STATS 551 Homework 4

Regression & Model Choice

Due date: 6:00 pm (EST) Mar. 26, 2018

Regression with many explanatory variables (5×12 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.

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

Table 15.2 *Data from a chemical experiment, from Marquardt and Snee (1975). The first three variables are experimental manipulations, and the fourth is the outcome measurement.*

1. Fit an ordinary linear regression model (that is, nonhierarchical with a uniform prior distribution on the coefficients), including a constant term and the nine explanatory variables above.

2. Fit a mixed-effects 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. 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. Repeat (1), but with a t_4 prior distribution on the nine variables.

5. Discuss other models for the regression coefficients.

Model selection (2×20 points). Diabetes data: A population of 532 women living near Phoenix, Arizona were tested for diabetes. Other information was gathered from these women at the time of testing, including number of pregnancies, glucose level, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age. This information appears in the file azdiabetes.dat. In this exercise we will be modeling the conditional distribution of glucose level (glu) as a linear combination of the other variables, excluding the variable diabetes.

1. Fit a regression model using the g-prior with $g = n, \nu_0 = 2, \sigma_0^2 = 1$. Obtain posterior confidence intervals for all of the parameters.

2. Perform the model selection and averaging procedure described in Section 9.3 in P. Hoff's book. Obtain $Pr(\beta_j \neq 0|y)$, as well as posterior confidence intervals for all of the parameters. Compare to the results in part 1.

Guideline for Submission: Submit R markdown (or jupyter notebook) with annotated code followed by results. Discussions about the results should follow the results.

Optional Reading. Read one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. Hoeting, Jennifer A., et al. "Bayesian model averaging: a tutorial." Statistical science (1999): 382-401.

2. Posada, David, and Thomas R. Buckley. "Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests." Systematic biology 53.5 (2004): 793-808.

3. Karlis, Dimitris, and Loukia Meligkotsidou. "Multivariate Poisson regression with covariance structure." Statistics and Computing 15.4 (2005): 255-265.

4. Genkin, Alexander, David D. Lewis, and David Madigan. "Large-scale Bayesian logistic regression for text categorization." Technometrics 49.3 (2007): 291-304.

5. Park, Trevor, and George Casella. "The bayesian lasso." Journal of the American Statistical Association 103.482 (2008): 681-686.

6. Yi, Nengjun, and Shizhong Xu. "Bayesian LASSO for quantitative trait loci mapping." Genetics 179.2 (2008): 1045-1055.

7. Chib, Siddhartha, and Edward Greenberg. "Bayes inference in regression models with ARMA (p, q) errors." Journal of Econometrics 64.1 (1994): 183-206.

8. Ng, Andrew Y., and Michael I. Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems. 2002.

9. Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g Prior Distributions". In Goel, P.; Zellner, A. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies in Bayesian Econometrics. 6. New York: Elsevier. pp. 233?243. ISBN 0-444-87712-6.

10. George, E.; Foster, D. P. (2000). "Calibration and empirical Bayes variable selection". Biometrika. 87 (4): 731?747. doi:10.1093/biomet/87.4.731.

11. Liang, F.; Paulo, R.; Molina, G.; Clyde, M. A.; Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection". Journal of the American Statistical Association. 103 (481): 410?423. doi:10.1198/016214507000001337.