Covariance and Correlation

Advanced statistical methods and models in experimental design

Bartosz Maćkiewicz

Correlation and Regression

We will explore two interrelated topics: correlation and regression.

Statisticians often distinguish between these two techniques, although the distinction is not always observed in practice. However, it is significant enough to merit brief consideration.

In problems of simple correlation and regression, the data consist of two observations from each of N subjects, one observation on each of the two variables under consideration. If we were interested in the correlation between the running speed of mice in a maze (Y) and the number of trials to reach some criterion (X) (both common measures of learning), we would obtain a running-speed score and a trials-to-criterion score from each subject.

Similarly, if we were interested in the regression of running speed (Y) on the number of food pellets per reinforcement (X), each subject would have scores corresponding to his speed and the number of pellets he received.

The difference between these two situations illustrates the statistical distinction between correlation and regression.

Correlation and Regression

In both cases, Y (running speed) is a *random variable*, beyond the experimenter's control. We don't know what the mouse's running speed will be until we carry out a trial and measure the speed.

In the former case, X is also a random variable, because the number of trials to criterion depends on how fast the animal learns, and this, too, is beyond the control of the experimenter. Put another way, a replication of the experiment would leave us with different values of both Y and X. In the food pellet example, however, X is a fixed variable. The number of pellets is determined by the experimenter (for example, 0, 1, 2, or 3 pellets) and would remain constant across replications.

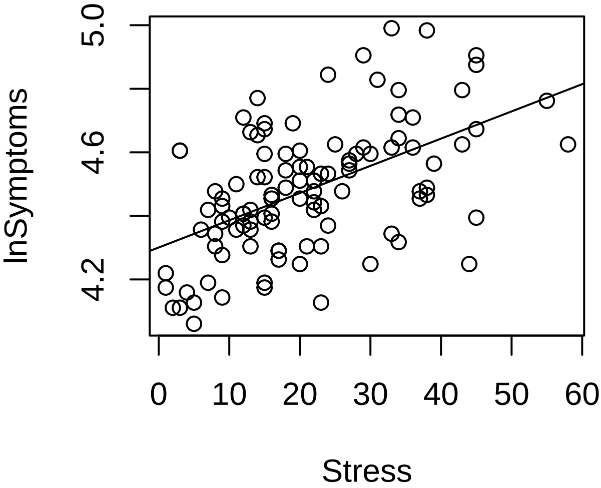
To some statisticians, the word **regression** is reserved for those situations in which **the value of X is fixed or specified by the experimenter before the data are collected**. In these situations, no sampling error is involved in X, and repeated replications of the experiment will involve the same set of X values. The word **correlation** is used to describe the situation in which **both X and Y are random variables**.

Example 1: Stress and mental health (`wagner1988.csv`)

Wagner, Compas, and Howell (1988) investigated the relationship between stress and mental health in first-year college students. Using a scale they developed to measure the frequency, perceived importance, and desirability of recent life events, they created a measure of negative events weighted by the reported frequency and the respondent's subjective estimate of the impact of each event. This served as their measure of the subject's perceived social and environmental stress. They also asked students to complete the Hopkins Symptom Checklist, assessing the presence or absence of 57 psychological symptoms.

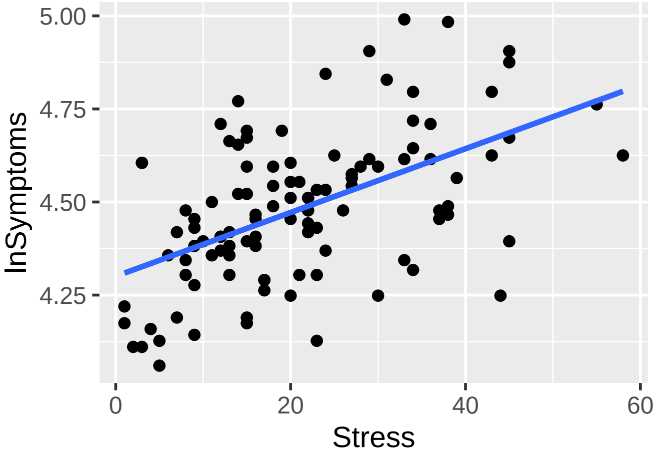
Scatterplot: visualization of bivariate data - base R

```
wagner <- read.csv("wagner1988.csv")
plot(lnSymptoms ~ Stress, data = wagner)
fit <- lm(lnSymptoms ~ Stress, data = wagner) # fit a linear model
abline(fit) # add regression line
```



