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 26th 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, $p \le n$). 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 $\hat{\beta}_\lambda^{\text{ridge}}$.

Solution

(i) In both ridge regression and Lasso regression, we obtain the estimators $\hat{\alpha}, \hat{\beta}$ by maximising an objective function of the form
$$L(\alpha, \beta) = \|Y - \alpha\mathbf{1}_n - X\beta\|_2^2 + \text{pen}(\beta),$$
where $\text{pen}(\beta)$ is a penalty/regularisation term that depends only on β (and a regularisation parameter). The optimal α choice is obtained by solving
$$0 = \frac{\partial L}{\partial \alpha}\bigg|_{\hat{\alpha},\hat{\beta}} = 2\mathbf{1}_n^\top(Y - \hat{\alpha}\mathbf{1}_n - X\hat{\beta}) = 2\mathbf{1}_n^\top(Y - \hat{\alpha}\mathbf{1}_n).$$
The final step is due to the centering of columns of X. Thus, we have $\hat{\alpha} = n^{-1}\sum_{i=1}^n Y_i$.
(ii) For ridge regression, we are minimising $\|Y - \alpha\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_2^2$. Differentiate with respect to β to obtain
$$0 = \frac{\partial L}{\partial \beta}\bigg|_{\hat{\alpha},\hat{\beta}_\lambda^{\text{ridge}}} = -2X^\top(Y - \hat{\alpha}\mathbf{1}_n - X\hat{\beta}_\lambda^{\text{ridge}}) + 2\lambda\hat{\beta}_\lambda^{\text{ridge}}.$$
Using the fact that $\mathbf{1}_n^\top X = 0$ and $X^\top X = I_p$, we have $\hat{\beta}_\lambda^{\text{ridge}} = \frac{1}{1+\lambda}X^\top Y$.
$\hat{\beta}_\lambda^{\text{Lasso}}$ minimises
$$\|Y - \hat{\alpha}\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1 = \|Y - \hat{\alpha}\mathbf{1}\|_2^2 + \|\beta\|_2^2 - 2(X^\top Y)^\top\beta + \|\beta\|_1,$$

where in the final step we have used both $\mathbf{1}^\top X = 0$ and $X^\top X = I_p$. Observe that different coordinates in β are decoupled in the final expression above:

$$\|\beta\|_2^2 - 2(X^\top Y)^\top \beta + \|\beta\|_1 = \sum_{j=1}^{p} \beta_j^2 - 2(X^\top Y)_j \beta_j + \lambda|\beta_j|. \tag{1}$$

Hence the Lasso optimisation amounts to solving (1) for every j. When $\pm\beta_j \geq 0$, the minimum is at $(X^\top Y)_j \mp \lambda/2$, respectively, either by inspecting the graphs or finding the minimum of the quadratic functions. Hence, overall the Lasso estimator $\hat{\beta}_\lambda^{\text{Lasso}}$ has coordinates given by the so-called "soft-thresholding" operation

$$\hat{\beta}_j^{\text{Lasso}} = \text{sgn}(\hat{\beta}_j) \max\{|\hat{\beta}_j| - \lambda/2, 0\},$$

where $\hat{\beta}_j := (X^\top Y)_j$ is the j^{th} coordinate of the OLS or MLE estimator $\hat{\beta}$.

(iii) From the expressions above, the gradient descent algorithm proceeds as follows:

- initialise $\hat{\alpha}^{(0)}$ and $\hat{\beta}^{(0)}$; and,
- set

$$\hat{\alpha}^{(t)} = \hat{\alpha}^{(t-1)} + 2\gamma_t(\mathbf{1}_n^\top Y - n\hat{\alpha}^{(t-1)})$$

and

$$\hat{\beta}^{(t)} = \hat{\beta}^{(t-1)} + 2\gamma_t \left(\lambda\hat{\beta}^{(t-1)} - X^\top(Y - \hat{\alpha}^{(t-1)}\mathbf{1}_n - X\hat{\beta}^{(t-1)})\right)$$

for $t = 1, 2, \ldots$ until numerical convergence.

