0.5423      R-Sq = 0.3%      R-Sq(adj) = 0.0%

Analysis of Variance

Source           DF      SS       MS         F        P
Regression        1  0.0276   0.0276      0.09    0.761
Residual Error   33  9.7055   0.2941
Total            34  9.7331
```

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

- y_i denotes the observed grade point average for student *i*
- $\hat{y}_i$ is the estimated regression line (solid line) and therefore denotes the estimated grade point average for the height of student *i*
- $\bar{y}$ represents the "no relationship" line (dashed line) between height and grade point average. It is simply the average grade point average 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 sums of squares summarize this point.



$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0.0276$

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 9.7055$

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 9.7331$

- The "**total sum of squares**," which again quantifies how much the observed grade point averages vary if you don't take into account height, is $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 9.7331$.
- The "**regression sum of squares**," which again quantifies how far the estimated regression line is from the no relationship line, is $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0.0276$.
- The "**error sum of squares**," which again quantifies how much the data points vary around the estimated $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 9.7055$.

In short, we have illustrated that the total variation in the observed grade point averages y (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 grade point averages is just due to random error. It appears as if very little of the variation can be attributed to the predictor height.

‹ 3.2 - Another Example of Slope Inference (/stat462/node/103) up (/stat462/node/80) 3.4 - Analysis of Variance: The Basic Idea › (/stat462/node/106)

STAT 462

Applied Regression Analysis

3.4 - Analysis of Variance: The Basic Idea

- Break down the total variation in *y* ("**total sum of squares**") into two components:
  - a component that is "due to" the change in *x* ("**regression sum of squares**")
  - 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 *x* and the response *y*.

Here is a simple picture illustrating how the distance $y_i - \bar{y}$ is decomposed into the sum of two distances, $\hat{y}_i - \bar{y}$ and $y_i - \hat{y}_i$. Roll your cursor over each of the three components of the equation at the bottom of the graphic below to see what each of the the values represents geometrically.

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

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SSTO Total sum of squares | **SSR** Regression sum of squares | **SSE** Error sum of squares

$$SSTO = SSR + SSE$$

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 *n*-1. Therefore, *n*-1 is the degrees of freedom associated with *SSTO*.
- Recall that the mean square error *MSE* is obtained by dividing *SSE* by *n*-2. Therefore, *n*-2 is the degrees of freedom associated with *SSE*.

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

$(n - 1)$	$=$	(1)	$+$	$(n - 2)$
degrees of freedom associated with **SSTO**		degrees of freedom associated with **SSR**		degrees of freedom associated with **SSE**

‹ 3.3 - Sums of Squares (/stat462/node/104) up (/stat462/node/80) 3.5 - The Analysis of Variance (ANOVA) table and the F-test › (/stat462/node/107)

STAT 462

Applied Regression Analysis

3.5 - The Analysis of Variance (ANOVA) table and the F-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/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/skincancer.txt)).

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

Recall that there were 49 states in the data set.

- The degrees of freedom associated with *SSR* will always be 1 for the simple linear regression model. The degrees of freedom associated with *SSTO* is *n*-1 = 49-1 = 48. The degrees of freedom associated with *SSE* is *n*-2 = 49-2 = 47. And the degrees of freedom add up: 1 + 47 = 48.
- The sums of squares add up: *SSTO* = *SSR* + *SSE*. That is, here: 53637 = 36464 + 17173.

Let's tackle a few more columns of the analysis of variance table, namely the "**mean square**" column, labled **MS**, and the *F*-statistic column, labeled **F**.

Definitions of mean squares

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

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}.$$

That is, we obtain the mean square error by dividing the error sum of squares by its associated degrees of freedom *n*-2. Similarly, we obtain the "**regression mean square (*MSR*)**" by dividing the regression sum of squares by its degrees of freedom 1:

$$MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}.$$

Of course, that means the regression sum of squares (*SSR*) and the regression mean square (*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 H_0: $\beta_1 = 0$ against the alternative hypothesis H_A: $\beta_1 \neq 0$.

