Paper 207-2019

Polynomial Regression for Modeling Curvilinear Data
A Biological Example

Elisha Johnston, Palos Verdes Peninsula High School

ABSTRACT

This paper provides an introduction to polynomial regression, which is useful for analyzing curvilinear data. I illustrate the power of the procedure with biological data simulated from my science fair research. I demonstrate an easy-to-implement approach for overcoming an issue common in polynomial regression, namely multicollinearity. I discuss a common criticism of polynomial regression (overfitting) and suggest as an alternative greater exploration of the interplay between substantive theory and statistical modeling. I write this paper as an introduction to the basics of polynomial regression for SAS programmers who are comfortable with multiple regression.

INTRODUCTION

After providing a conceptual overview of polynomial regression, I show how to fit a first, second, and third order polynomial regression models in SAS® to simulated dose-response data. For model selection, I discuss 2 general approaches. The first, more traditional in nature, emphasizes domain knowledge (substantive theory, in this case biological dose-response theory). The other, popularized by machine learning, emphasizes algorithmic decision-making. In discussing the distinctive perspectives of the two approaches, I seek to highlight the benefit of greater dialogue about the substantive and empirical insights as an alternative to discussions framed around overfitting.

POLYNOMIAL REGRESSION: THE BASICS

Cook and Weisberg (1999 p. 149) define polynomial regression with a single predictor, x, as

$$E(y|x) = \beta_0 + \beta_1 x \ + \beta_1 x^2 \ + \ \dots \ + \beta_d x^d$$

where E is the expected value of y at a given level of x, d is a positive integer which represents the degree of the polynomial, β_0 is the intercept and β_i is the coefficient of the polynomial term with the i^{th} power (i = 1,2,…,d). In stating that a polynomial of degree d has d+1 terms, Cook and Weisberg (1999 p. 149) indicate that polynomial regression is hierarchical. Those interested in the wider scope of two predictor polynomial regression may find helpful Cook and Weisberg (1999 p. 151-152). Croxford (p. 4) deals explicitly with a known limitation of polynomial regression, namely modeling extrema.

One challenge analysts often face in conducting polynomial regression is minimizing multicollinearity. Multicollinearity occurs when the independent variables are strongly linearly related (Fox 1991, p. 10). Analysts may use pairwise correlation coefficients as a simple diagnostic tool to surface multicollinearity (Kutner et al 2005, p. 289). A more powerful tool to assess for multicollinearity is the variance inflation factor (VIF), which quantifies the extent to which an independent variable is predictable by the other independent variables in a model (Fox 1991, p. 11). Kutner (et al 2005, p. 408) presents the general formula for VIF and an adaptation for polynomial regression is:

$$VIF_k = (1 - R_k^2)^{-1}$$

where VIF_k is the VIF for the k^{th} polynomial term, k = 1, 2, …, d (where d is the number of polynomial terms in the regression model), and R_k^2 is the coefficient of determination (R^2) for a regression of the k^{th} polynomial term on the other polynomial terms. R^2 is defined as:

$$R^2 = 1 - \frac{SSE}{SSTO}$$

where SSE is the residual sum of squares and SSTO is total sum of squares (Kutner et al 2004, p. 354).

A statistical rule of thumb that analysts use to identify a variable as multicollinear with others in a model is VIF≥10 (Kutner et al 2004, pp. 408-410; Myers 1990, p. 369). Consider the implication of the VIF formula for a model with independent variables *q*, *r*, and *s*, where the variable *q* has VIF = 10. Treating *q* as the dependent variable and regressing *q* on *r* and *s* would yield a model R^2 = 0.90 (indicating that the dependent variable *q* would have a high level of linear association with the two independent variables *r* and *s*).

One method of reducing multicollinearity is to mean center the independent variable x and fit a polynomial regression model on the mean centered variable. Mean centering provides the added benefit of bringing the intercept into the range of observed values. In this case, the intercept would become the mean of y centered to x = $\bar{x}$.

A criticism analysts sometimes confront in conducting polynomial regression is that they are overfitting a model to the data. Overfitting is less a precisely defined technical term and more of a general idea. Babyak (2004, p. 411) explains overfitting as overly responsive to the idiosyncratic characteristics of the data at hand. When an analyst selects a model that is overresponsive to the data, they run the risk that their model will depart from substantive theoretical concerns and not generalize to other settings (Mavrevski et al 2018).

Good statistical practice involves using a variety of tools to select a model (Kutner et al 2004, p. 349-375). One approach is based on information criteria, which connects information theory to random variable distributions (Christensen 2018, p. 1). The Bayesian Information Criterion (BIC) computes relative ranks of candidate models based on their distance from a "true model" (Christensen 2018, p. 2). The BIC is helpful in selecting a good model that does not overfit (Kutner et al 2004, p. 359-360); the algorithm does so by implementing a penalty for each added variable. This approach is consistent with the principle of Occam's razor, which suggests to use statistical models with fewer parameters. Shumway and Stoffer (2017, p. 50) defines BIC for regression models as

$$\log \hat{\sigma}_k^2 + \frac{k \log(n)}{n},$$

where log $\hat{\sigma}_k^2$ is the log maximum likelihood estimate of the variance of the error term, n is the sample size, and k is the number of regression coefficients. A lower BIC indicates a model is closer to the "true model" and thus preferable.

Statistical packages commonly produce the adjusted R^2 (R_{adj}^2), which is another model selection tool (Kutner et al 2004, p. 355-356). Defined above as part of VIF, R^2 represents the fraction of variation in the y-values that is explained by the independent variables (Ngo 2012 p. 2). For the purpose of model selection, statisticians have enhanced R^2 to R_{adj}^2 which includes a penalty for adding variables (consistent with Occam's razor). R_{adj}^2 is defined as

$$R_{adj}^2 = 1 - \left(\frac{n-1}{n-k}\right)\frac{SSE}{SSTO},$$

where k is the number of regression coefficients, n is the number of observations, SSE is the residual sum of squares, and SSTO is total sum of squares (Kutner et al 2004, p. 355-356). The coefficient of determination (R^2), almost always varies from 0 to 1, with higher values indicating better models. The adjusted coefficient of determination (R_{adj}^2) has a similar range.

