UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 1 June, 2015 9:00 am to 12:00 pm

PAPER 33

APPLIED STATISTICS

*Attempt no more than **FOUR** questions, with at most **THREE** from Section A.*

*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.

SECTION A

1

Suppose that $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ satisfies $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where X is a known $n \times p$ matrix ($p < n$), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is unknown, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ where $\varepsilon_1, \ldots, \varepsilon_n$ are independent normally distributed random variables with mean zero and variance σ^2 (unknown), and where ^{T} denotes the transpose. Assume that $X^T X$ is invertible. Find $\hat{\boldsymbol{\beta}}$ that minimises $R(\boldsymbol{\beta}) = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$. Define the fitted values $\hat{\mathbf{Y}}$ and show that $\hat{\mathbf{Y}} = H\mathbf{Y}$ for some suitable matrix H (which you should specify in terms of X) satisfying $H^T = H$ and $HH = H$. Define the residuals $\hat{\boldsymbol{\varepsilon}}$ and show that $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}^T G\mathbf{Y}$ for some matrix G satisfying $G^T = G$ and $GG = G$. Find $\mathrm{cov}(\hat{\boldsymbol{\varepsilon}}, \hat{\mathbf{Y}})$.

An investigation is carried out into the use of a windmill to generate electricity. There are 25 observations of the electrical output together with the corresponding values of the wind velocity. In the (edited) R output below, the electrical outputs and wind velocities are in `output` and `velocity`, respectively. Figure 1 shows a plot of the data and various residual plots. For these plots, `res1`, `res2` and `res3` contain the residuals from models `wind1.lm`, `wind2.lm` and `wind3.lm`, respectively. The corresponding fitted values are in `fit1`, `fit2` and `fit3`.

Write down the algebraic form of the model fitted in `wind1.lm`. Several numerical values in the output to the directive `anova(wind1.lm)` have been replaced by asterisks. Write down what these numerical values should be. Explain how the value of `Multiple R-squared` is calculated.

With reference to the plots, explain why the model wind2.lm is fitted to the data. Write down the algebraic form of the model fitted in `wind2.lm`. What hypothesis is being tested in the line marked `(A)` in the output? Carry out this test in detail, and give your conclusion.

Compare the models `wind2.lm` and `wind3.lm`. What further information or plots would help you to compare the adequacy of these two models?

```
> wind1.lm <- lm(output~velocity)
> anova(wind1.lm)
          Df Sum Sq Mean Sq F value    Pr(>F)
velocity   * 8.9296    *     160.26 7.546e-12
Residuals  * 1.2816  0.0557
> summary(wind1.lm)
 <output omitted>
Multiple R-squared:  0.8745
> velocity2 <- velocity*velocity
> wind2.lm <- lm(output~velocity+velocity2)
```

```
> summary(wind2.lm)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.155898   0.174650  -6.618 1.18e-06
velocity     0.722936   0.061425  11.769 5.77e-11
velocity2   -0.038121   0.004797  -7.947 6.59e-08    (A)
Residual standard error: 0.1227 on 22 degrees of freedom
Multiple R-squared:  0.9676
F-statistic: 328.3 on 2 and 22 DF,  p-value: < 2.2e-16
> velocity3 <- 1/velocity
> wind3.lm <- lm(output~velocity3)
> summary(wind3.lm)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9789     0.0449   66.34   <2e-16
velocity3    -6.9345     0.2064  -33.59   <2e-16
Residual standard error: 0.09417 on 23 degrees of freedom
Multiple R-squared:   0.98
F-statistic:  1128 on 1 and 23 DF,  p-value: < 2.2e-16
```

