Iowa State University

Applying Polynomial Regression to Dyadic Data

Author:

Nehemias Ulloa

Advisor:

Dr. Frederick O. Lorenz

April 15, 2017

1 Introduction

Dyadic data are often used by family researchers who are interested in comparing the attitudes, behaviors, and opinions of husbands and wives, parents and children, and romantic partners towards each other. The dyadic structure can come from inter-individual reporting (e.g. comparison of husband's report of his support toward his wife to the wife's report of husband's support), or it can come from intra-individual reporting (e.g. comparison of an observer's report of husband's hostility to an observer's report of the wife's hostility).

The most common methods in analyzing dyadic data is difference scores. Difference scores are used to see how well two variables or indices 'fit' together. Some of the most common difference scores use the algebraic, absolute, and squared difference between the two variables (X and Y) as shown by Equations 1, 2, and 3 respectively. These difference scores are often used to predict for an outcome variable (Z) as shown in the equations below,

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

$$Z = \beta_0 + \beta_1|X - Y| + \epsilon \tag{2}$$

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

There are many methodological issues associated with difference scores. Edwards (1993) describes many of these issues in his paper *Problems with the Use of Profile Similarity Indices in the Study of Congruence in Organizational Research* [2]. Perhaps the most important reason why statisticians avoid using difference scores is that it makes it difficult to identify the underlying mechanism. This happens because, in using difference scores, two distinct measures are turned into one entity so it becomes difficult to tease out the effect of one variable relative to another. These individual effects are what we interested in. Difference scores fail to tell the researcher which variables contribute the most to the response. We also lose information when using difference scores. There are assumptions used with difference scores namely, that the difference between two entities to the outcome variable are ssumed to be symmetric and that the outcome is constant at all points where the two entities are equal. For instance, consider the line $y = x$, the assumptions with difference scores would assume that as you move accross the line the response remains constant. It is very difficult assumption to hold true as we will see in our example. These

are assumptions that we need to look at by examining the absolute level of both measures and its direction.

In his paper, *On the Use of Polynomial Regression Equations as an Alternative to Difference Scores in Organizational Research*, Dr. Edwards suggested using "polynomial regression equations containing the...[two variables]...composing the difference and higher-order terms."(Edwards 1993)[1] as an alternative to difference scores. An example would be to take the two predictor variables (X, Y), their squares (X^2, Y^2), and their products (XY, X^2Y, XY^2). A successful application of polynomial regression is presented by Phillips, Diefenbach, Kronish, Negron, and Horowitz (2014) in their paper titled *The Necessity-Concerns Framework: a Multidimensional Theory Benefits from Multidimensional Analysis*[5]. They use polynomial regression to assess stroke survivors' concerns about medication and their beliefs effect on their adherence to medication whose goal is stroke prevention.

One potential advantage of polynomial regression is that it provides more information about the underlying process affecting the dyadic data. However, the polynomial models are grounded in more complex linear algebra that has not been overtly developed and displayed, and the results from the polynomial model will be more difficult to interpret. In this paper, our goal is to assess the strengths and limitations of polynomial regression. We will do this by first replicating the results of a published paper, and then we will apply the method to a new dyadic data set that compares wives and husbands impressions of each other's behavior. Before we jump into our examples, we develop the algebra of response surface methodology that polynomial regression uses in order to appreciate and understand the complexity. We will examine and interpet the predictive surface using 3D graphs. Response surfaces depend on indentifying principal axes, as we will in a connonical sense, and indentifying principal axes will require us to discuss stationarity points.

2 The Model

The linear regression model (Eq 4) is familiar to most researchers; it is often used in research due in part to the ease and familiarity of the interpretations of the parameters and the ease of computation. A slight tweak to simple linear models is to include two explanatory variables as in Eq 5. We call Eq 5 the linear difference model; this particular model does not contain an interaction nor a seperate parameter for each varaible. The lack of parameters is what leads to issues regarding interpretability of the difference. It is difficult to tease out which explanatory variable is causing more change to the response when you are only considering the difference of the two variables.

$$Z = \beta_0 + \beta_1 X + \epsilon \tag{4}$$

$$Z = \beta_0 + \beta_1 (X - Y) + \epsilon \tag{5}$$

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

An extension of the linear regression model is the quadratic regression model (Eq 6). It's form is similar to the form of a linear regression model with the addition of a quadratic term of the explanatory variable. It is very similar to the linear model in terms of computation and ease of interpret-ability. In this project, we consider a composition of these two ideas; a polynomial(quadratic) regression using two predictors and the interaction (Eq 6). We will be using the basic framework of response methodology to look two important concepts at stationary points and principal axes. Using the response surface methodology will allow us to take advantage of some key properties developed by that field. So looking how to develop the stationary points and principal axes will be key because they will be very useful in making inference on the predicted surface. In the pages that follow, we demonstrate how this framework has been used and then apply these methods our own data.

2.1 Form

Let us consider a model with 2 explanatory variables and a single response like Eq 6. We can write the model as follows:

$$z = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2 + \epsilon \tag{7}$$

And the fitted model:

$$\hat{z} = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 \tag{8}$$

The fitted model can be expressed using matrix notation as:

$$\hat{z} = b_0 + \mathbf{d}'\mathbf{b} + \mathbf{d}'\mathbf{B}\mathbf{d} \tag{9}$$

where

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

For furthur derivations, we will use matrix notation since it condenses the algebra.

We can fit a model like this in `R` using the `lm()` function. This setup corresponds to Eq 6 where `Z` is the response variable, and `X` and `Y` are the explanatory variables.

```
QuadFit <- lm(Z ~ X + Y + I(X^2) + I(X*Y) + I(Y^2), data=d)
```

2.2 Analyze/Properties

2.2.1 Stationary Points: How and Why?

As we hinted to earlier, polynomial regression draws on ideas from response surface methodology, starting with stationarity points. The stationary points are the points on the X and the Y - axes where the slope of the response (Z) is zero no matter which direction you take the derivative. This corresponds to find the minimum and maximum of a function from calculus; it is a similar concept except now we have a three dimensional surface rather than just a two-dimensional surface. We are interested in the stationarity points because they allow us to see which values of our explanatory variables provide the "best" fit for the response ("best" meaning minimum or maximum). The benefit of knowing the stationary points is it gives us a goal to reach or not reach depending on the situation. (We will see examples of the usefulness of stationary points later.) Using our intuition from calculus, we take the derivatives of Equation 8 with respect to x and y, set the derivatives equal to zero, and solve for x and y in terms of $\mathbf{b}$ and $\mathbf{B}$ to find the stationarity points (we'll refer to these points as x_0 and y_0).

To find the stationarity points, let's begin with taking the partial derivatives with respect to x and y

$$\begin{bmatrix} \frac{dz}{dx} = b_1 + 2b_3x + b_4y \\ \frac{dz}{dy} = b_2 + b_4x + 2b_5y \end{bmatrix} = \mathbf{b} + 2\mathbf{Bd}$$

Next, we set the derivatives to zero and solve for $\mathbf{d_0}$, where $\mathbf{d_0}$ is:

$$\begin{aligned}
\mathbf{d_0} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} &= -\frac{\mathbf{B}^{-1}\mathbf{b}}{2} \\
&= -\frac{1}{2}\frac{1}{b_5 b_3 - \frac{b_4^2}{2}} \begin{bmatrix} b_5 & -b_4/2 \\ -b_4/2 & b_3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\
&= -\frac{2}{4b_5 b_3 - b_4^2} \begin{bmatrix} b_5 b_1 - b_2 b_4/2 \\ -b_1 b_4/2 + b_2 b_3 \end{bmatrix} \\
&= \begin{bmatrix} \frac{b_2 b_4 - 2b_5 b_1}{4b_5 b_3 - b_4^2} \\ \frac{b_1 b_4 + 2b_2 b_3}{4b_5 b_3 - b_4^2} \end{bmatrix}
\end{aligned}$$

So our stationary points are

$$x_0 = \frac{b_2 b_4 - 2b_5 b_1}{4b_5 b_3 - b_4^2} \tag{10}$$

$$y_0 = \frac{b_1 b_4 + 2b_2 b_3}{4b_5 b_3 - b_4^2} \tag{11}$$

These points represent the values in our predictors that will optimize our response. Optimizing the response means either a minimum or maximum depending on the surface shape.

