

\*\*\*\*\*

## Lab 07: Polynomial Regression

\*\*\*\*\*

### OBJECTIVES:

1. Fit polynomial regression models using PROC GLM.
2. Use PROC GPLOT.

Polynomial regression is a statistical modeling technique to fit the curvilinear data that either shows a maximum or a minimum in the curve, or that could show a max or min if you extrapolated the curve beyond your data. The ability to determine a minimum or maximum point based on the experimental data is a useful application of polynomials. The simple polynomial regressions are multiple regression that use power terms of the independent variable ( $X_i$ ) with the form of  $Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_k X_i^k + e_i$ . Notice the subtle difference from multiple linear regression model, here the numbers 2, 3, ..., k represent the powers of the same variable.

When fitting polynomial regression models, if a particular model term is significant, all terms of lower order should be assumed significant and retained in the regression model. For this reason, the sequentially adjusted **Type I sums of squares** should be used when one attempts to test whether a polynomial model is as good as the one with a higher order term. There is also a null hypothesis for each equation that says that it does not fit the data significantly better than a horizontal line; in other words, that there is no relationship between the X and Y variables.

It should be noticed that the fully adjusted regression coefficients are still used to fit the polynomial regression, which usually leads to no practical explanation of regression coefficients. Further, extrapolation outside the range of the fitted experimental data is untenable, and should not be attempted. This is because the shape of the regression function, as predicted by the model, may not at all accurately represent reality outside the range of the experimental data.

All of the assumptions for regression apply to polynomial regression analyses. The assumptions of normality and homogeneity can and should be verified via use of customary diagnostic techniques (Shapiro-Wilk test, residual plots). Residual and influence statistics still work with these models.

Note that **proc glm** is used for today's lab rather than **proc reg**. While they are essentially the same for our analysis, there is one difference: when using **proc glm**, one can enter the higher order terms directly at the model statement, and does not need to create new variables in the data step for each of them, which on the contrary has to be done before **proc reg** works properly.

## LABORATORY INSTRUCTIONS

### Part I.

#### Housekeeping Statements

```
dm 'log; clear; output; clear';
options nodate nocenter pageno = 1 ls=78 ps=53;
```

```

title1 'EXST7014 lab 07, Name, Section#';
ods rtf file = 'c:/temp/lab07.rtf';
ods html file = 'c:/temp/lab07.html';

```

### Data set

The data set that we use contains the per capita state and local public expenditures and associated state demographic and economic characteristics for 48 states during the year of 1960. Detailed information can be found at <http://lib.stat.cmu.edu/DASL/Datafiles/pubexpendat.html>.

The variables in the dataset are:

**EX:** Per capita state and local public expenditures (\$)  
**ECAB:** Economic ability index, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted.  
**MET:** Percentage of population living in standard metropolitan areas  
**GROW:** Percent change in population, 1950-1960  
**YOUNG:** Percent of population aged 5-19 years  
**OLD:** Percent of population over 65 years of age  
**WEST:** Western state (1) or not (0)

In this lab, we will use MET, Percentage of population living in standard metropolitan area, as the independent variable and the expenditure (EX) as the dependent variable. Other variables are dropped from the dataset.

**Data** Expenditure;

**Input** EX ECAB MET GROW YOUNG OLD WEST STATE \$;

**Drop** ECAB GROW YOUNG OLD WEST;

**Cards;**

256	85.5	19.7	6.9	29.6	11.0	0	ME
275	94.3	17.7	14.7	26.4	11.2	0	NH
327	87.0	0.0	3.7	28.5	11.2	0	VT
297	107.5	85.2	10.2	25.1	11.1	0	MA
256	94.9	86.2	1.0	25.3	10.4	0	RI
312	121.6	77.6	25.4	25.2	9.6	0	CT
374	111.5	85.5	12.9	24.0	10.1	0	NY
257	117.9	78.9	25.5	24.8	9.2	0	NJ
257	103.1	77.9	7.8	25.7	10.0	0	PA
336	116.1	68.8	39.9	26.4	8.0	0	DE
269	93.4	78.2	31.1	27.5	7.3	0	MD
213	77.2	50.9	21.9	28.8	7.3	0	VA
308	108.4	73.1	22.2	28.0	8.2	0	MI
273	111.8	69.5	21.8	26.9	9.2	0	OH
256	110.8	48.1	18.3	27.5	9.6	0	IN
287	120.9	76.9	15.5	25.4	9.7	0	IL
290	104.3	46.3	14.9	27.4	10.2	0	WI
217	85.1	30.9	-7.4	30.0	9.3	0	WV
198	76.8	34.1	0.3	29.4	9.6	0	KY
217	75.1	45.8	8.1	28.9	8.7	0	TE
195	78.7	24.6	12.4	30.8	6.9	0	NC
183	65.2	32.2	12.9	32.9	6.3	0	SC