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.

Solution

(i) Let Y_i be the proportion of deaths in ith cage ($i = 1, \ldots, 20$) and $x_i^\top = (1, \mathbf{1}\{\text{new pesticide in cage } i\})$. The algebraic form is $n_i Y_i \overset{\text{ind}}{\sim} \text{Bin}(n_i, p_i)$, where $n_i = 30$ throughout and $\text{logit}\, p_i = x_i^\top \beta$ for some unknown $\beta \in \mathbb{R}^2$.

(ii) Zero-inflated binomial model: algebraically,

$$n_i Y_i \overset{\text{ind}}{\sim} \begin{cases} \delta_0 & \text{with probability } \pi_i, \\ \text{Bin}(n_i, p_i) & \text{with probability } 1 - \pi_i, \end{cases}$$

where $\text{logit}\, p_i = x_i^\top \beta$ and $\text{logit}\, \pi_i = x_i^\top \gamma$ for some unknown $\beta, \gamma \in \mathbb{R}^2$.

(iii) The likelihood of this new model is

$$L(\beta, \gamma; Y) = \prod_{i: Y_i = 0} \frac{e^{x_i^\top \gamma} + (1 + e^{x_i^\top \beta})^{-n_i}}{1 + e^{x_i^\top \gamma}} \prod_{i: Y_i > 0} \frac{1}{1 + e^{x_i^\top \gamma}} \binom{n_i}{n_i Y_i} \frac{e^{x_i^\top \beta n_i Y_i}}{(1 + e^{x_i^\top \beta})^{n_i}}.$$

For the augmented likelihood we introduce latent variables $Z_1, \ldots, Z_n$ such that

$$Z_i = \mathbf{1}\{Y_i \text{ is sampled from the zero population}\}.$$

Thus, $Z_i \overset{\text{iid}}{\sim} \text{Bin}(1, \pi_i), i = 1, \ldots, n$, and, writing $Z := (Z_1, \ldots, Z_n)^\top$,

$$\begin{aligned} L(\beta, \gamma; Y, Z) &= f(Y \mid Z, \beta, \gamma) f(Z \mid \gamma) \\ &= \prod_{i=1}^n \left(\binom{n_i}{n_i Y_i} \frac{\exp\left(n_i Y_i x_i^\top \beta\right)}{\left(1 + \exp(x_i^\top \beta)\right)^{n_i}} \right)^{1 - Z_i} \frac{\exp\left(Z_i x_i^\top \gamma\right)}{1 + \exp(x_i^\top \gamma)}. \end{aligned}$$

(iv) Let

$$\ell_0(\beta, \gamma; Y, Z) = \ell_1(\gamma; Y, Z) + \ell_2(\beta; Y, Z) + \sum_{i=1}^n (1 - Z_i) \log \binom{n_i}{n_i Y_i},$$

where $\ell_1(\beta; Y, Z) := \sum_{i=1}^n (1 - Z_i)\{n_i Y_i x_i^\top \beta - n_i \log(1 + e^{x_i^\top \beta})\}$ and $\ell_2(\gamma; Y, Z) := \sum_{i=1}^n \{Z_i x_i^\top \gamma - \log(1 + e^{x_i^\top \gamma})\}$. In the EM algorithm, we first initialise $\hat{\gamma}^{(0)}$ and $\hat{\beta}^{(0)}$. For $k \geq 0$, in the expectation step, we compute the expected value of ℓ_0 under $Z \mid Y, \hat{\gamma}^{(k)}, \hat{\beta}^{(k)}$, for which we need to compute Z_i given Y_i, assuming that $\hat{\gamma}^{(k)}$ and $\hat{\beta}^{(k)}$ are the true parameters, i.e.

$$Z_i^{(k)} = \begin{cases} 0 & \text{if } Y_i > 0 \\ \left\{1 + (1 + e^{x_i^\top \hat{\beta}^{(k)}})^{-n_i} e^{-x_i^\top \hat{\gamma}^{(k)}}\right\}^{-1} & \text{if } Y_i = 0. \end{cases}$$

Then, writing $Z^{(k)} := (Z_1^{(k)}, \ldots, Z_n^{(k)})^\top$, in the maximisation step we let $\hat{\beta}^{(k+1)} = \arg\max_\beta \ell_1(\beta; Y, Z^{(k)})$ and $\hat{\gamma}^{(k+1)} = \arg\max_\gamma \ell_2(\gamma; Y, Z^{(k)})$. We iterate between the expectation and maximisation steps until fulfilling our numerical convergence criterion.

