

# STAT 462

## Applied Regression Analysis

### Lesson 2: Simple Linear Regression (SLR) Model

#### Overview of this Lesson

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. This lesson introduces the concept and basic procedures of simple linear regression. We will also learn two measures that describe the strength of the linear association that we find in data.

#### Key Learning Goals for this Lesson:

- Distinguish between a deterministic relationship and a statistical relationship.
- Understand the concept of the least squares criterion.
- Interpret the intercept  $b_0$  and slope  $b_1$  of an estimated regression equation.
- Know how to obtain the estimates  $b_0$  and  $b_1$  using statistical software.
- Recognize the distinction between a population regression line and the estimated regression line.
- Summarize the four conditions that underlie the simple linear regression model.
- Know what the unknown population variance  $\sigma^2$  quantifies in the regression setting.
- Know how to obtain the estimate  $MSE$  of the unknown population variance  $\sigma^2$  using statistical software.
- Know that the coefficient of determination ( $r^2$ ) and the correlation coefficient ( $r$ ) are measures of **linear** association. That is, they can be 0 even if there is perfect nonlinear association.
- Know how to interpret the  $r^2$  value.
- Understand the cautions necessary in using the  $r^2$  value as a way of assessing the strength of the linear association.
- Know how to calculate the correlation coefficient  $r$  from the  $r^2$  value.
- Know what various correlation coefficient values mean. There is no meaningful interpretation for the correlation coefficient as there is for the  $r^2$  value.

- 
- 2.1 - What is Simple Linear Regression? (</stat462/node/91>)
  - 2.2 - What is the "Best Fitting Line"? (</stat462/node/92>)
  - 2.3 - The Simple Linear Regression Model (</stat462/node/93>)
  - 2.4 - What is the Common Error Variance? (</stat462/node/94>)
  - 2.5 - The Coefficient of Determination, r-squared (</stat462/node/95>)

- 2.6 - (Pearson) Correlation Coefficient  $r$  (</stat462/node/96>)
- 2.7 - Coefficient of Determination and Correlation Examples (</stat462/node/97>)
- 2.8 - R-squared Cautions (</stat462/node/98>)
- 2.9 - Simple Linear Regression Examples (</stat462/node/101>)

---

2.1 - What is Simple Linear Regression? ›

(</stat462/node/91>)

---

## STAT 462

## Applied Regression Analysis

## 2.1 - What is Simple Linear Regression?

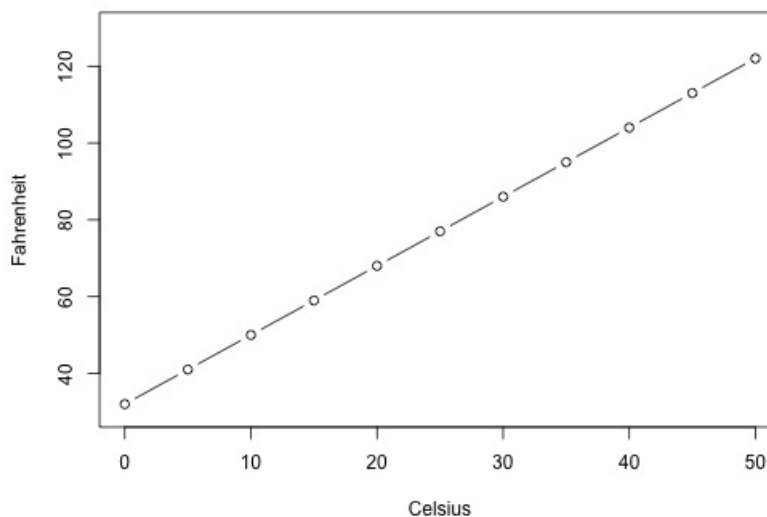
**Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted  $x$ , is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted  $y$ , is regarded as the **response**, **outcome**, or **dependent** variable.

Because the other terms are used less frequently today, we'll use the "**predictor**" and "**response**" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them. Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

### Types of relationships

Before proceeding, we must clarify what types of relationships we won't study in this course, namely, **deterministic** (or **functional**) **relationships**. Here is an example of a deterministic relationship.



Note that the observed  $(x, y)$  data points fall directly on a line. As you may remember, the relationship between degrees Fahrenheit and degrees Celsius is known to be:

$$F = \frac{9}{5}C + 32$$

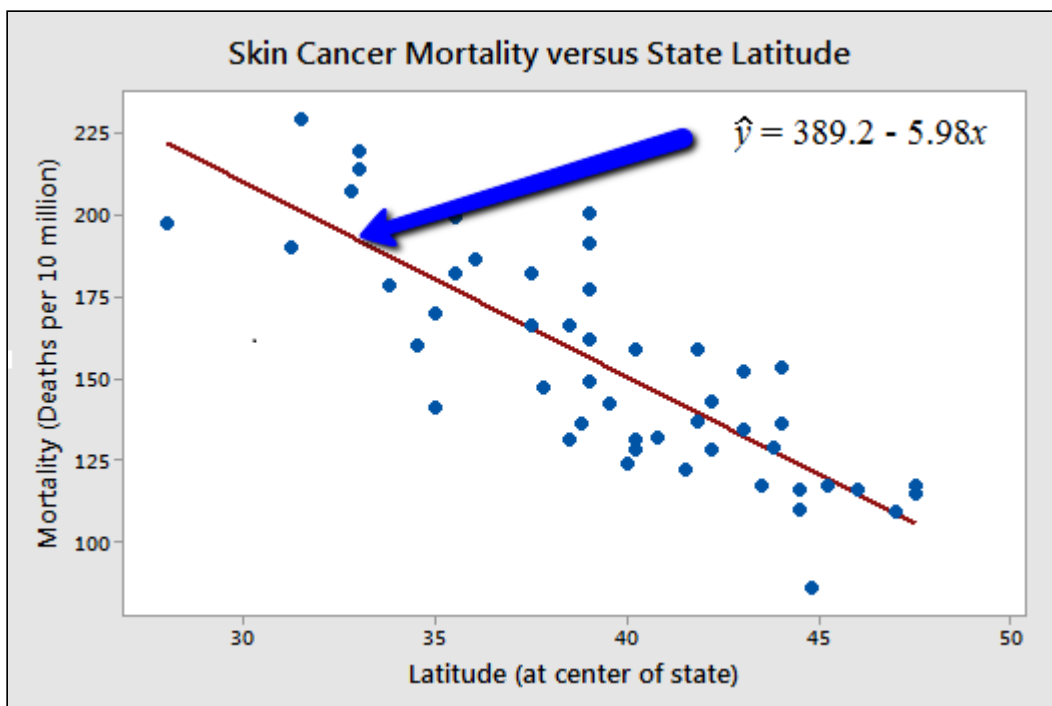
That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit *exactly*.

Here are some examples of other deterministic relationships that students from previous semesters have shared:

- Circumference =  $\pi \times \text{diameter}$
- Hooke's Law:  $Y = \alpha + \beta X$ , where  $Y$  = amount of stretch in a spring, and  $X$  = applied weight.
- Ohm's Law:  $I = V/r$ , where  $V$  = voltage applied,  $r$  = resistance, and  $I$  = current.
- Boyle's Law: For a constant temperature,  $P = \alpha/V$ , where  $P$  = pressure,  $\alpha$  = constant for each gas, and  $V$  = volume of gas.

For each of these deterministic relationships, the equation *exactly* describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in **statistical relationships**, in which the relationship between the variables is not perfect.

Here is an example of a statistical relationship. The response variable  $y$  is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable  $x$  is the latitude (degrees North) at the center of each of 49 states in the U.S. (skincancer.txt (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/skincancer.txt)) (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states, and Washington, D.C. is included in the data set even though it is not technically a state.)



You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter plot supports such a hypothesis. There appears to be a negative linear relationship between latitude and mortality due to skin cancer, but the relationship is not perfect. Indeed, the plot exhibits some "**trend**," but it also exhibits some "**scatter**." Therefore, it is a statistical relationship, not a deterministic one.

Some other examples of statistical relationships might include:

- Height and weight — as height increases, you'd expect weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.

- Vital lung capacity and pack-years of smoking — as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
- Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Okay, so let's study statistical relationships between one response variable  $y$  and one predictor variable  $x$ !

---

<a href="#">◀ Lesson 2: Simple Linear Regression (SLR) Model (/stat462/node/79)</a>	<a href="#">up (/stat462/node/79)</a>	<a href="#">2.2 - What is the "Best Fitting Line"? ▶ (/stat462/node/92)</a>
---	---------------------------------------	---

---

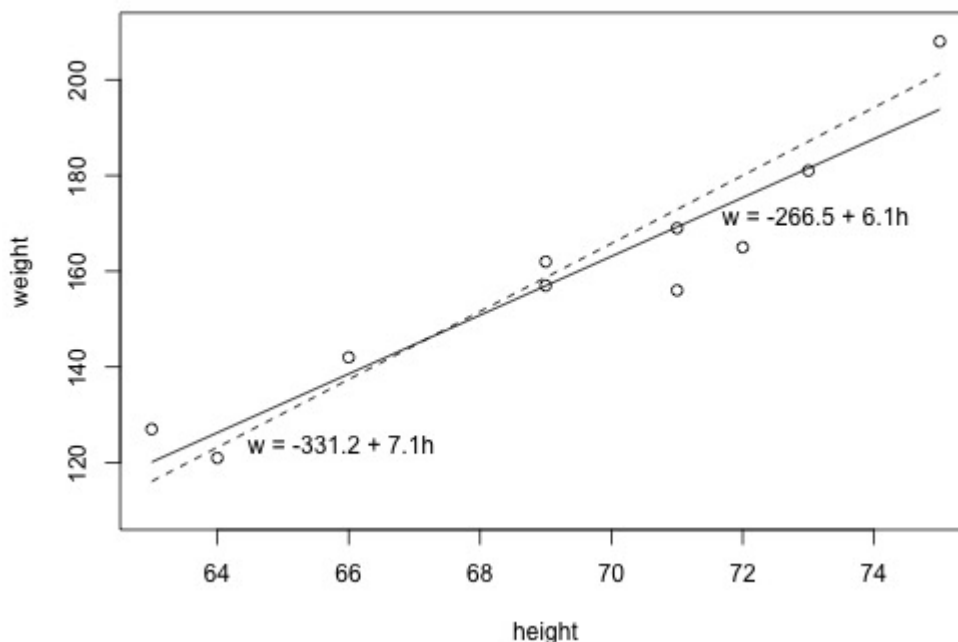
# STAT 462

## Applied Regression Analysis

### 2.2 - What is the "Best Fitting Line"?

Since we are interested in summarizing the trend between two quantitative variables, the natural question arises — "what is the best fitting line?" At some point in your education, you were probably shown a scatter plot of  $(x, y)$  data and were asked to draw the "most appropriate" line through the data. Even if you weren't, you can try it now on a set of heights ( $x$ ) and weights ( $y$ ) of 10 students, (student\_height\_weight.txt)

(/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/student\_height\_weight.txt) . Looking at the plot below, which line — the solid line or the dashed line — do you think best summarizes the trend between height and weight?



Hold on to your answer! In order to examine which of the two lines is a better fit, we first need to introduce some common notation:

- $y_i$  denotes the observed response for experimental unit  $i$
- $x_i$  denotes the predictor value for experimental unit  $i$
- $\hat{y}_i$  is the predicted response (or fitted value) for experimental unit  $i$

Then, the equation for the best fitting line is:


$$\hat{y}_i = b_0 + b_1x_i$$

Incidentally, recall that an **"experimental unit"** is the object or person on which the measurement is made. In our height and weight example, the experimental units are students.

Let's try out the notation on our example with the trend summarized by the line  $w = -266.53 + 6.1376 h$ . (Note that this line is just a more precise version of the above solid line,  $w = -266.5 + 6.1 h$ .) The first data point in the list indicates that student 1 is 63 inches tall and weighs 127 pounds. That is,  $x_1 = 63$  and  $y_1 = 127$ . Do you see this point on the plot? If we know this student's height but not his or her weight, we could use the equation of the line to predict his or her weight. We'd predict the student's weight to be  $-266.53 + 6.1376(63)$  or 120.1 pounds. That is,  $\hat{y}_1 = 120.1$ . Clearly, our prediction wouldn't be perfectly correct — it has some **"prediction error"** (or **"residual error"**). In fact, the size of its prediction error is  $127 - 120.1$  or 6.9 pounds.

You might want to roll your cursor over each of the 10 data points to make sure you understand the notation used to keep track of the predictor values, the observed responses and the predicted responses:

$i$	$x_i$	$y_i$	$\hat{y}_i$
1	63	127	120.1
2	64	121	126.3
3	66	142	138.5
4	69	157	157.0
5	69	162	157.0
6	71	156	169.2
7	71	169	169.2
8	72	165	175.4
9	73	181	181.5
10	75	208	193.8



Click to enable Adobe Flash Player

As you can see, the size of the prediction error depends on the data point. If we didn't know the weight of student 5, the equation of the line would predict his or her weight to be  $-266.53 + 6.1376(69)$  or 157 pounds. The size of the prediction error here is  $162 - 157$ , or 5 pounds.

In general, when we use  $\hat{y}_i = b_0 + b_1 x_i$  to predict the actual response  $y_i$ , we make a prediction error (or residual error) of size:

$$e_i = y_i - \hat{y}_i$$

A line that fits the data **"best"** will be one for which the  **$n$  prediction errors** — one for each observed data point — **are as small as possible in some overall sense**. One way to achieve this goal is to invoke the **"least squares criterion,"** which says to "minimize the sum of the squared prediction errors." That is:

- The equation of the best fitting line is:  $\hat{y}_i = b_0 + b_1 x_i$
- We just need to find the values  $b_0$  and  $b_1$  that make the sum of the squared prediction errors the smallest it can be.

