STATISTICAL LEARNING IN PRACTICE Part III / Lent 2020
Example Sheet 2 (of 4) By courtesy of Dr T. Wang

You have the option to submit your answers to questions 2 and 5 to be marked. If you wish your answers to be marked, please leave them in my pigeon-hole in the central core of the CMS by 11am on 26$^{\text{th}}$ February.

1. Consider the model
$$Y = \alpha \mathbf{1}_n + X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I).$$
   (i) Suppose $Y = (Y_1, \ldots, Y_n)^\top$. Show that in both ridge regression and Lasso regression, we estimate α by $\hat{\alpha} = n^{-1} \sum_{i=1}^n Y_i$.
   (ii) Assume that the design matrix $X \in \mathbb{R}^{n \times p}$ has zero mean columns and orthogonal design, i.e. $X^\top X = I_p$ (so, in particular, $p \leq n$). Derive explicit expressions for the ridge regression estimator $\hat{\beta}_\lambda^{\text{ridge}}$ and the Lasso regression estimator $\hat{\beta}_\lambda^{\text{Lasso}}$.
   (iii) Explain how gradient descent can be applied to find the ridge regression estimator $\hat{\beta}_\lambda^{\text{ridge}}$.

2. A company is investigating the effectiveness of a new pesticide. The researchers set up the following experiments. 30 adult whiteflies were put in each of 20 clip-on leaf cages. Each cage was attached to a different plant. 10 of these cages were irrigated with the new pesticide, while the other 10 were irrigated with an older product. The response variable for the experiment was the number of dead whiteflies after a week. The goal of the researchers was to investigate the probability of death for the whiteflies, with the hope that the new pesticide would lead to a higher death rate. They decided to fit a binomial regression model (a generalized linear model with a binomial distribution for the response) with the pesticide factor as predictor.

   (i) Write down the algebraic form of this model.

   After fitting the model, they realized that in the data there were many more zeros (no dead flies after a week) than expected from the model. They assumed that for some cages the pesticide had been washed away by some external factor before it could act. They ask you how it is possible to analyse these data taking into account this possibility.

   (ii) Suggest an appropriate model to address this problem.
   (iii) Write down the likelihood and the augmented likelihood of this model.
   (iv) Describe how this model can be fitted with an expectation-maximisation algorithm.

3. A researcher collected a dataset of 40 patients to analyse the recurrence of heart attack after a first episode. The variables in the dataset are:

   - `ha2`: a binary variable which assumes value 1 if the patient has a second heart attack after the first episode and 0 if the patient has no additional episodes.

- `anxiety`: a continuous variable which measures the level of anxiety of the patients.
- `treatment`: a binary variable which assumes value 1 if the patient completed an anger management treatment, 0 otherwise.

The date are analysed with the following `R` code:

```
model1 <- glm(ha2 ~ anxiety + treatment, family=binomial, data=heart.attack)
summary(model1)
```

```
# Coefficients:
#              Estimate  Std.  Error  z value  Pr(> |z|)
#  (Intercept)  -6.38342      2.50468   -2.549  0.01082  *
#      anxiety   0.13970      0.04819    2.899  0.00374  **
#    treatment  -2.73309      1.00548   -2.718  0.00656  **
#
# - - -
# Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance:  95.452 on ?  degrees of freedom
# Residual deviance:  69.753 on ?  degrees of freedom
#
# AIC: 135.75
#
# Number of Fisher Scoring iterations:  5
```

(i) Write down the algebraic form of the model that has been fitted and the estimates for the parameters of the model.

(ii) Give an interpretation of the role of anxiety and of the treatment in the probability of a second heart attack.

(iii) What are the degrees of freedom that have been substituted by question marks in the output?

(iv) How should the `R` syntax above be changed to fit the null model that corresponds to the 'null deviance' in the output? What is the corresponding AIC value for that model?

(v) After seeing the output, the researcher then fitted a quasibinomial model to the same dataset. Is there sufficient evidence for doing so? Which of the parameters will be significant at 5% level in the quasibinomial model?

4. Suppose we have a random intercept linear mixed effect model with a single covariate

$$Y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij},$$

where $b_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ is independent from $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, m$, $j = 1, \ldots, \ell_i$.

(i) Write down the log-likelihood for this model.