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?

Solution

(i) Let $Y_i =$`ha2`$_i$, i.e. whether patient i had a second heart attack or not, $i = 1, \ldots, n$. Then, the algebraic form of the model is $Y_i \overset{\text{ind}}{\sim} \text{Bin}(1, p_i)$ with $\text{logit}\, p_i = \beta_0 + \beta_1 x_{i1} +$

$\beta_2 x_{i2}$ and $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$ unknown, where p_i is the probability that patient i has a second heart attack, x_{i1} is the anxiety level of patient i and x_{i2} is the indicator variable for whether anger management treatment is completed for patient i. The estimated coefficients are $\hat{\beta}_0 = -6.38$, $\hat{\beta}_1 = 0.140$, $\hat{\beta}_2 = -2.73$.

(ii) For any fixed treatment, every unit increase in anxiety level increases the odds of having a second heart attack by $e^{0.140} = 1.15$ folds. For any fixed anxiety level, completing the anger management treatment changes the odds of having a second heart attack by a factor of $e^{-2.73} = 0.0650$.

(iii) 39 and 37.

(iv) in the `glm` call, change the first argument to `ha2 ~ 1`. Difference between log-likelihoods of `model1` and the null model is $(95.45 - 69.75)/2 = 12.85$, so difference between AIC is $2(12.85-2) = 21.70$. Hence the AIC for null model is $135.75+21.70 = 157.45$.

(v) Note that the residual deviance is approximated by $37\hat{\phi}$.

Under SDA asymptotics, the latter is asymptotically distributed as a χ^2_{37} distribution. The answer to the first question can go in two ways: justify why SDA does not apply here (this is a binary binomial model with a continuous covariate, so there cannot be large counts) so we do not have enough evidence to conclude that there is overdispersion; or, explicitly acknowledge this limitation but apply SDA asymptotics nonetheless and note that 69.75 is much larger than the 37 degrees of freedom, so there seems to be overdispersion.

After fitting the quasibinomial model, the estimated dispersion parameter is not given in this question but, by the observation above, the best approximate we have is $69.753/37 = 1.89$. If we treat this as the actual estimated dispersion parameter from the Pearson's residuals, and scaling the standard errors by $\sqrt{1.89}$, the t-score for anxiety and treatment are 2.11 and 1.98 respectively. Then, only `anxiety` is significant at 5% level.

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.

Solution

(i) Define

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i\ell_i} \end{pmatrix}, \quad X_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{i\ell_i} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Then marginally we have $Y_i \overset{\text{ind}}{\sim} N_{\ell_i}(X_i\beta, \Sigma_i)$ for $i = 1, \ldots, n$, for $\Sigma_i = \Sigma_i(\sigma^2, \tau^2) = \sigma^2 I_{\ell_i} + \tau^2 \mathbf{1}_{\ell_i}\mathbf{1}_{\ell_i}^\top$ (the notation $\mathbf{1}_{\ell_i}$ denotes an all-one vector of length ℓ_i). Thus, the log-likelihood is

$$\ell(Y; \beta, \sigma^2, \tau^2) = -\frac{1}{2}\sum_{i=1}^{m}\Big\{\ell_i \log(2\pi) + \log\det\Sigma_i + (Y_i - X_i\beta)^\top \Sigma_i^{-1}(Y_i - X_i\beta)\Big\}.$$

