DATA7202 Semester 1, 2018

Objectives

On completion of this laboratory session you should be able to:

1. Examine associations between a continuous response and a continuous explanatory variable using scatterplots and correlation
2. Examine associations between a continuous response and a categorical explanatory variable using tables and multiple box-plots and scatterplots.
3. Use R to fit a multiple regression model
4. Explain the output.
5. Interpret the results of multiple regression models.

Reading:

Telford, R.D. and Cunningham, R.B. Sex, sport and body-size dependency of hematology in highly trained athletes. (Telford91_hematology_Aust_athletes.pdf)

Case Study

Scientists at the Australian Institute of Sport continually monitor athletes to try to improve their performance and to prevent injuries and illness. Here we will try to find out what sorts of variables are related to blood hemoglobin concentration. Various hematological attributes are thought to vary among athletes from different types of sport. Hemoglobin concentration influences how efficiently the body uses oxygen and can only be measured via a blood sample. We will use multiple regression to try to predict hemoglobin concentration using a range of other variables including body mass index (BMI), gender and sport which can be more easily measured from athletes. We hope to answer the following general questions:
How well can we predict hemoglobin concentration given measurements on these other variables ?
Which variables are significantly associated with hemoglobin concentration, given the other variables ?
What can we interpret about the relationship between hemoglobin concentration and the significantly associated variables ?

The data for this case study are from a cross-sectional survey and were obtained from Richard Telford and Ross Cunningham of the Australian National University. Measurements were made on 102 male and 100 female athletes, with blood samples taken from a forearm vein at least 6 hours after a training sessions were completed. This is a subset of the data originally collected by Telford and Cunningham and we will only consider some of the measurement types they recorded.

The data set is stored in a comma-separated variable file called **ais4.xls** (or ais4.csv). It can be obtained from the Blackboard site under Datasets. You should open this file in Excel and then copy it all (**ctrl-A, ctrl-C**) into R.

<u>**Dataset**</u>: Data has been recorded for each subject on their:
sex: male or female
sport: basketball, field – e.g. shot put, gymnastics, netball, rowing, swimming, track (>400m), track (100-400m), tennis, or water polo. In the data file, these are labelled Bball, Field, Gym, Netball, Row, Swim, T400m, TSprint, Tennis, WPolo.
hemoglobin: hemoglobin concentration in the blood in grams per decilitre
BMI: Body Mass Index: weight (kg)/ square of height (m^2)
percentBF: percent body fat estimated from skin fold results

Analysis

Section 1: Exploratory Data Analysis:

Here we aim to gain an intuitive understanding of the data, looking for patterns, similarities and differences, to help guide subsequent analysis. Most of the conditions for a successful regression analysis can be checked via residual plots, which we will look at later on. Numerical summaries are also worthwhile, to obtain information on e.g. sample size. Given the number of subjects involved here, any numerical answer should be rounded to at most 3 significant figures.

1i) First you should find out how many observations we have for each combination of the categorical variables, calculate associated totals and produce a table.
Which combinations of categorical variables are missing entirely from the dataset? Also mention all the combinations of categorical variables that are present, but have a sample size of less than 10.

A sample size under 10 is not too bad, but any estimates associated with a group become less accurate as the number of observations in that group becomes smaller.

1ii) We would also like to know some basic summary statistics about the hemoglobin variable, i.e. mean and standard deviation. Use R to produce these.

We could also look at the continuous variables further via tables, but one tends to notice things faster when looking at each the distribution of each continuous variable. First, we will look at the overall data.

1iii) Produce a box plot of the hemoglobin variable.
Repeat this for each of the continuous variables. Note that we shouldn't put them all on a single plot because the values are not directly comparable.

1iv) Comment on whether or not you see any outliers likely to be data entry errors or other incorrect values (explain). Also comment on whether the data looks normal (normally distributed) or skewed or otherwise in any of the three boxplots. If any of the variables look skewed, mention in which direction (towards higher or lower values)? The skew direction is the one in which the data is more spread out, but otherwise use any terminology you like.

Now produce a **hemoglobin** box plot for each combination of the categorical variables.

1(v) Copy the graph into your report and give it a figure number and caption.

1(vi) What do you notice when looking at the hemoglobin box plots ? Specifically: does there seem to be much difference on average between the genders or between some sports within each gender ? (When comparing sports, it would be worth comparing against the rowers – we will use them as a baseline later on.)

Now we will look at the continuous variables to see if they seem to be directly associated with hemoglobin. It is worth remembering that the relationships will be different when variables are used in combination to predict hemoglobin levels, as we will see when running multiple regression.

First we will prepare scatterplots with (simple) linear regression lines to predict hemoglobin levels.

