Measuring the association between 2 variables

Lecture 7

8th August, 2007
So far ....

- We have looked at how we can plot 2 or more variables to see if a relationship MAY exist.

- We have used investigative plotting – comparison boxplots and scatterplots.
Can we quantify this possible relationship?

**Least squares regression** is a fundamental tool of statistics.

This is finding the equation of a "**straight line of best fit**"

This describes a linear relationship between 2 quantitative variables.
We can use the equation of the line to make predictions about Y.

It is a mathematical expression of the X-Y relationship over the range of X.

Extrapolations outside this range should only be made with extreme care – if at all.
- The relationship may not be linear outside the given range.

3. A good point is not to make extrapolations at all! Before showing 4. Why?
Example 7.1: CO$_2$ concentration data
Figure 7.1 Straight line fit to CO$_2$ uptake data

- Recall this from Lecture 3.
- Import data set C02uptake.txt.
- Plot a scatterplot.
- Add a least squares line.
The fitted line

- The general equation of a straight line is?
  
  \[ y = a + bx \]
  
  Or \[ y = mx + b \]
  
  Or \[ y = \beta_0 + \beta_1x \]

- Here \( y \) is the numbers in Uptake whilst \( x \) is replaced by Conc.

- The symbols \( \beta_0, \beta_1 \) are the intercept and slope of the linear relationship.

Ask as you go for all of this.

All these equations are the same. Different books use different forms of the equation. Be familiar with all of them. We will tend to use the last one with the Beta’s
For every data pair \((x_i, y_i)\) there is a fitted value which we denote by \(\hat{y}_i\).

Where \(\hat{y}_i = a + b \cdot x_i\).

The line of least squares fit is the line with intercept \(a\) and slope \(b\) such that
\[
\sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2
\]
is minimised. This is called the Sum of Squares of Residuals.

We name the \(\hat{y}_i\) a hat and say ‘y hat’.

Excuse my ‘funny’ hats. They are very hard to draw correctly in PowerPoint.

You will not be asked to calculate this SSR value by hand. We will get it later from Rcmdr output.
In practice

- We need to choose our \( a \) and \( b \) so that this sum is minimal.
  - This means we want the best line so the average distance of the points away from the line is the smallest.
  - We will see how we get these values in a later lecture.
  - In this example the equation of the line of best fit is \( \hat{y} = -2.04 + 0.025x \)
    Or Uptake = -2.04 + 0.025 concentration

Work through what each number is telling you: 

- \( a \) = intercept. Does it have a reasonable meaning here?
- \( b \) = slope: for every unit increase in concentration the rate of Uptake increases by 0.025 units.

How did we know the Uptake increases? The slope is positive. If the slope is negative then there would be a decrease in \( Y \) (Uptake)
Now we have the equation of this fitted line we can generalise about the experiment and do not need to refer to the data, just the estimated response.
Is a linear relationship always the best explanation?

- Example 7.2 USA car data package built into R - use
  Data < Data in packages < Read data set from..

Select **dataset** as the package and then scroll down to find **mtcars** to call it up.

- Fuel consumptions (mpg) for cars of different weights (wt) and engine Displacement (disp).
- More information about the makes included.

The dataset package has files starting with a capital letter 1st then the lowercase named files so scroll for a bit.

Some of the fields of information are:
<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cyl</th>
<th>disp</th>
<th>hp</th>
<th>drat</th>
<th>wt</th>
<th>qsec</th>
<th>vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21</td>
<td>6</td>
<td>160</td>
<td>110</td>
<td>3.9</td>
<td>2.62</td>
<td>16.46</td>
<td>0</td>
</tr>
<tr>
<td>Mazda RX4 Wag</td>
<td>21</td>
<td>6</td>
<td>160</td>
<td>110</td>
<td>3.9</td>
<td>2.875</td>
<td>17.02</td>
<td>0</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>108</td>
<td>93</td>
<td>3.85</td>
<td>2.32</td>
<td>18.61</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>6</td>
<td>258</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360</td>
<td>175</td>
<td>3.15</td>
<td>3.44</td>
<td>17.02</td>
<td>0</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225</td>
<td>105</td>
<td>2.76</td>
<td>3.46</td>
<td>20.22</td>
<td>1</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>8</td>
<td>360</td>
<td>245</td>
<td>3.21</td>
<td>3.57</td>
<td>15.84</td>
<td>0</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>4</td>
<td>146.7</td>
<td>62</td>
<td>3.69</td>
<td>3.19</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>
We want to examine the relationship between mpg and wt.

