Regression Analysis
Final Exam Review

Review your midterm exam. Anything covered by that exam or on that exam review is fair game.

When doing PCA/PCR, why is it typically important to standardize the predictor variables?

In PCA, what do the eigenvalues represent? *Answer: they're the variances along the PCs.* Also, know that the PCs point in the directions of the eigenvectors of the covariance matrix, which are the orthogonal directions that account for the most variation in the data.

Know how to read a scree plot to determine an appropriate number of PCs to retain in a model.

Be able to write down the objective functions to be minimized when doing

(a) least-squares regression:

$$\min_b (Y - Xb)^T (Y - Xb)$$

(b) ridge regression:

$$\min_b (Y - Xb)^T (Y - Xb) + \lambda \sum_{k \geq 1} b_k^2$$

(c) lasso regression:

$$\min_b (Y - Xb)^T (Y - Xb) + \lambda \sum_{k \geq 1} |b_k|$$

How does R go about finding the best tuning parameter in lasso or ridge regression? *Answer: cross-validation.*

Know the three link functions we've studied in doing binomial response regression (their names and formulas).

$$\begin{aligned} \text{logit link:} \quad \eta(x) &= \log \frac{x}{1-x} \\ \text{probit:} \quad \eta(x) &= \Phi^{-1}(x), \;\; \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ \text{complementary log-log:} \quad \eta(x) &= \log(-\log(1-x)) \end{aligned}$$

Know the formulas for Mallow's Cp, AIC, BIC (also called SBC), R^2, adjusted R^2, MSE, and SSE. Know that these are all standard goodness-of-fit measures.

$$\begin{aligned} \text{Mallow's } Cp &= \ldots \\ AIC &= \ldots \\ BIC &= \ldots \\ R^2_{adj} &= \ldots \\ R^2 &= \ldots \\ SSE &= \ldots \\ MSE &= \ldots \end{aligned}$$

Know how the method of best-subsets works- be able to write down the method in your own words.

Know generally how step-wise regression methods work, and that they are useful when there are too many variables to apply best-subsets, but can lead to mediocre (or worse) models.

In logistic/binomial response regression, what is deviance? Know the formula for it. What is the null deviance? What is the residual deviance?

$$D = 2 \log \frac{L_{large}}{L_{small}}$$

If the response is truly a binomial random variable for fixed predictor variable values, and if the number of trials at each predictor variable point is "large", what is the approximate distribution of the residual deviance? *Answer:* χ^2 *with* $d.f. = n - s$ *= number of data points - number of parameters.*

Do we prefer models with small or large deviance?

Given some data and logistic/binomial regression output from R, be able to use that output to predict probabilities of the response outcome based on combinations of the predictors. Also be able to calculate the odds ratio (or percentage to which the odds have changed) due to a certain factor level over the null model (Refer to the breast feeding example, pp 32-34 in the second Faraway text.

Know how to use the chi-square distribution to get a p-value for the residual deviance, and be able to interpret this.

For categorical predictors: given some data, be able to write down the model matrix. Remember, if you have a categorical variable that can take on one of

f factor levels, you should only use f-1 columns (or dummy variables) in your model matrix- else your model matrix will be overdetermined and inverting $X^T X$ will be impossible.

Example:

Animal	V1	V2	Response
Dog	9	12	32
Dog	8	1	31
Dog	2	7	30
Dog	1	7	29
Cat	3	3	30
Cat	0	9	29
Cat	4	6	31
Cat	6	2	32
Snake	1	1	41
Snake	6	6	39
Snake	2	-5	41
Snake	8	11	39

Define $d_1 = \mathbb{1}\{\text{dog}\}$; $d_2 = \mathbb{1}\{\text{cat}\}$. Then the model matrix can be written as

$$X = \begin{pmatrix} & d1 & d2 & V1 & V2 \\ 1 & 1 & 0 & 9 & 12 \\ 1 & 1 & 0 & 8 & 1 \\ 1 & 1 & 0 & 2 & 7 \\ 1 & 1 & 0 & 1 & 7 \\ 1 & 0 & 1 & 3 & 3 \\ 1 & 0 & 1 & 0 & 9 \\ 1 & 0 & 1 & 4 & 6 \\ 1 & 0 & 1 & 6 & 2 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 6 & 6 \\ 1 & 0 & 0 & 2 & -5 \\ 1 & 0 & 0 & 8 & 11 \end{pmatrix}$$

Again with categorical predictors... Be able to judge significance of model terms based on summary outputs for the four types of models at the bottom of page 209 in Faraway (see the example there in the following pages).

More on categorical predictors... Be able to judge significance of factors and model coefficients based on model output and scatterplots (like those made using ggplot2 on pages 212 and 215). Also, be able to predict response based on predictor variable values given model output.

Expect a lot of R output on the exam. Answering the questions will require you to be able to interpret the output.

Know about the bias-variance trade-off. What tends to happen to test/validation SSE as model flexibility increases? What tends to happen to training SSE as model flexibility increases?

