Chapter 8

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

8.1 Polynomial regression

- Used when the relationship between Y and the predictor(s) is curvilinear.
- **Example**: we might add a quadratic term to a simple linear model to get a parabolic mean

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \epsilon_i.$$

- We can no longer interpret β_1 and β_{11} "as usual." Cannot hold x_1 constant and increase x_1^2 by one unit, or vice-versa!
- Adding higher order terms in PROC REG is a pain; it needs to be done in the DATA step. For PROC GLM, you can specify a model such as `model outcome=age chol age*age age*chol;` directly.

Higher degree polynomials

- The degree of a polynomial is the largest power the predictor is raised to. The previous model is a 2nd degree polynomial giving a quadratic-shaped mean function.
- Here is a third-order (cubic) in one predictor:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11} x_{i1}^2 + \beta_{111} x_{i1}^3 + \epsilon_i.$$

- A polynomial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$ can have up to $k-1$ “turning points” or extrema.
- (p. 296) A $(k-1)$th-order polynomial can go through $(x_1, Y_1), \ldots, (x_k, Y_k)$ *exactly*!

General notes on fitting

- (p. 295) Predictors can be first centered by subtracting off the sample mean from each predictor, i.e. $x_{ij}^* = x_{ij} - \bar{x}_j$ is used as a predictor instead of x_{ij} where $\bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$. This reduces multicollinearity among, for example, x_{i1}, x_{i1}^2, x_{i1}^3, etc. This isn't always necessary.
- Polynomials of degree 4 (quartic) and higher should rarely be used; cubic and lower is okay. High-degree polynomials have unwieldy behavior and can provide extremely poor out of sample prediction. Extrapolation particularly dangerous (p. 294).
- A better option is to fit an "additive model" (discussed later); the degrees of freedom on the smoothers can mimic third, fourth degree polynomials while being better behaved.

Polynomial regression: more than one predictor

In the case of multiple predictors with quadratic terms, cross-product terms should also be included, at least initially.

Example: Quadratic regression, two predictors:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{1st order}} + \underbrace{\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2}}_{\text{2nd order}} + \epsilon_i .$$

- This is an example of a *response surface*, or parabolic surface (Chapter 30!)
- "Hierarchical model building," (p. 299) stipulates that a model containing a particular term should also contain all terms of lower order including the cross-product terms.
- Degree of cross-product term is obtained by summing power for each predictor. e.g. the degree of $\beta_{1123} x_{i1}^2 x_{i2} x_{i3}$ is $2 + 1 + 1 = 4$.

Hierarchical model building

"When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate...With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model. Thus, one would not drop the quadratic term of a predictor variable but retain the cubic term in the model. Since the quadratic term is of lower order, it is viewed as providing more basic information about the shape of the response function; the cubic term is of higher order and is viewed as providing refinements in the specification of the shape of the response function." – *Applied Statistical Linear Models* by Neter, Kutner, Nachtsheim, and Wasserman.

"It is not usually sensible to consider a model with interaction but not the main effects that make up the interaction." – *Categorical Data Analysis* by Agresti.

"Consider the relationship between the terms $\beta_1 x$ and $\beta_2 x^2$. To fit the term $\beta_0 + \beta_2 x^2$ without including $\beta_1 x$ implies that the maximum (or minimum) of the response occurs at $x = 0$...ordinarily there is no reason to suppose that the turning point of the response is at a specified point in the x-scale, so that the fitting of $\beta_2 x^2$ without the linear term is usually unhelpful.
A further example, involving more than one covariate, concerns the relation between a cross-term such as $\beta_{12} x_1 x_2$ and the corresponding linear terms $\beta_1 x_1$ and $\beta_2 x_2$. To include the former in a model formula without the latter two is equivalent to assuming the point $(0, 0)$ is a col or saddle-point of the response surface. Again, there is usually no reason to postulate such a property for the origin, so that the linear terms must be included with the cross-term." – *Generalized Linear Models* by McCullagh and Nelder.