Expected mean squares

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

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

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

$$E(MSE) = \sigma^2$$

These expected values suggest how to test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$:

- If $\beta_1 = 0$, then we'd expect the ratio *MSR*/*MSE* to equal 1.
- If $\beta_1 \neq 0$, then we'd expect the ratio *MSR*/*MSE* to be greater than 1.

These two facts suggest that we should use the ratio, *MSR*/*MSE*, to determine whether or not $\beta_1 = 0$.

Note that, because β_1 is squared in *E(MSR)*, we cannot use the ratio *MSR*/*MSE*:

- to test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 < 0$
- or to test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 > 0$.

We can only use MSR/MSE to test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$.

We have now completed our investigation of all of the entries of a standard analysis of variance table for simple linear regression. 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	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Residual error	*n*-2	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	*n*-1	$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$		

However, we will always let statistical software do the dirty work of calculating the values for us. Why is the ratio *MSR*/*MSE* labeled *F** in the analysis of variance table? That's because the ratio is known to follow an ***F*** **distribution** with 1 numerator degree of freedom and *n*-2 denominator degrees of freedom. For this reason, it is often referred to as the analysis of variance *F*-test. The following section summarizes the ANOVA *F*-test.

The ANOVA *F*-test for the slope parameter β_1

The null hypothesis is H_0: $\beta_1 = 0$.

The alternative hypothesis is H_A: $\beta_1 \neq 0$.

The test statistic is $F^* = \frac{MSR}{MSE}$.

As always, the *P*-value is obtained by answering the question: "What is the probability that we'd get an *F** statistic as large as we did, if the null hypothesis is true?"

The *P*-value is determined by comparing *F** to an *F* distribution with 1 numerator degree of freedom and *n*-2 denominator degrees of freedom.

In reality, we are going to let statistical software calculate the *F** statistic and the *P*-value for us. Let's try it out on some new examples!

‹ 3.4 - Analysis of Variance: The Basic Idea (/stat462/node/106) up (/stat462/node/80) 3.6 - Further SLR Evaluation Examples › (/stat462/node/108)

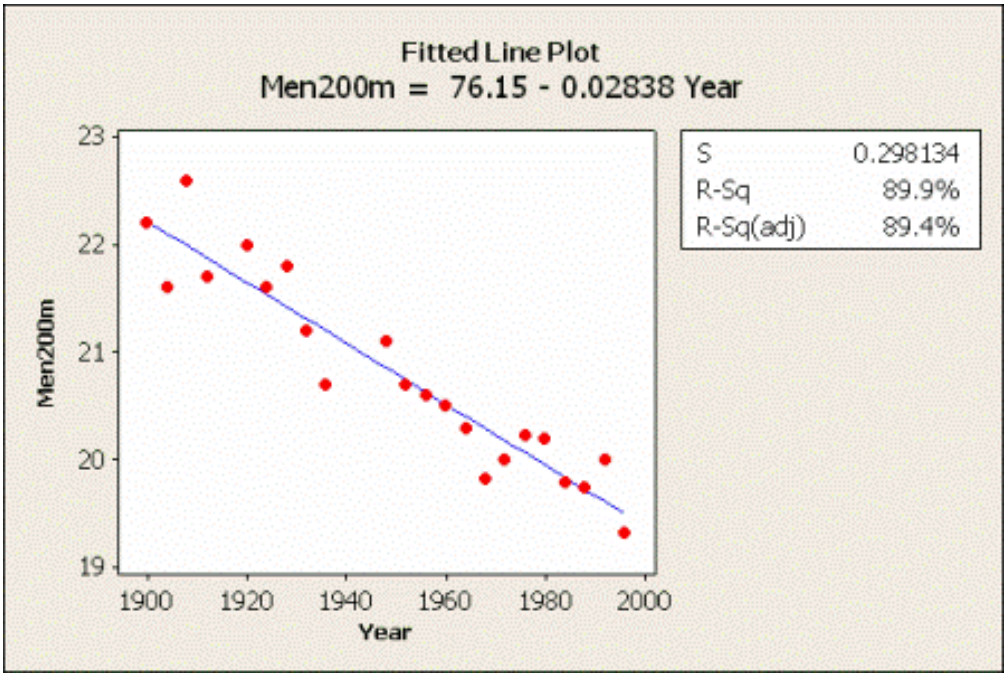
STAT 462

Applied Regression Analysis

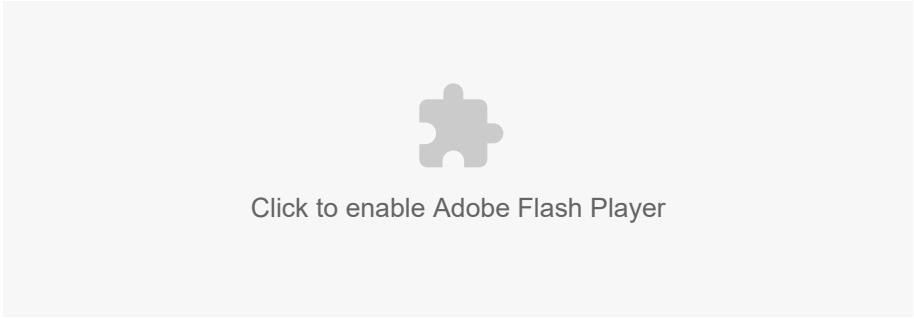
3.6 - Further SLR Evaluation Examples

Example 1: Are Sprinters Getting Faster?

