STOR 556: Spring 2019
Take-home Final Exam With Grade Scheme

Answer all questions.

This is a take-home exam that you are expected to do in your own time and hand in no later than **Noon Thursday May 2**. The exam should be submitted via the "Assignments" tab of the course sakai page.

Rules of the Exam. All course resources including text, personal notes and resources available through R or R-Studio are permitted. Your submitted answers should include full verbal answers to the questions, illustrated where appropriate by R code, tables or figures. Very long-winded answers are discouraged; greatest credit will be given for full but concise answers to the questions. Solutions may be submitted in R-Markdown but this is not required. (A fully acceptable alternative is if you submit a Word document into which you cut and paste R output as appropriate; however, I recommend you "save as" a pdf file for the final submission.) Other web resources may be used if fully acknowledged and referenced. Discussion among yourselves or with an outside party is not permitted; you are allowed to email the instructor if you find the question ambiguous or if you think there is an error, but the instructor will not give advice how to solve the problems.

The datasets have been previously posted (see the "Resources" tab in sakai for instructions how to download them); please download the data **first** and contact the instructor immediately if you have any problem with this step.

Please acknowledge you accept these conditions by copying out and signing:

PLEDGE: I will neither give nor receive unauthorized aid in this exam.

SIGNED: (A typed signature will be accepted)

1. The "dolphins" dataset documents number of dolphin bycatch (dolphins caught by accident in fishing nets) over six seasons (1989/90 through 1994/95) in two areas in New Zealand (North Taranaki and South Taranaki), for two types of gear (Bottom and Midwater) and for both day and night trawls. For each combination of season, area, gear type and time of day, the dataset documents the number of tows (fishing trawls observed) and number of dolphin bycatch. Five of the possible 48 rows of data are absent because the number of tows was 0.

   (a) A reasonable model is that the number of bycatch $y_{ijk\ell}$ in season i, area j, type of gear k and time of day ℓ is

   $$\begin{aligned} y_{ijk\ell} &\sim \text{Poisson}(\lambda_{ijk\ell} T_{ijk\ell}), \\ \log \lambda_{ijk\ell} &= \beta_0 + \beta_{S_i}\ \text{Season}_i + \beta_{A_j}\ \text{Area}_j + \beta_{G_k}\ \text{Gear}_k + \beta_{T_\ell}\ \text{Time}_\ell \end{aligned} \qquad (1)$$

   where $T_{ijk\ell}$ is the number of tows and each of the variables Season, Area, Gear and Time is treated as a factor variable, β_0 is the intercept and each of β_{S_i}, β_{A_j}, β_{G_k} and β_{T_ℓ} represents a regression coefficient for the corresponding variable. Fit the model (1) and display a table showing the regression coefficients and their standard errors. **[6 points]**

   (b) Can any of the terms from the model (1) be dropped? Test each of the terms in turn, using an appropriate χ^2 or F test, and state your conclusions. **[5 points]**

(c) Do the data show evidence of overdispersion? Use appropriate diagnostics and tests, and if necessary, repeat your calculations of part (b). **[5 points]**

(d) Would the model (1) be improved by adding interactions? Test, in particular, interactions among the Area, Gear and Time variables, and state your conclusions. **[4 points]**

(e) It can be seen that in one season (1990/91), there were no dolphin bycatch. To what extent does that year's data bias the conclusions? Refit the model without 1990/91, and describe any significant features that change. **[5 points]**

(f) For whatever model you finally accepted based on parts (a)–(e), investigate the data for (i) non-random patterns among the residuals, (ii) datapoints of high leverage, (iii) points of high influence. Overall, would you describe this as a satisfactory analysis, or if not, why not? **[7 points]** (**[32 points for the whole question]**)

2. The "glucose" dataset records the blood glucose levels in six subjects at various times before or after a test meal, starting 15 minutes before the meal and ending 6 hours later. There are six separate runs (using the same six subjects on different days), corresponding to different times of day that the meal is taken (6am, 10am, 2pm,...,2am). This is an example of a "repeated measures experiment" in which multiple measures are taken for the same subject at different times. It is expected that the pattern of responses will be different from one run to another, and there may also be a variation from one subject to another.