Polynomial model as an approximation to unknown surface

Real surface given by

$$y_i = f(x_{i1}, x_{i2}) + \epsilon_i.$$

First order approximation to $f(x_1, x_2)$ about some $(\bar{x}_1, \bar{x}_2)$ is

$$\begin{aligned}
f(x_1, x_2) &= f(\bar{x}_1, \bar{x}_2) + \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1}(x_1 - \bar{x}_1) + \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1}(x_2 - \bar{x}_2) \\
&\quad + \text{HOT}. \\
&= \left[f(\bar{x}_1, \bar{x}_2) - \bar{x}_1 \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1} - \bar{x}_2 \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_2} \right] \\
&\quad + \left[\frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1} \right] x_1 + \left[\frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_2} \right] x_2 + \text{HOT} \\
&= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{HOT}
\end{aligned}$$

$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is an approximation to unknown, infinite-dimensional $f(x_1, x_2)$ characterized by $(\beta_0, \beta_1, \beta_2)$.

2nd order Taylor's approximation

Now let $\mathbf{x} = (x_1, x_2)$ and

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})' D^2 f(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})' + \text{HOT}.$$

This similarly reduces to

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \text{HOT},$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ correspond to various (unknown) partial derivatives of $f(x_1, x_2)$. Depending on the shape of the true (unknown) $f(x_1, x_2)$, some or many of the terms in the approximation $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ may be unnecessary.

We work backwards via F tests *hierarchically* getting rid of HOT first to get at more general trends/shapes, e.g. the first order approximation.

When to include polynomial terms

- If the response vs. a predictor is curved in the initial scatterplot, this relationship *may or may not hold* when other predictors are added! It's better to examine residuals versus each predictor to see if, e.g. adding a quadratic term, might be useful.
- Added variable plots are a refined plot to help figure out if the "non-linear" pattern is there when other variables are added (Section 10.1)
- With lots of predictors, say $k \geq 5$, it is easier to pare down to important main effects first, look for possible pairwise interactions (if necessary), and then see if any of the residual plots look curved; if so, toss in a quadratic term.

When to include polynomial terms

- Sometimes people fit a higher-order model and then start "paring away" higher order terms with t and F-tests to get a simpler, more interpretable model. This is called *backwards elimination* (Chapter 9).
- The example on pp. 300–305 starts with a full quadratic function, then pares away higher order terms, finally leaving only the main effects as important.

8.2 Pairwise interactions among predictors

Recall that Taylor's theorem includes *cross product terms*.

An interaction model includes one or several *cross-product* terms.

Example: two predictors

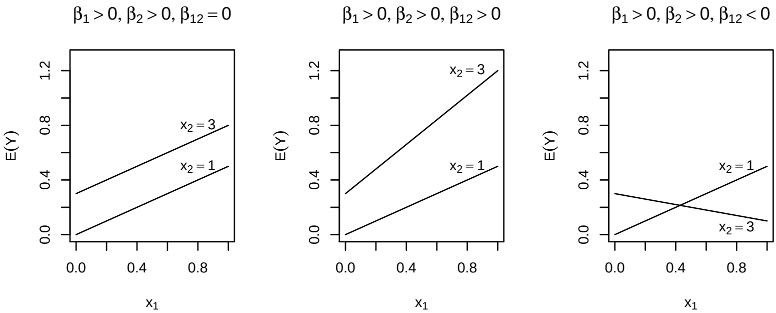
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

How does the mean change when we increase x_1 by unity?

$$\begin{array}{rcl} \text{at } x_1 & \Rightarrow & E(Y) = \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \\ \text{at } x_1 + 1 & \Rightarrow & E(Y) = \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_{12} (x_1 + 1) x_2 \\ \textbf{difference} & = & \beta_1 + \beta_{12} x_2 \end{array}$$

How the mean changes depends on the other variable.
An additive model is like a sheet of paper held "flat." A pairwise interaction is like twisting the two ends of the paper. Plots can show what's happening (pp. 310–311)

Interactions



Interactions

- Parallel lines indicate reflect no interaction between x_1 and x_2; non-parallel lines indicate an interaction.
- Including all pairwise (or higher) interactions complicates things tremendously.
- Need to pare them out via t-tests and/or F-tests.
- Book suggests fitting additive model, then looking at residuals e_i versus each two-way interaction; if there's a pattern you could include that interaction in the model.
- In my personal experience, scientists will often have "an idea" of which variables might interact, i.e. there's already some intuition there on their part. This can be helpful.
- Can also find pool of "good" main effects, then add interactions one at a time (forward selection!) using `proc glm`.