Some analysts use conceptual tools in their model selection process. One such conceptual tool is based on the well-known rule of thumb that a statistical model should have at least 10 to 15 observations per predictor (Babyak 2004, p. 411). In some cases, such as observational data with skewed and uncommon distributions, this rule of thumb would underestimate sample size for reliable statistical inference. In the case of balanced experimental data with a normally distributed dependent variable, 10 observations per predictor is reasonable. For the rest of this paper, I will refer to this as the sample size tool.

Weisberg (1985, p. 210) argues the single most important tool for selecting a model is the analyst's domain knowledge, which is also known as substantive theory or in this paper simply theory. To investigate the dose-response relationship, biology-related scientists employ a variety of models. Perhaps the most famous is "characteristic dose-response" (Trevan 1927), which involves applying multiple doses to estimate a dose that is lethal to 50% of the exposed population (Lethal Dose 50, abbreviated as LD_{50}). Wu (et al 2018) describes a more recent approach, namely model-based drug development. Here, scientists integrate clinical pharmacology, pharmacometric modeling, and human system biology to predict clinical responses in the form of a hyperbolic function. Wu (et al 2018) points out that a few biological scientists have utilized non-monotonic models such as the Quadratic but that the biological community is cautious in accepting these models because of their complexity. Motulsky and Christopoulos (2003, p. 41) argue: "Don't fit a complicated...model if a simpler model fits the data fine" (see also Friedman 2010 and Mavrevski et al 2018). As I discuss further in the next section, my science fair research investigates a therapy for which the scientific community considers that low doses will have little impact on outcome, moderate doses will lead to positive outcomes, and high doses will lead to negative outcomes. When considering the full range of low, moderate and high doses, one of the simplest models is Quadratic.

A BIOLOGICAL EXAMPLE: SIMULATED DATA

I introduce polynomial regression with a common type of biological data, namely *in vitro* data (also known as cell culture data). In 2017, I began investigating the hypothesis that when physicians inject hypertonic dextrose into arthritic joints, that hypertonic dextrose stimulates a patient's immune system to regenerate damaged cartilage (Topol 2017) and thereby slow the progression of osteoarthritis (OA). For short, I will refer to this therapy as hypertonic dextrose injections. Dr. Joseph Freeman, an associate professor of Biomedical Engineering at Rutgers University, pioneered the *in vitro* research design that scientists are now using to investigate hypertonic dextrose injections (Freeman et al 2011, Ekwueme et al 2018).

A prototypical *in vitro* research design for investigating hypertonic dextrose injections involves seeding 10,000 cells per well, treating cells for 24 hours, and then measuring cells at 31 hours after treatment initiation. To create a simulated data set, I start with patterns in my previously presented *in vitro* research (Johnston et al 2017; Johnston 2017; Johnston 2018; Johnston 2019), accentuating that part of the pattern that surfaces the tension between a second or third order model. I omit empirical complexities such as noise due to pipetting anomalies. Also, to improve understandability, I change the outcome variable from an indirect measure of cell count (absorbance via spectrophotometry) to a direct count of cells (such as would arise from flow cytometry). The experimental goal is to identify which dose of hypertonic dextrose leads to the greatest amount of cell proliferation. I name the custom created data set OATherapy_Trial1. The outcome variable is number of cartilage cells (CELLS). The predictor variable is the treatment dose of hypertonic dextrose (TRT).

I use a SAS input statement (Delwiche and Slaughter 2003, p. 39) to create the data set. The code below is that snippet for inputting the first observation of the data set (see **Appendix A** for the full data set, **Appendix B** for a data dictionary, Cody 2017, p. 2):

INPUTTING DATA

```
DATA OATherapy_Trial1;
 INPUT Cells Trt;
 CARDS;
12000 5
<44 more observations>
;
run;
```

To explore polynomial regression models, I create polynomial variables ranging from degree 2 to degree 3, in both a raw and centered form, via PROC SQL (Lafler 2005 p. 2, Lafler 2007 p. 2), with the computed columns being Trt_C, Trt2, Trt2_C, Trt3, and Trt3_C:

VARIABLE CREATION

```
PROC SQL;
 CREATE TABLE OATherapy_Trial1_Poly AS
 SELECT
 Cells
   , Trt
   , (Trt-Mean(Trt)) AS Trt_C
   , Trt*Trt AS Trt2
   , (Trt-Mean(Trt))*(Trt-Mean(Trt)) AS Trt2_C
   , Trt*Trt*Trt AS Trt3
   , (Trt-Mean(Trt))*(Trt-Mean(Trt))*(Trt-Mean(Trt)) AS Trt3_C
  FROM OATherapy_Trial1;
 quit;
```

DATA PROFILING

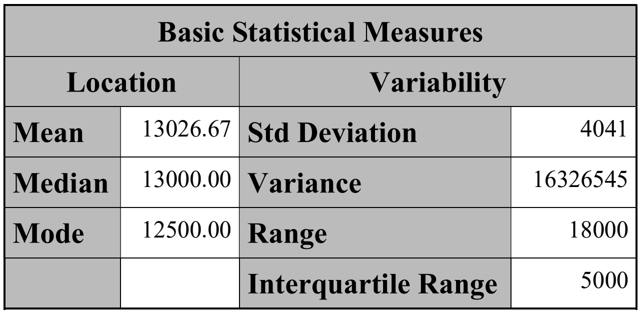
I begin profiling by the distribution of the dependent variable and investigate for potential problems such as outliers and missing data:

PROC UNIVARIATE ON DEPENDENT VARIABLE (CELLS)

```
PROC UNIVARIATE data = OATherapy_Trial1_Poly;
  var Cells;
  histogram Cells / normal kernel;
  QQPLOT Cells;
  run;
```

The “Basic Statistical Measures” output (Table 1) indicates that across the 45 records, the average number of cells (13,027) is very close to the median (13,000). The interquartile range is 5,000 cells and the standard deviation is 4,041 cells.

SAS OUTPUT from PROC UNIVARIATE



Basic Statistical Measures			
Location		**Variability**	
Mean	13026.67	**Std Deviation**	4041
Median	13000.00	**Variance**	16326545
Mode	12500.00	**Range**	18000
		Interquartile Range	5000

Table 1. BASIC STATISTICAL MEASURES on DEPENDENT VARIABLE (CELLS)

The “Distribution of Cells” output (Figure 1) shows the variable has a roughly normal distribution (skewed slightly left). Polynomial regression does not assume that the outcome variable has a normal distribution but such a distribution may lead to normally distributed residuals (which is a model assumption).

