STATISTICAL LEARNING IN PRACTICE Part III / Lent 2020
Example Sheet 1 (of 4) By courtesy of Dr T. Wang

You have the option to submit your answers to questions 2 and 3 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 preferably by 2pm on 11^{th} February, and no later than 11am on 12^{th} February.

1. Suppose $Y = X\beta + \epsilon$, where $X \in \mathbb{R}^{n\times p}$ has full rank $p \leq n$ and $\epsilon \sim N_n(0, \sigma^2 I_n)$.

   (i) Show that the maximum likelihood estimators for β and σ^2 are $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ and $\hat{\sigma}^2 = n^{-1}\|Y - X\hat{\beta}\|_2^2$ respectively.

   (ii) Show that the vector of residuals satisfies $R = (I - P)Y$, where $P = X(X^\top X)^{-1}X^\top$.

   Suppose from now on that $p \leq n - 1$.

   (iii) Let $x_i^\top$ be the ith row of X and $\hat{\beta}^{(i)}$ the maximum likelihood estimator of β using all observations but the ith one. Show that

   $$\hat{\beta} - \hat{\beta}^{(i)} = \frac{(X^\top X)^{-1}x_i R_i}{1 - P_{ii}} \quad \text{and} \quad (\hat{\beta} - \hat{\beta}^{(i)})^\top (X^\top X)(\hat{\beta} - \hat{\beta}^{(i)}) = \frac{R_i^2 P_{ii}}{(1 - P_{ii})^2}.$$

   [*Hint: use* $(A + uv^\top)^{-1} = A^{-1} - (1 + v^\top A^{-1}u)^{-1}A^{-1}uv^\top A^{-1}$.]

   (iv) Explain how the leave-one-out cross-validation error $\mathrm{err}_{\mathrm{CV}}$ can be computed for this linear model. Show that the leave-one-out cross-validation error can be represented as a weighted mean squared residuals in this case:

   $$\mathrm{err}_{\mathrm{CV}} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i^2}{(1 - P_{ii})^2}.$$

2. (i) Show that $\{\mathrm{Poi}(\lambda) : \lambda \in (0, \infty)\}$ is an exponential dispersion family.

   Suppose we observe data $Y_i \sim^{ind.} \mathrm{Poi}(e^{x_i^\top \beta})$ for $i = 1, \ldots, n$, with $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n\times p}$ of full rank $p \leq n$.

   (ii) Show that the observed Fisher information matrix (i.e. negative Hessian of the log-likelihood) is positive definite at every β. Write down explicit expressions for one Newton–Raphson and one Fisher scoring iteration starting from an initial point β. Would they have been equal if the link were not canonical? Justify your answer.

   (iii) Suppose a new observation is made with covariates $x_* \in \mathbb{R}^p$. Derive a (small-dispersion) asymptotic $1 - \alpha$ confidence interval for the associated mean response under the Poisson model.

   (iv) What are the residual deviance and deviance residuals for the maximum likelihood fitted model? Show that when the fitted values are close to the observed values, the deviance residuals can be approximated by the Pearson's residuals $r_i = \hat{\lambda}^{-1/2}(Y_i - \hat{\lambda}_i)$, where $\hat{\lambda}_i = e^{x_i^\top \hat{\beta}}$.

3. Question 6 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

4. Consider the following linear regression problem without an intercept term

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \quad \text{for } i = 1, \ldots, n.$$

   Suppose estimates of the regression parameters (β_1, β_2) of this model are obtained through the minimization of the sum-of-squares augmented with a ridge-type regularisation

$$\sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2 + \lambda(\beta_1^2 + \beta_2^2 + 2\nu\beta_1\beta_2),$$

   with tuning parameters $\lambda \in [0, \infty)$ and $\nu \in [-1, 1]$.

   (i) Write down the above optimisation problem in an equivalent constrained form. Sketch for both $\nu = 0$ and $\nu = 0.9$ the shape of the parameter constraint induced by the penalty above and describe in words the qualitative difference between both shapes.

