UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 7 June, 2018 9:00 am to 12:00 pm

PAPER 218

STATISTICAL LEARNING IN PRACTICE

*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 Researchers wanted to understand the effect of different diets on rat weight growth over time. They assigned each of 16 rats to one of three diets (coded 1, 2 and 3) and measured the body weight of each rat (in grams) on day 1 and every seven days thereafter until day 71. Their study data is recorded in the `BodyWeight` dataset and analysed with the following `R` code.

```
> BodyWeight
##     weight Time Rat Diet
## 1      240    1   1    1
## 2      250    8   1    1
## 3      255   15   1    1
## ...
## 174    550   50  16    3
## 175    553   57  16    3
## 176    569   64  16    3

> library(lme4)
> bw.lm1 <- lm(weight ~ Time + Diet, data = BodyWeight)
> bw.lm2 <- lm(weight ~ Time + Diet + Rat, data = BodyWeight)
> bw.lme1 <- lmer(weight ~ Time + Diet + (1 | Rat), data = BodyWeight, REML = TRUE)
> summary(bw.lme1)
## Linear mixed model fit by REML ['lmerMod']
## Formula: weight ~ Time + Diet + (1 | Rat)
##    Data: BodyWeight
##
## REML criterion at convergence: 1304.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5236 -0.5557 -0.0467  0.5667  3.0932
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Rat      (Intercept) 1337.88  36.577
##  Residual               66.85   8.176
## Number of obs: 176, groups:  Rat, 16
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 244.06890   13.00477  18.768
## Time          0.58568    0.03168  18.486
## Diet2       220.98864   22.44958   9.844
## Diet3       262.07955   22.44958  11.674
##
## Correlation of Fixed Effects:
##       (Intr) Time   Diet2
## Time  -0.082
## Diet2 -0.575  0.000
```

```
## Diet3 -0.575  0.000  0.333
```

(a) Which modelling assumption is violated for the model `bw.lm1` in this dataset? Why is model `bw.lm2` non-identifiable?

(b) Write down algebraically the model fitted in `bw.lme1` and estimated values of all parameters in the model. How do you interpret the fixed effect coefficient of `Time` in the `R` output?

(c) Let X be the design matrix of model `bw.lm1`. Let A be a matrix whose columns form an orthonormal basis of the orthogonal complement of the column space of X. Describe, with reference to matrix A, the objective function maximised by the random effect estimates for `Rat` and `Residual` in the `R` output.

(d) To test whether the random effect in model `bw.lme1` is necessary, they carried out a likelihood ratio test using the following `R` commands.

```
> test_stat <- 2*(logLik(bw.lme1) - logLik(bw.lm1))
> pval <- 1 - pchisq(test_stat, df = 1)
```

Why is the test carried out above is not valid? Explain in detail how a valid test can be carried out.

2 Suppose we have observations $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_1, \ldots, x_n \in \mathbb{R}^p$ are covariates and $y_1, \ldots, y_n \in \{-1, 1\}$ are associated class labels.

(a) Define *positive definite kernels* on $\mathbb{R}^p$. Show that the linear kernel $k : (x, x') \mapsto x^\top x'$ is a positive definite kernel on $\mathbb{R}^p$.

(b) Write down the optimisation problem solved by the soft-margin support vector machine with a linear kernel. What is the set of support vectors in this support vector machine? [You may assume that the optimisation problem has a unique optimiser.]

(c) Describe how leave-one-out cross-validation (with respect to the binary misclassification loss) can be implemented to choose the soft-margin tuning parameter.

(d) Show that the decision boundary of the support vector machine is unchanged if we remove any non-support vector from the training data. Hence or otherwise, show that the leave-one-out cross-validation error $\mathrm{err}_{\mathrm{CV}}$ satisfies

$$\mathrm{err}_{\mathrm{CV}} \leqslant \frac{s}{n},$$

where s is the number of support vectors.

