MTH6134 Statistical Modelling II
Lab sessions

Autumn 2023

Lab sessions built upon practicals provided by S Coad (formerly of QMUL).

Practical 1 - 2 and 4 October 2023 (Week 2)

This practical reminds you about the language and commands of `R` which you will practice with the package `RStudio`. This package should be available if you are using a university computer, or you may use your own laptop with your installation of the package. **If everything else fails**, use the site `rdrr.io/snippets/` to run `R` commands. This last option only requires a browser.

Open `RStudio` and from the menu `File>New File>R Script`, open a new script in which you will type commands (alternatively, use `Control+Shift+N`). Remember to save this file for your records and for later practice.

The data

In this practical we revisit linear regression that you used in Statistical Modelling I.

Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Below are data on powerboat registrations (x), in thousands, and the number of manatees killed by boats (y) in Florida in the years 1977 to 1987.

Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
x	447	460	481	498	513	512	526	559	585	614	645
y	13	21	24	16	24	20	15	34	33	33	39

Do the data provide any evidence of a linear relationship between the number of powerboat registrations and the number of manatees killed by boats?

Entering and plotting the Data

First enter the x and y values as vectors in `R`:

```
x <- c(447,460,481,498,513,512,526,559,585,614,645)
y <- c(13,21,24,16,24,20,15,34,33,33,39)
```

Then produce a scatterplot of the data using

```
plot(x,y,main="Plot of y against x")
```

Does the relationship between y and x seem to be linear?

Fitting the Model

Fit a simple linear regression model to the data by

```
manatee <- lm(y ~ x)
```

To see the details of the fitted model, we use

```
summary(manatee)
```

The values of the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$ can be seen from the output. Add the fitted line to your scatterplot by using the following commands:

```
plot(x,y,main="Fitted Line Plot")
abline(manatee)
```

To test $H_0 : \beta_1 = 0$, we look at the analysis of variance table:

```
anova(manatee)
```

What are your conclusions?

Checking the Assumptions

To check the assumptions of the model, we examine the residual plots. We first plot the standardised residuals against x:

```
stres <- rstandard(manatee)
plot(x,stres,main="Standardised Residuals against x")
```

Is there any reason to doubt the linearity of the model? Next, we plot the standardised residuals against the fitted values:

```
fits <- fitted(manatee)
plot(fits,stres,main="Standardised Residuals against Fitted Values")
```

Is there any reason to doubt that the variance is constant? Finally, we look at the Q-Q plot:

```
qqnorm(stres,main="Q-Q Plot")
qqline(stres)
```

Is there any reason to doubt the assumption that the errors are normally distributed?

Finishing the Session

Remember to save your files and log out.

Practical 2 - 9th (and 18th) October 2023 (Week 3)

This practical reminds you how to read data from a file into R and how to fit a multiple linear regression model.

Diabetic Data

The file `diabetic.csv` on the module webpage contains data for 20 male insulin-dependent diabetics who had been on a high carbohydrate diet for six months. Compliance with the regime is thought to be related to age (x_1), in years, body weight as a percentage of 'ideal' weight for height (x_2) and the percentage of calories as protein (x_3). The dependent variable is the percentage of total calories obtained from complex carbohydrates.

Load the data into R as follows. First, you have to tell R where you have saved the data. This is known as your working directory. You set this by telling R where the data are. For example, it is probably convenient to save the data onto the `G:` drive, in which case you use

```
setwd("G:")
```

You can check that it is correct by

```
getwd()
```

Now read in the data by

```
diabetic <- read.csv("diabetic.csv")
```

The data will have been read into R, but they are stored in four columns, and we need to allocate the columns of the matrix to y, x_1, x_2 and x_3. This is achieved by

```
y <- diabetic[,1]
x1 <- diabetic[,2]
x2 <- diabetic[,3]
x3 <- diabetic[,4]
```

Check that data has been correctly read by looking at the first few rows using

```
head(diabetic)
```

and examine the collection of pairwise scatterplots of data with

```
pairs(diabetic)
```

Briefly describe what you observe.

Model 1

Fit a multiple linear regression model to the data in which y is linearly related to each of the explanatory variables by

```
diabetic1 <- lm(y ~ x1 + x2 + x3)
```

Obtain the fitted model using `summary`. Save the fitted values and the standardised residuals. Assess the assumptions of normality and constant variance of the random errors by examining suitable residual plots. What are your conclusions?

