

Interpreting Regression

January 20, 2021

Contents

- Standard error of the estimate
- Homoscedasticity
- Using R to make interpretations about regression
- Questions

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.

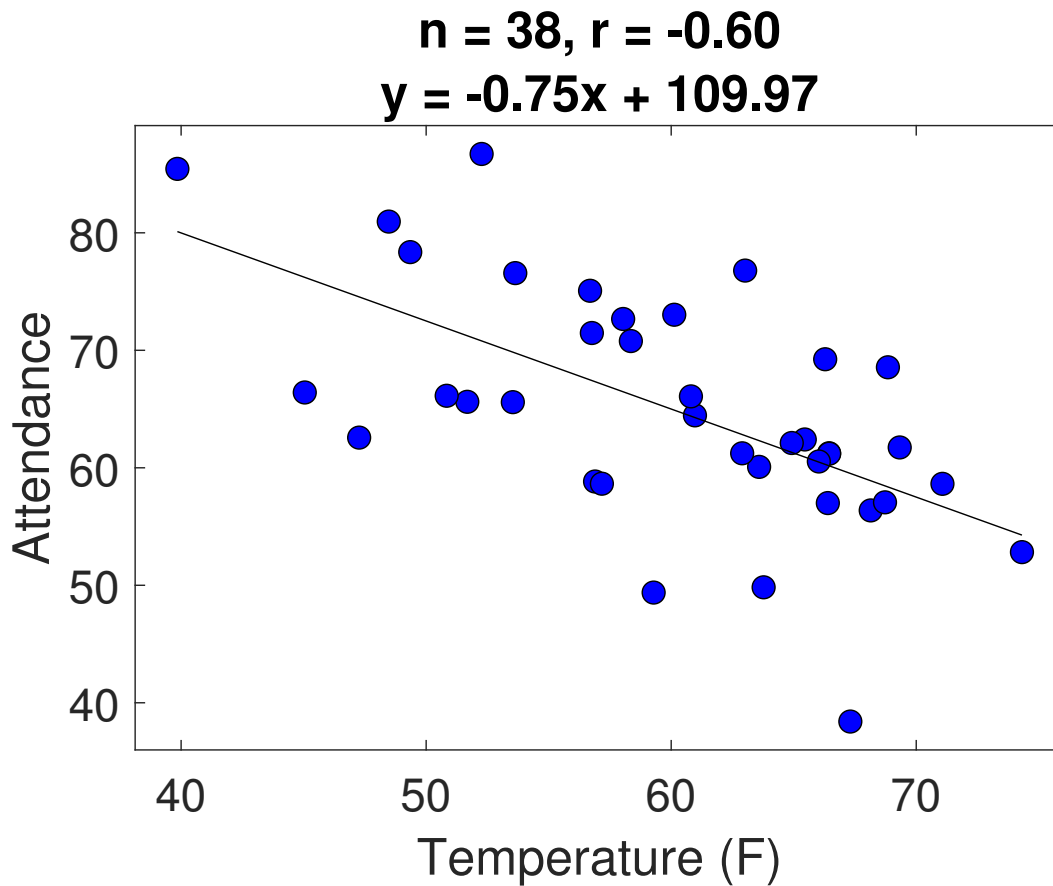
If you've gone through the `z_table` and the `normal_distribution` tutorials then you should be familiar with how to use normal distributions to calculate areas and scores. This comes in useful here where we'll use the assumption of homoscedasticity to estimate the distribution of scores around a regression 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 8 degrees, (2) attendance (y) is distributed with a mean of 65 and a standard deviation of 10 students, and (3) temperature and attendance correlates with a value of -0.6.

You can download the csv file containing the data for this tutorial here: `TemperatureAttendance.csv`

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.

Remember 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:

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

For our example predicting attendance with temperature, the correlation is $r = -0.6$ and the standard deviation for attendance is $s_y = 10$:

$$S_{yx} = 10 \sqrt{1 - (-0.6)^2} = 8$$

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.

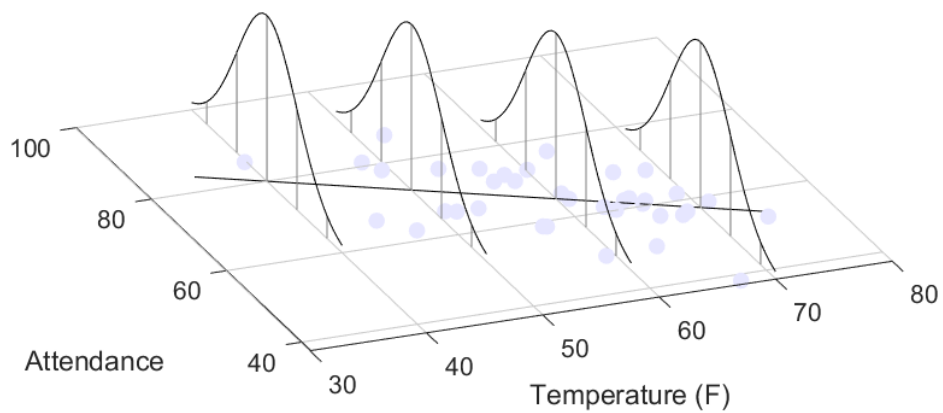
On the other hand, if the correlation is zero, then $S_{yx} = S_y$. That is, the standard deviation of the values around the regression line is the same as the standard deviation of the y -values. Again, this should make sense. If the correlation is zero, then the slope of the regression line is zero, 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.

