Lab 12 - Correlation and Linear Models

Carl Gladish

Nov 22, 2025

This lab contains some instructional material along with some questions. You are required to submit your answers in a Microsoft Word report produced by “knitting” an R Notebook to .docx format. Your Word file will contain *all* your R code *and* your written answers and charts.

Create your own R Markdown file (i.e., R notebook) called Lab_12_Notebook.Rmd.

Due date: 11:59pm, two school days from today (weekend days count as half)

Lab Objectives

- We will calculate r, the *linear correlation coefficient*
- We will test for statistical significance of r
- We will find lines of “best fit” (i.e., regression lines) and prediction intervals where appropriate.

For this lab, we will work with the dataset `survey` from the `MASS` library. To begin, load the required libraries and the data set.

```
library(MASS)
library(dplyr)
data(survey)
```

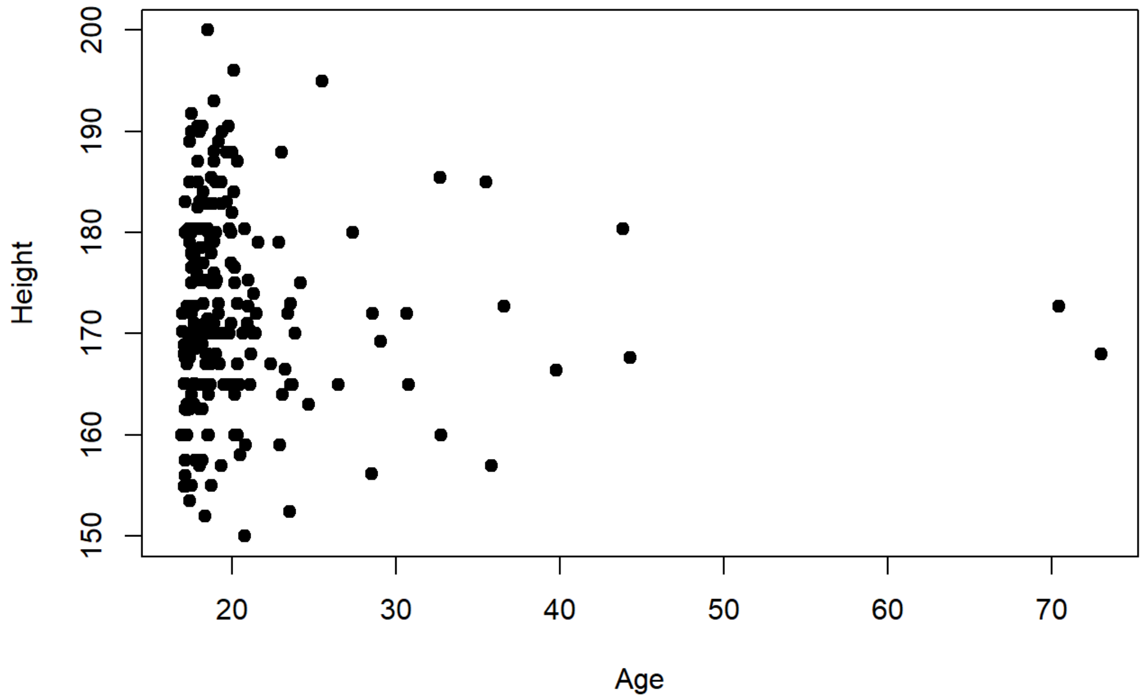
Reminder: The data set `survey` comes from a set of statistics students at the University of Adelaide. We could make the reasonable assumption that the sample is representative of all students at University of Adelaide.

Testing Strength of a Linear Correlation

We will first test the strength of the linear correlation between student **Age** and student **Height**. The students in this sample are adults and people tend to stop growing before they reach adulthood, so we would expect there to be no correlation between age and height.