Using our output from the `lm` function noted earlier in `R`, we can get the stationary points. Note we can use other cumputing programs as well as `R` (e.g. `SAS`, `MPlus`, etc.) The `R` code is as follows:

```
statpts <- function(lm){
  # Takes a quadratic lm function
  # coef(lm) gives a vector of all the parameter estimates in the linear model
  # So here we are grabbing the individual parameter estimates from that vector
  b0 <- as.numeric(coef(lm)[1])
  b1 <- as.numeric(coef(lm)[2])
  b2 <- as.numeric(coef(lm)[3])
  b3 <- as.numeric(coef(lm)[4])
  b4 <- as.numeric(coef(lm)[5])
  b5 <- as.numeric(coef(lm)[6])
```

```
  # Stationary Pts. using the formulas in Eq 10 & 11
  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)
  return(out)
}
```

2.2.2 Predicted Response at Stationary Points

Once we have the stationary points, we can get the predicted responses at the stationary points by simply plugging in the stationary points into Equation 8. The predicted response at the stationary points give us the optimized value for the response variable i.e. where the response surface reaches its maximum or minimum. In matrix notation, the predicted response is given as:

$$\begin{aligned}\hat{z}_0 &= b_0 + \mathbf{d_0'b} + \mathbf{d_0'Bd_0} \\ &= b_0 + \mathbf{d_0'b} - \frac{\mathbf{d_0'BB^{-1}b}}{2} \\ &= b_0 + \mathbf{d_0'b} - \frac{\mathbf{d_0'b}}{2} \\ &= b_0 + \frac{\mathbf{d_0'b}}{2}\end{aligned}$$

It is easy to get the predicted response in R:

```
x0 <- statpts(lm)$x0
y0 <- statpts(lm)$y0
predstatpts <- coef(lm)[1] + coef(lm)[2]*x0 + coef(lm)[3]*y0 + coef(lm)[4]*x0^2 + coef(lm)[5]*x0*y0 +
predstatpts
```

2.2.3 Principal Axes

Once stationarity points are established, they can be used to find our second key component, the principal axes of the model. The purpose of principal axes is to measure the amount of "bend" in certain direction at a certain point and make interpretations of the model easier by rotating the axes so that we can remove all the cross-product terms. In response surface methodology, and therefore polynonial regression, we focus on the principal axes in two directions based at the stationary points and the predicted reponse at the stationarity points. Khuri and Cornell (1996), *Response Surfaces: Designs and Analyses* [3], lays out the process of deriving what are commonly known as the canonical equations. The canonical equations come from transforming the response system to what is called its canonical form (this is the name of the form where the surface resides when its cross-product terms are removed); this process occurs because it represents the response centered around the stationarity points.

These canonical equations are then used in finding the principal axes. In order to complete this process, new axes (principal axes) are created which the surface will be on. This maybe useful in some applications, but it is often difficult to relate the response surface on the principal axes to our original variables. Essentially it becomes difficult to interpret how our original variables are affecting the response surface on the principal axes. So we focus on the process of transforming these canonical equations back to functions of the original variables i.e. transform our principal axes back to the original variables. By doing this, we get the benefits of the principal axes and the ease of interpretation using the original variables.

To solve for principal axes, we need to first solve for eigenvalues. These eigenvalues have meaning in the canonical setting(when the surface is algined on the principal axes) but for our purposes, we will not focus on those interpretations. The equations for solving for the eigenvalues comes from Khuri and Cornell (1996) and are as follows:

$$
\begin{aligned}
\Big|\mathbf{B}-\lambda\mathbf{I}\Big|=0 \quad &\Rightarrow \left|\begin{bmatrix} b_3 & b_4/2 \\ b_4/2 & b_5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = 0 \\
&\Rightarrow \left|\begin{bmatrix} b_3-\lambda & b_4/2 \\ b_4/2 & b_5-\lambda \end{bmatrix}\right| = 0 \\
&\Rightarrow det\begin{pmatrix} b_3-\lambda & b_4/2 \\ b_4/2 & b_5-\lambda \end{pmatrix} = 0 \\
&\Rightarrow (b_2-\lambda)(b_5-\lambda) - \Big(\frac{b_4}{2}\Big)\Big(\frac{b_4}{2}\Big) = 0 \\
&\Rightarrow b_3b_5 - b_3\lambda - b_5\lambda + \lambda^2 - \frac{b_4^2}{4} = 0 \\
&\Rightarrow \lambda^2 - (b_3+b_5)\lambda + (b_3b_5 - \frac{b_4^2}{2}) = 0 \\
&\Rightarrow \lambda = \frac{b_3+b_5 \pm \sqrt{\Big(b_3+b_5\Big)^2 - 4\Big(b_3b_5 - \frac{b_4^2}{4}\Big)}}{2}
\end{aligned}
$$

Now let λ_1 be the first eigenvalue:

$$
\begin{aligned}
\lambda_1 &= \frac{(b_3+b_5) + \sqrt{\Big(b_3+b_5\Big)^2 - 4\Big(b_3b_5 - \frac{b_4^2}{4}\Big)}}{2} \\
&= \frac{(b_3+b_5) + \sqrt{\big(b_3+b_5\big)^2 - 4b_3b_5 + b_4^2}}{2} \\
&= \frac{(b_3+b_5) + \sqrt{\big(b_3-b_5\big)^2 + b_4^2}}{2}
\end{aligned}
$$

and λ_2 be the second eigen value:

$$
\begin{aligned}
\lambda_2 &= \frac{(b_3+b_5) - \sqrt{\Big(b_3+b_5\Big)^2 - 4\Big(b_3b_5 - \frac{b_4^2}{4}\Big)}}{2} \\
&= \frac{(b_3+b_5) - \sqrt{\big(b_3+b_5\big)^2 - 4b_3b_5 + b_4^2}}{2} \\
&= \frac{(b_3+b_5) - \sqrt{\big(b_3-b_5\big)^2 + b_4^2}}{2}
\end{aligned}
$$

We will define a matrix $\mathbf{M}$ whose columns are the egienvectors associated with λ_1 and λ_2:

$$
\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 \end{bmatrix}
$$

To solve for the elements of $\mathbf{m}_i$ we use eigen vector associated with λ_i and then normalize it s.t. $\mathbf{m}_i^{'}\mathbf{m}_i = 1$. Here we complete that process using λ_1:

$$(\mathbf{B} - \lambda_1\mathbf{I})\mathbf{m_1} = \mathbf{0}$$

$$\begin{bmatrix} b_3 - \lambda_1 & b_4/2 \\ b_4/2 & b_5 - \lambda_1 \end{bmatrix} \begin{bmatrix} m_{11} \\ m_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$\Rightarrow$

$$b_3 m_{11} - \lambda_1 m_{11} + b_4 m_{21}/2 = 0$$

$$b_4 m_{11}/2 + b_5 m_{21} - \lambda_1 m_{21} = 0$$

Next, we need to normalize it i.e. $m_{11}^2 + m_{21}^2 = 1$. A simple way of doing this it to set $m_{11} = 1$ and solve for m_{21}:

$$\begin{aligned} m_{21} &= \frac{b_4}{2} \div (b_5 - \lambda_1) \\ &= \frac{b_4}{2(b_5 - \lambda_1)} \\ &= \frac{b_4}{2b_5 - b_3 - b_5 - \sqrt{(b_5 - b_3)^2 + b_4^2}} \text{ plug in } \lambda_1 \\ &= \frac{b_4}{b_5 - b_3 - \sqrt{(b_5 - b_3)^2 + b_4^2}} \end{aligned}$$

Now we begin our process of tranforming the principal axes back to the original variables. Let $z = d - d_0$ and define:

$$\begin{aligned} \mathbf{w} = \mathbf{m}'\mathbf{z} = \begin{bmatrix} \mathbf{m_1} \\ \mathbf{m_2} \end{bmatrix} (\mathbf{d} - \mathbf{d_0}) &= \begin{bmatrix} m_{11} & m_{21} \\ m_{12} & m_{22} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \\ &= \begin{bmatrix} m_{11}(x - x_0) + m_{21}(y - y_0) \\ m_{12}(x - x_0) + m_{22}(y - y_0) \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \end{aligned}$$

w_1 and w_2 are the variables the canonical equations use so in this next step, we will remove them.

Let $w_1 = 0$ and solve the first equation above:

$$\begin{aligned} m_{11}x - m_{11}x_0 + m_{21}y - m_{21}y_0 &= 0 \\ \rightarrow m_{21}y &= -x + x_0 + m_{21}y_0 \\ \rightarrow y &= -\frac{x}{m_{21}} + \frac{x_0}{m_{21}} + y_0 \\ \rightarrow y &= -P_{21}x + P_{21}x_0 + y_0 \end{aligned}$$

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

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 $P_{11} = \frac{b_5 - b_3 + \sqrt{(b_5-b_3)^2+b_4^2}}{b_4}$.

Now we have transformed our principal axes back to functions of our original variables. This will make them easier to plot and interpret. In a concave surface, the first principal axis (Eq 12) tells us where the surface has the lowest downward curvature. This is where the two explanatory variables create a maximized surface that decreases the least as we move along the line. The second principal axis is the opposite of the first principal axis. The second principal axis is the line in the surface where the outcomes decreases the most. It is the line in the surface where we the two explanotry variables create the greatest decrease in the response. The simplest case would be when the first principal axis is parallel to the line in the surface where $X = Y$ and the second principal axis is parallel to the line in the surface where $X = -Y$. If these axes are parallel to these lines, then our intuition about our explanatory variables and their congruence, or lack of, would be correct.

