Contents

In the tutorial on prediction we used the regression line to predict values of y for values of x. That is, the regression line is a way of using your data to predict what an average y value should be for a given value of x.

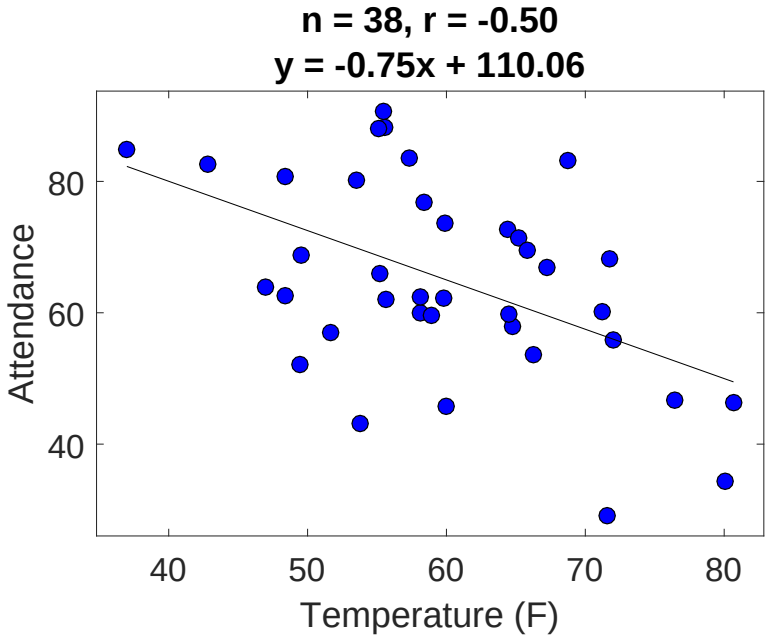
In this tutorial, we'll take this one step further by defining not just what the average value of y should be, but how those values of y should be distributed around the regression line.

If you've gone through the z_table and the normal_distribution tutorials then you should be familiar with how to use normal distributions to calculate areas and scores. This comes in useful here where we'll use the assumption of homoscedasticity to estimate the distribution of scores around a regression line.

Let's work with a made-up example. Suppose that we measure the class attendance and the outdoor temperature for the 38 lectures of psych 315 this quarter.

Here are some statistics for our made up data: (1) temperatures (x) are distributed with a mean of 60 and a standard deviation of 10 degrees, (2) attendance (y) is distributed with a mean of 65 and a standard deviation of 15 students, and (3) temperature and attendance correlates with a value of -0.5.

Here's a scatterplot and regression line of our data. For practice you could use the statistics above to derive the equation of the regression line.



Standard error of the estimate

This tutorial is all about how well we can use our data to predict y from x. Intuitively, if all of the data points fall nicely along the regression line, then we can be pretty confident that this line will provide an accurate estimate of y from x. However, if the line doesn't fit well, then the points will be scattered all over the line, and we won't be as confident about our prediction.

Rememeber from the tutorial on prediction that we defined how well the regression line fit the data with the **standard error of the estimate**, S_{yx}:

$$s_{yx} = \sqrt{\frac{\sum (y-y')^2}{n}}$$

where y' is the y-value of the regression line for each value of x.

We discussed in the prediction tutorial that s_{yx} can be thought of as the standard deviation of the distribution of scores around the regression line.

If you know the correlation, then there's an easier way of calculating the standard error of the estimate:

$S_{yx} = S_y\sqrt{1 - r^2}$

$S_{yx} = 15\sqrt{1 - (-0.5)^2} = 12.99$

Look closely at this equation. What happens with the correlation, r, is either near 1 or -1? S_{yx} gets close to zero. This should make sense; if the correlation is nearly perfect, then the data points are close to the line and therefore the standard deviation of the values around the line is near zero.

On the other hand, if the correlation is zero, then $S_{yx} = S_y$. That is, the standard deviation of the values around the regression line is the same as the standard deviation of the y-values. Again, this should make sense. If the correlation is zero, then the slope of the regression line is zero, which means that the regression line is simply $y' = \bar{y}$. In other words, if the correlation is zero, then the predicted value of y is just the mean of y. So it makes sense that the standard deviation around the regression line is just the standard deviation around the mean of y, which is s_y.