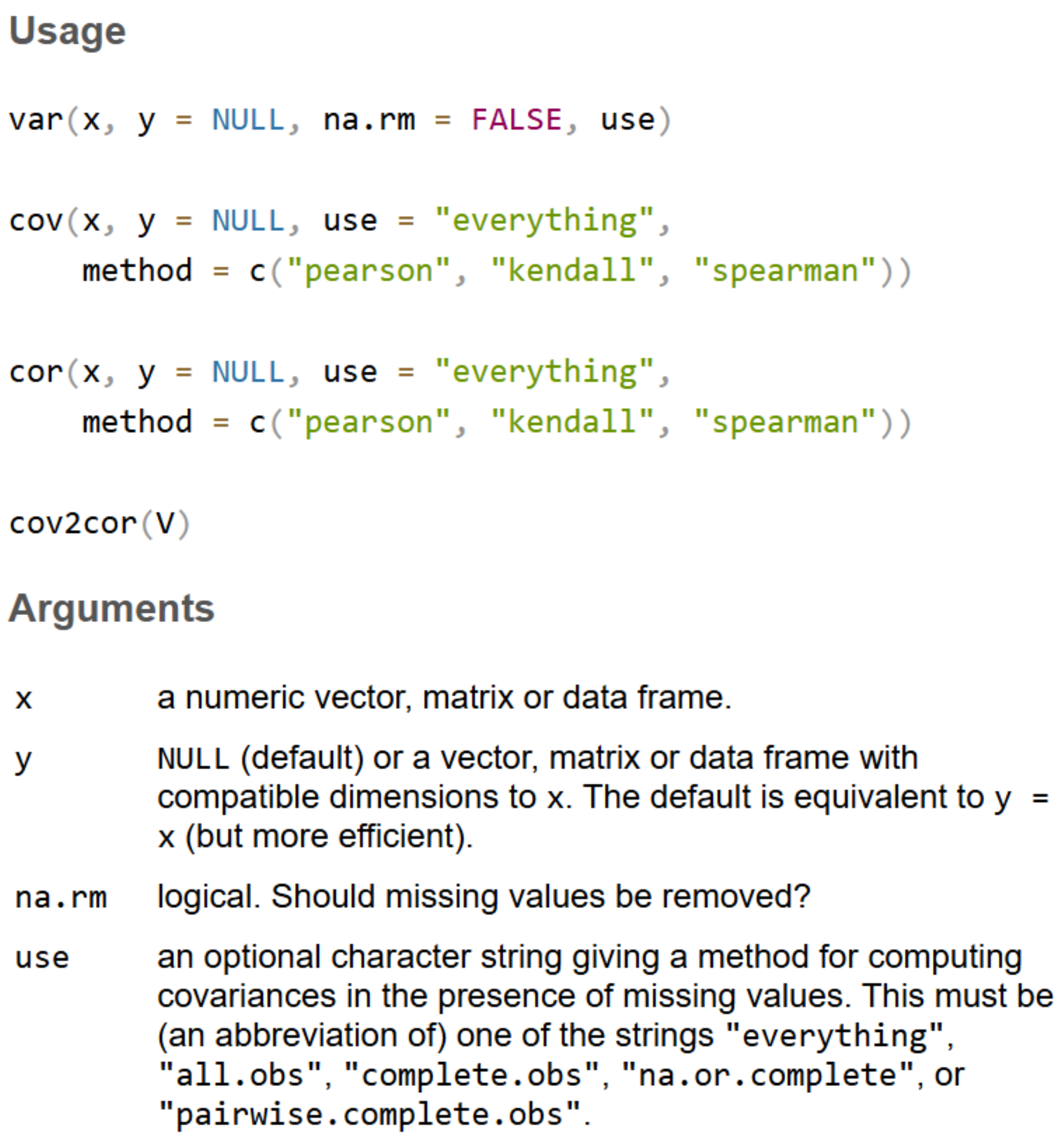
A scatter plot supports that theory:

```
plot(survey$Age, survey$Height, pch=19, type="p",
     xlab="Age", ylab="Height")
```



The function `cor` returns the linear correlation coefficient. You should open the helpfile for `cor` .

```
help(cor)
```



Usage

```
var(x, y = NULL, na.rm = FALSE, use)

cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

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"`.

Example usage:

```
x <- survey$Age
y <- survey$Height
r <- cor(x, y,  use="complete.obs")
r
```

```
## [1] -0.0372773
```

The most important arguments to `cor` are `x` and `y`, which provide the raw sample data for the two variables (X=**Age** and Y=**Height**).

The argument `use = "complete.obs"` tells R to ignore individuals where either **Age** or **Height** is NA.

The correlation between student **Age** and **Height** is -0.0373, which is quite close to zero. (It would be highly unlikely to have exactly $r = 0$ for real world data.)

Is this degree of correlation statistically significant? In other words, does the sample data give strong evidence that the *population* correlation ρ differs from zero?

To answer this, we run a formal hypothesis test on the hypotheses:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