- That is, we need to find the values  $b_0$  and  $b_1$  that minimize:

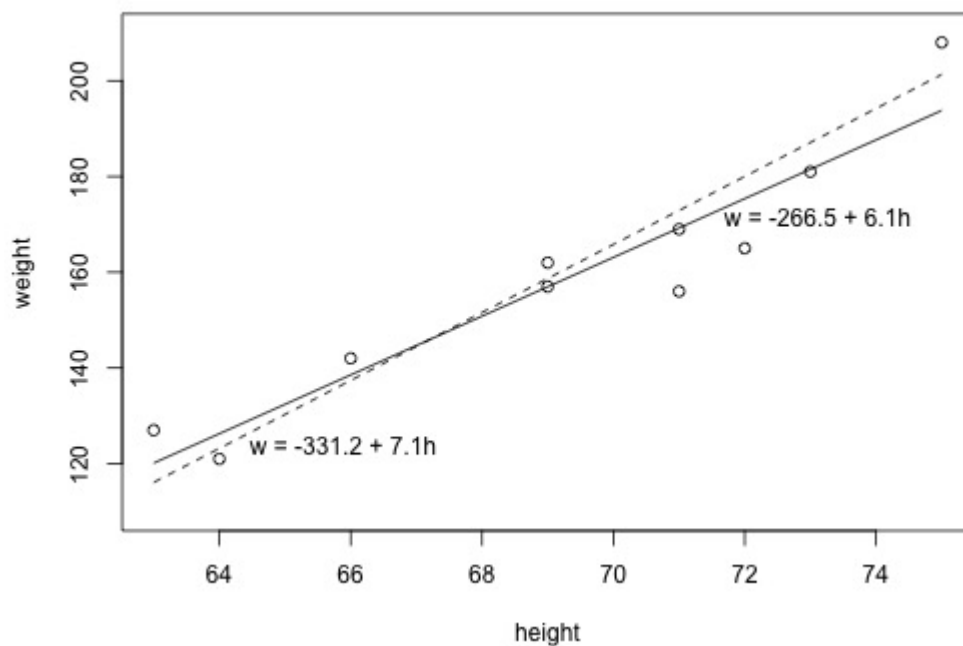
$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Here's how you might think about this quantity  $Q$ :

- The quantity  $e_i = y_i - \hat{y}_i$  is the prediction error for data point  $i$ .
- The quantity  $e_i^2 = (y_i - \hat{y}_i)^2$  is the squared prediction error for data point  $i$ .
- And, the symbol  $\sum_{i=1}^n$  tells us to add up the squared prediction errors for all  $n$  data points.

Incidentally, if we didn't square the prediction error  $e_i = y_i - \hat{y}_i$  to get  $e_i^2 = (y_i - \hat{y}_i)^2$ , the positive and negative prediction errors would cancel each other out when summed, always yielding 0.

Now, being familiar with the least squares criterion, let's take a fresh look at our plot again. In light of the least squares criterion, which line do you now think is the best fitting line?



Let's see how you did! The following two side-by-side tables illustrate the implementation of the least squares criterion for the two lines up for consideration — the dashed line and the solid line.

$w = -331.2 + 7.1 h$ (the dashed line)						$w = -266.53 + 6.1376 h$ (the solid line)					
$i$	$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$	$i$	$x_i$	$y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81	1	63	127	120.139	6.8612	47.076
2	64	121	123.2	-2.2	4.84	2	64	121	126.276	-5.2764	27.840
3	66	142	137.4	4.6	21.16	3	66	142	138.552	3.4484	11.891
4	69	157	158.7	-1.7	2.89	4	69	157	156.964	0.0356	0.001
5	69	162	158.7	3.3	10.89	5	69	162	156.964	5.0356	25.357
6	71	156	172.9	-16.9	285.61	6	71	156	169.240	-13.2396	175.287
7	71	169	172.9	-3.9	15.21	7	71	169	169.240	-0.2396	0.057



8	72	165	180.0	-15.0	225.00	8	72	165	175.377	-10.3772	107.686
9	73	181	187.1	-6.1	37.21	9	73	181	181.515	-0.5148	0.265
10	75	208	201.3	6.7	44.89	10	75	208	193.790	14.2100	201.924
					<u>766.5</u>						<u>597.4</u>

Based on the least squares criterion, which equation best summarizes the data? The sum of the squared prediction errors is 766.5 for the dashed line, while it is only 597.4 for the solid line. Therefore, of the two lines, the solid line,  $w = -266.53 + 6.1376h$ , best summarizes the data. But, is this equation guaranteed to be the best fitting line of all of the possible lines we didn't even consider? Of course not!

If we used the above approach for finding the equation of the line that minimizes the sum of the squared prediction errors, we'd have our work cut out for us. We'd have to implement the above procedure for an infinite number of possible lines — clearly, an impossible task! Fortunately, somebody has done some dirty work for us by figuring out formulas for the **intercept**  $b_0$  and the **slope**  $b_1$  for the equation of the line that minimizes the sum of the squared prediction errors.

The formulas are determined using methods of calculus. We minimize the equation for the sum of the squared prediction errors:

$$Q = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

(that is, take the derivative with respect to  $b_0$  and  $b_1$ , set to 0, and solve for  $b_0$  and  $b_1$ ) and get the "**least squares estimates**" for  $b_0$  and  $b_1$ :

$$b_0 = \bar{y} - b_1 \bar{x}$$

and:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Because the formulas for  $b_0$  and  $b_1$  are derived using the least squares criterion, the resulting equation —  $\hat{y}_i = b_0 + b_1 x_i$  — is often referred to as the "**least squares regression line**," or simply the "**least squares line**." It is also sometimes called the "**estimated regression equation**." Incidentally, note that in deriving the above formulas, we made no assumptions about the data other than that they follow some sort of linear trend.

We can see from these formulas that the least squares line passes through the point  $(\bar{x}, \bar{y})$ , since when  $x = \bar{x}$ , then  $y = b_0 + b_1 \bar{x} = \bar{y} - b_1 \bar{x} + b_1 \bar{x} = \bar{y}$ .

In practice, you won't really need to worry about the formulas for  $b_0$  and  $b_1$ . Instead, you are going to let statistical software, such as R or Minitab, find least squares lines for you.

One thing the **estimated regression coefficients**,  $b_0$  and  $b_1$ , allow us to do is to predict future responses — one of the most common uses of an estimated regression line. This use is rather straightforward:

A common use of the estimated regression line.	$\hat{y}_{i,wt} = -266.53 + 6.1376x_{i,ht}$
Predict (mean) weight of 66"-inch tall people.	$\hat{y}_{i,wt} = -266.53 + 6.1376(66) = 138.55$
Predict (mean) weight of 67"-inch tall people.	$\hat{y}_{i,wt} = -266.53 + 6.1376(67) = 144.69$

**Now, what does  $b_0$  tell us?** The answer is obvious when you evaluate the estimated regression equation at  $x = 0$ .

Here, it tells us that a person who is 0 inches tall is predicted to weigh -266.53 pounds! Clearly, this prediction is nonsense. This happened because we "**extrapolated**" beyond the "**scope of the model**" (the range of the  $x$  values). It is not meaningful to have a height of 0 inches, that is, the scope of the model does not include  $x = 0$ . So, here the intercept  $b_0$  is not meaningful. In general, if the "scope of the model" includes  $x = 0$ , then  $b_0$  is the predicted mean response when  $x = 0$ . Otherwise,  $b_0$  is not meaningful. There is more discussion of this here

(<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-to-interpret-the-constant-y-intercept>) .

**And, what does  $b_1$  tell us?** The answer is obvious when you subtract the predicted weight of 66"-inch tall people from the predicted weight of 67"-inch tall people. We obtain  $144.69 - 138.55 = 6.14$  pounds -- the value of  $b_1$ . Here, it tells us that we predict the mean weight to increase by 6.14 pounds for every additional one-inch increase in height. In general, we can expect the mean response to increase or decrease by  $b_1$  units for every one unit increase in  $x$ .

---

◀ 2.1 - What is Simple Linear Regression?  
(/stat462/node/91)

up  
(/stat462/node/79)

2.3 - The Simple Linear Regression Model ▶  
(/stat462/node/93)

---

## STAT 462

## Applied Regression Analysis

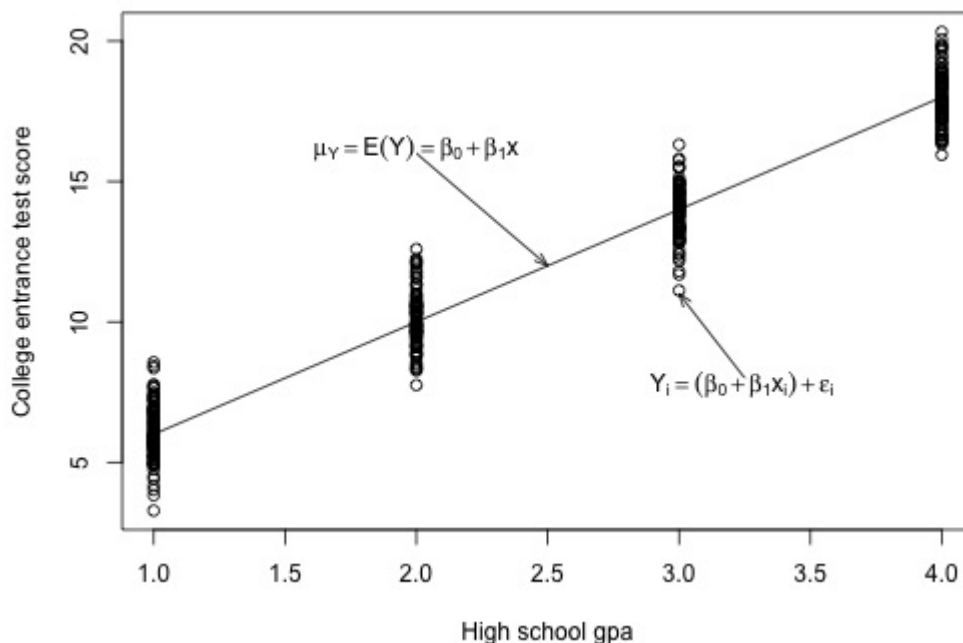
## 2.3 - The Simple Linear Regression Model

We have worked hard to come up with formulas for the intercept  $b_0$  and the slope  $b_1$  of the least squares regression line. But, we haven't yet discussed what  $b_0$  and  $b_1$  estimate.

What do  $b_0$  and  $b_1$  estimate?

Let's investigate this question with another example. Below is a plot illustrating a potential relationship between the predictor "high school grade point average (gpa)" and the response "college entrance test score." Only four groups ("subpopulations") of students are considered — those with a gpa of 1, those with a gpa of 2, ..., and those with a gpa of 4.

Let's focus for now just on those students who have a gpa of 1. As you can see, there are so many data points — each representing one student — that the data points run together. That is, the data on the entire subpopulation of students with a gpa of 1 are plotted. And, similarly, the data on the entire subpopulation of students with gpas of 2, 3, and 4 are plotted.



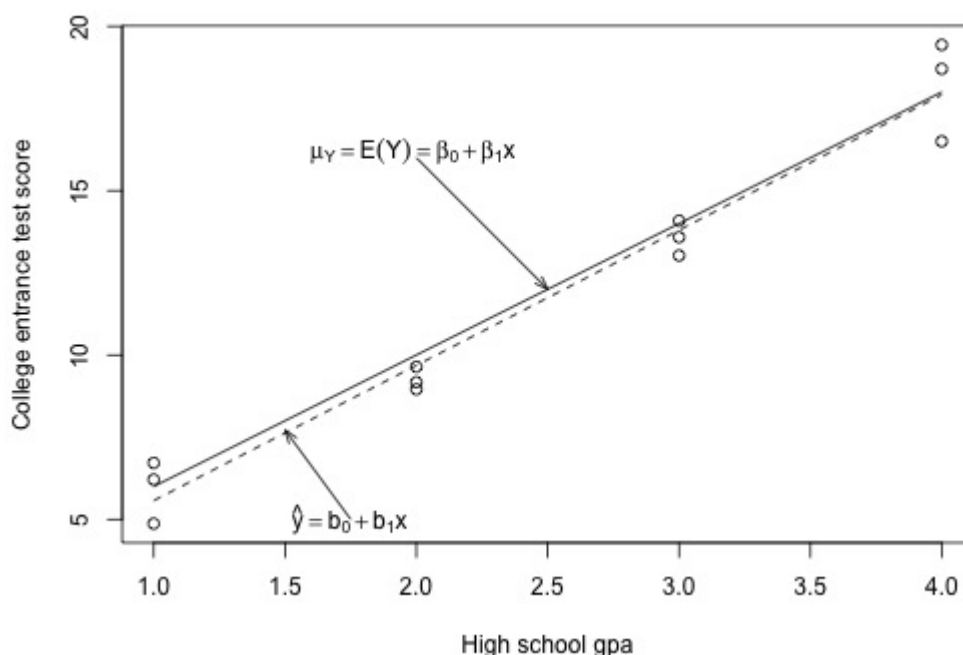
Now, take the average college entrance test score for students with a gpa of 1. And, similarly, take the average college entrance test score for students with a gpa of 2, 3, and 4. Connecting the dots — that is, the averages — you get a line, which we summarize by the formula  $\mu_Y = E(Y) = \beta_0 + \beta_1 x$ . The line — which is called the

"**population regression line**" — summarizes the trend *in the population* between the predictor  $x$  and the mean of the responses  $\mu_Y$ . We can also express the average college entrance test score for the  $i$ -th student,

$E(Y_i) = \beta_0 + \beta_1 x_i$ . Of course, not every student's college entrance test score will equal the average  $E(Y_i)$ . There will be some error. That is, any student's response  $y_i$  will be the linear trend  $\beta_0 + \beta_1 x_i$  plus some error  $\epsilon_i$ . So, another way to write the simple linear regression model is  $y_i = E(Y_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$ .

When looking to summarize the relationship between a predictor  $x$  and a response  $y$ , we are interested in knowing the population regression line  $\mu_Y = E(Y) = \beta_0 + \beta_1 x$ . The only way we could ever know it, though, is to be able to collect data on everybody in the population — most often an impossible task. We have to rely on taking and using a sample of data from the population to estimate the population regression line.