(a) For each of the six runs A–F, draw a trace plot that shows the pattern of glucose responses in all six subjects for that run. **[5 points]**

(b) Draw a second set of trace plots where you show the response of each subject in each run using one of a (i) linear, (ii) quadratic, (iii) cubic or (iv) quartic regression through all ten time points (in other words, including powers of time t up to t^4 in the case of quartic regression). State which one of these gives the best representation, and briefly justify your choice (formal tests, confidence intervals, etc. are not required). **[5 points]**

(c) Recast the dataset as a 360×4 matrix where the columns represent all the glucose levels, times ($t = -15,\ 0,\ 30, ..., 360$), subjects (1 through 6) and runs (A through F). (*Hint:* You may need some R command such as `y=as.vector(as.matrix(glucose[,3:12]))` to write all the glucose levels as a single column vector.) **[5 points]**

(d) Now try fitting the data as a single regression model where the covariates are (i) powers of t up to t^4 (you may need to rescale for numerical stability), (ii) Subject treated as a random effect, (iii) Run treated as a fixed effect. Check for interactions between t (or powers of t) and either Run or Subject, and don't forget that the variance of a random effect can under some circumstances be estimated as 0. What do you conclude? Where appropriate, use tests of hypotheses (such as the Kenward-Roger test, or a bootstrap test) to decide between different nested models. **[13 points]**

(e) Summarize your conclusions, with particular attention to whether there is evidence that patterns of glucose levels vary among the six Runs. **[5 points]** (**[33 points for the whole question]**)

3. The "indonesia" dataset records a number of outcomes from a children's health study in Indonesia. Variables include an ID number for each child (repeated up to 6 times), indicators of respiratory disease, xerophthalmia (used to mark vitamin A deficiency), age (in months, centered about 36 months), sex, season, height adjusted for age, and an indicator of stuntedness. There is also an age group variable "agegp" that groups the ages into four groups. Respiratory disease (coded 0 or 1) is the primary outcome of interest, and we are interested in how each of the other variables affects it. Since each child appears in the dataset multiple times, we must account for correlated observations by using either a random effects model or a generalized estimating equations approach.

   (a) Construct a 2×2 table relating incidence of respiratory disease to concurrent incidence of xerophthalmia. Repeat the same construction separately for each of the four age groups. What pattern do you notice? Is this an example of Simpson's paradox? **[7 points]**

   (b) Create a random effects model using the `glmer` command to related incidence of respiratory disease each of the other variables, treating `ID` as a random effect to allow for systematic variation from one child to another. (Don't include "agegp" in this analysis, since age is already included as a covariate.) Which variables are significant? **[5 points]**

   (c) Ae there alternative models that are superior to the omdel in (b)? Consider, in particular, (i) adding square or cubic terms in age, (ii) omitting any of the other variables that may not be significant. Summarize your conclusions. **[5 points]**

   (d) Based on your results to parts (b) and (c), what do you now say about the relationship between xerophthalmia and respiratory disease? **[3 points]**

   (e) For whatever model you previously decided was best, test the fit of the model using the Hosmer-Lemeshow test. What do you conclude from that? **[3 points]**

   (f) For whatever model you previously decided was best, try an alternative fit using (i) the PQL method, (ii) the generalized estimating equations approach, using your own judgement (or trial and error) to decide which correlation structure is appropriate. Do any of your conclusions change? **[6 points]**

   (g) Comment on the dependence of respiratory disease on age, using suitable plots to illustrate your conclusions. **[6 points]** (**[35 points for the whole question]**)

Solutions

Note: As with homework solutions produced during the course, these are not meant to be the unique "right answers" to the questions. Your answers may be different from mine in a number of ways and still be correct solutions.

1. (a) The idea is to treat $\log T_{ijk\ell}$ as an offset. Some sample code is as follows:

```
dolph=read.table('.../dolphins.txt',header=T)
m1=glm(Bycatch~offset(log(Tows))+Season+Area+Gear+Time,family=poisson,data=dolph)
summary(m1)
```

