2017 Prelim

Will Gertsch

8/19/2020

1

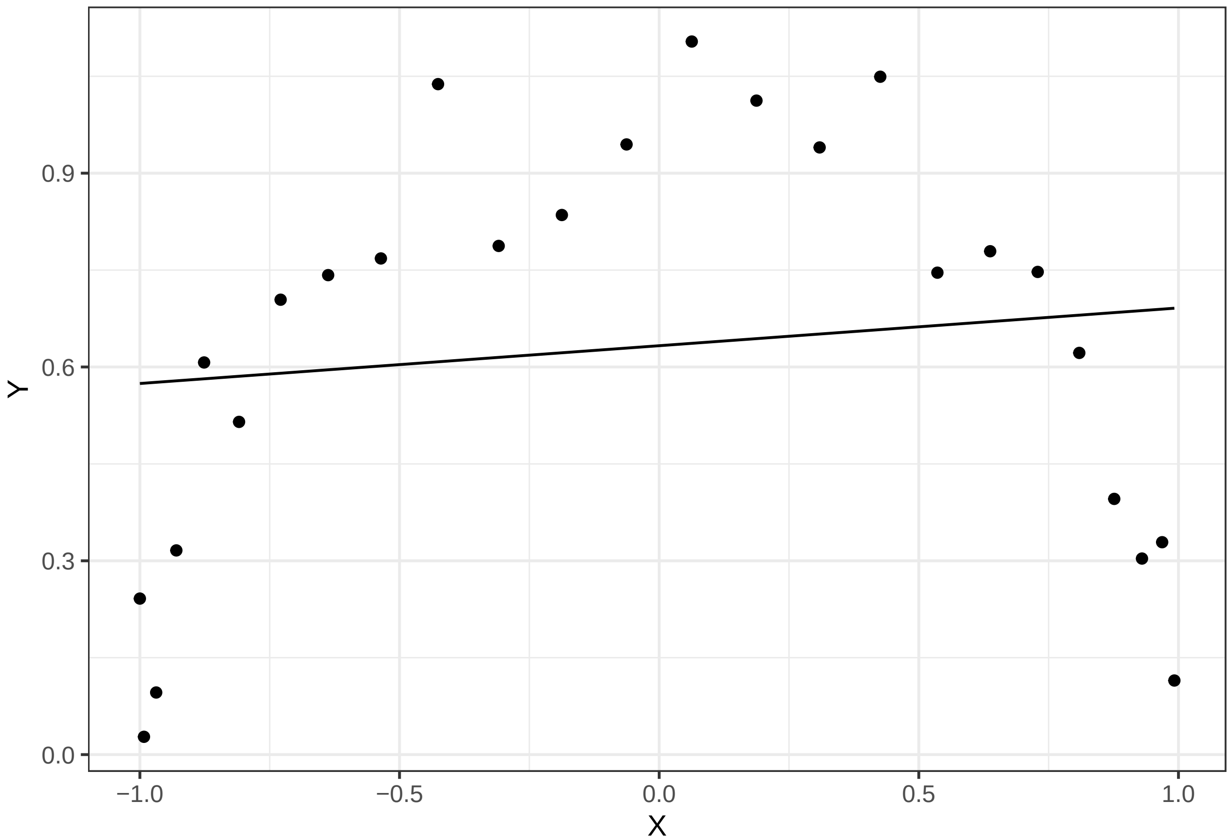
a

The line should be flat, similar to the residual plot since the model overestimates, underestimates, and then overestimates.

Let's fit the model and see if I am right.

```
U = seq(0.04, 1.00, 0.04)
X = cos(pi*U)
Y = sin(pi*U) + rnorm(25, 0, sqrt(.01))
mod <- lm(Y ~ X)

library(ggplot2)
ggplot(data = NULL, aes(x = X)) +
  geom_point(aes(y = Y)) +
  geom_line(aes(y = predict(mod))) +
  theme_bw()
```



The line is mostly flat, but most of the time has a slight positive slope.

b

Correlation is a measure of linear relationship between two variables. Since the relationship is quadratic on this interval, the correlation will be close to 0.

The actual value is

```
cor(X,Y)
```

```
## [1] 0.1308139
```

c

The parabola will point downwards so in the corresponding quadratic $ax^2 + bx + c$, a will be negative. Therefore β_2 will be negative.