(ii) Solving the maximum likelihood estimation problem, we have for $\hat{\Sigma}_i = \Sigma_i(\hat{\sigma}^2, \hat{\tau}^2)$ that

$$\begin{aligned} 0 = \frac{\partial \ell}{\partial \beta}\bigg|_{\hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2} &= \sum_{i=1}^{m} X_i^\top \hat{\Sigma}_i^{-1}(Y_i - X_i\hat{\beta}) \\ \implies \hat{\beta} &= \bigg(\sum_{i=1}^{m} X_i^\top \hat{\Sigma}_i^{-1} X_i\bigg)^{-1}\bigg(\sum_{i=1}^{m} X_i^\top \hat{\Sigma}_i^{-1} Y_i\bigg). \end{aligned}$$

Under a balanced design, $X_1 = \cdots = X_m$ and $\hat{\Sigma}_1 = \cdots = \hat{\Sigma}_m$ and the above expression simplifies to

$$\hat{\beta} = \frac{1}{m}\sum_{i=1}^{m}\big(X_1^\top \hat{\Sigma}_1^{-1} X_1\big)^{-1}\big(X_1^\top \hat{\Sigma}_1^{-1} Y_i\big).$$

We claim that

$$(X_1^\top \hat{\Sigma}_1^{-1} X_1)^{-1} X_1^\top \hat{\Sigma}_1^{-1} = (X_1^\top X_1)^{-1} X_1^\top. \tag{2}$$

To see this, note that both sides of (2) are matrices whose rows are in the row space of $X_1^\top$ (i.e. the column space of X_1). This is because $\hat{\Sigma}_1^{-1} = aI + b\mathbf{1}_\ell\mathbf{1}_\ell^\top$ for some $a, b \in \mathbb{R}$, so both $(1, \ldots, 1)^\top$ and $(x_{11} - \bar{x}_1, \ldots, x_{1\ell} - \bar{x}_1)^\top$ are eigenvectors of $\hat{\Sigma}_1^{-1}$ (where we use $\bar{x}_1$ to denote the average of $x_{11}, \ldots, x_{1\ell}$). Consequently, for any $v \in \mathbb{R}^\ell$ orthogonal to the column span of X_1, we have $(X_1^\top \hat{\Sigma}_1^{-1} X_1)^{-1} X_1^\top \hat{\Sigma}_1^{-1} v = 0 = (X_1^\top X_1)^{-1} X_1^\top v$. On the other hand, we have

$$(X_1^\top \hat{\Sigma}_1^{-1} X_1)^{-1} X_1^\top \hat{\Sigma}_1^{-1} X_1 = I_2 = (X_1^\top X_1)^{-1} X_1^\top X_1.$$

Thus the difference between LHS and RHS in (2) annihilates vectors both in the column span of X_1 and its orthogonal complement and must be zero. This verifies the claim in (2). Consequently,

$$\hat{\beta} = \frac{1}{m}\sum_{i=1}^{m}(X_1^\top X_1)^{-1}(X_1^\top Y_i) = \frac{1}{m}\sum_{i=1}^{m}\hat{\beta}^{(i)},$$

where $\hat{\beta}^{(i)}$ is the regression coefficients for regressing Y_i against X_i.

(iii) Yes. By essentially the same argument (noting that the vector $(z_{11}-\bar{z}_1, \ldots, z_{1\ell}-\bar{z}_1)^\top$ is also an eigenvector of $\hat{\Sigma}_1$).

(iv) Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}.$$

Then $Y \sim N_n(X\beta, \Sigma)$, where $\Sigma = \Sigma(\sigma^2, \tau^2) = \mathrm{diag}(\Sigma_1, \ldots, \Sigma_m)$ is a block diagonal matrix. Then $U := A^\top Y \sim N_n(0, A^\top \Sigma A)$. The restricted log-likelihood for σ^2, τ^2 is

$$\begin{aligned}\ell^{\mathrm{Res}}(\sigma^2, \tau^2; U) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}c\det(A^\top \Sigma A) - \frac{1}{2}U^\top (A^\top \Sigma A)^{-1} U \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(A^\top \Sigma A) - \frac{1}{2}Y^\top A(A^\top \Sigma A)^{-1}A^\top Y.\end{aligned}$$

