MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #9 (Linear and Logistic Regression)

Last modified: April 9, 2014

Reading from Chihara and Hesterberg

- Chapter 9. This chapter is full of ideas that might fit well into a final project. We will return to Chapters 7 and 8 later.

  You may ignore the details of sections 9.4.1, 9.4.2 (except for example 9.8), 9.5.2, and 9.6.1.

Proof of the Week

- None – we did it last week!

R scripts

- Script 9A-LinearRegression.R
  Topic 1 - Covariance and correlation for two random variables.
  Topic 2 - Least-squares regression for two random variables.
  Topic 3 - a maximum-likelihood approach
  Topic 4 - checking the MLE approach when the residuals really have a normal distribution.

- Script 9B-ResamplingRegression.R
  Topic 1 - getting a confidence interval for regression coefficients.
  Topic 2 - applying the bootstrap to linear regression.
  Topic 3 - Permutation test for lack of independence.

- Script 9C-Logistic regression.R
  Topic 1 - doing linear regression when the response is always zero or one.
  Topic 2 - a maximum likelihood approach.
  Topic 3 - logistic regression.

Mathematical notes

1. Covariance and correlation

   (a) The covariance of random variables X and Y is defined as
   $\mathrm{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]$
   Prove that $\mathrm{Cov}[X, Y] = E[XY] - E[X]E[Y]$.

   (b) The correlation coefficient of random variables X and Y is defined as

   $$\rho(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

   Prove that $|\rho(X, Y)| \leq 1$.

   (c) Prove that when calculating the sample correlation r, you can divide $\sum(x_i - \overline{x})(y_i - \overline{y})$ by $n, n - 1$, or 1 in the numerator, as long as you do the same thing in the denominator.

2. Least-squares regression

   You have values x_i of a "predictor" and matching values y_i of a "response." Your goal is to minimize the sum of squares of the prediction errors,

$$g(a,b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2.$$

   Prove that

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, a = \overline{y} - b\overline{x}.$$

3. The connection between correlation and the slope of the regression line.

   Define $ss_{xy} = \sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}); ss_x = \sum_{i=1}^{n}(x_i - \overline{x})^2; ss_y = \sum_{i=1}^{n}(y_i - \overline{y})^2.$

   Prove that $r^2 ss_y = b^2 ss_x$.

   An observation is y_i; a predicted observation is $\hat{y}_i = a + bx_i; \overline{y} = a + b\overline{x}$. Prove that the ratio of the variance of the predicted y's to the variance of the observed y's equals R-squared, the square of the sample correlation r.

   Prove that the ratio of the variance of the residuals $y - \hat{y}$ to the variance of the observed y's equals $1 - r^2$.

4. Maximum likelihood regression

   You have a fixed set of values, x_i, of a "predictor" variable.

   For each x_i, the response Y_i is a random variable whose expectation is $\mu_i = \alpha + \beta x_i$ and whose variance is σ^2. The residuals $Y_i - \mu_i$ are independent.

   Given a set of pairs of values $(x_1, Y_1), (x_1, Y_1), \cdots (x_n, Y_n)$, prove that the maximum-likelihood estimates of α and β are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}.$$

   while $\hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i-1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ (note: divide by $n-2$ for an unbiased estimator).

5. Logistic regression

   You have a fixed set of values, x_i, of a "predictor" variable. Each "response" variable Y_i is a Bernoulli random variable with parameter p_i.

   Assume that

$$p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}.$$

   (a) Prove that $\alpha + \beta x_i$ is equal to the "log odds" $\ln \frac{p_i}{1-p_i}$.

   (b) Prove that $0 < p_i < 1$.

   (c) Given a set of pairs of values $(x_1, Y_1), (x_1, Y_1), \cdots (x_n, Y_n)$, form the likelihood function $L(\alpha, \beta)$ and express its logarithm in terms of α and β. Do not attempt to maximize it!

Section problems

1. The last section problem from Module 8, which has already come under discussion.
2. Exercise 11 on page 295. Answers are on page 404. You can answer all these questions with only the information that is supplied, but you need the results on page 5.
3. Exercise 21 on page 297. Partial answers on page 404.
4. Exercise 33 on page 299. Answers on page 405. There are lots of similar examples in sports – perhaps this is a good topic for a final project! (I made Westinghouse Science Talent Search Top 40 in 1959 by dining linear regression of runs scored against the sum of total bases and runners reaching base, an idea that was rediscovered in 1984.)

Homework assignment This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 9 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. The next-to-last problem from Module 8.
2. The last problem from Module 8.
3. Exercise 12 on page 295. For part (c) look on page 261 of page 5 of the math notes.
4. (a) Exercise 14 on pages 295-296.

   (b) Exercise 24 on page 298.
5. Exercise 34 on pages 299-300 – very similar to the last section problem.