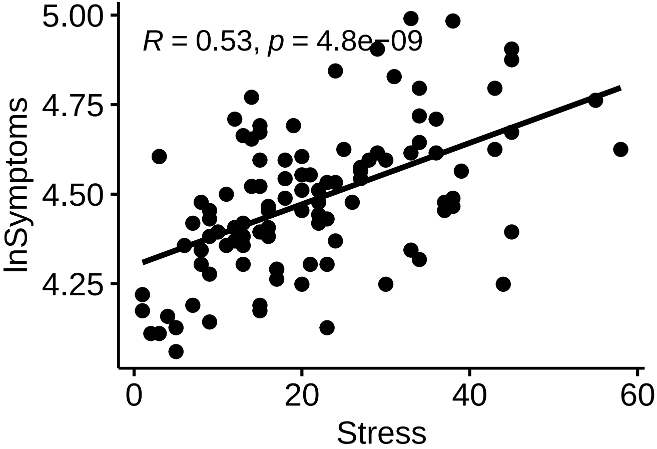
Scatterplot: visualization of bivariate data - ggplot2

```r
library(ggplot2)
ggplot(data = wagner) +
  aes(x = Stress, y = lnSymptoms) +
  geom_point() +
  geom_smooth(method = lm, se = F) # add regression line
```



Scatterplot: visualization of bivariate data - ggpubr

```r
library(ggpubr)
ggscatter(data = wagner, x = "Stress", y = "lnSymptoms",
          add = "reg.line", cor.coef = T)
```



The Covariance

The covariance is a number that reflects the degree to which two variables vary together.

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

The covariance is similar in form to the variance. If we changed all the Ys in the equation to Xs, we would have s_X^2 ; if we changed the Xs to Ys, we would have s_Y^2.

“Mechanics” of the Covariance

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case I: Positive Relationship

For the data on `Stress` and `lnSymptoms` we would expect that high stress scores will be paired with high symptom scores. Thus, for a stressed participant with many problems, both $(X - \bar{X})$ and $(Y - \bar{Y})$ will be positive and their product will be positive.
For a participant experiencing little stress and few problems, both $(X - \bar{X})$ and $(Y - \bar{Y})$ will be negative, but their product will again be positive. Thus, the sum of $(X - \bar{X})(Y - \bar{Y})$ will be large and positive, giving us a large positive covariance.

“Mechanics” of the Covariance

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case II: Negative Relationship

The reverse would be expected in the case of a strong negative relationship. Here, large positive values of $(X - \bar{X})$ most likely will be paired with large negative values of $(Y - \bar{Y})$, and *vice versa*. Thus, the sum of the products of the deviations will be large and negative, indicating a strong negative relationship.

“Mechanics” of the Covariance

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Case III: No relationship

Finally, consider a situation in which there is no relationship between X and Y. In this case, a positive value of $(X - \bar{X})$ will sometimes be paired with a positive value and sometimes with a negative value of $(Y - \bar{Y})$. The result is that the products of the deviations will be positive about half of the time and negative about half of the time, producing a near-zero sum and indicating no relationship between the variables.

The Covariance in R

```r
sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
  ) /
(nrow(wagner) - 1)
```

```
## [1] 1.336434
```

```r
cov(wagner$Stress, wagner$lnSymptoms)
```

```
## [1] 1.336434
```

The Pearson Correlation Coefficient

$$r = \frac{cov_{XY}}{s_X s_Y}$$

Because the maximum value of cov_{XY} can be shown to be $s_X s_Y$, it follows that the limits on r are ± 1.00. One interpretation of r, then, is that it is a measure of the degree to which the covariance approaches its maximum.

The Pearson Correlation Coefficient in R

```
(sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
  ) /
 (nrow(wagner) - 1)
) /
(sd(wagner$Stress) * sd(wagner$lnSymptoms))
```

```
## [1] 0.5286565
```

```
cor(wagner$Stress, wagner$lnSymptoms)
```

```
## [1] 0.5286565
```

Intepretation of r

Rules for interpretation of r ale domain specific.

