Prediction from regression

January 9, 2021

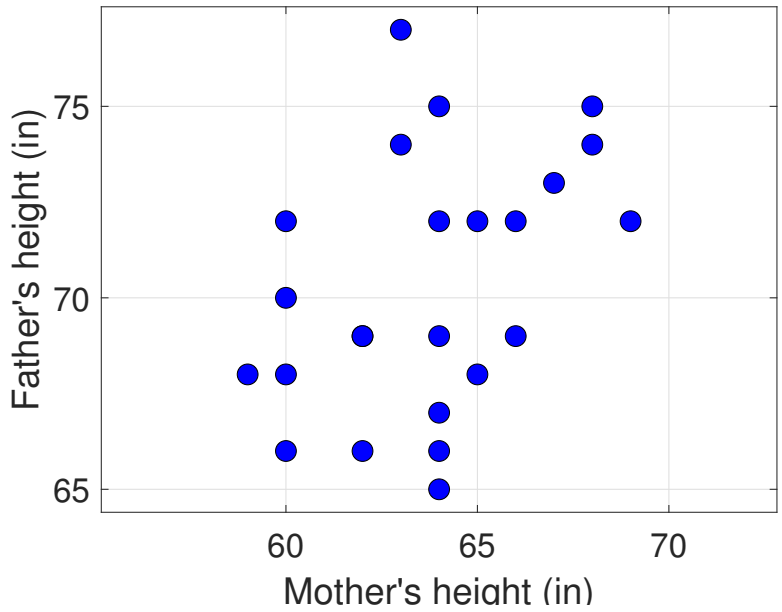
Contents

In the correlation tutorial we discussed how the Pearson correlation coeficient is a measure of how well a set of points are fit by a line. In this tutorial, we'll show how to find that best-fitting line, called the **regression line**. We'll then talk about how to use this line to predict values of y from arbitrary values of x.

We previously used as an example the relation between the heights of mothers and fathers of students that chose Green as their favorite color. We'll use the same example data set here.

You can download the csv file containing the data for this tutorial here: parentheight-Green.csv

Here's the scatterplot again:



We also calculated the Pearson correlation, r, which has the formula:

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum x)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum y)^2}{n}\right)}}$$

The summary statistics for this data set are:

x	y	xy	x^2	y^2
60	72	4320	3600	5184
63	74	4662	3969	5476
63	77	4851	3969	5929
64	75	4800	4096	5625
64	72	4608	4096	5184
65	72	4680	4225	5184
68	75	5100	4624	5625
68	74	5032	4624	5476
67	73	4891	4489	5329
66	72	4752	4356	5184
69	72	4968	4761	5184
66	69	4554	4356	4761
65	68	4420	4225	4624
64	69	4416	4096	4761
64	67	4288	4096	4489
64	66	4224	4096	4356
64	65	4160	4096	4225
62	66	4092	3844	4356
60	66	3960	3600	4356
62	69	4278	3844	4761
62	69	4278	3844	4761
60	68	4080	3600	4624
59	68	4012	3481	4624
60	70	4200	3600	4900

$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$
1529	1688	107626	97587	118978

Plugging these sums into the formula gives:

$$r = \frac{107626 - \frac{(1529)(1688)}{24}}{\sqrt{\left(97587 - \frac{(1529)^2}{24}\right)\left(118978 - \frac{(1688)^2}{24}\right)}} = 0.41$$

A first guess

Looking the scatterplot, you can probably make a guess at the best fitting line through the data. Now we'll define how to calculate that best fitting line.

First, we need to define what we mean by 'best fit'. Let's pick a line that looks like it might be a reasonable fit. Remember, lines can be defined by a choosing a point that it goes through, and a slope.

A reasonable guess that the line passes through the mean height of US women on the x-axis and the mean height of men on the y-axis, which is the point x= 64 and y = 70.

Let's also assume that for every inch that a mother is taller, the corresponding father is also an inch taller. That means that the line has a slope of 1. From these values we can find the equation of our firt guess of the best-fitting line.