The following data set (mens200m.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/mens200m.txt)) contains the winning times (in seconds) of the 22 men's 200 meter olympic sprints held between 1900 and 1996. (Notice that the Olympics were not held during the World War I and II years.) Is there a linear relationship between year and the winning times? The plot of the estimated regression line sure makes it look so!



To answer the research question, let's conduct the formal F-test of the null hypothesis H_0: $\beta_1 = 0$ against the alternative hypothesis H_A: $\beta_1 \neq 0$.



The analysis of variance table above has been animated to allow you to interact with the table. As you **roll your mouse over the blue numbers**, you are reminded of how those numbers are determined.

From a scientific point of view, what we ultimately care about is the *P*-value, which is 0.000 (to three decimal places). That is, the *P*-value is less than 0.001. The *P*-value is very small. It is unlikely that we would have obtained such a large *F** statistic if the null hypothesis were true. Therefore, we reject the null hypothesis H_0: $\beta_1 = 0$ in favor of the alternative hypothesis H_A: $\beta_1 \neq 0$. There is sufficient evidence at the $\alpha = 0.05$ level to conclude that there is a linear relationship between year and winning time.

Equivalence of the analysis of variance *F*-test and the *t*-test

As we noted in the first two examples, the *P*-value associated with the *t*-test is the same as the *P*-value associated with the analysis of variance *F*-test. This will always be true for the simple linear regression model. It is illustrated in the year and winning time example also. Both *P*-values are 0.000 (to three decimal places):

Predictor	Coef	SE Coef	T	P
Constant	76.153	4.152	18.34	0.000
Year	-0.0284	0.00213	-13.33	0.000

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.796	15.796	177.7	0.000
Residual Error	20	1.778	0.089		
Total	21	17.574			

The *P*-values are the same because of a well-known relationship between a *t* random variable and an *F* random variable that has 1 numerator degree of freedom. Namely:

$$(t^*_{(n-2)})^2 = F^*_{(1,n-2)}$$

This will always hold for the simple linear regression model. This relationship is demonstrated in this example as:

$$(-13.33)^2 = 177.7$$

In short:

- For a given significance level α, the *F*-test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is algebraically equivalent to the two-tailed *t*-test.
- We will get exactly the same *P*-values, so…
  - If one test rejects H_0, then so will the other.
  - If one test does not reject H_0, then so will the other.

The natural question then is ... when should we use the *F*-test and when should we use the *t*-test?

