Applying Polynomial Regression to Dyadic Data

Nehemias Ulloa

Department of Statistics
Iowa State University

April 20, 2017

Outline

- Intro and Motivation
- Model
- Recreate Dr. Phillips' Application
- Our Application

Introduction

Hypotheses:

- Family researchers comparing the attitudes, behaviors, and opinions of pairs
- Collect Dyadic data
  - Inter-individual reporting
  - Intra-individual reporting

Motivation

- Difference scores used to analyze dyadic data
- Difference scores allow to see how well they "fit" together
- Common types: algebraic, absolute, and squared difference

$$Z = \beta_0 + \beta_1(X - Y) + \epsilon \tag{1}$$
$$Z = \beta_0 + \beta_1|X - Y| + \epsilon \tag{2}$$
$$Z = \beta_0 + \beta_1(X - Y)^2 + \epsilon \tag{3}$$

Motivation

- Many methodological issues with Difference Scores
  1. Difficult to identify the underlying mechanism
  2. Problems with underlying assumptions

Motivation

- Any alternatives?
- Polynomial Regression
  e.g.

$$Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY + \epsilon \tag{4}$$

- Take the Squared Difference and expand it

$$(X - Y)^2 = X^2 + Y^2 - 2XY \tag{5}$$

- Expand on the ideas from Simple Linear Regression

The Model

- Theoretical model:

$$Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY + \epsilon \tag{6}$$

- Fitted model:

$$\hat{z} = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 \tag{7}$$

- Fitted model in matrix notation:

$$\hat{z} = b_0 + \mathbf{d}'\mathbf{b} + \mathbf{d}'\mathbf{B}\mathbf{d} \tag{8}$$

where

$$\mathbf{d} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_3 & b_4/2 \\ b_4/2 & b_5 \end{bmatrix}$$

Model

We can fit a model like this in R using the `lm()` function:

```
# Z - the response variable
# X - the first explanatory variable
# Y - the second explanatory variable
QuadFit <- lm(z ~ x + y + I(x^2) + I(x*y) + I(y^2), data=d)
```

Stationarity Points: How & Why?

- What are they?
  - Points where the slope is zero no matter which direction you take the derivative
- Values of our explanatory variables provide the "best" fit for the response
- How do you derive the stationary points?
  1. Take the derivatives of Equation 7 with respect to x and y
  2. Set the derivatives equal to zero
  3. Solve for x and y in terms of $\mathbf{b}$ and $\mathbf{B}$ to find the stationarity points
  4. Refer to these points as x_0 and y_0

Stationarity Points

$$x_0 = \frac{b_2 b_4 - 2 b_5 b_1}{4 b_5 b_3 - b_4^2} \tag{9}$$

$$y_0 = \frac{b_1 b_4 + 2 b_2 b_3}{4 b_5 b_3 - b_4^2} \tag{10}$$

These points represent the values in our predictors that will optimize our response (minimum or maximum depending on the surface shape).

```
statpts <- function(lm){
  # Stationary Pts.
  x0  <- (b2*b4 - 2*b1*b5)/(4*b3*b5 - b4^2)
  y0  <- (b1*b4 - 2*b2*b3)/(4*b3*b5 - b4^2)
  # Output
  out <- matrix(c(x0, y0), ncol=2)
  colnames(out) <- c("x0","y0")
  out <- data.frame(x0=x0, y0=y0)
```

Predicted Response at Stationarity Points

We then get the maximum (or minimum) predicted reponse ($\hat{z}_0$) by plugging in the Stationary Points to Eq 7:

$$\hat{z}_0 = b_0 + b_1 x_0 + b_2 y_0 + b_3 x_0^2 + b_4 x_0 y_0 + b_5 y_0^2 \tag{11}$$

```
x0 <- statpts(lm)$x0
y0 <- statpts(lm)$y0
predstatpts <- coef(lm)[1] + coef(lm)[2]*x0 + coef(lm)[3]*y0 +
```

Principal Axes

- Measure the amount of "bend" in two directions at the stationary points
- Make interpretations of the model easier by rotating the axes by removing all the cross-product terms (Ex: PCA)

In the Creative Componant,

- Brought together a complete picture of how to derive the Principal Axes.
  - Khuri and Cornell (1996) in *Response Surfaces: Designs and Analyses* lay out some pieces of how derive them but never completely laid out the complete process

