```
********************************************************************************
*  TITLE :      SAS GRAIN PRICE PROJECT ANALYSIS
*
*  DESCRIPTION: Final project for BIOS 7400 with Xiao Song, UGA, Spring 2022.
*               Simple analysis of grain price data.
*
*-------------------------------------------------------------------------------
*  JOB NAME:    analysis.SAS
*  LANGUAGE:    SAS v9.4 (on demand for academics)
*
*  NAME:        Zane Billings
*  DATE:        2022-04-22
*
*******************************************************************************;

FOOTNOTE "Job run by Zane Billings on &SYSDATE at &SYSTIME.";

TITLE 'ANALYSIS OF USDA HISTORICAL GRAIN PRICE DATA';

OPTIONS NODATE LS=95 PS=42;

LIBNAME HOME '/home/u59465388/SAS-Grain-Prices';

ODS GRAPHICS / WIDTH = 6in HEIGHT = 3in;

*******************************************************************************;
* Show the descriptor portion of the dataset;
*******************************************************************************;

TITLE2 "CONTENTS OF GRAINS DATASET";

PROC CONTENTS DATA = HOME.GRAINS;
RUN;

*******************************************************************************;
* Plot outcome time series;
*******************************************************************************;

FOOTNOTE; * Remove the footnote so it isn't on the graphs;

* Plot the time series of log grain price over time. This makes a separate
    time series line for each grain.;
TITLE2 "PRICE PER BUSHEL OF GRAINS OVER TIME";
PROC SGPLOT DATA = HOME.GRAINS;
    SERIES X = YEAR Y = LPE / GROUP = GRN;
RUN;

* Color the points of the time series by the President's political party. The
    default colors are already red and blue so we don't need to change them!
    Also plots a gray line underneath the points, since the JOIN option for
    the SCATTER statement will not connect the points in order.;
TITLE2 "GRAIN PRICES AND PRESIDENT'S POLITICAL PARTY OVER TIME";
PROC SGPANEL DATA = HOME.GRAINS;
    PANELBY GRN;
```

```
    SERIES X = YEAR Y = LPE / LINEATTRS = (COLOR = "GRAY") SMOOTHCONNECT;
    SCATTER X = YEAR Y = LPE / GROUP = PARTY
        MARKERATTRS = (SYMBOL = CIRCLEFILLED);
RUN;

ODS GRAPHICS / WIDTH = 6in HEIGHT = 6in;

* Make a boxplot of log price vs. president's political party. This ignores the
    time series information, but can tell us if either party has more high or
    low years compared to the other.;
TITLE2 "LOG PRICE DISTRIBUTION BY PRESIDENT'S POLITICAL PARTY";
PROC SGPANEL DATA = HOME.GRAINS;
    PANELBY GRN;
    HBOX LPE / GROUP = PARTY;
RUN;

ODS GRAPHICS / WIDTH = 6in HEIGHT = 9in;

* Make a scatterplot of the log price vs each covariate, ignoring the time
    series component of the data. There is not an easy way to connect the points
    like a phase portrait using PROC SGPLOT.;
* I divided this into two plots so they would fit on one page nicer. In the final
    manuscript they could be put side by side.;
TITLE2 "SCATTERPLOTS OF PRICE VS COVARIATES";
PROC SGSCATTER DATA = HOME.GRAINS;
    PLOT LPE * (ACR HVT LNR PRD YLD) / REG
        COLUMNS = 2 GROUP = GRN;
RUN;

PROC SGSCATTER DATA = HOME.GRAINS;
    PLOT LPE * (INFL PWR TEMP VALUE) / REG
        COLUMNS = 2;
RUN;

*******************************************************************************;
* Plots of covariates across time;
*******************************************************************************;

* Plot the time series of each covariate, to assess how they change. I split
    this one into two plots to prevent the plots being too small as before.;
TITLE2 "CHANGE IN COVARIATES ACROSS TIME";
PROC SGSCATTER DATA = HOME.GRAINS;
    PLOT (ACR HVT LNR PRD YLD) * YEAR /
        COLUMNS = 2 GROUP = GRN JOIN MARKERATTRS = (SIZE = 0);
RUN;

PROC SGSCATTER DATA = HOME.GRAINS;
    PLOT (INFL PWR TEMP VALUE) * YEAR /
        COLUMNS = 2 JOIN MARKERATTRS = (SIZE = 0);
RUN;