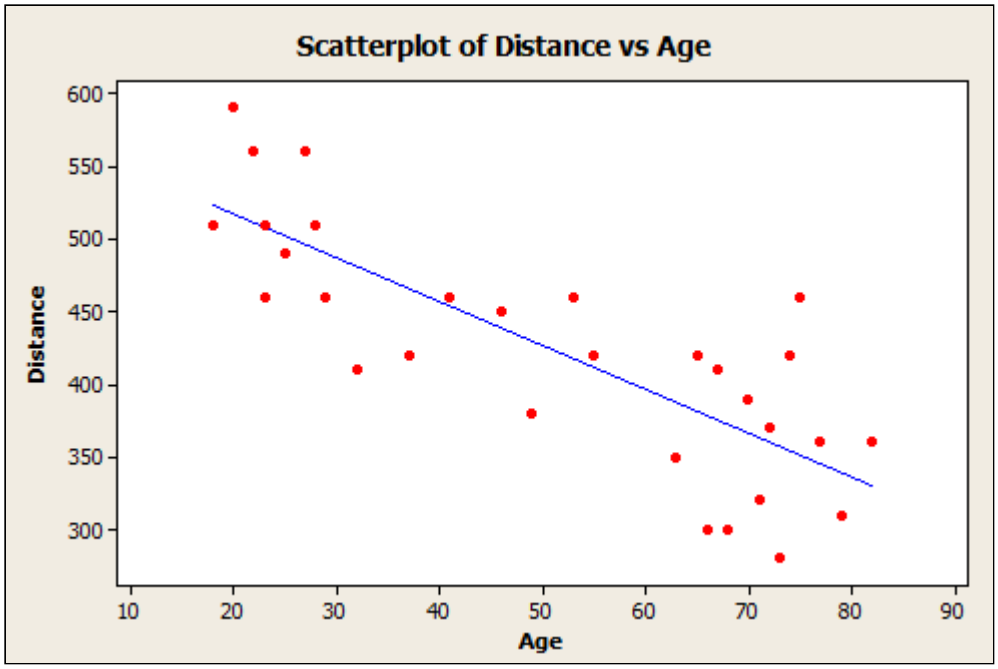
- The *F*-test is only appropriate for testing that the slope differs from 0 ($\beta_1 \neq 0$).
- Use the *t*-test to test that the slope is positive ($\beta_1 > 0$) or negative ($\beta_1 < 0$). Remember, though, that you will have to divide the reported two-tail *P*-value by 2 to get the appropriate one-tailed *P*-value.

The *F*-test is more useful for the multiple regression model when we want to test that more than one slope parameter is 0. We'll learn more about this later in the course!

Example 2: Highway Sign Reading Distance and Driver Age

The data are $n = 30$ observations on driver age and the maximum distance (feet) at which individuals can read a highway sign (signdist.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/signdist.txt)). (Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard).

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



Here is the accompanying regression output:

Regression Analysis: Distance versus Age

```
The regression equation is
Distance = 577 - 3.01 Age


Predictor      Coef  SE Coef      T      P
Constant    576.68    23.47  24.57  0.000
Age        -3.0068   0.4243  -7.09  0.000
```

Hypothesis Test for the Intercept (β_0)

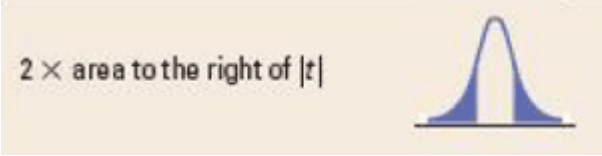
This test is rarely a test of interest, but does show up when one is interested in performing a regression through the origin (which we touched on earlier in this lesson). In the software output above, the row labeled Constant gives the information used to make inferences about the intercept. The null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

$$H_0 : \beta_0 = 0$$
$$H_A : \beta_0 \neq 0.$$

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 how the value of y changes when the value of x changes. Nevertheless, to test whether the population intercept is 0, the information from the software output would be used as follows:

1. The sample intercept is $b_0 = 576.68$, the value under **Coef**.
2. The standard error (SE) of the sample intercept, written as $\text{se}(b_0)$, is $\text{se}(b_0) = 23.47$, the value under SE Coef. The SE of any statistic is a measure of its accuracy. In this case, the SE of b_0 gives, very roughly, the average difference between the sample b_0 and the true population intercept β_0, for random samples of this size (and with these x-values).