We can easily create a function in `R` to take the `lm` output to get the principal axis (Eq 12 and 13):

```
paxis <- function(lm){
  # Takes a quadratic lm function just as the function before and just as before
  # coef(lm) gives a vector of all the parameter estimates in the linear model
```

```
  # So here we are grabbing the individual parameter estimates from that vector
  b0 <- as.numeric(coef(lm)[1])
  b1 <- as.numeric(coef(lm)[2])
  b2 <- as.numeric(coef(lm)[3])
  b3 <- as.numeric(coef(lm)[4])
  b4 <- as.numeric(coef(lm)[5])
  b5 <- as.numeric(coef(lm)[6])

  # 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)
}
```

Now that we have devloped some intuition about the model, we will look at some examples of their use. First, we will recreate an example of the model being used in a paper then we will apply this methodology to our own data.

3 Re-creation of Previous Results

To check our understanding of how to run the model and make sure that our results are consistent with previous works, we reproduced the results in Phillips et al. (2012), *Congruence research in behavioral medicine: methodological review and demonstration of alternative methodology* [4]. In their paper, they were interested in seeing what created good situations in which patients followed through with their

doctors orders. To study this they looked at patient-reported adherence a month after the doctor visit(Z_1) and its correspondence to the physician's rating of patient's health and the physician's estimate of how the patient would rate of patient's own health. We'll begin by first outlining the variables used here:

$Z_{1,i}$ – 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

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

There are two parts to applying polynomial regression. First they check whether or not it is appropriate to use polynomial regression, and the second, if appropriate, they would check any hypothesis they had. In this paper, they had no pre-formed hypothesis so they explored the surface to see what they could learn.

They start by determining whether or not they should use polynomial regression. Before we go through their process of appropriateness, we looked at their data which we show in Table 1. As we can see for patient-reported adherence a month after the doctor visit for patient i, most of the data is greater than 0 so it is not equally balanced on the scale. Most of the data for the second response variable, physician's percieved agreement with patient i on the illness treatment, lies between 3 and 5. It is surprising that a large part of the data lives in a small range.

Now, we will check the assumptions of normality for all the models by looking for points with high leverage and potential influential points. No points showed any major signs of concern although there was one point which consistently had a higher leverage and Cook's D value compared to the rest of the points in the data set. This is not a major issue, but rather something we need to keep in mind and

be aware of while the analysis is continued. Next, in the paper, four criterion are presented which help decide whether using a polynomial surface is appropriate.

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):

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

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

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

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

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

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

where W is a dummy variable equal to 0 when $X > Y$, 1 when $X < Y$, and randomly assigned 1 or 0 when $X = Y$.

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).

The first criterion checks the basic linear model assumption which is whether the model is better then a basic mean(intercept) only model (Eq 14). The second criterion checks whether the coefficients are lined up in the way we expect them too and whether they add anything to the model. The fourth criterion is a intertwined with the second in that they check whether a simple model or polynomial model should be used. It can very well happen that the variables are significant but none of the higher order terms are in which case the response surface would be a plane. The thrid criterion checks to see if there

is benefit in moving from the simpler difference models to more complex polynomial models.

We have replicated Table 2 found in Phillips et al. (2012) in Table 2 to help with checking these assumptions. We can see from Table 2 under the unconstrained model both outcomes explained a significant amount of variation in the data; so criterion 1 was satisfied for both outcomes (R^2 of 0.1,0.18 for responses Z_1 and Z_2 respectively). However, criterion 2 was 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. In Table 2, F_c is the F-ratio that tests the accuracy of the constraints imposed by the particular difference score model. This F-test is used in checking Criterion 3. The third Criterion was only met for the second outcome i.e. the physician perceptions (54.89) and not for patient adherence (0.84). Whereas F_h is the F-ratio that tests the difference in variance explained in the outcome by the unconstrained model and the higher-order terms for that model where the algebraic difference score model has higher order terms: X^2, XY, and Y^2, and the absolute value difference score model has higher order terms: X^2, XY, Y^2, WX^2, WXY, WY^2). This F-ratio will help us in checking the Fourth Criterion. Again we see that the fourth criterion is only statisfied for the outcome conerning physician perceptions of patient-agreement (17.39) and not for patient adherence (0.83). So by these criteria, we may 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.

Table 2: 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

Similarly, we also checked the criterion using the absolute-value difference score model. Again we can refer to Table 2 when checking the criterion. Like with the difference score model, the unconstrained model explained a significant amount of the variation for both outcomes (0.16,0.21). Again criterion 2 was only satisfied for the outcome dealing with patient adherence as the regression coefficients for physician perceptions were not going in opposite directions and were not both significant. For criterion 3, the unconstrained model regarding physician perception was the only unconstrained model that explained

the variation in the data better then the absolute-value difference score model(2.19 vs. 13.22), but it reversed for the test to see if a higher-order polynomial model(in this case a qudratic model) to explain the variation in the data better(2.18 for patient adherence but not for physician perceptions of patient-agreement 7.45). We can conclude that the absolute-value difference score model was not an adequate fit for either of the outcome variables, the difference score is an adequate model for the outcome concerning patient adherence, and that the polynomial regression would be the best model for the data using the physician perception outcome.

Now we will continue the analysis by exploring the data; this is used when there is not a specific hyothesis in mind. We can preform change of sum of squares F-tests to check that we should use a quadratic model. If we test linear model vs the quadratic model, we see that the quadratic model explains the data significantly better than the linear model with a p-value of 0.0065404, but when we do not see a significant improvement when using the cubic vs the quadratic (p-value 0.673799).

So the fitted model(Eq 19) is:

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

We can use our functions to get the stationary points and the principal axes.

```
statpts(lm2Square)

##          x0        y0
## 1 -2.481127 -2.003207

paxis(lm2Square)

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

So we can see that our stationary points are at $(-2.48, -2)$. The first and second principal axis are Equations 20 and 21, respectively:

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

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

The surface slice plots in of both principal axes in Figure 4. We can also look at the predicted surface plot using `Plotly`. We can follow the link to the `Plotly` website and look at and manipulate the predicted surface.

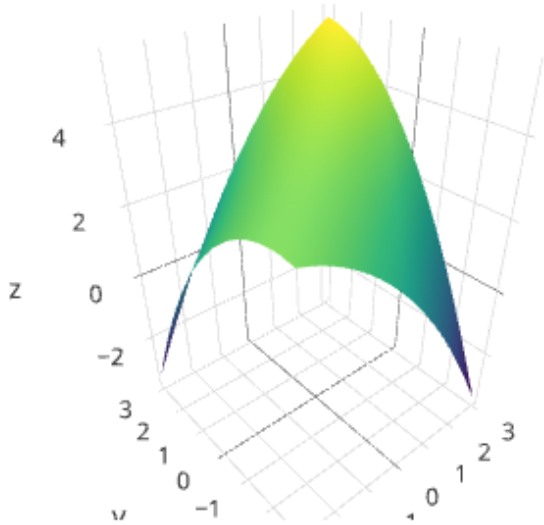

Figure 1: A screenshot of the predicted surface plot from the center perspective.

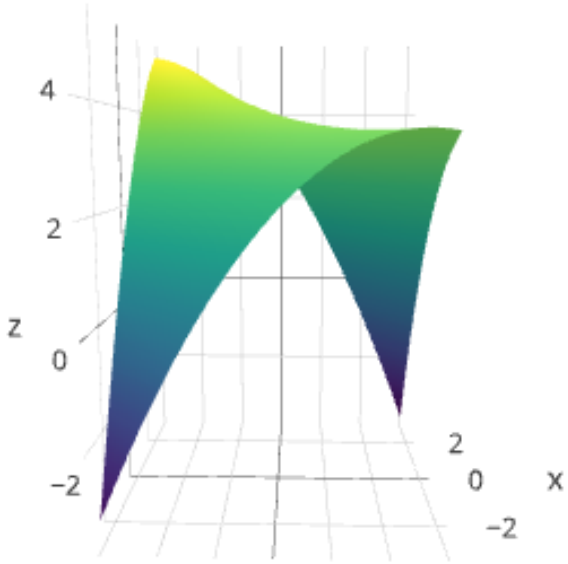

Figure 2: A screenshot of the predicted surface plot from the y axis perspective i.e. eliminating the x axis.

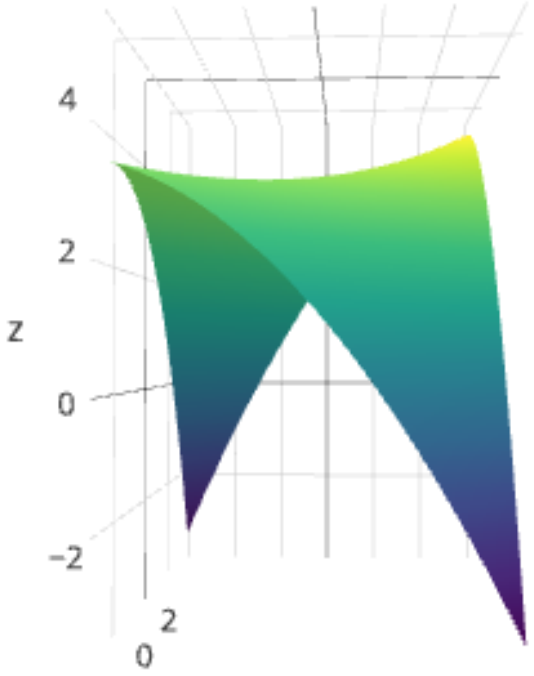

Figure 3: A screenshot of the predicted surface plot from the x axis perspective i.e. eliminating the y axis.

As we can see in the figures above, the first principal axes is the axis that makes the most sense. The support of the function lies where we have data. This line 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. As we can imagine, as the two predictor variables increase, the response increases as well. This is examplified in Figure 1. The response variable(Z) is highest when the explanatory variables(X and Y) are highest as well. This makes sense because when the physician's rating of patient's health and the physician's estimate of how patient would rate their own health agree i.e. the patient and physician are in agreement about the situation, then the patient is more likely to follow the doctor's orders in the month following the visit. We can also see this in the surface plot. Notice that the lowest points in the surface are when the physician rating and the physician's perception of the patient's own rating are at odds with each other (points (-2,3,-3) or (-2,-3,3) in Figure 1). This also makes sense; if the doctor and patient don't agree with what is happening to the patient, then the patient is much less likely to follow through the doctors orders. The surface allows us to see how the two predictor variables intereact and affect the response.

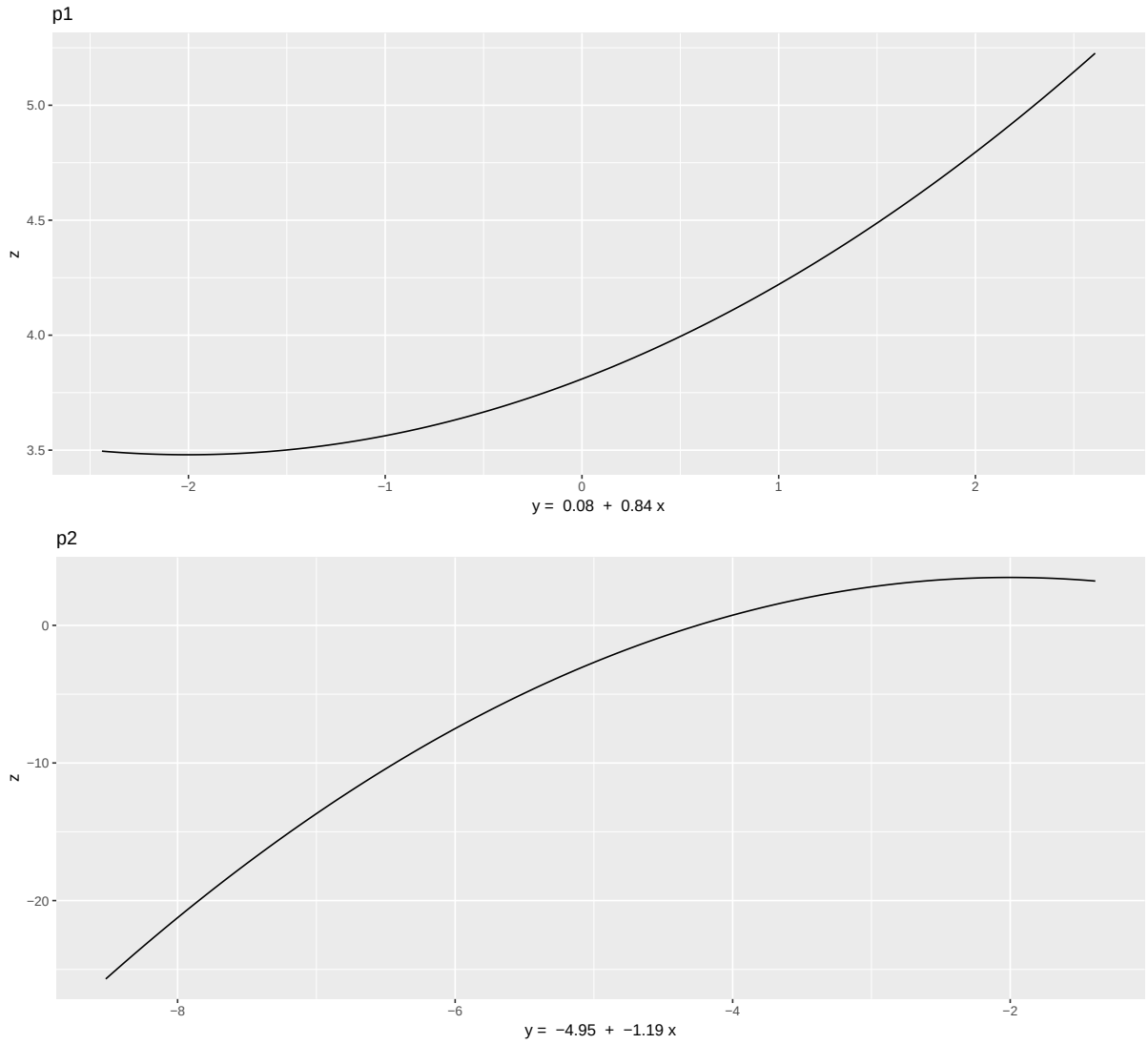

Figure 4: This Figure shows slices of the response surface taken at the two principal axes.

4 Our Example

For our example, we were interested in understanding the congruence between survey reports and observer reports. Our data comes from the Iowa Midlife Transitions Project (MTP); this longitudinal study conducted between 1991 and 2001 on families from eight counties in rural Iowa. The study was designed to look at the effects of the 1960's farm crisis on rural Midwestern families. Families in MTP either participated in the Iowa Youth and Families Project (IFYP) or the Iowa Single Parent Project (ISPP). The IFYP started in 1989 to follow families with at least 2 children and married parents; one of the requirements was the one of the children be in seventh grade in '89 while the other sibling was in a four year range in age of the child. The ISPP started in '91. It focused on recently divorced mothers with a minimum of two children, one of which was a ninth grader in the starting year and the other child was in a 4 year age range of the other child just like before. They had a 78% participation rate of all possible participants in the IFYP; similarly the ISSP had a 99% participation rate. While there were a myriad of economic variables available, we focused on family dynamic variables. Specifically we focused on the husband and wife relationship variables. Our variables of interest are divided into two sets.

Set 1:

$$Z : \text{Observed Relationship Quility of couple } i \text{ at Time Wave 0}$$
$$Y_1 : \text{Wife's Report of Husbund's Hostility for couple } i \text{ at Time Wave 0}$$
$$X_1 : \text{Husband's Report of Husbund's Hostility for couple } i \text{ at Time Wave 0}$$

Set 2:

$$Z : \text{Observed Relationship Quility of couple } i \text{ at Time Wave 0}$$
$$Y_2 : \text{Wife's Report of Relationship Instability of couple } i \text{ at Time Wave 0}$$
$$X_2 : \text{Husband's Report of Relationship Instability of couple } i \text{ at Time Wave 0}$$

Just like in Phillips et al. (2012), we also centered our variables to ease in interpretation and analysis. We will also check to see if the use of polynomial regression is justified in our case. To do this, we will check the same criterion as we did earlier, namely:

1. Check whether the unconstrained model(see below) explained a significant amount of the variance

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's	n
Z	-4.00	-1.00	1.00	0.76	2.00	4.00	148.00	407.00
X_1	-2.40	-1.80	-1.20	-1.17	-0.60	1.80	123.00	432.00
Y_1	-2.60	-2.05	-1.60	-1.52	-1.00	3.00	115.00	440.00
X_2	-1.50	0.85	1.50	1.04	1.50	1.50	115.00	440.00
Y_2	-1.50	0.90	1.50	1.08	1.50	1.50	123.00	432.00

Table 3: Summary statistics of response variable (Z) and explanotry variables (X_1,Y_1 and X_2,Y_2).

in the data.

