MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

4. Linear Regression

4.1 Introduction

So far our data have consisted of observations on a single variable of interest. We now look at what happens when we have additional information associated with each observation. For simplicity we will assume that this information consists of the value of a single real-valued variable, but the approach can easily be extended.

It is customary to use Y to denote the original variable of interest and to use x to denote the associated variable. In this notation, the data now consist of a set of n pairs of values $(x_1, y_1), \ldots, (x_n, y_n)$, corresponding to the n members of our sample, where y_i is the observed value of the original variable and x_i is the value of the associated variable for the ith member of the sample.

We are interested in whether there is a pattern to the relationship between the two variables which can be used to explain or predict values of the variable of interest in terms of the values of the associated variable. For example, we might have data on the heights and weights of a sample of students and be interested in how well height can be used to predict weight. Alternatively we might have data on the level of debt and the annual parental income for a sample of students, and may wish to investigate whether there was any dependence between the two variable, and, if so, what form the dependence took.

Note that the two variables play different roles, in that the original variable often *depends on* the associated variable. For example, changes in weight do not usually cause changes in height, but a change in height (through growth) is usually associated with an increase in weight. For that reason the variable of interest (our Y variable) is called the **response variable** (an old-fashioned term is the **dependent variable**) while the associated variable (our x variable) is known as the **predictor variable** or the **explanatory variable** (or the **independent variable**; again, best to avoid this term).

We also need to take account of the random variation in the relationship between the x and Y values. For example, if we took repeated samples, then even if the x values were kept the same the y values obtained would usually vary from sample to sample. Thus an appropriate framework is to assume that for each value x of the explanatory variable there is a corresponding *population* of values of Y with its own x-dependent distribution, and we call the function $g(x)$ given by $g(x) = \mathrm{E}(Y|x)$ the **regression** of Y on x. In this framework, our search for a simple functional explanation of the dependence of the (mean of the) Y variable on the x variable becomes a search for a simple expression for $\mathrm{E}(Y|x)$ which is valid over an appropriate range of x values.

The **simple linear regression model** says that relationship of $\mathrm{E}(Y|x)$ to x is of the form

$$\mathrm{E}(Y|x) = \alpha + \beta x$$

For this model, the basic questions of interest are:

- What are good estimates of the unknown parameters α and β (assuming the model is correct)?
- How well do the data fit the model and is there any evidence from the data that the model is not correct (i.e. systematic deviation from what we would expect if the model was correct)?
- What evidence is there that Y really does depend on x (i.e. that $\beta \neq 0$)?

4.6 Example – Leaning Tower of Pisa

Studies by engineers on the Leaning Tower of Pisa between 1975 and 1987 recorded the following data on the increasing tilt of the tower. Each tilt value in the table represents the difference between where a point on the tower would have been if the tower were straight and where it actually was in the corresponding year. The data are coded in tenths of a millimetre in excess of 2.9 metres, so the 1975 tilt of 642 represents an actual difference of 2.9642 metres. Only the last two digits of the year are shown. The data are contained in the Statistics 1 data frame `pisa`; the variables are called `year` and `tilt` respectively.

Year (x_i)	75	76	77	78	79	80	81	82	83	84	85	86	87
Tilt (y_i)	642	644	656	667	673	688	696	698	713	717	725	742	757
Fitted	637.78	647.10	656.42	665.74	675.05	684.37	693.69	703.01	712.33	721.65	730.97	740.29	749.60
Residuals	4.22	-3.10	-0.42	1.26	-2.05	3.63	2.31	-5.01	0.67	-4.65	-5.97	1.71	7.40

The summary statistics for the data set are:

$$n = 13 \quad \sum x_i = 1053 \quad \sum y_i = 9018 \quad \sum x_i^2 = 85475 \quad \sum y_i^2 = 6271714 \quad \sum y_i x_i = 732154$$

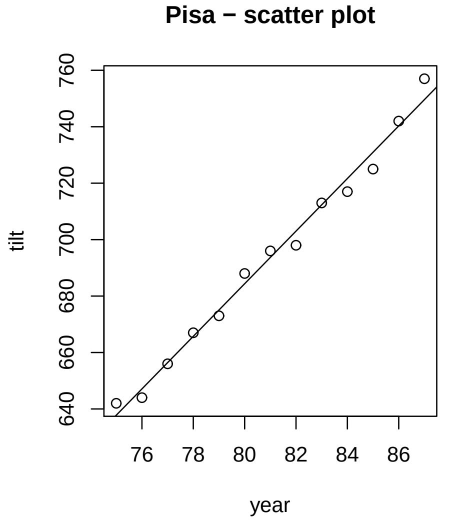
giving $\bar{x} = 81$, $\bar{y} = 693.6923$, $ss_{xx} = 182$, $ss_{xy} = 1696$ and $ss_{yy} = 15996.77$.
Thus the least squares estimates are