Homoscedasticity

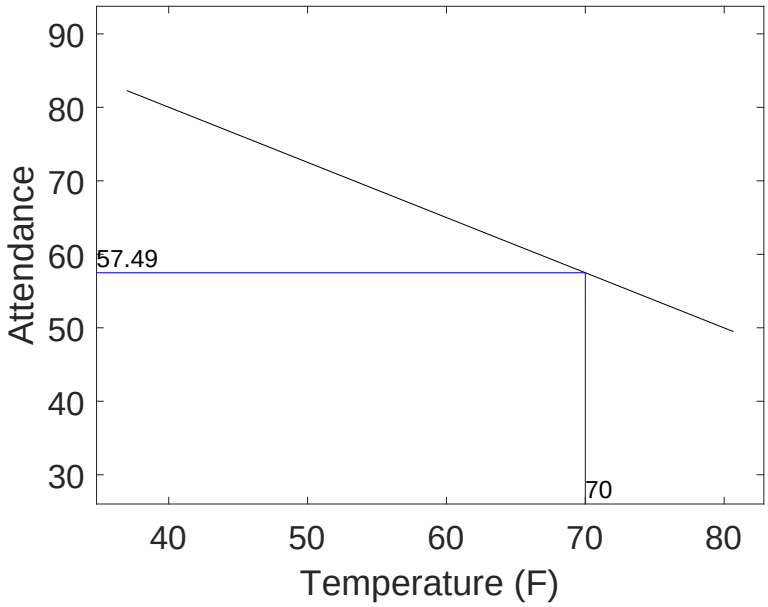
Next we'll make an important assumption: the distribution of scores above and below the regression line is normally distributed around y' with a standard deviation equal to the standard error of the estimate (s_{yx}). This assumption has a special name: **homoscedasticity**.

With homoscedasticity, we know everything about the distribution of scores above and below the regression line. We know it's normal, we know the mean (y'), and we know the standard deviation (s_{yx}). This means that we can answer questions about the proportion of scores that are expected to fall above and below the regression line.

Example 1: What is the expected attendance when the outdoor temperature is 70 degrees?

This is found simply by finding the y-value on the regression line for x = 70:

$y' = mx + b = (-0.75)(70) + 110.06 = 57.49$

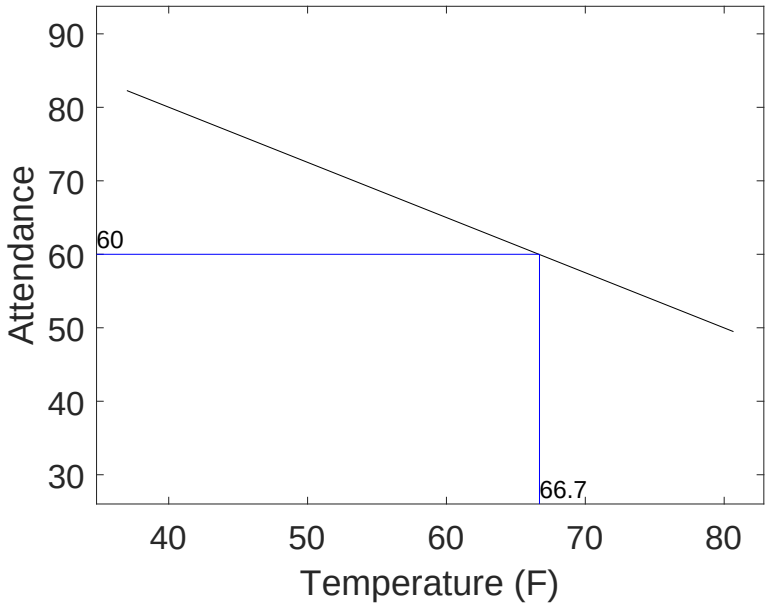


Example 2: What is the temperature for which the expected attendance is 60?

This requires us to use the regression line, but solving for x for when y' = 60:

