UNIVERSITY OF CAMBRIDGE

MAMA/218, NST3AS/218, MAAS/218

MAT3

MATHEMATICAL TRIPOS Part III

Tuesday, 6 June, 2023 9:00 am to 12:00 pm

PAPER 218

STATISTICAL LEARNING IN PRACTICE

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

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
Rough 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 In a survey 1308 individuals were asked the question: within the past 12 months, how many people they know personally that were victims of racial discrimination? A researcher was interested in how the answered number of people, given in the survey as `count`, is affected by the gender of individuals (1 for male, 0 for female) and if they belong to an ethnic minority (1 for yes, and 0 for no). The shortened `R` output of the analysis by the researcher is shown below.

```
> model1 <- glm(count ~ gender + minority,data=racism,family="poisson")
> summary(model1)
...
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2959     0.1161 -19.773   <2e-16 ***
gender       -0.1916     0.1473  -1.301    0.193
minority      1.7293     0.1466  11.796   <2e-16 ***
...
Null deviance: 962.8  on 1307  degrees of freedom
Residual deviance: 843.0  on 1305  degrees of freedom
AIC: 1122.3
...
```

(a) Write down algebraically the model fitted in `model1`, clearly defining all quantities. State the log-likelihood function and give the maximum likelihood estimates from the model output. Give an interpretation of the estimated `gender` coefficient.

The researcher then performed a statistical test and fitted a second model.

```
> X2 <- sum(residuals(model1, type="pearson")^2)
> 1-pchisq(X2,1305)
[1] 0
> model2 <- glm(count ~ gender + minority, data=racism, family="quasipoisson")
> summary(model2)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.2959     0.1522 -15.080   <2e-16 ***
gender       -0.1916     0.1931  -0.992    0.321
minority      1.7293     0.1922   8.996   <2e-16 ***
...
(Dispersion parameter for quasipoisson family taken to be 1.719346)

Null deviance: 962.8  on 1307  degrees of freedom
Residual deviance: 843.0  on 1305  degrees of freedom
AIC: NA
...
```

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) Explain the statistical test performed by the researcher. State the phenomenon behind the researcher's conclusion to fit the second model and state two possible causes for the phenomenon. Is the researcher's conclusion justified? Determine the value of `X2` from the model outputs.

(c) Which equation is solved by the estimated coefficients in `model2`? State a formula for the standard error of the `gender` coefficient in `model2`.

(d) Are the standard errors in `model2` trustworthy? Explain in detail how the parametric bootstrap can be used to compute the standard error for `gender` in `model1`.

The researcher then fitted a third model to the same data.

```
> model3 <- glmer(count ~ gender + (1 | minority),data=racism,family="poisson")
> summary(model3)
...
AIC      BIC   logLik deviance df.resid
1132.8   1148.3   -563.4   1126.8     1305
...
Random effects:
Groups   Name        Variance Std.Dev.
minority (Intercept) 0.7356   0.8577
Number of obs: 1308, groups:  minority, 2

Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4379     0.6141  -2.342   0.0192 *
gender       -0.1921     0.1469  -1.308   0.1910
...
```

(e) Explain why the researcher may have fitted `model3`. Is this model a good fit for the data?

(f) State two methods for deciding between `model1` and `model3` on this data. Can these methods also be used to decide between `model2` and `model3`?

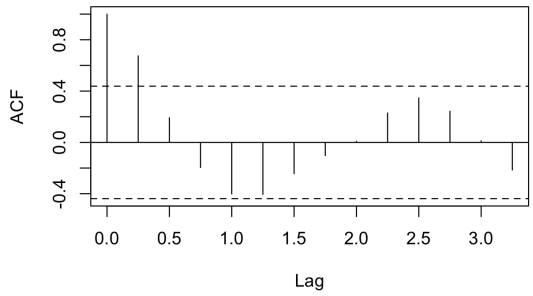
2 A company is interested in the number of calls to their customer service. In one file, they have collected the (logarithm of the) number of calls every 15 minutes between 7am and noon on one day, and analysed these data using the following `R` commands.

```
> acf(calls)
> auto.arima(calls, ic="aic")
Series: calls
ARIMA(1,0,1) with zero mean

Coefficients:
      ar1     ma1
      0.6997  0.9510
s.e.  0.1796  0.3287

sigma^2 = 0.4944:  log likelihood = -22.23
AIC=50.46   AICc=51.96   BIC=53.45
```

This is the output of the `acf` command.