   (ii) When $\nu = -1$ and $\lambda \to \infty$, the estimates of β_1 and β_2 (resulting from minimisation of the penalized loss function above) converge towards each other: $\lim_{\lambda\to\infty} \hat{\beta}_1(\lambda, -1) = \lim_{\lambda\to\infty} \hat{\beta}_2(\lambda, -1)$. Motivated by this observation a data scientists incorporates the equality constraint $\beta_1 = \beta = \beta_2$ explicitly into the model, and she estimates the 'joint regression parameter' β through the minimisation (with respect to β) of:

$$\sum_{i=1}^{n} (Y_i - \beta X_{i,1} - \beta X_{i,2})^2 + \rho\beta^2,$$

   with a tuning parameter $\rho \in [0, \infty)$. The data scientist is surprised to find that resulting estimate $\hat{\beta}(\rho)$ does not have the same limiting (in the penalty parameter) behaviour as the $\hat{\beta}_1(\lambda, -1)$, i.e. $\lim_{\rho\to\infty} \hat{\beta}(\rho) \neq \lim_{\lambda\to\infty} \hat{\beta}_1(\lambda, -1)$. Explain the misconception of the data scientist.

   (iii) Assume that (a) $n \gg 2$, (b) the unpenalised least squares estimates $(\hat{\beta}_1(0, 0), \hat{\beta}_2(0, 0))$ are equal to $(-2, 2)$, and (c) that the two covariates $X_1 = (X_{11}, \ldots, X_{n1})^\top$ and $X_2 = (X_{12}, \ldots, X_{n2})^\top$ are zero-centred, have unit variance, and are strongly negatively correlated. Consider $(\hat{\beta}_1(\lambda, \nu), \hat{\beta}_2(\lambda, \nu))$ for both $\nu = -0.9$ and $\nu = 0.9$. For which value of ν do you expect the sum of the absolute value of the estimates to be larger?

5. Recall that the negative binomial distribution $\text{NB}(r, p)$ models the total number of successes until $r \in \mathbb{N}$ failures have occurred in a sequence of i.i.d. Bernoulli trials each with a success probability $p \in [0, 1]$.

   (i) Write down the probability mass function of $Y \sim \text{NB}(r, p)$ and show that $\{\text{NB}(r, p) : r \in \mathbb{N}, p \in (0, 1)\}$ is an exponential dispersion family if and only if r is known.

(ii) Suppose $\nu \sim \text{Gamma}(\theta, \theta)$ for some $\theta \in \mathbb{N}$ and $Y \mid \nu \sim \text{Poi}(\lambda\nu)$ for some $\lambda > 0$. Show that $Y \sim \text{NB}(\theta, \frac{\lambda}{\lambda+\theta})$. What are the resulting parameters in the alternative parametrisation introduced in class? Make sure you are convinced about the validity of the generalisation of this reparametrisation to $\theta > 0$.

6. To understand the relationship between the number of typographical errors (dependent variables Y_i, $i = 1, \ldots, n$) and lengths of manuscripts in words (independent variables x_i, $i = 1, \ldots, n$), a researcher fitted both a Poisson model ω_1 and a negative binomial model ω_2 to the data.

   (i) Write down the two models algebraically.

   (ii) If the log-likelihoods for ω_1 and ω_2 are -193.4 and -192.2 respectively. What is the result of the likelihood ratio test between the null hypothesis that ω_1 is correct and the alternative hypothesis that model ω_2 is correct (at 5% level)? Which model will be selected based on AIC?

7. (*Exercise with* `R`) The dataset `ships` in the library `MASS` contains the number of incidents for a set of ships. Investigate the relationship between the number of incidents (response variable `incidents`) and the months of service (variable: `service`) and the type of the ships (variable: `type`). Ignore the other columns of the dataset. Explore the data graphically (it may be necessary to apply transformations to the data) and fit the appropriate model to predict the number of incidents.

   Remark: this type of question is <u>not</u> something that can happen in the exam, which is pen and paper only. However, it is a good exercise to become more familiar with `R` and to tackle a small data analysis example.