UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday 2 June, 2016 9:00 am 12:00 pm

PAPER 206

APPLIED STATISTICS

*Attempt no more than **FOUR** questions.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

A physician is studying the speed of growth of a tumour. Data are collected by measuring the size of tumours (`size`, in mm) after they have been allowed to grow in a laboratory growth medium for a certain number of days (`days`). Data are collected in three different laboratories and this is denoted in R with a factor `lab` with levels `A`,`B` and `C`.

(a) The physician decides to fit a linear model to the data:

```
> tumour1<-lm(size~days+days:lab)
> summary(tumour1)

##
## Call:
## lm(formula = size ~ days + days:lab)
##

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6464     5.3745   0.120    0.905
## days         10.1404     0.3677  27.576  < 2e-16 ***
## days:labB     5.0770     0.3739  13.580 2.58e-13 ***
## days:labC    -3.1296     0.3739  -8.371 7.48e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 26 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9842
## F-statistic: 601.5 on 3 and 26 DF,  p-value: < 2.2e-16
```

Write down the algebraic form of the model. What is the estimated daily increase in size for laboratory `A`, `B` and `C`? What is the estimate of the error variance?

(b) Check the diagnostic plots in Figure 1 and define the quantities that appear on the x and y axes of these plots. Do you spot any problems with the model assumptions?

(c) Write down the expression of a $(1-\alpha)$-level prediction interval for the size of a tumour after 10 days of growth in the laboratory B. Show that the probability of new observations lying within the interval is $(1-\alpha)$.

(d) The physician tries then to fit a second model which allows for different intercepts for the different laboratories:

```
> tumour2<-lm(size~days*lab)
> anova(tumour1,tumour2)

## Analysis of Variance Table
##
```

```
## Model 1: size ~ days + days:lab
## Model 2: size ~ days * lab
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 5921.7
## 2     24 5103.0  2    818.72 1.9253 0.1677
```

Explain the test that is carried out by the `anova` command, specifying the null and the alternative hypothesis, the expression of the test statistics and how the p-value is computed. Which model is preferable?

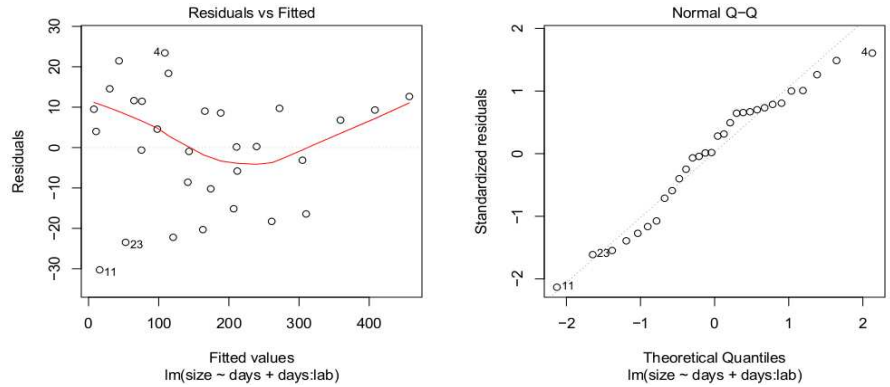

Figure 1: Diagnostics plots.

(e) A colleague of the physician suggests the following analysis.

```
> library(lme4)
> tumour3<-lmer(size~1+(0+days|lab))
> summary(tumour3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: size ~ days + (0 + days | lab)
##
## REML criterion at convergence: 253.7
##

## Random effects:
##  Groups   Name Variance Std.Dev.
##  lab      days  17.08    4.133
##  Residual      227.76   15.092
## Number of obs: 30, groups:  lab, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.6464     5.3745   0.120
## days         10.7895     2.4048   4.487
##
```

What is the difference with respect to model `tumour1`? Is this appropriate for the problem at hand? What is the estimated daily increase in size when averaged across the laboratories?

