STAT 525

Chapter 8
Quantitative and Qualitative Predictors

Dr. Qifan Song

Polynomial Regression

- Multiple regression using X_i^2, X_i^3, etc as additional predictors to fit/approximate a nonlinear regression
- Hierarchical approaching to fitting (model selection); usually no more than the third polynoial degree
- Generates quadratic, cubic polynomial relationships
- Can often lead to a multicollinearity problem
- Possible solution: centralize the predictors
  - Centering

$$\tilde{X}_{ij} = X_{ij} - \overline{X}_j,$$

  where $\overline{X}_j = \sum_i X_{ij}/n$

Example: Power Cell (p.300)

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment
  - Notice $\sum_i (X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2) = 0 \rightarrow cor(X_1, X_2) = 0$
  - Notice $cor(X_i, X_i^2)$ is a large value.
- Standardizing the explanatory variables
  - For example, we code the values of predictors as $x_{ij} = 0, \pm 1$
  - Notice $\sum x_{i1} x_{i1}^2 = 0 \rightarrow cor(x_1, x_1^2) = 0$
  - Notice $\sum x_{i2} x_{i2}^2 = 0 \rightarrow cor(x_2, x_2^2) = 0$

```
options nocenter;
data a1;
    infile 'U:\.www\datasets525\Ch07ta09.txt';
    input cycles chrate temp;

proc print data=a1;
run;

data a1; set a1;
    chrate2=chrate*chrate;
    temp2=temp*temp;
    ct=chrate*temp;

proc reg data=a1;
    model cycles=chrate temp chrate2 temp2 ct;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

Root MSE	32.37418	R-Square	0.9135
Dependent Mean	172.00000	Adj R-Sq	0.8271
Coeff Var	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	337.72149	149.96163	2.25	0.0741
chrate	1	-539.51754	268.86033	-2.01	0.1011
temp	1	8.91711	9.18249	0.97	0.3761
chrate2	1	171.21711	127.12550	1.35	0.2359
temp2	1	-0.10605	0.20340	-0.52	0.6244
ct	1	2.87500	4.04677	0.71	0.5092

- MULTICOLLINEARITY? Overall significant model but no individual variables significant

```
proc corr data=a1;
    var chrate temp chrate2 temp2 ct;
run; quit;
```

```
             Pearson Correlation Coefficients, N = 11
                   Prob > |r| under H0: Rho=0
             chrate       temp    chrate2      temp2         ct
chrate      1.00000    0.00000    0.99103    0.00000    0.60532
                        1.0000     <.0001     1.0000     0.0485
temp        0.00000    1.00000    0.00000    0.98609    0.75665
             1.0000                1.0000     <.0001     0.0070
chrate2     0.99103    0.00000    1.00000    0.00592    0.59989
             <.0001     1.0000                0.9862     0.0511
temp2       0.00000    0.98609    0.00592    1.00000    0.74613
             1.0000     <.0001     0.9862                0.0084
ct          0.60532    0.75665    0.59989    0.74613    1.00000
             0.0485     0.0070     0.0511     0.0084
```

- As anticipated, high correlation between X_1 and X_1^2 as well as X_2 and X_2^2

- SAS Codes: Standardizing

```
data a2; set a1;
    schrate=chrate; stemp=temp;
    keep cycles schrate stemp;

proc standard data=a2 out=a3 mean=0 std=1;
    var schrate stemp;

proc print data=a3;
run;

data a3; set a3;
    schrate2=schrate*schrate;
    stemp2=stemp*stemp;
    sct=schrate*stemp;