$y' = 60 = (-0.75)x + 110.06$

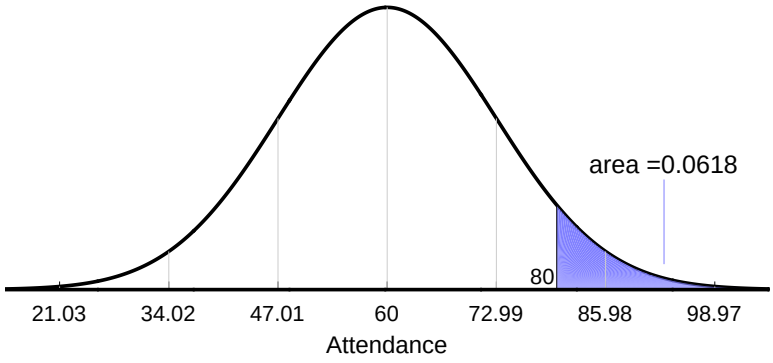
$x = \frac{(60-110.06)}{-0.75} = 66.7$



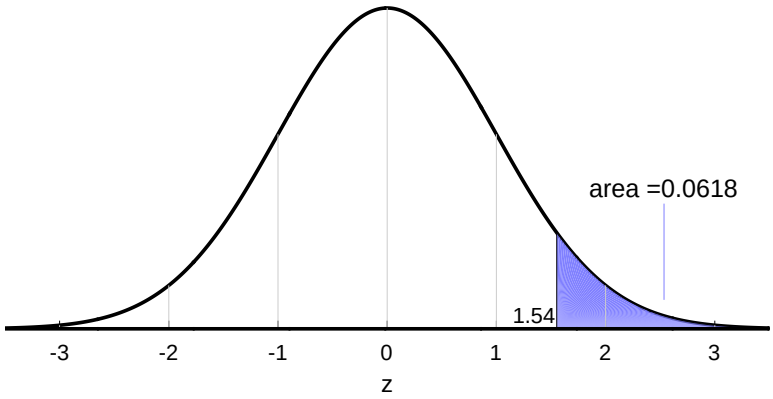
If the outdoor temperature is 66.7 degrees, we expect the attendance to be 60 students.

Example 3: On a day that the outdoor temperature is 66.7, what is the probability that 80 or more students will show up for class?

We know that the distribution of attendance for x = 66.7 will be normal with a mean of y'=60 and a standard deviation of S_{yx} = 12.99. So now we're reduced to a normal distribution problem of finding the area under the normal distribution of mean 60 and standard deviation of 12.99 above a value of 80:



The z-score is therefore $z = \frac{80-60}{12.99} = 1.54$:



Using Table A, the area above z = 1.54 is 0.0618. So the probability of 80 or more students showing up on a 66.7 degree day is 0.0618.

Example 4: What does the outdoor temperature have to be so that there is a 75% chance of 60 or more students showing up?

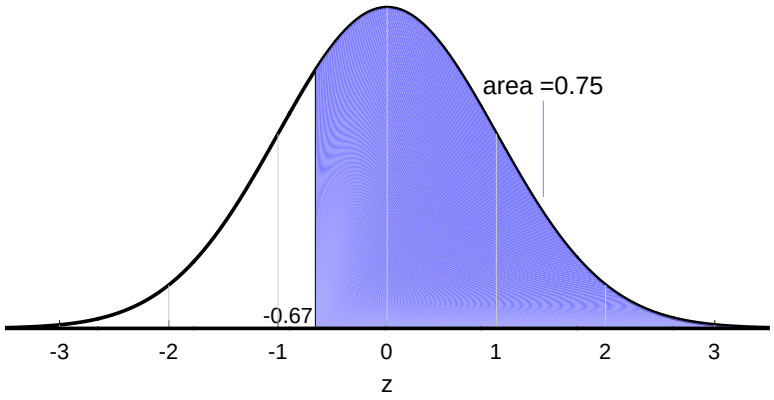
This is like the normal distribution tutorial problems where we convert from areas to scores. For those problems, we first find the z-score for the area, and then convert that to scores using:

$$x = (z)(\sigma) + \mu$$

But now, we have scores distributed normally around a regression line with mean y' and standard deviation S_{yx}, so this means that:

$$y = (z)(S_{yx}) + y'$$

For this problem, the z-score for which 75% of the standard normal distribution lies above is -0.67:

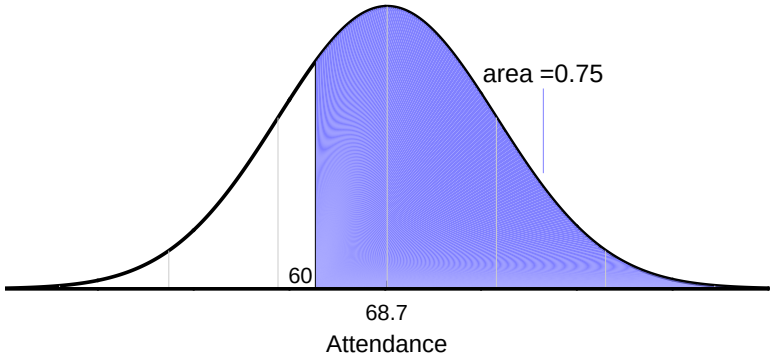


To convert to scores, we know that z = -0.67, we know that y = 60, and we know that $S_{yx} = 12.99$, so