(f) The physician is considering the possibility of errors being correlated for the observations taken in the same laboratory on different days. Assuming that the error correlation decreases exponentially with the time interval between the two observations, propose a modification to the model `tumour3` to address this problem and give the R commands to fit this new model.

2

Let us consider the model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ (^{T} denotes transpose), X is a known $n \times p$ matrix ($p < n$), $\boldsymbol{\beta}$ a vector of unknown parameters and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$, I_n being the $n \times n$ identity matrix and $X^T X$ assumed to be invertible.

(a) Derive the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ for $\boldsymbol{\beta}$ and σ^2 respectively and write down the distribution of $\hat{\boldsymbol{\beta}}$.

(b) Define the fitted values $\hat{\mathbf{Y}}$ and derive the expression of the hat matrix H such that $\hat{\mathbf{Y}} = H\mathbf{Y}$. What is the covariance matrix of $\hat{\mathbf{Y}}$? Show that H is a projection matrix and that the residuals vector $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated. Show that $\hat{\sigma}^2$ is biased and propose an alternative unbiased estimator for σ^2.

(c) In the (edited) R output below, `energy` is the yearly production of energy from photovoltaic systems, `sunny` is the number of sunny days in the year and `latitude` is the latitude of the systems location.

```
> photo1<-lm(energy~sunny*latitude)
> summary(photo1)

##
## Call:
## lm(formula = energy ~ sunny * latitude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65879 -0.18912  0.01026  0.16919  0.74891
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.619e+00  2.476e-01   6.538 3.01e-09 ***
## sunny           1.403e-02  2.269e-03   6.183 1.53e-08 ***
## latitude        1.438e-02  5.092e-03   2.823  0.00578 **
## sunny:latitude -4.872e-06  4.726e-05  -0.103  0.91811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2836 on 96 degrees of freedom
## Multiple R-squared:  0.7515, Adjusted R-squared:  0.7437
## F-statistic: 96.78 on 3 and 96 DF,  p-value: < 2.2e-16
```

Write down the algebraic form of the fitted model and the estimates of the parameters. What can you suggest to improve the model?

(d) Discuss the output that can be found in Figure 1, obtained with the following R code:

```
> library(MASS)
> boxcox(photo1)
```

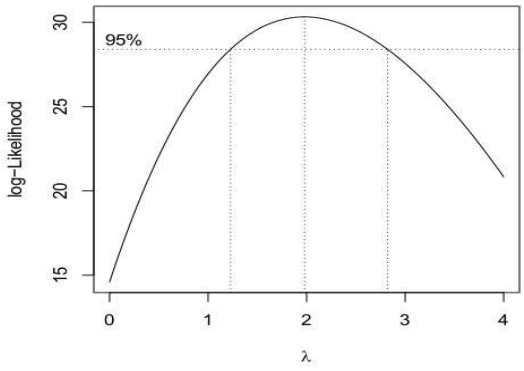

Figure 1: Output of the `boxcox` function.

Which transformation of the data is suggested?

(e) Which model checks would you need to carry out to support the inference based on the chosen model?

(f) Assume now that $\boldsymbol{\epsilon} \sim N(0, \Sigma)$ for some *known* covariance matrix Σ. Derive the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$. What is the expression of the covariance matrix of $\widehat{\boldsymbol{\beta}}$?

3

A group of biologists are investigating the biodiversity in different island environments. They collected data from 30 islands, measuring the number of species (`nspecies`), the average altitude of the island (`altitude`, in meters) and the maximum yearly temperature (`temperature`, in degree Celsius).