```
cor.test(survey$Age, survey$Height,
         alternative="two.sided")
```

```
##
##  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
```

The output is very similar to the output when calling the function `t.test`. The important outputs are:

- A confidence interval for ρ. In this case, we have 95% confidence that $-0.1721 < \rho < 0.0989$. This is consistent with zero correlation ($\rho = 0$).
- p-value $= 0.5921$. In this case, we fail to reject the null hypothesis. So there is insufficient evidence to support the claim that $\rho \neq 0$. (There is no correlation at the population level.)

Question 1

Create a function `CorrelationTest` that takes arguments:

- `X.data` = raw data for variable X
- `Y.data` = raw data for variable Y
- `X.label` = label for variable X
- `Y.label` = label for variable Y
- `alpha` = significance level

The function returns a sentence stating the appropriate conclusion about the population proportion between X and Y.

Example

```
CorrelationTest(survey$Age, survey$Height, "student age", "student height", 0.05)
```

(Output:) At the 5% significance level, we have insufficient evidence of a linear correlation between student age and student height. (p-value=0.5921)

Demonstrate that your implementation of `CorrelationTest` can reproduce the example above.

Correlated Variables

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 (**Wr.Hnd**) and the span of a student's non-writing hand (**NW.Hnd**).

Question 2

What is the correlation coefficient r for the spans of students' writing hand span **Wr.Hnd** and non-writing hand span **NW.Hnd**?

Question 3

Using your function `CorrelationTest` from 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 at the population level. Record the input and output.

Question 4

Create a scatter plot of **Wr.Hnd** (x-axis) vs **NW.Hnd** (y-axis). Give your scatter plot a meaningful title and meaningful axis labels.

Linear Models

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 linear model (i.e., regression line) and use it to make predictions. We use the function `lm` (“linear model”) to do this.

When calling the function `lm` we use a “model formula” expression like

$$Y \sim X$$

which tells R that we want a linear model where Y depends on X. Here we will get a model in which **NW.Hnd** depends on **Wr.Hnd**:

```
model <- lm(NW.Hnd~Wr.Hnd, data=survey)
model
```

```
## 
## Call:
## lm(formula = NW.Hnd ~ Wr.Hnd, data = survey)
## 
## Coefficients:
## (Intercept)       Wr.Hnd  
##     0.04859      0.99277
```

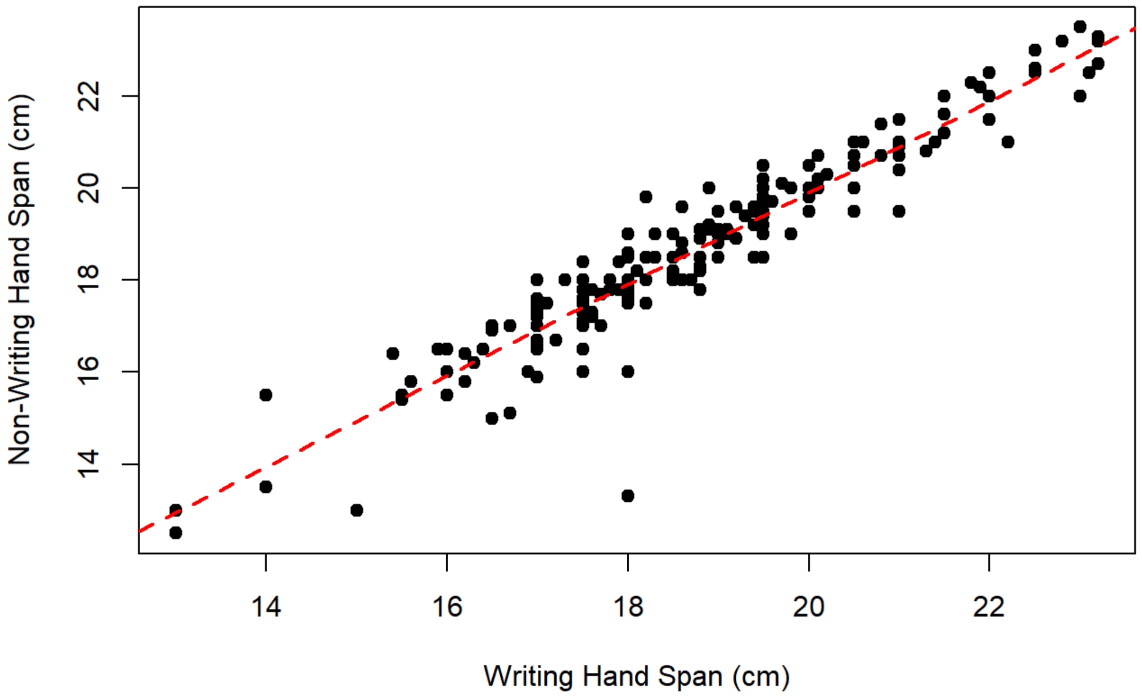
The linear model (i.e., “regression line” or “line of best fit”) is:

$$\hat{Y} = 0.04859 + 0.99277X$$

Here X denotes the student’s writing hand span and $\hat{Y}$ denotes the students *predicted* non-writing hand span.

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

