STOR 590: Spring 2020
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 **6:00 pm Thursday April 30**. 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 are posted under the "Resources" tab in sakai. Please download the data **first** and contact the instructor immediately if you have any problem with this step. The datafiles are all in "csv" format and can be loaded into R by typing a command of the form `soappads=read.csv('.../soappads.csv')` where you should insert your own path to identify the file.

If you don't understand or can't do one part of a question, feel free to attempt the later sections first and then go back to the one you skipped. There is no requirement to do the part-questions in the order they appear on the exam. If you feel that some part of a question does not have a clear-cut answer, give your best interpretation.

Please acknowledge you accept the 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 "soappads" dataset contains the result of an experiment to evaluate the quality of soap pads which differ in three respects: amount of detergent (d), coarseness (c) and solubility (s). Each has two possible levels resulting in eight possible treatment combinations labelled 0, d, c, s, dc, ds, cs, dcs according to which variables are applied in which specimen (0 means all three are at their base level). The experiment is conducted over four replicates (1 through 4), two days (1 or 2) and 16 judges (1–16). Each judge evaluates two types of soap pad on each of two days, and gives them a score of 1 through 5, where 1 is best. The ultimate objective of the study is to decide which of the eight possible treatments is best. The data is in the file "soappads.csv".

   (a) For each of the eight possible treatments, calculate the mean score given by the judges. Which treatment(s) come out best by this measure? **[4 points.]**

(b) A possible (fixed effects) model would be to take each of the variables `Judge`, `Replicate`, `Day` and `Treat` as factor variables, and fit a linear regression with `Score` as a continuous (normally distributed) response. Explain why that method would not work. **[4 points.]**

(c) Now consider the variant on part (b) where we drop `Replicate` and `Day` and consider `Judge` and `Treat` as fixed-effects factor variables. Fit the resulting analysis of variance model and say which treatment now shows up best in the sense of minimizing the mean score. Explain in words why this gives a different result from (a). **[6 points.]**

(d) Suppose in part (c) we treat `Judge` as a random effect instead of a fixed effect. Fit the model under this assumption and again say how to interpret the result in terms of which treatment is best. **[6 points.]**

(e) Now consider the alternative viewpoint in which a score of 1 or 2 is considered "Success" and any other score a failure. Refit the model of (c) as a logistic regression model and state which of the eight possible treatments gives the largest probability of a successful result. What difficulties arise in applying this model? **[6 points.]**

(f) Now return to your answer from (d) and state (i) whether you think the treatment you selected as best is indeed better than the others, and (ii) whether there is any statistically significant difference among the eight treatments. State clearly what assumptions or statistical tests you are using to support your answer. **[7 points.]**
[33 points for the whole question.]

2. The "spruce" dataset documents the growth of 79 spruce trees divided among four chambers, labelled 1, 2, 3, 4. Two of the chambers (1 and 2) have a high ozone environment, marked by `tx=1`, and the other two a low-ozone environment (`tx=0`). The `y` variable is the size of the tree (the logarithm of an estimate of total tree volume) and the `day` variable marks the day within the year (all measurements are taken during the summer of a single year). The variable `id` (values 1–79) is a indicator of which tree is which.

(a) Draw a line plot that shows the growth of each tree against time, in two panels where one panel represents the high-ozone environment and the other panel represents the low-ozone environment. Based on the plot, would you say that either environment is beneficial to (i) the overall size, (ii) the rate of growth, of a tree? **[4 points.]**

(b) Analyze the data using an ordinary linear regression, using `day`, `tx` and `chamber` as covariates, but ignoring the fact that there are repeated measures on each tree. Decide which (if any) of these variables should be treated as factor variables, and also which (if any) interactions are appropriate. Use the results of your analysis to answer the questions posed in part(a), and also draw suitable plots to illustrate (i) how well the model fits the data, and (ii) whether the model fit could be improved by a Box-Cox transformation. Summarize your conclusions. **[10 points.]**

(c) Now do a random effects analysis in which `id` is nested within `chamber`, both as random effects. Repeat the analysis of (b), calculating the estimates and standard errors of the fixed-effect terms that depend on `tx`. How do your conclusions compare with those of part (b)? **[7 points.]**

(d) Use the `PBmodcomp` and `exactRLRT` functions to test for the statistical significance of the two random effects, and report your conclusions. **[5 points.]**

(e) Using whichever random effects model you consider appropriate based on (d), conduct a more formal test of the statistical significance of the `tx` effects using a Kenward-Roger test. What is your conclusion? **[3 points.]**

(f) Combining all your answers to the preceding parts, what would you say are the advantages or disadvantages of doing a random effects analysis for this dataset? **[4 points.]**
[33 points for the whole question.]