This produces the following (edited) output:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.32782    0.59038 -12.412  < 2e-16 ***
Season1990/91  -19.66524 2439.00216  -0.008   0.9936
Season1991/92    1.81895    0.28503   6.382 1.75e-10 ***
Season1992/93    0.07391    0.39557   0.187   0.8518
Season1993/94   -0.93191    0.41132  -2.266   0.0235 *
Season1994/95   -0.13503    0.30923  -0.437   0.6623
AreaSouth        1.82245    0.41063   4.438 9.07e-06 ***
GearMidwater     1.91769    0.44308   4.328 1.50e-05 ***
TimeNight        2.17656    0.45138   4.822 1.42e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 335.096  on 42  degrees of freedom
Residual deviance:  42.078  on 34  degrees of freedom
AIC: 97.461
```

It looks as though each of the terms is statistically significant, though in the case of "Season," this is heavily influenced by the result for 1991/92.

(b) The command `drop1(m1,test='Chi')` leads to the following output:

```
       Df Deviance     AIC     LRT  Pr(>Chi)
<none>      42.078  97.461
Season  5  147.187 192.570 105.110 < 2.2e-16 ***
Area    1   71.303 124.685  29.225 6.446e-08 ***
Gear    1   68.061 121.443  25.983 3.445e-07 ***
Time    1   75.992 129.375  33.914 5.759e-09 ***
```

All terms are highly significant. Note that the alternative `drop1(m1,test='F')` generates somewhat larger p-values and the warning `F test assumes 'quasipoisson' family` — in other words, better to use `test='Chi'` at this point.

(c) We can compute sums of squares of either deviance or Pearson residuals, for example

```
sum(residuals(m1,type='deviance')^2)
```

produces deviance 42.1 while

```
sum(residuals(m1,type='pearson')^2)
```

produces $X^2 = 54.6$. The `df.residual` for this model is 34, which implies p-values of 0.16 and 0.014 when tested against the χ^2_{34} distribution (e.g. `pchisq(54.6,34,lower=F)`. So the Pearson X^2 statistic is statistically significant but the deviance statistic is not. This provides some (not conclusive) evidence that the distribution is overdispersed and we should estimate the overdispersion parameter to be $54.6/34 = 1.61$. Alternatively, repeat the model fit in (a) with `family=quasipoisson`. The previously quoted parameter estimates do not change but all the standard errors are increased by $\sqrt{1.61}$ which, in this case, is not sufficient to overturn our previous conclusions about statistical significance. (As noted at the end of part (b), you could also apply `drop1` with `test='F'`, which assumes the quasipoisson model, but all the variables are still strongly significant.)

(d) If you try each of the possible interaction terms individually, the only one that is statistically significant is `GearMidwater:TimeNight` which produces a negative coefficient and a p-value of 0.022. Also (the question didn't ask you to do this, but you will earn bonus points if you did) if you now repeat the Pearson X^2 calculation for overdispersion, you will find that it is not statistically significant ($X^2 = 46.1$, p=0.064). Therefore, the best from among these models seems to be the original model (1), with an additional `Gear:Time` interaction. This model is used as the basic model for the following parts (e) and (f), though if you had a different model at this stage of the question, you should still get similar answers to (e) and (f).

(e) Actually, it makes very little difference. One way to do it is some command such as