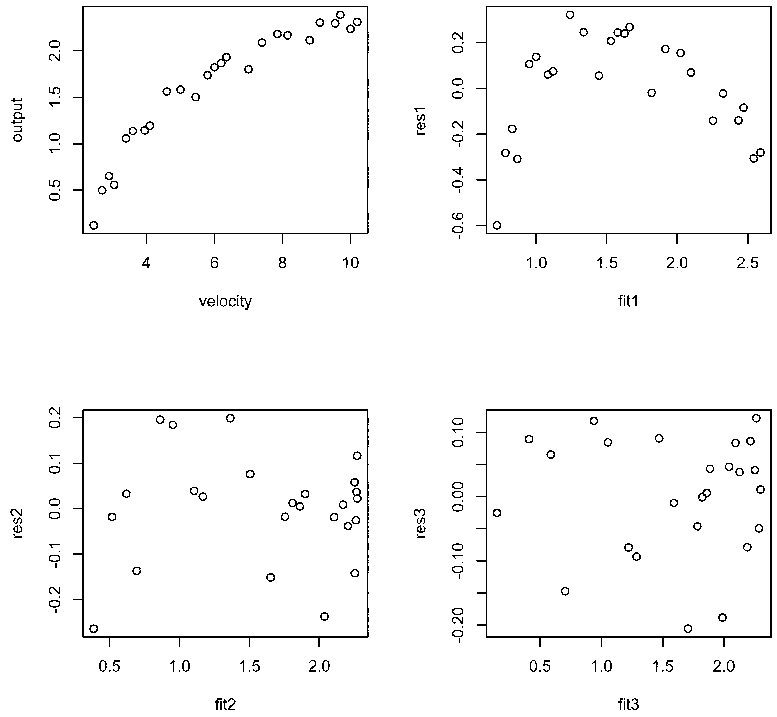


Figure 1: Plot of the data and various residual plots

2

A scientist carried out an experiment to investigate the effects of age and different processing methods on the recall of a list of words. There were fifty subjects in each of two age groups (`Younger` and `Older`). The subjects were randomly assigned to five processing groups, labelled `A`, `B`, `C`, `D` and `E`, in such a way that there were ten of each age group in each processing group. Each processing group was asked to process the words in a different way, then each subject was asked to write down as many words as they could remember, and the number of words correctly recalled was recorded for each subject. The first two and the last two lines of the data are shown below:

```
    Age        Process       Words
1   Younger    A              14
2   Younger    A              11
        <lines omitted>
99  Older      E               9
100 Older      E               6
```

The left-hand number is the subject number, `Age` is a factor with two levels `Younger` (between the ages of 18 and 30) and `Older` (between the ages of 55 and 65), `Process` is a factor with five levels `A`, `B`, `C`, `D` and `E`. The (edited) R output below refers to statistical analysis of the above data. The default constraints in R are used.

For the model `memory1.lm`, write down the algebraic form of the model, defining your notation and writing down all assumptions. For this model, find the estimated mean number of words correctly recalled by an older subject in group `A` and explain how to obtain a 95% confidence interval for this estimate.

Write down the algebraic form of the model fitted in `memory2.lm`. Carry out a statistical test to determine which of the two models is preferable, stating the null hypothesis, the test statistic, the null distribution of the test statistic, and whether or not the null hypothesis should be rejected. Explain the output to the directive `summary(memory2.lm)` in detail, and describe the effects of age and processing method on the correct recall of words.

The scientist later tells the statistician that groups `A` and `C` are in fact using methods of type `I`, groups `B` and `E` are using methods of type `II`, and group `D` is using a method of type `III`. Explain how to test whether this reduced number of categories of processing method is adequate.

```
> memory1.lm <- lm(Words~Age + Process)
> summary(memory1.lm)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.3500     0.7632  14.871  < 2e-16
AgeYounger          3.1000     0.6232   4.975 2.94e-06
ProcessB           -6.1500     0.9853  -6.242 1.24e-08
ProcessC            2.6000     0.9853   2.639  0.00974
ProcessD            2.7500     0.9853   2.791  0.00636
ProcessE           -5.6500     0.9853  -5.734 1.18e-07

> memory2.lm <- lm(Words ~ Age*Process)
> anova(memory2.lm)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Age          1  240.25  240.25 29.9356 3.981e-07
Process      4 1514.94  378.74 47.1911 < 2.2e-16
Age:Process  4  190.30   47.58  5.9279 0.0002793
Residuals   90  722.30    8.03
> summary(memory2.lm)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.0000     0.8959  12.279  < 2e-16
AgeYounger                   3.8000     1.2669   2.999  0.00350
ProcessB                    -4.0000     1.2669  -3.157  0.00217
ProcessC                     2.4000     1.2669   1.894  0.06139
ProcessD                     1.0000     1.2669   0.789  0.43201
ProcessE                    -4.1000     1.2669  -3.236  0.00170
AgeYounger:ProcessB         -4.3000     1.7917  -2.400  0.01846
AgeYounger:ProcessC          0.4000     1.7917   0.223  0.82385
AgeYounger:ProcessD          3.5000     1.7917   1.953  0.05387
AgeYounger:ProcessE         -3.1000     1.7917  -1.730  0.08702

Residual standard error: 2.833 on 90 degrees of freedom
Multiple R-squared:  0.7293
F-statistic: 26.93 on 9 and 90 DF,  p-value: < 2.2e-16
```

3

(a) Let Y be a Poisson random variable with density function

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}; \quad k = 0, 1, 2, \ldots, +\infty.$$

Show that Y belongs to an exponential dispersion family with respect to the unknown parameter λ. 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 company is responsible for the maintenance of plumbing systems in a small town. The company is interested in exploring if the monthly number of maintenance works (`monthly_works`) is related to the average temperature (`avg_temperature`) and precipitation (`avg_precipitation`) during the month. A Poisson regression model is fitted to analyze the relationship. The R code for the analysis and its output are given below.

Write down the algebraic form of the model fitted with the `glm` command and the estimates for the parameters of the model. Using the information in the summary, deduce the number of months in which data have been collected, i.e. the number of observations in the dataset.

Is the model a good fit for the data? Why?