Model 2

Fit a multiple linear regression model to the data in which y is linearly related to x_2 and x_3 by

```
diabetic2 <- lm(y ~ x2 + x3)
```

Obtain the fitted model and save values as before. Assess the assumptions of normality and constant variance of the random errors. Compare the two model fits. Which one is best and why?

Model 3

From looking at the initial results (Model 1), it is clear that x_3 is the most important term. Fit a univariate regression model explaining y in terms of x_3 only and compare it with the previous two models.

Finishing the Session

Remember to save your files and log out.

Practical 3 - 16th (and 25th) October 2023 (Week 4)

This practical introduces you to generalised linear models by modelling some binary response probabilities.

Beetle Data

In bioassays, the response may vary with a covariate x termed the *dose*. Data for a typical example involving a binary response are given in the file `beetle.csv` on the module webpage. Here, a certain number of beetles (r) are exposed to various concentrations of gaseous carbon disulphide, in milligrammes per litre, for five hours and the number of beetles killed (y) is recorded. The dose is the base 10 logarithm of the concentration.

The three columns of `beetle.csv` contain x, r and y in that order. Create variables `x`, `r` and `y` from the columns of the data. The proportion of beetles killed at each of the doses may be calculated using

```
p <- y/r
```

Now plot these proportions against the doses by

```
plot(x,p,main="Plot of p against x")
```

The purpose of this practical is to identify a model which provides a good description of the data.

Logistic Model

Fit a logistic model to the data in which the probability of a beetle being killed π_i at the ith dose x_i satisfies the logit link

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

by

```
beetle1 <- glm(p ~ x,family=binomial(link=logit),weights=r)
```

Note that, since a logit link is the default one for the binomial distribution, it is sufficient to put `family=binomial` as the second argument in this case.

Obtain the fitted model using `summary`. The standard errors of the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$ are the square roots of the diagonal elements of the inverse of the estimated Fisher information matrix. Under the null hypothesis that the logistic model describes the data, the residual deviance has an approximate χ^2_6 distribution, since there are eight doses and two parameters in the model. Does this model provide a good fit?

Probit Model

Fit a model to the data in which π_i satisfies the probit link

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i,$$

where Φ denotes the standard normal distribution function, by

```
beetle2 <- glm(p ~ x,family=binomial(link=probit),weights=r)
```

Obtain the fitted model as before. Compare the two model fits. Are they similar?

Extreme Value Model

Fit an extreme value model to the data in which π_i satisfies

$$\log\{-\log(1-\pi_i)\} = \beta_0 + \beta_1 x_i$$

by

```
beetle3 <- glm(p ~ x,family=binomial(link=cloglog),weights=r)
```

Here, `cloglog` means that a complementary log-log link is being used. Obtain the fitted model as before. Which of the three models provides the best description of the data?

Finishing the Session

Remember to save your files and log out.

Practical 4 - 23rd October 2023 (Week 5)

This practical shows you how to fit a Poisson regression model to count data.

Count Data

The data below are counts (y) observed at various values of a covariate (x).

y	2	3	6	7	8	9	10	12	15
x	-1	-1	0	0	0	0	1	1	1

Enter the x and y values as vectors in R. Call them `x` and `y`. Then produce a scatterplot of the data. Is the variance constant?

Linear Model

Fit a linear model to the data in which the mean response μ_i for covariate x_i satisfies

$$\mu_i = \beta_0 + \beta_1 x_i$$

by

```
count1 <- glm(y ~ x,family=poisson(link=identity))
```

Note that, since an identity link is not the default one for the Poisson distribution, it is necessary to put `family=poisson(link=identity)` as the second argument in this case.

Obtain the fitted model using `summary`. Under the null hypothesis that the linear model describes the data, the residual deviance has an approximate χ^2_7 distribution, since there are nine observations and two parameters in the model. Does this model provide a good fit?

Log-Linear Model

Fit a log-linear model to the data in which μ_i satisfies

$$\log(\mu_i) = \beta_0 + \beta_1 x_i$$

by

```
count2 <- glm(y ~ x,family=poisson(link=log))
```

Obtain the fitted model as before. Which of the two models provides the best description of the data?

Finishing the Session

Remember to save your files and log out.

Practical 5 - 30th October 2023 (Week 6)

This practical shows you how to assess whether a link function is appropriate and how to check the fit of a model.

Beetle Data Revisited

In Practical 3, you fitted three models to the beetle data. Each of these was based on a different link function. The purpose of this practical is to see which of these link functions is best, and then to examine the fitted values of the chosen model.

