

## 5 Bivariate data. Double the data, double the fun

We have so far considered mostly samples of a single measurement made on some objects. We'll now look at the *bivariate* or two-variable case. There are several sorts of questions that could be asked in various examples:

- A sample of people are surveyed to find their annual income and the number of years that they spent in school. Are these measurements related?
- The body mass indicator or BMI is computed by taking an individual's weight in *Kg* and dividing by the square of their height in *m*. Does the implied relationship, that weight is a linear function of the squared height, hold in a sample of people?
- A sample of individuals are categorized according to their genotypes at an assayed genetic locus, *AA*, *AB* or *BB*, and by their disease status *affected* or *unaffected*. Are these categorizations independent?
- Two wine tasters rank 12 bottles of wine in order of their judged quality. Do the two sets of ranks agree?

We'll see methods for analyzing data of all these types. However, the first step in looking at bivariate data is to plot it in some way. The *scatter plot* (see week1) is a good way when there are a spread of values. However, for categorical data this often just piles point invisibly on top of each other so a simple table is better.

Be aware also that the same data can be analyzed in different ways. For example measurement data, like height and weight above, can be categorized, for instance into light/heavy and tall/intermediate/short. Typically such categorizations will give more robust but less powerful analyses.

### 5.1 Covariance and correlation

The *population covariance* between two variables  $X$  and  $Y$  is defined as

$$\mathbf{C}(X, Y) = \mathbf{E} [(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))]$$

which can be estimated from a bivariate sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  using the *sample covariance*

$$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Note that the population and sample covariances of a variable  $X$  with itself are the population and sample variances of  $X$ .

The value of the covariance depends on the units that we use to measure  $X$  and  $Y$ . So, for instance, it then matters whether we measured height and weight in inches and pounds or metres and Kilograms. To avoid this and produce a unit free measure of co-variation we define the *correlation coefficient*

$$\rho_{x,y} = \frac{\mathbf{C}(X, Y)}{\sqrt{\mathbf{V}(X)\mathbf{V}(Y)}}$$

To estimate this with the sample correlation coefficient we compute

$$S_{x,x} = \sum_{i=1}^n (X_i - \bar{X})^2 \quad S_{x,y} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad S_{y,y} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

and hence

$$R = \frac{S_{x,y}}{\sqrt{S_{x,x}S_{y,y}}}$$

$R$  is the *sample correlation coefficient*.  $R^2$  is often reported as a measure of the strength of the correlation. Note that  $S_{x,y}$  like  $\rho$  and  $R$  can be positive or negative.  $\rho$  and  $R$  must lie between -1 and 1.

- $\rho$  or  $R$  close to 1 indicates an increasing linear relationship between the variables
- $\rho$  or  $R$  close to -1 indicates a decreasing linear relationship between the variables
- $\rho$  or  $R$  close to 0 indicates no linear relationship between the variables.

Note the careful use of *linear* above. A  $\rho$  of zero means that there is no simple linear relationship between the variables *not* that the variables are independent. Even when  $\rho$  is 0 there may be strong non linear relationships. An exception to this is when data is from a *bivariate Normal* distribution: in this case *uncorrelated* and *independent* are equivalent conditions.

See the R functions `cov` and `cor`.

## 5.2 Linear regression

### The method of least squares

We can estimate a linear relationship between  $X$  and  $Y$ . It may be the case that both  $X$  and  $Y$  are randomly observed, as in the heights and weights example above, but it can also be the case that the  $X$  values are chosen in a design and the  $Y$ s are random. In this latter case when we are trying to predict the  $Y$  values from the  $X$  values  $X$  is called the *explanatory* or *independent variable* and  $Y$  is the *dependent* or *response* variable. When both variables are random it makes sense to estimate  $a$  and  $b$  for either or both of

$$Y = a + bX \quad \text{and} \quad X = a + bY$$

However, in the designed case it will only make sense to regress the dependent variable on the explanatory: that is estimate  $a$  and  $b$  only for  $Y = a + bX$ . Note that in the random case the line that we estimate will depend on the order of regression.

The method that we use is called *least squares regression*. In order to estimate  $a$  and  $b$  in  $Y = a + bX$  we compare the observed values  $Y_i$  with the predicted values  $a + bX_i$ , and minimize the summed squared difference. That is: choose  $a$  and  $b$  to minimize

$$S = \sum_{i=1}^n [Y_i - (a + bX_i)]^2$$

By solving  $\frac{dS}{da} = 0$ ,  $\frac{dS}{db} = 0$  it is easy to show that  $S$  is minimized by

$$\begin{aligned}\hat{b} &= \frac{S_{x,y}}{S_{x,x}} \\ \hat{a} &= \bar{Y} - \hat{b}\bar{X}\end{aligned}$$

Note the use of  $a$  and  $b$  for the parameters and  $\hat{a}$  and  $\hat{b}$  for their estimates.