```
dolph1=dolph[dolph$Season!="1990/91",]
```

and then repeat the best of the preceding model fits using the dataframe `dolph1`. Many of the regression coefficients do not change at all and there are only slight changes in their standard errors.

(f) There are various plots you could raw and I only show one here for illustration (Figure 1). This one shows two highly influential observations, numbers 20 and 7, which also have the two highest numbers of bycatch (16 and 23 respectively). Other diagnostics also give reasons to question the suitability of the model but I think the main reason for that is that with two scenarios containing such large numbers of bycatch, and most of the others containing none, there is no model that would fit the data completely satisfactorily.

*Comments on student answers. A lot of students missed the need for an offset in part (a) — some students simply put in log Tows as a regular covariate, some put in Tows without the log transform (this doesn't work well, because the glm command by default uses a log link) and some neglected Tows altogether. Parts (b) and (c) were generally answered well, though a lot of students didn't try to make a formal test for overdispersion. For (d), some students tried the three-level interaction Area*Gear*Time, but on realizing*

that that model did not work well because of overfitting, did not explore the individual interactions further. In (f), a lot of students noticed there was a problem but didn't give a clear explanation why. One suggestion that a couple of students made was to try a zero-inflated Poisson model — I actually think this would be a good idea but I haven't tried it out.

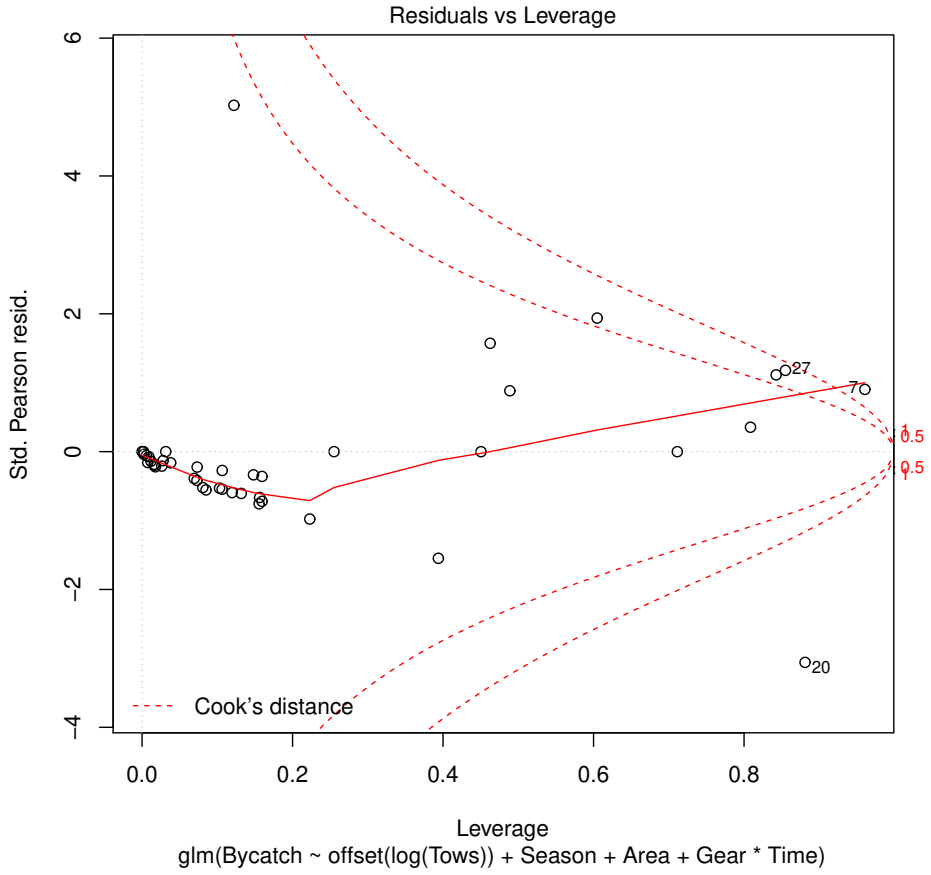

Figure 1: Influence and Leverage Plot for Model of Question 1(d)

2. (a) See Figure 2.

   (b) See Figure 3. These plots were drawn by simply fitting a quartic regression through the 10 time points for each of 36 independent datasets (six individuals in each of six runs). You would get similar-looking plots with a cubic regression. The agreement with Figure 2 is reasonable but could perhaps be improved (e.g. using splines instead of polynomial regression).

   (c) Obviously there are many ways to do this but one possible sequence is given at the top of page 8, along with the 15 lines of the resulting data frame given by `Y[1:15,]`.

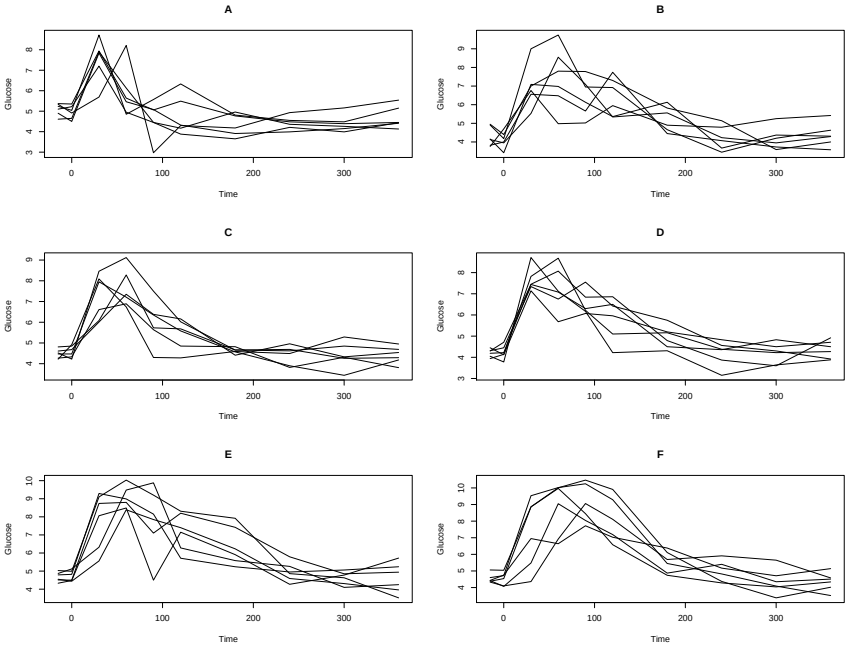

Figure 2: Trace Plots for Question 2(a)

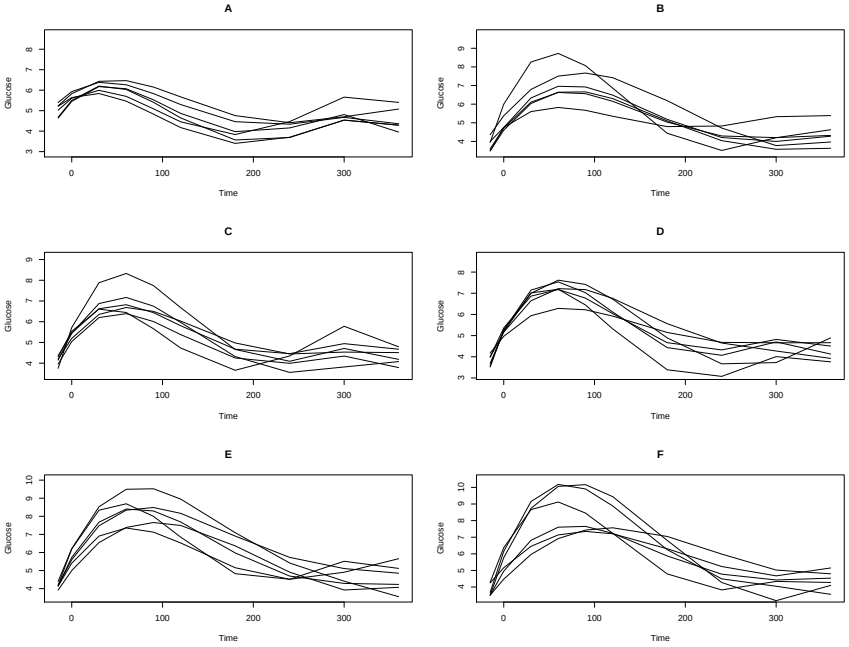

Figure 3: Trace Plots for Question 2(b)

