UNIVERSITY OF BRISTOL

School of Mathematics

Theory of Inference
MATH35600/MATHM0019
(Paper code MATH–35600)

Example Exam 2 hours 30 minutes

This paper contains FOUR questions. All answers will be used for assessment.

Calculators are not permitted in this examination.

The marking scheme is indicative and is intended only as a guide to the relative weighting of the questions.

Do not turn over until instructed.

1. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \mathbf{I}\sigma^2$. As usual $\mathbf{y}$ is a response vector, $\mathbf{X}$ a full rank $n \times p$ model matrix ($p < n$), $\boldsymbol{\beta}$ a parameter vector and $\boldsymbol{\epsilon}$ a vector of residual random variables. Consider the orthogonal decomposition $\mathbf{X}^\mathsf{T} = [\mathbf{L}, \mathbf{0}]\mathbf{U}$ where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an orthogonal matrix. Let $\mathbf{U}_f$ denote the first p rows of $\mathbf{U}$ and $\mathbf{U}_r$ denote the remaining $n - p$ rows.

   (a) Show that the least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \mathbf{L}^{-\mathsf{T}}\mathbf{U}_f\mathbf{y}$. [5 marks]

   (b) Show that the estimator, $\hat{\boldsymbol{\beta}}$, from part (a), is unbiased. [2 marks]

   (c) Show that $\hat{\boldsymbol{\beta}}$ from (a) is equivalent to $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. [3 marks]

   (d) Suppose you are working in an online advertising consultancy, and examining the relationship between daily time spent online t_i and advertising revenue generated a_i for a sample of web users. The model

$$a_i = \beta_0 + \beta_1 t_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

   seems to work well for the data, although $\hat{\sigma}^2$ is large, so $\hat{\beta}_1$ has quite high uncertainty. Your boss (very much an 'ideas man') has been reading up on experimental design and noticed that the design minimizing the variance of $\hat{\beta}_1$ would place all the t_i values at the extremes of their possible range. He therefore proposes to discard the data corresponding to the middle half of the t_i observations and refit the model, claiming that this will reduce the uncertainty of $\hat{\beta}_1$.

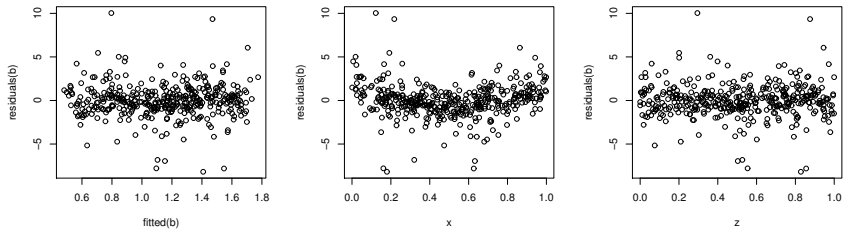
      i. Comment on this proposal with reference to the Gauss Markov theorem.

      ii. Your boss says he has never heard of the Gauss Markov theorem, stop blinding him with science. Provide a simpler explanation of why the proposal is unlikely to work.

   [5 marks]

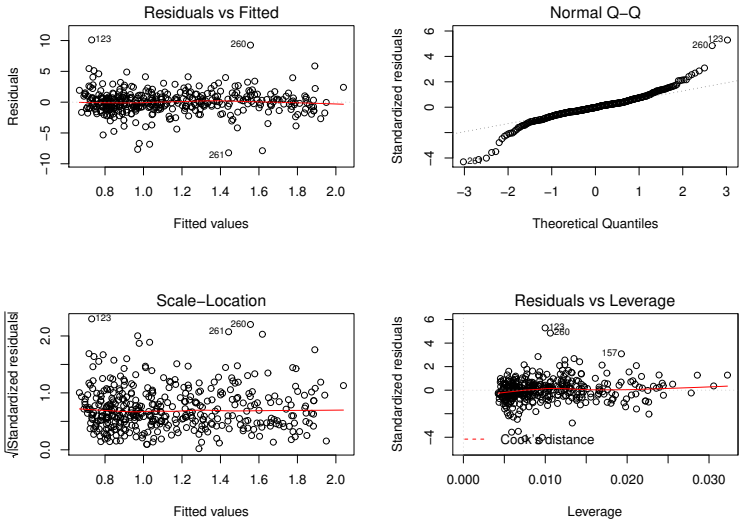
   (e) Write down the statistical model being fitted in the following code, and explain how and why you would modify it in the light of the plots produced.

```
b <- lm(y~x+z)
par(mfrow=c(1,3))
plot(fitted(b),residuals(b))
plot(x,residuals(b))
plot(z,residuals(b))
```



   [5 marks]

   (f) Assuming the model is corrected as you suggested above, the R default residual plots for the model, now have the appearance shown below.



