12K Statistical Modelling

For 31 days after the outbreak of the 2014 Ebola epidemic, the World Health Organization recorded the number of new cases per day in 60 hospitals in West Africa. Researchers are interested in modelling Y_{ij}, the number of new Ebola cases in hospital i on day $j \geqslant 2$, as a function of several covariates:

- `lab`: a Boolean factor for whether the hospital has laboratory facilities,
- `casesBefore`: number of cases at the hospital on the previous day,
- `urban`: a Boolean factor indicating an urban area,
- `country`: a factor with three categories, Guinea, Liberia, and Sierra Leone,
- `numDoctors`: number of doctors at the hospital,
- `tradBurials`: a Boolean factor indicating whether traditional burials are common in the region.

Consider the output of the following R code (with some lines omitted):

```
> fit.1 <- glm(newCases~lab+casesBefore+urban+country+numDoctors+tradBurials,
+ data=ebola,family=poisson)
> summary(fit.1)
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.253106   0.046473   5.446 5.14e-08 ***
labTRUE              0.024735   0.053463   0.463  0.64361
casesBefore          0.308299   0.007193  42.862  < 2e-16 ***
urbanTRUE           -0.027651   0.086295  -0.320  0.74865
countryLiberia       0.102070   0.033455   3.051  0.00228 **
countrySierra Leone -0.212426   0.036551  -5.812 6.18e-09 ***
numDoctors          -0.025615   0.004514  -5.675 1.39e-08 ***
tradBurialsTRUE     -0.007499   0.030669  -0.245  0.80684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Would you conclude based on the z-tests that an urban setting does not affect the rate of infection?

(b) Explain how you would predict the total number of new cases that the researchers will record in Sierra Leone on day 32.

We fit a new model which includes an interaction term, and compute a test statistic using the code:

```
> fit.2 <- glm(newCases~casesBefore+country+casesBefore:country
+numDoctors,data=ebola,family=poisson)
> fit.3 <- glm(newCases~lab+casesBefore+urban+country
+numDoctors,data=ebola,family=poisson)
> fit.2$deviance-fit.1$deviance
[1] -1.946486
> fit.3$deviance-fit.1$deviance
[1] 0.05980278
```

(c) For each statistic computed by the last two commands, say whether it is possible to derive an asymptotic distribution through Wilks' theorem, and if so, specify it.

(d) Under what conditions is the deviance of each model approximately chi-squared?