Consider the output of the `anova` command. What do you conclude about which variables should be considered in the prediction of the number of maintenance works?

```
> model<-glm(monthly_works ~ avg_temperature+avg_precipitations,
             family=poisson)

> summary(monthly_works)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   49.75   54.00   52.71   57.25   71.00

> summary(model)

Call: glm(formula = monthly_works ~ avg_temperature +
avg_precipitations,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9268  -0.7649  -0.1740   0.5438   1.8069

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.079374   0.085628  47.641  < 2e-16 ***
```

```
avg_temperature    -0.006162   0.001844  -3.342 0.000833 ***
avg_precipitations -0.002922   0.005988  -0.488 0.625582
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 37.969  on 23  degrees of freedom
Residual deviance: 23.527  on 21  degrees of freedom

AIC: 169.2

Number of Fisher Scoring iterations: 4

> 1-pchisq(37.969,23)
[1] 0.02566753

> 1-pchisq(23.527,21)
[1] 0.3165361


> anova(model,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: monthly_works

Terms added sequentially (first to last)

                   Df     Deviance Resid. Df Resid. Dev   Pr(>Chi)
NULL                                      23     37.969

avg_temperature    1          14.204      22     23.765   0.000164 ***

avg_precipitations 1           0.238      21     23.527   0.625677

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

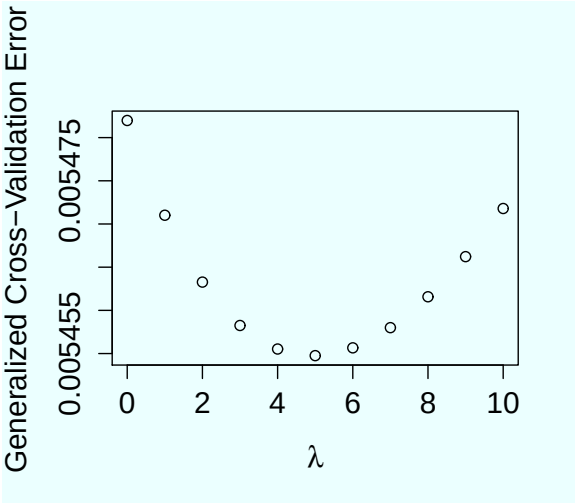
4

We are interested in investigating how a response variable Y is related to 6 predictor variables $X_1, \ldots, X_6$. The predictor variables have been centered and scaled prior to the analysis. Given that some of the predictor variables are highly correlated, a regularized linear model is considered. We define a matrix X whose columns contain prediction variables $X_1, \ldots, X_6$. A ridge regression is performed using the following R code.

```
> library(MASS)
> lambda = seq(0,10,1)
> model.ridge <- lm.ridge(Y ~ X,lambda=lambda)
```

(a) Write down the optimization problem whose solution provides the ridge regression estimator for the parameter vector $\boldsymbol{\beta}$ and specify the analytical expression of the solution as a function of λ and X.

(b) Looking at Figure 1, choose and justify an appropriate value for the parameter λ and report the corresponding estimates for the regression coefficients.



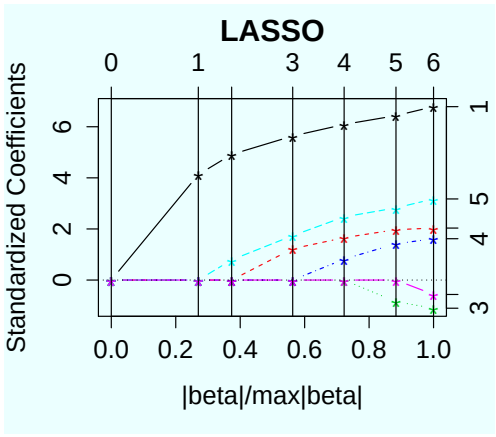

```
> model.ridge
    (Intercept)        X1        X2          X3         X4        X5          X6
 0 0.9348684 0.5876467 0.4180838 -0.01511242 0.11381222 0.7825645 -0.041183801
 1 0.9193437 0.5761795 0.4177029 -0.01453414 0.11168645 0.7700264 -0.032220247
 2 0.9051732 0.5653772 0.4171106 -0.01398452 0.10968893 0.7584556 -0.023970134
 3 0.8922286 0.5551772 0.4163366 -0.01346129 0.10780673 0.7477375 -0.016357926
 4 0.8803977 0.5455247 0.4154061 -0.01296245 0.10602865 0.7377744 -0.009318270
 5 0.8695819 0.5363716 0.4143406 -0.01248621 0.10434493 0.7284822 -0.002794344
 6 0.8596943 0.5276756 0.4131587 -0.01203099 0.10274706 0.7197886  0.003263490
 7 0.8506575 0.5193992 0.4118766 -0.01159535 0.10122756 0.7116305  0.008898781
 8 0.8424028 0.5115089 0.4105080 -0.01117799 0.09977980 0.7039536  0.014149896
 9 0.8348687 0.5039750 0.4090651 -0.01077776 0.09839796 0.6967100  0.019050763
10 0.8280000 0.4967709 0.4075584 -0.01039359 0.09707681 0.6898579  0.023631494
```

Figure 1: Top: Generalized cross-validation error as function of λ. Bottom: Table of the ridge trace for the selected values of λ.

A more interpretable model can be obtained using the LASSO, which allows one to select the relevant predictor variables.