The figure below illustrates homoscedasticity for our example with temperature and attendance. Shown is the scatterplot and regression line, along with a set of normal distributions centered at on the scatterplot that spread above and below the regression line with a standard deviation of $s_{xy} = 8$ students of attendance.

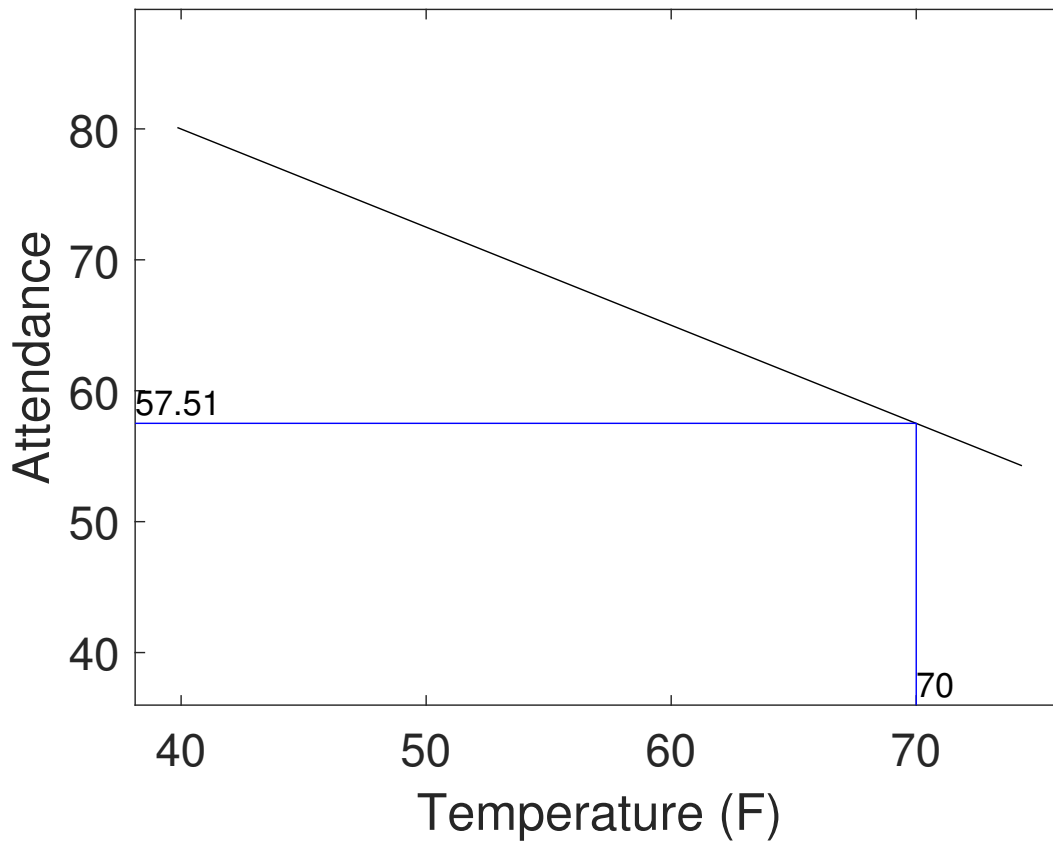


Here are some examples using regression, homoscedasticity and what we've learned about normal distributions to make predictions about the distribution of values above and below the regression line:

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 $x = 70$:

$$y' = mx + b = (-0.75)(70) + 109.97 = 57.51$$

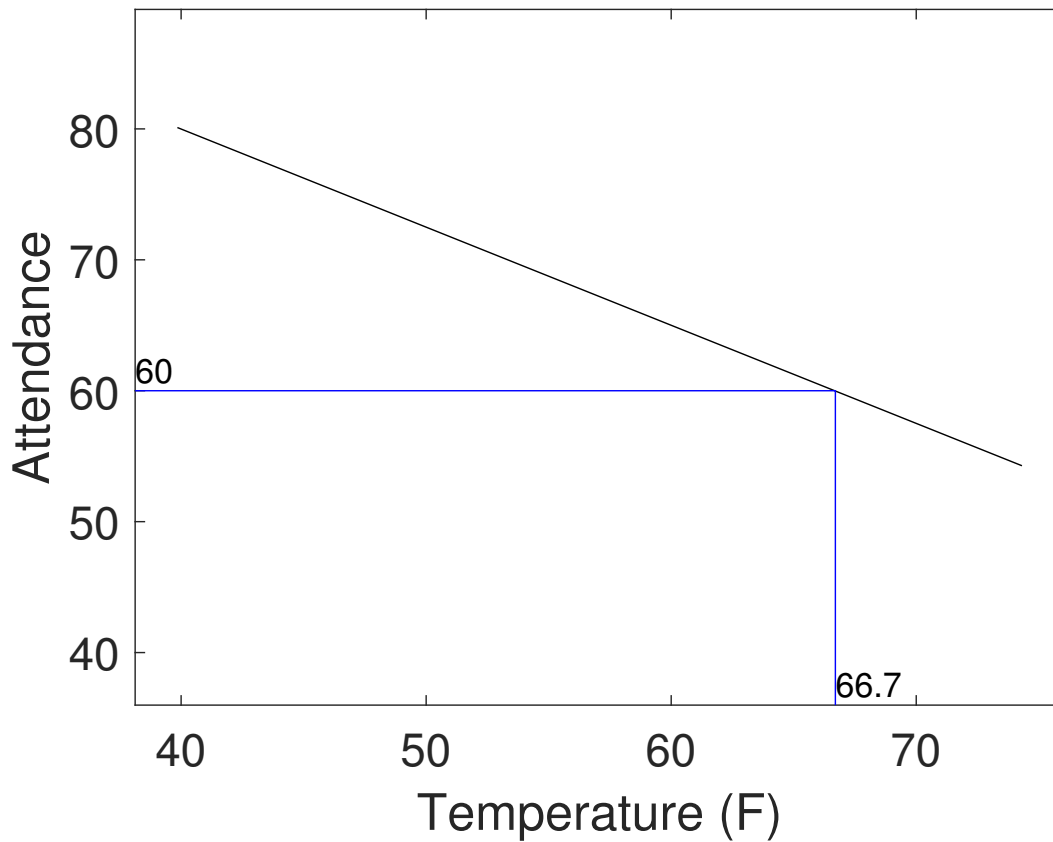


Example 2: What is the temperature for which the expected attendance is 60?