(a) State the definition of a (weakly) stationary process.

(b) What are the values plotted in the `acf(calls)`-figure? How is the dashed line computed? With reference to the plot, explain why the number of calls is unlikely to follow a white noise process. Suggest a possible ARMA(p, q) model supported by the plot.

(c) Write down algebraically the model fitted by the `auto.arima` function. What are the parameter maximum likelihood estimates? Construct a 95% confidence interval for the `ma1` parameter from the model output and explain shortly why this interval is likely too narrow.

(d) Consider now a white noise process $W \sim WN(0, \sigma_W^2)$, $\sigma_W^2 > 0$. Let $X = (X_t)_{t\in\mathbb{Z}}$ be a causal AR(1) process such that for $\phi \in \mathbb{R}$ and another white noise process $\varepsilon \sim WN(0, \sigma^2)$, $\sigma^2 > 0$,

$$X_t = \phi X_{t-1} + \varepsilon_t.$$

Let $\mathbb{E}[\varepsilon_t W_t] = 0$ for all $t \in \mathbb{Z}$ and consider the time series $Y = (Y_t)_{t\in\mathbb{Z}}$ with $Y_t = X_t + W_t$.

(i) What is the possible range of values ϕ for X to be causal? State the autocovariance function of X in terms of ϕ and σ^2.

[QUESTION CONTINUES ON THE NEXT PAGE]

(ii) Show that Y is stationary and find its autocovariance function.

(iii) Show that the autocovariance function of the time series $(U_t)_{t\in\mathbb{Z}}$ with $U_t = Y_t - \phi Y_{t-1}$ vanishes for lags $h > 1$.

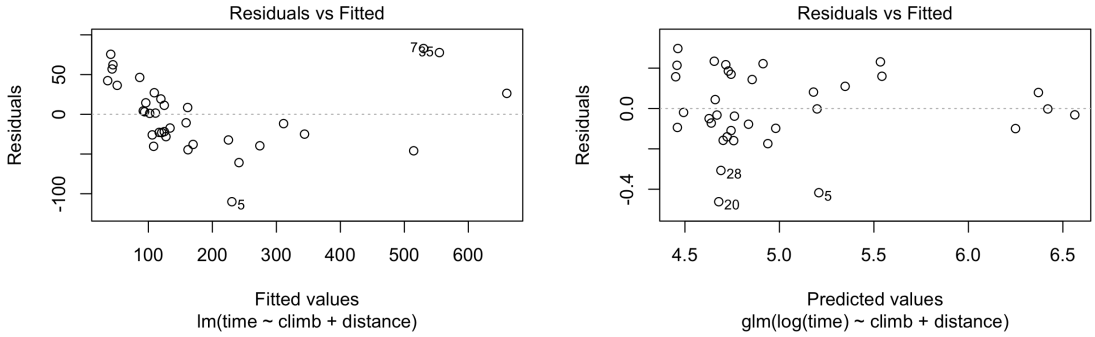
(iv) Conclude by (iii) that Y is an ARMA$(1,1)$ process. You may use without a proof the fact that a time series, whose autocorrelations vanish for $h > 1$, can be represented as an MA(1) process.

3 A runners association has collected record times for 35 races in the past year. A table contains for each race the record time (in minutes), as well as the distance of the race (in miles) and the cumulative climb (in thousands of feet). A data analyst was interested in explaining the record times based on the distance and the cumulative climb. The data analyst first standardised the distance and climb columns to have mean zero and unit standard deviation, and then fitted three models. A (shortened) version of the `R` output is shown below.

```
> model1 <- lm(time ~ climb + distance, data=races)
> model2 <- glm(log(time) ~ climb + distance, data=races)
> model3 <- glm(time ~ climb + distance, family = Gamma(link=log), data=races)
> summary(model1)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   186.51       7.50  24.867  < 2e-16 ***
climb          42.89      10.04   4.272 0.000162 ***
distance      125.75      10.04  12.525 7.04e-14 ***
...
> summary(model2)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.99902    0.03190 156.699  < 2e-16 ***
climb        0.22514    0.04271   5.272 9.01e-06 ***
distance     0.41135    0.04271   9.632 5.62e-11 ***
...
(Dispersion parameter for gaussian family taken to be 0.035621)
...
> summary(model3)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.01475    0.03046 164.609  < 2e-16 ***
climb        0.22332    0.04078   5.476 4.97e-06 ***
distance     0.40868    0.04078  10.021 2.15e-11 ***
...
> AIC(model2, model3)
df       AIC
model2  4 -12.52943
model3  4 336.38099
> plot(model1, which=1, add.smooth=FALSE)
> plot(model2, which=1, add.smooth=FALSE)
```