(ii) We say the design is *balanced* if $\ell_1 = \cdots = \ell_m =: \ell$ and $x_{1j} = \cdots = x_{mj}$ for all $j = 1, \ldots, \ell$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be maximum likelihood estimators for β_0 and β_1 in this balanced-design single-covariate linear mixed effect model. Show that

$$\hat{\beta}_0 = \frac{1}{m}\sum_{i=1}^{m} \hat{\beta}_0^{(i)} \quad \text{and} \quad \hat{\beta}_1 = \frac{1}{m}\sum_{i=1}^{m} \hat{\beta}_1^{(i)}, \tag{$\star$}$$

where $(\hat{\beta}_0^{(i)}, \hat{\beta}_1^{(i)})$ are OLS estimators for regressing $Y_{i1}, \ldots, Y_{i\ell}$ against $X_{i1}, \ldots, X_{i\ell}$.

(iii) Would $(\star)$ still hold if an additional covariate z_{ij} is included in the mixed effect model (i.e., if $Y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \beta_2 z_{ij} + \epsilon_{ij}$)?

(iv) 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, how the REML estimates of σ^2 and τ^2 can be obtained. Show that the REML estimates do not depend on the choice of A.

5. Question 1 of the 2017–2018 past paper. You may find it in the following link:

   https://www.maths.cam.ac.uk/postgrad/part-iii/files/pastpapers/2018/paper_218.pdf.

6. Contagious bovine pleuropneumonia (CBPP) is a major disease of cattle in Africa. Researchers wanted to study the within-heard spread of CBPP in newly infected herds. Blood samples were quarterly collected in 15 newly infected herds to determine the number of infected animals. They performed data analysis using the following R code.

```
head(cbpp)
 #     herd  incidence  size  period
 #  1     1          2    14       1
 #  2     1          3    12       2
 #  3     1          4     9       3
 #  4     1          0     5       4
 #  5     2          3    22       1
 #  6     2          1    18       2
cbpp.glmm <- glmer(incidence / size ~ period + (1 | herd), weights = size,
family = binomial, data = cbpp)
summary(cbpp.glmm)
 #    AIC    BIC  logLik  deviance  df.resid
 #  194.1  204.2   -92.0     184.1        51
 #
 # Scaled residuals:
 #      Min       1Q   Median       3Q      Max
 #  -2.3816  -0.7889  -0.2026  0.5142  2.8791
 #
 # Random effects:
```

```
#  Groups         Name  Variance  Std.Dev.
#    herd  (Intercept)    0.4123    0.6421
# Number of obs:  56, groups:  herd, 15
#
# Fixed effects:
#               Estimate  Std.  Error  z value   Pr(>|z|)
#  (Intercept)   -1.3983       0.2312   -6.048  1.47e-09  ***
#      period2   -0.9919       0.3032   -3.272  0.001068  **
#      period3   -1.1282       0.3228   -3.495  0.000474  ***
#      period4   -1.5797       0.4220   -3.743  0.000182  ***
# - - -
# Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
```

(i) Write down the algebraic form of the model fitted and estimated coefficients.

(ii) How do you interpret the fixed effect coefficients in the R output?

(iii) Describe how parametric bootstrap may be used to estimate the standard error of the random effect coefficient estimator.

7. (*Exercise with R*) Download the `leukaemia` dataset from the website:

   https://raw.githubusercontent.com/AJCoca/SLP19/master/

   This dataset contains tumour mRNA samples from 38 patients with leukaemia. The first column encodes the type of leukaemia: 27 acute lymphoblastic leukaemia (ALL) cases (code 0) and 11 acute myeloid leukaemia (AML) cases (code 1). The remaining columns contain gene expression levels for 3051 different genes measured.

   (i) Relate the leukemia subtype and the gene expression levels by a logistic regression model. Fit this model by means of regularised maximum likelihood, employing the ridge regularisation with penalty parameter $\lambda = 1$.

   (ii) Obtain the fits from the regression model. The fit is almost perfect. Could this be due to overfitting the data, or could it be that the biological information in the gene expression levels indeed determines the leukemia subtype almost perfectly? To discern between the two explanations for the almost perfect fit, randomly shuffle the responses. Refit the logistic regression model and obtain the fits. On the basis of this and the previous fit, which explanation is more plausible?

   (iii) Compare the fit of the logistic model with different penalty parameters, say $\lambda = 1$ and $\lambda = 1000$. How does λ influence the possibility of overfitting the data?

   (iv) Explain why a Lasso penalty might be more suitable for this dataset. Fit the Lasso regression in R. Explain how you can use cross-validation to choose the tuning parameter.