```
y=as.vector(as.matrix(t(gluc[,3:12])))
t1=rep(c(-15,0,30,60,90,120,180,240,300,360),36)
Run=c(rep('A',60),rep('B',60),rep('C',60),rep('D',60),rep('E',60),rep('F',60))
Sub=factor(rep(c(rep(1,10),rep(2,10),rep(3,10),rep(4,10),rep(5,10),rep(6,10)),6))
Y=cbind(y,t1,Run1,Subj)
```

First few lines of the resulting data frame:

```
      y      t1    Run1 Subj
 [1,] "4.9"  "-15" "A"  "1"
 [2,] "4.5"  "0"   "A"  "1"
 [3,] "7.84" "30"  "A"  "1"
 [4,] "5.46" "60"  "A"  "1"
 [5,] "5.08" "90"  "A"  "1"
 [6,] "4.32" "120" "A"  "1"
 [7,] "3.91" "180" "A"  "1"
 [8,] "3.99" "240" "A"  "1"
 [9,] "4.15" "300" "A"  "1"
[10,] "4.41" "360" "A"  "1"
[11,] "4.61" "-15" "A"  "2"
[12,] "4.65" "0"   "A"  "2"
[13,] "7.9"  "30"  "A"  "2"
[14,] "6.13" "60"  "A"  "2"
[15,] "4.45" "90"  "A"  "2"
```

(d) There are many possible models you could try but a good sequence would be