Note that for any other orthonormal $\tilde{A}$ whose column span the same space as A, we can write $\tilde{A} = AR$ for $R \in \mathcal{O}(n-p)$. Then $\det(A^\top \Sigma A) = \det(R^\top A^\top \Sigma A R)$. Also, $\Sigma^{1/2}A(A^\top \Sigma A)A^\top \Sigma^{1/2}$ is the projection onto column space of $\Sigma^{1/2}A$, which is the same as the column space of A (note $\Sigma^{1/2}$ is invertible) and does not depend on particular choices of the basis used in constructing A. Hence the MLEs for σ^2 and τ^2 do not depend on the particular choices of columns 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.

Solution

(a) `bw.lm1` is a normal linear model so it assumes independence between observations, which is clearly violated by observations of the same rat for different days.

`bw.lm2` is non-identifiable because the design matrix is not of full rank: the column corresponding to diet r (2 or 3) is the sum of the columns of those rats having such diet.

(b) Let Y_{ij} be the weight of the ith rat at the jth measurement. Then, the algebraic form of `bw.lme1` is

$$Y_{ij} = \beta_0 + b_i + \beta_1 x_j + \alpha_2 \mathbb{1}_{\{j\in D_2\}} + \alpha_3 \mathbb{1}_{\{j\in D_3\}} + \epsilon_{ij}, \qquad i = 1, \ldots, 16,\ j = 1, \ldots, 11,$$

where $x_j = 7j - 6$ is the time of the jth measurement (in days), D_2 and D_3 are set of rat indices on diet 2 and 3 respectively, $b_i \overset{\text{iid}}{\sim} N(0, \tau^2), \tau^2 > 0$, $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2), \sigma^2 > 0$ and the b_i's are independent from the ϵ_{ij}'s. The estimated values are $\hat{\beta}_0 = 244$, $\hat{\beta}_1 = 0.586$, $\hat{\alpha}_2 = 221$, $\hat{\alpha}_3 = 262$, $\hat{\tau}^2 = 1340$ and $\hat{\sigma}^2 = 66.9$.

The fixed effect coefficient for `Time` is interpreted as meaning that the increase in the mean weight of a given rat per day is 0.586 grams.

(c) See question 4 part (iv) (should define all quantities involved correctly, including the design matrix X, the vector of responses Y, the vector of all fixed effect parameters θ and the covariance matrix Σ).

(d) The test is not valid. There are two reasons: `bw.lme1` uses restricted maximum likelihood, so the log-likelihoods of the fitted models `bw.lme1` and `bw.lm1` are not comparable; and, even if the maximum likelihood had been used, the test statistic would not follow a χ^2_1 distribution as $\tau^2 = 0$, which formally gives rise to the null model `bw.lm1`, is on the boundary of the parameter space of `bw.lme1`.

To construct a valid test we first fit a linear mixed effect model using maximum likelihood: set `REML = FALSE` in the last argument of the `lmer` function used for `bw.lme1`, and call this new model `bw.lme2`. Then, we construct the likelihood ratio test statistic

$$T = 2\big\{\ell_{\texttt{lme2}}(\hat{\theta}_{\texttt{lme2}}, \hat{\tau}^2_{\texttt{lme2}}, \hat{\sigma}^2_{\texttt{lme2}}; Y) - \ell_{\texttt{lm1}}(\hat{\theta}_{\texttt{lm1}}, \hat{\sigma}^2_{\texttt{lm1}}; Y)\big\},$$

where $\ell_{\texttt{lme2}}$ and $\ell_{\texttt{lm1}}$ are log-likelihood functions for models `bw.lme2` and `bw.lm1`, respectively, $\hat{\theta}_{\texttt{lme2}}, \hat{\tau}^2_{\texttt{lme2}}, \hat{\sigma}^2_{\texttt{lme2}}$ are the MLEs for `bw.lme2`, and $\hat{\theta}_{\texttt{lm1}}, \hat{\sigma}^2_{\texttt{lm1}}$ are the MLEs for `bw.lm1`. We compute the p-value of this test statistic via parametric bootstrap:

- For $b = 1, \ldots, B$, draw $Y^{(b)} \overset{\text{iid}}{\sim} N(X\hat{\theta}_{\texttt{lm1}}, \hat{\sigma}^2_{\texttt{lm1}} I_n)$;
- Compute the likelihood ratio test statistic using $Y^{(b)}$ by

$$T^{(b)} = 2\left\{\sup_{\theta \in \mathbb{R}^4, \tau^2 > 0, \sigma^2 > 0} \ell_{\texttt{lme2}}(\theta, \tau^2, \sigma^2; Y^{(b)}) - \sup_{\theta \in \mathbb{R}^4, \sigma^2 > 0} \ell_{\texttt{lm1}}(\theta, \sigma^2; Y^{(b)})\right\}.$$

- compute the p-value as $B^{-1} \sum_{b=1}^{B} \mathbb{1}_{T^{(b)} \geq T}$.

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.

Solution

(i) Let Y_{ij} be the fraction of herd i affected in period j and n_{ij} the size of herd i in period j. Then the algebraic form is $n_{ij}Y_{ij} \sim \text{Bin}(n_{ij}, p_{ij})$ with $\log \frac{p_{ij}}{1-p_{ij}} = \mu + b_i + \beta_j$ for $i = 1, \ldots, 15$ and $j = 1, \ldots, 4$. Here $b_i \overset{\text{ind}}{\sim} N(0, \tau^2)$, $\mu, \tau^2\beta_j$ are unknown with the exception of the corner constraint $\beta_1 = 0$.

(ii) The `intercept` estimate of -1.40 means that the *mean* log-odds of infection in *a given herd* in period 1 is -1.40. The `period2` estimate of -0.992 means that for a given herd, the change in log-odds of infection from period 1 to period 2 is -0.992. Similarly, the change in log-odds of infection in a given herd from period 1 to period 3 is -1.13 and the change in log-odds of infection in a given herd from period 1 to period 4 is -1.58.

(iii) The parametric bootstrap procedure for estimating $\text{se}(\hat{\tau})$ (or $\text{se}(\hat{\tau}^2)$ fine too) is as follows:

- For each $r = 1, \ldots, B$,
  - We generate a bootstrap sample $(Y_{ij}^{(r)} : i = 1, \ldots, 15, j = 1, \ldots, 4)$ as follows: first simulate $b_1^{(r)}, \ldots, b_{15}^{(r)} \overset{\text{iid}}{\sim} N(0, \hat{\tau}^2)$ for $\hat{\tau}^2 = 0.412$, then simulate, independently from the $b_i^{(r)}$s,

$$n_{ij} Y_{ij}^{(r)} \mid b_i^{(r)} \overset{\text{ind}}{\sim} \text{Bin}(n_{ij}, \hat{p}_{ij}^{(r)}), \quad \hat{p}_{ij}^{(r)} = \text{expit}(\hat{\mu} + b_i^{(r)} + \hat{\beta}_j),$$

    where $\hat{\mu} = -1.40$, $\hat{\beta}_1 = 0$, $\hat{\beta}_2 = -0.992$, $\hat{\beta}_3 = -1.13$ and $\hat{\beta}_4 = -1.58$.
  - Fit a random intercept logistic regression model of responses $Y_{ij}^{(r)}$ against the covariate `period`, grouped by `herd`. (You can describe the model using an algebraic form as well) to obtain random the effect estimate $\hat{\tau}^{(r)}$.
  - Aggregate $\{\hat{\tau}^{(r)} : r = 1, \ldots, B\}$ to compute the standard error of the random effect coefficient estimator $\hat{\tau}^2$ (recall that the standard error of an estimator is the estimated standard deviation of the estimator)

$$\text{se}(\hat{\tau}) = \sqrt{\frac{1}{B} \sum_{r=1}^{B} \left(\hat{\tau}^{(r)} - \frac{1}{B} \sum_{s=1}^{B} \hat{\tau}^{(s)} \right)^2}.$$

7. (*Exercise with* `R`) Download the `leukaemia` dataset (.txt file) 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.