Simple Linear Regression

Overview

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.

Jan 22 2026

Phuc-Loi Luu, PhD

Email: loilp@bvtn.org.vn

Zalo: 0901802182

INSTITUTE FOR APPLIED RESEARCH IN HEALTH SCIENCES AND AGING (ARiHA) **– THONG NHAT HOSPITAL**

Content

https://online.stat.psu.edu/stat501/lesson/1

Objectives

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

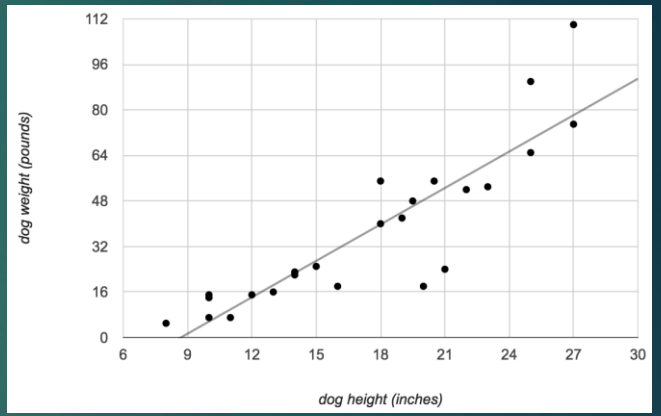
- Distinguish between a deterministic relationship and a statistical relationship.
- Understand the concept of the least squares criterion.
- Interpret the intercept b_0 and slope b_1 of an estimated regression equation.
- Know how to obtain the estimates b_0 and b_1 from Minitab's fitted line plot and regression analysis output.
- Recognize the distinction between a population regression line and the estimated regression line.
- Summarize the four conditions that comprise the simple linear regression model.
- Know what the unknown population variance σ^2 quantifies in the regression setting.
- Know how to obtain the estimated *MSE* of the unknown population variance σ^2 from Minitab's fitted line plot and regression analysis output.
- 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 a 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.

1.1 - What is Simple Linear Regression?

A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

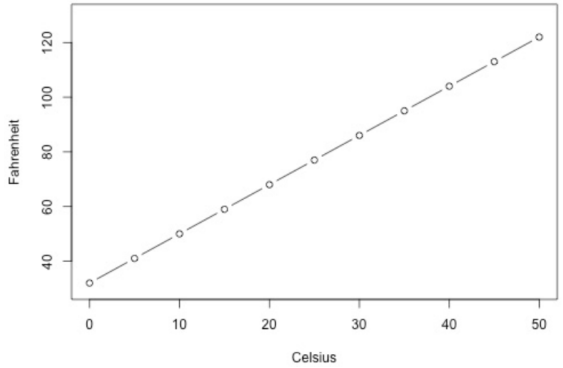
- One variable, denoted x (dog height), is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y (dog weight), is regarded as the response, outcome, or dependent variable.

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 in other arenas. 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: deterministic 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:

$$\text{Fahr} = \frac{9}{5}\text{Cels} + 32$$

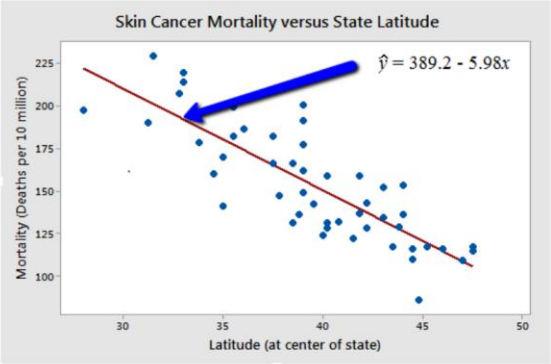
That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit *exactly*.

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

- Circumference = π × 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, α = constant for each gas, and *V* = volume of gas.

Types of relationships: statistical relationships

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 48 states in the United States (U.S. Skin Cancer data) (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.)

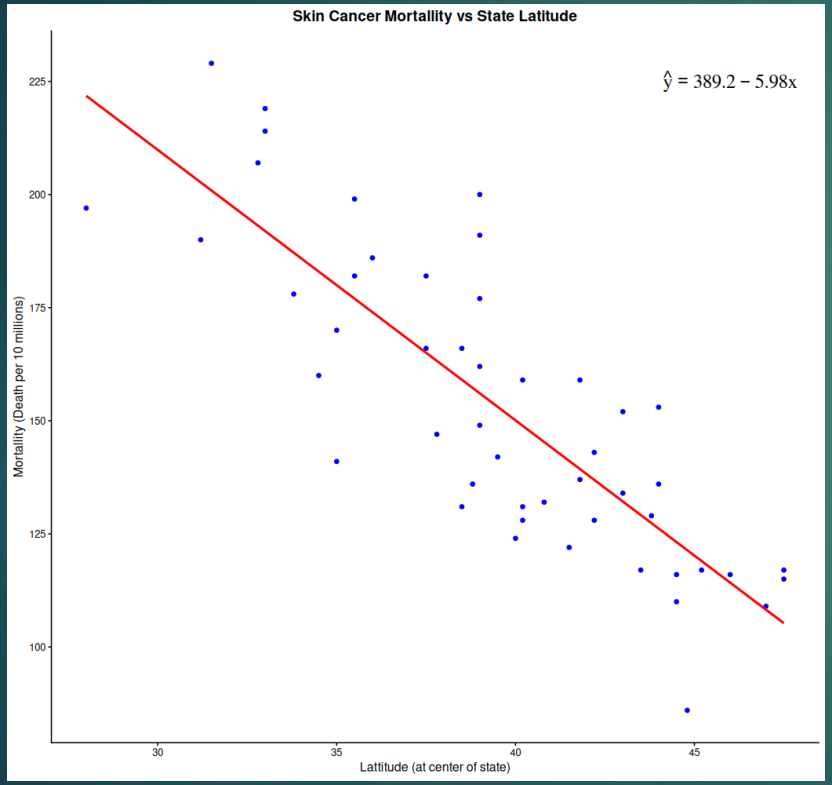


You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter plot supports such a hypothesis. There appears to be a negative linear relationship 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 the 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 the 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.

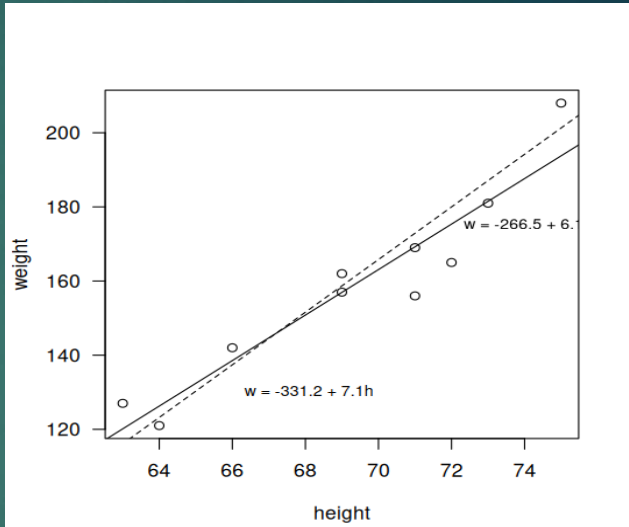
Types of relationships: statistical relationships




```
skin_cancer <- read.table ("US_Skin_cancer.tsv",header = T)

#Skin Cancer Mortallity versus State Lattitude
ggplot (skin_cancer, aes (x= Lat, y= Mort)) +
  geom_point (col= "blue") +
  labs (
    title = "Skin Cancer Mortallity vs State Latitude",
    x = "Lattitude (at center of state)",
    y = "Mortallity (Death per 10 millions)"
  ) +
  geom_smooth (method = "lm",col = "red",se = F) +
  theme_classic() +
  annotate("text", x = 46, y = 225,
           label = "hat(y) == 389.2 - 5.98*x",
           parse = TRUE, size = 6, family = "serif", color = "black")
+
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  scale_y_continuous(breaks = seq(100, 225, by = 25))
```

1.2 - 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 and Weight Dataset). 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?



