McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 5. October 17–20, 2017

Diagnostic Plots in Regression

In class we are currently covering various diagnostic plots to look for violations of the assumptions of a linear regression model. In these labs we shall see how such plots can be created in SAS.

1. Simple Linear Regression

In simple linear regression the following plots are useful

1. Scatterplot of the response and predictor variable with the fitted line.
2. Plot of the (studentized) residuals against the fitted values.
3. Normal QQ plot of the (studentized) residuals.

We can have a number of `Plot` commands within `PROC REG` to construct a number of these plots. These typically use keywords to extract or calculate the appropriate values. These keywords all end with a period. In this example we use the (internally) studentized residuals (`Student.`) and the fitted values (`Predicted.`). To construct normal quantile-quantile plots we need to compute the appropriate quantiles of the standard normal for the current value of n. SAS will do this automatically for us and the quantiles can be accessed using the keyword `nqq.`. The following SAS code (using the dataset in `Heights.txt`) does the regression and produces these plots.

```
PROC REG Data=S3A3.Heights;
  Model DHeight=MHeight;
  PLOT DHeight*MHeight;
  PLOT Student.*Predicted.;
  PLOT Student.*nqq.;
run;
quit;
```

None of the major assumptions of the model seem violated in this example. Note that I used the internally studentized residuals in the example but could have used either the raw residuals using the keyword `Residuals.` or the externally studentized residuals (`RStudent.`) as well.

2. Multiple Linear Regression

In multiple linear regression with a relatively small number of covariates, a useful plot before fitting the model is the scatterplot matrix which we saw in the third SAS lab. Here I illustrate this using the dataset from the file `Examination.txt` that you used in the second assignment.

```
PROC SGSCATTER Data=S3A3.Examination;
   Matrix F P1 P2;
run;
quit;
```

Once we have fitted the regression model we usually want the following plots

1. Plot of the observed against the fitted values.
2. Plot of the residuals against the fitted values.
3. Normal QQ Plot of the residuals.

These are produced exactly as they were for the simple linear regression case as in the following code.

```
PROC REG Data=S3A3.Examination;
  Model F=P1 P2;
  Plot F*Predicted.;
  Plot Student.*Predicted.;
  Plot Student.*nqq.;
run;
quit;
```

The Added Variable Plot

The added variable plot relates to the effect of adding a new covariate into a regression. This plot is called a *partial residual plot* in SAS and can be created using the `partial` option in the `Model` statement in `PROC REG` as follows

```
PROC REG DATA=Exams;
  Model F=P1 P2/partial;
run;
quit;
```

Exercises

1. For each of the Anscombe Datasets (`anscombe.txt`), examine the three diagnostic plots described in the first section above.

2. For the data in `CSData.txt` consider a model with response `GPA` and three predictor variables `HSM`, `HSS` and `HSE`. Fit the model and construct the appropriate diagnostic plots as described above. Also construct added variable plots for each of the predictor variables.