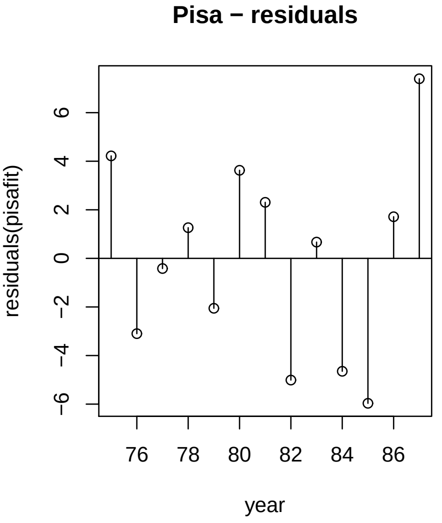
$$\hat{\beta} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{ss_{xy}}{ss_{xx}} = 9.3187 \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -61.1209$$

giving the fitted regression line $\quad y = \hat{\alpha} + \hat{\beta}x = -61.1209 + 9.3187x$
From this the fitted values and the residuals in the table can be calculated, using the formulae

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \qquad \hat{e}_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n.$$

A scatter plot of the data is shown on the left below, together with the fitted regression line. There seems to be quite a good fit of the straight line to the data. A plot of the residuals against the corresponding year is shown on the right. They appear to be fairly random, with equal numbers of $+$'ve and $-$'ve values and no obvious systematic pattern or systematic trend in variability.





4.7 Analysing linear regression models in R

R has a simple command `lm` for analysing linear regression models. This command produces an **R** object, containing a variety of numerical outputs which can be accessed using appropriate commands such as `coef`, `fitted`, `residuals`, `plot` and `summary`.

Assume the predictor (x) values are in a data array `xdata` and the response (y) values are in a data array `ydata`. We can perform an initial analysis with the commands:

```
> plot(xdata,ydata)
> xyoutput <- lm(ydata ~ xdata)
> coef(xyoutput)
```

The first line produces an initial scatter plot, the second line tells **R** to perform a linear regression with the response values in `ydata` and the predictor values in `xdata` and to store the output in the object `xyoutput`, and the third line produces a vector containing the least squares estimates of the intercept α and the slope β.

```
> plot(xdata,ydata); abline(coef(xyoutput))
```

will produce a scatter plot together with the fitted regression line – i.e. the line whose intercept a is the first value and whose slope b is the second value in the vector `coef(xyoutput)` .

```
> fitted(xyoutput)
> residuals(xyoutput)
```

will respectively output the vector of fitted values and the vector of residual values. Thus, for example, we can plot the residuals against the predictor values with the command:

```
> plot(xdata,residuals(xyoutput))
```

Towards the end of the course, we will look at other outputs such as `summary(xyoutput)`, which produces (among other things) estimates of σ^2 and of Var$(\hat{\alpha})$ and Var$(\hat{\beta})$.

For the Leaning Tower of Pisa example above, the predictor (year) values are in the variable `year` and the response (tilt) values in `tilt`, in the data frame `pisa`. I used the commands:

```
> attach(pisa); pisafit <- lm(tilt ~ year)
```

to perform the linear regression analysis and store the output in the object `pisafit`. I then inspected the scatter plot and the fitted line with the commands:

```
> plot(year,tilt);  abline(coef(pisafit))
```

and inspected the values of the least squares estimates with the command:

```
> coef(pisafit)
```

which gave output:

```
 (Intercept)       year
 -61.120879    9.318681
```

Finally I inspected the fitted values and the values of the residuals with the commands:

```
> fitted(pisafit)
> residuals(pisafit)
```

and plotted the residuals against the predictor (year) values with the command:

```
> plot(year, residuals(pisafit))
```

(for those who are interested, I used the `segments` command, specifically

```
> segments(year,0,,residuals(pisafit)); abline(h=0)
```

to add the extra lines – see `help(segments)`).

4.10 Normal linear regression

If we are prepared to assume a little more than in §4.2, we can make stronger statements about the least squares estimates.

It is sometimes reasonable to assume that the errors $\{e_i\}$ are normally distributed. We still have $\mathrm{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$, so the extra assumption is that $e_i \sim \mathrm{N}(0, \sigma^2)$, independently for $i = 1, 2, \ldots, n$.

This is equivalent to saying that $Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$, independently for $i = 1, 2, \ldots, n$.

Some of the consequences (not proved here) are:

- the least squares estimates $(\hat{\alpha}, \hat{\beta})$ are also the maximum likelihood estimates (so we have a second good reason to think they will be reasonable estimates)
- the estimates are themselves Normally distributed:

  $$\hat{\alpha} \sim \mathrm{N}(\alpha, \sigma^2[1/n + \overline{x}^2/ss_{xx}])$$

  $$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2/ss_{xx})$$

  (these facts will be useful near the end of the course, when we discuss hypothesis testing and confidence intervals in linear regression)

Note that we cannot check the assumption that $Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$ from the data by simply making a histogram, stem-and-leaf plot, or QQ plot of the data $y_1, y_2, \ldots, y_n$, since all the observations have *different* normal distributions. But we can carry out a check after the linear regression has been fitted, by looking at the residuals. Continuing the example in §4.7, typing

```
> qqnorm(residuals(xyoutput))
```

shows a Normal Q-Q plot of the residuals and helps check for non-Normality.

