Comprehensive/Qualifying Exam: Applied Statistics

Boston University, 2023

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 at the **Ph.D. level**, you need to answer all four questions.

1. Consider a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, for $i = 1, \ldots, n$, where the error terms ϵ_i's are assumed independent and identically distributed (i.i.d.) $\mathbf{N}(0, \sigma^2)$. We assume that $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$. Let $\hat{\beta}_0, \hat{\beta}_1$ denote the least squares estimators of β_0 and β_1 respectively. We know that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

We can also rewrite the model in matrix form. If we set

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \text{and} \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

then, the simple linear regression model above is equivalent to

$$y = X\beta + \epsilon.$$

And the least squares estimate of (β_0, β_1) is then

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1} X' y.$$

Show that the two expressions match.

2. We consider the multiple linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\beta} \in \mathbb{R}^p$. We assume that $\boldsymbol{\epsilon} \sim \mathbf{N}(0, \sigma^2 I_n)$ and that X has full column rank. We assume the model is correctly specified with true parameter values $\boldsymbol{\beta}_\star$, $\sigma^2_\star$. Here the intercept will not play any special role. Let $\boldsymbol{y}_{(i)} \in \mathbb{R}^{n-1}$ be the vector of responses obtained after removing the i-th response. Let $X_{(i)} \in \mathbb{R}^{(n-1) \times p}$ be the explanatory matrix obtained after removing the i-th row of X, that we denote $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\beta}}_{(i)}$ be the least squares estimate of the model $\boldsymbol{y}_{(i)} = X_{(i)}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{(i)}$. The i-th deleted residual is defined as

$$\hat{\epsilon}_{(i)} = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{(i)}.$$

(a) Show that $\mathsf{Var}(\hat{\epsilon}_{(i)}) = \sigma^2_\star \left(1 + \boldsymbol{x}_i (X'_{(i)} X_{(i)})^{-1} \boldsymbol{x}'_i\right)$.

(b) We know that $\hat{\epsilon}_{(i)} / \sqrt{\mathsf{Var}(\hat{\epsilon}_{(i)})}$ can be approximated by the i-th Studentized residual of the model defined as

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_i}}, \quad \text{where} \quad \hat{\sigma}^2_{(i)} = \frac{\|\boldsymbol{y}_{(i)} - X_{(i)} \hat{\boldsymbol{\beta}}_{(i)}\|^2}{n - 1 - p},$$

where h_i is the i-th leverage, and $\hat{\epsilon}_i$ is the i-th residual. Show that $t_i \sim T_{n-1-p}$.

3. Let. ϕ be a random variable taking values in $[-\pi, \pi)$ with pdf $f(\phi) = \frac{exp(\kappa cos(\phi_i - \mu_i))}{2\pi I_0(\kappa)}$, where $I_0(\kappa) = \frac{1}{2\pi}\int_{-\pi}^{\pi} exp(\kappa cos(x))dx$.

   (a) Let $y_i = \begin{pmatrix} cos(\phi) \\ sin(\phi) \end{pmatrix}$ be a unit vector and $\theta = \begin{pmatrix} \kappa cos(\mu) \\ \kappa sin(\mu) \end{pmatrix}$. Show that this distribution can be written in the form of the natural exponential family. Hint: recall that $cos(a-b) = cos(a)cos(b) + sin(a)sin(b)$.

   (b) What are the natural parameter, dispersion parameter, and cumulant function for this distribution? Make sure the cumulant function is expressed as a function of the natural parameter.

   (c) Express μ and κ as functions of the natural parameter.

   (d) Write down an expression for the mean of y as a function of the natural parameter and as a function of μ_i and κ.

4. An analysis was performed to determine how the number of traffic accidents occurring at various intersections through town are related to the average traffic volume at those intersections. The number of traffic accidents occurring over the period of 1 year at each of 100 intersections, `Accidents`, were recorded along with a standardized measure of traffic volume for each intersection, `Vol`.

   A Poisson GLM with log link was fit to `Accidents` as a function of `Vol` and `Vol^2`. Here is the summary of the fitted model:

```
                Estimate Std. Error z-value Pr(>|z|)
(Intercept)      0.04699    0.13011   0.361  0.71799
Vol              0.61983    0.17143   3.616  0.00030
I(Vol^2)        -0.37810    0.14170  -2.668  0.00762

Residual deviance: 186.78  on 97  degrees of freedom
```

   (a) Interpret the coefficients for `Vol` and `Vol^2`. If this fitted model is correct, how many accidents per year would we expect in an intersection with no traffic volume? At what value of `Vol` is the expected number of accidents maximized? What is the expected number of accidents at this maximizing value?

   (b) The researchers are worried that there is overdispersion in this model. The Pearson statistic for this fit has a value of `218.25`. Use this to estimate a dispersion parameter for this model. Is there significant overdispersion in this model? Compute 95% confidence intervals for the parameters associated with `Vol` and `Vol^2`, using your estimated dispersion.

   The researchers added a new categorical predictor describing each intersection in one of three safety categories with `C1` being the safest, `C2` being in-between, and `C3` being the least safe. A new Poisson GLM was fit to this model, yielding the following summary:

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -17.1760  1262.5576  -0.014  0.98915
Vol            0.6138     0.1671   3.673  0.00024
I(Vol^2)      -0.3116     0.1170  -2.664  0.00773
C2            16.0244  1262.5576   0.013  0.98987
C3            18.1158  1262.5576   0.014  0.98855

Residual deviance:  91.80  on 95  degrees of freedom
```

(c) Conduct a hypothesis test to determine whether this categorical predictor has a statistically significant influence on `Accidents`. Explain why the estimated `Std. Error` values in the summary table are so large.

(d) Assuming this new fitted model is correct, compute the expected number of accidents in an intersection of type `C1` with no traffic volume. What is the maximum expected number of accidents in an intersection of type `C3`?

(e) The Pearson statistic for this fit has a value of `95`. Is there evidence of overdispersion in this model? Based on the fits from both of these models, draw a conclusion about whether traffic volume has a significant influence on the number of accidents occurring at each intersection.