3. The test statistic is $t = b_0/\text{se}(b_0) = 576.68/23.47 = 24.57$, the value under T.
4. The p-value for the test is $p = 0.000$ and is given under P. The p-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 $p < 0.05$. Thus, we conclude that there is statistically significant evidence that the population intercept is not equal to 0.

So how exactly is the p-value found? For simple regression, the p-value is determined using a t distribution with $n - 2$ degrees of freedom (df), which is written as t_{n-2}, and is calculated as 2 × area past $|t|$ under a t_{n-2} curve. In this example, $df = 30 - 2 = 28$. The p-value region is the type of region shown in the figure below. The negative and positive versions of the calculated t provide the interior boundaries of the two shaded regions. As the value of t increases, the p-value (area in the shaded regions) decreases.



Hypothesis Test for the Slope (β_1)

This test can be used to test whether or not x and y are linearly related. The row pertaining to the variable **Age** in the software output from earlier gives information used to make inferences about the slope. The slope directly tells us about the link between the mean y and x. When the true population slope does not equal 0, the variables y and x are linearly related. When the slope is 0, there is not a linear relationship because the mean y does not change when the value of x is changed. The null and alternative hypotheses for a hypotheses test about the slope are written as:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0.$$

In other words, the null hypothesis is testing if the population slope is equal to 0 versus the alternative hypothesis that the population slope is not equal to 0. To test whether the population slope is 0, the information from the software output is used as follows:

1. The sample slope is $b_1 = -3.0068$, the value under **Coef** in the **Age** row of the output.
2. The SE of the sample slope, written as se(b_1), is se(b_1) = 0.4243, the value under SE **Coef**. Again, the SE of any statistic is a measure of its accuracy. In this case, the SE of b1 gives, very roughly, the average difference between the sample b_1 and the true population slope β_1, for random samples of this size (and with these x-values).
3. The test statistic is $t = b_1/\text{se}(b_1) = -3.0068/0.4243 = -7.09$, the value under T.
4. The p-value for the test is $p = 0.000$ and is given under P.
5. The decision rule at the 0.05 significance level is to reject the null hypothesis since our $p < 0.05$. Thus, we conclude that there is statistically significant evidence that the variables of Distance and Age are linearly related.

As before, the p-value is the region illustrated in the figure above.

Confidence Interval for the Slope (β_1)

A confidence interval for the unknown value of the population slope β_1 can be computed as

$$\text{sample statistic} \pm \text{multiplier} \times \text{standard error of statistic}$$

$$\rightarrow b_1 \pm t^* \times \text{se}(b_1).$$

In simple regression, the t^* multiplier is determined using a t_{n-2} distribution. The value of t^* is such that the confidence level is the area (probability) between $-t^*$ and $+t^*$ under the t-curve. To find the t^* multiplier, you can do one of the following:

1. A table such as the one in the textbook can be used to look up the multiplier.
2. Alternatively, software like Minitab can be used.

95% Confidence Interval

In our example, $n = 30$ and $df = n - 2 = 28$. For 95% confidence, $t^* = 2.05$. A 95% confidence interval for β_1, the true population slope, is:

$-3.0068 \pm (2.05 \times 0.4243)$
-3.0068 ± 0.870
or about − 3.88 to − 2.14.

Interpretation: With 95% confidence, we can say the mean sign reading distance decreases somewhere between 2.14 and 3.88 feet per each one-year increase in age. It is incorrect to say that with 95% *probability* the mean sign reading distance decreases somewhere between 2.14 and 3.88 feet per each one-year increase in age. Make sure you understand why!!!