```
>library(lars)
>model.lasso<-lars(X,Y,type="lasso")
>plot(model.lasso)
```

The plot of the LASSO coefficients can be found in Figure 2.




```
> coef(model.lasso)
          X1        X2          X3          X4        X5          X6
[1,] 0.0000000 0.0000000  0.00000000  0.00000000 0.0000000  0.0000000
[2,] 0.3573071 0.0000000  0.00000000  0.00000000 0.0000000  0.0000000
[3,] 0.4257203 0.0000000  0.00000000  0.00000000 0.1947693  0.0000000
[4,] 0.4881368 0.2520802  0.00000000  0.00000000 0.4284335  0.0000000
[5,] 0.5274163 0.3407763  0.00000000  0.05757327 0.6094759  0.0000000
[6,] 0.5571803 0.4053950 -0.01160599  0.09986493 0.6963550  0.0000000
[7,] 0.5876467 0.4180838 -0.01511242  0.11381222 0.7825645 -0.0411838
```

Figure 2: Top: Plot of the lasso coefficients as function of the fraction of the maximum l_1 norm. Bottom: Table of the estimates of the coefficients at each step of the lasso algorithm, corresponding to the vertical lines numbered from 0 to 6 on the plot in the top panel.

(c) Write down the optimization problem whose solution provides the LASSO estimator for the parameter vector $\boldsymbol{\beta}$.

(d) The fraction of the maximum l_1 norm of the coefficients which minimizes the 10-fold cross-validation error is 0.8. Looking at Figure 2, which predictors should be included in the model?

SECTION B

5

(a) Let T be a positive continuous survival time random variable, with probability density function $f(t)$, survivor function $S(t)$, and hazard function $h(t)$, at $t > 0$. Denote the cumulative hazard function by $H(t) = \int_0^t h(u)du$. Derive

(i) an expression for the mean of T, μ_T, in terms of $S(t)$;

(ii) an expression for $S(t)$ in terms of $H(t)$; and

(iii) the distribution of the random variable $H(T)$.

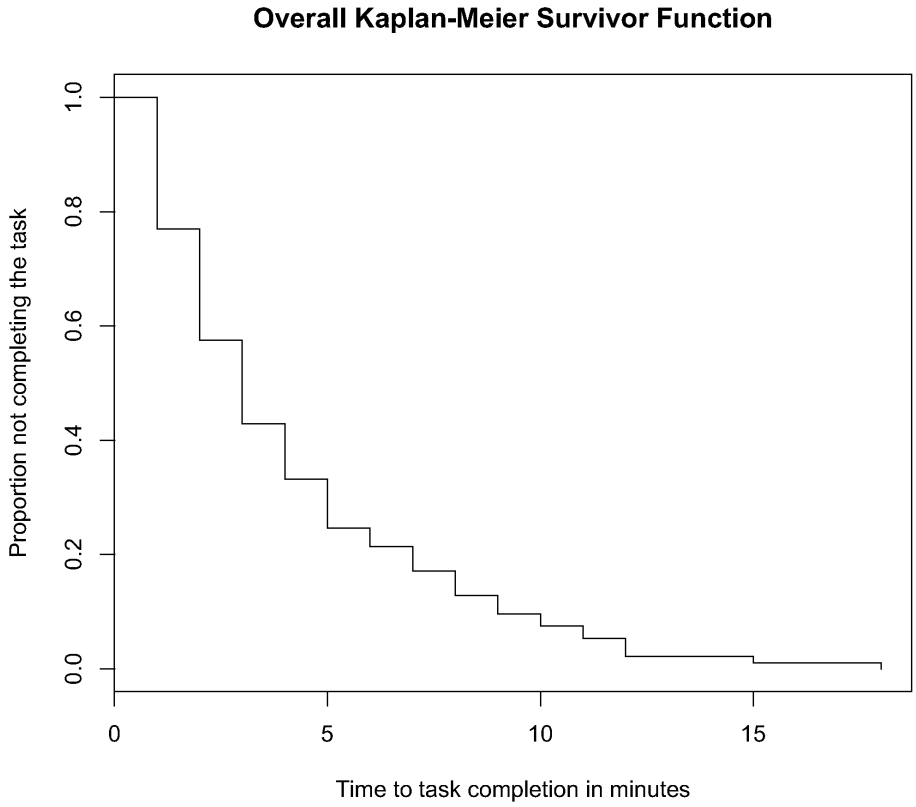
(b) In a study investigating the time in minutes to complete a particular computer task, one hundred adults aged 21 years old and over were randomly assigned to either perform the task in a controlled environment where Handel's *Messiah* oratorio was being played or in the same controlled environment but where Elgar's *The Kingdom* was played instead. The data collected are in the form (T_i, D_i), where T_i represents either the time to completion of the task, if observed, or the time to giving up on the task for the ith individual, and is recorded in the variable `comptime` within the data-set, `task.dat`. D_i takes the value 0 or 1 depending on whether the task was uncompleted (`status`= 0) or observed to be completed (`status`= 1). Also recorded in the data-set `task.dat` are the variables `group` (coded 0 for *The Kingdom* and 1 for *Messiah*), `age` (the age of the subject in years) and `gender` (coded 0 for female and 1 for male).

The Researchers responsible for analysing the data decide that it is appropriate to use survival analysis techniques. They begin by producing an overall Kaplan-Meier curve to describe the completion time data. They then proceed to fit a Cox proportional hazards model and conduct diagnostics. The plots, R code and edited R output from their analyses are displayed on the subsequent pages.

(i) From the overall Kaplan-Meier estimator of the survivor function for time to completion, work out the appropriate estimates for the median and mean completion times. What would be the impact on these estimates if the last observed completion time of 18 minutes was actually a censored observation, whilst all the other data remain the same?

(ii) Interpret the output from the model `task.coxph` in the context of the study.

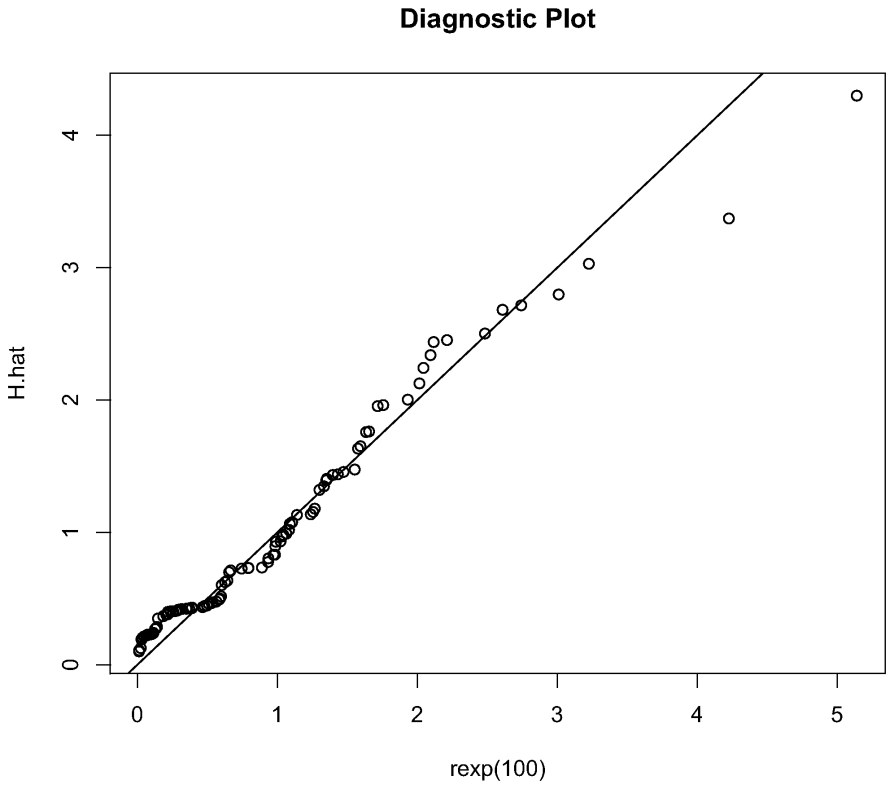
(iii) Explain why the Researchers have run the R command `cox.zph(task.coxph)` and produced the diagnostic plot shown. What can be concluded about the Cox model fitted?