Identify the one problem evident in the plots, commenting on the seriousness or otherwise of this. How might you use a computer intensive approach to check whether the identified problem was adversely affecting confidence intervals calculated for the parameters of the fitted model? Write example R code for obtaining a 95% confidence interval for the coefficient associated with z in the model using this computer intensive method. [5 marks]

2. (a) The department of health wishes to commission research on the effects of cycling and running on health, in order to decide on recommendations for how people can improve their cardiovascular health (health of heart and rest of blood handling system). A call for proposals is issued and you are asked to compare the following two proposals (the first of which is somewhat cheaper). Write brief, statistically based, advise on which study the department should fund, making sure to explain your choice.

   i. The research wing of large internet company proposes to enlist 20,000 runners and cyclists owning personal health monitors (fitbits, etc) to log their weekly running and/or cycling activity, monitored data on heart rate and body fat, and (with consent) to gather other information about them from their online profiles and activity. They will then model these data using the latest methods in order to establish the effect of exercise on health.

   ii. A university public health department proposes to enlist 600 participants aged 40-60, initially taking less than 30 minutes vigorous exercise a week, to take place in a study. The participants will be randomly allocated to a control (no additional exercise), cycling or running treatment, with those on the cycling or running treatment receiving a tailored programme of running or cycling for 2 hours per week. Standardized measures of cardiovascular function and body fat mass will be measured at enrolment, at the end of the study, and at points in between, along with other data such as subject age, sex, weight, etc. The effects of the treatments and covariates on cardio function and fat mass will be analysed using linear models. [10 marks]

   (b) Briefly explain, with an example, the meaning of *confounding* in a statistical model. [5 marks]

(c) Explain concisely how *randomization* is used to set up experiments in a way that avoids confounding problems. [4 marks]

(d) Define an *instrumental variable*, and give a brief mathematical explanation of how they can be used to circumvent confounding problems in observational data analysed with linear models. [6 marks]

3. (a) By considering a Taylor expansion of the derivative of the log likelihood, derive a large sample approximation for the covariance matrix of the maximum likelihood estimator of a parameter vector $\boldsymbol{\theta}$. You can assume that the MLE is consistent and the likelihood sufficiently regular, and you can use standard results on the expectation of derivatives of the log likelihood without proof. [5 marks]

(b) Explain how the Cramer-Rao lower bound relates to the result from (a). [3 marks]

(c) Explain whether the result from (a) is useful for finding the approximate variance of a variance parameter estimator $\hat{\sigma}^2$ when the true value of σ^2 is 0 (or very close to zero). Suggest an alternative approach you might take in this case. [2 marks]

(d) Briefly explain Newton's method for finding maximum likelihood estimates and how it relates to part (a), above. [3 marks]

(e) An electrical goods store chain runs a customer help line, and wants to ensure that it has enough staff available to meet demand. The number of calls received in any given week varies widely, and it is hoped to develop models to better predict demand. One possible model is that the number of calls in one week depends on the number of sales in the previous week, and data are available on weekly numbers of calls, together with sales volume in the previous week, for one year.

i. The base model for the weekly call data is that the number of calls per week is a Poisson random variable with mean θ. Write an R function, `l1`, to evaluate the negative log likelihood of θ, given that 52 weeks of call data are stored in a vector `calls`. The first argument of `l1` should be variable containing a value for θ. [3 marks]

ii. A second model is suggested in which the weekly calls are still realizations of Poisson random variables, but now the expected number of calls each week is given by a 'baseline' number of calls plus some small coefficient multiplied by the total number of sales the previous week. Both the baseline number and the small coefficient are intrinsically positive quantities, so the model is written as:

$$\texttt{calls}_i \sim \text{Poi}(\lambda_i) \quad \text{where} \quad \lambda_i = \mathrm{e}^{\beta_0} + \mathrm{e}^{\beta_1} \times \texttt{sales}_i.$$