$$60 = (-0.67)(12.99) + y'$$

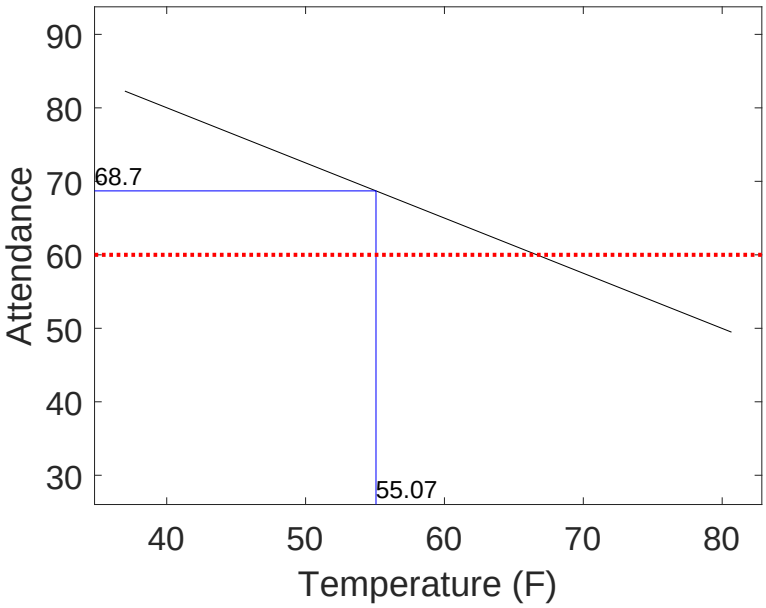
Solving for y':

$$y' = 60 - (-0.67)(12.99) = 68.7$$



Now that we know y', we can find the corresponding value of x (temperature) on the regression line like we did for Example 2 above:

$x = \frac{(68.7-110.06)}{-0.75} = 55.07$ degrees.



In summary, when it's 55.07 degrees outside, then the distribution of attendance will have a mean of 68.7 and a standard deviation of 12.99. For this distribution of attendance, there is a 75 percent chance of 60 or more students showing up.

Questions

It's your turn again. Here are 30 problems and answers.

1) Let the mean of x be 78, the mean of y be 78
the standard deviation of x be 1.2 and the standard deviation of y be 1
and the correlation between x and y be 0.87

Find the value of x for which 62% of the values lie above y = 77.8.
Answer: x = 77.4

2) Let the regression line for y on x be y' = -1.08x + 144.91
and the standard error of the estimate be Syx = 0.22

For x = 63.2, find the value of y for which 62% of the values lie above.

Answer: y = 76.6

3) Let the mean of x be 103, the mean of y be 95
the standard deviation of x be 1.7 and the standard deviation of y be 1.6
and the correlation between x and y be 0.05

For x = 102.4, find the value of y for which 91% of the values lie above.

Answer: y = 93.2

4) Let the regression line for y on x be y' = -1.02x + 123.47
and the standard error of the estimate be Syx = 0.79

For x = 85.5, find the value of y for which 74% of the values lie below.

Answer: y = 36.8

5) Let the mean of x be 47, the mean of y be 35
the standard deviation of x be 1.8 and the standard deviation of y be 2
and the correlation between x and y be 0.97

For x = 47, find the value of y for which 72% of the values lie below.

Answer: y = 35.4

6) Let the regression line for y on x be y' = -0.19x + 55.53

Find the value of x for which y' = 36.9

Answer: x= 98.1

7) Let the regression line for y on x be y' = 0.86x + 35.74
and the standard error of the estimate be Syx = 0.84

Find the value of x for which 91% of the values lie below y = 59.9.

Answer: x = 26.8

8) Let the regression line for y on x be y' = -0.01x + 109.42
and the standard error of the estimate be Syx = 1.5

Find the value of x for which 27% of the values lie below y = 108.1.

Answer: x = 42.1

9) Let the regression line for y on x be y' = 1.13x + -16.33
and the standard error of the estimate be Syx = 0.84

Find the value of x for which 53% of the values lie below y = 38.

Answer: x = 48.0

10) Let the mean of x be 87, the mean of y be 93
the standard deviation of x be 1.1 and the standard deviation of y be 1.8
and the correlation between x and y be -0.75

Find the value of x for which 2% of the values lie below y = 90.9.