Principal Axes

Basic idea in their derivation:

- Derive canonical equations to get surface in what is known as canonical form
- Transform canonical equations back onto original variables
- These are the Principal Axes

$$y = -P_{21}x + P_{20} \tag{12}$$

where $P_{20} = y_0 + P_{21}x_0$ and $P_{21} = \frac{b_5 - b_3 - \sqrt{(b_5 - b_3)^2 + b_4^2}}{b_4}$.
Using similar algebra we can find the second principal axes:

$$y = -P_{11}x + P_{10} \tag{13}$$

where $P_{10} = y_0 + P_{11}x_0$ and $P_{11} = \frac{b_5 - b_3 + \sqrt{(b_5 - b_3)^2 + b_4^2}}{b_4}$.

Principal Axes

Interpretations:

1. In a concave surface:
   - First principal axis (Eq 12): lowest downward curvature where two explanatory variables create a maximized surface that decreases the least
   - Second principal axis (Eq 13): greatest downward curvature where two explanatory variables create the greatest decrease in the response
2. In a convex surface:
   - Flip the interpretations of the two Principal Axes

```
paxis <- function(lm){
...
  # Stationary Pts.
  x0  <- (b2*b4 - 2*b1*b5)/(4*b3*b5 - b4^2)
  y0  <- (b1*b4 - 2*b2*b3)/(4*b3*b5 - b4^2)
  # First Principal axis
  p11 <- (b5 - b3 + sqrt((b3 - b5)^2 + b4^2))/b4 #slope
  p10 <- y0 - p11*x0 #intercept
  # Second Principal axis
  p21 <- (b5 - b3 - sqrt((b3 - b5)^2 + b4^2))/b4
  p20 <- y0 - p21*x0
  # Output
  out <- data.frame(p10=p10, p11=p11, p20=p20, p21=p21)
  return(out)
}
```

Re-creation

An example of this methodology being applied is found in Phillips et al. (2012), *Congruence research in behavioral medicine: methodological review and demonstration of alternative methodology*.
Interested in seeing what created good situations in which patients followed through with their doctors orders
The variables used in this example are:

$Z_{1,i}$ – patient-reported adherence a month after the dr. visit for patient i

$Z_{2,i}$ – physician's percieved agreement with patient i on the illness treatment

X_i – physician's rating of patient i's health

Y_i – physician's estimate of how patient i would rate of their own health

Re-creation

Here in Table 1, some summary statistics are presented of the response and explanatory variables used.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	n
Z_1	-2.48	-0.20	0.30	-0.00	0.53	0.54	225.00	177.00
Z_2	1.00	3.25	4.00	3.88	4.62	5.00	151.00	251.00
X	-2.00	-1.00	0.00	0.28	1.00	2.00	112.00	290.00
Y	-2.00	-1.00	0.00	0.18	1.00	2.00	111.00	291.00

Table: Summary statistics of response variables(Z_1 and Z_2) and explanatry variable X and Y.

Re-creation

Before we get into whether or not we should use polynomial regression, they first check some stastics for obvious non-normailty:

1. High Leverage
2. Influential Points

No points were of major concern but one point consistently had a higher leverage and Cook's D value.

Re-creation

The models being used in the criteria are listed below (note that these may have been listed previously but are relisted here for ease of reading):

$$\text{Null Model: } Z = \beta_0 + \epsilon \tag{14}$$

$$\text{Diff Model: } Z = \beta_0 + \beta_1(X - Y) + \epsilon \tag{15}$$

$$\text{Uncon Diff Model: } Z = \beta_0 + \beta_1 X + \beta_2 Y + \epsilon \tag{16}$$

$$\text{Abs Diff Model: } Z = \beta_0 + \beta_1 |X - Y| + \epsilon \tag{17}$$

$$\text{Uncon Abs Diff Model: } Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 W + \beta_4 WX + \beta_5 WY + \epsilon \tag{18}$$

$$\text{Abs Diff Model: } Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY + \epsilon \tag{19}$$

where W is a dummy variable equal to 0 when $X > Y$, 1 when $X < Y$, and randomly assigned 1 or 0 when $X = Y$, and $\epsilon \sim N(0, \sigma_e^2)$

Re-creation

The four criterion are:

1. Check whether the unconstrained model (Eq 16 and 18) explained a significant amount of the variance in the data (compared to Eq 14).
2. Check whether the regression coefficients for the unconstrained model are in expected directions and whether they are significant to the model.
3. Check whether the unconstrained models (Eq 16 and 18) is significantly better than thier constrained models (Eq 15 and 17).
4. Check whether a higher order model i.e. polynomial regression (Eq 19) is significantly better than the constrained model (Eq 15).