$\texttt{calls}_i$ is the number of calls in week i; $\texttt{sales}_i$ is the number of sales in the week before week i and the β_j are parameters to be estimated. Write an R function, `l2`, to evaluate the negative log likelihood of the parameters given vectors of `sales` and `calls` data. The first argument of `l2` should be a vector `b` containing the values for β_0 and β_1, in that order. [3 marks]

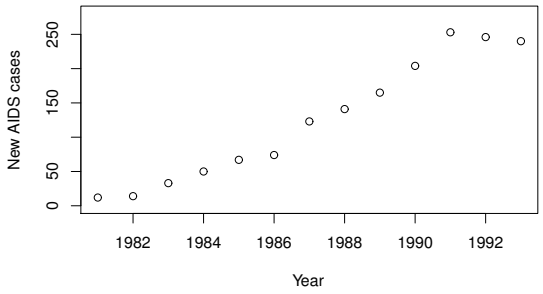
iii. The following is an R session using the two likelihood functions `l1` and `l2`. Explain, briefly, what each R command is doing, what the results mean statistically, and what large sample distributional result has been used. [3 marks]

```
> fit0 <- optim(150,l1,method="BFGS",calls=calls)
> fit0$par
[1] 119.8077
```

```
> fit1 <- optim(c(5,-6),l2,method="BFGS",calls=calls,sales=sales)
> fit1$par
[1]  3.606867 -4.677395
> lrt <- 2*(fit0$val-fit1$val)
> lrt
[1] 15.95610
> 1-pchisq(lrt,df=1)
[1] 6.48286e-05
```

iv. Explain the theoretical problem with part (e), and how you might fix it. [3 marks]

4. The following plot shows new AIDS cases in Belgium at the start of the epidemic.



(a) Consider the model $y_i \sim \text{Poi}\{n_0 \exp(rt_i)\}$ where y_i is the number of AIDS cases in year t_i after 1980. Use R code to write a Metropolis Hastings sampler for n_0 and r, given that the y_i and t_i values are in vectors `y` and `t` respectively. Assume that the prior distribution for n_0 is a gamma distribution with shape and scale parameters 4 and 0.4 respectively (`dgamma(x,4,.4)` evaluates such a density in R), while the prior for r is $N(0, 0.1^2)$. Normal random walk proposals are fine. [10 marks]

(b) Describe 2 diagnostic plots you would examine and what they can tell you. [2 marks]

(c) The following code analyses the same dataset using JAGS called from R using the `rjags` package. Write down mathematically the model being used in this case (you do not need to worry about knowing the exact parameterizations being used for the priors, which have been chosen to be vague/uninformative). [5 marks]

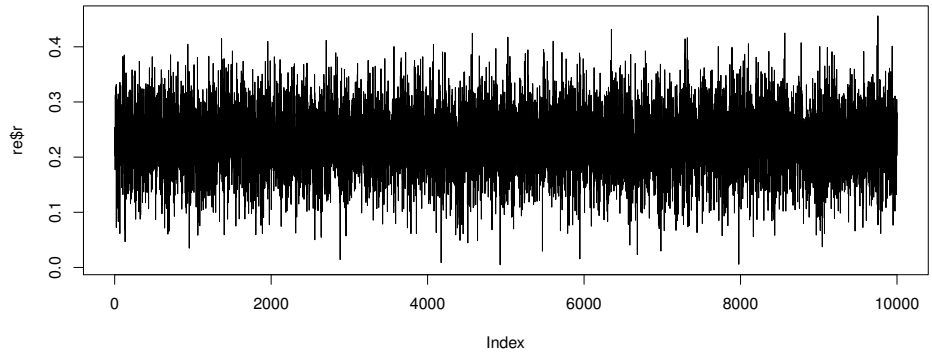
```
library(rjags)
setwd("~sw283/lnotes/brinference/exam-eg")
dat <- list(y=y)
jal <- jags.model("exam.jags",data=dat,inits=list(n=log(y)),n.adapt=1000)
re <- jags.samples(jal,c("n","r"),n.iter=10000)
```

The file `"exam.jags"` contains

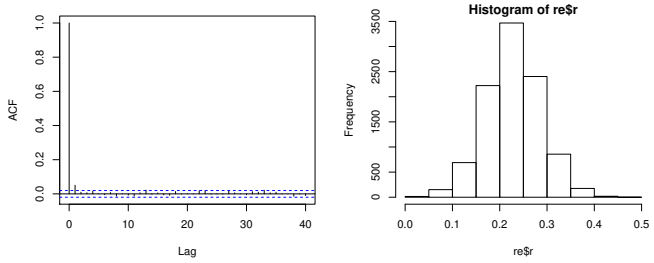
```
model {
  n[1] ~ dlnorm(2.5,1)
  for (i in 2:13) { n[i] ~ dnorm(n[i-1] + r,tau)}
  for (i in 1:13) { y[i] ~ dpois(exp(n[i]))}
  r ~ dnorm(.2,100)
  tau ~ dgamma(1.0,.1)
}
```