```
> task.surv <- survfit(Surv(comptime,status)~1,data=task.dat)
> plot(task.surv,conf.int=F,xlab="Time to task completion in minutes", ylab=
  "Proportion not completing the task", main="Kaplan-Meier Survivor Function")
> summary(task.surv)
Call: survfit(formula = Surv(comptime, status) ~ 1, data = task.dat)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      23   0.7700  0.0421      0.69178       0.8571
    2     75      19   0.5749  0.0498      0.48512       0.6814
    3     55      14   0.4286  0.0502      0.34067       0.5392
    4     40       9   0.3322  0.0481      0.25007       0.4412
    5     31       8   0.2464  0.0442      0.17337       0.3503
    6     23       3   0.2143  0.0422      0.14572       0.3151
    7     20       4   0.1714  0.0388      0.11002       0.2671
    8     16       4   0.1286  0.0345      0.07598       0.2176
    9     12       3   0.0964  0.0305      0.05191       0.1791
   10      9       2   0.0750  0.0272      0.03684       0.1527
   11      7       2   0.0536  0.0233      0.02286       0.1255
   12      5       3   0.0214  0.0150      0.00544       0.0843
   15      2       1   0.0107  0.0107      0.00153       0.0752
   18      1       1   0.0000     NaN           NA           NA
```