$Z = \beta_0 + \epsilon$ vs. $Z = \beta_0 + \beta_1 X + \beta_2 Y + \epsilon$

$Z = \beta_0 + \epsilon$ vs. $Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 W + \beta_4 WX + \beta_5 WY + \epsilon$

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 model is significantly better than the constrained model.

$Z = \beta_0 + \beta_1(X - Y) + \epsilon$ vs $Z = \beta_0 + \beta_1 X + \beta_2 Y + \epsilon$

$Z = \beta_0 + \beta_1|X - Y| + \epsilon$ vs $Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 W + \beta_4 WX + \beta_5 WY + \epsilon$

4. Check whether a higher order model i.e. polynomial regression is significantly better than the constrained model.

$Z = \beta_0 + \beta_1(X - Y) + \epsilon$ vs $Z = \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 X^2 + \beta_4 Y^2 + \beta_5 XY + \epsilon$

We also checked that our assumptions of normality and constant variance were not violated using the same methods that Dr. Phillips used: high leverage and influential points via Cook's D. We do not find any obvious violations in assumptions for either of the two examples.

We created a table similar to Table 2 in Table 4 to help us collect our information about the confirmatory process which will in turn help us examine the criterion.

Looking at criterion 1, the unconstrained model, in both outcomes, explained a significant amount of variation in the data ($R^2 = .22$, pval of $2.3632091 \times 10^{-23}$ and $R^2 = .11$, pval of $5.6530671 \times 10^{-12}$, respectively). For criterion 2, the regression coefficients for both models were consistent with each other which is what we would expect. The difference in direction between the two models has to deal with the

fact that one would be minimizing and the other is maximizing. Additionally, for both of the unconstrained algebraic models, the regression coefficients are significant. However, the regression coefficients for the unconstrained absolute difference models were not all significant leading us to believe that it would not be agood fit for the data. Criterion 3 and 4 were both met by both examples/models. All the F-ratios in Table 4 are significant (Criterion 3: 112.54^{***}, 50.41^{***}, and Criterion 4: 28.47^{***}, 14.48^{***}). They both suggest using something greater than a simple difference model although that should not be an unconstrained absolute difference model.

Table 4: This table follows the same format as Table 2, in that it summarizes the confimatory approach to polynomial regression. Here we see that the 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		
Relationship Quality w/ Hostility	0.03	0	-0.54^{***}	-0.66^{***}				0.22^{***}	112.54^{***}	28.47^{***}
Relationship Quality w/ Instability	-0.09	0	0.35^{*}	0.35^{***}				0.11^{***}	50.41^{***}	14.48^{***}
	$\lvert X-Y\rvert$	R^2	X	Y	W	WX	WY	R^2		
Relationship Quality w/ Hostility	-0.44^{**}	0.02^{**}	-0.42^{*}	-0.74^{**}	-0.09	-0.23	0.04	0.22^{***}	25.81^{***}	13.12^{***}
Relationship Quality w/ Instability	-0.75^{***}	0.05^{***}	-0.1	0.74^{***}	-0.4	0.59	-0.24	0.12^{***}	7.95^{***}	5.36^{***}

After checking our criterion, we find that it is appropriate to use more than difference scores on both examples. If we examine closer, our first model suggests using a linear model between to the two predictors and the response (p-val 0.636177 i.e. $H0 : \beta_3 = \beta_4 = \beta_5 = 0$ in Eq 19). In our second model which examines the relationship between observed relationship quality of the marriage and the precieved relationship instabililty has moderate evidence of using a quadratic model for the relationship (p-val 0.073613 i.e. $H0 : \beta_3 = \beta_4 = \beta_5 = 0$ in Eq 19) and has very weak evidence for using a cubic model on the relationship (p-val 0.0945018). Given this information, we will proceed with using the quadratic polynomial regression model on our second example. While we could reach for using a cubic model, in this case given the complexity of the cubic polynomial regression it is best to stick with quadratic regression which already has enough challenges.

Recall the fitted quadratic polynomial for our example using precieved relationship instabililty as a predictor for relationship quality is:

$$\hat{z} = -0.52 + 0.25x + 0.23y + 0.02x^2 + 0.09xy + 0.37y^2 \tag{22}$$

We can see the estimates of the model and their standard errors in Table 5. As we can see only two of the variables are found to be significant. While this is not ideal, for the sake of this example we will continue on.

	b_0	b_1	b_2	b_3	b_4	b_5
Estimate	-0.52	0.25	0.23	0.02	0.09	0.37
Std. Error	0.26	0.19	0.21	0.21	0.21	0.20
Pr(¿—t—)	0.05	0.19	0.28	0.94	0.68	0.06

Table 5: Model estimates and SE for polynomial regression model using relationship instability ratings as a explanatory variables.

Using `Plotly`, we can look at and manipulate the predicted surface by following this link. I have included some screenshots here so you can get a feel for the convexity of the surface which will help in the interpretations of the key components in the surface.

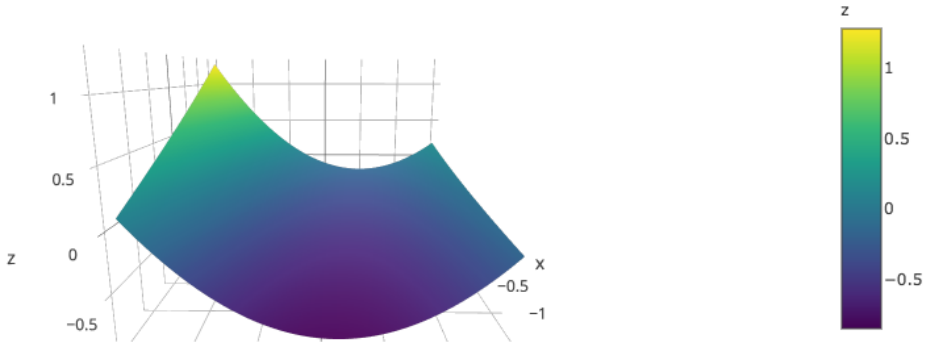

Figure 5: A screenshot of the predicted surface plot from the center perspective.

Just as we did earlier, we will use our functions to get the stationary points and the principal axes.