(a) The biologists fit three different models and the edited R output can be found below.

```
> species1<-glm(nspecies~altitude*temperature,family=poisson)
> summary(species1)


##
## Call:
## glm(formula = nspecies ~ altitude * temperature, family = poisson)
##

## Coefficients:

##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.7689520  0.7369307   3.757 0.000172
## altitude               0.0428051  0.0372897   1.148 0.251008
## temperature            0.0163103  0.0247408   0.659 0.509736
## altitude:temperature -0.0008147  0.0012722  -0.640 0.521927

##
## (Dispersion parameter for poisson family taken to be 1)
##

##     Null deviance: 101.163  on 29  degrees of freedom
## Residual deviance:  82.916  on 26  degrees of freedom
## AIC: 253.65

> 1-pchisq(82.916,26)

## [1] 7.394536e-08

> species2<-glm(nspecies~altitude+temperature,family=poisson)
> summary(species2)

##
## Call:
## glm(formula = nspecies ~ altitude + temperature, family = poisson)
##

## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.211783   0.255187  12.586  < 2e-16
## altitude    0.019091   0.004530   4.214 2.51e-05
## temperature 0.001145   0.007219   0.159    0.874

## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 101.163  on 29  degrees of freedom
## Residual deviance:  83.326  on 27  degrees of freedom
## AIC: 252.06

>1-pchisq(83.326,27)

## [1] 1.171784e-07

> species3<-glm(nspecies~altitude,family=poisson)
> summary(species3)

##
## Call:
## glm(formula = nspecies ~ altitude, family = poisson)
##

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.24926    0.09578  33.923  < 2e-16
## altitude     0.01898    0.00447   4.245 2.18e-05

## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 101.163  on 29  degrees of freedom
## Residual deviance:  83.351  on 28  degrees of freedom
## AIC: 250.08

>1-pchisq(83.351,28)

2.0935e-07
```

Explain why the biologists decided to fit the model `species2` and the model `species3`. Write down the algebraic form of the final model.

(b) Derive the expression of the deviance of the final model. Is the model a good fit for the data? Why or why not?

(c) Consider now the following R output

```
> phi=(1/(28)*sum((nspecies-species3$fitted.values)^2/(species3$fitted.values)))
> phi
## [1] 3.099079
> species_new<-glm(nspecies~altitude,family=quasipoisson)
```

Describe why there is the need to fit this new model and write down the estimated relationship between the average number of species and altitude. Provide a 95% approximate confidence interval for the coefficient associated with altitude (recall that the 0.975 quantile of a standard normal distribution is $Z_{0.025} = 1.96$). Does this suggest dropping altitude from the model?

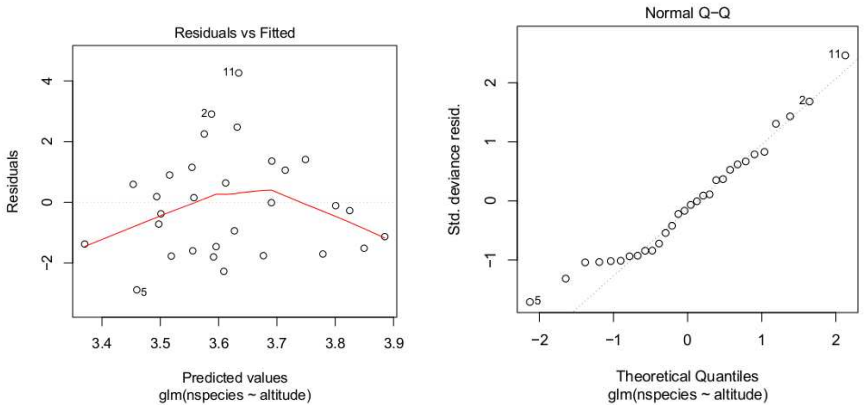

Figure 1: Diagnostics plots.

(d) Write the R code that would be needed to fit a negative binomial model for the number of species with altitude as predictor. Explain why this is not a generalized linear model.

(e) Looking at the diagnostics plots in Figure 1, is it a good idea to fit the following model? Why?

```
> library(mgcv)
> species_gam<-gam(nspecies~s(altitude,bs="cr"),family=quasipoisson)
> plot(species_gam)
```

Considering the plot of the estimated regression function in Figure 2, why is this better than including a quadratic term in the model?