Re-creation

1. First criterion checks the basic linear model assumption i.e. is the model better than a basic mean(intercept) only model (Eq 14)?
2. Second criterion checks whether the coefficients are lined up in the way we expect them too and whether they add anything to the model
3. Thrid criterion checks if there is benefit to using complex polynomial models compared to simpler difference models
4. Fourth criterion checks whether a simple model or polynomial model should be used

Re-creation

Table: This table is a replication of the Table 2 in Phillips et al. (2012) where $*p < 0.05$; $**p < 0.01$; $***p < .001$

Outcome	Constrained Model		Unconstrained Model						F_c	F_h
	$(X-Y)$	R^2	X	Y				R^2		
Patient Adherence	-0.35^{***}	0.1^{***}	-0.33^{**}	0.39^{***}				0.1^{**}	0.84	0.83
Physician perception of patient agreement	-0.05	0.001	0.1	0.27^{**}				0.18^{***}	54.89^{***}	17.39^{**}
	$\lvert X-Y\rvert$	R^2	X	Y	W	WX	WY	R^2		
Patient Adherence	-0.41^{***}	0.1^{***}	-0.46^{**}	0.46^{**}	0.14	0.57^{*}	-0.45	0.16^{**}	2.19	2.18^{*}
Physician perception of patient agreement	-0.32^{**}	0.04^{**}	-0.06	0.44^{**}	0.1	0.51^{**}	-0.57^{*}	0.21^{***}	13.22^{***}	7.45

1. Criterion 1: Unconstrained model for both outcomes explained a significant amount of variation in the data (R^2 of 0.1,0.18 for responses Z_1 and Z_2 respectively).
2. Criterion 2: Only satisfied for the outcome dealing with patient adherence(coefficients of -.33 and .39) as the regression coefficients for physician perceptions (.1 and .27) were not going in opposite directions and were not both significant
3. Criterion 3: Met for the second outcome i.e. the physician perceptions ($F_c = 54.89$) and not for patient adherence ($F_c = 0.84$)
4. Criterion 4: Satisfied for the outcome conerning physician perceptions of patient-agreement ($F_h = 17.39$) and not for patient adherence ($F_h = 0.83$).

Consider using polynomial regression for the outcome dealing with physician perceptions while it looks as if the model using patient adherence as its outcome will be sufficiently explained using an algebraic difference score model.

- Repeat above looking at the absolute difference
- Conclude the absolute-value difference score model was not an adequate fit for either of the outcome variables
- Difference score is an adequate model for the outcome concerning patient adherence
- Polynomial regression would be the best model for the data using the physician perception outcome

Then they test what kind of polynomial regression:

- Quadratic model vs linear model with a p-value of 0.0065404: Quadratic model explains the data better
- Cubic vs quadratic (p-value 0.673799): Cubic does not explain the data significantly better

So the fitted model(Eq 19) via OLS:

$$\hat{z} = 3.82 + 0.13x + 0.18y - 0.15x^2 + 0.43xy - 0.22y^2$$

Stationary points are at $(-2.48, -2)$
The first and second principal axis are Equations 20 and 21, respectively:

$$y = 0.08 + 0.84x \tag{20}$$

$$y = -4.95 - 1.19x \tag{21}$$

Use our functions to get the stationary points and the principal axes in prev slide

```
statpts(lm2Square)

##           x0         y0
## 1 -2.481127 -2.003207

paxis(lm2Square)

##           p10       p11       p20       p21
## 1 0.08433488 0.8413686 -4.952125 -1.18854
```

The surface slice plots in of both principal axes in Figure 1.