Assessing a Link Function

To assess whether a logit link function might be appropriate, plot the empirical logits

$$\log\left(\frac{y_i + 0.5}{r_i - y_i + 0.5}\right)$$

against the doses. Note that the number 0.5 added in both numerator and denominator attempts to avoid problems when $y_i = 0$ or r_i. The plot is produced by

```
el <- log((y + 0.5)/(r - y + 0.5))
plot(x,el,main="Plot of Empirical Logits against x")
```

and should be approximately linear.

The possibility of a probit link function can be examined by plotting the empirical probits

$$\Phi^{-1}\left(\frac{y_i + 0.5}{r_i + 1}\right)$$

against the doses. This time, the plot is produced by

```
ep <- qnorm((y + 0.5)/(r + 1))
plot(x,ep,main="Plot of Empirical Probits against x")
```

Is this plot similar to the first one?

Similarly, to see whether the complementary log-log link function is adequate, plot the empirical log-log values

$$\log\left\{-\log\left(\frac{r_i - y_i + 0.5}{r_i + 1}\right)\right\}$$

against the doses. Here, the plot is produced by

```
ell <- log(-log((r - y + 0.5)/(r + 1)))
plot(x,ell,main="Plot of Empirical Log-Log Values against x")
```

Which of the three link functions appears to work best?

Checking the Fit

Fit your chosen model and call it `beetle`. Obtain the fitted values using

```
fits <- fitted(beetle)
print(fits)
```

Plot both the proportions and the fitted values against the doses by

```
plot(x,p,col="blue",main="Plot of p against x")
lines(x,p,col="blue")
points(x,fits,col="red")
lines(x,fits,col="red")
```

The blue points in the plot correspond to the observed proportions and the red ones to the fitted values. What are your conclusions?

Finishing the Session

Remember to log out.

Practical 6 - 13 November 2023 (Week 8)

This practical introduces you to one of the different residuals that can be examined for generalised linear models.

Cloth Data

The file `cloth.csv` on the module webpage gives the length (x), in metres, and the number of defects (y) for 32 pieces of cloth. Since the latter variable is discrete and its discreteness is effectively restricted to very few values when x is small, this suggests that the mean number of defects increases with x.

The two columns of `cloth.csv` contain x and y in that order. Then produce a scatterplot of the data. Is the variance constant?

Fitting the Model

Fit a Poisson regression model to the data in which the mean number of defects μ_i for length x_i satisfies

$$\mu_i = \beta x_i.$$

Note that, in order to exclude the intercept, the specified model is `y ~ x - 1`. Another possibility is `y ~ 0 + x`.

Obtain the fitted model using `summary`. What is the maximum likelihood estimate of β? Does this model provide a good fit?

Checking the Fit

Obtain the fitted values using the `R` function `fitted()` and plot both the numbers of defects and the fitted values against the lengths.

Examining the Residuals

Compute the Pearson and deviance residuals using the `R` function `residuals()`, and compute the Anscombe residuals. Is there any reason to doubt the model assumptions?

Finishing the Session

Remember to log out.

Practical 7 - 22 November 2023 (Week 9)

Rat Data

The effects of the dose of poison x, in milligrammes, and the method of delivery w on the probability of survival were examined in a study of rats. The two methods of delivery were as a solid with food or as a liquid in water. For each combination of dose and method of delivery, a certain number of rats r were used and the number who survived y is recorded.

The data are given in the file `rat.csv`, with four columns containing x, w, r and y in that order. Read these data into `R`, compute the proportion of rats who survived and define w as a qualitative variable or factor.

Plot proportions against doses by method of delivery where black points correspond to the delivery of the poison as a solid with food and the red ones as a liquid in water. Add a legend for clarity. What are your conclusions?

Model 1

Fit a logistic regression model to the data in which the probability of a rat surviving π_{jk} at the k-th dose x_k with the j-th method of delivery satisfies

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta_j x_k.$$

This model allows a different intercept and slope for each method of delivery. Does this model provide a good fit?

Model 2

Fit a logistic regression model to the data in which π_{jk} satisfies

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta x_k$$

The slope is the same for the two methods of delivery. Fit the model and compare the residual deviance with that of Model 1. Are the regression lines parallel?

Model 3

Fit a logistic regression model to the data in which π_{jk} satisfies

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \beta x_k$$

Both intercept and slope are the same for the two methods of delivery. Fit the model and compare the residual deviance with that of Model 2. Is there a difference between the methods of delivery?