(d) Following on from part (c), the following R code is run. Briefly explain what is being done and the statistical interpretation of the output. [5 marks]

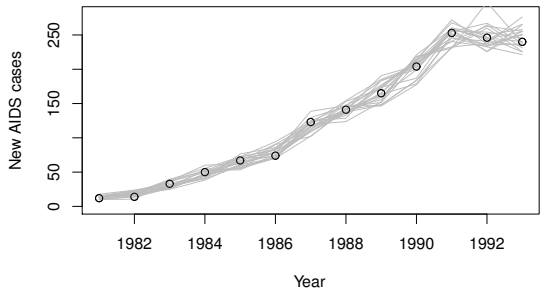
```
plot(re$r,type="l")
```




```
par(mfrow=c(1,2),mar=c(5,5,1,1))
acf(re$r[1,,1],main="");hist(re$r)
```




```
plot(t+1980,y,xlab="Year",ylab="New AIDS cases",ylim=c(0,280))
for (i in 1:20*100+1000) {
  lines(t+1980,exp(re$n[,i,1]),col="grey")
}
points(t+1980,y)
```



(e) Suppose that you want to compare the models from parts (a) and (c) statistically. Briefly explain the approach that might be most appropriate here. [3 marks]

End of examination.

Solutions Theory of Inference Example Exam

1. (a) The basic idea is that the squared Euclidian length of vector $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ is unchanged by rotation/reflection by an orthogonal matrix, so

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{U}\mathbf{y}-\mathbf{U}\mathbf{X}\boldsymbol{\beta}\|^2 = \left\| \begin{pmatrix} \mathbf{U}_f\mathbf{y} \\ \mathbf{U}_r\mathbf{y} \end{pmatrix} - \begin{pmatrix} \mathbf{L}^T \\ \mathbf{0} \end{pmatrix} \boldsymbol{\beta} \right\|^2 = \|\mathbf{U}_f\mathbf{y}-\mathbf{L}^T\boldsymbol{\beta}\|^2 + \|\mathbf{U}_r\mathbf{y}\|^2.$$

   The rightmost expression is clearly minimized by $\hat{\boldsymbol{\beta}} = \mathbf{L}^{-T}\mathbf{U}_f\mathbf{y}$.

   (b) $E(\hat{\boldsymbol{\beta}}) = \mathbf{L}^{-T}\mathbf{U}^f E(\mathbf{y}) = \mathbf{L}^{-T}\mathbf{U}_f\mathbf{X}\boldsymbol{\beta} = \mathbf{L}^{-T}\mathbf{L}^T\boldsymbol{\beta} = \boldsymbol{\beta}$.

   (c) $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L}\mathbf{U}_f\mathbf{y} = \mathbf{L}_f^{-T}\mathbf{U}_f\mathbf{y}$.

   (d) i. The suggestion would produce a linear unbiased estimator of the model that is not the least squares estimator (it is basically a weighted least squares estimate, with weights of zero or one depending on the value of t_i). The Gauss Markov theorem tells us that this can not result in estimators with lower variance than the un-weighted least squares estimator, for this model.

   ii. The bosses suggestion amounts to discarding information — unless there is really something wrong with the model or the discarded information, that can not reduce uncertainty!

   (e) The model being fitted is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i,$$

   where the ϵ_i are i.i.d. $N(0,\sigma^2)$. The plot of residuals against x_i shows a quadratic pattern, so the model should probably be modified to

$$y_i = \beta_0 + \beta_1 x_i \beta_2 x_i^2 + \beta_3 z_i + \epsilon_i.$$

   (f) The plots look ok, except that the residuals seem to have more extreme values than the normal assumption would imply (the residuals are heavy tailed – QQ plot makes this very clear). This is probably not as important as violating the independence or constant variance assumptions would be, but might have some affect. To check, we could try bootstrapping to get confidence intervals for the model coefficients. Here is some example code:

```
n.rep <- 1000
b <- rep(0,n.rep)
dat <- data.frame(y=y,x=x,z=z)
n <- length(y)
for (i in 1:n.rep) {
  ii <- sample(1:n,n,replace=TRUE) ## sample indices of data with replacement
  b[i] <- coef(lm(y~x+I(x^2)+z),data=dat[ii,])[4] ## beta_3 for this rep
}
quantile(b,c(.025,0.975)) ## 95% CI
```