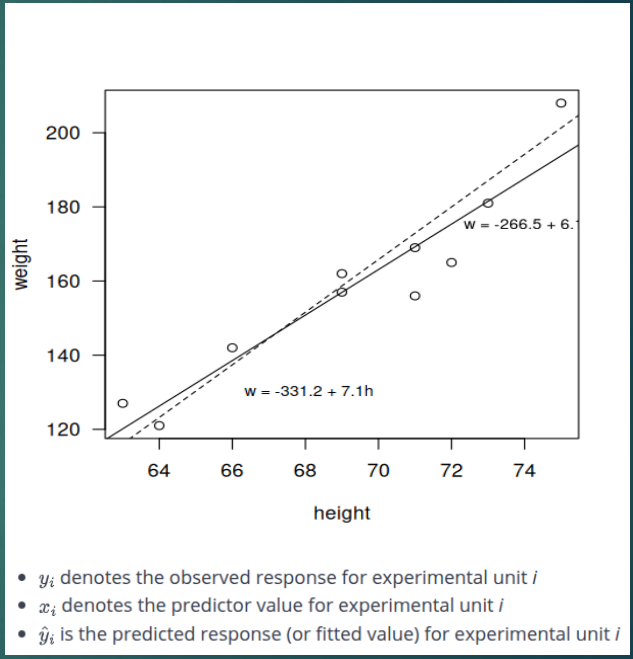
- y_i denotes the observed response for experimental unit *i*
- x_i denotes the predictor value for experimental unit *i*
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit *i*

THE BEST FITTING LINE EQUATION

$$\hat{y}_i = b_0 + b_1 x_i$$

1.2 - What is the "Best Fitting Line"?

```
#Code for the plot:
student <- read.table ("student_hw.tsv",header = T)
plot(student$ht, student$wt,
     pch = 1,
     xlab = "height",
     ylab = "weight",
     las = 1)
model1 <- lm(wt ~ ht, data = student)
abline(model1, lty = 1)

abline(a = -331.2, b = 7.1, lty = 2)

text(x = 66, y = 130, labels = "w = -331.2 + 7.1h", pos = 4, cex =
0.8)
text(x = 72, y = 175, labels = "w = -266.5 + 6.1h", pos = 4, cex =
0.8)
```



- 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*

1.2 - What is the "Best Fitting Line"?

- 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:

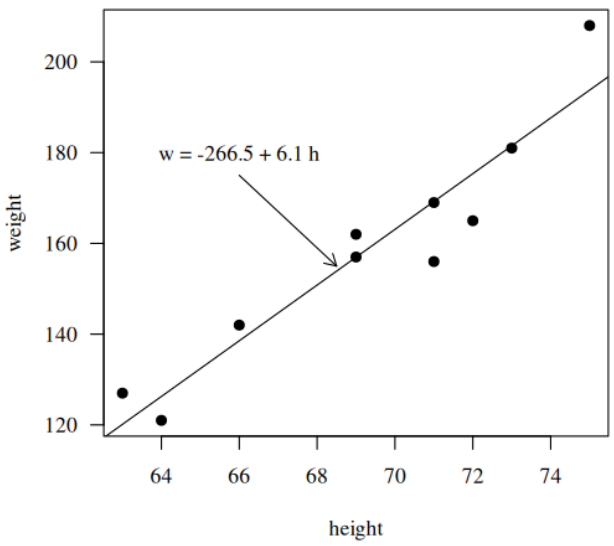
$$\hat{w} = -266.53 + 6.1376h.$$

- Note! that this line is just a more precise version of the above solid line

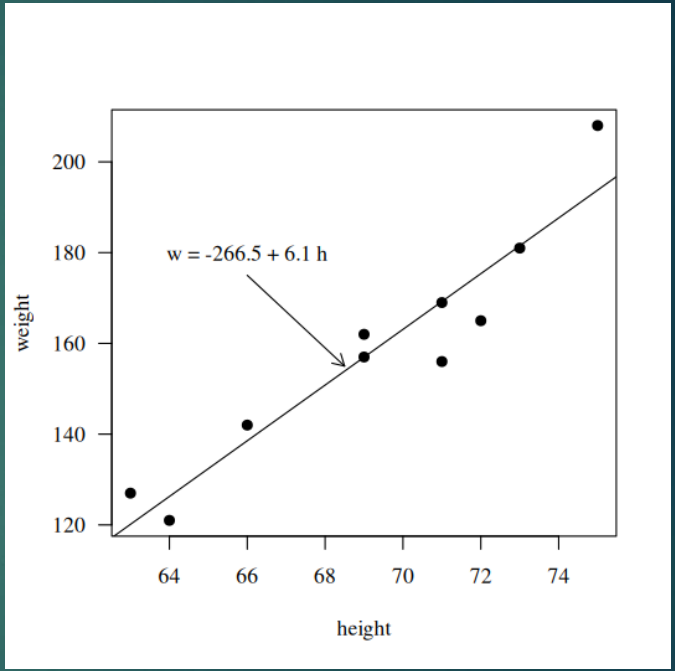
$$w = -266.5 + 6.1h$$

- 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 in 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



```
plot(student$ht, student$wt,
     pch = 19,
     xlab = "height",
     ylab = "weight",
     las = 1,
     family = "serif")

abline(lm(student$wt ~ student$ht), col = "black")

text(x = 66, y = 180, labels = "w = -266.5 + 6.1 h",
family = "serif")

arrows(x0 = 66, y0 = 175, x1 = 68.5, y1 = 155, length =
0.1)
```



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 which make the sum of the squared prediction errors the smallest they can be.
- That is, we need to find the values b_0 and b_1 that minimize:

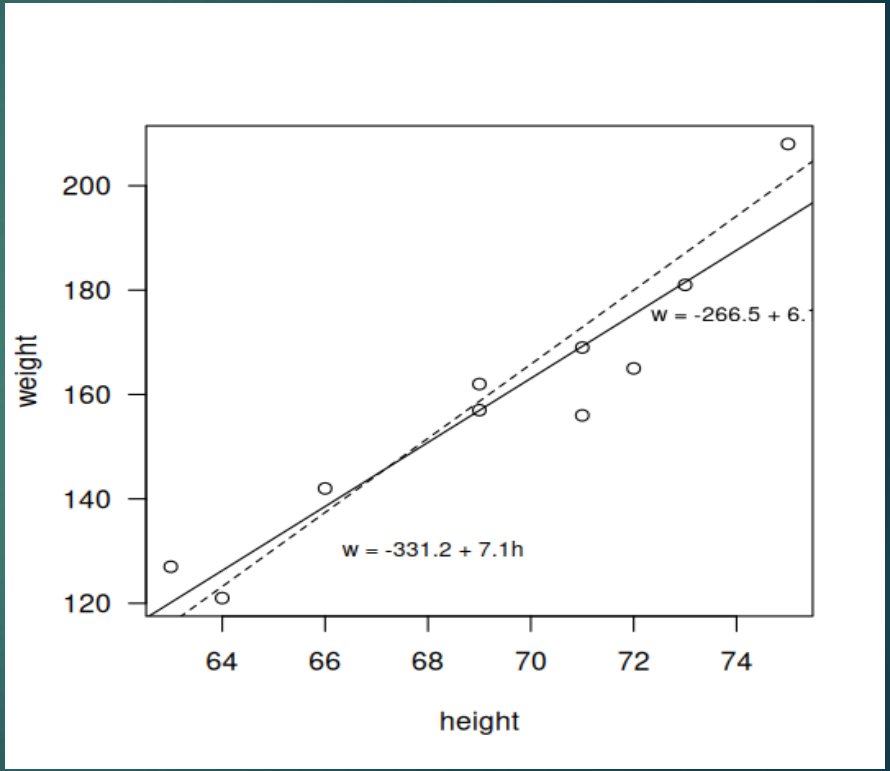
$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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?



$\hat{w} = -331.2 + 7.1h$ (the dashed line)

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
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					766.5

$\hat{w} = -266.53 + 6.1376h$ (the solid line)

i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					597.4

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.

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:

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

In practice, you won't really need to worry about the formulas for b_0 and b_1. Instead, you are are going to let statistical software, such as Minitab, find least squares lines for you. But, we can still learn something from the formulas — for b_1 in particular.

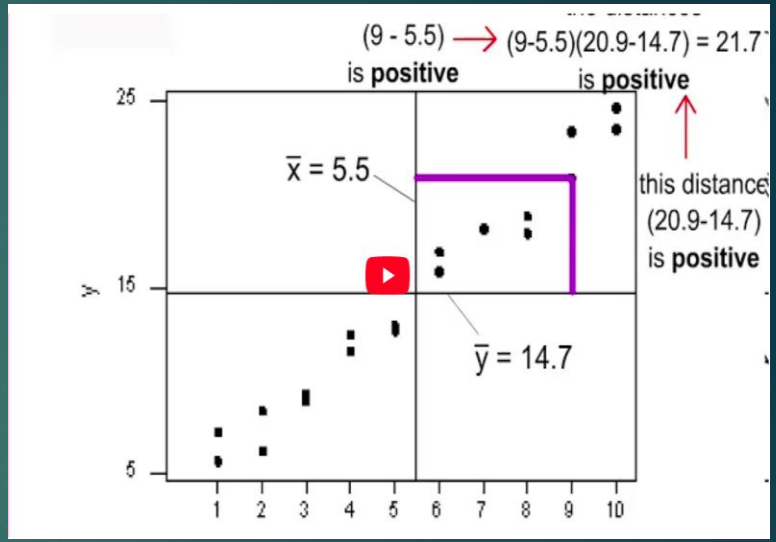
If you study the formula for the slope b_1:

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

you see that the denominator is necessarily positive since it only involves summing positive terms. Therefore, the sign of the slope b_1 is solely determined by the numerator. The numerator tells us, for each data point, to sum up the product of two distances — the distance of the *x*-value from the mean of all of the *x*-values and the distance of the *y*-value from the mean of all of the *y*-values. Let's see how this determines the sign of the slope b_1 by studying the following two plots.