Compare your results with Examples 4.1 and 4.2 in the lecture notes, and make sure you can reproduce it.

Finishing the Session

Remember to log out.

Practical 8 - 27 November 2023 (Week 10)

This practical uses Poisson regression for count data.

Chronic medical conditions data

It has been observed that women who live in country areas tend to have fewer consultations with general practitioners (family physicians) than women who live near a wider range of health services. It is not clear whether this is because they are healthier or because structural factors, such as shortage of doctors, higher costs of visits and longer distances to travel, act as barriers to the use of general practitioner (GP) services. The table below shows the numbers of chronic medical conditions (for example, high blood pressure or arthritis) reported by samples of women living in large country towns (town group) or in more rural areas (country group) in an area. All the women were in the same age group, had the same socio-economic status and had three or fewer GP visits in the past year. The question of interest is: do women who have similar levels of use of GP services in the two groups have the same need as indicated by their number of chronic medical conditions?

Group	Numbers of chronical medical conditions
Town	0 1 1 0 2 3 0 1 1 1 1 2 0 1 3 0 1 2 1 3 3 4 1 3 2 0
Country	2 0 3 0 0 1 1 1 1 0 0 2 2 0 1 2 0 0 1 1 1 0 2

Enter the data into `R` defining variable `y` for counts and `x` for the categorical variable groups. Plot the data. Is there an apparent difference between the two groups? What type of plot is most effective for this type of data?

Model fitting

The Poisson distribution provides a plausible way of modelling these data as they are counts and within each group the sample mean and variance are approximately equal. Compute these summary statistics per group to verify the claim.

Let Y_{jk} be the number of conditions for the k-th woman in the j-th group, where $j = 1$ for the town group and $j = 2$ for the country group and $k = 1, \ldots, K_j$ with $K_1 = 26$ and $K_2 = 23$. The Y_{jk} are assumed to be independent random variables $\text{Poi}(\alpha_j)$, where the mean α_j is the expected number of conditions.

The question of interest can be formulated as a test of $H_0 : \alpha_1 = \alpha_2 = \alpha$ against alternative $H_1 : \alpha_1 = \alpha_2$. The model fitting approach to testing H_0 is to fit two models, one assuming H_0 is true, that is $Y_{jk} \sim \text{Poisson}(\alpha)$ and the other assuming H_0 is not true, so that $Y_{jk} \sim \text{Poisson}(\alpha_j)$ where $j = 1$ or 2.

Testing H_0 against H_1 involves comparing how well these two models fit the data. If they are about equally good then there is little reason for rejecting H_0. However if the second model is clearly better, then H_0 would be rejected in favor of H_1.

Analyze the data using the `glm` function and postprocessing the output as required.

Extras

In your analyses, you probably used the syntax `glm(y~x,family=poisson)` for fitting the model with two means. Here you are going to explore in detail the model outputs and relating these to the link function. In particular, look at the output of the model: coefficients, fitted values and deviances for each of the following analyses `glm(y~x,family=poisson)` and `glm(y~x,family=poisson(link="identity"))` and respond to the following items for each case.

1. determine how the fitted values relate to the means per group;
2. how do the coefficients relate as well to the means per group. In your analysis you need to consider the link used.
3. Look at the deviances between analyses: they concide. Why?

Repeat the analyses using `y~as.integer(x)` instead of `y~x`.
What happens when you use the syntax `y~x-1`?

Finishing the Session

Remember to log out.

Practical 9 - 4 December 2023 (Week 11)

In this practical we use log-linear models for count data in a contingency table.

Cancer Data

In a cross-sectional study of bone cancer, the type of cancer and the site were recorded for 300 patients. The contingency table below shows the number of patients (y) with each combination of type of cancer and site.

	Site				
Type	Head	Arms	Body	Legs	Total
I	21	13	42	8	84
II	10	26	20	35	91
III	30	34	32	29	125
Total	61	73	94	72	300

Do the data provide any evidence that there is an association between type of cancer and site?

Entering the Data

First enter the y values as a vector in R column by column:

```
c1 <- c(21,10,30)
c2 <- c(13,26,34)
c3 <- c(42,20,32)
c4 <- c(8,35,29)
y <- c(c1,c2,c3,c4)
```

Then generate the levels of the row and column factors by

```
row <- gl(n = 3,k = 1,length=12)
column <- gl(n = 4,k = 3,length=12)
```

Note that the first and second arguments of `gl` are the numbers of levels and replications, respectively. Check that the levels have been correctly generated.