4.11 Visual assessment of the quality of fit of a linear regression model to data

One way of assessing the fit of a model is by examining a plot of the residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ (plotted against the predictor values $x_1, x_2, \ldots, x_n$ or the fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$).

If the model in §4.2 is correct, then $e_1, e_2, \ldots, e_n$ is a random sample from a distribution with expectation 0 and variance σ^2. We cannot observe or calculate $e_1, e_2, \ldots, e_n$, but we can look at their estimates $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ instead. What we should see:

- no systematic pattern in the size or sign of the residuals
- and, if we assume normally distributed errors as in §4.10, additionally:
- a roughly symmetric distribution of the residuals about 0
- no extreme outliers (residuals $\geq 3\hat{\sigma}$ or $\leq -3\hat{\sigma}$, say)

If what we see departs from this ideal, we may be able to judge from the pattern we can see how to change the model so that it does fit for example, we might allow the error variance σ^2 to depend on x, or we could include a quadratic term in the model, like $E(Y|x) = \alpha + \beta x + \gamma x^2$. But this is beyond the scope of this unit.

4.12 Examples of Lack of Fit

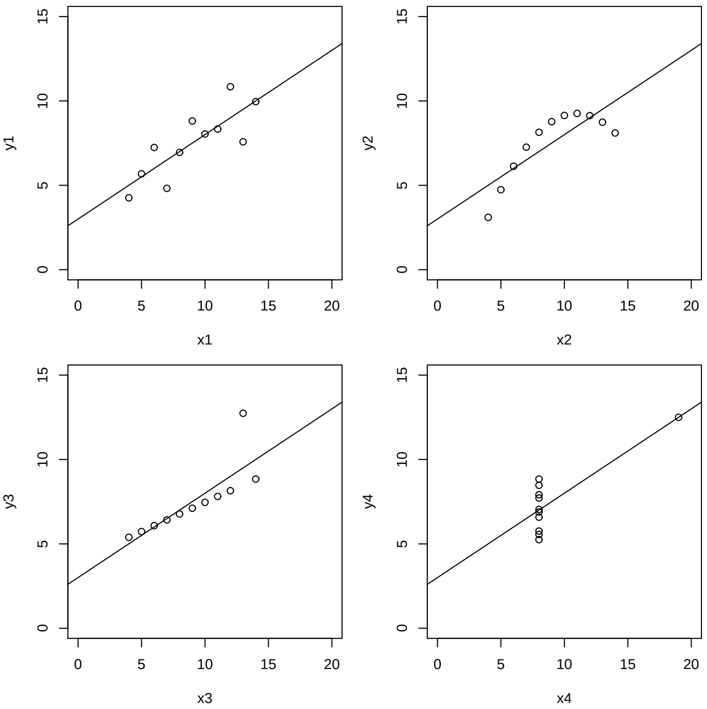
In linear regression examples, you should always plot the points on a scatter plot, draw in the estimated regression line, and also plot the residuals. This may enable you to see by eye (i) if the basic linear model is incorrect; (ii) if there are any unusual observations or outliers, which may perhaps have been wrongly recorded; (iii) if the regression line is especially sensitive any of the observations. This information may not be at all apparent just from the summary data values.

The example below, due to Anscombe, brings out this point clearly. It consists of four artificial data sets, each of 11 data pairs, with the same values of the relevant summary statistics. Thus each data set gives rise to exactly the same regression line and exactly the same inferences for α, β and σ^2. The data are contained in the Statistics 1 data set `anscombe`.

Data Set 1	x values	10	8	13	9	11	14	6	4	12	7	5
	y values	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
Data Set 2	x values	10	8	13	9	11	14	6	4	12	7	5
	y values	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data Set 3	x values	10	8	13	9	11	14	6	4	12	7	5
	y values	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Data Set 4	x values	8	8	8	8	8	8	8	19	8	8	8
	y values	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

The summary statistics for each data set are (approximately):
$n = 11$ $\sum x_1 = 99$ $\sum y_i = 82.5$ $\sum x_i^2 = 1001$ $\sum y_i^2 = 660$ $\sum y_i x_i = 797.5$.

From the scatter plots with the fitted regression lines, we see immediately that there is a lack of fit for data sets 2, 3 and 4: in data set 2 the relationship between x and y is quadratic rather than linear so the simple linear model is incorrect; in data set 3 the simple linear regression model is correct, but a very clear regression line is distorted by the effect of a single outlier; in data set 4, the regression line is particularly sensitive to the y value for the single observation taken at $x = 19$ and it is impossible to tell from this choice of x values whether or not a simple linear regression model is suitable.


y1
x1
y2
x2
y3
x3
y4
x4
0
5
10
15
20