- **When is the slope $b_1 > 0$**? Do you agree that the trend in the following video is positive — that is, as x increases, y tends to increase? If the trend is positive, then the slope b_1 must be positive. Let's see how!

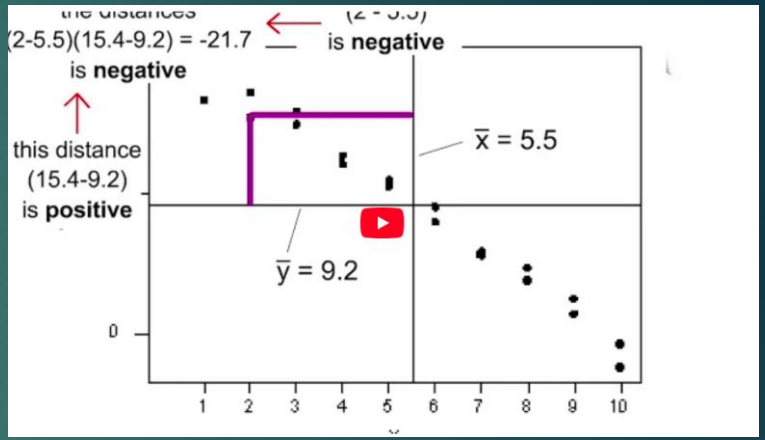
- Watch the following video and note the following:
  - Note that the product of the two distances for the first highlighted data point is positive. In fact, the product of the two distances is positive for any data point in the upper right quadrant.
  - Note that the product of the two distances for the second highlighted data point is also positive. In fact, the product of the two distances is positive for any data point in the lower left quadrant.



https://youtu.be/4fJgjoUAG8c

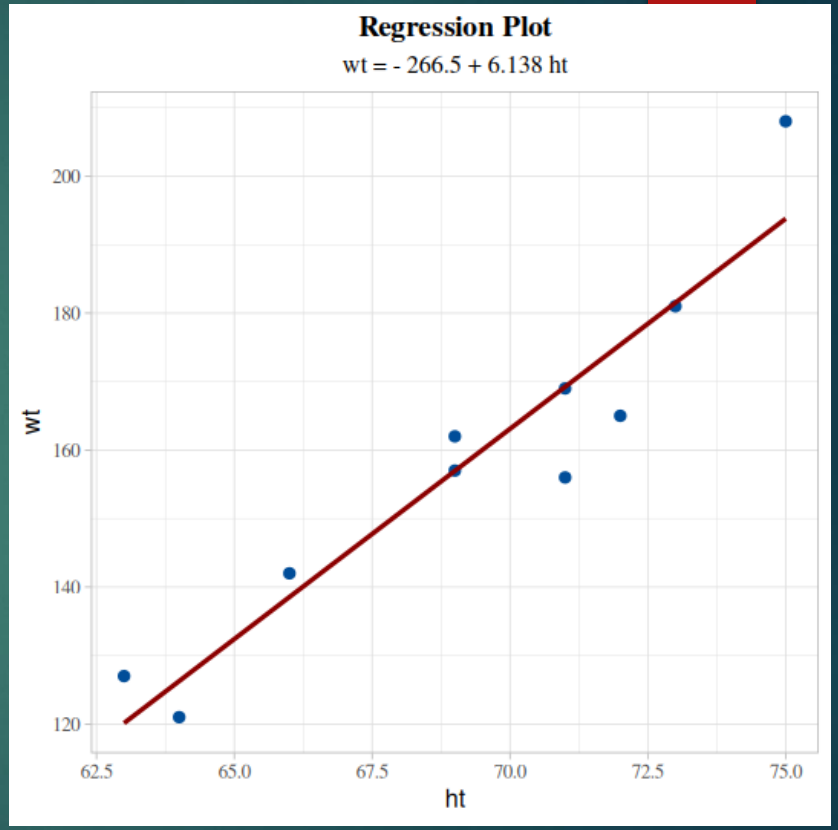
- **When is the slope $b_1 < 0$?** Now, do you agree that the trend in the following plot is negative — that is, as x increases, y tends to decrease? If the trend is negative, then the slope must be negative. Let's see how!

- Watch the following video and note the following:
  - Note that the product of the two distances for the first highlighted data point is negative. In fact, the product of the two distances is negative for any data point in the upper left quadrant.
  - Note that the product of the two distances for the second highlighted data point is also negative. In fact, the product of the two distances is negative for any data point in the lower right quadrant.

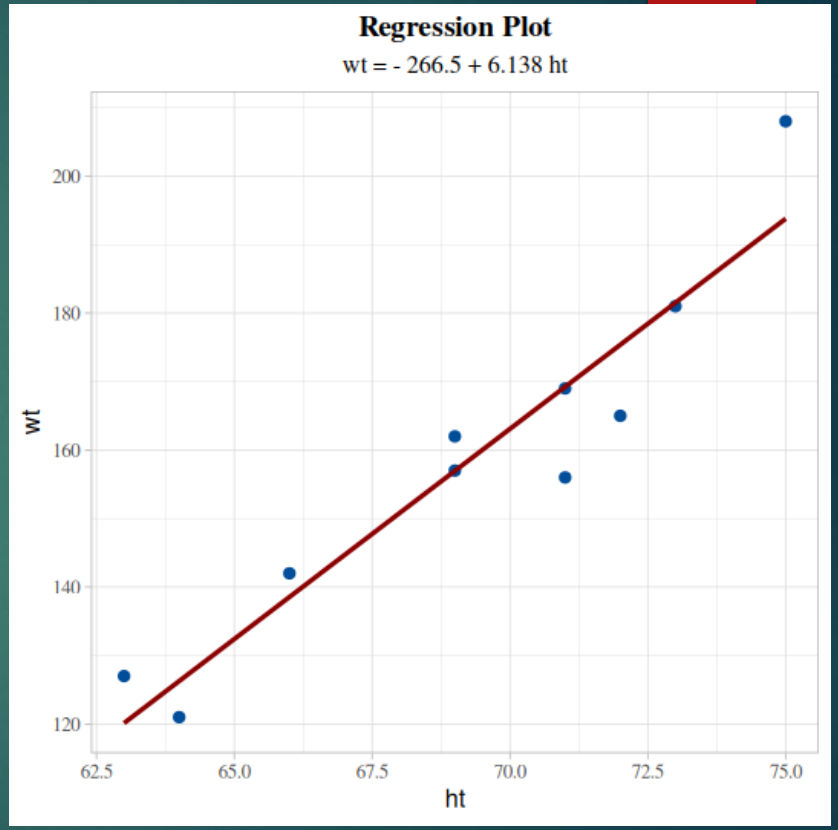


https://youtu.be/rE5vXk6HiDw

Now that we finished that investigation, you can just set aside the formulas for $\mathbf{b_0}$ and $\mathbf{b_1}$. Again, in practice, you are going to let statistical software, find the least squares lines for you. We can obtain the estimated regression equation in two different places in R. The following plot illustrates where you can find the least squares line (below the "Regression Plot" title).



```
ggplot(student, aes(x = ht, y = wt)) +
  geom_point(color = "#004d99", size = 2) +
  geom_smooth(method = "lm", se = FALSE, color =
"#8b0000", linewidth = 1) +
  labs(
    title = "Regression Plot",
    subtitle = "wt = - 266.5 + 6.138 ht",
    x = "ht",
    y = "wt"
  ) +
  theme_light() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold",
family = "serif"),
    plot.subtitle = element_text(hjust = 0.5, family =
"serif"),
    axis.text = element_text(family = "serif")
  )
```



lessR package's output

library (lessR)
reg(wt ~ ht, data = student,brief = T)

Note that the estimated values -266.534 ($\mathbf{b_0}$) and 6.138 ($\mathbf{b_1}$) appear in the "BASIC ANALYSIS" table under the columns labeled "Estimate". In this output, the intercept is labeled as "(Intercept)" and the slope for height is labeled as "ht".Additionally, note that the value obtained by minimizing the sum of the squared prediction errors, 597.386, appears in the "Analysis of Variance" table in the row labeled "Residuals" and under the column labeled "Sum Sq" (which stands for Sum of Squares).

```
  BASIC ANALYSIS

-- Estimated Model for wt

             Estimate     Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)  -266.534      51.032   -5.223    0.001    -384.214    -148.854
         ht     6.138       0.735    8.347    0.000       4.442       7.833

-- Model Fit

Standard deviation of wt: 25.385

Standard deviation of residuals:  8.6414 for df=8
95% range of residuals:  39.8541 = 2 * (2.306 * 8.6414)

R-squared: 0.897    Adjusted R-squared: 0.884    PRESS R-squared: 0.816

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 69.666     df: 1 and 8     p-value:  0.000

-- Analysis of Variance

            df    Sum Sq   Mean Sq   F-value   p-value
Model        1  5202.214  5202.214    69.666     0.000
Residuals    8   597.386    74.673
wt           9  5799.600   644.400
```

lessR package's output

library (lessR)
reg(wt ~ ht, data = student,brief = T)