3. The "schiz" dataset concerns the progress of schizophrenia patients. There are five variables: `ID` is the patient id, `Y` is the symptom indicator (1 means symptoms observed, 0 means no symptoms), `MONTH` is the number of months since the patient was admitted to hospital (0 through 11), `GENDER` is the patient's gender (1=female, 0=male) and `AGE` is an indicator of the patient's age (1 if age < 20, 0 if age ≥ 20). The patients were all admitted to hospital in month 0 and there is a general decline over time of the proportion of patients showing symptoms.

(a) For each month (0 through 11), calculate the proportion of patients showing symptoms in that month and draw a scatterplot. Briefly describe the shape of the scatterplot. **[3 points.]**

(b) On top of the scatterplot you drew in (a), show a fitted smooth curve using (i) the Bowman-Azzalini method (kernel regression with cross-validation choice of bandwidth), (ii) a regression splines model with 3DF. How many DF are needed in the regression splines model for the appearance of the curve to be approximately the same as that of the Bowman-Azzalini estimator? **[7 points.]**

(c) The main question of interest here is the influence of age and gender on the proportion of patients showing symptoms. First, do an analysis using the `glm` command with `MONTH`, `AGE` and `GENDER` all treated as a factor variables. Qualitatively describe the resulting conclusions, e.g. do women show symptoms more frequently than men, and does the effect vary according to age? Do the results change substantially if you use a quasibinomial model instead of binomial? **[5 points.]**

(d) Now repeat the same analysis, treating `ID` as a random effect, using the Gauss-Hermite quadrature method. In what respects do your conclusions differ from those of part (c)? **[5 points.]**

(e) Repeat the analysis, treating `ID` as a random effect, using the GEE approach with AR1 correlation structure. In what respects do your conclusions differ from those of parts (c) and (d)? **[5 points.]**

(f) Repeat the analysis, treating `ID` as a random effect, using the INLA approach. Use 95% posterior intervals as a measure of statistical significance for each of the fixed effects. In what respects do your conclusions differ from those of parts (c) through (e)? **[5 points.]**

(g) Summarize your conclusions. How sensitive are the results of the analysis to the choice of analysis method? **[4 points.]**
[34 points for the whole question.]

SOLUTIONS AND COMMENTS

1. (a) Assuming the data frame is called `soap`, the following commands

```
library(dplyr)
soap %>%
  group_by (Treat) %>%
  summarise(rate=mean(Score)) %>%
  xtabs(formula=rate~Treat)
```

produce the table

```
Treat
    0     c    cs     d    dc   dcs    ds     s
3.000 2.500 2.500 2.750 3.250 3.500 3.000 3.125
```

The treatments `c` and `cs` are best from the point of view of minimizing mean score.

(b) If you try to fit the model, say

```
m1=lm(Score~Treat+factor(Judge)+factor(Day)+factor(Replicate),soap)
```

none of the `Replicate` variables gives an estimate (all are `NA`). The reason is a fundamental issue of model identifiability: each judge is involved in only one replicate, and if you estimate a fixed effect for `Judge`, there are no degrees of freedom to estimate `Replicate` as well. There is nothing specifically wrong with including `Day` in the model, but it is clear that this variable is not statistically significant.

(c) Modify the model statement to

```
m1=lm(Score~Treat+factor(Judge),soap)
```

then the treatment variables turn out to be

```
               Estimate Std. Error t value Pr(>|t|)
Treatc          -0.14583    0.67497  -0.216  0.83001
Treatcs         -0.45833    0.57045  -0.803  0.42634
Treatd           1.04167    0.67497   1.543  0.13045
Treatdc          1.02083    0.57045   1.790  0.08092 .
Treatdcs         0.83333    0.62490   1.334  0.18971
Treatds          0.10417    0.57045   0.183  0.85601
Treats           0.93750    0.67497   1.389  0.17235
```

In this case `cs` clearly has the smallest (i.e. best) coefficient, though based on the standard errors, it seems unlikely this is significantly different from several others. The reason this is different from (a) is because each judge only evaluated two treatments: therefore, if we ignore the `Judge` effect, the analysis will be biased because of the variability between judges. Including `Judge` as a variable in the model removes this bias.