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

$$y' = 60 = (-0.75)x + 109.97$$

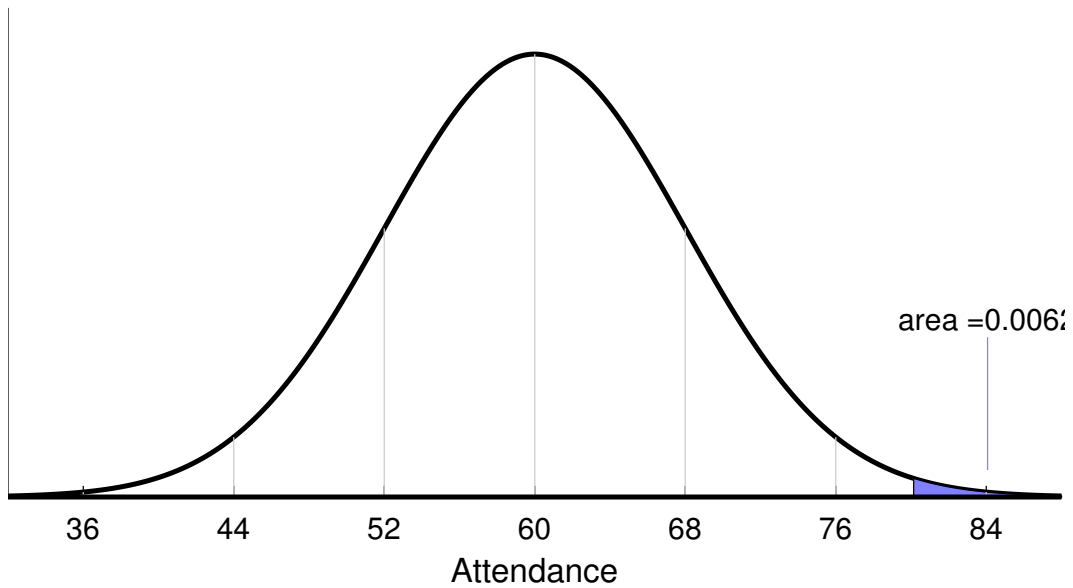
$$x = \frac{(60 - 109.97)}{-0.75} = 66.7$$



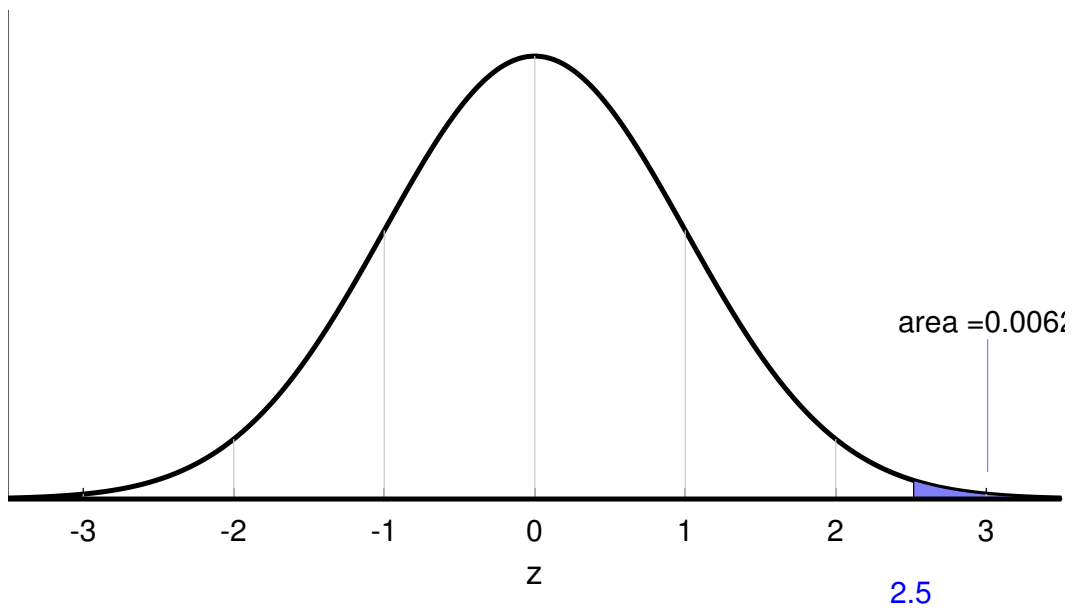
If the outdoor temperature is 66.7 degrees, we expect the attendance to be 60 students.

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 $S_{yx} = 8$. 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 8 above a value of 80:



The z-score is therefore $z = \frac{80-60}{8} = 2.5$:



Using Table A, the area above $z = 2.5$ is 0.0062. So the probability of 80 or more students showing up on a 66.7 degree day is 0.0062.

Example 4: This is a challenging one: 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.

