Comprehensive/Qualifying Exam: Applied Statistics

Boston University, 2022

Instructions: This is a closed book exam. You are not allowed a crib sheet or a calculator. Please answer problems 1–2 and 3–4 in separate blue books. ALL answers need to include an explanation, even if this is not explicitly asked in the question.

To pass the exam at the **Master's level**, you need answer 3 questions. If you answer more than three problems, the lowest three scores will be used to compute your total. If you do not want us to grade any part of your answer, please cross it out completely.

To pass the exam the **Ph.D. level**, you need to answer all four questions.

1. Suppose your data follows the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \ , \tag{1}$$

where $E[\mathbf{e}|\mathbf{X}] = 0$, $\mathrm{Var}(\mathbf{e}|\mathbf{X}) = \mathbf{I}$ and $\mathbf{e} \sim \mathcal{N}(0, \mathbf{I})$. Assume that the data matrix $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is full column rank with $n > p+1$ and $\lambda_{max}(\mathbf{X}^T\mathbf{X})$ is uniformly bounded. You decide to apply the ridge regression optimization

$$\hat{\boldsymbol{\beta}}_{ridge} := \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{p+1}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_2^2,$$

where $\lambda \geq 0$.

a) Show that the ridge regression optimization is

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}.$$

What do you conclude about the estimate $\hat{\boldsymbol{\beta}}_{ridge}$ as λ becomes large ?

b) Suppose that $\lambda > 0$. Is $\hat{\boldsymbol{\beta}}_{ridge}$ a biased estimator ? If not, why ?

c) Suppose your data follows model (1) but instead $\mathrm{Var}(\mathbf{e}|\mathbf{X}) = \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix. You decide to apply ridge regression to this data. Show that

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ridge}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{C}\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

2. We consider the multiple linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and where the regression errors satisfy $\mathbf{e} \sim \mathbf{N}(0, \sigma^2 I_n)$. The parameters of the model are $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$, with true values $\boldsymbol{\beta}_\star$, $\sigma^2_\star$. We assume that the first column of X represents the intercept, and X is of rank p.

(a) Give the expressions of the least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_\star$ and the unbiased estimator $\hat{\sigma}^2$ of $\sigma^2_\star$.

(b) Under the assumptions of the model what is the distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$? Using these distributions, explain how to construct a 95% confidence interval for $\boldsymbol{\beta}_{\star,1}$, the first component of $\boldsymbol{\beta}_\star$.

(c) For some $1 \leq p_1 < p$, suppose that we partition X as $X = (\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$, and $\mathbf{X}_2 \in \mathbb{R}^{n \times p_2}$, where $p_1 + p_2 = p$. We assume that X_1 is of rank p_1. Suppose now that we remove all the predictors in X_2 and fit the model (under the same error distribution assumption)

$$\mathbf{Y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{e}. \tag{3}$$

We partition the true regression coefficient as $\boldsymbol{\beta}_\star = \begin{pmatrix} \boldsymbol{\beta}_{1\star} \\ \boldsymbol{\beta}_{2\star} \end{pmatrix}$.

i. Show that regardless of $\boldsymbol{\beta}_{\star 2}$, the R^2 of model (2) is typically larger than the R^2 of model (3). We recall that the R^2 of a linear regression model (with intercept) with fitted values $\hat{\mathbf{Y}}$ is given by

$$R^2 = 1 - \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

ii. Show that if $\boldsymbol{\beta}_{\star 2} = 0$, then Mallow's C_p will prefer model (3) to the full model (2). We recall that Mallow's C_p for a sub-model of model (2) with fitted values $\hat{\mathbf{Y}}$, and p_1 columns is

$$C_p = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|_2^2}{\hat{\sigma}^2} - (n - 2p_1),$$

where $\hat{\sigma}^2$ is computed from the full model.

3. Suppose you want to regress n observations $Y_i \overset{\text{ind}}{\sim} \mathsf{G}(\mu_i, \nu_i)$, $i = 1, \ldots, n$, where the means $\mu_i = \mathbb{E}[Y_i]$ are related to a set of regressors X by $g(\mu_i) = \mathbf{x}_i^\top \beta$, and the precisions ν_i are related to another set of regressors Z (more on that below). The gamma log-likelihood is given by (up to a constant that does not depend on μ or ν):

$$\ell(\mu, \nu; \mathbf{y}) = \sum_{i=1}^n \nu_i \left(\log \frac{y_i}{\mu_i} - \frac{y_i}{\mu_i} \right) + \nu_i \log \nu_i - \log \Gamma(\nu_i),$$

where $\Gamma(\cdot)$ is the gamma function.

