McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 11, November 28 – December 1, 2017

Weighted Least Squares

Weighted least squares can be done either using a transformation model as we saw in class or using the `weight` statement in `PROC REG` or `PROC GLM`. The data in `HeteroData.txt` were simulated to have variance proportional to x_i^2. The following code will fit the raw model and its diagnostic plots from which we note that there is clear violation of the homogeneity of variance assumption. We then fit the weighted model in each of these two ways. Note that the role of intercept and slope in the transformed model have been reversed as we saw in class.

```
PROC REG Data=S3A3.hetero plots=none;
    Model y=x;
    Plot y*x;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;

Data S3A3.hetero;
    Set S3A3.hetero;
    yprime=y/x;
    xprime=1/x;
    weights=1/x**2;
run;

PROC REG Data=S3A3.hetero plots=none;
    Model yprime=xprime;
    Plot yprime*xprime;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;

PROC REG Data=S3A3.hetero plots=none;
    Model y=x;
    weight weights;
    Plot y*Predicted.;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
quit;
```

Whenever we know the relative variances (weights) of the observations we can also use either transformation or weighted least squares as shown in class notes. The dataset `physics.txt` contains data from experiments examining the strong interaction force which holds nuclei of particles together. The experiment used collisions of beams unstable elementary particles. The beams had different energies (momentum) denoted by s and the response was the scattering cross-section y. A theory in physics says that

$$E(Y \mid s) \;=\; \beta_0 + \frac{\beta_1}{\sqrt{s}}$$

The data consists of the values of $x = 1/\sqrt{s}$, the average y observed for each level of s and also the standard deviations of y for each level of s. Since the experiments consisted of huge numbers of particles at each momentum these standard deviations can be considered proportional to the true values of $\mathrm{SD}(Y \mid s)$ and so these can be used to construct weights.

First we examine a model relating x to y without accounting for unequal variances.

```
PROC REG Data=S3A3.Physics plots=none;
    Model y=x;
    Plot y*x;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
quit;
```

Next we use the transformed model $Z \;=\; \beta_0 x_1 + \beta_1 x_2 + \varepsilon_z$ where $z = y/SD$, $x_1 = 1/SD$, $x_2 = x/SD$ and $\varepsilon_z = \varepsilon/\mathrm{SD}$. We note that $\mathrm{Var}(\varepsilon \mid x) \;=\; SD^2$ then $\mathrm{Var}(\varepsilon_z \mid x) \;=\; 1$ so we have common variance. Note that there is no intercept in this model.

```
Data S3A3.Physics;
    Set S3A3.Physics;
    z=y/SD;
    x1=1/SD;
    x2=x/SD;
run;

PROC REG Data=S3A3.Physics plots=none;
* Adding / noint to the model statement fits a model with no intercept;
    Model z=x1 x2 / noint;
    Plot z*x2;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
quit;
```

Finally we use weighted least squares with weights equal to the $1/SD^2$.

```
Data S3A3.Physics;
    Set S3A3.Physics;
    w=1/SD**2;
run;
```

```
PROC REG Data=S3A3.Physics plots=none;
    Model y=x;
    Weight w;
    Plot y*x;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
quit;
```

From the output of this analysis we note that there is some evidence of deviation from linearity. If we transform the response variable we would have to also change the weights so this is not really desireable. Instead we can transform the covariate which is often useful when dealing with a lack of linearity. In this case using x^2 as the covariate works better.

```
Data S3A3.physics;
    Set S3A3.physics;
xsquare=x**2;
run;

PROC REG Data=S3A3.Physics plots=none;
    Model y=xsquare;
    Weight w;
    Plot y*xsquare;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
run;
quit;
```

Exercise 1: The `Physics1.txt` dataset is from the same set of experiments except that a different particle was examined. Repeat the above analysis for this dataset.

Unknown Weights and Two-Stage Estimation

When the weights are unknown we need to estimate them from the residuals. In general this is only possible when the variability is associated with a categorical variable. In that case we can run an unweighted regression first and use the estimated variability of the residuals in each category of the categorical variable to estimate the weights. It is common in such analyses to rescale the weights in such a way that that the weighted sum of squared errors will still estimate an appropriate quantity. This is most often done using the mean squared error from the original unweighted model or something proportional to it such as the average squared residual as done here and in the textbook. The following code will do this analysis for the Education expenditure data in Table 7.3 of your textbook. For simplicity I will just focus on the 48 states excluding Alaska and Hawaii, the data are in `Education75a.txt` on the website. I have imported it into the SAS dataset `S3A3.ED75a` for this exercise.

```
PROC REG Data=S3A3.ED75a plots=none;
    Model Y=X1 X2 X3;
    Plot Y*Predicted.;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
```

```
    Output Out=S3A3.ED75a_out
         Predicted=Fitted
         Student=Res_stud
         Residual=Res;
run;
quit;
```

From the residual plot we see that there is a non-constant variability problem. a plot of the residuals by region shows that this could be caused by differences in the variability in the four regions.

```
PROC GPLOT Data=S3A3.ED75a_out;
    PLOT Res_stud*Region;
run;
```

We can use `PROC Means` to get the variances of the residuals by region. Note that the book used the raw residuals but here we use the studentized residuals instead. In general this should not have a large effect unless there are the leverages are very different.

```
PROC MEANS Data=S3A3.ED75a_out NWAY;
    CLASS Region;
    VAR Res_stud;
    OUTPUT Out=temp
          Mean=Mean
          Var=Var;
run;

PROC PRINT Data=temp;
run;
```

Next we store the 4 variances in variables and then construct a new column in our dataset containing these residual variances.

```
Data temp;
    set temp;
    IF (_N_=1) THEN call symput('var1', Var);
    IF (_N_=2) THEN call symput('var2', Var);
    IF (_N_=3) THEN call symput('var3', Var);
    IF (_N_=4) THEN call symput('var4', Var);
run;

Data S3A3.ED75a;
    Set S3A3.ED75a;
    IF (Region=1) THEN var=&var1;
    IF (Region=2) THEN var=&var2;
    IF (Region=3) THEN var=&var3;
    IF (Region=4) THEN var=&var4;
run;
```

We shall use the sample variance of the studentized residuals to rescale the within-region variances

```
PROC MEANS Data=S3A3.Ed75a_out;
     Var Res_stud;
     Output Out=temp1
          Mean=Mean
    Var=Var;
run;

PROC PRINT Data=temp1;
run;

Data temp1;
    set temp1;
    Call symput('vare', Var);
run;

Data S3A3.ED75a;
    Set S3A3.ED75a;
    w=&vare/var;
run;
```

Finally we can fit the weighted regression using these weights.

```
PROC REG Data=S3A3.ED75a plots=none;
    Model Y=X1 X2 X3;
    Weight w;
    Var Region;
    Plot Y*Predicted.;
    Plot Student.*Predicted.;
    Plot Student.*nqq.;
    PLOT Student.*Region;
run;
quit;
```

Exercise 2:

(a) Construct the dummy variables needed to include the categorical variable Region in the regression. Is there still non-constant variance when we include these indicator variables as well as X1, X2 and X3 in the unweighted model?

(b) Using the model in (b), recalculate the weights as in our example above and refit the weighted regression. Is there still heteroscedasticity?