McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 9, November 14–17, 2017

Categorical and Continuous Covariates

We can combine continuous covariates in a single linear model to give different lines for different levels of a categorical variable in a process that is sometimes called **Analysis of Covariance (ANCOVA)**. The simplest model has parallel lines for each level of the categorical variable or each cell in the cross-tabulation if there are more than one categorical variables. As before we can use `PROC REG` (with dummy variables) or `PROC GLM` (with `class` variables) to do this. I will assume that S3A3.Salary is the original salary dataset from `Salary.txt` and S3A3.Salary1 is the same dataset with added dummy variables E1, E2 and E3 as constructed in a data step in SAS lab 7.

```
PROC REG Data=S3A3.Salary1 plots=none;
   Model Salary=E2 E3 Exp;
run;

PROC GLM Data=S3A3.Salary;
   CLASS Educ;
   Model Salary=Educ Exp /solution;
run;
```

It is important to take be able to extract the fitted lines from the output of `PROC REG` or `PROC GLM` for these models. In this case the lines are

$$\begin{aligned}
\widehat{\text{Salary}} &= 10474.00 + 548.61 \times \text{Exp} && \text{if Education level is 1 (High school diploma)} \\
\widehat{\text{Salary}} &= 13695.12 + 548.61 \times \text{Exp} && \text{if Education level is 2 (University degree)} \\
\widehat{\text{Salary}} &= 15254.14 + 548.61 \times \text{Exp} && \text{if Education level is 3 (Post-graduate degree)}
\end{aligned}$$

The levels of the categorical variable(s) in this setup only affect the intercepts, not the slope. The following will give 6 parallel lines for each cell in a table of education level and management.

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

For non-parallel lines we need to use interactions between categorical and continuous variables. The most general model for the salary example is to have 6 unrelated lines. The code to fit this using `PROC GLM` is

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

The six lines from this model are

$$\begin{aligned}
\widehat{\text{Salary}} &= 9481.39 + 496.12 \times \text{Exp} && \text{for non-managers with a high school diploma only} \\
\widehat{\text{Salary}} &= 13417.12 + 504.68 \times \text{Exp} && \text{for managers with a high school diploma only} \\
\widehat{\text{Salary}} &= 10808.44 + 502.02 \times \text{Exp} && \text{for non-managers with a university degree} \\
\widehat{\text{Salary}} &= 19806.86 + 489.11 \times \text{Exp} && \text{for managers with a university degree} \\
\widehat{\text{Salary}} &= 11189.45 + 502.36 \times \text{Exp} && \text{for non-managers with a post-graduate degree} \\
\widehat{\text{Salary}} &= 18283.84 + 492.51 \times \text{Exp} && \text{for managers with a post-graduate degree}
\end{aligned}$$

From the Type III sum of squares table we see that we can omit the final term which will give an additive structure in the slopes. In fact we can compare the last two models and see that the F statistic to test the hypothesis of 6 parallel lines is only $F = 0.16$ calculated using the method for comparing nested models given in Chapter 3. If parallel lines are sufficient then this would be an observation from an $F_{5,34}$ distribution. The critical value for this distribution can be found using the methods shown in SAS Lab 3 but that is not necessary in this case. It is clear from Table A.4 in the book that the 5% critical value will be around 2.5 (true value is 2.49) and so the p-value is much larger than 0.05. We therefore do not reject H_0 and conclude that parallel lines are sufficient.

Exercise Use the data in `Respiratory.txt` to model how Forced Expiratory Volume at 1 second `FEV1` relates to `Height` taking race and gender into account.

1. Fit the full model with a separate intercept and slope for each of the 4 cells and give the regression lines.
2. Fit a reduced model which gives each of the cells the same slope but different intercepts and test if this model is significantly worse than the full model.
3. Test if we need to include race in the full model.