OUTPUT from PROC UNIVARIATE

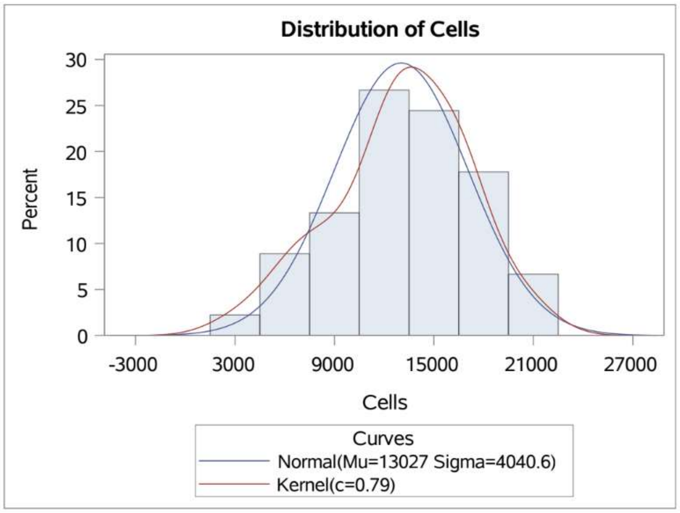


Figure 1. DISTRIBUTION of DEPENDENT VARIABLE (CELLS)

By including the Q-Q Statement in the PROC UNIVARIATE code, I utilize SAS to create a normal quantile-quantile plot that compares the distribution of the dependent variable CELLS to a theoretically normally distributed variable with the same mean and standard deviation as CELLS. Figure 2 shows that CELLS is approximately normally distributed because the data points are falling on a relatively straight line. Further, the plot shows that there are no outliers. The Missing Value Section of PROC UNIVARIATE did not display, indicating no missing values.

SAS OUTPUT from PROC UNIVARIATE

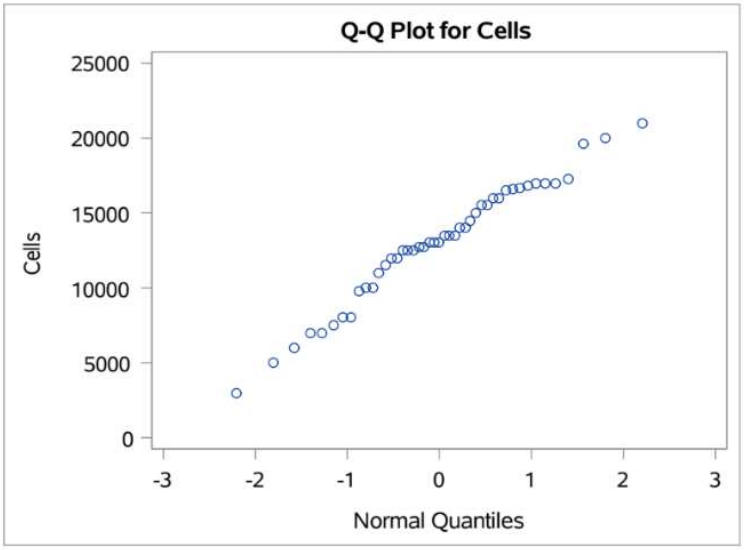


Figure 2. CHECKING the DISTRIBUTION of the DEPENDENT VARIABLE (CELLS)

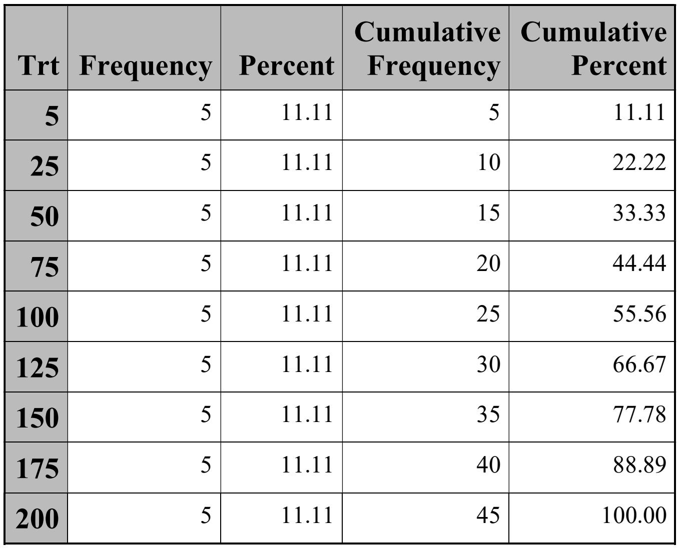
I continue profiling by examining the independent variable, confirming no missing values and that the dosing schedule is plausible with PROC FREQ for the independent variable Trt (Cody 2017 p. 4):

PROC FREQ for TREATMENT

```
PROC FREQ data = OATherapy_Trial1_Poly;
  TABLES Trt / missing;
  run;
```

The independent variable TREATMENT has nine arms, each with the same number of values (five) and no missing values (Table 2).

SAS OUTPUT from PROC FREQ



Trt	Frequency	Percent	Cumulative Frequency	Cumulative Percent
5	5	11.11	5	11.11
25	5	11.11	10	22.22
50	5	11.11	15	33.33
75	5	11.11	20	44.44
100	5	11.11	25	55.56
125	5	11.11	30	66.67
150	5	11.11	35	77.78
175	5	11.11	40	88.89
200	5	11.11	45	100.00

Table 2. ASSESING for MISSING in the INDEPENDENT VARIABLE (TREATMENT)

Next, I carry out the same data profiling with the remaining variables (Trt2, Trt3, Trt_C, Trt2_C, and Trt3_C):

PROC FREQ for Multiple Variables

```
PROC FREQ data= OATherapy_Trial1_Poly;
  TABLES Trt2 Trt3 Trt_C Trt2_C Trt3_C / missing;
run;
```

I find that SAS successfully created the variables: no missing values, each treatment level has 5 replicates and the observed values conform to a polynomial structure (output not shown).

