

Multiple Regression Analysis

The purpose of this exercise is to consider some of the steps in carrying out regression analyses with SAS. Here, we will use a data set of $n=31$ cases designed to study the relation between physical fitness (measured by oxygen consumption in a standard task) and other measures

case	oxy	age	weight	runtime	rstpulse	runpulse	maxpulse
1	44.609	44	89.47	11.37	62	178	182
2	45.313	40	75.07	10.07	62	185	185
3	54.297	44	85.84	8.65	45	156	168
4	59.571	42	68.15	8.17	40	166	172
5	49.874	38	89.02	9.22	55	178	180
....							

Here, OXY is the outcome measure; people who are **more** fit use **less** oxygen in this task. RUNTIME is the time to run a given distance on a treadmill; RSTPULSE, RUNPULSE and MAXPULSE are measures of pulse rate at rest, average during the run, and maximum during the run. AGE and WEIGHT are obvious control variables.

1. The data is stored in `N:\data\fitnessd.sas`. Read it into SAS using

```
%include data(fitnessd);
```
2. As always, it is useful to get an overview of the data by plotting. By now, you should be familiar with scatterplot matrices, either using the `%scatmat` macro, or SAS/INSIGHT (Analyze -> Scatter). The `%scatmat` macro has the advantage of being able to add useful annotations; SAS/INSIGHT allows interactive use.

```
%scatmat(data=fitness,
          var=oxy age weight runtime rstpulse runpulse maxpulse,
          interp=rl, anno=ellipse);
```


3. From your graphical analysis:
 - a. Which predictor variable(s) seem to be most strongly related to OXY?
 - b. Do any of the predictors seem to be very highly related to each other?
 - c. Which predictor variable(s) seem necessary, on logical grounds, to include in any predictive model?
4. Let's start fitting a regression model with using all variables, to see what we get. (We will consider problems in model selection—which variables to include in MRA models in the lecture.)

```
proc reg data=fitness;
  model oxy=age weight runtime rstpulse runpulse maxpulse;
run;
```

5. Examine the (default) printed output

- a. Do the signs of the regression coefficients seem to make sense?
 - b. Are there any variables that seem not to contribute to predicting OXY? (Hint: consider the $Pr > |t|$ column under parameter estimates.)
 - c. Ask yourself: what else should I know to decide if this model provides an adequate description of fitness for these variables?
6. For today, we will just confine ourselves to getting some additional output that helps to interpret & diagnose a given model. We'll use options on the MODEL statement and ask for some simple plots:

```
proc reg data=fitness;
    model oxy=age weight runtime rstpulse runpulse maxpulse
        / P R vif;
    plot residual. * predicted.;
    plot residual. * NQQ.;
run;
```

7. By default, SAS produces all output in HTML form, with a default style. You can get output in other formats or change the style using ODS statements. You might want to change the working directory used by SAS, by clicking on the “Change current folder” icon at the bottom of the outer SAS window  or else specify the full path in the file= option on the ODS statement.

```
ods pdf file='fitness-tutorial.pdf' style=journal;
proc reg data=fitness;
    model oxy=age weight runtime rstpulse runpulse maxpulse
        / R vif;
run;

ods pdf close;
```

The option R provides some analyses of residuals, including Cook's D measure of influence. VIF gives variance inflation factors for the predictors. We'll explore regression diagnostics and other topics starting next week.

8. Finally, let's examine a whole set of potential models, using R^2 selection, one of several model selection options.

```
proc reg data=fitness;
    model oxy=age weight runtime rstpulse runpulse maxpulse
        / selection=rsquare cp best=4;
run;
```