3 A research team in financial company wanted to understand risk factors for loan defaults. They looked at $n = 3000$ loans issued in 2005 and collected information on client age (in years), yearly income (in £1000s), loan amount (in £1000s) and whether or not a default had occurred (coded 1 for yes and 0 for no) within 10 years of issuance. They first fit the following model in `R`.

```
> head(loan)
##   default income  age amount
## 1       0   59.2 49.5   31.5
## 2       0   81.8 47.3    5.9
## 3       1   57.6 45.9   33.7
## 4       0   25.0 53.1   13.3
## 5       1   76.4 23.3   29.5
## 6       0   63.2 39.7   20.3
> model1 <- glm(default~income+age+amount, family="binomial", data=loan)
> summary(model1)
## Call:
## glm(formula = default ~ income + age + amount, family = "binomial",
##       data = loan)
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7126  -0.5868  -0.5440  -0.4876   2.1978
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.322701   0.204388  -6.472  9.7e-11 ***
## income      -0.005916   0.002292  -2.581 0.009858 **
## age         -0.008580   0.003833  -2.239 0.025179 *
## amount       0.012981   0.003787   3.428 0.000609 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2480.1  on 2999  degrees of freedom
## Residual deviance: 2462.7  on 2996  degrees of freedom
##
## Number of Fisher Scoring iterations: 4
```

(a) Write down algebraically the model fitted in `model1` and estimated coefficients. Compute an approximate 95% confidence interval of the coefficient for `age`.

(b) Suppose that the company wanted to use `model1` as a classifier and classify a loan as risky if the predicted probability of default is at least $p \in (0, 1)$. Write down the decision boundary for this classifier.

The research team then fitted a neural network model using the following commands.

```
> library(keras)
```

```
> x <- model.matrix(~income+age+amount-1, data=loan)
> y <- model.matrix(~as.factor(default)-1, data=loan)
> layer1 <- layer_dense(units = 2, activation = 'sigmoid', input_shape = dim(x)[2])
> layer2 <- layer_dense(units = 2, activation = 'softmax')
> model2 <- keras_model_sequential(list(layer1, layer2))
> compile(model2, optimizer='sgd', loss='categorical_crossentropy', metrics='acc')
> fit(model2, x, y, batch_size=1, epochs=5)
```

(c) Draw the architecture of the neural network in `model2` and write down the model algebraically.

(d) Describe how stochastic gradient ascent can be carried out to maximise the log-likelihood, specifying explicitly how relevant gradients with respect to model coefficients are computed. [You may assume that model coefficients have been appropriately initialised.]

4 In a clinical trial, 107 patients suffering from epilepsy were randomised to receive either the anti-epileptic drug Progabide or a placebo. A researcher was interested to understand how the post-treatment seizure counts over a one-week period (`seizure`) depend on the age of patients (`age`), baseline seizure count of the patients prior to the study (`baseline`) and treatment option (`treatment`). She fitted the following model in `R`.

```
> head(epilepsy)

##   age baseline treatment seizure
## 1  33       16 Progabide       8
## 2  16       56   Placebo      14
## 3  27       22   Placebo       6
## 4  30       55   Placebo      25
## 5  35       32 Progabide       5
## 6  26       87   Placebo       9

> epilepsy.glm <- glm(seizure ~ age + treatment * baseline, family = "poisson",
      data = epilepsy)
> summary(model1)

## Call:
## glm(formula = seizure ~ age + treatment * baseline, family = "poisson",
##     data = epilepsy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2697  -1.3423  -0.2103   0.6166   4.4431
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.1215275  0.1742522   6.436 1.22e-10 ***
## age                         0.0090311  0.0051408   1.757 0.078962 .
## treatmentProgabide         -0.3592068  0.1117255  -3.215 0.001304 **
## baseline                    0.0186824  0.0011610  16.092  < 2e-16 ***
## treatmentProgabide:baseline 0.0008682  0.0013767   0.631 0.528303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1014.69  on 106  degrees of freedom
## Residual deviance:  319.86  on 102  degrees of freedom
## AIC: 681.15
##
## Number of Fisher Scoring iterations: 5
```

(a) Let Y_i be the post-treatment seizure count for patient i. Write down the algebraic form of the model for $Y_1, \ldots, Y_n$ that has been fitted in `epilepsy.glm`. Write down

the log-likelihood function for this model and the corresponding maximum likelihood estimates. How do you interpret the estimated coefficient for the interaction term `treatmentProgabide:baseline`?