Next, I use PROC SGSCATTER to look at the Correlation Matrix for the non-centered and centered variables:

PROC SGSCATTER to Produce Correlation Matrices

```
ODS RTF FILE='/folders/myfolders/CORRMATRIX';
title "Scatter Plot Matrix";
PROC SGSCATTER DATA=OATherapy_Trial1_Poly;
      MATRIX Trt Trt2 Trt3 /
      diagonal=(histogram kernel) ellipse;
run;
ODS RTF CLOSE;
ODS RTF FILE='/folders/myfolders/CORRMATRIX';
title "Scatter Plot Matrix";
PROC SGSCATTER DATA=OATherapy_Trial1_Poly;
      MATRIX Trt_C Trt2_C Trt3_C /
      diagonal=(histogram kernel) ellipse;
run;
ODS RTF CLOSE;
```

Pairwise relationships between the non-centered variables (Trt, Trt2 and Trt3) are highly correlated as indicated by the narrow ellipses in Figure 3 (scatterplot matrices on the left). Pairwise relations between centered variables (Trt_C, Trt2_C, and Trt3_C) are less correlated as indicated by the rounder ellipses in Figure 3 (scatterplot matrices on the right). Centering reduces, but does not completely eliminate, correlation – the centered variables Trt_C and Trt3_C still appear to have a strong linear association. Below, I use the more comprehensive VIF too to formally assess if this correlation calls into question the regression model estimates.

SAS OUTPUT from PROC SGSCATTER

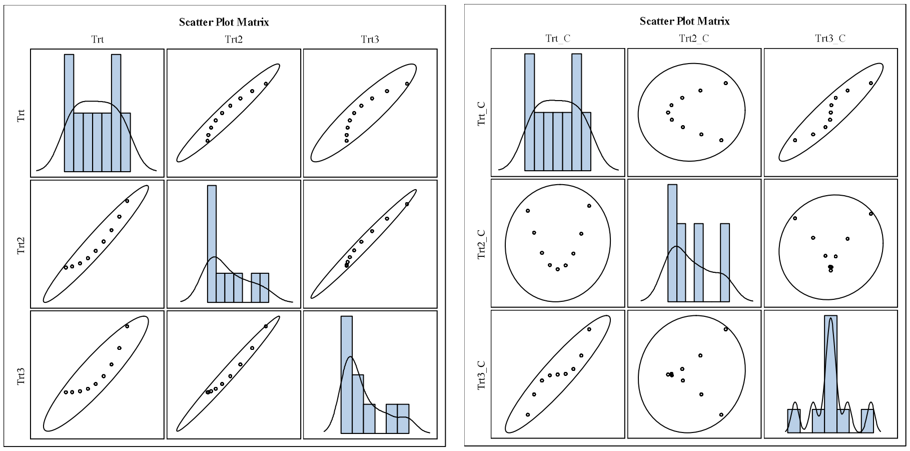


Figure 3. CORRELATION MATRICES on INDEPENDENT VARIABLES

For my final step of profiling, I visually examine the relationship between the independent and dependent variable with PROC SGPLOT:

PROC SGPLOT: TREATMENT vs. CELLS

```
PROC SGPLOT DATA = OATherapy_Trial1_Poly;
  scatter x = Trt y = Cells;
  run;
```

The scatter plot hints the data has a curvilinear form (Figure 4).

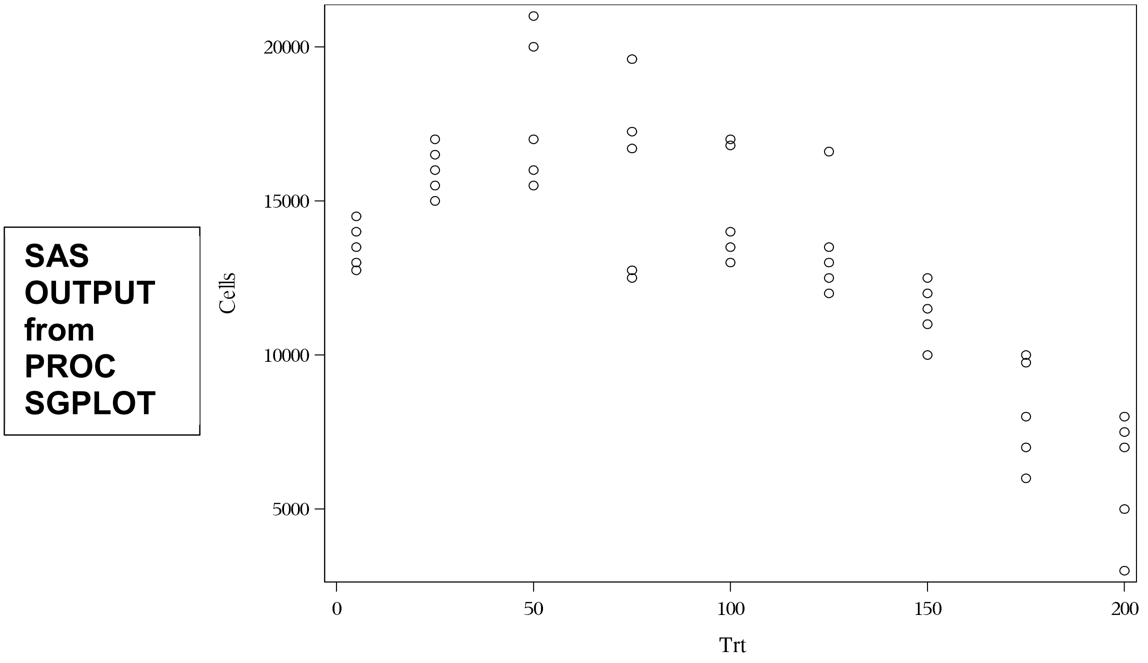


Figure 4. PROC SGPLOT SCATTER PLOT of TRT vs. CELLS

EXAMPLE 1: FIRST ORDER POLYNOMIAL REGRESSION

The following SAS code produces a first order polynomial regression model:

PROC REG for FIRST ORDER MODEL

```
PROC REG DATA = OATherapy_Trial1_Poly
    Plots = diagnostics (stats = (default bic));
  MODEL Cells = Trt;
  run;
```

The first degree polynomial regression model characterizes dose as having a negative linear relationship with cell number (Figure 5, top). The residual vs. dose plot clearly indicates that the first order model does not fit the data well because the residuals exhibit a clear systematic trend (Figure 5, bottom).