222	73.0	46.0	14.4	30.0	7.4	0	GA
283	80.9	65.6	77.2	25.5	11.2	0	FL
217	69.4	45.6	7.0	30.5	8.0	1	AL
231	57.4	8.6	0.5	32.1	8.7	1	MS
329	95.7	51.3	14.4	28.8	10.4	1	MN
294	100.2	33.2	5.3	27.3	11.9	1	IA
232	99.1	57.9	9.8	25.6	11.7	1	MO
369	93.4	10.6	2.9	30.2	9.3	1	ND
302	88.2	12.7	4.6	28.9	10.5	1	SD
269	99.1	37.6	6.8	26.6	11.6	1	NB
291	102.2	37.4	13.7	26.8	11.0	1	KS
323	86.0	5.0	21.9	30.3	7.4	1	LA
198	68.6	19.1	-6.2	29.4	10.9	1	AR
282	84.9	43.9	6.4	27.4	10.7	1	OK
246	98.8	63.4	24.1	28.8	7.8	1	TX
309	86.2	27.6	39.4	31.5	5.4	1	NM
309	90.2	71.4	74.3	29.7	6.9	1	AZ
334	97.6	22.6	13.4	28.9	9.7	1	MT
284	93.9	0.0	13.3	30.7	8.7	1	ID
454	125.8	0.0	13.7	29.1	7.8	1	WY
344	98.0	6.8	31.5	28.0	9.0	1	CO
307	92.5	67.5	28.7	31.9	6.7	1	UT
333	100.4	63.1	19.9	27.5	9.8	1	WA
343	98.0	50.4	15.7	27.7	10.4	1	OR
421	205.0	74.2	77.8	25.6	6.4	1	NV
380	112.6	86.5	48.5	26.2	8.8	1	CA

;

```
Proc print data=Expenditure;
run;
```

```
Proc gplot data=Expenditure;
Plot EX*MET;
Title2 "Plot of raw data";
Symbol1 c=red v=dot;
Run;
```

**Proc gplot:** It can produce high-resolution graphic plot, and improve the appearance of the plot by defining plot symbols. **C=red** specifies the color of the symbol. **V=dot** would place a dot for each data point. Take time to exam the plot, can you spot a maximum or minimum point?

## Part II.

### **Fitting a Polynomial Regression Model with a cubic term of MET:**

```
Proc glm data=Expenditure;
Title2 "Polynomial regression model with cubic term";
Model EX = MET MET*MET MET*MET*MET / ss1 ss2;
Output out=outdata1 p=yhat1 r=resid1;
Run;
```

```
Proc plot data=outdata1;
Title3 'residual plot';
Plot resid1*yhat1;
Run;
```

```
Proc univariate data=outdata1 normal;
Title3 'univariate procedure on residuals';
Var resid1;
Run;
```

Notice that **PROC GLM** is used for polynomial regression rather than **PROC REG**. While they are essentially the same for the analysis, there is one difference. When using **PROC GLM**, one can enter the higher order terms directly at the model statement, does not need to create new variables in the data step for each of them, which on the contrary has to be done before **PROC REG** works properly. Carefully exam the F-tests of parameter estimates. Keep in mind that the sequentially adjusted **type I sums of squares** are used for polynomial regression.

### Part III.

#### **Fitting a Polynomial Regression Model with a Quadratic term of MET:**

```
Proc glm data=Expenditure;
Title2 "Polynomial regression model with quadratic term";
Model EX = MET MET*MET / ss1 ;
Output out=outdata2 p=yhat2 r=resid2;
Run;
```

```
Proc plot data=outdata2;
Title2 'residual plot';
Plot resid2*yhat2;
Run;
```

```
Proc univariate data=outdata2 normal;
Title3 'univariate procedure on residuals';
Var resid2;
Run;
ods rtf close;
ods html close;
Quit;
```

**LAB ASSIGNMENT**

Your assignment is to perform necessary analysis using SAS and answer the following questions (Please do not print all the output. Only print the graphs and tables that you think are relevant to your answers).

1. Describe the trend in the scatter plot of the raw data: what is the relationship between variables EX and MET?
2. Fit a polynomial regression model with cubic term of MET. When you decide whether the cubic term and quadratic term should be included in the model, do you use the Type I SS or Type II SS? Why?
3. Is the cubic effect significant? How about quadratic and linear effects?
4. Based on your answers to the above questions and the SAS output, which polynomial model do you consider the best? Write down the polynomial model with the estimated coefficient values. Do you keep the linear term in the model? Why?
5. Now assume that there is a state where 80 percent of its residents live in standard metropolitan areas. Use the best model to predict the per capita public expenditure of this state. Is there any problem in doing so?

\*Remember to attach your SAS log with your lab report.