(b) Write down the algebraic form of the null model associated with 'null deviance' in the `R` output. Show that the null deviance is equal to

$$2\sum_{i=1}^{n} Y_i \log(Y_i/\bar{Y}),$$

where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$.

(c) Is the model `epilepsy.glm` a good fit? Support your conclusion with a relevant goodness-of-fit test, stating without proof any relevant theory used. The researcher decided to improve the model fit by using a quasi-Poisson model to account for overdispersion. How is the dispersion parameter estimated in a quasi-Poisson model? If the estimated dispersion parameter is 3.24, is the treatment effect of Progabide statistically significant at 5% level under the quasi-Poisson model?

5 A meteorologist wanted to model mean daily temperature (in degrees Celsius) at a location in Cambridge for $n = 31$ days in December. Let X_t denote the mean temperature measurement on the tth day. From prior experience, she wanted to fit a stationary Gaussian AR(1) model with a non-zero mean.

(a) Show that the Yule–Walker estimator of the autoregressive coefficient in this model is equal to the lag 1 sample autocorrelation.

(b) Write down the likelihood function of relevant parameters involved in this model.

While handling data, the meteorologist accidentally misplaced the data file for 2 Dec and she now only had data for the remaining 30 days.

(c) What is the distribution of X_2 conditional on $X_1, X_3, X_4, \ldots, X_n$?

(d) Describe how she could use an expectation-maximisation algorithm to impute the missing data and find (local) maximum likelihood estimators of the AR(1) model parameters. Specify in your description how the expression of the expectation step can be computed.

6 A statistician was interested in performing a bivariate linear regression

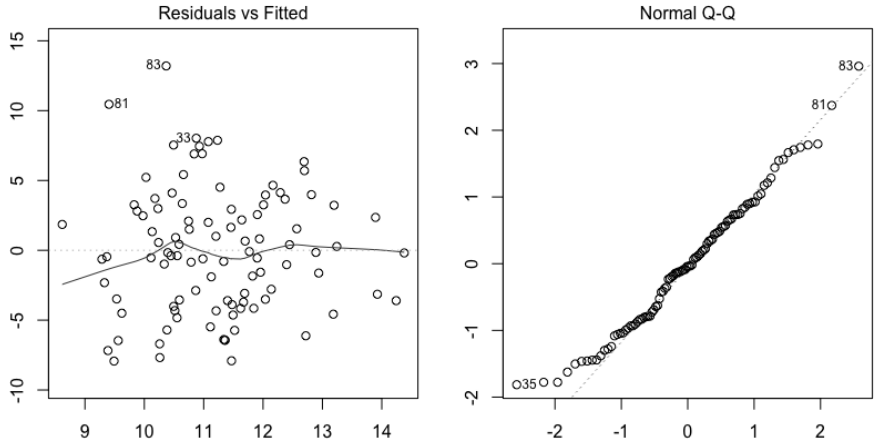
$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$$