- R-squared (0.897): This means that approximately 89.7% of the variation in weight can be explained by height.
- P-value (0.000): Since the p-value for ht is less than 0.05, height is a statistically significant predictor of weight in this student group.

```
  BASIC ANALYSIS

-- Estimated Model for wt

             Estimate     Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)  -266.534      51.032   -5.223    0.001    -384.214    -148.854
         ht     6.138       0.735    8.347    0.000       4.442       7.833

-- Model Fit

Standard deviation of wt: 25.385

Standard deviation of residuals:  8.6414 for df=8
95% range of residuals:  39.8541 = 2 * (2.306 * 8.6414)

R-squared: 0.897    Adjusted R-squared: 0.884    PRESS R-squared: 0.816

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 69.666     df: 1 and 8     p-value:  0.000

-- Analysis of Variance

            df    Sum Sq   Mean Sq   F-value   p-value
Model        1  5202.214  5202.214    69.666     0.000
Residuals    8   597.386    74.673
wt           9  5799.600   644.400
```

Although we've learned how to obtain the "**estimated regression coefficients**" b_0 and b_1, we've not yet discussed what we learn from them. One thing they 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:

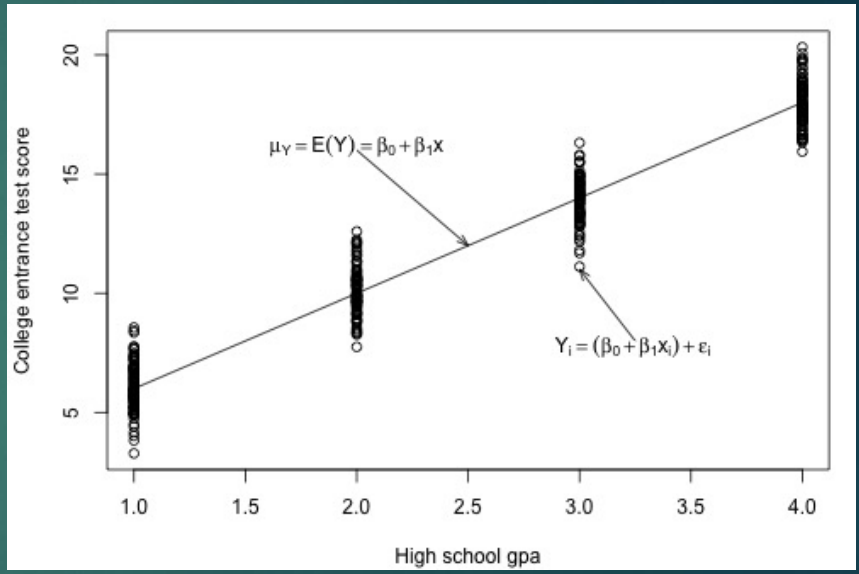
- A common use of the estimated regression line: $\hat{y}_{i,wt} = -267 + 6.14x_{i,ht}$
- Predict (mean) weight of 66"-inch tall people: $\hat{y}_{i,wt} = -267 + 6.14(66) = 138.24$
- Predict (mean) weight of 67"-inch tall people: $\hat{y}_{i,wt} = -267 + 6.14(67) = 144.38$

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 -267 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 information about this in a blog post on the Minitab Website.

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 144.38 - 138.24 = 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*.

1.3 - The Simple Linear Regression Model

- We have worked hard to come up with formulas for the intercept $\mathbf{b_0}$ and the slope $\mathbf{b_1}$ of the least squares regression line. But, we haven't yet discussed what and estimate.
- What do $\mathbf{b_0}$ **and** $\mathbf{b_1}$ **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 five 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.

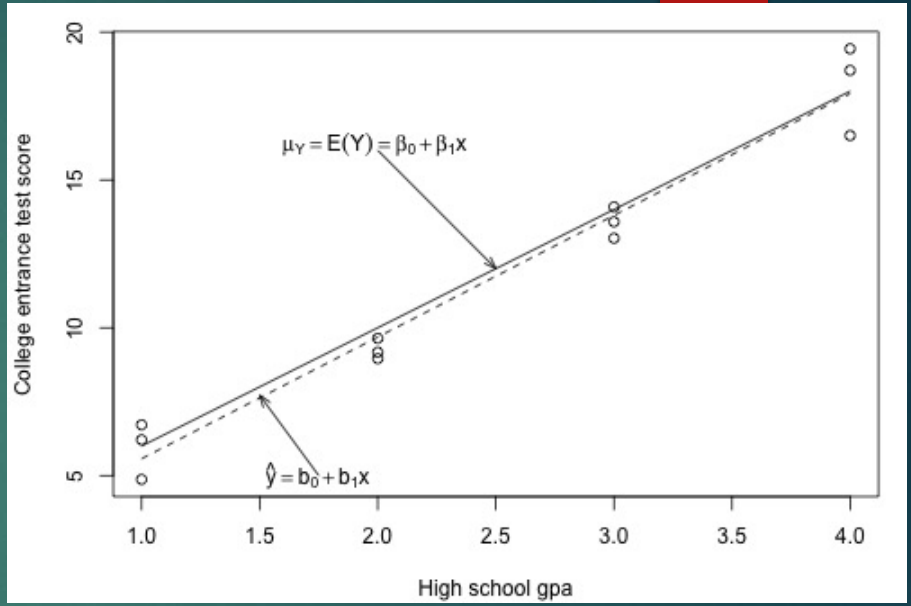


Now, take the average college entrance test score for students with a GPA of 1. And, similarly, take the average college entrance test score for students with a GPA of 2, 3, and 4. Connecting the dots — that is, the averages — you get a line, which we summarize by the formula $\mu_Y = \mathrm{E}(Y) = \beta_0 + \beta_1 x$. 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 μ_Y. We can also express the average college entrance test score for the i^{th} student, $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$. Of course, not every student's college entrance test score will equal the average $\mathrm{E}(Y_i)$. There will be some errors. That is, any student's response y_i will be the linear trend $\beta_0 + \beta_1 x_i$ plus some error ϵ_i. So, another way to write the simple linear regression model is $y_i = \mathrm{E}(Y_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

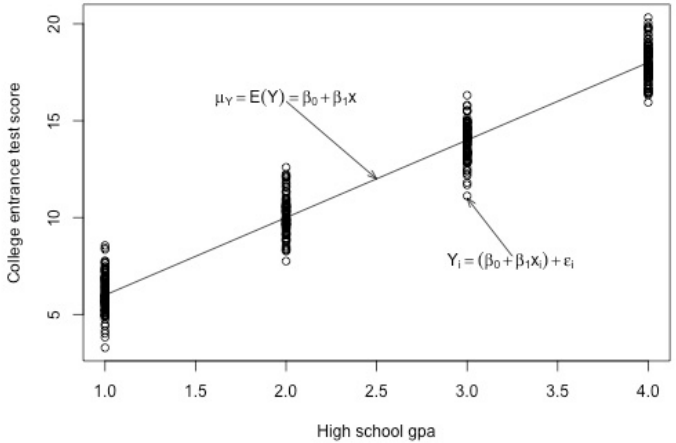
When looking to summarize the relationship between a predictor *x* and a response *y*, we are interested in knowing the population regression line $\mu_Y = \mathrm{E}(Y) = \beta_0 + \beta_1 x$. The only way we could ever know it, though, is to be able to collect data on everybody in the population — 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 β_0 and the sample slope b_1 estimates the population slope β_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 B_1. And, we will use the sample intercept b_0 to learn about the population intercept B_0. In order to draw any conclusions about the population parameters B_0 and B_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.
- In order to draw any conclusions about the population parameters B_0 and B_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.



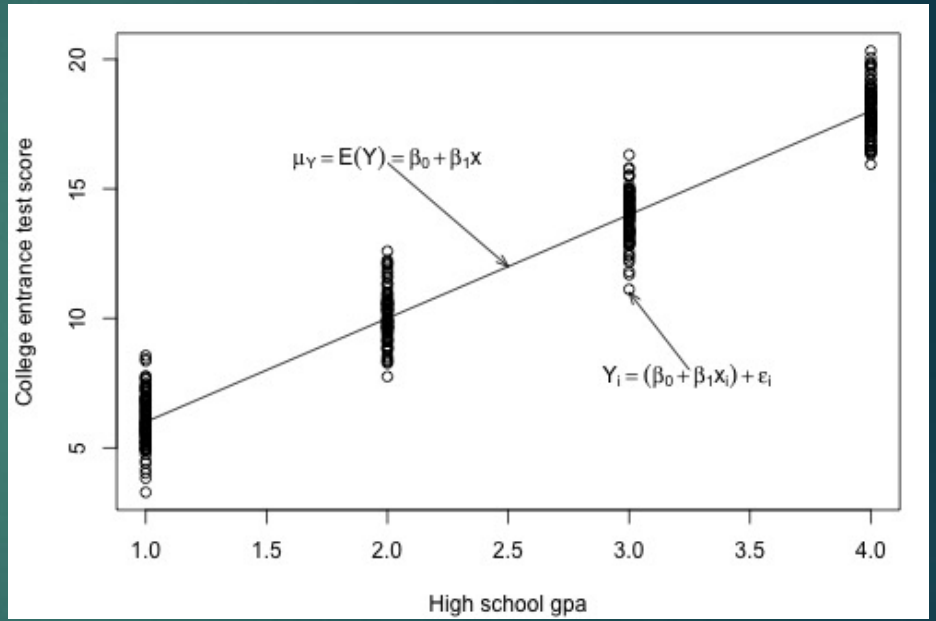
First, notice that when we connected the averages of the college entrance test scores for each of the subpopulations, it formed a line. Most often, we will not have the population of data at our disposal as we pretend to do here. If we didn't, do you think it would be reasonable to assume that the mean 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, ϵ_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 that comprise "**the simple linear regression model**:"