If (x_C, y_C) is the point that the line goes through, and m is the slope, then the equation of the line is:

$$y - y_c = m(x - x_c)$$

This can be rearranged into the 'slope-intercept' formula:

$$y' = mx + [y_c - mx_c]$$

y' is used here instead of y because in statistics the ' (prime) symbol is used to indicate a predicted value. Our line has the equation:

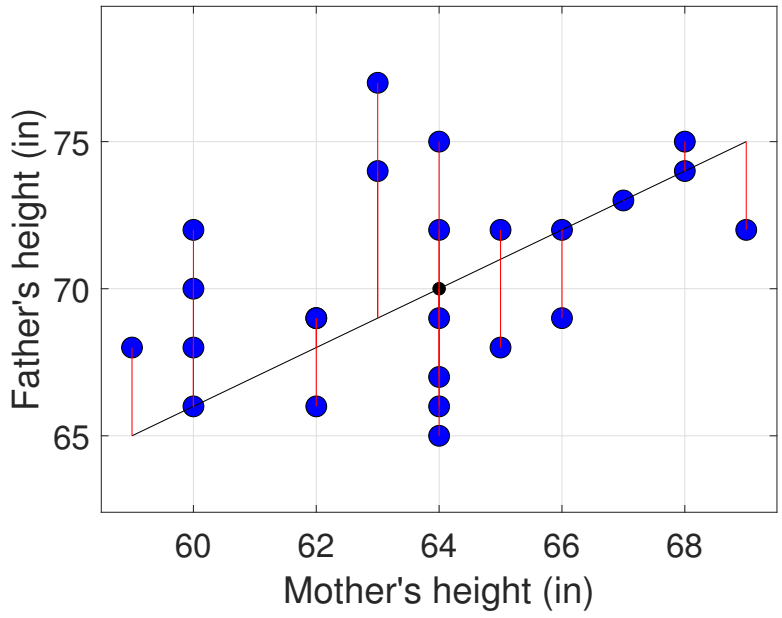
$$y' = 1x + [70 - (1)(64)]$$

Or

$$y' = 1x + 6$$

We can draw this line by picking two points on the line and drawing a line through them. We already have one point: (64,70). We can pick a second point, say, 2 inches to the right, x = 66. The y' value for this point is $y' = (1)(66) + 6 = 72$.

Here's that line drawn through the data:



Residuals

At first glance it looks like a reasonable fit. But to find the actual best-fitting line, we need to define what it means to be a good fit. With regression, the best fitting line is the defined as the line that minimizes the sums of squared deviation from the line to the data.

I've drawn red lines showing these deviations, called **residuals**, between the y value for each point and the value of y'.

Here's a table showing the values of x, y, y', the residuals, and the residuals squared:

x	y	$y' = 1x + 6$	$y - y'$	$(y - y')^2$
60	72	66	6	36
63	74	69	5	25
63	77	69	8	64
64	75	70	5	25
64	72	70	2	4
65	72	71	1	1
68	75	74	1	1
68	74	74	0	0
67	73	73	0	0
66	72	72	0	0
69	72	75	-3	9
66	69	72	-3	9
65	68	71	-3	9
64	69	70	-1	1
64	67	70	-3	9
64	66	70	-4	16
64	65	70	-5	25
62	66	68	-2	4
60	66	66	0	0
62	69	68	1	1
62	69	68	1	1
60	68	66	2	4
59	68	65	3	9
60	70	66	4	16

Our measure of the best fit is the sums of these squared residuals:

$$\sum (y - y')^2 = 36 + 25 + ... + 16 = 269$$

This equation should remind you of our definitions of variance and standard deviation. Remember that if $\bar{y}$ is the mean if the values of y, then the sums of squared deviation from

the mean is:

$$SS_y = \sum (y - \bar{y})^2$$

This was used to calculuate the standard deviation:

$$s_y = \sqrt{\frac{SS_y}{n}} = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

We will now define an analogous statistic like the standard deviation, but for the fit of the regression line. It's called the **standard error of the estimate** and is given the symbol s_{yx}:

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

You can think of it as the average deviation of the line from the data.

For our example,

$$s_{yx} = \sqrt{\frac{269}{24}} = 3.35 \text{ inches.}$$

You can see how this is a sensible statistic for measuring how well the data fits the line. First, it's in the same units as our y-axis variable (inches). Second, s_{xy} is small when the line fits the data well.

Unless our initial guess happens to be right, we'll should be able to find a line that has a sum of squared residuals that is less than 3.35. The regression line is the line that has the smallest value of s_{yx}.