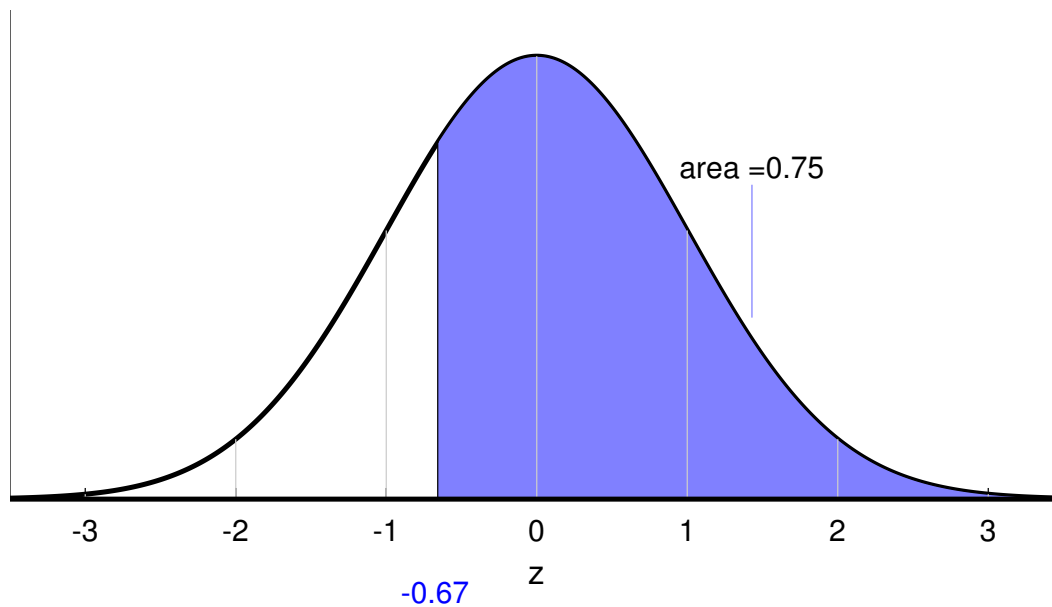
For those problems, we first find the z-score for the area, and then convert that to scores using:

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

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:

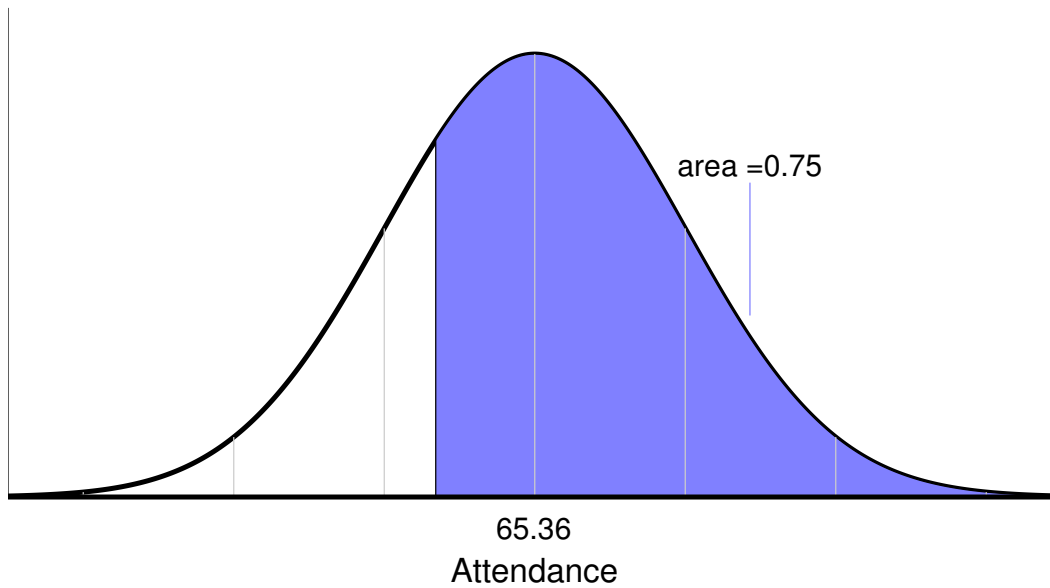


To convert to scores, we know that $z = -0.67$, we know that $y = 60$, and we know that $S_{yx} = 8$, so