*******************************************************************************;
* Univariate analyses;
*******************************************************************************;

```

```
* Univariate analysis of main outcome (log price) by grain type;
TITLE2 "UNIVARIATE SUMMARY OF GRAIN DATA OVER TIME";

ODS GRAPHICS / WIDTH = 4in HEIGHT = 4in;

PROC UNIVARIATE DATA = HOME.GRAINS PLOTS;
    VAR LPE;
    CLASS GRN;
RUN;

********************************************************************************;
* Bivariate analyses of price and covariates, ignoring time;
********************************************************************************;

TITLE2 "BIVARIATE CORRELATIONS ACROSS NUMERICAL VARIABLES";
* Correlations -- check to see which covariates are correlated with the outcome,
    and which are correlated with each other and should not be modeled
    together.;
PROC CORR PEARSON SPEARMAN DATA = HOME.GRAINS;
    VAR LPE ACR HVT LNR PRD YLD INFL PWR TEMP VALUE;
    BY GRN;
RUN;

TITLE2 "SUMMARY STATISTICS BY PRESIDENTIAL PARTY AND GRAIN";
* Mean difference in LPE by party -- proc corr does not have a point biserial
    option, so we can check the difference/overlap in means and standard errors
    to assess if party seems to impact log price for any of the grains.;
PROC MEANS DATA = HOME.GRAINS MEAN STDERR MEDIAN RANGE NWAY;
    VAR LPE;
    CLASS PARTY;
    BY GRN;
RUN;

********************************************************************************;
* Simple and multiple OLS regression models;
********************************************************************************;

TITLE2 "SIMPLE LINEAR REGRESSION MODELS";

* Model stratified by grain only;
PROC GLM DATA = HOME.GRAINS PLOTS = ALL;
    CLASS GRN;
    MODEL LPE = GRN / NOINT;
RUN;

* Write a macro to fit all regression models of the form
        MODEL COVAR GRN COVAR * GRN
    without having to type out all of the PROC GLM statements. This model will
    be parametrized without an intercept, and will generate all appropriate
    diagnostic plots for the model.;

%MACRO ALLSIMPLE(DAT = , RESP = , PRED = );
    %LET N = %SYSFUNC(COUNTW(&PRED));
    %DO I = 1 %TO &N;
        PROC GLM DATA = &DAT PLOTS = ALL;
```

```
            CLASS GRN;
            MODEL &RESP = %SCAN(&PRED, &I) | GRN / NOINT;
        RUN;
    %END;
%MEND;

%ALLSIMPLE(
    DAT = HOME.GRAINS,
    RESP = LPE,
    PRED = ACR PRD INFL TEMP PWR YEAR
);

* Fit the same model that was used as before, but with party as a covariate.
    Party needs to be in the class statement, and is the only categorical
    variable, so it wasn't worth modifying the above macro to use party
    correctly and I did it manually.;
PROC GLM DATA = HOME.GRAINS PLOTS = ALL;
    CLASS GRN PARTY;
    MODEL LPE = GRN | PARTY / NOINT;
RUN;

TITLE2 "1866 FULL MODEL";
* 1866 FULL MODEL: this model includes all non-correlated predictors that were
    measured in 1866.;
PROC MIXED DATA = HOME.GRAINS PLOTS = ALL;
    CLASS GRN PARTY;
    MODEL LPE =
        GRN HVT PRD INFL PWR YEAR PARTY
        GRN*HVT GRN*PRD GRN*INFL GRN*PWR GRN*YEAR GRN*PARTY /
        NOINT SOLUTION
    ;
RUN;

TITLE2 "1880 FULL MODEL";
* FULL MODEL WITH TEMP (1880 MODEL): this model is the same as the previous
    model, but also includes the temperature anomaly information. Consequently,
    it only uses data from 1880 onwards (even less for sorghum).;
PROC MIXED DATA = HOME.GRAINS PLOTS = ALL;
    CLASS PARTY GRN;
    MODEL LPE =
        GRN HVT PRD INFL PWR YEAR PARTY TEMP
        GRN*HVT GRN*PRD GRN*INFL GRN*PWR GRN*YEAR GRN*PARTY TEMP*PARTY/
        NOINT SOLUTION
    ;
RUN;

********************************************************************************;
* GLS multiple regression analysis;
********************************************************************************;

* Take the better fitting (by AIC) of the two previous models, and run a model
    that can account for correlation using generalized least squares.
