QE ID (given by Angie)

August 2019 Qualifying Exam in Linear Models

August 5, 2019

Instruction:

- This is a closed-book test.
- There are three questions; each has multiple parts.
- Answer each question as fully as possible.
- Show and justify all steps of your solutions.
- All answers should be simplified and should be free of matrix and vector expressions.
- Refer clearly to any known results that you are using, stating such results precisely.
- Show how the assumptions of a result you are using are satisfied in your application of the result.
- Indicate how the assumptions given in the question are used in the solution.
- Write your solutions on the blank sheets of paper that are provided.
- Write your QE ID number (given to you by Angie) on all answer sheets. **DO NOT** put your name, UTD ID, or any other identifying information on any of your answer sheets.
- On each sheet, identify which question and part is being answered.
- Begin each question on a new sheet.
- When finished, arrange your sheets in order, number each sheet, and be sure that your QE ID number (given by Angie) is on each sheet.
- Although the notations used in Q1, Q2, Q3, and Q4 are similar, they are independent, standalone problems.
- The total possible points is 100.

Q1 Consider a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i, i = 1, \ldots, n,$$

where the ϵ_i's are independent normal random variables with mean 0 and variance σ^2. Suppose we observed $n = 10, \sum_{i=1}^n X_i = \sum_{i=1}^n X_i^3 = 0$.

(a) (10 pts) Derive the maximum likelihood estimator for β_1.

(b) (10 pts) Derive a $\alpha-$level t test for $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$.

(c) (10 pts) Suppose $\sigma = 1$, construct the 95% confidence interval for β_1.

Q2 Consider a weighted straight line model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

where ϵ_i's are independent normal random variables with mean 0 and variance $i\sigma^2$.

(a) (10 pts) Derive the weighted least squares estimators for β_0 and β_1.

(b) (10 pts) Derive a α-level t test for $H_0 : \beta_0 = 0$ against $H_a : \beta_0 \neq 0$.

(c) (10 pts) The model (1) reduces to

$$Y_i = \beta_1 X_i + \epsilon_i, i = 1, \ldots, n,$$

when $\beta_0 = 0$. Derive the weighted least squares estimators for β_1 in the reduced model.

Q3 An important property of a generalized linear model is that the score function depends only on the mean and variance of the response variable. More specifically, let

$$E(Y_i) = g^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta}), i = 1, \ldots, n, \tag{2}$$

be the mean function denoted by $\boldsymbol{\mu}_i(\boldsymbol{\beta})$, where Y_i is the response variable, $\boldsymbol{X}_i$ is the covariate vector, $\boldsymbol{\beta}$ is the vector of regression coefficient, and $g(\cdot)$ is the link function. Then the score function for (2) has the form:

$$S(\beta) = \sum_{i=1}^n \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} V_i^{-1} \left[Y_i - \mu_i(\boldsymbol{\beta}) \right], \tag{3}$$

where V_i is the variance structure. Setting $S(\beta) = 0$ is the generalized estimating equation.

(20 pts) Show that the score function under the identity link, e.g., $g(t) = t$, is in the form of (3). To receive full credit, please identify both the mean function and the variance structure.

Q4 (20 pts) Define a linear model

$$Y_{ij} = \mu_i + \epsilon_{ij}, \ i = 1, 2, 3, \ j = 1, \ldots, n,$$

where ϵ_{ij}'s are independent normal random variables with mean 0 and variance σ^2. Derive an F-test for testing $H_0 : \mu_1 = 3\mu_2$ against $H_a :$ not H_0.

Qualifying Exam August 2019 — Statistical Methods

Instructions

- NOTE: You are not allowed to use internet or email except for downloading the data and emailing the final report (optional). To use these for any other purpose, ask the proctor.
- Go to http://www.utdallas.edu/~mchen/QE and download two datasets: (1) graduateEarnings.txt and (2) bullet.txt. Let the proctor know if you have any problems with this step.
- You can use any software of your choice. You can use the lab machines or your own laptop.
- Your report should clearly explain the steps, results, conclusions, and justification for the conclusions. Also, include your codes (with brief comments explaining each step) and outputs (ONLY relevant parts; highlighted wherever possible). Do not attach the parts of the output that were not used in answering questions.
- Submit a report (written or typed), hard copy or by **email to Angie.Bustamante@utdallas.edu**. If you choose to email, then attach only one single PDF file with the whole report. DO NOT email separate files for codes or outputs.
- **Write your QE ID number (given to you by Angie) on all answer sheets. DO NOT put your name, UTD ID, or any other identifying information on any of your answer sheets. DO NOT email your exam to any one other than Angie.**