SAS
OUTPUT
from
PROC
REG

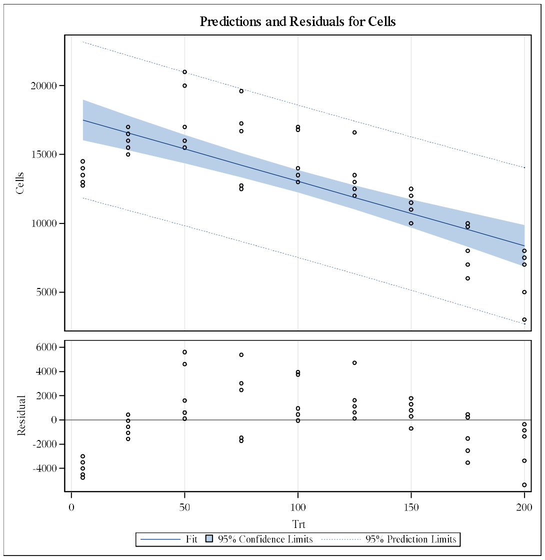
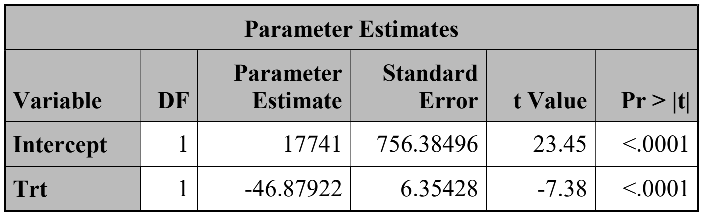


Figure 5. FIRST ORDER POLYNOMIAL REGRESSION and RESIDUAL PLOT

The “Parameter Estimates” table shows the regression equation’s intercept is estimated as 17,741 cells (Table 3), indicating that with no glucose at all, the cells would grow from 10,000 cells (at seeding) to 17,741 cells (31 hours after treatment initiation). The slope is -46.9, indicating that each 1 mM increase in dose is associated with a decrease in proliferation of 47 cells. Scaling to a larger, more meaningful increment, the model indicates that an increase in treatment dose of 25 mM would be associated with an increase in proliferation of 1,175 cells. According to this model, the effective dose is 0 mM. The output also provides a table with information on the parameters’ estimate, degrees of freedom, standard error, t value, and statistical significance. The above SAS code provides a separate output window that I summarize here in text: R^2_{adj} = 0.55 and BIC = 715.7.

SAS
OUTPUT
from
PROC
REG



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17741	756.38496	23.45	<.0001
Trt	1	-46.87922	6.35428	-7.38	<.0001

Table 3. FIRST ORDER POLYNOMIAL REGRESSION PARAMETER ESTIMATES

The quantile-quantile residual plot shows that the residuals form a relatively straight line, indicating that these residuals have a distribution close to normal (Figure 6).

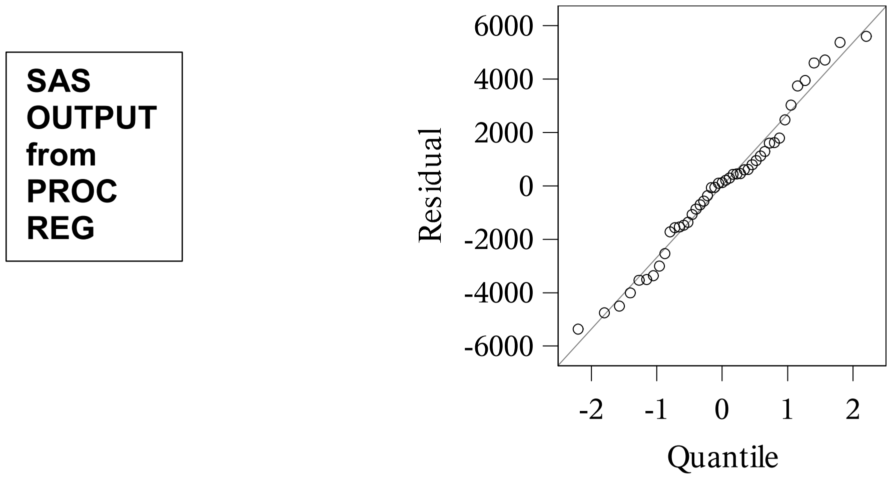


Figure 6. FIRST ORDER POLYNOMIAL REGRESSION RESIDUAL Q-Q PLOT

EXAMPLE 2: SECOND ORDER POLYNOMIAL REGRESSION

The SAS code produces a second order polynomial regression model:

PROC REG for SECOND ORDER MODEL

```
  PROC REG DATA = OATherapy_Trial1_Poly plots =
predictions (X = Trt)
    MODEL Cells = Trt Trt2 / vif;
    run;
```

The above SAS code provides a separate output window that I summarize here in text: BIC = 686.0 and R^2_{adj} = 0.77.

Visually, the second degree polynomial model indicates that at the lowest dose of hypertonic dextrose, proliferation is relatively low (Figure 7). As the dose increases to moderate amounts, the proliferation increases. However, as the dose further increases, the proliferation decreases (due to therapy-induced cell death). The residual vs. dose plot indicates that the model may not fit the data at the lowest dose level because the residual at the lowest dose level are below the zero horizontal line.