(d) After loading `lme4`, the model statement

```
m2=lmer(Score~Treat+(1|Judge),soap)
```

fits the same model treating `Judge` as a random effect. The standard deviations for `Judge` and the residual are quite similar (0.71 and 0.88) showing that the inter-judge variability is similar to the within-judge variability (so we can't ignore the former). The table of fixed effects includes

	Estimate	Std. Error	t value
Treatc	-0.39711	0.56464	-0.703
Treatcs	-0.52037	0.52145	-0.998
Treatd	0.43058	0.56464	0.763
Treatdc	0.69957	0.52145	1.342
Treatdcs	0.59511	0.53401	1.114
Treatds	0.06286	0.52145	0.121
Treats	0.47691	0.56464	0.845

which again shows treatment `cs` as best, `c` as second best, but overall less variability among treatments than in the model of (c). The random effects model makes more sense here because you have no particular interest in these specific judges; what you are trying to learn is how different people will evaluate the soap pads when they are released to the general public. Therefore, it is appropriate to treat the judges as if they were randomly sampled from some larger population.

(e) A suitable way of doing this is by writing

```
soap$S=ifelse(soap$Score<=2,1,0)
m3=glm(S~Treat+factor(Judge),family=binomial,soap)
```

However, this produces the warning message "glm.fit: fitted probabilities numerically 0 or 1 occurred" and the coefficients are very unstable (large coefficients and extremely large standard errors). This model makes it appear as if `s` is the treatment with the largest effect. In this case, the largest (positive) effect would be interpreted as best, because it maximizes the probability of success (defined as a score of 1 or 2).

There are some other models you could try, such as the bias-reduction estimation method of Section 2.7, which may be implemented through the model

```
library(brglm)
m4=brglm(S~Treat+factor(Judge),family=binomial,soap)
summary(m4)
```

This also produces a set of results in which `s` has the largest effect, though it is only very slightly ahead of `cs`.

You could also try a random effect model in this case, such as

```
library(lme4)
m5=glmer(S~Treat+(1|Judge),family=binomial,soap)
summary(m5)
```

but this also produces unstable parameter estimates and the warning "Model is nearly unidentifiable."

You could also try the GEE method, for example

```
library(geepack)
m6=geeglm(S~Treat,id=Judge,corstr='exchangeable',family=binomial,soap)
```

but this also produces unstable estimates (though again with `cs` as the treatment with the largest individual coefficient).

Probably the `brglm` is the most satisfactory of these different models but the basic difficulty is that once the scores are reduced to binary responses, there isn't enough variability in the data to fit all the treatment and judge effects simultaneously. It's not necessary to try all these models to reach that conclusion.

(f) You could try fitting the model of (d) without the treatment effect at all, and then doing a Kenward-Roger test, for example

```
library(pbkrtest)
m7=lmer(Score~(1|Judge),soap)
KRmodcomp(m2,m7)
```

which produces a p-value of 0.21, implying that none of the `Treat` effects are significant. This suggests the answer to (ii) is no, and therefore, the answer to (i) is presumably no as well. Directly, from the estimates and standard errors of the `cs` and `c` coefficients, it is clear that the difference between these treatments is not statistically significant.

[There are other things you could try, such as the "Helmert contrasts" idea that we saw in Chapter 11, but this leads to essentially the same conclusion.]

. Comments on student solutions. In general, I would say that there were some fairly persistent errors. A fairly common issue is students simply not answering the question as asked, e.g. if you are being asked which treatment is best, you should give an explicit answer to that question, or else state why the question cannot be answered in the form stated.

Part (b): many students gave vague answers that didn't pinpoint the precise nature of the problem, which is that the model is simply not estimable in the form given. This is most clearly seen by the NAs for the Replicate coefficients.

Parts (c) and (d): Generally well answered, though some students mixed up best and worst effects.

Part (e): a lot of students didn't make any comment about the warning that some estimated probabilities were exactly 0 or 1, which is clearly telling you something and not to be ignored. Some students went further into this and pointed out specific difficulties, for example, every instance of `Treat=='dcs'` led to an original Score of 3 or 4, which translates to `S=0` under the recoding of part (e). Therefore, a logistic regression fails to produce a meaningful estimate for this treatment.

Part (f): I needed to see some form of a hypothesis test here, though it could be Kenward-Roger (the most common choice) or a bootstrap test, or a fixed-effects ANOVA; all three lead to acceptance of the null hypothesis that there is no treatment effect.

2. (a) See Figure 1. There are many ways to draw a plot looking like this, but here is one possible code.

