Research Journal (Multivariate Regression)

Aleksandra Orlova
Fall 2006

Week 9/11 - 9/15

• What was done this week:

Read the project draft.

Started studying of linear regression and regression analysis with researching general information on these topics on the Internet (*http://en.wikipedia.org/wiki/Regression*). Read simple examples of using linear regression to estimate expected values of unknown parameters.

Studied regression using J.E. Freud's *Mathematical Statistics with applications*, including the following topics: Linear regression; Method of least squares; Normal Regression analysis; Multiple linear regression (estimation of one random variables using known values of multiple variables).

Searched materials on Multivariate Regression.

• Questions and difficulties:

Questions about the project draft:

* formulation with a fixed basis;
* cost of computing (r).

1. Had troubles understanding how Normal Correlation Analysis is set up.
2. Construction of confidence intervals for linear regression coefficients (why we use chi-squared distribution to estimate coefficients).

• What needs to be done next:

Meet with Dr. Lin to figure out questions 1 and 2.
Do some practice problems on linear regression before turning to multivariate cases.
Continue stuyding regression using M. Bilodeau *Theory of multivariate statistics*. Start reading multivariate regression.
Acquire more materials on Multivariate Regression.

Week 9/18 - 9/22

• What was done this week:

Figured out the unanswered questions from the last week (correlation analysis; confidence intervals).

Summarized the methods of bivariate and multiple linear regression; least squares and normal analysis regression methods.

Studied the multiple regression method in M. Bilodeau *Theory of multivariate statistics*.

• What needs to be done next:

Continue to study multivariate regression in detail. Summarize ideas of applying it to the studied method.

Week 9/25 - 9/29

• What was done this week:

Studied different methods of evaluating how good a given regression model is. Summarized ideas for different assumptions (Gauss-Markov, normal, general multiple regression case).

• Questions and difficulties:

Deriving formulas for multiple regression correlation coefficient. Studied this material in C. Rencher's *Linear Models in Statistics*.

• What needs to be done next:

Continue researching the multiple regression correlation coefficient (R and R^2) and its properties. Study in detail any method that can be of interest. Start *Model Validation and Diagnosis* for the multivariate regression model.

Week 10/2 - 10/6

• What was done this week:

Studied the cross-validation technique for validation a regression model:

- common types (holdout, K-fold, leave-one-out) and their advantages and disadvantages (source - Smith, M., *Neural Networks for Statistical Modeling*);

- prediction sum of squares residuals (source - Walpole, R., and Myers, R., *Probability and Statistics for Engineers and Scientists*).

Briefly studied another statistic (C_p) that is used as a model validation technique.

• Questions and difficulties:

Disadvantages of the leave-one-out cross-validation.

Computational costs for different types of cross-validation

• What needs to be done next:

Study other techniques for model validation and diagnosis.

Week 10/9 - 10/13

• What was done this week:
Briefly studied model validation and diagnostics methods for linear multiple regression (outliers, the Hat matrix, influential observations and leverage - potentially influential outliers).
Outlined different methods for choosing best models and algorithms for selecting subsets of variables.

• Questions and difficulties:
Using f-distribution values as the termination criteria for stepwise regression, forward and backward elimination. Why we choose $n - p$ or $n - p - 1$ degrees of freedom for these values.

• What needs to be done next:
Study Adjusted R-Square technique for model validation; Leaps and bounds method for selecting subset of predictors.
Research the performance measures mentioned in the Benchmarking paper (classification error; decomposition into model bias, model variance and data variance).

Week 10/16 - 10/20

• What was done this week:
Studied the adjusted R-square statistic: another method of evaluation of a multiple regression model.
Studied the general ideas of nonlinear regression methods in Rencher's *Linear Models in Statistics*. Acquired reference literature mentioned in the book.
Briefly read about logistic, loglinear and Poisson regression models.

• What needs to be done next:
Continue to research different nonlinear regression models.

Week 10/23 - 10/27

• What was done this week:

Researched different approaches to validation of nonlinear regression models, including residual mean square minimization, Variance/Covariance matrix for parameters, plots of observed vs. predicted values and fitted function. Researched methods of model validation that are used in Statistics module of the Orange Software. Currently, mean squared error is the only evaluation function provided for regression. The software also provides different measures of quality of logistic models, such as classification accuracy, ROC statistics, sensitivity and specificity. The program implements cross validation (leave-one-out) technique for testing learning algorithms.

• Questions and difficulties:

Found contradictory information on using the R-square statistic for nonlinear regression models.

Goodness-of-fit Chi-square test. Some sources suggest that the least squares criterion for the best fit is realized by minimizing χ^2. Or it can be compared to the chi-square distribution to determine the goodness of fit. It is still unclear how exactly it should be compared to determine the goodness of fit. Some sources give a slightly different formula for this statistic (using the maximum likelihood function) and suggest that it can only be used for probit and logistic regression models.

• What needs to be done next:

Continue to research different parameter estimation methods for nonlinear models.

Research the Brier score (the average deviation between predicted probabilities for a set of events and their outcomes); Chi-squared goodness-of-fit test.