Homework 5

name: your name

INSTRUCTIONS:

- Submit the solutions to all questions as a single PDF file on Gradescope.
- Please show all your work to receive full credit.
- All answers must be typeset preferably using LaTeX (the template is provided, so please use it). Start each exercise on a new page and write your answers inside the `exercise` environments provided.
- For all coding questions, or questions for which you rely on any computations please include the code and the outputs that reproduce your analyses in your PDF (you can use Ctrl+P on your Colab notebook, save as PDF and merge with your submission document). Don't forget to run all cells before submitting.
  If a question only involves coding, write "Code provided below." in the solutions box and insert the PDF pages with the relevant code on the next page.
- When submitting on Gradescope, please indicate which pages of your solution are associated with each problem.
- You are free to discuss homework solutions with other students, but write-ups must be done independently. Be sure to give credit to any students you collaborate with.

Failure to comply with these instructions will result in a deduction of points.

Optional Questions

1. Do you have any confusion or questions about the previous lectures?
2. Any suggestions or thoughts about the course?

Suggestions

Exercise 1

Traffic Study on Residential Streets [30 points]. A survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. The following table displays the number of bicycles and other vehicles recorded in the study.

Type of street	Bike route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154
Fairly busy	yes	8/29, 35/415, 31/425, 19/42, 38/180, 47/675, 44/620, 44/437, 29/47, 18/462
Fairly busy	no	10/557, 43/1258, 5/499, 14/601, 58/1163, 15/700, 0/90, 47/1093, 51/1459, 32/1086
Busy	yes	60/1545, 51/1499, 58/1598, 59/503, 53/407, 68/1494, 68/1558, 60/1706, 71/476, 63/752
Busy	no	8/1248, 9/1246, 6/1596, 9/1765, 19/1290, 61/2498, 31/2346, 75/3101, 14/1918, 25/2318

Now restrict, your attention to the first four rows of the table: the data on residential streets.

1. **[8 points]** Let $y_1, \ldots, y_{10}$ and $z_1, \ldots, z_8$ be the observed proportion of traffic that was on bicycles in the residential streets with bike lanes and with no bike lanes, respectively (so $y_1 = 16/(16 + 58)$ and $z_1 = 12/(12+13)$ for example). Set up a model so that the y_i's are independent and identically distributed given parameters θ_y and the z_i's are independent and identically distributed given parameters θ_z.

2. **[4 points]** Set up a prior distribution that is independent in θ_y and θ_z.

3. **[10 points]** Draw 1000 simulations from the posterior distribution. (Hint: θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)

4. **[8 points]** Let $\mu_y = E(y_i|\theta_y)$ be the mean of the distribution of the y_i's; μ_y will be a function of θ_y. Similarly, define μ_z. Using your posterior simulations from 3, plot a histogram of the posterior simulations of $\mu_y - \mu_z$, the expected difference in proportions in bicycle traffic on residential streets with and without bike lanes.

Solution

Exercise 2

Regression with many explanatory variables [40 points]. Table 15.2 displays data from a designed experiment for a chemical process. In using these data to illustrate various approaches to selection and estimation of regression coefficients, Marquardt and Snee (1975) assume a quadratic regression form; that is, a linear relation between the expectation of the untransformed outcome, y, and the variables x_1, x_2, x_3, their two-way interactions, x_1x_2, x_1x_3, x_2x_3, and their squares, x_1^2, x_2^2, x_3^2.

Table 1: Data from a chemical experiment, from Marquardt and Snee (1975). The first three variables are experimental manipulations, and the fourth is the outcome measurement.

Reactor temperature (°C), x_1	Ratio of H_2 to n-heptane (mole ratio), x_2	Contact time (sec), x_3	Conversion of n-heptane to acetylene (%), y
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	19.5

Data is available here: http://www.stat.columbia.edu/~gelman/book/data/factorial.asc

1. **[5 points]** Fit an ordinary linear regression model (OLS frequentist model), including a constant term and the nine explanatory variables above.

2. **[10 points]** Fit a Bayesian linear regression model with a uniform prior distribution on the constant term and a shared normal prior distribution on the coefficients of the nine variables above. If you use iterative simulation in your computations, be sure to use multiple sequences and monitor their joint convergence.

3. **[10 points]** Discuss the differences between the inferences in (1) and (2). Interpret the differences in terms of the hierarchical variance parameter. Do you agree with Marquardt and Snee that the inferences from (1) are unacceptable?

4. **[8 points]** Repeat (1), but with a t_4 prior distribution on the nine variables.

5. **[7 points]** Discuss other models for the regression coefficients.

Solution

Exercise 3

Hierarchical model [30 points]. This is a question that intends to help you with your final project. Find a data set on Kaggle or UCI Machine Learning Repository (you can reuse the dataset you chose from the previous homework or choose a new one), is there any group structures in the dataset? Can youå expand your one-parameter model in the previous homeworks into a hierarchical model? Interpret the results from both model fits; what do they tell you about the dataset?

Compare the one-parameter model and the hierarchical model fit; consider performing posterior predictive check. Which model fit the data better?

This question will be graded based on effort, as opposed to correctness. You will get full mark if substantial effort is demonstrated.

Solution