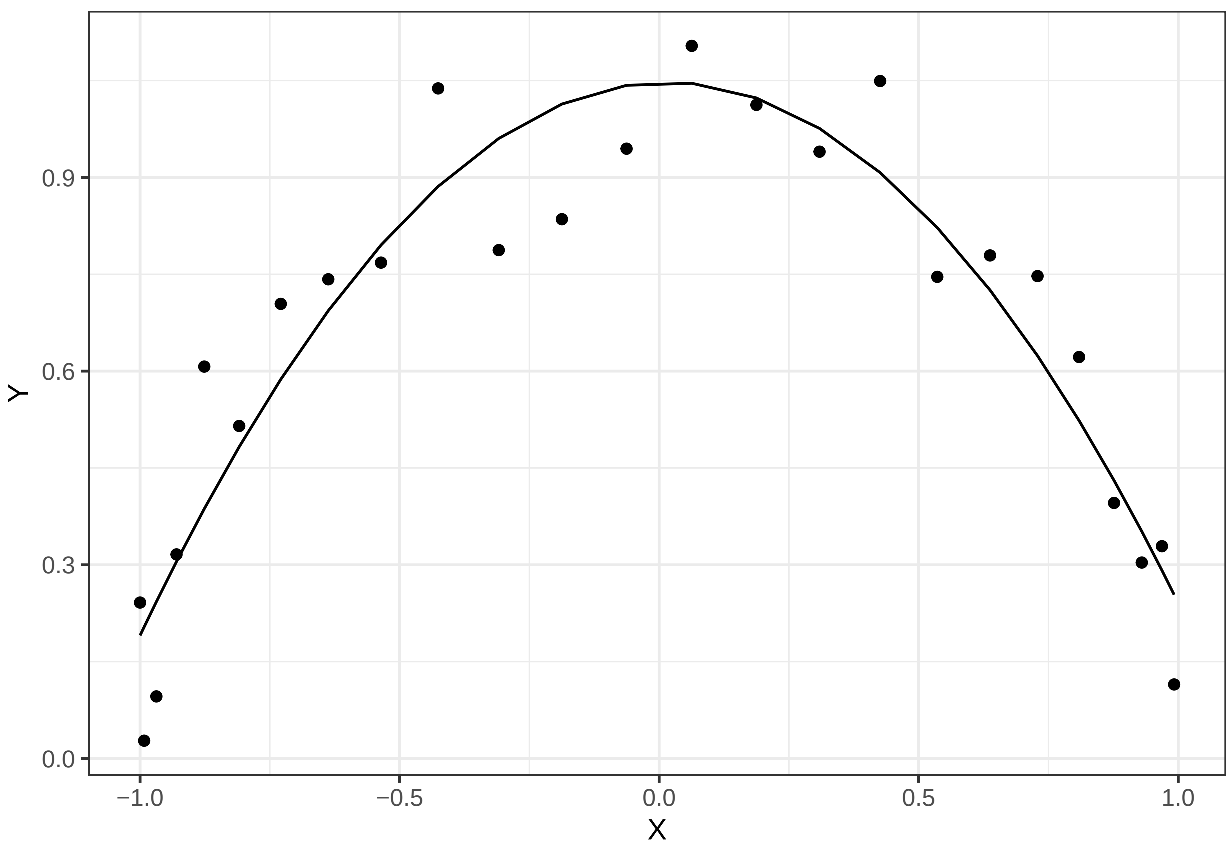
Let see if that is true.

```
mod2 <- lm(Y ~ X + I(X^2))
summary(mod2)
```

```
##
## Call:
```

```
## lm(formula = Y ~ X + I(X^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17837 -0.07613  0.01103  0.05793  0.22029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.04753    0.04039  25.934  < 2e-16 ***
## X            0.02510    0.03312   0.758    0.457
## I(X^2)      -0.83188    0.06614 -12.578  1.6e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1165 on 22 degrees of freedom
## Multiple R-squared:   0.88,  Adjusted R-squared:  0.8691
## F-statistic: 80.67 on 2 and 22 DF,  p-value: 7.425e-11
```

```r
ggplot(data = NULL, aes(x = X)) +
  geom_point(aes(y = Y)) +
  geom_line(aes(y = predict(mod2))) +
  theme_bw()
```



d Standard deviation is (roughly) the average of the distances from the mean. It appears that the assumptions of linear regression are being met. Therefore the distribution of the residuals should be $N(0, \sigma^2)$. Since the residuals are mostly within ± 0.2 of 0, the standard deviation should be close to 0.1.

Let's test that

```r
sd(residuals(mod2))
```

```
## [1] 0.1115786
```

e

It depends on the range of your data. A quadratic will work fine when only looking at half the period of the trig function. Trying to model more than half a period with a quadratic will be a bad idea.