$$60 = (-0.67)(8) + y'$$

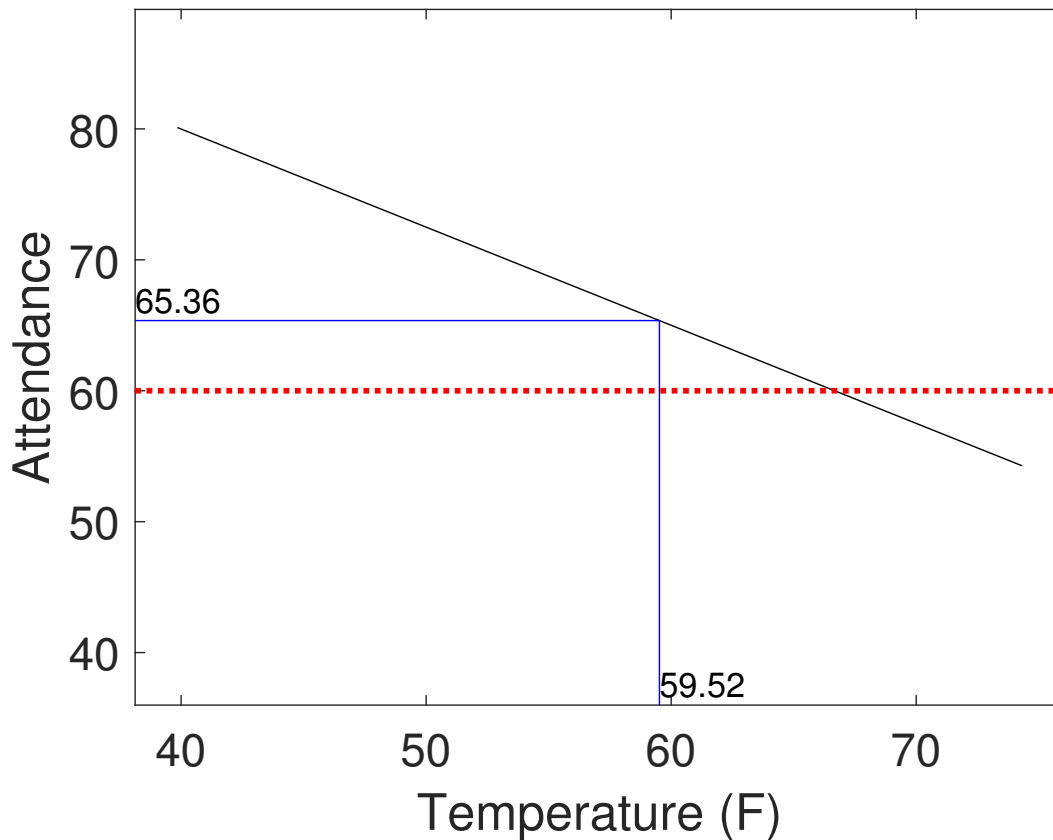
Solving for y' :

$$y' = 60 - (-0.67)(8) = 65.36$$



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:

$$x = \frac{(65.36 - 109.97)}{-0.75} = 59.52 \text{ degrees.}$$



In summary, when it's 59.52 degrees outside, then the distribution of attendance will have a mean of 65.36 and a standard deviation of 8. For this distribution of attendance, there is a 75 percent chance of 60 or more students showing up.

Using R to make interpretations about regression

The following script shows how to use R to do the examples above:

The R commands shown below can be found here: [Interpretation.R](#)

```
# Interpretation.R
#
# This script implements the examples in the 'interpretation
# of regression' tutorial.

# First, we'll load in the data from the tutorial:
mydata <-read.csv("http://www.courses.washington.edu/psy315/datasets/TemperatureAttendance.csv")
# We'll follow the procedure in the 'prediction.R' script to find
# the correlation, slope, intercept and syx for this data set.
```

```

# Define 'x' to be the temperature and y be the attendance:
x <- mydata$Temperature
y <- mydata$Attendance
# The scatterplot can now be done by plotting x vs y:
plot(x,y,
      xlab = "Temperature",
      ylab = "Attendance",
      pch = 19,
      col = "blue",
      cex = 1)
# draw the regression line
abline(lm(y~x))
# Calculate the correlation:
r <- cor(x,y,use = "complete.obs")
print(r)
[1] -0.5995988

# The slope of the regression line is:
m <- r*sd(y)/sd(x)
# The intercept is:
b <- mean(y) - m*mean(x)
# Calculate the standard error of the estimate, syx:
syx <- sd(y)*sqrt(1-r^2)
print(syx)
[1] 8.003007

# We're now ready to do the examples in the tutorial.

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

# For this we just find the value on the regression line
# where x = 70:
print(70*m+b)
[1] 57.50502

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

# For this we find x for y=60 on the regression line
# if  $y = m*x+b$ , then  $x = (y-b)/m$ 
print((60-b)/m)
[1] 66.67113

###
# Example 3: On a day that the outdoor temperature is

```

```

# 66.7, what is the probability that 80 or more students
# show up for class

# We know from Example 2 that if the temperature is 66.7
# then attendance will be distributed normally with a mean
# of 60 and the standard deviation will be syx.

# The probability of 80 or more showing up can be found
# using 'pnorm' (see NormalDistribution.R):
print(1-pnorm(80,60,syx))
[1] 0.006226152

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

# The z-score for which 75% of the area lies above is:
z <- -qnorm(.75)
# The attendance that is z standard deviations above 60 is:
yprime <- 60-syx*z
# The temperature for this value of y is (see Example 2:)
print((yprime-b)/m)
[1] 59.46905

```

Questions

It's your turn again. Here are 24 problems and answers, along with R code.

1) Let the regression line for y on x be $y' = 0.749x + 19.6$

Find the value of y' for which $x' = 77.76$

$$y' = (\text{slope})(x') + \text{intercept} = (0.749)(77.76) + 19.6 = 77.84$$

Answer: $y' = 77.84$

```
# y = slope*x + intercept
yprime <- 0.749*77.76 + 19.6
print(yprime)
[1] 77.84224
```

2) Let the mean of x be 13, the mean of y be 15
the standard deviation of x be 1.52 and the standard deviation of y be 1.83
and the correlation between x and y be 0.06

Find the value of y' for which $x' = 13.52$

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (0.06)\left(\frac{1.83}{1.52}\right) = 0.072$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 15 - (0.072)(13) = 14.06$$

$$y' = (\text{slope})(x') + \text{intercept} = (0.072)(13.52) + 14.06 = 15.03$$

Answer: $y' = 15.03$

```
slope <- 0.06*(1.83/1.52)
print(slope)
[1] 0.07223684
intercept <- 15 - slope*13
print(intercept)
[1] 14.06092
yprime <- 0.072*13.52 + 14.06
print(yprime)
[1] 15.03344
```