8.3 Categorical predictors

Let's say we wish to include variable "cat," a categorical variable that takes on values cat $\in \{1, 2, \ldots, I\}$. We need to allow each level of cat $= x$ to affect $E(Y)$ differently. This is accomplished by the use of dummy variables.

In PROC GLM, categorical variables are defined through the `CLASS CAT;` statement and all dummy variables are created and handled internally. PROC REG doesn't do this.

Type 3 tests are used to see whether an entire categorical predictor can be dropped from the model (all of the dummy variables at once).

Creating zero-one dummies

Define $z_1, z_2, \ldots, z_{I-1}$ as follows:

$$z_j = \left\{ \begin{array}{cc} 1 & \text{cat} = j \\ 0 & X \neq j \end{array} \right\}$$

This sets class $\text{cat} = I$ as baseline (baseline is last alpha-numeric level). Say $I = 3$, then the model is

$$E(Y) = \beta_0 + \beta_1 z_1 + \beta_2 z_2.$$

which gives

$$\begin{array}{lll} E(Y) = \beta_0 + \beta_1 & \text{when} & \text{cat} = 1 \\ E(Y) = \beta_0 + \beta_2 & \text{when} & \text{cat} = 2 \\ E(Y) = \beta_0 & \text{when} & \text{cat} = 3 \end{array}$$

β_1 and β_2 are *offsets to baseline* mean.

Collapsing levels

- Sometimes a researcher is interested in whether levels can be pooled for a categorical predictor.
- The table of regression coefficients only provide offsets to baseline, and so only allow to to test whether the *baseline* can be collapsed with the $(I-1)$ other levels.
- Adding a `contrast` statement to PROC GLM will allow us to test whether we can collapse levels. For example, for $I=3$ we can add `contrast ‘collapse 2 and 3’ CAT 0 1 -1;`
- If we instead create dummy variables (d1 and d2) by hand in the data step, in PROC REG we can use `test d1-d2=0;`

Interaction between two categorical variables

A two-way interaction is defined by multiplying the variables together; if one or both variables are categorical then all possible pairings of dummy variables are considered.

Example: Say we have two categorical predictors, $X = 1, 2, 3$ and $Z = 1, 2, 3, 4$. An additive model is

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 I\{X = 1\} + \beta_2 I\{X = 2\} \\ &\quad + \beta_3 I\{Z = 1\} + \beta_4 I\{Z = 2\} + \beta_5 I\{Z = 3\}. \end{aligned}$$

The model that includes an interaction between X and Z adds $(3 - 1)(4 - 1) = 6$ additional dummy variables accounting for all possible ways, i.e. all levels of Z, the mean can change from $X = i$ to $X = j$. The new model is rather cumbersome:

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 I\{X = 1\} + \beta_2 I\{X = 2\} \\ &\quad + \beta_3 I\{Z = 1\} + \beta_4 I\{Z = 2\} + \beta_5 I\{Z = 3\} \\ &\quad + \beta_6 I\{X = 1\} I\{Z = 1\} + \beta_7 I\{X = 1\} I\{Z = 2\} \\ &\quad + \beta_8 I\{X = 1\} I\{Z = 3\} + \beta_9 I\{X = 2\} I\{Z = 1\} \\ &\quad + \beta_{10} I\{X = 2\} I\{Z = 2\} + \beta_{11} I\{X = 2\} I\{Z = 3\}. \end{aligned}$$

8.5 Categorical & quantitative interaction

Consider the following model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

where x_1 is size of firm and $x_2 = 1$ if stock firm and $x_{i2} = 0$ otherwise (one continuous and one categorical with two levels). Then

$$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 \text{ stock firm,}$$
$$E(Y) = \beta_0 + \beta_1 x_1 \text{ otherwise,}$$

Have different intercepts and different slopes. If we instead fit an additive model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

then

$$E(Y) = (\beta_0 + \beta_2) + \beta_1 x_1 \text{ stock firm,}$$
$$E(Y) = \beta_0 + \beta_1 x_1 \text{ otherwise.}$$