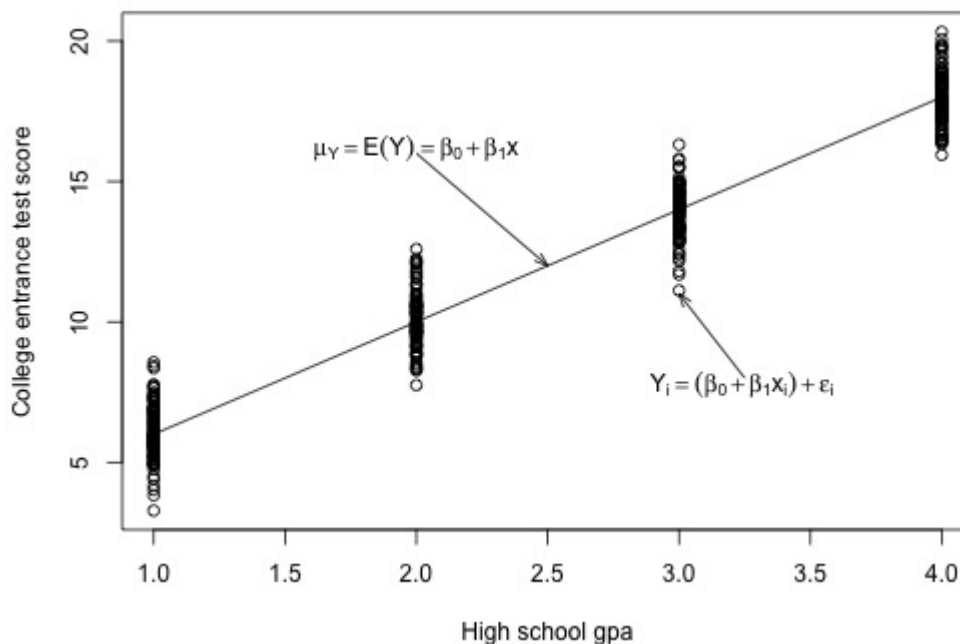
Let's take a sample of three students from each of the subpopulations — that is, three students with a gpa of 1, three students with a gpa of 2, ..., and three students with a gpa of 4 — for a total of 12 students. As the plot below suggests, the least squares regression line  $\hat{y} = b_0 + b_1 x$  through the sample of 12 data points estimates the population regression line  $\mu_Y = E(Y) = \beta_0 + \beta_1 x$ . That is, the sample intercept  $b_0$  estimates the population intercept  $\beta_0$  and the sample slope  $b_1$  estimates the population slope  $\beta_1$ .



The least squares regression line doesn't match the population regression line perfectly, but it is a pretty good estimate. And, of course, we'd get a different least squares regression line if we took another (different) sample of 12 such students. Ultimately, we are going to want to use the sample slope  $b_1$  to learn about the parameter we care about, the population slope  $\beta_1$ . And, we will use the sample intercept  $b_0$  to learn about the population intercept  $\beta_0$ .

In order to draw any conclusions about the population parameters  $\beta_0$  and  $\beta_1$ , we have to make a few more assumptions about the behavior of the data in a regression setting. We can get a pretty good feel for the assumptions by looking at our plot of gpa against college entrance test scores.

First, notice that when we connected the averages of the college entrance test scores for each of the subpopulations, it formed a line. Most often, we will not have the population of data at our disposal as we pretend to do here. If we didn't, do you think it would be reasonable to assume that the mean college entrance test scores are **linearly related** to high school grade point averages?



Again, let's focus on just one subpopulation, those students who have a gpa of 1, say. Notice that most of the college entrance scores for these students are clustered near the mean of 6, but a few students did much better than the subpopulation's average scoring around a 9, and a few students did a bit worse scoring about a 3. Do you get the picture? Thinking instead about the errors,  $\epsilon_i$ , most of the errors for these students are clustered near the mean of 0, but a few are as high as 3 and a few are as low as -3. If you could draw a probability curve for the errors above this subpopulation of data, what kind of a curve do you think it would be? Does it seem reasonable to assume that the errors for each subpopulation are **normally distributed**?

Looking at the plot again, notice that the spread of the college entrance test scores for students whose gpa is 1 is similar to the spread of the college entrance test scores for students whose gpa is 2, 3, and 4. Similarly, the spread of the errors is similar, no matter the gpa. Does it seem reasonable to assume that the errors for each subpopulation have **equal variance**?

Does it also seem reasonable to assume that the error for one student's college entrance test score is **independent** of the error for another student's college entrance test score? I'm sure you can come up with some scenarios — cheating students, for example — for which this assumption would not hold, but if you take a random sample from the population, it should be an assumption that is easily met.

We are now ready to summarize the four conditions or assumptions that underlie "**the simple linear regression model**:"

- The mean of the response,  $E(Y_i)$ , at each value of the predictor,  $x_i$ , is a **Linear function** of the  $x_i$ .
- The errors,  $\epsilon_i$ , are **Independent**.
- The errors,  $\epsilon_i$ , at each value of the predictor,  $x_i$ , are **Normally distributed**.
- The errors,  $\epsilon_i$ , at each value of the predictor,  $x_i$ , have **Equal variances** (denoted  $\sigma^2$ ).

Do you notice what the first letters that are colored in blue spell? "**LINE**." And, what are we studying in this course? Lines! Get it? You might find this mnemonic a useful way to remember the four conditions that make up what we call the "simple linear regression model." Whenever you hear "simple linear regression model," think of these four conditions!

An equivalent way to think of the first (linearity) condition is that the mean of the error,  $E(\epsilon_i)$ , at each value of the predictor,  $x_i$ , is **zero**. An alternative way to describe all four assumptions is that the errors,  $\epsilon_i$ , are independent normal random variables with mean zero and constant variance,  $\sigma^2$ .

---

<a href="#">◀ 2.2 - What is the "Best Fitting Line"? (/stat462/node/92)</a>	<a href="#">up (/stat462/node/79)</a>	<a href="#">2.4 - What is the Common Error Variance? ▶ (/stat462/node/94)</a>
---	---	---

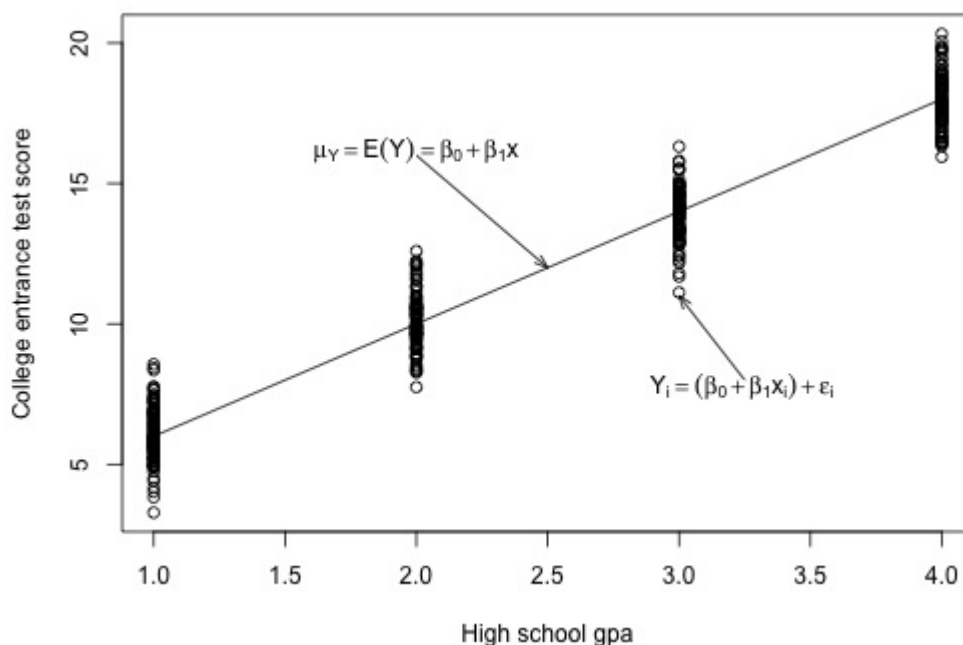
---

## STAT 462

## Applied Regression Analysis

## 2.4 - What is the Common Error Variance?

The plot of our population of data suggests that the college entrance test scores for each subpopulation have equal variance. We denote the value of this common variance as  $\sigma^2$ .

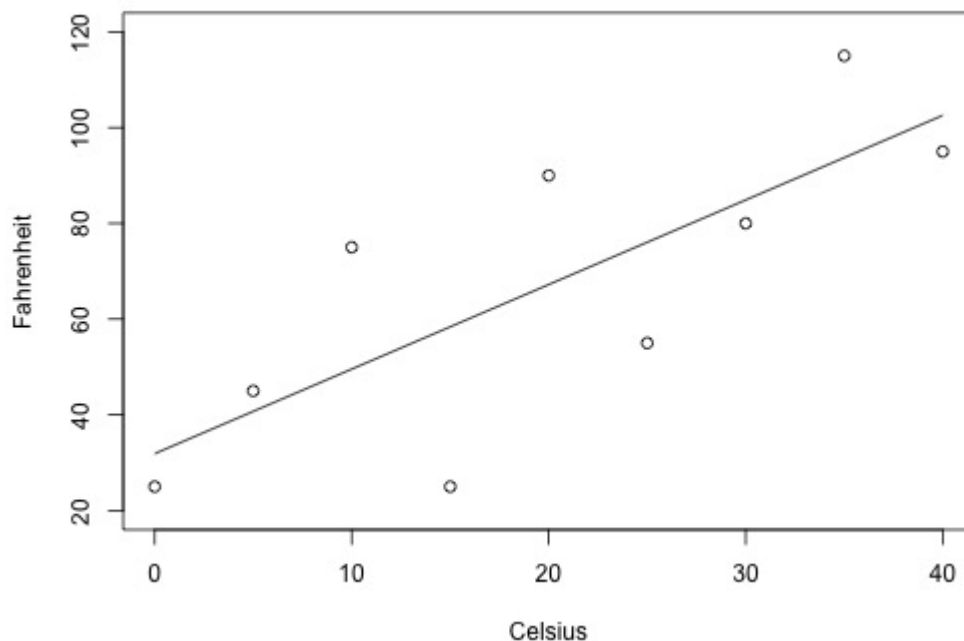


That is,  $\sigma^2$  quantifies how much the responses ( $y$ ) vary around the (unknown) mean population regression line  $\mu_Y = E(Y) = \beta_0 + \beta_1 x$ .

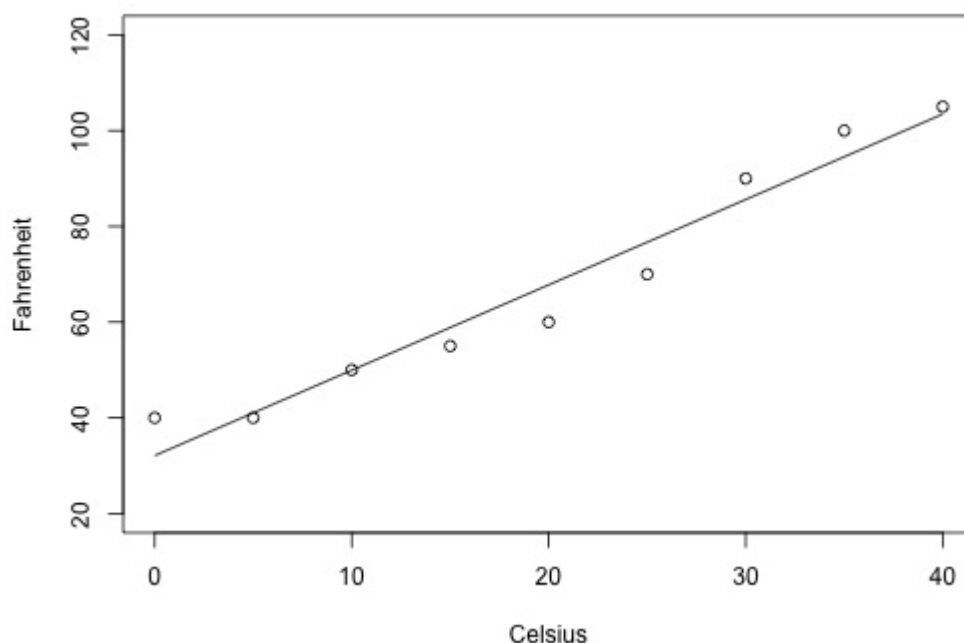
Why should we care about  $\sigma^2$ ? The answer to this question pertains to the most common use of an estimated regression line, namely predicting some future response.

Suppose you have two brands (A and B) of thermometers, and each brand offers a Celsius thermometer and a Fahrenheit thermometer. You measure the temperature in Celsius and Fahrenheit using each brand of thermometer on ten different days. Based on the resulting data, you obtain two estimated regression lines — one for brand A and one for brand B. You plan to use the estimated regression lines to predict the temperature in Fahrenheit based on the temperature in Celsius.

Will this thermometer brand (A) yield more precise future predictions ...?



... or this one (B)?