```
library(pbkrtest)
m0=lmer(y~(t1+I(t1^2)+I(t1^3)+I(t1^4))+(1|Sub))
m1=lmer(y~(t1+I(t1^2)+I(t1^3)+I(t1^4))+Run+(1|Sub))
m2=lmer(y~(t1+I(t1^2)+I(t1^3)+I(t1^4))*Run+(1|Sub))
KRmodcomp(m1,m0)
KRmodcomp(m2,m1)
```

which fits a random effects model without Run (m0), with Run as an additive effect (m1), and with the full Run $\times$ nonlinear Time interaction (m2). Both tests result in a p-value of $< 10^{-9}$ so there is very strong evidence that the Run and Time interaction is real. However, all three models give the warning `singular fit` and if you look at the output you can see that the random effect for Subject has variance 0 (as the question warned you might happen). So, it seems, there is no evidence for a Subject effect, but some of you did not stop there and looked for a Subject $\times$ Time interaction — two further tests you could do are

```
m3=lmer(y~(t1+I(t1^2)+I(t1^3)+I(t1^4))*Run+(1+t1+I(t1^2)+I(t1^3)+I(t1^4)|Sub))
PBmodcomp(m3,m1)
m4=lmer(y~(t1+I(t1^2)+I(t1^3)+I(t1^4))*Run+(1+t1|Sub))
PBmodcomp(m4,m1)
```

The first of these (test for a nonlinear Time interaction) results in a p-value of about 0.16 (after about 2 hours running on my laptop — this is a slow model to do a bootstrap test because the full 5-dimensional random effect model has to be re-estimated on each bootstrap sample). The second test (linear Time interaction) produced a p-value of 0.009 suggesting that this interaction *is* statistically significant though it's a little hard to interpret. (In theory you could also test for a Subject × Run interaction — it's conceivable that some people would be more susceptible than others to indigestion following a late-night meal! — but I didn't test this and I don't think anyone in the class did either.)

So, in summary, I think the key points are:

- You definitely need a nonlinear Time effect that could be modeled with either a cubic or quartic polynomial (quartic is preferred);
- There is a Run effect and in fact the Run × Time interaction is significant so long as you include higher powers of Time;
- There doesn't appear to be a Subject random effect though this is a little ambiguous given the finding of a significant interaction between Subject and linear Time.

(e) See Figure 4. There appears to be no statistically significant difference among subjects, but there is a clear difference among the six runs, with peak glucose levels being lowest in run A and highest in runs E and F.

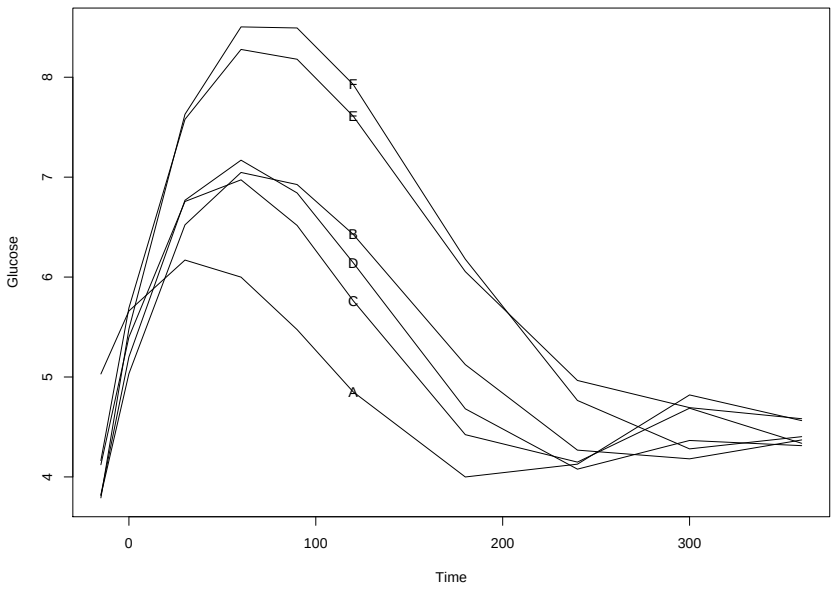

Figure 4: Trace Plots from Final Model

Note about student solutions. A lot of students apparently didn't know about the need for the `I(...)` *notation when including models with higher-order terms — either that, or define separate variables to represent time*^{2}*, time*^{3}*, etc. Even if you didn't know this in advance, it should have been very obvious that the model output was not doing what you wanted —*

there's no point handing in page after page of R model output if you've never looked and tried to interpret it!

3. (a) The overall 2×2 table is

		Xerophthalmia	
		No	Yes
Respiratory Disease	No	1045	48
	Yes	100	7

which shows a respiratory disease rate of $\frac{100}{1045+100} = 0.087$ among the no-Xerophthalmia group and $\frac{7}{48+7} = 0.127$ among the Xerophthalmia group, which seems to indicate higher respiratory disease among the Xerophthalmia group (the original test hypothesis). However if we recalculate the table for each of the four age groups we get

		Xerophthalmia							
		Age Group 1		Age Group 2		Age Group 3		Age Group 4	
		No	Yes	No	Yes	No	Yes	No	Yes
Respiratory Disease	No	254	1	250	13	271	13	270	21
	Yes	27	0	49	4	14	3	10	0

which shows a zero rate of respiratory disease in the xerophthalmia group for age groups 1 and 4 (contradicting the test hypothesis). It's not an example of Simpson's paradox because the respiratory disease rate is still higher in the xerophthalmia group for groups 2 and 3, but it does illustrate that there is a clear age dependence in pattern of responses so this does support taking age into account.

Note about student solutions. Many students calculated the conditional probabilities the wrong way round, as the probability of xerophthalmia given respiratory disease. It leads to the same conclusion regarding Simpson's paradox, but the question very clearly stated that we were interested in xerophthalmia as a possible causal factor in respiratory disease and not than the other way round. Also, it's only really Simpson's paradox if all four of the within-age-group comparisons go the opposite way to the aggregated comparison — that's why I say it's not an example of Simpson's paradox, but the underlying point, that it's better to include the age group variable when this is relevant, is true here and was one of the intended points of the question.

(b) A simple form of the `glmer` command including all the variables as linear fixed effect, plus a single random effect, is