Draw a scatterplot with a least squares line fitted
Figure 7.2 Adding a least squares line to mtcars data: plot of linear model for mpg ~ wt

Does this straight line look like it is a good fit to the data?

Note that I have Y ~ X in the title and that is how R puts the order when using a ~ (tilda).
What is this smooth line indicating?
This line looks as though it fits the data better than the straight line. This infers that the relationship may be curvilinear rather than linear.

At the moment which is the best model is being done by inspection. There are analytical methods for comparing the models but that is for later.
Tables: Qualitative variables

- Some variables are obviously categorical
  - Gender, degree programs, occupation

- Others are formed by grouping values of quantitative variables into classes
  - Age classes – 25-34, 35-44 etc
- We use counts (=frequencies) or percents (=relative frequencies) of individuals that fall into the different categories to compare the distribution of variables.

- Each individual will fall into each category.

- We present the information in a 2-way Contingency table.

- When row or column totals differ we need to use percentages to allow for a useful comparison.

2. Also called a frequency table.
Example 7.3 Titanic data set from R

- The 4 columns of this data set are
  - PClass => the cabin category (1st, 2nd or 3rd)
  - Age
  - Gender
  - Survived => 1 = yes; 0 = no

- Which variables are categorical?

Even though this data set is in R it can't be loaded into Rcmdr so we have reset it and you will find it in the STAT100 web page Data sets used in lectures.

1. View data set
   PClass = passenger class
2. PClass, Gender and Survived
Table 7.2 Some of the Titanic data

<table>
<thead>
<tr>
<th>PClass</th>
<th>Age</th>
<th>Sex</th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>29</td>
<td>female</td>
<td>1</td>
</tr>
<tr>
<td>1st</td>
<td>2</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>1st</td>
<td>30</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>1st</td>
<td>25</td>
<td>female</td>
<td>0</td>
</tr>
<tr>
<td>1st</td>
<td>1</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>1st</td>
<td>47</td>
<td>male</td>
<td>1</td>
</tr>
<tr>
<td>1st</td>
<td>63</td>
<td>female</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>30</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>28</td>
<td>female</td>
<td>1</td>
</tr>
<tr>
<td>2nd</td>
<td>18</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>34</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>32</td>
<td>female</td>
<td>1</td>
</tr>
<tr>
<td>2nd</td>
<td>57</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>18</td>
<td>male</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>23</td>
<td>male</td>
<td>0</td>
</tr>
</tbody>
</table>

There is a lot of data as it covers every passenger that was on the Titanic.
Import the data set from the web page and as Survived values are presently numeric we need to convert this variable to a factor.

Data < Manage variables < Convert numeric ..

Use numbers for the factor levels.
Two-way tables

- Use **Statistics < Contingency tables < Two-way tables**
  
  - Use PClass as the row variable and Survived as the Column variable
  
  - Check **No percentages** and deselect Chi-square test.

We will get to this test later in the unit.
You notice that each person could be classified in 2 ways – what passenger class they were in and whether they survived.

### Table 7.3 Survival frequencies for Passenger classes

<table>
<thead>
<tr>
<th>PClass</th>
<th>Survived 0</th>
<th>Survived 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>87</td>
<td>139</td>
</tr>
<tr>
<td>2nd</td>
<td>116</td>
<td>96</td>
</tr>
<tr>
<td>3rd</td>
<td>240</td>
<td>78</td>
</tr>
</tbody>
</table>

This table indicates that the majority of 1st class passengers survived while ~ ¾ of the 3rd class passengers perished.

We may infer (tentatively at this stage) that the frequency of Survived is related to PClass.
Because there were different numbers of people in each of the passenger classes it would be more convenient to convert the frequencies to percentages or relative frequencies.

In Rcmdr recall the contingency table but check the row percentages.
### Table 7.4 Relative frequencies of Survival by PClass

<table>
<thead>
<tr>
<th>Survived</th>
<th>PClass</th>
<th>0</th>
<th>1</th>
<th>Total Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
<td>38.5</td>
<td>61.5</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>54.7</td>
<td>45.3</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>75.5</td>
<td>24.5</td>
<td>100</td>
</tr>
</tbody>
</table>

You can check you have done this correctly because the rows total 100.
Interpreting this then: Of the 1\textsuperscript{st} class passengers \(\sim62\%\) survived compared to \(\sim25\%\) only of the 3\textsuperscript{rd} class passengers.