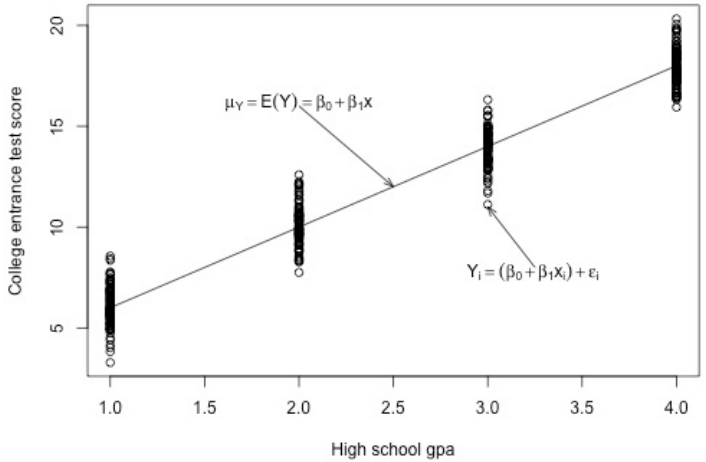
- **L inear Function:** The mean of the response, $\mathrm{E}(Y_i)$, at each value of the predictor, x_i, is a Linear function of the x_i.
- **I ndependent:** The errors, ϵ_i, are Independent.
- **N ormally Distributed:** The errors, ϵ_i, at each value of the predictor, x_i, are Normally distributed.
- **E qual variances (denoted σ^2):** The errors, ϵ_i, at each value of the predictor, x_i, have Equal variances (denoted σ^2).

Do you notice what the first highlighted letters 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, $\mathrm{E}(\epsilon_i)$, at each value of the predictor, x_i, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵ_i, are independent normal random variables with mean zero and constant variance, σ^2.

1.4 - 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 σ^2.

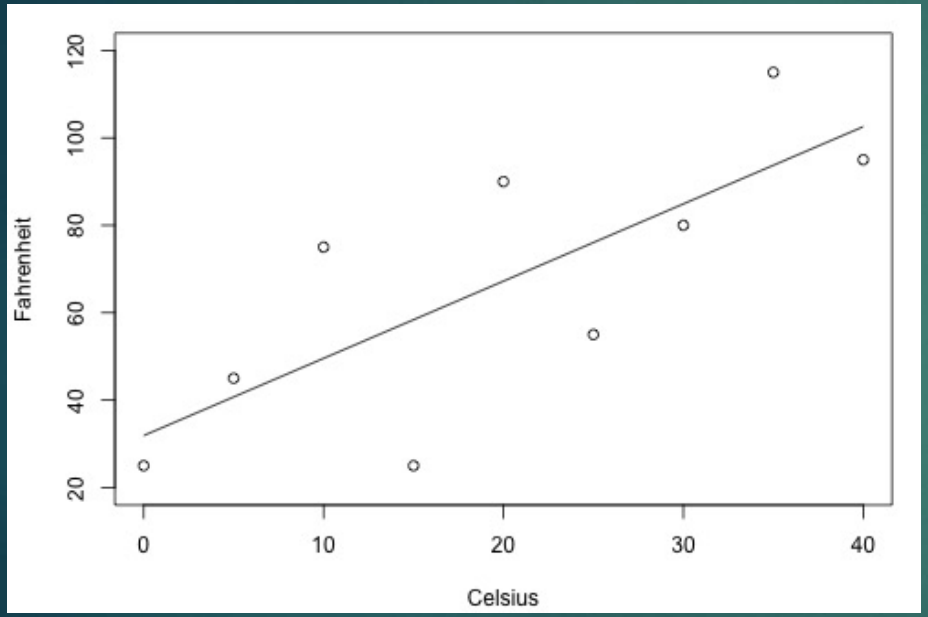


That is, σ^2 quantifies how much the responses (y) vary around the (unknown) mean population regression line $\mu_Y = E(Y) = \beta_0 + \beta_1 x$.

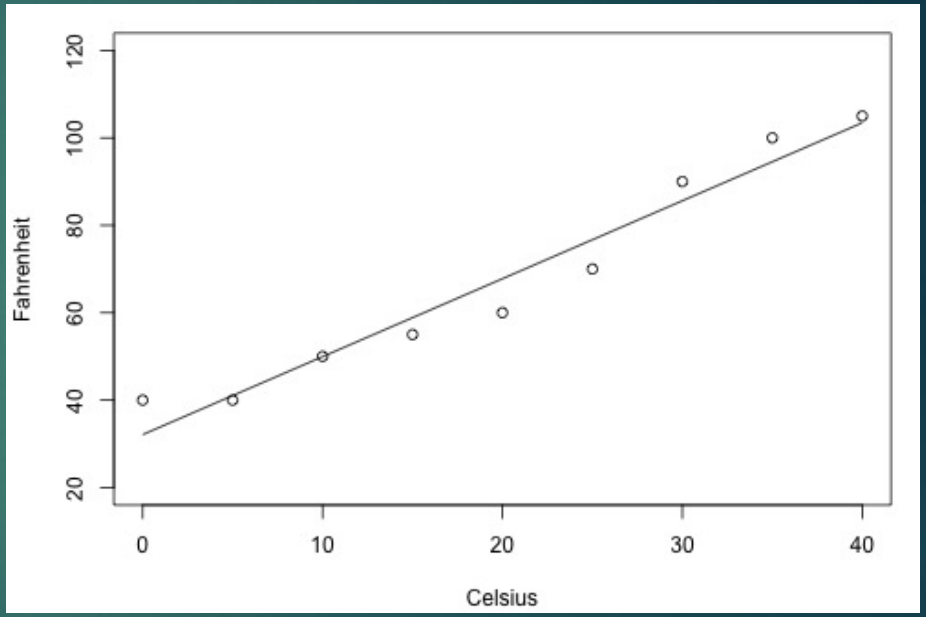
Why should we care about σ^2? The answer to this question pertains to the most common use of an estimated regression line, namely predicting some future response.

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 thermometer brand (A) yield more precise future predictions or brand (B)?



Brand A



Brand 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.

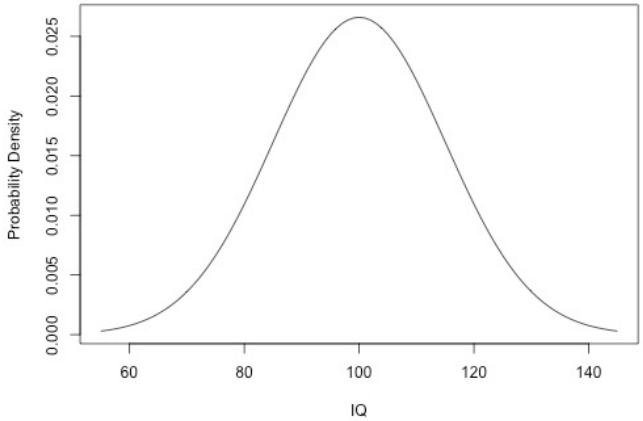
Celsius

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, σ^2 quantifies this variance in the responses. Will we ever know this value σ^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 σ^2in the simple linear regression setting, it is helpful to recall the formula for the estimate of the variance of the responses, σ^2, when there is only one population.

The following is a plot of the (one) 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?

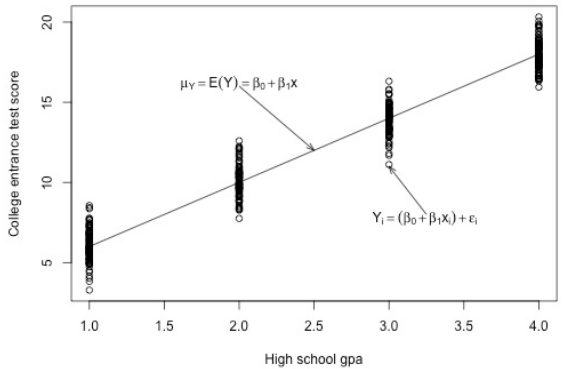


Sample Variance

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

The sample variance estimates σ^2, the variance of 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 μ. But, we don't know the population mean μ, 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 μ.

Now let's extend this thinking to arrive at an estimate for the population variance σ^2 in the simple linear regression setting. Recall that we assume that σ^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 μ_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 $\hat{y}_i = b_0 + b_1 x_i$.