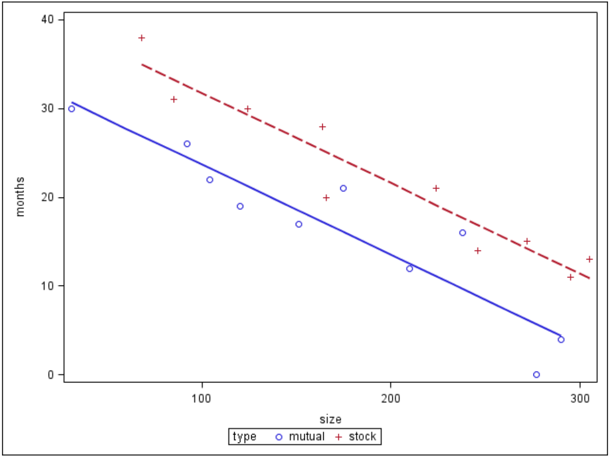
These are two *parallel* lines; the slope is the same. β_2 is how much better (or worse) stock firms do *at any firm size* x_1.

Insurance innovation

An economist studied 10 mutual firms and 10 stock firms to relate speed Y (months) with which an insurance innovation is adopted to size (total assets) of insurance firm x_1 and type x_2, stock or mutual.

```
data insurance;
input months size typedum type$ @@;  sizetype=size*typedum;
datalines;
  28  164  1 stock  15   272  1 stock   11  295  1 stock   38   68  1 stock
  31   85  1 stock  21   224  1 stock   20  166  1 stock   13  305  1 stock
  30  124  1 stock  14   246  1 stock   17  151  0 mutual  26   92  0 mutual
  21  175  0 mutual 30    31  0 mutual  22  104  0 mutual   0  277  0 mutual
  12  210  0 mutual 19   120  0 mutual   4  290  0 mutual  16  238  0 mutual;
proc sgscatter; plot months*size / reg group=type; run;
```

Insurance innovation



Do these look parallel?

Insurance innovation

```
proc reg data=insurance;
  model months=size typedum; run;

proc glm data=insurance order=data; class type;
  model months=size type / solution; run;
-----------------------------------------------------------------------------------
The REG Procedure
                        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
typedum       1        8.05547        1.45911       5.52      <.0001
-----------------------------------------------------------------------------------
The GLM Procedure

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
size                         1     1358.613335     1358.613335     130.94    <.0001
type                         1      316.245973      316.245973      30.48    <.0001

                                           Standard
Parameter                Estimate             Error    t Value    Pr > |t|
Intercept             33.87406904 B      1.81385830      18.68      <.0001
size                  -0.10174212        0.00889122     -11.44      <.0001
type      stock        8.05546921 B      1.45910570       5.52      <.0001
type      mutual       0.00000000 B       .                 .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the
normal equations.  Terms whose estimates are followed by the letter 'B' are not uniquely estimable.
```

Insurance innovation

In class we will

- Look at standard ODS diagnostic plots from both PROC REG and PROC GLM.
- Test whether a quadratic in firm size is necessary.
- Test whether an interaction between firm size and type is necessary.
- Interpret the model.

Two more examples from this chapter...

```
data power; * pp. 300-305 QUESTION: is first order model okay?
input cycles rate temp @@;
datalines;
  150  0.6  10   86  1.0  10   49  1.4  10  288  0.6  20  157  1.0  20
  131  1.0  20  184  1.0  20  109  1.4  20  279  0.6  30  235  1.0  30
  224  1.4  30
;

data soap; * pp. 330-333 QUESTION: are the lines the same?;
input scrap speed line @@;
datalines;
  218  100  1  248  125  1  360  220  1  351  205  1  470  300  1  394  255  1
  332  225  1  321  175  1  410  270  1  260  170  1  241  155  1  331  190  1
  275  140  1  425  290  1  367  265  1  140  105  0  277  215  0  384  270  0
  341  255  0  215  175  0  180  135  0  260  200  0  361  275  0  252  155  0
  422  320  0  273  190  0  410  295  0
;
```