According to Funder and Ozer (2019) (see `effectsize::interpret_r` for more details and different sets of rules):

- $r < 0.05$ - Tiny
- $0.05 \leq r < 0.1$ - Very small
- $0.1 \leq r < 0.2$ - Small
- $0.2 \leq r < 0.3$ - Medium
- $0.3 \leq r < 0.4$ - Large
- $r \geq 0.4$ - Very large

Guess the correlation

http://guessthecorrelation.com/ - an useful tool for developing intuition about correlation

Statistical test for correlation

The most common hypothesis that we test for a sample correlation is that the correlation between X and Y in the population, denoted ρ (rho), is zero. This is a meaningful test because the null hypothesis being tested is really the hypothesis that X and Y are linearly independent. Rejection of this hypothesis leads to the conclusion they are not independent and there is some linear relationship between them.

It can be shown that when $\rho = 0$, for large N, r will be approximately normally distributed around zero. (It is not normally distributed around its mean when $\rho \neq 0$) A legitimate t test can be formed from the ratio

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

which is distributed as t on $N - 2$ df.

Statistical test for correlation in R

```r
covariance <- sum(
  (wagner$Stress - mean(wagner$Stress)) *
  (wagner$lnSymptoms - mean(wagner$lnSymptoms))
                 ) / (nrow(wagner) - 1)
covariance
## [1] 1.336434
correlation <- covariance /
               (sd(wagner$Stress) * sd(wagner$lnSymptoms))
correlation
## [1] 0.5286565
t_stat <- correlation * sqrt(nrow(wagner) - 2) /
          sqrt(1 - correlation**2)
t_stat
## [1] 6.381819
pval <- pt(t_stat, nrow(wagner) - 2, lower.tail = F)
pval
## [1] 2.413743e-09
```

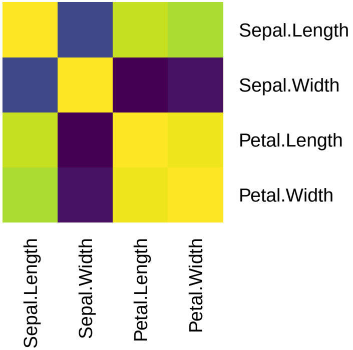
Statistical test for correlation in R

```r
cor.test(~ Stress + lnSymptoms, data = wagner)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Stress and lnSymptoms
## t = 6.3818, df = 105, p-value = 4.827e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3765970 0.6529758
## sample estimates:
##       cor
## 0.5286565
```

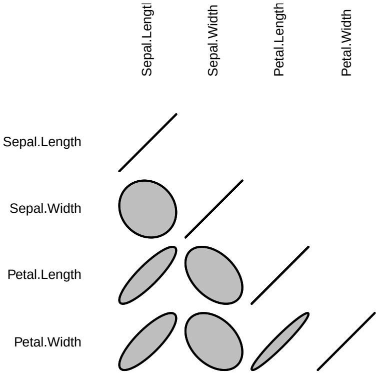
Visualization of correlations in multivariate datasets - base R

```r
library(viridis)
correlations <- cor(iris[, 1:4])
heatmap(correlations, symm = T, Rowv = NA, col = viridis(256),
        cexRow = 0.7, cexCol = 0.7) # graphical parameters, ignore
```



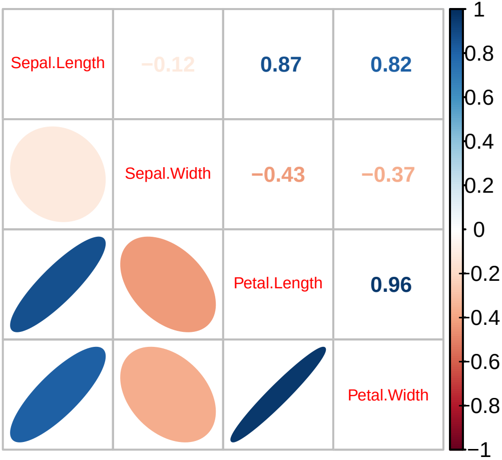
Visualization of correlations in multivariate datasets - ellipse

```
library(ellipse)
plotcorr(correlations, type = "lower", diag = T,
         cex.lab = 0.4, cex = 0.2, mar = c(4,4,4,4)) # ignore
```



Visualization of correlations in multivariate datasets - corrplot

```r
library(corrplot)
corrplot.mixed(correlations, lower = "ellipse", upper = "number",
               cl.cex = 0.7, number.cex = 0.7, tl.cex = 0.5) # ignore
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width