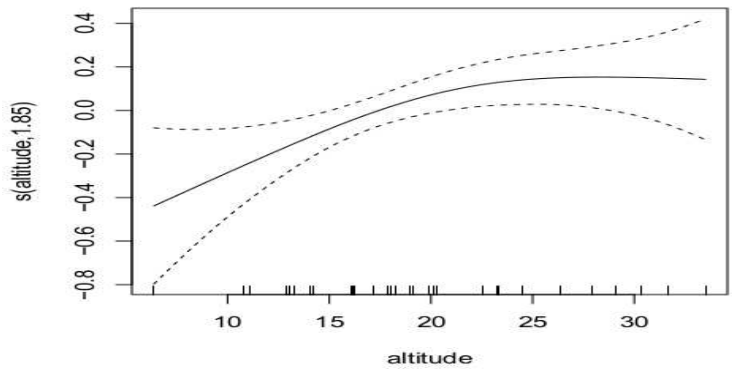

Figure 2: Estimated regression function.

4

(a) Let Y be a Bernoulli random variable, with $P[Y = 1] = p$ and $P[Y = 0] = 1-p$. Show that Y belongs to an exponential dispersion family. Identify the natural parameter and the dispersion parameter. Use the results about the mean and variance of the exponential dispersion family to compute $E[Y]$ and $Var(Y)$.

(b) A health agency is investigating the probability of success of a type of surgery. The dataset imported in R contains the outcome of the surgery (`surg` is 1 if successful, 0 else), the age of the patient (`age`) and a measure of the frailty of the patient (`frail`, this is a factor coded 0 or 1 and it is assessed by clinicians before the surgery).

Consider the following (edited) R output.

```
> surgery_model<-glm(surg~age+frail,family=binomial,
                       weights = rep(1,length(surg)))
> summary(surgery_model)

##
## Call:
## glm(formula = surg ~ age + frail, family = binomial, weights = rep(1,
##     length(surg)))

##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.25375    2.08083   2.044  0.04093 *
## age         -0.08942    0.03635  -2.460  0.01390 *
## frail       -3.15336    1.08296  -2.912  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 84.542  on 99  degrees of freedom
## Residual deviance: 61.376  on 97  degrees of freedom
## AIC: 67.376
##
```

Write down the algebraic form of the model and the estimates of the parameters. Infer from the output how many patients are included in the dataset. State the definition of the Aikake Information Criterion and, from that, derive the expression of its estimator for this specific model.

(c) Consider the output of the following `anova` command.

```
> anova(surgery_model,test="Chisq")

## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: surg
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     99     84.542
## age    1   6.5187        98     78.023   0.01067 *
## frail  1  16.6474        97     61.376 4.501e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Assuming that there is no problem with the diagnostics plots, what can you conclude about which variables affect the probability of success of the surgery?

(d) A second model is fitted with the commands

```
> library(mgcv)
> surgery_model2<-gam(surg~s(age,bs="cr")+frail,family=binomial,
                              weights = rep(1,length(surg)))
```

Write down the algebraic form of this second model. Looking at Figure 1, is this an improvement with respect to the first model? Why?

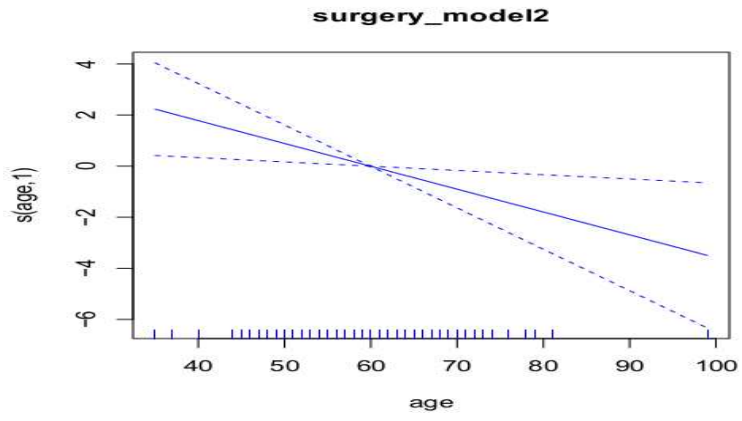

Figure 1: Estimated regression function

5

Consider the model

$$Y_i = f(X_i) + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.

(a) Explain how the regression function f can be estimated using local polynomial regression with first degree polynomials.

(b) Define the penalized minimization problem (with smoothing parameters $\lambda > 0$) whose solution provides the cubic smoothing spline estimator $\hat{f}_\lambda$. Derive the expression of the smoothing operator S_λ such that $\hat{\mathbf{Y}} = S_\lambda \mathbf{Y}$, where $\hat{Y}_i = \hat{f}_\lambda(X_i)$. What are viable strategies for the choice of λ?

Data have been collected about wages, years of education and years of working experience in an European city. These data have been imported in R in the variables `wage`, `education` and `experience` respectively.

(c) Consider the following R code