Problems

1. Let $X_1, \ldots, X_n$ be a random sample of size n from a Poisson (λ) distribution, where $\lambda > 0$ is an unknown parameter. As an approximate large-sample $100(1-\alpha)\%$ confidence interval for λ, we often consider two choices. One is
$$\left[\bar{X} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\bar{X}}, \quad \bar{X} + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\bar{X}}\right]$$
and the other is
$$\left[\left(\sqrt{\bar{X}} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2, \quad \left(\sqrt{\bar{X}} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2\right],$$
where z_α denotes the upper αth percentile of a $N(0,1)$ distribution. This exercise involves examining the accuracy of the two intervals by estimating their coverage probabilities using Monte Carlo simulation for various combinations of n and λ. For this investigation, take confidence level $1-\alpha = 0.95$, $n = 5, 10, 30$, and $\lambda = 1, 2, 4$. This means we have a total of $3 * 3 = 9$ settings of (n, λ) to investigate. [30 points]
   (a) For a given setting of (n, λ), compute Monte Carlo estimates of coverage probabilities of the two confidence intervals by simulating appropriate data, using them to construct the two confidence intervals, and repeating the process 1,000 times. Use your QE ID as the starting seed value and use the same data to compute both the intervals. [15 points]
   (b) Repeat (a) for the remaining combinations of (n, λ). Present an appropriate summary of the results. [5 points]
   (c) Interpret the results in (b). Be sure to mention which interval you would recommend in which situation. Provide justification for all your conclusions. [10 points]
2. Use the data in the tab-separated-values file "graduateEarnings.txt". Cost of higher education has risen rapidly. Is college worth the expense? Money magazine collected the data containing the following information:

#	Variable	Description
1	School	School name
2	Public	Public (1), Private (0)
3	Location	School location
4	Earn	Early career earnings
5	SAT	Average SAT score
6	ACT	Average ACT score
7	Price	Net price of a degree
8	Price_with_aid	Net price less average amount of aid
9	need_fraction	Fraction of students receiving need-based aid
10	merit_aided	Fraction of students receiving merit-based grants

We would investigate best predictors of earnings. [35 points]

(a) Visually inspect the relationship between earnings (Earn) and net price (Price). Fit a linear model to predict earnings based on net price. Is it a good model? Why or why not? [6 points]

(b) Improve your model by adding other important variables if possible. Check assumptions of the model and state your conclusions. Interpret the model coefficients and r-squared. [10 points]

(c) Comparing your model in (b) with the simple one in (a), has the coefficient of Price changed? Explain your findings. [4 points]

(d) One student fit a model Earn ~ 1 + Price + Public + ACT + SAT, and by looking at the t-tests of the coefficients, he concludes that ACT is useful in predicting earnings while SAT is unimportant. Comment on this conclusion. [5 points]

(e) An educator conjectures that students whose ACT scores are above 25 will gain less in earnings than those with ACT below 25 for every dollar they spend in college education. Perform a statistical analysis to test this hypothesis. Note you want to control for all other variables that you deem important. Clearly state the hypotheses associated with this test (you can perform a two-sided test) and interpret the result. [10 points]

3. Use the data in "bullet.txt". Ballistic panels, or bullet resistant panels, are protective shields designed to absorb the energy and prevent or reduce the damage of a bullet or similar high velocity projectiles. The dataset is from a published paper. The experiment is to test a cloth ballistic panel with different layers of a fabric. Three types of 0.22 caliber bullets are fired at various initial velocities. The variables are: [35 points]

bulltype: Bullet Type (1=Rounded, 2=Sharp, 3=FSP);
layers: Number of layers in the panel (2,6,13,19,25,30,35,40);
v50: The velocity at which approximately half of a set of projectiles penetrate the fabric panel (m/sec).
[*Hint: do a transformation of (v50/100)^2 and use it as your response variable.*]

(a) Draw plots to explore the relationship between the response and number of layers for each of the three types of bullets. [5 points]

(b) The authors of the paper split the data to 3 subsets, one for each bullet type, and subsequently fitted 3 independent regression models. They want to know whether the 3 slopes are equal. Repeat their analysis and construct a 95% confidence interval for the slope in each model. [5 points]

(c) If you were involved in the study, what suggestions would you give to improve the statistical analysis above? Fit your model. Verify the assumptions used with the model. Compare your results with the analysis in (b). [10 points]

(d) The researchers would like to test the equality of the 3 slopes. Conduct a hypothesis testing on the parameters of an appropriate model. Use $\alpha = 0.05$. Also construct 95% confidence intervals for pairwise comparisons. What are your conclusions? [10 points]

(e) A graduate student suggests to treat the number of layers as a categorical variable. Do you think this is a good idea? Why or why not? [5 points]