Applied Linear Statistical Models
Assignment 6

This assignment focuses on the analysis of binomial data based on the logistic regression model. You should write your report and do your analysis in a `.Rmd` or a `.Rnw` file. Please upload the resulting `.html` or the `.pdf` in Canvas before midnight on the 4th of November (Friday). The assignment counts for 8.33% of the final grade in the course.

The data

Here we will analyze data on low birth weight. These data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986. The dataset can be found in the book *Applied Logistic Regression* by Hosmer and Lemeshow (2000), Wiley, New York. The dataset is available in R as `birthwt` and it is within the package `MASS`, see

`https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/birthwt.html`

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Four variables were believed to be of importance, namely; age; weight of the subject at her last menstrual period; race; and the number of physician visits during the first trimester of pregnancy.

The assignment

The response variable is

$$y_i = \begin{cases} 1 \text{ if the } i\text{-th baby's weight is less than or equal to 2500 g,} \\ 0 \text{ if the } i\text{-th baby's weight is more than 2500 g,} \end{cases}$$

where $i = 1, ..., n$, n is the number of mothers, and $n = 189$. Here, we assume that the y_i's are independent Bernoulli random variables, that is, $y_i \sim \text{Bin}(1, \mu_i)$, $i = 1, ..., n$, where μ_i is the probability of the i-th baby weighing is less than or equal to 2500 g. The following generalized linear model is proposed,

$$\eta_i = g(\mu_i) = \log(\mu_i) - \log(1 - \mu_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij},$$

for $i = 1, ..., n$, and thus,

$$\mu_i = g^{-1}(\eta_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij}\right)}.$$

1. Fit a logistic regression model to the low birth weight dataset by using the `glm` function in `R`. Use all the explanatory variables. Report a summary of the estimates.

2. Use the p-values to identify three explanatory variables that should potentially be removed from the model. Once these three variable have been identified fit all 8 combination of models with these three explanatory variables either in or not in the model. Use AIC to determine the best model out of these 8 models. Report the summary of the selected model.

3. Draw a normal probability plot of the deviance residuals. Do the deviance residuals appear to follow a normal distribution? Draw the deviance residuals versus all the explanatory variables. Are any outliers found when looking at these plots? The deviance residuals are evaluated by the `glm` function.

4. Interpret all the parameters in the model selected in 2.

5. A particular mother has the following explanatory variables;
   `age` $= 33$ year old;
   `lwt` $= 107$ lb;
   `race` $= 2$ (black);
   `smoke` $= 0$ (doesn't smoke);
   `ptl` $= 0$, (no previous premature labours);
   `ht` $= 1$ (hypertension present);
   `ui` $= 0$ (uterine irritability not present);
   `ftv` $= 0$ (number of physician visits during the first trimester is zero).

   What is the probability that the mother will have a baby weighing less than 2500 g? Compute a 95% confidence interval for this probability based on the model selected in 2. Here an estimate of the covariance of the estimator for β is needed for the calculations. *Hint*: The '`vcov()`' function is helpful here.