3) Let the regression line for y on x be $y' = 0.506x + 43.41$

Find the value of y' for which $x' = 16.95$

$$y' = (\text{slope})(x') + \text{intercept} = (0.506)(16.95) + 43.41 = 51.99$$

Answer: $y' = 51.99$

```
# y = slope*x + intercept
yprime <- 0.506*16.95 + 43.41
print(yprime)
[1] 51.9867
```

4) Let the regression line for y on x be $y' = 0.297x + 0.28$

Find the value of x' for which $y' = 18.91$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{18.91 - 0.28}{0.297} = 62.72$$

Answer: $x' = 62.72$

```
# x' = (y'-intercept)/slope
xprime <- (18.91-0.28)/0.297
print(xprime)
[1] 62.72727
```

5) Let the mean of x be 86, the mean of y be 36
the standard deviation of x be 1.7 and the standard deviation of y be 1.91
and the correlation between x and y be -0.91

Find the value of x' for which $y' = 36.78$

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (-0.91)\left(\frac{1.91}{1.7}\right) = -1.022$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 36 - (-1.022)(86) = 123.93$$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{36.78 - 123.93}{-1.022} = 85.27$$

Answer: $x' = 85.27$

```
slope <- -0.91*(1.91/1.7)
print(slope)
[1] -1.022412
intercept <- 36 - slope*86
print(intercept)
[1] 123.9274
# x' = (y'-intercept)/slope
xprime <- (36.78-intercept)/slope
print(xprime)
[1] 85.2371
```

6) Let the regression line for y on x be $y' = -0.331x + 137.46$

Find the value of x' for which $y' = 109.03$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{109.03 - 137.46}{-0.331} = 85.88$$

Answer: $x' = 85.88$

```
# x' = (y'-intercept)/slope
xprime <- (109.03-137.46)/-0.331
print(xprime)
[1] 85.89124
```

7) Let the regression line for y on x be $y' = -0.738x + 137.76$
and the standard error of the estimate be $S_{yx} = 0.91$

For $x = 84.61$, find the value of y for which 37 percent of the values lie above.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.738)(84.61) + 137.76 = 75.32$

For 37 percent of area above, $z = 0.33$

$y = (z)(s_{yx}) + y' = (0.33)(0.91) + 75.32 = 75.62$

Answer: $y = 75.62$

```
# y' = slope*x' + intercept
yprime <- -0.738*84.61 + 137.76
print(yprime)
[1] 75.31782
z <- -qnorm(0.37)
print(z)
[1] 0.3318533
#y = z*syx + yprime
y<- 0.33*0.91+75.32
print(y)
[1] 75.6203
```

8) Let the mean of x be 54, the mean of y be 87
the standard deviation of x be 1.88 and the standard deviation of y be 1.27
and the correlation between x and y be -0.04

For $x = 53.62$, find the value of y for which 76 percent of the values lie above.

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (-0.04)\left(\frac{1.27}{1.88}\right) = -0.027$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 87 - (-0.027)(54) = 88.46$$

$$s_{yx} = s_y\sqrt{1 - r^2} = (1.27)\sqrt{1 - (-0.04)^2} = 1.27$$

$$\text{regression line: } y' = (\text{slope})(x') + \text{intercept} = (-0.027)(53.62) + 88.46 = 87.01$$

For 76 percent of area above, $z = -0.71$

$$y = (z)(s_{yx}) + y' = (-0.71)(1.27) + 87.01 = 86.11$$

Answer: $y = 86.11$

```
slope <- -0.04*(1.27/1.88)
print(slope)
[1] -0.02702128
intercept <- 87 - slope*54
print(intercept)
[1] 88.45915
syx <- 1.27*sqrt(1-(-0.04)^2)
print(syx)
[1] 1.268984
# y' = slope*x' + intercept
yprime <- slope*53.62 + 88.46
print(yprime)
[1] 87.01112
z <- -qnorm(0.76)
print(z)
[1] -0.7063026
#y = z*syx + yprime
y<- z*syx+yprime
print(y)
[1] 86.11483
```

9) Let the regression line for y on x be $y' = 0.857x + 35.87$
and the standard error of the estimate be $S_{yx} = 0.82$

For $x = 27.87$, find the value of y for which 9 percent of the values lie above.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (0.857)(27.87) + 35.87 = 59.75$

For 9 percent of area above, $z = 1.34$

$y = (z)(s_{yx}) + y' = (1.34)(0.82) + 59.75 = 60.85$

Answer: $y = 60.85$

```
# y' = slope*x' + intercept
yprime <- 0.857*27.87 + 35.87
print(yprime)
[1] 59.75459
z <- -qnorm(0.09)
print(z)
[1] 1.340755
#y = z*syx + yprime
y<- 1.34*0.82+59.75
print(y)
[1] 60.8488
```

10) Let the regression line for y on x be $y' = -0.645x + 100.72$
and the standard error of the estimate be $S_{yx} = 1.77$

For $x = 60.23$, find the value of y for which 99 percent of the values lie below.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.645)(60.23) + 100.72 = 61.87$

For 99 percent of area below, $z = 2.33$

$y = (z)(s_{yx}) + y' = (2.33)(1.77) + 61.87 = 65.99$

Answer: $y = 65.99$

```
# y' = slope*x' + intercept
yprime <- -0.645*60.23 + 100.72
print(yprime)
[1] 61.87165
z <- qnorm(0.99)
print(z)
[1] 2.326348
#y = z*syx + yprime
y<- 2.33*1.77+61.87
print(y)
[1] 65.9941
```

11) Let the mean of x be 19, the mean of y be 105
the standard deviation of x be 1.49 and the standard deviation of y be 1.27
and the correlation between x and y be -0.85

For $x = 19.54$, find the value of y for which 50 percent of the values lie below.