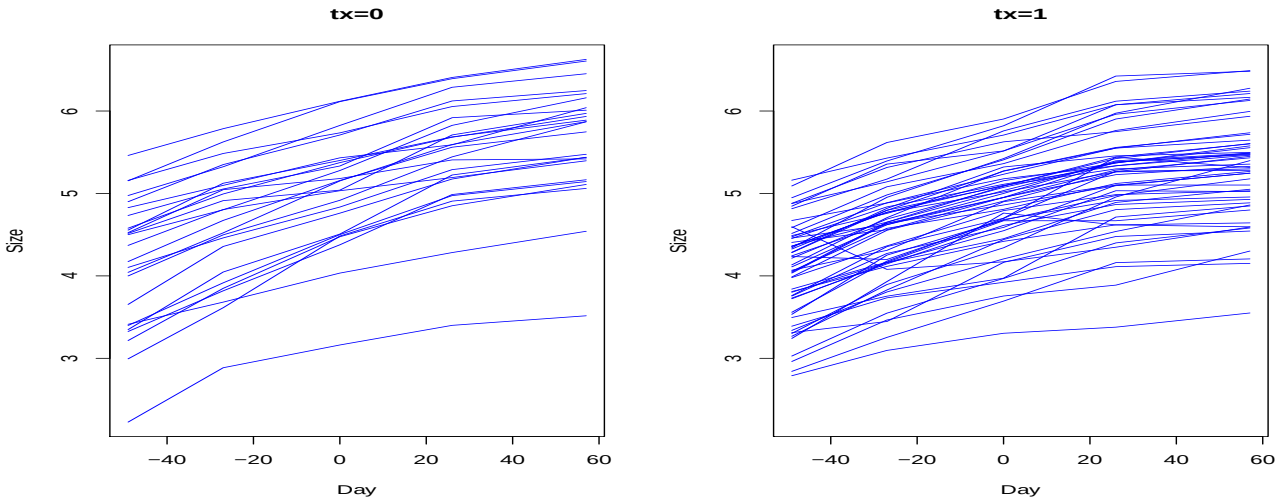

Figure 1: *Growth curves for spruce data with* `tx=0` *(left plot) and* `tx=1` *(right plot).*

```
par(mfrow=c(1,2))
plot(spruce$day,spruce$y,type='n',xlab='Day',ylab='Size',main='tx=0')
for(i in 1:length(unique(spruce$id))){
t=min(spruce$tx[spruce$id==i])
if(t==0)lines(spruce$day[spruce$id==i],spruce$y[spruce$id==i],col='blue')}
plot(spruce$day,spruce$y,type='n',xlab='Day',ylab='Size',main='tx=1')
for(i in 1:length(unique(spruce$id))){
t=min(spruce$tx[spruce$id==i])
if(t==1)lines(spruce$day[spruce$id==i],spruce$y[spruce$id==i],col='blue')}
```

The figure shows that the `tx=0` trees are smaller at the beginning but they seem to have caught up by the end, implying that the main effect of `tx` is in the slope rather than the intercept. However, this is not a clear-cut interpretation and other reasonable answers will be accepted.

As many students pointed out, this plot can be drawn more elegantly using ggplot, for example,

```
ggplot(spruce, aes(x=day, y=y,group=id)) + geom_line() + facet_wrap(~ tx)
```

or if you want something more colorful, for example,

```
ggplot(spruce, aes(x = day, y , color = id, group = id)) + geom_point()
 + geom_line() + facet_grid(~ tx)
```

(b) Some possible code is

```
lm0=lm(y~day*factor(tx)+factor(chamber),spruce)
lm1=lm(y~day*factor(tx),spruce)
lm2=lm(y~day+factor(tx),spruce)
anova(lm0,lm1)
```

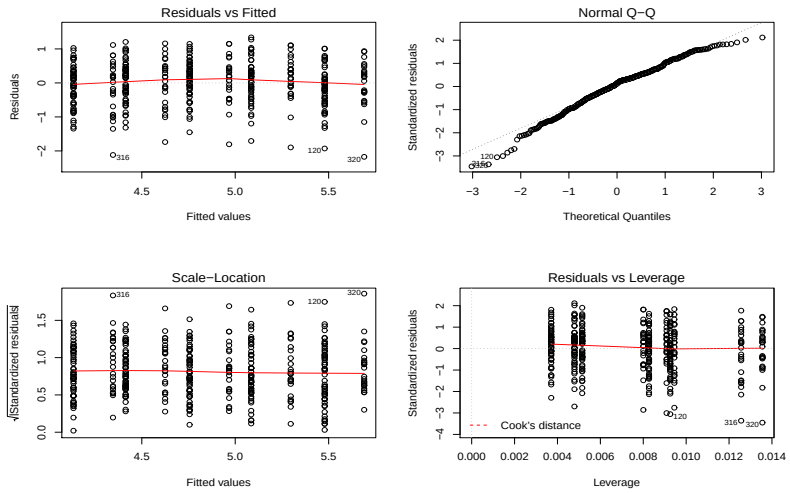

Figure 2: *Diagnostic plots for the model* `lm2`.

```
anova(lm1,lm2)
```

where the first `anova` statement confirms that `chamber` does not have a significant effect separate from `tx` (p=0.18) and the second shows that the model with interaction between `day` and `tx` is not significant (p=0.24). Indeed the `summary(lm1)` output includes

```
                  Estimate Std. Error t value Pr(>|t|)
day                0.01414    0.00151    9.36   <2e-16 ***
factor(tx)1       -0.20851    0.06861   -3.04   0.0025 **
day:factor(tx)1 -0.00213    0.00183   -1.17   0.2432
```

showing that the influence of `tx=1` is statistically significant on the intercept of the straight-line fit (p=0.0025) but not on the slope (p=0.2432).