Finding the best fitting line through the data

To define a line we need a point and a slope. Finding a point on the regession line is easy: the best-fitting line passes through the mean of the x and y values, $(\bar{x}, \bar{y})$, which is the point (63.71,70.33) in our example.

The slope of the best-fitting line is:

$$m = r(\frac{s_y}{s_x})$$

where r is the correlation and s_y and s_x are our friends, the population standard deviations of y and x:

$$s_x = \sqrt{\frac{\sum (x-\bar{x})^2}{n}},\ s_y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

For this example, $s_x = 2.77$ and $s_y = 3.33$, so the slope is $m = (0.41)\frac{3.33}{2.77} = 0.49$.

Does it make sense that the slope of the regression line is related to the correlation? We'll have more to say about this in the tutorial on interpretation.

Putting it all together, using the point-slope formula, the equation of the best fitting line is:

$$y' - \bar{y} = r(\frac{s_y}{s_x})(x - \bar{x})$$

Which can be rearranged in the form of the slope-intercept formula as:

$$y' = r(\frac{s_y}{s_x})x + [\bar{y} - r(\frac{s_y}{s_x})\bar{x}]$$

For our example, the slope-intercept form of the regression line is:

$$y' - 70.33 = 0.49(x - 63.71)$$

And the slope-intercept form is:

$$y' = 0.49x + [70.33 - (0.49)(63.71)] = 0.49x + 39.11$$

Here's a table of the residuals for the best fitting line:

x	y	$y' = 0.49x + 39.11$	$y - y'$	$(y - y')^2$
60	72	68.51	3.49	12.1801
63	74	69.98	4.02	16.1604
63	77	69.98	7.02	49.2804
64	75	70.47	4.53	20.5209
64	72	70.47	1.53	2.3409
65	72	70.96	1.04	1.0816
68	75	72.43	2.57	6.6049
68	74	72.43	1.57	2.4649
67	73	71.94	1.06	1.1236
66	72	71.45	0.55	0.3025
69	72	72.92	-0.92	0.8464
66	69	71.45	-2.45	6.0025
65	68	70.96	-2.96	8.7616
64	69	70.47	-1.47	2.1609
64	67	70.47	-3.47	12.0409
64	66	70.47	-4.47	19.9809
64	65	70.47	-5.47	29.9209
62	66	69.49	-3.49	12.1801
60	66	68.51	-2.51	6.3001
62	69	69.49	-0.49	0.2401
62	69	69.49	-0.49	0.2401
60	68	68.51	-0.51	0.2601
59	68	68.02	-0.02	0.0004
60	70	68.51	1.49	2.2201

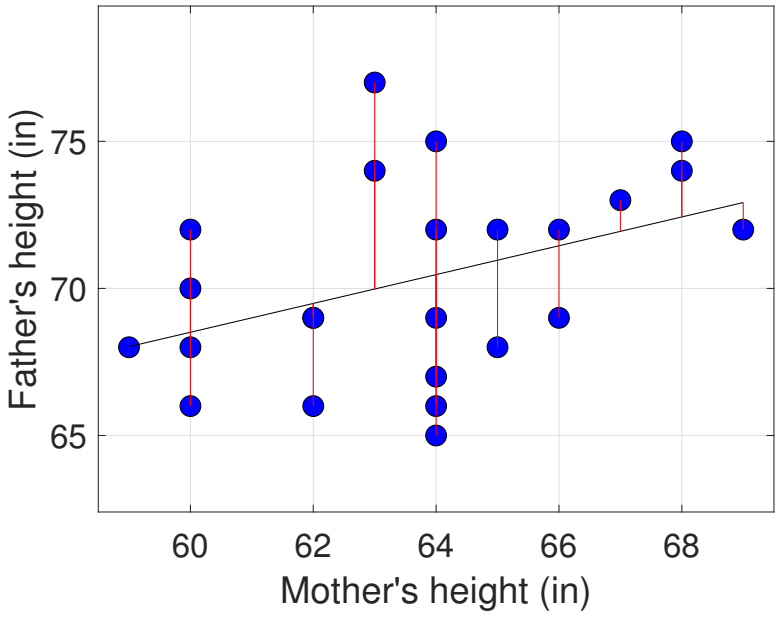
The sum of squared resdiduals is:

$$\sum (y - y')^2 = 12.1801 + 16.1604 + ... + 2.2201 = 213.2153$$

and the standard error of the estimate is: $s_{yx} = \sqrt{\frac{213.22}{24}} = 2.98$ inches.