    This model assumes exchangeable correlations between each of the time points.;
TITLE2 "GENERALIZED LEAST SQUARES MODEL";
PROC MIXED DATA = HOME.GRAINS PLOTS = ALL;
```

```
    CLASS GRN;
    MODEL LPE = HVT PRD INFL PWR YEAR GRN GRN*HVT GRN*PRD /
        NOINT SOLUTION CHISQ;
    REPEATED;
RUN;

********************************************************************************;
* Simple forecasting;
********************************************************************************;

* Now instead of just using regression models, we can try to fit a more
    flexible forecasting model using PROC ARIMA.
    First, we need a time variable that is actually a SAS date, so we create
    that first.;
DATA TS_DAT;
    SET HOME.GRAINS;
    T = MDY(1, 1, YEAR);
RUN;

* Next we use the IDENTIFY modeling stage. We check up to 30 lags in the
    first ARIMA modeling stage, and also explicitly test for stationarity at
    the first 10 differences using the random walk with drift test. We
    also use the SCAN method, which is a heuristic for identifying
    candidate ARIMA models.;
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE NLAG = 30 SCAN STATIONARITY = (RW = 10);
    BY GRN;
    TITLE2 "ARIMA TESTS";
RUN;

* Next we use the ESTIMATE modeling stage. We fit several different ARIMA
    models to the data in order to see which fits our time series the best,
    and if any have white noise as the error term.;
* One PROC ARIMA can contain multiple ESTIMATE statements, but I split these
    into multiple PROC steps to make the output easier to read.;
* We are basically fitting all of these models to get the AIC and see which is
    the best fit.;

* Model 1: AR(1);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE;
    ESTIMATE P = 1;
    BY GRN;
    TITLE2 "AR(1)";
RUN;

* MODEL 2: AR(2);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE;
    ESTIMATE P = 2;
    BY GRN;
    TITLE2 "AR(2)";
RUN;

* MODEL 3: MA(1);
```

```
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE;
    ESTIMATE Q = 1;
    BY GRN;
    TITLE2 "MA(1)";
RUN;

* MODEL 4: ARMA(1, 1);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE;
    ESTIMATE P = 1 Q = 1;
    BY GRN;
    TITLE2 "ARMA(1, 1)";
RUN;

* MODEL 5: ARIMA(1, 1, 0);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE(1);
    ESTIMATE P = 1;
    BY GRN;
    TITLE2 "ARIMA(1, 1, 0)";
RUN;

* MODEL 6: ARIMA(1, 1, 1);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE(1);
    ESTIMATE P = 1 Q = 1;
    BY GRN;
    TITLE2 "ARIMA(1, 1, 1)";
RUN;

* MODEL 7: ARIMA(0,0,0) (WHITE NOISE);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE;
    ESTIMATE P = 0 Q = 0;
    BY GRN;
    TITLE2 "ARIMA(0, 0, 0)";
RUN;

* MODEL 8: ARIMA(0,1,0) (RANDOM WALK);
PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE(1);
    ESTIMATE P = 0 Q = 0;
    BY GRN;
    TITLE2 "ARIMA(0, 1, 0)";
RUN;

* Finally, we use the best fitting model to make some simple forecasts in
    the FORECAST modeling stage. We also identify outliers of the best
    fitting model.;

PROC ARIMA DATA = TS_DAT;
    IDENTIFY VAR = LPE(1);
    ESTIMATE P = 1 Q = 1;
    OUTLIER;

```

```
    FORECAST LEAD = 10 INTERVAL = YEAR ID = T OUT = GRAIN_FC;
    BY GRN;
    TITLE2 "FORECASTING";
RUN;
```