Based on the model `lm2`, we show the standard diagnostic plots (Figure 2) and a Box-Cox plot (Figure 3). The diagnostic plots seem okay except that the QQ-plot deviates from a straight line at both ends, suggesting a transformation may be necessary. The Box-Cox plots suggests a transformation of the form $y \to y^\lambda$ with $\lambda \approx 1.5$. (However, for the remaining parts of the question, we keep y on its original scale.)

(c) There are various possible ways you could organize the calculations at this point but one way to start a random effects analysis is via

```
library(lme4)
lm3=lmer(y~factor(tx)*day+(1|chamber)+(1|id:chamber),spruce)
lm4=lmer(y~factor(tx)*day+(1|id:chamber),spruce)
```

The `summary(lm3)` shows that the standard deviation of the chamber effect is much smaller than that of the id:chamber effect, so we drop that term in `lm4`. Part of the resulting `summary(lm4)` then gives

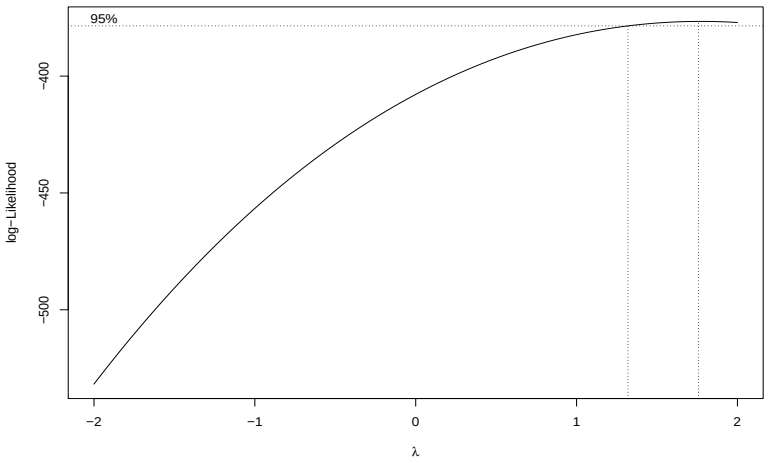

Figure 3: *Box-Cox plot for the model* `lm2`.

```
 Groups   Name        Variance Std.Dev.
Random effects:
 id       (Intercept) 0.3698   0.608
 Residual             0.0376   0.194
Fixed effects:
                  Estimate Std. Error t value
factor(tx)1     -0.208511   0.148592   -1.40
day              0.014142   0.000462   30.61
factor(tx)1:day -0.002135   0.000559   -3.82
```

This shows that the `id` random effect (the between-tree variation) in much larger than the residual (within-tree) variation, so we keep that in the model.

Looking at the fixed effects, it seems the `tx` coefficient is not significant but the interaction with `day` is significant. In other words, `tx` influences the slope but not the intercept. This is the other way round from the answer given in (b).

(d) `PBmodcomp` is appropriate for testing `lm3` against `lm4` but when I tried it myself there were many "Model failed to converge" error messages and a reported p-value of 1. This is probably because the estimated variance due to chamber is so small as to be effectively zero, so we should drop that term. You can't use `PBmodcomp` to test for a single random effect but `exactRLRT(lm4)` is designed exactly for this situation and leads to `p-value <2e-16`, confirming that the random effect due to `id` is statistically significant (and, therefore, contradicting the model in part (b) of the question).

(e) Again you have some flexibility deciding exactly what models to fit but a possible sequence is

```
lm5=lmer(y~factor(tx)+day+(1|id),spruce)
lm6=lmer(y~day+factor(tx):day+(1|id),spruce)
```

```
lm7=lmer(y~day+(1|id),spruce)
lm8=lmer(y~(1|id),spruce)
KRmodcomp(lm4,lm5)
KRmodcomp(lm4,lm6)
KRmodcomp(lm6,lm7)
KRmodcomp(lm6,lm8)
```

which look at first dropping the interaction between `tx` and `day`, and then either or both of the main effects. These results show: `lm4` cannot be reduced to `lm5` (p-value 0.00016); it could possibly reduced to `lm6` (p-value for `lm4` against `lm6` is 0.16) but any further reduction to `lm7` or `lm8` is impossible. This confirms that the `tx:day` is highly significant but you have discretion whether to drop the main effect for `tx`. In other words, the slopes of the growth curves are different between `tx=1` and `tx=0`, but the intercepts may not be.