```
> library(gam)
> wage_model1<-gam(wage~ lo(experience,span=1,degree=1)
                       + lo(education,span=1,degree=1))
> wage_model2<-gam(wage~ s(experience,spar=0.8)+s(education,spar=1.2))
```

Write down the algebraic form of the model that has been considered and explain what is the difference between the fitted models `wage_model1` and `wage_model2`. Looking at the estimated regression functions in Figure 1, do you see any qualitative difference between the two fitted models? Can you suggest how to tune the smoothing parameter to improve the fit, if necessary?

(d) Describe the algorithm used to estimate the regression functions in the model `wage_model2`.

(e) Look at the (edited) summary of the fitted model `wage_model2`:

```
##
## Call: gam(formula = wage ~ s(experience, spar = 0.8) + s(education,
##     spar = 1.2))

##     Null Deviance: 55843310 on 99 degrees of freedom
## Residual Deviance: 906437.7 on 87.7644 degrees of freedom

## AIC: 1221.47
##

## Anova for Parametric Effects
```

```
##                                Df   Sum Sq  Mean Sq F value    Pr(>F)
## s(experience, spar = 0.8)  1.000 45464839 45464839 4402.06 < 2.2e-16 ***
## s(education, spar = 1.2)   1.000  5242047  5242047  507.55 < 2.2e-16 ***
## Residuals                 87.764   906438    10328
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                            Npar Df Npar F   Pr(F)
## (Intercept)
## s(experience, spar = 0.8)     8.8 45.447 < 2e-16 ***
## s(education, spar = 1.2)      0.4  6.016 0.03923 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What are the effective degrees of freedom associated to each predictor? And the total effective degrees of freedom of the model? What is the estimate of the error variance in this case?