99% Confidence Interval

For 99% confidence, $t^* = 2.76$. A 99% confidence interval for β_1 , the true population slope is:

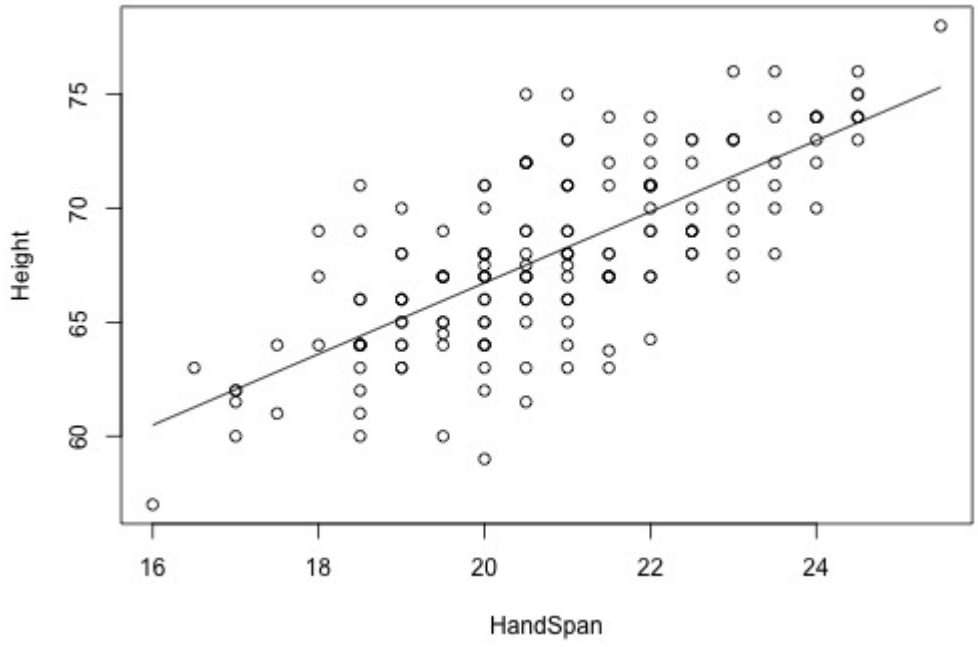
$-3.0068 \pm (2.76 \times 0.4243)$
-3.0068 ± 1.1711
or about -4.18 to -1.84.

Interpretation: With 99% confidence, we can say the mean sign reading distance decreases somewhere between 1.84 and 4.18 feet per each one-year increase in age. Notice that as we increase our confidence, the interval becomes wider. So as we approach 100% confidence, our interval grows to become the whole real line.

As a final note, the above procedures can be used to calculate a confidence interval for the population intercept. Just use b_0 (and its standard error) rather than b_1.

Example 3: Handspans Data

Stretched handspans and heights are measured in centimeters for $n = 167$ college students (handheight.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/handheight.txt)). We'll use y = height and x = stretched handspan. A scatterplot with a regression line superimposed is given below, together with results of a simple linear regression model fit to the data.




```
lm(formula = Height ~ HandSpan)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.5250     2.3160   15.34   <2e-16 ***
HandSpan      1.5601     0.1105   14.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.744 on 165 degrees of freedom
Multiple R-squared:  0.5469,    Adjusted R-squared:  0.5442
F-statistic: 199.2 on 1 and 165 DF,  p-value: < 2.2e-16

Analysis of Variance Table
```

```
Response: Height
           Df Sum Sq Mean Sq F value    Pr(>F)
HandSpan    1 1500.1 1500.06  199.17 < 2.2e-16 ***
Residuals 165 1242.7    7.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Some things to note are:

- The residual standard deviation S is 2.744 and this estimates the standard deviation of the errors.
- r^2 = (SSTO-SSE) / SSTO = SSR / (SSR+SSE) = 1500.1 / (1500.1+1242.7) = 1500.1 / 2742.8 = 0.547 or 54.7%. The interpretation is that handspan differences explain 54.7% of the variation in heights.
- The value of the F statistic is F = 199.2 with 1 and 165 degrees of freedom, and the p-value for this F statistic is 0.000. Thus we reject the null hypothesis $H_0 : \beta_1 = 0$ because the p-value is so small. In other words, the observed relationship is statistically significant.

‹ 3.5 - The Analysis of Variance (ANOVA) table and the F-test (/stat462/node/107) | up (/stat462/node/80) | 3.7 - Decomposing The Error When There Are Replicates › (/stat462/node/111)