(f) The conclusion of (c), reinforced by the tests of parts (d) and (e), show that the difference among trees is real, and when taken into account, that the effect of the ozone treatment is on the slopes much more than on the intercepts. Therefore, the random effects analysis in this example is essential to understanding the true nature of the ozone effect.

Comments about student solutions:

In general, this question was well answered — most students were successful in configuring the various R functions needed to answer the question, but as usual, the problems tended to be ones of interpretation more than computational implementation.

Part (a): given the error on the originally posted version of the exam (where I mixed up which of `tx=0` or `1` represented high ozone) I didn't penalize anybody for confusing the high-ozone and low-ozone regimes, but the essential point was to note that the two plots don't look the same.

Part (b): the main point here was that if you account for `tx`, you don't need to include `chamber` as a factor variable as well, but it turns out it does depend to some extent on the order in which you introduce the `tx` and `chamber` variables into the equation — if `chamber` goes first, then a first pass at the regression does not show `tx` as significant. So some students stated (wrongly, but understandably) that `tx` was not significant and could be dropped. Apart from that, the most frequent error with this question was not testing for interactions, though if you do, it turns out that none of the interactions is significant in this model — but this point is important for part (c), see further comments below.

About the Box-Cox transformation, most students correctly drew the plot (Figure 3) but some failed to interpret this correctly — the confidence interval for λ (shows by the vertical lines in Figure 3) does *not* include 1, so a transformation *will* improve the fit. However, students who tried this generally noted that the only feature of the analysis which is improved by transforming is the qq-plot for the residuals, and it makes very little difference to the more substantial issues about the influence of ozone on tree growth. Combined with the rather odd nature of the transformation (taking the original volume measurement to

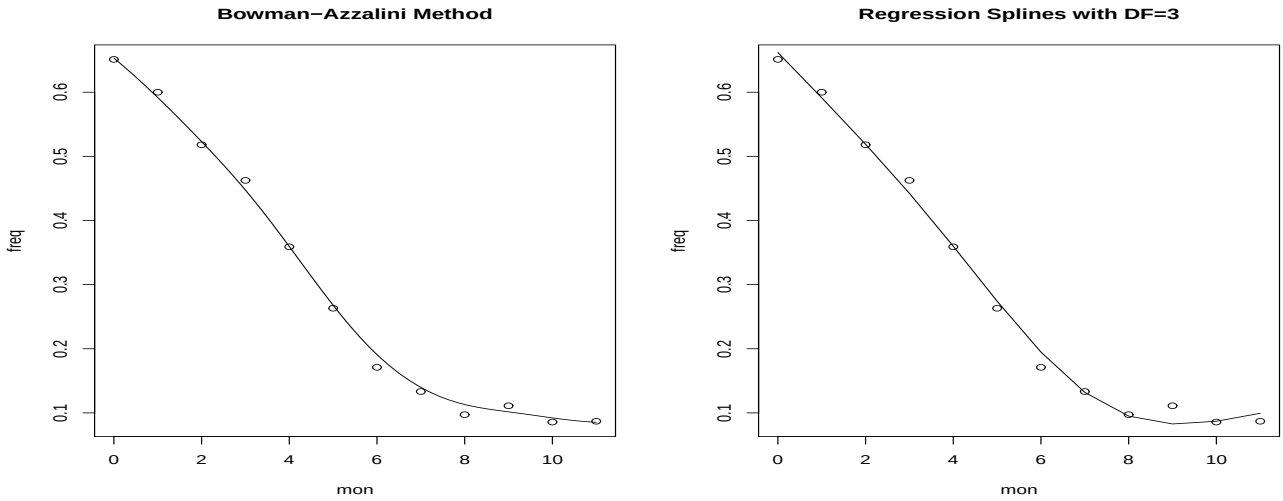

Figure 4: *Fitted smooth curves in question 3.*

(log volume)$^{\lambda}$) making it rather hard to interpret, I think it was reasonable not to persist with the transformation for the rest of the question.

In part (c), the main issue was how to handle the day:tx interaction effect. If you test for this in the simple linear regression in part (b), you find it's not significant, but include it in the random effects model, and it is! This is critical to the whole interpretation of the question, because if you want to know whether the slope of the growth curve varies depends on the ozone concentration, this is precisely the parameter you need to test! However the way the question was graded, you could miss this point and still not lose a whole hatful of grading points. Some students treated `day` as a factor variable, which was not my intention, but it's not wrong either and still raises the question of how to handle the interaction between `day` and `tx`.

