Statistics and Scientific Method: Rwanda, March 2022

Exercises

A. A useful result in probability theory

For any discrete-valued random variable X with probability distribution $p(x)$, and any function $g(X)$, the *expectation* of $g(X)$ is $\mathrm{E}[g(X)] = \sum_x g(x)p(x)$, and the *variance* of X is $\mathrm{Var}(X) = \mathrm{E}[(X - E[X])^2]$.

For any two discrete-valued random variables, X and Y, with joint probability distribution $p(x,y) = \mathrm{Prob}(X = x, Y = y)$, recall that $p(x,y)$ can be factorised in two equivalent ways, $p(x,y) = p(x)p(y|x) = p(y)p(x|y)$, where $p(x|y)$ and $p(y|x)$ are the *conditional distributions* of X given Y and of Y given X. The *expectation* of any function $h(X,Y)$ is

$$\mathrm{E}[h(X,Y)] = \sum_x \sum_y h(x,y)p(x,y).$$

Show that:

1. $\mathrm{E}[X] = \mathrm{E}_Y[\mathrm{E}_X[X|Y]]$
2. $\mathrm{Var}(X) = \mathrm{Var}_Y(E_X[X|Y]) + \mathrm{E}_Y[\mathrm{Var}_X(X|Y)]$,

where the subscripts indicate the random variable with respect to which each expectation is taken.

Note. The same results hold for continuous-valued random variables, with probabiity density functions replacing probability distirbutions.

B. The asthma data

Data from the clinical trial of two drugs for the treatment of chronic asthma are given in the file `asthmadata.csv`

The study-design was a two-period crossover. Patients 1 to 7 received Formoterol in the first time-period, Salbutamol in the second. Patients 8 to 13 received Salbutamol in the first time-period, Formoterol in the second.

1. Calculate the differences, $d_i : i = 1, ..., 13$ between PEF after administraton of"F" or "S" for each child.
2. Calculate a 95% confidence interval for the mean difference. Does this give evidence that amongst the population of asthmatic children from which the trial subjects were drawn, the mean difference is statistically and/or clinically significant?
3. Implement the following `R` code:

   ```
   attach(asthma)
   d<-F-S
   fit<-lm(d~1); summary(fit)
   ```

   How do the results relate to your solution to the previous exercise?
4. The asthma trial was designed to control both for variation between children and variation between time-periods in the PEF response, but the above analysis ignores any effect of time-period. How would you extend the linear model set out in the previous exercise to take time-period into account?
5. Re-analyse the asthma data using your extended model and summarise your conclusions regarding the relative efficacy of Formoterol and Salbutamol.

6. **Challenge question**. Write an R function that implements a randomisation-based method to test the hypothesis that there is no difference between the two treatments, whilst alowing for possible time-period effects. Compare the result with the result from exercise A5.

C. A first look at the PANSS data

An extract from the schizophrenia dataset is given in the file `PANS50data.csv`. You should see that there are 50 patients each with between 1 and 6 measurements of PANSS, a questionnaire-based instrument for measuring overall severity of mental health symptoms. Measurements were intended to be made 0, 1, 2, 4, 6 and 8 weeks after recruitment but some patients dropped out before 8 weeks.

1. Plot the PANSS responses against time.
2. Add a line-graph of the *observed mean* PANSS response at times 0, 1, 2, 4, 6 and 8.
3. Develop a model for the mean PANSS response as a polynomial function of time, and add your fitted polynomial to the plot.
4. **Challenge question** What assumptions is the model you developed in the previous exercise making, and do the data support these?

D. Re-visiting the PANSS data

1. Plot the PANSS responses against time.
2. Add a set of line-graphs, one for each patient
3. What do you notice about:
   (a) the rank-order of the 50 lines over time?
   (b) the position in the graph of the *last* recorded PANSS measurement on each patient
4. The following code re-fits a model for the mean PANSS response as a (quadratic) polynomial function of time, with a random intercept for each patient. The data-set must first be stored in `R` as a data-frame, `PANSS50`.

```
library(nlme)
fit<-lme(PANSS~week+I(week^2),random=~1|ID,data=PANSS50,na.action=na.omit)
summary(fit)
```

   (a) Make a line-graph of the observed mean PANSS responses at weeks 0, 1, 2, 4, 6 and 8.
   (b) Add two smooth curves showing the fitted mean PANSS response as a polynomial function of time, according to the models you fitted in Lab 2 and in Lab 3
   (c) Discuss the similarities and differences between the two fitted models

E. A first look at the Kericho malaria data

The file `kerichodata.csv` contains a time series of monthly incident malaria cases amongst the workforce of Kericho tea-plantation, Kenya, with some potentially relevant explanatory variables. To a first approximation, the size of the workforce can be assumed to be constant over the time-period covered by the data.

	Year:	self-explanatory
	Month:	self-explantory
	Cases:	number of incident cases
Columns are:	Rain:	average rainfall in mm/day
	minT:	minimum temperature in degrees C
	maxT:	maximum temperature in degrees C
	VCAP:	a proxy measure of mosquito abundance

1. Load the data into a data-frame called “kericho” and run the following R code.

```
y<-kericho$malaria.cases
x<-1:length(y)
par(mfrow=c(2,1)
plot(x,y,type="l",xlab="month",ylab="cases")
plot(x,log(y),type="l",xlab="month",ylab="log-cases")
```

   Comment on the general features of the two plots. Can you suggest a *statistical* and a *biological* reason why you might prefer to analyse the data on the log-transformed scale?