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (-0.85)\left(\frac{1.27}{1.49}\right) = -0.724$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 105 - (-0.724)(19) = 118.77$$

$$s_{yx} = s_y\sqrt{1 - r^2} = (1.27)\sqrt{1 - (-0.85)^2} = 0.67$$

$$\text{regression line: } y' = (\text{slope})(x') + \text{intercept} = (-0.724)(19.54) + 118.77 = 104.62$$

For 50 percent of area below, $z = 0$

$$y = (z)(s_{yx}) + y' = (0)(0.67) + 104.62 = 104.62$$

Answer: $y = 104.62$

```
slope <- -0.85*(1.27/1.49)
print(slope)
[1] -0.7244966
intercept <- 105 - slope*19
print(intercept)
[1] 118.7654
syx <- 1.27*sqrt(1-(-0.85)^2)
print(syx)
[1] 0.669014
# y' = slope*x' + intercept
yprime <- slope*19.54 + 118.77
print(yprime)
[1] 104.6133
z <- qnorm(0.5)
print(z)
[1] 0
#y = z*syx + yprime
y<- z*syx+yprime
print(y)
[1] 104.6133
```

12) Let the regression line for y on x be $y' = -0.065x + 69.56$
and the standard error of the estimate be $S_{yx} = 1.28$

For $x = 101.09$, find the value of y for which 94 percent of the values lie below.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.065)(101.09) + 69.56 = 62.99$

For 94 percent of area below, $z = 1.55$

$y = (z)(s_{yx}) + y' = (1.55)(1.28) + 62.99 = 64.97$

Answer: $y = 64.97$

```
# y' = slope*x' + intercept
yprime <- -0.065*101.09 + 69.56
print(yprime)
[1] 62.98915
z <- qnorm(0.94)
print(z)
[1] 1.554774
#y = z*syx + yprime
y<- 1.55*1.28+62.99
print(y)
[1] 64.974
```

13) Let the regression line for y on x be $y' = -1.257x + 202.37$

Find the value of y' for which $x' = 86.9$

$$y' = (\text{slope})(x') + \text{intercept} = (-1.257)(86.9) + 202.37 = 93.14$$

Answer: $y' = 93.14$

```
# y = slope*x + intercept
yprime <- -1.257*86.9 + 202.37
print(yprime)
[1] 93.1367
```

14) Let the mean of x be 73, the mean of y be 84
the standard deviation of x be 1.69 and the standard deviation of y be 1.87
and the correlation between x and y be 0.45

Find the value of y' for which $x' = 73.57$

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (0.45)\left(\frac{1.87}{1.69}\right) = 0.498$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 84 - (0.498)(73) = 47.65$$

$$y' = (\text{slope})(x') + \text{intercept} = (0.498)(73.57) + 47.65 = 84.29$$

Answer: $y' = 84.29$

```
slope <- 0.45*(1.87/1.69)
print(slope)
[1] 0.497929
intercept <- 84 - slope*73
print(intercept)
[1] 47.65118
yprime <- 0.498*73.57 + 47.65
print(yprime)
[1] 84.28786
```

15) Let the regression line for y on x be $y' = 0.02x + 44.2$

Find the value of y' for which $x' = 41.04$

$$y' = (\text{slope})(x') + \text{intercept} = (0.02)(41.04) + 44.2 = 45.02$$

Answer: $y' = 45.02$

```
# y = slope*x + intercept
yprime <- 0.02*41.04 + 44.2
print(yprime)
[1] 45.0208
```

16) Let the regression line for y on x be $y' = -0.13x + 49.18$

Find the value of x' for which $y' = 37.05$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{37.05 - 49.18}{-0.13} = 93.3$$

Answer: $x' = 93.3$

```
# x' = (y'-intercept)/slope
xprime <- (37.05-49.18)/-0.13
print(xprime)
[1] 93.30769
```

17) Let the mean of x be 28, the mean of y be 25
the standard deviation of x be 1.47 and the standard deviation of y be 1.29
and the correlation between x and y be 0.14

Find the value of x' for which $y' = 25.09$

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (0.14)\left(\frac{1.29}{1.47}\right) = 0.123$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 25 - (0.123)(28) = 21.56$$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{25.09 - 21.56}{0.123} = 28.68$$

Answer: $x' = 28.68$

```
slope <- 0.14*(1.29/1.47)
print(slope)
[1] 0.1228571
intercept <- 25 - slope*28
print(intercept)
[1] 21.56
# x' = (y'-intercept)/slope
xprime <- (25.09-intercept)/slope
print(xprime)
[1] 28.73256
```

18) Let the regression line for y on x be $y' = 0.743x + 61.44$

Find the value of x' for which $y' = 106.21$

Since $y' = (\text{slope})(x') + \text{intercept}$

$$x' = \frac{y' - \text{intercept}}{\text{slope}} = \frac{106.21 - 61.44}{0.743} = 60.26$$

Answer: $x' = 60.26$

```
# x' = (y'-intercept)/slope
xprime <- (106.21-61.44)/0.743
print(xprime)
[1] 60.25572
```

19) Let the regression line for y on x be $y' = -0.489x + 71.93$
and the standard error of the estimate be $S_{yx} = 1.33$

For $x = 94.17$, find the value of y for which 29 percent of the values lie above.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.489)(94.17) + 71.93 = 25.88$

For 29 percent of area above, $z = 0.55$

$y = (z)(s_{yx}) + y' = (0.55)(1.33) + 25.88 = 26.61$

Answer: $y = 26.61$

```
# y' = slope*x' + intercept
yprime <- -0.489*94.17 + 71.93
print(yprime)
[1] 25.88087
z <- -qnorm(0.29)
print(z)
[1] 0.5533847
#y = z*syx + yprime
y<- 0.55*1.33+25.88
print(y)
[1] 26.6115
```

20) Let the mean of x be 37, the mean of y be 10
the standard deviation of x be 1.13 and the standard deviation of y be 1.09
and the correlation between x and y be -0.17

For $x = 37.16$, find the value of y for which 97 percent of the values lie above.