```
statpts(lm2Square)

##          x0       y0
## 1 -10.18558 0.920306

paxis(lm2Square)

##        p10      p11        p20        p21
## 1 82.92166 8.050729 -0.3448691 -0.1242124
```

The stationary points are located at $(-10.19, 0.92)$; due to our convex nature of the surface, these points are the levels of relationship instability that will minimize the relationship quality in a marriage.

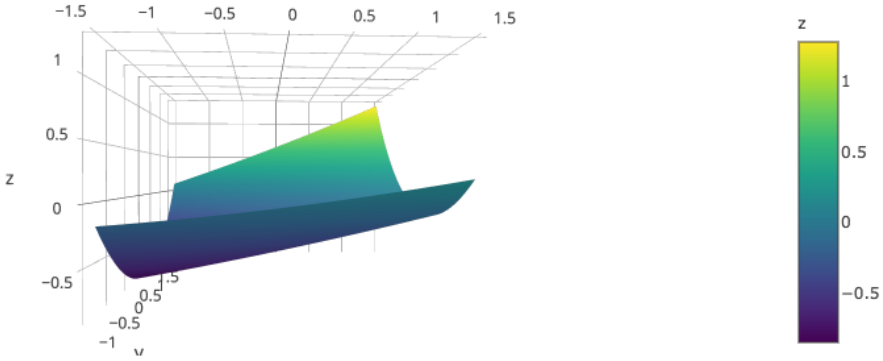

Figure 6: A screenshot of the predicted surface plot from the y axis perspective i.e. eliminating the x axis.

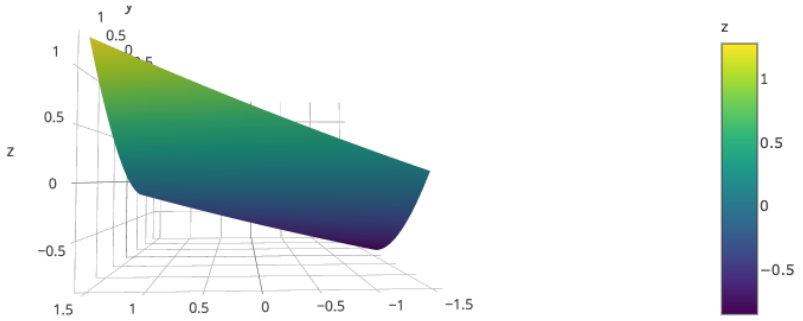

Figure 7: A screenshot of the predicted surface plot from the x axis perspective i.e. eliminating the y axis.

The predicted respnse at the stationary point is -1.705 whic is low but not too much. It is interesting that relationship quality seems to be minimzed when the wife's percieved relationship instability is really low while the husbands percieved raltionship instability is on the higher side.

4.1 Principal Axes (Interpretations come with the plots)

Likewise we can look at the principal axes:

$$y = 82.92 + 8.05x \tag{23}$$

$$y = -0.34 - 0.12x \tag{24}$$

We can look at surface slice plots in of both principal axes in Figure 8.

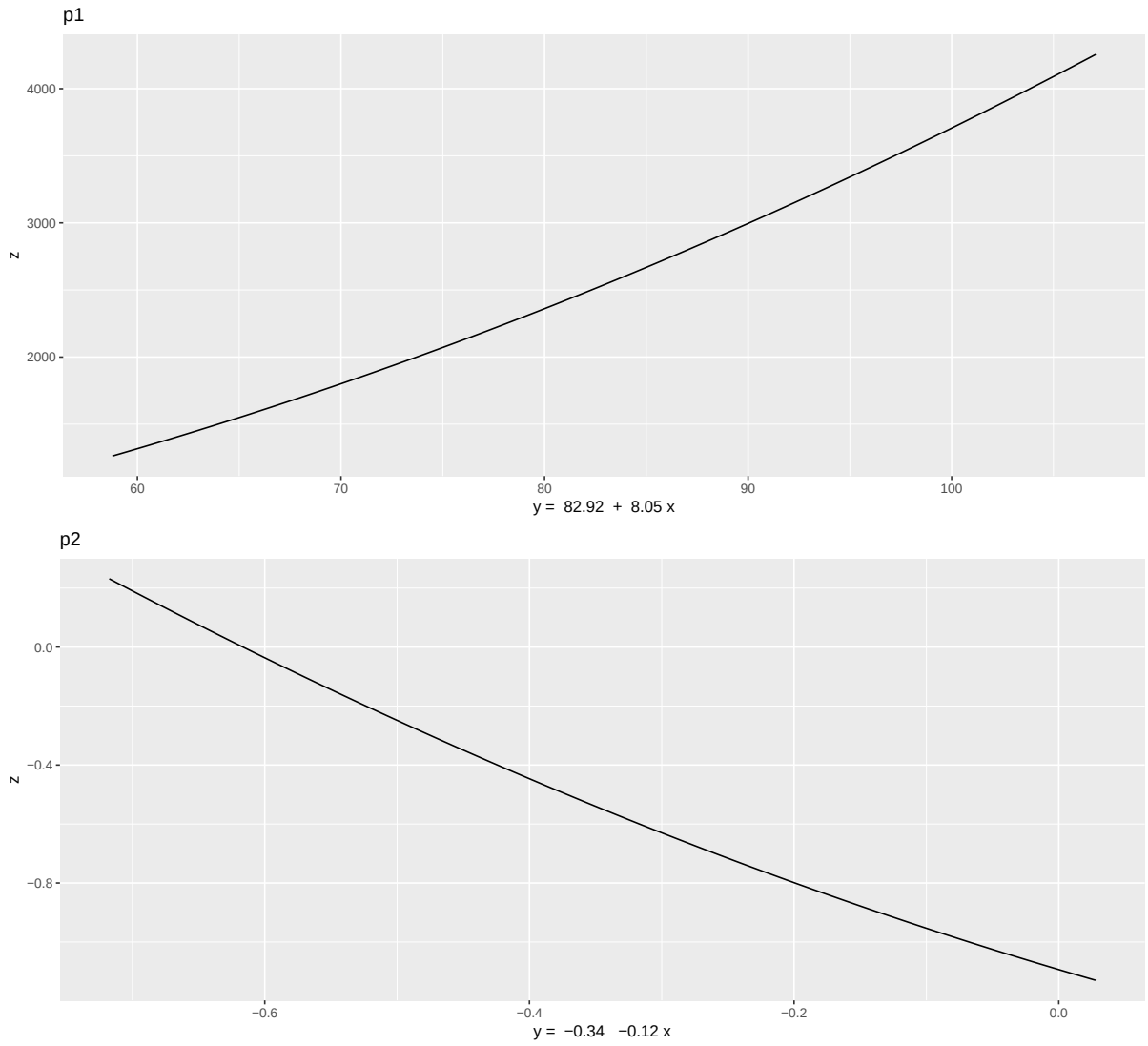

Figure 8: This Figure shows slices of the response surface taken at the two principal axes.

Because we have a convex surface, our interpetations of the principal axis are flipped; the first principal axis, Equation 23, represents the line in the surface where the outcomes decreases the most and the second principal axis, Equaion 24, is where the surface has the lowest downward curvature. So the first principal axis shows us where relationship qualtiy can go downhil rather fast. It shows high level of percieved

relationship instability by the wife. And likewise, the second princpal axis shows that relationship quality does not decrease that much when the wife's percieved relationship instability hangs around 0 (in the middle of the observed range). There is an issue with the first principal axis, namely that the ranges we would have to make inferences on are out of the bounds of data. This is an issue because it predicts way outside of the scope of our data, so we can't really trust any inference made on the first principal axis similar to how the second principal axis in Phillips(2012) study was out of the bounds of the data. From our surface, we can gather most integral to relationship quality is the wife's perception of instability. When the wife's rating of the relationship instability is held constant and one move along the husband's rating of relationship instability the observed relationship quality does no change too much. If the opposite is considered i.e. hold husband's rating of relationship instability the observed relationship quality and move along the wife's rating of relationship instability the observed relationship quality there is a great change in predicted observed relationship quality.

There are some issues with our data that affect inferences. As we can see from Figure 9, our predictor variables are very left skewed. The inferences we can make when the relationship has a high percieved instability are fairly well rounded, but we are lacking data in the low percieved relationship instability realm as see in Figure 10. The method does seem to be allow us to easily visualize the realm of possible combinations of the predictor variables and their effect on the predicted response. Ultimately, I feel that the predictor variables need to be fairly well distributed in order to make proper inferences.

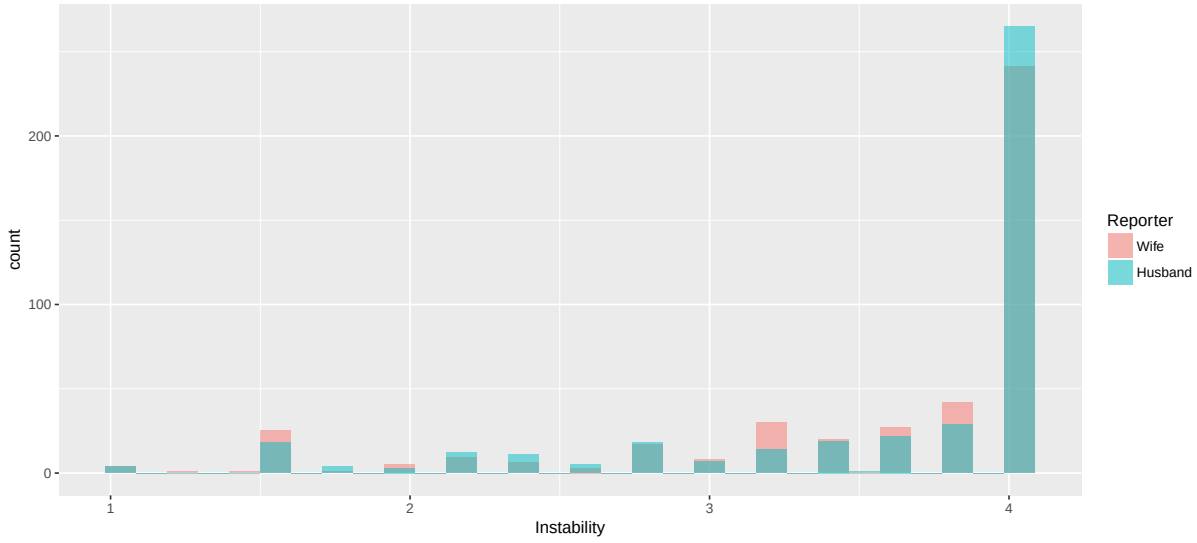

Figure 9: This Figure shows the histograms of our two predictor variables plotted on top of each other to show skewness.

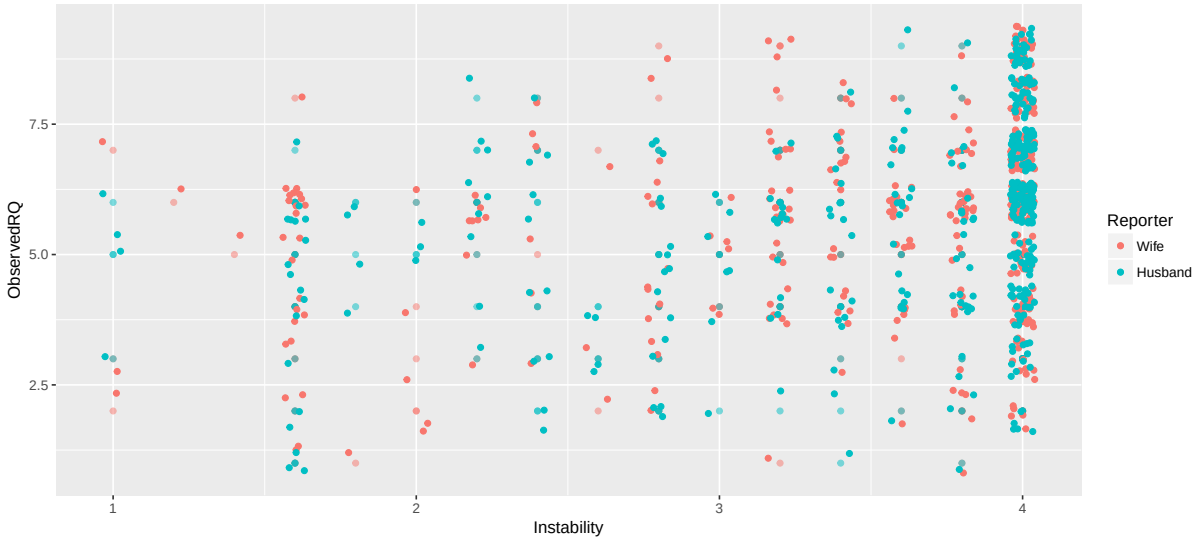

Figure 10: This Figure shows a scatterplot of Instability against Observed Relationship Quality colored by who reported the instability rating.

References

[1] J. R. Edwards, *On the use of polynomial regression equations as an alternative to difference scores in organizational research*, Academy of Management Journal, 36 (1993), pp. 1577–1613.

[2] ——, *Problems with the use of profile similarity indices in the study of congruence in organizational research*, Personnel Psychology, 46 (1993), pp. 641–665.

[3] A. Khuri and J. Cornell, *Response Surfaces: Designs and Analyses: Second Edition*, Statistics: A Series of Textbooks and Monographs, Taylor & Francis, 1996.

[4] L. A. Phillips, *Congruence research in behavioral medicine: methodological review and demonstration of alternative methodology*, Journal of Behavioral Medicine, 36 (2012), pp. 61–74.

[5] L. A. Phillips, M. A. Diefenbach, I. M. Kronish, R. M. Negron, and C. R. Horowitz, *The necessity-concerns framework: a multidimensional theory benefits from multidimensional analysis*, Annals of Behavioral Medicine, 48 (2014), pp. 7–16.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

A Source code

Presented here is all the R code used in the Creative Component:

```
QuadFit <- lm(Z ~ X + Y + I(X^2) + I(X*Y) + I(Y^2), data=d)
statpts <- function(lm){
  # Takes a quadratic lm function
  # coef(lm) gives a vector of all the parameter estimates in the linear model
  # So here we are grabbing the individual parameter estimates from that vector
  b0 <- as.numeric(coef(lm)[1])
  b1 <- as.numeric(coef(lm)[2])
  b2 <- as.numeric(coef(lm)[3])
  b3 <- as.numeric(coef(lm)[4])
  b4 <- as.numeric(coef(lm)[5])
  b5 <- as.numeric(coef(lm)[6])

  # Stationary Pts. using the formulas in Eq 10 & 11
  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)
  return(out)
}
x0 <- statpts(lm)$x0
y0 <- statpts(lm)$y0
predstatpts <- coef(lm)[1] + coef(lm)[2]*x0 + coef(lm)[3]*y0 + coef(lm)[4]*x0^2 + coef(lm)[5]*x0*y0 +
predstatpts
paxis <- function(lm){
  # Takes a quadratic lm function just as the function before and just as before
  # coef(lm) gives a vector of all the parameter estimates in the linear model
  # So here we are grabbing the individual parameter estimates from that vector
  b0 <- as.numeric(coef(lm)[1])
  b1 <- as.numeric(coef(lm)[2])
  b2 <- as.numeric(coef(lm)[3])
  b3 <- as.numeric(coef(lm)[4])
```

```
  b4 <- as.numeric(coef(lm)[5])
  b5 <- as.numeric(coef(lm)[6])

  # 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)
}
d <- read.spss("../Phillips_Data/Congruence PRIM.sav", to.data.frame=TRUE)

###### Data Setup ######
d$pre_curr_health <- as.numeric(d$pre_curr_health)
d$Quant_MD_rate_health <- as.numeric(d$Quant_MD_rate_health)
d$Quant_MD_rate_own <- as.numeric(d$Quant_MD_rate_own)
d$MiddleCentered_precurrhealth <- d$pre_curr_health - 3
d$MiddleCentered_MDratehealth  <- d$Quant_MD_rate_health - 3
d$MiddleCentered_MDrateown     <- d$Quant_MD_rate_own - 3



# Z: patient-reported adherence a month after the dr. visit
# X: physician's rating of patient's health
# Y: physician's est of how the patient would rate of patient's own health
d$z  <- d$PatientAdh_Avg
```

```r
d$x <- d$MiddleCentered_MDratehealth
d$y <- d$MiddleCentered_MDrateown



d$w <- d$x
d$w[which(!is.na(d$w))] <- 1 # These are all the points where X < Y
d$w[which(!is.na(d$w) & d$x > d$y)] <- 0
d$w[which(!is.na(d$w) & d$x == d$y)] <- sample(c(0,1),length(d$w[which(!is.na(d$w) & d$x == d$y)]), r



# Z: physician perceptions of patient-agreement
# X:
# Y:
d$z2 <- d$PhysicianSharedModels
d$x2 <- d$MiddleCentered_MDratehealth
d$y2 <- d$MiddleCentered_MDrateown
d$w2 <- d$x2
d$w2[which(!is.na(d$w2))] <- 1 # These are all the points where X < Y
d$w2[which(!is.na(d$w2) & d$x2 > d$y2)] <- 0
d$w2[which(!is.na(d$w2) & d$x2 == d$y2)] <- sample(c(0,1),length(d$w[which(!is.na(d$w2) & d$x2 == d$y

Phillipsdf <- data.frame(Z1=c(summary(d$z),n=sum(!is.na(d$z))),
                         Z2=c(summary(d$z2),n=sum(!is.na(d$z2))),
                         X=c(summary(d$x),n=sum(!is.na(d$x))),
                         Y=c(summary(d$y),n=sum(!is.na(d$y)))
)
colnames(Phillipsdf) <- c("$Z_1$","$Z_2$","$X$","$Y$")
print(xtable(t(Phillipsdf),caption = c("Summary statistics of response variables($Z_1$ and $Z_2$) and
# Z: patient-reported adherence a month after the dr. visit
# X: physician's rating of patient's health
# Y: physician's est of how the patient would rate of patient's own health


# z = x+y
lm <- lm(z~x+y, data=d)
```

```
lmR2 <- summary(lm)$r.squared
lmcoef <- coef(lm)


# z = x-y
lmDiff <- lm(z~I(x-y), data=d)
lmDiffR2 <- summary(lmDiff)$r.squared
lmDiffcoef <- coef(lmDiff)


# z = |x-y|
lmAbs <- lm(z~I(abs(x-y)), data=d)
lmAbsR2 <- summary(lmAbs)$r.squared
lmAbscoef <- coef(lmAbs)


# z = x+y+x^2+xy+y^2
lmSquare <- lm(z~x+y+I(x^2)+I(x*y)+I(y^2), data=d)


# z = x+y+w+wx+wy
lmW <- lm(z~x+y+w+I(w*x)+I(w*y), data=d)
lmWR2 <- summary(lmW)$r.squared
lmWcoef <- coef(lmW)


# z = x+y+w+wx+wy
lmWSquare <- lm(z~x+y+w+I(x^2)+I(x*y)+I(y^2)+I(w*x^2)+I(w*x*y)+I(w*y^2), data=d)




# Z: patient-reported adherence a month after the dr. visit
# X:
# Y:


# z = x+y
lm2 <- lm(z2~x2+y2, data=d)
lm2R2 <- summary(lm2)$r.squared
lm2coef <- coef(lm2)
```

```
# z2 = x2-y2
lm2Diff <- lm(z2~I(x2-y2), data=d)
lm2DiffR2 <- summary(lm2Diff)$r.squared
lm2Diffcoef <- coef(lm2Diff)

# z2 = |x2-y2|
lm2Abs <- lm(z2~I(abs(x2-y2)), data=d)
lm2AbsR2 <- summary(lm2Abs)$r.squared
lm2Abscoef <- coef(lm2Abs)

# z2 = x2+y2+x2^2+x2y2+y2^2
lm2Square <- lm(z2~x2+y2+I(x2^2)+I(x2*y2)+I(y2^2), data=d)

# z2 = x2+y2+x2^2+x2y2+y2^2
lm2Cube <- lm(z2~x2+y2+I(x2^3)+I(y2^3)+I(x2^2*y2)+I(x2*y2^2)+I(x2*y2)+I(x2^2)+I(y2^2), data=d)

# z2 = x2+y2+w+wx2+wy2
lm2W <- lm(z2~x2+y2+w+I(w*x2)+I(w*y2), data=d)
lm2WR2 <- summary(lm2W)$r.squared
lm2Wcoef <- coef(lm2W)

# z2 = x2+y2+w+wx2+wy2
lm2WSquare <- lm(z2~x2+y2+w+I(x2^2)+I(x2*y2)+I(y2^2)+I(w*x2^2)+I(w*x2*y2)+I(w*y2^2), data=d)
# Check the leverage (maybe one point of concern)
lev <- hat(model.matrix(lmW))
plot(lev)

# Basically see the same
cook <- cooks.distance(lmW)
plot(cook,ylab="Cooks distances")

statpts(lm2Square)
paxis(lm2Square)
```

```
principal_plot <- function(x){
  require(ggplot2)
  require(reshape2)
  x_temp <- seq(-3,3,length.out= 1000)
  y_p1 <- x$p10 + x$p11*x_temp
  y_p2 <- x$p20 + x$p21*x_temp
  y <- melt(data.frame(y_p1, y_p2))
  temp <- cbind(x = rep(x_temp,2), y)
  ggplot(data=temp, aes(x=x, y=value)) + geom_line() + facet_grid(variable~., scales = "free_y")
}


principal_plot_surface <- function(x, lm){
  require(ggplot2)
  require(reshape2)
  require(gridExtra)
  x_temp <- seq(-3,3,length.out= 1000)
  y_p1 <- x$p10 + x$p11*x_temp
  y_p2 <- x$p20 + x$p21*x_temp

  x2 <- x_temp
  y2 <- y_p1
  pred1 <- predict(lm, data.frame(x2,y2,x2^2,x2*y2,y2^2))

  x2 <- x_temp
  y2 <- y_p2
  pred2 <- predict(lm, data.frame(x2,y2,x2^2,x2*y2,y2^2))

  a <- ggplot() + geom_line(aes(x=y_p1, y=pred1)) + labs(x=paste("y = ", round(x$p10,2), " + ", round
  b <- ggplot() + geom_line(aes(x=y_p2, y=pred2)) + labs(x=paste("y = ", round(x$p20,2), " + ", round
  grid.arrange(a, b)
}
principal_plot_surface(paxis(lm2Square), lm2Square)
# Make Predictions
n     <- 100
```