```
> task.coxph <- coxph(Surv(comptime,status)~age+gender+group,data=task.dat,
   method="breslow")
> summary(task.coxph)
Call:
coxph(formula = Surv(comptime, status) ~ age + gender + group,
    data = task.dat, method = "breslow")

  n= 100, number of events= 96

            coef exp(coef)  se(coef)      z Pr(>|z|)
age     0.006536  1.006558  0.010261  0.637 0.524136
gender -0.603547  0.546868  0.219039 -2.755 0.005861
group   0.755308  2.128266  0.226498  3.335 0.000854
---

       exp(coef) exp(-coef) lower .95 upper .95
age       1.0066     0.9935    0.9865    1.0270
gender    0.5469     1.8286    0.3560    0.8401
group     2.1283     0.4699    1.3653    3.3176

> cox.zph(task.coxph)
          rho  chisq     p
age    0.0718 0.5429 0.461
gender 0.0622 0.3791 0.538
group  0.0153 0.0221 0.882
GLOBAL     NA 0.8135 0.846

> task.survph <- survfit(task.coxph)
> summary(task.survph) # Baseline Survivor Function Table
Call: survfit(formula = task.coxph)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    100      23 0.812584 0.03617     7.45e-01       0.8867
    2     75      19 0.632534 0.04674     5.47e-01       0.7311
    3     55      14 0.491275 0.04991     4.03e-01       0.5995
    4     40       9 0.388888 0.05006     3.02e-01       0.5005
    5     31       8 0.290836 0.04786     2.11e-01       0.4015
    6     23       3 0.246232 0.04676     1.70e-01       0.3573
    7     20       4 0.190023 0.04367     1.21e-01       0.2982
    8     16       4 0.138653 0.03879     8.01e-02       0.2399
    9     12       3 0.100551 0.03389     5.19e-02       0.1947
   10      9       2 0.074829 0.02976     3.43e-02       0.1632
   11      7       2 0.048641 0.02456     1.81e-02       0.1308
   12      5       3 0.018487 0.01423     4.09e-03       0.0836
   15      2       1 0.006598 0.00879     4.85e-04       0.0898
   18      1       1 0.000794 0.00209     4.58e-06       0.1377
```