Compare this value to standard error of the estimate that we calculated from our first guess (3.35). It's smaller. In fact, it will be smaller than any other combination of slope and y-intercept you can think of.

Here's the scatterplot with the best-fitting line:



Using the regression line to predict y from x

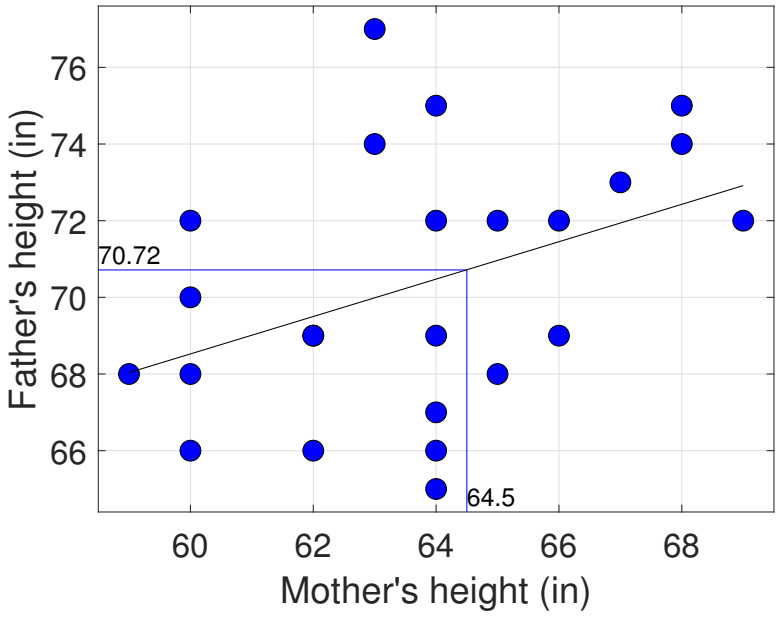
You can think of this regression line as a summary of our the relation between x and y. More specifically, if the correlation, r, is near plus or minus 1 then the regression line summarizes the linear relation between x and y.

This is useful because it lets us predict what y should be for arbitrary values of x. For example, we can guess what the height of a student's father should be, on average, for a student whose mother is 64.5 inches tall.

All we do is plug this value of x = 64.5 into the equation of the regression line:

$y'(64.5) = (0.49)(64.5) + 39.11 = 70.715$ inches.

Here's an illustration on the scatterplot:



IQ example

It is known that the correlations between IQ's of identical twins is 0.8. We also know that IQ's are distributed with a mean of 100 and standard deviation of 15.

What is the regression line that predicts the IQ of one twin based on the IQ of his/her sibling?

The equation of the regression line is:

$$y' - \bar{y} = r(\frac{s_y}{s_x})(x - \bar{x})$$

Which for our example is:

$$y' - 100 = 0.8(\frac{15}{15})(x - 100)$$

Or:

$$y' = 0.8x + 100 - (0.8)(100) = 0.8x + 20$$

We can now ask: What is the expected IQ of a twin whose sibling has an IQ of 115?

We find this by plugging in 115 into our regression equation:

$$y'(115) = 0.8(115) + 20 = 112$$

Are you surprised? You might expect that if the IQ's of siblings to have the same mean and standard deviations, then the expected IQ's of siblings should be the same. But they're not. A sibling that has a higher than average IQ of 115 will, on average, have an IQ that's

lower than their sibling: 112. However, the expected IQ is still higher than the mean IQ of 100.

A similar thing happens to for siblings that have IQ's that are lower than average, say an IQ of 85:

$$y'(85) = 0.8(85) + 20 = 88$$

If a twin has a below average IQ of 85 the expected IQ of their sibling is higher: 88. But still lower than the mean of 100.

This seems impossible - how can one twin be expected, on average, to have a different IQ than his or her sibling? And isn't the choice if which sibling you choose as the 'x' sibling and which is the 'y' arbitrary?

