**Lab 10 – Correlation and Regression**

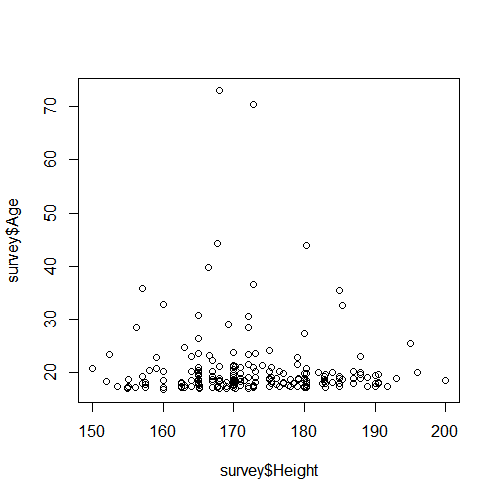
**To submit: answers to all numbered questions. When the question asks you to write code or create graphs, submit the code and/or graphs in the Word document as part of your answer. Also submit a single .R file that contains all of your code.**

In this lab, we will be testing the strength of linear correlations, then then finding lines of best fit and prediction intervals where appropriate. We will be working with the dataset **survey** from the **MASS** library.

Load and view **survey**. The first correlation we will be testing is the strength of the linear correlation between student age and student height. Since the students in **survey** are all adults, and people tend to stop growing before they reach adulthood, we would expect there to be no linear correlation (and in fact no correlation at all) between age and height.

A scatterplot supports that theory:

> plot(survey$Height, survey$Age)



The **cor** function returns the linear correlation coefficient. Here is what the help file has to say about **cor**:

cor(x, y = NULL, use = "everything",

method = c("pearson", "kendall", "spearman"))

cov2cor(V)

**Arguments**

|  |  |
| --- | --- |
| x | a numeric vector, matrix or data frame. |
| y | NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient). |
| na.rm | logical. Should missing values be removed? |
| use | an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs","complete.obs", "na.or.complete", or "pairwise.complete.obs". |
| method | a character string indicating which correlation coefficient (or covariance) is to be computed. One of"pearson" (default), "kendall", or "spearman": can be abbreviated. |
| V | symmetric numeric matrix, usually positive definite such as a covariance matrix. |

As usual, we don’t need all this information. In our case, all we need to do is provide R with the (x,y) data – in this case, height and age – and tell R how to handle missing values. Some students provided only a height, or only an age. We will ignore those, and find the correlation coefficient only for complete pairs of observations. We instruct R to do this with the argument **use=”complete.obs”**.

> cor(survey$Height, survey$Age, use="complete.obs")

[1] -0.0372773

That is, the correlation between student age and height is -0.0372773 – pretty close to zero. (Note: it is extremely unusual to find pairs of data for which r is exactly zero.)

This seems far enough from 1 to suggest that there is no linear correlation between student age and student height, but let’s run a formal significance test anyway with the command **cor.test**:

> cor.test(survey$Height, survey$Age)

Pearson's product-moment correlation

data: survey$Age and survey$Height

t = -0.5367, df = 207, p-value = 0.5921

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.17212114 0.09893785

sample estimates:

cor

-0.0372773

Notice that the output is very similar to the output of the **t.test** function:

* Like the **t.test**, **cor.test** uses a default confidence level of 0.95. We can specify a different confidence level the way we did last week.
* The alternative hypothesis is that the true correlation between student height and age is not zero. As before, the **p.value** output is particularly important for evaluating this claim. If there were no correlation between student height and student age, we would expect to get an r-value at least as far from 0 as the one we got (-0.0372773), 59.21% of the time. That’s pretty likely! Clearly, it’s safe to say that there’s no evidence of a linear correlation between student age and student height. (In general, if the p-value is greater than alpha, which in our case is 1-0.95=0.05, it’s safe to say there’s no evidence of a linear correlation.)
* The **cor.test** command also provides us with a confidence interval. Here we see that we are 95% sure that the population correlation coefficient ρ is between -0.17212114 and 0.09893795. The exact values here are not important. What matters for us is that the confidence interval ρ contains zero. That is…we can’t say with 95% confidence that ρ isn’t around zero. That’s equivalent to saying that we don’t have evidence of a linear correlation.