```
> round(task.survph$surv,4)
 [1] 0.8126 0.6325 0.4913 0.3889 0.2908 0.2462 0.1900 0.1387 0.1006 0.0748
[11] 0.0486 0.0185 0.0066 0.0008
> H0.hat <- -log(task.survph$surv)
> round(H0.hat,4)
 [1] 0.2075 0.4580 0.7108 0.9445 1.2350 1.4015 1.6606 1.9758 2.2971 2.5925
[11] 3.0233 3.9907 5.0210 7.1385
> task.survph$time
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 15 18
>
> task.coxph$y
  [1]  2   1   1   8   1   3   1   4   3   2   8   1   2   1+  8   1   3   1
 [19]  5   1   1+  1   1   4   7   3   3   2+  3+  2  12   6   9   3   6   4
 [37]  3   1   1   2   1   3   2   3   2   4   6  10  11   5   4  12   5   3
 [55]  4   1   1  11   1   1   4   5   2   2   3   4   2   1  15   8   4  18
 [73]  3   5   5   3   2   5  10   1   1   7   1   2   9   2   2   2   1   2
 [91]  7   7   2   5   2   2   1   9  12   3
> position <- match(task.coxph$y[,1],task.survph$time)
> position
  [1]  2  1  1  8  1  3  1  4  3  2  8  1  2  1  8  1  3  1  5  1  1  1  1  4  7
 [26]  3  3  2  3  2 12  6  9  3  6  4  3  1  1  2  1  3  2  3  2  4  6 10 11  5
 [51]  4 12  5  3  4  1  1 11  1  1  4  5  2  2  3  4  2  1 13  8  4 14  3  5  5
 [76]  3  2  5 10  1  1  7  1  2  9  2  2  2  1  2  7  7  2  5  2  2  1  9 12  3
> H.hat <- H0.hat[position]*exp(task.coxph$linear.predictor)
>
> qqplot(rexp(100),H.hat,main="Diagnostic Plot")
> abline(a=0,b=1)
```


Diagnostic Plot
H.hat
rexp(100)
0
1
2
3
4
5

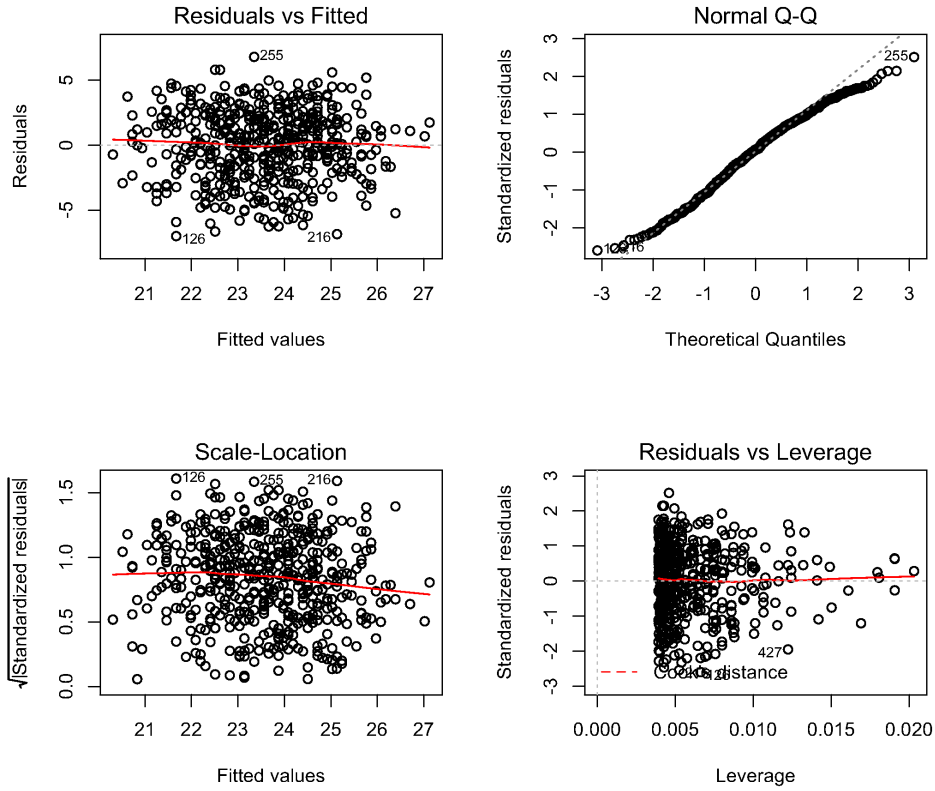
6

(a) Let $y_1, \ldots, y_n$ be realisations of n independent random variables from a finite normal mixture distribution with k components and k Gaussian densities where the means are different but the variances are equal. Explain how an E-M algorithm can be used to obtain the maximum likelihood estimates of the unknown parameters (i.e. the mixture probabilities, the means and the variance). Explicit expressions for the parameter updates in the M-step of the algorithm to compute the maximum likelihood estimates are required.

(b) A Research Nurse collects data on body mass index (defined as the ratio of weight in kg to height squared in m^2), age and gender from 500 patients in a GP practice. He is interested in determining whether patients, after taking account of their age (in years) and gender, can be clustered into different groups based on their body mass index (bmi). He realises that he does not know how he should go about addressing this problem and therefore approaches a member of staff at the Statistical Laboratory, who has many years of experience in applied research, with the anonymised data collected, `bmi.dat`. The Statistician begins by fitting a linear regression model, `bmi.lm`, to `bmi` adjusting for `age` and `gender`. After examining the results produced from this linear model and performing diagnostic checks using `plot(bmi.lm)`, the Statistician decides to examine the residuals further.

By looking at the provided R code, edited R output and plots from the Statistician's analysis, discuss what was done by the Statistician. You need to give sensible explanations for the analysis strategy adopted by the Statistician, including whether the underlying assumptions being made are justifiable. (**Detailed derivations of the underlying techniques are not required**.)

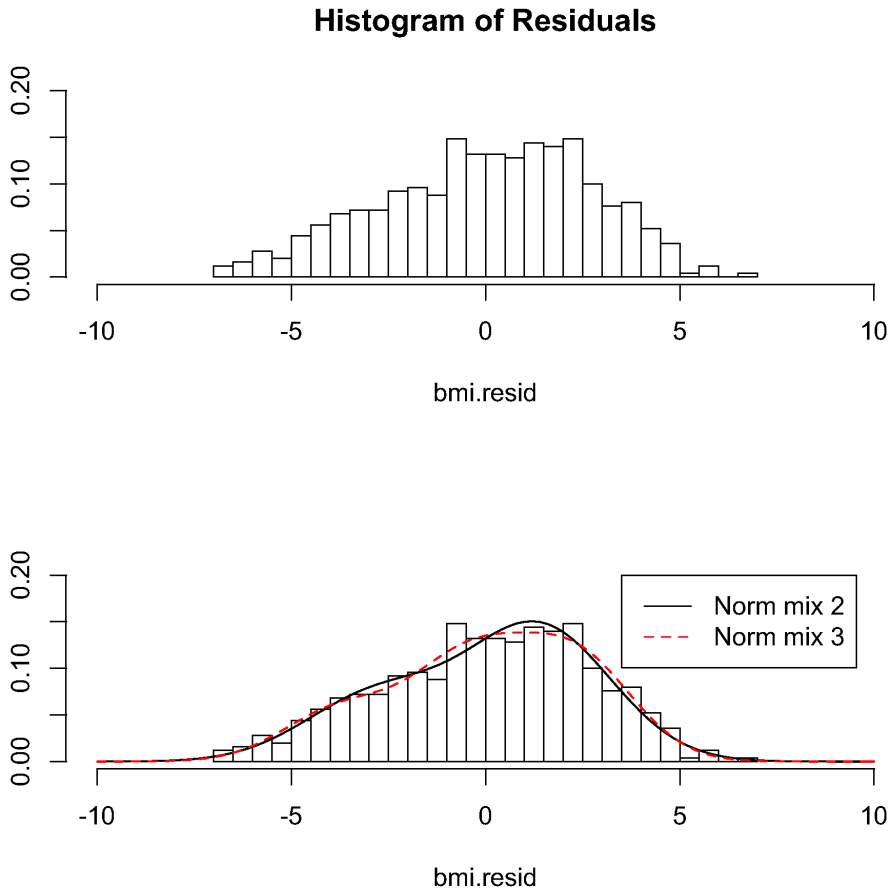
(c) Which of the three models used for analysing the residuals best fit these data? You need to justify your answer.




