McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 10, November 21–24, 2017

Transformations

It is relatively easy to fit transformations in SAS. All that is really necessary is to use a `Data` step to construct the transformed variables and then use either `PROC REG` or `PROC GLM` to fit the model. The functions `Log` and `Sqrt` are particularly useful. Also `**` is used to raise a variable by a power (not necessarily integer). When looking for a transformation, it is important to use the standard diagnostic plots to guide the process. The aim of the process is to find the transformation(s) which make the diagnostic plots as close as possible to what we expect if the three main assumptions (linearity, homogeneity of variance, normality) are at least approximately satisfied. We will look at an example modelling stopping distance of a car as a function of its speed. The data file is called `Stopping.txt` on the website.

The first step is to look at the model without any transformations. This is the easiest to interpret so should be used unless it violates any of the assumptions.

```
PROC REG Data=S3A3.Stopping;
  Model Distance=Speed;
  Plot Distance*Speed;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;
```

We now use a `Data` step to add some transformed variables to the dataset. We will first transform the response variable using various power transformations.

```
 Data S3A3.Stopping;
  Set S3A3.Stopping;
  RecDist=1/Distance;
  LogDist=Log(Distance);
  SqrtDist=Sqrt(Distance);
  Dist2=Distance**2;
run;

PROC REG Data=S3A3.Stopping;
  Model Dist2=Speed;
  Plot Dist2*Speed;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;

PROC REG Data=S3A3.Stopping;
```

```
  Model SqrtDist=Speed;
  Plot SqrtDist*Speed;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;

PROC REG Data=S3A3.Stopping;
  Model LogDist=Speed;
  Plot LogDist*Speed;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;

PROC REG Data=S3A3.Stopping;
  Model RecDist=Speed;
  Plot RecDist*Speed;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;
```

Use this code to verify that the square root transformation for the response is the one which does the best job of satisfying the standard assumptions of the linear model.

Transformations of the covariate(s) are done similarly. We would use another `Data` step to add transformed versions of speed to the dataset. If we use the square root transformation for the response variable as found above, we can compare different transformations of the covariate and select the one that minimizes the mean squared error (maximizes R^2). We would still need to check the diagnostic plots for this model to ensure that the assumptions are still met.

Exercise: For the dataset `Stopping.txt` described above, verify that the identity transformation of speed is the power transformation which minimizes the mean squared error for models with response variable the square root of stopping distance. You need only look at the standard power transformations described on Page 6-15 of my Lecture notes.

Exercise: The file `Cedar.txt` on the website contains measurements of the diameter at breast height in millimetres (labelled Dbh) and total height in decimetres of 138 Atlantic White Cedar trees (taken from a much larger dataset with different species of trees).

a Fit the linear model without any transformation and examine the diagnostic plots.

b Consider transformations of the covariate, Dbh. Show that, of the transformations X, $\sqrt{X}$, $\log(X)$, $X^{-0.5}$ and X^{-1}, the log transformation is the transformation which gives the best fitting model as measured by R^2 and $\hat{\sigma}^2$ but that the homogeneity of variance assumption is suspect.

c Show that also taking the log of the response variable, Height, solves this issue and all model assumptions seem valid.