Parts (d) and (e) involved the bootstrap and Kenward-Roger test and were generally well answered. For part (d), I gave credit to anyone who correctly identified that the `chamber` random effect is not statistically significant, but the `id` random effect is, regardless of whether you treat it as a solo random effect or nested within `chamber`.

3. (a) The scatterplots and fitted straight lines are both shown in Figure 4. Purely based on the scatterplot, you can say that the frequency of symptoms drops sharply until about month 8, but then it levels off (and never gets to 0).

   (b) See Figure 4. The two curves differ a bit at the right-hand end where the regression spline curves up again (not seen in the scatterplot). However the same curve with DF=4 (not shown here) shows very close agreement. Therefore, I would say you need DF=4.

   (c) A model statement such as

   ```
   m1=glm(Y~factor(MONTH)+factor(AGE)*factor(GENDER),family=binomial,schiz)
   ```

   will fit the desired model: the critical part of the output is

```
                             Estimate Std. Error z value Pr(>|z|)
factor(AGE)1                   -0.252      0.241   -1.05  0.29534
factor(GENDER)1                -1.235      0.222   -5.57  2.5e-08 ***
factor(AGE)1:factor(GENDER)1    1.100      0.348    3.16  0.00157 **
```

which shows that the age effect (on its own) is not statistically significant, but the gender effect definitely is, and there appears to be an age:gender interaction as well. However, this ignores the repeated measurements aspect of the analysis. Doing the same analysis with the quasibinomial family makes no difference, since the overdispersion parameter is only 1.028.

(d) You can try

```
m3=glmer(Y~factor(MONTH)+factor(AGE)*factor(GENDER)+(1|ID),nAGQ=25,
family=binomial,schiz)
```

but when I did this I got a "Model failed to converge" message. The result, however, showed a significant negative effect for gender (meaning women are less likely to show symptoms than men), but no effect for age, nor an age:gender interaction. Relevant parts of the model output are here:

```
                             Estimate Std. Error z value Pr(>|z|)
factor(AGE)1                   -0.391      0.829   -0.47   0.6369
factor(GENDER)1                -1.957      0.695   -2.81   0.0049 **
factor(AGE)1:factor(GENDER)1    1.876      1.149    1.63   0.1025
```

(e) The model statement

```
library(geepack)
m4=geeglm(Y~factor(MONTH)+factor(AGE)*factor(GENDER),id=ID,corstr='ar1',
family=binomial,schiz)
```

shows no error message, a statistically significant negative effect due to gender (p=0.018), and no age effect of age:gender interaction. See below:

```
                             Estimate Std.err  Wald Pr(>|W|)
factor(AGE)1                   -0.190   0.559  0.12   0.7340
factor(GENDER)1                -0.968   0.409  5.59   0.0180 *
factor(AGE)1:factor(GENDER)1    0.949   0.715  1.76   0.1840
```

(f) Some possible code to run INLA is

```
library(INLA)
formula=Y~factor(MONTH)+factor(AGE)*factor(GENDER)+f(ID,model='iid')
m7=inla(formula,family='binomial',data=schiz)
```

The relevant "Fixed effects" portion of the output (`summary(m7)`) is

```
                              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
factor(AGE)1                -0.519 0.839     -2.191   -0.513      1.119 -0.501   0
factor(GENDER)1             -2.017 0.699     -3.431   -2.005     -0.675 -1.980   0
factor(AGE)1:factor(GENDER)1  2.026 1.161     -0.217    2.011      4.360  1.980   0
```

In the Bayesian/INLA context, we typically interpret a "statistically significant" result as meaning that the 95% interval between the 0.025quant and the 0.975quant does *not* include zero. By this interpetation, `factor(GENDER)1` is statistically significant (negative), and neither `factor(AGE)1` nor the interaction `factor(AGE)1:factor(GENDER)1` is statistically significant, consistent with the results you got using the Gauss-Hermite and GEE methods.

(g) The results of the Gauss-Hermite, GEE and INLA approaches are all basically the same, in particular noting the significance of the GENDER effect but not of AGE or the AGE:GENDER interaction. The actual coefficients and levels of statistical significance vary from one method to another, but this is to be expected. The error message with the Gauss-Hermite method (which you also get with the Laplace method, by the way) should be noted, but doesn't seem to affect the conclusions. The overall conclusion is that while the frequency of symptoms in both men and women declines over the 11 months, women are less likely to show symptoms than men, and age does not seem to be a factor. The result of part (c) is most likely deficient because this did not take account of the correlation among repeated measurements on the same individual.