[QUESTION CONTINUES ON THE NEXT PAGE]

The output of the diagnostic plots is as follows:



(a) State algebraically the three fitted models `model1`, `model2` and `model3`.

(b) Discuss if any of the assumptions for fitting the first two models are violated according to the diagnostic plots. Which model seems to fit the data better?

(c) Define the AIC criterion for a statistical model, introducing all necessary notation.

(d) Why would the data analyst conclude from the AIC values in the `R` output that `model2` fits the data better than `model3`? Explain why the AIC value for the transformed data in `model2` and the conclusion of the data analyst are not correct. Suggest a corrected AIC value for `model2` from its model output.

(e) Show that when the fitted values are close to the observed values in `model3`, then the estimated coefficients in `model2` and `model3` are close. [Hint: You may use without proof that a Gamma GLM has variance function $V(\mu) = \mu^2$, where μ is the mean function, and weights $w_i = 1$.]

(f) Why should we recommend `model3` to the data analyst instead of transforming the response variables as in `model2`?

4 An advertiser wanted to understand how different web design decisions influence the effectiveness of an online advertisement. They showed the advertisement to all users visiting the website while varying the font typeface (cagegorical variable `font`, with two levels `sanserif` and `serif`), display style (variable `display`, with two levels `banner` and `popup`) and seriousness of writing (`writing`, numerical variable taking values between 1 and 10). For each user, the advertiser recorded whether the advertisement was clicked. Here is a snippet of the data.

```
> head(ad)
    click  font       display   writing
1    yes   serif      banner      3
2    no    sanserif   banner      8
3    no    sanserif   banner      1
4    no    serif      popup       7
5    no    serif      popup       2
6    no    serif      popup       9
```

The advertiser then used the following commands in `R`:

```
> x <- model.matrix(~font*display, data=ad)[, -1]
> y <- model.matrix(~click-1, data=ad)
> ad.fit1 <- keras_model_sequential() %>%
layer_dense(units = 2, activation = 'relu', input_shape = dim(x)[2]) %>%
layer_dense(units = 2, activation = 'softmax') %>%
compile(optimizer='sgd', loss='categorical_crossentropy') %>%
fit(x, y, batch_size=1, epochs=5)
> ad.fit2 <- keras_model_sequential() %>%
layer_dense(units = 2, activation = 'softmax', input_shape = dim(x)[2]) %>%
compile(optimizer='sgd', loss='categorical_crossentropy') %>%
fit(x, y, batch_size=1, epochs=5)
```

(a) Sketch a diagram of the neural network fitted in `ad.fit1`. Write down algebraically the neural network model, clearly defining all necessary quantities, associating `sanserif` and `serif` to 0 and 1, `banner` and `popup` to 0 and 1, and `no` and `yes` to 0 and 1. How many parameters are there in all?

(b) State the fitted neural network classifier and the loss function used in fitting `ad.fit1`. Let $\hat{\beta}$ be the vector of all estimated coefficients. Show that $\hat{\beta}$ is not unique if no restrictions are imposed on it.

(c) State a logistic classifier giving the same predictions as the classifier fitted in `ad.fit2`. Does this solve the corresponding non-uniqueness mentioned in (b)? Justify your answer. Discuss why the logistic classifier is preferable over `ad.fit2`.

(d) Assume that all weights in the neural network in `ad.fit1` are initialised to be equal to 1 and that stochastic gradient descent is applied with constant learning rate $\gamma = 1$ and without randomly shuffling the data. What are the values of the parameters after one iteration (i.e., after one batch in the first epoch)?