proc reg data=a3;
    model cycles=schrate stemp schrate2 stemp2 sct;
run; quit;
```

- Note that this standardizes values so the mean is zero and variance 1. The key aspect is centering to remove the collinearity.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

Root MSE	32.37418	R-Square	0.9135
Dependent Mean	172.00000	Adj R-Sq	0.8271
Coeff Var	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	162.84211	16.60761	9.81	0.0002
schrate	1	-43.24831	10.23762	-4.22	0.0083
stemp	1	58.48205	10.23762	5.71	0.0023
schrate2	1	16.43684	12.20405	1.35	0.2359
stemp2	1	-6.36316	12.20405	-0.52	0.6244
sct	1	6.90000	9.71225	0.71	0.5092

Remarks

- If we only care the final fitting with all polynomial terms, the two models are exactly the same. As a consequence, the overall F test/statistics are the same.
- But the individual coefficient estimator and inferences are completely different. After centralization, the two "main" effects are significant, implying that a linear model appears reasonable. Could do a general linear test (`test schrate2,stemp2, sct;`).
- Another possible way of handling the multicollinearity in polynomial regression is using orthogonal polynomial (STAT 514).

Interaction Models

- With several explanatory variables, we need to consider the possibility that the effect of one variable depends on the value of another variable
- Model this relationship as the product of predictors
- Special Cases:
  - One binary (Y/N) and one continuous
  - Two continuous predictors

Interaction Models: Special Case #1

- $X_1 = 0$ or 1 identifying two groups
- X_2 is a continuous variable

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- When $X_1 = 0$ (Group 1)

$$Y_i = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$$

- When $X_1 = 1$ (Group 2)

$$Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_{i2} + \varepsilon_i$$

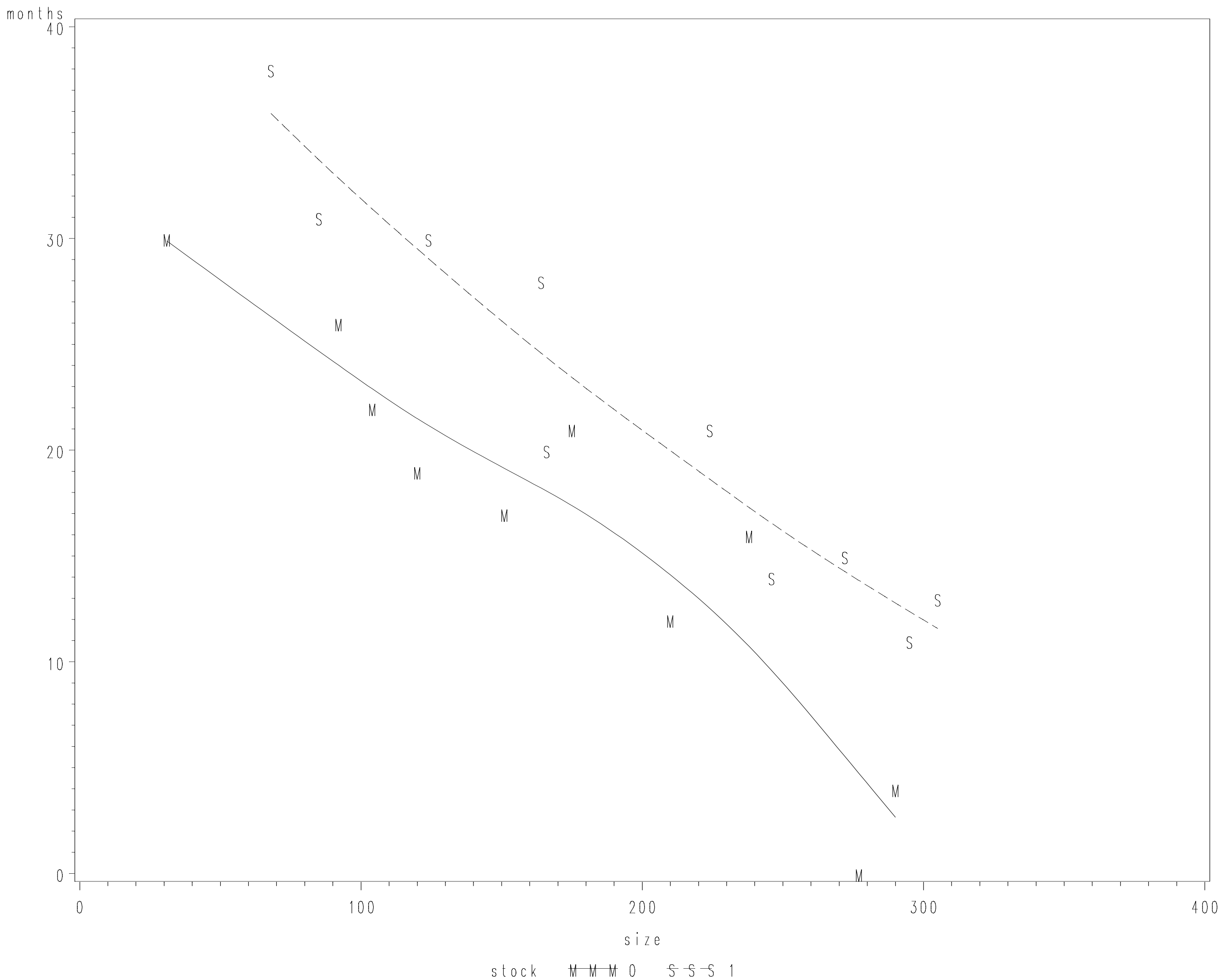
- Results in two regression lines
- β_2 is the slope for Group 1
- $\beta_2 + \beta_3$ is the slope for Group 2
- Similar relationship for the intercepts
- Three Hypotheses of Interest
  - $H_0 : \beta_1 = \beta_3 = 0$: regression lines are the same
  - $H_0 : \beta_1 = 0$: intercepts are the same
  - $H_0 : \beta_3 = 0$: slopes are the same

Example: Insurance Innovation (p.316)

- Y is the number of months for an insurance company to adopt an innovation
- X_1 is the size of the firm
- X_2 is the type of firm
  - $X_2 = 0 \rightarrow$ mutual fund firm
  - $X_2 = 1 \rightarrow$ stock firm
- Do stock firms adopt innovation faster? Is this true regardless of size?

