# Statistics 203
# Introduction to Regression and Analysis of Variance
# Assignment #3
# Due Tuesday, March 1

Prof. J. Taylor

USE R FOR ALL CALCULATIONS. PROVIDE COPIES OF YOUR CODE IN THE ASSIGNMENT.

Q. 1) (**MP, 9.26**) Consider the model

$$y_i = \theta_1 - \theta_2 e^{-\theta_3 x_i} + \varepsilon_i, \qquad 1 \leq i \leq n.$$

This is called the Mitcherlich equation, and it is often used in chemical engineering. For example, $y_i$ may be yield and $x_i$ may be reaction time.

(a) Is this a nonlinear regression model?

(b) Graph the expectation function for the parameter values $\theta_1 = 0.5, \theta_2 = -0.1$ and $\theta_3 = 0.1$. Discuss the shape of the function.

(c) Graph the expectation function for the parameter values $\theta_1 = 0.5, \theta_2 = 0.1$ and $\theta_3 = 0.1$. Discuss the shape of the function. Compare the shape with the shape in part (b).

(d) The file

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/chlorine.table

contains the fraction of active chlorine in a chemical product as a function of time after manufacturing. Plot the data and fit the Mitcherlich law to the data, including the fitted curve on the original plot.

(e) Provide approximate confidence intervals for the parameters.

Q. 2) Consider the one-sample problem: $Y_i \sim N(\mu, \sigma^2), 1 \leq i \leq n$ with the $Y_i$'s i.i.d. The MLE is of course

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

If we constrain $|\mu|^2 \leq C$ and transform the problem to a penalized minimization problem we had to solve

$$\widehat{\mu}_\lambda = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \mu)^2 + \lambda \mu^2.$$

(a) Find a design matrix $X$ such that

$$\widehat{\mu} = (X^t X)^{-1} X^t Y.$$

(b) Show that

$$\widehat{\mu}_\lambda = \frac{\widehat{\mu}}{1 + \lambda/n}.$$

(c) Find a design matrix $X(\lambda)$ and a data vector $Y(\lambda)$ such that

$$\widehat{\mu}_\lambda = (X(\lambda)^t X(\lambda))^{-1} X(\lambda)^t Y(\lambda).$$

(HINT: $Y(\lambda)$ WILL GENERALLY HAVE TO BE OF LENGTH $n+1$ OR GREATER – I.E. YOU NEED TO ADD AN OBSERVATION TO THE ORIGINAL $Y$, AS WELL AS AN ENTRY TO THE ORIGINAL $X$.)

(d) Generalize this to the constrained regression problem for a vector of non-negative constraints $\lambda = (\lambda_0, \ldots, \lambda_{p-1})$

$$\widehat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}} \left( \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j X_{ij})^2 + \sum_{j=0}^{p-1} \lambda_j \beta_j^2 \right).$$

(e) Write a function in `R` that takes two arguments, one the output of `lm`, the other a vector $\lambda$ of length $p$ as above and returns $\widehat{\beta}_\lambda$. (HINT: THE FUNCTION `model.matrix` WILL LIKELY BE USEFUL).

Q. 3) Consider the viral load data from Assignment # 2 found at

http://www-stat/~jtaylo/courses/stats203/data/vl.table

(a) As it seems that the variance of viral load depends on their GSS, use weighted least squares (WLS) with appropriately chosen weights to refit the model. Has this improved the diagnostic plots for this model? (HINT: YOU WILL HAVE TO USE THE PEARSON RESIDUALS $r_i = (Y_i - \widehat{Y}_i) * \sqrt{w_i}$ FOR THE DIAGNOSTIC PLOTS. THE DIAGNOSTIC PLOTS CAN ALSO HELP YOU CHOOSE APPROPRIATE WEIGHTS.

(b) Does this significantly affect the results compared to ordinary least squares (OLS), in terms of confidence intervals, or $p$-values? Report both the "weighted" and "unweighted" confidence intervals. Which do you feel are more accurate?

Q. 4) (**NKNW, 14.10**) A marketing research firm was engaged by an automibile manufacturer to conduct a pilot study to examine the feasibility of using logistic regression for ascertaining the likelihood that a family will purchase a new car during the next year. A random sample of 33 suburban families was selected. Data an annual family income and the current age of the oldest family automobile were obtained. A follow-up interview conducted 12 months later was used to determine whether the famillt actually purchased a new car or did not purchase a new car. The data can be found at

http://www-stat.stanford.edu/~jtaylo/courses/stats203/data/car.table

(a) Using a logistic regression model, find the MLEs of the parameters $\beta_0, \beta_{income}, \beta_{age}$.

(b) State the response function.

(c) Find $\exp(\widehat{\beta}_{income})$ and $\exp(\widehat{\beta}_{age})$ and interpret them.

(d) What is the estimated probability that a family with annual income of 50,000$ and an oldest car of 3 years will purchase a new car next year?

(e) Plot the standard diagnostic plots – are there any outliers, anything unusual?

(f) Use a partial deviance test to test whether the age of oldest family automobile can be dropped from the regression model; use $\alpha = 0.15$. What is the approximate $p$-value?

(g) Test whether the two-factor interaction effect between annual family income and age of oldest automobile should be added to the regression model containing family income and age of oldest automobile as first-orer terms; use $\alpha = 0.05$. What is the approximate $p$-value?

(h) Repest the previous test using Pearson's $X^2$ instead of partial deviance.