Mean square error

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

The mean square error estimates σ^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 μ_Y, we effectively estimate two parameters — the population intercept β_0 and the population slope β_1. That is, we lose two degrees of freedom.

lessR package's output

library (lessR)
reg(wt ~ ht, data = student,brief = T)

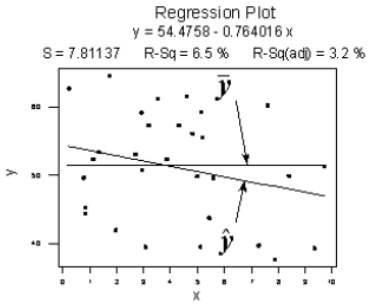
- In practice, we let statistical software like R (using lessR) calculate the mean square error (MSE) for us
- For your student height and weight data, the quantity emphasized in the box, S = 8.6414, is the square root of MSE.
- Height is a very strong predictor for this group. Because your R-squared is 0.897, you can conclude that 89.7% of the differences in student weights are explained by their heights.

```
  BASIC ANALYSIS

-- Estimated Model for wt

             Estimate     Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)  -266.534      51.032   -5.223    0.001    -384.214    -148.854
         ht     6.138       0.735    8.347    0.000       4.442       7.833

-- Model Fit

Standard deviation of wt: 25.385

Standard deviation of residuals:  8.6414 for df=8
95% range of residuals:  39.8541 = 2 * (2.306 * 8.6414)

R-squared: 0.897    Adjusted R-squared: 0.884    PRESS R-squared: 0.816

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 69.666     df: 1 and 8     p-value:  0.000

-- Analysis of Variance

            df    Sum Sq   Mean Sq   F-value   p-value
Model        1  5202.214  5202.214    69.666     0.000
Residuals    8   597.386    74.673
wt           9  5799.600   644.400
```

1.5 - The Coefficient of Determination, R^2

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:



$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 119.1$$

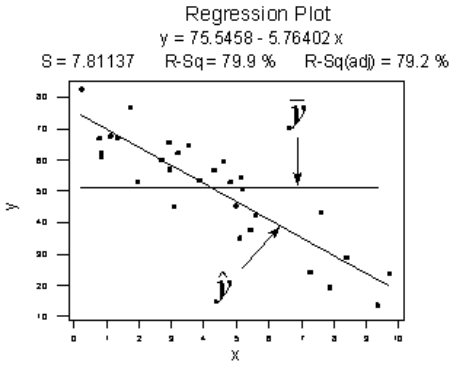
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1708.5$$

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 1827.6$$

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, $\bar{y}$.

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:



$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 6679.3$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 1708.5$$

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 8487.8$$

The sums of squares for this data set 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 Minitab's fitted line plot.

The previous two examples have suggested how we should define the measure formally.

Coefficient of determination

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 the video 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:

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

Characteristics of R^2

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 variations in *y*!
- If R^2 = 0, the estimated regression line is perfectly horizontal. The predictor *x* accounts for *none* of the variations in *y*!

Interpretation of R^2

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 ×100 percent of the variation in *y* is reduced by taking into account predictor *x*"

or:

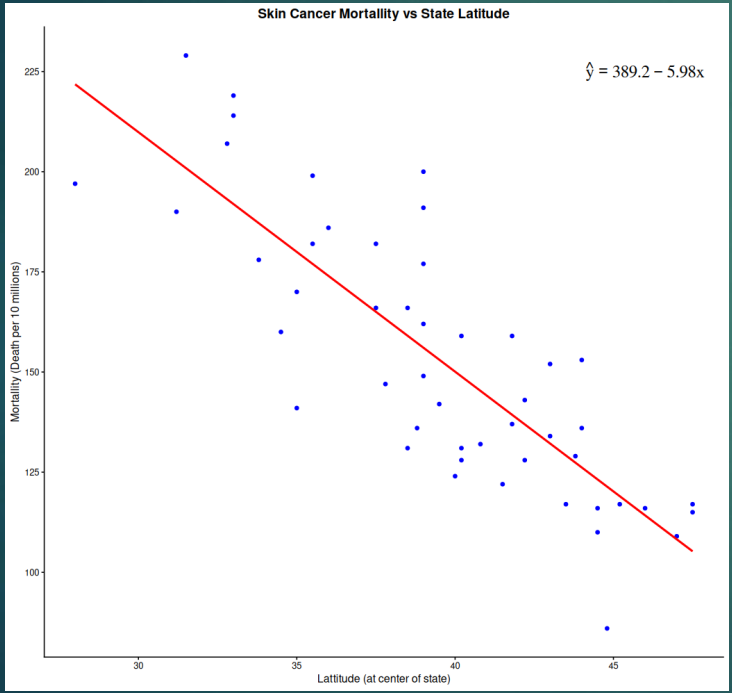
"R^2 ×100 percent of the variation in *y* is 'explained by the variation in predictor *x*."

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 ×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% merely 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 (Skin Cancer Data). Any statistical software that performs a simple linear regression analysis will report the r-squared value for you

library (lessR)
reg (Mort ~ Lat, data = skin_cancer,brief = T)




```
  BASIC ANALYSIS

-- Estimated Model for Mort

             Estimate     Std Err   t-value   p-value    Lower 95%   Upper 95%
(Intercept)   389.189      23.812    16.344     0.000      341.285     437.094
       Lat    -5.978       0.598    -9.990     0.000       -7.181      -4.774

-- Model Fit

Standard deviation of Mort: 33.4282

Standard deviation of residuals:  19.115 for df=47
95% range of residuals:  76.909 = 2 * (2.012 * 19.115)

R-squared: 0.680    Adjusted R-squared: 0.673    PRESS R-squared: 0.651

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 99.797     df: 1 and 47     p-value:  0.000

-- Analysis of Variance

            df     Sum Sq    Mean Sq   F-value   p-value
Model        1  36464.200  36464.200    99.797     0.000
Residuals   47  17173.065    365.384
Mort        48  53637.265   1117.443
```

- We can say that 68% (R-squred) 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 due to or explained by latitude.

```
  BASIC ANALYSIS

-- Estimated Model for Mort

             Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)   389.189     23.812   16.344    0.000     341.285     437.094
        Lat    -5.978      0.598   -9.990    0.000      -7.181      -4.774

-- Model Fit

Standard deviation of Mort: 33.4282

Standard deviation of residuals:  19.115 for df=47
95% range of residuals:  76.909 = 2 * (2.012 * 19.115)

R-squared: 0.680    Adjusted R-squared: 0.673    PRESS R-squared: 0.651

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 99.797     df: 1 and 47     p-value:  0.000

-- Analysis of Variance

            df     Sum Sq    Mean Sq   F-value   p-value
Model        1  36464.200  36464.200    99.797     0.000
Residuals   47  17173.065    365.384
Mort        48  53637.265   1117.443
```

1.6 - (Pearson) Correlation Coefficient, r

The correlation coefficient, *r*, is directly related to the coefficient of determination R^2 in an 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:

Correlation Coefficient, r

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

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 are inches × pounds. Similarly, the units for the denominator are inches × 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 the video. (https://youtu.be/137kwf_zENk)
- The correlation coefficient r is a unitless measure. If you don't see why this must be true, view the video (https://youtu.be/MqhzZgt9Qxs)
- 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. Here's what R-base's output looks like for the skin cancer mortality and latitude example (Skin Cancer Data):

```
> cor.test (skin_cancer$Mort, skin_cancer$Lat)

        Pearson's product-moment correlation

data:  skin_cancer$Mort and skin_cancer$Lat
t = -9.9898, df = 47, p-value = 0.0000000000003309
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8976036 -0.7073128
sample estimates:
       cor
-0.8245178
```

The output tells us that the correlation between skin cancer mortality and latitude is -0.825 for this data set. Note that it doesn't matter the order in which you specify the variables

The output tells us that the correlation between skin cancer mortality and latitude is still -0.825. What does this correlation coefficient tell 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 true 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:

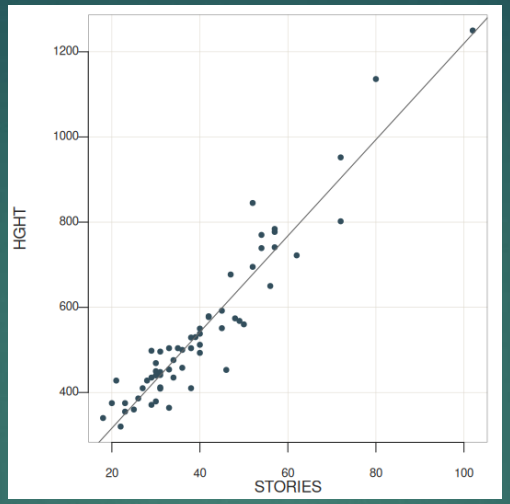
- The relationship is negative. As the latitude increases, the skin cancer mortality rate decreases (linearly).
- The relationship is quite strong (since the value is pretty close to -1)

1.7 - Some Examples

1.7.1. Building Storage

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 (Building Stories data).