5 Suppose we are given $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$. An elastic net fits coefficients to this data set for $\lambda_1, \lambda_2 \geqslant 0$ by

$$\hat{\beta}^E \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left(\|Y - X\beta\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right).$$

(a) Discuss how the solutions behave as i) $\lambda_1 \to 0$ or $\lambda_1 \to \infty$ for λ_2 fixed, ii) $\lambda_2 \to 0$ for λ_1 fixed, iii) $\lambda_1, \lambda_2 \to 0$. Is the solution $\hat{\beta}^E$ unique when $\lambda_2 > 0$?

(b) Describe how coordinate descent can be applied to compute $\hat{\beta}^E$.

(c) Suppose now that $X^\top X = I_p$. Compute a solution $\hat{\beta}^E$. [Hint: Let $\hat{\beta}^E = (1 + \lambda_2)^{-1/2}\hat{\beta}$, where $\hat{\beta}$ solves a penalised least-squares problem in an extended space with response $\tilde{Y} \in \mathbb{R}^{n+p}$ and ℓ_1-penalty only. You may find the function $S_\lambda(u) = \operatorname{sign}(u)\max(|u| - \lambda/2, 0)$ useful.]

A researcher wanted to understand from a data set how a clinical indicator of prostate cancer depends on patient characteristics and levels of a number of prostate-specific antigens. For this purpose, the researcher fitted three different elastic nets (`m1` with $\lambda_1 = 0$ and $\lambda_2 = 1$, `m2` with $\lambda_1 = 1$ and $\lambda_2 = 0$, and `m3` with $\lambda_1 = \lambda_2 = 1/2$), and obtained the following coefficient estimates for $\hat{\beta}^E$.

```
> coefficients
          m1    m2    m3
lcavol   0.330 0.307 0.336
lweight  0.515 0.000 0.389
age     -0.004 0.000 0.000
lbph     0.111 0.000 0.035
svi      0.541 0.000 0.369
lcp      0.015 0.000 0.000
gleason  0.063 0.000 0.000
pgg45    0.004 0.000 0.002
> cor(svi, lcavol)
[1] 0.54
```

(d) Name the estimators for the coefficients in `m1` and `m2`. Use the table, the correlation output and your result in (c) to explain how the coefficients estimated in `m3` compare to the estimates in `m1` and `m2`.

(e) State briefly procedures for comparing elastic nets with different λ_1 and λ_2, and for obtaining confidence intervals for the non-zero coefficients in `m3`.

6 Suppose we are given observations $(X_i, Y_i)_{i=1}^n$, where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. Let $x \in \mathbb{R}^p$ be a covariate we want to classify as 0 or 1.

(a) State the regression functions in the context of this classification problem. Define the Bayes classifier and the Bayes decision boundary.

(b) State the kNN classifier for $k \geqslant 1$ and discuss how the choice of k relates to its bias, its variance and the smoothness of its decision boundary.

(c) Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be an inner product space with induced norm $\|\cdot\|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}^{1/2}$ and let $\phi : \mathbb{R}^p \to \mathcal{H}$ be a feature map such that $\sum_{i=1}^n \phi(X_i) = 0$. Consider the optimisation problem

$$\hat{u} \in \operatorname*{argmax}_{u \in \mathcal{H}, \|u\|_{\mathcal{H}}=1} \frac{1}{n} \sum_{i=1}^n \langle u, \phi(X_i) \rangle_{\mathcal{H}}^2.$$

(i) Assuming that $\phi(x) = x$ and $\mathcal{H} = \mathbb{R}^p$, state an equivalent optimisation problem solved by PCA and compute the solution $\hat{u}$ in terms of the data X_i.

(ii) For general ϕ, you may assume without proof that any solution satisfies $\hat{u} = \sum_{i=1}^n \hat{\alpha}_i \phi(X_i)$ for $\hat{\alpha} \in \mathbb{R}^n$. With the matrix $K = (\langle \phi(X_i), \phi(X_j) \rangle_{\mathcal{H}})_{i,j=1}^n$, show that

$$\hat{\alpha} \in \operatorname*{argmax}_{\alpha \in \mathbb{R}^n,\, \alpha^\top K \alpha = 1} \alpha^\top K^2 \alpha,$$

and compute the solution $\hat{\alpha}$ explicitly in terms of K.

(iii) Using (ii), discuss why PCA can be computed efficiently for the transformed features $\phi(X_i)$ and how this can be used to improve the kNN classifier. How is this related to the kernel trick?

END OF PAPER