1. Create a function **correlationtest** that reads in paired (x,y) data, an optional significance level, and descriptions of the two variables, and returns a sentence that informs the user whether there is evidence of a linear correlation. You may wish to modify your Lab 10 code for this. Example input and output for the height and weight data:

> correlationtest(survey$Age, survey$Height, 0.05, "student age", "student height")

At alpha = 0.05 we have no evidence of a linear correlation between student age and student height .

> correlationtest(survey$Age, survey$Height,, "student age", "student height")

At alpha = 0.05 we have no evidence of a linear correlation between student age and student height .

Now let’s look at some data that is actually correlated. We would expect a strong correlation between the span of a student’s writing hand (**survey$Wr.Hnd**) and the span of a student’s non-writing hand (**survey$NW.Hnd**),

1. What is the correlation coefficient r for the spans of student hands?
2. Using the function you wrote in Question 1, confirm that there is a linear correlation between the span of a student’s writing hand and the span of a student’s non-writing hand. Your output should look like this:

> correlationtest(survey$NW.Hnd, survey$Wr.Hnd, 0.05, "the span of a student's writing hand", "the span of a student's non-writing hand")

At alpha = 0.05 we have evidence of a linear correlation between the span of a student's writing hand and the span of a student's non-writing hand .

1. Create a scatterplot of **NW.Hnd** vs **Wr.Hnd**. Give your scatterplot a meaningful title and meaningful axis labels.

Now that we have established a linear correlation between the span of a student’s writing hand and the span of a student’s non-writing hand, let’s find a line of best fit and use it to make predictions. We use the **lm** function (“linear model”) to do this.

> hands=lm(formula=NW.Hnd~Wr.Hnd, data=survey)

> hands

Call:

lm(formula = NW.Hnd ~ Wr.Hnd, data = survey)

Coefficients:

(Intercept) Wr.Hnd

0.04859 0.99277

This tells us that the best point estimate for the span of the non-writing hand for a student whose writing hand has a span of *x* is given by the equation .

We can plot the line of best fit on the scatterplot:

> abline(hands)

The line should now appear on your scatterplot. (If the line does not appear, make sure that NW.Hnd is on the y-axis and Wr.Hnd is on the x-axis.)

1. Find the best point estimate for the span of the non-writing hand for a student whose writing hand has a span of 20 cm.

We can use the **predict** command to get a prediction interval corresponding to our answer to Question 5. We will use the output of the **lm** command, which we have already stored in the variable **hands**. This will provide the coefficients of the equation for the line of best fit, as well as the standard error and the other values that are involved in creating a prediction interval. We also need to supply the value of **Wr.Hnd** for which we are finding a value of **NW.Hnd**. In this case, that value is 20 cm. For some reason known only to the writer of the **predict** function, we need to enter this value as a dataframe, rather than as the single number 20.

> WrHnd=data.frame(Wr.Hnd=20)

> WrHnd

Wr.Hnd

1 20

Here, **WrHnd** is a 1x2 array containing the index and value of its only entry, the span of the writing hand for which we’re creating a prediction interval. (Look, I’m as annoyed as you are by this nonsense. Sadly, I don’t make the rules.)

Finally, we can generate our prediction interval:

> predict(hands, WrHnd, interval="predict")

fit lwr upr

1 19.90393 18.66754 21.14032

The default confidence level is again 95%. (We could use the optional **level** argument to change this if we want, but we will not worry about that in this lab.) This tells us that we are 95% sure that if the span of a student’s writing hand is 20 cm, the span of their non-writing hand is between 18.66754 cm and 21.14032 cm.

Now let’s investigate the correlation between another pair of data: a student’s height, and the span of their writing hand.

1. Find the correlation coefficient for a student’s height, and the span of their writing hand. How does this number compare to the answer you got for Question 2 (larger, smaller, around the same)? Is this what you would expect? Explain.
2. Using the function you wrote in Question 1, determine whether there is a linear correlation between the height of a student and the span of a student’s writing hand. Provide your output.
3. Create a scatterplot of **Wr.Hnd** vs **Height**. Give your scatterplot a meaningful title and meaningful axis labels, and display the line of best fit directly on the graph.
4. At 95% confidence, predict the span of the writing hand for a student who is 173 cm tall. Include a sentence stating your conclusion.
5. How does your prediction interval compare to the prediction interval we got when we predicted the span for a non-writing hand when we were given the span of a writing hand? (Specifically – is this prediction interval narrower or wider than the last one?) Explain why this is the case, making reference to other values you have computed in this lab.