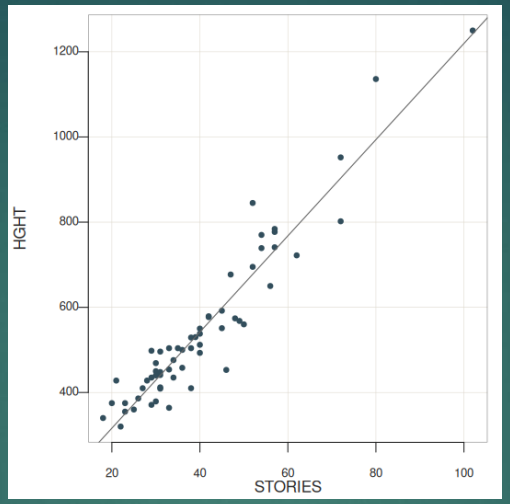
```
BASIC ANALYSIS

-- Estimated Model for HGHT

            Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)    90.310     20.962    4.308    0.000      48.349     132.270
    STORIES    11.292      0.484   23.310    0.000      10.323      12.262

-- Model Fit

Standard deviation of HGHT: 186.2125

Standard deviation of residuals:  58.3259 for df=58
95% range of residuals:  233.5041 = 2 * (2.002 * 58.3259)

R-squared: 0.904    Adjusted R-squared: 0.902    PRESS R-squared: 0.895

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 543.377     df: 1 and 58     p-value:  0.000

-- Analysis of Variance

            df       Sum Sq      Mean Sq   F-value   p-value
Model        1  1848520.326  1848520.326   543.377     0.000
Residuals   58   197311.007     3401.914
HGHT        59  2045831.333    34675.107
```




```
> cor.test (building$STORIES, building$HGHT)

        Pearson's product-moment correlation

data:  building$STORIES and building$HGHT
t = 23.31, df = 58, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9182732 0.9702830
sample estimates:
      cor
0.9505549
```

- R reports that R-squared = 0.904 and r = 0.951.
- The positive sign of r (cor) tells us that the relationship is positive — as the 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-squared 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.

```
BASIC ANALYSIS

-- Estimated Model for HGHT

            Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)    90.310     20.962    4.308    0.000      48.349     132.270
    STORIES    11.292      0.484   23.310    0.000      10.323      12.262

-- Model Fit

Standard deviation of HGHT: 186.2125

Standard deviation of residuals:  58.3259 for df=58
95% range of residuals:  233.5041 = 2 * (2.002 * 58.3259)

R-squared: 0.904    Adjusted R-squared: 0.902    PRESS R-squared: 0.895

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 543.377     df: 1 and 58     p-value:  0.000

-- Analysis of Variance

            df       Sum Sq      Mean Sq   F-value   p-value
Model        1  1848520.326  1848520.326   543.377     0.000
Residuals   58   197311.007     3401.914
HGHT        59  2045831.333    34675.107
```




```
> cor.test (building$STORIES, building$HGHT)

        Pearson's product-moment correlation

data:  building$STORIES and building$HGHT
t = 23.31, df = 58, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9182732 0.9702830
sample estimates:
      cor
0.9505549
```

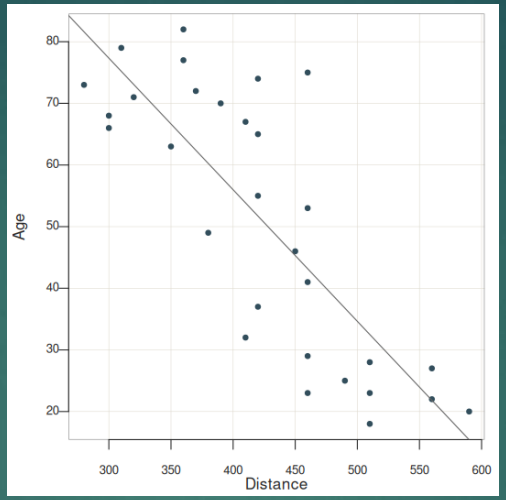
```
building <- read.table ("Buil_Storage.tsv", header = T)

library(lessR)

reg(HGHT ~ STORIES, data = building,brief = T)

cor.test (building$STORIES, building$HGHT)
```

1.7 - Some Examples

1.7.2. Drivers and Age

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 (Driver Age and Distance data).

```
BASIC ANALYSIS

-- Estimated Model for Age

             Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)   141.386     12.983   10.890    0.000     114.791     167.982
   Distance    -0.214      0.030   -7.086    0.000      -0.275      -0.152

-- Model Fit

Standard deviation of Age: 21.7763

Standard deviation of residuals:  13.2602 for df=28
95% range of residuals:  54.3245 = 2 * (2.048 * 13.2602)

R-squared: 0.642    Adjusted R-squared: 0.629    PRESS R-squared: 0.604

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 50.211     df: 1 and 28     p-value:  0.000

-- Analysis of Variance

            df    Sum Sq   Mean Sq   F-value   p-value
Model        1  8828.688  8828.688    50.211     0.000
Residuals   28  4923.312   175.833
Age         29 13752.000   474.207
```



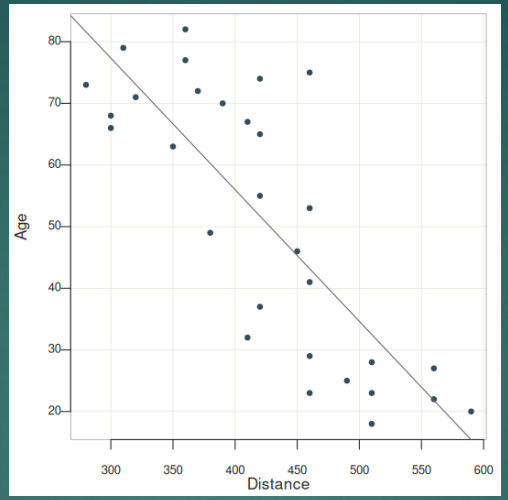

```
> cor.test (age_dist$Age, age_dist$Distance)

        Pearson's product-moment correlation

data:  age_dist$Age and age_dist$Distance
t = -7.086, df = 28, p-value = 0.0000001041
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9013320 -0.6199255
sample estimates:
       cor
-0.8012447
```

age_dist <- read.table ("Age_Dist.tsv",header=T)

reg(Age ~ Distance, data = age_dist,brief = T)

cor.test (age_distAge, age_distDistance)

```
BASIC ANALYSIS

-- Estimated Model for Age

             Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)   141.386     12.983   10.890    0.000     114.791     167.982
   Distance    -0.214      0.030   -7.086    0.000      -0.275      -0.152

-- Model Fit

Standard deviation of Age: 21.7763

Standard deviation of residuals:  13.2602 for df=28
95% range of residuals:  54.3245 = 2 * (2.048 * 13.2602)

R-squared: 0.642    Adjusted R-squared: 0.629    PRESS R-squared: 0.604

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 50.211     df: 1 and 28     p-value:  0.000

-- Analysis of Variance

            df    Sum Sq   Mean Sq   F-value   p-value
Model        1  8828.688  8828.688    50.211     0.000
Residuals   28  4923.312   175.833
Age         29 13752.000   474.207
```




```
> cor.test (age_dist$Age, age_dist$Distance)

        Pearson's product-moment correlation

data:  age_dist$Age and age_dist$Distance
t = -7.086, df = 28, p-value = 0.0000001041
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9013320 -0.6199255
sample estimates:
       cor
-0.8012447
```

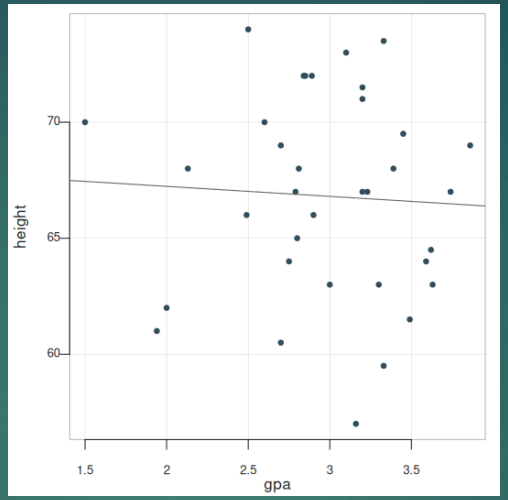
- R reports that R-squared = 0.642 and r = -0.801.
- The negative sign of r tells us that the relationship is negative — as driving age increases, seeing distance decreases — as we expected.
- Because r is fairly close to -1, it tells us that the linear relationship is fairly strong, but not perfect.
- The R-squared value tells us that 64.2% of the variation in the seeing distance is reduced by taking into account the age of the driver.