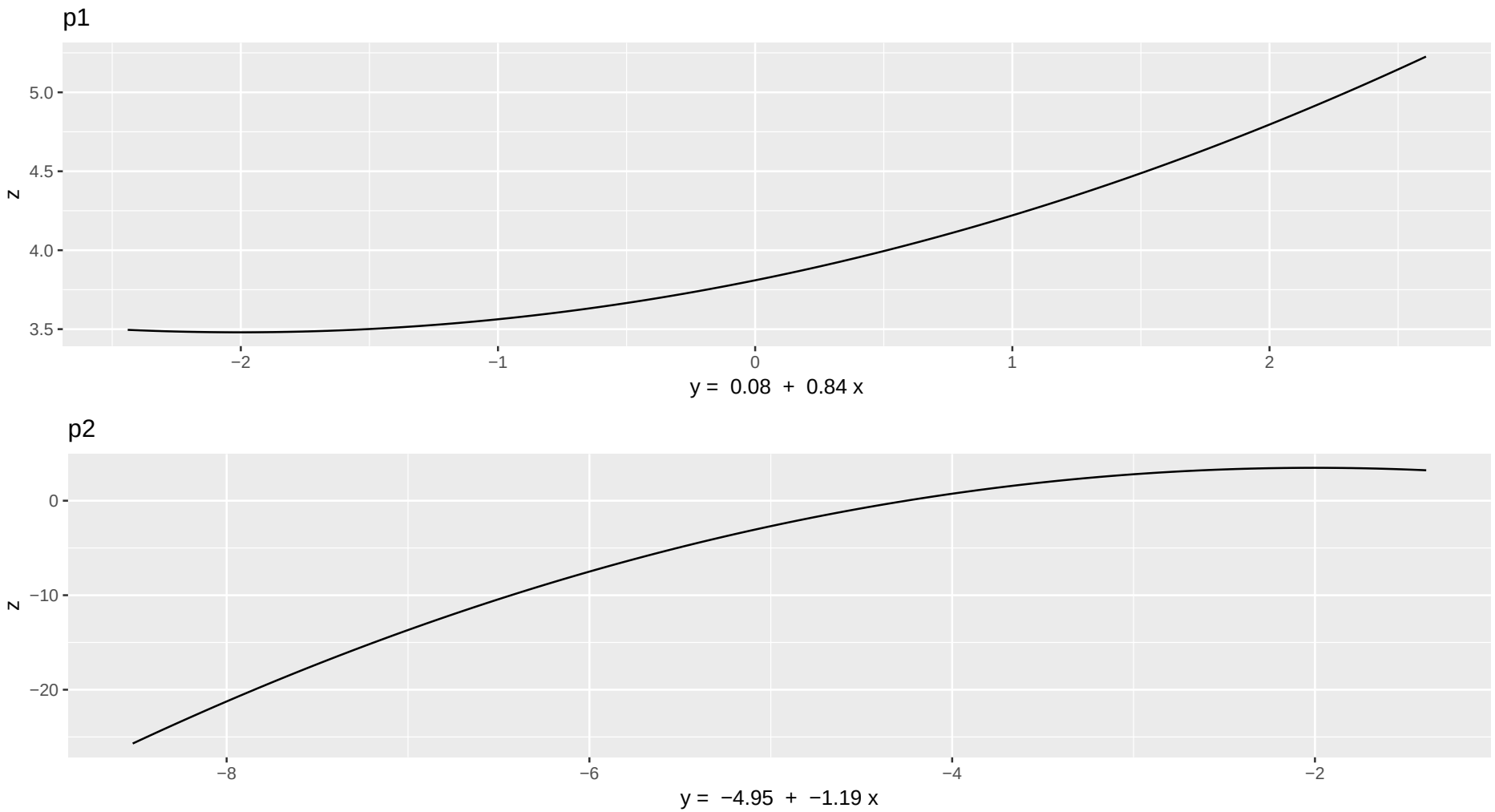

Figure: This Figure shows slices of the response surface taken at the two principal axes.

To look at the predicted surface plot, we use `Plotly`.
We can follow the link to the `Plotly` website and look at and manipulate the predicted surface.

Conclusions

- First principal axes is the axis that makes the most sense: it represnts the path in which the values of patients preferences for shared decision-making and the values of patients experiences of shared decision-making along their entire scales that optimize patient satisfaction
- This makes sense; when doctor and patient agree with what is going on, the patient is most likely to follow the orders and vice-versa
- Second principal axes doesn't have any real interpretations because it lies outside of the scope of the data i.e. it doesn't lie on our predictive surface