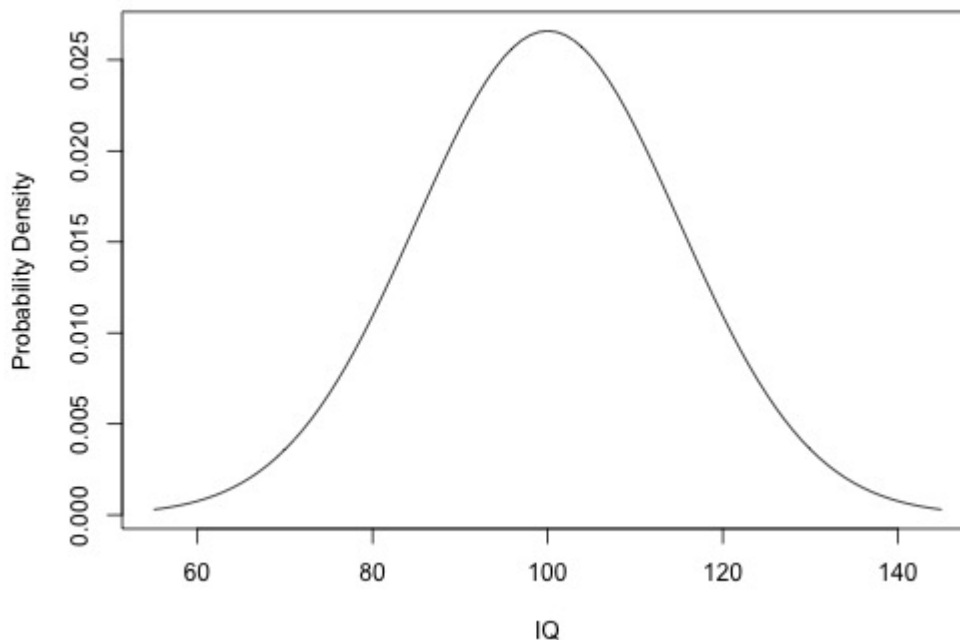
As the two plots illustrate, the Fahrenheit responses for the brand B thermometer don't deviate as far from the estimated regression equation as they do for the brand A thermometer. If we use the brand B estimated line to predict the Fahrenheit temperature, our prediction should never really be too far off from the actual observed Fahrenheit temperature. On the other hand, predictions of the Fahrenheit temperatures using the brand A thermometer can deviate quite a bit from the actual observed Fahrenheit temperature. Therefore, the brand B thermometer should yield more precise future predictions than the brand A thermometer.

To get an idea, therefore, of how precise future predictions would be, we need to know how much the responses ( $y$ ) vary around the (unknown) mean population regression line  $\mu_Y = E(Y) = \beta_0 + \beta_1 x$ . As stated earlier,  $\sigma^2$  quantifies this variance in the responses. Will we ever know this value  $\sigma^2$ ? No! Because  $\sigma^2$  is a population parameter, we will rarely know its true value. The best we can do is estimate it!



To understand the formula for the estimate of  $\sigma^2$  in the simple linear regression setting, it is helpful to recall the formula for the estimate of the variance of the responses,  $\sigma^2$ , when there is only one population.

The following is a plot of a population of IQ measurements. As the plot suggests, the average of the IQ measurements in the population is 100. But, how much do the IQ measurements vary from the mean? That is, how "spread out" are the IQs?

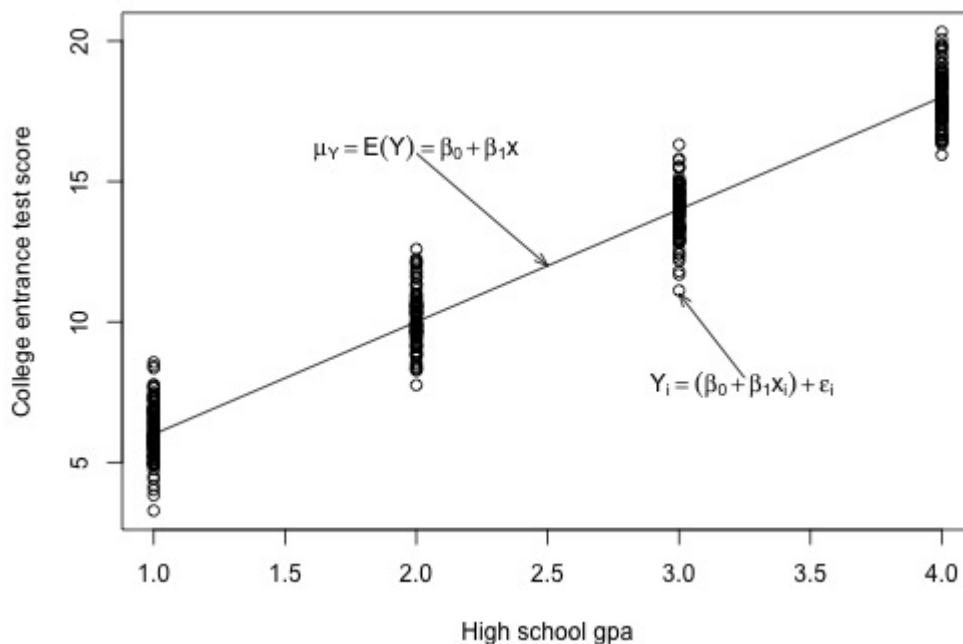


The **sample variance**:

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

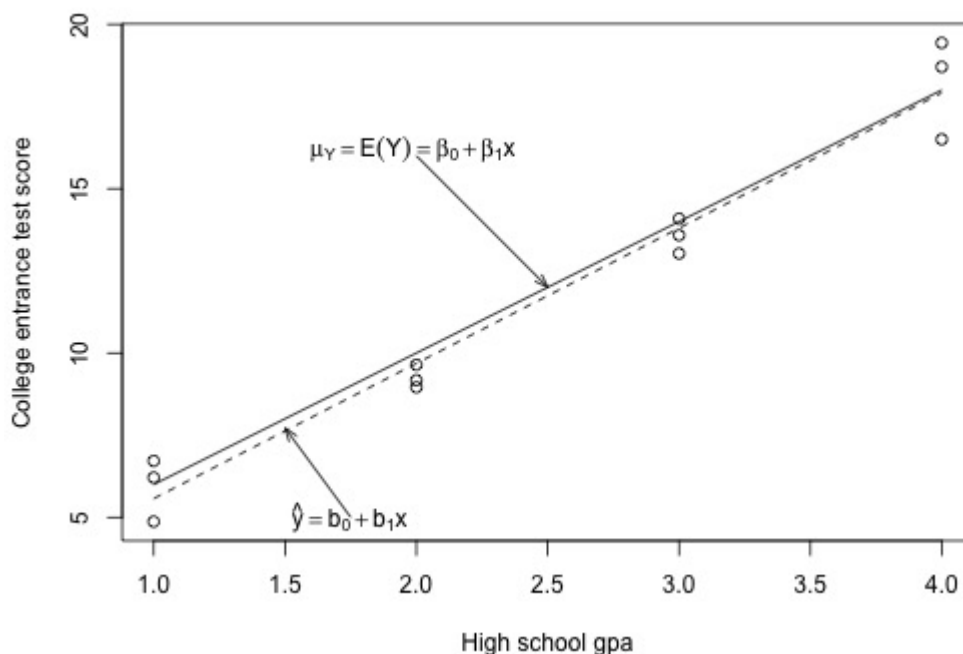
estimates  $\sigma^2$ , the variance of the one population. The estimate is really close to being like an average. The numerator adds up how far each response  $y_i$  is from the estimated mean  $\bar{y}$  in squared units, and the denominator divides the sum by  $n-1$ , not  $n$  as you would expect for an average. What we would really like is for the numerator to add up, in squared units, how far each response  $y_i$  is from the unknown population mean  $\mu$ . But, we don't know the population mean  $\mu$ , so we estimate it with  $\bar{y}$ . Doing so "costs us one degree of freedom". That is, we have to divide by  $n-1$ , and not  $n$ , because we estimated the unknown population mean  $\mu$ .

Now let's extend this thinking to arrive at an estimate for the population variance  $\sigma^2$  in the simple linear regression setting. Recall that we assume that  $\sigma^2$  is the same for each of the subpopulations. For our example on college entrance test scores and grade point averages, how many subpopulations do we have?



There are four subpopulations depicted in this plot. In general, there are as many subpopulations as there are distinct  $x$  values in the population. Each subpopulation has its own mean  $\mu_Y$ , which depends on  $x$  through

$\mu_Y = E(Y) = \beta_0 + \beta_1 x$ . And, each subpopulation mean can be estimated using the estimated regression equation  $\hat{y}_i = b_0 + b_1 x_i$ :



The **mean square error**:

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

estimates  $\sigma^2$ , the common variance of the many subpopulations.

How does the mean square error formula differ from the sample variance formula? The similarities are more striking than the differences. The numerator again adds up, in squared units, how far each response  $y_i$  is from its estimated

mean. In the regression setting, though, the estimated mean is  $\hat{y}_i$ . And, the denominator divides the sum by  $n-2$ , not  $n-1$ , because in using  $\hat{y}_i$  to estimate  $\mu_Y$ , we effectively estimate two parameters — the population intercept  $\beta_0$  and the population slope  $\beta_1$ . That is, we lose two degrees of freedom.

In practice, we will let statistical software, such as R or Minitab, calculate the mean square error ( $MSE$ ) for us. For the sample of 12 high school GPAs and college test scores,

$$MSE = \frac{8.678}{12 - 2} = 0.8678.$$

As well as displaying  $MSE$ , software typically also displays  $S = \sqrt{MSE}$ , which estimates  $\sigma$  and is known as the *regression standard error* or the *residual standard error*. For the sample of 12 high school GPAs and college test scores,  $S = \sqrt{0.8678} = 0.9315$ .

---

◀ 2.3 - The Simple Linear Regression Model  
(/stat462/node/93)

up  
(/stat462/node/79)

2.5 - The Coefficient of Determination, r-squared ▶ (/stat462/node/95)

---

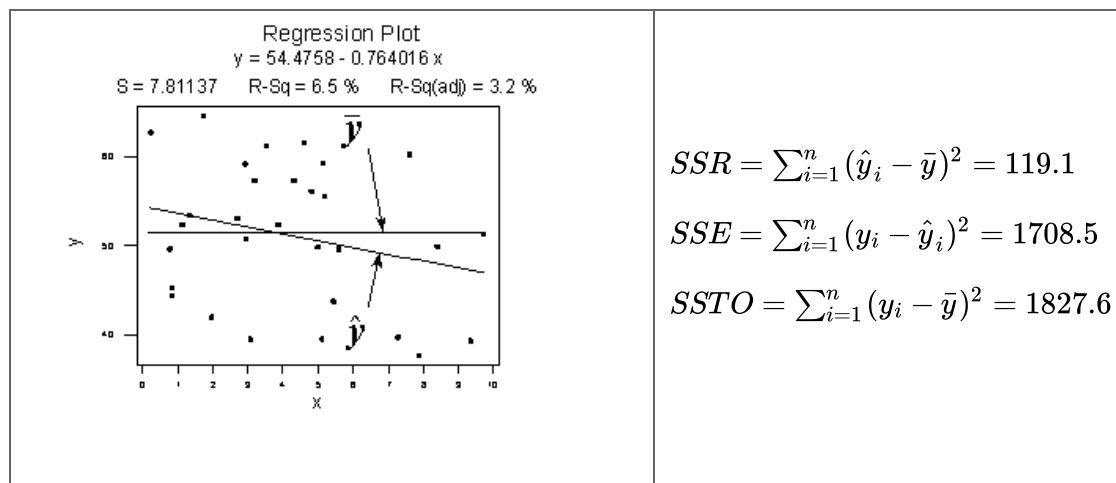
## STAT 462

## Applied Regression Analysis

## 2.5 - The Coefficient of Determination, r-squared

Let's start our investigation of the coefficient of determination,  $r^2$ , by looking at two different examples — one example in which the relationship between the response  $y$  and the predictor  $x$  is very weak and a second example in which the relationship between the response  $y$  and the predictor  $x$  is fairly strong. If our measure is going to work well, it should be able to distinguish between these two very different situations.

Here's a plot illustrating a very weak relationship between  $y$  and  $x$ . There are two lines on the plot, a horizontal line placed at the average response,  $\bar{y}$ , and a shallow-sloped estimated regression line,  $\hat{y}$ . Note that the slope of the estimated regression line is not very steep, suggesting that as the predictor  $x$  increases, there is not much of a change in the average response  $y$ . Also, note that the data points do not "hug" the estimated regression line:

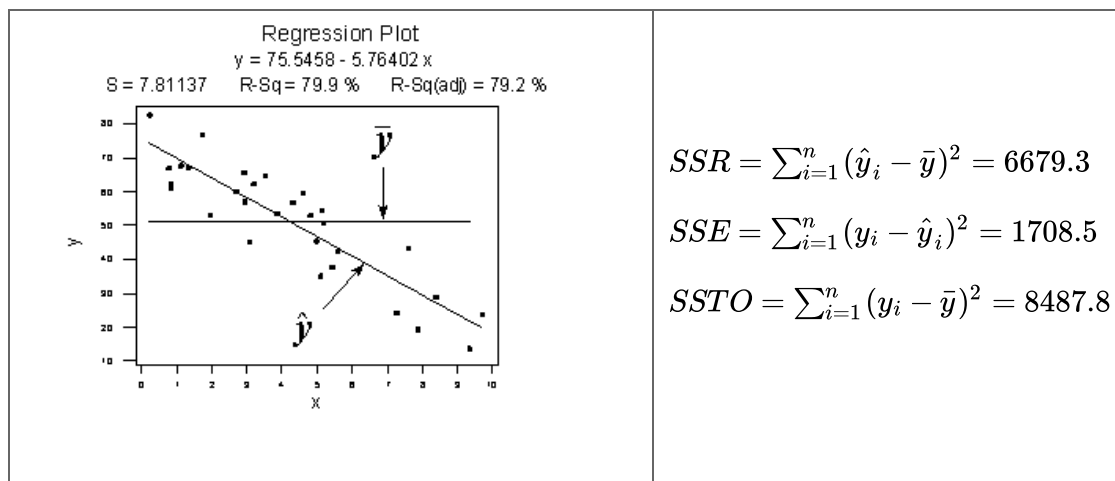


The calculations on the right of the plot show contrasting "sums of squares" values:

- SSR is the "regression sum of squares" and quantifies how far the estimated sloped regression line,  $\hat{y}_i$ , is from the horizontal "no relationship line," the sample mean or  $\bar{y}$ .
- SSE is the "error sum of squares" and quantifies how much the data points,  $y_i$ , vary around the estimated regression line,  $\hat{y}_i$ .
- SSTO is the "total sum of squares" and quantifies how much the data points,  $y_i$ , vary around their mean,  $\bar{y}$ .