1.7 - Some Examples

1.7.3. Height and GPA

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 (Height and GPA data)

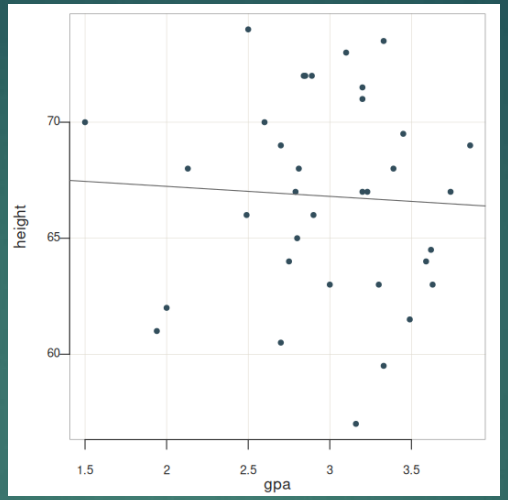
```
  BASIC ANALYSIS

-- Estimated Model for height

            Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)    68.098      4.256   16.000    0.000      59.439      76.757
        gpa    -0.432      1.410   -0.306    0.761      -3.301       2.437

-- Model Fit

Standard deviation of height: 4.3404

Standard deviation of residuals:  4.3995 for df=33
95% range of residuals:  17.9016 = 2 * (2.035 * 4.3995)

R-squared: 0.003    Adjusted R-squared: -0.027    PRESS R-squared: -0.115

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 0.094     df: 1 and 33     p-value:  0.761

-- Analysis of Variance

           df    Sum Sq   Mean Sq   F-value   p-value
Model       1     1.816     1.816     0.094     0.761
Residuals  33   638.727    19.355
height     34   640.543    18.839
```




```
> cor.test (gpa$height, gpa$gpa)

        Pearson's product-moment correlation

data:  gpa$height and gpa$gpa
t = -0.30628, df = 33, p-value = 0.7613
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3797501  0.2850630
sample estimates:
        cor
-0.05324126
```

```
gpa <- read.table
("height_gpa.tsv",header = T)

reg(height ~ gpa, data = gpa,brief = T)

cor.test (gpa$height, gpa$gpa)
```

```
BASIC ANALYSIS

-- Estimated Model for height

            Estimate    Std Err  t-value  p-value   Lower 95%   Upper 95%
(Intercept)    68.098      4.256   16.000    0.000      59.439      76.757
        gpa    -0.432      1.410   -0.306    0.761      -3.301       2.437

-- Model Fit

Standard deviation of height: 4.3404

Standard deviation of residuals:  4.3995 for df=33
95% range of residuals:  17.9016 = 2 * (2.035 * 4.3995)

R-squared: 0.003    Adjusted R-squared: -0.027    PRESS R-squared: -0.115

Null hypothesis of all 0 population slope coefficients:
  F-statistic: 0.094     df: 1 and 33     p-value:  0.761

-- Analysis of Variance

           df    Sum Sq   Mean Sq   F-value   p-value
Model       1     1.816     1.816     0.094     0.761
Residuals  33   638.727    19.355
height     34   640.543    18.839
```




```
> cor.test (gpa$height, gpa$gpa)

        Pearson's product-moment correlation

data:  gpa$height and gpa$gpa
t = -0.30628, df = 33, p-value = 0.7613
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3797501  0.2850630
sample estimates:
        cor
-0.05324126
```

- Rreports that R-squared = 0.003 and r = -0.053.
- 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-suared 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 the number of hours studied, if predicting a student's grade point average is important to us.

1.8 - R-squared Cautions

- Unfortunately, the coefficient of determination R-squared 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.

1.8 - R-squared Cautions

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

2. A large R-squared 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.

3. The coefficient of determination R-squared and the correlation coefficient r can both be greatly affected by just one data point (or a few data points).

1.8 - R-squared Cautions

4. Correlation (or association) does not imply causation.
5. Ecological correlations — correlations that are based on rates or averages — tend to overstate the strength of an association.
6. A "statistically significant" R-squared value does not imply that the slope B_1 is meaningfully different from 0.
7. A large R-squared value does not necessarily mean that a useful prediction of the response y_{new}, or estimation of the mean response, can be made. It is still possible to get prediction intervals or confidence intervals that are too wide to be useful.

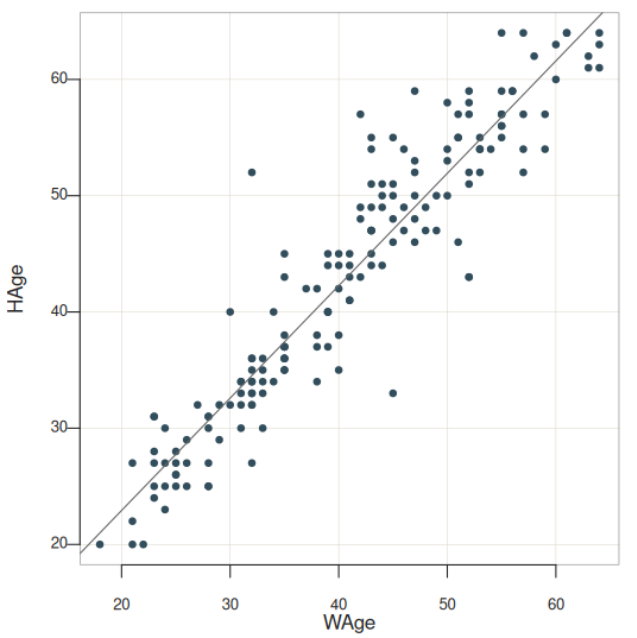
1.9 - Hypothesis Test for the Population Correlation Coefficient

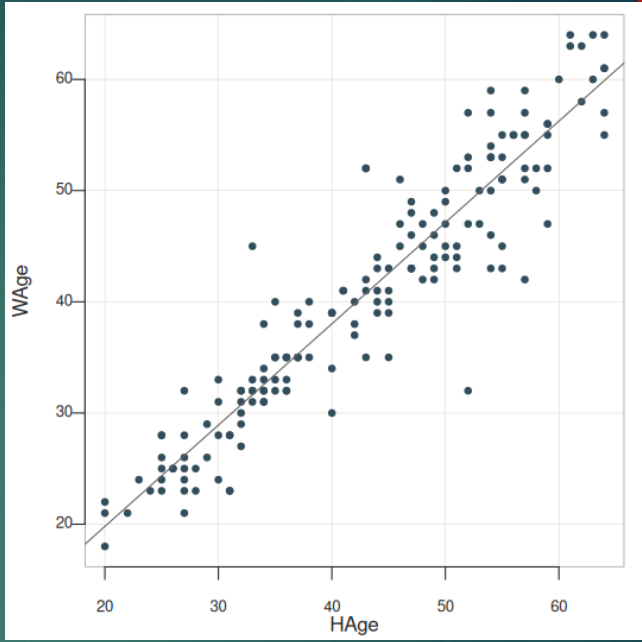
There is one more point we haven't stressed yet in our discussion about the correlation coefficient *r* and the coefficient of determination R^2 — namely, the two measures summarize the strength of a linear relationship *in samples only*. If we obtained a different sample, we would obtain different correlations, different R^2 values, and therefore potentially different conclusions. As always, we want to *draw conclusions about populations*, not just samples. To do so, we either have to conduct a hypothesis test or calculate a confidence interval. In this section, we learn how to conduct a hypothesis test for the population correlation coefficient ρ (the greek letter "rho").

In general, a researcher should use the hypothesis test for the population correlation ρ to learn of a linear association between two variables, when it isn't obvious which variable should be regarded as the response. Let's clarify this point with examples of two different research questions.

Consider evaluating whether or not a linear relationship exists between skin cancer mortality and latitude. We can perform either of the following tests:

- t-test for testing $H_0 : B_1 = 0$
- ANOVA F-test for testing $H_0 : B_1 = 0$





By contrast, suppose we want to evaluate whether or not a linear relationship exists between a husband's age and his wife's age (Husband and Wife data). In this case, one could treat the husband's age as the response or vice versa

In cases such as these, we answer our research question concerning the existence of a linear relationship by using the t-test for testing the population correlation coefficient

Steps for Hypothesis Testing for ρ

Step 1: Hypotheses

First, we specify the null and alternative hypotheses:

- **Null hypothesis** $H_0: \rho = 0$
- **Alternative hypothesis** $H_A: \rho \neq 0$ or $H_A: \rho < 0$ or $H_A: \rho > 0$

Step 2: Test Statistic

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

Test statistic: $t^* = \dfrac{r\sqrt{n-2}}{\sqrt{1-R^2}}$

Step 3: P-Value

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.

Step 4: Decision

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. We conclude that "there is sufficient evidence at theα 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*."

Final Note

One final note ... as always, we should clarify when it is okay to use the *t*-test for testing $H_0: \rho = 0$? The guidelines are a straightforward extension of the "LINE" assumptions made for the simple linear regression model. It's okay:

- When it is not obvious which variable is the response.
- When the (*x*, *y*) pairs are a random sample from a bivariate normal population.
  - For each *x*, the *y*'s are normal with equal variances.
  - For each *y*, the *x*'s are normal with equal variances.
  - Either, *y* can be considered a linear function of *x*.
  - Or, *x* can be considered a linear function of *y*.
- The (*x*, *y*) pairs are independent

Thank you for your listening

Q & A

Phuc-Loi Luu, PhD

Email: loilp@bvtn.org.vn

Zalo: 0901802182