You can repeat this using column percentages and interpret it in words..
Extension to more than 2 categories

- There are actually 3 categories in this data set.
- Include Sex

**Statistics < Contingency Tables < Multi-way**

Choose PClass as rows, Survived as Columns and Sex as control variable.

3. The control variable groupings are also called layers.
### Table 7.5 PClass and Survival by Sex

, , Sex = female , , Sex = male

<table>
<thead>
<tr>
<th></th>
<th>Survived</th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClass</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1st</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td>2nd</td>
<td>10</td>
<td>75</td>
</tr>
<tr>
<td>3rd</td>
<td>56</td>
<td>46</td>
</tr>
</tbody>
</table>

Note the 2 commas before the sex? The 1st one represents the row variable (PClass) and the 2nd one the column variable (survived)
We can combine these into a **flat table**

Edit the line

```
.Table <- xtabs(~PClass+Survived+Sex, data=titanic)
```

To read

```
.Table <- ftable(xtabs(~PClass+Survived+Sex, data=titanic))
```

Enter under it **.Table**

Highlight both lines and submit.

The full stop in front of Table is important!

Or add the `ftable` command to the present `.Table` line with `xtabs`, then highlight both lines and submit.
### Table 7.6 PClass and Survival by Sex

<table>
<thead>
<tr>
<th>Sex</th>
<th>female</th>
<th>male</th>
</tr>
</thead>
<tbody>
<tr>
<td>PClass</td>
<td>Survived</td>
<td>Survived</td>
</tr>
<tr>
<td>1st</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>96</td>
</tr>
<tr>
<td>2nd</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>75</td>
</tr>
<tr>
<td>3rd</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>46</td>
</tr>
</tbody>
</table>

Lynette McLean 28
Example 7.4 University admissions

- Data set from R => UCBAAdmissions
  - 3-way array of frequencies of admissions and rejections for Males and Females in 6 departments at the University of California, Berkley, USA.
  - The dimensions are in

<table>
<thead>
<tr>
<th>Table 7.7 Dimensions for UCBAAdmissions data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>dimension</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>
We want to know the percentage of males admitted in each department.
First you will need to load the package “abind” from Packages in R.
Type these lines into Rcmdr and submit all at once.

```
data(UCBAdmissions)
UCBAdmissions
Madmissions <- UCBAdmissions[,1,]
colPercents(Madmissions)
```

The data line calls up the data set.

The 2nd line UCBAdmissions calls up all the data with the 3\textsuperscript{rd} dimension Dept as the control variable.

The 3\textsuperscript{rd} line creates an internal file for the male data only. Inside the square brackets before the 1\textsuperscript{st} comma is nothing indicating we want all the information on the 1\textsuperscript{st} dimension, Admit. The 1 in the 2\textsuperscript{nd} space means we only want the information on the 1\textsuperscript{st} level inside that dimension Gender. The empty value after the 2\textsuperscript{nd} comma means we want all of the departments.

If we had wanted the female data we would have used [,2,].

Rcmdr hint: If you get an error message “could not find function "abind" “ then go back to the R GUI and Load Package ‘abind’. Then resubmit your code. What has happened is that Rcmdr couldn’t combine the column percentages into a table.
Of course we could have used Tinn-R to save this code. If you do then you need to use the following as the last line.

\textbf{colPercents(UCBAAdmissions)}

To include a save information command in your .r code file use `sink("ucba-out.txt")` at the beginning and `sink()` at the end. (See next slide)

To have Rcmdr or R read this file in the script window use `source("ucba.r")`
Code source file for UCBAdmissions example for use in R

```r
#---ucba.r
sink("ucba-out.txt")
data(UCBAdmissions)
UCBAdmissions
Madmissions <- UCBAdmissions[,1,]
print(colPercents(Madmissions))
sink()
```

Hash lines – those starting with # are comment lines for you and R ignores them.
Code source file for UCBAdmissions example for use in Rcmdr

```r
#---ucba.r
# 1st load abind from R
sink("ucba-out.txt")
data(UCBAdmissions)
print(UCBAdmissions)
print(UCBAdmissions)
Madmissions <- UCBAdmissions[,1,]
colPercents(Madmissions)
sink()
```
If we were using this table in a report should we modify it?

How?

From class:
Round numbers to 2 sig. no.s – here 62, 63 etc
Rearrange columns – swap A for B
If know names of depts use them.
Leave out rejected row – superfluous.
Could improve by adding female acceptances