```
> bmi.lm <- lm(bmi~age+gender,data=bmi.dat)
> summary(bmi.lm)

Call:
lm(formula = bmi ~ age + gender, data = bmi.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9939 -1.8781  0.1901  2.0450  6.7899

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.01675    0.52180  51.776  < 2e-16 ***
age         -0.10483    0.01232  -8.506  < 2e-16 ***
gender       1.47031    0.24186   6.079 2.41e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.704 on 497 degrees of freedom
Multiple R-squared:  0.18,Adjusted R-squared: 0.1767
F-statistic: 54.53 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
> par(mfrow=c(2,2))
> plot(bmi.lm)  # command to produce the residual plots

> bmi.resid <- residuals(bmi.lm)
> summary(lm(bmi.resid~1))

Call:
lm(formula = bmi.resid ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9939 -1.8781  0.1901  2.0450  6.7899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.986e-17  1.207e-01       0        1

Residual standard error: 2.698 on 499 degrees of freedom

> logLik(lm(bmi.resid~1))
'log Lik.' -1205.25 (df=2)

> library(MASS)
> par(mfrow=c(1,2))
> truehist(bmi.resid,20,xlim=c(-10,10),ymax=0.2,col=0,main="Histogram of Residuals")

# E-M algorithm for Mixture of Normals with equal variances
> EM.Normeqvar(bmi.resid,pi=c(0.3,0.7),mu=c(-2,2),sigma2=3,iterations=100)
[1]  "pi1"     "pi2"    "mu1"      "mu2"     "sigma2"    "log-likelihood"
[1]  0.4007    0.5992   -2.4642    1.6477    3.2052      -1195.4134
[1]  0.3916    0.6084   -2.4932    1.6046    3.2649      -1194.9732
.
.
.
[1]  0.3317    0.6683   -2.7844    1.3819    3.4175      -1193.8265
[1]  0.3317    0.6683   -2.7844    1.3819    3.4175      -1193.8265

> EM.Normeqvar(bmi.resid,pi=c(0.2,0.4,0.4),mu=c(-3.7,-0.45,2.3),sigma2=2,iterations=100)
[1]  "pi1"    "pi2"    "pi3"   "mu1"    "mu2"    "mu3"    "sigma2"  "log-likelihood"
[1]  0.2050   0.3922   0.4029  -3.7564  -0.4556  2.3546   2.0585    -1192.3087
[1]  0.2044   0.3934   0.4022  -3.7545  -0.4559  2.3541   2.0729    -1192.2953
.
.
.
[1]  0.2051   0.3926   0.4023  -3.7306  -0.4417   2.3329   2.1448   -1192.2715
[1]  0.2051   0.3926   0.4023  -3.7306  -0.4417   2.3329   2.1448   -1192.2715

> x <- seq(-10,10,0.2)
```

```
> mixdens2 <- 0.3317*dnorm(x,-2.7844,sqrt(3.4175))+0.6683*dnorm(x,1.3819,sqrt(3.4175))
> mixdens3 <- 0.2051*dnorm(x,-3.7306,sqrt(2.1448))+0.3926*dnorm(x,-0.4417,sqrt(2.1448))
  +0.4023*dnorm(x,2.3329,sqrt(2.1448))
> bmi.mix2 <- list(x=x,y=mixdens2)
> bmi.mix3 <- list(x=x,y=mixdens3)
> truehist(bmi.resid,20,xlim=c(-10,10),ymax=0.2,col=0)
> lines(bmi.mix2,col=2,lty=2)
> lines(bmi.mix3,col=4,lty=3)
> legend(3.5,0.2,c("Norm mix 2", "Norm mix 3"),
   lty=c(1,2),col=c(1,2))
```



END OF PAPER