## The residuals

An important aspect of any estimation of the relationship between  $X$  and  $Y$  is to look at the *residuals*, that is the left-overs we have after we have estimated the relationship. There is a residual for each bivariate observation:

$$Y_i - (\hat{a} + \hat{b}X_i)$$

Implicit in the method of least square is the assumption that, if the relationship is true then **the residuals are an independent sample from a  $N(0, \sigma^2)$  distribution** for some fixed value of  $\sigma^2$ . So we should plot the residuals to see whether they look Normal. We should also plot the residuals against  $X$  to see whether there seems to be any trend left in the data.

We can estimate  $\sigma^2$ , the *residual variance* using

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - (\hat{a} + \hat{b}X_i)]^2 \quad (\text{Why } n-2 \text{ ?})$$

The variances of our estimators all depend on  $\sigma^2$ .

- $\mathbf{V}(\hat{a}) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{X}^2}{S_{x,x}} \right]$   
so that

$$\frac{a - \hat{a}}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{x,x}}}} \sim t_{n-2}$$

from which we get all tests and confidence intervals for  $a$ .

- $\mathbf{V}(\hat{b}) = \sigma^2 \frac{1}{S_{x,x}}$   
so that

$$\frac{b - \hat{b}}{\hat{\sigma} \sqrt{\frac{1}{S_{x,x}}}} \sim t_{n-2}$$

Testing whether  $b = 0$  is equivalent to testing whether there is a linear relationship dependence of  $Y$  on  $X$ .

- $\mathbf{V}(\hat{y}) = \mathbf{V}(\hat{a} + \hat{b}x) = \sigma^2 \left[ \frac{1}{n} + \frac{(x-\bar{X})^2}{S_{x,x}} \right]$   
so that

$$\frac{y - \hat{y}}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\bar{X})^2}{S_{x,x}}}} \sim t_{n-2}$$

Also notice that in general  $\hat{a}$  and  $\hat{b}$  are not independent. Their covariance is given by

$$\mathbf{C}(\hat{a}, \hat{b}) = -\sigma^2 \frac{\bar{X}}{S_{x,x}}$$

Finally, notice that we can write the residual sum of squares as

$$\sum_{i=1}^n [Y_i - (\hat{a} + \hat{b}X_i)]^2 = S_{y,y}(1 - R^2)$$

which explains why  $R^2$  is a good measure of the relatedness of  $X$  and  $Y$ , or a measure of *what proportion of the variation in  $Y$   $X$  explains*.

## Computation

The R function `lsfit` computes the estimates  $\hat{a}$  and  $\hat{b}$  and also the residuals. These are returned in a structure with elements `coefficients` and `residuals`. You should also write a function to compute sums of squares. For example:

```
function(x,y=x)
{
    sum ( (x - mean(x)) * (y - mean(y)) )
}
```

so that  $S_{x,y} = \text{sumsq}(x,y)$  and  $S_{x,x} = \text{sumsq}(x)$  and so on. The rest is up to you!

## Data transformations

We have looked at a simple linear relationship  $Y = a + bX$ . However, modeling  $Y = a + bX^2$  is also linear regression. All we have to do is make the transformation  $Z = X^2$  and estimate the linear relationship  $Y = a + bZ$  as before. Thus linear regression is far more flexible than it first appears. Choosing data transformations takes some guesswork and experience, but a good place to start is by plotting  $X$  by  $Y$ ,  $X$  by  $\log Y$ ,  $\log X$  by  $Y$  and/or  $\log X$  by  $\log Y$ .

## Prediction

Note that above when we found  $\mathbf{V}\hat{y}$  we were deriving the variance of the mean of  $Y$  as a function of  $X$ . If we were asked to predict the value of  $Y$  for any given value of  $x$  then we would have to account for the variation of  $Y$  about this mean.

In this case we would use the same estimate of

$$\hat{Y} = \hat{a} + \hat{b}x$$

but the variance is now

$$\mathbf{V}(\hat{Y}) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{S_{x,x}} \right]$$

Thus the variance of this predictive estimate of  $Y$  must not only account for the uncertainty that we have about the mean of the  $Y$ s but also the inherent uncertainty in the  $Y$ s. For tests and confidence intervals for predictions use

$$\frac{Y(x) - \hat{Y}(x)}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{S_{x,x}}}} \sim t_{n-2}$$

Note that as  $n \rightarrow \infty$  the variance of the estimate of the **fit** goes to zero, but the variance of the estimate of the **prediction** goes to  $\sigma^2$ . Prediction is always risky, even with good data.