```
data a1;
    infile 'U:\.www\datasets525\Ch8ta02.txt';
    input months size stock;
```

```
/* Scatterplot */
symbol1 v=M i=sm70 c=black l=1; symbol2 v=S i=sm70 c=black l=3;
proc sort data=a1; by stock size;
proc gplot data=a1;
    plot months*size=stock/frame;
run;
```



- Investigate the model: `months = size stock size*stock`
- Test whether different types of firms adopt innovation at different paces regardless of size

```
data a1; set a1;
    sizestoc=size*stock;

proc reg data=a1;
    model months=size stock sizestoc;
    test stock, sizestock;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.41904	501.47301	45.49	<.0001
Error	16	176.38096	11.02381		
Corrected Total	19	1680.80000			

Root MSE	3.32021	R-Square	0.8951
Dependent Mean	19.40000	Adj R-Sq	0.8754
Coeff Var	17.11450		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.83837	2.44065	13.86	<.0001
size	1	-0.10153	0.01305	-7.78	<.0001
stock	1	8.13125	3.65405	2.23	0.0408
sizestoc	1	-0.00041714	0.01833	-0.02	0.9821

Test 1 Results for Dependent Variable months

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	158.12584	14.34	0.0003
Denominator	16	11.02381		

- Investigate the model: `months = size stock`

```
proc reg data=a1;
    model months=size stock;
run;
```