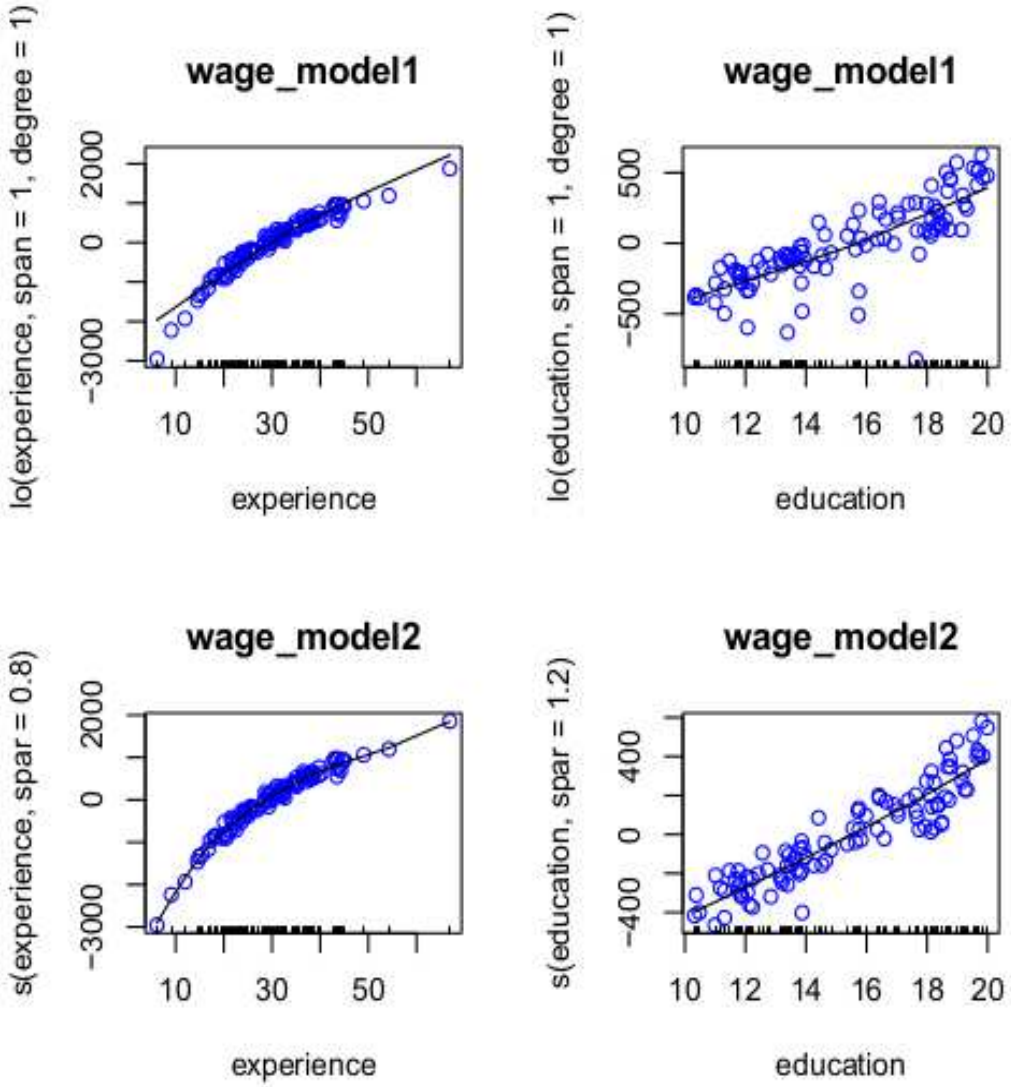

Figure 1: Estimated regression function.

6

A researcher collected data about a manufacturing process for aluminum bars. The researcher is interested in the relationship between the stirring rate applied in the furnace and the yield strength of the aluminium bars. Data are collected from three different furnaces and they are imported in R in the variables `strength` (yield strength of the bar in MPa), `stir` (the stirring rate in the furnace in rotations per minute) and `furnace` (a factor coded as A, B and C indicating the furnace the bar comes from).

(a) Consider the following R output:

```
> library(lme4)
> strength_model<- lmer(strength~ stir+(1|furnace)+(0+stir|furnace))
> summary(strength_model)

## Linear mixed model fit by REML [ lmerMod ]

## Formula: strength ~ stir + (1 | furnace) + (0 + stir | furnace)

## Random effects:

 ##    Groups       Name      Variance  Std.Dev.
 ##   furnace   (Intercept)  48.73800   6.9813
 ##  furnace.1      stir      0.01316   0.1147
 ##  Residuals                36.29950   6.0249

## Number of obs: 30, groups: furnace, 3

## ## Fixed effects:

 ##                Estimate  Std.  Error  t value
 ##  (Intercept)  277.6636     4.7690      58.22
 ##      stir      0.5601      0.1102      5.08
```

Write down the algebraic form of of the model that has been fitted and the estimates of the parameters and derive the marginal formulation of the model. Comment on the diagnostics plots in Figure 1, do they show any problems with the model assumptions?

(b) Derive an expression for the estimator of the conditional modes of the random effects. [You may use without proof any results you require about the multivariate normal distribution.]