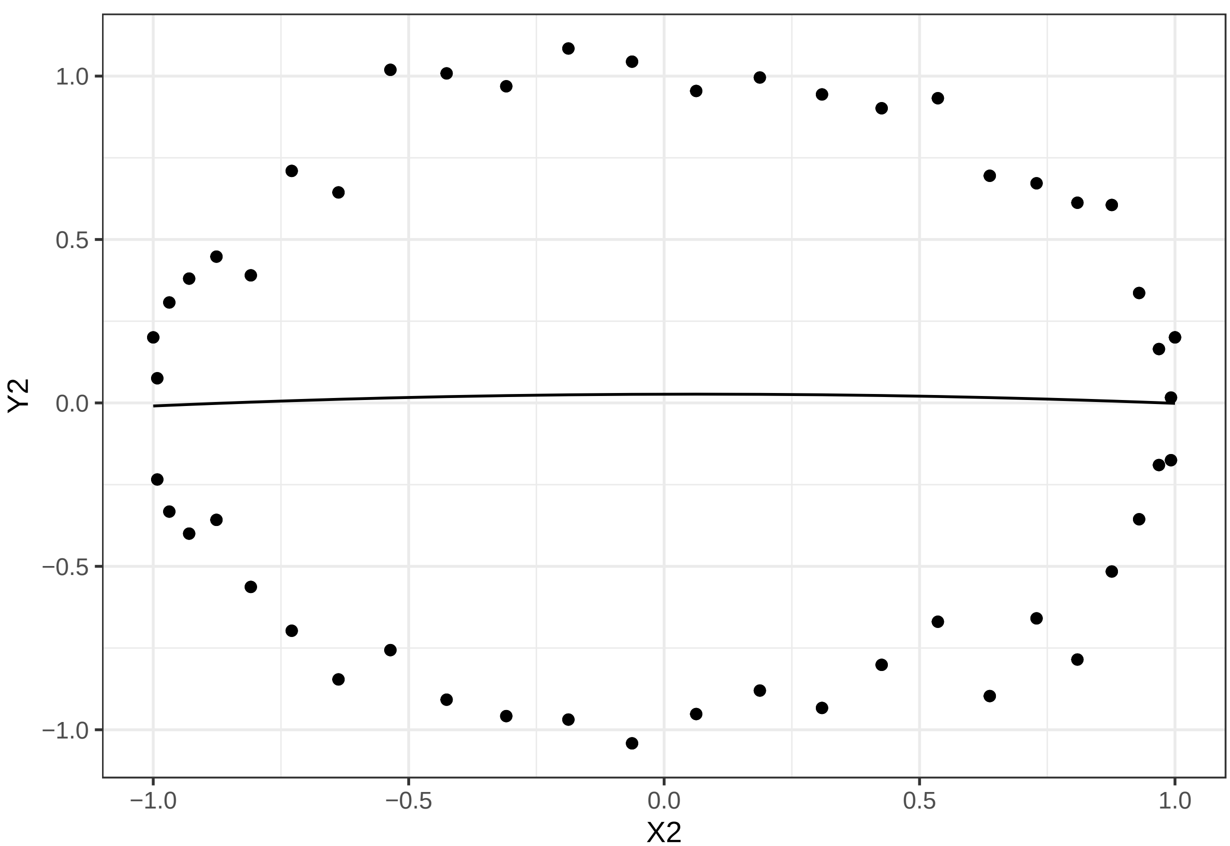
Let's see what happens

```r
U2 = seq(-2, 2, 0.04)
X2 = cos(pi*U2)
Y2 = sin(pi*U2) + rnorm(25, 0, sqrt(.01))
```

```
## Warning in sin(pi * U2) + rnorm(25, 0, sqrt(0.01)): longer object length is not
## a multiple of shorter object length
```

```r
mod3 <- lm(Y2 ~ X2 + I(X2^2))

library(ggplot2)
ggplot(data = NULL, aes(x = X2)) +
  geom_point(aes(y = Y2)) +
  geom_line(aes(y = predict(mod3))) +
  theme_bw()
```



2

a

$$R^2 = \frac{SSM}{SST} = \frac{2016.12237}{8704.81686} = 0.23161$$

About 23% of the variation in MFSI can be explained with a linear relationship with the predictors.

b

`age10` is the age in years divided by 10. Therefore, every 1 unit increase in `age10`, an increase of 10 years, is associated with a .68 decrease in MFSE holding all the other predictors constant. Divide the coefficient by 10 to get the yearly effect.

c

`log2daysout` is the log base 2 of number of days since date of diagnosis. For every doubling in the number of days since diagnosis, the MFSI increases by 1.040066, holding the other predictors constant.

d

The coefficient of marital status is 1.890183 while the coefficient of history of MDD is -.1695029. The interaction effect between the two is 3.48601. This suggests that being married increases the MFSI while MDD has a very small, statistically insignificant, decrease in MFSI. The interaction effect suggests that the effect of a history of MDD greatly increases when also married.

e

Interaction is not the same thing as correlation. Just because there is an interaction effect between 2 variables does not mean that they are related.

f

Not sure that these are techinically nested models, but let's assume that they are.

$$F^* = \frac{\frac{2026.35534-2016.12237}{11-10}}{\frac{6678.46151}{249}} = 0.3815264$$

Under H_0, $F^* \sim F(1, 249) \equiv t(249)$. Since the degrees of freedom are so large, we can make the approximation $t_{.975}(249) \approx Z_{.975} = 1.96$. Therefore, we fail to reject H_0 and conclude that the dummy variable model provides little additional explaination of the variation in MFSI compared to the original model.

3

4

5