```
xp    <- seq(-3,3,length.out= n)
y2    <- seq(-3,3,length.out= n)
preds <- matrix(rep(0, n))

for(i in 1:n){
  x2 <- rep(xp[i], times=n)
  preddat <- data.frame(x2, y2, x2^2, x2*y2, y2^2)
  preds   <- cbind(preds, as.numeric(predict(lm2Square, preddat)))
}

predictions <- list(x=xp,y=y2, z=preds[,-1])
### Make Plots
p <- with(predictions, plot_ly(x=x, y=y, z=z, type="surface"))
plotly_POST(p, filename="PhillipsExSurface")
# Link to plot
# https://plot.ly/~nulloa1/145/
d <- read.csv("../CurrentData/ModData_6_9.csv", header=TRUE)

###### Start Analysis ######
# Using Hostility as a predictor.

x <- d$HHhostW0 - 4
y <- d$WHhostW0 - 4
z <- d$OHrqW0 - 5

w <- x
w[which(!is.na(w))] <- 1 # These are all the points where X < Y
w[which(!is.na(w) & x > y)] <- 0
w[which(!is.na(w) & x == y)] <- sample(c(0,1),length(w[which(!is.na(w) & x==y)]), replace=TRUE)

# z = (x-y)
lmDiff <- lm(z~I(x-y))
lmDiffR2 <- summary(lmDiff)$r.squared
lmDiffcoef <- coef(lmDiff)
```

```
# z = |x-y|
lmAbs <- lm(z ~ I(abs(x-y)))
lmAbsR2 <- summary(lmAbs)$r.squared
lmAbscoef <- coef(lmAbs)

# z = x+y+w+wx+wy
lmW <- lm(z~x+y+w+I(w*x)+I(w*y))
lmWR2 <- summary(lmW)$r.squared
lmWcoef <- coef(lmW)

# z = x+y+w+wx+wy
lmWSquare <- lm(z~x+y+w+I(x^2)+I(x*y)+I(y^2)+I(w*x^2)+I(w*x*y)+I(w*y^2))

# z = x+y
lm <- lm(z~x+y)
lmR2 <- summary(lm)$r.squared
lmcoef <- coef(lm)

# z = x+y+x^2+xy+y^2
lmSquare <- lm(z~x+y+I(x^2)+I(x*y)+I(y^2))

# z = x+y+x^2+xy+y^2+x^3+x^2y+xy^2+y^3
lmCube <- lm(z~x+y+I(x^2)+I(x*y)+I(y^2)+I(x^3)+I((x^2)*y)+I(x*y^2)+I(y^3))

# Using Instability as a predictor.
z <- d$OHrqW0 - 5
x <- d$WWinstH0 - 2.5
y <- d$HHinstW0 - 2.5
w <- x
w[which(!is.na(w))] <- 1 # These are all the points where X < Y
w[which(!is.na(w) & x > y)] <- 0
w[which(!is.na(w) & x == y)] <- sample(c(0,1),length(w[which(!is.na(w) & x==y)]), replace=TRUE)
```

```
# z = (x-y)
lm2Diff <- lm(z ~ I(x-y))
lm2DiffR2 <- summary(lm2Diff)$r.squared
lm2Diffcoef <- coef(lm2Diff)

# z = x+y
lm2 <- lm(z ~ x + y)
lm2R2 <- summary(lm2)$r.squared
lm2coef <- coef(lm2)

# z = |x-y|
lm2Abs <- lm(z ~ I(abs(x-y)))
lm2AbsR2 <- summary(lm2Abs)$r.squared
lm2Abscoef <- coef(lm2Abs)

# z = x+y+x^2+xy+y^2
lm2Square <- lm(z ~ x + y + I(x^2) + I(x*y) + I(y^2))

# z = x+y+x^2+xy+y^2+x^3+x^2y+xy^2+y^3
lm2Cube <- lm(z~x+y+I(x^2)+I(x*y)+I(y^2)+I(x^3)+I((x^2)*y)+I(x*y^2)+I(y^3))

# z = x+y+w+wx+wy
lm2W <- lm(z~x+y+w+I(w*x)+I(w*y))
lm2WR2 <- summary(lm2W)$r.squared
lm2Wcoef <- coef(lm2W)

# z = x+y+w+wx+wy
lm2WSquare <- lm(z~x+y+w+I(x^2)+I(x*y)+I(y^2)+I(w*x^2)+I(w*x*y)+I(w*y^2))
# Check the leverage (maybe one point of concern)
lev <- hat(model.matrix(lm1))
plot(lev)

# Basically see the same
cook <- cooks.distance(lm1)
```

```
plot(cook,ylab="Cooks distances")


# Nothing of concern. maybe one point.
ourdf <- data.frame(Z=c(summary(d$OHrqW0 - 5),n=sum(!is.na(d$OHrqW0))),
                    X1=c(summary(d$HHhostW0 - 4),n=sum(!is.na(d$HHhostW0))),
                    Y1=c(summary(d$WHhostW0 - 4),n=sum(!is.na(d$WHhostW0))),
                    X2=c(summary(d$WWinstH0 - 2.5),n=sum(!is.na(d$WWinstH0))),
                    Y2=c(summary(d$HHinstW0 - 2.5),n=sum(!is.na(d$HHinstW0)))
)
colnames(ourdf) <- c("$Z$","$X_1$","$Y_1$","$X_2$","$Y_2$")
print(xtable(t(ourdf),caption = c("Summary statistics of response variable ($Z$) and explanotry varia
MEdf <- t(summary(lm2Square)$coefficients[,c(1,2,4)])
colnames(MEdf) <- c("$b_0$","$b_1$","$b_2$","$b_3$","$b_4$","$b_5$")
print(xtable(MEdf,caption = c("Model estimates and SE for polynomial regression model using relations
# Make Predictions
n     <- 100
xp    <- seq(-1.5,1.5,length.out= n)
y     <- seq(-1.5,1.5,length.out= n)
preds <- matrix(rep(0, n))

for(i in 1:n){
  x <- rep(xp[i], times=n)
  preddat <- data.frame(x, y, x^2, x*y, y^2)
  preds   <- cbind(preds, as.numeric(predict(lm2Square, preddat)))
}

our_predictions <- list(x=xp,y=y, z=preds[,-1])
### Make Plots
p <- with(our_predictions, plot_ly(x=x, y=y, z=z, type="surface"))
plotly_POST(p, filename="OurExPlot")
# Link to plot
# https://plot.ly/~nulloa1/147/
statpts(lm2Square)
paxis(lm2Square)
```

```
x0 <- statpts(lm2Square)$x0
y0 <- statpts(lm2Square)$y0
predstatpts <- coef(lm2Square)[1] + coef(lm2Square)[2]*x0 + coef(lm2Square)[3]*y0 + coef(lm2Square)[4
require(ggplot2)
require(reshape2)
require(gridExtra)
x_temp <- seq(-3,3,length.out= 1000)
y_p1 <- paxis(lm2Square)$p10 + paxis(lm2Square)$p11*x_temp
y_p2 <- paxis(lm2Square)$p20 + paxis(lm2Square)$p21*x_temp

x <- x_temp
y <- y_p1
pred1 <- predict(lm2Square, data.frame(x,y,x^2,x*y,y^2))
x <- x_temp
y <- y_p2
pred2 <- predict(lm2Square, data.frame(x,y,x^2,x*y,y^2))

a <- ggplot() + geom_line(aes(x=y_p1, y=pred1)) + labs(x=paste("y = ", round(paxis(lm2Square)$p10,2),
b <- ggplot() + geom_line(aes(x=y_p2, y=pred2)) + labs(x=paste("y = ", round(paxis(lm2Square)$p20,2),
grid.arrange(a, b)

skewcomb <- data.frame(d$WWinstH0,d$HHinstW0)
skewcomb <- data.frame(melt(skewcomb, by="d.OHrqW0"), rep(d$OHrqW0,2))
levels(skewcomb$variable) <- c("Wife", "Husband")
names(skewcomb) <- c("Reporter","Instability","ObservedRQ")
ggplot(skewcomb, aes(x=Instability, fill=Reporter)) + geom_histogram(alpha=0.5, position="identity")
ggplot(skewcomb, aes(x=Instability, y=ObservedRQ, color=Reporter)) + geom_point(alpha=0.5) + geom_jit
```