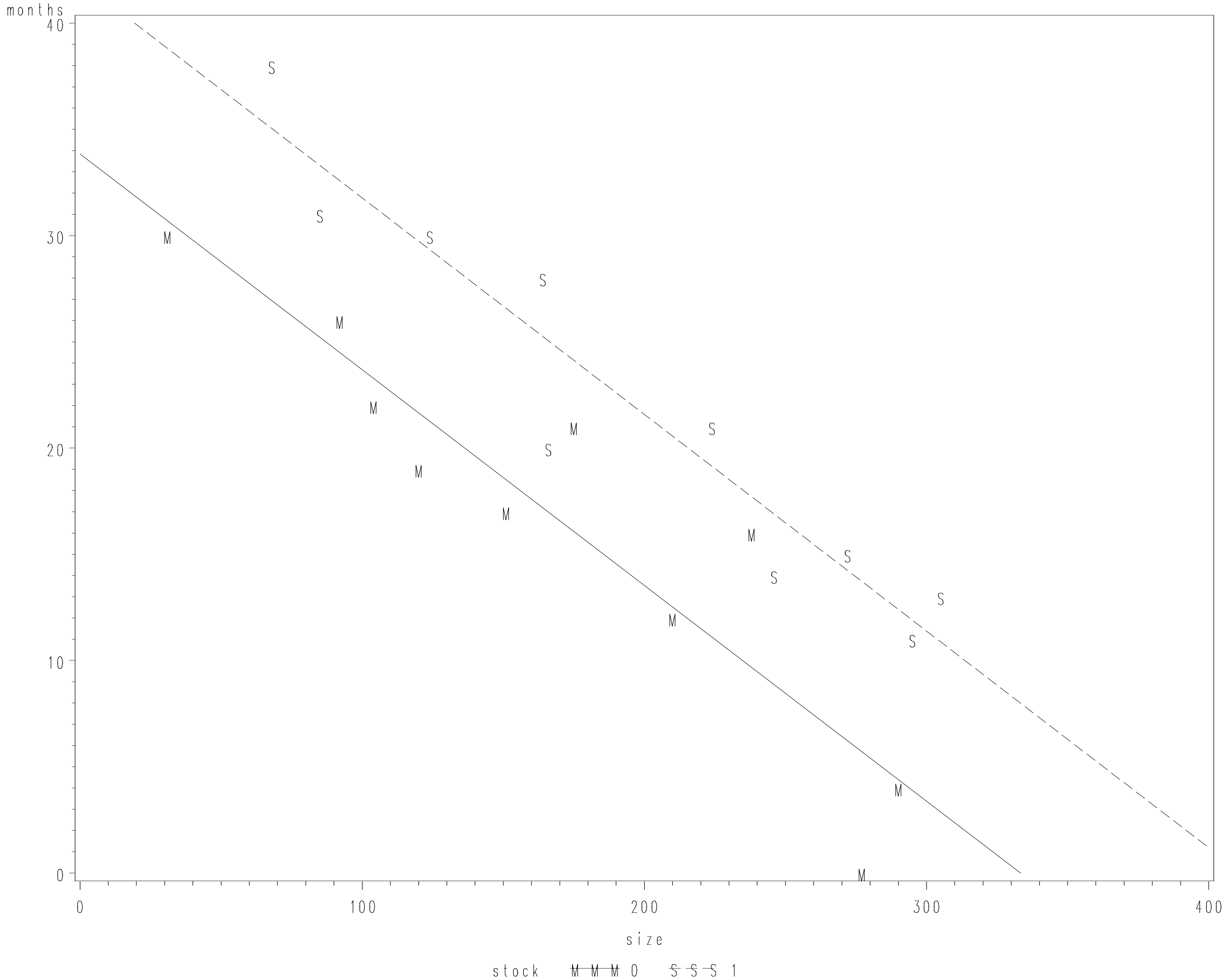
```
                   Analysis of Variance
                         Sum of         Mean
Source              DF      Squares       Square    F Value   Pr > F
Model                2   1504.41333   752.20667      72.50   <.0001
Error               17    176.38667    10.37569
Corrected Total     19   1680.80000

Root MSE               3.22113    R-Square     0.8951
Dependent Mean        19.40000    Adj R-Sq     0.8827
Coeff Var             16.60377

                  Parameter Estimates
                 Parameter   Standard
Variable     DF   Estimate      Error  t Value     Pr > |t|
Intercept     1   33.87407    1.81386    18.68       <.0001
size          1   -0.10174    0.00889   -11.44       <.0001
stock         1    8.05547    1.45911     5.52       <.0001
```