### 5.3 Contingency tables

If we categorize objects according to two different criteria we can summarize the data in a simple two dimensional table. This is called a *contingency table*. Higher dimensional contingency tables are also possible. For instance, some specific data for the genetic example mentioned in the introduction might be:

		Genotype		
		AA	AB	BB
Disease status	Affected	7	24	15
	Unaffected	8	16	9

The hypothesis we which to test is

$H_0$  : The categorizations are independent  $v$   $H_1$  : there is some dependence

To do this we note that if the categories are independent then

$$P(\text{an item is in the } (i, j)\text{th box}) = P(\text{it is in the } i\text{th row}) \times P(\text{it is in the } j\text{th column})$$

Then we can compare the observed counts in each box with those that we would predict when this condition is true. Note that the best estimate of  $P(\text{it is in the } i\text{th row})$  is simply the proportion of items in that row, and similarly for the columns. So to find the counts expected under  $H_0$  we first find the row, column and grand totals.

7	24	15	46
8	16	9	33
15	40	24	79

Then the expected count in the first box would be

$$\frac{46}{79} \times \frac{15}{79} \times 79 = 8.73$$

and so on to give the following table of expected values:

8.73	23.29	13.97
6.26	16.71	10.03

So it remains to find some statistics that measures the difference between the observed table and the expected table, and to find the distribution of this statistics when  $H_0$  is true. It turns out that this is a case where the  $\chi^2$  *goodness of fit statistic* is appropriate. This is generally a good statistic to compare observed and expected values under various hypotheses and takes the form

$$\sum_{\text{categories}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The distribution of this statistics is approximately  $\chi^2$  for a large enough set of observations. For a contingency table with  $r$  rows and  $c$  columns the degrees of freedom is  $(r - 1)(c - 1)$ .

Although only the total is needed for a test, it is good practice to look at the individual  $(\text{observed} - \text{expected})^2/\text{expected}$  terms and a table of signs showing which get bigger and which get smaller.

0.343	0.022	0.076
0.483	0.030	0.106

-	+	+
+	-	-

For this data the observed test statistic is 1.06. For this test we particularly look for upper tail probabilities to indicate values that deviate from the null hypothesis. In this case we have that

```
pchisq(1.06, 2, lower.tail=FALSE)
```

equals 0.588 so that we cannot reject  $H_0$ . When we can reject  $H_0$  these last two tables give us an indication of what alternatives can be conjectured. For instance, although the above data suggests independence of phenotype and genotype, the pattern of +s and -s is what we would expect from a dominantly expressed disease.

There is occasionally reason to check for values in the lower tail. This would indicate that that *the data fits the model too well*. For example, if we suspect that a researcher has cheated by inventing data to correspond with their theory instead of generating real data it might be revealed by this sort of test. It is generally thought that Gregor Mendel's published results for his pea crossing experiments are too good to be real, and that he probably at least censored some unfavourable data. Since he was brilliant and right and was working before statistics had been invented I think he has been forgiven.

The  $\chi^2_{(r-1)(c-1)}$  distribution is an approximation that works for reasonably large numbers of observation. This approximation is also unreliable if the expected value for one or more cells is small. A rule of thumb is that cells should have an expected value of 4 or more. For small numbers we can avoid using this approximation because complete enumeration of all possible tables giving more extreme test statistics is possible, in which case this is called *Fisher's exact test*. See `fisher.test()` in R. We will also consider simulation as a possibility: see the work sheet for this week.

## 5.4 Agreement in ranks

In some cases we will not have bivariate measurements but two rankings of objects that we wish to compare for concordance. There are also situations where we have numerical values but they are arbitrary or not comparable and so we might be better off disregarding the values and keeping only the order that they imply. This is often the case when subjective judgments are being compared. For example, two wine tasters may allocate scores on a scale of 1 to 100 for a set of wines. The tasters are unlikely to give the exactly the same meaning to the same score: is my idea of a 70 the same as yours? Nor are they likely to use the scale in a consistent way: is the difference between an 80 and a 90 on my scale the same as the difference between a 10 and a 20 on my scale? However, although the numerical values may not be usable the order of preference is probably reliable.

If we now let  $(X_i, Y_i) \dots (X_n, Y_n)$  be the pair of ranks we can calculate the sample correlation covariance  $R$  as before and note that, since

$$\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i = \frac{n(n+1)}{2}$$

it simplifies somewhat to give

$$R = 1 - \frac{6 \sum_{i=1}^n (X_i - Y_i)^2}{n(n^2 - 1)} = r_s$$

which is also known as *Spearman's rank correlation coefficient*. Note that simply using  $\text{cor}(x, y)$  when  $x$  and  $y$  are ranks gives us this result so we don't need to compute it this way.