Note that  $SSTO = SSR + SSE$ . The sums of squares appear to tell the story pretty well. They tell us that most of the variation in the response  $y$  ( $SSTO = 1827.6$ ) is just due to random variation ( $SSE = 1708.5$ ), not due to the regression of  $y$  on  $x$  ( $SSR = 119.1$ ). You might notice that  $SSR$  divided by  $SSTO$  is  $119.1/1827.6$  or  $0.065$ . Do you see where this quantity appears on the above fitted line plot?

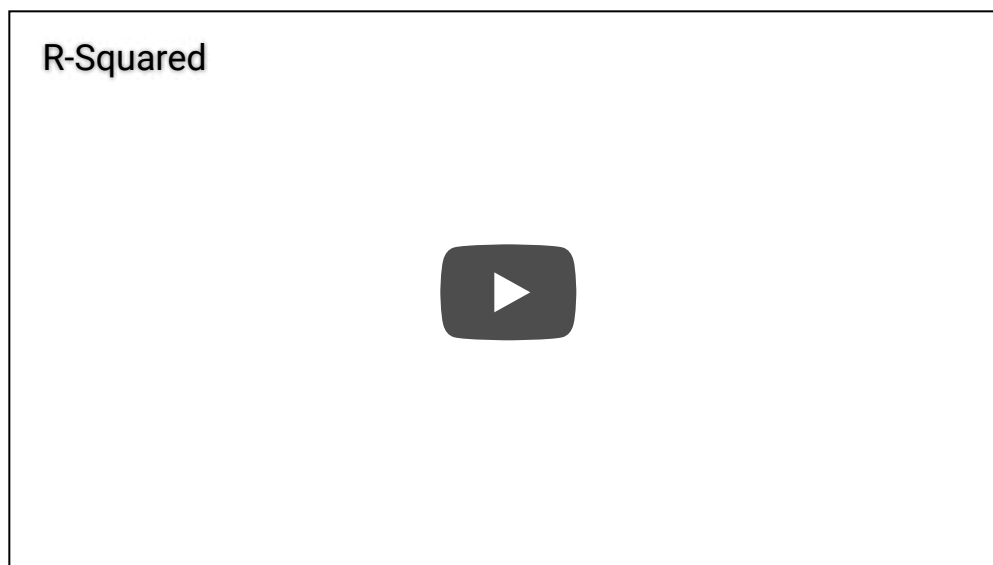
Contrast the above example with the following one in which the plot illustrates a fairly convincing relationship between  $y$  and  $x$ . The slope of the estimated regression line is much steeper, suggesting that as the predictor  $x$  increases, there is a fairly substantial change (decrease) in the response  $y$ . And, here, the data points do "hug" the estimated regression line:



The sums of squares for this dataset tell a very different story, namely that most of the variation in the response  $y$  ( $SSTO = 8487.8$ ) is due to the regression of  $y$  on  $x$  ( $SSR = 6679.3$ ) not just due to random error ( $SSE = 1708.5$ ). And,  $SSR$  divided by  $SSTO$  is  $6679.3/8487.8$  or  $0.799$ , which again appears on the fitted line plot.

The previous two examples have suggested how we should define the measure formally. In short, the "**coefficient of determination**" or "**r-squared value**," denoted  $r^2$ , is the regression sum of squares divided by the total sum of squares. Alternatively, as demonstrated in this screencast below, since  $SSTO = SSR + SSE$ , the quantity  $r^2$  also equals one minus the ratio of the error sum of squares to the total sum of squares:

$$r^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$



Here are some basic characteristics of the measure:

- Since  $r^2$  is a proportion, it is always a number between 0 and 1.

- If  $r^2 = 1$ , all of the data points fall perfectly on the regression line. The predictor  $x$  accounts for *all* of the variation in  $y$ !
- If  $r^2 = 0$ , the estimated regression line is perfectly horizontal. The predictor  $x$  accounts for *none* of the variation in  $y$ !

We've learned the interpretation for the two easy cases — when  $r^2 = 0$  or  $r^2 = 1$  — but, how do we interpret  $r^2$  when it is some number between 0 and 1, like 0.23 or 0.57, say? Here are two similar, yet slightly different, ways in which the coefficient of determination  $r^2$  can be interpreted. We say either:

" $r^2 \times 100$  percent of the variation in  $y$  is reduced by taking into account predictor  $x$ "

or:

" $r^2 \times 100$  percent of the variation in  $y$  is "explained by" the variation in predictor  $x$ ."

Many statisticians prefer the first interpretation. I tend to favor the second. The risk with using the second interpretation — and hence why "explained by" appears in quotes — is that it can be misunderstood as suggesting that the predictor  $x$  *causes* the change in the response  $y$ . Association is not causation. That is, just because a dataset is characterized by having a large  $r$ -squared value, it does not imply that  $x$  *causes* the changes in  $y$ . As long as you keep the correct meaning in mind, it is fine to use the second interpretation. A variation on the second interpretation is to say, " $r^2 \times 100$  percent of the variation in  $y$  is accounted for by the variation in predictor  $x$ ."

Students often ask: "what's considered a large  $r$ -squared value?" It depends on the research area. Social scientists who are often trying to learn something about the huge variation in human behavior will tend to find it very hard to get  $r$ -squared values much above, say 25% or 30%. Engineers, on the other hand, who tend to study more exact systems would likely find an  $r$ -squared value of just 30% unacceptable. The moral of the story is to read the literature to learn what typical  $r$ -squared values are for your research area!

Let's revisit the skin cancer mortality example (skincancer.txt

(/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/skincancer.txt) ). Any statistical software that performs simple linear regression analysis will report the  $r$ -squared value for you, which in this case is 67.98% or 68% to the nearest whole number.

We can say that 68% of the variation in the skin cancer mortality rate is reduced by taking into account latitude. Or, we can say — with knowledge of what it really means — that 68% of the variation in skin cancer mortality is "explained by" latitude.

---

◀ 2.4 - What is the Common Error Variance?  
(/stat462/node/94)

up  
(/stat462/node/79)

2.6 - (Pearson) Correlation Coefficient  $r$  ▶  
(/stat462/node/96)

---

# STAT 462

## Applied Regression Analysis

### 2.6 - (Pearson) Correlation Coefficient $r$

The correlation coefficient  $r$  is directly related to the coefficient of determination  $r^2$  in the obvious way. If  $r^2$  is represented in decimal form, e.g. 0.39 or 0.87, then all we have to do to obtain  $r$  is to take the square root of  $r^2$ :

$$r = \pm \sqrt{r^2}$$

The sign of  $r$  depends on the sign of the estimated slope coefficient  $b_1$ :

- If  $b_1$  is negative, then  $r$  takes a negative sign.
- If  $b_1$  is positive, then  $r$  takes a positive sign.

That is, the estimated slope and the correlation coefficient  $r$  always share the same sign. Furthermore, because  $r^2$  is always a number between 0 and 1, the correlation coefficient  $r$  is always a number between -1 and 1.

One advantage of  $r$  is that it is unitless, allowing researchers to make sense of correlation coefficients calculated on different data sets with different units. The "unitless-ness" of the measure can be seen from an alternative formula for  $r$ , namely:

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

If  $x$  is the height of an individual measured in inches and  $y$  is the weight of the individual measured in pounds, then the units for the numerator is inches  $\times$  pounds. Similarly, the units for the denominator is inches  $\times$  pounds. Because they are the same, the units in the numerator and denominator cancel each other out, yielding a "unitless" measure.

Another formula for  $r$  that you might see in the regression literature is one that illustrates how the correlation coefficient  $r$  is a function of the estimated slope coefficient  $b_1$ :

$$r = \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \times b_1$$

We are readily able to see from this version of the formula that:

- The estimated slope  $b_1$  of the regression line and the correlation coefficient  $r$  always share the same sign. If you don't see why this must be true, view this screencast.

Same sign



- The correlation coefficient  $r$  is a unitless measure. If you don't see why this must be true, view this screencast.

Unitless



- If the estimated slope  $b_1$  of the regression line is 0, then the correlation coefficient  $r$  must also be 0.

That's enough with the formulas! As always, we will let statistical software such as R or Minitab do the dirty calculations for us. For the skin cancer mortality and latitude example (skincancer.txt (<https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/skincancer.txt>)), the correlation between skin cancer mortality and latitude is -0.825. It doesn't matter the order in which you specify the variables, so the correlation between latitude and skin cancer mortality is also -0.825. What does this correlation coefficient tells us? That is, how do we interpret the Pearson correlation coefficient  $r$ ? In general, there is no nice practical operational interpretation for  $r$  as there is for  $r^2$ . You can only use  $r$  to make a statement about the strength of the linear relationship between  $x$  and  $y$ . In general:

- If  $r = -1$ , then there is a perfect negative linear relationship between  $x$  and  $y$ .
- If  $r = 1$ , then there is a perfect positive linear relationship between  $x$  and  $y$ .
- If  $r = 0$ , then there is no linear relationship between  $x$  and  $y$ .

All other values of  $r$  tell us that the relationship between  $x$  and  $y$  is not perfect. The closer  $r$  is to 0, the weaker the linear relationship. The closer  $r$  is to -1, the stronger the negative linear relationship. And, the closer  $r$  is to 1, the



stronger the positive linear relationship. As is the case for the  $r^2$  value, what is deemed a "large" correlation coefficient  $r$  value depends greatly on the research area.

So, what does the correlation of -0.825 between skin cancer mortality and latitude tell us? It tells us:

- The relationship is negative. As the latitude increases, the skin cancer mortality rate decreases (linearly).
- The relationship is quite strong (since the value is pretty close to -1)

---

<a href="#">◀ 2.5 - The Coefficient of Determination, r-squared (/stat462/node/95)</a>	<a href="#">up (/stat462/node/79)</a>	<a href="#">2.7 - Coefficient of Determination and Correlation Examples ▶ (/stat462/node/97)</a>
--	---------------------------------------	--

---

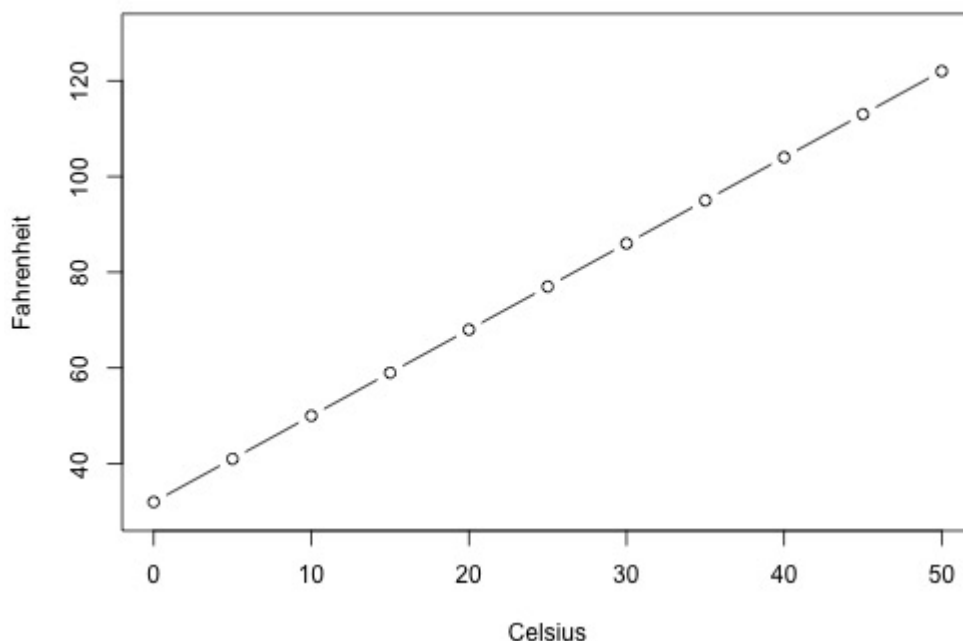
## STAT 462

## Applied Regression Analysis

## 2.7 - Coefficient of Determination and Correlation Examples

Let's take a look at some examples so we can get some practice interpreting the coefficient of determination  $r^2$  and the correlation coefficient  $r$ .

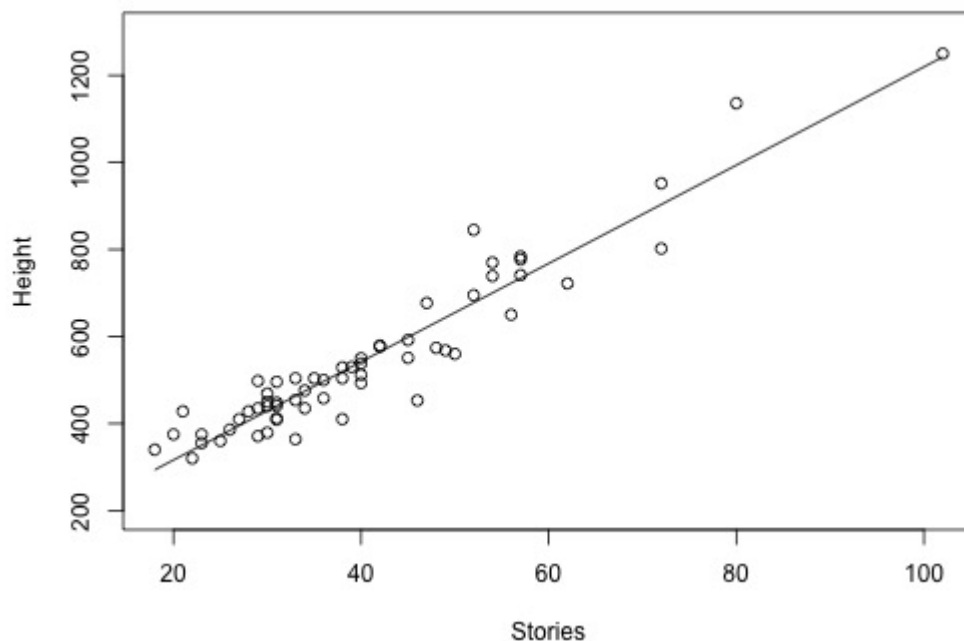
**Example 1.** How strong is the linear relationship between temperatures in Celsius and temperatures in Fahrenheit? Here's a plot of an estimated regression equation based on  $n = 11$  data points:



Statistical software reports that  $r^2 = 100\%$  and  $r = 1.000$ . Both measures tell us that there is a perfect linear relationship between temperature in degrees Celsius and temperature in degrees Fahrenheit. We know that the relationship is perfect, namely that  $\text{Fahrenheit} = 32 + 1.8 \times \text{Celsius}$ . It should be no surprise then that  $r^2$  tells us that 100% of the variation in temperatures in Fahrenheit is explained by the temperature in Celsius.