Answer: x = 86.6

11) Let the regression line for y on x be y' = 0.83x + 41.21
and the standard error of the estimate be Syx = 1.06

For x = 82.5, find the value of y for which 77% of the values lie above.

Answer: y = 108.9

12) Let the regression line for y on x be y' = 0.73x + -17.91
and the standard error of the estimate be Syx = 1.17

For x = 94.8, find the value of y for which 27% of the values lie below.

Answer: y = 50.6

13) Let the regression line for y on x be y' = 0.12x + 21.6
and the standard error of the estimate be Syx = 1.29

For x = 28.3, find the value of y for which 80% of the values lie below.

Answer: y = 26.1

14) Let the mean of x be 85, the mean of y be 65
the standard deviation of x be 1.5 and the standard deviation of y be 2
and the correlation between x and y be 0.78

Find the value of y' for which x = 85.2

Answer: y'= 65.2

15) Let the regression line for y on x be y' = -0.64x + 33.66

Find the value of y' for which x = 22.5

Answer: y'= 19.3

16) Let the mean of x be 34, the mean of y be 28
the standard deviation of x be 1.7 and the standard deviation of y be 1.9
and the correlation between x and y be -0.36

Find the value of y' for which x = 33.6

Answer: y'= 28.2

17) Let the mean of x be 49, the mean of y be 60

the standard deviation of x be 1.7 and the standard deviation of y be 1.4
and the correlation between x and y be -0.7

For x = 49.2, find the value of y for which 59% of the values lie below.

Answer: y = 59.9

18) Let the mean of x be 25, the mean of y be 108
the standard deviation of x be 2 and the standard deviation of y be 1.6
and the correlation between x and y be -0.18

Find the value of x for which y' = 108.1

Answer: x= 25.3

19) Let the mean of x be 106, the mean of y be 23
the standard deviation of x be 1.5 and the standard deviation of y be 1.5
and the correlation between x and y be -0.6

Find the value of y' for which x = 105.4

Answer: y'= 23.4

20) Let the regression line for y on x be y' = -0.05x + 39.24

Find the value of y' for which x = 89.7

Answer: y'= 34.8

21) Let the mean of x be 24, the mean of y be 105
the standard deviation of x be 1.9 and the standard deviation of y be 1.4
and the correlation between x and y be -0.18

Find the value of y' for which x = 24.1

Answer: y'= 105.0

22) Let the regression line for y on x be y' = -0.67x + 66.45
and the standard error of the estimate be Syx = 0.75

For x = 45.9, find the value of y for which 22% of the values lie above.

Answer: y = 36.3

23) Let the mean of x be 103, the mean of y be 75
the standard deviation of x be 1.5 and the standard deviation of y be 1.8
and the correlation between x and y be -0.57

Find the value of x for which y' = 75.8

Answer: x= 102.4

24) Let the regression line for y on x be y' = 0.51x + 39.18
and the standard error of the estimate be Syx = 1.15

Find the value of x for which 76% of the values lie below y = 71.3.

Answer: x = 61.5

25) Let the mean of x be 92, the mean of y be 79

the standard deviation of x be 1.6 and the standard deviation of y be 1.4
and the correlation between x and y be 0.4

Find the value of x for which 99% of the values lie below y = 81.9.

Answer: x = 91.7

26) Let the mean of x be 61, the mean of y be 20
the standard deviation of x be 1.3 and the standard deviation of y be 1.5
and the correlation between x and y be -0.17

For x = 61, find the value of y for which 42% of the values lie above.

Answer: y = 20.1

27) Let the regression line for y on x be y' = 0.46x + 52.42
and the standard error of the estimate be Syx = 1.73

Find the value of x for which 74% of the values lie below y = 91.

Answer: x = 81.5

28) Let the mean of x be 17, the mean of y be 81
the standard deviation of x be 1.5 and the standard deviation of y be 1.5
and the correlation between x and y be -0.02

Find the value of x for which y' = 81

Answer: x= 16.5

29) Let the mean of x be 99, the mean of y be 64
the standard deviation of x be 1.9 and the standard deviation of y be 1.9
and the correlation between x and y be -0.72

Find the value of x for which y' = 64

Answer: x= 99.0

30) Let the mean of x be 61, the mean of y be 98
the standard deviation of x be 1.4 and the standard deviation of y be 1.3
and the correlation between x and y be -0.12

For x = 61.8, find the value of y for which 53% of the values lie above.

Answer: y = 97.9