Fitting the Model

Fit a log-linear model to the data by

```
cancer <- glm(formula=y ~ row + column,family = poisson)
```

This model assumes that there is no association between type of cancer and site. Obtain the fitted model using `summary`. Does this model provide a good fit?

The value of Pearson's goodness-of-fit test statistic can also be calculated by

```
observed <- cbind(c1,c2,c3,c4)
chisq.test(observed)
```

Are the values of the two statistics similar? Can you explain the difference?

Examining the Residuals

To assess the fit of the model, obtain the Pearson and deviance residuals. Which observations are making the largest contributions to the two measures of goodness of fit?

Compare your answers with Example 5.2 in the lecture notes, and make sure you can reproduce the results.

Extra

Fit the model

```
cancer2 <- glm(formula=y ~ row * column,family = poisson)
```

Look at the output of `summary(cancer2)`. Can you explain the results? Which is the model being fitted?

Three-dimensional contingency table

For an extra challenge, this exercise shows you how to analyse a three-dimensional contingency table when the row totals are fixed.

Hernia Data

In a retrospective case-control study of hernias and occupations, a group of hernia patients were matched with a group of controls who were similar with respect to age and socio-economic class. The contingency table below shows the number of patients (y) in each group with each combination of type of hernia and type of occupation.

		Occupation		
		Manual	Sedentary	Total
Inguinal	Cases	25	35	60
	Controls	20	40	60
Other	Cases	37	18	55
	Controls	19	36	55

Do the data provide any evidence that there is an association between type of occupation and disease status, case or control, given type of hernia?

Entering the Data

First enter the y values as a vector in R. Call it `y`. Then generate the levels of the occupation, hernia and group factors. Check that the levels have been correctly generated.

Model 1

Fit a log-linear model to the data in which it is assumed that there is no association between type of occupation and disease status given type of hernia by

```
retro1 <- glm(formula = y ~ occ + hernia + group + occ : hernia
              + hernia : group,  family = poisson)
```

Note that, because the row totals are fixed, the model must always include the terms `hernia`, `group` and `hernia : group`. Obtain the fitted model using `summary`. Does this model provide a good fit?

Model 2

Fit a log-linear model to the data which allows for an association between type of occupation and disease status given type of hernia by

```
retro2 <- glm(formula = y ~ occ + hernia + group + occ : hernia
              + occ : group + hernia : group, family = poisson)
```

Obtain the fitted model as before. Compare the residual deviance with that of Model 1. Is the extent of the association between type of occupation and disease status the same for both hernia types?

Examining the Residuals

To assess the fit of the two models, obtain the deviance residuals for the two models. Is Model 2 an improvement over Model 1?

Finishing the Session

Remember to log out.

Practical 10 - 11 December 2023 (Week 12)

This practical shows you how to fit an exponential regression model to survival data.

Leukaemia Data

The survival time of patients with leukaemia (t), in weeks, was recorded. Patients were classified into two groups (w); the white blood cell count (x), in thousands, is a covariate. The data are given in the file `leukaemia.csv` on the module webpage.

The file `leukaemia.csv` contains columns x, w and t. Read these variables into R calling them `x`, `w` and `t` and compute the logarithms of the survival times and the white blood cell counts. Then produce a scatterplot of these data by group. Do they provide evidence of linear relationships between the logarithms of the survival times and the logarithms of the white blood cell counts for each group?

Model 1

Fit an exponential regression model to the data in which the mean survival time μ_{jk} of a patient with white blood cell count x_k in the jth group satisfies

$$\log(\mu_{jk}) = \alpha_j + \beta_j \log(x_k)$$

Note that, since a reciprocal link is the default one for the gamma distribution, it is necessary to specify the log link in R. When looking at the results we must also set the dispersion parameter is set to one because an exponential distribution is being assumed for the survival times. Does this model provide a good fit?

Model 2

Fit an exponential regression model to the data in which μ_{jk} satisfies

$$\log(\mu_{jk}) = \alpha_j + \beta \log(x_k)$$

Here, the slope is the same for the two groups. Obtain the fitted model as before. Compare the residual deviance with that of Model 1. Are the lines parallel?

Model 3

Fit an exponential regression model to the data in which μ_{jk} satisfies

$$\log(\mu_{jk}) = \alpha + \beta \log(x_k)$$

This time, both the intercept and the slope are the same for the two groups. Obtain the fitted model as before. Compare the residual deviance with that of Model 2. Is there a difference between the groups?

Finishing the Session

Remember to log out.