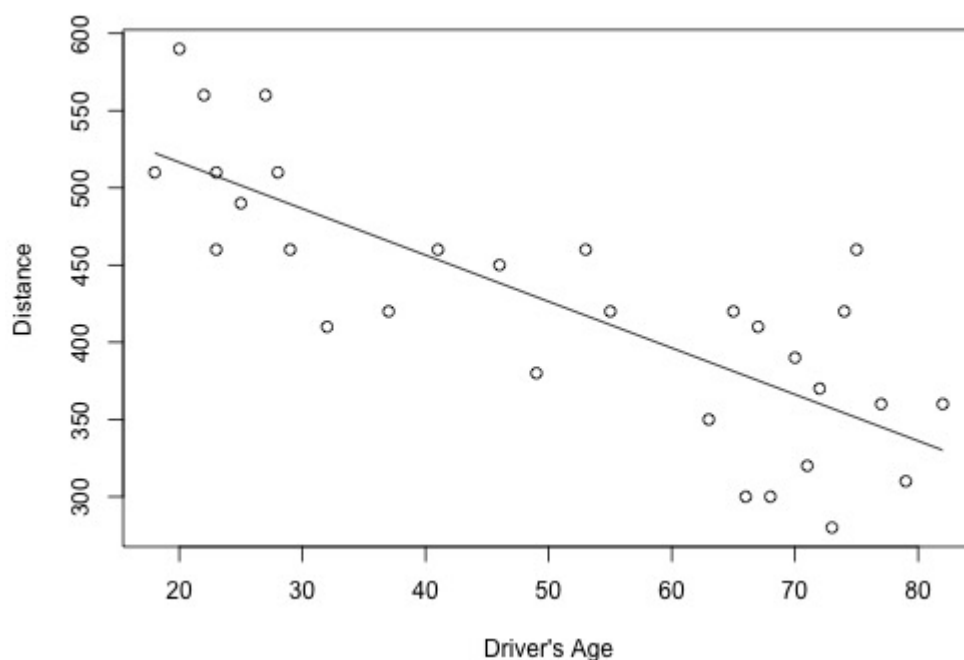
**Example 2.** How strong is the linear relationship between the number of stories a building has and its height? One would think that as the number of stories increases, the height would increase, but not perfectly. Some statisticians compiled data on a set of  $n = 60$  buildings reported in the 1994 World Almanac (bldgstories.txt

(/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/bldgstories.txt)). Statistical software reports  $r^2 = 90.4\%$  and  $r = 0.951$  and produced the following plot:



The positive sign of  $r$  tells us that the relationship is positive — as number of stories increases, height increases — as we expected. Because  $r$  is close to 1, it tells us that the linear relationship is very strong, but not perfect. The  $r^2$  value tells us that 90.4% of the variation in the height of the building is explained by the number of stories in the building.

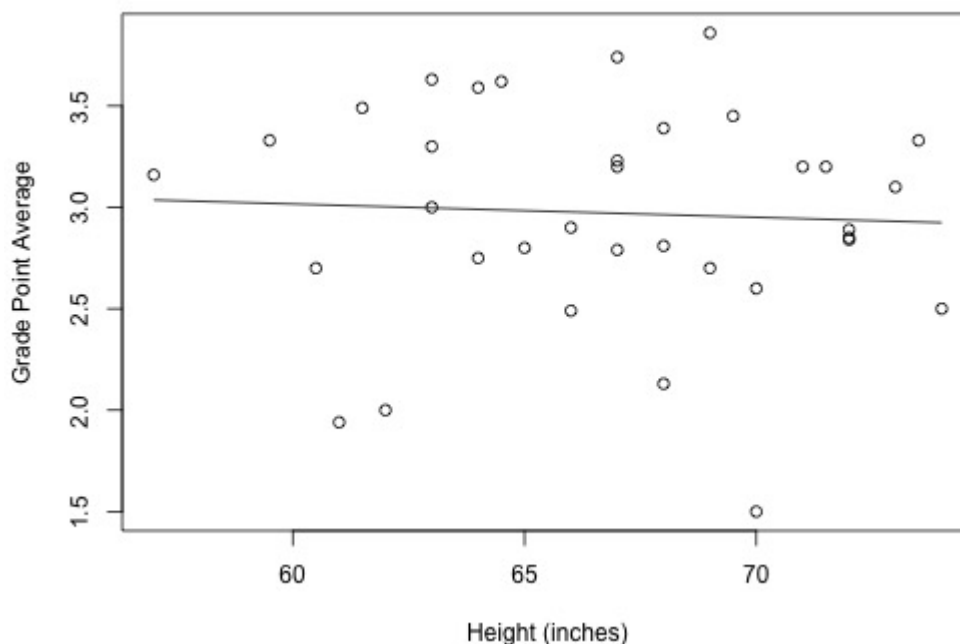
**Example 3.** How strong is the linear relationship between the age of a driver and the distance the driver can see? If we had to guess, we might think that the relationship is negative — as age increases, the distance decreases. A research firm (Last Resource, Inc., Bellefonte, PA) collected data on a sample of  $n = 30$  drivers (signdist.txt (/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/signdist.txt) ). Statistical software reports that reports that  $r^2 = 64.2\%$  and  $r = -0.801$  and produced the following output:



The negative sign of  $r$  tells us that the relationship is negative — as driving age increases, seeing distance decreases — as we expected. Because  $r$  is fairly close to -1, it tells us that the linear relationship is fairly strong, but not

perfect. The  $r^2$  value tells us that 64.2% of the variation in the seeing distance is reduced by taking into account the age of the driver.

**Example 4.** How strong is the linear relationship between the height of a student and his or her grade point average? Data were collected on a random sample of  $n = 35$  students in a statistics course at Penn State University (heightgpa.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/heightgpa.txt) ). Statistical software reports that  $r^2 = 0.3\%$  and  $r = -0.053$  and produced the following output:



Because  $r$  is quite close to 0, it suggests — not surprisingly, I hope — that there is next to no linear relationship between height and grade point average. Indeed, the  $r^2$  value tells us that only 0.3% of the variation in the grade point averages of the students in the sample can be explained by their height. In short, we would need to identify another more important variable, such as number of hours studied, if predicting a student's grade point average is important to us.

---

◀ 2.6 - (Pearson) Correlation Coefficient  $r$   
(/stat462/node/96)

up  
(/stat462/node/79)

2.8 - R-squared Cautions ▶ (/stat462/node/98)

---

## STAT 462

## Applied Regression Analysis

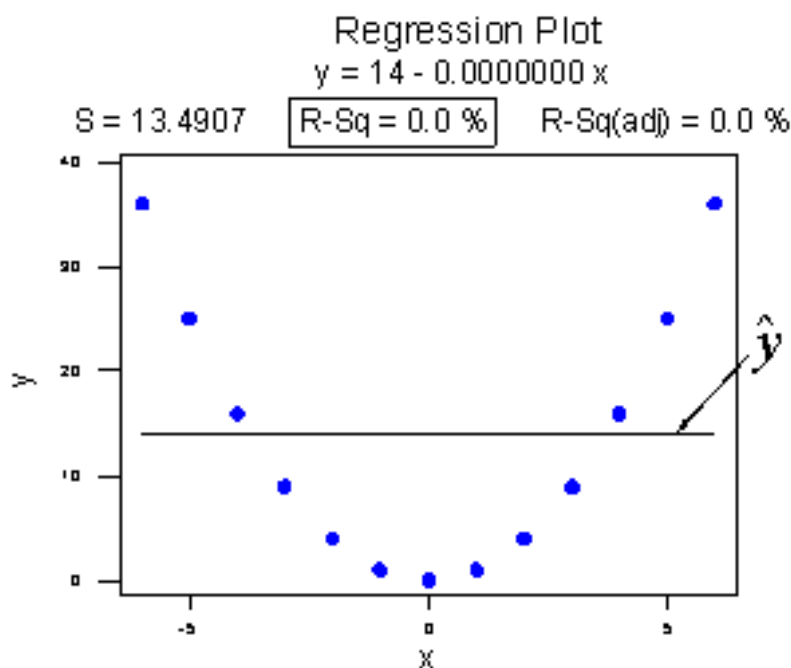
## 2.8 - R-squared Cautions

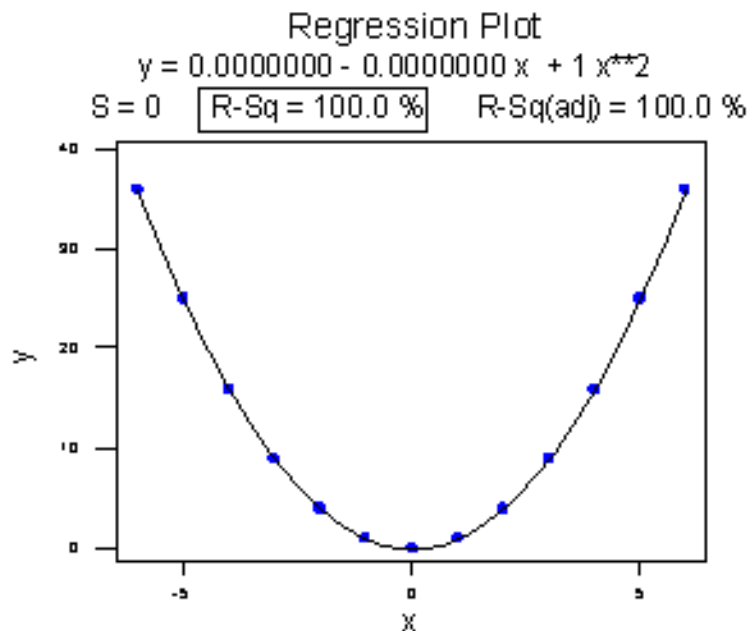
Unfortunately, the coefficient of determination  $r^2$  and the correlation coefficient  $r$  have to be the most often misused and misunderstood measures in the field of statistics. To ensure that you don't fall victim to the most common mistakes, we review a set of seven different cautions here. Master these and you'll be a master of the measures!

## Cautions # 1

**The coefficient of determination  $r^2$  and the correlation coefficient  $r$  quantify the strength of a *linear* relationship. It is possible that  $r^2 = 0\%$  and  $r = 0$ , suggesting there is no linear relation between  $x$  and  $y$ , and yet a perfect curved (or "curvilinear" relationship) exists.**

Consider the following example. The upper plot illustrates a perfect, although curved, relationship between  $x$  and  $y$ , and yet  $r^2 = 0\%$  and  $r = 0$ . The estimated regression line is perfectly horizontal with slope  $b_1 = 0$ . If you didn't understand that  $r^2$  and  $r$  summarize the strength of a *linear* relationship, you would likely misinterpret the measures, concluding that there is no relationship between  $x$  and  $y$ . But, it's just not true! There is indeed a relationship between  $x$  and  $y$  — it's just not linear.





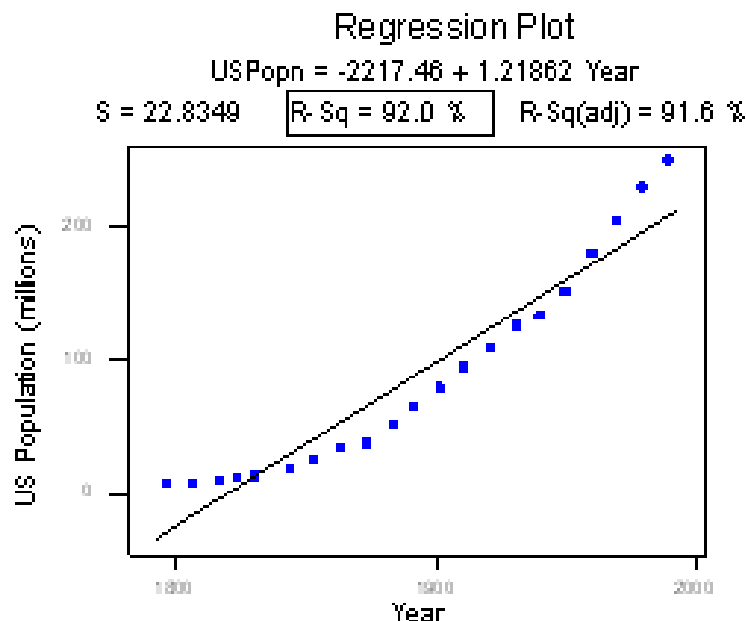
The lower plot better reflects the curved relationship between  $x$  and  $y$ . There is a "quadratic" curve through the data for which  $R^2 = 100\%$ . What is this all about? We'll learn when we study multiple linear regression later in the course that the coefficient of determination  $r^2$  associated with the simple linear regression model for one predictor extends to a "multiple coefficient of determination," denoted  $R^2$ , for the multiple linear regression model with more than one predictor. (The lowercase  $r$  and uppercase  $R$  are used to distinguish between the two situations. Statistical software typically doesn't distinguish between the two, calling both measures " $R^2$ .") The interpretation of  $R^2$  is similar to that of  $r^2$ , namely " $R^2 \times 100\%$  of the variation in the response is explained by the predictors in the regression model (which may be curvilinear)."

In summary, the  $R^2$  value of 100% and the  $r$  value of 0 tell the story of the second plot perfectly. The multiple coefficient of determination  $R^2 = 100\%$  tells us that all of the variation in the response  $y$  is explained in a curved manner by the predictors  $x$  and  $x^2$ . The correlation coefficient  $r = 0$  tells us that if there is a relationship between  $x$  and  $y$ , it is not linear.