2. Run the following R code.

```
par(mfrow=c(1,1))
y<-log(kericho$malaria.cases)
smooth<-lowess(x,y,f=1/3)
lines(smooth$x,smooth$y)
residuals<-y-smooth$y
```

   (a) Comment on the general features of the plot. What is the `lowess()` function doing?

   (b) Investigate the autocorrelation function of the time series of `residuals`. How can you explain the pattern of the residual autocorrelation function?

F. A time series model for the Kericho malaria data

1. Use the `lm()` function to fit a model to the log-transformed time series, `y`, that accounts for both the long-term trend in malaria incidence and the pattern of seasonal variation about the long-term trend.

2. investigate whether adding any or all of Rain, minT, maxT or VCAP to the model materially improves the fit.

3. Investigate the autocorrelation function of the time series of `residuals` from your preferred model and comment on its features.

4. **Challenge question** Your current model for the log-transformed case numbers can be written as

$$Y_t = x_t'\beta + Z_t,$$

   where x_t is a vector whose elements will depend on whch explanatory variables you have chosen to include in your model, and the residuals, Z_t, are assumed to be independent, Normally distributed with mean zero and variance σ^2. Suppose instead that the Z_t are generated by a model of the form,

$$Z_t = \alpha Z_{t-1} + W_t,$$

   where $-1 < \alpha < 1$ and the W_t are Normally distributed with mean zero and variance τ^2.

(a) Show that the Z_t are Normally distributed with mean zero and variance $\sigma^2 = \tau^2/(1-\alpha^2)$

(b) Show that $\text{Corr}(Z_t, Z_{t-u}) = \alpha^{|u|}$.

(c) Find the log-likelihood function for the model and show that *if the value of α is known*, the maximum likelihood estimates of β and σ^2 can be written explicitly as

$$\hat{\beta}(\alpha) = (X'R(\alpha)^{-1}X)^{-1}X'R(\alpha)^{-1}Y,$$

$$\hat{\sigma}^2(\alpha) = n^{-1}\{Y - X'\hat{\beta}(\alpha)\}'R(\alpha)^{-1}\{Y - X'\hat{\beta}(\alpha)\},$$

where n is the length of the vector $Y = (Y_1, ..., Y_n)$ and X and $R(\alpha)$ are suitable matrices, and that

$$\text{Var}\{\hat{\beta}(\alpha)\} = \sigma^2(X'R(\alpha)^{-1}X)^{-1}$$

(d) Show how you can use these results to obtain the maximum likelihood estimate of α when its value is unknown, using a one-dimensional numerical search

(e) Apply your function to the Kericho malaria data, and hence find the maximum likelihood estimates of β, σ^2 and α, and approximate standard errors for the elements of $\hat{\beta}$.

(f) Compare your maximum likelihood estimates of β, and their approximate standard errors, with the results you obtained prevously using the `lm()` function, and comment on any differences between the two sets of results.

5. You could fnd an `R` package to fit the time series model of the previous exercise, but working through it from first principles is good practice for deriving and implementing maximum likelihood estimation when a packaged solution isn't available. Here's another (simpler) example.

Suppose that $Y_1, ..., Y_n$ are independent randon variables, each Y_i is binomially distributed with number of trials m_i and probability of success p_i, where the values of p_i are given by

$$\log\{p_i/(1-p_i)\} = \alpha + \beta(x_i - \bar{x}),$$

$x_1, ..., x_n$ are the observed values of an explanatory variable and $\bar{x} = (\sum_{i=1}^{n} x_i)/n$

(a) Write an `R` function to simulate a data-set from this model. i.e. choose a set of values $m = (m_1, ..., m_n)$, $x = (x_1, ..., x_n)$, parameter values α, β and use the `R` function `rbinom()` to generate the vector $y = (y_1, ..., y_n)$.

(b) Write an `R` function to evaluate the log-likelihood function, $L(\alpha, \beta)$.

(c) Use the `R` function `optim()` to find the maximum likelihood estimates of α and β and their approxiate standard errors.

(d) Compare your answers with the result of the following `R` commands:

```
fit<-glm(cbind(y,m-y)~x,family=binomial)
summary(fit)
```

G. Onchocerciasis prevalence in Liberia

The file `LiberiaRemoData.csv` contains the following information from 90 rural communities in Liberia:

lat:	latitude
long:	longitude
ntest:	numberof people tested for onchocerciasis
npos:	number testing positive
elevation:	height above sea-level (metres)
log-elevation:	log(elevation)

1. Calculate the *empirical logit* transformation, $y = \log\{(\text{npos} + 0.5)/(\text{ntest} - \text{npos} + 0.5)\}$, and display the 90 values of y on a map.

2. Develop a regression model for y, considering lat, long, elevation and log-elevation as potential covariates.

3. Display the 90 residuals from your proposed regresion model on a map. Can you see any spatial pattern in the map?

4. **Challenge question** Install the `R` package `PrevMap` and develop a model to predict onchocerciasis logit-prevalence, allowing for a component of spatial variation in prevalence that cannot be explained by the available covariates.