STA 302 H1F / 1001 HF – Fall 2011 – A. Gibbs

Assignment 3

Estimating percentage of body fat from simple body measurements

Due: Tuesday, November 29 at 12:00 noon.
Include both your name and student number on your assignment. Underline your surname.

You should plan to hand in your assignment in lecture on that day. Assignments will be accepted before lecture starts and at the end of lecture. No late assignments will be accepted without a valid reason. If you cannot attend lecture on that day, you must drop off your assignment before noon in my office, SS 5016A. Do not drop off the assignment after noon. If my office door is closed do not knock. Assignments will not be accepted after noon. Assignments will not be accepted if they are handed in to anyone other than Dr. Gibbs. There is no tutorial on Tuesday, November 29.

Presentation of solutions is important. In particular, it is inappropriate to hand in pages of SAS output without explanation or interpretation. The only SAS output you need to submit with assignments is relevant plots. Quote relevant numbers from your SAS output as part of your solutions. You do not need to hand in your SAS code. The marking scheme includes a mark out of 3 for presentation of solutions.

The Data:
The data for this assignment are available on the course website under Assignments or at
`www.utstat.utoronto.ca/alisong/Teaching/1112/Sta302/assignments.html`.
You can read more about the data at R.W. Johnson (1992), Fitting Percentage of Body Fat to Simple Body Measurements, *Journal of Statistics Education*, volume 14, number 1. The problems in the data noted in this article have been fixed.

The data are measurements on 252 men. The data include several easily obtained body measurements and a measurement of the percentage of body fat. Calculating percentage of body fat is laborious, so we would like an equation to predict it from the easily obtained measurements.

The variables you are given are:
- Identification number
- Percentage of body fat (calculated using Siri's equation)
- Age (years)
- Weight (pounds)
- Height (inches)
- Adiposity index (or BMI) = weight/height^2 (kg/m^2
- Neck circumference (cm)
- Chest circumference (cm)
- Abdomen circumference (cm)
- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Extended biceps circumference (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

In this assignment, you will examine some methods for selecting a subset of the predictor variables to best model the percentage of body fat.

Use SAS to the the analysis for the following.

1. Look at scatterplots and correlations of the response with the explanatory variables. For which explanatory variables is there evidence of a linear relationship with the percentage of body fat? Is there evidence of any linear relationships between the potential explanatory variables?

   The following SAS code might be useful for this question:

   ```
   ods graphics on;

   proc corr plots=matrix(nvar=all);
     var  bodyfat age weight height adiposity neck chest abdomen hip thigh knee
              ankle bicep forearm wrist;
   run;

   ods graphics off;
   ```

   In the output from `proc corr`, the first number that is printed for each pair of observations is the correlation. The second number is the p-value for the two-sided test with null hypothesis that the correlation between the two variables is 0.

2. Fit a multiple regression model with all 14 explanatory variables. Look at the estimated regression coefficients and the p-values for the t-tests for these coefficients. Are any of these surprising?

3. A common goal in regression is to find a parsimonious model, that is the simplest model that adequately describes the dependent variable. One method for model selection choses the "best" model from all possible subsets of the predictor variables. Different criteria for "best" are possible. We will consider three:

   - Adjusted R^2
   - AIC (Akaike's Information Criterion)

     $$\text{AIC} = n \ln\left(\frac{SSE}{n}\right) + 2(p+1)$$

     The first term in AIC is a measure of the lack of fit of the model while the second is a penalty term for additional paramters in the model. The model with the smallest AIC is considered the "best".
   - SBC (Schwarz's Bayesian Criterion)

     $$\text{SBC} = n \ln\left(\frac{SSE}{n}\right) + (p+1)\ln(n)$$

     SBC is a Bayesian modification of AIC. The model with the smallest SBC is considered the "best".

The following SAS code will list the models with the 20 highest values of Adjusted R^2, as well as the values of AIC and SBC for those models. (There are many other possible models but this is enough to look at and I checked that these 20 will include the "best" models for all three criteria.)

```
proc reg;
  model deptvar = indeptvars / selection = adjrsq aic sbc best=20;
```

`deptvar` is the dependent variable and `indeptvars` is a list of the independent variables, with a space between each one.

Which explanatory variables are in the chosen models? Are there any surprises?

4. Another commonly-used method to find a parsimonious model is stepwise regression. This method starts with a model with no explanatory variables. In forwards steps, the variable that contributes most to the model is added as long as it achieves a specified significance level (we'll use 0.1). In backwards steps, the variable that contributes least to the model is removed as long is its p-value is greater than a specified level (we'll use 0.1). Stepwise regression alternates forwards steps with backwards steps. The idea is to end up with a model where no variables are redundant given the other variables in the model. When doing stepwise regression, SAS uses F-tests rather than t-tests for the coefficients of the variables. These F-test statistics are just the square of the corresponding t-test statistics and the tests are equivalent. The SAS code for stepwise regression is

```
proc reg;
  model deptvar = indeptvars / selection=stepwise slentry=.1 slstay=.1;
```

The final model chosen is given near the end of the output from the stepwise procedure, before the "Summary of Stepwise Selection".

Which explanatory variables does stepwise regression select? Are there any inconsistencies with what you saw in questions 1, 2, and 3?

5. A useful practice is to randomly split the data into a training dataset and a test dataset. The explanatory variables to include in the model are chosen from the training data set, and the resulting model is then evaluated by fitting it to the test data set. Here is some code to randomly split the data into two datasets.

```
data onehalf otherhalf;   /*create 2 data sets called onehalf and otherhalf */
  set originaldata; /* originaldata is what you called the data when you first read it in */
  onehalf=0;
  retain k 126 n 252;
  if ranuni(302) <= k/n then do;    /* 302 is the seed to the random number generator */
    k=k-1;                           /* use this seed so that we all get the same answer */
    onehalf=1;
    output onehalf;
  end;
  else output otherhalf;
  n=n-1;
  drop k n;
```

(a) Using the `onehalf` dataset, find the model chosen by stepwise regression. Then fit a model with these chosen variables to the data in the `otherhalf` dataset. Compare the estimated regression coefficients and p-values for the tests that these coefficients are zero from the model fit to the `onehalf` dataset to the model fit to the `otherhalf` dataset.

(b) Repeat part (a) reversing the roles of the two datasets. That is, now use `otherhalf` as the training data and `onehalf` as the test data.

6. One of my favourite websites is the *Little Handbook of Statistical Practice* by Gerald Dallal. You can find it at `http://www.jerrydallal.com/LHSP/LHSP.HTM`. Go there and read "The Most Important Lesson You'll Ever Learn About Multiple Linear Regression Analyses" (at `http://www.jerrydallal.com/LHSP/important.htm`). How has the analysis in this assignment illustrated Dallal's point?

Marking scheme: You will be given a mark out of 3 for clear and neat presentation of solutions. Each of the questions is worth 3 marks. 3 marks will be awarded for complete, correct answers, or answers with only very minor problems. Good answers that are unclear or have some mistakes or are missing some aspects of the solution will be awarded 2 marks. Poor answers that have some value will be awarded one mark. Note that sometimes an answer awarded 3 marks will not be perfect. You should always look at the solutions.

The total number of marks for the assignment is 21.