```
m1=glmer(resp~sex+height+costime+sintime+xero+age+stunted+(1|ID),
family=binomial,data=indonesia)
```

and leads to the following table of fixed effects coefficients:

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.723773   0.233129 -11.684  < 2e-16 ***
sex         -0.441407   0.264398  -1.669 0.095022 .
height      -0.048634   0.027142  -1.792 0.073152 .
```

```
costime     -0.592070   0.174497  -3.393 0.000691 ***
sintime     -0.162792   0.174975  -0.930 0.352178
xero         0.601636   0.486438   1.237 0.216154
age         -0.033919   0.007443  -4.557 5.18e-06 ***
stunted      0.206444   0.449154   0.460 0.645783
```

though you also get the `Model failed to converge` warning message as one student pointed out to me in a query that I relayed to the class. However, in this case the results seem to make sense despite the warning. It indicates that `costime` (a measure of the time of year) and `age` are clearly significant, the others do not appear to be. For a test of the random effect, you could try `confint(m1)` which produces the output:

```
Computing profile confidence intervals ...
                  2.5 %       97.5 %
.sig01       0.41730560  1.332254146
(Intercept) -3.22334563 -2.302168349
sex         -0.97824582  0.071594910
height      -0.10423258  0.002802248
costime     -0.94414705 -0.257987603
sintime     -0.51277725  0.175696904
xero        -0.42859310  1.502869830
age         -0.04919991 -0.019765885
stunted     -0.69527031  1.081721165
```

The non-zero lower bound for `.sig01` confirms the statistical significance of this variable.

(c) After some trial and error, I selected a model that included age squared as well as age, but omitted the terms `xero` and `stunted`. I didn't omit `sintime` because this variable and `costime` are parallel variables that should be included or excluded together. This led to the model

```
m1=glmer(resp~sex+height+costime+sintime+age+I(age^2)+(1|ID),
family=binomial,data=indonesia)
```

with results

```
Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.2476649  0.2474792  -9.082  < 2e-16 ***
sex         -0.5384301  0.2605235  -2.067 0.038760 *
height      -0.0457808  0.0224814  -2.036 0.041711 *
costime     -0.5927228  0.1734100  -3.418 0.000631 ***
sintime     -0.1514486  0.1731264  -0.875 0.381690
age         -0.0298284  0.0080055  -3.726 0.000195 ***
I(age^2)    -0.0010727  0.0004067  -2.638 0.008349 **
```

The results show that the `age` variable is significant at least as far as the quadratic term; `sex` and `height` appear to have a mild effect. *(Many students omitted* `sintime` *as well and that's OK — I don't see a real objection to that. Some students also included age*^{3} *as a covariate — the statistical significance of this is less clear but I didn't take points*

off for anyone who did that. As in question 2(d), some students didn't understand the need for `I(..)` *with the age-squared and age-cubed terms.)*

(d) None of the models fitted in part (b) show a statistically significant effect for xerophthalmia so we conclude this variable does not significant affect respiratory outcomes. (This may be too hasty a conclusion: the number of xerophthalmia cases in the dataset is quite small, only about 4.6% of all cases, and the sample may not be large enough to establish this effect.)