The response variable and covariates were collected in the dataset `dat` with each row representing an observation. Columns `x1` and `x2` in the dataset were both standardised to have mean 0 and Euclidean norm $\sqrt{n}$. A snippet of her analysis in `R` is shown below.

```
> head(dat)
##       y    x1    x2
## 1 16.08 -0.46 -0.41
## 2  7.01 -1.59 -1.49
## 3 10.54  0.00 -0.16
## 4  7.96  0.31  0.18
## 5 13.52  1.72  1.65
## 6 12.84 -0.34 -0.22

> cor(dat)
##            y        x1        x2
## y  1.0000000 0.2517297 0.2412650
## x1 0.2517297 1.0000000 0.9813846
## x2 0.2412650 0.9813846 1.0000000

> model1 <- lm(y~x1, data=dat)
> model2 <- lm(y~x2, data=dat)
> model3 <- lm(y~x1+x2, data=dat)
> plot(model3, which=c(1,2))

> summary(model1)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.2219     0.4476  25.070   <2e-16 ***
## x1            1.1586     0.4500   2.575   0.0115 *
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(model2)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.2224     0.4488  25.003   <2e-16 ***
## x2            1.1101     0.4511   2.461   0.0156 *
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(model3)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.2217     0.4497  24.953   <2e-16 ***
```

```
## x1            1.8663     2.3538   0.800     0.43
## x2           -0.7209     2.3532  -0.306     0.76
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(model2, model3)
## output omitted
```

The output of the diagnostic plots in the above code are as follows.



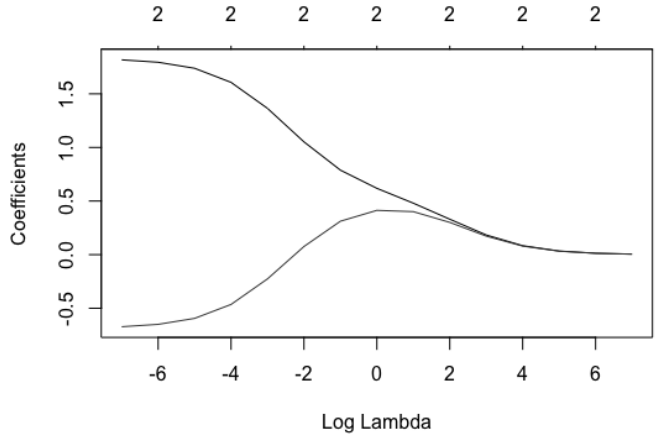
(a) How are the x- and y-coordinate values of points in the two diagnostic plots computed? Do you spot any violation of modelling assumptions in `model3` from these plots?

(b) Why are covariates `x1` and `x2` insignificant in the bivariate model `model3`, despite being significant individually in univariate regressions `model1` and `model2`?

(c) Explain the test that is carried out by the `anova(model2, model3)` command, specifying the null and alternative hypotheses, the expression for the test statistic and how the p-value is computed. Write down numerical values for both the test statistic and p-value of this test.

The statistician then fitted a ridge regression using the following `R` commands.

```
library(glmnet)
X <- model.matrix(y~x1+x2, data=dat)
model4 <- glmnet(X, dat$y, alpha=0, lambda=exp(7:(-7)))
plot(model4, xvar='lambda')
```



(d) Write down the optimisation problem solved by `model4` in a penalised form in terms of a regularisation parameter λ. The solution path of the ridge regression is shown in the figure above. What are the asymptotes of the two curves as the horizontal axis of the figure approaches $+\infty$ and $-\infty$?

(e) Let $\hat{\beta}^{\mathrm{r}}_{1,\lambda}$ and $\hat{\beta}^{\mathrm{r}}_{2,\lambda}$ be the ridge regression estimator of β_1 and β_2. Write down expressions for $\mathrm{var}(\hat{\beta}^{\mathrm{r}}_{1,\lambda})$ and $\mathrm{var}(\hat{\beta}^{\mathrm{r}}_{2,\lambda})$ (you do not need to carry out the computation). Are they useful in constructing confidence intervals for β_1 and β_2? Why or why not?

(f) Suggest a suitable criterion for selecting among linear models `model1`, `model2` and `model3`. What is a suitable criterion for selecting the tuning parameter λ in the ridge regression model?

END OF PAPER