1(vii) Add a y-axis label to the plot. Repeat all this for **percentbf** as a predictor of **hemoglobin** (create the plot, copy it over and give it a figure number and caption). Note: you will have to disable each of the previous scatter plots (or edit and alter them) to avoid them all being shown at once.

We will now look at the correlations between these continuous predictor variables and the hemoglobin response variable.

Produce the correlation values for all possible pairs of these variables and the associated p-values (null hypothesis: correlation = 0; alt: correlation ≠ 0).
1(viii) Record the correlation between hemoglobin and each continuous variable in words, along with the p-value. Based on this information, which of these variables do you think is likely to provide the better predictor of hemoglobin level ? Explain why ?

Section 2: Multiple Regression

In this section, we will try to fit a multiple regression model for hemoglobin concentration based on the all the available predictors, i.e. 2 continuous variables and 2 categorical variables.
You may have noticed that the variable **sex** is listed as either “female” or “male” in the dataset. These need these to be converted into numerical form before they can be used in multiple regression.

We also need to process the sport variable. 10 sports are represented in the dataset, so we need to create a set of indicator variables instead.

We need to convert the single sport variable into 10 sports ‘indicator’ variables (sp1-sp10) which can be analysed. Every subject has a set of these 10 variables recorded, but all of them are given a value of 0 except the sport this person specializes in, which is given a 1. So a subject whose main sport is Basketball (Sport1) would have the following set of indicator values recorded: 1 0 0 0 0 0 0 0 0 0 . Another subject whose sport is Rowing (Sport 5) would have the following set of indicator values recorded: 0 0 0 0 1 0 0 0 0 0 . Check that you have been able to do this conversion.

The same principle of one categorical indicator variable being redundant applies to any categorical variable. Here the other categorical variable is the sport variable, which has 10 possible values, i.e. the dataset contains data on athletes from 10 different sports. One of these 10 indicator variables is not needed, so we should leave one out. But which one ? It is actually ok to leave any one of the sports indicator variables out, but then some of what is reported will be relative to the sport that got left out. It becomes the baseline for comparison against the other sports. The main things that are affected are coefficients and their confidence intervals and the p-values. The adjusted R^2 value would be the same no matter which sport you leave out. One could argue for making various sports the baseline, but here I would like you to make *Rowing (sport 5)* the baseline sport (leaving it out of the list of variables in the regressions) because we happen to have the most data recorded for rowers, with plenty of rowing subjects from both genders.

Now we’ll try using multiple regression to predict **hemoglobin** concentrations in athletes on the basis of the other variables, i.e. **bmi, gend1, percentbf** and **sp1-sp10** (except for **sp5**: rowing).
Choose hemoglobin as the Dependent variable; choose as “Independent” variables: **bmi percentbf gend1 sp1 sp2 sp3 sp4 sp6 sp7 sp8 sp9 sp10**.

Note the p-value for the whole multiple regression equation, i.e. of **hemoglobin = β_0 (constant) + β_1*bmi + β_2*percentbf + β_3*gend1 + β_4*sp1 + … + β_{12}*sp10** , where the β’s are the coefficients of the explanatory (or “independent” or “predictor”) variables.

The null hypothesis is that $\beta_0 = \beta_1 = \beta_2 = ... = \beta_{12}$**.** The alternative hypothesis is that at least one of the β's is non-zero.

2(i) Record the p-value for the regression equation and state whether or not the multiple regression is significant at the 5% level.

The next thing to be interested in is the **adjusted R-squared** value, which is an estimate of the proportion of variance in the response variable (here: hemoglobin), which has been explained or accounted for by the combination of predictor variables. Whenever you are using more than one predictor variable in regression (always in multiple regression), the adjusted R-squared value is a better estimate than the listed R-squared value because it takes into account the number of predictor variables used. Without this adjustment, there is a tendency for the basic R^2 value to artificially fall as you add more predictor variables, even if these aren't any use in prediction.

2(ii) Record the adjusted R-square value. What proportion of the variance in hemoglobin levels has been explained via this multiple regression ?

2(iii) Refer back to Q1(ii) and the sample standard deviation for hemoglobin ($\mathbf{s_y}$). Given this value and the adjusted R-squared result, what would you expect the standard deviation of the residuals ($\mathbf{s_{residual}}$) to be after using this multiple regression ?

Note: a residual is the difference between the true hemoglobin concentration for a subject and the predicted hemoglobin concentration from the multiple regression equation. So we will in theory have a residual for each subject in the dataset when using this multiple regression (i.e. with the coefficients it found).
You can assume that (1- adjusted R^2 value) = $\mathbf{s_{residual}^2}$ / $\mathbf{s_y^2}$. Please show some working.

We will continue our analysis of this dataset in the next lab session.