(e) The Hosmer-Lemeshow test does not show a good result; for example, for the last model fitted

```
library(ResourceSelection)
hoslem.test(indonesia$resp,fitted(m1), g = 20)
hoslem.test(indonesia$resp,fitted(m1), g = 50)
```

the first test leads to a p-value of $\approx 10^{-6}$ and second one a p-value of 0.009. I don't know exactly what the problem is: possibly the model is still not fitting the age effect accurately enough. *These particular p-values are specific to the model in part (c), but if you used a different model from that, you should get similar p-values.*

(f) As an example of the GEE and PQL methods, I tried the model fits

```
library(geepack)
modgeep=geeglm(resp~sex+height+costime+sintime+age+I(age^2)+xero,id=ID,
family=binomial,corstr='ar1',data=indonesia)
summary(modgeep)
library(MASS)
modpql=glmmPQL(resp~sex+height+costime+sintime+age+I(age^2)+xero,
random=~1|ID,family=binomial,data=indonesia)
summary(modpql)
```

where in both cases I restored the variable `xero` to get an alternative test of this component, but none of the significant results changed: in both cases, `xero` is not significant but the dependence on age and age^2 clearly is.

The Hosmer-Lemeshow test seems to accept the fit for the GEE model but not for the PQL model; for example,

```
hoslem.test(indonesia$resp,fitted(modgeep), g = 20)
hoslem.test(indonesia$resp,ilogit(fitted(modpql)), g = 20)
```

produces a p-value 0.2 in the first case and 0.003 in the second. *However: some students tried the second test without the* `ilogit` *transformation and got a negative value of X^2 and apparent p-value of 1! This is another example of the need to always look at the output critically.*

(g) A grouping of age into 25 categories and corresponding means of the fitted respiratory disease rates is shown in Figure 5, together with a fited curve from the quadratic model of part (c). There is a clear peak of respiratory disease around age –15 (21 months in actual age) after which it declines.

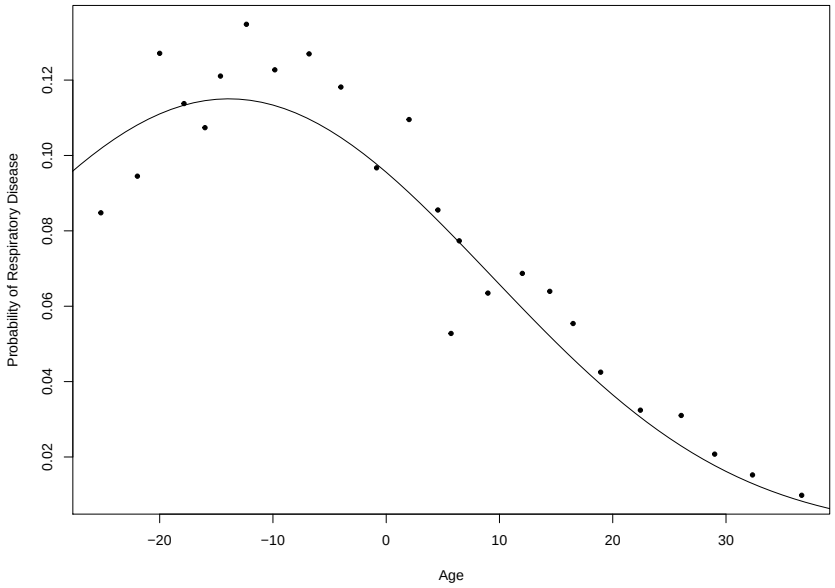

Figure 5: Grouped Means for Age and Respiratory Disease and Fitted Curve

Many other kinds of plots were used in this question and I accepted these so long as they did address the question of interest.