But according to the formula for the regression line, the only way to predict equal IQs between twins is to have a slope of 1. That would require the correlation to be 1 (since the standard deviations of x and y are the same).

A correlation less than 1 means that there are factors influencing the measured IQ of one sibling that are independent from the IQ of the other. These independent factors include day-to-day variability in the cognitive state of the test-taker, or the fact that not all twin siblings have had the exact same life experience.

So if you deliberately choose a twin with a high IQ, the IQ score measured on that day occurred due to a variety of factors which overall worked out in favor of that twin. Some of those factors will have nothing to do with genetics or heritability. That twin's sibling will, on average, not be as lucky. Therefore, on average, that twin's IQ will be closer to the mean.

This phenomenon is aptly named **regression to the mean**, which we'll discuss later in 'Interpretative aspects of correlation and regression'.

Using regression lines to make predictions in R

On of R's main uses by statisticians is for regression. The following script shows how to use R to draw regression lines and use it to find y from x:

The R commands shown below can be found here: Prediction.R

```
# Prediction.R

# Calculating the regression line in R is easy.  Here we'll
# work through the example in the 'prediction' tutorial

# First we'll load in the survey data:
survey <-read.csv("http://www.courses.washington.edu/psy315/datasets/Psych315W21survey.csv")
# And find the students that chose 'Green' as their favorite color:
students.green <- survey$color=="Green"
```

```
# to simply things for later, let's define 'x' to be the mother's height
# and y be the father's height:
x <- survey$mheight[students.green]
y <- survey$pheight[students.green]
# There might be 'NA's in either x or y, so we'll find where they are
# and take them out:
goodvals <- !is.na(x) & !is.na(y)
x <- x[goodvals]
y <- y[goodvals]
# The scatterplot can now be done by plotting x vs y:
plot(x,y,
     xlab = "Mother's height",
     ylab = "Father's height",
     pch = 19,
     col = "blue",
     cex = 2)
# Now we'll calculate the slope and intercept for the regression line:

# We'll need the correlation:
r <- cor(x,y,use = "complete.obs")
print(r)
[1] 0.4061526
n <- length(x)
# The function 'sd' in R uses the sample standard deviation formula
# which has the 'n-1' in the demominator.  The way to turn this into'
# the sample standard deviation is to mulitply the population deviation
# by sqrt((n-1)/n).  We'll do this for both x and y:
sy <- sd(y)*sqrt((n-1)/n)
sx <- sd(x)*sqrt((n-1)/n)
# The slope, m, of the regression line is:
m <- r*sy/sx
print(m)
[1] 0.4878738

# where the 'sd' function is standard deviation.

# The intercept, b, is:
b <- mean(y) - m*mean(x)
print(b)
[1] 39.25171

# where 'mean' is the mean (of course)

# We can draw the regression line on the scatterplot by
# using R's 'abline' function, which takes in the
# intercept and slope as inputs:
abline(b,m)
```

```
# From the slope and intercept we can find the residuals
# which are the deviations of each data point from the
# regression line.  First we find the values on the
# regression line for each value of x:
yprime <- m*x+b
head(yprime)
[1] 68.52413 68.52413 70.47563 69.98776 72.91500 71.45138

# The residuals are the differences between yprime and y:
residuals <- yprime -y
head(residuals)
[1] -3.4758653 -1.4758653  3.4756299 -4.0122439  0.9149988 -0.5486226

# The standard error of the estimate is the standard deviation
# of the residuals:
syx <- sqrt(sum(residuals^2)/n)
print(syx)
[1] 2.980587

# The other way to calculate syx is using sy and r:
print(sy*sqrt(1-r^2))
[1] 2.980587

# where 'sqrt' is the square root, and r^2 is r squared.
# Is this the same number?  I hope so.

# The regresion line can be used to predict y from x.
# For example, our best guess at what the height of
# the father for a student who's mother's height is 64.5
# inches is:
xplot <-64.5;
yplot <- xplot*m+b;
print(yplot)
[1] 70.71957

# To plot the data point on the graph we'll use the 'points'
# function intead of the 'plot' function.  'plot' will erase
# our current graph and start over.
points(xplot,yplot,
       pch = 16,    # symbol type.  16 is a filled circle
       col = 'red', # color of symbol
       cex = 2)     # size of symbol
```