SAS OUTPUT from PROC REG

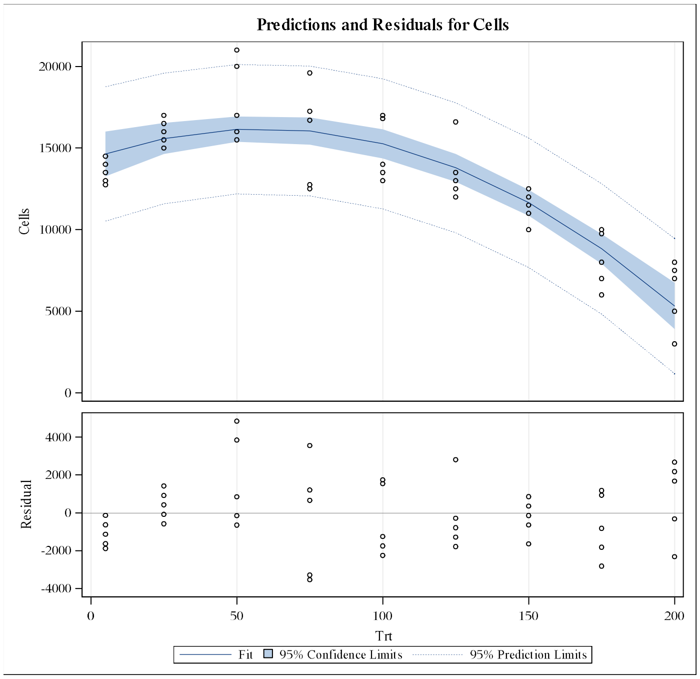
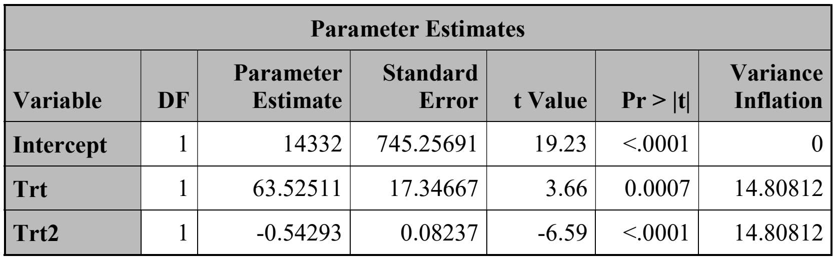


Figure 7. SECOND ORDER POLYNOMIAL REGRESSION and RESIDUAL PLOT

The “Parameter Estimates” output (Table 4) indicates that the Trt and Trt2 variables are both significant, characterizing dose as having a quadratic relationship with cell number. The intercept is 14,332, suggesting that with no treatment, the cells would grow from 10,000 cells (at seeding) to 14,332 cells (31 hours after treatment initiation). The slope varies by treatment level (positive from 0 to 59 mM and then subsequently negative). The maximum value of the quadratic is 59 mM, which represents the effective dose. Both TRT and TRT2 have a VIF which equals 14.8, indicating multicollinearity.

SAS OUTPUT from PROC REG



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	14332	745.25691	19.23	<.0001	0
Trt	1	63.52511	17.34667	3.66	0.0007	14.80812
Trt2	1	-0.54293	0.08237	-6.59	<.0001	14.80812

Table 4. PARAMETER ESTIMATES for SECOND ORDER POLYNOMIAL REGRESSION

The quantile-quantile residual plot (Figure 8) shows that the residuals form a relatively straight line, indicating the data are relatively normally distributed.

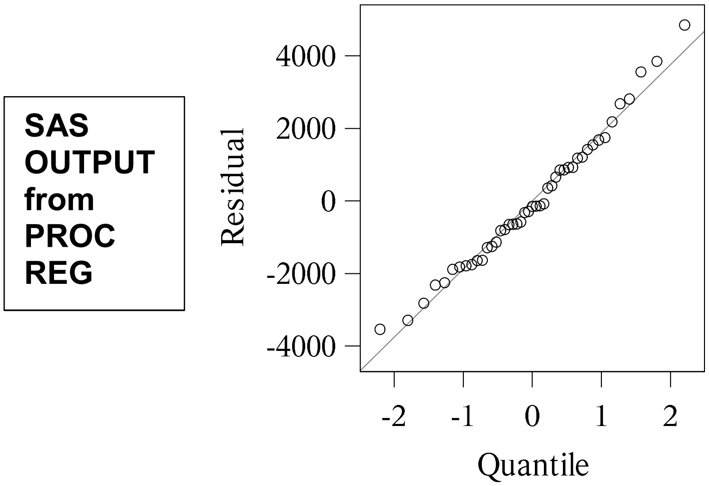


Figure 8. SECOND ORDER POLYNOMIAL REGRESSION RESIDUAL Q-Q PLOT

To address the multicollinearity, I rerun the second order model with the centered variables and find that the VIF for both the TRT_C and TRT2_C variables is 1.00167. All predictors remain significant. For the sake of easy understanding and brevity, I continue to use the native metric (instead of centered).

The second order polynomial regression model is clearly superior to the first order polynomial regression model. Visually, the second degree polynomial model fits the data much better than the first order polynomial model. A comparison of the second order residual vs. dose plot (Figure 7, bottom) to the same first order plot (Figure 5, bottom) indicates the second order model fits the data better than the first order model. Consequently, the second order model has a much higher R^2_{adj} (0.77 compared to 0.55). The BIC also indicates the second order model fits the data better than the first order model (686 compared to 716). Further, the second order model is more consistent with dose-response theory than the first order model in that the second order polynomial model identifies a meaningful effective dose.

EXAMPLE 3: THIRD ORDER POLYNOMIAL REGRESSION

The following SAS code produces a third order polynomial regression model:

PROC REG for THIRD ORDER MODEL

```
PROC REG DATA = OATherapy_Trial1_Poly
  plots = predictions (X = Trt)
  MODEL Cells = Trt Trt2 Trt3/ vif;
  run;
```

At first glance, the prediction graph (Figure 9, top) suggests the third degree polynomial model does not fit the data much better than the second order model. Similar to the second order model, the third order model indicates that at the lowest dose of hypertonic dextrose, proliferation is relatively low. As the dose increases to moderate amounts, the proliferation increases. However, as the dose further increases, the proliferation decreases. However, a closer look at the residual vs. dose plot (Figure 9, bottom) suggests the third order model may fit the data better than the second order model because the residuals at all dose levels are randomly scattered around the zero horizontal line.