```
plot(survey$Wr.Hnd, survey$NW.Hnd, pch=19, type="p",
     xlab="Writing Hand Span (cm)",
     ylab="Non-Writing Hand Span (cm)")
abline(model$coefficients, col="red",lty="dashed",lwd=2)
```



Question 5

Use the regression line equation to find the best point estimate for the span of the non-writing hand for a student whose writing hand has a span of 20.0 cm.

Prediction Intervals

We can use `predict` to get a *prediction interval* rather than a point estimate.

Let's make a series of predictions of **NW.Hnd** for the values:

- **Wr.Hnd** $= 20.0$
- **Wr.Hnd** $= 21.0$
- **Wr.Hnd** $= 22.0$
- **Wr.Hnd** $= 23.0$

First create a data frame that contains these predictor variable values:

```
predictor.vals <- data.frame( Wr.Hnd=c(20.0, 21.0, 22.0, 23.0))
```

Then run

```
model <- lm(NW.Hnd~Wr.Hnd, data=survey)
predict(model, predictor.vals)
```

```
##        1        2        3        4
## 19.90393 20.89670 21.88947 22.88223
```

This provides point estimates for the four values of **NW.Hnd**. To get prediction intervals, use

```
predict(model, predictor.vals, interval="predict")
```

```
##        fit      lwr      upr
## 1 19.90393 18.66754 21.14032
## 2 20.89670 19.65760 22.13580
## 3 21.88947 20.64619 23.13274
## 4 22.88223 21.63333 24.13114
```

For instance, we find that if a student's **Wr.Hnd** is $20.0\,\mathrm{cm}$, then their **NW.Hnd** is predicted to be between $18.67\,\mathrm{cm}$ and $21.14\,\mathrm{cm}$, with 95% confidence.

Question 6

You will now investigate the correlation between another pair of variables: a student's **Height**, and the span of their non-writing hand, **NW.Hnd**.

Find the correlation coefficient r for a student's **Height** and the span of their non-writing hand, **NW.Hnd**.

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.

Question 7

Using the function `CorrelationTest` you wrote in Question 1, determine whether there is a linear correlation at the population level between the **Height** of a student and **NW.Hnd**, the span of a student's non-writing hand.

Question 8

Create a scatter plot of **Height** (x-axis) vs **NW.Hnd** (y-axis). Give your scatter plot a meaningful title and meaningful axis labels. Plot the line of best fit directly on the graph.

Question 9

At 95% confidence, predict the span of the writing hand for a student who is $173\,\mathrm{cm}$ tall. Include a sentence stating your conclusion.

Question 10

Is the prediction interval for **NW.Hnd** narrower or wider when using **Height** as the predictor variable (i.e. x-axis variable) compared to when using **Wr.Hnd** as the predictor variable? Explain why, making reference to other values you have computed in this lab.

Question 11

Calculate the 95% prediction interval for **NW.Hnd** with **Height** $= 173\,\mathrm{cm}$ using the formulas

$$\hat{Y} - E < Y < \hat{Y} + E$$

$$E = t_{\alpha/2} \cdot S_e \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{X}\right)^2}{(n-1)s_x^2}}; \quad \mathrm{df} = n - 2$$

$$S_e = \sqrt{\frac{(\Sigma y)^2 - a\,(\Sigma y) - b\,(\Sigma xy)}{n-2}}$$