To assess whether the data are correlated we test the hypothesis

$$H_0 : \rho_{x,y} = 0 \quad v \quad H_1 : \rho_{x,y} \neq 0$$

We will clearly reject  $H_0$  if  $|r_s|$  is big and as in last week's work we can use a Normal, large sample, approximation

$$r_s \sqrt{\frac{n-2}{1-r_s^2}} \sim t_{n-2}$$

implemented in the R function

```
function(rawx,rawy)
{
  x = rank(rawx)
  y = rank(rawy)
  n = length(x)
  rs = cor(x,y)

  t = rs * sqrt((n-2)/(1-rs^2))
  if (t > 0) t = -t
  pval = pt(t,n-2) + 1-pt(-t,n-2)

  list( stat = rs, expect = 0, pval = pval)
}
```

or use simulation as in this R function:

```
function(rawx,rawy,s=1000)
{
  x = rank(rawx)
  y = rank(rawy)
  t = cor(x,y)
  n = length(x)

  e = 0
  k = 0
  r = 1:n
  for (i in 1:s)
  {
    u = cor(r,sample(r))      # or cor(sample(r),sample(r))
    if ( abs (e - u) >= abs (e - t) )
      k = k+1
  }

  list( stat = t, expect = e, pval = k/s, nsims = s )
}
```

Note that whether based on raw data or just the ranks the sample correlation coefficient is an attempt to estimate the underlying population correlation. Hence for large data sets  $\text{cov}(x,y)$  and  $\text{cov}(\text{rank}(x),\text{rank}(y))$  should give similar estimates of  $\rho_{x,y}$ .

Also note that we have seen another example where replacing the data by their ranks gives a more robust analysis method.

## 5.5 Worksheet

- \* For the `heights1` and `weight` data from week 1:
  - Plot the data.
  - Find and draw the regression line of weight on height.
  - Find and draw the regression line of height on weight. Where do these two lines cross?
  - For the regression of  $y = \text{weight}$  against  $x = \text{height}$  find 95% confidence intervals for the intercept and slope.
  - For the same regression find 95% confidence bounds for the fit of  $y$  as a function of  $x$ . Draw these on the plot. What does the shape of these bounding lines tell you?
  - Use appropriate plots to verify that the residuals are Normally and independently distributed.
  - Based on this data give a 95% confidence interval for the predicted weight of an individual who is 168 cm tall, but who has not yet been weighed.
- Download the file called `runtimes` from the web page. The data here shows the times taken by two algorithms to solve a computing problem. The first column gives the size of the problem the next two columns give the time in seconds that was required by each method to solve the problem. Analyse the data so as to compare the methods.
- Download the file called `cholest` from the web page. The data here result from a small study to test whether overnight fasting affects total cholesterol readings. Twenty individuals were randomly allocated into fasting and non-fasting groups. On the morning of the test 20 blood samples were collected. The researcher then assayed the cholesterol with a kit provided. They concluded from the data given here that fasting reduced the cholesterol readings. Why would they conclude that? Do you agree?
- Read the help page for the `chisq.test` function in R. How do you obtain the expected table, the test statistic, degrees of freedom, p-value and individual contribution to the test statistic using this function?
- Simulate 1000 random 2 dimensional contingency tables under the null hypothesis that there is no association. Use `chisq.test` to show that the test statistic really does have a  $\chi^2$  distribution. Or, equivalently, show that the p-values have a Uniform distribution under  $H_0$ . Arrange with other people to try different numbers of observations so that we see results for a broad range of situations.
- \* Two researchers produced tables of data to assess whether a genetic locus affects the ability of a person to metabolize a new drug. Their results are shown below. What would you conclude from the data?

Researcher 1				Researcher 2			
Genotype				Genotype			
	AA	AB	BB		AA	AB	BB
Metabolizer	17	36	18	Metabolizer	12	31	21
Non-metabolizer	7	16	8	Non-metabolizer	12	21	3

7. \* Bryce and I have both ranked the 12 students in last year's class in order from the most to least pleasant person. Our ranks are given in the table below. The first line of the table gives the randomly assigned identifier of each person. Is there any evidence that Bryce and I agreed?

Person	1	2	3	4	5	6	7	8	9	10	11	12
Bryce	8	10	7	12	11	4	2	3	9	6	1	5
Alun	1	2	7	4	6	9	5	10	8	3	12	11

8. The  $\chi^2$  test for contingency tables is an approximation. How would you use simulation to assess the p-value for small numbers of data?