2. (a) The first study is not really suitable for the health departments purposes. The data will be observational data, in which people essentially report what they are doing anyway. It is likely that the study will recruit only health conscious people who will have all sorts of other lifestyle factors influencing their health (which are unlikely to all

be found via online profiling). i.e. there will almost surely be a massive problem with confounding here, and it will not be possible to say much about whether cycling and running cause health improvements. The second study has a much smaller sample size, but properly targets the question of whether increasing exercise will improve health. Randomization of subjects to treatments, with a proper control group, enables inferences about causality to be made. The relatively small sample size might mean that we fail to detect small effects, but the health department probably doesn't want to promote a scheme if the benefits are only small.

(b) A confounding variable is one that is correlated with both the response of interest and covariates of interest. Unobserved confounding variables seriously compromise our ability to infer the causal effect of one variable on another. For example if we are examining the relationship between obesity and cancer rates, there are many confounding variables such as poor diet and poverty that may be associated with both.

(c) If we randomly allocate experimental units (e.g. subjects) to different treatments, then we automatically avoid any possibility of correlation between unobserved covariates and treatment. In effect all variability in the response attributable to other covariates can now be modelled as random variability. Hence there is no problem with hidden confounders.

(d) An instrumental variable is a variable that is correlated with the observed covariates (aka predictor variables) of interest, but is independent of the confounders. It has no direct effect on the response (being correlated with it only because of its correlation with the covariates of interest). Mathematically, we suppose that $\mathbf{y}$ is generated by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_x + \mathbf{H}\boldsymbol{\beta}_h + \boldsymbol{\epsilon}$$

where we have observed $\mathbf{X}$ but not $\mathbf{H}$, and $\mathbf{X}^T\mathbf{H} \neq \mathbf{0}$. If we try to estimate $\boldsymbol{\beta}_x$ by fitting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}$ we will get a biased estimate. Suppose that we have some instrumental variables giving a model matrix $\mathbf{Z}$ (of as high a rank as $\mathbf{X}$). We have $\mathbf{Z}^T\mathbf{H} \simeq \mathbf{0}$ and $\mathbf{Z}^T\mathbf{H} \neq \mathbf{0}$. We then replace $\mathbf{X}$ by a version in which each column of $\mathbf{X}$ is regressed on $\mathbf{Z}$. That is we replace $\mathbf{X}$ by $\mathbf{AX}$ where $\mathbf{A} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})\mathbf{Z}^T$, and fit the model $\mathbf{y} = \mathbf{AX}\boldsymbol{\beta}_x + \boldsymbol{\epsilon}$ to get

$$\hat{\boldsymbol{\beta}}_x = (\mathbf{X}^T\mathbf{AX})^{-1}\mathbf{X}^T\mathbf{Ay}$$

Taking expectations:

$$E(\hat{\boldsymbol{\beta}}_x) = (\mathbf{X}^T\mathbf{AX})^{-1}\mathbf{X}^T\mathbf{AX}\boldsymbol{\beta}_x + (\mathbf{X}^T\mathbf{AX})^{-1}\mathbf{X}^T\mathbf{AH}\boldsymbol{\beta}_h \simeq \boldsymbol{\beta}_x$$

since $\mathbf{AH} \simeq \mathbf{0}$ follows from $\mathbf{Z}^T\mathbf{H} \simeq \mathbf{0}$.

3. (a) Taylor expansion around the MLE gives

$$\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\hat{\theta}} \simeq \left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t} + \left.\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\theta_t} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)$$

with equality in the large sample limit, for which $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \to 0$. From the definition of $\hat{\boldsymbol{\theta}}$, the left-hand side is $\mathbf{0}$. So assuming $\boldsymbol{\mathcal{I}}/n$ is constant (at least in the $n \to \infty$ limit), then as the sample size tends to infinity,

$$\frac{1}{n}\left.\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\theta_t} \to -\frac{\boldsymbol{\mathcal{I}}}{n}, \quad \text{while} \quad \left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t}$$

is a random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\mathcal{I}}$. Therefore in the large sample limit,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \sim \boldsymbol{\mathcal{I}}^{-1} \left. \frac{\partial l}{\partial \boldsymbol{\theta}} \right|_{\theta_t},$$

implying that $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = 0$ and (using standard results on transformation of covariance matrices) var $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = \boldsymbol{\mathcal{I}}^{-1}$.