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (-0.17)\left(\frac{1.09}{1.13}\right) = -0.164$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 10 - (-0.164)(37) = 16.07$$

$$s_{yx} = s_y\sqrt{1 - r^2} = (1.09)\sqrt{1 - (-0.17)^2} = 1.07$$

$$\text{regression line: } y' = (\text{slope})(x') + \text{intercept} = (-0.164)(37.16) + 16.07 = 9.98$$

For 97 percent of area above, $z = -1.88$

$$y = (z)(s_{yx}) + y' = (-1.88)(1.07) + 9.98 = 7.97$$

Answer: $y = 7.97$

```
slope <- -0.17*(1.09/1.13)
print(slope)
[1] -0.1639823
intercept <- 10 - slope*37
print(intercept)
[1] 16.06735
syx <- 1.09*sqrt(1-(-0.17)^2)
print(syx)
[1] 1.074134
# y' = slope*x' + intercept
yprime <- slope*37.16 + 16.07
print(yprime)
[1] 9.976418
z <- -qnorm(0.97)
print(z)
[1] -1.880794
#y = z*syx + yprime
y<- z*syx+yprime
print(y)
[1] 7.956193
```

21) Let the regression line for y on x be $y' = -0.408x + 41.89$
and the standard error of the estimate be $S_{yx} = 1.81$

For $x = 34.31$, find the value of y for which 11 percent of the values lie above.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.408)(34.31) + 41.89 = 27.89$

For 11 percent of area above, $z = 1.23$

$y = (z)(s_{yx}) + y' = (1.23)(1.81) + 27.89 = 30.12$

Answer: $y = 30.12$

```
# y' = slope*x' + intercept
yprime <- -0.408*34.31 + 41.89
print(yprime)
[1] 27.89152
z <- -qnorm(0.11)
print(z)
[1] 1.226528
#y = z*syx + yprime
y<- 1.23*1.81+27.89
print(y)
[1] 30.1163
```

22) Let the regression line for y on x be $y' = -0.153x + 60.96$
and the standard error of the estimate be $S_{yx} = 1.12$

For $x = 77.92$, find the value of y for which 50 percent of the values lie below.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (-0.153)(77.92) + 60.96 = 49.04$

For 50 percent of area below, $z = 0$

$y = (z)(s_{yx}) + y' = (0)(1.12) + 49.04 = 49.04$

Answer: $y = 49.04$

```
# y' = slope*x' + intercept
yprime <- -0.153*77.92 + 60.96
print(yprime)
[1] 49.03824
z <- qnorm(0.5)
print(z)
[1] 0
#y = z*syx + yprime
y<- 0*1.12+49.04
print(y)
[1] 49.04
```

23) Let the mean of x be 95, the mean of y be 69
the standard deviation of x be 1.15 and the standard deviation of y be 1.59
and the correlation between x and y be 0.91

For $x = 95.49$, find the value of y for which 56 percent of the values lie below.

$$\text{slope} = r\left(\frac{s_y}{s_x}\right) = (0.91)\left(\frac{1.59}{1.15}\right) = 1.258$$

$$\text{intercept} = \bar{y} - (\text{slope})(\bar{x}) = 69 - (1.258)(95) = -50.53$$

$$s_{yx} = s_y\sqrt{1 - r^2} = (1.59)\sqrt{1 - (0.91)^2} = 0.66$$

$$\text{regression line: } y' = (\text{slope})(x') + \text{intercept} = (1.258)(95.49) + -50.53 = 69.6$$

For 56 percent of area below, $z = 0.15$

$$y = (z)(s_{yx}) + y' = (0.15)(0.66) + 69.6 = 69.7$$

Answer: $y = 69.7$

```
slope <- 0.91*(1.59/1.15)
print(slope)
[1] 1.258174
intercept <- 69 - slope*95
print(intercept)
[1] -50.52652
syx <- 1.59*sqrt(1-(0.91)^2)
print(syx)
[1] 0.6592271
# y' = slope*x' + intercept
yprime <- slope*95.49 + -50.53
print(yprime)
[1] 69.61303
z <- qnorm(0.56)
print(z)
[1] 0.1509692
#y = z*syx + yprime
y<- z*syx+yprime
print(y)
[1] 69.71255
```

24) Let the regression line for y on x be $y' = 0.224x + 12.58$
and the standard error of the estimate be $S_{yx} = 1.96$

For $x = 50.65$, find the value of y for which 25 percent of the values lie below.

regression line: $y' = (\text{slope})(x') + \text{intercept} = (0.224)(50.65) + 12.58 = 23.93$

For 25 percent of area below, $z = -0.67$

$y = (z)(s_{yx}) + y' = (-0.67)(1.96) + 23.93 = 22.62$

Answer: $y = 22.62$

```
# y' = slope*x' + intercept
yprime <- 0.224*50.65 + 12.58
print(yprime)
[1] 23.9256
z <- qnorm(0.25)
print(z)
[1] -0.6744898
#y = z*syx + yprime
y<- -0.67*1.96+23.93
print(y)
[1] 22.6168
```