(a) Defining the i-th deviance component as

$$d_i = -2 \left(\log \frac{y_i}{\mu_i} - \frac{y_i - \mu_i}{\mu_i} \right),$$

show that, given d_i, the log-likelihood above can be written as a function of d_i and depends only on ν_i.

(b) Now, show that the distribution of d_i belongs to the exponential family and so we can regress d_i on Z with $h(\mathbb{E}[d_i]) = \mathbf{z}_i^\top \gamma$. Identify the canonical parameter θ and the cumulant function $b(\theta)$ of this distribution.

(c) Using the cumulant function, find the mean $\tau_i = \mathbb{E}[d_i]$ as a function of the precision ν_i. Using the fact that $\log \nu_i > \psi(\nu_i)$ since $\nu_i > 0$, show that $\tau_i > 0$, as expected. Here $\psi(x) = d(\log \Gamma(x))/dx$ is the digamma function.

(d) Next, obtain the variance of d_i as a function of the precision ν_i. Now, using the fact that $\psi_1(\nu_i) > \nu_i^{-1}$ since $\nu_i > 0$, deduce that the variance of d_i is positive, as expected. The trigamma function ψ_1 is $\psi_1(x) = d\psi(x)/dx$.

(e) Describe a numerical procedure to obtain maximum likelihood estimates of β and γ. In particular, how are you exploiting the fact that computing d_i requires only y_i and μ_i but not ν_i?

4. In a study of game system preferences, a number of users were asked to rank three choices: PC (1), PlayStation (2), and Xbox (3). Each user was also asked to report their `age`, number of `hours` they play per week, and which systems they `own`. Suppose now that the probability of a user selecting the j-th system is π_j, $j = 1, \ldots, K = 3$. The probability of the ranking $\mathbf{r} = (r_1, \ldots, r_K)$ is then modeled as a sequence of successive choices,

$$\mathbb{P}(\mathbf{r}) = \prod_{j=1}^{K} \frac{\pi_{r_j}}{\sum_{l=j}^{K} \pi_{r_l}}.$$

For example, the ranking $3 > 1 > 2$, that is, $r_1 = 3$, $r_2 = 1$, and $r_3 = 2$, is

$$\mathbb{P}(3 > 1 > 2) = \frac{\pi_3}{\pi_3 + \pi_1 + \pi_2} \times \frac{\pi_1}{\pi_1 + \pi_2} \times \frac{\pi_2}{\pi_2} = \pi_3 \frac{\pi_1}{\pi_1 + \pi_2}.$$

The first-choice probabilities π are in turn represented as in a multinomial model with the predictors described above, with an intercept for each choice (having PC, the first level, as reference) and interactions with `hours` and `age`, but not for `own`.

(a) Show that with the canonical link for the multinomial distribution the probability of a ranking can be expressed as

$$\mathbb{P}(\mathbf{r}) = \prod_{j=1}^{K} \frac{\exp\{\mathbf{x}_{r_j}^\top \beta\}}{\sum_{l=j}^{K} \exp\{\mathbf{x}_{r_l}^\top \beta\}}.$$

where $\mathbf{x}_j$ has the regressors for choice j.

(b) Here is the summary from fitting the model described above:

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept):PlayStation	1.962800	1.765005	1.1121	0.266110
(Intercept):Xbox	2.547272	1.671729	1.5237	0.127575
own	0.929688	0.292422	3.1793	0.001476

```
hours:PlayStation       -0.094577   0.046603 -2.0294 0.042416
hours:Xbox              -0.142441   0.048780 -2.9201 0.003500
age:PlayStation         -0.059558   0.086579 -0.6879 0.491510
age:Xbox                -0.064226   0.080645 -0.7964 0.425796

Deviance: 304.57
```

Interpret the coefficient for `own`. In particular, explain why the probability of selecting any choice as top rank is the same if the user owns all systems or none of them.

(c) Here is a simpler model fit after removing `age` as a covariate.

```
                         Estimate Std. Error z-value  Pr(>|z|)
(Intercept):PlayStation  0.771083   0.324045  2.3796 0.0173337
(Intercept):Xbox         1.250263   0.356881  3.5033 0.0004595
own                      0.928585   0.291919  3.1810 0.0014678
hours:PlayStation       -0.096381   0.045745 -2.1069 0.0351234
hours:Xbox              -0.143285   0.048152 -2.9757 0.0029236

Deviance: 305.30
```

Conduct a test to check if this simpler model is adequate. State the test statistic and its distribution under the null.

(d) A new user is considering buying their first gaming system. What is their most likely ranking of the systems? What is the estimated probability of this ranking?

(e) Fitting these rank-ordered models can be done as in a regular multinomial regression, but requires expanding the data to a new format. How would you design a specialized Newton method (or iteratively reweighted least squares) to avoid this extra amount of work and memory? In particular, how would the score calculation change?