## Caution # 2

**A large  $r^2$  value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.**

Consider the following example in which the relationship between year (1790 to 1990, by decades) and population of the United States (in millions) is examined:



The correlation between year and population is 0.959. This and the  $r^2$  value of 92.0% suggest a strong linear relationship between year and U.S. population. Indeed, only 8% of the variation in U.S. population is left to explain after taking into account the year in a linear way! The plot suggests, though, that a curve would describe the relationship even better. That is, the large  $r^2$  value of 92.0% should not be interpreted as meaning that the estimated regression line fits the data well. (Its large value does suggest that taking into account year is better than not doing so. It just doesn't tell us that we could still do better.)

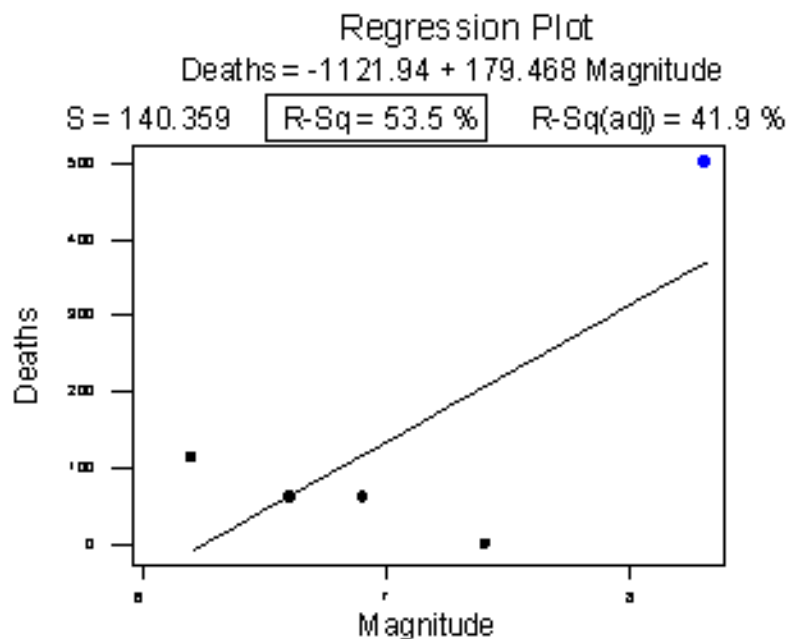
Again, the  $r^2$  value doesn't tell us that the regression model fits the data well. This is the most common misuse of the  $r^2$  value! When you are reading the literature in your research area, pay close attention to how others interpret  $r^2$ . I am confident that you will find some authors misinterpreting the  $r^2$  value in this way. And, when you are analyzing your own data make sure you plot the data — 99 times out of a 100, the plot will tell more of the story than a simple summary measure like  $r$  or  $r^2$  ever could.

### Caution # 3

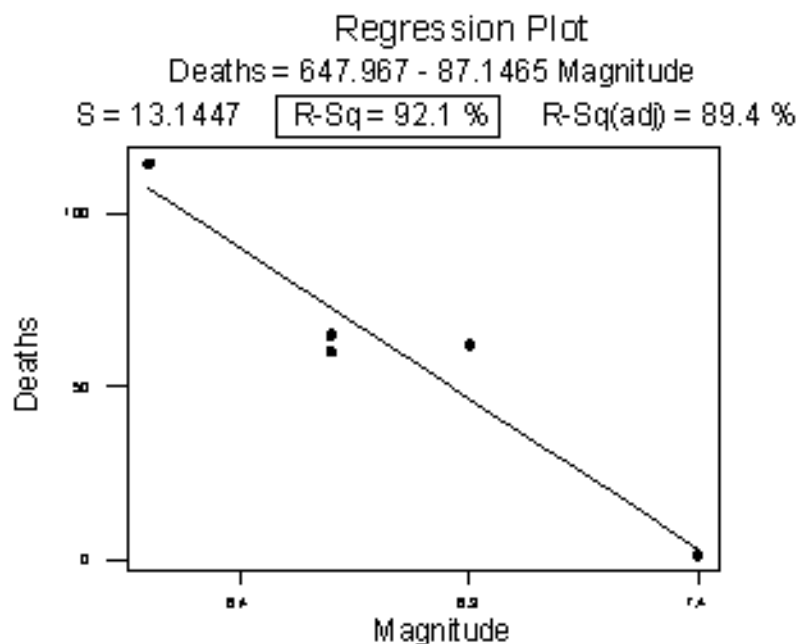
**The coefficient of determination  $r^2$  and the correlation coefficient  $r$  can both be greatly affected by just one data point (or a few data points).**

Consider the following example in which the relationship between the number of deaths in an earthquake and its magnitude is examined. Data on  $n = 6$  earthquakes were recorded, and the upper fitted line plot below was obtained. The correlation between deaths and magnitude is 0.732. The slope of the line  $b_1 = 179.5$  and the correlation of 0.732 suggest that as the magnitude of the earthquake increases, the number of deaths also increases. This is not a surprising result. Therefore, if we hadn't plotted the data, we wouldn't notice that one and only one data point (magnitude = 8.3 and deaths = 503) was making the values of the slope and the correlation positive.

#### Original plot



### Plot with unusual point removed



The second plot is a plot of the same data, but with the one unusual data point removed. The correlation between deaths and magnitude with the one unusual point removed is -0.960. Note that the estimated slope of the line changes from a positive 179.5 to a negative 87.1 — just by removing one data point. Also, both measures of the strength of the linear relationship improve dramatically —  $r$  changes from a positive 0.732 to a negative 0.960, and  $r^2$  changes from 53.5% to 92.1%.

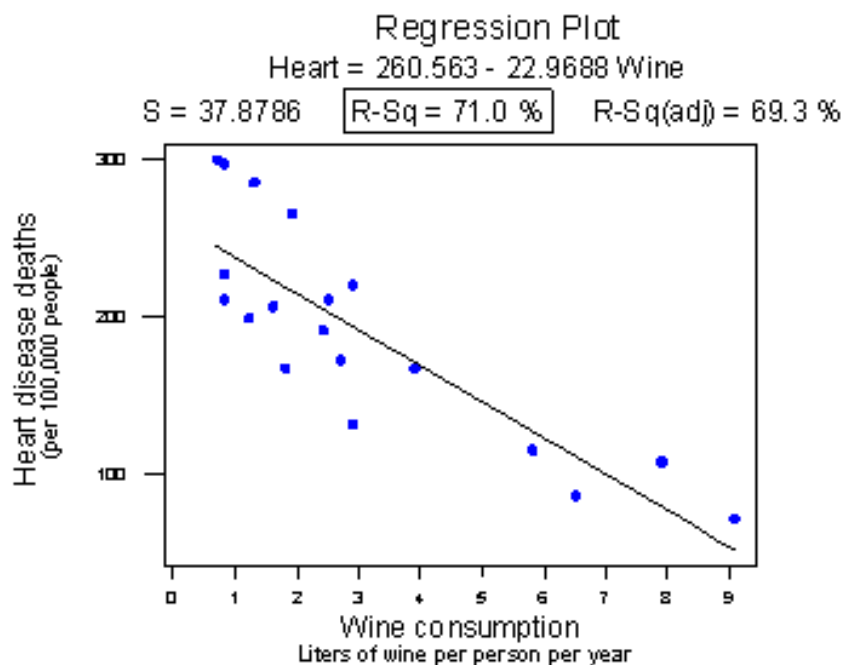
What conclusion can we draw from these data? Probably none! The main point of this example was to illustrate the impact of one data point on the  $r$  and  $r^2$  values. One could argue that a secondary point of the example is that a data set can be too small to draw any useful conclusions.



## Caution # 4

### Correlation (or association) does not imply causation.

Consider the following example in which the relationship between wine consumption and death due to heart disease is examined. Each data point represents one country. For example, the data point in the lower right corner is France, where the consumption averages 9.1 liters of wine per person per year and deaths due to heart disease are 71 per 100,000 people.



Statistical software reports that the  $r^2$  value is 71.0% and the correlation is -0.843. Based on these summary measures, a person might be tempted to conclude that he or she should drink more wine, since it *reduces* the risk of heart disease. If only life were that simple! Unfortunately, there may be other differences in the behavior of the people in the various countries that really explain the differences in the heart disease death rates, such as diet, exercise level, stress level, social support structure and so on.

Let's push this a little further. Recall the distinction between an experiment and an observational study:

- An **experiment** is a study in which, when collecting the data, the researcher controls the values of the predictor variables.
- An **observational study** is a study in which, when collecting the data, the researcher merely observes and records the values of the predictor variables as they happen.

The primary advantage of conducting experiments is that one can typically conclude that differences in the predictor values is what *caused* the changes in the response values. This is not the case for observational studies.

Unfortunately, most data used in regression analyses arise from observational studies. Therefore, you should be careful not to overstate your conclusions, as well as be cognizant that others may be overstating their conclusions.

## Caution # 5

**Ecological correlations — correlations that are based on rates or averages — tend to overstate the strength of an association.**

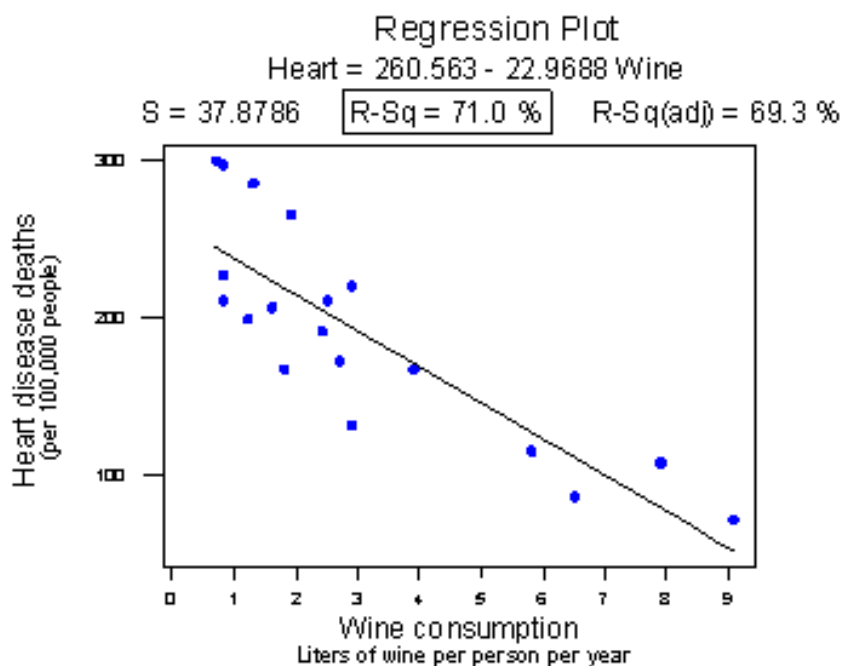
Some statisticians (Freedman, Pisani, Purves, 1997) investigated data from the 1988 Current Population Survey in order to illustrate the inflation that can occur in ecological correlations. Specifically, they considered the relationship between a man's level of education and his income. They calculated the correlation between education and income in two ways:

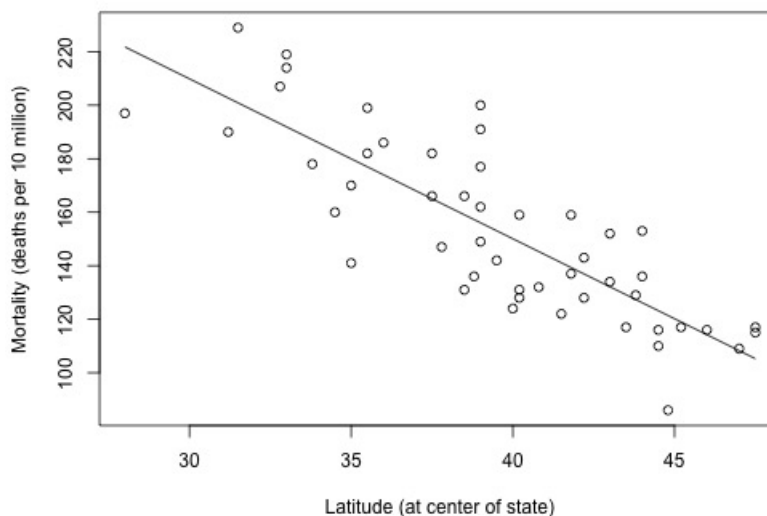
- First, they treated individual men, aged 25-64, as the experimental units. That is, each data point represented a man's income and education level. Using these data, they determined that the correlation between income and education level for men aged 25-64 was about 0.4, not a convincingly strong relationship.
- The statisticians analyzed the data again, but in the second go-around they treated nine *geographical regions* as the units. That is, they first computed the average income and average education for men aged 25-64 in each of the nine regions. They determined that the correlation between the average income and average education for the sample of  $n = 9$  regions was about 0.7, obtaining a much larger correlation than that obtained on the individual data.

Again, ecological correlations, such as the one calculated on the region data, tend to overstate the strength of an association. How do you know what kind of data to use — aggregate data (such as the regional data) or individual data? It depends on the conclusion you'd like to make.

If you want to learn about the strength of the association between an individual's education level and his income, then by all means you should use individual, not aggregate, data. On the other hand, if you want to learn about the strength of the association between a school's average salary level and the schools graduation rate, you should use aggregate data in which the units are the schools.

We hadn't taken note of it at the time, but you've already seen a couple of examples in which ecological correlations were calculated on aggregate data:





The correlation between wine consumption and heart disease deaths of -0.843 is an ecological correlation. The units are countries, not individuals. The correlation between skin cancer mortality and state latitude of -0.825 is also an ecological correlation. The units are states, again not individuals. In both cases, we should not use these correlations to try to draw a conclusion about how an *individual's* wine consumption or suntanning behavior will affect their *individual risk* of dying from heart disease or skin cancer. We shouldn't try to draw such conclusions anyway, because "association is not causation."

## Caution # 6

**A "statistically significant"  $r^2$  value does not imply that the slope  $\beta_1$  is meaningfully different from 0.**

This caution is a little strange as we haven't talked about any hypothesis tests yet. We'll get to that soon, but before doing so ... a number of former students have asked why some article authors can claim that two variables are "significantly associated" with a  $P$ -value less than 0.01, but yet their  $r^2$  value is small, such as 0.09 or 0.16. The answer has to do with the mantra that you may recall from your introductory statistics course: "statistical significance does not imply practical significance."

