McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 8, November 7–10 2017

In this lab we shall cover the use of `PROC REG` and `PROC GLM` to fit models which have categorical variables as the covariates.

One-Way ANOVA

The one-way ANOVA procedure is used when we wish to test the equality of two or more group means. Again this can be recast as a linear regression model and fitted with `PROC REG`. We can also use `PROC GLM` or `PROC ANOVA` with class variables to do the test.

We shall use the Salary survey data from the book (Table 5.1, datafile Salary.txt) to illustrate these methods. The variable Salary is the individual's salary and Educ is a categorical variable recording their education level as in three categories. I assume that this has been imported as usual into the SAS dataset S3A3.Salary.

First we shall use `PROC REG` as we did before. This procedure, however, will not work directly with categorical variables and so we must directly define the dummy variables mentioned in class.

```
Data S3A3.Salary1;
    set S3A3.Salary;
   * Construct the dummy variables with 0-1 coding;
    if Educ=1 then E1=1; else E1=0;
    if Educ=2 then E2=1; else E2=0;
    if Educ=3 then E3=1; else E3=0;
run;

PROC REG Data=S3A3.Salary1 plots=none;
    * In this model I will use Category 1 as the reference;
    Model Salary=E2 E3;
run;
```

`PROC GLM` in SAS can be used for fitting general linear models including models with categorical covariates. It works quite similarly to `PROC REG`.

```
PROC GLM Data=S3A3.Salary;
    Class Educ;
```

```
    Model Salary=Educ /solution ;
run;
```

The `Class` statement defines those variables which should be considered categorical in the model. This statement **must** come before the `Model` statement. Any such variables may only appear as covariates in the model. By default `PROC GLM` does not output the estimates but adding `/solution` in the `Model` statement tells it to do so. When you do this you will notice that there is a warning statement saying that the X'X matrix is singular. This is because including all levels of the categorical variable gives a design matrix in which the sum of the dummy variable columns equals the intercept column. The warning can be ignored. One of the levels will have an estimate of 0 and no standard error, t-value or p-value. This is the level that `PROC GLM` used as the reference level.

What plots are output depends on whether there are only categorical covariates, only continuous covariates or both. In this case the default is to produce boxplots of the response variable by levels of the categorical covariate. This plot (and any others for different models) can be suppressed by adding `plots=none` in the initial `PROC GLM` statement.

Finally we can use `PROC ANOVA`.

```
PROC ANOVA Data=S3A3.Salary;
    Class Educ;
    Model Salary=Educ;
    Means Educ;
run;
```

By default, `PROC ANOVA` does not output the estimates since it is primarily used for testing and there is no `solution` option for the `Model` statement. The `Means` statement will output the mean and standard deviation of the response variable for each level of the specified covariate. Note that by default `PROC ANOVA` also outputs boxplots of the response variable for each level of the covariate and this can be again be suppressed by adding `plots=none` to the `PROC ANOVA` statement.

More than one categorical variable

Multiple categorical variables can be included in the model by either including another class variable in `PROC GLM` or by including dummy variables for the other categorical variable(s) also in `PROC REG`.

```
PROC GLM Data=S3A3.Salary plots=none;
    Class Educ Manage;
```

```
    Model Salary=Educ Manage /solution;
run;

PROC REG Data=S3A3.Salary1 plots=none;
    * Manage is already coded as a 0-1 variable;
    Model Salary=E2 E3 Manage;
run;
```

It is also possible to use `PROC ANOVA` but I will not use this method since it is not possible in this case to directly find the estimates of the coefficients in the linear model. In general `PROC ANOVA` is only used to test if categorical variables are related to the response but not to actually fit a linear model.

Interactions of Categorical Variables

The structure given in the previous section is an additive model as explained in class. To fit separate means to each cell in the cross-tabulation of the levels of categorical variables we need to use interactions. Interactions are formed as products of dummy variables. Although it is possible to do these in `PROC REG`, it is not advisable since it is easy to miss including some of the required products making the output difficult to interpret. A better solution is to use `PROC GLM` and class variables.

```
PROC GLM Data=S3A3.Salary;
    Class Educ Manage;
    Model Salary=Educ Manage Educ*Manage / solution;
run;
```

The `Educ*Manage` term in the model statement is the interaction term.

To test if the interaction is needed, we just use the model with the interaction as the full model and the one without as the reduced model and proceed exactly as in multiple regression to compare a reduced model against a full model. We can also use the appropriate F test printed out as part of the **Type III Sums of Squares** table. Note that SAS also outputs a Type I Sums of Squares table but we should not use this in general. the difference between the two is that the order of including covariates matters for the Type I sums of squares but not for the Type III sums of squares.

Exercise: An experiment was conducted to compare the life (in hours) of two different brands of batteries. The batteries were used in three different devices; a radio, digital camera and a portable DVD player. The data are given in the file `Batteries.txt`. Conduct an analysis to answer the following questions.

1. Is there an overall difference in lifetime between the battery brands? This is a model using only brand as the covariate.

2. What happens to the coefficient estimate, standard error and p-value for brand if we take device into account? Explain.

3. Is there any evidence that of an interaction between brand and device? What happens to the main effect of brand if we have an interaction term in the model? Explain.