(c) The researcher then considers a second model. Below you can find the R commands and an edited output.

```
> strength_model2<- lmer(strength~ 1+(1|furnace)+(0+stir|furnace))
> anova(strength_model2,strength_model)
```

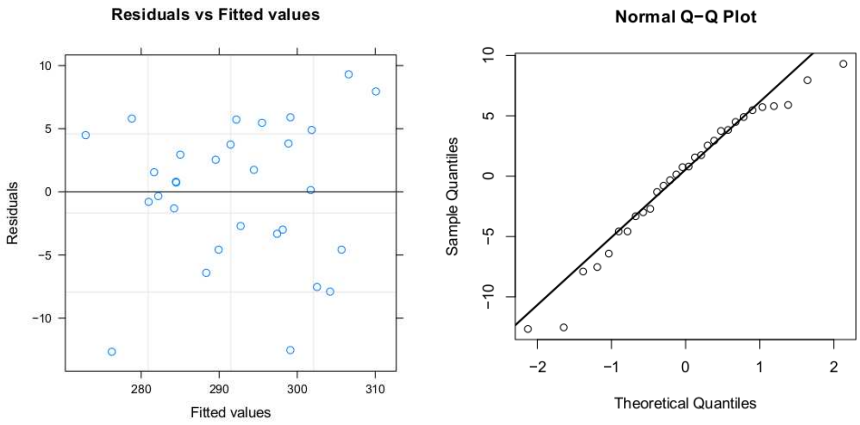

Figure 1: Diagnostics plots.

```
## refitting model(s) with ML (instead of REML)
## Models:

## strength_model2: strength ~ 1 + (1 | furnace) + (0 + stir |furnace)
## strength_model: strength ~ stir + (1 | furnace) + (0 + stir | furnace)

##                  Df    AIC    Chisq   Chi Df  Pr(>Chisq)

## strength_model2  4 216.18

## strength_model   5 210.22    7.9588     1    0.004786
```

Describe the restricted maximum likelihood procedure for parameters estimation and explain why it is necessary to refit the models using maximum likelihood. Can the researcher drop the fixed effect associated with the stirring rate from the model?

(d) The researcher also considers a third model

```
> strength_model3<- lmer(strength~ stir+(1|furnace))
> anova(strength_model3,strength_model)
```

Explain why the output of the `anova` command cannot be trusted in this case and suggest an alternative method to compare the models `strength_model3` and `strength_model1`.

(e) The researcher is worried that a linear function is not appropriate to describe the relationship between stirring rate and yield strength and decides to fit the fixed effects nonparametrically.

```
> library(mgcv)
> strength_model4<-gamm(strength~s(stir,bs="cr"),random=list(furnace=~1))
> plot(strength_model4$gam)
```

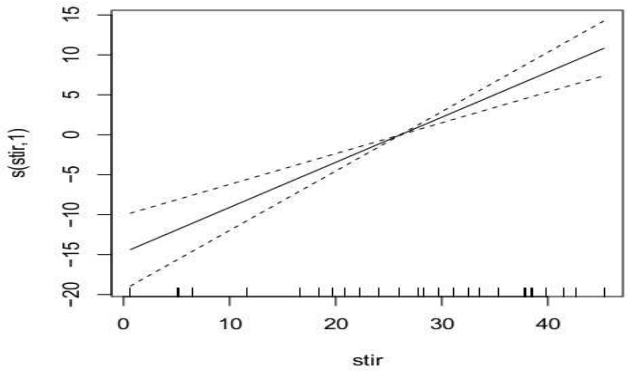

Figure 2: Estimated regression function.

Looking at Figure 2, is the nonparametric approach needed? Considering the following output, which model should the researcher choose?

```
> AIC(strength_model,strength_model3,strength_model4$lme)

##                       df      AIC

## strength_model        5   208.3243
## strength_model3       4   206.6408
## strength_model4$lme 5   210.3425
```

END OF PAPER