Comments on student solutions:

My intention here, as the above solution will make clear, was that you would look at the gender:age interaction as well as the main effects due to month. (It's possible there might also be an interaction of either gender or age with month, but I did not look at that, and nor did any of the students.) In the event, it was clear as soon as I started grading that most students had not considered interactions at all, so I decided *not* to take account of that in the grading. In other words, you could get full credit for following the requested sequence of analyses without interactions, provided you did all the other parts of the analysis correctly. In this analysis, the differences among models are primarily seen in the coefficient and standard error of the gender variable, illustrated by the following table:

Model	Estimate	S.E.	p-value
glm	–0.804	0.168	1.8×10^{-6}
Gauss-Hermite	–1.299	0.571	0.023
GEE	–0.611	0.329	0.063
INLA	–1.289	0.558	0.021

where the p-value for the INLA approach is derived from a normal approximation and not (so far as I am aware) from anything directly calculated by INLA.

To me, the biggest contrast to come out of these model comparisons is not in the variation of the estimate itself (which varies within the range of standard errors for the Gauss-Hermite, GEE and INLA appoaches) but the fact that all three estimates have a much larger standard error than the glm approach, so that our interpretation of the gender effect changes from "almost certain" (p-value about 10^{-6}) using the glm approach to somewhat questionable (p-values in the 0.02–0.06 range) by the other approaches. (And for those few students who did consider the age:gender interaction, this goes from a p-value of less than 0.002 in the glm approach to between 0.09 and 0.19 in the other approaches, i.e. highly significant to

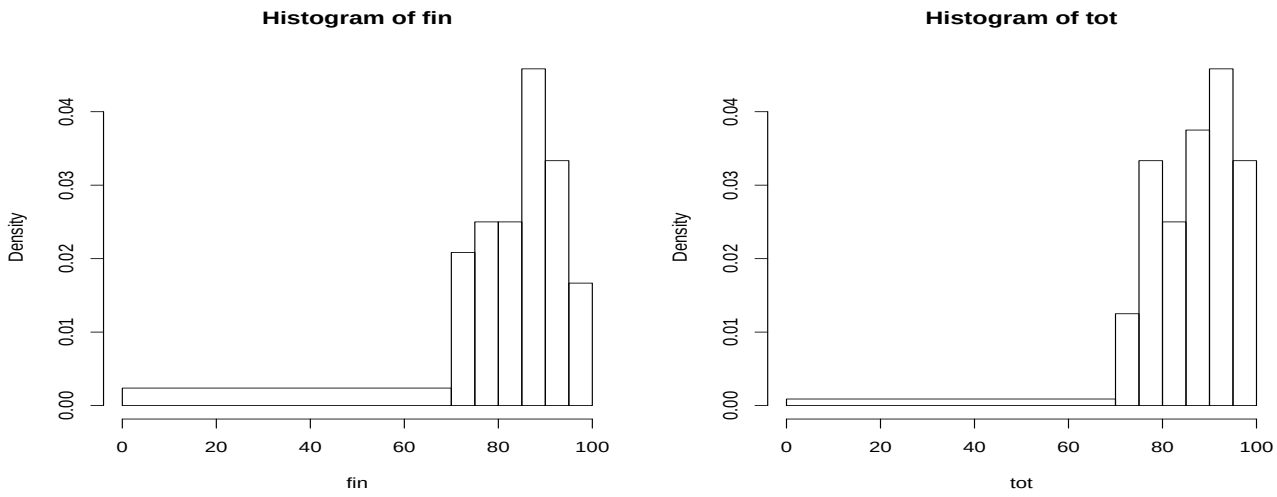

Figure 5: *Histograms for Final Exam and Total Score in the Course.*

not significant at all.) It is important to appreciate that the three latter approaches are *qualitatively* trying to do the same thing, which is take account of the repeated measures on each individual, whereas the glm approach ignores that feature of the data. Therefore, I feel that those students who said that all four approaches give similar answers missed the main point of the question, which is that the standard errors under any of the repeated measures analyses are substantially higher than those under the glm analysis, and this has implications for how one interprets the fixed effects.

Part (b) was generally well answered — many students argued for a higher DF (between 5 and 9) for the spline approach, but I gave credit for that so long as it was based on an actual comparison of the curve under different values for the DF.

In part (d), the majority of students overlooked the warning about non-convergence, which I believe is present for this example in any of the analyses by `glmer` — the results seem credible, but the fact that the algorithm apparently didn't converge is a definite warning sign that should be noted.

Summary of overall scores in the exam:

Mean 81.5, standard deviation 13.6, median 84, first and third quartiles at 74 and 90.2.

For the total scores in the course (25% based on best 7 homeworks, 25% midterm, 50% final), the summary statistics were:

Mean 85.8, standard deviation 10.1, median 87.9, first and third quartiles at 78.7 and 92.9.

Histograms for the final exam ("fin") and total scores ("tot"): see Figure 5.