SAS OUTPUT from PROC REG

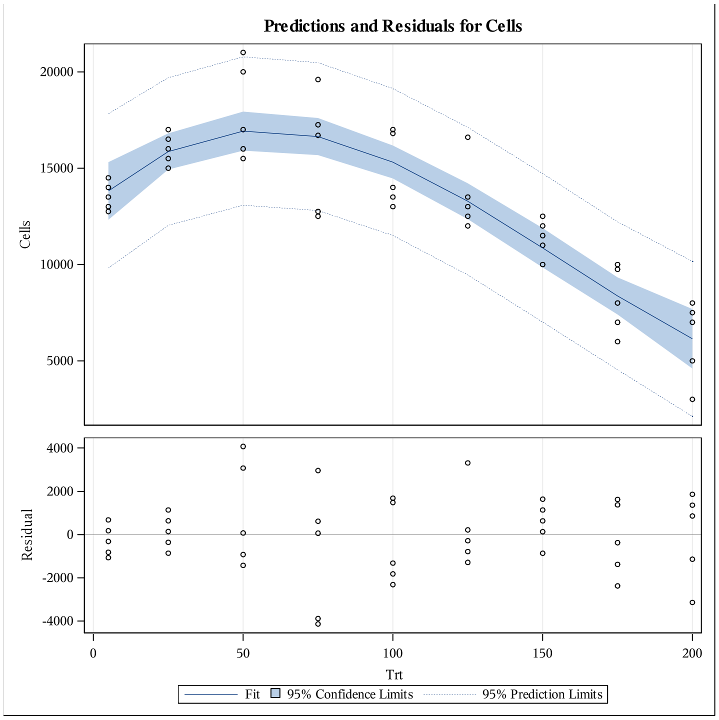


Figure 9. THIRD ORDER POLYNOMIAL REGRESSION and RESIDUAL PLOT

The "Parameter Estimates" output (Table 6) indicates that the Trt, Trt2, and Trt3 variables are all significant. This type of polynomial model is known as cubic. The intercept is 13,125, suggesting that with no treatment, the cells would grow from 10,000 cells (at seeding) to 13,125 cells (31 hours after treatment initiation). The slope varies by treatment level (positive from 0 to 56 mM and then subsequently negative). The maximum value of the quadratic is 56 mM, which represents the effective dose. TRT, TRT2, and TRT3 all have a VIF greater than 10, indicating multicollinearity.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	13125	891.00055	14.73	<.0001	0
Trt	1	147.24887	40.70820	3.62	0.0008	89.45236
Trt2	1	-1.59539	0.47403	-3.37	0.0017	537.88308
Trt3	1	0.00342	0.00152	2.25	0.0298	217.10246

Table 5. PROC REG: PARAMETER ESTIMATES for THIRD ORDER POLYNOMIAL REGRESSION

The quantile-quantile residual plot shows that the residuals form a relatively straight line, indicating the data are relatively normally distributed (Figure 10).

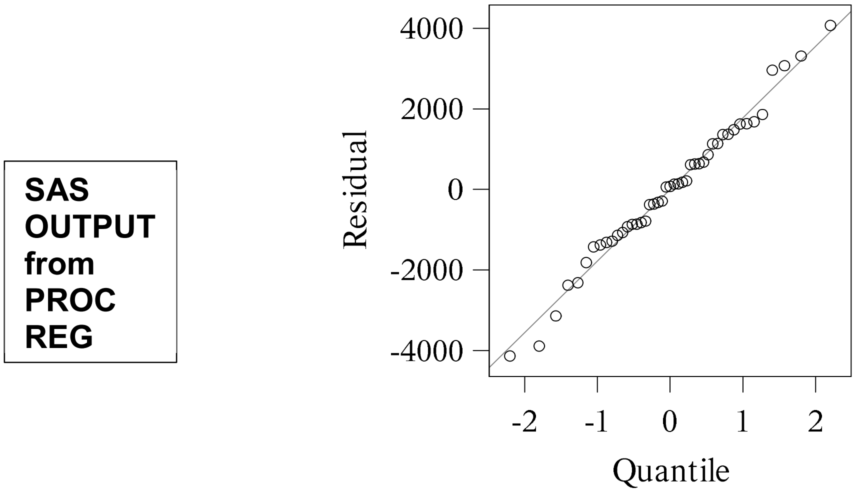


Figure 10. THIRD ORDER POLYNOMIAL REGRESSION RESIDUAL Q-Q PLOT

To address the multicollinearity, I rerun the third order model with the centered variables, finding the VIF for TRT_C, TRT2_C, and TRT3_C to be less than 10. All predictors remain significant.

COMPARING THE SECOND AND THIRD ORDER MODELS

The statistical community holds (at least) two positions on model selection. As discussed above, one group prioritizes domain (substantive) knowledge, even though that knowledge may not be well established. Another group prefers to base model selection decisions on concrete, objective, standardized, and transparent statistical algorithms, like those practicing machine learning. In some instances, the two positions harmoniously lead to a similar model choice. In other situations, such as the biological example presented in this paper, the two positions may constructively and productively interact on the interplay between substantive theory and statistical models.

Statisticians prioritizing substantive theory as the most important criterion for model selection would likely select the second order model because a quadratic dose-response curve is consistent with dose-response theory. Further, they would likely argue that the quadratic model fits the data almost as well as the cubic model. The biological argument would be that hypertonic dextrose from 0 to 100 mM acts on cells in a consistent way to produce a symmetric distribution of proliferation. With regards to the cubic term, these statisticians would likely wonder: what is the biological rationale for arguing that the treatment has a different impact at the lowest dose level, which is the control?

Statisticians prioritizing algorithmic decision-making would likely select the third order model because the third order model performs better on a number of metrics. In terms of fit statistics, the third order BIC is lower (683 compared to 686, Table 6) and the R_{adj}^2 is higher (0.79 compared to 0.77). These statisticians would likely also argue that the third order model regression diagnostics are better in that this model fits the data at all dose levels, especially the left tail (citing Figure 9 vs. Figure 7).

Polynomial Model	Predicted Effective Dose	BIC	R_{adj}^2	Theory	Sample Size
First	0 mM	716	.55	-	+
Second	59 mM	686	.77	++	+
Third	56 mM	683	.79	+	+

Table 6. SUMMARY of MODEL SELECTION FINDINGS

Constructive dialogue between these 2 groups of statisticians would begin with the recognition that both models identify an effective dose in the range of 56-59 mM hypertonic dextrose. Statisticians prioritizing domain knowledge would likely argue that the incremental gain in BIC and R_{adj}^2 is very small and that the steeper left tail could be due to chance variation specific to the control dose that comprises the left tail extrema. Statisticians prioritizing algorithmic decision-making would likely argue that the empirical asymmetry possibly surfaces a biological process meriting further theoretical consideration. They would likely further argue that including a cubic term leads to a lower effective dose (56 mM vs 59 mM), which is in line with larger medical opinion favoring conservative dosing. Such dialogue illustrates the way in which polynomial regression may stimulate constructive and meaningful dialogue about the intersection between substantive biological theory and statistical modeling that might otherwise remain unnoticed and undiscussed with traditional regression.