- No apparent interaction here. The slope of the line does not depend on the type of firm.

```
/* Constant Slope */
symbol1 v=M i=rl c=black;       symbol2 v=S i=rl c=black;
proc gplot data=a1;
    plot months*size=stock/frame;
run; quit;
```



Interaction Models: Special Case #2

- X_1 and X_2 are continuous variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Can be written

$$Y_i = \beta_0 + (\beta_1 + \beta_3 X_{i2}) X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + (\beta_2 + \beta_3 X_{i1}) X_{i2} + \varepsilon_i$$

- The coefficient of one explanatory variable depends on the value of the other explanatory variable

- Cannot discuss each predictor individually

- Use contour plot for visual interpretation:

Contour plot in SAS

```
/* Contour plot for Y=10+X1-3*X2+5X1*X2 */
data a1;
   do x1=0 to 1 by 0.1;
      do x2=0 to 1 by 0.1;
         y=10+1*x1-3x2+5x1*x2;
         output;
      end;
   end;
proc gcontour data=a1;
    plot x1*x2=y;
run;
```

Qualitative Predictors

Coding a Variable with Two Classes

- The "Insurance Innovation" example: $X_1 =$ size of firm
- Either stock firm or mutual fund firm in the study $\Longrightarrow$ a qualitative predictor with two classes,

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ , if mutual fund} \\ 0 \text{ , otherwise} \end{cases}$$

  – Cannot include both indicators in the model

  – The model below contains perfectly collinear columns

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

  – SAS will drop the last column

- The additive model below is appropriate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  – interpretation:

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \end{aligned}$$

Alternative Coding of a Two-Class Variable

- An alternative coding

$$X_2 = \begin{cases} 1 & \text{, if stock firm} \\ -1 & \text{, if mutual fund} \end{cases}$$

- The additive model is still appropriate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  – interpretation:

$$\begin{array}{ll} E\{Y_i\} = (\beta_0 + \beta_2) + \beta_1 X_{i1} & \qquad \text{if stock firm} \\ E\{Y_i\} = (\beta_0 - \beta_2) + \beta_1 X_{i1} & \qquad \text{if mutual fund} \end{array}$$

  – β_0 is an "average" intercept of regression line

  – will use this coding for Analysis of Variance

Coding a Variable with Three Classes

- First option: extend the indicator

$$X_2 = \begin{cases} 0 \text{ , if mutual fund} \\ 1 \text{ , if stock firm} \\ 2 \text{ , if foreign firm} \end{cases}$$

- The model below is still appropriate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - interpretation: enforces an equal change in $E\{Y\}$ for each extra indicator

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \\ E\{Y_i\} &= (\beta_0 + 2\beta_2) + \beta_1 X_{i1} && \text{if foreign firm} \end{aligned}$$

  - may be too restrictive

Alternative Coding of a Three-Class Variable

- Second option:

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ , if foreign firm} \\ 0 \text{ , otherwise} \end{cases}$$

- The model below contains two indicators

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

  - interpretation:

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \\ E\{Y_i\} &= (\beta_0 + \beta_3) + \beta_1 X_{i1} && \text{if foreign firm} \end{aligned}$$

  - also more flexibility in presence of interactions X_1X_2 and X_1X_3

Constrained Regression

- At times may want to put constraints on regression coefficients
  - $\beta_1 = 5$
  - $\beta_1 = \beta_2$
- Can do this in SAS by redefining explanatory variables
  - Page 268, redefine so model is $\tilde{Y}$ vs X_2
- Can also use `RESTRICT` statement in `PROC REG`

```
PROC REG DATA=test;
    MODEL Y=X1 X2 X3;
RUN; QUIT;
```

  - Restrict β_1=5: `RESTRICT X1=5;`
  - Restrict $\beta_1 = \beta_2$: `RESTRICT X1=X2;`

Chapter Review

- Polynomial Regression
- Interaction Models
- Qualitative Predictors
- Constrained Regression