In general, the larger the data set, the easier it is to reject the null hypothesis and claim "statistical significance." If the data set is very large, it is even possible to reject the null hypothesis and claim that the slope  $\beta_1$  is not 0, even when it is not practically or meaningfully different from 0. That is, it is possible to get a significant  $P$ -value when  $\beta_1$  is 0.13, a quantity that is likely not to be considered meaningfully different from 0 (of course, it does depend on the situation and the units). Again, the mantra is "statistical significance does not imply practical significance."

## Caution # 7

**A large  $r^2$  value does not necessarily mean that a useful prediction of the response  $y_{\text{new}}$ , or estimation of the mean response  $\mu_Y$ , can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.**

We'll learn more about such prediction and confidence intervals in Lesson 4.



# STAT 462

## Applied Regression Analysis

### 2.9 - Simple Linear Regression Examples

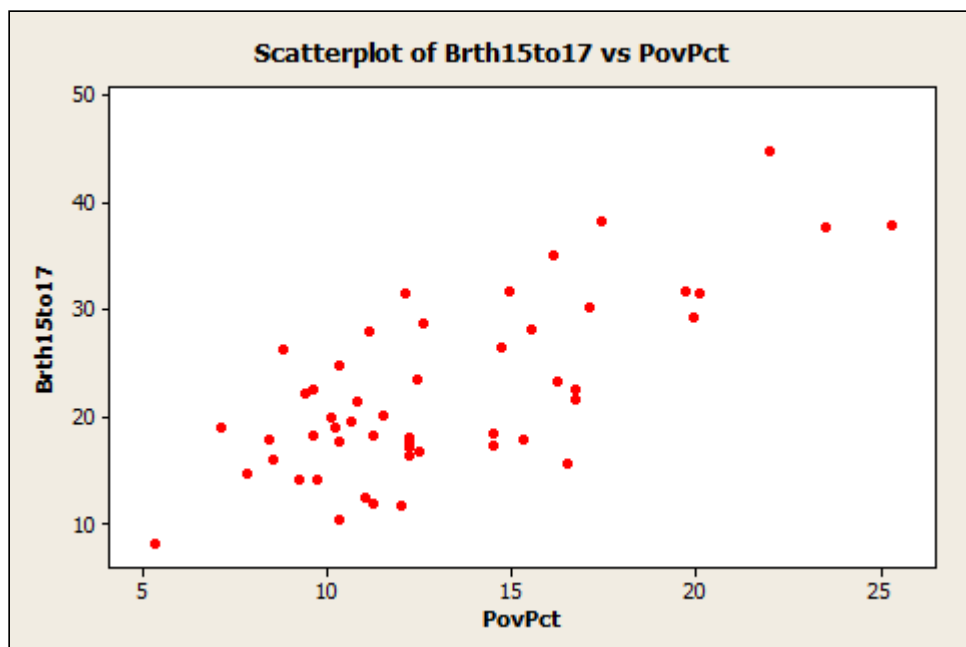
#### Example 1: Teen Birth Rate and Poverty Level Data

This dataset of size  $n = 51$  are for the 50 states and the District of Columbia in the United States (poverty.txt

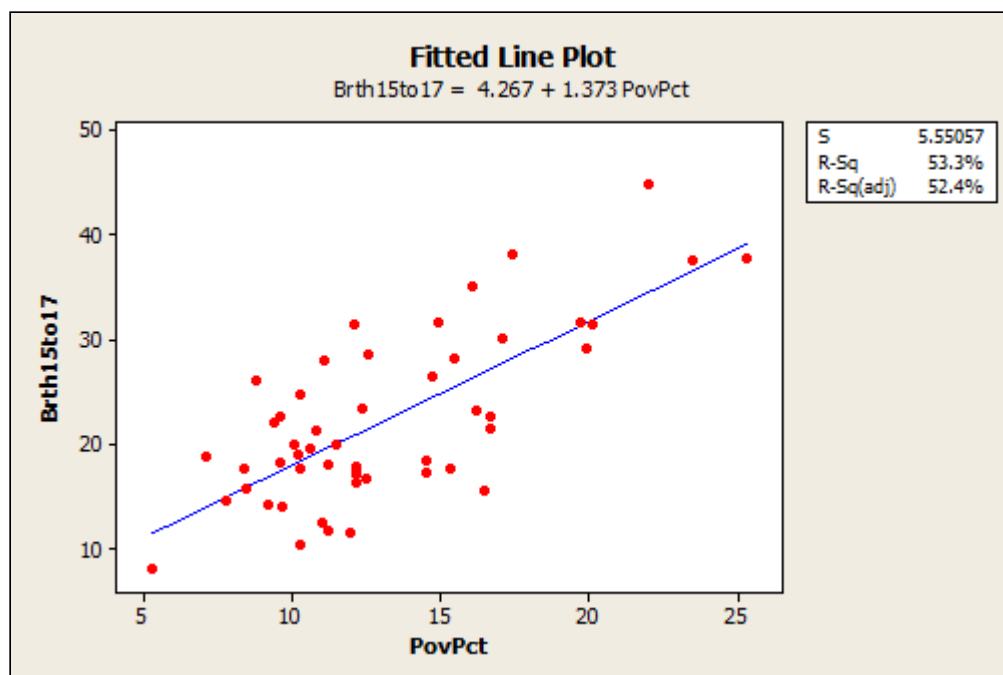


(/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/poverty.txt)). The variables are  $y$  = year 2002 birth rate per 1000 females 15 to 17 years old and  $x$  = poverty rate, which is the percent of the state's population living in households with incomes below the federally defined poverty level. (Data source: *Mind On Statistics*, 3rd edition, Utts and Heckard).

The plot of the data below (birth rate on the vertical) shows a generally linear relationship, on average, with a positive slope. As the poverty level increases, the birth rate for 15 to 17 year old females tends to increase as well.



The following plot shows a regression line superimposed on the data.



The equation of the fitted regression line is given near the top of the plot. The equation should really state that it is for the “average” birth rate (or “predicted” birth rate would be okay too) because a regression equation describes the average value of  $y$  as a function of one or more  $x$ -variables. In statistical notation, the equation could be written  $\hat{y} = 4.267 + 1.373x$ .

- The interpretation of the slope (value = 1.373) is that the 15 to 17 year old birth rate increases 1.373 units, on average, for each one unit (one percent) increase in the poverty rate.
- The interpretation of the intercept (value=4.267) is that if there were states with poverty rate = 0, the predicted average for the 15 to 17 year old birth rate would be 4.267 for those states. Since there are no states with poverty rate = 0 this interpretation of the intercept is not practically meaningful for this example.

In the graph with a regression line present, we also see the information that  $s = 5.55057$  and  $r^2 = 53.3\%$ .

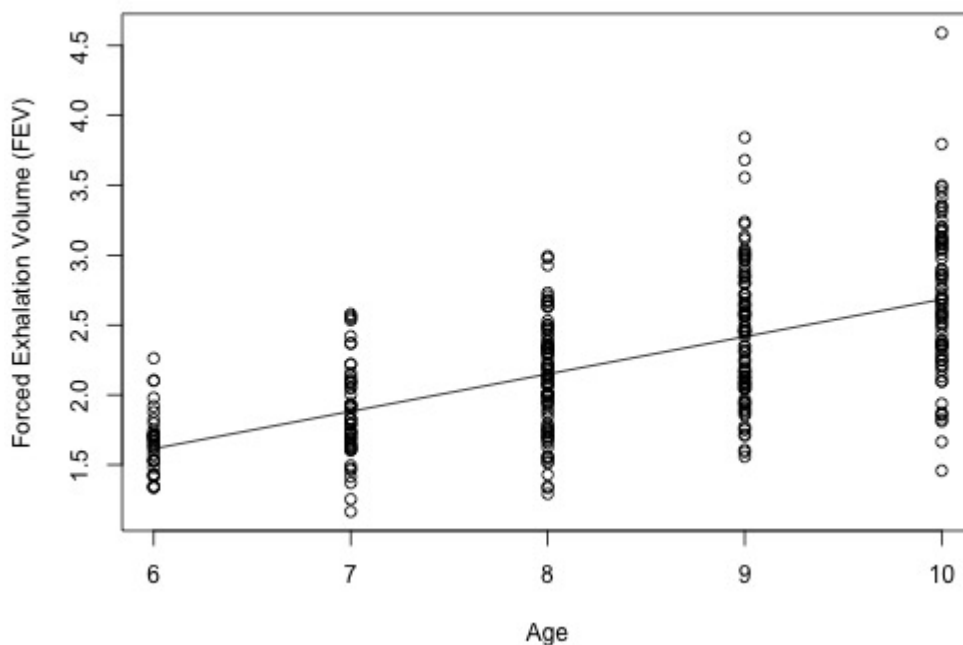
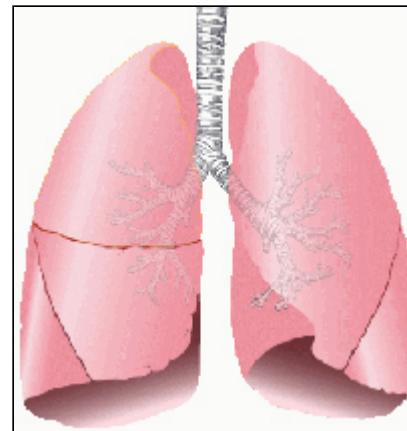
- The value of  $s$  tells us roughly the standard deviation of the differences between the  $y$ -values of individual observations and predictions of  $y$  based on the regression line.
- The value of  $r^2$  can be interpreted to mean that poverty rates “explain” 53.3% of the observed variation in the 15 to 17 year old average birth rates of the states.

The  $R^2$  (adj) value (52.4%) is an adjustment to  $R^2$  based on the number of  $x$ -variables in the model (only one here) and the sample size. With only one  $x$ -variable, the adjusted  $R^2$  is not important.

## Example 2: Lung Function in 6 to 10 Year Old Children

The data are from  $n = 345$  children between 6 and 10 years old. The variables are  $y$  = forced exhalation volume (FEV), a measure of how much air somebody can forcibly exhale from their lungs, and  $x$  = age in years. (Data source: The data here are a part of dataset given in Kahn, Michael (2005). “An Exhalent Problem for Teaching Statistics” (<http://www.amstat.org/publications/jse/v13n2/datasets.kahn.html>)”, *The Journal of Statistical Education*, 13(2).

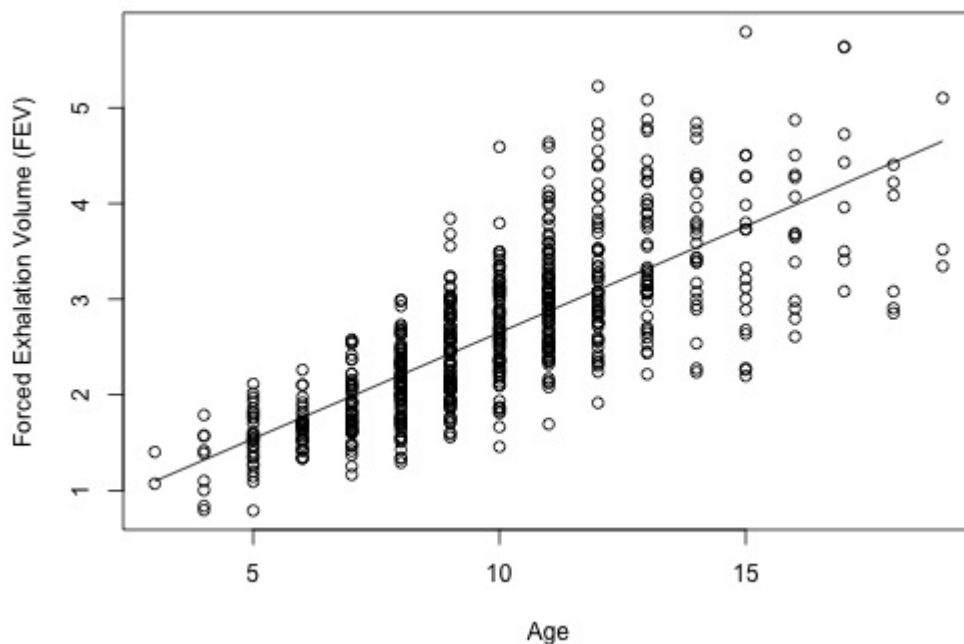
Below is a plot of the data with a simple linear regression line superimposed.



- The estimated regression equation is that average  $FEV = 0.01165 + 0.26721 \times \text{age}$ . For instance, for an 8 year old we can use the equation to estimate that the average  $FEV = 0.01165 + 0.26721 \times (8) = 2.15$ .
- The interpretation of the slope is that the average FEV increases 0.26721 for each one year increase in age (in the observed age range).

An interesting and possibly important feature of these data is that the variance of individual y-values from the regression line increases as age increases. This feature of data is called **non-constant variance**. For example, the FEV values of 10 year olds are more variable than FEV value of 6 year olds. This is seen by looking at the vertical ranges of the data in the plot. This may lead to problems using a simple linear regression model for these data, which is an issue we'll explore in more detail in Lesson 4.

Above, we only analyzed a subset of the entire dataset. The full dataset (fev\_dat.txt ([/stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/fev\\_dat.txt](http://stat462/sites/onlinecourses.science.psu.edu/stat462/files/data/fev_dat.txt))) is shown in the plot below:



As we can see, the range of ages now spans 3 to 19 years old and the estimated regression equation is  $FEV = 0.43165 + 0.22204 \times \text{age}$ . Both the slope and intercept have noticeably changed, but the variance still appears to be non-constant. This illustrates that it is important to be aware of how you are analyzing your data. If you only use a subset of your data that spans a shorter range of predictor values, then you could obtain noticeably different results than if you had used the full dataset.

---

◀ 2.8 - R-squared Cautions (/stat462/node/98)

up  
(/stat462/node/79)

---