CONCLUSION

This paper shows the ability of polynomial regression to surface empirical puzzles that those favoring substantive theory and statistical modeling will approach from different perspectives. In this paper, I simulated dose-response data, corrected for multi-collinearity, and presented a concrete dilemma. The dilemma is what should an analyst do when a statistically significant cubic term provides only a small, incremental improvement in R_{adj}^2 and BIC? Scientists prioritizing substantive theory would likely argue to omit the cubic term, while those favoring algorithmic decision-making would likely argue to include the cubic term. Constructive dialogue between scientists from these groups may enable fruitful collaboration that advances both biological theory and mathematical (and statistical) modeling (Friedman 2010).

REFERENCES

Babyak M. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. Psychosomatic Medicine 66:411-421.

Christensen W. (2018). Model Selection Using Information Criteria (Made Easy in SAS®), SAS Global Forum 2018, Conference Proceedings.

Cody R. (2017). Cody's Data Cleaning Techniques Using SAS. SAS Institute, Inc. Cary, NC.
Cook R.D. and Weisberg S. (1999). Applied Regression Including Computing and Graphics. John Wiley & Sons, New York.

Croxford R. (2017). Continuous Predictors in Regression Analyses, SAS Global Forum 2017.

Delwiche L.D. and Slaughter S.J. (2003). The Little SAS Book. SAS Institute Inc. Cary, NC.

Fox J. (1991). Regression Diagnostics. Sage Publications, Newbury Park, CA.

Friedman A. (2010). What Is Mathematical Biology and How Useful Is It? American Mathematical Society.

Horstman J.M. (2019). Getting Started with the SGPLOT Procedure, SAS Global Forum 2019.

Jablonski K. and Guagliardo M. (2016). Data Analysis Plans: A Blueprint for Success Using SAS. SAS Institute Inc. Cary, NC.

Johnston ED, Andrali SS, Kochan A, Johnston M, Lovick J (2017). Prolotherapy-Induced Cartilage Regeneration: Investigating Cellular-Level Mechanisms of Action with Mouse Preosteoblast Cells. JSM Biochem Mol Biol 4(3): 1032.

Johnston ED (2017). The Molecular Mechanisms of Regenerating Cartilage to Reduce Chronic Pain: Phenol-Glucose-Glycerin Upregulates FGF-2. Abstract of a Project Presented at the California State Science and Engineering Fair. Los Angeles, CA.

Johnston ED (2018). Assessing the Biochemistry of Dextrose-Based Cartilage Regeneration. Abstract of a Project Presented at the Palos Verdes Peninsula Science and Engineering Fair. Rolling Hills Estates, CA.

Johnston ED (2019). Transforming in vitro Studies of Hypertonic Dextrose Injections for Osteoarthritis: A Wide Range Investigation of Effective Dose with a Physiologically Relevant Model. Abstract of a Project Presented at the International Science and Engineering Fair. Phoenix, AZ.

Kutner, Nachtsheim, Neter, and Li. (2004). Applied Linear Statistical Models. 5th edition. McGraw-Hill. Madison, WI.

Lund B. (2016). Finding and Evaluating Multiple Candidate Models for Logistic Regression, SAS Global Forum 2016, Conference Proceedings.

Mavrevski R., Traykov M., Trenchev I., Trencheva M. (2018). Approaches to modeling of biological experimental data with Graphpad prism software. WSEAS Transactions on Systems and Control.

McGahan CE. (2009). Building Multiple Linear Regression Models – Food for Thought, PNWSUG 2009, Conference Proceedings, Pacific Northwest SAS Users Group, Inc.

Myers R. (1990). Classical and Modern Regression with Applications 2nd Edition. Duxbury Press (Thomson Learning); Pacific Grove, CA.

Shumway R, and Stoffer D. (2011). Time Series Analysis and Its Applications 3rd Edition. Springer. New York City, NY.

Trevan J.W. (1927). The error of determination of toxicity. *Proc R. Soc. Lond. B 1927, 101, 483-514.*

Wu J., Banerjee A, Jin B., Menon S. , Martin S., Heatherington A.C. (2018). Clinical dose-response for a broad set of biological products: A model-based meta-analysis. *Stat Methods Med Res. 2018 Sep;27(9):2694-2721.*

ACKNOWLEDGMENTS

I express a big thank you to many people. Kirk Lafler for serving as my mentor. Kidane Ghebrehawariat, PhD (Biostatistics) and Michael Johnston, PhD (Quantitative Sociology) for providing intensive coaching on statistics and writing. Elizabeth Li and Ron Fehd for reading an early draft of my manuscript. Susan Slaughter and Matthew Slaughter for reading multiple drafts of my manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Elisha Johnston
Johnston_elisha_dani@student.smc.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX A

```
DATA Example_Parabolic;
 INPUT Trt Cells;
 CARDS;
 5 12000
 5 11500
 5 12500
 5 13000
 5 11000
 25 17000
 25 16500
 25 17500
 25 18000
 25 16000
 50 25000
 50 25500
 50 24500
 50 26000
 50 24000
 100 15000
 100 15500
 100 14500
 100 16000
 100 14000
 200 5000
 200 5500
 200 4500
 200 6000
 200 4000
 ;
 run;
```

APPENDIX B

Name	Definition	Data Type	Unit
Cells	Number of Cartilage Cells	Numeric	Cells
Trt	Treatment (Dose of hypertonic dextrose)	Numeric	mM
Trt_C	Trt mean centered	Numeric	mM
Trt^2	Trt squared	Numeric	mM^2
Trt^2_C	Trt_C squared	Numeric	mM^2
Trt^3	Trt cubed	Numeric	mM^3
Trt^3_C	Trt_C cubed	Numeric	mM^3