(b) The CR lower bound says that $\boldsymbol{\mathcal{I}}^{-1}$ is the lower bound on the covariance matrix of an unbiased estimator. So in the large sample limit MLEs are minimum variance unbiased estimators.

(c) It is not, since we can not reasonably use a Taylor expansion about a point at the edge of the parameter space in this way.

(d) Newton's method operates by successively maximizing quadratic approximations to the log likelihood, where the quadratic approximation is based on the second order Taylor expansion at the current best estimate of the MLE. Hence at convergence Newton's method automatically provides the observed version of $\boldsymbol{\mathcal{I}}$, which can then be used for (large sample approximate) interval estimation for $\boldsymbol{\theta}$.

(e) i.
```
l1 <- function(b,calls)
{ -sum(log(dpois(calls,b)))
}
```

ii.
```
l2 <- function(b,calls,sales)
{ b<-exp(b)
  E.calls <- b[1]+b[2]*sales
  -sum(log(dpois(calls,E.calls)))
}
```

iii. The code maximizes the likelihood of both models, computed a GLRT statistic and corresponding p-value. This strongly suggests that the second model is better than the first.

iv. The null model is restricting a parameter of the alternative to the edge of the feasible parameter space, invalidating the distributional result. Either shift to using AIC here, or re-formulate the second model without the restriction on the slope parameter.

4. (a) Obviously solution is not unique. Here is an acceptable one.

```
lf <- function(theta,t,y) {
## log joint density of parameters and data
## initial pop prior is dgamma(.,4,.4) r prior is N(.2,.1^2)
  if (theta[1]<=0) return(-Inf)  ## zero prob according to prior
  mu <- theta[1]*exp(t*theta[2]) ## expected number of cases
  sum(dpois(y,mu,log=TRUE)) + dgamma(theta[1],4,.4,log=TRUE) +
                              dnorm(theta[2],.2,.1)
} ## lf

n.rep <- 100000 ## chain length
th <- matrix(0,2,n.rep) ## 2 row matrix to hold n_0 and r sims
th[,1] <- c(12,.2) ## initial values
ll0 <- lf(th[,1],t,y) ## initial log joint density
```

```
psd <- c(1,.01) ## proposal standard deviations (needs tuning)
accept <- 0 ## acceptance counter

for (i in 2:n.rep) { ## MH loop
  th[,i] <- th[,i-1] +rnorm(2)*psd ## proposal
  ll1 <- lf(th[,i],t,y)  ## joint density of proposal
  if (runif(1)<exp(ll1-ll0)) { ## MH accept
    ll0 <- ll1;accept <- accept + 1
  } else { ## reject
    th[,i] <- th[,i-1]
  }
} ## MH loop end
accept/n.rep ## acceptance rate
```

(b) Plot simulated n_0 and r against iteration number, to get an impression of convergence and degree of autocorrelation. Plot ACFs for r and n_0 to examine the autocorrelation length of the chains. Some adjustment of the proposal might be appropriate if there are high correlations at long lags (high 'correlation length').

(c) The model is

$$Y_t \sim \text{Poi}(N_t), \quad N_t = N_{t-1}e^{r+z_t}, \quad z_t \sim N(0, \tau)$$

($\tau = \sigma^{-2}$ is a precision parameter). A gamma prior is used for τ and a Gaussian prior for r, while (oddly) $\log N_1$ is given a log normal distribution.

(d) Firstly the chain for r is examined. Convergence appears very quick, and mixing is good (autocorrelation appears low in this plot). Next an ACF for r is plotted, confirming low auto-correlation. The histogram then shows the approximate posterior distribution of r (mostly lying between 0.1 and 0.35). Finally 20 example draws from the posterior distribution of N are overlaid on the raw case data. These latter plots are plausible, but it is possible that the model is a bit too variable in the later stages.

(e) Ideally we might look at the Bayes factor, but this would be problematic here since we have been told that the priors in (c) are chosen to be uninformative, rather than being precise statements of prior knowledge. BIC is difficult to compute in this case (we would need to integrate out the N_ts). So probably DIC is the least bad option for a formal criterion.